Communications of the Korean Statistical Society
2009, Vol. 16, No. 5, 851–858

Spatial Prediction Based on the Bayesian Kriging with Box-Cox Transformation

Jungsoon Choi[a], Man Sik Park[1,a]

[a]Department of Biostatistics, Korea University

Abstract

In the last decades, there has been much interest in climate variability because its change has dramatic effects on humanity. Especially, the precipitation data are measured over space and their spatial association is so complicated. So we should take into account such a spatial dependency structure while analyzing the data. However, in linear models for analyzing the data, data sets show severely skewed distribution. In the paper, we consider the Box-Cox transformation to satisfy the normal distribution prior to the analysis, and employ a Bayesian hierarchical framework to investigate the spatial patterns. The data set we considered is monthly average precipitation of the third quarter of 2007 obtained from 347 automated monitoring stations in Contiguous South Korea.

Keywords: Precipitation, Bayesian kriging, Box-Cox transformation.

1. Introduction

In the last decades, there has been much interest in climate variability because its change has dramatic effects on humanity. More recently, the drought and the massive rainfall in many regions have caused severe effects such as numerous death, destruction of crops, loss of houses, and so on (Park and Heo, 2009). Thus, it is very important to analyze the climate data in order to understand climate changes and predict its future changes.

The rainfall data are observed over space and the data have complicated spatial dependency. So we should take into account such a spatial dependency structure while analyzing the data. In this paper, we suggest a spatial modeling of the rainfall data in South Korea to investigate the spatial patterns of the rainfall data and then predict rainfall values at all locations of interest.

There are two approaches in statistical modeling for Geostatistical (point-referenced) data such as rainfall data: classical and Bayesian methods. Bayesian approaches are becoming very popular for spatial modeling in recent years (Le and Zidek, 1992; Handcock and Stein, 1993, Brown *et al.*, 1994; Handcock and Wallis, 1994; De Oliveira *et al.*, 1997; Ecker and Gelfand, 1997; Diggle *et al.*, 1998; Karson *et al.*, 1999). Spatial data in climate and environmental sciences have variabilities over space, and Bayesian approaches provide a natural framework to understand the spatial dependency structure. In the Bayesian approach, the most common computing tool is Markov Chain Monte Carlo(MCMC) method (see, for example, Robert (1996) for a review) due to its ability to make an inference from posterior distributions. We consider a Bayesian framework using MCMC method for the rainfall data to investigate their spatial patterns.

In linear models, one of the common assumptions is the normal distribution of the variable of interest. However, in most real application, data sets show non-normality characteristics such as

[1] Corresponding author: Research Professor, Department of Biostatistics, College of Medicine, Korea University, 5-1 Anam-Dong, Seungbuk-Gu, Seoul 136-705, Korea. E-mail: man.sik.park@gmail.com

skewed distributions and heavy tailed distributions. One of the natural ways to model non-normal data is to use a transformation. Box and Cox (1964) proposed a family of transformation, which is known as the Box-Cox transformation. This approach provides reasonable approximations to the normal distribution. De Oliveira *et al.* (1997), Smith *et al.* (2003) and Cressie *et al.* (2006) developed spatial models using the Box-Cox transformation. In the paper, we use the Box-Cox transformation to satisfy the normal distribution for climate data prior to the analysis.

In Section 2 we present a spatial model with the Box-Cox transformation for the precipitation data and the Bayesian kriging method. In Section 3, we describe the precipitation data in South Korea we used in this study, provide the results based on the Bayesian hierarchical framework, and compare its performance with the one from classical modeling approaches. Finally, in Section 4, we present conclusions and general discussion.

2. The Box-Cox Transformation and Bayesian Kriging

Point-referenced data (or geostatistical data) are called when a location $\mathbf{s} \in \mathbb{R}^2$ varies continuously over a fixed region, $D \subset \mathbb{R}^2$. One of main interests in point-referenced data is to model the spatial distribution of measurements for taking into account spatial correlation and for predicting values at new locations, which is called a kriging (See the books by Banerjee *et al.* (2004) and Schabenberger and Gotway (2004) for other spatial data types and additional discussion).

