Spatial Trends and Spatial Extremes in South Korean Ozone

Seokhoon Yun
University of Suwon, Department of Applied Statistics
Suwon, Kyonggi-do 445-743
South Korea
syun@mail.suwon.ac.kr

Richard L. Smith
University of North Carolina, Department of Statistics
Chapel Hill, NC 27599-3260
U.S.A.
rls@email.unc.edu

1. Introduction

Hourly ozone data are available for 73 stations in South Korea from January, 1988 to August, 1998. We are interested in detecting trends in both the mean levels and the extremes of ozone, and in determining how these trends vary over the country. The latter aspect means that we also have to understand the spatial dependence of ozone. In this connection, therefore, we examine the following features:

(a) Determining trends in mean ozone levels at individual stations and combination across stations;

(b) Determining trends in extreme ozone levels at individual stations and combination across stations;

(c) Spatial modeling of trends in mean and extreme ozone levels.

To determine trends in mean ozone levels, one way is to use the mean of hourly ozone measured from 9 a.m. to 5 p.m. for each day. In the present analysis, however, we focus only on daily maxima of hourly ozone since a lot of missing values for that time period were found in the hourly ozone data.

2. Determining trends in mean and extreme ozone levels at individual stations

To investigate a temporal trend in mean ozone levels at a single station, the mean of daily ozone maxima was computed for each month from January, 1988 to August, 1998, for each

of the 73 stations. In case the number of daily ozone maxima in a month is less than 15, the corresponding monthly mean was not computed, being treated as a missing value. Let Y_t denote the monthly mean ozone for month t at a single station. The model considered first is a general linear regression with seasonal components and AR(1) errors, of form

$$\begin{aligned} Y_t &= \beta_0 + \beta_1 t + \beta_2 \cos(2\pi t) + \beta_3 \sin(2\pi t) + \beta_4 \cos(4\pi t) + \beta_5 \sin(4\pi t) + \epsilon_t, \\ \epsilon_t &= \rho \epsilon_{t-h} + \delta_t, \end{aligned} \tag{1}$$

where δ_t's are iid $N(0, \sigma^2)$ disturbances and $|\rho| < 1$. We adopt a base time interval of one year, so that $h = 1/12$ is the length of one month. The parameter β_1 is a linear trend (in ppb per year) and the seasonal components correspond to one-year and six-month cycles respectively.

Using the method of maximum likelihood, model (**??**) was fitted to each of the 73 stations. The results show that there is enormous variability in the estimates of the trend parameter β_1 over the stations: from -2.88 to +3.66 with a mean of .728 and a standard deviation of 1.134. Although the overall mean seems to be slight, when compounded over the $10\frac{2}{3}$ years of the data series, it results an overall increase of about 7.8 ppb ($10\frac{2}{3} \times .728 = 7.765$). Our particular attention is given to Seoul city having 20 stations in which the range of the estimates of β_1 was -1.01 to +3.66 with a mean of 1.088 (slightly higher than the national mean) and standard deviation 1.082. In particular, the station with the highest trend in the nation locates in Seoul city. The results also show that there is very strong evidence of seasonal effects and of positive correlations among the monthly mean ozone levels over time.

The t statistic (parameter estimate divided by standard error) for β_1 was also computed for each station. According to the results, the number of stations for which $t > 2$ is 28 out of the 73 stations, which is substantially greater than the number that would have been expected by chance ($.025 \times 73 = 1.825$), assuming approximate normality of the parameter estimates. Moreover, the number of stations for which $t > 0$ is 56 out of 73, which is substantially larger than would seem plausible by chance alone if there were no overall trend.

Since the monthly mean ozone levels show great variability season to season, a separate analysis was carried out for each of the four seasons (winter, spring, summer and fall). Each season consists of three consecutive months and is considered to continue to the same season of the following year. Thus, winter season, for example, consists of December, January and February in the period 1988-1998. The model fitted to each season is a simple linear regression with AR(1) errors, i.e. $Y_t = \beta_0 + \beta_1 t + \epsilon_t$, with ϵ_t defined as in (**??**), where we adopt a base time interval of one season of three consecutive months, so that $h = 1/3$ is the length of one month.

In each season, the maximum likelihood estimate of the trend parameter β_1 was computed for each of the 73 stations. The range of the estimates and their mean and standard deviation are included in Table 1. Results of t statistics for β_1 at individual stations are also given in Table 1. As expected, summer season has the highest overall mean of 1.116, which implies that there was an overall increase of about 12.3 ppb ($11 \times 1.116 = 12.276$) in summer season during

Table 1. Estimates and t statistics for β_1 in each season

Season	Range of Estimates	Mean	Standard Deviation	$t > 2$	$t > 1$	$t > 0$	$t < 0$	$t < -1$	$t < -2$
Winter	-3.08∼+3.53	.437	1.179	24	35	52	21	10	7
Spring	-3.36∼+5.57	.792	1.741	21	31	48	25	10	6
Summer	-2.28∼+3.92	1.116	1.299	27	43	60	13	5	2
Fall	-3.97∼+2.87	.189	1.243	7	23	44	29	9	3

the 11 years of the data series. In that season, the number of stations for which $t > 2$ and for which $t > 0$ are 27 and 60, respectively, out of 73, which are both substantially larger than would have been expected by chance. The results in Table 1 also show that there were overall positive trends in all seasons.

