A Bayesian latent process spatiotemporal regression model for areal count data

Abstract

Model-based approaches for the analysis of areal count data are commonplace in spatiotemporal analysis. In Bayesian hierarchical models, a latent process is incorporated in the mean function to account for dependence in space and time. Typically, the latent process is modelled using a conditional autoregressive (CAR) prior. The aim of this paper is to offer an alternative approach to CAR-based priors for modelling the latent process. The proposed approach is based on a spatiotemporal generalization of a latent process Poisson regression model developed in a time series setting. Spatiotemporal dependence in the autoregressive model for the latent process is modelled through its transition matrix, with a structured covariance matrix specified for its error term. The proposed model and its parameterizations are fitted in a Bayesian framework implemented via MCMC techniques. Our findings based on real-life examples show that the proposed approach is at least as effective as CAR-based models.

Keywords: Autoregressive latent process; Bayesian inference; Conditional autoregressive prior; Markov Chain Monte Carlo; Spatiotemporal areal count data

1 Introduction

The development of Bayesian hierarchical models for analyzing spatiotemporal areal data, driven by an upsurge in the availability of data (Lee et al., 2016), has been an increasingly active area in recent times. Spatiotemporal areal count data result mostly from the need to collect and monitor administrative-level information relating to health, socioeconomic or demographic outcomes. Such spatially aggregated data play an important role in providing an evidence base for public health policies and interventions as they represent broader geographical scales and are not affected by confidentiality issues often encountered with individual-level data. The analysis of areal data focuses mainly on accounting for the underlying spatial and temporal autocorrelation, identifying local and global changes in spatial patterns and trends over time and quantifying the association between the outcome and covariate factors.

Several spatiotemporal models for analyzing areal data have been developed in the statistical literature (see Bernardinelli et al., 1995; Waller et al., 1997; Knorr-Held & Besag, 1998 and Xia & Carlin, 1998, for some of the earliest approaches). These models are widely applied in disease mapping studies for modelling different epidemiological outcomes. Their spatial analogues are, however, more widely studied and applied and most of the available spatiotemporal models are in fact direct extensions of well-known spatial models. In the spatial context, the outcome variables $\boldsymbol{Y} = (Y_1, \ldots, Y_n)$ corresponding to a set of n contiguous, non-overlapping areal units are regressed on area-level covariate information, $\mathbf{z} = (\mathbf{z}_1', \ldots, \mathbf{z}_n')$. Customarily, the expected counts, $E_1, \ldots, E_n$, calculated using internal or external standardization are included in the mean function of the model to account for sampling variability and demographic differences between the areas (Lee, 2011; Banerjee et al., 2015). In general, the spatial models take the form:

$$Y_i | X_i, \mathbf{z}_i \sim \text{Poisson}(E_i \exp(\mathbf{z}_i'\boldsymbol{\beta} + X_i)), \quad i = 1, \ldots, n, \tag{1}$$

where $\boldsymbol{\beta}$ is a $p-$dimensional set of regression coefficients, $\mu_i = \exp(\mathbf{z}_i'\boldsymbol{\beta} + X_i)$ is the relative risk for area i, and the X_i's constitute the latent process (also referred to as area-specific random effects - we use the terms 'latent process' and 'random effects' interchangeably) used to model residual spatial autocorrelation in the data. In (1), $\boldsymbol{X} = (X_1, \ldots, X_n)$ is typically represented by a conditional autoregressive (CAR) prior. Besag (1974) proposed the intrinsic CAR prior in which $\boldsymbol{X}$ is modelled as having a zero-mean multivariate normal distribution with a covariance matrix usually specified

as the generalized inverse of the singular precision matrix $\tau(\boldsymbol{D}-\boldsymbol{W})$, where $\boldsymbol{D}$ and $\boldsymbol{W}$ are $n \times n$ matrices that characterize the neighbourhood structure of the data and τ^{-1} is a conditional variance parameter. Specifically, $\boldsymbol{W}$ is a binary neighbourhood matrix (see Section 2.1) and $\boldsymbol{D}$ is a diagonal matrix whose ith entry is equal to the number of neighbours for the corresponding areal unit. A very popular CAR prior used in disease mapping studies is the Besag-York-Mollie (BYM) model (Besag et al., 1991) which is an extension of the intrinsic CAR model that includes an additional area-specific latent variable modelled as exchangeable or an unstructured term. Other commonly used CAR priors include those proposed by Leroux et al. (2000), which includes an autoregressive parameter that measures the strength of spatial autocorrelation in the latent process; see Lee (2013) for a review of these models.

Direct spatiotemporal extensions of spatial CAR models often proceed by decomposing the latent process in (1) into a sum of spatial, temporal and spatiotemporal interaction terms (see, e.g., Blangiardo & Cameletti, 2015, Chapter 7). With the variables now indexed by space ($i = 1, \ldots, n$) and time ($t = 1, \ldots, T$), the latent process has a general form which can be expressed as: $X_{t,i} = a_i + b_t + c_{t,i}$, where the terms are indexed according to the components they represent. For example, the specification in Waller et al. (1997) is a direct extension of the BYM model to the spatiotemporal case where spatial dependence is modelled separately for each time point so that $a_i = b_t = 0$ and $c_{t,i}$ is a sum of spatial (CAR) and independent random effects at time t. In many of these spatiotemporal models (see Anderson & Ryan, 2017, for a review), separate latent variables/random effects are used to capture spatial and temporal dependence. This implicitly suggests that dependence in space and time is separable and may, thus, be unrealistic for many real-world processes. A few of the proposed models (e.g. Lee & Lawson, 2014; Rushworth et al., 2014) capture spatiotemporal dependence using a single latent process but spatial autocorrelation in these models is characterized using CAR priors - a consequence of the extension from the spatial to the spatiotemporal setting. In particular, in the model proposed by Rushworth et al. (2014) which is similar in structure to the models developed here, the latent process characterizing spatiotemporal autocorrelation is modelled as a multivariate first-order autoregressive process with a single temporal autoregressive parameter and spatial autocorrelation induced by a precision matrix based on the CAR prior proposed by Leroux et al. (2000). In general, Bayesian estimation is the state-of-the-art in these spatiotemporal models and computational approaches such as Markov Chain Monte Carlo (MCMC) (Lee et al., 2016) and Integrated Nested Laplace Approximation (INLA) (Blangiardo

& Cameletti, 2015) are commonly used.

In this paper, we introduce an alternative, flexible class of models for analyzing spatiotemporal areal count data as a generalization of the parameter-driven model of Zeger (1988). Zeger's model is an extension of the standard Poisson regression model for independent outcomes to the time series setting. Using the same notation as in (1), the model can be expressed as:

$$Y_t|X_t, \mathbf{z}_t, \sim \text{Poisson}(\exp(\mathbf{z}_t^T\boldsymbol{\beta} + X_t)), \tag{2}$$

where $\{X_t\}$ is a first-order autoregressive latent process (e.g. $X_t|X_{t-1} \sim N(\alpha X_{t-1}, \tau^{-1})$) used to account for serial correlation and overdispersion in the observed data, $\{Y_t\}$. The term 'parameter-driven' arises from the use of a latent process to account for dependence in time as against the introduction of lagged values of the outcome variable in the conditional mean of the model. With a suitable moment restriction on X_t such as $E(\exp(X_t)) = 1$, the terms in $\exp(\mathbf{z}_t^T\boldsymbol{\beta})$ can be interpreted just as in a standard Poisson regression model. Other assumptions of the model, modifications and implementation details can be found in Zeger (1988) and Utazi (2017). McShane et al. (1997) generalized Zeger's model to the spatial setting to analyze point-referenced data, modelling spatial autocorrelation in the latent process using a parametric correlation function from the Matérn family (Banerjee et al., 2015, Chapter 2). In contrast, our spatiotemporal extension applies to areal data. In the proposed model, as in parameter-driven models, a latent process is used to account for spatiotemporal dependence in the observed counts. We exploit the natural ordering in time to specify a first-order vector autoregression for the latent process, with structured transition matrices and a constrained error covariance structure to characterize different levels of spatiotemporal dependence and address the problem of high dimensionality. Our modelling efforts are geared towards characterising spatial dependence using the autoregressive structure of the latent process as against using its covariance structure or second-order moments ubiquitous in the literature. The structures imposed on the transition matrices are mostly based on notions of spatial dependence defined through an adjacency matrix or a monotonically decreasing distance function. The proposed models are fully implemented in a Bayesian framework. We use MCMC methods to estimate the parameters of the model including the latent process via data augmentation. A key contribution is the development of a Metropolis-Hastings (M-H) step relative to an independence update for estimating the missing values of the latent

process.

The rest of this paper is organized as follows. In Section 2, we present the general form of the proposed model. The parameterizations of the general model are explored in Section 2.1. Section 3 discusses Bayesian estimation of the models. In Section 4, the methodology is illustrated using Ohio lung cancer mortality and Georgia low birth weight datasets and compared with CAR-based models.

2 Model Development

We begin by writing down the general form of the proposed model. The model assumes that conditional on the latent process $X_{t,i}$ and a $p-$dimensional covariate vector $\mathbf{z}_{t,i}$, the observed counts $Y_{t,i}$, $(i = 1, \ldots, n; \quad t = 1, \ldots, T)$ are distributed as Poisson with mean $E_{t,i}\mu_{t,i}$ denoted by

$$Y_{t,i}|X_{t,i}, \mathbf{z}_{t,i} \sim \text{Poisson}(E_{t,i} \exp(\mathbf{z}'_{t,i}\boldsymbol{\beta} + X_{t,i})), \tag{3}$$

where $E_{t,i}$ is the expected count for area i at time t (as defined previously) and $\boldsymbol{\beta} \in \mathbb{R}^p$ are uniform regression coefficients for all areas. In the model, $\mu_{t,i} = \exp(\mathbf{z}'_{t,i}\boldsymbol{\beta}_i + X_{t,i})$ is the relative risk for area i at time t - a quantity of significant epidemiological importance. A crude measure of disease risk based on the expected counts, $E_{t,i}$, is the standardized incidence ratio (SIR) given by SIR $= Y_{t,i}/E_{t,i}$. The SIR does not account for spatial autocorrelation and often exhibits large variability (Wakefield, 2007). Hence, model-based approaches for estimating disease risk are preferable.

