The Korean Journal of Applied Statistics (2009)
22(2), 425–433

Bayesian Spatial Modeling of Precipitation Data

Tae-Young Heo[1] · Man Sik Park[2]

[1]Department of Data Information, Korea Maritime University;
[2]Department of Biostatistics, Korea University

(Received December 2008; accepted February 2009)

Abstract

Spatial models suitable for describing the evolving random fields in climate and environmental systems have been developed by many researchers. In general, rainfall in South Korea is highly variable in intensity and amount across space. This study characterizes the monthly and regional variation of rainfall fields using the spatial modeling. The main objective of this research is spatial prediction with the Bayesian hierarchical modeling (kriging) in order to further our understanding of water resources over space. We use the Bayesian approach in order to estimate the parameters and produce more reliable prediction. The Bayesian kriging also provides a promising solution for analyzing and predicting rainfall data.

Keywords: Precipitation, Bayesian kriging, Markov Chain Monte Carlo.

1. Introduction

Climate variability, its extremes and possible future changes have a strong impact on mankind. Some countries have already faced devastating experiences of the ongoing climate change phenomena. The twin but opposite phenomena of El Nino (drought) and La Nina (massive rainfall) have caused numerous deaths and destruction of crops (Park and Heo, 2008). The analysis of the environmental/climate data is very important both from an environmental perspective and for understanding the climate change.

The need for accurate modeling of the rainfall data set is vital. The amount of monthly rainfall and its seasonal distribution are crucial factors for understand-ing the spatial distribution of different ecological units, regardless of the scale of analysis (Bailey, 1998). The modeling of extreme rainfall or local severe storm is very important in the design of water-related structures, in agriculture, weather modification, and monitoring climate changes phenomena (Svensson and Rakhecha, 1998).

Spatial data, especially, geostatistical (point-referenced) data, are becoming increasingly utilized in the study of many scientific fields due to the accessibility of data monitoring systems (Eom *et al.*, 2006). When data are available from the underlying spatial process, computationally efficient method is needed for analysis. Markov Chain Monte Carlo(MCMC) is a very powerful tool often used for the Bayesian analysis. Especially, geostatistical approaches are usually considered quite sensible when treating precipitation data.

This work was supported by the research fund of Korea Maritime University.

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

Geostatistical models have been used very often in both classical and Bayesian framework. Considerable work has been done in the area of modeling spatially correlated data in a Bayesian perspective; see 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) and Diggle *et al.* (1998).

We now suggest a spatial model with application to precipitation data using Bayesian framework. In Section 2, we build a spatial model and introduce the Bayesian kriging approach. Section 3 contains the real application and results with precipitation data in South Korea. Conclusions and discussion are presented in Section 4.

2. Spatial Model and Bayesian Kriging

Geostatistical methods are applied to the point-referenced data, where a location at which a variable of interest is measured, varies continuously over a fixed spatial region. Modeling approaches from geostatistics are based on (semi)variogram modeling and spatial interpolation method, which is well known as kriging. In order to present geostatistical approaches, the stationarity condition in spatial statistics needs to be introduced. The spatial process, $\{Y(\mathbf{s}), \mathbf{s} \in \mathcal{D} \subset \mathbb{R}^2\}$ where $\mathcal{D}$ is a fixed subset of 2-dimensional Euclidean space, $\mathbb{R}^2$. Kriging can predict the value at an unobserved location, $\mathbf{s}_0$, given observations of the process $\{Y(\mathbf{s}), \mathbf{s} \in \mathcal{D}\}$. Kriging is also called the linear unbiased predictor with minimum variance.

