Spatio-Temporal Modeling of Ground-Level Ozone Data

Hsin-Cheng Huang
Institute of Statistical Science
Academia Sinica
Taipei 115, Taiwan
hchuang@stat.sinica.edu.tw

Nan-Jung Hsu
Institute of Statistics
National Tsing Hua University
Hsin-Chu 300, Taiwan
njhsu@stat.nthu.edu.tw

1. Introduction

We propose an autoregressive space-time model for ground-level ozone data, which statistically models the spatio-temporal variation of hourly ozone concentration, and the association between ozone concentration and meteorological variables. The proposed model has a nonseparable spatio-temporal covariance structure that is both nonstationary in time and space. It assumes that the ozone concentration at some location $\boldsymbol{s}$ and time t is directly influenced by past values at neighboring locations via a weight function (Wikle and Cressie 1999, 2000). This weight function is further assumed to have space-time dynamics caused by wind speed and wind direction.

The proposed methodology allows us to understand patterns of transport of ozone. It also allows us to obtain optimal spatio-temporal prediction at any location and time with given meteorological conditions using the space-time Kalman filter (Huang and Cressie, 1996; Wikle and Cressie 1999).

2. The Proposed Model

Let $\{S(\boldsymbol{s},t) : \boldsymbol{s} \in D,\ t = 1,2,\ldots\}$ be a spatio-temporal process defined on a spatial region of interest D, with $|D| > 0$, and time points $t \in I\!N \equiv \{1,2,\ldots\}$. Assume that $S(\boldsymbol{s},t)$ can be decomposed into the following:

$$S(\boldsymbol{s},t) = \mu(\boldsymbol{s},t) + Y(\boldsymbol{s},t) + \nu(\boldsymbol{s},t); \quad \boldsymbol{s} \in D,\ t \in I\!N, \tag{1}$$

where $\mu(\boldsymbol{s},t)$ is a deterministic mean process, $Y(\boldsymbol{s},t)$ is a zero-mean, L_2-continuous, spatio-temporal, Gaussian process, and $\nu(\boldsymbol{s},t)$ is a spatial stationary and temporal independent Gaussian process, representing small-scale spatial variation. The component $Y(\boldsymbol{s},t)$ is assumed to evolve according to the state equation:

$$Y(\boldsymbol{s},t) = \int_D w_{\boldsymbol{s}}(\boldsymbol{x}(\boldsymbol{s},t),\boldsymbol{u})Y(\boldsymbol{u},t-1)d\boldsymbol{u} + \eta(\boldsymbol{s},t); \quad \boldsymbol{s} \in D,\ t \in I\!N, \tag{2}$$

where $\boldsymbol{x}(\boldsymbol{s},t) \equiv (x_1(\boldsymbol{s},t), x_2(\boldsymbol{s},t))'$ is a vector of wind speed $x_1(\boldsymbol{s},t)$ and wind direction $x_2(\boldsymbol{s},t)$ at location $\boldsymbol{s}$ and time t, $\eta(\boldsymbol{s},t)$ is a spatially dependent error process, and $w_{\boldsymbol{s}}(\boldsymbol{x}(\boldsymbol{s},t),\boldsymbol{u})$ is a weight function that depends on $\boldsymbol{x}(\boldsymbol{s},t)$. This weight function represents the relationship between the ozone concentration at location $\boldsymbol{s}$ and time t and that at location $\boldsymbol{u}$ and time $t-1$ under the wind condition $\boldsymbol{x}(\boldsymbol{s},t)$. In general, the weight function changes over time, hence $Y(\boldsymbol{s},t)$ is temporally nonstationary. It can be assumed to be stationary over time if $\boldsymbol{x}(\boldsymbol{s},t) = \boldsymbol{0}$ for all $\boldsymbol{s} \in D$ and $t \in I\!N$, under which the spatial covariance structure of $\eta(\boldsymbol{s},t)$ is assumed. This model is extended from a spatio-temporal model proposed by Wikle and Cressie (1999, 2000), where the weight function does not depend on $\boldsymbol{x}(\boldsymbol{s},t)$.