In the second level of the hierarchical model in (3), the latent process, $\{X_{t,i}\}$, is assumed to follow a first-order autoregressive model expressed as

$$X_{t,i} = \textstyle\sum_{j=1}^{n} a_{ij} X_{t-1,j} + \xi_t + \epsilon_{t,i}, \tag{4}$$

where $\xi_t \sim N(0, \tau_\xi^{-1})$, a temporal random effect, and $\epsilon_{t,i} \sim N(0, \tau^{-1})$, an error term, are assumed to be independent. Let $\boldsymbol{A} = \{a_{ij}\}$ be an $n \times n$ transition matrix or matrix of autoregressive coefficients, $\mathbf{1}$ an $n \times 1$ vector of 1's, and $\boldsymbol{\epsilon}_t = (\epsilon_{t,1}, \ldots, \epsilon_{t,n})'$ iid Gaussian (non-Gaussian errors are possible) with a positive definite covariance matrix, $\Sigma = \tau^{-1} I_n$. In a compact vector notation, (4) can be written as

$$\boldsymbol{X}_t = \boldsymbol{A}\boldsymbol{X}_{t-1} + \xi_t \mathbf{1} + \boldsymbol{\epsilon}_t, \quad \boldsymbol{\epsilon}_t \sim \text{N}_n(\mathbf{0}, \Sigma). \tag{5}$$

In (5), it is assumed that $\boldsymbol{X}_1 \sim N(\mathbf{0}, \Sigma_X)$, where Σ_X is the covariance matrix of $\boldsymbol{X}_t$ which can be expressed using the 'vec' operator as $\text{vec}(\Sigma_X) = (I_{n^2} - \boldsymbol{A} \otimes \boldsymbol{A})^{-1}\text{vec}(\tau_\xi^{-1}\mathbf{1}\mathbf{1}' + \Sigma)$, where '$\otimes$' denotes the Kronecker product (see supplementary material). To ensure that $\{\boldsymbol{X}_t\}$ is second-order stationary in time, we assume that the roots of $\det(I_n - \boldsymbol{A}z) = 0$ lie outside the unit circle, or equivalently, the eigenvalues of $\boldsymbol{A}$ have moduli less than one. This implies that $\xi_t\mathbf{1} + \boldsymbol{\epsilon}_t$ is independent of $\boldsymbol{X}_k$ for any $k < t$ (see Lütkepohl, 2005). The independent error terms, $\xi_t\mathbf{1}$ and $\boldsymbol{\epsilon}_t$, can be viewed as an additive decomposition of some general zero-mean error term, say $\boldsymbol{\psi}_t$, with covariance matrix, $\Sigma_\psi = \tau_\xi^{-1}\mathbf{1}\mathbf{1}' + \Sigma$. While the temporal random effect ξ_t represents time-specific variability common to all the areal units, the Gaussian zero-mean error term $\boldsymbol{\epsilon}_t$, as modelled by Σ, captures additional homogenous spatial variability. Thus, the combination of $\xi_t\mathbf{1}$ and $\boldsymbol{\epsilon}_t$ results in a structured conditional covariance matrix for $\boldsymbol{X}_t$. A further discussion on this choice of covariance matrix for $\boldsymbol{\psi}_t$ is given in Section 2.1. Note that a more general structure could be easily assumed for Σ_ψ if desired by replacing the vector of 1's in (5) with a different deterministic vector, $\boldsymbol{C} = (c_1, \ldots, c_n)$, which could represent area-specific weights applied to ξ_t. For example, these weights could be a function of the number of neighbours an area has. In addition, the covariance matrix Σ could be generalized to include area-specific variance parameters.

The latent process $\{\boldsymbol{X}_t\}$ captures both overdispersion and spatiotemporal dependence in the observed counts, as demonstrated by the moment properties of the model provided in the supplementary material, and can also be viewed as a surrogate for unmeasured spatially varying covariate factors. Spatiotemporal dependence in $\{\boldsymbol{X}_t\}$ is clearly expressed by the dynamical structures in (4) and (5), both of which show that the current value of the process for a given areal unit $X_{t,i}$ depends not only on its own past value but also on the past values of the process at other areal units. This dependence structure is, however, modified by the $\boldsymbol{A}$ matrix, with the contribution of the $(t-1)$th value of the jth areal unit to the tth value of the ith area scaled by a_{ij}. Note that (5) can be considered as a special case of the STAR(1) model (see Pfeifer & Deutsch, 1980) when all forms of spatial ordering are eliminated from the latter. Note also that higher order temporal lags could be considered if desirable.

2.1 Parameterizing the latent spatiotemporal process, $\{\boldsymbol{X}_t\}$

In its general form, the latent process model in equation (5) is often said to be 'saturatedly parameterized' (Davis et al., 2012). The matrix of autoregressive coefficients,

$\boldsymbol{A}$, has n^2 parameters to be estimated. In practice, when n is small, the model can be estimated with less uncertainty. However, with large (and even moderate) n as is often the case with areal data, the number of parameters can be prohibitively large and obtaining reliable estimates for these parameters and model interpretation can be challenging (see Cressie & Wikle, 2011, p.384). This is the so-called problem of 'curse of dimensionality'.

Parameterizing space-time models of the type given in model (5) typically involves trade-offs between the compositions/structures given to $\boldsymbol{A}$ and Σ_ψ - the covariance matrix for the general error term, $\boldsymbol{\psi}_t$ (Wikle et al., 1998). However, the $\boldsymbol{A}$ matrix is often considered 'the most critical part' (Xu & Wikle, 2007) of (5) owing to its role in capturing dynamical spatiotemporal interactions. Hence, we model spatial dependence using $\boldsymbol{A}$. We expect that if spatial interactions are adequately captured using this transition matrix, then the structure of Σ_ψ is greatly simplified. Hence the specification: $\Sigma_\psi = \tau_\xi^{-1}\mathbf{1}\mathbf{1}' + \Sigma$, with homogenous within- and between-area covariances (see Wikle *et al.* for a related discussion).

Known approaches used in reducing the dimensionality of the $\boldsymbol{A}$ matrix in (5) include making it sparse through the use of spatial neighborhood assumptions such as $[\boldsymbol{A}_{ij}] = 0$ except when the centroids of areas i and j are within a given distance of each other (Cressie & Wikle, 2011). Here, different parameterizations of $\boldsymbol{A}$ can be introduced with some flexibility subject to the stationarity constraint. Some ideas explored include the possibility of spatial independence, spatial homogeneity/heterogeneity assumptions which entail the use of a uniform set of parameters/ area-specific parameters to describe spatial dependence and the use of distances and spatial adjacency for characterizing spatial dependence. These will help refine spatial dependence in $\boldsymbol{A}$ and consequently reduce the number of parameters to be estimated. The specification in (3) is the same in all the proposed parameterizations and is omitted for conciseness. Also, we use the words 'parameterization' and 'model' interchangeably.

The first parameterization that we consider is that which allows for spatial independence in the $\boldsymbol{A}$ matrix. This parameterization is useful for analyzing processes with weak spatial dependence where prior knowledge could suggest this, as well as for the purpose of comparison with more complex models to preserve parsimony. This parameterization assumes that $\boldsymbol{A} = \text{diag}(\boldsymbol{\omega}) = \text{diag}(\omega_1, \ldots, \omega_n)$ and is given by

$$\text{Model 1}: \quad \boldsymbol{X}_t = \text{diag}(\boldsymbol{\omega})\boldsymbol{X}_{t-1} + \xi_t\mathbf{1} + \boldsymbol{\epsilon}_t. \tag{6}$$

In (6), temporal dependence is captured using a separate autoregressive parameter, ω_i, for each area. Similar parameterizations (e.g. $\omega_i = \omega \quad \forall i$) are often used in models that focus on modeling spatial dependence using the error covariance matrix, Σ_ψ (e.g. Rushworth et al., 2014). We note that although a spatial independence structure has been imposed on $\boldsymbol{A}$, the temporal random effect ξ_t ensures that Model 1 cannot be separated into individual time series models for each area.

To model spatial dependence in $\boldsymbol{A}$, we first consider a refinement of the fully parameterized $\boldsymbol{A}$ in which the present structure allows the current value of $X_{t,i}$ to depend not only on its past values but also on the past values of every other area. This dependence can be much improved in the spatial context since neighbouring areas are expected to be more similar and have a stronger effect on each other than areas that are further apart. Thus, spatial neighbourhood structures as defined by a binary neighbourhood matrix, $\boldsymbol{W}_{n\times n}$ ($\boldsymbol{W}_{ij} = 1$ if areas i and j are neighbours or if $i = j$; and 0 otherwise) can be utilized to make $\boldsymbol{A}$ sparse. The second model proposed is given by

$$\text{Model 2}: \quad \boldsymbol{X}_t = \boldsymbol{A}_{I(W_{ij}=1)}\boldsymbol{X}_{t-1} + \xi_t\mathbf{1} + \boldsymbol{\epsilon}_t; \tag{7}$$

where $I(.)$ is an indicator function and $\boldsymbol{A}$ has the specification in (5). This model results in a sparse transition matrix which could be robustly estimated with small n and moderate or large T. However, with even a moderate n, the number of parameters to be estimated may be large. Against this backdrop, the next model that we consider replaces the remaining heterogenous parameters in $\boldsymbol{A}$ in (7) with a single autoregressive parameter. This can be expressed as

$$\text{Model 3}: \quad \boldsymbol{X}_t = (\alpha\boldsymbol{W})\boldsymbol{X}_{t-1} + \xi_t\mathbf{1} + \boldsymbol{\epsilon}_t; \tag{8}$$

In addition to defining spatial proximity as given in $\boldsymbol{W}$, distance-based matrices have also been used to characterize spatial dependence in covariance matrices of latent process priors for areal data (Xia et al., 1997; Pascutto et al., 2000). Thus, we define a distance-based weighting matrix given by: $\boldsymbol{D}_\phi = \exp(-\phi d)$, where d is a matrix of Euclidean distances between the centroids of the areal units and ϕ ($\phi > 0$) is a decay parameter that determines the rate at which spatial dependence decreases with increasing distance. In principle, any other suitable monotonically decreasing function can be used in $\boldsymbol{D}_\phi$. It is possible to include a sense of direction in the weighting matrix (see Stoffer, 1986) in which case d_{ij} may not necessarily be equal to d_{ji}, but

we consider Euclidean distances only. Based on this proposition, Model 4, given by

$$\text{Model 4}: \quad \boldsymbol{X}_t = (\alpha \boldsymbol{D}_\phi)\boldsymbol{X}_{t-1} + \xi_t \mathbf{1} + \boldsymbol{\epsilon}_t, \tag{9}$$

results from substituting $\boldsymbol{D}_\phi$ for $\boldsymbol{W}$ in (8). This model is somewhat a generalization of Model 1 since independence can be attained as $\phi \to \infty$.

