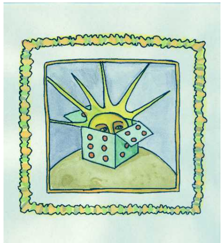
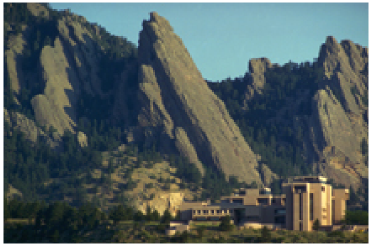
STATISTICAL MODELS FOR QUANTIFYING THE SPATIAL DISTRIBUTION OF SEASONALLY DERIVED OZONE STANDARDS

Eric Gilleland
Douglas Nychka

Geophysical Statistics Project
National Center for Atmospheric Research





Supported by the U.S. National Science Foundation DMS

Outline

- U.S. EPA NAAQS Standard for Ozone
- Data
- The Problem
- Strategies
  - Space-Time Approach[1]
  - Geostatistical Approach
  - Spatial Extreme Value Approach[2]
- Conclusions
- Future and ongoing work

[1] *Accepted to Special Issue of Environmetrics*
[2] *Related to work on* `extRemes`*:*
(`http://www.isse.ucar.edu/extremevalues/evtk.html`*)*

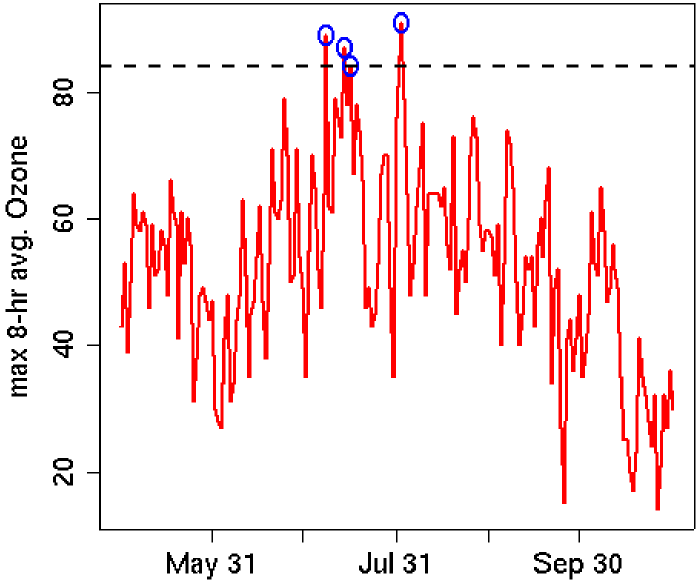
Background for Ozone: Air Quality Standards

As required by the Clean Air Act (CAA) of 1971, the EPA has established standards, known as the *National Ambient Air Quality Standards (NAAQS)*, to monitor and control ambient concentrations for six principal air pollutants (also referred to as criteria pollutants):

- carbon monoxide (CO),
- lead (Pb),
- nitrogen dioxide (NO_2),
- *ground-level Ozone (O_3)*,
- particulate matter (PM) and
- sulfur dioxide (SO_2)

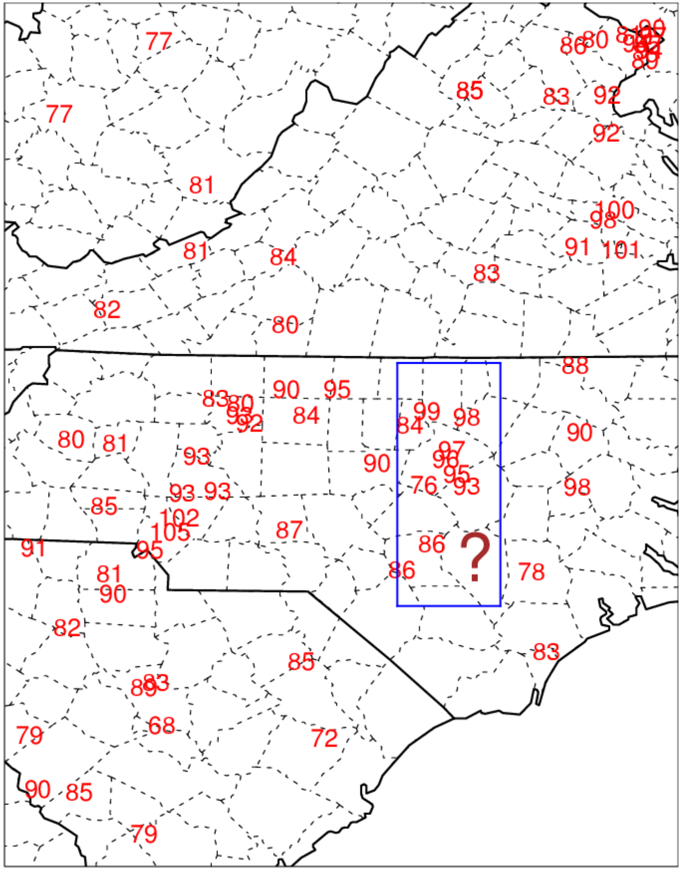
NAAQS for Ground-level Ozone

In 1997, the U.S. EPA changed the NAAQS for regulating ground-level Ozone levels to one based on the *fourth-highest daily maximum 8-hr. averages (FHDA)* of an Ozone season (184 days). Compliance is met when the FHDA over a three year (season) period is below 84 ppb.



