2011 2nd International Conference on Environmental Science and Technology
IPCBEE vol.6 (2011) © (2011) IACSIT Press, Singapore

Spatio-Temporal Interpolation of CO Concentration using a Likelihood-Based Method

Firoozeh Rivaz
Department of Statistics
Shahid Beheshti University, GC
Tehran, Iran
f_rivaz@sbu.ac.ir

Abstract—**Air pollution in large cities due to the increase of population growth by increasing the industrialization progress, has created problems for many residents. So, interpolation of air pollution across space and time is of immense help for sustaining the inhabitants' health level. In this paper, motivated by the statistical analysis of carbon monoxide which is one of the most hazardous air polluting agents in Tehran, we adopt the likelihood approach. In this setting, we assign a prior distribution to the model parameters and use posterior simulations for Markov chain Monte Carlo approximation of likelihood. Then, the maximum likelihood estimates are obtained through the Newton Raphson method.**

Keywords-air pollution; spatio-temporal data; Maximum likelihood estimates; Monte Carlo method

I. Introduction

Air pollution in large cities due to the increase of population growth by increasing the industrialization progress, has created problems for many residents. Tehran, capital of Iran, is one of the polluted cities in the world. Due to its geographical location, the ensnared condition as surrounded by mountain ranges, and also lack of perennial winds, the smoke and other particulate matters produced from the daily life do not vanish in the air. Air pollution and its health consequences has been a major concern for residents, planners and decision makers. Therefore, the demand for spatio-temporal models to assess progress in air quality has grown rapidly over the past decade. This paper concentrates on space-time interpolation of carbon monoxide which is one of the most important agents responsible for the high pollution in Tehran.

For space-time analysis, the Bayesian approach has been seen gaining popularity (e.g., [1],[3],[5],[6],[7]). Bayesian inference proceeds by summarizing the posterior distribution which, after observing data, reflects the uncertainty in the model parameters. Computing the posterior distribution has become feasible particularly, with the advent of the Markov chain Monte Carlo (MCMC) algorithms that are applied when the target statistical distribution contains a high dimensional integral. Notingly, although the Bayesian inferences are computationally feasible, they largely depend on the choice of the prior distributions which usually have a strong unpleasant influence on inferences ([4]).

On the other hand, with regard to high dimensionality of likelihood function, the numerical identification of the maximum likelihood estimates is so difficult. To overcome this problem, assigning a prior distribution to the model parameters and writing likelihood function as a posterior expectation, we provide the maximum likelihood estimation of model parameters through the MCMC output from posterior distribution. Hence, the maximum likelihood estimates are obtained through the Newton Raphson method. Compare to the Bayesian inferences, these inferences are stable so far as the prior distributions are concerned.

The article is organized as follows. Section 2 introduces the statistical model. Section 3 illustrates the new approach for determining maximum likelihood estimation of model parameters. In Section 4, we apply this method in order to analyze a data set related to CO concentration in Tehran city.

II. Statistical Model

Let $Y(\cdot,\cdot) = \{Y(s,t); s \in \mathcal{R}^d, t \in T\}$, $d \leq 1$ be a Gaussian random field with mean $E[Y(s,t) = f'(s,t)\beta$ and covariance function

$$Cov[Y(s_1,t_1), Y(s_2,t_2)] = \sigma^2 \rho(s_1 - s_2, t_1 - t_2; \theta),$$

where $s_1, s_2 \in \mathcal{R}^d$ and $t_1, t_2 \in T$. Where $f(s,t) = (f_1(s,t), \cdots, f_p(s,t))'$ denotes the space-time dependent covariates, $\beta = (\beta_1, \cdots, \beta_p)'$ is unknown regression parameters vector, $\sigma^2 = Var[Y(s,t)]$ is the fixed variance of the random field, $\rho(s_1 - s_2, t_1 - t_2; \theta)$ is the stationary space-time correlation function with parameter vector θ. Suppose that $Z = (Z(s_1,t_1), \cdots, Z(s_{n_S}, t_{n_T}))'$ be a $n_S n_T$-vector represents the data measured at the sampling locations $s_1, \cdots, s_{n_S} \in D$ and time instants $t_1, \cdots, t_{n_T} \in T$ such that

$$Z(s_i,t_j) = Y(s_i,t_j) + \epsilon(s_i,t_j); i = 1, \cdots, n_S, j = 1, \cdots, n_T \quad (1)$$