To explore further the characterization of spatial dependence without utilizing too many parameters, we extend Models 3 and 4 using area-specific parameters as in Model 1. Two sets of parameters: $\boldsymbol{\gamma}_{n\times 1} = (\gamma_i; i = 1, \ldots, n)$ and $\boldsymbol{\delta}_{n\times 1} = (\delta_j; j = 1, \ldots, n)$ are used to generate a non-symmetric matrix $\boldsymbol{\gamma}\boldsymbol{\delta}'$ which is then modified using the spatial matrices to form a transition matrix for the latent process. With the neighbourhood matrix, $\boldsymbol{W}$, we obtain

$$\text{Model 5}: \quad \boldsymbol{X}_t = (\boldsymbol{\gamma}\boldsymbol{\delta}' \circ \boldsymbol{W})\boldsymbol{X}_{t-1} + \xi_t \mathbf{1} + \boldsymbol{\epsilon}_t; \tag{10}$$

whereas using the weighting matrix, $\boldsymbol{D}_\phi$, we have

$$\text{Model 6}: \quad \boldsymbol{X}_t = (\boldsymbol{\gamma}\boldsymbol{\delta}' \circ \boldsymbol{D}_\phi)\boldsymbol{X}_{t-1} + \xi_t \mathbf{1} + \boldsymbol{\epsilon}_t. \tag{11}$$

These models in (10) and (11) introduce a higher level of complexity in the transition matrix using more parameters but a great reduction in the dimensionality of $\boldsymbol{A}$ in the full model in (5) is achieved, from n^2 to $\leq 2n+1$ parameters. The asymmetrical form of the transition matrices implies that areas i and j can influence each other differently. This relationship is better seen by examining the noncompact forms of the models (e.g., $X_{t,i} = \gamma_i \sum_{j=1}^{n} \delta_j \exp(-\phi d_{ij}) X_{t-1,j} + \xi_t + \epsilon_{t,i}$) as in equation (4). More explicitly, γ_i scales the dependence of the current value of the ith area on all past values (with the latter pre-modified by the δ_j's and the entries of the spatial matrices), while $\delta_j (j = i)$ partly determines the influence of its past value on all current values. There are non-identifiability issues with the joint estimation of $\boldsymbol{\gamma}$ and $\boldsymbol{\delta}$, requiring a constraint to be placed on either of these parameters. For example, we could fix the δ_j's or constrain the γ_i's to sum to a known constant. As in model (5), it is necessary that Models 1-6 represent stationary processes. For Model 1, this requires that $|\omega_i| < 1 \; \forall \; i$. For other models, we assume that the eigenvalues of the corresponding transition matrices have moduli less than unity as in (5).

3 Bayesian analysis

In this section, we present a Bayesian approach implemented using MCMC methods for estimating the parameters of the proposed models. Let $\boldsymbol{Y}_{(T\times n)}$ denote the data matrix of the observed counts, $\boldsymbol{X}_{(T\times n)}$ the values of the latent process and $\boldsymbol{Z}_{(T\times np)}$, a rectangular array of covariate data. Also, let $\boldsymbol{\xi} = (\xi_1, \ldots, \xi_T)$ and $\boldsymbol{\theta}$ denote the parameters of the general model in (3). The likelihood function of the model is not available in closed form, but only as a high-dimensional integral over the latent process, $\boldsymbol{X}$ (and $\boldsymbol{\xi}$). Therefore, we work with the augmented data likelihood function (Tanner & Wong, 1987) which facilitates the numerical evaluation of the integral in an MCMC framework. This is given by

$$\begin{aligned}
\pi(\boldsymbol{Y}, \boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{\xi}|\boldsymbol{\theta}) =& p_1(\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{\theta}) \times p_2(\boldsymbol{X}|\boldsymbol{\xi}, \boldsymbol{\theta}) \times p_3(\boldsymbol{\xi}|\boldsymbol{\theta}) \\
&\propto \prod_{i=1}^{n}\prod_{t=1}^{T}\left\{\exp\left(X_{t,i}Y_{t,i} + \mathbf{z}'_{t,i}\boldsymbol{\beta}Y_{t,i} - E_{t,i}e^{\mathbf{z}'_{t,i}\boldsymbol{\beta}+X_{t,i}}\right)\right\} \\
&\quad \times \prod_{i=1}^{n}\prod_{t=2}^{T}\left\{\tau^{1/2}\exp\left(-\frac{\tau}{2}\left[X_{t,i} - \textstyle\sum_{j=1}^{n} a_{ij}X_{t-1,j} - \xi_t\right]^2\right)\right\} \\
&\quad \times p(\boldsymbol{X}_1) \times \prod_{t=1}^{T}\tau_\xi^{1/2}\exp\left(-\frac{\tau_\xi}{2}\xi_t^2\right), \qquad (12)
\end{aligned}$$

where $p(\boldsymbol{X}_1)$ is as defined in (5). Note that the likelihood functions of models (6) - (11) can be obtained from (12) by replacing a_{ij} with the corresponding terms. For example, this would be $\alpha\exp(-\phi d_{ij})$ for model (9).

Assuming prior independence, we consider proper priors for the parameters common to all the models which, in some cases, are also conjugate priors. For τ, τ_ξ and β_j $(j = 0, \ldots, p-1)$ we choose Gamma(a_τ, b_τ), Gamma(a_ξ, b_ξ) and N(u_β, v_β^{-1}) respectively. The prior for the a_{ij}'s is $a_{ij} \sim N(u_a, v_a^{-1})$, $i, j = 1, \ldots, n$. For the parameters of the parameterizations of the general model, we use the following prior distributions: $\omega_i \sim N(u_\omega, v_\omega^{-1})$, $\alpha \sim N(u_\alpha, v_\alpha^{-1})$, $\gamma_i \sim N(u_\gamma, v_\gamma^{-1})$, $\delta_j \sim N(u_\delta, v_\delta^{-1})$ and $\phi \sim$ Gamma(a_ϕ, b_ϕ). We have chosen Gaussian priors for a_{ij}, ω, α, the γ_i's and the δ_j's which can be truncated to lie in $\mathbb{R}^+$ under the assumption of positive dependence between the areas.

With these choices of prior distributions, it is straightforward to write down the posterior distribution. The conditional posterior distributions of the parameters and the MCMC schemes to draw samples from them are given in the Appendix. We focus

here on estimating the missing values of the latent process, $\{\boldsymbol{X}_t\}$. Noting that the value of the process for areal unit i at time t, $X_{t,i}$, depends on $\boldsymbol{X}_{t-1}$ from the past, $\boldsymbol{X}_{t+1}$ into the future and the current values of the process for other areas $\boldsymbol{X}_{t,-i}$, its conditional posterior distribution is given by

$$\begin{aligned}\pi(X_{t,i}|\boldsymbol{X}_{t-1},\boldsymbol{X}_{t+1},\boldsymbol{X}_{t,-i},\boldsymbol{\theta},Y_{t,i},\boldsymbol{\xi}) \propto \\ \exp\left(-\frac{\tau}{2}\left[X_{t,i}-\textstyle\sum_{j=1}^{n} a_{ij}X_{t-1,j}-\xi_t\right]^2\right) \\ \times \exp\left(-\frac{\tau}{2}\textstyle\sum_{k=1}^{n}\left[X_{t+1,k}-a_{ki}X_{t,i}-\sum_{j\neq i,j=1}^{n} a_{kj}X_{t,j}-\xi_{t+1}\right]^2\right) \\ \times \exp\left(X_{t,i}Y_{t,i}-E_{t,i}e^{\mathbf{z}'_{t,i}\boldsymbol{\beta}+X_{t,i}}\right), \quad t=2,\ldots,n-1. \end{aligned} \tag{13}$$

Equation (13) shows how the interdependence among $X_{t,1}\ldots X_{t,n}$ can be exploited when estimating any given $X_{t,i}$. Clearly, (13) does not have a standard form if we do not wish to approximate any of the exponents and in the MCMC context, we cannot use a Gibbs move to obtain samples from it. Hence, we propose an updating scheme relative to independence M-H sampling procedure. Motivated by the Markovian structure of the latent process, the proposal density is given by

$$\begin{aligned}p(X_{t,i}|\boldsymbol{X}_{t-1},\boldsymbol{X}_{t+1},\boldsymbol{X}_{t,-i}),\boldsymbol{\theta},\boldsymbol{\xi}) \sim \\ N\left(\frac{\sum_{k=1}^{n} a_{ki}U_k(t)+\sum_{j=1}^{n} a_{ij}X_{t-1,j}+\xi_t}{1+\sum_{k=1}^{n} a_{ki}^2}, \frac{1}{\tau(1+\sum_{k=1}^{n} a_{ki}^2)}\right), \end{aligned} \tag{14}$$

where $U_k(t) = X_{t+1,k} - \sum_{\substack{j=1\\ j\neq i}}^{n} a_{kj}X_{t,j} - \xi_{t+1}$, and the corresponding acceptance probability is

$$\alpha_x(X_{t,i}\to X'_{t,i}) = \min\left\{1, \frac{\exp(X'_{t,i}Y_{t,i}-E_{t,i}e^{\mathbf{z}'_{t,i}\boldsymbol{\beta}+X'_{t,i}})}{\exp(X_{t,i}Y_{t,i}-E_{t,i}e^{\mathbf{z}'_{t,i}\boldsymbol{\beta}+X_{t,i}})}\right\}, \tag{15}$$

where $X'_{t,i}$ denotes the proposed value of $X_{t,i}$. Observe that both the proposal distribution in (14) and the acceptance probability in (15) were constructed from (13). Equation (15) reveals the role of the $Y_{t,i}$'s in providing feedback for the estimation of the $X_{t,i}$'s. The idea behind the choice of (14) is that if this feedback mechanism is obliterated by excluding the $Y_{t,i}$'s from the model, the acceptance probability in (15) becomes equal to unity and this M-H sampling procedure reduces to a Gibbs step in which samples are obtained from (14) to update $\{\boldsymbol{X}_t\}$. These proposals were investi-

gated and successfully used in Utazi (2014) and Utazi (2017) in both spatiotemporal and temporal settings. We also found that this updating scheme outperformed random walk M-H updates in pilot studies, attaining convergence quickly and yielding better parameter estimates.