Data

Five seasons of daily Ozone data (1995 to 1999) at 72 locations.



The Problem

Although such a standard makes sense from a health (and environment) standpoint, it presents a challenging statistical problem.

Goal

To draw spatial inference for the FHDA at unobserved locations.

Although it is straightforward to build spatial models for the daily Ozone field, the extension to the fourth-highest order statistic is not so simple. *Gaussianity? Covariance?*

The Three Approaches

Daily Model

- Determine an AR model for **every** location, even the unobserved ones. [more]
- Using spatially-coherent shocks, simulate every day of an Ozone season. [more]
- Build up the distribution of the FHDA.

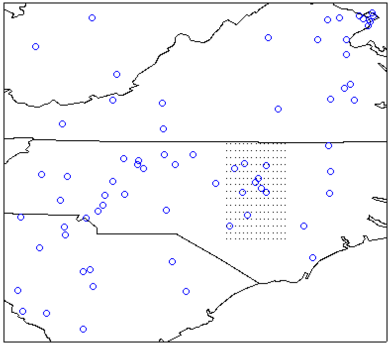
Seasonal Model

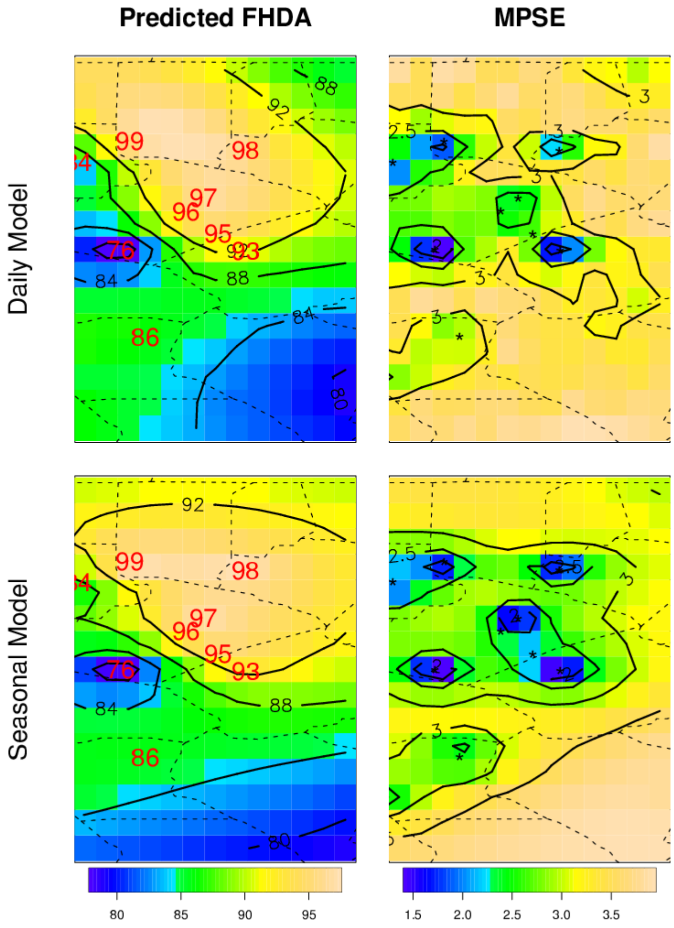
- Straightforward application of kriging to FHDA.

Extremes Model

- Uses a Generalized Pareto, lots more to it.

Daily and Seasonal Predicted FHDA (1997)

