Statistical methods for environmental data

Peter Guttorp
University of Washington
USA
peter@stat.washington.edu
www.stat.washington.edu/peter

Outline

Lecture 1: Space-time modeling of air quality data

- Kriging, nonstationary covariance, singular value decomposition

Lecture 2: Extremes, air quality standards, and climate trends

- Hypothesis testing, Rice's formula

Lecture 3: Compositional data in the environment

- Algebra of compositions, logistic normal distribution, spatial autoregression

1. Space-time modeling of air quality data

The spatial problem

Given observations at n locations
$Z(s_1),\dots,Z(s_n)$
estimate

$Z(s_0)$ **(the process at an unobserved site)**

or $\int_A Z(s)\,d\nu(s)$ **(an average of the process)**

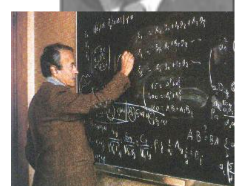
In the environmental context often time series of observations at the locations.

Acknowledgements

Joint work with
Paul Sampson
Wendy Meiring
Doris Damian

Some history

Regression (Galton, Bartlett)



Mining engineers (Krige 1951, Matheron, 60s)

Spatial models (Whittle, 1954)

Forestry (Matérn, 1960)

Objective analysis (Grandin, 1961)

More recent work Cressie (1993), Stein (1999)

A Gaussian formula

If $\begin{pmatrix} \mathbf{X} \\ \mathbf{Y} \end{pmatrix} \sim \mathbf{N}\left(\begin{pmatrix} \mu_{\mathbf{X}} \\ \mu_{\mathbf{Y}} \end{pmatrix}, \begin{pmatrix} \Sigma_{\mathbf{XX}} & \Sigma_{\mathbf{XY}} \\ \Sigma_{\mathbf{YX}} & \Sigma_{\mathbf{YY}} \end{pmatrix} \right)$

then $(\mathbf{Y} \mid \mathbf{X}) \sim \mathbf{N}(\mu_{\mathbf{Y}} + \Sigma_{\mathbf{YX}}\Sigma_{\mathbf{XX}}^{-1}(\mathbf{X} - \mu_{\mathbf{X}}),$

$\Sigma_{\mathbf{YY}} - \Sigma_{\mathbf{YX}}\Sigma_{\mathbf{XX}}^{-1}\Sigma_{\mathbf{XY}})$

Simple kriging

Let $X = (Z(s_1),\ldots,Z(s_n))^T$, $Y = Z(s_0)$, so that

$\mu_X=\mu 1_n$, $\mu_Y=\mu$,

$\Sigma_{XX}=[C(s_i-s_j)]$, $\Sigma_{YY}=C(0)$, and

$\Sigma_{YX}=[C(s_i-s_0)]$.

Then

$$p(X) \equiv \hat{Z}(s_0) = \mu + \left[C(s_i - s_0)\right]^T \left[C(s_i - s_j)\right]^{-1} \left(X - \mu 1_n\right)$$

This is the best unbiased linear predictor when μ and C are known (simple kriging).

The prediction variance is

$$m_1 = C(0) - \left[C(s_i - s_0)\right]^T \left[C(s_i - s_j)\right]^{-1} \left[C(s_i - s_0)\right]$$

Some variants

Ordinary **kriging (unknown μ)**

$$p(X) \equiv \hat{Z}(s_0) = \hat{\mu} + \left[C(s_i - s_0)\right]^T \left[C(s_i - s_j)\right]^{-1} \left(X - \hat{\mu}1_n\right)$$

where

$$\hat{\mu} = \left(1_n^T \left[C(s_i - s_j)\right]^{-1} 1_n\right)^{-1} 1_n^T \left[C(s_i - s_j)\right]^{-1} X$$

Universal **kriging ($\mu(s)=A(s)\beta$ for some spatial variable A)**

$$\hat{\beta} = (\left[A(s_i)\right]^T \left[C(s_i - s_j)\right]^{-1} \left[A(s_i)\right])^{-1}$$

$$\left[A(s_i)\right]^T \left[C(s_i - s_j)\right]^{-1} X$$

Still optimal for known C.