The conditional posterior distributions of $X_{1,i}$ and $X_{T,i}$ are special cases. For convenience, we use the vector notation to write the proposal density of $\boldsymbol{X}_1$ as

$$p(\boldsymbol{X}_1|\boldsymbol{X}_2,\boldsymbol{\theta}) \sim \boldsymbol{N}\left[(\Sigma_X^{-1}+\boldsymbol{A}'\Sigma_\psi^{-1}\boldsymbol{A})^{-1}(\boldsymbol{X}_2'\Sigma_\psi^{-1}\boldsymbol{A})', (\Sigma_X^{-1}+\boldsymbol{A}'\Sigma_\psi^{-1}\boldsymbol{A})^{-1}\right]. \quad (16)$$

For $X_{T,i}$, we have

$$p(X_{T,i}|\boldsymbol{X}_{T-1},\xi_T,\boldsymbol{\theta}) \sim N\left[\sum_{j=1}^{n} a_{ij}X_{T-1,j}+\xi_T, \tau^{-1}\right]. \quad (17)$$

In the MCMC algorithm, proposed values of $\boldsymbol{X}_2,\ldots,\boldsymbol{X}_T$ are drawn in block for each t using the compact forms of (14) and (17) for improved speed and convergence. These are provided in the Appendix. The estimation of missing values in the observed counts $\{Y_{t,i}\}$ is easy as these can be straightforwardly drawn in the algorithm. Details of forward predictions in time using the proposed models are also discussed in the Appendix.

3.1 Model choice and evaluation

Many approaches have been proposed for model choice in a Bayesian framework (Kass & Raftery, 1995; Spiegelhalter et al., 2002; Gelman et al., 2014). In this work, we consider two model choice criteria that appear to be popular in the literature for areal data models for determining the best model/parameterization from the set of models developed here for any given set of data. Also, as a diagnostic tool to assess the residuals of the best-fitting model, we employ the spatiotemporal version of the Moran's I statistic (see Anderson & Ryan, 2017). These are discussed as follows.

Deviance Information Criterion (DIC)

The DIC (Spiegelhalter et al., 2002) is a Bayesian generalization of the Akaike Information Criterion (AIC) that is based on two quantities: the posterior distribution of the deviance statistic given by $D_{\boldsymbol{\theta}} = -2\log p(\boldsymbol{Y}|\hat{\boldsymbol{\theta}})$, where $p(\boldsymbol{Y}|\hat{\boldsymbol{\theta}})$ is the likelihood function $p_1(.)$ given in (12) with dependence on other variables suppressed, and the effective number of parameters p_D given by $p_D = 2(\log p(\boldsymbol{Y}|\hat{\boldsymbol{\theta}}) - E_{\boldsymbol{\theta}}(\log p(\boldsymbol{Y}|\hat{\boldsymbol{\theta}})))$,

which replaces the actual number of parameters used in AIC. The posterior expectation of the deviance provides a measure of fit of the model whereas p_D summarizes its complexity. For each of the models, the DIC can be calculated as

$$DIC = -2 \log p(\boldsymbol{Y}|\hat{\boldsymbol{\theta}}) + 2p_{DIC} \tag{18}$$

The smaller the DIC, the better the fit.

Watanabe-Akaike information criterion (WAIC)

The WAIC is a fully Bayesian criterion for model choice introduced by Watanabe (2010). The WAIC is based on pointwise calculations and can be viewed as a Bayesian approximation of cross-validation (Vehtari et al., 2015). The fit of the model is measured by $D_{WAIC} = \sum_{t=1}^{T} \sum_{i=1}^{n} E_{\boldsymbol{\theta}}(p(Y_{t,i}|\boldsymbol{\theta}))$ while the effective number of parameters is $p_{WAIC} = 2 \sum_{t=1}^{T} \sum_{i=1}^{n} (\log(E_{\boldsymbol{\theta}}(p(Y_{t,i}|\boldsymbol{\theta}))) - E_{\boldsymbol{\theta}}(\log p(Y_{t,i}|\boldsymbol{\theta})))$. In these expressions, $p(.)$ is the likelihood function $p_1(.)$ in (12). The WAIC can be expressed as

$$WAIC = -2(\hat{D}_{WAIC} - \hat{p}_{WAIC}). \tag{19}$$

In (19), all expectations are with respect to the posterior distribution of the parameters and the smaller the WAIC, the better the fit.

Spatiotemporal Moran's I

We measure residual autocorrelation in the fitted models using the MoranST statistic (Anderson & Ryan, 2017) - a spatiotemporal extension of the Moran's I statistic (Moran, 1950). The MoranST statistic is given by

$$\text{MoranST} = \frac{Tn \sum_{t=1}^{T} \sum_{i=1}^{n} \sum_{l=1}^{T} \sum_{k=1}^{n} \tilde{\boldsymbol{W}}_{(ti,lk)} (r_{ti} - \bar{r})(r_{lk} - \bar{r})}{\sum_{t=1}^{T} \sum_{i=1}^{n} \sum_{l=1}^{T} \sum_{k=1}^{n} \tilde{\boldsymbol{W}}_{(ti,lk)} \sum_{t=1}^{T} \sum_{i=1}^{n} (r_{ti} - \bar{r})^2}, \tag{20}$$

where r_{ti} is the residual for areal unit i at time t (i.e. the observed value minus the fitted value), $\bar{r}$ is the overall mean of the residuals and $\tilde{\boldsymbol{W}}_{(ti,lk)}$ is spatial weight defined

by

$$\tilde{\boldsymbol{W}}_{(ti,lk)} = \begin{cases} \boldsymbol{W}_{ik}, & \text{if} \quad t = l, \\ 1, & \text{if} \quad i = k \quad \text{and} \quad |t-l| = 1, \\ 0, & \text{otherwise}, \end{cases} \tag{21}$$

where $\boldsymbol{W}$ is the adjacency matrix defined in Section 2.1. Equation (20) considers pairs of contemporaneous observations which are neighbours in space as defined by $\boldsymbol{W}$ and and first-order neighbours in time at the same areal unit to measure spatiotemporal dependence. Similar to the Moran's I, a MoranST value close to 1 indicates a strong positive dependence whereas a value of 0 indicates the absence of dependence in space and time.

4 Applications

We illustrate the methodology developed in this paper with two data sets: county-level lung cancer mortality in Ohio and low birth weight incidence in Georgia, both in USA. Each data is described in full in the following subsections, together with details of model fitting, model choice and the results of the best fitting model and comparisons with CAR-based approaches.

4.1 Ohio lung cancer mortality data

This data concerns lung cancer deaths in $n = 88$ counties of Ohio, USA, for the years 1968 to 1988. This dataset was originally analyzed by Devine (1992) but has since been analyzed by others including Waller et al. (1997), Xia et al. (1997), Lawson (2009) and Blangiardo & Cameletti (2015). The data and the corresponding shapefile were downloaded from the website: `https://sites.google.com/a/r-inla.org/stbook/`. The number of deaths for each year and county, $Y_{t,i}$ and the corresponding expected number of deaths, $E_{t,i}$, among other details, were obtained. For an exploratory assessment of the mortality burden of the counties, the standardized mortality ratio (SMR) estimates, i.e. ratios of the observed and expected counts, for selected years are plotted in the top panel of Figure 1. High risk areas (SMR > 1) have darker colours whereas low risk areas (SMR < 1) are coloured in lighter shades. The plotted values

range between 0 and 2.56. Further exploratory analysis revealed an evidence of trend in many counties. Hence, we include a trend term (t/T) as a covariate in the analysis.

The priors placed on the parameters common to all the models were: Gamma(1,1) on τ_ξ, Gamma(2,1) on τ and $N(\mathbf{0}, 10^3 I_p)$ on $\boldsymbol{\beta}$. Others were: $a_{ij}, \omega_i, \alpha, \gamma_i, \delta_j \sim N(0,1)$ and $\phi \sim$ Gamma$(20,1)$, with the prior on ϕ chosen based on the distances between the centroids of the areas to encourage localized dependence and stationarity in $\boldsymbol{A}$. For identifiability reasons, we set $\delta_j = 0.5 \ \forall \ j$ (other values are possible) in Models 5 and 6, as discussed earlier. Each time, the MCMC algorithm was run for 10,000 iterations after a burn-in period of 10,000 iterations. Convergence was assessed by visual inspection of the trace plots of the parameters and by using other MCMC diagnostics such as Geweke diagnostic (Geweke, 1992). In Table 1, we summarize the DIC and WAIC values of the fitted models. Model 5 has the minimum values of both criteria and is, hence, the best-fitting model.

Table 1: Model choice criteria for Ohio lung cancer mortality data

Model	DIC	WAIC
Model 1	11877	11758
Model 2	11788	11681
Model 3	12234	12143
Model 4	11929	11768
Model 5	11778	11642
Model 6	11792	11681

Table 2: Posterior means, standard deviations and 95% credible intervals of the parameters of Model 5

Parameter	Mean	SD	95% CI
τ_ξ	5.2485	1.5993	(2.6433, 8.9726)
τ	25.0260	2.2877	(20.8904, 29.9214)
γ_1(Trumbull)	0.1597	0.0284	(0.1043, 0.2141)
γ_{46}(Wayne)	0.2918	0.0221	(0.2466, 0.3293)
γ_{88}(Butler)	0.1583	0.0360	(0.0852, 0.2277)
β_0	0.5440	0.0174	(0.4971, 0.5717)
β_1	0.0884	0.0305	(0.0202, 0.1332)

The posterior estimates of the parameters of Model 5 - the parameters of the latent process and the regression coefficients - are reported in Table 2. The 95% credible intervals of the reported γ_i's (and those not reported) do not include zero.

This confirms the presence of significant spatiotemporal dependence in the data. The estimate of β_1 provides evidence of increasing trend in mortality for the period under study (see Blangiardo & Cameletti, 2015, for similar results). Also, the estimated values of τ_ξ and τ are both significant and these show that ξ_t is the more dominant component of ψ_t.

