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

Pallab K. Sinha[1] Shubham Niphadkar[1] Siuli Mukhopadhyay[1,2,*]

[1]Department of Mathematics, Indian Institute of Technology Bombay, Maharashtra, India.
[2]National Disease Modelling Consortium, 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 various other sociodemographic factors such as live births, population density, vaccination coverage, and disruption in disease surveillance have been taken into account while modeling multiple infection time series. For estimating these multiregion chain-binomial models, an appropriate likelihood function maximization approach is proposed. Simulation results considering effects of seasonal pattern in disease outbreaks, out-of-sync outbreaks in connected geographical regions, variations in interventions have been considered to depict realistic disease scenarios. Forecasting of spatial infections have also been studied. A real-world application based on measles counts from adjoining spatial regions is presented to motivate the proposed modeling approach.

Keywords: Epidemiological models, GL-ARMA models, Highest Density Regions, Predictive likelihood, Spatio-temporal model

1 Introduction

Spatio-temporal models of infectious diseases enable a better understanding of the disease dynamics by mapping its progression simultaneously 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 (Grenfell et al. (2001)), and most recently the Covid-19 pandemic (Kianfar et al. (2022)). This spatial propagation of infectious diseases can be mainly attributed to the movement of infectious individuals from one region to another. 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 models, such as 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.

Surveillance networks at the national or sub-national levels are usually set up to provide daily or monthly counts of particular diseases from several adjoining regions. These multiple time series of disease counts often show dependencies when the regions share geographical boundaries, are connected with each other through efficient public transport systems, or are historically connected through migration. The intermixture of infectious individuals from one region with healthy susceptibles from another region causes spatial spread of the disease. However, in conjunction with the spread of disease that is associated

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

Code repository: https://github.com/pallabksinha/Spatio-temporal-modelling

with the mixing of individuals between regions, we also often observe that factors, such as region-specific healthcare interventions, sometimes have an effect on the burden of the disease in neighboring regions. For example, when two regions A and B share a common boundary, a high vaccine coverage in region A can cause a decrease in the flow of infectious individuals or at the same time increase the inflow of vaccinated individuals from region A to B, consequently reducing the vaccine preventable disease in B. This happens as the total number of disease cases in region B is a function of its own local disease cases and the disease cases due to the spatial effect of A.

To model the spatial transmission of infections caused by infectious individuals from one region coming into effective contact with the susceptibles of another region, we propose a multi-region chain binomial framework. This spatial chain binomial model is able to successfully incorporate the effect of region-specific covariates or programmatic interventions such as socio-demographic factors, coverages of vaccination campaigns or MDA rounds for negelected tropical diseases in transmission. Using a distance-based contiguity matrix as popularly used in the spatial literature, we study the infection at a particular region as a function of local infections, infections in nearby regions and covariate effects. To enable the inclusion of the covariate effects on spatial transmission of infections we propose a more statistical version of the classical stochastic model for discrete-time epidemics or the so-called chain binomial model (Bailey (1957), Ferrari et al. (2005)). The region-specific mean infection at a particular time point is related through a link function to a linear predictor containing nearby infections and covariates as in GLARMA models of Fokianos and Kedem (2004). Simulations are used to study the effect of region-specific variations in spatial distances, infection counts, disease transmission rates, and vaccine coverages. The biweekly city level measles data set from the United Kingdom (Fine and Clarkson (1982)) is used to illustrate the proposed modelling approach. Comparison of models are carried out by using various criterion functions such as Pearson residuals, chi-squared statistic, deviance, and Akaike Information criterion (AIC). Prediction and forecasting of disease counts, both in and out of samples, based on the chosen model is done and forecast intervals are plotted based on the predictive likelihood method (Mukhopadhyay and Sathish (2019)) and the highest density regions of Hyndman (1995).

The foundational SIR model, introduced by Kermack and McKendrick in 1927 (Kermack and McKendrick (1927)), remains a cornerstone of epidemiological modeling, with its compartmental framework serving as the basis for more advanced approaches that address the intricate and stochastic nature of disease spread. Zakary et al. (2017) proposed a multiregion version of the discrete SIR epidemic model which tracks infectious diseases spreading across multiple geographical regions, incorporating population movement between regions. The popular TSIR model (Grenfell et al. (2001)) widely used in the modeling of measles by integrating time series methods with an SIR epidemic model is studied for spatial spread of the disease by Bjørnstad et al. (2002) and Xia et al. (2004). Bjørnstad et al. (2002) studied the spread of measles in connected regions assuming that the force of infection depends on the number of resident infectious hosts, the number of immigrant infections, and the transmission rate. While, Xia et al. (2004) integrated the TSIR model with a gravity-based spatial interaction model to account for the transmission of measles between different communities. Another series of models, namely the endemic-epidemic models (Held et al. (2005)) or the HHH model have often been used to model infectious diseases. The fundamental idea in the HHH model is to divide the 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. These models by Meyer and Held (2017) also proposed a spatio-temporal version where the disease case counts at a specific location and time were assumed to be either a Poisson or a Negative binomial random variable with the corresponding mean as a combination of autoregressive terms, influx of cases from other locations and a combination of time and unit specific factors. Bauer and Wakefield (2018) also used a negative binomial random variable to model the count data. They incorporated a endemic random component, neighborhood random effects and endemic conditional auto-regressive (CAR) spatial random effects in their model. More recently, Bracher and Held (2022) employed different weighting scheme such as shifted poisson weights, a linear decay of weights and geometric weight in the context of a HHH model. Lu and Meyer (2023) extended the HHH model to a zero-inflated HHH model to tackle the case of zero counts.

The spatial versions of the TSIR models (MTSIR) do not take into account the effect of region specific covariates that may influence the dynamics of the disease process. While the HHH model being more statistical in nature, consider covariate effects but do not consider the epidemiological information of the disease, such as number of susceptibles and birth-and-death rates in their infection modeling. Our proposed multi-region chain binomial model proposes an efficient way to combine epidemiogical 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 provides a brief overview of the multivariate chain binomial model, along with details on parameter estimation and forecasting of out of sample incidences. In Section 3, a detailed simulation study is conducted to illustrate various scenarios. Section 4 validates the proposed model using a real-world dataset comprising measles case counts from various cities in the United Kingdom during the period 1944–1966. Finally, Section 5 concludes the paper with some closing remarks.