For investigation of a trend in extreme ozone levels at an individual station, the threshold-based method of Smith (1989) was applied to the time series of daily ozone maxima for each of the 73 stations. Threshold methods are based on fitting stochastic models to the exceedances over a fixed high threshold u, say. Extreme values in a time series typically appear in clusters due to its local dependence. In the present analysis, clusters were defined by the property that two threshold exceedances within three days of each other are considered part of the same cluster. Following Smith (1989), the two-dimensional point process $\{(T_i, Y_i)\}$, where T_i is the time of the ith cluster maximum and Y_i is the value, may be approximated by a nonhomogeneous Poisson process with intensity measure $\lambda(\cdot)$ defined by

$$\lambda((t_1, t_2] \times (y, \infty)) = \int_{t_1}^{t_2} V(y; \xi_t, \mu_t, \sigma_t)\, dt,\ 0 \le t_1 < t_2,\ y \ge u, \tag{2}$$

where $V(y; \xi, \mu, \sigma) = \{1 + \xi(y - \mu)/\sigma\}_+^{-1/\xi}$, $x_+ = \max\{x, 0\}$ and ξ_t, μ_t, σ_t represent respectively a shape parameter, location parameter and scale parameter for time t. Under this model, if we observe the time series on a time interval $(0, T^*]$ and if we observe N cluster maxima at time $T_1, ..., T_N$, then the likelihood function is given by

$$L = \exp\left(- \int_0^{T^*} V(u; \xi_t, \mu_t, \sigma_t)\, dt \right) \prod_{i=1}^{N} v(Y_i; \xi_{T_i}, \mu_{T_i}, \sigma_{T_i}), \tag{3}$$

where $v(y; \xi, \mu, \sigma) = -\partial V(y; \xi, \mu, \sigma)/\partial y$. In practice, the integral in (**??**) is replaced by a sum of form $\int_0^{T^*} V(u; \xi_t, \mu_t, \sigma_t)\, dt \approx h \sum_t V(u; \xi_t, \mu_t, \sigma_t)$, where the sum is over days t and h is the length of one day. We adopt a base time interval of one year, so that $h = 1/365$. If there are missing data, the integral in (**??**) is replaced by an integral over the available period of data.

With the data series of daily ozone maxima for a single station, the model adopted is of form

$$\xi_t = \xi_0,\ \mu_t = \mu_0 e^{\nu_t},\ \sigma_t = \sigma_0 e^{\nu_t}, \tag{4}$$

where ξ_0, μ_0, σ_0 are constants and

$$\nu_t = \beta_1 t + \beta_2 \cos(2\pi t) + \beta_3 \sin(2\pi t) + \beta_4 \cos(4\pi t) + \beta_5 \sin(4\pi t). \tag{5}$$

For any $q \approx 1$ with $q < 1$, let $y_T(q)$ denote the q-level quantile of the distribution of the annual maximum of daily ozone maxima in a one-year time period $(T, T+1]$. Then, under model (??)-(??), it can be seen that $y_{T+1}(q) = e^{\beta_1} y_T(q)$, i.e. the linear trend β_1 is interpretable as an "inflation factor" associated with the extreme quantiles of the annual maxima.

Using the method of maximum likelihood, model (??)-(??) was fitted to each of the 73 stations. In each station, the analysis was repeated, varying the threshold u from the 95th percentile of the empirical distribution of daily ozone maxima to the 98th percentile. For the 95th-percentile threshold, successful model fits were obtained for 65 stations, while for higher-percentile thresholds (the 96th, 97th, 98th) the number of stations with successful fits decreased (62, 59, 47 respectively), which might be due to smaller number of exceedances. Summary statistics for the trend parameter β_1 are given in Table 2. For the 95th-percentile

Table 2. Estimates and t statistics for β_1 in extreme value model

Threshold	Range of Estimates	Mean	Standard Deviation	$t > 2$	$t > 1$	$t > 0$	$t < 0$	$t < -1$	$t < -2$
95th	-.065∼+.067	.0108	.0260	22	31	44	21	11	5
96th	-.065∼+.075	.0117	.0276	20	31	43	19	12	5
97th	-.063∼+.070	.0121	.0284	20	31	38	21	9	4
98th	-.051∼+.076	.0118	.0271	14	25	30	17	7	2

threshold, the overall mean of .0108 corresponds to a rise of approximately 1.1% ($e^{.0108} = 1.01086$) per year in the extreme quantiles, which results an overall increase of about 12.2% ($e^{10\frac{2}{3} \times .0108} = 1.1221$) during the $10\frac{2}{3}$ years of the data series. For higher thresholds, the overall means are slightly higher. The t statistics also reveal evidence of overall positive trends for all thresholds considered.