2.1. Spatial model with the Box-Cox transformation

We let $Z(\mathbf{s})$ be a spatial process at location $\mathbf{s} \in D$, where $D = \{\mathbf{s}_1, \mathbf{s}_2, \ldots, \mathbf{s}_n\} \subset \mathbb{R}^2$, and we assume that $Y(\mathbf{s})$ is a transformation of the observed data at location $\mathbf{s}$, $Z(\mathbf{s})$. So, in order to make $Z(\mathbf{s})$ satisfy the normality assumption, we model $Y(\mathbf{s})$ as

$$Y(\mathbf{s}) = f(Z(\mathbf{s}), \lambda),$$

where $f(\cdot, \lambda)$ is a family of transformation with the power parameter λ. We use the Box-Cox transformation (Box and Cox, 1964) as follows;

$$Y(\mathbf{s}) = \begin{cases} \dfrac{Z(\mathbf{s})^\lambda - 1}{\lambda}, & \text{if } \lambda \neq 0, \\ \log(Z(\mathbf{s})), & \text{if } \lambda = 0. \end{cases} \tag{2.1}$$

We model the process $Y(\mathbf{s})$ transformed by (2.1) as

$$Y(\mathbf{s}_i) = \mu(\mathbf{s}_i) + \epsilon(\mathbf{s}_i), \quad i = 1, \ldots, n, \tag{2.2}$$

where the spatial mean function $\mu(\mathbf{s}_i)$ represents the large-scale variation and the process $\epsilon(\mathbf{s}_i)$ is the small-scale variation and assumed as be a Gaussian process. In the study, we assumed that the mean function $\mu(\mathbf{s}_i)$ is modeled as a function of geographic information (*e.g.*, longitude and latitude).

For the vector $\mathbf{Y} = (Y(\mathbf{s}_1), \ldots, Y(\mathbf{s}_n))^T$, the model in (2.2) can be written as

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \tag{2.3}$$

where $\mathbf{X} = (X(\mathbf{s}_1), \ldots, X(\mathbf{s}_n))^T$ is the $n \times p$ covariate matrix and $\boldsymbol{\beta}$ is a $p \times 1$ coefficient vector corresponding to the covariate matrix $\mathbf{X}$. The error process vector $\boldsymbol{\epsilon} = (\epsilon(\mathbf{s}_1), \ldots, \epsilon(\mathbf{s}_n))^T$ is assumed to follow the normal distribution as follows;

$$\boldsymbol{\epsilon} \sim N\left(\mathbf{0}, \tau^2 \mathbf{I}_n + \sigma^2 \mathbf{H}(\phi)\right), \tag{2.4}$$

where $\mathbf{I}_n$ is an identity matrix of size of n and $\mathbf{H}(\phi)$ is the spatial correlation matrix with a parameter ϕ. In Equation (2.4), we consider the three parameters in variance-covariance matrix of the error process vector, which are nugget (τ^2), partial sill (σ^2) and range (ϕ). The range parameter, ϕ means the maximum distance at which the spatial correlation between any pair of data in the study region ceases. The nugget parameter, τ^2 indicates a non-spatial variance due to measurement error. The sill is the total variance in the study and the partial sill, τ^2 means a spatial variance, which is calculated as the sill minus the nugget.

2.2. Bayesian Kriging

Combining the equations (2.3) and (2.4), we can easily derive the distribution for $\mathbf{Y}$ as

$$\mathbf{Y}|\boldsymbol{\theta} \sim N\left(\mathbf{X}\boldsymbol{\beta}, \tau^2\mathbf{I}_n + \sigma^2\mathbf{H}(\phi)\right), \tag{2.5}$$

where $\boldsymbol{\theta} = (\boldsymbol{\beta}, \tau^2, \sigma^2, \phi)^T$. To determine a prior distribution on $\boldsymbol{\theta}$, we first assume that $\boldsymbol{\beta}, \tau^2, \sigma^2$ and ϕ are mutually independent. So the prior distribution for $\boldsymbol{\theta}$ is denoted as

$$p(\boldsymbol{\theta}) = p(\boldsymbol{\beta})\, p\left(\tau^2\right) p\left(\sigma^2\right) p(\phi).$$

