Statistica Sinica **25** (2015), 189-204
doi:http://dx.doi.org/10.5705/ss.2013.212w

BAYESIAN SPATIAL-TEMPORAL MODELING OF ECOLOGICAL ZERO-INFLATED COUNT DATA

Xia Wang[1], Ming-Hui Chen[2], Rita C. Kuo[3] and Dipak K. Dey[2]

[1] *University of Cincinnati,* [2] *University of Connecticut*
and [3] *Lawrence Berkeley National Laboratory*

Abstract: A Bayesian hierarchical model is developed for count data with spatial and temporal correlations as well as excessive zeros, uneven sampling intensities, and inference on missing spots. Our contribution is to develop a model on zero-inflated count data that provides flexibility in modeling spatial patterns in a dynamic manner and also improves the computational efficiency via dimension reduction. The proposed methodology is of particular importance for studying species presence and abundance in the field of ecological sciences. The proposed model is employed in the analysis of the survey data by the Northeast Fisheries Sciences Center (NEFSC) for estimation and prediction of the Atlantic cod in the Gulf of Maine - Georges Bank region. Model comparisons based on the deviance information criterion and the log predictive score show the improvement by the proposed spatial-temporal model.

Key words and phrases: Bayesian hierarchical modeling, deviance information criterion, log predictive score, spatial dynamic modeling, zero-inflated Poisson.

1. Introduction

An ecological survey often involves a collection of counts of individuals in one or more species over a study region across years. The intention of the survey is to estimate and predict the evolution of species distribution over the region. However, the spatial coverage is usually sparse and the survey locations are scattered throughout the study region and vary from year to year. Thus, the survey locations are rarely repeated frequently across years. Models including spatial and temporal components are of great importance in making inference on the species distribution using the field survey data.

Another unique characteristic of ecological count data is the interpretation of a zero count at a given time and location. A zero count may indicate that the species is truly absent at the given location, or it may be a result of incomplete survey coverage and imperfect detection. The zero-inflated Poisson (ZIP) model is a natural choice in modeling such count data. A mixture of Bernoulli and Poisson processes fits this scenario nicely, in which the Bernoulli process captures the true presence of the species while the Poisson process accounts for the abundance of the species when it is present. The ZIP model has been considered in many applications in the literature. Agarwal, Gelfand, and Citron-Pousty (2002) applied the ZIP model to fit isopod nest burrows data in which a spatial pattern was modeled in the Poisson process. Fei and Rathbun (2006) used a ZIP model in an oak regeneration study, modeled the spatial correlations in the Bernoulli process, and assumed the Poisson processes were independent across locations.

There is a rich literature on spatial-temporal modeling of zero-inflated count data. Wikle and Anderson (2003) applied the ZIP model using the Bayesian hierarchical spatial-temporal approach in the analysis of the 1953-1995 U.S. tornado report counts data. They assumed spatially varying temporal trend and ENSO effect along with spatially correlated random processes. Fernandes, Schmidt, and Migon (2009) discussed the zero-inflated spatial–temporal processes for continuous non-negative values and count data with point-referenced or areal spatial structure. They assumed the random process in both the Bernoulli and Poisson regression models as spatially correlated but independent across time. In both studies, the temporal pattern is modeled by the temporal covariates instead of random processes. Ver Hoef and Jansen (2007) developed ZIP and hurdle models with space-time errors to investigate haul-out patterns of harbor seals on glacial ice. They assumed that the spatial and temporal random effects are additive.

There are at least two aspects that need further investigation for spatial-temporal modeling of zero-inflated count data. The first is about less restrictive assumptions on the spatial-temporal correlation structure. Different structures on spatial-temporal random processes have been proposed for modeling counts data (Zhuang and Cressie (2012)). However, most of the spatial-temporal models for zero-inflated count data either relied on temporal or spatial-temporal covariates to model the dynamic evolution, or assumed that the spatial-temporal random processes are not only separable but also additive, which may not be desirable (Banerjee, Carlin, and Gelfand (2004)). Given the complexity of the ecological system, it is common that some influential temporal or spatial-temporal covariates may not be observed or available. Thus, it is more desirable to include both spatial and temporal random effects to systematically account for the unexplained spatial-temporal variations. The second aspect is about more scalable modeling of massive ecological data. Ecological data can be "big" both spatially and temporally. The proposed model is more computationally efficient than for existing models. Although the dimension reduction technique we use has been developed in the literature, its application has not been fully investigated for modeling and analyzing large ecological count data with more relaxed assumptions on the spatial-temporal correlation structure.

In the current study, we develop a model for zero-inflated count data that provides flexibility in modeling spatial patterns in a dynamic manner and also improves the computational efficiency via dimension reduction. Salazar et al. (2011) studied temperature data from a group of regional climate models using the spatial dynamic factor model approach and the spatial loading matrix was constructed based on the Gaussian predictive process approach of Banerjee et al. (2008). Built on their work, which focused solely on Gaussian data, we extend this approach to non-Gaussian zero-inflated count data and propose a hierarchial spatial-temporal ZIP model. More generally, the proposed methodology offers a modeling framework that is particularly suitable and computationally scalable to large ecological count data.

The rest of the paper is organized as follows. In Section 2, we present the detailed development of the proposed Bayesian spatial-temporal model. In Section 3, we discuss prior specification, Bayesian computation, and model comparison criteria. A detailed analysis of the Atlantic cod data is carried out in Section 4. Along with the proposed model, a sequence of models have been tested with different specifications on the spatial and temporal random errors. Models are compared using the deviance information criterion (DIC) and the log predictive score. We conclude the paper with a brief discussion on potential further improvements of the proposed model and future work in Section 5.

2. The Model

Suppose count data $y_{t,i}$ are collected from N uniform grid locations in the area of interest over T survey years. Let $E_{t,i}$ be the binary indicator of whether the species of interest is truly present at grid i in year t, $i = 1, \ldots, N$ and $t = 1, \ldots, T$. The presence status is unobservable if $y_{t,i} = 0$, and may be influenced by a rich collection of environmental variables. The data model is then usually specified as $\text{Prob}(Y_{t,i} = 0 | E_{t,i}) = 1$ if $E_{t,i} = 0$ and $\text{Prob}(Y_{t,i} = y_{t,i} | E_{t,i}) = \text{Poisson}(y_{t,i} | \lambda_{t,i})$ if $E_{t,i} = 1$, where $\text{Poisson}(y_{t,i} | \lambda_{t,i})$ is the probability mass function of a Poisson random variable $Y_{t,i}$ with $\mathbb{E}(Y_{t,i}) = \lambda_{t,i}$, $E_{t,i} = 1$ with probability $p_{t,i}$ and $E_{t,i} = 0$ with probability $1 - p_{t,i}$. It is assumed that, conditioned on $p_{t,i}$, the $E_{t,i}$'s are independent Bernoulli random variables with $\mathbb{E}(E_{t,i}) = p_{t,i}$. Given $E_{t,i} = 1$, the $Y_{t,i}$'s are conditionally independent.

