An Overview of Environmental Statistics

Richard L. Smith
Department of Statistics and Operations Research
University of North Carolina
Chapel Hill, N.C., U.S.A.

Theme Day in Environmental Statistics
55th Session of ISI
Sydney, April 6, 2005

http://www.stat.unc.edu/postscript/rs/isitutorial.pdf

Environmental Statistics is by now an extremely broad field, involving application of just about every technique of statistics.

Examples:

- Pollution in atmosphere and water systems
- Effects of pollution on human health and ecosystems
- Uncertainties in forecasting climate and weather
- Dynamics of ecological time series
- Environmental effects on the genome

and many more.

Some common themes:

- Most problems involve *time series* of observations, but also *spatial sampling*, often involving irregular grids

- *Design* of a spatial sampling scheme or a monitor network is important

- Often the greatest interest is in *extremes*, for example
  - Air pollution standards often defined by number of crossings of a high threshold
  - Concern over impacts of climate change often focussed on climate extremes. Hence, must be able to characterize likely frequencies of extreme events in future climate scenarios

- Use of *numerical models* — not viewed in competition, but how we can use statistics to improve the information derived from models

I have chosen here to focus on three topics that have applications across several of these areas:

I. Spatial and spatio-temporal statistics

- Interpolation of an air pollution or meteorological field
- Comparing data measured on different spatial scales
- Assessing time trends in data collected on a spatial network

II. Network design

- Choosing where to place the monitors to satisfy some optimality criterion related to prediction or estimation

III. Extreme values

- Probabilities of extreme events
- Time trends in frequencies of extreme events
- Assessing extremes on different spatial scales

TOPIC I: SPATIAL AND SPATIO-TEMPORAL STATISTICS

I.1. Spatial covariances

I.2. Model identification and estimation

I.3. Prediction and interpolation

I.4. Spatial-temporal models

I.5. Example: Interpolation of fine particulate matter over the U.S.

Major references

Cressie (1993)

Stein (1999)

Chilès and Delfiner (1999)

Banerjee, Carlin and Gelfand (2004)

and numerous others that have appeared over the past couple of years.

My own course notes (Smith 2001):

http://www.stat.unc.edu/postscript/rs/envstat/env.html

Software

S-PLUS Spatial Statistics module

SAS — PROC MIXED (for ML or REML estimation of variogram models) plus several more specialized spatial statistics procedures

R in combination with the geoR and geoRglm libraries (http://www.est.ufpr.br/geoR)

The “Fields” package from NCAR (http://www.cgd.ucar.edu/stats/Software/Fields)

My own programs and data sets (http://www.stat.unc.edu/postscript/rs/envstat/env2.html)

I.1. Spatial covariances

Basic structure: A stochastic process $\{Y(s),\ s \in D\}$, $D \subseteq \mathcal{R}^d$, usually though not necessarily $d = 2$.

Mean function

$$\mu(s) = \mathsf{E}\{Y(s)\}, \quad s \in D.$$

Covariance function

$$C(s_1, s_2) = \mathsf{Cov}\{Y(s_1), Y(s_2)\}.$$

Y is *Gaussian* if all joint distributions are multivariate normal.

Y is *second-order stationary* if $\mu(s) \equiv \mu$ and

$$\mathsf{Cov}\{Y(s_1), Y(s_2)\} = C(s_1 - s_2),$$

for all $s_1 \in D,\ s_2 \in D$, where $C(s)$ is $\mathsf{Cov}\{Y(s), Y(0)\}$.

The Variogram. Assume $\mu(s)$ is a constant, which we may without loss of generality take to be 0, and then define

$$\text{Var}\{Y(s_1) - Y(s_2)\} = 2\gamma(s_1 - s_2).$$

This makes sense only if the left hand side depends on s_1 and s_2 only through their difference $s_1 - s_2$. Such a process is called *intrinsically stationary*. The function $2\gamma(\cdot)$ is called the *variogram* and $\gamma(\cdot)$ the *semivariogram*.

Intrinsic stationarity is weaker than second-order stationarity. However, if the latter holds we have

$$\gamma(h) = C(0) - C(h).$$

We shall usually assume second-order stationarity though some applications require the wider class of intrinsically stationary models (particulate matter example later).

Isotropy. Suppose the process is intrinsically stationary with semivariogram $\gamma(h)$, $h \in \mathcal{R}^d$. If $\gamma(h) = \gamma_0(||h||)$ for some function γ_0, i.e. if the semivariogram depends on its vector argument h only through its length $||h||$, then the process is *isotropic*.

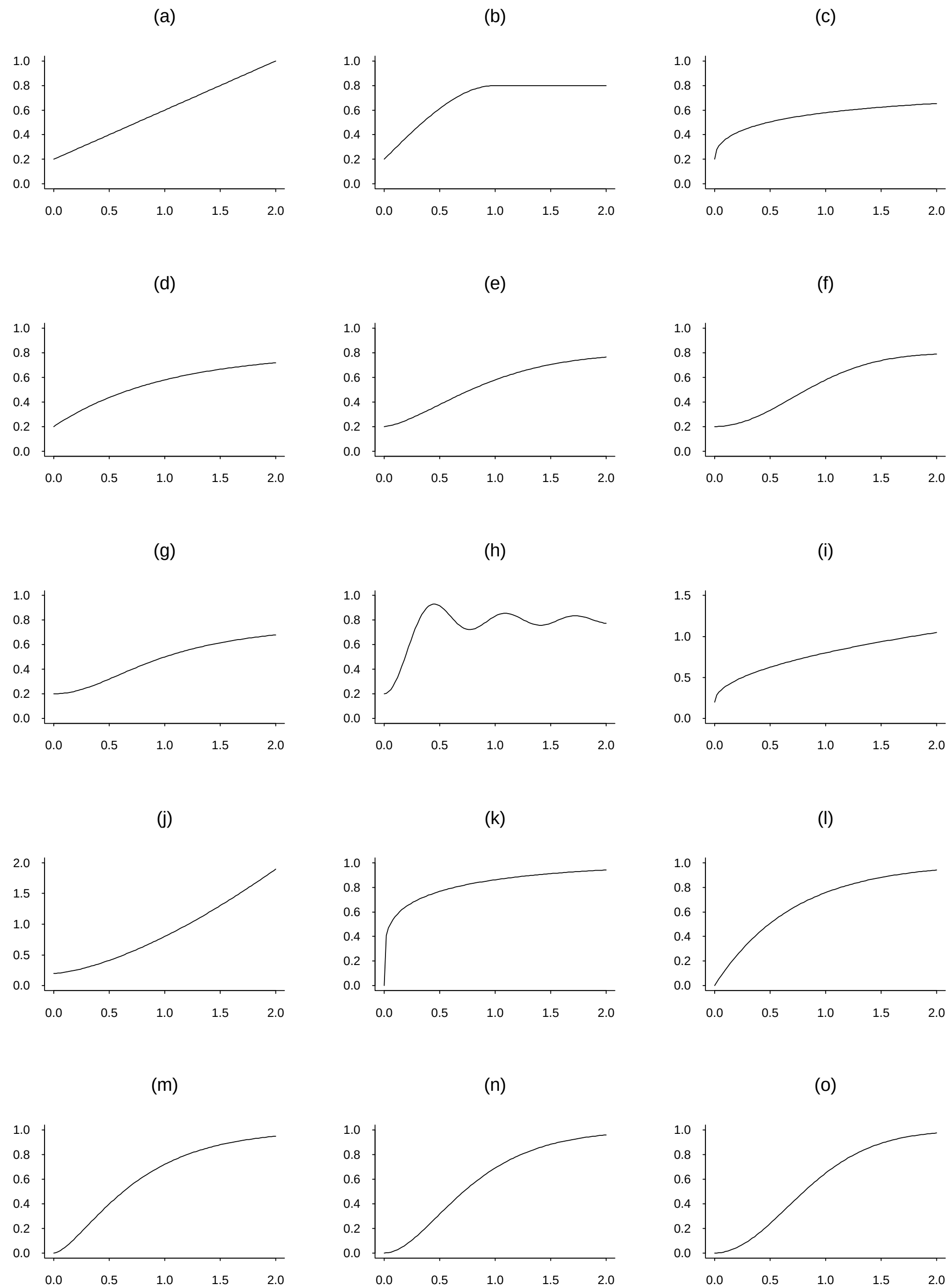
Specification of functional forms for covariances and variograms limited by *positive definiteness*: for any finite set of points $s_1, ..., s_n$ and arbitrary real coefficients $a_1, ..., a_n$ we must have

$$\sum_i \sum_j a_i a_j C(s_i, s_j) \geq 0.$$

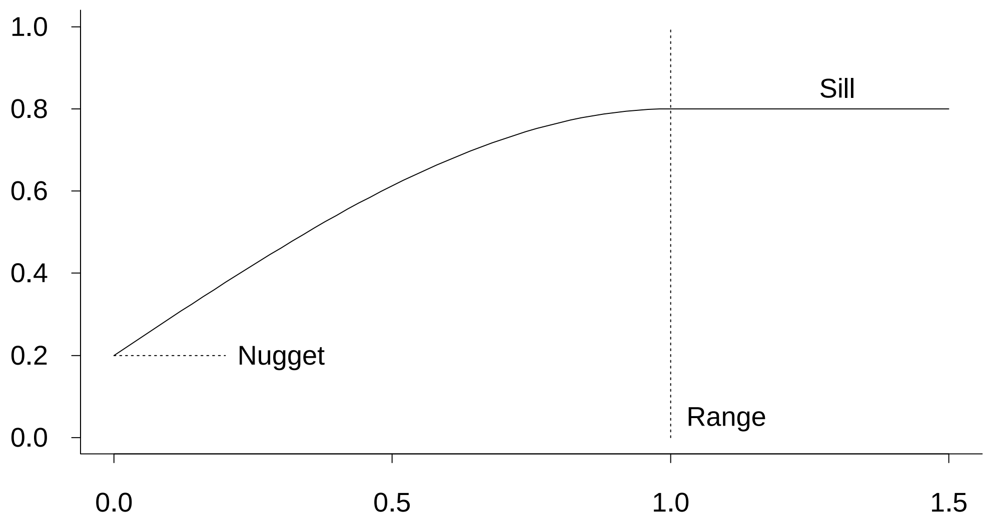
Corresponding conditions for variogram: if $\sum a_i = 0$,

$$\sum_i \sum_j a_i a_j \gamma(s_i - s_j) \leq 0.$$

More complete characterizations follow through spectral representations (Stein, Fuentes and others)



Some examples of variogram functions



The typical “nugget–sill-range” shape of a (stationary) variogram

We give some examples of specific functional forms for a stationary isotropic variogram γ_0 or covariance function C_0

1. *Exponential-power form:*

$$\gamma_0(t) = \begin{cases} 0 & \text{if } t = 0, \\ c_0 + c_1(1 - e^{-|t/R|^p}) & \text{if } t > 0. \end{cases}$$

Here $0 < p \leq 2$. $p = 1$ is called *exponential*, $p = 2$ is *Gaussian*.

2. *Spherical:* (for d=1,2,3,)

$$\gamma_0(t) = \begin{cases} 0 & \text{if } t = 0, \\ c_0 + c_1 \left\{ \frac{3}{2}\frac{t}{R} - \frac{1}{2}(\frac{t}{R})^3 \right\} & \text{if } 0 < t \leq R, \\ c_0 + c_1 & \text{if } t \geq R. \end{cases}$$

3. *Power law:*

$$\gamma_0(t) = \begin{cases} 0 & \text{if } t = 0, \\ c_0 + c_1 t^{\lambda} & \text{if } t > 0. \end{cases}$$

Valid if $0 \leq \lambda < 2$. $\lambda = 1$ is *linear variogram*. This case is *not* second-order stationary.

4. *Matérn:*

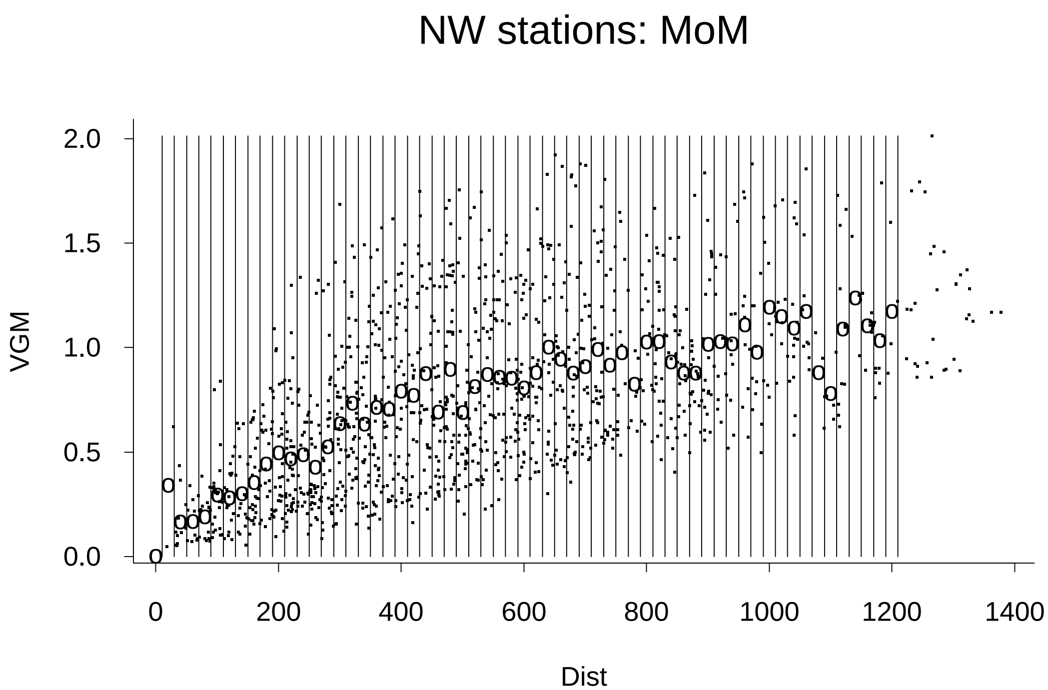
$$C_0(t) = \frac{1}{2^{\theta_2 - 1}\Gamma(\theta_2)} \left(\frac{2\sqrt{\theta_2}t}{\theta_1}\right)^{\theta_2} \mathcal{K}_{\theta_2}\left(\frac{2\sqrt{\theta_2}t}{\theta_1}\right).$$

$\theta_1 > 0$ is the spatial scale parameter and $\theta_2 > 0$ is a shape parameter. $\Gamma(\cdot)$ is the usual gamma function while $\mathcal{K}_{\theta_2}$ is the modified Bessel function of the third kind of order θ_2. $\theta_2 = \frac{1}{2}$ corresponds to the exponential form of semivariogram, and the limit $\theta_2 \to \infty$ results in the Gaussian form.