where $\epsilon(\cdot,\cdot)$ is a white noise process and specifically assumed to follow $N(0, \sigma^2\alpha^2)$ independently. By the stated assumptions, Z has multivariate normal distribution

$$Z \sim N_{n_S n_T}(X\beta, \sigma^2 R_\eta), \quad (2)$$

where $X = (f'(s_1,t_1), \cdots, f'(s_{n_S}, t_{n_T}))'$ is the known full rank $n_S n_T \times p$ matrix, $n_S n_T > p$ and $R_\eta = \Sigma_\theta + \alpha^2 I$ is a positive definite $n_S n_T \times n_S n_T$ matrix with $\eta = (\theta, \alpha^2)$, $\Sigma_\theta = \left(\rho(s_i - s_j, t_i - t_j; \theta)\right)$ and the identity matrix I. Applying the noisy observed data, we are intended in predicting noiseless random field $Y(\cdot,\cdot)$ at arbitrary location s_0 and time t_0. It can be shown that the predictive distribution of $Y(s_0, t_0)$ given z and φ is normally distributed as

$$Y(s_0, t_0) \sim N(\mu_1, \sigma^2 \rho_1)$$

where

$$\mu_1 = f'(s_0, t_0)\beta + r_\theta' {R_\eta}^{-1}(z - X\beta), \tag{3}$$

$\rho_1 = 1 - r_\theta' {R_\eta}^{-1} r_\theta$ and $r_\theta = \left(\rho(s_i - s_0, t_i - t_0; \theta)\right)$. As observed, space-time predictor is depend on unknown parameters.

III. Maximum likelihood estimation

With respect to the unobserved random vector y, the likelihood function of the model parameters $\varphi = (\beta, \sigma^2, \eta)$ based on the observed data $z = (z(s_1, t_1), \cdots, z(s_{n_S}, t_{n_T}))$, is obtained by marginalizing, leading to the likelihood,

$$L(\varphi; \boldsymbol{z}) = \int f(\boldsymbol{z}, \boldsymbol{y}|\varphi) d\boldsymbol{y}. \tag{4}$$

The ML estimates of parameters $\hat{\varphi}$ are the value of φ which maximize the likelihood function. Indeed, the calculations of the likelihood function and ML estimates involve the computational task of high dimension integration. We now explain Monte Carlo approximation to the likelihood. Assuming $\varphi^* \sim \pi(\varphi^*)$ and $f(\boldsymbol{z}) = \int f(\boldsymbol{z}|\varphi^*) d\varphi^*$, we have

$$\begin{aligned} \frac{L(\varphi;\boldsymbol{z})}{f(\boldsymbol{z})} &= \int\int f(\boldsymbol{z}, \boldsymbol{y}|\varphi) \frac{\pi(\varphi^*)}{f(\boldsymbol{z})} d\boldsymbol{y} d\varphi^* \\ &= \int\int \frac{f(\boldsymbol{z},\boldsymbol{y}|\varphi)}{f(\boldsymbol{z},\boldsymbol{y}|\varphi^*)} \pi(\varphi^*, \boldsymbol{y}|\boldsymbol{z}) d\boldsymbol{y} d\varphi^* \\ &= \tilde{E}\left(\frac{f(\boldsymbol{z},\boldsymbol{y}|\varphi)}{f(\boldsymbol{z},\boldsymbol{y}|\varphi^*)} \middle| \boldsymbol{z}\right), \end{aligned} \tag{5}$$

where $\tilde{E}(\cdot\,|\boldsymbol{z})$ is the expectation with respect to the conditional distribution $\pi(\varphi^*, \boldsymbol{y}|\boldsymbol{z})$. Then the Monte Carlo maximum likelihood estimates $\hat{\varphi}$ can be calculated by maximizing

$$L_M(\eta; \boldsymbol{z}) \approx \frac{1}{M} \sum_{i=1}^{M} \frac{f(y_i|\varphi)}{f(y_i|{\varphi_i}^*)} \tag{6}$$

where the values $({\varphi_i}^*, y_i)$, $i = 1, \cdots, M$, are generated from $\pi(\varphi^*, \boldsymbol{y}|\boldsymbol{z})$. Note that Rivaz ${\it et.al}$ (2010) discussed how one can simulate samples from the posterior distribution using MCMC methods. Since the identity (5) holds for any prior distribution, the inferences are invariant to the choice prior.