The process models on $p_{t,i}$ and $\lambda_{t,i}$ are given in the framework of the generalized linear mixed model as $g(p_{t,i}) = \mathbf{x}'_{t,i}\boldsymbol{\beta}_{t,i} + w_{t,i}$ (binary part) and $\log(\lambda_{t,i}) = \tilde{\mathbf{x}}'_{t,i}\boldsymbol{\alpha}_{t,i} + \tilde{w}_{t,i}$(count part), where g is the link function for the binary regression, $\mathbf{x}_{t,i}$ and $\tilde{\mathbf{x}}_{t,i}$ are the vectors of covariates, which may be spatially and temporally related, $\boldsymbol{\beta}_{t,i}$ and $\boldsymbol{\alpha}_{t,i}$ are the vectors of the corresponding regression coefficients, and $w_{t,i}$ and $\tilde{w}_{t,i}$ are the random components.

Models for the count data differ in their specifications on the regression coefficients $\boldsymbol{\beta}_{t,i}$ and $\boldsymbol{\alpha}_{t,i}$ and the random components $w_{t,i}$ and $\tilde{w}_{t,i}$. Some studies have considered spatial random components in either the binary part

(Fei and Rathbun (2006)) or the count part (Agarwal, Gelfand, and Citron-Pousty (2002)). Spatial-temporal models have also been developed to accommodate the inherent spatial and temporal nature of the data. Wikle and Anderson (2003) included spatially varying coefficients on the year index and yearly ENSO effect based on previous evidence. They assumed spatially and temporally independent random errors $w_{t,i}$ and temporally independent and spatially correlated errors $\tilde{w}_{t,i}$. Similarly, in the work of Fernandes, Schmidt, and Migon (2009), temporally independent spatial random errors were introduced in the count part. Instead of using the temporal random errors, they modeled the temporal dynamics by including indicators of lag variables in $\mathbf{x}_{t,i}$ and $\tilde{\mathbf{x}}_{t,i}$. The regression coefficients $\boldsymbol{\beta}_{t,i}$ and $\boldsymbol{\alpha}_{t,i}$ were temporally and spatially invariant. The model of Ver Hoef and Jansen (2007) included spatial and temporal random errors in both the binary part and the count part. The spatial and temporal processes were assumed to be additive and independent from each other, $w_{t,i} = e_{1,t} + e_{2,i}$, where $e_{1,t}$ is specified by a first-order autocorrelation process and $e_{2,i}$ by a conditional autoregressive (CAR) model. The two random components $e_{1,t}$ and $e_{2,i}$ were independent from each other. Similar structures were assumed on $\tilde{w}_{t,i}$. This additive assumption may not be appropriate when the spatial correlation is likely to change over time.

Our proposed model is constructed with more relaxed assumptions on the spatial-temporal random errors that also allows for efficient dimension reduction. Assume that there exists a latent process, $\mathbf{Z}_t = (Z_{t,1}, \ldots, Z_{t,N})'$ across the N grids in year t, assumed to be independent across years. We assume that the $\mathbf{Z}_t$'s are temporally independent in the belief that the probability of presence or absence, unlike abundance, is relatively stable over time. In applications where the probability in the binary part is likely to be temporally correlated, a dynamic pattern in $\mathbf{Z}_t$ can be modeled as in Ver Hoef and Jansen (2007). By including the latent process, the absence of a species is generated corresponding to the cases where the latent variable falls below a threshold (Albert and Chib (1993)). Without loss of generality, the threshold is set at 0, $E_{t,i} = 1$ if $Z_{t,i} > 0$ and $E_{t,i} = 0$ if $Z_{t,i} \leq 0$. Thus the sign of the latent random variable $Z_{t,i}$ indicates the true presence or absence status of the species. The distribution of the latent variable may depend on certain observable and unobservable environmental factors. Let $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_N)'$ be the $N \times p$ covariate matrix including an intercept. Then, we take $Z_{t,i} = \mathbf{x}_i'\boldsymbol{\beta} + \omega_{t,i}$, where $\boldsymbol{\beta}$ is a $p \times 1$ vector of regression coefficients, and $\Omega_t = (\omega_{t,1}, \ldots, \omega_{t,N})$ is used to incorporate the unobservable and spatially correlated environmental factors that influence the presence of the species in year t. We employ the CAR model specified in Cressie (1993): $\Omega_t \sim \mathrm{MVN}(\mathbf{0}, \sigma^2(\mathbf{I} - \phi\mathbf{W})^{-1})$, where $\mathrm{MVN}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ is a multivariate normal distribution with a mean vector $\boldsymbol{\mu}$ and a variance-covariance matrix $\boldsymbol{\Sigma}$, σ^2 is the spatial nugget parameter, ϕ is the spatial range parameter, and $\mathbf{W}$ is the adjacency matrix. The diagonal elements of $\mathbf{W}$ are $\mathbf{w}_{ii} = 0$, while the off-diagonal elements $\mathbf{w}_{i\ell} = 1$ if grids i and ℓ are neighbors and $\mathbf{w}_{i\ell} = 0$ if they are not $(i \neq \ell)$. The neighborhoods of two grids are defined according to a second-order neighbor definition (Banerjee, Carlin, and Gelfand (2004)). To make $(\mathbf{I} - \phi\mathbf{W})^{-1}$ nonsingular, we assume $\phi \in (1/\theta_{(1)}, 1/\theta_{(N)})$, where $\theta_{(1)} < \theta_{(2)} < \ldots < \theta_{(N)}$ are the ordered eigenvalues of $\mathbf{W}$. To ensure identifiability, σ^2 is set to 1 (De Oliveira (2000); Fei and Rathbun (2006)). Thus, the latent process $\mathbf{Z}_t$ is the multivariate normal $\mathrm{MVN}(\mathbf{X}'\boldsymbol{\beta}, (\mathbf{I} - \phi\mathbf{W})^{-1})$.