The data $\{Z(\boldsymbol{s}_i,t) : i = 1,\ldots,n,\ t = 1,\ldots,T\}$ are observed (perhaps incompletely) for n monitoring sites and T time points according to the measurement equation:

$$Z(\boldsymbol{s},t) = S(\boldsymbol{s},t) + \varepsilon(\boldsymbol{s},t); \quad \boldsymbol{s} \in D,\ t \in I\!N, \tag{3}$$

where $\varepsilon(\boldsymbol{s},t)$ is a Gaussian white-noise process with variance σ_ε^2, representing the measurement error.

When $\boldsymbol{x}(\boldsymbol{s},t)=\mathbf{0}$ for all $\boldsymbol{s}\in D$ and $t\in I\!N$, $Y(\boldsymbol{s},t)$ is temporally stationary. Therefore, for each time t, $Y(\boldsymbol{s},t)$ can be decomposed into:

$$Y(\boldsymbol{s},t)=\sum_{k=1}^{\infty}a_k(t)\phi_k(\boldsymbol{s}),$$

where $a_1(t),a_2(t),\ldots$, are uncorrelated random variables with variances $\lambda_1\geq\lambda_2\geq\cdots$, respectively, and $\{\phi_k(\cdot):k\in I\!N\}$ are orthonormal functions. That is,

$$\int_D\phi_k(\boldsymbol{s})\phi_l(\boldsymbol{s})d\boldsymbol{s}=\delta_{kl}\equiv\begin{cases}1; & \text{if } k=l,\\ 0; & \text{otherwise.}\end{cases}$$

The expansion is called the Karhunen-Loève expansion (Loève, 1945; Karhunen, 1947), and the set of orthonormal functions are called the empirical orthogonal functions in geophysics, which satisfy the Fredholm integral equation (Freiberger and Grenander, 1965; Holmström, 1977):

$$\int_D C(\boldsymbol{s},\boldsymbol{s}')\phi_k(\boldsymbol{s}')d\boldsymbol{s}'=\lambda_k\phi_k(\boldsymbol{s});\quad k\in I\!N, \tag{4}$$

where $C(\boldsymbol{s},\boldsymbol{s}')\equiv cov(Y(\boldsymbol{s},t),Y(\boldsymbol{s}',t))$. In practice, data are observed only at n monitoring sites, the expansion is usually truncated at some $K\leq n$:

$$Y(\boldsymbol{s},t)=\boldsymbol{a}_t'\boldsymbol{\phi}(s)\equiv\sum_{k=1}^{K}a_k(t)\phi_k(\boldsymbol{s}), \tag{5}$$

where $\boldsymbol{a}_t\equiv(a_1(t),\ldots,a_K(t))'$ and $\boldsymbol{\phi}(\boldsymbol{s})\equiv(\phi_1(\boldsymbol{s}),\ldots,\phi_K(\boldsymbol{s}))'$. The functions $\{\phi_1(\cdot),\ldots,\phi_K(\cdot)\}$ are usually assumed to be linear combinations of some generation functions $\{e_1(\cdot),\ldots,e_n(\cdot)\}$ with $e_k(\boldsymbol{s}_i)=\delta_{ki}$. Two simple examples of $\{e_i(\cdot)\}$ are piecewise constant functions on Voronoi polygons or facetlike linear functions on Delaunay triangles of $\{\boldsymbol{s}_1,\ldots,\boldsymbol{s}_n\}$ (Obled and Creutin, 1974).

