Modeling the Spatial Behavior of the Meteorological Drivers of Extreme Ozone

Brook Russell

Joint Work with: Dan Cooley (CSU), Will Porter (MIT), Brian Reich (NCSU), Colette Heald (MIT)

EPA STAR Grant RD-83522861-0
Where in the World is Clemson University?

Credit: Google Maps
CU’s Snow Family Outdoor Fitness and Wellness Center

Credit: Clemson University Website
1. Introduction
   ▶ Motivation
   ▶ Project Goals
2. Optimizing Tail Dependence
   ▶ Procedure
     ▶ Optimization
     ▶ Model selection
   ▶ ATL/CHAR ozone analysis
3. Spatial Analysis
   ▶ Model and inference
   ▶ Analysis of EPA Regions 3 and 4
4. Summary
Why Study Ozone?

Credit: my iPhone
Project Goals

Goals:

1. Develop and apply EV methods to describe the meteorological conditions for extreme ozone events
2. Model the meteorological drivers of ozone spatially

Credit: TN Dept. of Health Website
Outline

1. Introduction
   ▶ Motivation
   ▶ Project Goals
2. Optimizing Tail Dependence
   ▶ Procedure
     ▶ Optimization
     ▶ Model selection
     ▶ ATL/CHAR ozone analysis
3. Spatial Analysis
   ▶ Model and inference
   ▶ Analysis of EPA Regions 3 and 4
4. Summary
Motivation

Atlanta, GA: Air Temp vs Ground-level Ozone
Identifying the ‘Perfect Storm’

First Goal: Develop EV methods to describe the meteorological conditions for extreme ozone events (i.e., the ‘perfect storm’)

- **Model Fitting:** Given a fixed set of covariates, estimate coefs. in linear combination that yield highest degree of tail dependence with response (ozone)

- **Model Selection:** Find subset of covariates whose linear combination has highest degree of tail dependence with response (ozone)
What is tail dependence?

- $X_1$ and $X_2$ (with common marginals) are tail dependent if
  \[ \lim_{u \to x^+} P(X_2 > u | X_1 > u) > 0 \]

- $x^+$ is upper endpoint of support

- ‘Typical’ dependence metrics may not be useful
Simulated Data

- \( \hat{\rho} \approx .75 \) for both data sets
- Tail dependence is not the same
Multivariate Regular Variation (MVRV)

- MVRV: Probabilistic framework for describing joint tail (implies joint tail decays like power function)
- Useful to transform to pseudo-polar coordinates: $R = \|Z\|$ and $W = \|Z\|^{-1}Z$
- Angular measure $H$ describes distribution of ‘directions’
- In asymptotic dependence case $H$ completely describes tail dependence
- Our method utilizes the bivariate regular variation framework
Why is MVRV Right Useful for Modeling Tail Dependence?

- Defined in terms of tail, says nothing about distn’s ‘bulk’
- Allows for extrapolating further into the tail
- Theoretical justification; directly tied to multivariate max-stable (extreme value) distributions with heavy tails
- Allows for asymptotic dependence
We propose $\gamma$ as a way to measure tail dependence

$$\gamma = \int_0^1 |2w - 1|dH(w)$$

‘EDM-like’ parameter (Resnick (2004) and Larsson and Resnick (2012))

Smaller values of $\gamma$ imply higher levels of extremal dependence
Estimating $\gamma$

- We propose $\hat{\gamma}$ as an estimator of $\gamma$

$$
\hat{\gamma} = \int_0^1 |2w - 1|d\hat{H}(w) = \left(\sum_{t=1}^{n} \delta_t\right)^{-1} \sum_{t=1}^{n} \delta_t \frac{|x_t - y_t|}{x_t + y_t}
$$

- $\delta : (0, \infty) \rightarrow [0, 1]$ is non-decreasing weighting function
- Traditional MVRV analyses use ‘hard threshold’ ($\delta_t$ is indicator function)
- We propose a smoothed threshold
Marginal effects and dependence handled separately:

1. Transform marginals to Unit Fréchet
2. Work in ‘psuedo-polar’ coordinates
3. Estimate $H$ using angular components for ‘large’ points
4. Claims regarding tail dependence can be based on $\hat{H}$ (difficult to describe, use summary metrics instead)
Transformed Simulated Data