The binary part of the model under a Probit link function is $\mathrm{Probit}[\mathbb{E}(E_{t,i} = 1)] = \Phi^{-1}(p_{t,i}) = \mathbf{x}'_{t,i}\boldsymbol{\beta} + \omega_{t,i} = Z_{t,i}$, where $\Phi^{-1}(\cdot)$ is the inverse of the cumulative standard normal distribution function. As $Y_{t,i}$ is the count at grid i in year t and $Y_{t,i} \sim \mathrm{Poisson}(\lambda_{t,i}|E_{t,i} = 1)$, the count part of the model is

$$\log[\mathbb{E}(Y_{t,i}|E_{t,i} = 1)] = \log(\lambda_{t,i}) = \tilde{\mathbf{x}}'_{t,i}\boldsymbol{\alpha} + \mathbf{D}'_i\boldsymbol{\gamma}_t, \tag{2.1}$$

where the basis function $\mathbf{D}$ is constructed using the predictive process of Banerjee et al. (2008): $\mathbf{D} = [\mathbf{D}(\mathbf{s})]_i = \mathbf{V}(\mathbf{s})'\mathbf{H}^{-1}$, $\mathbf{V}(\mathbf{s}) = \tau^2(\mathbf{v}(\mathbf{s}, \mathbf{s}_1^*; \boldsymbol{\phi}), \ldots, \mathbf{v}(\mathbf{s}, \mathbf{s}_M^*; \boldsymbol{\phi}))$, $\mathbf{H}_{lk} = \tau^2\mathbf{v}(s_l^*, s_k^*; \boldsymbol{\phi})$, $s_1^*, \ldots, s_M^*$ are the selected knots in the surveyed area, and $\mathbf{v}(\cdot;\cdot)$ is a valid correlation function.

In the count part of ZIP model, the spatial correlation is only estimated using the data from grids with $Z_{t,i} > 0$. The CAR model as specified in the binary part can lead to one or a group of isolated grids. These grids do not have any neighborhoods and thus are assumed spatially independent from the rest of the region. We find this unsatisfactory. The previous spatial-temporal ZIP models used either a continuous correlation function (Wikle and Anderson (2003); Fernandes, Schmidt, and Migon (2009)) or a CAR model with an arbitrary cutoff distance to define neighborhood (Ver Hoef and Jansen (2007)) in the count part. We propose to use the Matérn correlation function specified as $(\Gamma(\phi_2)2^{(\phi_2-1)})^{-1}$ $(2\phi_2^{1/2}d(\mathbf{s},\mathbf{s}')/\phi_1)^{\phi_2}\mathcal{K}_{\phi_2}(2\phi_2^{1/2}d(\mathbf{s},\mathbf{s}')/\phi_1)$, where $\mathcal{K}_{\phi_2}$ is a modified Bessel function of the second kind of order ϕ_2, $d(\mathbf{s},\mathbf{s}')$ is the Euclidean distance between two locations $\mathbf{s}$ and $\mathbf{s}'$, ϕ_1 is the range parameter which measures how fast the correlation decays with distance, and ϕ_2 is the smoothness parameter that measures the degree of smoothness of the spatial process. The higher the value of ϕ_2, the smoother the spatial process would be. A continuous correlation structure makes it possible to investigate "hot spots" or "cold spots" effects.

The evolution of $\boldsymbol{\gamma}_t$ is specified as $\boldsymbol{\gamma}_t = \rho\boldsymbol{\gamma}_{t-1} + \mathbf{v}_t$, where $\mathbf{v}_t \sim N(0, \mathbf{H})$ and $-1 < \rho < 1$. An interesting result was derived based on the spectral decomposition of $\mathbf{H}$ (Salazar et al. (2011)): $\mathbf{H} = \tau^2\mathbf{P}\boldsymbol{\Lambda}\mathbf{P}$, where $\mathbf{P}$ is an orthogonal matrix and $\boldsymbol{\Lambda}$ is a diagonal matrix with the eigenvalues of $\mathbf{H}/\tau^2$ as the diagonal elements. Letting $\boldsymbol{\gamma}_t = \mathbf{P}\boldsymbol{\xi}_t$ for all t, we have

$$\mathbf{D}(\mathbf{s})\mathbf{P}\boldsymbol{\xi}_t = \boldsymbol{\Psi}(\mathbf{s})\boldsymbol{\xi}_t,\ \ \boldsymbol{\xi}_t \sim N(\rho\boldsymbol{\xi}_{t-1}, \tau^2\boldsymbol{\Lambda}), \tag{2.2}$$

where $\boldsymbol{\Psi}(\mathbf{s}) = \mathbf{D}(\mathbf{s})\mathbf{P} = \mathbf{V}(\mathbf{s})'\mathbf{H}^{-1}\mathbf{P}$ and $\boldsymbol{\xi}_t = \rho\boldsymbol{\xi}_{t-1} + \boldsymbol{\upsilon}_t$ with $\boldsymbol{\upsilon}_t \sim N(0, \tau^2\boldsymbol{\Lambda})$ and $\boldsymbol{\xi}_0 \sim N(\mathbf{m}_0, \mathbf{C}_0)$. As a result, the temporal changes in spatial patterns can be modeled as M independent processes $\tilde{\boldsymbol{\xi}}_i = \{\xi_{1,i}, \xi_{2,i}, \ldots, \xi_{t,i}, \ldots, \xi_{T,i}\}$, $i = 1, \ldots, M$.

The appropriate temporal correlation structure in $\boldsymbol{\gamma}_t$ (also $\boldsymbol{\xi}_t$) can be explored using autocorrelation function (ACF) and partial autocorrelation function (PACF) plots. Given the sparsity of our data in each individual grid across years, it is impossible to examine the temporal sequence of counts at a given location. We investigated, for example, the average number of Atlantic cod caught in a single tow in the Gulf of Maine - Georges Bank region during 1970-2008 (See Figure S2 in Supplementary Material). The PACF plot suggests that the first order autocorrelation may exist in the temporal sequence of mean count of fish caught in a tow averaged over all the locations. Thus, we chose to use the first-order polynomial model structure. This type of dynamic model has been commonly used in applications (West and Harrison (1997)).