In this paper, we consider the priors of the parameters in the variance-covariance matrix shown in (2.5), τ^2, σ^2 and ϕ as

$$p(\tau^2) \sim \text{IG}(a_1, b_1), \quad p(\sigma^2) \sim \text{IG}(a_2, b_2) \quad \text{and} \quad p(\phi) \sim \text{Unif}(c, d),$$

where $\text{IG}(a, b)$ denotes the Inverse-Gamma distribution with mean $b/(a-1)$ and variance $b^2/(a-1)^2(a-2)$. The prior distribution of $\boldsymbol{\beta}$, $p(\boldsymbol{\beta})$ is assumed to be noninformative priors. Inference on the parameter vector $\boldsymbol{\theta}$ is based on its posterior distribution as follows;

$$f(\boldsymbol{\theta}|\mathbf{Y}) \propto f(\mathbf{Y}|\boldsymbol{\theta}) \times p(\boldsymbol{\theta}).$$

Then, we calculate the values of $\mathbf{Y}$ at a new location $\mathbf{s}_0$. The predictive posterior distribution of $Y(\mathbf{s}_0)$ given the observed information, $\mathbf{Y}, \mathbf{X}$ and the covariates at location $\mathbf{s}_0$, $X(\mathbf{s}_0)$ is represented as

$$p(Y(\mathbf{s}_0)|\mathbf{Y}, \mathbf{X}, X(\mathbf{s}_0)) = \int p(Y(\mathbf{s}_0)|\mathbf{Y}, \mathbf{X}, X(\mathbf{s}_0), \boldsymbol{\theta})\, p(\boldsymbol{\theta}|\mathbf{Y}, \mathbf{X})\, d\boldsymbol{\theta}.$$

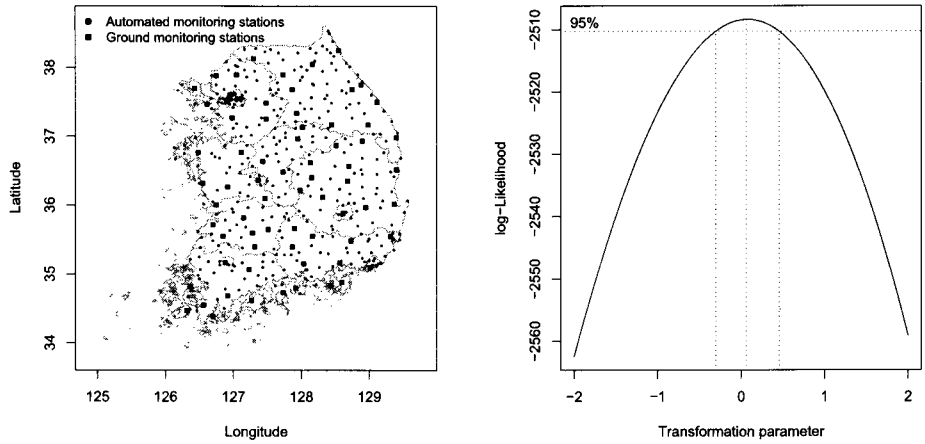
The estimation of the parameters and the prediction given the data are obtained by Markov chain Monte Carlo(MCMC) algorithm.

3. Real Application

We apply our Bayesian model proposed in Section 2 to the real data. In addition, Maximum likelihood(ML) and Restricted Maximum likelihood(REML) methods are also applied to the real data to compare their performances. The software that we used for the analysis is R (R Development Core Team, 2008), especially, "geoR" (Ribeiro and Diggle, 2001) and "spBayes" package (Finley *et al.*, 2008) in R.

3.1. Data resources

Precipitation measurements from two different monitoring networks in South Korea for the third quarter of 2007 (July, August and September) are used in this study. The first source of the precipitation



(a) Location information (b) The Box-Cox transformation

Figure 1: *Location information of the networks and the Box-Cox transformation of the power parameter.*

data is from the ground monitoring station network. There are 69 monitoring stations in South Korea. The second source of the precipitation data has been obtained from the automated monitoring station network, which have the 347 automated monitoring stations (see Heo and Park (2009) for the description of the networks in detail). Figure 1(a) shows the two monitoring stations in South Korea. We use monthly averaged values for the precipitation data obtained at the automated monitoring stations when modeling, and then we use the precipitation data from the ground monitoring station network in order to assess the performance of the model. The unit of the precipitation data is millimeter, and longitude and latitude coordinates for the monitoring stations taking into account the curvature of the earth are used.