2 A new chain binomial model

We start by considering a single time series of disease counts. The number of infections at time $(t+1)$ is modelled as in Ferrari et al. (2005),

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

where S_t is the number of individuals at risk or susceptible to the disease at time t. The probability of infection, π_{t+1} is usually chosen to be an appropriate non-decreasing function of number of infections at time t, i.e., I_t, between 0 and 1. The generative equation for the number of susceptible at time t is,

$$S_{t+1} = S_t - I_{t+1},$$

or, $S_t = S_0 - \sum_{i=1}^{t} I_i$, here S_0 is the initial size of the susceptible population. From Vynnycky and White (2010), we know $S_{t+1} = S_t - I_{t+1} = (1-\lambda_t)S_t$, where $\lambda_t = \mu I_t$ and μ is the effective rate of contact. In a chain binomial setup, Ferrari et al. (2005) proposed using $\pi_{t+1} = 1 - e^{-\mu I_t}$.

In this article however, we propose a more statistical approach and model the probability of infection using a logit-link function. Thus, given the past information, H_t, till time t, the conditional distribution of I_{t+1} is:

$$I_{t+1}|H_t \sim Bin(S_0 - \sum_{i=1}^{t} I_i, \pi_{t+1})$$

where

$$logit(\pi_t) = \eta_t = \mathbf{x}_t'\boldsymbol{\beta} + \sum_{d=1}^{p} \phi_d I_{t-d} = z_t\boldsymbol{\nu},$$

with $\boldsymbol{\nu} = (\beta, \phi_1, \ldots, \phi_p)'$, $z_t = (x_t', I_{t-1}, \ldots, I_{t-p})$ and $p < t$.

In the above equation, $\mathbf{x}'_t$ is assumed to be of the form, $A_t + B_t + C_t$, where A_t represents 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 auto-regressive part of the linear predictor is represented by the p terms, $\sum_{d=1}^{p} \phi_d I_{t-d}$. Note we deviate from a Markovian type of dependence as in the conventional chain binomial model and assume that infections at time t may depend on infections till time $(t-p)$. Based on the above assumptions, $H_t = \{C_t, S_0, I_1, \ldots, I_t\}$. 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 varying 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 multivariate version of the chain binomial model

We extend the above model to capture the spatial spread of a disease. Suppose we have m spatial regions and disease counts from each of these regions have been measured. We propose a multivariate version of the chain binomial model as follows,

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

where I_t^j, S_t^j and π_t^j denote the same quantities as before, for region j. Using a multiple SIR type of framework as in Zakary et al. (2017), we write $S_{t+1}^j = S_t^j - I_{t+1}^j = (1-\lambda_t^j)S_t^j$, where the region-specific force of infection is, $\lambda_t^j = \sum_{j'=1}^m \mu_{jj'} I_t^{j'}$. Using these relations, the generative equation for the jth susceptible population is updated to, $S_t^j = S_0^j - \sum_{i=1}^t I_i^j$.

As before, we model the region-specific probability of infection to be a function of endemic effects and past disease counts,

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

Counts from different spatial regions j and j' at time t are assumed to be conditionally independent given the past, $H_t^j = \{\mathbf{x}_t^j, S_0^j, \mathbf{I}_t, \mathbf{I}_{t-1}, \mathbf{I}_{t-2}, \ldots, \mathbf{I}_1\}$. By $\mathbf{I}_t$ we denote the vector of infections recorded from the m regions at time t, thus $\mathbf{I}_t = (\mathrm{I}_t^1, \ldots, \mathrm{I}_t^m)'$. For simplicity and to avoid overfitting, ϕ_d^j may be assumed to be same for all j'. The connection between spatial regions is then introduced through spatial contiguity weights $w_{jj'}$ (Anselin (2022)) as,

$$logit(\pi_t^j) = \eta_t^j = \mathbf{x}_t^{'j}\boldsymbol{\beta}^j + \sum_{d=1}^{p}\phi_d^j \sum_{j'=1}^{m} w_{jj'} I_{t-d}^{j'} = Z_t^j \boldsymbol{\nu}^j, \tag{2}$$

These weights can be modeled as a decreasing function of the distance between the centroids of the spatial regions j and j'. One popular choice of weight is $w_{jj'} = (\frac{1}{1+d_{jj'}})^\rho$, where ρ is estimated from the data. In practice the weights are normalized before usage.

2.2 Parameter Estimation

We use the method of maximum likelihood for estimating the parameters of the multivariate model. Since the covariates are time dependent, we maximize the partial log-likelihood function as in Fokianos and Kedem (2004),

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

to obtain the MLE of $\boldsymbol{\nu} = (\boldsymbol{\nu}^1, \ldots, \boldsymbol{\nu}^m)$. 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 $Z_t^j \boldsymbol{\nu}^j$. The Fisher scoring algorithm is used to find $\boldsymbol{\nu}$ iteratively as,

$$\hat{\boldsymbol{\nu}}^{(w)} = \hat{\boldsymbol{\nu}}^{w-1} + G_n^{-1}(\hat{\boldsymbol{\nu}}^{w-1}) S_n(\hat{\boldsymbol{\nu}}^{w-1}) \tag{3}$$

where w denotes the wth iteration step in the estimation process.