Similar to Wikle and Cressie (1999, 2000), we may model the weight function $w_{\boldsymbol{s}}(\boldsymbol{x},\cdot)$ in terms of K orthonormal functions $\{\phi_1(\cdot),\ldots,\phi_K(\cdot)\}$. However, the function space generated by these orthonormal functions is typically not flexible enough to model the weight function for all $\boldsymbol{s}\in D$ under different wind conditions. Therefore, we consider a larger function space $\mathcal{F}$ on D generated by another sequence of orthonormal functions $\{f_l(\cdot),\ldots,f_L(\cdot)\}$ such that $e_i(\cdot)\in\mathcal{F}$. Write

$$w_{\boldsymbol{s}}(\boldsymbol{x},\boldsymbol{u})=\sum_{l=1}^{L}b_l(\boldsymbol{s},\boldsymbol{x})f_l(\boldsymbol{u}), \tag{6}$$

where $b_1(\boldsymbol{s},\boldsymbol{x}),\ldots,b_L(\boldsymbol{s},\boldsymbol{x})$ are the corresponding coefficients of $f_1(\cdot),\ldots,f_L(\cdot)$, respectively, and write

$$\phi_k(\cdot)=\sum_{l=1}^{L}d_{kl}f_l(\cdot);\quad k=1,\ldots,K. \tag{7}$$

Evaluating (2) on a set of fine grid points $\{\boldsymbol{s}_1^*,\ldots,\boldsymbol{s}_N^*\}$, from (5), (6), and (7), we have

$$\boldsymbol{F}^*\boldsymbol{D}\,\boldsymbol{a}_t=\boldsymbol{B}_t^*\boldsymbol{D}\,\boldsymbol{a}_{t-1}+\boldsymbol{\eta}_t^*, \tag{8}$$

where $\boldsymbol{F}^*$ and $\boldsymbol{B}_t^*$ are $N\times L$ matrices whose (i,l)-th components are $f_l(\boldsymbol{s}_i^*)$ and $b_l(\boldsymbol{s}_i^*,\boldsymbol{x}(\boldsymbol{s}_i^*,t))$, respectively, $\boldsymbol{D}$ is an $L\times K$ matrix whose (l,k)-th component is d_{kl}, and $\boldsymbol{\eta}_t^*\equiv(\eta(\boldsymbol{s}_1^*,t),\ldots,\eta(\boldsymbol{s}_N^*,t))'$. From (1), (3), (8), a reduced state-space model is obtained:

$$\boldsymbol{Z}_t = \boldsymbol{\mu}_t+\boldsymbol{F}\boldsymbol{D}\,\boldsymbol{a}_t+\boldsymbol{\nu}_t+\boldsymbol{\epsilon}_t;\quad t\in I\!N, \tag{9}$$

$$\boldsymbol{a}_t = \boldsymbol{H}_t^*\boldsymbol{a}_{t-1}+\boldsymbol{J}^*\boldsymbol{\eta}_t^*;\quad t=2,3,\ldots, \tag{10}$$

where $\boldsymbol{F}$ is an $n\times L$ matrix whose (i,l)-th component is $f_l(\boldsymbol{s}_i)$, $\boldsymbol{H}_t^*\equiv(\boldsymbol{D}'(\boldsymbol{F}^*)'\boldsymbol{F}^*\boldsymbol{D})^{-1}\boldsymbol{D}'(\boldsymbol{F}^*)'\boldsymbol{B}_t^*\boldsymbol{D}$, $\boldsymbol{J}^*\equiv(\boldsymbol{D}'(\boldsymbol{F}^*)'\boldsymbol{F}^*\boldsymbol{D})^{-1}\boldsymbol{D}'(\boldsymbol{F}^*)'$, $\boldsymbol{Z}_t\equiv(Z(\boldsymbol{s}_1,t),\ldots,Z(\boldsymbol{s}_n,t))'$, $\boldsymbol{\mu}_t\equiv(\mu(\boldsymbol{s}_1,t),\ldots,\mu(\boldsymbol{s}_n,t))'$, $\boldsymbol{\nu}_t\equiv(\nu(\boldsymbol{s}_1,t),\ldots,\nu(\boldsymbol{s}_n,t))'$, and $\boldsymbol{\epsilon}_t\equiv(\varepsilon(\boldsymbol{s}_1,t),\ldots,\varepsilon(\boldsymbol{s}_n,t))'$. Note that $var(\boldsymbol{J}^*\boldsymbol{\eta}_t^*)=\boldsymbol{\Lambda}-\boldsymbol{H}^*\boldsymbol{\Lambda}(\boldsymbol{H}^*)'$, which is needed in computing the optimal prediction of $\boldsymbol{a}_t$, where $\boldsymbol{\Lambda}$ is a diagonal matrix with diagonal elements $\lambda_1,\ldots,\lambda_K$, $\boldsymbol{H}^*\equiv(\boldsymbol{D}'(\boldsymbol{F}^*)'\boldsymbol{F}^*\boldsymbol{D})^{-1}\boldsymbol{D}'(\boldsymbol{F}^*)'\boldsymbol{B}^*\boldsymbol{D}$, and $\boldsymbol{B}^*$ is an $N\times L$ matrix whose (i,l)-th components is $b_l(\boldsymbol{s}_i^*,\mathbf{0})$.

