Spatial Methods for Modeling Extreme and Rare Events

Samuel A. Morris

North Carolina State University

August 22, 2016

Outline

- Brief overview of theory for extremes
- Three principal contributions:
  1. A spatio-temporal model with flexible tails, asymptotic spatial dependence, and computation on the order of Gaussian models for large space-time datasets (Biometrics)
  2. Modeling spatial rare binary events with a max-stable extension to the GEV link function (JABES)
  3. Empirical basis functions to explore and model extremal spatial dependence (AOAS)

Motivation

- Average behavior is important to understand, but it does not paint the whole picture
    - e.g. When constructing river levees, engineers need to be able to estimate a 100-year or 1000-year flood levels
    - e.g. Probability of ambient air pollution exceeding a certain threshold level
- Estimating the probability of rare events is challenging because these events are, by definition, rare
- Spatial extremes is promising because it borrows information across space
- Spatial extremes is also useful for estimating probability of extremes at sites without data

Motivation

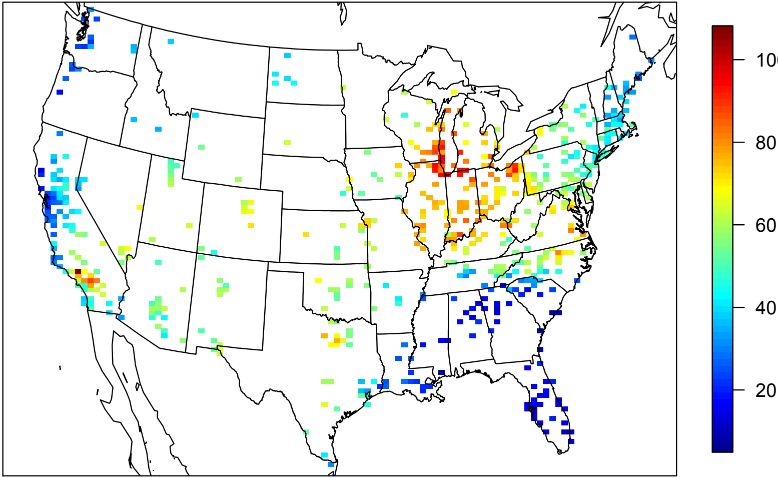


Figure: Max 8-hour ozone measurements on July 10, 2005

Motivation

Ozone compliance for Clean Air Act (EPA)

- Annual fourth-highest daily maximum 8-hour concentration, averaged over 3 years, not to exceed 75 ppb
- Annual fourth-highest is the 99th percentile for the year
- Common objectives are
    - To interpolate to unmonitored sites
    - Detect changes in extremes over time
    - Study meteorological conditions that lead to extreme events

Defining extremes

- Key in extreme value analysis is to define extremes
- Typically done in one of two ways
  - Block maxima (red dots)
  - Values over threshold considered extreme

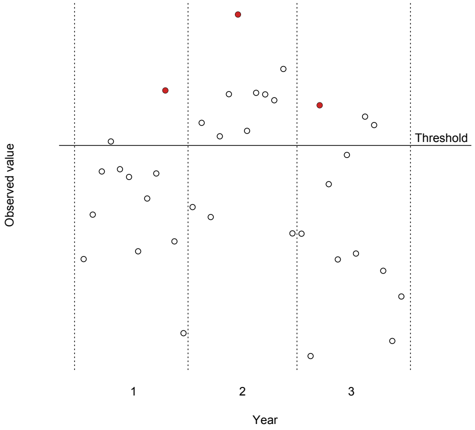


Figure: Hypothetical monthly data

Non-spatial analysis: Block maxima

Fisher-Tippett-Gnedenko theorem

- Let $X_1, \ldots, X_n$ be i.i.d.
- Consider the block maximum $M_n = \max(X_1, \ldots, X_n)$
- If there exist normalizing sequences $a_n > 0$ and $b_n \in \mathcal{R}$ such that

$$\frac{M_n - b_n}{a_n} \xrightarrow{\mathrm{d}} G(z)$$

  then $G(z)$ follows a generalized extreme value distribution (GEV) (Gnedenko, 1943)

