Journal of Biometrics & Biostatistics
ISSN: 2155-6180
Zheng and Charnigo, J Biom Biostat 2012, 3:6
DOI: 10.4172/2155-6180.1000e112
Editorial Open Access

On Selecting Spatial-Temporal Autologistic Regression Models for Binary Lattice Data

Yanbing Zheng[1]* and Richard Charnigo[1,2]

[1]Department of Statistics, University of Kentucky, 725 Rose St, Lexington, KY 40536, USA

[2]Department of Biostatistics, University of Kentucky, 725 Rose St, Lexington, KY 40536, USA

Introduction

In many biological and physical sciences, rapid advances in technical capabilities have dramatically increased the amount of data that are collected across space and over time. Spatial-temporal models are important tools for the analysis of spatial data collected repeatedly over time and have been applied to a wide range of problems, including modeling patterns in lung cancer [1], breast cancer [2], birth defects [3], and West Nile virus [4]; see also Cressie [5], Rue and Held [6], and Schabenberger and Gotway [7]. In particular, for binary data that are observed on a spatial lattice over time, spatial-temporal autologistic regression models relate binary responses to covariates while accounting for spatial and temporal dependence simultaneously [8,9].

Spatial-Temporal Autologistic Regression Model

Let y_{it} denote the response variable such that $y_{it} = 0$ or 1 at site i and time t, where $i = 1,...,n$ and $t = 1,...,m$. Let $y_{1t} = (y_{1t},...,y_{nt})'$ denote the binary responses on the spatial lattice for a given time point t. We specify the joint distribution of $y = (y'_{s+1},...,y'_m)'$ via conditional distributions,

$$p(y_t \,|\, y_{t'} : t' = t-1, t-2, ...) = p(y_t \,|\, y_{t'} : t' = t-1, ..., t-s),$$

for $t = s+1,...,m$. Further, for a given time point t, we assume that the response variable follows an autologistic model

$$p(y_{it} \,|\, y_{jt} : j \neq i, y_{t'} : t' = t-1, ..., t-s = p(y_{it} \,|\, y_{jt} : j \in N(i), y_{it'} : j \in \{i \cup N(i)\}, t' \in \{t-1, ..., t-s\}),$$

where $y_{it} \,|\, y_{jt} : j \in N(i), y_{it'} : j \in \{i \cup N(i)\}, t' \in \{t-1,...,t-s\} \sim Bernoulli(p_{it})$, and

$$\log it(p_{it}) = \sum_{j=1}^{p} \beta_j x_{jit} + \sum_{k=1}^{q} \theta_k \sum_{j \in N_k(i)} y_{jt} + \sum_{l=1}^{s} \alpha_l y_{i(t-1)} + \sum_{k=1}^{q} \sum_{l=1}^{s} \theta_k^l \sum_{j \in N_k(i)} y_{j(t-1)}.$$

Here x_{jit} denotes the j^{th} covariate at site i and time t, $\beta = (\beta_1,...,\beta_p)'$ are regression coefficients $\theta = (\theta_1,...,\theta_q)'$ are spatial autoregressive coefficients, $\alpha = (\alpha_1,...,\alpha_s)'$ are temporal autoregressive coefficients, and $\theta^l = (\theta_1^l,...,\theta_q^l)'$ for $l = 1,...,s$ are spatial-temporal interactive coefficients. For a given site i, we can partition the neighborhood $N(i) = U_{k=1}^{q} N_k(i)$. For example, in the bark beetle infestation example of Zhu et al. [8], the study region is a regular square grid. Then we can define $N_k(i)$, the k^{th}-order neighbors of a given site i, to contain the k nearest neighbors in terms of distance, for $k = 1,...,q$. Taking $q = 2$ for example, we note that $\theta_1 \neq 0, \theta_2 = 0$ corresponds to spatial autocorrelation along the north-south and west-east directions, while $\theta_1 = 0, \theta_2 \neq 0$ corresponds to spatial autocorrelation along the northwest-southeast and northeast-southwest directions. Furthermore, to account for anisotropy, we could further partition $N_k(i)$ by direction as in Zhu et al. [10]. In general, the magnitude of θ_k reflects not only the extent but also the direction of spatial autocorrelation.