I.2. Model identification and estimation

Assume a process $\{Y(s),\ s \in D\}$ observed at a finite number of points $s_1, ..., s_N$.

The *sample variogram* is often used as an initial guide to the form of spatial model. It can be drawn as either a *variogram cloud*, or a *binned variogram*.



Binned variogram versus "variogram cloud" for temperature stations in the northwest US

Fitting parametric models

Sample variogram not negative definite: therefore, not acceptable as an estimate of population variogram

Solution: fit a parametric model

- Curve fitting to the variogram,
- Maximum likelihood (ML),
- Restricted maximum likelihood (REML),
- Bayesian estimators.

Maximum likelihood estimation (Mardia and Marshall 1984)

Assume Gaussian process. General model (includes regression terms):

$$\begin{aligned} Y &\sim \mathcal{N}(X\beta, \Sigma), \\ \Sigma &= \alpha V(\theta), \end{aligned}$$

X a $n \times q$ matrix of covariates, α a scale parameter and $V(\theta)$ determined by θ, parameters of spatial model.

Maximum likelihood estimation reduces to minimizing the neg. log. profile likelihood

$$\ell^*(\theta) = \text{const } + \frac{n}{2} \log \frac{G^2(\theta)}{n} + \frac{1}{2} \log |V(\theta)|.$$

where $G^2(\theta) = (Z - X\widehat{\beta})^T V(\theta)^{-1}(Z - X\widehat{\beta})$,
$\widehat{\beta} = (X^T V(\theta)^{-1} X)^{-1} X^T V(\theta)^{-1} Y$ the GLS estimator of β.

Restricted maximum likelihood

Let $W = A^T Y$ be a vector of $n-q$ linearly independent contrasts, i.e. the $n-q$ columns of A are linearly independent and $A^T X = 0$, then we find that

$$W \sim \mathcal{N}(0, A^T \Sigma A).$$

The density of W is taken to define the neg log likelihood function. After some manipulation, this reduces to minimizing

$$\ell_W^*(\theta) = \text{const} + \frac{n-q}{2} \log \frac{G^2(\theta)}{n} + \frac{1}{2} \log |X^T V(\theta)^{-1} X| + \frac{1}{2} \log |V(\theta)|.$$

Bayesian interpretation: this is the integrated likelihood at θ assuming a uniform prior on β (Harville 1974)

Advantages of REML over MLE

1. Asymptotic theory: although MLE and REMLE are first-order equivalent, evidence suggests that REML performs better when evaluated using second-order asymptotics (Smith and Zhu 2004, preliminary work)

2. Much closer correspondence with Bayesian theory, e.g. the "reference prior" for a Bayesian approach (Berger, de Oliveira and Sansó, 2001) turns out to coincide with the Jeffreys prior derived from the restricted likelihood

3. For models which are intrinsically stationary but not second-order stationary, REML estimation works almost without modification

I.3. Prediction and interpolation

Suppose we have the same universal kriging model as before but extended to include some variable y_0 that we want to predict:

$$\begin{pmatrix} Y \\ y_0 \end{pmatrix} \sim \mathcal{N}\left[\begin{pmatrix} X\beta \\ x_0^T\beta \end{pmatrix}, \begin{pmatrix} \Sigma & \tau \\ \tau^T & \sigma_0^2 \end{pmatrix}\right]$$

where x_0 are new covariates corresponding to y_0, σ^2 is the variance of y_0 and τ is a vector of cross-covariances.

Note that y_0 is not restricted to being a single unobserved element of the random field but could also be, for example, either a spatial or a temporal average of the random field.

Traditional specification of *best linear unbiased prediction*: find a predictor $\widehat{y}_0 = \lambda^T Y$ to minimize $E\left\{(\widehat{y}_0 - y_0)^2\right\}$ subject to $E\{\widehat{y}_0 - y_0\} = 0$.

In vector-matrix notation, the problem is:

Find λ to minimize

$$V_0 = \lambda^T \Sigma \lambda - 2\lambda^T \tau + \sigma_0^2$$

subject to

$$X^T \lambda = x_0.$$

The solution is

$$\lambda = \Sigma^{-1}\tau + \Sigma^{-1}X(X^T\Sigma^{-1}X)^{-1}(x_0 - X^T\Sigma^{-1}\tau),$$

and the corresponding MSPE is

$$V_0 = (x_0 - X^T\Sigma^{-1}\tau)^T(X^T\Sigma^{-1}X)^{-1}(x_0 - X^T\Sigma^{-1}\tau) - \tau^T\Sigma^{-1}\tau + \sigma_0^2.$$

I.4. Spatial-temporal models

The direct generalization of spatial statistics to spatial-temporal data is based on finding classes of spatial-temporal covariance functions that obey the positive definiteness property, for which the preceding theories of estimation, interpolation etc., go through directly.

We concentrate here on two specific classes, *separable models* and the *dissociated processes*. These are the simplest cases of spatio-temporal models and can be viewed as the basic building blocks from which more complicated models may be built.

The separable model is defined by

$$C(h, u) = C_0(h)\gamma(u)$$

where $C(h, u)$ denotes the covariance between two space-time coordinates with spatial separation h and temporal separation u, $C_0(h)$ is a pure spatial covariance and $\gamma(u)$ is a temporal autocovariance. Since we may always transfer a constant between the functions C_0 and γ, there is no loss of generality in assuming $\gamma(0) = 1$, in other words, that γ is a temporal autocorrelation function.

The special case where $\gamma(u) = 0$ for all $u \neq 0$ was called the *repeated measurements model* by Mardia and Goodall (1993). I prefer *dissociated processes* to avoid the confusion with traditional repeated measurements models.

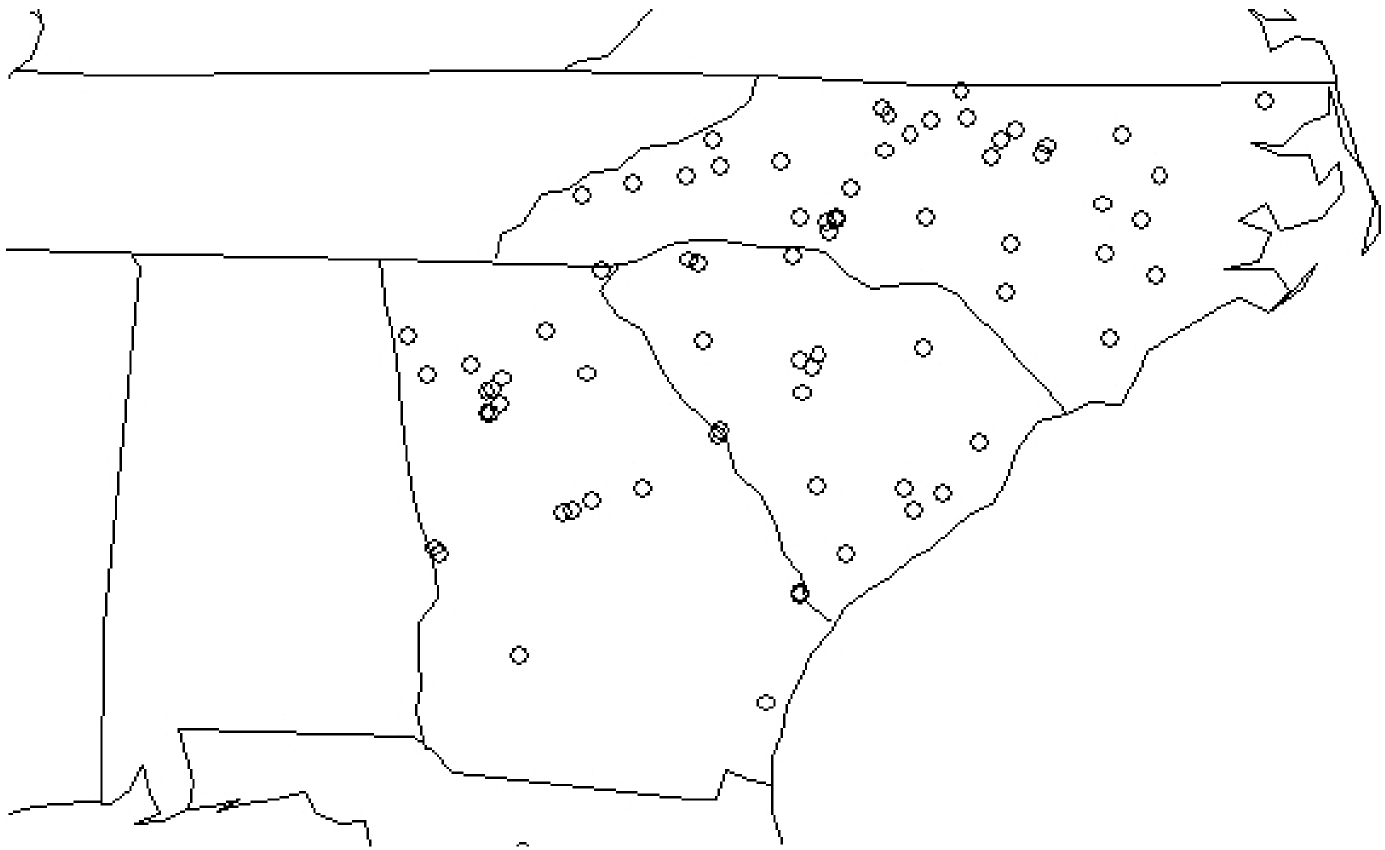
I.5. Example: Interpolation of fine particulate matter over the U.S.

Ref: Smith, Kolenikov and Cox (2003)

A new set of air pollution standards, first proposed in 1997, is finally being implemented by the U.S. Environmental Protection Agency (EPA). One of the requirements is that the mean level of fine particulate matter ($PM_{2.5}$) at any location should be no more than 15 μg/m^3. A network of several hundred monitors has been set up to assess this.

The present study is based on 1999 data for a small portion of this network, 74 monitors in North Carolina, South Carolina and Georgia. We converted the raw values to weekly averages, but even so more than $\frac{1}{4}$ of the data are missing. The EPA also recorded a "land-use" variable, classified as one of five types of land-use: agricultural (A), commercial (C), forest (F), industrial (I) and residential (R).

Map of 74 Stations



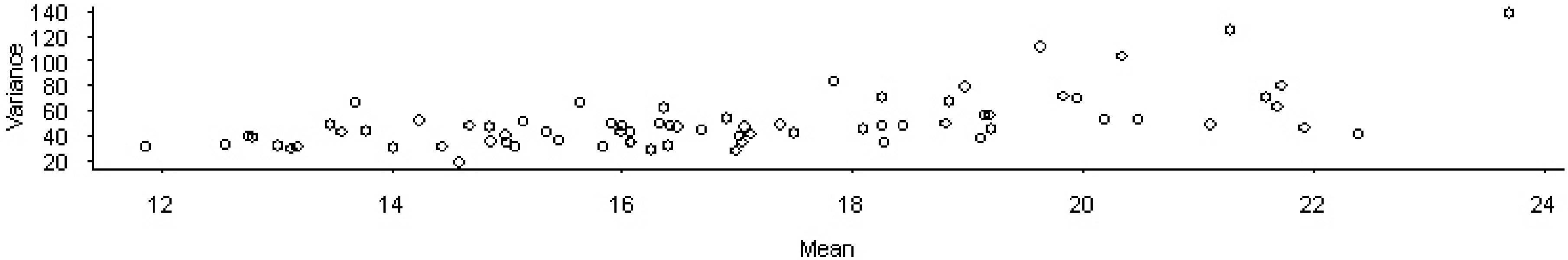
Preliminary analyses

We made a number of decisions based on plots of the data:

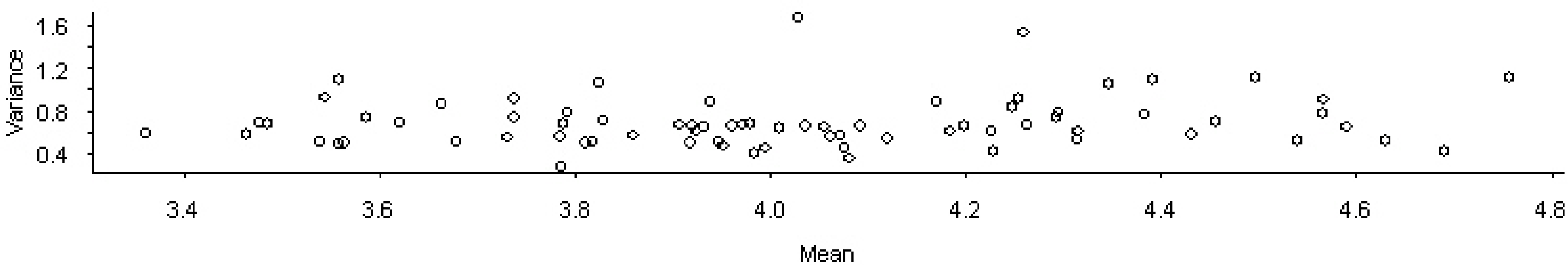
- Use square roots of $PM_{2.5}$ to stabilize variances (approximately)
- Time trend assumed to be common across all stations
- No temporal correlation once time-trend is removed from data
- There is spatial correlation — suggests a dissociated model
- The form of the spatial correlation looks like a linear or power-law variogram — different from traditional second-order stationary models