- This motivates the use of the GEV for block maximum data

Non-spatial analysis: Block maxima

- GEV distribution

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

  where
  - $\mu \in \mathcal{R}$ is a location parameter
  - $\sigma > 0$ is a scale parameter
  - $\xi \in \mathcal{R}$ is a shape parameter
    - Unbounded above if $\xi \geq 0$
    - Bounded above by $(\mu - \sigma)/\xi$ when $\xi < 0$

- Challenges:
  - Lose information by only considering maximum in a block
  - Underlying data may not be i.i.d.

Non-spatial analysis: Peaks over threshold

Pickands-Balkema-de Haan theorem

- Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} F$
- If there exist normalizing sequences $a_T > 0$ and $b_T \in \mathcal{R}$ such that for any $x \geq 0$, as $T \to \infty$

$$\Pr\left(\frac{X - b_T}{a_T} > x \mid X > T\right) \overset{\text{d}}{\to} H(x),$$

  where T is a thresholding value, then $H(x)$ follows a generalized Pareto distribution (GPD) (Balkema and de Haan, 1974)

Non-spatial analysis: Peaks over threshold

Select a threshold, T, and use the GPD to model the exceedances

$$H(y) = P(Y < y) = \begin{cases} 1 - \left[1 - \xi\left(\frac{y-T}{\sigma}\right)\right]^{-1/\xi} & \xi \neq 0 \\ 1 - \exp\left\{\frac{y-T}{\sigma}\right\} & \xi = 0 \end{cases}$$

where

- $\sigma > 0$ is a scale parameter
- $\xi \in \mathcal{R}$ is a shape parameter
  - Unbounded above if $\xi \geq 0$
  - Bounded above by $(T - \sigma)/\xi$ when $\xi < 0$

Non-spatial analysis: Peaks over threshold

- The GPD is related to GEV distribution through

$$H(y) = 1 + \log[G(y)], \quad y \geq T$$

- Challenges:
  - Sensitive to threshold selection
  - Temporal dependence between observations (e.g. flood levels don't dissipate overnight)

Max-stable processes for spatial data

- Consider i.i.d. spatial processes $x_j(\mathbf{s})$, $j = 1, \ldots, J$
- Let $M_J(\mathbf{s}) = \bigvee_{j=1}^{J} x_j(\mathbf{s}_i)$ be the block maximum at site $\mathbf{s}$
- If there exists normalizing sequences $a_J(\mathbf{s})$ and $b_J(\mathbf{s})$ such that for all sites, $\mathbf{s}_i$, $i = 1, \ldots, d$,

$$\frac{M_J(\mathbf{s}) - b_J(\mathbf{s})}{a_J(\mathbf{s})} \xrightarrow{\mathrm{d}} G(\mathbf{s})$$

  then $G(\mathbf{s})$ is a max-stable process (Smith, 1990)

- Therefore, max-stable processes are the standard model for block maxima

Multivariate representations

- Marginally at each site, observations follow a GEV distribution
- For a finite collection of sites the representation for the multivariate GEV (mGEV) is

$$\Pr(\mathbf{Z} \leq \mathbf{z}) = G^*(\mathbf{z}) = \exp[-V(\mathbf{z})]$$

$$V(\mathbf{z}) = d \int_{\Delta_d} \bigvee_{i=1}^{d} \frac{w_i}{z_i} H(\mathrm{d}w)$$

where

- $V(\mathbf{z})$ is called the exponent measure
- $\Delta_d = \{\mathbf{w} \in \mathcal{R}_+^d \mid w_1 + \cdots + w_d = 1\}$
- H is a probability measure on Δ_d
- $\int_{\Delta_d} w_i H(\mathrm{d}w) = 1/d$ for $i = 1, \ldots, d$

