Assessing exceedance of ozone standards: space-time modeling of fourth highest ozone concentrations

Veronica J. Berrocal

University of Michigan
Department of Biostatistics
Introduction

• By the Clean Air Act, the U.S. Environmental Protection Agency (EPA) is required to monitor, set and revise national ambient air quality standards (NAAQS) for six air pollutants, among which ozone.

• Exposure to elevated levels of ozone has been associated with increased risks of cardiovascular and respiratory disease.

• The current NAAQS for ozone (revised in 2008) is stated as follows: “The annual fourth-highest daily maximum 8-hour concentration of ozone, averaged over 3 years should not exceed 75 ppb”.

• Currently, exceedances of the ozone NAAQS are determined using monitoring sites data.
Introduction

- Sites with 3-year rolling average of the fourth-highest ozone concentration over years 2001-2003 above the NAAQS for ozone.
Introduction

- Sites with 3-year rolling average of the fourth-highest ozone concentration over years 2002-2004 above the NAAQS for ozone.
Introduction

- Sites with 3-year rolling average of the fourth-highest ozone concentration over years 2003-2005 above the NAAQS for ozone.
Introduction

• Sites with 3-year rolling average of the fourth-highest ozone concentration over years 2004-2006 above the NAAQS for ozone.
Introduction

- Sites with 3-year rolling average of the fourth-highest ozone concentration over years 2005-2007 above the NAAQS for ozone.
Introduction

- Sites with 3-year rolling average of the fourth-highest ozone concentration over years 2006-2008 above the NAAQS for ozone.
• Monitoring sites are sparsely located → information on exceedances of the ozone NAAQS is very limited.

• The EPA uses a second source of information to estimate the concentration of air pollutants: the output from numerical air quality models.

• Numerical air quality models are mathematical models that represent various diffusion, chemical and atmospheric processes and obtain estimates of the average concentration of air pollutants over grid cells via numerical methods.
Daily observations $Y(s_1, t), \ldots, Y(s_n, T)$ of 8-hour max ozone concentration at $n=598$ sites in the Eastern United States during years 2001-2008.
• **Daily** observations $Y(s_1, t), \ldots, Y(s_n, T)$ of 8-hour max ozone concentration at $n=598$ sites in the Eastern United States during years 2001-2008.
Daily observations $Y(s_1, t), \ldots, Y(s_n, T)$ of 8-hour max ozone concentration at $n=598$ sites in the Eastern United States during years 2001-2008

we can derive the annual fourth highest ozone concentration $M_Y(s_1, \tilde{t}), \ldots, M_Y(s_n, \tilde{t})$ at each site for years $\tilde{t} = 2001, \ldots, 2008$
Output \( \{X(B_1, t), \ldots, X(B_g, t)\}, \ldots, \{X(B_1, T), \ldots, X(B_g, T)\} \) of daily 8-hour maximum ozone concentration over the Eastern United States provided by a numerical air quality model (CMAQ) at \( g = 40,044 \) 12-km grid cells during years 2001-2008.
• Output \( \{X(B_1, t), \ldots, X(B_g, t)\}, \ldots, \{X(B_1, T), \ldots, X(B_g, T)\} \) of daily 8-hour maximum ozone concentration over the Eastern United States provided by a numerical air quality model (CMAQ) at \( g = 40,044 \) 12-km grid cells during years 2001-2008.
Output $\{X(B_1, t), \ldots, X(B_g, t)\}, \ldots, \{X(B_1, T), \ldots, X(B_g, T)\}$ of daily 8-hour maximum ozone concentration over the Eastern United States provided by a numerical air quality model (CMAQ) at $g = 40,044$ 12-km grid cells during years 2001-2008.

we can derive the annual fourth highest ozone concentration $M_X(B_1, \tilde{t}), \ldots, M_X(B_g, \tilde{t})$ over the Eastern United States for years $\tilde{t} = 2001, \ldots, 2008$. 