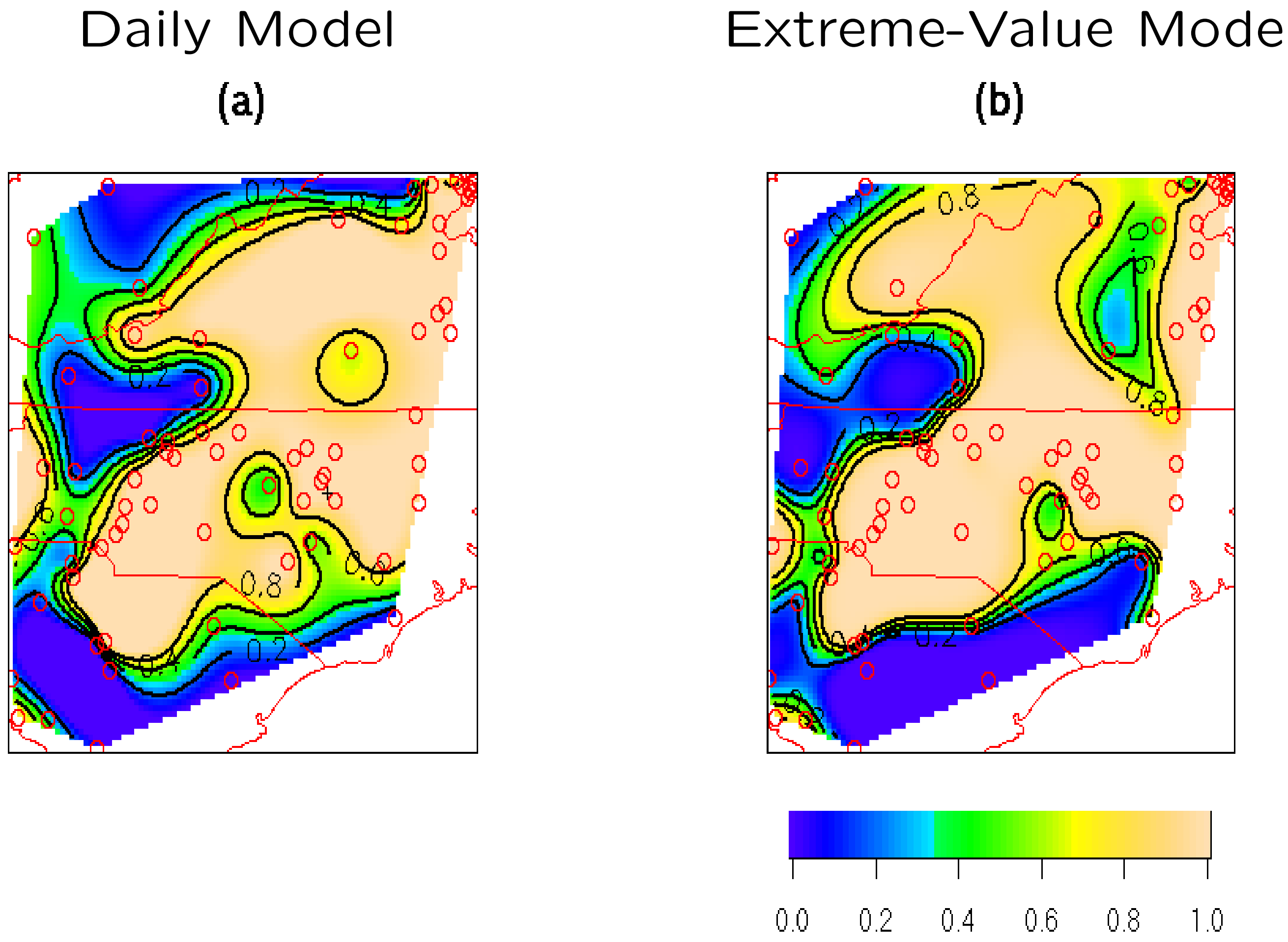




Comparing the Daily and Seasonal models

		TPS	Daily	Seasonal	
				Variogram	Correlation
MPSE	1995	2.23	2.67	5.68	5.27
	1996	2.49	2.85	5.96	5.90
	1997	2.91	3.01	6.41	6.02
	1998	2.75	2.93	5.35	4.85
	1999	4.34	2.94	6.76	6.22
CV RMSE	1995	5.34	4.73	5.19	5.33
	1996	5.61	4.84	5.51	5.68
	1997	6.27	4.59	6.03	6.05
	1998	5.00	3.25	4.98	4.93
	1999	6.25	4.91	6.47	6.30