2.1. Spatial model

Let $\mathbf{y} = \{Y(\mathbf{s}), \mathbf{s} \in \mathcal{D}\}$ be a spatial process that is observed at locations, $\zeta = \{\mathbf{s}_1, \ldots, \mathbf{s}_n\}$. We assume the following additive decomposition

$$Y(\mathbf{s}_i) = \mu(\mathbf{s}_i) + \epsilon(\mathbf{s}_i), \quad i = 1, \ldots, n, \tag{2.1}$$

where $Y(\mathbf{s}_i)$ represents (a function of) the measured outcome at location $\mathbf{s}_i$, $\mu(\mathbf{s}_i)$ is the large-scale variation (mean function) and $\epsilon(\mathbf{s}_i)$ represents an error process or small-scale variation. When we explain the large-scale variation with some geographic location information (*e.g.*, longitude and latitude), we can formulate the model shown in (2.1) as

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \tag{2.2}$$

where $\mathbf{X} = \{X_j(\mathbf{s}_i)\}_{i=1,\ldots,n; j=1,\ldots,p}$ is the covariate matrix, $\boldsymbol{\beta}$ is a $p \times 1$ column vector of coefficients for the covariate matrix, and $\boldsymbol{\epsilon} = (\epsilon(\mathbf{s}_1), \ldots, \epsilon(\mathbf{s}_n))^T \sim N_n(\mathbf{0}, \Sigma_{\boldsymbol{\theta}_1})$ denotes a spatial error process associated with parameter vector, $\boldsymbol{\theta}_1$. We can decompose the spatial error process, $\boldsymbol{\epsilon}$ in (2.2) into the two processes, with which (2.2) are expressed as

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}^0 + \boldsymbol{\epsilon}^1, \tag{2.3}$$

where $\boldsymbol{\epsilon}^0 \sim N_n(\mathbf{0}, \tau^2\mathbf{I})$ is an $n \times 1$ vector of independent and identically distributed vector of measurement errors and $\boldsymbol{\epsilon}^1 \sim N_n(\mathbf{0}, \sigma^2\mathbf{H}_\phi)$ is an $n \times 1$ random vector capturing the spatial correlation under the stationary condition. Here, $\boldsymbol{\theta}_1 = (\tau^2, \sigma^2, \phi)^T$, $\Sigma_{\boldsymbol{\theta}_1} = \tau^2\mathbf{I} + \sigma^2\mathbf{H}_\phi$ and $\mathbf{H}_\phi$ is the correlation matrix with a parameter, ϕ.

2.2. Bayesian kriging

In order to use a Bayesian approach, we need to specify a prior distribution on $\boldsymbol{\theta}^T = (\boldsymbol{\beta}^T, \boldsymbol{\theta}_1^T)$, $\pi(\boldsymbol{\theta})$. Common choice for the prior distribution is

$$\pi(\boldsymbol{\theta}) = \pi(\boldsymbol{\beta})\pi(\boldsymbol{\theta}_1) = \prod_{j=0}^{p} \pi(\beta_j) \times \pi\left(\sigma^2\right) \pi\left(\tau^2\right) \pi(\phi),$$

where $\pi(\boldsymbol{\beta})$ will be a noninformative prior, and Inverse-Gamma distribution is commonly used for $\pi(\sigma^2)$ and $\pi(\tau^2)$. Specification of $\pi(\phi)$ depends on choice of the correlation function ρ, but Uniform distribution is usually regarded as $\pi(\phi)$. Then, the parameter estimates are obtained from the posterior density $\pi(\boldsymbol{\theta}|\mathbf{y})$ as follows:

$$\pi(\boldsymbol{\theta}|\,\mathbf{y}) \propto f(\mathbf{y}|\,\boldsymbol{\theta}) \times \pi(\boldsymbol{\theta}),$$

where $f(\mathbf{y}|\boldsymbol{\theta})$ is the multivariate normal distribution as

$$\mathbf{y}|\,\boldsymbol{\theta} \sim N_n\left(\mathbf{X}\boldsymbol{\beta},\, \tau^2\mathbf{I} + \sigma^2\mathbf{H}_\phi\right).$$

To make inference on the parameters, we need to the marginal posterior densities, for example,

$$\pi(\phi|\,\mathbf{y}) = \iiint \pi\left(\boldsymbol{\beta},\, \sigma^2,\, \tau^2,\, \phi\,|\,\mathbf{y}\right)\, d\boldsymbol{\beta}\, d\sigma^2\, d\tau^2.$$

In general, the marginal posterior densities are not expressed in a closed form. Therefore, numerical integration or Markov Chain Monte Carlo(MCMC) technique is required.