Mean-Variance Plots

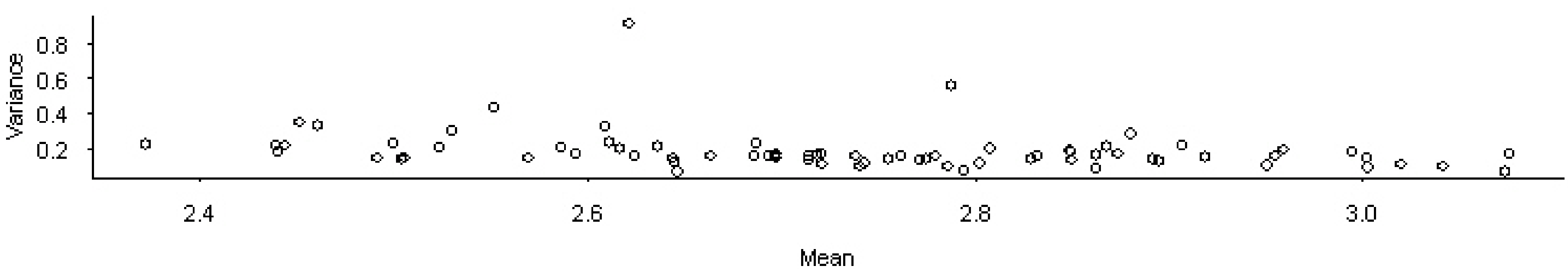
Original Data



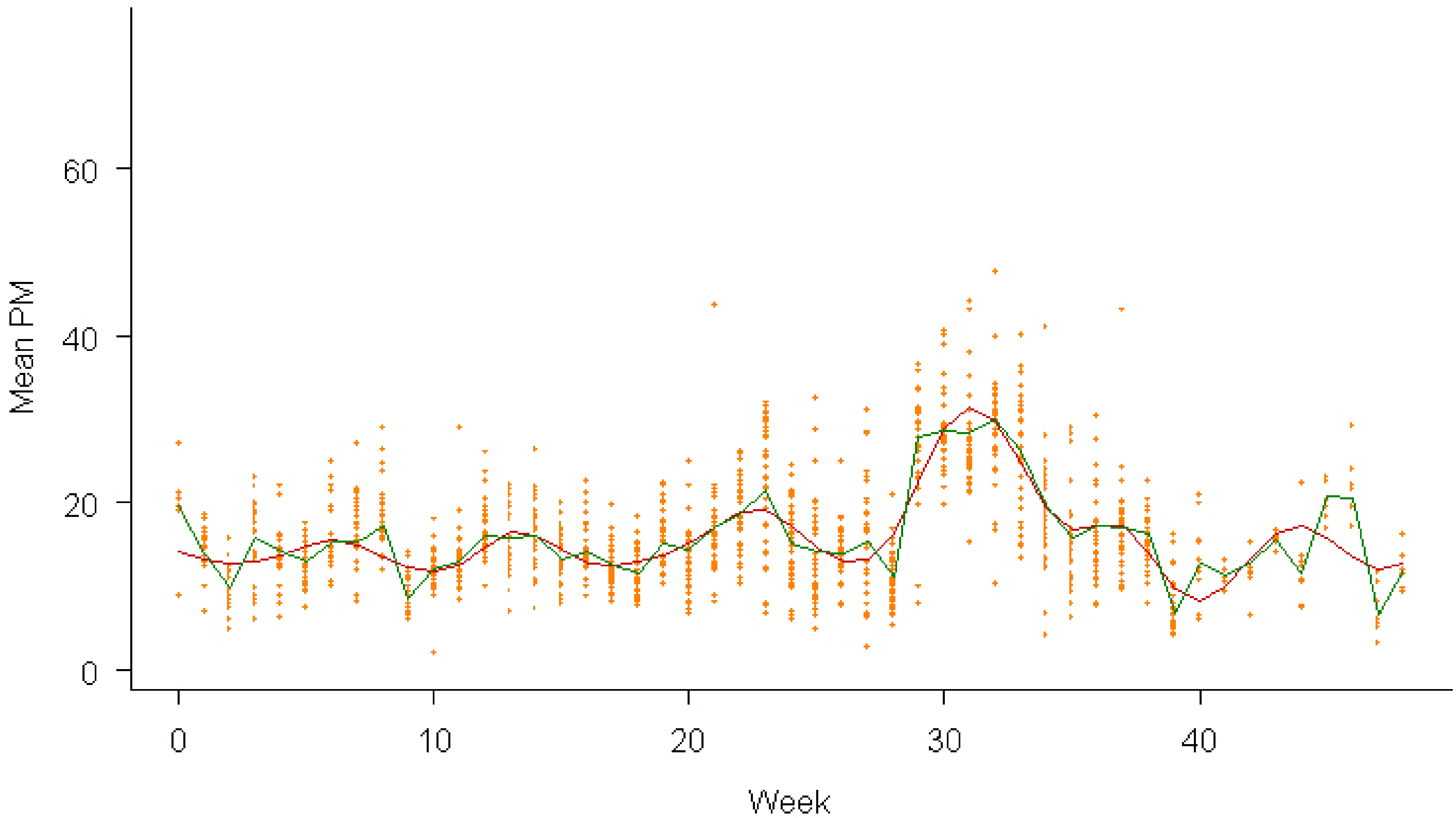
Square Root Transform



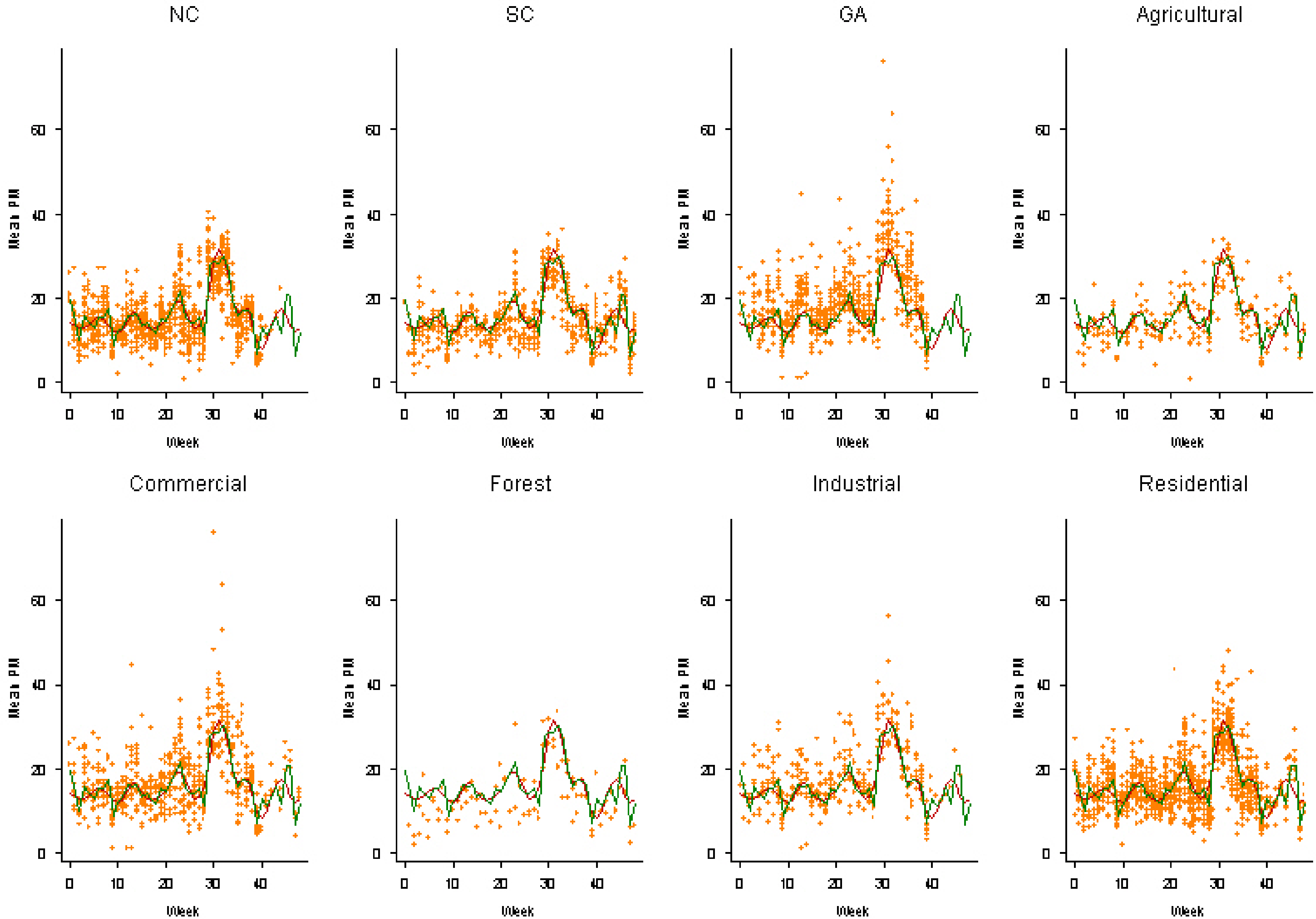
Logarithmic Transform



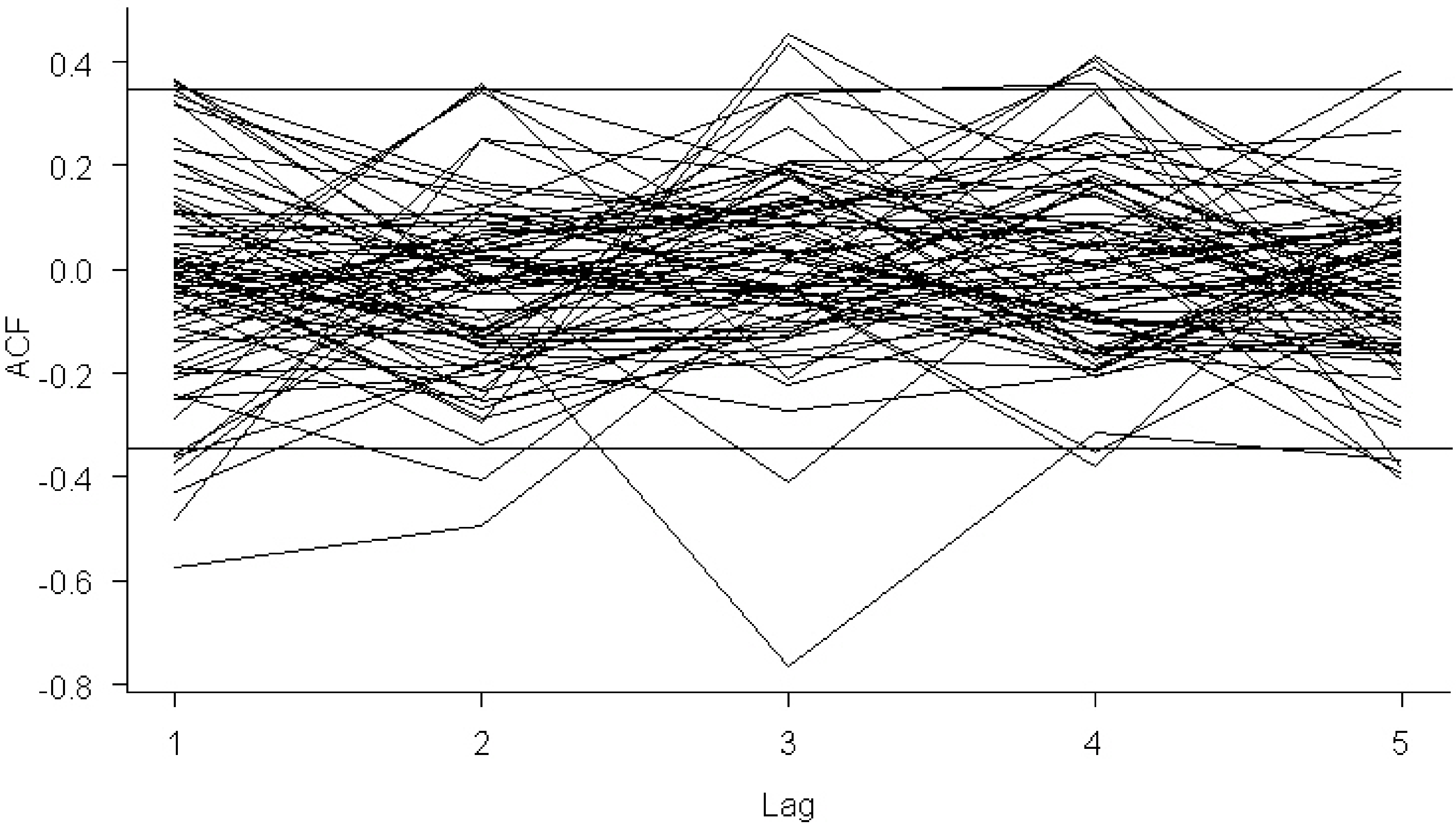
Time Trend Fits to Entire Data Set



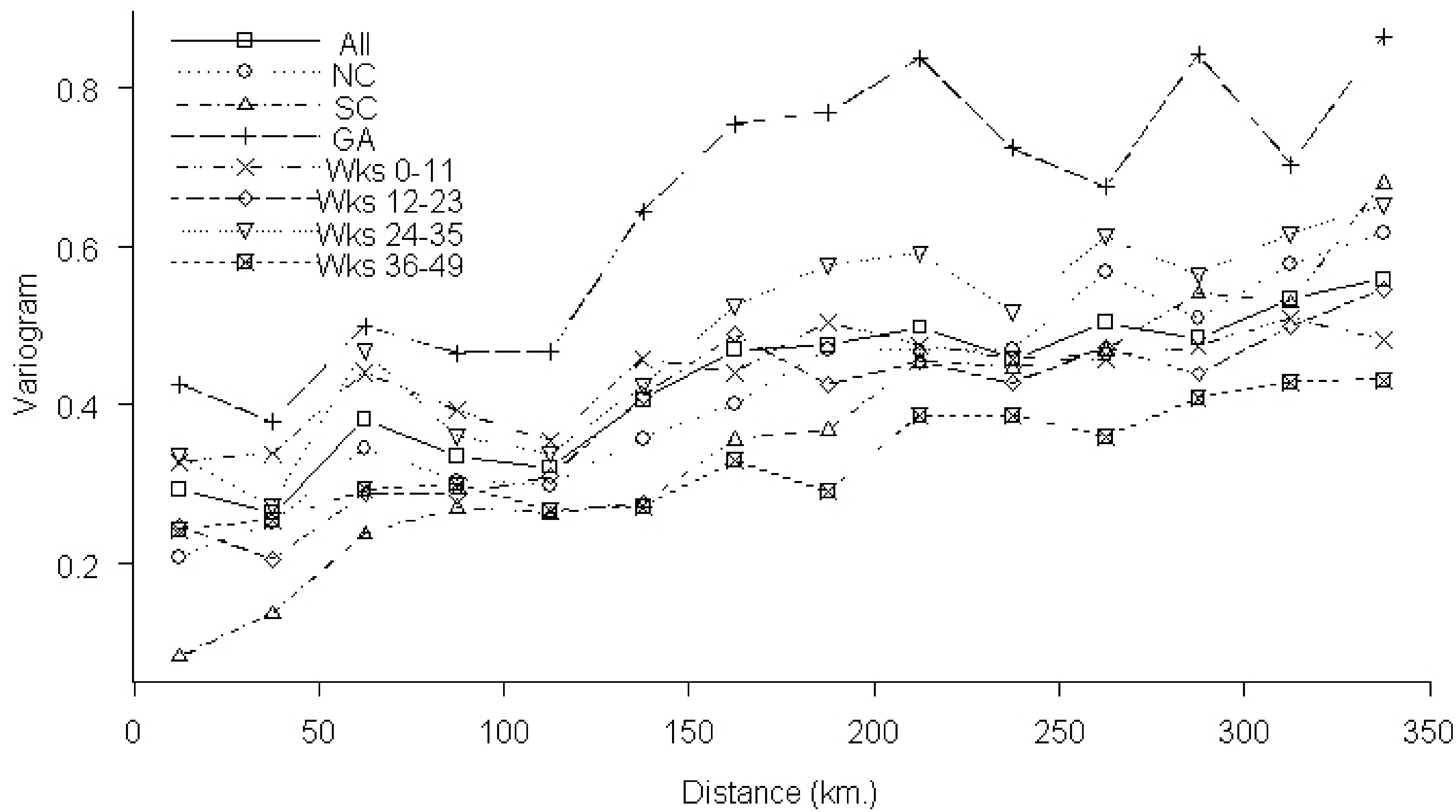
Overall Time Trends with Selected Subsets of Data



Autocorrelation Plots for 74 Stations



Variogram Plots for Selected Subsets of Data



Basic model

$$y_{xt} = w_t + \psi_x + \theta_x + \eta_{xt}$$

in which y_{xt} is the square root of $PM_{2.5}$ in location x in week t, w_t is a week effect, ψ_x is the spatial mean at location x (in practice, estimated through a thin-plate spline representation), θ_x is a land-use effect corresponding to the land-use as site x, and η_{xt} is a random error.

We fit the power law variogram to $\{\eta_{xt}\}$ for each time point t

$$\gamma(h) = \begin{cases} 0 & \text{if } h = 0, \\ \theta_0 + \theta_1 h^\lambda & \text{if } h > 0, \end{cases}$$

where $\theta_0 > 0$, $\theta_1 > 0$, $0 \leq \lambda < 2$. MLE of λ is 0.92 with standard error 0.097 (close to $\lambda = 1$, linear variogram)

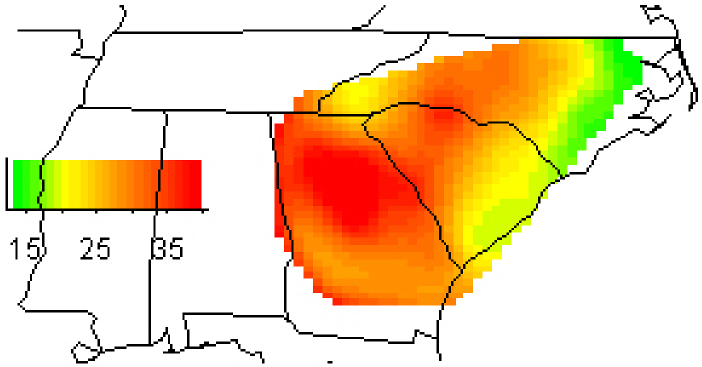
Results