Probability of exceeding the standard

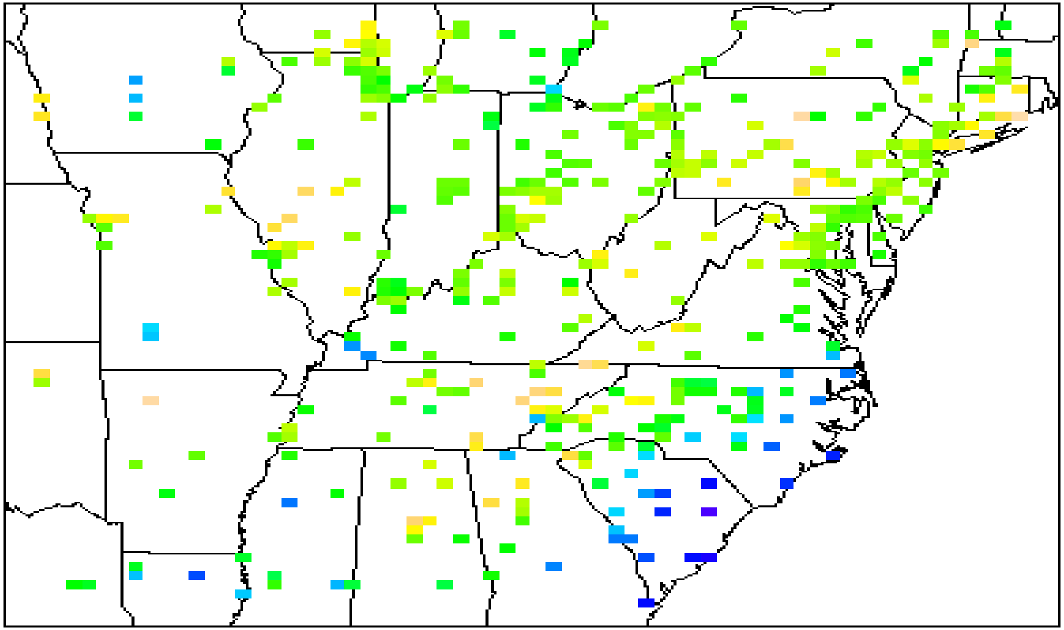


Conclusions

- Simplicity of the seasonal model approach is desirable.
- Daily model yields consistently lower MSE from cross-validation.
- Daily model can account for “complicated” spatial features without resorting to non-standard techniques.
- Daily MPSE is consistently too optimistic.
- Extreme value models good alternative to modelling the tail of distributions.
- Two very different approaches yield similar results

Future and Ongoing Work

- Apply models to network design issues.
- Extending Extremes Toolkit (`extRemes`) to have spatial model.
- Extend model to entire eastern United States (*Nonstationarity of Shocks?*).



That's all folks!

Spatial Extremes

Given a spatial process, $Z(\mathbf{x})$, what can be said about

$$\Pr\{Z(\mathbf{x}) > z\}$$

when z is large?

Spatial Extremes

Given a spatial process, $Z(\mathbf{x})$, what can be said about

$$\mathsf{Pr}\{Z(\mathbf{x}) > z\}$$

when z is large?

Note:
This is not about dependence between $Z(\mathbf{x})$ and $Z(\mathbf{x}')$–this is another topic!

Spatial Extremes

Given a spatial process, $Z(\mathbf{x})$, what can be said about

$$\Pr\{Z(\mathbf{x}) > z\}$$

when z is large?

Note:
This is not about dependence between $Z(\mathbf{x})$ and $Z(\mathbf{x}')$–this is another topic!

Spatial structure on parameters of distribution (not FHDA).

Generalized Pareto Distribution (GPD)

Exceedance Over Threshold Model

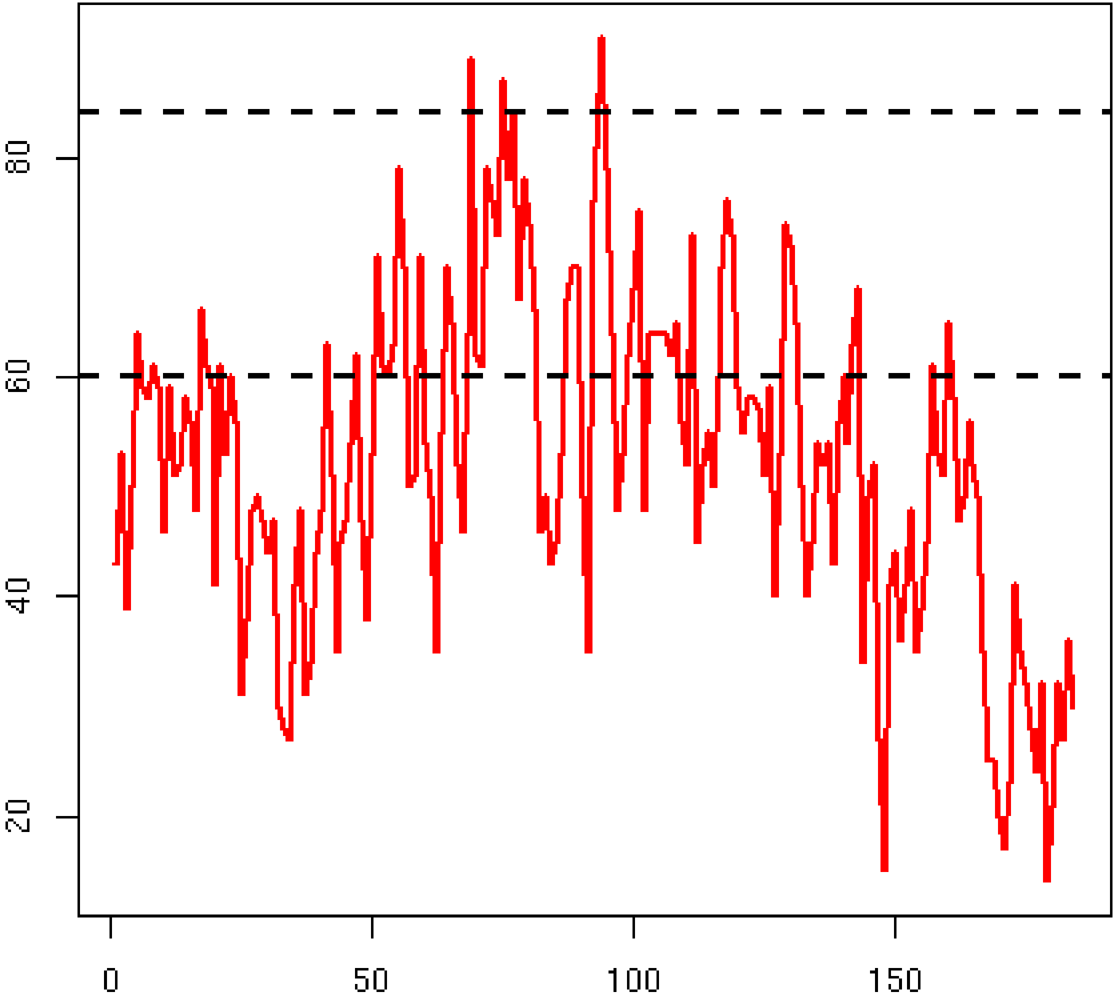
For X random (with cdf F) and a (large) threshold u