Veronica J. Berrocal  
MBI 2015
GEV distribution

• For scalar random variables, the classical theory of extreme values states that if $Y_1, \ldots, Y_T \overset{iid}{\sim} F$ and there exist sequences $\{a_T\} > 0$ and $\{b_T\}$ such that

$$\frac{\max(Y_1, \ldots, Y_T) - b_T}{a_T} \xrightarrow{d} G$$
GEV distribution

• For scalar random variables, the classical theory of extreme values states that if $Y_1, \ldots, Y_T \overset{iid}{\sim} F$ and there exist sequences $\{a_T\} > 0$ and $\{b_T\}$ such that

$$\frac{\max(Y_1, \ldots, Y_T) - b_T}{a_T} \xrightarrow{d} G$$

then $G$ is the Generalized Extreme Value (GEV) distribution with parameters $(\mu, \sigma, \xi)$:

$$G(y) = \begin{cases} 
\exp \left[ - \left\{ 1 + \xi \left( \frac{y - \mu}{\sigma} \right) \right\}^{-\frac{1}{\xi}} \right] & \xi \neq 0 \\
\exp \left[ - \exp \left\{ - \frac{y - \mu}{\sigma} \right\} \right] & \xi = 0
\end{cases}$$
GEV distribution

- For scalar random variables, the classical theory of extreme values states that if $Y_1, \ldots, Y_T \overset{iid}{\sim} F$ and there exist sequences $\{a_T\} > 0$ and $\{b_T\}$ such that

$$\frac{\max(Y_1, \ldots, Y_T) - b_T}{a_T} \overset{d}{\to} G$$

then $G$ is the Generalized Extreme Value (GEV) distribution with parameters $(\mu, \sigma, \xi)$:

$$G(y) = \begin{cases} 
\exp\left[-\left\{1 + \xi\left(\frac{y - \mu}{\sigma}\right)\right\}^{\frac{-1}{\xi}}\right] & \xi \neq 0 \\
\exp\left[-\exp\left\{-\frac{y - \mu}{\sigma}\right\}\right] & \xi = 0
\end{cases}$$

- $\mu$ location parameter
- $\sigma > 0$ scale parameter
- $\xi$ shape parameter: $\xi < 0$ gives the reverse Weibull distribution; $\xi = 0$ gives the Gumbel distribution; $\xi > 0$ gives the Fréchet distribution.
GEV distribution

GEV density function
location=10 and scale=1

Shape=-1; Reverse
Weibull
Shape=1; Frechet
Shape=0; Gumbel
K-th largest order statistic

- Under the conditions that provide an asymptotic GEV distribution for the maximum, one can obtain the asymptotic distribution of the \( k \)-th largest order statistic.

- If \( w = 1 + \xi \cdot \frac{(y-\mu)}{\sigma} \), the cdf of the \( k \)-th largest order statistic is:

\[
F^{(k)}(w) = \exp(-w^{-\frac{1}{\xi}}) \sum_{j=0}^{k-1} \frac{w^{-j}}{j!}
\]
K-th largest order statistic

• Under the conditions that provide an asymptotic GEV distribution for the maximum, one can obtain the asymptotic distribution of the \( k \)-th largest order statistic.

• If \( w = 1 + \xi \cdot \left( \frac{y-\mu}{\sigma} \right) \), the cdf of the \( k \)-th largest order statistic is:

\[
F(k)(w) = \exp(-w^{-\frac{1}{\xi}}) \sum_{j=0}^{k-1} \frac{w^{-\frac{j}{\xi}}}{j!}
\]

• Hence for \( k = 4 \), we can derive the distribution of the fourth largest order statistic.

• We call this distribution, the **FHEV** distribution with parameters \( \xi, \sigma \) and \( \mu \).

• A multivariate version of this distribution is not available.
How we model the fourth highest ozone concentrations

- We adopt a hierarchical model formulation and assume *independence* among sites given spatial latent processes.
How we model the fourth highest ozone concentrations

- We adopt a hierarchical model formulation and assume independence among sites given spatial latent processes: if $M_Y(s)$ denotes the fourth highest ozone concentration at site $s \in S$ then

$$M_Y(s) | \mu(s), \sigma(s), \xi(s) \sim \text{FHEV} (\mu(s), \sigma(s), \xi(s))$$

with
- $\mu(s)$ Gaussian process
- $\sigma(s) > 0$ log-Gaussian process
- $\xi(s)$ Gaussian process
Process model

- We model $\mu(s)$ as a Gaussian process whose mean depends on the numerical model output.
- We assume that $\mu(s)$ has an exponential covariance function with no nugget effect.
- We assume that $\sigma(s)$ is constant in space.
- We assume that $\xi(s)$ is constant in space.
We split the 598 monitoring sites into: 550 training sites and 48 validation sites.

Considered the following values for the shape parameter $\xi$: -1.0, -0.5, -0.1, 0, 0.1, 0.5
Results

- Spatial map of the probability that sites in the Eastern United States are not in compliance with the NAAQS for ozone in years 2001-2003 as provided by downscaler model $\tilde{M}_2$. 
Results

- Spatial map of the probability that sites in the Eastern United States are **not in compliance** with the NAAQS for ozone in years 2006-2008 as provided by downscaler model $\tilde{M}_2$. 