Before analyzing the precipitation data, we first check the normality of the precipitation data from the automated monitoring stations. Figure 1(b) shows the log-likelihood of the transformation parameter, λ. It appears that taking the natural logarithm would be appropriate in sense that the zero value precisely belongs to the 95% confidence interval of λ.

3.2. Modeling framework

To construct the mean function $\mu(\mathbf{s}_i)$ in (2.2) using the transformed data, we investigate the association between the measurements and their geographic information. From Figure 2, we can see that there is a quadratic association between the precipitation measurements and the longitude and there is a linear association between the measurements and the latitude. Thus, we model the mean function $\mu(\mathbf{s}_i)$ as

$$\mu(\mathbf{s}_i) = \beta_0 + X_1(\mathbf{s}_i)\beta_1 + X_2(\mathbf{s}_i)\beta_2 + X_1^2(\mathbf{s}_i)\beta_3 + X_2^2(\mathbf{s}_i)\beta_4 + X_1(\mathbf{s}_i)X_2(\mathbf{s}_i)\beta_5, \tag{3.1}$$

where $X_1(\mathbf{s}_i)$ and $X_2(\mathbf{s}_i)$ are longitude and latitude at each station $\mathbf{s}_i$, respectively. To construct the spatial dependency structure in the model, we use an exponential correlation function for the matrix $\mathbf{H}(\phi) = \{h_{ij}(\phi)\}$, where

$$h_{ij}(\phi) = \exp\left(-\frac{\|\mathbf{s}_i - \mathbf{s}_j\|}{\phi}\right), \quad i, j = 1, \ldots, n.$$

Here $\|\mathbf{s}_i - \mathbf{s}_j\|$ is the Euclidean distance between the locations, $\mathbf{s}_i$ and $\mathbf{s}_j$.

