A multi-region discrete time chain binomial model for infectious disease transmission

Pallab K. Sinha Siuli Mukhopadhyay*

Department of Mathematics, Indian Institute of Technology Bombay, Maharashtra, India.

Abstract

Conventional mathematical models of infectious diseases often overlook the spatial spread of the disease, focusing only on local transmission. However, spatial propagation of various diseases has been observed mainly due to the movement of infectious individuals from one geographical region to another. In this work, we propose a multiregion discrete-time chain binomial framework to model the dependencies between the multiple infection time series from neighboring regions. It is assumed that infection counts in each region at various time points are governed not only by local transmissions but also by interactions of individuals between spatial units. The effects of intervention strategies such as vaccination campaigns used in disease control and sociodemographic factors such as live births and population density were taken into account while modeling multiple infection time series. For estimating these multiregion chain-binomial models, an appropriate likelihood function maximization approach is proposed. A predictive likelihood based method is used to generate short- and long term forecasts of disease counts in connected spatial units. In-depth analyses of two real world datasets and simulation studies—incorporating seasonal patterns, asynchronous outbreaks across connected regions, and varying intervention strategies—are used to motivate the proposed modeling framework and demonstrate its ability to reproduce realistic outbreak dynamics.

Keywords: Distance based methods, Epidemiological models, Highest Density Regions, Predictive likelihood, Spatio-temporal model

1 Introduction

Spatio-temporal models of infectious diseases are essential to improve our understanding of epidemic dynamics as they are able to capture the simultaneous evolution of the disease over time and space. The spatial propagation of various infectious diseases has been documented multiple times throughout history in the epidemiological literature, for example, the 14th century plague in Europe, smallpox in the 1500s, measles in the United Kingdom in 1900s (Grenfell et al. (2001)), and most recently the Covid-19 pandemic (Kianfar et al. (2022)). This spatial propagation of infectious diseases can be attributed to the intermixture of infectious individuals from one region with healthy susceptibles from another region. In this work, our aim is to propose a novel spatio-temporal epidemiological model that can accurately predict and forecast counts of an infectious disease, in a multi-region setup. We integrate traditional epidemiological model, the chain binomial framework, with statistical techniques to develop a new approach that offers a comprehensive view of the disease dynamics over space and time.

Epidemic chain models of infectious disease transmission, such as the discrete-time chain binomial model, have been widely used in epidemiological studies to understand transmission of various diseases such as measles in a school setting (En'Ko (1989)), household chickenpox transmission (Abbey (1952)), and ebola, swine fever and foot and mouth diseases (Ferrari et al., 2005). In these models (Bailey (1957), Becker (2017)), the binomial formulation arises from the assumption that each potential exposure during an epidemic is an independent Bernoulli trial, resulting in either infection or continued susceptibility. The Reed–Frost model, a well-known example of a chain binomial model, assumes that the number of new infections in time t depends on the number of infectious and susceptible individuals at the previous time point $t-1$. The chain structure of these models makes them suitable for parameter estimation using maximum likelihood methods. Recently, chain binomial models have been utilized in the epidemic

*Corresponding author
Email address: siuli@math.iitb.ac.in (Siuli Mukhopadhyay)

modeling of COVID-19, where (Lefèvre et al., 2021) formulated a generalized Reed–Frost model that included two classes of infections (symptomatic vs asymptomatic) for testing and vaccination, However, prior to this work, epidemic chain models have not been utilized to examine the spatial dynamics of infectious disease transmission.

To model the spatial transmission of infections caused by movement of infectious individuals from one region coming into effective contact with the susceptibles of another region, we propose a new multi-region chain binomial framework. This model is based on the epidemiological principle that new infections are generated from the corresponding susceptible populations and previous infection counts. To enhance the statistical rigor of the proposed approach, we assume that the probability of new infections for each spatial unit are modeled through an appropriately chosen link function to depend on the lagged local and inter-regional incidence and region-specific time-dependent covariate effects. Instead of a simple auto regressive order 1 (AR(1)) structure as in conventional chain models, we allow for AR(p) dependence. To address the rapid growth in model complexity as the number of spatial units increases, we incorporate spatial contiguity matrices and k-nearest-neighbor structures to impose sparsity on inter-regional transmission effects.

In-depth simulations are conducted to study the effect of spatial distances, inter-regional infection counts, disease transmission rates, and vaccine coverages on the performance of the proposed method under various scenarios depicting real life disease epidemics. We also illustrate the performance of the proposed model using two real world data sets. The first dataset consists of city-level measles incidence data from the United Kingdom spanning the period 1944–1966. This dataset represents the pre-vaccination era and has been extensively analyzed in the epidemiological and statistical literature. Seasonal variation, the baby boom effect, and city-specific live birth rates are known to be important covariates influencing measles dynamics in this period. Accordingly, we assume 2% to 20% of the city-specific populations to represent the corresponding susceptible populations. To demonstrate the applicability of the proposed method in the post-vaccination era, we further analyze a more recent district-level measles dataset from an Indian state covering the period 2014–2020. Due to the effect of vaccination campaigns, the second data set allows us to adjust the susceptible population of each district by accounting for vaccinated cohorts over time. For each real data set, comparison and subsequent selection of models are carried out by using various criterion functions such as mean squared errors based on various error residuals, chi-squared statistic, deviance, and Akaike Information criterion (AIC). We also use the predictive likelihood method (Mukhopadhyay and Sathish (2019)) to simultaneously forecast disease counts in all spatial units involved for both data sets. Both short-term and long-term forecasting performance are examined, and the corresponding summary metrics are shown to be related to population density and public transport efficiency within the spatial units considered. Although both of our real-world datasets are based on measles epidemics, the proposed approach is applicable to a broad class of directly transmitted diseases in which infection results in either long-term immunity or death, thereby requiring consideration of only susceptible and infectious compartments. Furthermore, diseases that progress rapidly without long latent periods, and therefore do not need explicit modeling of host birth and death dynamics in the equations (Ferrari et al., 2005), can also be effectively modeled.

Various spatio-temporal models have been proposed in the literature to study the dynamics of disease epidemics. Building on the SIR model of (Kermack and McKendrick, 1927), Hethcote (2000), Sattenspiel and Dietz (1995), Zakary et al. (2017) proposed a multiregion version of the discrete SIR model that accounts for the spread of infectious diseases across multiple geographical regions, incorporating the movement of the population between them. A time-series adaptation of the SIR model, known as the TSIR model (Grenfell et al., 2001), has been used extensively to model the spatial spread of diseases. Bjørnstad et al. (2002) examined measles transmission in connected regions, modeling the force of infection as a function of resident infectious hosts, immigrant infections, and the transmission rate, while, Xia et al. (2004) integrated the TSIR model with a gravity-based spatial interaction framework to describe measles transmission between different communities. Another series of models, namely the endemic-epidemic models (Held et al. (2005)) or the HHH models have been used to study infectious diseases over space and time. These models separate infection dynamics into two components: an endemic component handling the influx of new infections from external sources, and an epidemic component covering the contagious nature of the disease by assuming the expected number of transmissions to be a function of earlier infections in time. Meyer and Held (2017) studied this framework in a spatio-temporal setting, modeling case counts at specific locations and times as Poisson or negative binomial variables, with means determined by autoregressive effects, spatial transmission, and time- or unit-specific factors. Similarly, Bauer and Wakefield (2018) modeled counts using a negative binomial distribution and incorporated random endemic, neighborhood, and spatial CAR effects. More recently, Bracher and Held (2022) explored

alternative weighting schemes—such as shifted Poisson, linear decay, and geometric weights—while Lu and Meyer (2023) introduced a zero-inflated version of the HHH model to address excess zeros. It should be noted that the spatial versions of the TSIR models are not flexible enough to model AR(p) dependence on local or inter region counts or include the effect of region specific covariates which may influence the dynamics of the disease process. HHH models on the other hand, consider covariate effects but do not incorporate the epidemiological information of the disease, i.e., the generation of new cases through effective contact between infectious and susceptible individuals. In contrast, our proposed multi-region chain binomial model proposes an efficient way to combine epidemiological aspects of a disease with statistical modeling. It allows for the incorporation of various covariate factors that influence disease outbreaks in a statistically rigorous manner. In addition, the chain binomial model provides a more granular understanding of individual infection events, enabling precise estimation of transmission probabilities and the identification of key drivers of outbreaks. This level of detail is especially important for designing targeted interventions and understanding the early spread of infectious diseases.

The remainder of the manuscript is structured as follows. Section 2 introduces the chain binomial model. In Section 3 we provide details on the proposed multivariate chain binomial model, along with parameter estimation, model selection statistics and a forecasting algorithm. In Sections 4 and 5, we give simulation results and analyse two real-world datasets, respectively, to illustrate the application of the proposed approach to model disease epidemics over space and time. Finally, Section 6 provides closing remarks.