$$\Pr\{X > x | X > u\} = \frac{1 - F(x)}{1 - F(u)}$$

Then for $x > u$ (u *large*), the GPD is given by

$$\frac{1 - F(x)}{1 - F(u)} \approx [1 + \frac{\xi}{\sigma}(x - u)]^{-1/\xi}$$

Extreme Value Distributions: GPD



Fitting the GPD to data

Method

Maximum Likelihood is easy to evaluate and maximize.

Threshold Selection: Variance vs. Bias

Trade-off between a low enough u to have enough data (low variance), but high enough for the limit model to be a reasonable approximation (low bias).

Confidence Intervals

The parameter distributions are generally skewed. So, the best method for finding confidence intervals (or sets) are based on the likelihood value (or surface) and a χ^2 critical value.

A Hierarchical Spatial Model

Observation Model:

$y(\mathbf{x}, t)$ surface Ozone at location $\mathbf{x}$ and time t

$$[y(\mathbf{x}, t)|\sigma(\mathbf{x}), \xi(\mathbf{x}), u, y(\mathbf{x}, t) > u]$$

Spatial Process Model:

$$[\sigma(\mathbf{x}), \xi(\mathbf{x}), u|\boldsymbol{\theta}]$$

Prior for hyperparameters:

$$[\boldsymbol{\theta}]$$

A Hierarchical Spatial Model

Assume extreme observations to be *conditionally independent* so that the joint pdf for the data and parameters is

$$\prod_{i,t} [y(\mathbf{x}_i, t) | \sigma(\mathbf{x}), \xi(\mathbf{x}), u, y(\mathbf{x}_i, t) > u] \; [\sigma(\mathbf{x}), \xi(\mathbf{x}), u | \boldsymbol{\theta}] \; [\boldsymbol{\theta}]$$

t indexes time and i stations.

Shortcuts and Assumptions

- $\xi(\mathbf{x}) = \xi$ (i.e., shape is constant over space). *Justified by univariate fits*.
- Assume $\sigma(\mathbf{x})$ is a Gaussian process with isotropic Matérn covariance function.
- Fix Matérn smoothness parameter at $\nu = 2$, and let the range be very large–leaving only λ (ratio of variances of nugget and sill).

More on $\sigma(\mathbf{x})$

λ is the only hyper-parameter–use an uninformative prior for it.

$$\sigma(\mathbf{x}) = P(\mathbf{x}) + e(\mathbf{x}) + \eta(\mathbf{x})$$

with P a linear function of space, e a smooth spatial process, and η white noise (nugget).

- As $\lambda \longrightarrow \infty$, the posterior surface tends toward just the linear function.
- As $\lambda \longrightarrow 0$, the posterior surface will fit the data more closely.

log of joint distribution

$$\sum_{i=1}^{n} \ell_{\mathsf{GPD}}(y(\mathbf{x}_i, t), \sigma(\mathbf{x}_i), \xi) -$$

log of joint distribution

$$\sum_{i=1}^{n} \ell_{\mathsf{GPD}}(y(\mathbf{x}_i, t), \sigma(\mathbf{x}_i), \xi) -$$

$$\lambda(\boldsymbol{\sigma} - \mathbf{X}\boldsymbol{\beta})^T K^{-1}(\boldsymbol{\sigma} - \mathbf{X}\boldsymbol{\beta})/2 - \log(|\lambda K|) + C$$