Some special cases of the above spatial-temporal autologistic regression models (Cf. Reyes [11]) are as follows:

- Spatial independence: $\theta_1 = \cdots = \theta_q = 0$ and all $\theta_k^l = 0, k = 1,...,q, l = 1,...,s$
- Temporal independence: $\alpha_1 = \cdots = \alpha_s = 0$ and all $\theta_k^l = 0, k = 1,...,q, l = 1,...,s$
- Spatial-temporal separable neighborhood structure: all $\theta_k^l = 0, k = 1,...,q, l = 1,...,s$
- Spatial-temporal non-separable neighborhood structure: some $\theta_k^l \neq 0, k = 1,...,q, l = 1,...,s$

In what follows, for simplicity we focus on the spatial-temporal separable neighborhood structure.

Model Selection

Some interesting statistical problems for autologistic regression models include how to select covariates and determine an appropriate spatial and temporal neighborhood structure. For example, in studying the impact of climate change on bark beetle infestation of pine forests in North America, some of the most important scientific objectives are to identify and quantify the effects of environmental conditions (e.g. climate change) on bark beetle infestation. Also of great interest is describing the extent and direction of bark beetle dispersal [12]. Judicious selection of covariates and spatial-temporal neighborhood structure permits fulfillment of the aforementioned scientific objectives.

For binary spatial-temporal lattice data, there is not a consensus on how to perform model selection. Particularly regarding spatial-temporal neighborhood structure, this lack of consensus has resulted in researchers employing creative but ad-hoc methods for which the statistical properties are not fully understood. For example, Zhu et al. [13] selected covariates using backward elimination based on t-ratios of the parameter estimates under a pre-specified spatial and temporal neighborhood structure for their analysis of the southern pine beetle outbreak in North Carolina, United States. Zhu et al. [8] pre-selected the spatial and temporal neighborhood structure without including covariates using the AIC and then, once the neighborhood structure was specified, chose covariates for their analysis of the mountain pine beetle outbreak in British Columbia, Canada. Using pre-selected covariates, Bandyopadhyay et al. [9] employed a Bayesian paradigm

***Corresponding author:** Yanbing Zheng, Department of Statistics, University of Kentucky, 725 Rose St, Lexington, KY 40536, USA, E-mail: yanbing.zheng@uky.edu

Received August 07, 2012; **Accepted** August 08, 2012; **Published** August 13, 2012

Citation: Zheng Y, Charnigo R (2012) On Selecting Spatial-Temporal Autologistic Regression Models for Binary Lattice Data. J Biom Biostat 3:e112. doi:10.4172/2155-6180.1000e112

to compare several different spatial dependence structures for dental caries data. As these examples suggest, covariates and neighborhood structure are usually not selected simultaneously, since examining all possible combinations of covariates and neighborhood structure may be prohibitively time-consuming.

In the remainder of this editorial, we discuss some possibilities for selection of covariates and spatial-temporal neighborhood structure, based on the premise of determining which regression and autoregressive coefficients are non-zero. One idea would be to consider a penalized log-likelihood function via adaptive LASSO [14],

$$Q(\eta) = l(\eta) - n(m-s)\sum_{j=1}^{p}\lambda_j\left|\beta_j\right| - n(m-s)\sum_{k=1}^{q}\tau_k\left|\theta_k\right| - n(m-s)\sum_{l=1}^{s}\zeta_l\left|\alpha_l\right|,$$

where $\{\lambda_j\}_{j=1}^{p}$ are regularization parameters for the regression coefficients $\beta, \{\tau_k\}_{k=1}^{q}$ correspond to the spatial autoregressive coefficients θ, and $\{\zeta_l\}_{l=1}^{s}$ pertain to the temporal autoregressive coefficients Here is the likelihood function. However, for the spatial-temporal autologistic regression model, there is no explicit representation of the likelihood function. One possibility would be to replace the likelihood function by the pseudolikelihood function [15]. Another would be to use the Monte Carlo likelihood function (see, e.g. Geyer and Thompson [16], Huffer and Wu [17]), which consistently estimates the likelihood function but is computationally intensive.