The fitted model was used to construct a predicted surface, with estimated root mean squared prediction error (RMSPE), for each week of the year and also for the average over all weeks. The latter is of greatest interest in the context of EPA standards setting.

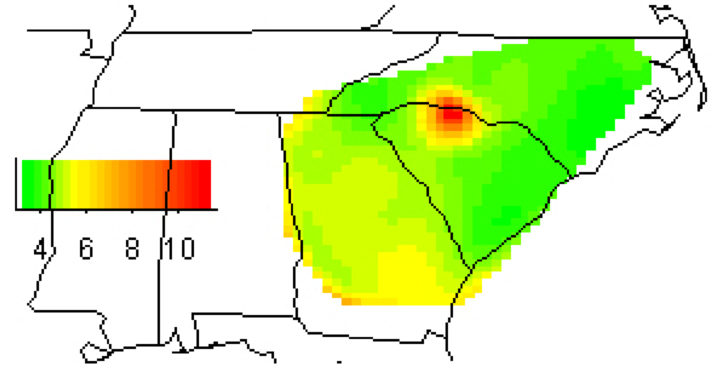
We show the predicted surface and RMSPE for week 33 (the week with highest average $PM_{2.5}$) and overall for the annual mean. We also show the estimated probability that any particular location exceeds the 15 μg/m^3 annual mean standard.

Predicted PM2.5 Surfaces and RMSPEs

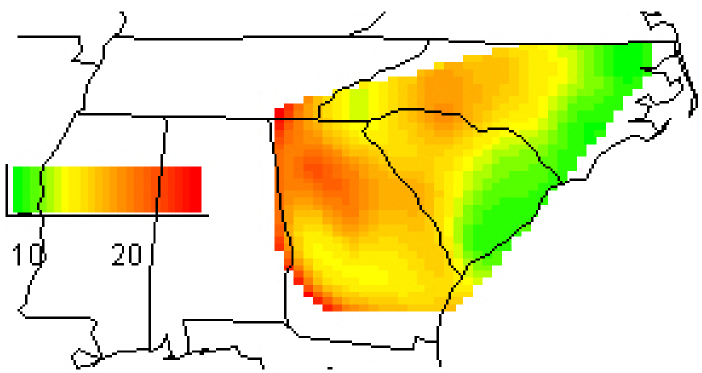
Predicted Surface for Week 33



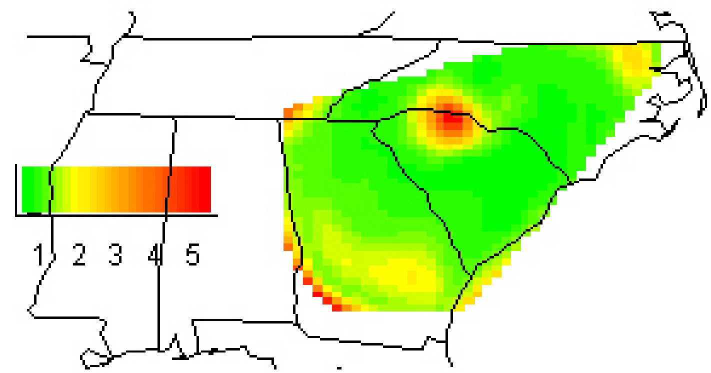
RMSPE for Week 33



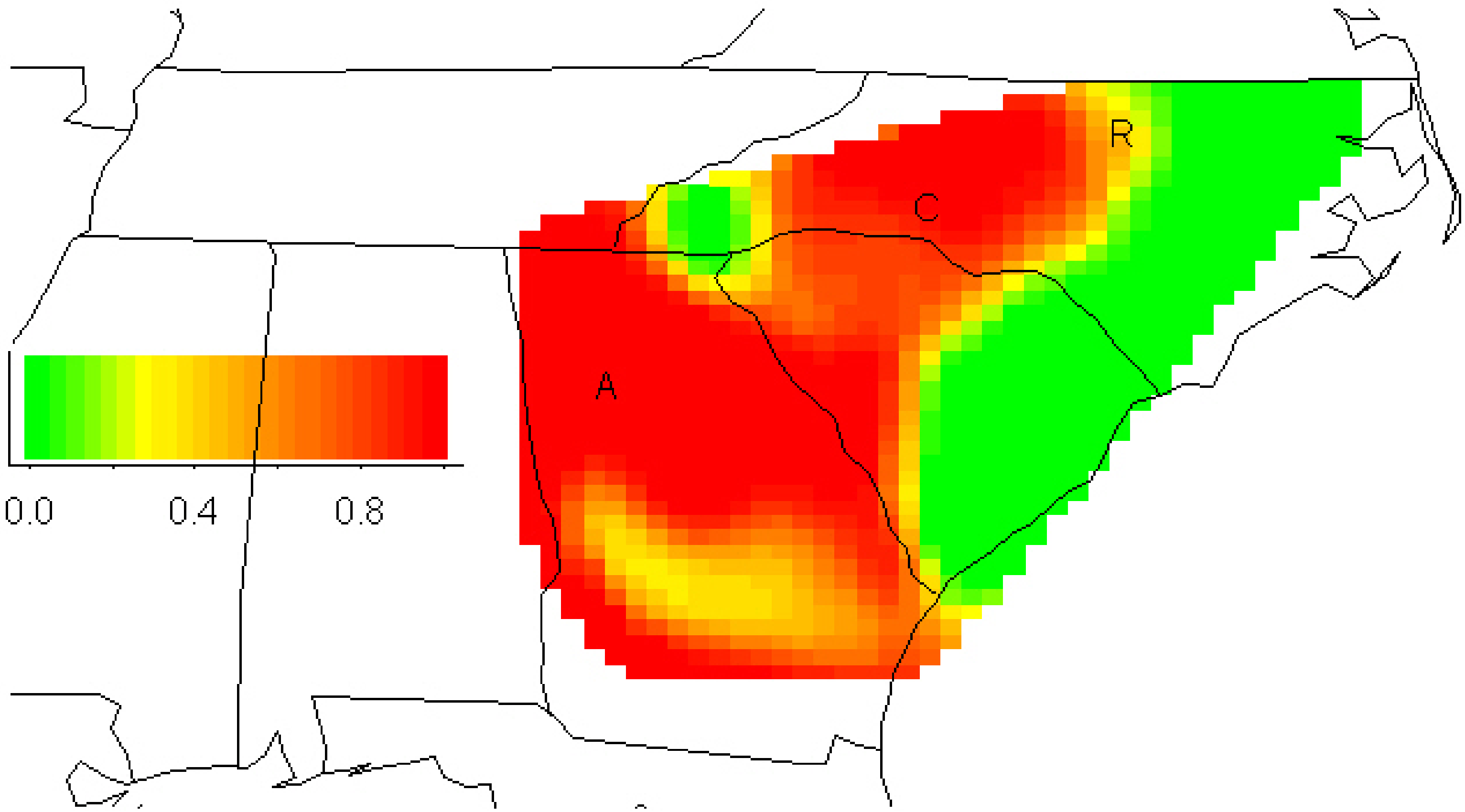
Predicted Surface for Annual Average



RMSPE for Annual Average



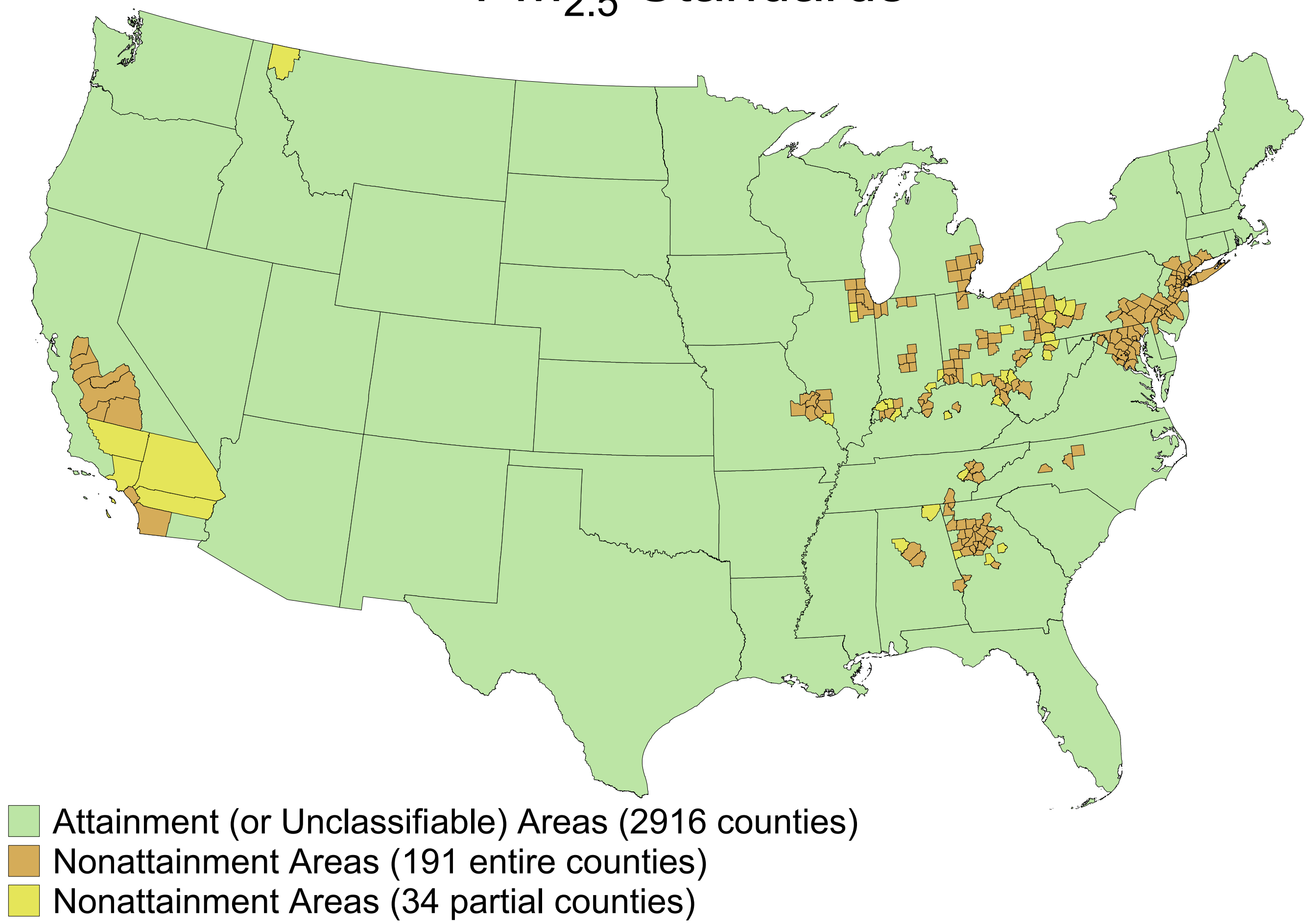
Probability of Exceeding Standard



It can be seen that substantial parts of the region, including the western portions of North and South Carolina and virtually the whole of the state of Georgia, appear to be in violation of the standard. Of the three major cities marked on the last figure, Atlanta and Charlotte are clearly in the "violation" zone; Raleigh is on the boundary of it.