Let $\mathbf{Y} = (y_{1,1}, \ldots, y_{1,N}, \ldots, y_{T,1}, \ldots, y_{T,N})'$ and $\mathbf{Z} = (\mathbf{Z}_1', \mathbf{Z}_2', \ldots, \mathbf{Z}_T')'$. Then the likelihood function of $(\mathbf{Y}, \mathbf{Z})$ is given by

$$
\begin{aligned}
&L(\mathbf{Y}, \mathbf{Z}|\boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{\xi}_t, \rho, \phi, \phi_1, \phi_2, \tau^2) \\
&= (2\pi)^{-NT/2}|\boldsymbol{\Sigma}(\phi)|^{-T/2} \exp\Big\{ -\frac{1}{2}\sum_{t=1}^{T}(\mathbf{Z}_t - \mathbf{X}'\boldsymbol{\beta})'\boldsymbol{\Sigma}^{-1}(\phi)(\mathbf{Z}_t - \mathbf{X}'\boldsymbol{\beta})\Big\} \\
&\quad\times \prod_{t=1}^{T}\prod_{i=1}^{N}\left[I(Z_{t,i} \le 0)I(y_{t,i} = 0) + I(Z_{t,i} > 0)(y_{t,i}!)^{-1}\exp\{\eta_{t,i}y_{t,i} - \exp(\eta_{t,i})\}\right],
\end{aligned}
$$

where $\eta_{t,i} = \tilde{\mathbf{x}}_{t,i}'\boldsymbol{\alpha} + \boldsymbol{\Psi}(s_i; \phi_1, \phi_2)\boldsymbol{\xi}_t$. The data augmentation method of introducing the auxiliary variables $\mathbf{Z}$ in the model avoids high dimensional integral challenges and facilitates efficient Markov chain Monte Carlo (MCMC) computation (Chib and Greenberg (1998); Chen, Dey, and Shao (1999); Pettitt, Weir, and Hart (2002); Fei and Rathbun (2006)).

3. Prior Specification, Posterior Computation and Model Assessment

In the Matérn correlation function, the smoothness parameter ϕ_2 is typically assigned a uniform prior $U(0, 2)$ because the data cannot inform about the smoothness in the higher order (Finley et al. (2009)). We assume the parameter $\phi_2 = 1$ to avoid the weak identification problem. The subclass of the Matérn correlation function with $\phi_2 = 1$ was introduced by Whittle (1954) as the "elementary" model for two-dimensional fields. It has been commonly used in hydrology (Handcock and Stein (1993)).

The parameters that need to be estimated in the proposed model include $\boldsymbol{\beta}, \boldsymbol{\alpha}, \phi, \tau^2, \phi_1$, and ρ. The priors for these parameters are specified as follows:

$\boldsymbol{\beta} \sim N(0, g_{\boldsymbol{\beta}}(\mathbf{X}'\mathbf{X})^{-1})$, where $g_{\boldsymbol{\beta}} = 1,000$; $\boldsymbol{\alpha} \sim N(0, g_{\boldsymbol{\alpha}} I)$, where $g_{\boldsymbol{\alpha}} = 1,000$; $\phi \sim U(\phi_{\min}, \phi_{\max})$, which ensures the positive definiteness of the variance matrix $(\mathbf{I} - \phi\mathbf{W})^{-1}$; $\tau^2 \sim IG(c, d)$, where $c = 2$ and $d = 1$; $\phi_1 \sim IG(2, h)$, where $h = \max(d(\mathbf{s}, \mathbf{s}')/(-2\log(0.05))$, and $\rho \sim U(-1, 1)$. The hyper-parameters are chosen to ensure that relatively noninformative priors are used in Bayesian estimation.

As discussed in West and Harrison (1997), the initial prior on $\boldsymbol{\xi}_0$ contains a concrete interpretation of the final state vector for the historical data if it is a summary of information from the past. When there is no such information and interpretation, the model may be equivalently initialized by specifying a normal prior for the first state vector. We adopt the second approach here since there is no prior information available on $\boldsymbol{\xi}_0$ in our application. For the initial prior $\boldsymbol{\xi}_1 \sim N(\mathbf{m}_1, \mathbf{C}_1)$, the hyperparameters $\mathbf{m}_1$ and $\mathbf{C}_1$ are specified as a $M \times 1$ vector $\mathbf{0}$ and an $M \times M$ diagonal covariance matrix $10^6 \cdot \mathbf{I}$.

The posterior estimates of the parameters are computed via MCMC sampling. The detailed development of the MCMC sampling algorithm is given in Section S1 in Supplementary Material. The convergence of MCMC chains is tested using the `R boa` package (Smith (2007)).

The deviance information criterion (DIC) is considered for model comparison. Using the methods in Hadfield (2010), the parameters in the deviance function include the Bernoulli probabilities and the Poisson means $(p_{t,i}, \lambda_{t,i})$ for $i = 1, \ldots, N$ and $t = 1, \ldots, T$. The deviance function is negative two times the logarithm of the likelihood, $-2\sum_{t=1}^{T}\sum_{i=1}^{N}\log\{(1 - p_{t,i})I(y_{t,i} = 0) + p_{t,i}\text{Poisson}(y_{t,i}|\lambda_{t,i})\}$.

Further model comparison and evaluation can be performed based on the predictive model assessment (Czado, Gneiting, and Held (2009)) for dynamic models. In particular, the logarithm predictive scores are used in the comparison of spatial-temporal models. These scores are calculated using the methods proposed by (Abanto-Valle and Dey (2014)) via the particle learning algorithm. This utilizes a resample-propagate scheme together with a particle set that includes state-sufficient statistics in the estimation of both static and state parameters in the dynamic linear model (Carvalho et al. (2010)). The details of the particle learning algorithm and the log predictive score are discussed in Section S2 in Supplementary Material.

4. Application: Abundance of Atlantic Cod in the Gulf of Maine - Georges Bank Region

In this study, we are particularly interested in the Atlantic cod population in the Gulf of Maine - Georges Bank region. The Atlantic cod (*Gadus morhua*) is an important species in many of the world's ocean systems from an economic, ecological, and cultural perspective. It is a mainstay of the commercial fishery in New England, and its history can be traced back to the 1600s. In fact, fishing for the Atlantic cod was such a key industry for much of coastal Massachusetts that a carving of a codfish named the Sacred Cod is hung in the House of Representatives' chamber of the Massachusetts State House. The cod also plays an important role in the ocean ecosystems, as mentioned in Link et al. (2009), by the interactions with its prey, predators, and competitors. The cod population experienced dramatic changes with a collapse in the 1990s. It has not been fully recovered even with the cessation of fishing. The cod species is labeled VU (vulnerable) on the International Union for Conservation of Nature Red List of Threatened Species.