The partial score function is,

$$S_n(\boldsymbol{\nu}) \equiv \nabla(\boldsymbol{\nu}) = \sum_{t=1}^{n}\sum_{j=1}^{m} Z_{t-1}^j \frac{exp(Z_{t-1}^{j'}\nu)}{[1+exp(Z_{t-1}^{j'}\nu)]^2} \frac{I_t^j - \mu_t^j}{(\sigma_t^j)^2}$$

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

$$G_n(\boldsymbol{\nu}) = \sum_{t=1}^{n} Z_{t-1}' W(\boldsymbol{\nu}) Z_{t-1},$$

where $W(\boldsymbol{\nu})$ is a diagonal matrix with elements, $\left(\frac{exp(Z_t^{j'}\boldsymbol{\nu}^j)}{[1+exp(Z_{t-1}^{j'}\boldsymbol{\nu}^j)]^2}\right)^2 \frac{1}{(\sigma_t^j)^2}$, $\mu_t^j = S_{t-1}^j \frac{e^{Z_t^{j'}\boldsymbol{\nu}^j}}{1+e^{Z_t^{j'}\boldsymbol{\nu}^j}}$, 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 Model evaluation and averaging

Various forms of the above multi-region chain binomial model was fitted to simulated and a real datasets of measles counts. For evaluating these models we used popular goodness of fit statistics from model selection literature. Criteria like, model deviance, Chi-squared criterion, Akaike information criterion, Bayesian Schwarz information criterion, various residuals (such as working, Pearson etc.) and mean square errors based on these residuals were used.

2.4 Forecasting using a Predictive Likelihood

Forecasting future disease incidences in each of the m regions is of interest. Suppose that we have incidence values at n time points from each of the m regions, i.e., we have past information, H_n^j for all j. Using the predictive likelihood method (Mukhopadhyay and Sathish (2019)) the estimated density of I_{n+1}^j is,

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

where $k(\cdot)$ is a normalizing constant and $LL_p(I_{n+1}^j|H_n^j) = \prod_{t=1}^{n+1} f_{\hat{\nu}^j}(I_t^j|H_{t-1}^j)$. The forecast value of $I_{n+1}^j = \arg\max_{I_{n+1}^j \in M} \hat{p}(I_{n+1}^j)$, where $M = [\min_t I_t^j, \max_t I_t^j]$. To estimate the prediction interval for $\hat{p}(I_{n+1}^j)$, we use the highest density region of Hyndman (1995). The $(1-\alpha)100\%$ HDR for I_{n+1}^j is the region R_α^j, where $R_\alpha^j = \{I_{n+1}^j : \hat{p}(I_{n+1}^j) \geq f_\alpha\}$, where f_α is the largest constant such that $Pr(I_{n+1}^j \in R_\alpha^j)$ is at least $(1-\alpha)$.

3 Simulation Studies

Results from several simulation studies are used to evaluate the proposed disease model. 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 to control and eliminate the disease. The first dose of the vaccine is scheduled usually at 9-12 months and the second dose is recommended, usually at 15-18 months children. Measles have been modeled by various researchers, namely, Grenfell et al. (2001), Chen et al. (2012), Thakkar et al. (2019).

For generating biweekly measles counts we use the following susceptible-infection equations of the MSIR model (Keeling and Rohani (2002), Zakary et al. (2017)) with various choices of parameters:

$$S_{t+1}^j = S_t^j - \sum_{k=1}^{n} \beta_{jk} S_t^j I_t^k \tag{4}$$

$$I_{t+1}^j = I_t^j + \sum_{k=1}^{n} \beta_{jk} S_t^j I_t^k - \gamma I_t^j, \quad j = 1,2,3 \quad \text{and} \quad t = 0,\ldots,160 \tag{5}$$

The recovery rate, γ, assumed to be $1/14$ usually for measles is taken, β_{jk}'s are the transmission rates between two locations j and k.

Within region transmission rate is considered to be influenced by region specific factors and yearly seasonality,

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

The β_{jk}'s representing transmission rates are modeled to depend on the distance between two spatial regions j and k as:

$$\beta_{jk}(t) = \beta_{jj}(t) w_{jk},$$

where w_{jk} is some weight attached to represent the spatial distance between the regions j and k.

We consider three scenarios with different spatial weight matrices, W. In all these scenarios, it is assumed that there are three regions with transmission parameters and initial incidences, as in the following two cases:

- **C1:** $\bar{\beta}_{jj} = 0.5$, $I_j(1) = 6$, j=1(1)3
- **C2:** $\bar{\beta}_{jj} = 0.5$, $I_j(1) = 6$; $j = 1,2$; $\bar{\beta}_{33} = 0.55$, $I_3(1) = 34$

The initial susceptible population S_0^j in all three regions for all scenarios was taken as 10^6.

First Scenario: S1

Three spatial regions are considered, where regions 2 and 3 are chosen close to each other and equidistant from region 1. One representation of such a scenario is through the following spatial weight matrix, W:

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

Figure 1 plots the disease counts for both cases. For C1, the simulated case counts of regions 2 and 3 are seen to be similar to each other and higher than region 1. This happens because of the compounding effect of disease transmission between the two closely connected regions. Similar results are seen in C2. Here, the counts in region 3 are highest due to higher transmission rate than other regions and different initial condition. These high counts in region 3 are seen to affect the nearby region 2 while not affecting region 1 much.

Second Scenario: S2

All three regions are now assumed to be equidistant from each other,

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

When the regions are equidistant from each other, we see from Figure 1, that I^3 has similar influence on I^1 and I^2 for both cases.

Third Scenario: S3

Interventions such as vaccination strategies play an important role in disease control. To take into account the effect of vaccination, we use a vaccinated birth cohort as in Chen et al. (2012). To adjust the susceptible population in the equations of the MSIR model accordingly, we assume:

$$S_{t+1}^j = S_t^j - \sum_{k=1}^{n} \beta_{jk} S_t^j I_t^k + X_t^j \tag{6}$$

$$I_{t+1}^j = I_t^j + \sum_{k=1}^{n} \beta_{jk} S_t^j I_t^k - \gamma I_t^j, \quad j = 1, 2, 3 \text{ and } t = 0, 1, \ldots, 160 \tag{7}$$

where

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

For the jth region, the birth cohort is indicated by B_t^j, while the first and second vaccination coverages are respectively V_{1t}^j and V_{2t}^j. Note that the efficacy of doses 1 and 2 of measles vaccination has been assumed to be, respectively, 85% and 99% from Chen et al. (2012). We considered parameter values similar to those in C1 and C2, and the same distance matrix as in Scenario 1. Vaccination was introduced exclusively in Region 3, with first-dose coverage increasing linearly from 0.70 to 0.90 and second-dose coverage from 0.40 to 0.70 over biweeks. In each region, the birth cohort was chosen to remain constant over time, $B_j^t = 10^3$. In Figure 1 we note that although vaccination was administered only in Region 3, closely connected or spatially coupled Region 2 also exhibited a similar decrease in case counts compared to Region 1. This reduction can be attributed to the lower influx of infected individuals who migrate from Region 3 to Region 2, thus reducing disease transmission.