The actual EPA nonattainment regions suggest a rather different picture.

Attainment and Nonattainment Areas in the U.S. $PM_{2.5}$ Standards



Current nonattainment areas (Source: EPA website, 12/18/2004).

Postscript on spatial and spatio-temporal statistics

There are, of course, many more kinds of models than the ones I have presented here. I could have included whole chapters about any of the following:

- Nonstationary models
- Lattice models and their use in modern Bayesian methods
- Spatial-temporal models other than dissociated and separable models
- Spatial GLMs (fitted by geoRglm module in R)

TOPIC II: NETWORK DESIGN

II.1. Overview of different approaches

II.2. Predictive and estimative criteria

II.3. A new combined predictive-estimative approach

II.4. Example

II.1. Overview of different approaches

The problem: We would like to monitor some environmental variable over some region of interest.

Examples include both air and water pollution, meteorological observing stations, fish and wildlife surveys, and many others.

The problem is where to place the monitors.

Traditional criteria for design of experiments, or for the design of sample surveys, do not allow for spatial correlation among the design points.

Some methodological approaches to design of a network with spatial correlation:

- Approaches based on information theory and entropy (Zidek and co-authors)

- Approaches based on the theory of optimal design (e.g. W. Müller (2000))

- Space-filling designs, e.g. Nychka and Saltzman (1998)

- Bayesian approaches, e.g. by P. Müller (1999)

Here I outline an approach (due to Stein and Zhu) that draws explicit contrast between *design for prediction* and *design for estimation*, and some recent work on a unified approach.

II.2. Predictive and estimative criteria

Initial discussion follows Zhu (2002), Zhu and Stein (2004a,2004b)

Recall earlier universal kriging model where we indicate explicitly that the covariance model depends on unknown parameters θ:

$$\begin{pmatrix} Y \\ y_0 \end{pmatrix} \sim \mathcal{N}\left[\begin{pmatrix} X\beta \\ x_0^T\beta \end{pmatrix}, \begin{pmatrix} \Sigma(\theta) & \tau(\theta) \\ \tau^T(\theta) & \sigma_0^2(\theta) \end{pmatrix}\right]$$

Universal kriging (assuming θ known) leads to the following expression for the mean square prediction error (MSPE) of y_0:

$$V_0 = (x_0 - X^T\Sigma^{-1}\tau)^T(X^T\Sigma^{-1}X)^{-1}(x_0 - X^T\Sigma^{-1}\tau) - \tau^T\Sigma^{-1}\tau + \sigma_0^2.$$

Of course, this depends on θ.

Predictive approaches:

For some y_0 of interest and known θ, choose the design to minimize V_0.

In practice, a family of y_0's and θ is unknown, but resolve the latter issue either through a weighted minimax approach, or averaging with respect to some prior distribution for θ.

Estimative approaches:

Choose the design to optimize the estimation of θ, via some criterion like the determinant of the Fisher information matrix.

These are contrasting criteria, e.g. the predictive approach favors space-filling designs while the estimative approach often leads to designs with clusters of neighboring points.

Combined approaches (Zhu and Stein)

Harville and Jeske (1992) and Zimmerman and Cressie (1992) proposed the following correction to the mean squared prediction error:

$$V_1 \;=\; \mathsf{E}\left\{(y_0 - \hat{\lambda}^T Y)^2\right\} \;\approx\; V_0 + \mathrm{tr}\left\{\mathcal{I}^{-1}\left(\frac{\partial\lambda}{\partial\theta}\right)^T \Sigma \left(\frac{\partial\lambda}{\partial\theta}\right)\right\}$$

where $\mathcal{I}$ is the observed information matrix for θ. This formula corrects for the error in specifying the kriging weights λ.

So one possibility is to use V_1 (rather than V_0) as a design criterion. However, this still doesn't allow for error in estimating the MSPE (important for prediction intervals).

The error in estimating V_0 depends on the quantity

$$V_2 \;=\; \left(\frac{\partial V_0}{\partial \theta}\right)^T \mathcal{I}^{-1} \left(\frac{\partial V_0}{\partial \theta}\right).$$

This suggest that some linear combination of V_1 and $\frac{V_2}{V_0}$ would best measure the overall uncertainty. Zhu and Stein suggested

$$V_3 \;=\; V_1 + \frac{1}{2} \cdot \frac{V_2}{V_0}$$

as a suitable combined criterion. However, it's not clear exactly why this particular linear combination is appropriate.

II.3. A new combined predictive-estimative approach

Assume the objective is to construct a two-sided $100(1-\alpha)\%$ prediction interval for y_0, conditional on Y, with θ and β unknown.

Among all possible designs, we select the one that leads to the smallest expected length of prediction interval, subject to the constraint that the coverage probability be $1-\alpha$.

Via second-order asymptotics, Smith and Zhu (preprint, 2004) show that such a prediction interval can be calculated from Bayesian principles, and the expected length criterion leads to

$$V_4 \;=\; V_1 + \frac{1}{4}\left\{\Phi^{-1}\left(1-\frac{\alpha}{2}\right)\right\}^2 \frac{V_2}{V_0}$$

This has the unusual feature that the design might depend on the desired coverage probability of a prediction interval.

II.4. Example

Suppose we are considering redistributing 38 $PM_{2.5}$ monitors in North Carolina.

Assume the objective is to estimate population-weighted daily average. Daily data from 2000. Assume individual days' data are independent replications of the model

$$\text{Cov}(y_i, y_j) = \begin{cases} \theta_1^2 & \text{if } i = j, \\ \theta_3 \theta_1^2 e^{-d_{ij}/\theta_2} & \text{if } i \neq j, \end{cases}$$

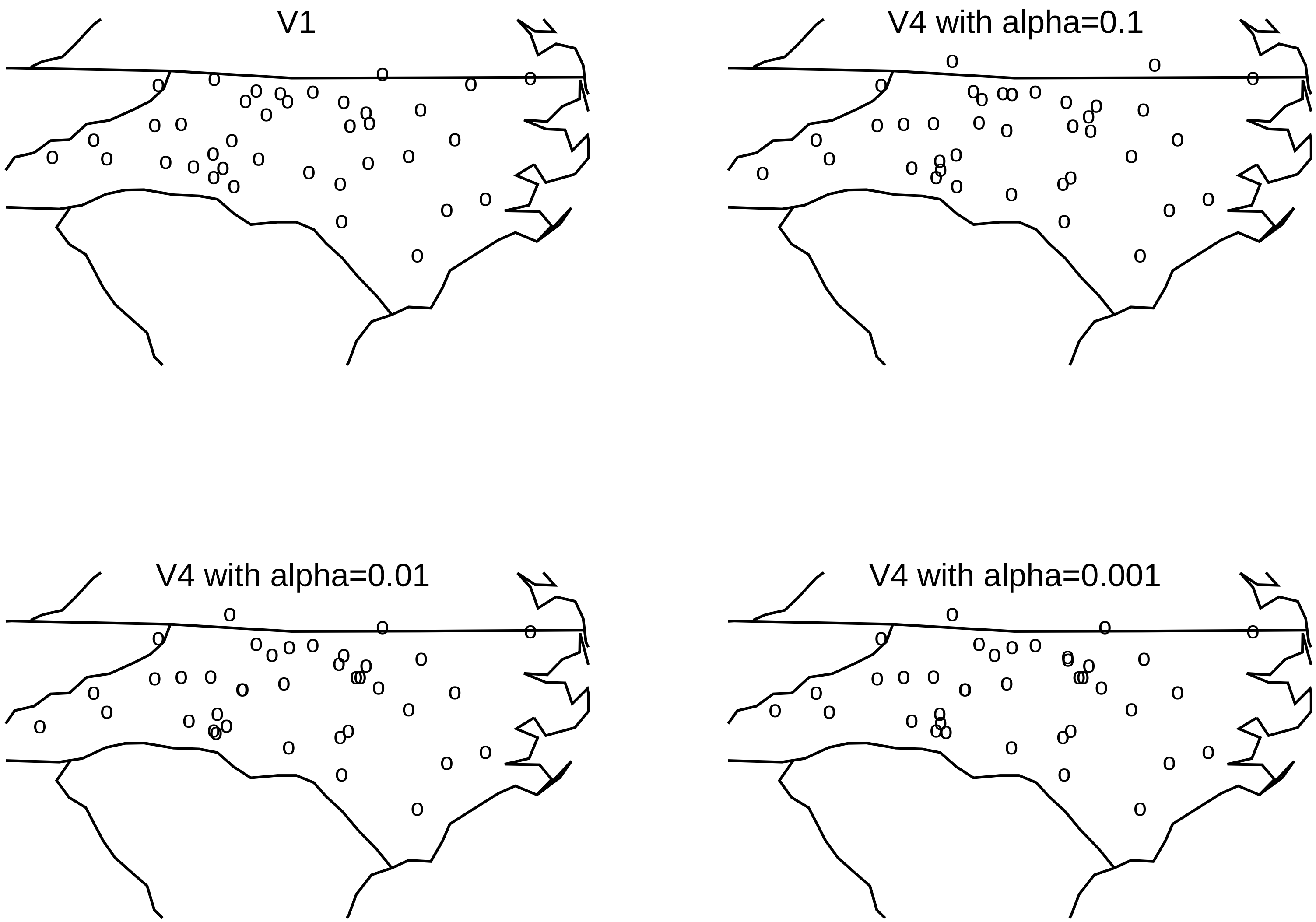
with y_i, y_j the $PM_{2.5}$ at locations i and j, d_{ij} is distance (units of 100 km.), and we estimated $\theta_1 = 6.495$, $\theta_2 = 4.019$, $\theta_3 = .9423$. Treat this as the true model, but assume $\theta_1, \theta_2, \theta_3$ would have to be re-estimated on any given day.

Population-weighted averages were calculated using data from the 2000 U.S. census for the 809 zip code tabulation areas (ZCTA) in North Carolina. Select 38 ZCTA out of 809 to place the monitoring station to give most accurate prediction of the total population PM2.5 exposure defined as

$$y_0 = \sum_i p_i y_i,$$

where p_i is the population at the i'th ZCTA, and y_i is the PM2.5 level there. V_1 and V_4 with two-sided tail probabilities $\alpha = 0.1, 0.01, 0.001$ are used as design criteria, and a simulated annealing algorithm is used to find the designs given in the following figure:

Optimal Designs Under Four Criteria



Four designs selected using simulated annealing and the V_4 criterion (calculations due to Zhengyuan Zhu)

All four designs tend to place monitors in regions of high population density (as does the current EPA network) but it is noticeable that the criterion V_4, especially for smaller α, tends to favor a network with clusters of nearby monitors, reflecting the role such clusters play in ensuring good estimation of model parameters.

TOPIC III: EXTREME VALUES

III.1. Introduction and motivation

III.2. Basics of extreme value theory

III.3. Application: Insurance data

III.4. Trends in U.S. rainfall extremes

References: Coles (2001), Smith (2003)

III.1. Introduction and motivation

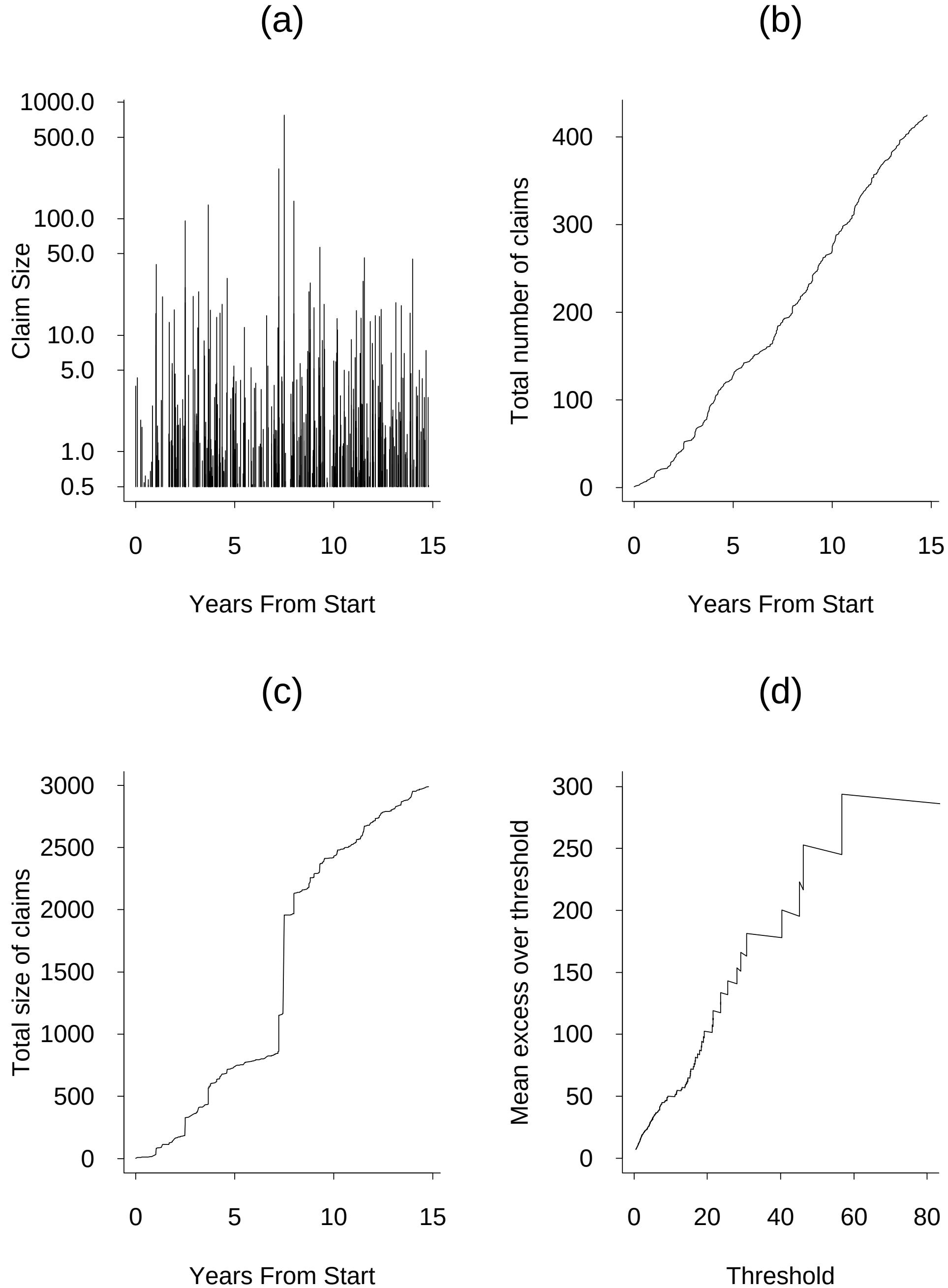
From a paper by Smith and Goodman (2000):

We consider a dataset consisting of all insurance claims experienced by a large international oil company over a threshold 0.5 during a 15-year period — a total of 393 claims.