The Gulf of Maine - Georges Bank region has very complex geology and oceanology. It is one of the most varied and productive marine ecosystems in the world. The cods in these areas are relatively residential and no apparent intensive migration patterns are observed (O'Brien et al. (2005)). A better understanding of the cod's population distribution would help management of the marine protected areas in the region.

4.1. The data

We use the survey data collected by the Northeast Fisheries Sciences Center (NEFSC). NEFSC has been carrying out a standardized research survey since Fall, 1963. Nowadays the survey is on a regular basis four times a year for spring, summer, fall and winter seasons. The survey area ranges from the Gulf of Maine to Cape Hatteras, NC. Approximately 350-400 stations are surveyed during each survey season, with locations selected by a stratified random sampling design to assure that the number of stations allocated to strata are roughly in proportion to area. Samples are collected in depths of 27 meters to 350 meters with 4 depth zones. Data recorded on site include the species caught, weight, counts of fish, surface and bottom water temperature, bottom depth of the tow, along with many other variables. Previous studies on this survey data have been conducted mainly based on the aggregated data either across time or space (Drinkwater (2005); Fogarty et al. (2008); Kuo, Auster, and Parent (2010)). The current study is the first to apply a Bayesian spatial-temporal model in understanding the presence and abundance of Atlantic cods stocks.

The study area is divided into 1,325 ($N = 1,325$) 10 km by 10 km grids. The survey data used here were collected in the fall of 1970-2008 ($T = 39$). The ocean geographical characteristics included in the model are the average depth of the ocean and the depth standard deviation (see Figure S1) that show wide variation in the area. The latitude and the sampling year are considered in the model. All the covariates are standardized by their corresponding means and standard deviations to improve MCMC convergence. The average depth is originally measured with respect to the sea level and the smaller the value (which is negative), the deeper the sea. To ease interpretation, we use its positive value and the larger the value of the average depth (which is now positive), the deeper the sea.

There are 4,863 tows in the study region 1970-2008, out of which 2,746 tows did not have Atlantic cod (56.47% zeros). Zero counts may not be a true indicator of the absence of the fish at a given grid. They may be a result of incomplete coverage and imperfect detection.

The temporal pattern can be roughly examined in Figure S2, which shows the average number of fish caught in a tow during 1970-2008. An examination of this sequence of counts suggests a first-order autocorrelation model.

The approximate spatial pattern is shown by aggregating counts at each grid over years. Figure S3(a) shows the average number of fish caught in each tow at each grid through 1970-2008. Based on this measure, Atlantic cod are relatively more abundant in the northeast, west and to the north of the Georges Bank in the study area. Figure S3(b) shows the standard deviation of the number of fish caught during the study period. Clearly, the standard deviation increases as the average number of fish caught increases. Also, there are a few locations that were surveyed only once. Statistical modeling allows information borrowing from neighborhood locations for inference on those locations. A tow at Grid 88 had 350 fish caught in 2002, and at most 1 caught in all other years. In the likelihood calculation, the Poisson probability of this observation is zero. It was treated as an outlier and set to 1.

4.2. The results

Besides the proposed model, a sequence of existing ZIP models have been fitted with different model specifications on spatial and temporal random errors. Model comparisons show that the incorporation of spatial-temporal random processes benefits the goodness of fit in modeling the data.

A simple zero-inflated Poisson model is first fitted for comparison without considering the spatial and temporal correlations (Model 1). A total of 4,863 tows sampled at 1,222 grids were used in estimating the regression coefficients in the binary and count parts. The posterior estimates of the model parameters and the DIC values are reported in Table 1 under the column for Model 1. Figure S4(a) and S4(b) shows the estimated posterior mean count at each grid and the standard deviation of the estimate. The standard deviations range from 0 to 1.54. Out of 1,222 grids, 1,217 grids have standard deviations less than 0.6. Thus, in Figure S4(b), the standard deviation is capped at 0.7 in the figure for legibility. The blank grids are those that were not surveyed during the study period. There are consistently over- and under-estimated areas as shown in the residual plots (Figure S4(c)).

Model 2 includes the survey year as a covariate in the Poisson part of the model to capture the general temporal trend. As shown in Table 1, the expected count of fish caught deceases between year 1970 and 2008. Also, the inclusion of the time effect in the mean decreases the DIC value, which implies a better fit to the data.

Starting from Model 3, a CAR spatial correlation structure was assumed for the latent process $Z_{t,i}$. The DIC value decreases from 41,891.21 to 40,954.23, implying that it improves the goodness of fit by modeling the spatial correlation in the probability of presence. As noted in Reich, Hodges, and Zadnik (2006) and Hughes and Haran (2013), there is a potential confounding problem for the spatial generalized linear mixed model with CAR random effects, when the fixed effects include spatially related covariates. The examination of the posterior estimates of the regression coefficients in the binary part in Models 1−3 indicates that confounding is not a problem in this study. When the inflation of variance and bias occur, the methods proposed in Hughes and Haran (2013) can be employed to deal with the confounding problem for large ecological data.

Models 4−7 include the spatial correlation in the Poisson count part, with the temporal correlation modeled in Model 5 and Model 7. In modeling the spatial correlation, $M = 16$ and 32 knots were selected (see Figure S5). The posterior estimates are shown in Table 2. The smaller DIC values show that it improves model fitting by modeling spatial and temporal correlations in the counts of fish, and that it provides better inference from the model.

To further check how the spatial dynamic model can improve statistical inference, Figure 1 provides a snapshot of the survey data in year 1978, 1988, 1998, and 2008. The maximum numbers of fish caught in these four years were 132, 115, 35, and 201, respectively. The count of fish caught was capped at 40 in the figure.

Figures 2 and S6 show the posterior mean counts of fish at each grid and the standard deviations over the above 4 years estimated under Model 3, Model 6 and Model 7. The mean count of the fish is $p_{t,i}\lambda_{t,i}$ and the corresponding variance is $p_{t,i}\lambda_{t,i}+p_{t,i}(1-p_{t,i})\lambda_{t,i}^2$. The posterior mean count plot shows that the distribution of the Atlantic cods is related to the ocean geographic characteristics (mean depth and the standard deviation), while there is an obvious decreasing trend in the abundance of the fish. The results under Model 3 and Model 6 suggest a similar spatial pattern of the fish abundance across years except for the overall decreasing trend (the results under Model 4 are similar and thus are omitted here). In addition to the temporal correlation, spatial patterns under Model 7 were allowed to be different from year to year (the results from Model

Table 1. Posterior estimates under Models 1−3. DIC is the deviance information criterion; −2llike is the deviance evaluated at the posterior means of parameters; p_D is the effective number of model parameters in DIC. Model 1: the ZIP model on data combined across years and locations. Model 2: the ZIP model on the data combined across locations with year as a covariate (time) in the count part. Model 3: the ZIP model with the spatial correlation on the binary part and with time as a covariate in the count part.