Data Set 1 (Unit Frechet Scale)

Data Set 2 (Unit Frechet Scale)
Our Model Fitting Process

\[ Y = X^\beta \]
Form of the Data

Response | Covariates
---|---
$Y_1$ | $X_{1,1}, X_{1,2}, \ldots, X_{1,p}$
$Y_2$ | $X_{2,1}, X_{2,2}, \ldots, X_{2,p}$
$\vdots$ | $\vdots$
$Y_t$ | $X_{t,1}, X_{t,2}, \ldots, X_{t,p}$
$\vdots$ | $\vdots$
$Y_n$ | $X_{n,1}, X_{n,2}, \ldots, X_{n,p}$

- Number of covariates $p$ is large, many unimportant
- Response $Y_t$ is not necessarily regularly varying
- Covariates $X_{t,j}$ not regularly varying, different marginals, supports

Q: How do we fit into the regular variation framework?
Model Fitting Process

Fitting Into RV Framework:

- Transform $Y_t$ to unit Fréchet (denote $Y_t^{**} \sim \text{Fréchet}(1)$)
- Need $X_t^T \beta$ to be unit Fréchet for any $\beta$

Our Solution: A Two-Step Transformation

- **Step 1**
  - $X_{t,j} \rightarrow X_{t,j}^{*}$ which are $N(0, 1)$ for all $j = 1, \ldots, k$
  - Linear combination will be approximately Gaussian
  - Constrain $\beta$ s.t. $X_t^* \beta \sim N(0, 1)$

- **Step 2**
  - Transform linear combination to unit Fréchet
  - Under constraint, $X_t^{**}(\beta) \sim \text{Fréchet}(1)$
Assume \((X_t^{**}(\beta), Y_t^{**})\) is bivariate regularly varying

Constrained optimization:

\[
\beta^* = \arg\min_{\{\beta \in \mathbb{R}^k : \beta^T \Sigma \beta = 1\}} \int_0^1 |2w - 1| dH_\beta(w)
\]

Difficult optimization \(\Rightarrow\) smooth threshold
Need for Smooth Threshold
Optimization Surface
Illustrating Model Fitting Procedure
What combinations of meteorological drivers are associated with the most extreme ozone levels?
Covariates: met data from NARR reanalysis, approx 60 available, we start with a list of 20-30 suspects

▶ reanalysis: climate-model-like data, daily correspondence
▶ no emissions/pollution data

Response: max of 8-hour moving average (daily)
Transforming between marginals relatively straightforward when they are continuous—precipitation isn’t

Solution: add another linear combination which appears only when precipitation is non-negligible

\[ X_t^*(\beta) = \beta_1 X_{t,1} + \ldots + \beta_p X_{t,p} + \]
\[ \mathbb{I}(X_{t,\text{precip}} > \epsilon)(\phi_0 + \phi_1 X_{t,1} + \ldots + \phi_{p'} X_{t,p'}) \].

\[ X^*(\beta) \] is mixture of two normals:
\[ N(0, 1) \text{ and } N((\phi_0, \sigma^2(\phi))) \]

Still has known marginal for any \( \beta \) and \( \phi \)
### Step 1: All Four Variable Models