Seven different "claim types"

Total of all 393 claims: 2989.6

10 largest claims: 776.2, 268.0, 142.0, 131.0, 95.8, 56.8, 46.2, 45.2, 40.4, 30.7.



(a) Plot of raw data. (b) Cumulative number of claims *vs.* time. (c) Cumulative claim amount *vs.* time. (d) Mean excess plot.

Questions of interest

- Estimate probabilities of extreme claims

- Are the extremes associated with particular types of claims?

- Is there any evidence of a time trend?

- If so, are the trends in any way associated with climate change? (Almost certainly not for this particular dataset, but in many insurance-related questions, this is asked and potentially of great interest.)

III.2. Basics of extreme value theory

We start with the *extreme value limit laws* (Fisher and Tippett 1928; Gnedenko 1943)

Let $X_1, X_2, ...,$ be independent identically distributed (IID) random variables with distribution function F.

Let $M_n = \max(X_1, ..., X_n)$. Then

$$\Pr\{M_n \le x\} = F^n(x) \to 0$$

for any x such that $F(x) < 1$.

To obtain interesting results *renormalize*: Find $a_n > 0$, b_n,

$$\begin{aligned} \Pr\left\{\frac{M_n - b_n}{a_n} \le x\right\} &= F^n(a_n x + b_n) \\ &\to G(x) \end{aligned}$$

where G is a nondegenerate limiting distribution function.

The Three Extreme Value Types

Type I (Gumbel)

$$\Lambda(x) = \exp(-e^{-x}), \ -\infty < x < \infty.$$

Type II (Fréchet)

$$\Phi_\alpha(x) = \begin{cases} 0, & \text{if } x \le 0, \\ \exp(-x^{-\alpha}), & \text{if } x \ge 0 \ (\alpha > 0). \end{cases}$$

Type III (Weibull)

$$\Psi_\alpha(x) = \begin{cases} \exp\{-(-x)^\alpha\}, & \text{if } x \le 0 \ (\alpha > 0), \\ 1, & \text{if } x \ge 0. \end{cases}$$

Generalized EV Distribution:

$$G(x) = \exp\left[-\left\{ 1 + \xi \frac{x-\mu}{\psi} \right\}_+^{-1/\xi} \right]$$

($x_+ = \max(x, 0)$) where $-\infty < \mu < \infty$, $0 < \psi < \infty$, $-\infty < \xi < \infty$.
The limit $\xi \to 0$ corresponds to the Gumbel case.

Exceedances Over Thresholds

Exceedances over a high threshold u.

$$\begin{aligned} F_u(y) &= \Pr\{X \le u + y \mid X > u\} \\ &= \frac{F(u+y) - F(u)}{1 - F(u)}. \quad (y > 0) \end{aligned}$$

Look for scaling constants $\{c_u\}$ so that as $u \uparrow \omega_F = \sup\{x : F(x) < 1\}$,

$$F_u(zc_u) \to H(z)$$

where H is nondegenerate. In that case, H must be of form

$$H(z) = \begin{cases} 1 - \left(1 + \frac{\xi z}{\sigma}\right)_+^{-1/\xi}, & \text{if } \xi \neq 0, \\ 1 - e^{-z/\sigma}, & \text{if } \xi = 0, \end{cases}$$

where $\sigma > 0$ and $-\infty < \xi < \infty$.

This is the *Generalized Pareto Distribution* (Pickands 1975).

Statistical Approaches

Peaks Over Thresholds

Basic idea: fix a high threshold u say, and fit the Generalized Pareto distribution (GPD) to exceedances over the threshold.

May need separate analysis to model the probability of crossing the threshold as a function of covariates, e.g. logistic regression.

Extensions of the basic methodology:

- Selecting the threshold
- Incorporating covariates
- Dependence in the time series

Statistical Approaches, Continued

Point process approach

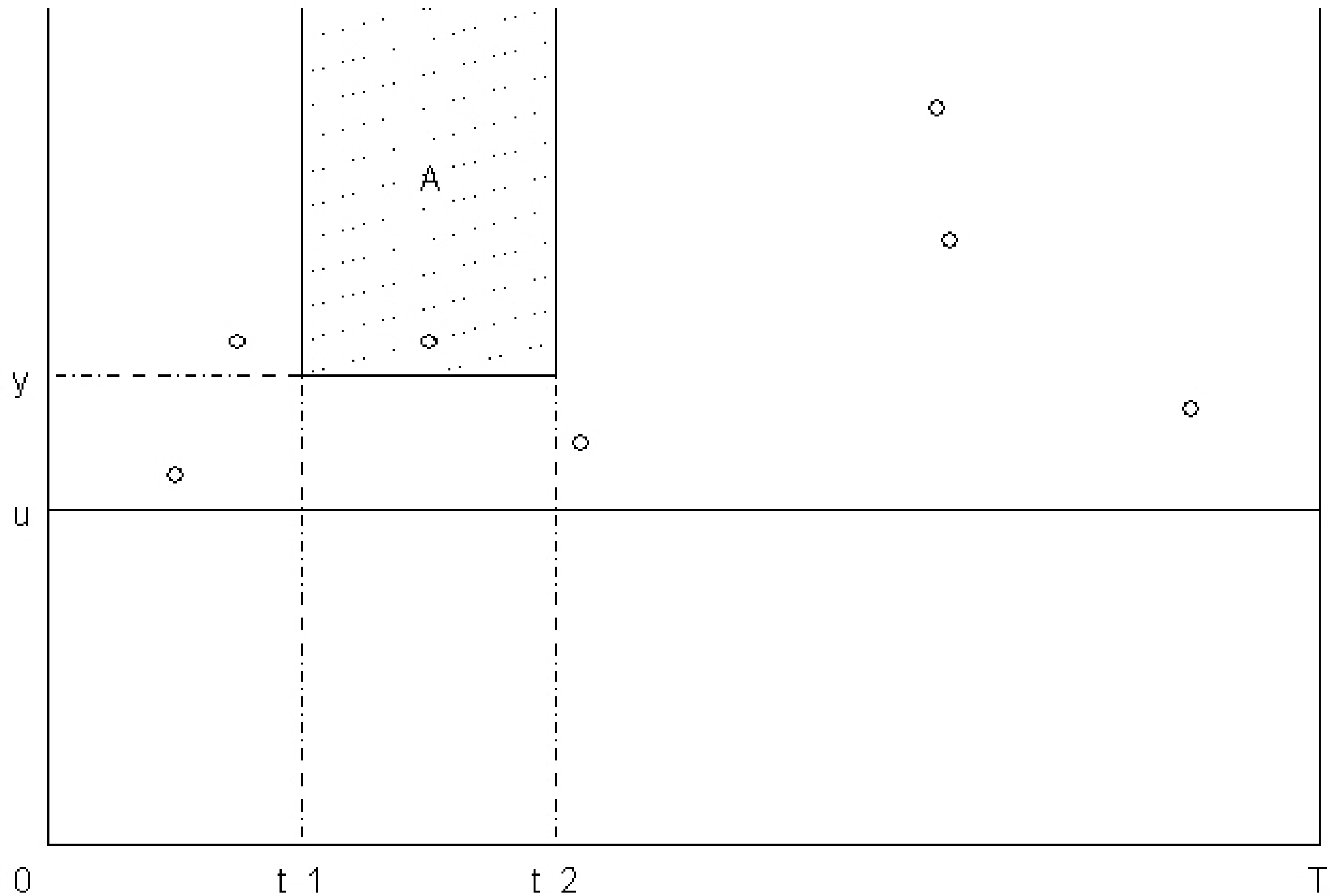
The expected number of exceedances in a box of the form of A is assumed to be

$$\Lambda(A) = (t_2 - t_1)\Psi(y; \mu, \psi, \xi)$$

where

$$\Psi(y; \mu, \psi, \xi) = \left(1 + \xi \frac{y - \mu}{\psi}\right)_+^{-1/\xi}.$$

In practice, allow parameters to depend on covariates.



A
y
u
0
t_1
t_2
T

Diagnostics

Testing threshold exceedance rate

Assume probability of crossing threshold in a small time interval $(t, t+dt)$ is of the form $\lambda(t)dt$. Exceedances at $T_1, T_2, ...$

$$\Pr\{T_k - T_{k-1} > h | T_1, ..., T_{k-2}, T_{k-1} = t\} \;=\; \exp\left\{-\int_t^{t+h} \lambda(s)ds\right\}$$

independently of $\{T_1, ..., T_{k-1}\}$ $(T_0 = 0)$.

Alternatively,

$$Z_k = \int_{T_{k-1}}^{T_k} \lambda(s)ds, \quad k = 1, 2, ...,$$

are independent exponentially distributed with mean 1.

In practice, use discrete analog of Z_k.

Testing distribution of excesses

High value $Y_t > u$ at $t = T_k$.

$$W_k = \frac{1}{\xi_{T_k}} \log \left(1 + \xi_{T_k} \frac{Y_{T_k} - u}{\psi_{T_k}} \right)$$

or in the case $\xi_{T_k} = 0$,

$$W_k = \frac{Y_{T_k} - u}{\psi_{T_k}},$$

then the $\{W_k\}$ have independent exponential distributions with mean 1, if the model is correct.

Uses of the Z and W statistics

- Plot Z_k and W_k against time T_k to look for trends

- QQ plots of ordered Z_k and W_k to test for distribution

- Correlation plots to look for time series dependence

III.3. Application: Insurance data

GPD fits to various thresholds:

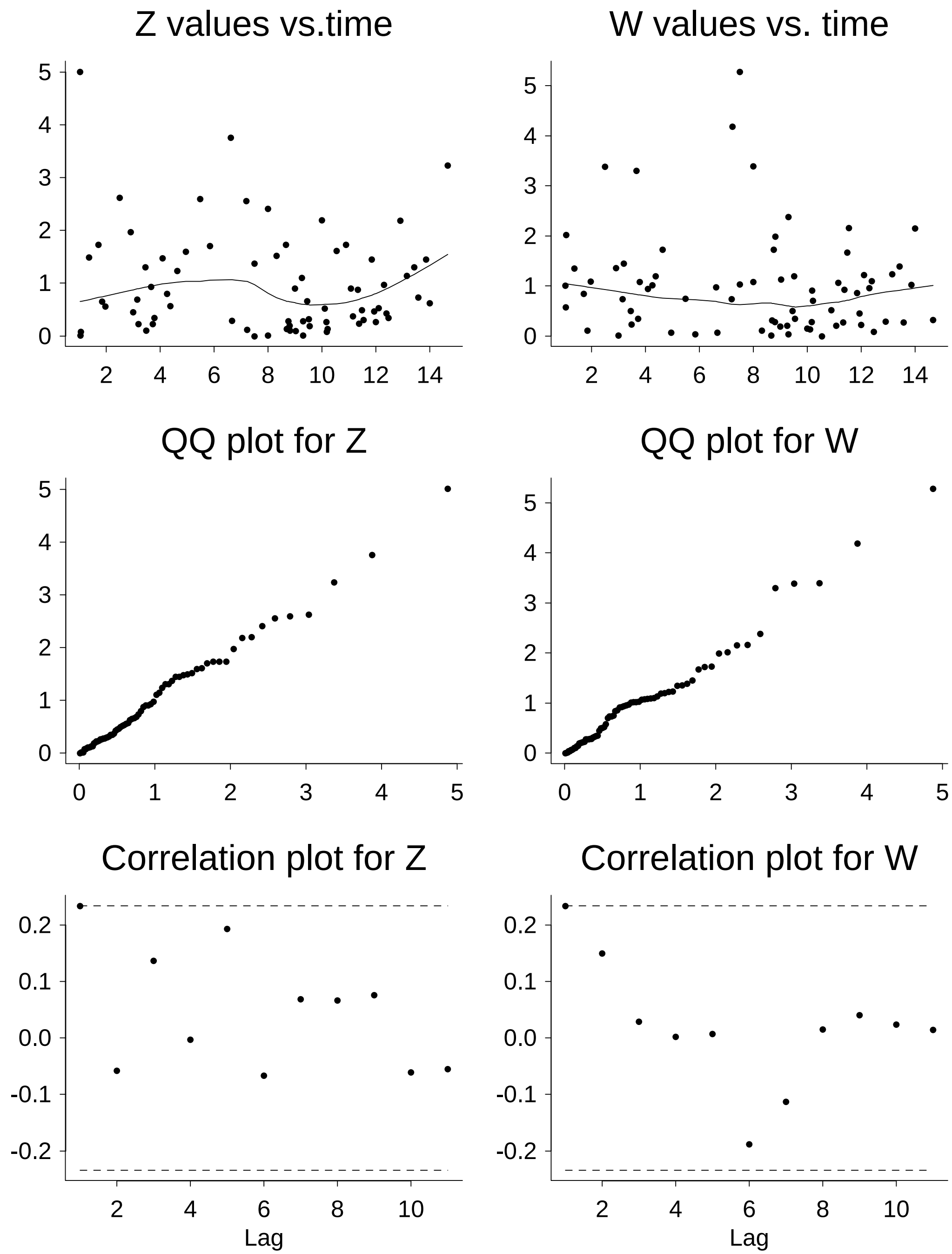
u	N_u	Mean Excess	σ	ξ
0.5	393	7.11	1.02	1.01
2.5	132	17.89	3.47	0.91
5	73	28.9	6.26	0.89
10	42	44.05	10.51	0.84
15	31	53.60	5.68	1.44
20	17	91.21	19.92	1.10
25	13	113.7	74.46	0.93
50	6	37.97	150.8	0.29