2 The chain binomial model - single region

We start by introducing the conventional chain binomial model and show steps to add statistical rigour. Suppose, a time series of disease counts have been obtained from a single region. The number of infections at time $(t+1)$ is modelled as:

$$I_{t+1} \sim Bin(S_t, \pi_{t+1})$$

where S_t is the number of people at risk or susceptible to the disease at time t. The probability of infection π_{t+1} is generally chosen as a suitable nondecreasing function of the number of infections at time t, that is, I_t, between 0 and 1. The susceptibles at time $t+1$ is

$$S_{t+1} = (1-\lambda_t)S_t = S_t - I_{t+1},$$

where λ_t is the force of infection and can be written as $\lambda_t = \mu I_t$, μ is the effective rate of contact among infectious and susceptible individuals of that region. Thus, we obtain a generative equation $S_t = S_0 - \sum_{i=1}^{t} I_i$, where S_0 is the initial size of the susceptible population.

We deviate from a Markovian type of dependence as in (Ferrari et al., 2005), and assume that the probability of infections at time t may depend on infections until time lag p as in an AR(p) process. Also, we use a link function approach to incorporate the effect of long term disease trend, seasonal fluctuations, and time dependent covariates to model the probability of new infection.

Consider that the conditional distribution of $I_{t+1}|H_t$ is assumed to follow a $Bin(S_0 - \sum_{i=1}^{t} I_i, \pi_{t+1})$ where $H_t = \{\mathbf{x}_t, S_0, I_t, \ldots, I_1\}$. A logit link function is used as follows.

$$\text{logit}(\pi_t) = \eta_t = \mathbf{x}_t^T \boldsymbol{\beta} + \sum_{d=1}^{p} \phi_d g_d(I_{t-d}) = \mathbf{z}_t^T \boldsymbol{\nu},$$

with $\boldsymbol{\nu} = (\boldsymbol{\beta}, \phi_1, \ldots, \phi_p)^T, \mathbf{z}_t = (\mathbf{x}_t, I_{t-1}, \ldots, I_{t-d})$, and $g_d(\cdot)$ is any known function for each d (Fokianos and Kedem (2004)).

In the above equation, $\mathbf{x}_t^T$ is assumed to be of the form $A_t + B_t + C_t$, where A_t represents a possible long-term temporal trend component, B_t the seasonal components, and C_t may involve time dependent covariates such as vaccination coverages over time, live births, etc. The autoregressive part of the linear predictor is represented by the p terms, $\sum_{d=1}^{p} \phi_d g_d(I_{t-d})$. Similar forms of η_t have been used in the endemic-epidemic models as described in Held et al. (2005). If required, lag terms corresponding to time dependent covariates may also be introduced. As an example, see lag terms of weather variables in the vector-borne disease model in Mukhopadhyay et al. (2019).

2.1 A new multi-region chain binomial model

To extend the conventional chain binomial model to a multi region model we make use of the multiple SIR type of framework as discussed in Zakary et al. (2017). Suppose that there are m spatial regions and we observe disease counts I_t^j, from the jth location, where $j = 1, \ldots, m$. The susceptible population in the region j can then be written as $S_{t+1}^j = (1 - \lambda_t^j)S_t^j = S_t^j - I_{t+1}^j$, where $\lambda_t^j = \sum_{j'=1}^m \mu_{jj'} I_t^{j'}$ is the region-specific force of infection and $\mu_{jj'}$ is the effective rate of contact between susceptible individuals in the region j with infectious individuals j'. Thus, the multi-region chain binomial model is as follows,

$$I_{t+1}^j \sim Bin(S_t^j, \pi_{t+1}^j),\ j = 1, \ldots, m,$$

where the generative equation for the susceptible population j^{th} is updated to $S_t^j = S_0^j - \sum_{i=1}^t I_i^j$, and π_{t+1}^j now represents the probability of infection at time $(t+1)$ in the region j.

As the m incidence counts observed are from neighboring spatial regions, they may be correlated with each other. Thus, we need to accommodate these dependencies over both time and space while modelling the probability of new infections in region j. We assume that the probability of new infection in region j, π_t^j, has AR(p) dependence not only on the disease counts of the same region j, i.e., $\{I_{t-1}^j, \ldots, I_{t-p}^j\}$, but also on the counts of the neighboring regions of j, that is, $\{I_{t-1}^{j'}, \ldots, I_{t-p}^{j'}\}$, where $j' \in N_j$, N_j is a spatial neighborhood of j. Thus, the conditional distribution of $I_{t+1}^j | H_t^j$ is as follows:

$$I_{t+1}^j | H_t^j \sim Bin(S_t^j, \pi_{t+1}^j),\ j = 1, \ldots, m, \tag{1}$$

where $H_t^j = \{\mathbf{x}_t^j, S_0^j, I_t^j, \ldots, I_1^j, I_t^{j'}, \ldots, I_{t-p}^{j'}\}$, for all $j' \in N_j$.

Using a logit link function we model π_t^j as

$$\text{logit}(\pi_t^j) = \eta_t^j = (\mathbf{x}_t^j)^T \boldsymbol{\beta}^j + \sum_{d=1}^p \phi_d^{jj} g_d(I_{t-d}^j) + \sum_{d=1}^p \sum_{j' \in N_j} \phi_d^{jj'} g_d(I_{t-d}^{j'}), \tag{2}$$

Here, we propose two choices to define N_j.

1. Choice 1: $N_j = \{j' : j' \neq j\}$, that is, assume that all surrounding regions may affect the counts of region j irrespective of sharing boundaries or being close to each other.

2. Choice 2: $N_j = \{j' : j'$ chosen using k-nearest neighbour rule$\}$, for pre-specified values of k.

For Choice 1, note that even for 3 spatial regions and a AR(2) dependence, for each j we have 6 different $\phi_d^{jj'}$s, i.e. total 18 parameters. To reduce the number of inter region parameters, $\phi_d^{jj'}$, we assume that $\phi_d^{jj'} = \phi_d^j w_{jj'}$, where $w_{jj'}$ are distance-based weights. Thus, we introduce a jth region specific effect ϕ_d^j and weight it by its distance from its neighbour j'. We may model $w_{jj'}$ as a decreasing function of the distance between the centroids of the spatial regions j and j', one popular choice is $w_{jj'} = (\frac{1}{1+d_{jj'}})^\rho$, and ρ may be estimated from the data. For estimating the decay parameter, ρ, it is possible to use shifted poisson weights, linear decay or geometric weighting scheme as mentioned in Bracher and Held (2022). However, for computational simplicity, we have fixed the value of ρ to 2 in this manuscript. Note, $d_{jj} = 0$. For our computations, we also tried another set of weights $w_{jj'} = \exp(-d_{jj'})$. See Section 4 for details. Thus, using the distance-based weights the logit-AR(p) model is as follows for region j

$$\text{logit}(\pi_t^j) = \eta_t^j = (\mathbf{x}_t^j)^T \boldsymbol{\beta}^j + \sum_{d=1}^p \phi_d^j \sum_{j'=1}^m w_{jj'} g_d(I_{t-d}^{j'}) = (\mathbf{Z}_t^j)^T \boldsymbol{\nu}^j, \tag{3}$$

2.2 Parameter Estimation

We use the maximum likelihood method for estimating the parameters, $\boldsymbol{\nu} = (\boldsymbol{\nu}^1, \ldots, \boldsymbol{\nu}^m)$,

$$\begin{aligned} l(\boldsymbol{\nu}) &= \sum_{t=1}^n \sum_{j=1}^m \log f_{\boldsymbol{\nu}}(I_t^j | H_{t-1}^j) \\ &= \sum_{t=1}^n \sum_{j=1}^m log \left(\begin{array}{c} S_0^j - \sum_{i=1}^{t-1} I_i^j \\ I_t^j \end{array} \right) + I_t^j \log \frac{\pi_t^j}{1 - \pi_t^j} + (S_0^j - \sum_{i=1}^{t-1} I_i^j) \log(1 - \pi_t^j). \end{aligned}$$

Using the logit link function, $\pi_t^j = \frac{e^{\eta_t^j}}{1+e^{\eta_t^j}}$, where η_t^j can be written as $(\mathbf{Z}_t^j)^T \boldsymbol{\nu}^j$. The Fisher scoring algorithm is used to find $\boldsymbol{\nu}$ iteratively as,

$$\hat{\boldsymbol{\nu}}^{(\omega)} = \hat{\boldsymbol{\nu}}^{(\omega-1)} + G_n^{-1}(\hat{\boldsymbol{\nu}}^{(\omega-1)}) S_n(\hat{\boldsymbol{\nu}}^{(\omega-1)}) \tag{4}$$

where ω denotes the ωth iteration step in the estimation process.