Universal kriging variance

simple kriging variance

$$E\left(\hat{Z}(s_0) - Z(s_0)\right)^2 = \boxed{m_1} +$$

$$\left(A(s_0) - [A(s_i)^T [C(s_i - s_j)]^{-1} [C(s_i - s_0)]\right)^T$$
$$\times\left(\left[A(s_i)\right]^T \left[C(s_i - s_j)\right]^{-1} \left[A(s_i)\right]\right)^{-1}$$
$$\times\left(A(s_0) - [A(s_i)^T [C(s_i - s_j)]^{-1} [C(s_i - s_0)]\right)$$

variability due to estimating β

The (semi)variogram

$$\gamma(\|h\|) = \frac{1}{2}\mathrm{Var}(Z(s+h) - Z(s)) = C(0) - C(\|h\|)$$

Intrinsic stationarity

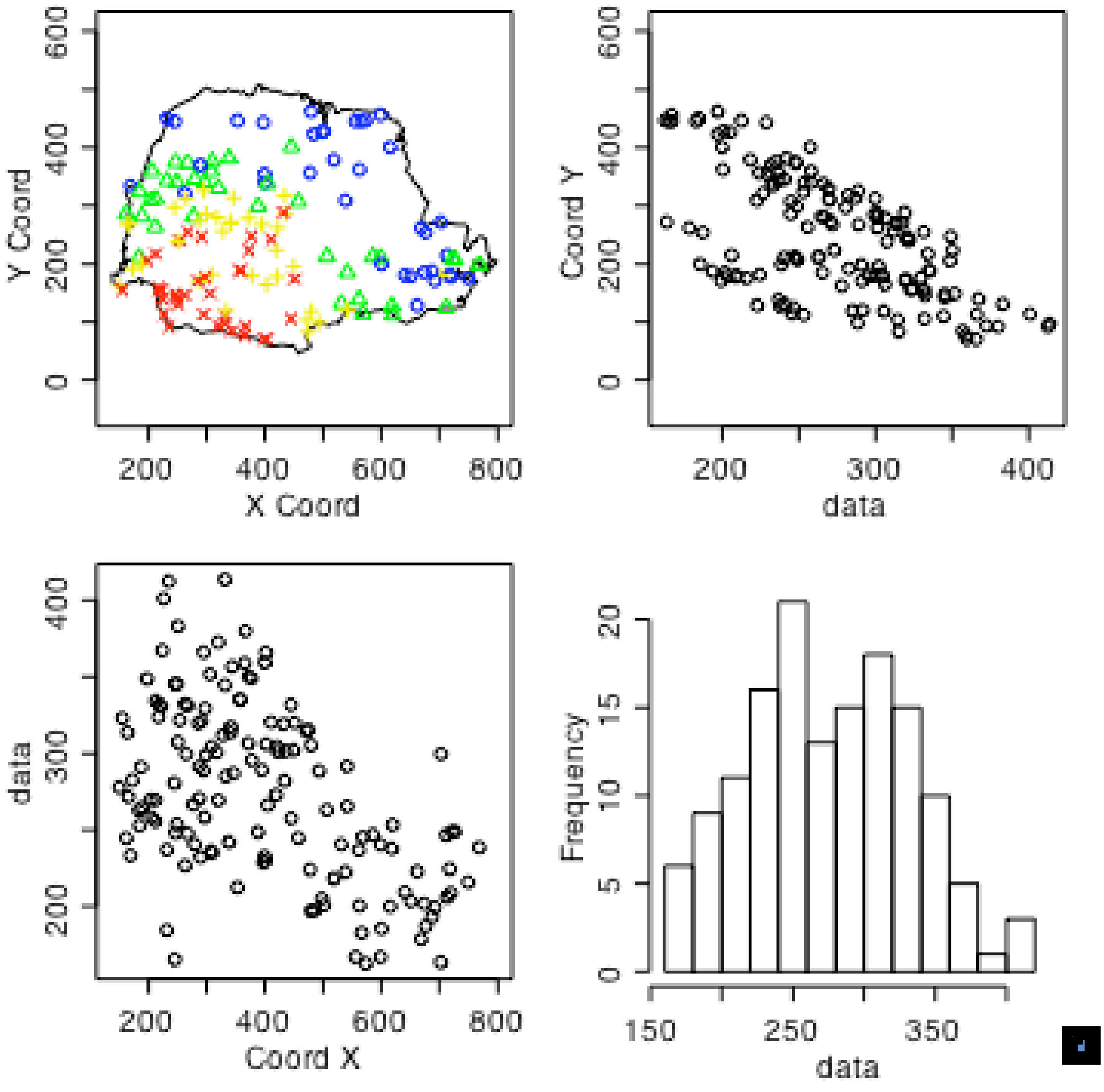
Weaker assumption (C(0) needs not exist)