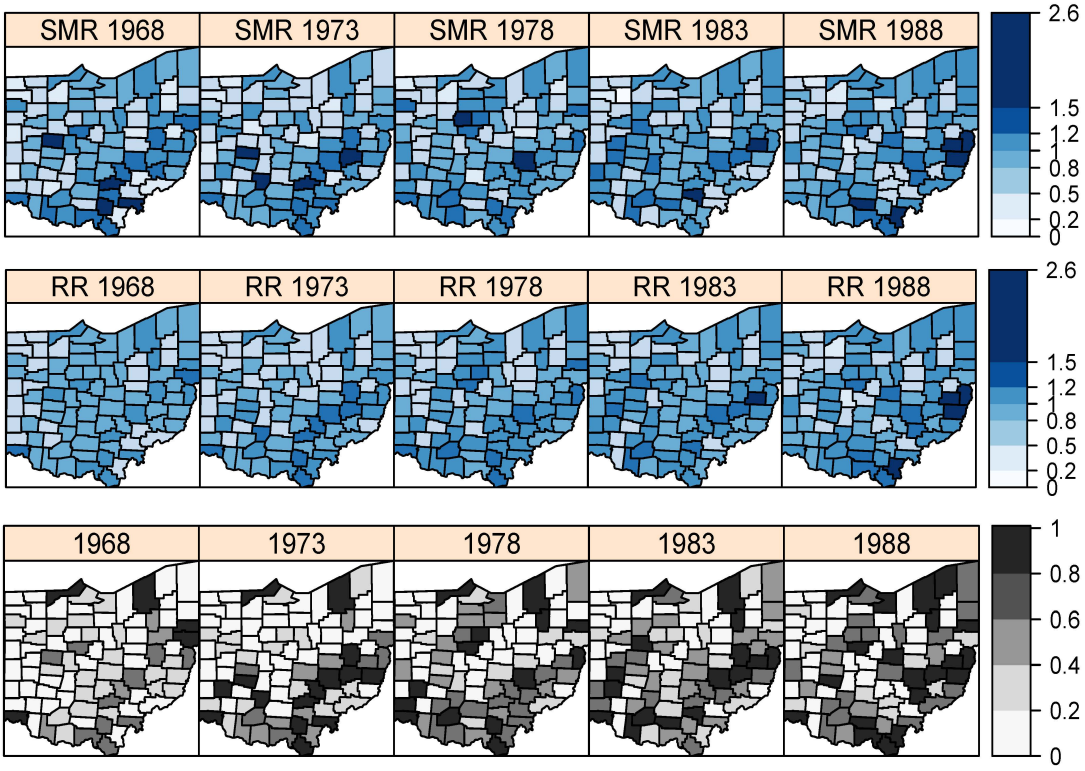

Figure 1: Ohio lung cancer mortality data: (top) Standardized Mortality Ratio (SMR) estimates, (middle) Relative Risk (RR) estimates $E(\mu_{t,i}|\boldsymbol{Y})$ and (bottom) corresponding posterior probabilities $p(\mu_{t,i} > 1|\boldsymbol{Y})$.

In the middle panel of Figure 1, the relative risk estimates of Model 5 for selected years are mapped. These generally appear to be smoother than the SMRs, ranging between 0.45 and 1.72. The associated probabilities of the risk estimates are displayed in the bottom panel of the figure. The high risk areas (i.e. areas with $p(\mu_{t,i} > 1|\boldsymbol{Y}) \geq 0.8$) are mostly concentrated in the southeastern part of the state over the years with pockets of other high risk areas springing up in an inconsistent manner in other areas. Two counties - Hamilton and Cuyahoga - consistently had posterior probabilities (i.e. $p(\mu_{t,i} > 1|\boldsymbol{Y})) \geq 0.8$ throughout the study period. These constitute hotspot areas of

high mortality. As reported in Table 3, the residuals of the Model 5 had a MoranST value of 0.0026, indicating the absence of spatial autocorrelation.

Table 3: MoranST values of the residuals of Model 5, other proposed models and some CAR models

Proposed models	MoranST	CAR models	MoranST
Model 1	0.0574	CAR-1	0.2204
Model 2	-0.0025	CAR-2	-0.0053
Model 3	-0.0357	CAR-3	-0.0350
Model 4	0.0001	CAR-4	-0.0393
Model 5	0.0026	CAR-5	-0.0359
Model 6	0.0073		

Additional analyses were carried out to compare the best-fitting model (i.e. Model 5) with some existing CAR-based models used for analyzing spatiotemporal areal data. The first CAR model examined (CAR-1), a modification of that proposed by Bernardinelli et al. (1995), characterizes spatiotemporal dependence in its mean function using spatially-varying linear time trends. In the model, two sets of spatial random effects - area-specific intercept and slope terms - both modelled using the CAR prior proposed by Leroux et al. (2000) are used to estimate a separate but correlated linear trend for each areal unit. The second CAR model considered (CAR-2, see Knorr-Held (2000) and Lee et al. (2016)) decomposes spatiotemporal variation in the data into an overall spatial effect, an overall temporal effect and a space-time interaction term. As in CAR-1, the spatial and temporal random effects are modelled using Leroux et al.'s CAR prior whereas the interaction terms are modelled as being independent. The next model, termed CAR-3 (Rushworth et al., 2014), shares the same structure as Models 1-5 proposed here in that it captures spatiotemporal dependence using a single set of random effects/latent process modelled using a first-order multivariate autoregressive process. More explicitly, the $\boldsymbol{A}$ matrix is modelled as ωI_n, where ω is a uniform autoregressive parameter, while Σ_ψ is modelled as a product of a variance parameter and the correlation matrix proposed by (Leroux et al., 2000). In the fourth model (CAR-4, Lee & Lawson (2016)), a single set of spatiotemporal random effects having the same structure as CAR-3 characterizes spatiotemporal dependence. Additionally, this model incorporates a piecewise constant term to model step changes in the mean response with the aim of identifying clusters of areas of elevated or reduced risks. The last CAR model examined, CAR-5, was proposed by Rushworth et al. (2017). This

is similar in structure to CAR-3 but replaces the usual binary neighbourhood matrix in the correlation matrix with a modified version in which the non-zero elements of the matrix are treated as parameters with support on the unit interval to allow adaptive levels of spatial smoothing. These CAR models were chosen because they could be implemented using an open software. The models were fitted using CARBayesST package in R (Lee et al., 2016) with standard model specifications.

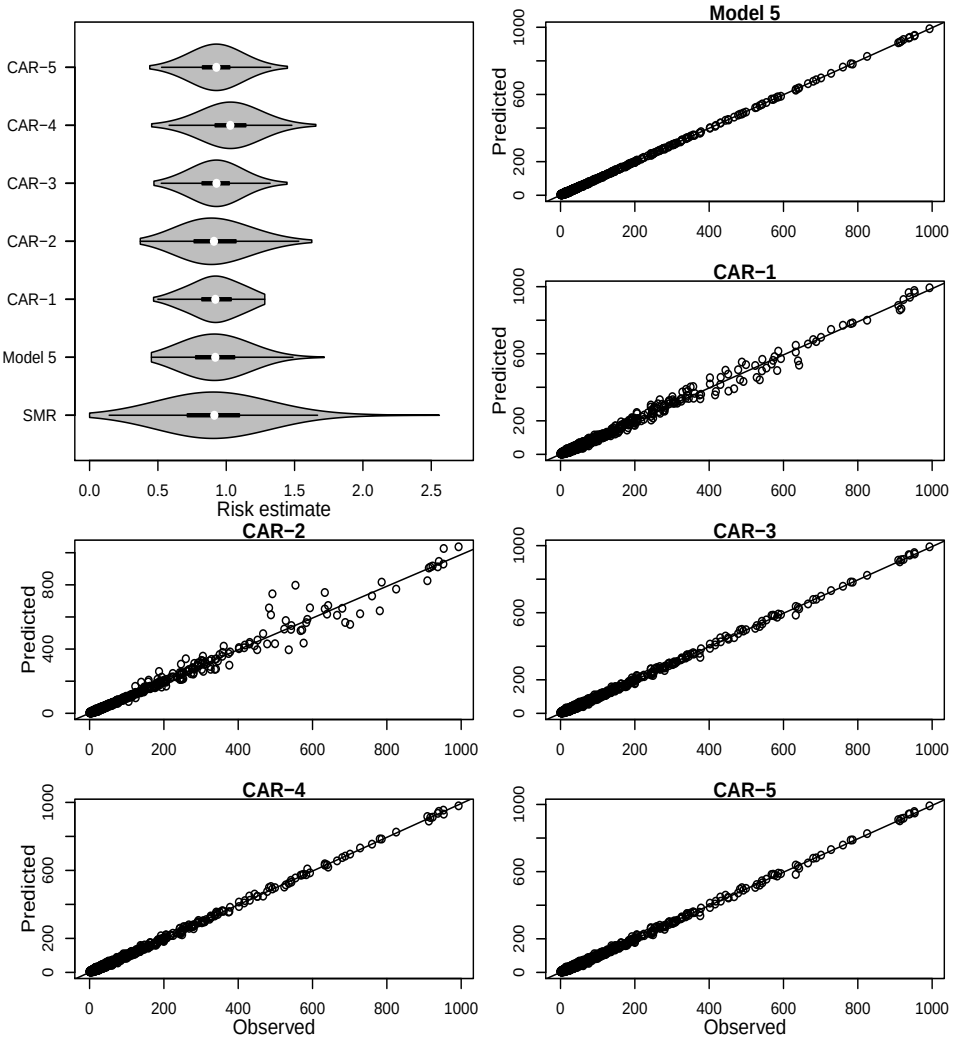

Figure 2: Ohio lung cancer mortality data: Violin plots of Standardized Mortality Ratio (SMR) and Relative Risk (RR) estimates from Model 5 and the CAR models and plots of observed versus predicted values from all the models.

In Figure 2, relative risk estimates obtained from the CAR models are compared with those of Model 5 using violin plots. In general, these plots show that Model 5 performed equally well as the CAR models in shrinking the risk estimates towards their mean value. However, upon examining the plots of the observed versus predicted values from these models, it is clear that Model 5 outperforms some of the CAR models (CAR-1 and CAR-2) by producing less biased predictions. The MoranST values of the proposed models (including Model 5) and the CAR models are reported in Table 3. These indicate that spatiotemporal autocorrelation is adequately accounted for in these models, with the exception of CAR-1 which appears to show some residual autocorrelation.

4.2 Georgia low birth weight data

Low birthweight (LBW), i.e. a live birth weighing less than 2500g, has been associated with greater risk of infant mortality. As in many countries around the world, LBW has been a significant public health issue in the USA, with rates in Georgia being among the worst (Tian et al., 2013). In this section, we undertake an analysis of low birth weights data for the 159 counties of the state of Georgia, USA, for the years 2000-2010. The data were downloaded from `https://sites.google.com/a/r-inla.org/stbook/` and were analyzed in Blangiardo & Cameletti (2015). Other versions of the data covering different spatial and temporal scales have been analyzed by Lawson (2009), Kirby et al. (2011) and Tian et al. (2013). For each county, both the observed and expected counts were available with the latter calculated by using the internal standardization method. The Standardized Incidence Ratio (SIR) estimates, ranging between 0.00 and 3.04 are mapped in Figure 3 (top panel) for selected years. These provide an exploratory visualization of the patterns in the data.

Due to the short time period covered by the data, we include only an intercept term in the regression function and regard this analysis as purely illustrative. The prior distributions used in the analysis were similar to those given in Section 4.1. Here again, the MCMC algorithms were run for 10,000 iterations each time after a burn in period of 10,000 iterations. Convergence diagnostics were also carried out as before. There is a good agreement between the DIC and WAIC values of the models reported in Table 4. Model 4 is the best fitting model with the lowest values of both criteria. Posterior estimates of the parameters of this model are shown in Table 5. The estimates of the parameters of the latent process generally show the presence of significant spatiotemporal correlation in the data. The middle and bottom panels of

Figure 3 map the relative risk estimates and the corresponding posterior probabilities of being greater than one, respectively. The RR estimates (min. = 0.45, max. = 2.30) from the model generally appear to be smoother and less variable than the SIRs.