The model shown in (2.3) can equivalently be expressed in the following hierarchical framework:

$$\begin{aligned}\mathbf{y}|\,\boldsymbol{\theta}, \boldsymbol{\epsilon}^1 &\sim N_n\left(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}^1,\, \tau^2\mathbf{I}\right)\\ \boldsymbol{\epsilon}^1|\,\sigma^2, \phi &\sim N_n\left(\mathbf{0},\, \sigma^2\mathbf{H}(\phi)\right),\end{aligned}$$

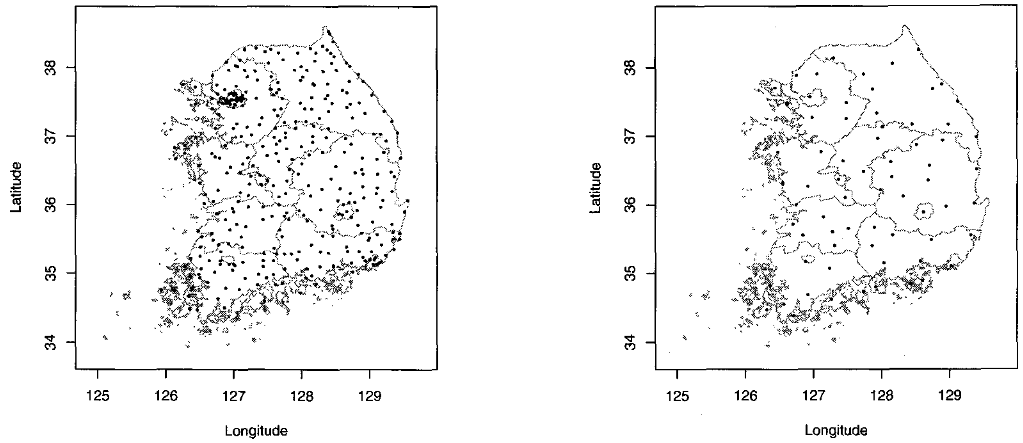
where σ^2 and ϕ may be viewed as hyper-parameters in Bayesian context. We now have an interest in estimating the spatial model of $\boldsymbol{\epsilon}^1|\mathbf{y}$ and obtaining spatial predictions of $\epsilon^1(\mathbf{s}_0)|\mathbf{y}$ at a new location $\mathbf{s}_0$. Finally, the Bayesian kriging is the prediction at a new location $\mathbf{s}_0$ by using the covariate values at the location $\mathbf{x}(\mathbf{s}_0)$ and the set of covariates for the observed locations, $\mathbf{X}$. So, for each of the new locations, we can calculate the predictive posterior values from the density

$$\pi(Y_0|\mathbf{y}, \mathbf{X}, \mathbf{x}(\mathbf{s}_0)) = \int \pi(Y_0|\mathbf{y}, \boldsymbol{\theta}, \mathbf{x}(\mathbf{s}_0))\pi(\boldsymbol{\theta}|\mathbf{y}, \mathbf{X})\, d\boldsymbol{\theta}.$$

We may estimate $\pi(Y_0|\,\mathbf{y}, \mathbf{X}, \mathbf{x}(\mathbf{s}_0))$ using MCMC method for most priors on $\boldsymbol{\theta}$ in practice. More details are found in Cressie (1993) and Banerjee (2004).

3. Real Application

We summarized the spatial modeling based on the Bayesian kriging approach in Section 2. In this section, we try to apply the Bayesian estimation procedure as well as other two popular estimation ones named Maximum likelihood(ML) and Restricted Maximum likelihood(REML) ones to real data and make the kriging (interpolation) maps. Finally, we compare their performances by means of the validation.