Variables	Model 1		Model 2		Model 3	
	est.	SD	est.	SD	est.	SD
Binary Part						
intercept	-0.090	0.023	-0.076	0.024	0.015	0.052
mean depth	-0.216	0.029	-0.202	0.029	-0.214	0.039
(mean depth)^2	-0.141	0.037	-0.136	0.038	-0.148	0.047
sd.of depth	0.134	0.020	0.128	0.020	0.123	0.026
latitude	0.481	0.024	0.485	0.024	0.568	0.034
ϕ	-	-			0.1252	0.00013
Count Part						
intercept	1.609	0.016	1.558	0.017	1.559	0.016
mean depth	-1.095	0.018	-1.099	0.019	-1.101	0.018
(mean depth)^2	-0.133	0.021	-0.152	0.022	-0.154	0.022
sd.of depth	0.347	0.009	0.352	0.009	0.351	0.009
latitude	-0.064	0.010	-0.066	0.010	-0.065	0.010
time	-	-	-0.117	0.008	-0.114	0.008
ϕ_1	-	-				
τ^2	-	-				
ρ	-	-				
DIC	42110.66		41891.21		40954.10	
−2llike	42089.87		41868.23		39943.15	
p_D	10.39		11.49		505.48	

5 are similar). This flexibility provides a much better fit of the model (the DIC value reduces from 34,165.00 under Model 6 to 20,769.90 under Model 7). The results under Model 7 indicate that, despite the overall decreasing trend, some areas may have a significant increase in the abundance (such as the west area in year 2008).

The comparisons between Models 4 and 6 and between Models 5 and 7 show that the models with the 32 knots fit the data better than those with the 16 knots. The values of the log predictive score criterion are -5,968.998 under Model 5 and -5,760.557 under Model 7. These results indicate that the 32-knot model also has a better one-year ahead predictive performance. The models with different numbers of knots (i.e., 16, 32, 56, 64, 150) have been examined. The results suggest that the model performance does not necessarily always improve with a

Table 2. Posterior estimates under Models 4−7. DIC is the deviance information criterion; −2llike is the deviance evaluated at the posterior means of parameters; p_D is the effective number of model parameters in DIC; L_p is the log predictive score. Model 4: the ZIP model with spatial correlation on the binary part and the count part and with time as a covariate in the count part. For the predictive process, 16 knots were selected as shown in the left panel of Figure S5. Model 5: the ZIP model with spatial correlation on the binary part and with spatial-temporal correlation in the count part and with time as a covariate in the count part. Knots were selected as in Model 4. Model 6: Similar to Model 4 except that 32 knots were selected as shown in the right panel of Figure S5. Model 7: Similar to Model 5 except that 32 knots were selected as in Model 6.

Variables	Model 4		Model 5		Model 6		Model 7	
	est.	SD	est.	SD	est.	SD	est.	SD
Binary Part								
intercept	0.085	0.048	0.268	0.051	0.128	0.048	0.540	0.052
mean depth	-0.262	0.039	-0.254	0.045	-0.253	0.039	-0.274	0.051
$(\text{mean depth})^2$	-0.281	0.050	-0.233	0.055	-0.346	0.051	-0.335	0.057
sd.of depth	0.218	0.032	0.195	0.034	0.235	0.033	0.216	0.039
latitude	0.289	0.039	0.358	0.046	0.171	0.044	0.201	0.057
ϕ	0.1250	0.0002	0.1251	0.0002	0.1250	0.0002	0.1248	0.0004
Count Part								
intercept	-0.783	0.219	0.359	0.169	-1.350	0.132	-0.885	0.088
mean depth	-0.921	0.021	-1.019	0.024	-0.960	0.023	-1.052	0.028
$(\text{mean depth})^2$	-0.086	0.024	-0.132	0.030	0.038	0.023	0.121	0.026
sd.of depth	0.246	0.012	0.218	0.014	0.244	0.012	0.235	0.016
latitude	-0.680	0.112	-0.493	0.082	0.106	0.097	0.056	0.120
time	-0.167	0.008	-0.640	0.131	-0.173	0.008	0.080	0.178
ϕ_1	92.330	5.580	66.351	1.446	67.677	0.859	62.103	0.127
τ^2	6.190	2.220	8.061	0.753	2.575	0.662	15.949	1.111
ρ			0.419	0.031			0.153	0.022
DIC	34918.73		25086.17		34165.00		20769.90	
-2llike	33961.71		23431.25		33210.19		18796.66	
p_D	478.51		827.46		477.41		986.62	
L_p			-5968.998				-5760.557	

larger number of knots. More importantly, the computation stability becomes problematic when the number of knots becomes large. The determination of the optimal number of knots is challenging, and under investigation currently.