Table 4: Model choice for Georgia low birth weight data

Model	DIC	WAIC
Model 1	11842	11685
Model 2	11934	11788
Model 3	12137	11949
Model 4	11803	11624
Model 5	11955	11786
Model 6	11875	11739

Table 5: Posterior means, standard deviations and 95% credible intervals of the parameters of Model 4

Parameter	Mean	SD	95 % CI
α	0.4717	0.0333	(0.4058, 0.5361)
ϕ	7.4813	0.4404	(6.7026, 8.3735)
τ_ξ	6.8346	2.6231	(2.6917, 12.9211)
τ	16.5504	1.8466	(13.3152, 20.3702)
β_0	-0.0114	0.0141	(-0.0345, 0.0272)

Throughout the study period, counties with higher LBW rates were found to be concentrated in the area stretching from the southwestern part of the state up to its middle area, as these figures reveal. Similar findings were also observed in Tian et al. (2013). Fulton, Hancock, Spalding, Bibb, Muscogee, Clay and Dougherty counties consistently had posterior probabilities (i.e. $p(\mu_{t,i} > 1|\boldsymbol{Y})) \geq 0.8$. These counties constitute hotspot areas of high LBW rates. As reported in Table 6, the residuals of the Model 4 had a MoranST value of 0.0165, indicating the absence of spatial autocorrelation.

As in Section 4.1, we performed a comparative analysis of the data using some CAR-based models described previously. The results obtained are displayed in Figure 4. The violin plots indicate that all the models show similar behaviour in shrinking the risk estimates although the best model (Model 4) appears to have identified slightly greater numbers of higher and lower risk areas in comparison with the CAR models.

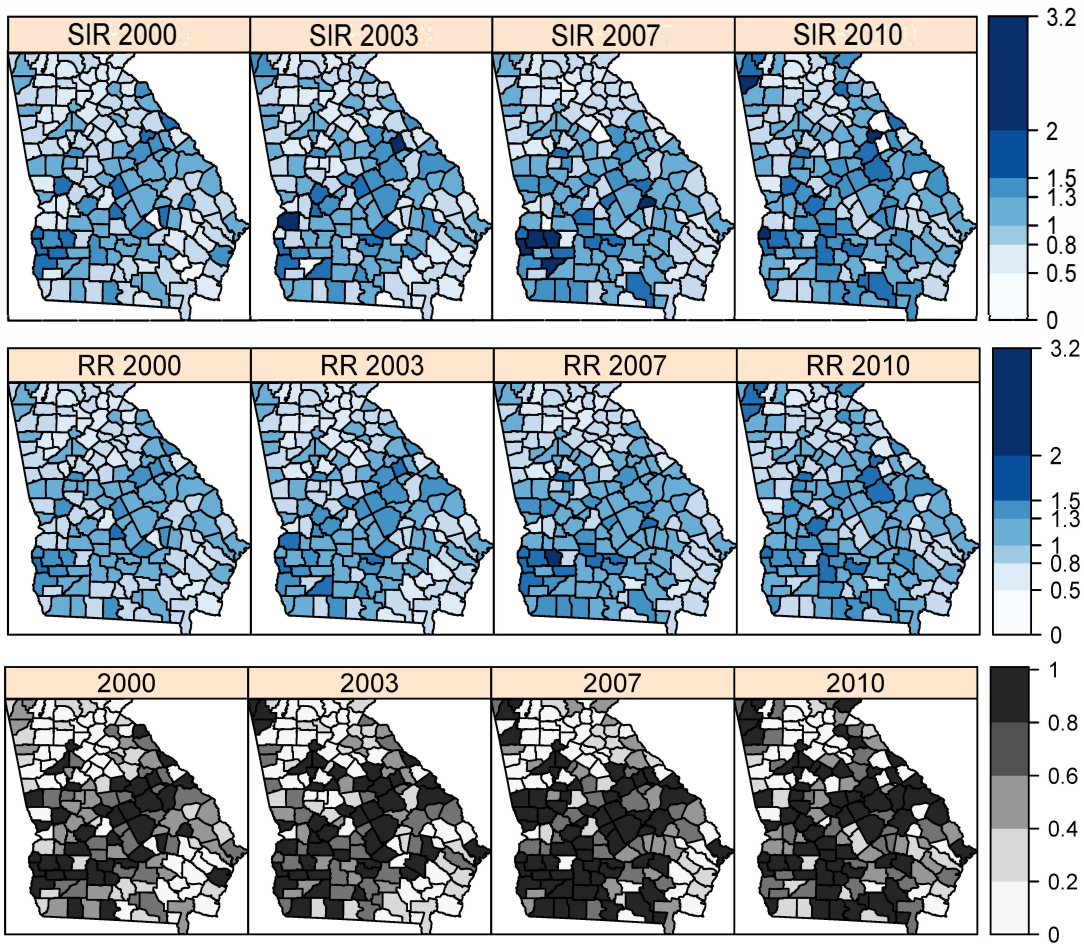

Figure 3: Georgia low birth weight data: (top) Standardized Incidence Ratio (SIR) estimates, (middle) Relative Risk (RR) estimates $E(\mu_{t,i}|\boldsymbol{Y})$ and (bottom) corresponding posterior probabilities $p(\mu_{t,i} > 1|\boldsymbol{Y})$.

Table 6: MoranST values of the residuals of the models

Proposed models	MoranST	CAR models	MoranST
Model 1	0.0153	CAR-1	0.2179
Model 2	0.0576	CAR-2	0.0371
Model 3	-0.0405	CAR-3	-0.0606
Model 4	0.0165	CAR-4	-0.0301
Model 5	0.0697	CAR-5	-0.0575
Model 6	-0.0349		

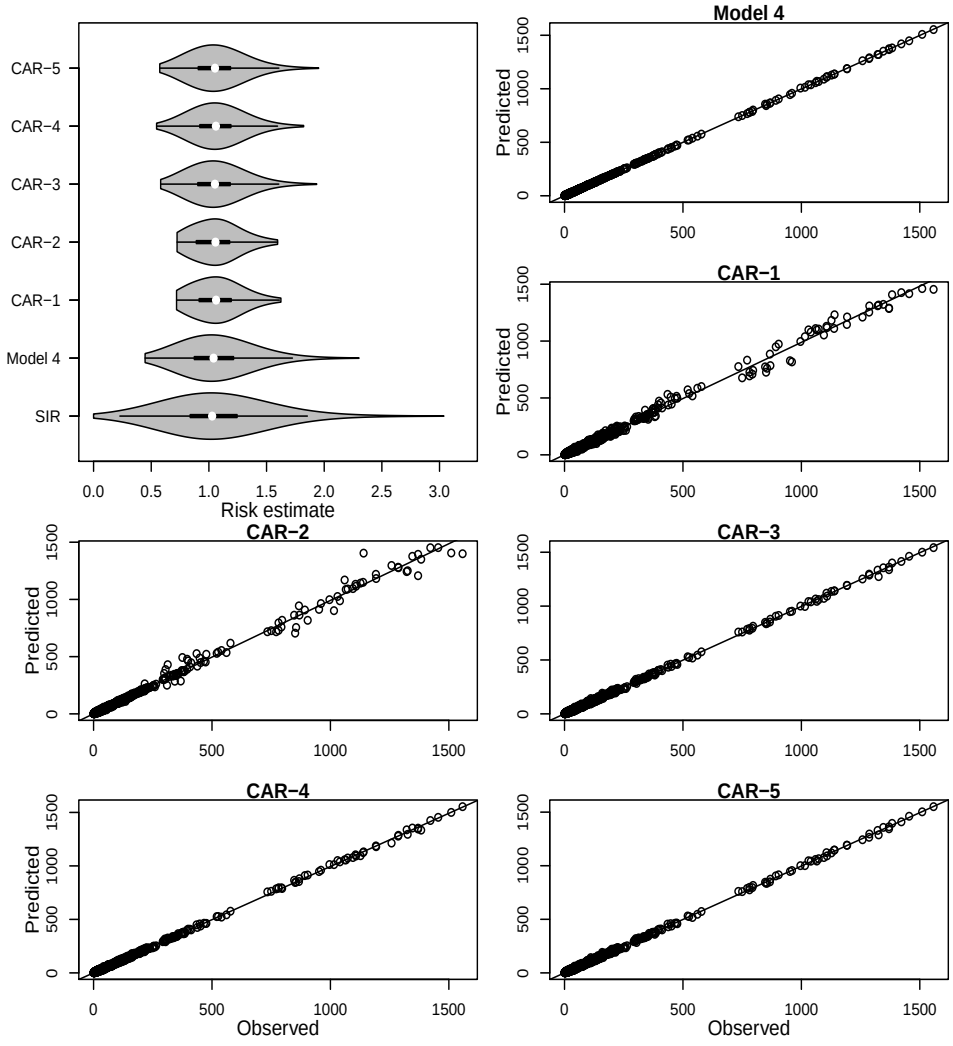

Figure 4: Georgia low birth weight data: Violin plots of Standardized Incidence Ratio (SIR) and Relative Risk (RR) estimates from Model 5 and the CAR models and plots of observed versus predicted values from the models.

The plots of observed versus predicted values of the models reveal that Model 4 and other CAR models outperformed CAR-1 and CAR-2, both of which produced more biased predictions. The MoranST values of the residuals of the proposed models and those of CAR models given in Table 6 generally indicate that spatiotemporal autocorrelation has been properly accounted for by these models with the exception of CAR-1 with a relatively higher MoranST value.

5 Discussion

In this paper, we have presented a new methodology for the analysis of spatiotemporal areal count data. In the proposed models, a new approach that characterizes spatiotemporal dependence dynamically using the transition matrix of the first-order autoregressive latent process as against its the error covariance matrix was explored. A great deal of flexibility was achieved through the development of various parameterizations of the transition matrix. Parameter estimation in the proposed models is shown to be straightforward and we have demonstrated, using two illustrative examples, that the methodology is as effective as, and in some cases more effective than, alternative approaches which characterize spatial and spatiotemporal correlation using CAR-based priors. In particular, the goodness-of-fit evaluation done in Section 4 points to the fact that the proposed models intrinsically achieve a good balance between smoothing (variance reduction) and bias in relative risk estimation. This implies that these models are not only well-suited for identifying the overall spatial pattern in the risk estimates but also for highlighting the presence of heterogeneity resulting from clustering of high or low risk areas or individual areas with distinctive risks. This is a desirable attribute in disease mapping studies (see Best et al., 2005); however, a more comprehensive study involving a variety of set-ups not considered here will be necessary to provide a broad evaluation of the models in this regard.