log of joint distribution

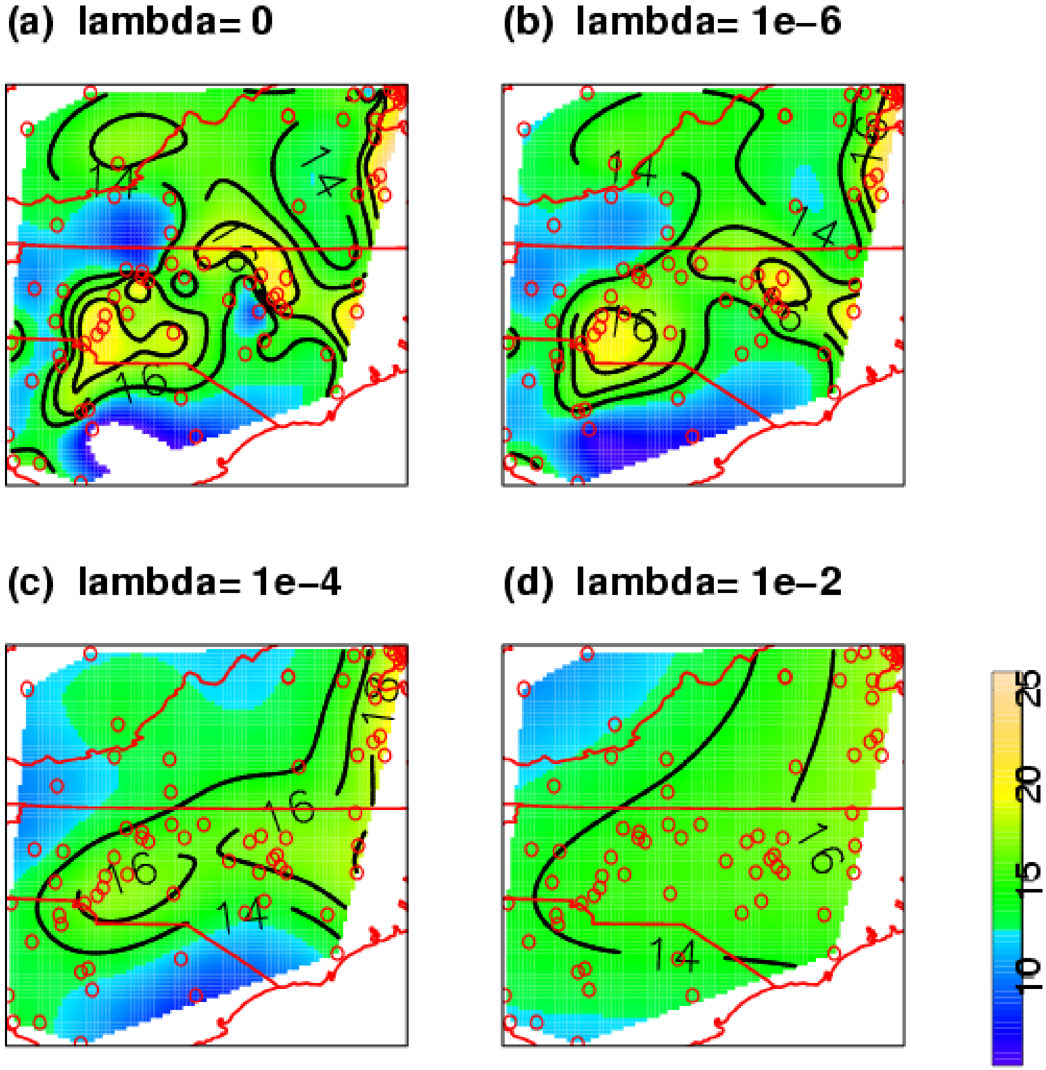
$$\sum_{i=1}^{n} \ell_{\mathsf{GPD}}(y(\mathbf{x}_i, t), \sigma(\mathbf{x}_i), \xi) -$$

$$\lambda(\boldsymbol{\sigma} - \mathbf{X}\boldsymbol{\beta})^T K^{-1}(\boldsymbol{\sigma} - \mathbf{X}\boldsymbol{\beta})/2 - \log(|\lambda K|) + C$$