3. Spatial modeling of trends in mean and extreme ozone levels

The results of section 2 appear to confirm overall positive trends in mean and extreme ozone levels, but are nevertheless hard to interpret because of the enormous spatial variability in the estimates of the trend parameters. In this section, we explore a variant of the usual hierarchical model as one way of spatially smoothing the β_1 estimates obtained in section 2, by assuming the existence of an underlying smooth spatial field.

Let $\beta_1(s)$ denote a temporal trend of interest which is assumed to vary smoothly as a function of spatial location s lying in some domain $\mathcal{S}$. We assume that for each s_i of a fixed subset of spatial locations $\{s_1, ..., s_n\}$, we observe a time series $Y(s_i, t)$, where t is time, whose distribution depends on $\beta_1(s_i)$ as well as possibly other nuisance parameters, which we shall denote

by ϕ. Suppose, for each spatial location s_i, we calculate an estimate of $\beta_1(s_i)$, which we denote by $\hat{\beta}_1(s_i)$, based just on the time series $Y(s_i, t)$. This may be based on any model appropriate for that time series. Since most statistical methods lead to approximately normal distributions of estimators in large samples, we may assume $\hat{\beta}_1(s_i) = \beta_1(s_i) + \eta(s_i)$, where $(\eta(s_1), ..., \eta(s_n))$ is a multivariate normal vector of errors with mean zero and known covariance matrix W. In the present study, we shall assume W to be diagonal with entries determined by the standard errors of the maximum likelihood analyses in section 2. Assuming W to be diagonal contains an implicit assumption that the time series $Y(s_i, t)$, $i = 1, ..., n$, are independent. We also assume that the random field $\{\beta_1(s), s \in \mathcal{S}\}$ is Gaussian with mean and covariance functions given by a finite-parameter model with parameters θ. In particular, the mean vector and covariance matrix of $(\beta_1(s_1), ..., \beta_1(s_n))$ may be written by $\mu_1(\theta)$ and $\sum_1(\theta)$ respectively. Since $\eta(s_i)$ represents measurement error while $\beta_1(s_i)$ reflects the inherent randomness of the environment, it is reasonable to assume that $(\beta_1(s_1), ..., \beta_1(s_n))$ and $(\eta(s_1), ..., \eta(s_n))$ are independent.

With these assumptions, the model now becomes

$$(\hat{\beta}_1(s_1), ..., \hat{\beta}_1(s_n)) \sim N(\mu_1(\theta), \sum_1(\theta) + W)$$

from which the parameters θ may be estimated by the method of maximum likelihood. Moreover, once the parameters θ are estimated, it is then possible to reconstruct smoothed estimates of $\beta_1(s)$, $s \in \mathcal{S}$, by kriging. It remains to specify parametric models for $\mu_1(\theta)$ and $\sum_1(\theta)$. In the present study, we assume that the mean of $\beta_1(s_i)$ is a cubic polynomial function of the two-dimensional vector s_i and the covariance matrix $\sum_1(\theta)$ is of either the Gaussian structure with

$$\mathrm{Cov}(\beta_1(s_i), \beta_1(s_j)) = \theta_2 \exp\left(-\frac{||s_i - s_j||^2}{\theta_1^2}\right)$$

for the trends in mean ozone levels, where $\theta = (\theta_1, \theta_2)$ and $||\cdot||$ is Euclidean distance, or the Matérn structure with

$$\mathrm{Cov}(\beta_1(s_i), \beta_1(s_j)) = \frac{\theta_2}{2^{\theta_3 - 1}\Gamma(\theta_3)} \left(\frac{2\sqrt{\theta_3}||s_i - s_j||}{\theta_1}\right)^{\theta_3} K_{\theta_3}\left(\frac{2\sqrt{\theta_3}||s_i - s_j||}{\theta_1}\right)$$

for the trends in extreme ozone levels, where $\theta = (\theta_1, \theta_2, \theta_3)$ and $K_{\theta_3}(\cdot)$ is the modified Bessel function of the third kind of order θ_3 (Handcock and Stein (1993) gave a detailed account of the Matérn covariance function). The detailed results of the analysis will be given in the talk.

REFERENCES

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

Smith, R.L. (1989). Extreme value analysis of environmental time series: an application to trend detection in ground-level ozone (with discussion). *Statistical Science*, **4**, 367-393.