Journal of the Korean Statistical Society 43 (2014) 339–353

Contents lists available at ScienceDirect

Journal of the Korean Statistical Society

journal homepage: www.elsevier.com/locate/jkss

Maximum likelihood estimation for generalized conditionally autoregressive models of spatial data

M. Kyung[a,*], S.K. Ghosh[b]

[a] Department of Statistics, Duksung Women's University, Seoul, Republic of Korea

[b] Department of Statistics, North Carolina State University, Raleigh, NC, USA

ARTICLE INFO

Article history:
Received 15 April 2013
Accepted 28 October 2013
Available online 2 December 2013

AMS 2000 subject classifications:
62H11
62J99

Keywords:
Conditionally autoregressive models
Kernel estimator
Lattice data
Maximum likelihood estimation
Spatial analysis

ABSTRACT

Conditionally autoregressive (CAR) models are often used to analyze a spatial process observed over a lattice or a set of irregular regions. The neighborhoods within a CAR model are generally formed deterministically using the inter-distances or boundaries between the regions. To accommodate directional and inherent anisotropy variation, a new class of spatial models is proposed that adaptively determines neighbors based on a bivariate kernel using the distances and angles between the centroid of the regions. The newly proposed model generalizes the usual CAR model in a sense of accounting for adaptively determined weights. Maximum likelihood estimators are derived and simulation studies are presented for the sampling properties of the estimates on the new model, which is compared to the CAR model. Finally the method is illustrated using a data set on the elevated blood lead levels of children under the age of 72 months observed in Virginia in the year of 2000.

1. Introduction

Given a set of geographical regions, observations collected over regions nearer to each other tend to have similar characteristics as compared to distant regions. In geoscience, this feature is known as *Tobler's first law* (Miller, 2004). From a statistical perspective, given a set of geographical regions, the autocorrelation between the observations collected from nearer regions tends to be higher than those that are distant. Thus, this spatial process observed over a lattice or a set of irregular regions is usually modeled using autoregressive models.

A real data set of estimating the rate per thousand of children under the age of 72 months with elevated blood lead levels observed in Virginia in the year 2000, is an example of a data set observed over irregular regions from Schabenberger and Gotway (2005). Details with various model fittings are discussed in Schabenberger and Gotway (2005), and we will reanalyze this data set using our proposed models.

Let $\mathcal{S}$ be the study region of interest, which can be split into the well defined set of disjoint and exhaustive sub-areas $S_1, \ldots, S_n$ such that $\mathcal{S} = \cup_{i=1}^{n} S_i$ and $S_i \cap S_j = \emptyset$ for $i \neq j$. Suppose $Y_i = Y(S_i)$ denotes some form of aggregated response collected over S_i. We consider a generalized linear model for the aggregated responses

$$\begin{aligned} &\mathrm{E}[\boldsymbol{Y}|\boldsymbol{Z}] = g(\boldsymbol{Z}) \quad \text{and} \\ &\boldsymbol{Z} = \boldsymbol{\mu} + \boldsymbol{\eta}, \end{aligned} \tag{1}$$

where $\boldsymbol{Y} = (Y_1, \ldots, Y_n) = (Y(S_1), \ldots, Y(S_n))$, $\boldsymbol{Z} = (Z_1, \ldots, Z_n) = (Z(S_1), \ldots, Z(S_n))$, $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_n) = (\mu(S_1), \ldots, \mu(S_n))$ and $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_n) = (\eta(S_1), \ldots, \eta(S_n))$. Here, $g(\cdot)$ is a suitable link function, $\boldsymbol{\mu}$ represents a vector of

* Corresponding author. Tel.: +82 29018336; fax: +82 29018337.
E-mail addresses: mkyung@duksung.ac.kr, minjung.kyung@gmail.com (M. Kyung), sujit.ghosh@ncsu.edu (S.K. Ghosh).

1226-3192/$ – see front matter © 2013 The Korean Statistical Society. Published by Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.jkss.2013.10.005

large-scale variations (trends over geographical regions) and $\boldsymbol{\eta}$ denotes a vector of small-scale variations (spatial random effects) with mean 0 and variance–covariance matrix $\boldsymbol{\Sigma}$.

The large scale variations on each sub-region S_i, μ_i, are usually modeled as a function of some explanatory variables (e.g., latitudes, longitudes and other areal level covariates) using a parametric or semiparametric regression model (see van der Linde, Witzko, & Jockel, 1995). However, developing a suitable model for the spatial random effects η_i is a more challenging issue because the η_is are spatially correlated and are modeled as a latent (random) variable. Also, for any model specification, the positive definiteness condition of the covariance structure needs to be satisfied. Estimations of such spatial covariances are usually based on choosing suitable parametric forms. This means that the $n \times n$ covariance matrix $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}(\boldsymbol{\omega})$ is assumed to be a deterministic function of a finite dimensional parameter $\boldsymbol{\omega}$. Thus, it should lead to a positive definite matrix for any sample size n and for all allowable parameter values $\boldsymbol{\omega}$, where $\boldsymbol{\omega}$ is estimated from data.