(a) Automated monitoring stations$(n=347)$ (b) Ground monitoring stations $(m = 69)$

Figure 3.1. Geographic location of meteorological monitoring stations

3.1. Data

Before going further to the real application based on Section 2, we briefly explain the sources of the real data. We consider two different but relevant monitoring networks; the ground monitoring station network and the automated monitoring station one. Both of the networks have been installed for the preparation against flood (or drought) and the appropriate usage of water resources. We have the 69 ground monitoring stations in South Korea, the geographical domain considered in this study (Figure 3.1). The automated monitoring network make real-time meteorological observations at isolated or remote places, for example, inaccessible mountainous area and unapproachable islands. About 347 automated monitoring stations are located in South Korea (Figure 3.1).

Now we employ the Bayesian universal kriging method for analyzing the precipitation data, which were measured at the automated monitoring stations. The values with unit of millimeter are the monthly averages obtained for July, August, and September, 2007. Once we construct the spatial model with the automated monitoring stations data, we evaluate validity of the model by means of another precipitation data set, which was made from the 69 ground monitoring stations. In order identify the large-scale variation $\mu(\mathbf{s}_i)$ shown in (2.1), we first checked the relations between location information and the precipitation measurement. As can be seen from Figure 3.2, the precipitation values are related to the longitude (quadratic association) and the latitude (linear association). Hence we considered the second-order polynomial function of location information as the large-scale variation, that is,

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

where $X_1(\mathbf{s}_i)$ and $X_2(\mathbf{s}_i)$ are longitude and latitude of station $\mathbf{s}_i$, respectively. For the correlation matrix, $\mathbf{H}_\phi = \{h(\phi)_{ij}\}$ of ϵ^1, we considered the following exponential correlation structure

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

where $||\mathbf{s}_i - \mathbf{s}_j||$ denotes the Euclidean distance between locations, i and j.

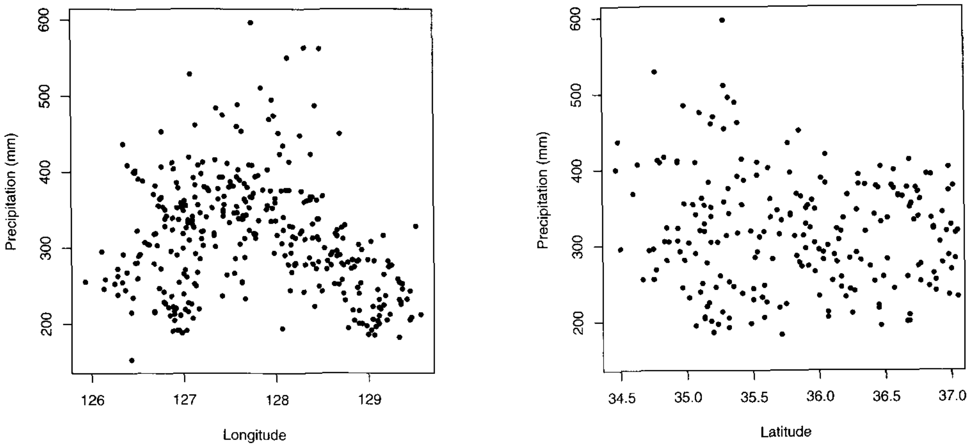


Figure 3.2. Scatterplots of precipitation measurements and location information.

3.2. Bayesian modeling

The Bayesian hierarchical modeling is achieved using some functions in "spBayes" package (Finley *et al.*, 2008) in R (R Development Core Team, 2008). Noninformative priors were assigned to all hyperparameters. There are two types of parameters: 1) regression coefficients, $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)'$ in the mean function, and 2) partial sill(σ^2), nugget (τ^2), and range (ϕ) in variance-covariance matrix. We assume that the measurement error process and spatial process are mutually independent, that is $\epsilon^0 \perp \epsilon^1$. The prior distributions are specified as follows:

$$\begin{aligned} \beta_i &\sim \text{flat prior} \quad i = 0, 1, \ldots, p, \\ 1/\phi &\sim \text{Uniform}\left(d_M^{-1}, 1\right), \\ \sigma^2 &\sim \text{Inverse-Gamma}(0.001, 0.001) \quad \text{and} \\ \tau^2 &\sim \text{Inverse-Gamma}(0.001, 0.001), \end{aligned} \tag{3.1}$$

where $d_M = 500\text{km}$ denotes the maximum of distances between the automated monitoring stations. Using "spBayes" package, we obtained samples from the posterior densities of the parameters. To check the convergence problem, we used several graphical techniques, such as the trace plot and autocorrelation plot. More general discussion of MCMC convergence monitoring is available in Carlin and Louis (2000). We considered one chain with burn-in of 15,000 and 5,000 updates. Initial parameter vector is set to $\boldsymbol{\theta} = (0, 0, 0, 0, 0, 0, 2500, 1000, 25)^T$.

3.3. Result

Table 3.1 displays the parameter estimates and the 95% confidence intervals obtained from the three estimation method considered in this study: Maximum likelihood, Restricted ML, Bayesian hierarchical estimation methods. The parameter estimates in the large-scale variation are quite similar regardless of the estimation methods while the confidence intervals are slightly different. The Bayesian estimation method provides the confidence intervals for the parameters in the spatial covariance structure, $\Sigma_{\theta_1} = \tau^2\mathbf{I} + \sigma^2\mathbf{H}_\phi$. The posterior means of the parameters using the Bayesian approach are quite similar with the estimates from the other two estimation methods. In case of the spatial range parameter, ϕ, all the estimation methods show that the precipitation value at

Table 3.1. Parameter estimates and 95% confidence intervals

	Maximum Likelihood			Restricted ML			Bayesian		
	Est.	95% CI		Est.	95% CI		Est.†	95% CI‡	
β_0	354.4	353.0	355.7	354.2	352.4	356.0	354.1	308.9	396.7
β_1	−0.1456	−0.1371	−0.1371	−0.1404	−0.1512	−0.1297	−0.1375	−0.3540	0.1037
β_2	−0.0410	−0.0472	−0.0348	−0.0420	−0.0500	−0.0340	−0.0452	−0.2344	0.1256
β_3	−0.0054	−0.0055	−0.0053	−0.0052	−0.0053	−0.0051	−0.0052	−0.0073	−0.0026
β_4	−0.0002	−0.0003	0.0001	−0.0003	−0.0004	−0.0002	−0.0003	−0.0018	0.0010
β_5	0.0023	0.0022	0.0024	0.0022	0.0021	0.0023	0.0022	0.0002	0.0040
σ^2	2370.7			2661.4			2981.3	1795.2	6924.4
τ^2	954.4			1076.3			1036.8	484.6	1645.5
ϕ	20.0			27.6			32.8	14.9	131.3

Notes: Est.: estimate; †: Posterior mean; ‡: 95% Bayesian confidence interval.

a monitoring station is influenced only by neighboring stations, which are located within a small distance. Moreover, the upper limit of the 95% confidence interval is at most 131.3, which are about 25% of maximum distance. The local downpours or local severe storm are the characteristics of precipitation in summer, South Korea. Therefore, we guess the range of spatial correlation is not too long.