The score function is,

$$S_n(\boldsymbol{\nu}) \equiv \nabla(\boldsymbol{\nu}) = \sum_{t=1}^{n} \sum_{j=1}^{m} \mathbf{Z}_{t-1}^j \frac{\exp{(\mathbf{Z}_{t-1}^j)^T \boldsymbol{\nu}^j}}{[1+\exp{(\mathbf{Z}_{t-1}^j)^T \boldsymbol{\nu}^j}]^2} \frac{I_t^j - \mu_t^j}{(\sigma_t^j)^2}$$

and the conditional covariance matrix $G_n(\boldsymbol{\nu})$ is given by:

$$G_n(\boldsymbol{\nu}) = \sum_{t=1}^{n} \mathbf{Z}_{t-1}^T W(\boldsymbol{\nu}) \mathbf{Z}_{t-1},$$

where $W(\boldsymbol{\nu})$ is a diagonal matrix with elements $\left(\frac{\exp{(\mathbf{Z}_{t-1}^j)^T \boldsymbol{\nu}^j}}{[1+\exp{(\mathbf{Z}_{t-1}^j)^T \boldsymbol{\nu}^j}]^2} \right)^2 \frac{1}{(\sigma_t^j)^2}$, $\mu_t^j = S_{t-1}^j \frac{\exp{(\mathbf{Z}_{t-1}^j)^T \boldsymbol{\nu}^j}}{[1+\exp{(\mathbf{Z}_{t-1}^j)^T \boldsymbol{\nu}^j}]}$, and $(\sigma_t^j)^2 = S_{t-1}^j \pi_t^j (1-\pi_t^j)$. Fokianos and Kedem (1998) showed that $\sqrt{n}(\hat{\boldsymbol{\nu}} - \boldsymbol{\nu}) \to \mathcal{N}_c(0, G^{-1})$ where G is the limiting term of G_n and c is the dimension of $\boldsymbol{\nu}$.

2.3 Susceptible population

The above MLE procedure depends on the values of the initial susceptible populations, S_0^j. Determining a suitable initial susceptible population S_0, that best describes the dynamics of the disease, is an important step to correctly analyze disease epidemics. In the epidemiological literature, it is generally assumed that S_0 is the proportion not immune of the total population, i.e., $S_0 \approx (1-p)N$ where p denotes the immunity coverage and N denotes the population (Anderson and May (1991), Keeling and Rohani (2002)). Bjørnstad (2022) used from 2% to 20% of the total population as a candidate of the initial susceptible population. However, other methods such as estimating S_0 using MLE as in Ferrari et al. (2005) have also been used. Very recently, Madden et al. (2024) used the TSIR model to get estimates of susceptibles in their neural network model setup. In this work, we follow the practices of the standard epidemic modeling literature and take $S_0^j, j = 1, \ldots, m$ to be varying fractions of the region specific populations based on immunity or vaccination coverage. The results represented in later sections are based on 20% of the population densities, as is usually done for measles (Bjørnstad (2022), Ferrari et al. (2008)).

In vaccine preventable diseases (VPDs), particularly measles, the generative equation for S_t^j is further updated as

$$S_t^j = S_0^j - \sum_{i=1}^{t} I_i^j + \sum_{i=1}^{t} X_i^j,$$

where X_t^j represents the non-immune birth cohort at time t and region j due to vaccination. For a vaccine preventable disease, if the number of scheduled doses, the efficacy of each dose, and coverages are known, the computation of X_t^j can be based on this information. Suppose, we have a VPD with one dose of vaccine. Then the non-immune cohort is Chen et al. (2012):

$$X_t^j = \begin{cases} B_t^j & \text{vaccination is not available in region } j \\ B_t^j[1 - xV_t^j] & \text{vaccination is available in region } j, \end{cases} \tag{5}$$

where B_t^j is the birth cohort, V_t^j is the coverage of vaccination, and the efficacy of the vaccine is x.

2.4 Model Evaluation

Various forms of the above multiregion chain binomial model were fitted to simulated and real data sets of disease counts. For evaluating these models, we used popular goodness-of-fit statistics from the model selection literature like model deviance, chi-squared criterion, AIC, Bayesian Schwarz information criterion (BIC), various residuals (such as working, Pearson etc.) and mean square errors based on these residuals (Kedem and Fokianos (2005)).

2.5 Forecasting using a Predictive Likelihood

Once the model is selected, it is of interest to forecast future disease incidences and obtain intervals for these forecasts in the m regions. Based on past information up to the time point n, H_n^j, for all j regions, we estimate the density of the predicted value I_{n+1}^j (one step ahead) using the predictive likelihood method (Mukhopadhyay and Sathish (2019)) as

$$\hat{p}(I_{n+1}^j) = k(H_n^j) LL_p(I_{n+1}^j | H_n^j),$$

where $k(\cdot)$ is a normalizing constant and $LL_p(I_{n+1}^j|H_n^j) = \prod_{t=1}^{n+1} f_{\hat{\boldsymbol{\nu}}^j}(I_t^j|H_{t-1}^j)$. The forecast value of I_{n+1}^j denoted by $\tilde{I}_{n+1}^j$, is $\arg\max_{I_{n+1}^j \in M} \hat{p}(I_{n+1}^j)$, where $M = [\min_t I_t^j, \max_t I_t^j]$. Note, H_n^j involves the counts of the jth region as well as $j' \in N_j$. To generate, a set of q one-step ahead forecasts, the following steps are used:

Step 1: Generate the first forecast for all j regions; For $\omega = 1$ and $j = 1, \ldots, m$ using the predictive likelihood,

$$\hat{p}(I_{n+\omega}^j) = k(H_{n+\omega-1}^j) LL_p(I_{n+\omega}^j | H_{n+\omega-1}^j),$$

Step 2: To generate $\tilde{I}_{n+\omega}^j$, for $\omega = 2, \ldots, q$ and $j = 1, \ldots, m$ we use the predictive likelihood

$$\hat{p}(I_{n+\omega}^j) = k(\hat{H}_{n+\omega-1}^j) LL_p(I_{n+\omega}^j | \hat{H}_{n+\omega-1}^j),$$

where, $\hat{H}_{n+\omega-1}^j = \{\mathbf{x}_{n+\omega-1}^j, S_0^j, \tilde{I}_{n+\omega-1}^j, \ldots, \tilde{I}_{n+1}^j, I_n^j, \ldots, I_1^j, \tilde{I}_{n+\omega-1}^{j'}, \ldots, \tilde{I}_{n+1}^{j'}, I_n^{j'}, \ldots, I_1^{j'}\}$ for all $j' \in N_j$.

To estimate the prediction interval for a set of one-step forward forecasts, we use the region of highest density of Hyndman (1995). The $(1-\alpha)100\%$ HDR for $I_{n+\omega}^j$, $\omega = 1, \ldots, q$ is the region R_α^j, where $R_\alpha^j = \{I_{n+\omega}^j : \hat{p}(I_{n+\omega}^j) \geq f_\alpha\}$, where f_α is the largest constant such that $Pr(I_{n+\omega}^j \in R_\alpha^j)$ is at least $(1-\alpha)$.

3 Simulation Studies

We perform simulations in two different scenarios to show the applicability of our proposed method to track disease epidemics over space and time when interventions are available for disease control. We assume that the disease being modeled is the biweekly counts of measles. Measles is a highly contagious, vaccine-preventable viral disease. The World Health Organization recommends administering two doses of measles vaccines at 9-12 months and 16-24 months to control and eliminate the disease. Measles have been modeled previously by various researchers, namely, Grenfell et al. (2001), Chen et al. (2012), Thakkar et al. (2019).

For generating biweekly counts we use the following susceptible-infection equations of the multiregion SIR model (Zakary et al. (2017)) with various choices of parameters. It is assumed that there are 3 spatially connected regions and 160 biweekly incidences are generated using the following equations,

$$\begin{aligned} S_{t+1}^j &= S_t^j - \sum_{k=1}^{n} \mu_{jk} S_t^j I_t^k \\ I_{t+1}^j &= I_t^j + \sum_{k=1}^{n} \mu_{jk} S_t^j I_t^k - \gamma I_t^j, \quad j = 1,2,3 \quad \text{and} \quad t = 0, \ldots, 160, \end{aligned}$$

where the recovery rate, γ, is assumed to be $1/14$ day^{-1} and μ_{jk}'s are the disease transmission rates between two locations j and k. The transmission rates are taken to be influenced by region-specific factors such as the number of susceptibles and annual seasonality,

$$\mu_{jj}(t) = \frac{\bar{\mu}_{jj}}{S_j(0)} \left(1 + 0.5 \sin \frac{2\pi t}{26}\right),$$

here $\bar{\mu}_{jj}$ are the region-specific mean transmission rates. We model the μ_{jk}'s to depend on the distance between two spatial regions j and k as: $\mu_{jk}(t) = \mu_{jj}(t) w_{jk}$, where w_{jk}'s are spatial weights.