For example, several geostatistical models are available for point-reference observations assuming that the spatial process is weakly stationary and isotropic (see Cressie, 1993). Several extensions to model nonstationary and anisotropic processes have also been developed (see Fuentes, 2002, 2005; Fuentes & Smith, 2001; Higdon, 1998; Higdon, Swall, & Kern, 1999; Hughes-Oliver, Heo, & Ghosh, 2008; Paciorek & Schervish, 2006). Once a valid model for $\boldsymbol{\mu}$ and $\boldsymbol{\eta}$ is specified, parameter estimates can be obtained using the maximum likelihood method, weighted least squares, or Bayesian methods (see Schabenberger & Gotway, 2005). Once the point-referenced data are aggregated to the areal regions (S_i's), the process representing the aggregated data is modeled using integrals of spatial continuous processes (Journel & Huijbregts, 1978). In our paper, the focus is the estimation of the covariance parameter $\boldsymbol{\omega}$ with a model chosen for $\boldsymbol{\Sigma}$.

There are two distinct approaches to develop models for spatial covariance based on areal data. A suitably aggregated geostatistical model directly specifies a deterministic function of the elements of the $\boldsymbol{\Sigma}$ matrix. Alternatively the Conditional Autoregressive (CAR) models specify a deterministic function of elements of the inverse of the covariance, $\boldsymbol{\Sigma}^{-1}(\boldsymbol{\omega})$. Over the past decades, there have been several attempts to explore the possible connections between these approaches of spatial modeling (see e.g. Griffith & Csillag, 1993; Hrafnkelsson & Cressie, 2003; Rue & Tjelmeland, 2002). Song, Fuentes, and Ghosh (2008) proposed that these Gaussian geostatistical models can be approximately represented by Gaussian Markov Random Fields (GMRFs) and vice versa by using spectral densities. However, so far most of the GMRFs that are available in literature do not specifically take into account the anisotropic nature of areal data. Lindgren, Rue, and Lindström (2011) showed that, using an approximate stochastic weak solution to linear stochastic partial differential equations, for some Gaussian fields in the Matérn class, an explicit link for any triangulation of $\mathbb{R}^d$, between Gaussian fields and Gaussian Markov Random Fields, formulated as a basis function representation could be provided. Finite element methods have been used to give a principled construction of Matérn-type Gaussian Models in a variety of spatial settings for a sparse inverse covariance or precision matrix.

In practice, statistical practitioners are accustomed to the exploration of relationships among variables, modeling these relationships with regression and classification models, testing hypotheses about regressors and treatment effects, developing meaningful contrasts, and so forth (Schabenberger & Gotway, 2005). For these spatial linear models, we usually assume the correlated relationship among sub-regions is a hidden structure and study how a particular region is influenced by its "neighboring regions" (Cliff & Ord, 1981). Thus, generalized linear mixed models for the area aggregate data are usually considered with the latent spatial process. In these models, the latent spatial process η_i's can be treated as a random effect, and to model it, conditionally autoregressive models (Besag, 1974, 1975; Cressie & Chan, 1989) and Simultaneously Autoregressive (SAR) models (Ord, 1975) have been used widely.

To explain the latent spatial process using suitably formed neighbors, Gaussian CAR models have been used as random effects within generalized mixed effects models (Breslow & Clayton, 1993). The Gaussian CAR process has the merit that, under fairly general regularity conditions, (e.g., positivity conditions etc.) lower dimensional conditional Gaussian distributions uniquely determine joint Gaussianity of the spatial CAR processes. Thus, maximum likelihood (ML) and Bayesian estimates can be easily obtained. However, one of the major limitations of the CAR model is that the neighbors are formed using some form a distance metric, or based on adjacency, but the effect of the direction is completely ignored. To explain different effects depending on direction, in recent years, there have been some attempts to use different CAR models for different parts of the region. For instance, Reich, Hodges, and Carlin (2007) presented a novel model for periodontal disease and used separate CAR models for separate jaws. White and Ghosh (2008) used a stochastic parameter within the CAR framework to determine effects of the neighbors. Also, Kyung and Ghosh (2010) proposed a Directional CAR model which captures anisotropy by assuming different autocorrelation between the neighbors in different directions. Recently, Lee and Mitchell (2013) proposed a new methodology to capture localized spatial correlation by estimating the elements of the neighborhood matrix (w_{kj}) as 1 or 0 as long as areas share a common border, rather than assuming that they are fixed as 1. Because the elements of the neighborhood matrix determine the partial correlations between the random effects in areas, if $w_{kj} = 1$ the random effects are smoothed over in the modeling process, whereas if $w_{kj} = 0$ they are conditionally independent. Generally, the weights assigned to neighbors might be a function of other features of the lattice sites. This means that the weights of neighbors might be different from each other by various features. For example, the weights can be decided by the distance between the centroids of the sub-regions, then the closer neighbors have more weights to explain the small scale variation of sub-region i. Also, if the underlying spatial process is anisotropic, the magnitude of autocorrelation between the neighbors might be different in different directions. Thus, we propose an "adaptive" method for the spatial random effect η_i's to obtain the neighboring regions using the centroids of S_i's and we call it a generalized conditional autoregressive (GCAR) model.