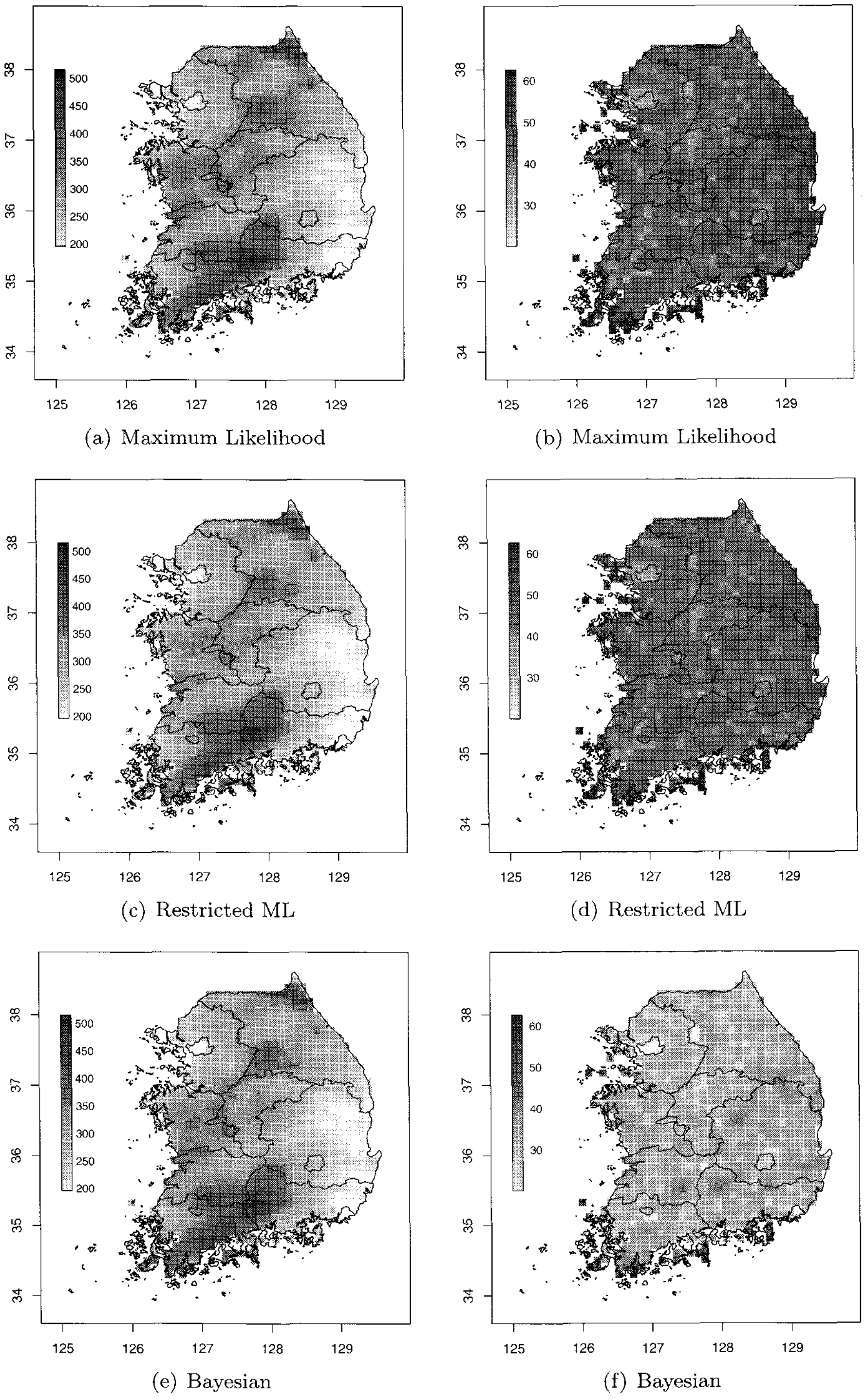
Figure 3.3 illustrates the prediction maps based on the estimates displayed in Table 3.1. In order to construct the prediction map for each estimation method, we preassigned the lattice grid points covering the spatial domain. In the Bayesian method, we used the mean of predictive posterior values at each point. As can be seen from Figure 3.3(a), (c), (e), the maps of kriging estimates look quite similar. The prediction maps show that the eastern region of Gyoungsang province and Seoul have smaller precipitation than average and south-western region of Gyoungsang province and north-eastern region of Kangwon province are predicted to have the largest precipitation. Because the rainy season over Korea, called Changma, continues for a month from late June to late July. A short period of rainfall comes in early September when the monsoon front retreats back to the north. In case of the kriging standard deviation (Figure 3.3(b), (d), (f)), the Bayesian method has much smaller variation for the prediction than the Maximum likelihood-based estimation methods. The natural reason is that the posterior values for the parameters are calculated and refined via an MCMC simulation.