3. Statistical Procedure

In this section, a detailed procedure for modeling ozone data from Taiwan Air Quality Monitoring Network is described. The data we use consist of hourly ozone measurements for year 1999 at 15 stations in Taipei area. Meteorological variables such as wind speed, wind direction, and temperature are also recorded hourly at each station.

Similar to Carroll *et al.* (1997) in analyzing ozone exposure, we use the square-root transformation of the ozone data, and assume the trend component $\mu(\boldsymbol{s},t)$ to be

$$\mu(\boldsymbol{s},t) = \alpha_{month} + \beta_{hour} + \gamma_1 x_0(\boldsymbol{s},t) + \gamma_2[x_0(\boldsymbol{s},t)]^2 + \gamma_3 x_1(\boldsymbol{s},t) + \gamma_4[x_1(\boldsymbol{s},t)]^2; \quad \boldsymbol{s} \in D,\ t \in I\!N,$$

where α_{month} accounts for month effects, β_{hour} accounts for hour effects, and $x_0(\boldsymbol{s},t)$ and $x_1(\boldsymbol{s},t)$ are the temperature and the wind speed at location $\boldsymbol{s}$ and time t, respectively. The trend parameters are estimated using the least-squares method. After subtracting the estimated trend component from the data, we then use the residuals to fit the zero-mean stochastic component $Y(\boldsymbol{s},t) + \nu(\boldsymbol{s},t) + \varepsilon(\boldsymbol{s},t)$.

The weight function $w_{\boldsymbol{s}}(\boldsymbol{x},\boldsymbol{u})$ is assumed to be

$$w_{\boldsymbol{s}}(\boldsymbol{x},\boldsymbol{u}) = \begin{cases} \dfrac{\theta_0}{c}\left[1 - \left(\dfrac{(u_1^*)^2}{(\theta_1+\theta_2 x_1)^2} + \dfrac{(u_2^*)^2}{\theta_1^2}\right)^{1/2}\right]; & \text{if } \dfrac{(u_1^*)^2}{(\theta_1+\theta_2 x_1)^2} + \dfrac{(u_2^*)^2}{\theta_1^2} \le 1,\ u_1^* \ge 0, \\ \dfrac{\theta_0}{c}\left[1 - \left(\dfrac{(u_1^*)^2}{(\theta_1 \exp(-\theta_3 x_1))^2} + \dfrac{(u_2^*)^2}{\theta_1^2}\right)^{1/2}\right]; & \text{if } \dfrac{(u_1^*)^2}{(\theta_1 \exp(-\theta_3 x_1))^2} + \dfrac{(u_2^*)^2}{\theta_1^2} \le 1,\ u_1^* < 0, \\ 0; & \text{otherwise,} \end{cases}$$

where

$$\boldsymbol{u}^* \equiv \begin{pmatrix} u_1^* \\ u_2^* \end{pmatrix} = \begin{pmatrix} \cos x_2 & \sin x_2 \\ -\sin x_2 & \cos x_2 \end{pmatrix} (\boldsymbol{u} - \boldsymbol{s}),$$