In both scenarios, $10,000$ simulations were performed and the proposed model in Section 2 was fitted. The mean MSE and MAE values are listed in Table 1. Note that in S1 and S3, 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. In S2, when the three regions are spatially uncoupled, the MAE and RMSE values for C1 are the same due to similar transmission rates, while for C2, they vary. The distribution of some the estimated parameters (standardised parameters) and the residual analysis for Scenario 3 and Case 1 are shown in Figure 2 and Figure 3, respectively.

4 Real Data Example: Analysis of Measles Counts

To illustrate our proposed model, we fit a multivariate chain binomial model to biweekly case reports of measles from three urban cities, Coventry, Leicester and Birmingham, in the United Kingdom recorded between the years 1944-1966. This data set have been used by various authors (Grenfell et al. (2001), Xia et al. (2004)) for proposing different models to track the transmission of measles. The baby boom effect in 1946 and the effect of live births on measles cases are factored in our model. From Figure 4, we note that the measles counts in each city varies from each other. We also notice a seasonal pattern in the case reports, and introduce harmonic terms to model the seasonality in counts. Xia et al. (2004) observed that infections in some cities frequently became extinct, however the disease persists over the years due to its dependence on episodic reintroduction and possible interaction with infectious cases from neighbouring spatial regions. 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 5% level) were observed, implying existence of spatial correlation among the disease counts of the three cities. Using various forms of equation (1) we try to select the best model for understanding the spatial coupling between these three UK cities and fit the measles data. The measles counts from 1944-1965 are used for model prediction, while the data from year 1966 is used for forecasting the disease progress.

All our models contain a linear time trend, seasonality (either annual and/or biennial), and effects of covariates, such as live births and baby boom. For region j, we assume at most lag 2 terms of the infection counts, while for $j \neq j'$ only lag 1 terms of infection counts in region j' are taken in the model. We used $g(I_t^j)$ in (1) where $g(\cdot)$ was assumed to be either identity, logarithm or the logit function, however the best fitting results were obtained for the logarithm function. The model in (1) was also extended to include error terms (Fokianos and Kedem (2004)) as follows:

$$logit(\pi_t^j) = \eta_t^j = \mathbf{x}_t^{'j}\boldsymbol{\beta}^j + \sum_{d=1}^{p}\phi_d^j \sum_{j'=1}^{m} w_{jj'} I_{t-d}^{j'} + \sum_{d=1}^{q}\theta_d^j e_{t-d}^j,$$

where e_t^j are the residuals measured in the original scale. The different types of models considered are listed in Table 1. For models 1-12, the spatial weights, $w_{jj'}$, are $(\frac{1}{1+d_{jj'}})^2$. We also consider a second set of 12 models with spatial weights $e^{-d_{jj'}}$, and call these second set, models 13-24. Thus, for each of the three UK cities we have 24 models. Several goodness of fit (GOF) measures were used to choose the final models for the 3 cities. The selected model forms for the three cities are listed in equations 8-10.

Suppose we denote the three cities as: C:Coventry , L:Leicester, B:Birmingham.

For Coventry, the best model is M9:

$$\begin{aligned} \text{logit}\pi_t^{\text{C}} &= \eta_t^{\text{C}} = \beta_0^{\text{C}} + \beta_1^{\text{C}} cos(\frac{2\pi t}{26}) + \beta_2^{\text{C}} cos(\frac{2\pi t}{52}) + \beta_3^{\text{C}} t + \beta_4^{\text{C}} B_t^{\text{C}} + \beta_5^{\text{C}} M_t \\ &+ \phi_1^C (w_{\text{CC}} I_{t-1}^C + w_{\text{CL}} I_{t-1}^L + w_{\text{CB}} I_{t-1}^B) + \theta_1^C e_{t-1}^C + \theta_2^C e_{t-2}^C \end{aligned} \tag{8}$$

with $w_{jj'} = (\frac{1}{1+d_{jj'}})^2$.

For Leicester, the best model is M21.

$$\begin{aligned} \text{logit}\pi_t^{\text{L}} &= \eta_t^{\text{L}} = \beta_0^{\text{L}} + \beta_1^{\text{L}} cos(\frac{2\pi t}{26}) + \beta_2^{\text{L}} cos(\frac{2\pi t}{52}) + \beta_3^{\text{L}} t + \beta_4^{\text{L}} B_t^{\text{L}} + \beta_5^{\text{L}} M_t \\ &+ \phi_1^L (w_{\text{LL}} I_{t-1}^L + w_{\text{LC}} I_{t-1}^C + w_{\text{LB}} I_{t-1}^B) + \theta_1^L e_{t-1}^L + \theta_2^L e_{t-2}^L \end{aligned} \tag{9}$$

with $w_{jj'} = \exp(-d_{jj'})$.

For Birmingham, the best model is again M9:

$$\begin{aligned} \text{logit}\pi_t^{\text{B}} &= \eta_t^{\text{B}} = \beta_0^{\text{B}} + \beta_1^{\text{B}} cos(\frac{2\pi t}{26}) + \beta_2^{\text{B}} cos(\frac{2\pi t}{52}) + \beta_3^{\text{B}} t + \beta_4^{\text{B}} B_t^{\text{B}} + \beta_5^{\text{B}} M_t \\ &+ \phi_1^B (w_{\text{BB}} I_{t-1}^B + w_{\text{BL}} I_{t-1}^L + w_{\text{BC}} I_{t-1}^C) + \theta_1^B e_{t-1}^B + \theta_2^B e_{t-2}^B \end{aligned} \tag{10}$$

with $w_{jj'} = (\frac{1}{1+d_{jj'}})^2$.

Tables 3 and 4 give the parameter estimates, standard errors and significance values along with the GOF values. Model fitting (in-sample) results versus the observed counts for the three cities are plotted in Fig 5, while forecasted counts for 1966 (26 biweekly measles counts) and 95% HDR intervals are plotted in Fig 6.