Multivariate GEV challenges

- Only a few closed-form expressions for $V(\mathbf{z})$ exist
- Two common forms for $V(\mathbf{z})$
  - Symmetric logistic (Gumbel, 1960)

$$V(\mathbf{z}) = \left[\sum_{i=1}^{n}\left(\frac{1}{z_i}\right)^{1/\alpha}\right]^{\alpha}$$

  - Asymmetric logistic (Coles and Tawn, 1991)

$$V(\mathbf{z}) = \sum_{l=1}^{L}\left[\sum_{i=1}^{n}\left(\frac{w_{il}}{z_i}\right)^{1/\alpha_l}\right]^{\alpha_l}$$

where $w_{il} \in [0,1]$ and $\sum_{l=1}^{L} w_{il} = 1$

Extremal dependence: χ statistic

- Correlation is the most common measure of dependence
  - Focuses on the center and not tails
  - This makes it irrelevant for extreme value analysis
- Extreme value analysis focuses on the χ statistic (Coles et al., 1999), a measure of extremal dependence given by

$$\chi(h) = \lim_{c \to \infty} \Pr[Y(\mathbf{s}) > c \mid Y(\mathbf{t}) > c]$$

  where $h = ||\mathbf{s} - \mathbf{t}||$

- If $\chi(h) = 0$, then observations are asymptotically independent at distance h

Existing challenges

- Multivariate max-stable models have nice features, but they are
  - Computationally challenging (e.g, the asymmetric logistic has $2^{n-1}(n+2) - (2n+1)$ free parameters)
  - Joint density only available in low dimensions
- Some recent approaches
  - Bayesian hierarchical model (Reich and Shaby, 2012)
  - Pairwise likelihood approach (Huser and Davison, 2014)
- Many opportunities to explore new methods

Back to a Gaussian process model

- The max-stable process is an elegant approach, but does that mean it's the right model?
- In reality, it is only an approximation
- There are less complicated approximations
- For example, we could model daily data as a Gaussian process (GP)
- If the goal is spatial interpolation, perhaps this is competitive

GP - asymptotic independence

- A GP leads to simple interpretation and computing, but asymptotic independence
- If $Y(\mathbf{s}_1)$ and $Y(\mathbf{s}_2)$ are bivariate normal then $\chi(\mathbf{s}_1, \mathbf{s}_2) = 0$, i.e., asymptotic independence
- This suggests Kriging will not capture extremes
- But so much is known for the Gaussian case: nonstationarity, multivariate, numerical approximations, ...
- Rather than toss it out, can we patch it up?

Spatial skew-t process

A spatial skew-t process (Azzalini and Capitanio, 2014) resembles a GP but exhibits asymptotic dependence

$$\begin{aligned} Y_t(\mathbf{s}) &= \mathbf{X}(\mathbf{s})^\top \boldsymbol{\beta} + \lambda \sigma_t |z_t| + \sigma_t v_t(\mathbf{s}) \\ z_t &\sim \text{Normal}(0, 1) \\ \sigma_t^2 &\sim \text{InvGamma}(a/2, b/2) \\ v_t &\sim \text{Spatial GP} \end{aligned}$$

- Location: $\mathbf{X}(\mathbf{s})^\top \boldsymbol{\beta}$
- Scale: $b > 0$
- Skewness: $\lambda \in \mathcal{R}$
- Degrees of freedom: $a > 0$

Good properties

- Flexible t marginal distribution with four parameters including the degrees of freedom which allows for heavy tails ($a = 1$ gives a Cauchy)

- Computation on the order of a GP; the only extra steps are z_t and σ_t which have conjugate full conditionals

- Asymptotic dependence: $\chi(\mathbf{s}_1, \mathbf{s}_2) > 0$ for all $\mathbf{s}_1$ and $\mathbf{s}_2$