<table>
<thead>
<tr>
<th>Model</th>
<th>CV</th>
<th>Wind Speed</th>
<th>DSWRF</th>
<th>Precip</th>
</tr>
</thead>
<tbody>
<tr>
<td>Atlanta 1</td>
<td>.5398</td>
<td>.45 (.25)</td>
<td>- .44 (.36)</td>
<td>.53 (.64)</td>
</tr>
<tr>
<td>Atlanta 2</td>
<td>.5508</td>
<td>Wind Speed</td>
<td>Rel Hum</td>
<td>Precip</td>
</tr>
<tr>
<td></td>
<td></td>
<td>.59</td>
<td>-.46</td>
<td>-.28</td>
</tr>
<tr>
<td>Atlanta 3</td>
<td>.5512</td>
<td>Wind Speed</td>
<td>DSWRF</td>
<td>CAPE</td>
</tr>
<tr>
<td></td>
<td></td>
<td>.80</td>
<td>-.39</td>
<td>.34</td>
</tr>
<tr>
<td>Atlanta 4</td>
<td>.5513</td>
<td>Wind Speed</td>
<td>HPBL 7AM</td>
<td>CAPE</td>
</tr>
<tr>
<td></td>
<td></td>
<td>.87</td>
<td>-.36</td>
<td>-.24</td>
</tr>
<tr>
<td>Atlanta 5</td>
<td>.5519</td>
<td>HPBL 7AM</td>
<td>Rel Hum</td>
<td>Precip</td>
</tr>
<tr>
<td></td>
<td></td>
<td>.53</td>
<td>-.39</td>
<td>-.48</td>
</tr>
<tr>
<td>Charlotte 1</td>
<td>.5452</td>
<td>Wind Speed</td>
<td>DSWRF</td>
<td>Precip</td>
</tr>
<tr>
<td></td>
<td></td>
<td>.44 (.33)</td>
<td>-.50</td>
<td>.53 (.61)</td>
</tr>
<tr>
<td>Charlotte 2</td>
<td>.5770</td>
<td>Wind Speed</td>
<td>NW Wind</td>
<td>DSWRF</td>
</tr>
<tr>
<td></td>
<td></td>
<td>.54</td>
<td>-.50</td>
<td>.12</td>
</tr>
<tr>
<td>Charlotte 3</td>
<td>.5772</td>
<td>Wind Speed</td>
<td>DSWRF</td>
<td>Rel Hum</td>
</tr>
<tr>
<td></td>
<td></td>
<td>.52</td>
<td>-.44</td>
<td>.34</td>
</tr>
<tr>
<td>Charlotte 4</td>
<td>.5774</td>
<td>Wind Speed</td>
<td>E Wind</td>
<td>DSWRF</td>
</tr>
<tr>
<td></td>
<td></td>
<td>.54</td>
<td>-.46</td>
<td>-.10</td>
</tr>
<tr>
<td>Charlotte 5</td>
<td>.5781</td>
<td>Wind Speed</td>
<td>DSWRF</td>
<td>TCDC</td>
</tr>
<tr>
<td></td>
<td></td>
<td>.55</td>
<td>-.51</td>
<td>.46</td>
</tr>
</tbody>
</table>
What we learned from Atlanta and Charlotte analyses:

1. Four ‘core’ covariates important: air temp, dswrf, wind speed, and precip.
2. Other variables play secondary role: rel. hum., hpbl, and clouds
3. Local effect of wind direction
Outline

1. Introduction
   - Motivation
   - Project Goals

2. Optimizing Tail Dependence
   - Procedure
     - Optimization
     - Model selection
   - ATL/CHAR ozone analysis

3. Spatial Analysis
   - Model and inference
   - Analysis of EPA Regions 3 and 4

4. Summary
Motivation for Spatial Analysis

1. Understand how primary ozone drivers change spatially
2. Decrease uncertainty associated with parameter estimates

- Solution: MV Spatial Model
- Study Region: EPA Regions 3 and 4

Credit: EPA Website
Different from Other Spatial Extremes Analyses

- We model **drivers** of extreme ozone spatially, differs fundamentally from other spatial extremes studies.
- Previous spatial extremes studies can be (roughly) be divided into two categories:
  1. Model marginal extreme behavior spatially
     - Popular approach: build hierarchical model where parameters of GEV or GPD vary spatially (Cooley et al. (2007), Sang and Gelfand (2009), Dyrrdal et al. (2015)).
     - Useful when interest is in quantities that can be expressed as a function of GEV or GPD parameters (return levels).
  2. Model the spatial extent of extreme events, such as storms
     - Fit max-stable process to block maximum data (Kabluchko et al., 2009)
     - Fit Pareto process to threshold exceedances (Ferreira et al., 2014)
Spatial Model

No likelihood, use hierarchical model with “two-step” inference procedure:

- For each of the six parameters in the common model \((i = 1, \ldots, 6)\) at location \(s \in D \subset \mathbb{R}^2\), assume

\[
\beta_i(s) = \alpha_i + \eta_i(s)
\]

- \(\alpha_i\): constant
- \(\eta_i(s)\): spatially correlated random effects

- Use coregionalization model (Wackernagel, 2003)

\[
\eta(s) = A\delta(s)
\]

- \(\delta_i(s)\): indep. second-order stationary Gaussian processes with mean 0 and variance 1
Choose $A$ to be lower diagonal matrix, suggested by Finley et al. (2008)