A spatial weight matrix, $W = ((w_{jk}))$,

$$W = \begin{bmatrix} 1 & e^{-5} & e^{-5} \\ e^{-5} & 1 & e^{-.005} \\ e^{-5} & e^{-.005} & 1 \end{bmatrix}$$

is used. We work with the following parameters choices:

- Case 1: $\bar{\mu}_{jj} = 0.5$, $I_0^j = 6$; $j = 1, 2, 3$ and $S_0^j = 10^6, \forall j$.
- Case 2: $\bar{\mu}_{jj} = 0.5$, $I_0^j = 6$; $j = 1, 2$; $\bar{\mu}_{33} = 0.55$, $I_0^3 = 34$ $S_0^j = 10^6, \forall j$.

As interventions such as vaccination strategies play an important role in measles control, we also introduce a vaccinated birth cohort as discussed in Section 2.3. The susceptible population in the equations of the MSIR model is adjusted as: $S_{t+1}^j = S_t^j - \sum_{k=1}^n \mu_{jk} S_t^j I_t^k + X_t^j$ where

$$X_t^j = \begin{cases} B_t^j & \text{vaccination is unavailable in region } j \\ B_t^j[1 - 0.85V_{1t}^j(1 - V_{2t}^j) - 0.99V_{1t}^j V_{2_t}^j] & \text{vaccination is available in region } j, \end{cases}$$

The birth cohort was chosen to remain constant over time $B_j^t = 10^3, \forall j$. Thus, for a region j where vaccination is available, X_t^j denotes the birth cohort who remains unvaccinated, thus adding to the susceptible population of the region.

For each case, we consider two scenarios: In scenario 1 (without vaccination): counts are generated without the effect of vaccination in any of the three spatial regions. In scenario 2 (with vaccination): we assume that vaccination is available in Region 3 only. The coverages of the two doses are, V_{1t}^j increasing linearly from 0.70 to 0.90 and V_{2t}^j increasing linearly from 0.40 to 0.70 over the biweeks (t).

After fitting the proposed model to the data generated from the MSIR model, the results are plotted in Figure 1. In both cases we see that counts of regions 2 and 3 behave similarly as these are the spatially coupled regions. Also, for case 2, as the initial case counts of region 3 is higher, it increases the counts of the spatially coupled regions 2 and 3 to be higher than in case 1, while region 1 remains unaffected in both cases. We also note that for both parameter choices (cases 1 and 2), the effect of vaccination reduces the measles counts in the spatially related regions 2 and 3, while region 1 remains unaffected. This reduction may be attributed to the spatial proximity of regions 2 and 3, as vaccination available in region 3 may lead to a lower influx of infected individuals to migrate from Region 3 to Region 2, thus reducing disease transmission.

For both without and with vaccination setups, $10,000$ simulations were performed and the proposed model in Section 2 was fitted. The mean squared error (MSE) and mean squared absolute deviation (MAE) values are listed in Table 1. Note that the error statistics values are similar for regions 2 and 3. This is due to the fact that in these scenarios we assume that Regions 2 and 3 are spatially coupled and have comparable incidence patterns. The distribution of the estimated parameters (standardized parameters) and the residual analysis were also studied; however, we do not report them here.

4 Real Data Example: Analysis of Measles Counts in UK

To illustrate our proposed model on real data, we fitted it to biweekly case reports of measles recorded between the years 1944-1966 from seven urban cities, Birmingham, Coventry, Leicester, Derby, Nottingham, Wolverhampton and Walsall, in the United Kingdom (UK). The UK measles data set has previously been analyzed by various authors, namely (Grenfell et al. (2001), Xia et al. (2004), Madden et al. (2024)). We chose the seven cities due to their spatial proximities (distance) on the UK map. From an initial spatial correlation analysis, a Moran's-I value (Anselin (1995)) of 3.0277 and a p-value of 0.0012 (significant at the 5% level) were observed, implying existence of spatial correlation among the disease counts (averaged over time) of the seven cities. During the period 1944 to 1966, these seven English Midlands cities (see Fig. 2) were connected through a complex and evolving public transport network Eye (2016) that facilitated the frequent movement of individuals among the cities. The railway system served as the primary inter-city connector, with Birmingham operating as the central hub thus, justifying the highest case counts of Birmingham. For the location of the cities and the strength of the public transport connection between them, see Figure 2, while for distances between cities, see Table 2. From Figure 2, as Birmingham is the most strongly connected to all cities, outbreaks in all cities seem to be synchronized or follow. These seven cities also differ in their population status, with Birmingham having the highest population. We divide the cities that exclude Birmingham into two groups (high/low) according to their population densities and plot the measles counts to compare the incidences and pattern with Birmingham in Fig. 3.

The measles data was divided into a training set and a test set. The training sets were chosen as the period 1945 (January) to 1966 (June) and thereafter were the test data. Using various forms of equation

3, our objective was to select the best model to understand the spatial coupling between these seven UK cities, study the transmission within and between cities through $\phi_d^{jj'}, j, j' = 1, \ldots, 7$, and the region specific trend, seasonality and covariates which affect the spread of the disease. For this we studied long term trend effect, seasonality (annual/biennial) and covariates such as live births in 1946 and baby boom effect in the model. To select the lag order d, we tested different values of p ranging from 1 to 7, and ranked the models according to their BIC values, the lowest BIC (1286.25) was obtained for p=1. We checked various forms of $g(\cdot)$ in equation 3; and, the best fitting results were obtained for the logarithm function. The models chosen based on the training set for each of the 7 cities are described in Table 3. These models were chosen with respect to various goodness of fitness statistics as mentioned in section 2.4. To see some of the values of the GOF statistics for the chosen models, refer to Figure 4a.

Incidence counts across all spatial units constitute a multivariate time series with inter-dependencies; however, conditional on H_t^j, the incidence time series for each region can be treated as independent as shown in equation 1. Consequently, we applied the GLM function from the base stats package in R to each city separately, using the model specification described above. Estimates of model parameters, standard errors, and p-values for each of the seven cities are listed in Table 4. From Table 4, we observe an increasing trend in the incidence of measles in all cities. Strong seasonal patterns, both semi-annual and annual, are evident across all locations, indicating periodic oscillations in measles transmission. The covariates related to demographic dynamics, specifically live births and the post-World War-II baby boom, demonstrate significant associations with the incidence of measles. The transmission parameter, ϕ_1^{jj}, modeling the effect of disease transmission within the city was significant in all cities. We see that the odds of infection increase by a factor of 2.50 for Birmingham, 2.49 for Nottingham, 2.39 for Leicester, 2.43 for Coventry, 2.29 for Wolverhampton, 2.41 for Derby, 2.48 for Walsall. To find $\phi_1^{jj'}, j \neq j'$, we used $\phi_1^{jj'} = \phi_1^{jj} w_{jj'}$ for two different weight functions, $w_{jj'} = (\frac{1}{1+d_{jj'}})^2$, and $w_{jj'} = \exp(-d_{jj'})$.

Short- and long term forecasting was performed in the range of 1 month, 6 months, and 12 months for the test data along with 95% intervals using the method described in section 2.5. We note that both in-sample and forecast performance generally deteriorated with city population size (Figure 4). The accuracy of the forecast is evaluated using the root mean square error (RMSE) defined as RMSE= $\sqrt{\frac{1}{h}\sum_{t=1}^{h}(I_t^j - \hat{I}_t^j)^2}$ (Hyndman and Koehler (2006)). An HDR confidence interval for RMSE was obtained based on the HDR confidence interval for the predicted disease counts as in section 2.5.

5 Analysis of Measles Counts in West Bengal

Since measles is a vaccine-preventable disease, its outbreaks can be largely controlled through immunization. In most parts of the world, measles is now close to eradication due to widespread vaccination and cumulative immunity developed through both vaccination and prior infection. West Bengal, a state in India with an estimated population of 99.36 million, introduced the first dose of the measles-containing vaccine (MCV1) in 1985 and the second dose (MCV2) in 2010 as part of India's measles elimination program.

We extracted monthly aggregated measles cases, and immunization information for first and second doses of measles in all districts of West Bengal from March 2014 to April 2020 from the Health Management Information System (HMIS) database, which is publicly available and downloadable at https://hmis.mohfw.gov.in/#!/standardReports. Although the state currently has 23 districts, it underwent several administrative boundary changes during the study period. The Barddhaman district was divided into Purba Barddhaman and Paschim Barddhaman, and two new districts, Jhargram and Kalimpong, were created. In addition, Alipurduar was separated from Jalpaiguri and established as an independent district in 2015. For district specific measles cases and vaccination coverages, see Figures 5 and 6.

As the first dose of the measles vaccine is administered at 9 months of age and the second dose at 16 months of age, to account for the vaccination effect, we again took the unvaccinated cohort at time t in district j as:

$$X_t^j = B_t^j\left[(1 - 0.85\mathrm{V}_1^j) + (1 - 0.99V_1^jV_2^j)\right]$$

where B_t^j denotes the birth cohort in the time period t in district j. The respective vaccine coverages, V_1^j and V_2^j are taken from Figure 6. As discussed in Section 2.3, the generative equation for the number of susceptibles at time t is updated to $S_t^j = S_0^j - \sum_{i=1}^t I_i^j + \sum_{i=1}^t X_i^j$.