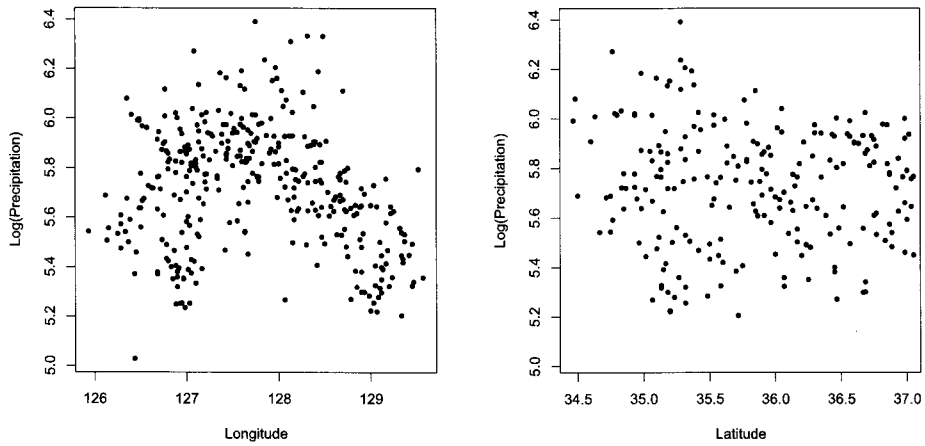


Figure 2: *Scatterplots of precipitation measurements and location information.*

Table 1: Parameter estimates and 95% confidence intervals

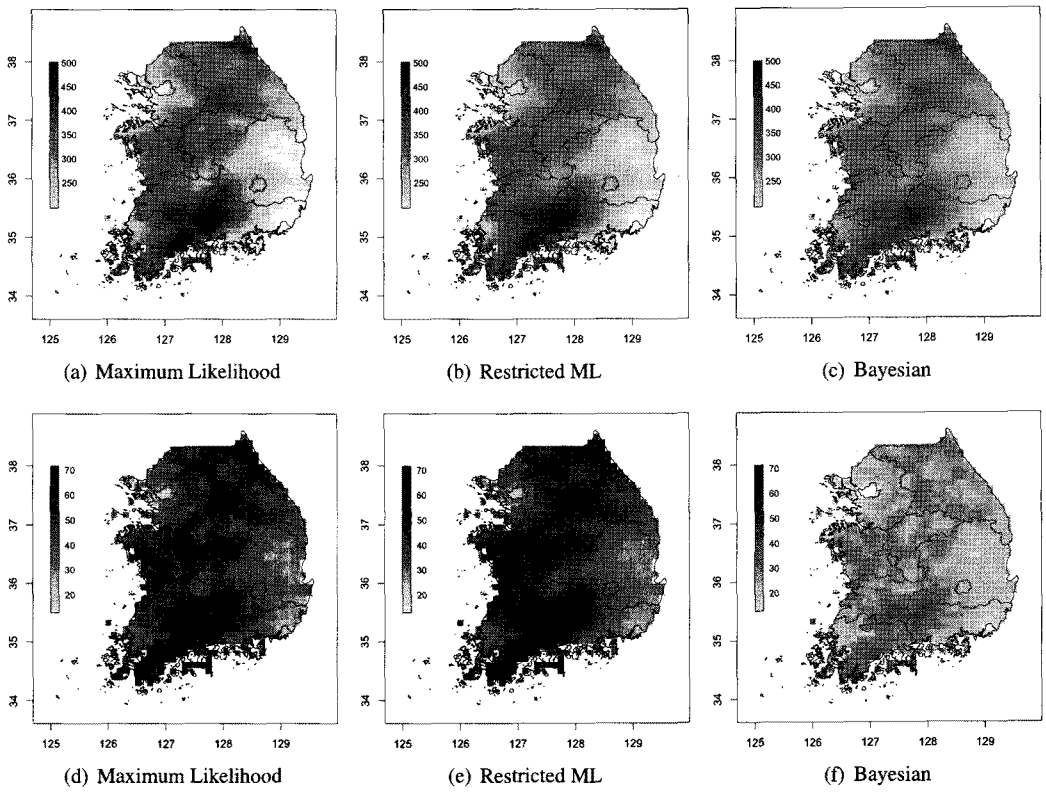
	Maximum Likelihood			Restricted ML			Bayesian		
	Est.	95% CI		Est.	95% CI		Est.†	95% CI‡	
β_0	5.856	5.841	5.871	5.862	5.858	5.866	5.861	5.751	5.970
β_1	-3.424^4	-4.115^4	-2.733^4	-4.881^4	-5.139^4	-4.623^4	-4.607^4	-1.072^3	1.717^4
β_2	-1.726^4	-2.284^4	-1.168^4	-1.105^4	-1.292^4	-9.173^5	-1.151^4	-5.860^4	3.459^4
β_3	-1.519^5	-1.575^5	-1.463^5	-1.789^5	-1.818^5	-1.759^5	-1.733^5	-2.372^5	-1.040^5
β_4	-2.193^6	-2.551^6	-1.835^6	-1.072^6	-1.232^6	-9.117^7	-1.288^6	-5.125^6	2.278^6
β_5	6.830^6	6.343^6	7.316^6	8.232^6	8.002^6	8.463^6	7.985^6	2.582^6	1.310^5
σ^2	0.045			0.019			0.021	0.013	0.032
τ^2	0.015			0.012			0.013	0.008	0.017
ϕ	99.807			21.843			30.712	16.160	57.428

Notes: Est.: estimate; †: Posterior mean; ‡: 95% Bayesian confidence interval; $a^b \equiv a \times 10^{-b}$.