IV. Interpolation of CO Concentration

This section deals with interpolation of CO concentration using our methodology. The data set are weekly averages of carbon monoxide (CO) in ppm at 11 monitoring sites, geographically distributed across Tehran city, the capital of Iran, during 2004. In order to get a main feature of the data, some exploratory analysis have been performed. The plots of variances versus mean of sites and weeks showed heteroscedasticity suggesting the use of a logarithm transformation to stablize the variance over both sites and weeks. In addition, the normal plot of the square root of data confirm that the data distribution is more closer in agreement with the Gaussian distribution. It must be noted that there is a large-scale spatial trend from marginal areas to the city center. It seems that a bivariate quadratic polynomial surface would be a reasonable structure for spatial trend. Then, we choose the best joint spatio-temporal correlation model using the AIC criterion; the product-sum model ([2]) has the smallest AIC value. Then, based on every fifth draw from an MCMC chain of length 100000 with a burn-in of 50000, which was more than sufficient for convergence, we obtain $\hat{\varphi}$. Weekly predictions of CO for the 1st and the 4th weeks of January 2005 are shown in Fig.1. As observed from the predicted maps, the mean levels for the 4th week are larger than the 1st week due to the effect of the fireworks during revolution victory celebrations. In addition, the predictions are high in northern area of the city for both weeks. The main reason for this higher concentration is arrival of cold winds and city wide snow fall. Since the northern area of the city are surrounded by the Alborz Mountains, it has the freezing weather compare to the central and southern areas. So, the abundance of household fuel consumption for the heating purposes there caused the higher CO in the northern area.

References

[1] S. Banerjee, B. P. Carlin, and A. E. Gelfand, Hierarchical Modeling and Analysis for Spatial Data, Boca Raton, FL: Chapman and Hall/CRC, 2004.

[2] L. De Cesare, D. E. Myers, and D. Posa, “Estimating and Modeling Space-Time Correlation Structures,” Statistics and Probability Letters, vol. 51, pp. 9-14, 2001.

[3] M. B. Hooten and C. K. Wikle, C. K., “A Hierarchical Bayesian Non-Linear Spatio-Temporal Model for The Spread of Invasive Species with Application to the Eurasian Collared-Dove,” Environmental and Ecological Statistics, vol. 15, pp. 59-70, 2008.

[4] F. Rivaz, M. Mohammadzadeh, and M. Jafari Khaledi, “Spatio-Temporal Modeling and Prediction of CO Concentrations in Tehran City, ” Journal of Applied Statistics, to appear, 2011.

[5] S. K. Sahu, A. E. Gelfand, and D. M. Holland, “Spatio-Temporal Modeling of Fine Particulate Matter,” Journal of Agricultural, Biological and Environmental Statistics, vol. 11, pp. 61-86, 2006.

[6] S. K. Sahu, A. E. Gelfand, and D. M. Holland, “High Resolution Space-Time Ozone Modeling for Assessing Trends,” Journal of the American Statistical Association, vol. 102, pp. 1221-1234, 2007.

[7] S. K. Sahu and P. Challenor, “A Space-Time Model for Joint Modeling of Ocean Temperature and Salinity Levels As Measured by Argo Floats,” Environmetrics, vol. 19, pp. 509-528, 2008.

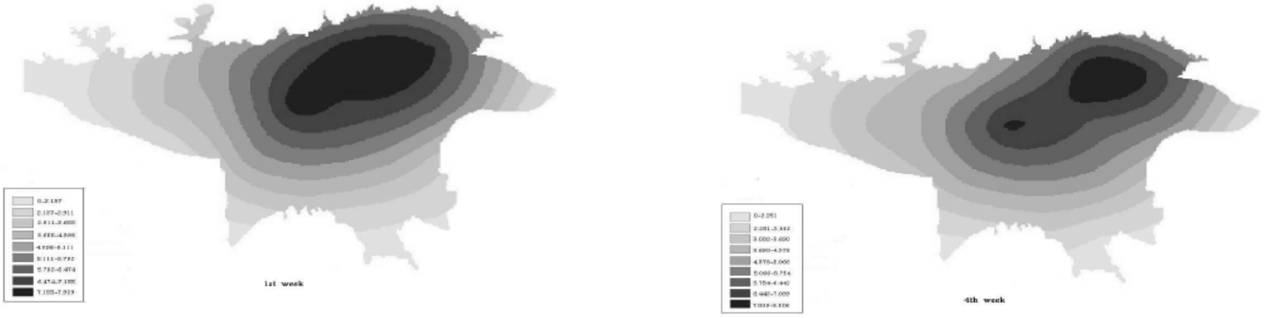

Figure 1. Predicted CO concentrations in Tehran city.