Covariance matrix at fixed location:

\[ \text{Cov}(\beta(\cdot)) = \text{Cov}(\eta(\cdot)) = AA^T \]

Assume isotropy, use exponential covariance function

\[ \text{Cov}(\delta_j(s), \delta_j(s')) = \exp(-\|s - s'\|/\rho_j) \]
Cross Covariances

For fixed $i, j \in \{1, \ldots, 6\}$ and $s, s' \in D$:

$$\text{Cov}(\eta_i(s), \eta_j(s')) = \sum_{k=1}^{\min\{i,j\}} a_{ik} a_{jk} \exp\left(-\frac{||s - s'||}{\rho_k}\right)$$  \hspace{1cm} (1)

Special Cases:

- For $i = j$ and $s = s'$ (1) reduces to $\sum_{k=1}^{i} a_{ik}^2$; corresponds to diagonal elements of $AA^T$

- When $i \neq j$ and $s = s'$ (1) reduces to $\sum_{k=1}^{\min\{i,j\}} a_{ik} a_{jk}$; corresponds to off diagonal elements of $AA^T$

- For $i = j$ and $s \neq s'$ (1) simplifies to

$$\sum_{k=1}^{i} a_{ik}^2 \exp\left(-\frac{||s - s'||}{\rho_k}\right)$$
Selecting the Common Model

- Need identical covariates at all locations, call this the ‘common model’
- Exploratory analysis to choose common model
- Must include primary drivers at each location
- Include core covariates
  - Air temp
  - DSWRF
  - Wind speed
  - Precipitation indicator – filter data
- Use three additional covariates
  - Relative humidity
  - Clouds – create new variable
  - TKE
Several cloud variables in the NARR

DSWRF and DLWRF also correlated with clouds

We seek new variable which is linear combination of five of the cloud variables in the NARR:

\[ \mathbf{x} = [x_{cdcon}, x_{cdlyr}, x_{lcdc}, x_{mcdc}, x_{hcdc}]^T \]

We want the parameter vector \( \mathbf{a} \) (unit length) such that:

- \( \text{Var}(\mathbf{a}^T \mathbf{x}) \) is maximized
- \( \text{Cov}(x_{dswrf}, \mathbf{a}^T \mathbf{x}) = 0 \)

Estimate \( \mathbf{a} \) at several locations, similar at all locations.

We define the new cloud variable:

\[ x_{\text{new.cloud}} = 0.47x_{cdcon} - 0.45x_{cdlyr} - 0.37x_{lcdc} + 0.46x_{mcdc} + 0.48x_{hcdc} \]
160 stations give reasonable spatial/temporal coverage
At each location fit common model, obtain:
  - Point estimates for each parameter
  - Bootstrap based estimate of covariance matrix
Air temperature: less important in South
Cloud variable: positive in South, negative in North
TKE: more important in South
Obtain parameter estimates, $\tilde{\beta}_i$ ($i = 1, \ldots, 6$), at each of the 160 stations

We assume

$$\tilde{\beta}_i(s_l) = \beta_i(s_l) + \epsilon_i(s_l)$$

$(\epsilon_1(s_l), \epsilon_2(s_l), \ldots, \epsilon_6(s_l))^T$ represents estimation error.

Also assume:

1. $(\epsilon_1(s_l), \epsilon_2(s_l), \ldots, \epsilon_6(s_l))^T \sim N(\mathbf{0}, \Sigma(s_l))$
2. $(\epsilon_1(s_l), \epsilon_2(s_l), \ldots, \epsilon_6(s_l))^T$ is independent of
   $(\epsilon_1(s_{l'}), \epsilon_2(s_{l'}), \ldots, \epsilon_6(s_{l'}))^T$ for all $s_l \neq s_{l'}$
3. $(\epsilon_1(s_l), \epsilon_2(s_l), \ldots, \epsilon_6(s_l))^T$ is independent of $\eta_i(s_l)$
Second stage of inference: use $\tilde{\beta}_i$ to estimate $\alpha$, $A$, and $\rho$

Utilize sequential maximum likelihood

At each iteration we first maximize $\alpha_i$ (for $i = 1, \ldots, 6$), then $A$, then $\rho_i$ (for $i = 1, \ldots, 6$)
Parameter Estimation