For setting up the spatial neighborhoods, both choices N1 and N2 were used. In the nearest neighbor method (N2) k was taken to be 3. As before, we tested for trend and seasonality terms in our model.

We have applied the model described in section 2.1, with three nearest neighbor districts as a spatial cluster (N_2 choice). For model training, data from April 2014 to September 2019 were used, while forecasting was done for October 2019 to March 2020. The prediction model for each district was chosen as described based on several GOFs as described in Section 4. The BIC value was reported to be lowest for $p = 1$ (4329.40). In Figure 7, we show the medium-term forecast of six months with respect to two types of spatial neighbor choices, N1 and N2. For N_1, we used $w_{jj'} = \exp(-d_{jj'})$. From Figure 7, we note that the forecast results based on the prediction model with choice N1 outperformed the model that only involved 3 nearest districts (N_2). This is likely to happen due to the well-developed connectivity between the districts of West Bengal. The movement of individuals may not be restricted to the immediate neighboring units and may extend across all districts. Noting that West Bengal's 3,825-kilometer rail network, mainly run by the Eastern and South Eastern Railway zones, connects its districts efficiently. Important routes like the Howrah–Barddhaman–Asansol and Malda–New Jalpaiguri lines handle key east-west and north-south travel. The widespread Kolkata Suburban Railway serves as a vital daily link for eight surrounding districts. This rail network is complemented by a robust road system, with major National Highways such as NH-34, NH-2, NH-31, and NH-55 ensuring strong connectivity throughout the state.

Based on the estimated values of N_1 choice, we sought to identify the dominant inter-district connections influencing measles transmission. Since the choice of N_1 assumes a fully connected network in which all districts are linked, we introduced sparsity by applying a thresholding procedure to the estimated interaction parameters $\phi_d^{jj'}$. The estimated values of $\phi_d^{jj'}$, for all district pairs j, j', ranged from 0 to 0.237, with a median value of 0.014. We classified a connection as dominant if $\phi_d^{jj'} > 0.014$; otherwise, the corresponding parameter was set to zero. This thresholding approach enabled the identification of key transmission pathways while reducing weaker connections.

The number of dominant edges associated with each district is illustrated in Figure 8. The districts in the central region of West Bengal exhibit greater connectivity, likely reflecting stronger transportation networks and greater livelihood opportunities that facilitate population movement. In contrast, northern districts show comparatively fewer dominant connections, consistent with migration patterns in which individuals tend to move toward the more economically active central regions of the state.

6 Concluding Remarks

In this study, we developed a novel multivariate model to predict and forecast disease counts based on surveillance data. Our approach integrates advanced time series techniques, specifically the generalized linear auto regressive model, within an epidemiological framework. The model selection process involved evaluating multiple candidate models to identify the one that performed the best.

A key limitation of our model is its inability to account for under-reporting, a common issue in surveillance data. Detailed information on under-reporting is generally unavailable. As a future direction, we are currently working to develop a model that uses sero-surveillance data to understand the size of the susceptible population and consequently estimate the underreported cases. While modelling infectious diseases, substantial heterogeneity in incidence is often observed across age groups. This phenomenon is particularly pronounced in childhood infections, where younger children tend to be more vulnerable to infection than older children. In future work, we aim to incorporate such age-specific heterogeneity into our spatio-temporal modelling framework and investigate its impact on parameter estimation as well as short- and long-term forecasting performance.

Conflict of Interest

The authors declare that they have no potential conflict of interest.

Disclaimer

The work/opinion is based on research findings by the authors and not the opinion of the government.

Acknowledgments

Funding for this study was provided by the Gates Foundation (INV-044445). We thank Dr. Shubham Niphadkar for his valuable inputs in this project.

CASE	Error Statistics	Region		
		1st Region	**2nd Region**	**3rd Region**
S1C1	**MAE**	0.9577307	2.36089	2.36089
	RMSE	0.9729221	4.917786	4.917786
S1C2	**MAE**	1.023573	16.71269	18.3528
	RMSE	1.192331	39.52522	42.682
S2C1	**MAE**	1.013975	5.118036	5.103921
	RMSE	1.05589	13.22069	13.18252
S2C2	**MAE**	1.044087	17.28936	19.01157
	RMSE	1.215637	40.7225	43.99615

Table 1: Error Statistics for Different Regions Under Various Simulation Scenarios in Section 3. Scenarios 1 and 2 are indicated as S1 and S2. Cases 1 and 2 are indicated as C1 and C2.

	Coventry	Leicester	Birmingham	Derby	Nottingham	Wolverhampton	Walsall
Coventry	0	36.18591	25.40883	55.24824	64.78265	45.08821	36.47235
Leicester	36.18591	0	53.24797	38.19645	36.04171	67.42161	56.88780
Birmingham	25.40883	53.24797	0	54.57648	70.57708	20.10414	13.78967
Derby	55.24824	38.19645	54.57648	0	20.32999	56.87155	48.41575
Nottingham	64.78265	36.04171	70.57708	20.32999	0	75.93764	66.73816
Wolverhampton	45.08821	67.42161	20.10414	56.87155	75.93764	0	10.53635
Walsall	36.47235	56.88780	13.78967	48.41575	66.73816	10.53635	0

Table 2: Distance in km among the seven UK cities for Section 4.

City	Linear Trend	Seasonality		Weights	Covariates	
		Annual	Bienniality		$\mathbf{B_t}$	$\mathbf{M_t}$
Birmingham	Yes	Yes	Yes	$\exp(-d_{jj'})$	Yes	Yes
Coventry	Yes	Yes	Yes	$\exp(-d_{jj'})$	Yes	Yes
Derby	Yes	Yes	Yes	$\exp(-d_{jj'})$	Yes	Yes
Leicester	Yes	Yes	Yes	$\frac{1}{(1+d_{jj'})^{\rho}}$	Yes	Yes
Nottingham	Yes	Yes	Yes	$\exp(-d_{jj'})$	Yes	Yes
Walsall	Yes	Yes	Yes	$\frac{1}{(1+d_{jj'})^{\rho}}$	Yes	Yes
Wolverhampton	Yes	Yes	Yes	$\frac{1}{(1+d_{jj'})^{\rho}}$	Yes	Yes

Table 3: Description of best model for each of the seven UK cities in Section 4.

(a) Birmingham

Variable	Estimate	SE	p-value
Intercept	−13.2100	0.0314	<.001
Endemic-Epidemic	0.9195	0.0026	<.001
Bi-weekly Birth Counts	0.0003	0.0000	<.001
Annual Seasonality	0.1238	0.0039	<.001
Biennial Seasonality	0.1943	0.0046	<.001
Trend	0.0004	0.0000	<.001
Baby Boom Effect	0.0839	0.0069	<.001

(b) Coventry

Variable	Estimate	SE	p-value
Intercept	−12.1300	0.0349	<.001
Endemic-Epidemic	0.8880	0.0041	<.001
Bi-weekly Birth Counts	−0.0012	0.0002	<.001
Annual Seasonality	0.0574	0.0065	<.001
Biennial Seasonality	0.1527	0.0084	<.001
Trend	0.0004	0.0000	<.001
Baby Boom Effect	0.0631	0.0137	<.001

(c) Derby

Variable	Estimate	SE	p-value
Intercept	−10.9500	0.0731	<.001
Endemic-Epidemic	0.8819	0.0061	<.001
Bi-weekly Birth Counts	0.0025	0.0006	<.001
Annual Seasonality	0.0690	0.0099	<.001
Biennial Seasonality	0.1836	0.0113	<.001
Trend	0.0003	0.0000	<.001
Baby Boom Effect	0.0731	0.0187	<.001

(d) Leicester

Variable	Estimate	SE	p-value
Intercept	−11.3000	0.0599	<.001
Endemic-Epidemic	0.8729	0.0040	<.001
Bi-weekly Birth Counts	0.0029	0.0003	<.001
Annual Seasonality	0.1925	0.0627	<.001
Biennial Seasonality	0.1907	0.0073	<.001
Trend	0.0002	0.0000	<.001
Baby Boom Effect	0.0600	0.0132	<.001

(e) Nottingham

Variable	Estimate	SE	p-value
Intercept	−11.8800	0.0599	<.001
Endemic-Epidemic	0.9146	0.0044	<.001
Bi-weekly Birth Counts	0.0015	0.0002	<.001
Annual Seasonality	0.0966	0.0066	<.001
Biennial Seasonality	0.0344	0.0075	<.001
Trend	0.0005	0.0000	<.001
Baby Boom Effect	0.0974	0.0131	<.001