5. Discussion

The selection of the number and locations of knots is important in modeling

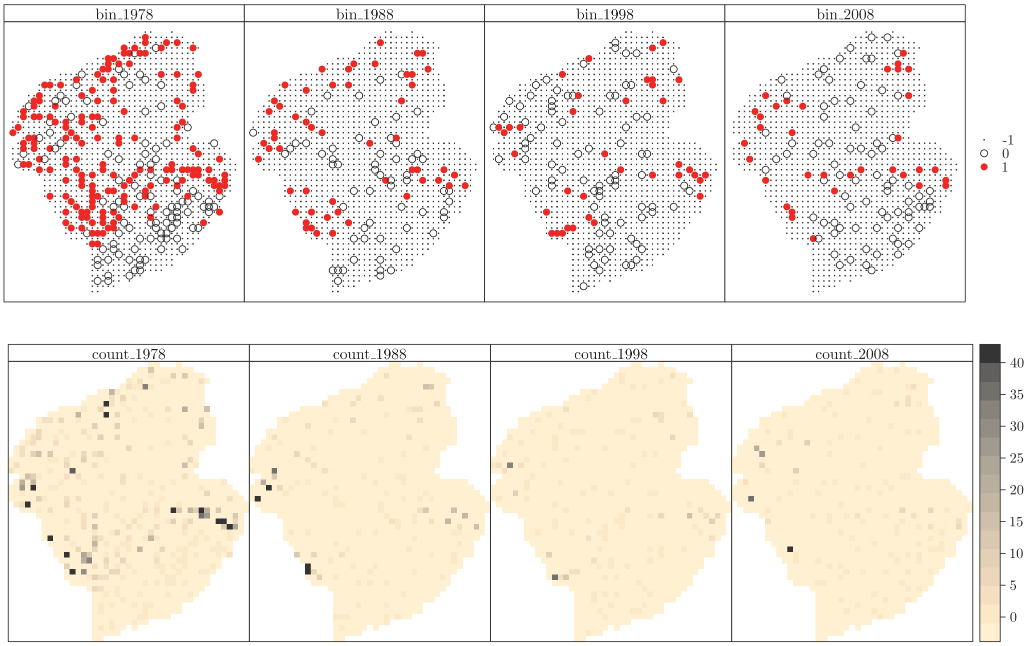

Figure 1. A 40-year snapshot of survey locations, presence and abundance for the Atlantic cods. Upper panel: surveyed grids with fish caught (solid circle (1)), surveyed grids without fish caught (empty circle (0)) and grids that were not surveyed (dot (-1)). Lower panel: counts of fish caught in the surveyed grids.

the spatial data. The results discussed in Section 4 indicate that it has an impact on the model inference. We chose to use evenly spaced knots with arbitrarily selected locations in this study. This may have missed some spatial structure that exists with a distance smaller than the distance between two knots. It is possible that the selected locations do not provide the optimal approximation of the parent process. To further improve the spatial-temporal modeling, the selection of the optimal number and their optimal locations of knots need to be investigated. Finley et al. (2009) designed an algorithm to achieve approximately optimal spatial placement of knots by minimizing spatially averaged prediction variance. It may also be possible to estimate the number and locations of knots using the reversible jump MCMC algorithm (Lopes, Gamerman, and Salazar (2011)).

The NEFSC survey collects data on multiple species simultaneously. It provides an opportunity to further investigate how the presence and abundance of different species are correlated. Multivariate spatial-temporal modeling is expected to be helpful in studying this aspect of the ecological system. It is possible to extend the proposed methodology to model multivariate zero-inflated count

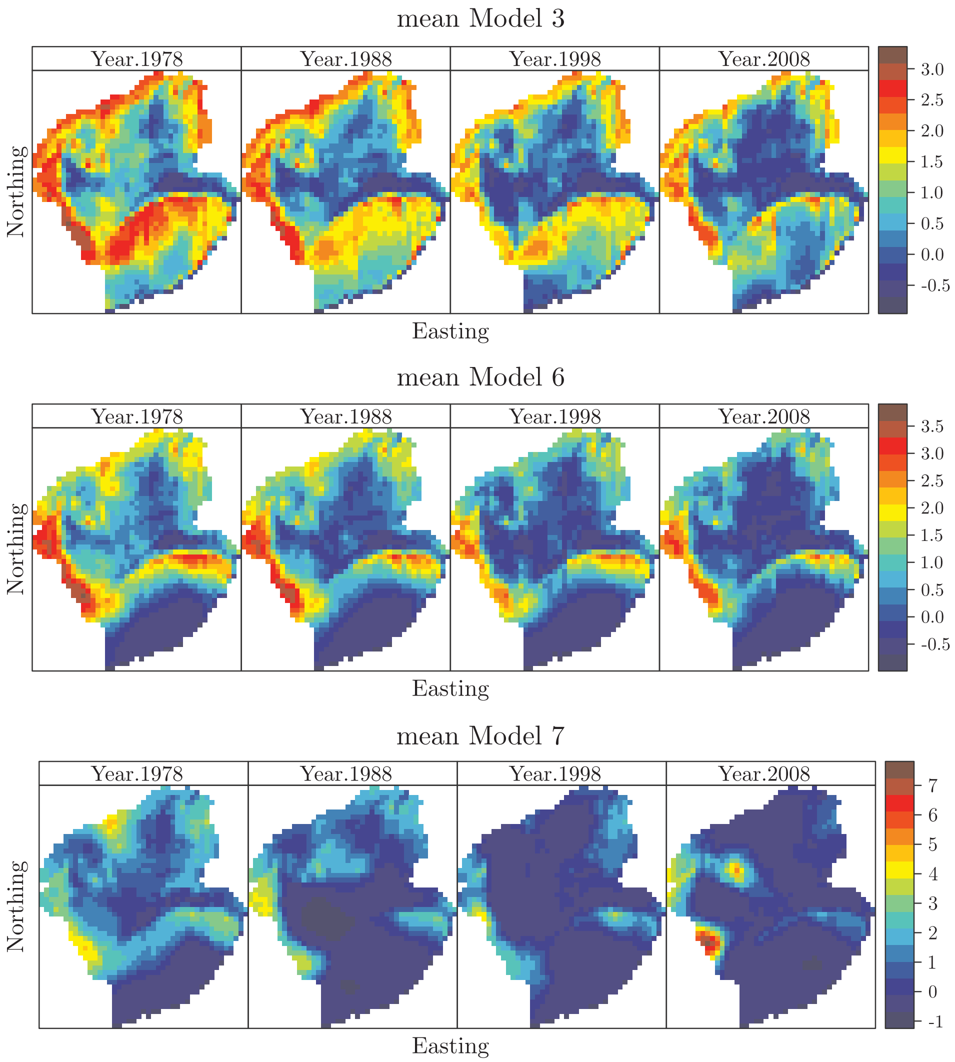

Figure 2. The logarithm of posterior mean count from Model 3, Model 6 (32 knots spatial), Model 7 (32 knots spatial-temporal) at each grid.

data. However, such an extension is not straightforward, requiring a much more in-depth investigation from both theoretical and computational perspectives.

Discussion on other potential further improvements can be found in Section S3 in Supplementary Material.

Acknowledgement

We thank the Guest Editor and the referee for their helpful comments and suggestions, which have led to a much improved version of the paper. Dr.

Chen's research was partially supported by US National Institutes of Health under Grants #GM 70335 and #CA 74015. Dr. Wang thanks the travel support provided by the Faculty Development Council at the University of Cincinnati. This project benefits from the Statistical and Applied Mathematical Sciences Institute (SAMSI) 2009-2010 and 2012-2013 programs.

References

Abanto-Valle, C. A. and Dey, D. K. (2014). State space mixed models for binary responses with scale mixture of normal distributions links. *Comput. Statist. Data Anal.* **71**, 274-287.