From Table 3 we note that an increasing trend is detected in the disease counts for all cities. We also get significant seasonality (both half yearly and yearly) in Leicester and Birmingham, while only yearly seasonality is significant in Coventry. The covariate, live birth, has a significant effect on the measles counts of Leicester and Birmingham, while the baby boom effect is seen to be significant in Coventry and Birmingham. Transmission parameter (ϕ) modelling the effect of within and between city disease transmission is significant in all three cities, implying spatial coupling in measles counts. We see a unit increase in the weighted auto-regressive components, accounting for both local and spatially connected city effects, results in an increase in the odds of infection by a factor of 2.44, 2.36, and 2.49 for Coventry, Leicester, and Birmingham, respectively. The MA components (lags 1 and 2) are significant for all three cities. The best models for each region consistently exhibit lower goodness-of-fit statistics compared to alternative models, as reported in Table 4. The reported MAE values for the 26 weeks forecasting scenario is 27.16, 28.16 and 54.56 for Coventry, Leicester and Birmingham, respectively.

5 Conclusions

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 GL-ARMA model, within an epidemiological framework. The model selection process involved evaluating multiple candidate models to identify the one that performed the best. In addition to evaluating various goodness-of-fit statistics, we performed a residual analysis to ensure the adequacy and validity of the model. All model calculations were performed using the statistical software R.

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. 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.

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

This paper benefited significantly from feedback from Niket Thakkar and Matthew J. Ferrari. Funding for this study was provided by the Gates Foundation (NV-044445).

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**	27.34489	27.34489	27.34489
	RMSE	55.32937	55.32937	55.32937
S2C2	**MAE**	22.09476	22.09476	25.5314
	RMSE	70.79094	70.79094	82.92279
S3C1	**MAE**	1.013975	5.118036	5.103921
	RMSE	1.05589	13.22069	13.18252
S3C2	**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

Model	Seasonality		AR Terms		Covariates		Linear Trend	MA Terms
	Annual	Biennial	I^{j}_{t-d}	$I^{j'}_{t-d}$	B_t	M_t		
1/7	Yes	No	1	1	Yes	Yes	Yes	No/Yes
2/8	Yes	No	1	1	Yes	Yes	Yes	No/Yes
3/9	Yes	Yes	1	1	Yes	Yes	Yes	No/Yes
4/10	Yes	Yes	1	1	Yes	Yes	Yes	No/Yes
5/11	Yes	Yes	1,2	1	Yes	Yes	Yes	No/Yes
6/12	Yes	Yes	1,2	1	Yes	Yes	Yes	No/Yes

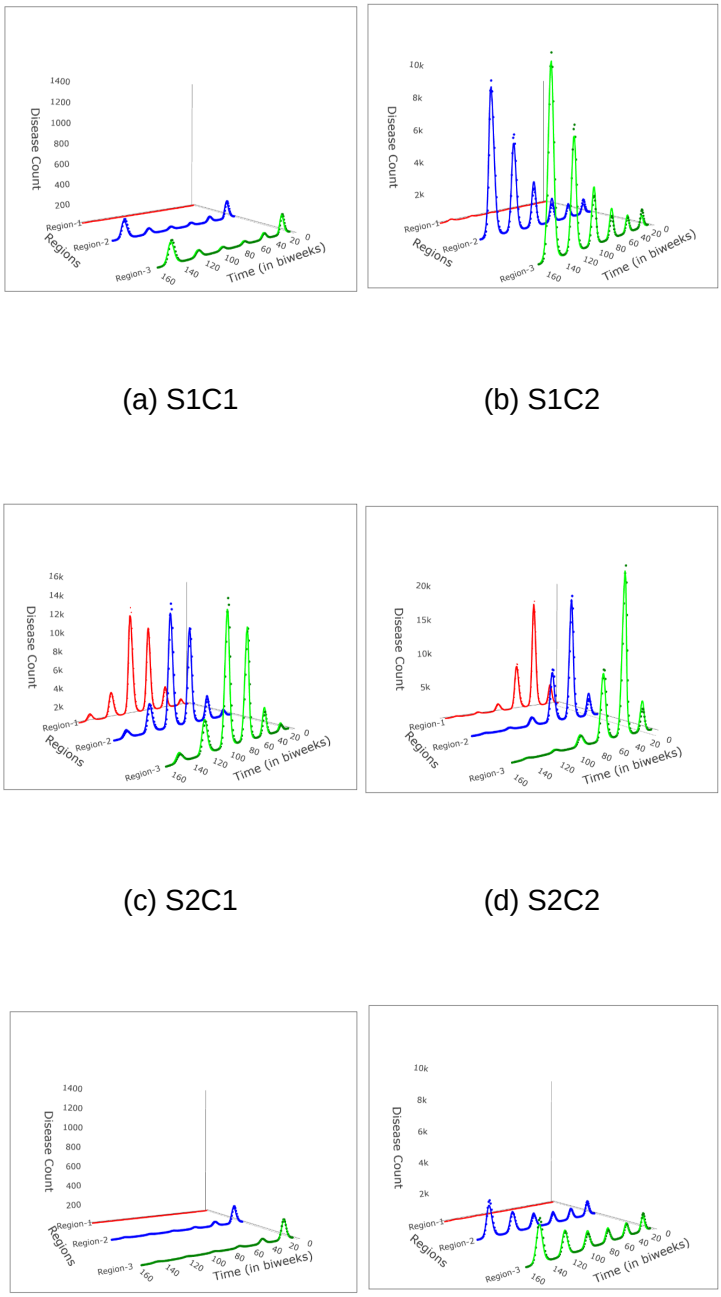
Table 2: First set of models used to fit the measles counts in the jth city, $j = 1, 2, 3$ corresponding to the data in Section 4