and $c \equiv \pi[\theta_1 + \theta_2 x_1 + \theta_1 \exp(-\theta_3 x_1)]\theta_1/6$ is the normalizing constant so that $\int w_{\boldsymbol{s}}(\boldsymbol{x},\boldsymbol{u})d\boldsymbol{u} = \theta_0$. Parameter $\theta_1 > 0$ represents the range of ozone transport from neighboring locations when there is no wind, parameter $\theta_2 > 0$ represents the increase for the range of ozone transport in the upwind direction to an increase of unit wind speed, and parameter $\theta_3 > 0$ controls the decrease of the range of ozone transport in the downwind direction as wind speed increases. Figure 1 shows the weight functions $w_{\boldsymbol{s}}(\boldsymbol{x},\cdot)$ for different values of wind speed ($x_1 = 0,1,2,3$) at $\boldsymbol{s} = \boldsymbol{0}$, where $x_2 = \pi/4$ (i.e., northeast wind), $\theta_0 = 0.5$, $\theta_1 = \theta_2 = 1$, and $\theta_3 = 0.05$. Note that when there is no wind (i.e., $x_1 = 0$), $w_{\boldsymbol{s}}(\boldsymbol{x},\cdot)$ is a cone function.

We use facetlike linear functions on Delaunay triangles of $\{\boldsymbol{s}_1,\ldots,\boldsymbol{s}_n\}$ as our orthonormal functions $\{\phi_1(\cdot),\ldots,\phi_K(\cdot)\}$, and use tensor-product cubic B-splines on fine grid cells as our orthonormal basis functions $\{f_1(\cdot),\ldots,f_L(\cdot)\}$. We assume the covariance function of the small-scale spatio-temporal process $\nu(\boldsymbol{s},t)$ to be $cov(\nu(\boldsymbol{s},t),\nu(\boldsymbol{s}',t')) = \sigma_\nu^2 \exp(-\|\boldsymbol{s}-\boldsymbol{s}'\|/r_\nu)\delta_{tt'}$.

If all the model parameters are known, the reduced state-space model (9) and (10) allows us to compute the optimal mean-squared-error predictor $\hat{\boldsymbol{a}}_{t|t} \equiv E(\boldsymbol{a}_t|\boldsymbol{Z}_1,\ldots,\boldsymbol{Z}_t)$ and the prediction variance $\Sigma_{t|t} \equiv E\left[(\boldsymbol{a}_t - \hat{\boldsymbol{a}}_{t|t})(\boldsymbol{a}_t - \hat{\boldsymbol{a}}_{t|t})'\right]$ recursively using the Kalman-filter algorithm based on the data $\boldsymbol{Z}_1,\ldots,\boldsymbol{Z}_t$. The optimal predictor of $S(\boldsymbol{s},t) = \mu(\boldsymbol{s},t) + \boldsymbol{a}_t'\boldsymbol{\phi}(\boldsymbol{s}) + \nu(\boldsymbol{s},t)$ at any location $\boldsymbol{s}$ and time t based on the data $\boldsymbol{Z}_1,\ldots,\boldsymbol{Z}_t$ is given by

$$\hat{S}(\boldsymbol{s},t) = \mu(\boldsymbol{s},t) + \hat{\boldsymbol{a}}_{t|t}'\boldsymbol{\phi}(s) + (\boldsymbol{C}_\nu(\boldsymbol{s}))'(var(\boldsymbol{Z}_t))^{-1}(\boldsymbol{Z}_t - \boldsymbol{\mu}_t),$$

where $\boldsymbol{C}_\nu(\boldsymbol{s}) \equiv (cov(\nu(\boldsymbol{s},t),\nu(\boldsymbol{s}_1,t)),\ldots,cov(\nu(\boldsymbol{s},t),\nu(\boldsymbol{s}_n,t)))'$.