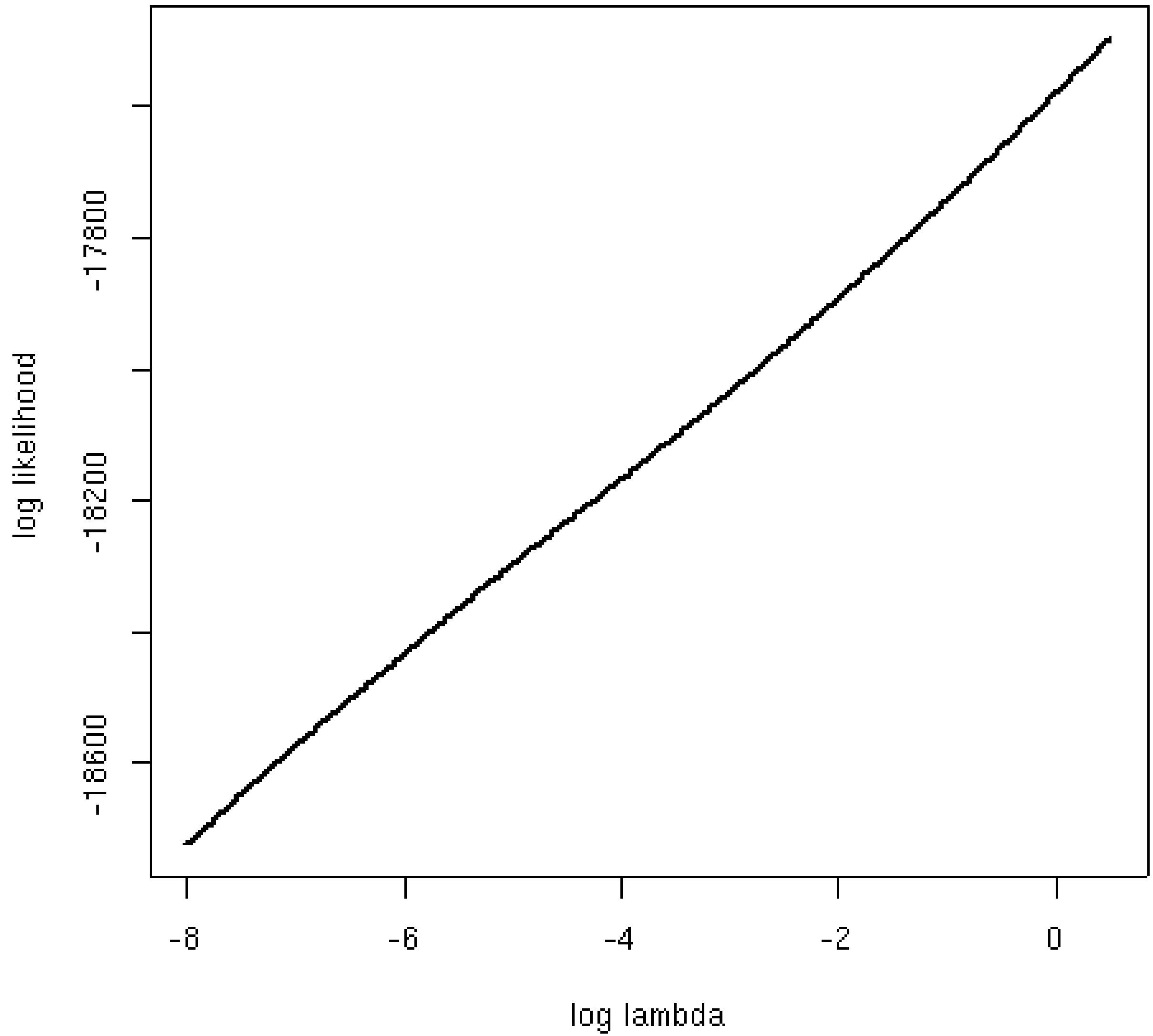
K is the covariance for the prior on σ at the observations.

This is a penalized likelihood:

The penalty on $\boldsymbol{\sigma}$ results from the covariance and smoothing parameter λ.



(a) lambda= 0
(b) lambda= 1e−6
(c) lambda= 1e−4
(d) lambda= 1e−2
14
16
10
15
20
25


-17800
-18200
-18600
log likelihood
-8
-6
-4
-2
0
log lambda

Inference for λ: Try something different.

Fit a thin plate spline to the MLE (of σ) from the univariate fitting, and determine λ by cross-validation:

Inference for λ: Try something different.

Fit a thin plate spline to the MLE (of σ) from the univariate fitting, and determine λ by cross-validation:

1. For fixed λ, fit TPS to all but one location (do this for each location).

Inference for λ: Try something different.

Fit a thin plate spline to the MLE (of σ) from the univariate fitting, and determine λ by cross-validation:

1. For fixed λ, fit TPS to all but one location (do this for each location).

2. Predict the value at the omitted location.

Inference for λ: Try something different.

Fit a thin plate spline to the MLE (of σ) from the univariate fitting, and determine λ by cross-validation:

1. For fixed λ, fit TPS to all but one location (do this for each location).
2. Predict the value at the omitted location.
3. Obtain residuals between the prediction and observation.

Inference for λ: Try something different.

Fit a thin plate spline to the MLE (of σ) from the univariate fitting, and determine λ by cross-validation:

1. For fixed λ, fit TPS to all but one location (do this for each location).

2. Predict the value at the omitted location.

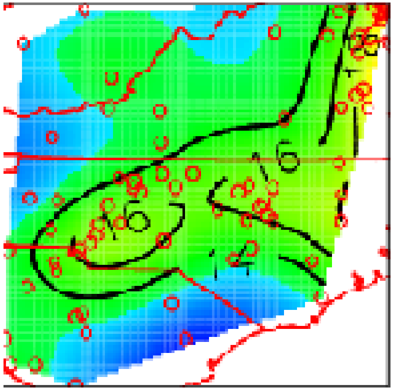
3. Obtain residuals between the prediction and observation.