Agarwal, D. K., Gelfand, A. E. and Citron-Pousty, S. (2002). Zero-inflated models with application to spatial count data. *Environ. Ecol. Stat.* **9**, 341-355.

Albert, J. H. and Chib, S. (1993). Bayesian analysis of binary and polychotomous response data. *J. Amer. Statist. Assoc.* **88**, 669-679.

Banerjee, S., Carlin, B. P. and Gelfand, A. E. (2004). *Hierarchical Modeling and Analysis for Spatial Data.* Chapman & Hall/CRC, Boca Raton.

Banerjee, S., Gelfand, A., Finley, A. and Sang, H. (2008). Gaussian predictive process models for large spatial datasets. *J. Roy. Statist. Soc. Ser. B* **70**, 825-848.

Carvalho, C. M., Johannes, M. S., Lopes, H. F. and Polson, N. G. (2010). Particle learning and smoothing. *Statist. Sci.* **25**, 88-106.

Chen M.-H., Dey D. K. and Shao, Q. M. (1999). A new skewed link model for dichotomous quantal response data. *J. Amer. Statist. Assoc.* **94**, 1172-1186.

Chib, S. and Greenberg, E. (1998). Analysis of multivariate probit models. *Biometrika* **85**, 347-361.

Cressie, N. (1993). *Statistics for Spatial Data.* Revised edition. Wiley, New York.

Czado, C., Gneiting, T. and Held, L. (2009). Predictive model assessment for count data. *Biometrics* **65**, 1254-1261.

De Oliveira, V. (2000). Bayesian prediction of clipped Gaussian random fields. *Comput. Statist. Data Anal.* **34**, 299-314.

Drinkwater, K. F. (2005). The response of Atlantic cod (*Gadus morhua*) to future climate change. *ICES J. Mar. Sci.* **62**, 1327-1337.

Fei, S. and Rathbun, S. L. (2006). A spatial zero-inflated Poisson model for oak regeneration. *Environ. Ecol. Statist.* **13**, 406-426.

Fernandes, M. V., Schmidt, A. M. and Migon, H. S. (2009). Modelling zero-inflated spatio-temporal processes. *Statist. Model.* **9**, 3-25.

Finley, A. O., Sang, H., Banerjee, S. and Gelfand, A. E. (2009). Improving the performance of predictive process modeling for large datasets. *Comput. Statist. Data Anal.* **53**, 2873-2884.

Fogarty, M., Incze, L., Hayhoe, K., Mountain, D. and Manning, J. (2008). Potential climate change impacts on Atlantic cod (Gadus morhua) off the northeastern USA. *Mitig. Adapt. Strat. Gl.* **13**, 453-466.

Hadfield, J. D. (2010). MCMC methods for multi-response generalized linear mixed models: The MCMCglmm R package. *J. Statist. Softw.* **33**, 1-22.

Handcock, M. S. and Stein, M. L. (1993). A Bayesian analysis of kriging. *Technometrics* **35**, 403-420.

Hughes, J. and Haran, M. (2013). Dimension reduction and alleviation of confounding for spatial generalized linear mixed models. *J. Roy. Statist. Soc. Ser. B* **75**, 139-159.

Kuo, C. Y., Auster, P. and Parent, J. (2010). Variation in planning-unit size and patterns of fish diversity: Implicaitons for design of marine protected areas. available at `http://sanctuaries.noaa.gov`.

Link, J. S., Bogstad, B., Sparholt, H. and Lilly, G. R. (2009). Trophic role of Atlantic cod in the ecosystem. *Fish Fish.* **10**, 58-87.

Lopes, H. F., Gamerman, D. and Salazar, E. (2011). Generalized spatial dynamic factor models. *Comput. Statist. Data Anal.* **55**, 1319-1330.

O'Brien, L., Gregory, L. R., Mayo, R. K., Hunt and J. J. (2005). Gulf of Maine and Georges Bank (NAFO Subareas 5 and 6). In *Spawning and Life History Information for North Atlantic Cod Stocks* (Edited by K. Brander), 95-104. Prepared by the ICES/GLOBEC Working Group on Cod and Climate Change, ICES Cooperative Research Report No. 274.

Pettitt, A. N., Weir, I. S. and Hart, A. G. (2002). A conditional autoregressive Gaussian process for irregularly spaced multivariate data with application to modelling large sets of binary data. *Statist. Comput.* **12**, 353-367.

Reich, B. J., Hodges, J. S. and Zadnik, V. (2006). Effects of residual smoothing on the posterior of the fixed effects in disease-mapping models. *Biometrics* **62**, 1197-1206.

Salazar, E., Sansó, B., Finley, A., Hammerling, D., Steinsland, I., Wang, X. and Delamater, P. (2011). Comparing and blending regional climate model prediction for the American southwest. *J. Agric. Biol. Environ. Statist.* **16**, 586-605.

Smith, B. J. (2007). boa: An R package for MCMC output convergence assessment and posterior inference. *J. Statist. Softw.* **21**, 1-37.

Ver Hoef, J. M. and Jansen, J. K. (2007). Space-time zero-inflated count models of harbor seals. *Environmetrics* **18**, 697-712.

West, M. and Harrison, J. (1997). *Bayesian Forecasting and Dynamic Models.* 2nd edition. Springer, New York.

Whittle, P. (1954). On stationary processes in the plane. *Biometrika* **41**, 434-449.

Wikle, C. K. and Anderson, C. J. (2003). Climatological analysis of tornado report counts using a hierarchical Bayesian spatiotemporal model. *J. Geophys. Res.- Atmos.* **108**, 9005-9019.

Zhuang, L. and Cressie, N. (2012). Spatio-temporal modeling of sudden infant death syndrome data. *Statist. Methodol.* **9**, 117-143.

Department of Mathematical Sciences, University of Cincinnati, 2815 Commons Way Cincinnati, OH 45221-0025, USA.

E-mail: xia.wang@uc.edu

Department of Statistics, University of Connecticut, 215 Glenbrook Road, U-4120 Storrs, CT 06269-4120, U.S.A.

E-mail: ming-hui.chen@uconn.edu

Joint Genome Institute, Lawrence Berkeley National Laboratory, Berkeley, CA 94720, USA.

E-mail: chykuo@gmail.com

Department of Statistics, University of Connecticut, 215 Glenbrook Road, U-4120 Storrs, CT 06269-4120, U.S.A.

E-mail: dipak.dey@uconn.edu

(Received July 2013; accepted February 2014)