For the Bayesian hierarchical framework, we assign noninformative priors to all the parameters. The regression coefficients, $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_5)$, in the mean function shown in (3.1) have flat priors. For the partial sill, σ^2 and the nugget, τ^2, an inverse gamma prior, IG(0.001, 0.001) is used. The inverse of the range, $1/\phi$, has a uniform prior, Unif$(1/d_M, 1)$, where $d_M = 500km$ is the maximum of distance between the automated monitoring stations used in the study. One chain is run for 20,000 MCMC samples from the posterior densities of the parameters, of which the first 15,000 are discarded as burn-in. For all MCMC sequences, we conducted several MCMC convergence diagnosis techniques such as autocorrelation functions, and trace plots.

3.3. Statistical results

Table 1 presents the parameter estimates and their 95% confidence intervals for the three different estimation methods: Maximum likelihood, Restricted ML and Bayesian methods. The Bayesian approach among the three estimation methods only provides the confidence intervals for the spatial parameters. The parameter estimates except the intercept parameter which are obtained from the Maximum likelihood method are slightly different with the estimates from the other two estimation



(a) Maximum Likelihood (b) Restricted ML (c) Bayesian

(d) Maximum Likelihood (e) Restricted ML (f) Bayesian

Figure 3: *Maps of kriging estimates (first row) and the standard errors.*

methods. The confidence intervals from the Restricted ML and Bayesian methods are also slightly different. From the results for the spatial range parameter ϕ, we can see that the precipitation measurements have a spatial dependency, so the precipitation value at a monitoring station is correlated with the precipitation values at neighboring stations. Based on the results obtained from the Bayesian method, the upper limit of the 95% confidence interval is 57.428, which is about 10% of maximum distance. It thus seems that the range of the spatial correlation of the precipitation data in summer, South Korea is short.

Figure 3 shows the maps of the predicted precipitation values and their standard errors for the three estimation methods considered in the study. We obtained the prediction values at the lattice grid points over South Korea and produced the prediction maps. In the Bayesian method, we used the mean of the posterior predictive distribution. From Figure 3(a), (b) and (c), we can see the maps of the kriging estimates, showing the similar estimates for the methods. The precipitation is low in Seoul and the eastern region of Gyoungsang province, while the precipitation is high in south-western region of Gyoungsang province and south-eastern region of Jeolla province and north-eastern region of Kangwon province. Unlike the maps of the kriging estimates, the maps of the kriging standard errors in Figure 3 (d), (e) and (f) do not look similar. Overall, the Bayesian method has the smallest variation for the prediction among the methods.

In order to illustrate the performance of the models proposed in the study using the different estimation methods, we used the 69 ground monitoring stations. We predict the prediction values at the monitoring stations and compare them with the observed values. We also consider the models

Table 2: Comparison of predictions and observations

	With Transformation						Without Transformation			
	OLS		ReML		Bayesian		ReML		Bayesian	
	Est.	*P.*	Est.	*P.*	Est.	*P.*	Est.	*P.*	Est.	*P.*
Intercept	38.01	0.418	67.74	0.066	66.10	0.082	150.20	<.001	152.20	<.001
Slope	0.90	<.001	0.79	<.001	0.80	<.001	0.54	<.001	0.53	<.001
MSPE	54.94		52.66		53.11		54.05		54.17	
Mean.Res.	−6.73		0.38		−1.38		0.54		0.51	
Median.Res.	2.82		2.47		1.72		5.60		4.83	

Notes. Est.: estimate; *P.*: *P*-value.; Mean.Res.: Mean(Residuals); Median.Res.: Median(Residuals)

without the Box-Cox transformation for the estimation methods (Heo and Park, 2009) to see the need of the transformation for the rainfall data. In addition, we consider the prediction based on the ordinal least squares estimation method without spatial correlation structure. For the comparison, a linear relationship between the observed values, $Z(\mathbf{s}_i)$, and the prediction values, $\widehat{Z}(\mathbf{s}_i)$ is considered as

$$\widehat{Z}(\mathbf{s}_i) = a_0 + a_1 Z(\mathbf{s}_i) + e(\mathbf{s}_i), \quad i = 1, \ldots, 69.$$

For each method, we obtain the intercept(a_0) and the slope(a_1) and we also compute the Mean Square Prediction Error(MSPE)as follows;

$$\text{MSPE} = \frac{1}{69} \sum_{i=1}^{69} \left\{\widehat{Z}(\mathbf{s}_i) - Z(\mathbf{s}_i)\right\}^2$$

and mean (median) of the residuals. In the Table 2, the slope estimates for the models with the transformation are much closer to one than them for the models without the transformation. The estimation methods with spatial correlation structure among the models with the transformation are better than the least squares estimation method in terms of the MSPE and mean (median) of the residuals. Thus, the estimation methods with spatial correlation structure and transformation perform well.