Coventry			
Parameter	**Estimate**	**Std. Error**	**p-value**
β_0^C	-12.0074	0.03444	$< .001$
β_1^C	-0.0053	0.00696	0.446
β_3^C	0.00062	0.00004	$< .001$
β_4^C	0.00014	0.00016	0.383
β_5^C	0.14306	0.01313	$< .001$
ϕ_1^C	0.89055	0.00388	$< .001$
θ_1^C	0.00052	0.00003	$< .001$
θ_2^C	0.00023	0.00003	$< .001$
Leicester			
β_0^L	-11.4601	0.06598	$< .001$
β_1^L	0.18352	0.00682	$< .001$
β_3^L	0.00039	0.00004	$< .001$
β_4^L	-0.00181	0.00028	$< .001$
β_5^L	-0.01946	0.01296	0.133
ϕ_1^L	0.86039	0.00435	$< .001$
θ_1^L	0.00065	0.00002	$< .001$
θ_2^L	0.00031	0.00003	$< .001$
Birmingham			
β_0^B	-13.1573	0.0305	$< .001$
β_1^B	0.24505	0.0042	$< .001$
β_3^B	0.00059	0.00002	$< .001$
β_4^B	-0.00039	0.00003	$< .001$
β_5^B	0.1193	0.00683	$< .001$
ϕ_1^B	0.91615	0.00251	$< .001$
θ_1^B	0.0002	0.00001	$< .001$
θ_2^B	0.00008	0.00001	$< .001$

Table 3: Parameter estimates, standard errors and p-values for of the models fitted to the measles counts from Coventry, Leicester, and Birmingham, respectively corresponding to the data in Section 4

Coventry			
MSE$_{PR}$	**MSE$_{WR}$**	**AIC**	**D**
17.99	0.76	11740.73	94.58
Leicester			
18.49	0.52	12310.45	97.74
Birmingham			
29.06	0.18	19418.99	125.64

Table 4: Goodness of fit statistics for the three cities, corresponding to the data set in Section 4



(a) S1C1 (b) S1C2

(c) S2C1 (d) S2C2

(e) S3C1 (f) S3C2

Figure 1: Comparison of measles incidences in the three regions for the three different simulation scenarios (S1-S3) and cases (C1-C2). In each plot, (•), (•), (•) represent simulated counts of region 1, 2, and 3 respectively, and (—), (—), and (—) represent the predicted counts.

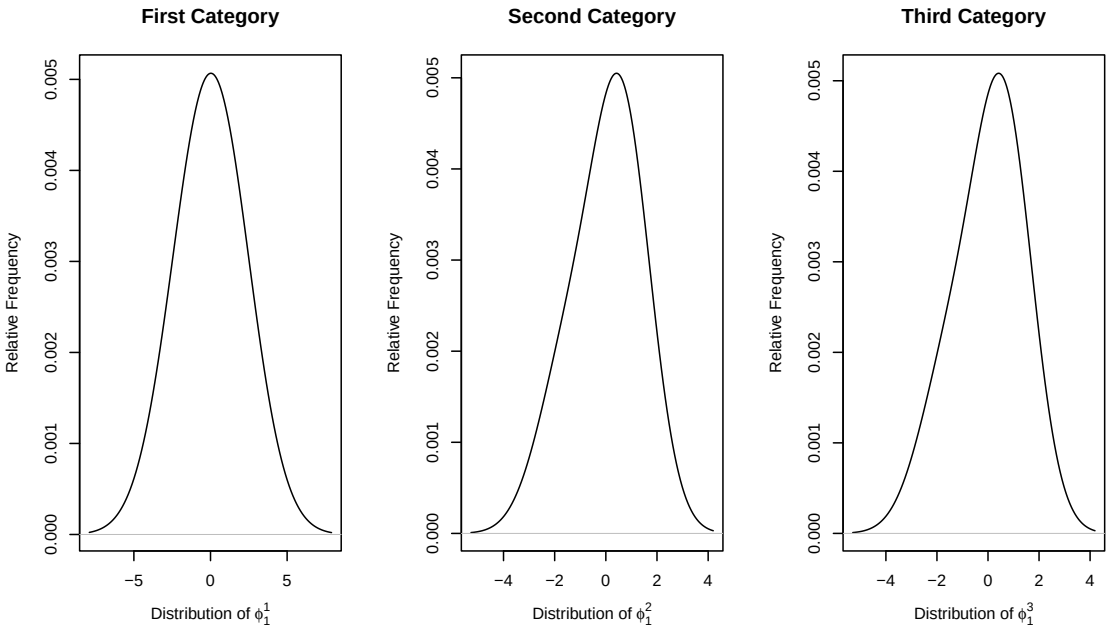


Figure 2: Distribution of standardized ϕ_i^j, results are averaged over simulation runs