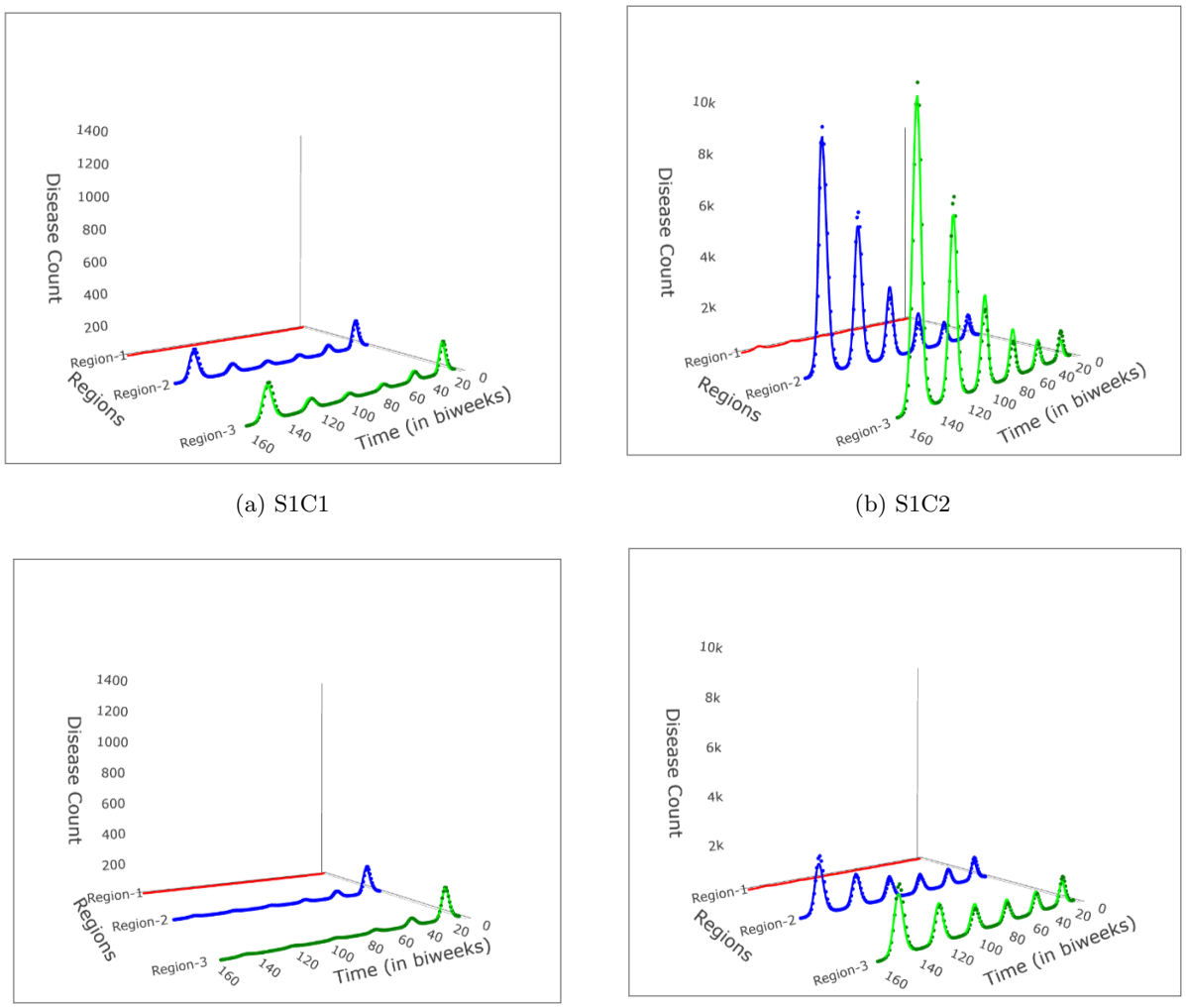
(f) Walsall

Variable	Estimate	SE	p-value
Intercept	−11.0800	0.0682	<.001
Endemic-Epidemic	0.9100	0.0062	<.001
Bi-weekly Birth Counts	0.0016	0.0007	0.0223
Annual Seasonality	0.1912	0.0126	<.001
Biennial Seasonality	0.1807	0.0128	<.001
Trend	0.0007	0.0000	<.001
Baby Boom Effect	0.0240	0.0205	0.2428

(g) Wolverhampton

Variable	Estimate	SE	p-value
Intercept	−10.8500	0.0535	<.001
Endemic-Epidemic	0.8304	0.0065	<.001
Bi-weekly Birth Counts	0.0028	0.0004	<.001
Annual Seasonality	0.1446	0.0109	<.001
Biennial Seasonality	0.3037	0.0129	<.001
Trend	0.0004	0.0000	<.001
Baby Boom Effect	0.0108	0.0185	0.5580

Table 4: Parameter estimates, standard errors and p-values for of the models fitted to the measles counts from all the concerned cities in Section 4.



(a) S1C1

(b) S1C2

(c) S2C1

(d) S2C2

Figure 1: Comparison of measles incidences in the three regions for the two simulation scenarios (S1-S2) and cases (C1-C2). In each plot, (•), (•), (•) represent simulated counts of region 1, 2, and 3 respectively, and (—), (—), and (—) represent the estimated counts using the proposed approach in Section 2.

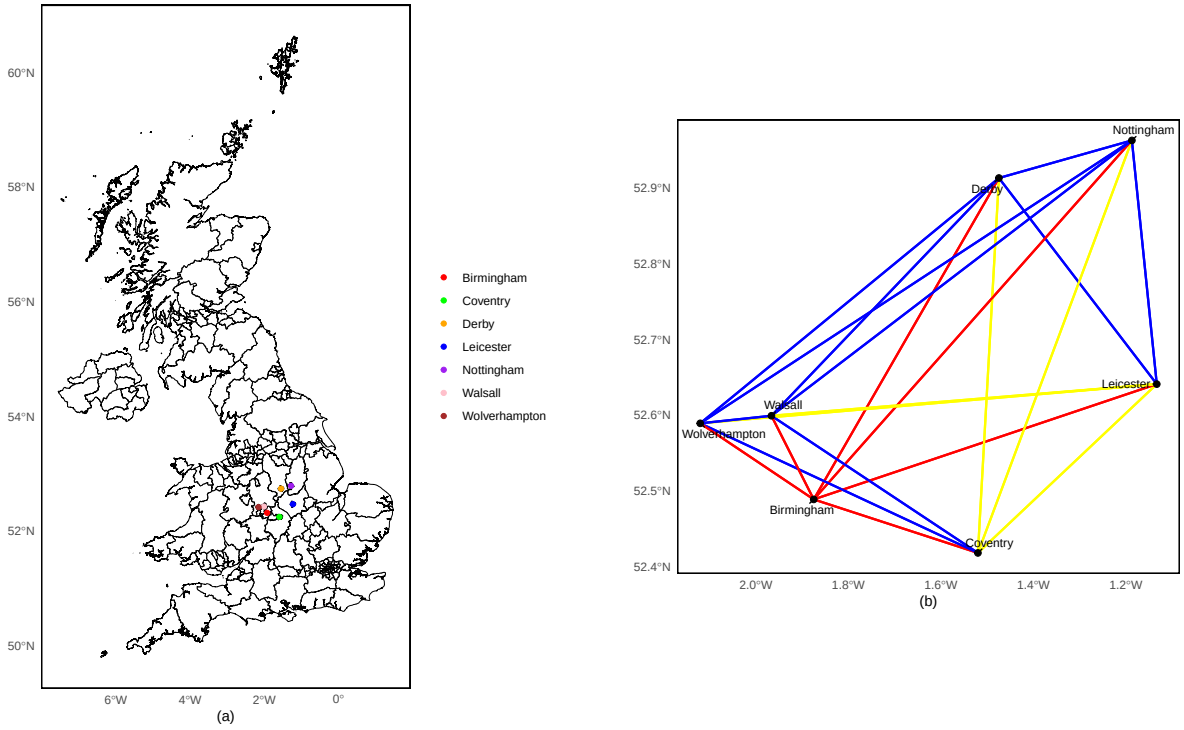


Figure 2: Left panel (a) of the figure shows the location of the seven cities in the UK map and the right panel (b) shows the strength of connection between the seven UK cities. In (b), the red lines represent strong connections, blue lines represent moderate connections and yellow lines represent weak connections.

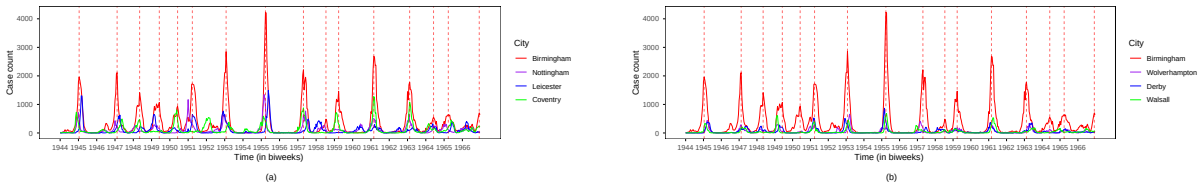


Figure 3: Outbreak comparisons between Birmingham and other cities on the basis of raw incidences in Section 4. The cities in (a) are the high population density cities while in (b) we have the low population density cities.