4. Conclusions

In the paper, we suggested a spatial model with the Box-Cox transformation for the precipitation data in South Korea. We compared the Bayesian method with the common estimation methods (Maximum Likelihood and Restricted ML). From the results, we found that the model used for the spatial dependency structure performs well. We also found that the Box-Cox transformation is needed for the normality assumption in linear models. Thus, the spatial linear regression model based on the Bayesian hierarchical framework with the transformation might be preferred for real data without the normality assumption.

References

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

Box, G. E. P. and Cox, D. R. (1964). An analysis of transformations, *Journal of the Royal Statistical Society, Series B,* **26**, 211–246.

Brown, P. J., Le, N. D. and Zidek, J. V. (1994). Multivariate spatial interpolation and exposure to air pollutants, *Canadian Journal of Statistics,* **22**, 489–509.

Cressie, N., Frey, J., Harch, B. and Smith, M. (2006). Spatial prediction on a river network, *Journal of Agricultural, Biological, and Environmental Statistics,* **11**, 127–150.

De Oliveira, V., Kedem, B. and Short, D. A. (1997). Bayesian prediction of transformed Gaussian random fields, *Journal of the American Statistical Association,* **92**, 1422–1433.

Diggle, P. J., Tawn, J. A. and Moyeed, R. A. (1998). Model-based geostatistics (with discussion), *Applied Statistics,* **47**, 299–350.

Ecker, M. D. and Gelfand, A. E. (1997). Bayesian variogram modeling for an isotropic spatial process, *Journal of Agricultural, Biological, and Environmental Statistics,* **2**, 347–369.

Finley, A. O., Banerjee, S. and Carlin, B. P. (2008). spBayes: Univariate and Multivariate Spatial Modeling, R package version 0.1-0.

Handcock, M. S. and Stein, M. L. (1993). A Bayesian analysis of kriging, *Technometrics,* **35**, 403–410.

Handcock, M. S. and Wallis, J. R. (1994). An approach to statistical spatio-temporal modeling of meteorological fields, *Journal of the American Statistical Association,* **89**, 368–378.

Heo, T. Y. and Park, M. S. (2009). Bayesian spatial modeling of precipitation data, *Korean Journal of Applied Statistics,* **22**, 425–433.

Karson, M. J., Gaudard, M., Linder, E. and Sinha, D. (1999). Bayesian analysis and computations for spatial prediction (with discussion), *Environmental and Ecological Statistics,* **6**, 147–182.

Le, N. D. and Zidek, J. V. (1992). Interpolation with uncertain spatial covariance: A Bayesian alternative to Kriging, *Journal of Multivariate Analysis,* **43**, 351–374.

Park, M. S. and Heo, T. Y. (2009). Seasonal spatial-temporal model for rainfall data of South Korea, *Journal of Applied Sciences Research,* **5**, 565–572.

R Development Core Team (2008). R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing. Vienna, Austria, ISBN 3-900051-07-0.

Ribeiro, P. J. and Diggle, P. J. (2001). geoR: A package for geostatistical analysis, *R-NEWS,* **1**, 15–18.

Roberts, G. O. (1996). *Markov Chain Concepts Related to Sampling Algorithms, in Markov Chain Monte Carlo in Practice*, edited by W. R. Gilks, S. Richardson and D. J. Spiegelhalter. Chapman & Hall/CRC, London, 45–57.

Schabenberger, O. and Gotway, C. A. (2004). *Statistical Methods for Spatial Data Analysis*, Chapman & Hall/CRC, Florida.

Smith, R. L., Kolenikov, S. and Cox, L. H. (2003). Spatiotemporal modeling of PM2.5 data with missing values, *Journal of Geophysical Research,* **108**, STS 11-1.

Received July 2009; Accepted August 2009