A topic of interest in spatiotemporal modelling often ignored in models for areal data is the modelling of the dependence of responses on covariates using spatially-varying or area-specific regression coefficients. Such specifications not only introduce greater flexibility in modelling but also facilitate the quantification of covariate effects for each areal unit - an important attribute that is appealing from an epidemiological perspective - and production of risk maps for individual covariate factors. An approach for accounting for spatial autocorrelation in the regression parameters explored in Assunção (2003) is the use of a Markov random field prior. Correlated normal priors with the elements of the correlation matrix specified using a suitable parametric function of the distances between the centroids of the areas, such as the exponential decay function used in Section 2.1, could also be used. The uniform regression coefficients used in the proposed models is a special case of spatially-varying coefficients and it is straightforward to extend these models along these lines. We plan to take up this idea in future work. However, to demonstrate the utility of this approach using the datasets analyzed in Section 4, spatially-varying covariate data such as county-level

socioeconomic variables as in Kirby et al. (2011) will need to be included in the analysis. We anticipate that the selected models for the latent process may change when such covariates are included in the analysis as these help to account for residual autocorrelation in the data. Nevertheless, spatially-varying regression parameters may not be relevant in all cases. Where prior knowledge suggests that covariate effects do not vary over the study region, a uniform set of regression parameters should be used to preserve parsimony.

Finally, as an additional future line of development, generalizations of the methodology to other outcome distributions such as binomial and Gaussian distributions will be considered. This will improve the applicability of the methodology.

Competing Interests

The authors declare no competing interests.

References

Anderson, C. & Ryan, L. M. (2017). A comparison of spatio-temporal disease mapping approaches including an application to ischaemic heart disease in New South Wales, Australia. *International Journal of Environmental Research and Public Health*, 14(2), 146.

Assunção, R. M. (2003). Space varying coefficient models for small area data. *Environmetrics*, 14(5), 453–473.

Banerjee, S., Carlin, B. P., & Gelfand, A. E. (2015). *Hierarchical Modeling and Analysis for Spatial Data*. CRC Press.

Bernardinelli, L., Clayton, D., Pascutto, C., Montomoli, C., Ghislandi, M., & Songini, M. (1995). Bayesian analysis of space-time variation in disease risk. *Statistics in Medicine*, 14(21-22), 2433–2443.

Besag, J. (1974). Spatial interaction and the statistical analysis of lattice systems (with discussion). J. Roy. Statist. Soc., Ser. B, 36, 192–236.

Besag, J., York, J., & Mollié, A. (1991). Bayesian image restoration, with two applications in spatial statistics. *Annals of the Institute of Statistical Mathematics*, 43(1), 1–20.

Best, N., Richardson, S., & Thomson, A. (2005). A comparison of Bayesian spatial models for disease mapping. *Statistical Methods in Medical Research*, 14, 35–59.

Blangiardo, M. & Cameletti, M. (2015). *Spatial and Spatio-temporal Bayesian Models in R-INLA*. John Wiley & Sons.

Cressie, N. & Wikle, C. (2011). *Statistics for Spatio-Temporal Data*. John Wiley & Sons, New York.

Davis, R. A., Zang, P., & Zheng, T. (2012). Sparse vector autoregressive modeling. *arXiv preprint arXiv:1207.0520*.

Devine, O. J. (1992). Empirical Bayes and constrained empirical Bayes methods for estimating incidence rates in spatially aligned areas. Unpublished PhD thesis, Department of Biostatistics, Emory University.

Gelman, A., Hwang, J., & Vehtari, A. (2014). Understanding predictive information criteria for Bayesian models. *Statistics and Computing*, 24(6), 997–1016.

Geweke, J. (1992). Evaluating the accuracy of sampling-based approaches to calculating posterior moments. In J. M. Bernardo, J. Berger, A. P. Dawid, & J. F. M. Smith (Eds.), *Bayesian Statistics 4* (pp. 169–193). Oxford: Oxford University Press.

Kass, R. E. & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, 90(430), 773–795.

Kirby, R. S., Liu, J., Lawson, A. B., Choi, J., Cai, B., & Hossain, M. (2011). Spatio-temporal patterning of small area low birth weight incidence and its correlates: A latent spatial structure approach. *Spatial and Spatio-temporal Epidemiology*, 2(4), 265 – 271.

Knorr-Held, L. (2000). Bayesian modelling of inseparable space-time variation in disease risk. *Statistics in Medicine*, 19, 2555–2567.

Knorr-Held, L. & Besag, J. (1998). Modelling risk from a disease in space and time. *Statistics in Medicine*, 17, 2045–2060.

Lawson, A. (2009). *Bayesian Disease Mapping. Hierarchical Modeling in Spatial Epidemiology*. CRC Press, Boca Raton, Florida.

Lee, D. (2011). A comparison of conditional autoregressive models used in Bayesian disease mapping. *Spatial and Spatio-temporal Epidemiology*, 2(2), 79–89.

Lee, D. (2013). CARBayes: An R package for Bayesian spatial modeling with conditional autoregressive priors. *Journal of Statistical Software*, 55(13), 192–236.

Lee, D. & Lawson, A. (2014). Cluster detection and risk estimation for spatio-temporal health data. *ArXiv*.

Lee, D. & Lawson, A. (2016). Quantifying the spatial inequality and temporal trends in maternal smoking rates in Glasgow. *Annals of Applied Statistics*, 10, 1427–1446.

Lee, D., Rushworth, A., & Napier, G. (2016). CARBayesST version 2.2: An R package for spatio-temporal areal unit modelling with conditional autoregressive priors. *Technical Report, University of Glasgow*.

Leroux, B. G., Lei, X., & Breslow, N. (2000). Estimation of disease rates in small areas: A new mixed model for spatial dependence. In M. E. Halloran & D. Berry (Eds.), *Statistical Models in Epidemiology, the Environment and Clinical Trials* (pp. 179–191). Springer:New York, NY.

Lütkepohl, H. (2005). *New Introduction to Multiple Time Series Analysis*. Springer-Verlag, Berlin, DE.

McShane, L. M., Albert, P. S., & Palmatier, M. A. (1997). A latent process regression model for spatially correlated count data. *Biometrics*, 53(2), 698–706.

Moran, P. A. P. (1950). Notes on continuous stochastic phenomena. *Biometrika*, 37(1/2), 17–23.

Pascutto, C., Wakefield, J. C., Best, N. G., Richardson, S., Bernadinelli, L., Staines, A., & Elliot, P. (2000). Statistical issues in the analysis of disease mapping data. *Statistics in Medicine*, 19, 2493–2519.

Pfeifer, P. E. & Deutsch, S. J. (1980). A three-stage iterative procedure for space-time modeling. *Technometrics*, 22(1), 35–47.

R Core Team (2014). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Rushworth, A., Lee, D., & Mitchell, R. (2014). A spatio-temporal model for estimating the long-term effects of air pollution on respiratory hospital admissions in Greater London. *Spatial and Spatio-temporal Epidemiology*, 10, 29 – 38.

Rushworth, A., Lee, D., & Sarran, C. (2017). An adaptive spatiotemporal smoothing model for estimating trends and step changes in disease risk. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 66(1), 141–157.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., & Van Der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(4), 583–639.

Tanner, M. A. & Wong, W. H. (1987). The calculation of posterior distributions by data augmentation. *Journal of the American Statistical Association*, 82(398), 528–540.

Tian, J., Tu, W., Tedders, S., & Chen, D. (2013). A spatial–temporal analysis of low birth weight prevalence in Georgia, USA. *GeoJournal*, 78(5), 885–895.

Utazi, C. E. (2014). *Spatio-Temporal Modelling of Partially Observed Processes*. PhD thesis, Lancaster University.

Utazi, C. E. (2017). Bayesian single changepoint estimation in a parameter-driven model. *Scandinavian Journal of Statistics*, 44(3), 765–779. 10.1111/sjos.12274.

Vehtari, A., Gelman, A., & Gabry, J. (2015). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *ArXiv e-prints*.

Wakefield, J. (2007). Disease mapping and spatial regression with count data. *Biostatistics*, 8(2), 158–183.

Waller, L. A., Carlin, B. P., Xia, H., & Gelfand, A. E. (1997). Hierarchical spatio-temporal mapping of disease rates. *Journal of the American Statistical Association*, 92(438), 607–617.

Watanabe, S. (2010). Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory. *J. Mach. Learn. Res.*, 11, 3571–3594.

Wikle, C. K., Berliner, L. M., & Cressie, N. (1998). Hierarchical Bayesian space-time models. *Environmental and Ecological Statistics*, 5(2), 117–154.

Xia, H. & Carlin, B. P. (1998). Spatio-temporal models with errors in covariates: Mapping Ohio lung cancer mortality. *Statistics in Medicine*, 17(18), 2025–2043.

Xia, H., Carlin, B. P., & Waller, L. A. (1997). Hierarchical models for mapping Ohio lung cancer rates. *Environmetrics*, 8, 107–120.

Xu, K. & Wikle, C. K. (2007). Estimation of parameterized spatio-temporal dynamic models. *Journal of Statistical Planning and Inference*, 137(2), 567–588.

Zeger, S. L. (1988). A regression model for times series of counts. *Biometrika*, 75(4), 621–9.

Appendix

Conditional posterior distributions of the parameters of the proposed models

Here, we present the conditional posterior distributions of the parameters of the models developed in Section 2.1. We begin with the parameters common to all the models.