Point process approach:

u	N_u	μ	$\log\psi$	ξ
0.5	393	26.5 (4.4)	3.30 (0.24)	1.00 (0.09)
2.5	132	26.3 (5.2)	3.22 (0.31)	0.91 (0.16)
5	73	26.8 (5.5)	3.25 (0.31)	0.89 (0.21)
10	42	27.2 (5.7)	3.22 (0.32)	0.84 (0.25)
15	31	22.3 (3.9)	2.79 (0.46)	1.44 (0.45)
20	17	22.7 (5.7)	3.13 (0.56)	1.10 (0.53)
25	13	20.5 (8.6)	3.39 (0.66)	0.93 (0.56)

Standard errors are in parentheses

Insurance Data with Threshold 5



Conclusions for this example

- Either the GPD or the point process model fits very well, but the point process model is easier to interpret because the parameters are stable across different thresholds

- No evidence of an overall time trend (and no connection with climate change)

- However, there is evidence of a *type of claim* effect and a Bayesian hierarchical analysis (Smith and Goodman 2000) shows the predicted probabilities of extreme events change quite a bit if these are taken into account

III.4. Trends in U.S. rainfall extremes

Data base: 187 stations of daily rainfall data from HCN network. Most stations start from 1910 but this analysis is restricted to 1951–1997 during which coverage percentage is fairly constant.

The analysis will assume that for each station, the data may be described by a point-process model with parameters (μ_t, ψ_t, ξ_t) dependent on time t.

From this we shall estimate a "trend in extremes" for each station.

Then we combine information across stations in a spatial analysis.

Models

Model 1:

$$\mu_t = \mu_0 + v_t, \ \psi_t = \psi_0, \ \xi_t = \xi_0.$$

Model 2:

$$\mu_t = \mu_0 e^{v_t}, \ \psi_t = \psi_0 e^{v_t}, \ \xi_t = \xi_0.$$

Regression term:

$$v_t = \sum x_{tj}\beta_j$$

where regressors may be

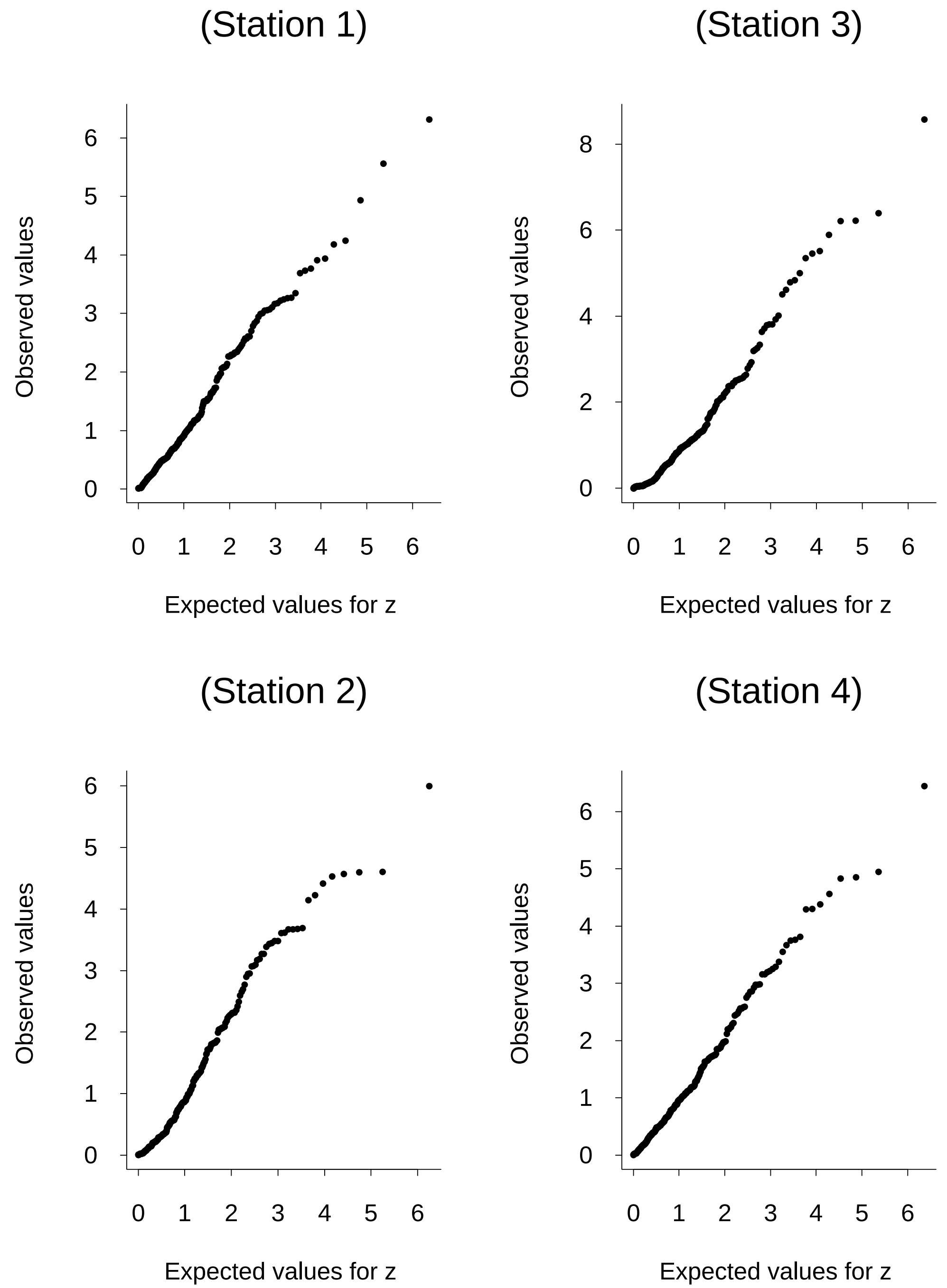
- Linear time trend in first covariate ($x_{t1} = t$):
- Seasonal terms ($\cos\omega t, \ \sin\omega t$)
- External signals, e.g. El Niño

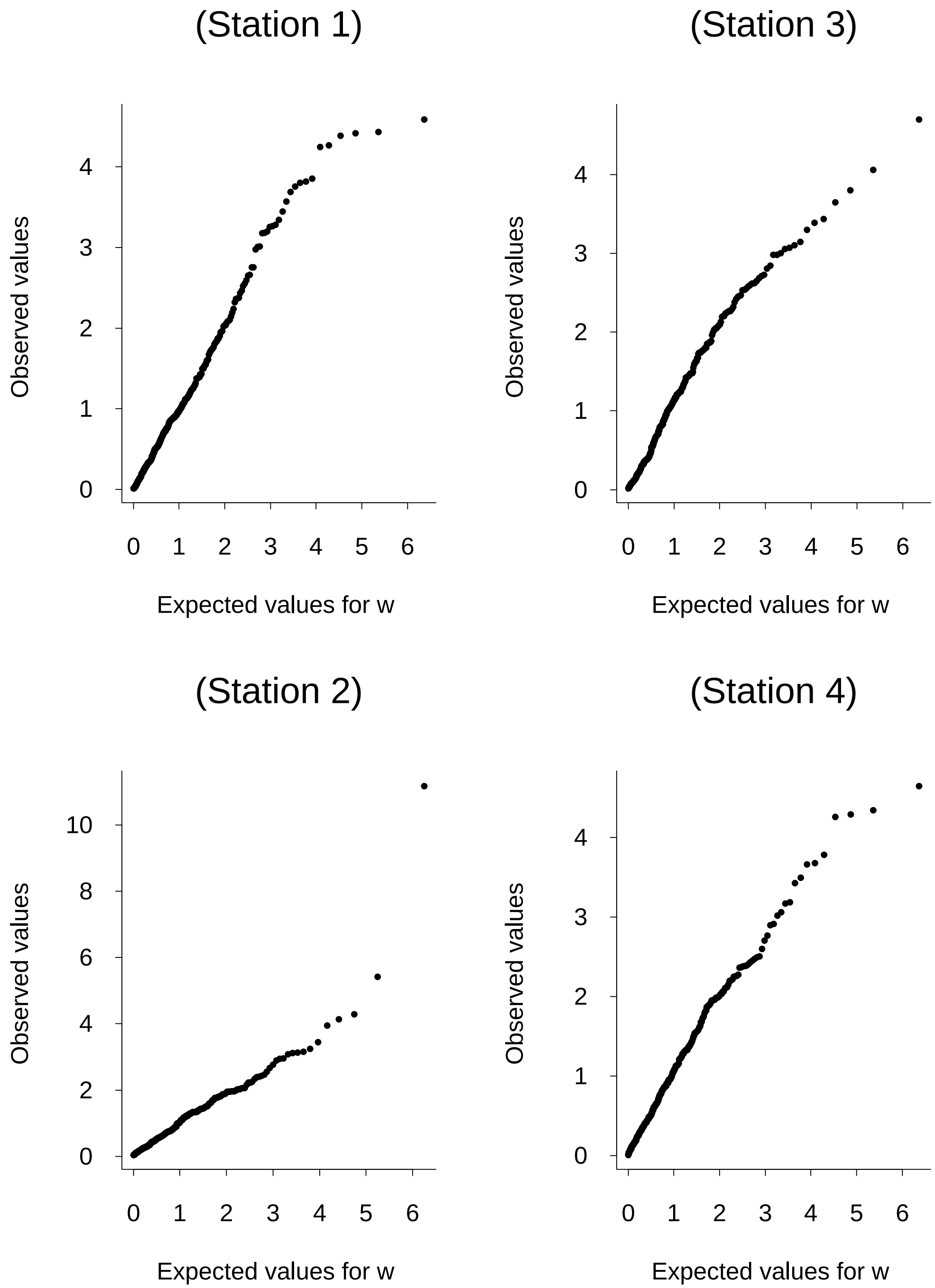
Results of single-station analyses

For the overall analysis, model 2 was adopted, though it is not clear that it fits better than model 1.

As an example of diagnostics, we show QQ plots for the Z-statistics and W-statistics of four stations. Note outlier in W-plot for Station 2 (Gunnison, CO).

The main focus was on the parameter β_1, measured separately for each station, representing the overall rate of increase in extreme rainfall quantiles.


(Station 1)
(Station 3)
(Station 2)
(Station 4)
Observed values
Expected values for z


(Station 1)
(Station 3)
(Station 2)
(Station 4)
Observed values
Expected values for w
0
1
2
3
4
5
6
8
10

Combining results

	β_1	ξ
$t > 2$	25	74
$t > 1$	73	134
$t > 0$	125	162
$t < 0$	59	22
$t < -1$	21	5
$t < -2$	10	1

Summary table of t statistics (estimate divided by standard error) for extreme value model applied to 187 stations and 98% threshold.

Question: How to integrate the results from 187 stations in a meaningful way?