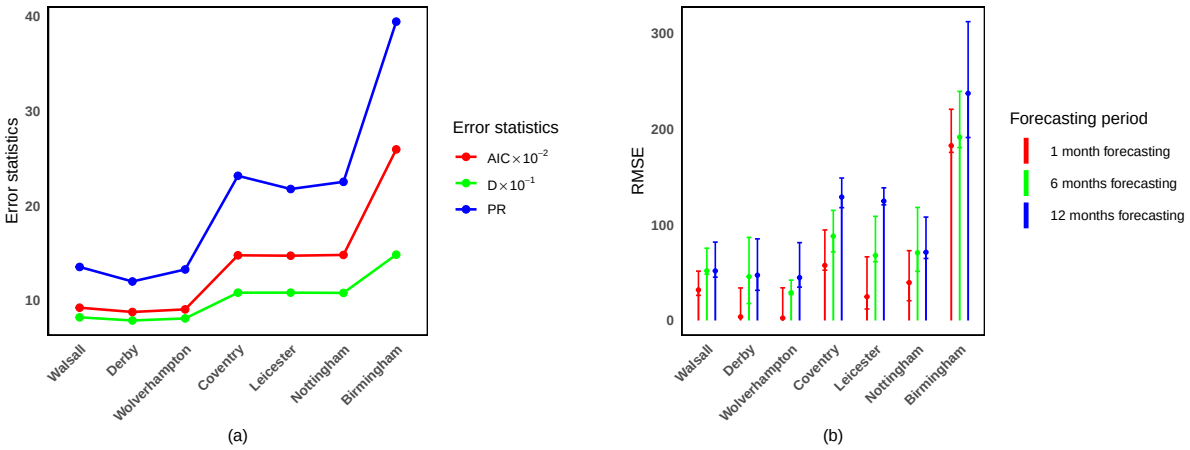


Figure 4: Prediction and forecasting performance for the seven cities. The left panel (a) shows three GOF statistics, Akaike Information Criterion (AIC), Deviance (D) and Pearson residuals (PR) based on training data. In the right panel (b), we show the forecasting results based on test data. Each dot indicates the RMSE value, and the surrounding whiskers represent the confidence interval of the RMSE. Cities are ordered on the x-axis according to their population.

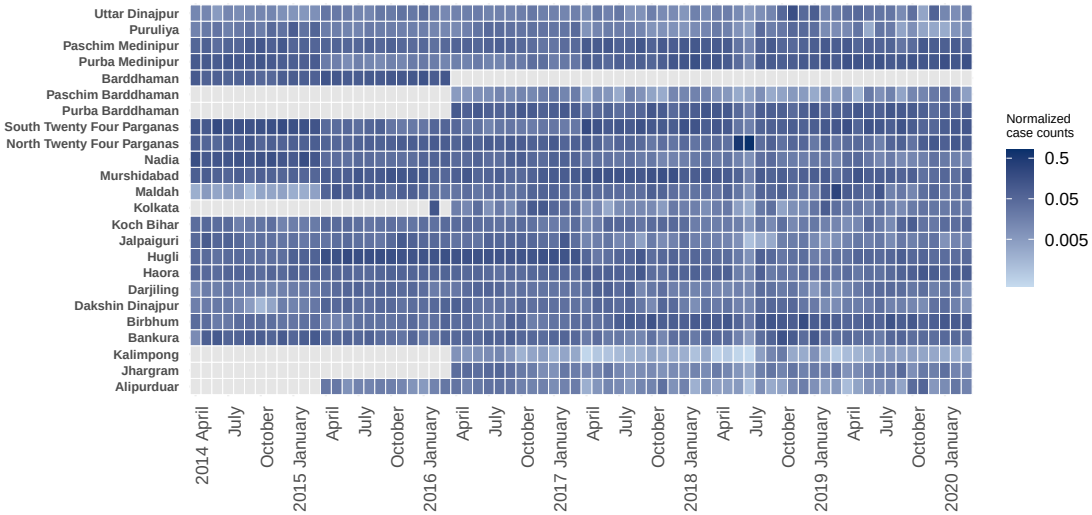


Figure 5: Normalized measles incidence rates (in monthly scale) for districts in West Bengal (2014-2020) corresponding to Section 5. The grey color indicates either districts that did not exist at the time (e.g., Jhargram, Kalimpong, Alipurduar, Purba Barddhaman and Paschim Barddhaman) or areas with missing data (e.g., Kolkata).

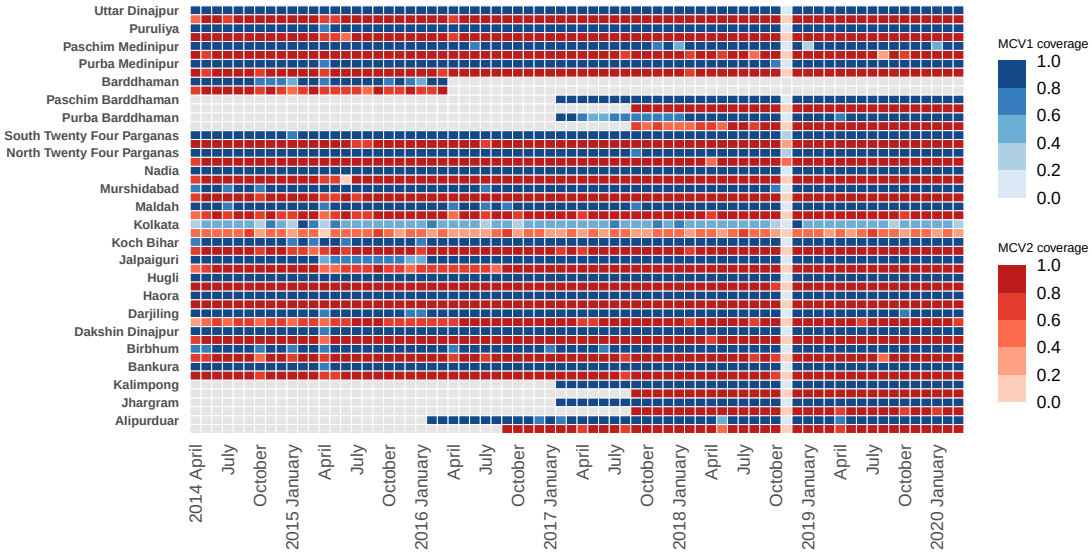


Figure 6: District wise vaccination coverage in West Bengal from 2014 to 2020 corresponding to Section 5. Grey shading indicates periods prior to the initiation of vaccination or missing vaccination data for specific districts. Imputation was used for the missing coverages.

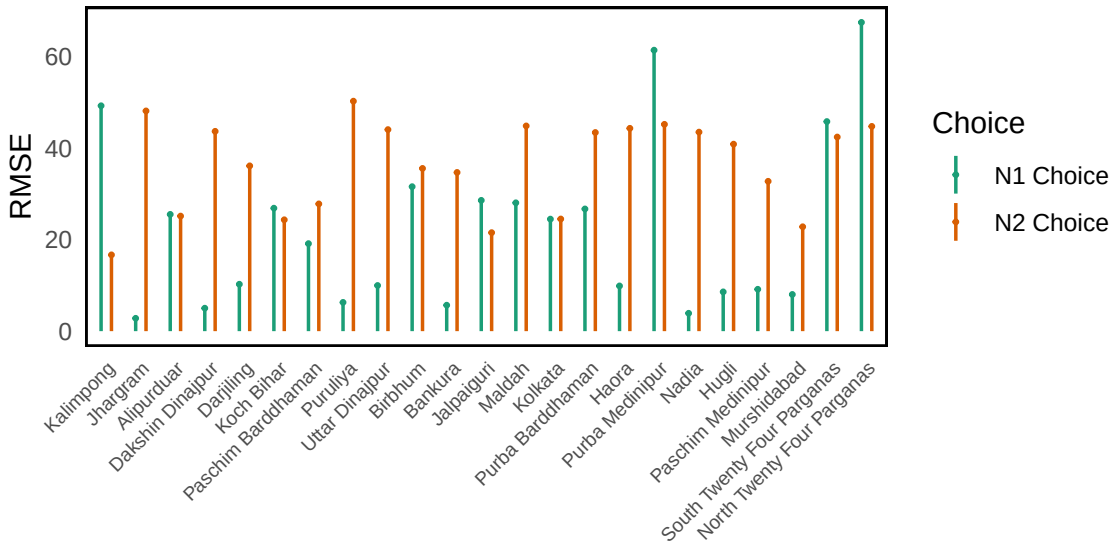


Figure 7: Comparison of forecasting performances between two choices N1 and N2 for West Bengal district level measles data corresponding to Section 5. Districts are ordered according to their population size.