In order to check the validity of the model suggested in (2.3) based on the estimation methods, we predict the precipitation values at each of the 69 ground monitoring stations and compare them with the observed values. The prediction based on the ordinary least squares estimation method is also considered. For the comparison, we considered the following relationship between the predicted values, $\{\widehat{Y}(\mathbf{s}_i)\}$ and the observed ones, $\{Y(\mathbf{s}_i)\}$:

$$\widehat{Y}(\mathbf{s}_i) = \alpha_0 + \alpha_1 Y(\mathbf{s}_i) + e(\mathbf{s}_i), \quad i = 1, \ldots, m(= 69). \tag{3.2}$$

We then computed the intercept (α_0) and the slope (α_1) for each estimation method. By using the residuals $\{\widehat{e}(\mathbf{s}_i)\}$ from the model in (3.2), we obtained the following statistics:

$$\text{MSPE} = \frac{1}{m}\sum_{i=1}^{m} \widehat{e}(\mathbf{s}_i)^2$$



(a) Maximum Likelihood

(b) Maximum Likelihood

(c) Restricted ML

(d) Restricted ML

(e) Bayesian

(f) Bayesian

Figure 3.3. Maps of kriging estimates (first column) and the standard deviations.

Table 3.2. Comparison of predictions and observations

	OLS		ML		ReML		Bayesian	
	Est.	P-value	Est.	P-value	Est.	P-value	Est.	P-value
Intercept	192.15	<0.001	155.05	<0.001	150.18	<0.001	152.19	<0.001
Slope	0.40	<0.001	0.52	<0.001	0.54	<0.001	0.53	<0.001
MSPE	56.17		54.24		54.05		54.17	
Mean(Error)	−1.37		0.68		0.54		0.51	
Median(Error)	7.67		6.91		5.60		4.83	

Notes. Est.: estimate

and mean (median) of the residuals. From Table 3.2, we know that, in terms of precision(MSPE) and unbiasedness (mean(median) of the errors), the estimation methods taking spatial correlation structure into account outperforms the least squares estimation method in that the former have less MSPE and biases closer to zero than the latter. We also know that the slope estimates from the estimation methods focused on in this paper are closer to one, which is an ideal situation under the assumption that the two station network captures the same underlying spatial process although all the estimates for intercept are quite far from zero. Among the estimation methods with spatial correlation structure, the Bayesian method produces slightly more reliable predicted values in terms of the bias.

4. Conclusions

In this paper, we employed the Bayesian kriging method for analyzing the precipitation data along with the commonly used estimation methods. We found that the spatial correlation(dependency) structure should be incorporated the better prediction outcomes even though the spatial range is not quite long. In conclusion, the Bayesian estimation method provides some additional information that the maximum-likelihood-based estimation methods do not and, hence, can be an alternative for the spatial data analysis.

References

Bailey, R. G. (1998). *Ecoregions: The Ecosystem Geography of the Oceans and Continents*, Springer, New York.

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

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.

Carlin, B. P. and Louis, T. A. (2000). *Bayes and Empirical Bayes Methods for Data Analysis*, 2nd Edition, Chapman & Hall/CRC, Boca Raton.

Cressie, N. A. C. (1993). *Statistics for Spatial Data*, John Wily & Sons, New York.

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**, 299–326.

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.

Eom, J. K., Park, M. S., Heo, T. Y. and Huntsinger, L. F. (2006). Improving the prediction of annual average daily traffic for non-freeway facilities by applying spatial statistical method, *Transportation Research Record*, **1968**, 20–29

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.

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. (2008). Seasonal spatial-temporal model for rainfall data of South Korea, *Journal of Applied Sciences Research*, accepted.

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.

Svensson, C. and Rakhecha, P. R. (1998). Estimation of probable maximum precipitation for dams in the Hongru river catchment, China, *Theoretical and Applied Climatology*, **59**, 79–91.