Spatial integration of time trend parameter

$\beta_1(s)$: true but unobserved spatial field, indexed by location s

$\widehat{\beta}_1(s)$: estimate of $\beta_1(s)$ at site s

2-stage model: universal kriging model with measurement error

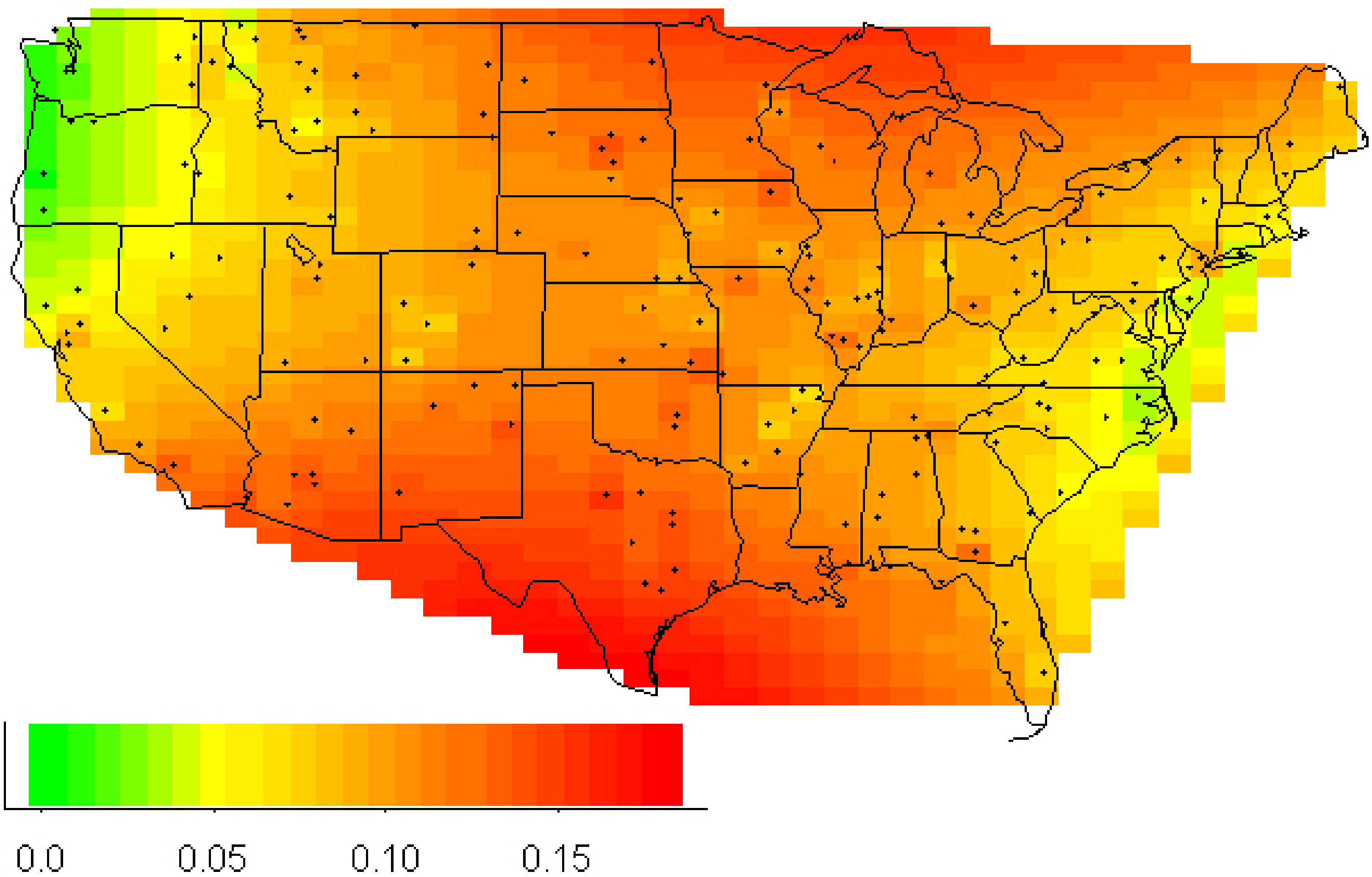
$$\begin{aligned} \beta_1 &\sim \mathcal{N}(X\beta, \Sigma), \\ \widehat{\beta}_1 \, |\beta_1 &\sim \mathcal{N}(Z, W). \end{aligned}$$

Combined:

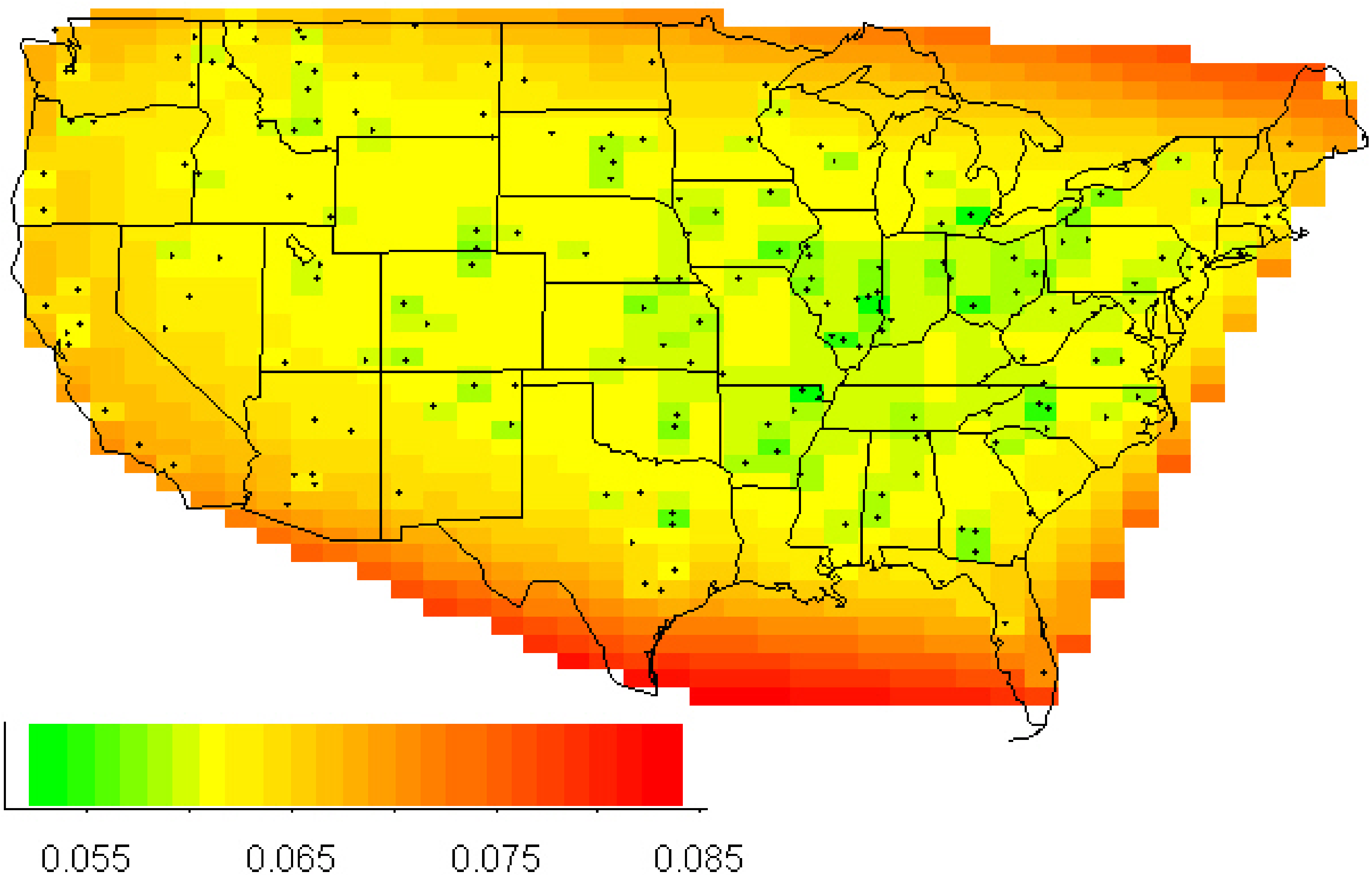
$$\widehat{\beta}_1 \sim \mathcal{N}(X\gamma, \Sigma + W].$$

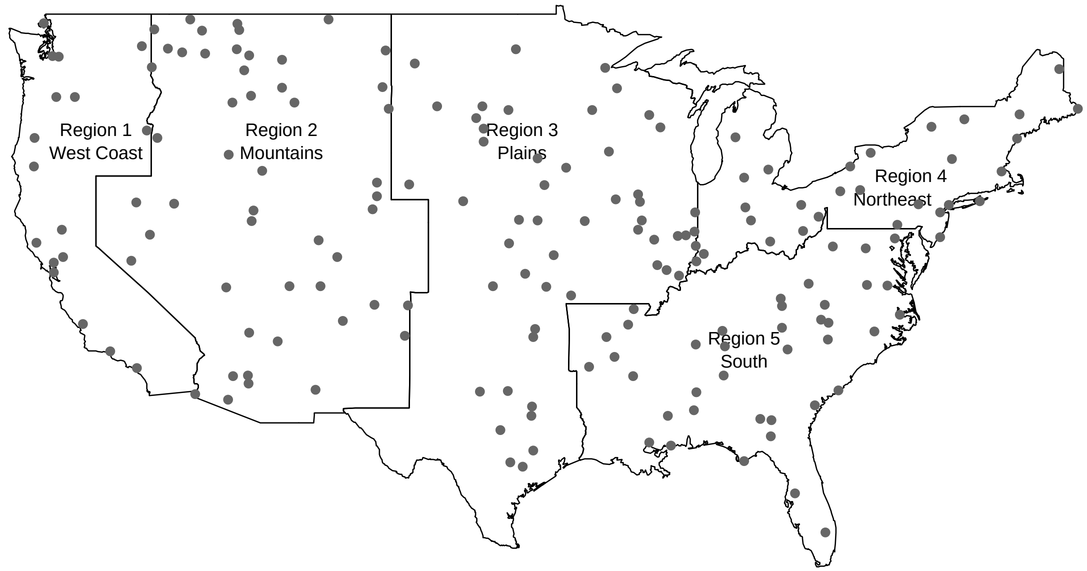
In rainfall application, Σ taken as a exponential or Matérn spatial covariance function. Possible systematic variation of β_1 across space taken into account by $X\gamma$ term (in practice, a quadratic polynomial in latitude and longitude)

Interpolated rainfall trend



S.E. of interpolated trend





Five regions for regional analysis

Results of Regional Analysis

Region	Extreme Rainfall Trend	(S.E.)
1	.055	.024
2	.092	.017
3	.115	.014
4	.097	.016
5	.075	.013
All	.094	.007

Table represents mean trend (average spatially smoothed value of trend parameter $\widehat{\beta}_1$) over each of five regions, and over whole country. Also shown is the estimated standard error of the regional average by the spatial smoothing technique.

Conclusions and Summary

- Although the estimated β_1 at a single site is generally not significant, when combined across all sites, there is clear significant evidence of positive β_1 (meaning, positive trend in the frequency of extreme events)

- There are, however, significant differences across regions

- The challenge for the future is to reconcile these results with those of weather forecasting model reanalyses (e.g. by NCEP, ECMWF) and with climate models. If this exercise is successful, we can hope to use the model for probabilistic prediction of extreme rainfall events in future climate change scenarios (connection with Claudia Tebaldi's talk earlier today)

REFERENCES

Banerjee, S., Carlin, B.P. and Gelfand, A.E. (2004), *Hierarchical Modeling and Analysis for Spatial Data*. Chapman and Hall/CRC, Boca Raton, FL.

Berger, J.O., De Oliveira, V. and Sansó, B. (2001), Objective Bayesian analysis of spatially correlated data. *Journal of the American Statistical Association* **96**, 1361–1374.

Chilès, J.-P. and Delfiner, P. (1999). *Geostatistics: Modeling Spatial Uncertainty*. John Wiley, New York.

Coles, S.G. (2001) *An Introduction to Statistical Modeling of Extreme Values.* Springer Verlag, New York.

Cressie, N. (1993), *Statistics for Spatial Data*. 2nd edition, Wiley, New York.

Fisher, R.A. and Tippett, L.H.C. (1928), Limiting forms of the frequency distributions of the largest or smallest member of a sample. *Proc. Camb. Phil. Soc.* **24**, 180–190.

Gnedenko, B.V. (1943), Sur la distribution limite du terme maximum d'une série aléatoire. *Ann. Math.* **44**, 423-453. (In French; an English translation is contained in the book *Breakthroughs in Statistics I: Foundations and Basic Theory*, edited by S. Kotz and N.L. Johnson, Springer Verlag, New York (publ. 1992).

Harville, D.A. (1974), Bayesian inference for variance components using only error contrasts. *Biometrika* **61**, 383-385.

Harville, D.A. and Jeske, D.R. (1992), Mean squared error of estimation or prediction under a general linear model. *J. Amer. Statist. Assoc.* **87**, 724-731.

Mardia, K.V. and Goodall, C.R. (1993), Spatial-temporal analysis of multivariate environmental monitoring data. In *Multivariate Environmental Statistics*, eds. G.P. Patil and C.R. Rao, Elsevier Science Publishers, pp. 347–386.

Mardia, K.V. and Marshall, R.J. (1984), Maximum likelihood estimation of models for residual covariance in spatial regression. *Biometrika* **71**, 135-146.

Müller, P. (1999), Simulation based optimal design (with discussion). In *Bayesian Statistics 6*, edited by J.M. Bernardo, J.O. Berger, A.P. Dawid and A.F.M. Smith. Oxford University Press, 459–474.

Müller, W.V. (2000), *Collecting Spatial Data: Optimum Design of Experiments for Random Fields.* Second edition, Physica Verlag, Heidelberg.

Nychka, D. and Saltzman, N. (1998), Design of air quality networks. In *Case Studies in Environmental Statistics*, eds. D. Nychka, W. Piegorsch and L.H. Cox, Lecture Notes in Statistics number 132, Springer Verlag, New York, pp. 51–76.

Pickands, J. (1975), Statistical inference using extreme order statistics. *Ann. Statist.* **3**, 119-131.

Smith, R.L. (2001), *Environmental Statistics*. Presented as a CBMS lecture series at the University of Washington, June 2001. Currently under revision for book publication. Current version available from http://www.stat.unc.edu/postscript/rs/envnotes.pdf

Smith, R.L. (2003), Statistics of extremes, with applications in environment, insurance and finance. Chapter 1 of, *Extreme Values in Finance, Telecommunications and the Environment*, edited by B. Finkenstadt and H. Rootzen, Chapman and Hall/CRC Press, London, pp. 1–78.

Smith, R.L. and Goodman, D. (2000), Bayesian risk analysis. Chapter 17 of *Extremes and Integrated Risk Management*, edited by P. Embrechts. Risk Books, London, 235–251.

Smith, R.L, Kolenikov, S. and Cox, L.H. (2003), Spatio-temporal modeling of PM2.5 data with missing values. *J. Geophys. Res., 108* (D24), 9004, doi:10.1029/2002JD002914, 2003.

Smith, R.L. and Zhu, Z. (2004), Asymptotic theory for kriging with estimated parameters and its application to network design. Preliminary version, online at
http://www.stat.unc.edu/postscript/rs/supp5.pdf

Stein, M.L. (1999), *Interpolation of Spatial Data: Some Theory of Kriging.* Springer Verlag, New York.

Zhu, Z. (2002), *Optimal Sampling Design and Parameter Estimation of Gaussian Random Fields.* Ph.D thesis, Department of Statistics, University of Chicago.

Zhu, Z. and Stein, M.L. (2004a), Spatial sampling design for parameter estimation of the covariance function. *Journal of Statistical Planning and Inference*, In press.

Zhu, Z. and Stein, M.L. (2004b), Two-step sampling design for prediction with estimated parameters. Submitted for publication.

Zimmerman, D.L. and Cressie, N. (1992), Mean squared prediction error in the spatial linear model with estimated covariance parameters. *Ann. Inst. Statist. Math.* **44**, 27-43.