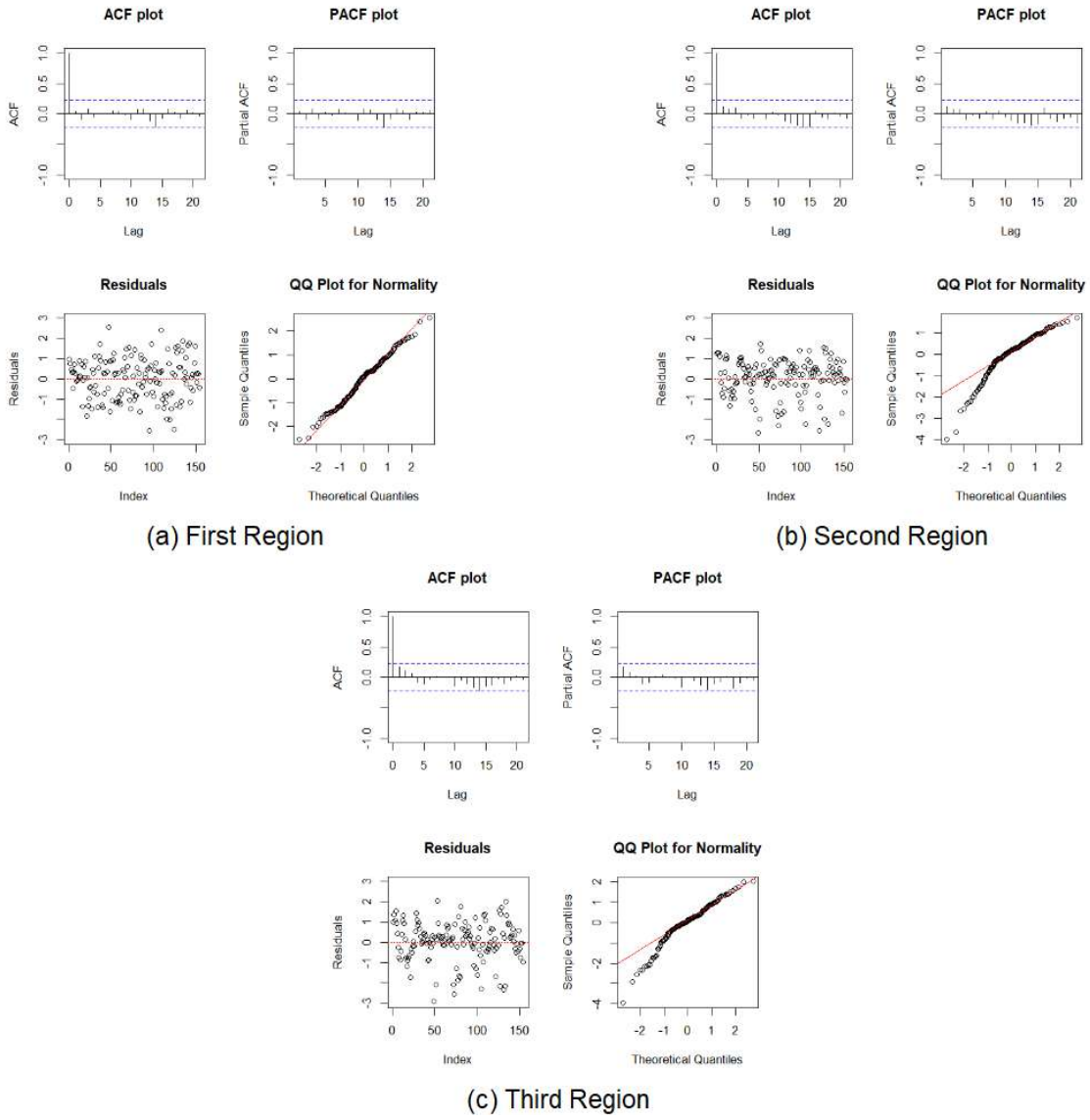


Figure 3: Randomized quantile residual (RQR) plots for three regions in simulation scenario 3 Case 1 as discussed in Section 3

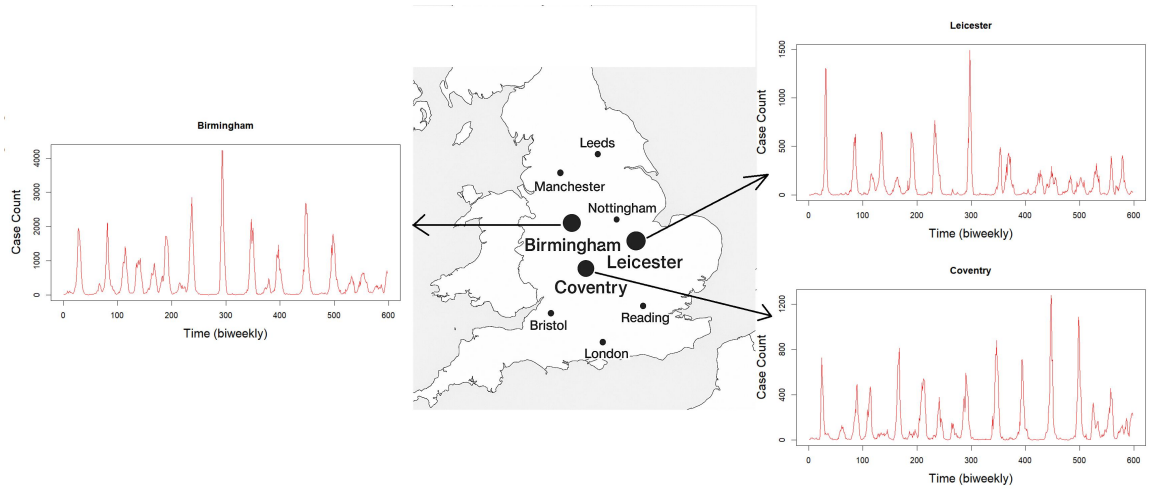


Figure 4: Biweekly measles case reports in the three UK cities from 1944-1966 and a map showing locations of the three cities