The model parameters are given by $\boldsymbol{\theta} \equiv \left(\lambda_1,\ldots,\lambda_K,\theta_0,\theta_1,\theta_2,\theta_3,\sigma_\nu^2,r_\nu,\sigma_\varepsilon^2\right)'$. When they are unknown, they can be estimated using the maximum likelihood method. Note that the likelihood function $p(\boldsymbol{Z}_1;\boldsymbol{\theta})\prod_{t=2}^T p(\boldsymbol{Z}_t|\boldsymbol{Z}_{t-1};\boldsymbol{\theta})$ can also be computed recursively using the Kalman-filter algorithm.

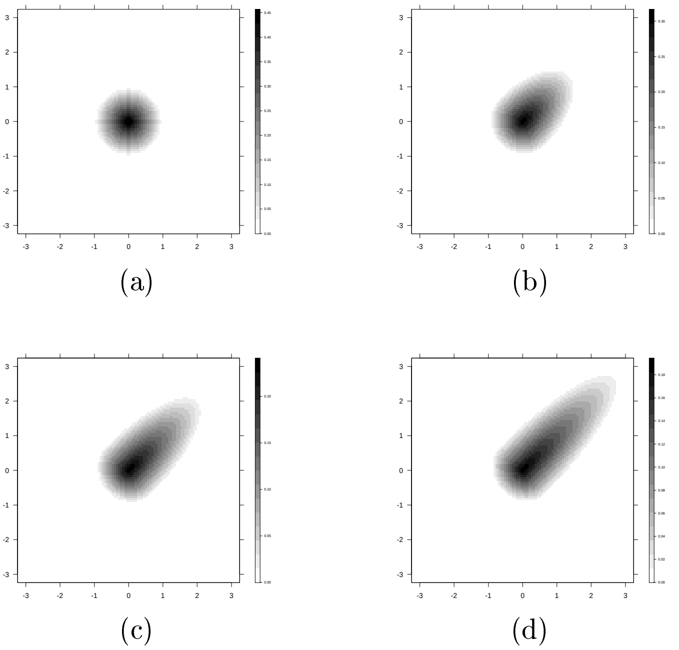

Figure 1: The weight functions $w_{\boldsymbol{s}}(\boldsymbol{x}, \cdot)$ for different values of wind speed at $\boldsymbol{s} = \mathbf{0}$. (a) $x_1 = 0$, (b) $x_1 = 1$, (c) $x_1 = 2$, and (d) $x_1 = 3$, where $x_2 = \pi/4$, $\theta_0 = 0.5$, $\theta_1 = \theta_2 = 1$, and $\theta_3 = 0.05$.

References

Carroll, R. J., Chen, R., George, E. I., Li, T. H., Newton, H. J., Schmiediche, H., and Wang, N. (1997). Ozone exposure and population density in Harris county, Texas. *Journal of the American Statistical Association*, **92**, 392-415.

Freiberger, W. and Grenander, U. (1965). On the formulation of statistical meteorology. *Review of the International Statistical Institute*, **33**, 59-86.

Holmström, I. (1977). Optimization of atmospheric models. *Tellus*, **29**, 415-427.

Huang, H.-C. and Cressie, N. (1996). Spatio-temporal prediction of snow water equivalent using the Kalman filter. *Computational Statistics and Data Analysis*, **22**, 159-175.

Karhunen, K. (1947). Über lineare Methoden in der Wahrscheinlichkeitsrechnung. *Annales Academiase Scientiorum Fennicae, Series AI*, **37**, 1-79.

Loève, M. (1945). Fonctions Aléatores de second ordre. *Comptes Rendus, Académie des Sciences*, Paris, **220**, 469.

Obled, C. and Creutin, J. D. (1986). Some developments in the use of empirical orthogonal functions for mapping meteorological fields. *Journal of Climate and Applied Meteorology*, **25**, 1189-1204.

Wikle, C. K. and Cressie, N. (1999). A dimension reduction approach to space-time Kalman filtering. *Biometrika*, **86**, 815-829.

Wikle, C. K. and Cressie, N. (2000). Space-time statistical modeling of environmental data. *Quantifying Spatial Uncertainty in Natural Resources*, Ann Arbor Press, 213-235.