- $\hat{\alpha}$ has useful interpretation
  
  $\hat{\alpha} = (0.46, 0.45, -0.23, -0.12, 0.07, -0.19)^T$

- $\hat{\rho}$ and $\hat{A}$ do not have a great deal of interpretability by themselves

- Can be used to estimate attributes of spatial process
Parameter Surfaces: Estimated via Universal co-Kriging

Takeaways:

- Air temp most important in areas with highest emissions, perhaps due to increased sensitivity
- Air temp, DSWRF highly correlated
- Effect of DSWRF and Wind Speed similar everywhere
- RH, TKE more important in South
- Contrast between High/Low clouds important in FL and GA
Estimating Uncertainty: Square Root MSPE

Takeaways:

- More variability in locations not close to stations (as expected)
- DSWRF, Wind Speed have least amount of variability associated with it (play similar role throughout study area?)
- RH, Cloud Var, Air Temp have higher degree of variability relative to magnitude of their parameter ests.
Reducing Uncertainty

Takeaways:

- ATL and CHAR estimates have a high degree of uncertainty.
- Secondary Goal: Reduce uncertainty by pooling spatially.
- All coefficients see some reduction.
- Some coefficients/regions see more reduction.
1. Introduction
   ▶ Motivation
   ▶ Project Goals
2. Optimizing Tail Dependence
   ▶ Procedure
     ▶ Optimization
     ▶ Model selection
   ▶ ATL/CHAR ozone analysis
3. Spatial Analysis
   ▶ Model and inference
   ▶ Analysis of EPA Regions 3 and 4
4. Summary
Summary

- Started with Information from ATL/CHAR analysis
  - Find core four are important in optimizing tail dependence
  - Others play secondary role
- Spatially Model Drivers of Extreme Ozone
  - Use hierarchical model with two-step inference technique
  - Primary drivers vary spatially


Estimate covariance matrix:

\[
\hat{\text{Cov}}((\beta_1(s), \ldots, \beta_6(s))^T) = \hat{A}\hat{A}^T
\]

Estimated correlation matrix:

\[
\hat{\text{Cor}}(\beta(s)) = \begin{pmatrix}
1.00 & 0.40 & 0.68 & 0.45 & -0.82 & 0.20 \\
0.40 & 1.00 & 0.82 & 0.17 & 0.09 & 0.73 \\
0.68 & 0.82 & 1.00 & 0.08 & -0.39 & 0.47 \\
0.45 & 0.17 & 0.08 & 1.00 & -0.41 & 0.65 \\
-0.82 & 0.09 & -0.39 & -0.41 & 1.00 & 0.08 \\
0.20 & 0.73 & 0.47 & 0.65 & 0.08 & 1.00
\end{pmatrix}
\]
Define $\hat{\mathcal{C}}_i(d) := \sum_{k=1}^{i} \hat{a}_{ik}^2 \exp\{-d/\hat{\rho}_k\}$

- Estimated sills: $\hat{C}_i(0) = \sum_{k=1}^{i} \hat{a}_{ik}^2$

<table>
<thead>
<tr>
<th>Est. Sill</th>
<th>Air Temp</th>
<th>.033</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>DSWRF</td>
<td>.005</td>
</tr>
<tr>
<td></td>
<td>RH</td>
<td>.030</td>
</tr>
<tr>
<td></td>
<td>Wnd Spd</td>
<td>.004</td>
</tr>
<tr>
<td></td>
<td>Cloud</td>
<td>.020</td>
</tr>
<tr>
<td></td>
<td>TKE</td>
<td>.013</td>
</tr>
</tbody>
</table>
Find smallest $d$ such that $\hat{C}_i(d)/\hat{C}_i(0) \leq .05$

<table>
<thead>
<tr>
<th></th>
<th>Est. Eff. Range</th>
</tr>
</thead>
<tbody>
<tr>
<td>Air Temp</td>
<td>383.80 km</td>
</tr>
<tr>
<td>DSWRF</td>
<td>151.30 km</td>
</tr>
<tr>
<td>RH</td>
<td>295.10 km</td>
</tr>
<tr>
<td>Wnd Spd</td>
<td>302.10 km</td>
</tr>
<tr>
<td>Cloud</td>
<td>339.80 km</td>
</tr>
<tr>
<td>TKE</td>
<td>229.60 km</td>
</tr>
</tbody>
</table>