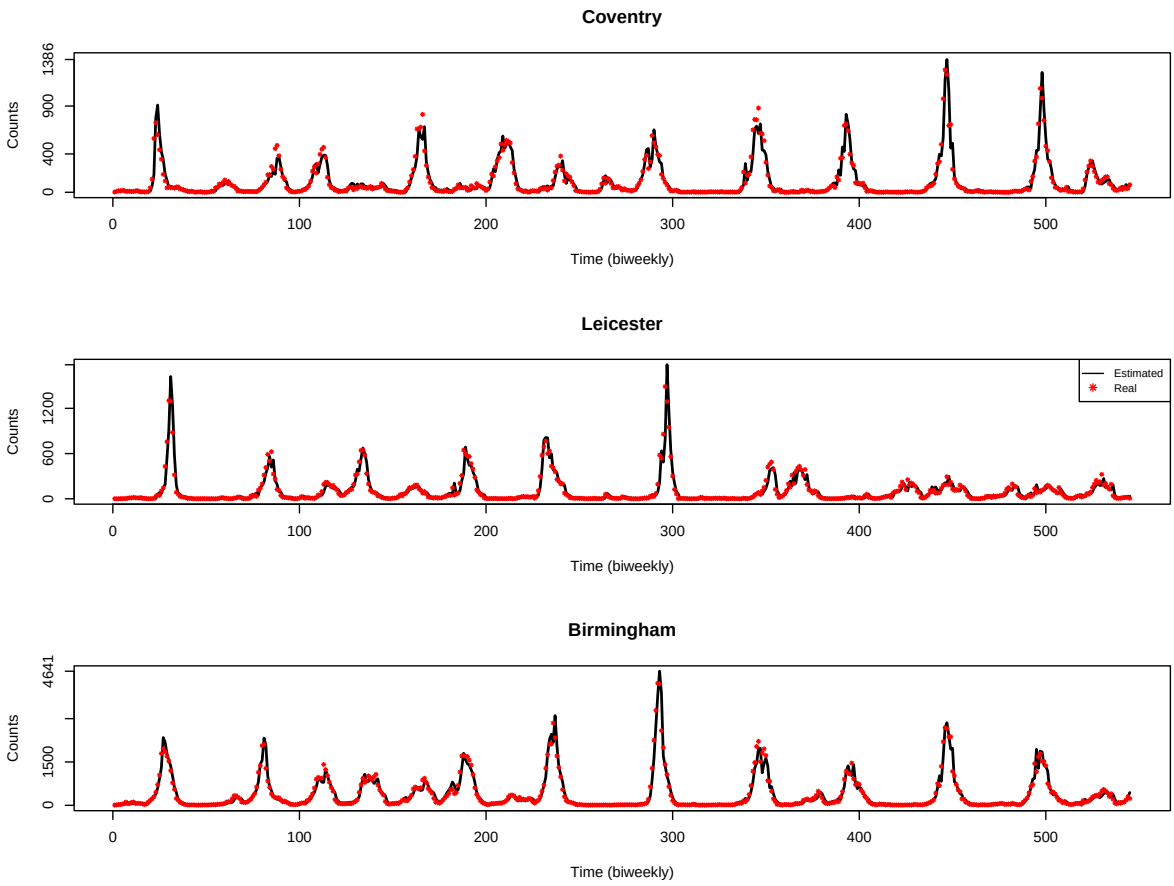


Figure 5: Observed biweekly measles counts vs fitted values for the three UK cities from 1944-65.

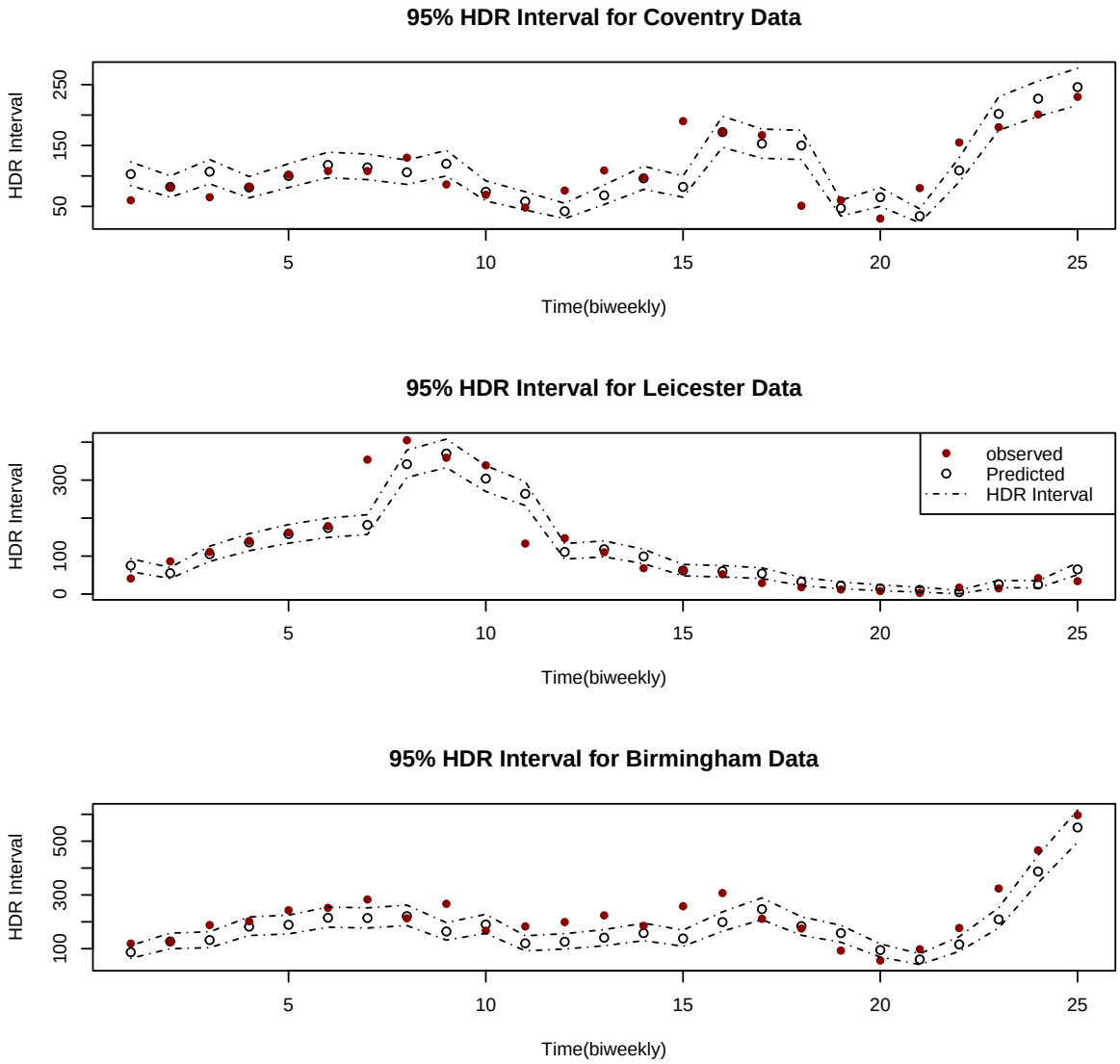


Figure 6: Forecasted measles counts and the 95% highest density regions from 1965–1966 for the UK three cities.

References

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

Anselin, L. (2022). Spatial econometrics. *Handbook of spatial analysis in the social sciences*, pages 101–122.

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.

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.

Ferrari, M. J., Bjørnstad, O. N., and Dobson, A. P. (2005). Estimation and inference of r0 of an infectious pathogen by a removal method. *Mathematical biosciences*, 198(1):14–26.

Fine, P. E. and Clarkson, J. A. (1982). Measles in england and wales—i: an analysis of factors underlying seasonal patterns. *International journal of epidemiology*, 11(1):5–14.

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.

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

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.

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.

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.

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.

Vynnycky, E. and White, R. (2010). *An introduction to infectious disease modelling*. OUP oxford.

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.