To maximize $Q(\eta)$, one possibility is to deploy a Newton-Raphson (NR) type algorithm based on a local quadratic approximation (LQA). The LQA algorithm has been used widely and shown to produce reliable results in practice, even for dependent data [18]. However, this algorithm is slow, and a coefficient shrunk to 0 during the iteration of the algorithm remains at 0 throughout all subsequent iterations. Other methods may be considered for non-Gaussian distributions. For example, Madigan and Ridgeway [19] considered LARS-type algorithms for logistic regression, while Genkin et al. [20] proposed Bayesian logistic regression with a Laplace prior for large-scale text categorization. Park and Hastie [21] developed a path algorithm for variable selection in a generalized linear model based on a predictor-corrector method. We conclude this editorial by calling for further research on efficient variable and neighborhood structure selection for autologistic regression models, which will equip scientists with more advanced statistical tools for exploring and analyzing spatial-temporal lattice data.

References

1. Richardson S, Abellan JJ, Best N (2006) Bayesian spatio-temporal analysis of joint patterns of male and female lung cancer risks in Yorkshire (UK). Stat Methods Med Res 15: 385-407.
2. Jin X, Carlin BP (2005) Multivariate parametric spatiotemporal models for county level breast cancer survival data. Lifetime Data Anal 11: 5-27.
3. Earnest A (2010) Addressing issues in sparseness, ecological bias and formulation of the adjacency matrix in Bayesian spatio-temporal analysis of disease counts. Queensland University of Technology.
4. Hartley DM, Barker CM, Le Menach A, Niu T, Gaff HD, et al. (2012) Effects of temperature on emergence and seasonality of West Nile virus in California. Am J Trop Med Hyg 86: 884-894.
5. Cressie NAC (1993) Statistics for Spatial Data. 2nd edn, J Wiley.
6. Rue H, Held L (2005) Markov Random Field: Theory and Application. Chapman and Hall, London.
7. Schabenberger O, Gotway CA (2004) Statistical Methods for Spatial Data Analysis. Chapman and Hall/CRC.
8. Zhu J, Zheng Y, Carroll AL, Aukema BH (2008) Autologistic regression analysis of spatial-temporal binary data via Monte Carlo maximum likelihood. J Agric Biol Environ Stat 13: 84-98.
9. Bandyopadhyay D, Reich BJ, Slate EH (2009) Bayesian modeling of multivariate spatial binary data with applications to dental caries. Stat Med 28: 3492-3508.
10. Zhu J, Huang HC, Reyes PE (2010) On selection of spatial linear models for lattice data. J R Stat Soc Series B Stat Methodol 72: 389-402.
11. Reyes PE (2010) Selection of spatial and spatial-temporal linear models for lattice data. University of Wisconsin, Madison.
12. Aukema BH, Carroll AL, Zheng Y, Zhu J, Raffa KF, et al. (2008) Movement of outbreak populations of mountain pine beetle: Influences of spatiotemporal patterns and climate. Ecography 31: 348-358.
13. Zhu J, Huang HC, Wu J (2005) Modeling spatial-temporal binary data using Markov random fields. J Agric Biol Environ Stat 10: 212-225.
14. Zou H (2006) The adaptive lasso and its oracle properties. J Am Stat Assoc 101: 1418-1429.
15. Besag J (1975) Statistical analysis of non-lattice data. Statistician 24: 179-195.
16. Geyer CJ, Thompson EA (1992) Constrained Monte Carlo maximum likelihood for dependent data. Journal of the Royal Statistical Society. Series B (Methodological) 54: 657-699.
17. Huffer FW, Wu H (1998) Markov Chain Monte Carlo for autologistic regression models with application to the distribution of plant species. Biometrics 54: 509-524.
18. Wang H, Li G, Tsai CL (2007) Regression coefficient and autoregressive order shrinkage and selection via the lasso. J Royal Stat Soc Series B Stat Methodol 69: 63-78.
19. Madigan D, Ridgeway G (2004) Discussion of "least angle regression" by Efron et al. Ann Stat 32: 465-469.
20. Genkin A, Lewis DD, Madigan D (2007) Large-scale Bayesian logistic regression for text categorization. Technometrics 49: 291-304.
21. Park MY, Hastie T (2007) L_1-regularization path algorithm for generalized linear models. J Royal Stat Soc B 69: 659-677.