Bad properties and remedies

- Modeling all the data (bulk and extreme) can lead to poor tail probability estimates if the model is misspecified
- Long-range dependence: $\chi(\mathbf{s}_1, \mathbf{s}_2) > 0$ for all $\mathbf{s}_1$ and $\mathbf{s}_2$ even if $\mathbf{s}_1$ and $\mathbf{s}_2$ are far apart
- This occurs because all sites share z_t and σ_t
- Remedies:
    - We use a censored likelihood to focus on the tails
    - We propose a local skew-t process

Censored likelihood

- Censored likelihood: We censor the data

$$\tilde{Y}_t(\mathbf{s}) = \begin{cases} T & \text{for } Y_t(\mathbf{s}) \leq T \\ Y_t(\mathbf{s}) & \text{for } Y_t(\mathbf{s}) > T \end{cases}$$

- Censoring is handled using standard Bayesian imputation methods

- The threshold T is chosen by cross-validation

- If T is moderately extreme in the distribution (e.g. $q(0.75)$), set $\lambda = 0$

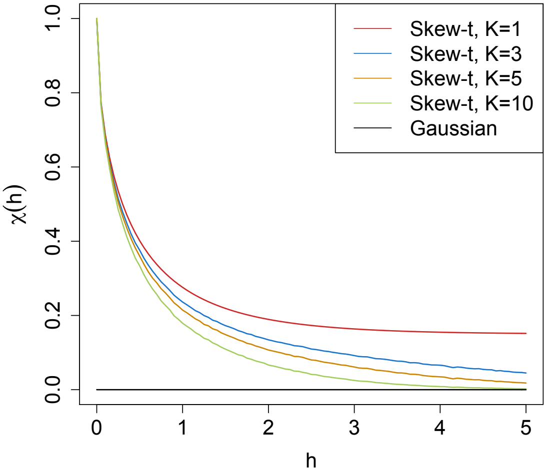
Local skew-t process

- Let the knots $\mathbf{v}_{t1}, ..., \mathbf{v}_{tK}$ follow a homogeneous Poisson process over the domain of interest (in practice we fix K)
- Associated with each is
  - $z_{tk} \sim \text{Normal}(0, 1)$
  - $\sigma^2_{tk} \sim \text{InvGamma}(a/2, b/2)$
- The knots partition the domain if we assign location $\mathbf{s}$ to subregion $k = \arg\min_l ||\mathbf{s} - \mathbf{v}_{tl}||$
- If $\mathbf{s}$ is in subregion k then

$$Y_t(\mathbf{s}) = \mathbf{X}(\mathbf{s})^T \boldsymbol{\beta} + \lambda \sigma_{tk} |z_{tk}| + \sigma_{tk} v_t(\mathbf{s})$$

- The marginal distribution remains a t, but partitioning breaks long-range spatial dependence

χ-statistic by $h = ||\mathbf{s}_1 - \mathbf{s}_2||$



Temporal dependence

- It may not be reasonable to assume that observations are temporally independent (e.g. flooding, high temperatures)
- Temporal dependence is handled through the z_{tk}, σ_{tk} and $\mathbf{v}_{tl}$
- Method:
  - Use a copula to transform parameters to *nice* space (i.e. $\mathcal{R}$)
  - AR(1) structure imposed on parameters in transformed space
  - Transform back to original parameter space to preserve skew-t

Results of a simulation study

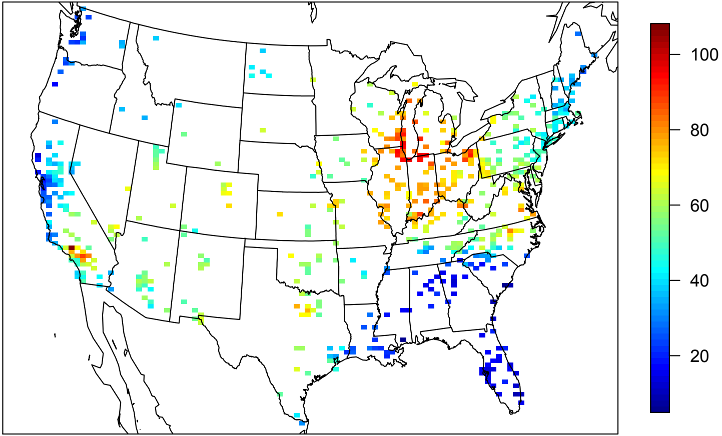
In terms of Brier scores for spatial prediction:

- Data generated as a GP:
  - skew-t is close to GP
  - max-stable is 15% – 30% worse than GP
- Data generated as a skew-t with multiple partitions:
  - skew-t is 15% better than GP
  - max-stable is 30% worse than GP
- Data generated as asymmetric logistic (max-stable):
  - skew-t is close to GP
  - max-stable performs 10% better than GP
- Data generated as Brown-Resnick (max-stable):
  - skew-t performs 40% – 60% better than GP
  - max-stable performs 40% – 60% better than GP

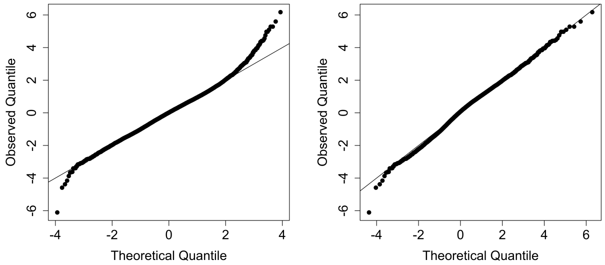
Application to ozone

- The USEPA has an extensive network of ozone monitors throughout the US
- We will analyze ozone for 31 days in July, 2005 at $n = 1,089$ stations
- Currently the EPA regulates the annual 99^{th} percentile
- Our objective is to map the probability of an extreme ozone event

Ozone on July 10



Q-Q plots

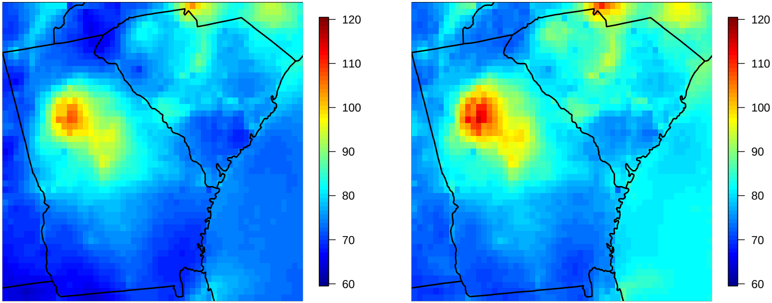


Gaussian Q-Q plot (left) and skew-t with $a = 10$ and $\lambda = 1$ Q-Q plot (right)

Cross-validation

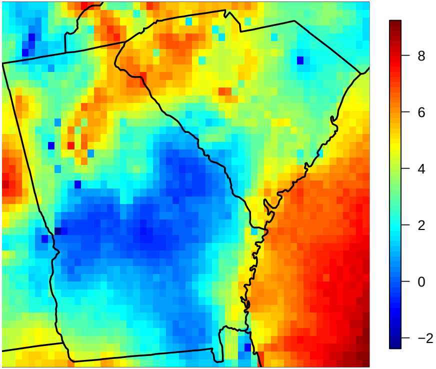
- We split the sites into training and testing
- We found that $K = 15$ knots and censoring at T equal to the median with no time series gave the best results
- Results were not sensitive to these tuning parameters
- This model was 5% more accurate (Brier score) than GP
- The max-stable model fit was 15% less accurate than GP

Fitted 99^{th} percentile - Gaussian



Gaussian (left) Symmetric-t, 10 knots, T = 75, Time series (right)

Difference (Thresholded t - Gaussian)



Difference between Symmetric-t, 10 knots, T = 75 and Gaussian

Questions

- Questions?
- Thank you for your attention.
- Acknowledgment: This work was funded by EPA STAR award R835228