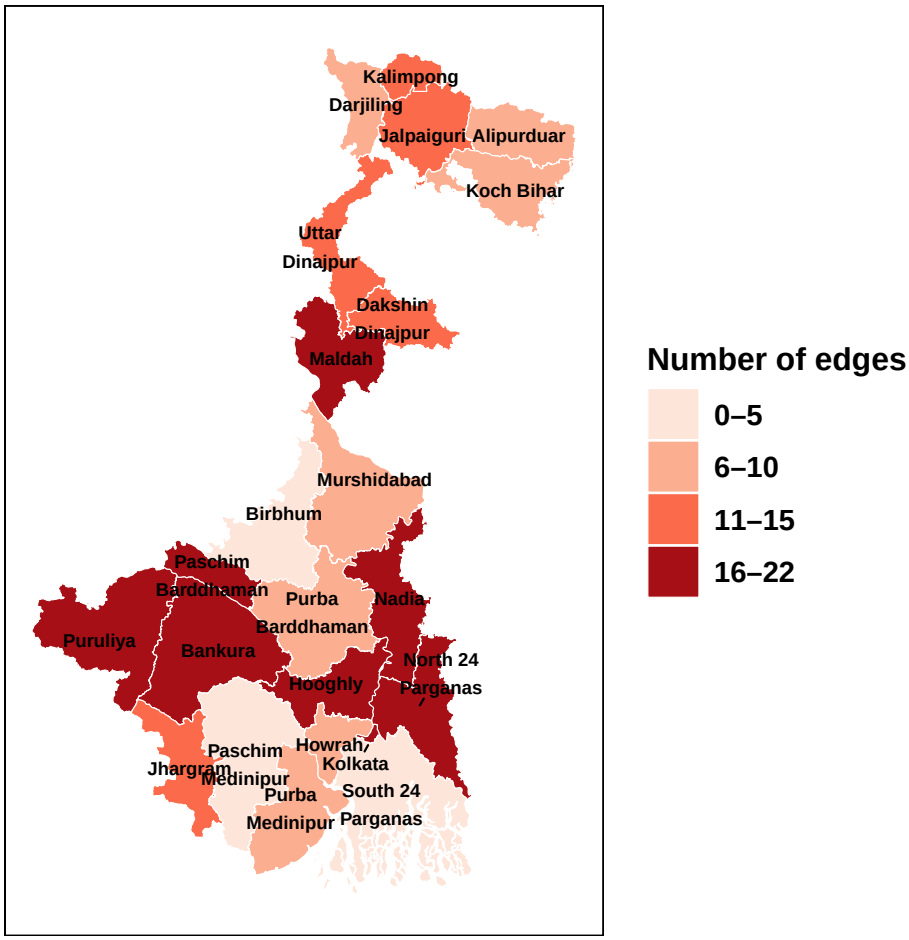


Figure 8: Connection strength of each district in West Bengal corresponding to Section 5.

References

Abbey, H. (1952). An examination of the Reed-Frost theory of epidemics. *Human biology*, 24(3):201.

Anderson, R. M. and May, R. M. (1991). *Infectious diseases of humans: dynamics and control*. Oxford university press.

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

Bailey, N. T. (1957). *The mathematical theory of epidemics*. Charles Griffin & Co., Ltd.

Bauer, C. and Wakefield, J. (2018). Stratified space–time infectious disease modelling, with an application to hand, foot and mouth disease in china. *Journal of the Royal Statistical Society Series C: Applied Statistics*, 67(5):1379–1398.

Becker, N. G. (2017). *Analysis of infectious disease data*. Chapman and Hall/CRC.

Bjørnstad, O. N. (2022). *Epidemics: models and data using R*. Springer Nature.

Bjørnstad, O. N., Finkenstädt, B. F., and Grenfell, B. T. (2002). Dynamics of measles epidemics: estimating scaling of transmission rates using a time series SIR model. *Ecological Monographs*, 72(2):169–184.

Bracher, J. and Held, L. (2022). Endemic-epidemic models with discrete-time serial interval distributions for infectious disease prediction. *International Journal of Forecasting*, 38(3):1221–1233.

Chen, S., Fricks, J., and Ferrari, M. J. (2012). Tracking measles infection through non-linear state space models. *Journal of the Royal Statistical Society Series C: Applied Statistics*, 61(1):117–134.

En'Ko, P. (1989). On the course of epidemics of some infectious diseases.

Eye, H. (2016). The black country's lost railway network. Accessed: 2025-11-04.

Ferrari, M. J., Bjørnstad, O. N., and Dobson, A. P. (2005). Estimation and inference of R_0 of an infectious pathogen by a removal method. *Mathematical Biosciences*, 198(1):14–26.

Ferrari, M. J., Grais, R. F., Bharti, N., Conlan, A. J., Bjørnstad, O. N., Wolfson, L. J., Guerin, P. J., Djibo, A., and Grenfell, B. T. (2008). The dynamics of measles in sub-saharan africa. *Nature*, 451(7179):679–684.

Fokianos, K. and Kedem, B. (1998). Prediction and classification of non-stationary categorical time series. *Journal of Multivariate Analysis*, 67(2):277–296.

Fokianos, K. and Kedem, B. (2004). Partial likelihood inference for time series following generalized linear models. *Journal of Time Series Analysis*, 25(2):173–197.

Grenfell, B. T., Bjørnstad, O. N., and Kappey, J. (2001). Travelling waves and spatial hierarchies in measles epidemics. *Nature*, 414(6865):716–723.

Held, L., Höhle, M., and Hofmann, M. (2005). A statistical framework for the analysis of multivariate infectious disease surveillance counts. *Statistical Modelling*, 5(3):187–199.

Hethcote, H. W. (2000). The mathematics of infectious diseases. *SIAM Review*, 42(4):599–653.

Hyndman, R. J. (1995). Highest-density forecast regions for nonlinear and non-normal time series models. *Journal of Forecasting*, 14(5):431–441.

Hyndman, R. J. and Koehler, A. B. (2006). Another look at measures of forecast accuracy. *International journal of forecasting*, 22(4):679–688.

Kedem, B. and Fokianos, K. (2005). *Regression models for time series analysis*. John Wiley & Sons.

Keeling, M. J. and Rohani, P. (2002). Estimating spatial coupling in epidemiological systems: a mechanistic approach. *Ecology Letters*, 5(1):20–29.

Kermack, W. O. and McKendrick, A. G. (1927). 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*, 115(772):700–721.

Kianfar, N., Mesgari, M. S., Mollalo, A., and Kaveh, M. (2022). Spatio-temporal modeling of covid-19 prevalence and mortality using artificial neural network algorithms. *Spatial and Spatio-temporal Epidemiology*, 40:100471.

Lefèvre, C., Picard, P., Simon, M., and Utev, S. (2021). A chain binomial epidemic with asymptomatics motivated by covid-19 modelling. *Journal of Mathematical Biology*, 83(5):54.

Lu, J. and Meyer, S. (2023). A zero-inflated endemic–epidemic model with an application to measles time series in Germany. *Biometrical Journal*, 65(8):2100408.

Madden, W. G., Jin, W., Lopman, B., Zufle, A., Dalziel, B., E. Metcalf, C. J., Grenfell, B. T., and Lau, M. S. (2024). Deep neural networks for endemic measles dynamics: Comparative analysis and integration with mechanistic models. *PLoS computational biology*, 20(11):e1012616.

Meyer, S. and Held, L. (2017). Incorporating social contact data in spatio-temporal models for infectious disease spread. *Biostatistics*, 18(2):338–351.

Mukhopadhyay, S. and Sathish, V. (2019). Predictive likelihood for coherent forecasting of count time series. *Journal of Forecasting*, 38(3):222–235.

Mukhopadhyay, S., Tiwari, R., Shetty, P., Gogtay, N., and Thatte, U. (2019). Modeling and forecasting Indian malaria incidence using generalized time series models. *Communications in Statistics: Case Studies, Data Analysis and Applications*, 5(2):111–120.

Sattenspiel, L. and Dietz, K. (1995). A structured epidemic model incorporating geographic mobility among regions. *Mathematical Biosciences*, 128(1-2):71–91.

Thakkar, N., Gilani, S. S. A., Hasan, Q., and McCarthy, K. A. (2019). Decreasing measles burden by optimizing campaign timing. *Proceedings of the National Academy of Sciences*, 116(22):11069–11073.

Xia, Y., Bjørnstad, O. N., and Grenfell, B. T. (2004). Measles metapopulation dynamics: a gravity model for epidemiological coupling and dynamics. *The American Naturalist*, 164(2):267–281.

Zakary, O., Rachik, M., and Elmouki, I. (2017). On the analysis of a multi-regions discrete SIR epidemic model: an optimal control approach. *International Journal of Dynamics and Control*, 5:917–930.