4. Choose λ that minimizes the sum of squares of these residuals.

Inference for λ: Try something different.

Fit a thin plate spline to the MLE (of σ) from the univariate fitting, and determine λ by cross-validation:

1. For fixed λ, fit TPS to all but one location (do this for each location).
2. Predict the value at the omitted location.
3. Obtain residuals between the prediction and observation.
4. Choose λ that minimizes the sum of squares of these residuals.

Resulting surface looks like 

Space-Time Approach: Daily Model

Let $Y(\mathbf{x}, t)$ denote the daily 8-hr max Ozone for m sites over n time points. Consider,

$$Y(\mathbf{x}, t) = \mu(\mathbf{x}, t) + \sigma(\mathbf{x})u(\mathbf{x}, t),$$

where $u(\mathbf{x}, t)$ is a de-seasonalized zero mean, unit variance space-time process, *i.e.*

$$u(\mathbf{x}, t) = \rho(\mathbf{x})u(\mathbf{x}, t-1) + \varepsilon(\mathbf{x}, t),$$

where $|\rho(\mathbf{x})| < 1$, the spatial shocks, $\varepsilon(\mathbf{x}, t)$, are independent over time, but spatially correlated with covariance function

$$\text{Cov}(\varepsilon(\mathbf{x}, t), \varepsilon(\mathbf{x}', t)) = \sqrt{1 - \rho^2(\mathbf{x})}\sqrt{1 - \rho^2(\mathbf{x}')}\psi(d(\mathbf{x}, \mathbf{x}'))$$

Note that $\mu(\cdot, \cdot)$, $\sigma(\cdot)$ and $\rho(\cdot)$ are spatial fields. [back]

Space-Time Approach: Daily Model

Algorithm to predict FHDA at unobserved location, $\mathbf{x}_0$.

1. Simulate data for an entire Ozone season
   (a) Interpolate spatially from $u(\mathbf{x}, 1)$ to get $\hat{u}(\mathbf{x}_0, 1)$.
   (b) Also interpolate spatially to get $\hat{\rho}(\mathbf{x}_0)$, $\hat{\mu}(\mathbf{x}_0, \cdot)$ and $\hat{\sigma}(\mathbf{x}_0)$.
   (c) Sample shocks at time t from $[\varepsilon(\mathbf{x}_0, t) | \varepsilon(\mathbf{x}, t)]$.
   (d) Propagate AR(1) model.
   (e) Back transform $\hat{Y}(\mathbf{x}_0, t) = \hat{u}(\mathbf{x}_0, t)\hat{\sigma}(\mathbf{x}_0) + \hat{\mu}(\mathbf{x}_0, t)$
2. Take fourth-highest value from Step 1.
3. Repeat Steps 1 and 2 many times to get a sample of FHDA at unobserved location.

[back]

Space-Time Approach: Daily Model

Distribution for the AR(1) shocks

$[\varepsilon(\mathbf{x}_0, t) | \varepsilon(\mathbf{x}, t)]$ (Step 1c) given by

$$\text{Gau}(\mathbf{M}, \mathbf{\Sigma})$$

with

$$\mathbf{M} = \mathbf{k}'(\mathbf{x}_0, \mathbf{x})\mathbf{k}^{-1}(\mathbf{x}, \mathbf{x})\varepsilon(\mathbf{x}, t)$$

and

$$\mathbf{\Sigma} = \mathbf{k}'(\mathbf{x}_0, \mathbf{x}_0) - \mathbf{k}'(\mathbf{x}_0, \mathbf{x})\mathbf{k}^{-1}(\mathbf{x}, \mathbf{x})\mathbf{k}(\mathbf{x}, \mathbf{x}_0),$$

where $k(\mathbf{x}, \mathbf{y})$ represents the covariance between two spatial locations.

[back]

Geostatistical Approach: Seasonal Model

Covariance

Estimate a covariance function for the FHDA field, and use it to predict an unobserved location.

$$\widehat{Y}(\mathbf{x}_0) = \mathbf{k}'(\mathbf{x}_0, \mathbf{x})\mathbf{k}^{-1}(\mathbf{x}, \mathbf{x})\mathbf{Y}$$

where $\mathbf{Y}$ is the observed FHDA, $\mathbf{k}(\mathbf{x}, \mathbf{y})$ is the covariance between two locations $\mathbf{x}$ and $\mathbf{y}$. This has variance,

$$k(\mathbf{x}_0, \mathbf{x}_0) - \mathbf{k}'(\mathbf{x}_0, \mathbf{x})\mathbf{k}^{-1}(\mathbf{x}, \mathbf{x})\mathbf{k}(\mathbf{x}, \mathbf{x}_0)$$

Geostatistical Approach: Seasonal Model

Covariance

Two types of covariance: ψ_v and ψ_m. back