The conditional posterior distribution of τ, the precision parameter for $\epsilon_{t,i}$, is the Gamma distribution given by

$$\pi(\tau|\boldsymbol{\theta}_{-\tau}) \sim \text{Gamma}\left(a_\tau + \frac{n(T-1)}{2}, \frac{1}{2}\sum_{i=1}^{n}\sum_{t=2}^{T}\left[X_{t,i} - \sum_{j=1}^{n} a_{ij}X_{t-1,j} - \xi_t\right]^2 + b_\tau\right). \tag{22}$$

For the autoregressive parameters a_{ik}, $i,k \in \{1,\ldots,n\}$, we have that

$$\pi(a_{ik}|\boldsymbol{\theta}_{-a_{ik}}) \sim N\left(\frac{\tau\sum_{t=2}^{T} X_{t-1,k}U_t(ik) + u_a v_a}{\tau\sum_{t=2}^{T} X_{t-1,k}^2 + v_a}, \frac{1}{\tau\sum_{t=2}^{T} X_{t-1,k}^2 + v_a}\right), \tag{23}$$

where $U_t(ik) = X_{t,i} - \sum_{\substack{j=1 \\ j\neq k}}^{n} a_{ij}X_{t-1,j} - \xi_t$. In Model 2, these parameters are only updated when $W_{ik} = 1$. The conditional posterior distribution of τ_ξ, the precision parameter for ξ_t, is

$$\pi(\tau_\xi|\boldsymbol{\theta}_{-\tau_\xi}) \sim \text{Gamma}\left(\frac{T}{2} + a_\xi, \frac{\sum_{t=1}^{T}\xi_t^2}{2} + b_\xi\right). \tag{24}$$

The posterior distribution of $\boldsymbol{\xi}$ factors into $T-1$ independent posteriors for each ξ_t. That is, for $t = 2, \ldots, T$, we have that

$$\pi(\xi_t|\boldsymbol{\theta}_{-\xi_t}) \sim N\left(\frac{\tau\sum_{i=1}^{n}(X_{t,i} - \sum_{j=1}^{n} a_{ij}X_{t-1,j})}{\tau_\xi + n\tau}, \frac{1}{\tau_\xi + n\tau}\right). \tag{25}$$

The conditional posterior distribution of the uniform regression coefficients, $\boldsymbol{\beta}$, is given by

$$\begin{aligned}\pi(\boldsymbol{\beta}|\boldsymbol{\theta}_{-\boldsymbol{\beta}}) \propto \quad & \exp\left(\sum_{i=1}^{n}\sum_{t=1}^{T}\left[\mathbf{z}'_{t,i}\boldsymbol{\beta}Y_{t,i} - E_{t,i}e^{\mathbf{z}'_{t,i}\boldsymbol{\beta}+X_{t,i}}\right]\right) \\ & \times \exp\left(-\frac{v_\beta}{2}\sum_{j=0}^{p-1}(\beta_j - u_\beta)^2\right).\end{aligned} \tag{26}$$

The conditional distributions of τ, τ_ξ, ξ_t and a_{ik} in equations (22) - (25) are in standard forms. These are updated using Gibbs moves. For $\boldsymbol{\beta}$, we use a random walk Metropolis step. By making appropriate substitutions for a_{ij} in (22), (24) and (25), the conditional posterior distributions of these parameters in Models 1 - 6 can be obtained.

The conditional posterior distribution of the site-specific autoregressive parameter in model (6), ω_i, is given by

$$\pi(\omega_i|\boldsymbol{\theta}_{-\omega_i}) \sim N\left(\frac{\tau\sum_{t=2}^{T} X_{t-1,i}(X_{t,i} - \xi_t) + u_\omega v_\omega}{\tau\sum_{t=2}^{T} X_{t-1,i}^2 + v_\omega}, \frac{1}{\tau\sum_{t=2}^{T} X_{t-1,i}^2 + v_\omega}\right). \tag{27}$$

The full conditional distribution of α in Models 3 and 4 is

$$\pi(\alpha|\boldsymbol{\theta}_{-\alpha}) \sim N\left(\frac{\tau\sum_{i=1}^{n}\sum_{t=2}^{T} U_{i,t-1}(X_{t,i} - \xi_t) + u_\alpha v_\alpha}{\tau\sum_{i=1}^{n}\sum_{t=2}^{T}(U_{i,t-1})^2 + v_\alpha}, \frac{1}{\tau\sum_{i=1}^{n}\sum_{t=2}^{T}(U_{i,t-1})^2 + v_\alpha}\right); \tag{28}$$

where $U_{i,t-1}$ equals $\sum_{j=1}^{n} W_{ij}X_{t-1,j}$ in Model 3 and $\sum_{j=1}^{n} e^{-\phi d_{ij}}X_{t-1,j}$ in Model 4.

In the models in equations (10) and (11), the γ_i's and δ_j's are to be estimated. For the γ_i's, we have

$$\pi(\gamma_i|\boldsymbol{\theta}_{-\gamma_i}) \sim N\left(\frac{\tau\sum_{t=2}^{T} U_{i,t-1}(Y_{t,i} - \xi_t) + u_\gamma v_\gamma}{\tau\sum_{t=2}^{T} U_{i,t-1}^2 + v_\gamma}, \frac{1}{\tau\sum_{t=2}^{T} U_{i,t-1}^2 + v_\gamma}\right), \tag{29}$$

where $U_{i,t-1} = \sum_{j=1}^{n}\delta_j W_{ij}Y_{t-1,j}$ in (10) and $\sum_{j=1}^{n}\delta_j e^{-d_{ij}\phi}Y_{t-1,j}$ in (11). For the δ_j's,

we have

$$\pi(\delta_j|\boldsymbol{\theta}_{-\delta_j}) \sim N\left(\frac{\tau\sum_{i=1}^{n}\gamma_i C_{ij}\sum_{t=2}^{T}U_{i,t}+u_\delta v_\delta}{\tau\sum_{i=1}^{n}\gamma_i^2 C_{ij}^2\sum_{t=2}^{T}X_{t-1,j}^2+v_\delta}, \frac{1}{\tau\sum_{i=1}^{n}\gamma_i^2 C_{ij}^2\sum_{t=2}^{T}X_{t-1,j}^2+v_\delta}\right), \tag{30}$$

where $U_{i,t} = X_{t,i}X_{t-1,j} - \xi_t X_{t-1,j} - \gamma_i X_{t-1,j}\sum_{\substack{k=1\\k\neq j}}^{n}\delta_k C_{ik}X_{t-1,k}$ and $C_r = W_r$ in model (10) and $e^{-d_r\phi}$ in (11). Note that under the prior assumption of positive dependence between the areas, equations (23), (27), (28), (29), and (30) will be truncated distributions with support in $\mathbb{R}^+$. Note also that with a large n, block updates could be explored for the area-specific parameters for increased computational efficiency. For the decay parameter ϕ in models (9) and (11,) we have that

$$\begin{aligned}\pi(\phi|\boldsymbol{\theta}_{-\phi}) \propto &\exp\left(-\frac{\tau}{2}\sum_{i=1}^{n}\sum_{t=2}^{T}\left[X_{t,i}-\sum_{j=1}^{n}C_{ij}e^{-\phi d_{ij}}X_{t-1,j}-\xi_t\right]^2\right)\\ &\times\phi^{a_\phi-1}\exp(-b_\phi\phi),\end{aligned} \tag{31}$$

where C_{ij} equals α in equation (9) and $\gamma_i\delta_j$ in (11). All the model-specific parameters in equations (27) - (30) are updated using Gibbs steps while the conditional distribution in (31) is sampled using a random walk Metropolis step.

As highlighted in Section 3, the compact forms of equations (14) and (17) for updating $\boldsymbol{X}_2,\ldots,\boldsymbol{X}_{T-1}$ and $\boldsymbol{X}_T$ are

$$\begin{aligned}p(\boldsymbol{X}_t|\boldsymbol{X}_{t-1},\boldsymbol{X}_{t+1},\boldsymbol{\theta}) \sim N\Big((\Sigma_\psi^{-1}+\boldsymbol{A}'\Sigma_\psi^{-1}\boldsymbol{A})^{-1}(\boldsymbol{X}_{t-1}'\boldsymbol{A}'\Sigma_\psi^{-1}+\boldsymbol{X}_{t+1}'\Sigma_\psi^{-1}\boldsymbol{A})',\\ (\Sigma_\psi^{-1}+\boldsymbol{A}'\Sigma_\psi^{-1}\boldsymbol{A})^{-1}\Big), \quad \text{and}\\ p(\boldsymbol{X}_T|\boldsymbol{X}_{T-1},\boldsymbol{\theta}) \sim N(\boldsymbol{A}\boldsymbol{X}_{T-1},\Sigma_\psi),\end{aligned}$$

respectively. Finally, the stationarity condition discussed in Section 2 can be straightforwardly imposed when updating the autoregressive parameters in each case. All the MCMC algorithms for the proposed models were coded in **R** (R Core Team, 2014).

Prediction

We consider predictions in time for the latent process $\{\boldsymbol{X}_t\}$ and the observed counts $\{\boldsymbol{Y}_t\}$. First, the predictive distribution of $\{\boldsymbol{X}_t\}$ is derived as follows.

By iterating on the second level of equation (5) from time $t = T$ to time $t = T+h$,

we have that

$$\boldsymbol{X}_{T+h} = \boldsymbol{A}^h \boldsymbol{X}_T + \sum_{i=0}^{h-1} \boldsymbol{A}^i(\boldsymbol{\epsilon}_{T+h-i} + \xi_{T+h-i}\mathbf{1}). \tag{32}$$

From (32), it is easy to see that

$$E(\boldsymbol{X}_{T+h}|\boldsymbol{X}_T) = \boldsymbol{A}^h \boldsymbol{X}_T \tag{33}$$

and

$$\Sigma_X(h) = Var(\boldsymbol{X}_{T+h}|\boldsymbol{X}_T) = \sum_{i=0}^{h-1} \boldsymbol{A}^i[\tau_\xi^{-1}\mathbf{1}\mathbf{1}' + \Sigma](\boldsymbol{A}^i)'. \tag{34}$$

Using the vec operator, we can simplify (34) further to obtain

$$\text{vec}(\Sigma_X(h)) = (I_{n^2} - (\boldsymbol{A} \otimes \boldsymbol{A})^h)(I_{n^2} - (\boldsymbol{A} \otimes \boldsymbol{A}))^{-1}\text{vec}(\tau_\xi^{-1}\mathbf{1}\mathbf{1}' + \Sigma). \tag{35}$$

The h-step ahead predictive distribution of $\{\boldsymbol{X}_t\}$ for area i is therefore given by

$$X_{T+h,i}|\boldsymbol{X}_T \sim N\left([\boldsymbol{A}^h \boldsymbol{X}_T]_i, [\Sigma_X(h)]_i\right), \tag{36}$$

where $[.]_i$ denotes the ith element or the ith diagonal element of the corresponding vector or matrix. Conditional on (36), for $i = 1, \ldots, n$, $Y_{T+h,i}$ has a Poisson distribution with an unconditional mean given by

$$E[Y_{T+h,i}|\boldsymbol{Y}, \boldsymbol{\theta}] = E_{T+h,i}\exp\left(\mathbf{z}'_{T+h,i}\boldsymbol{\beta}_i + 0.5[\Sigma_X(h)]_i\right) E[\exp([\boldsymbol{A}^h \boldsymbol{X}_T]_i)|\boldsymbol{X}, \boldsymbol{\theta}]. \tag{37}$$

Equation (37) is straightforward to evaluate in an MCMC algorithm.