Kriging predictions can be expressed in terms of the variogram instead of the covariance.

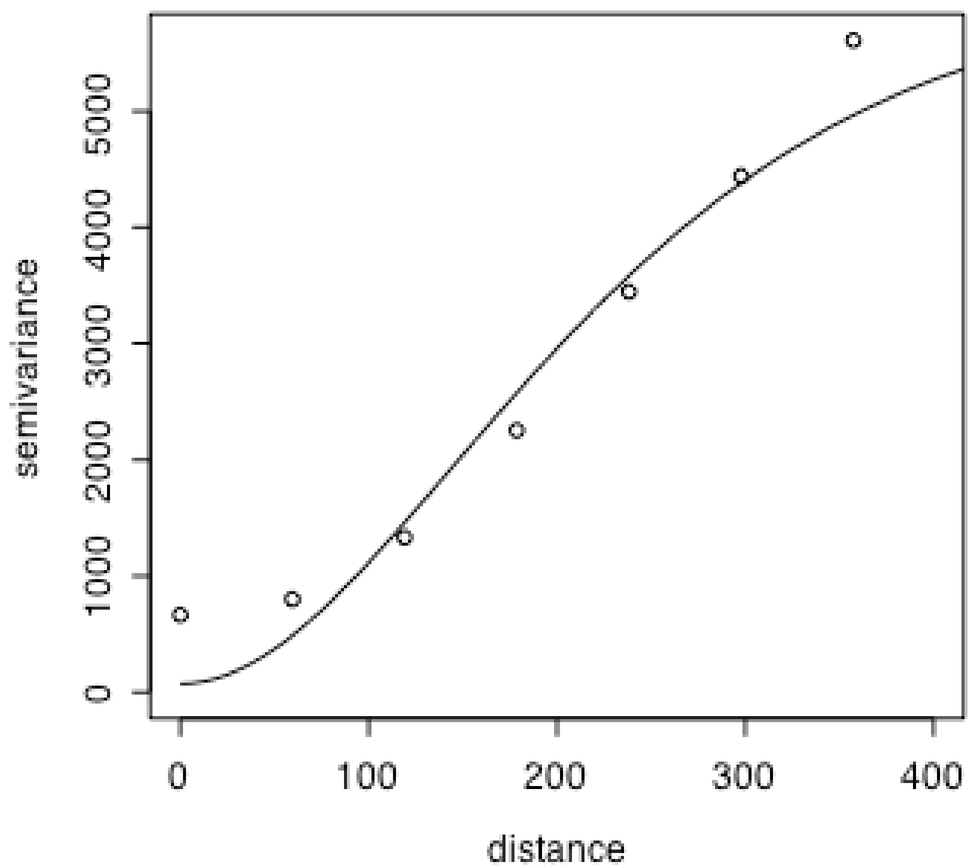
Parana rainfall

- Built-in geoR data set
- Average rainfall over different years for May-June (dry-season)
- 143 recording stations throughout Parana State, Brazil

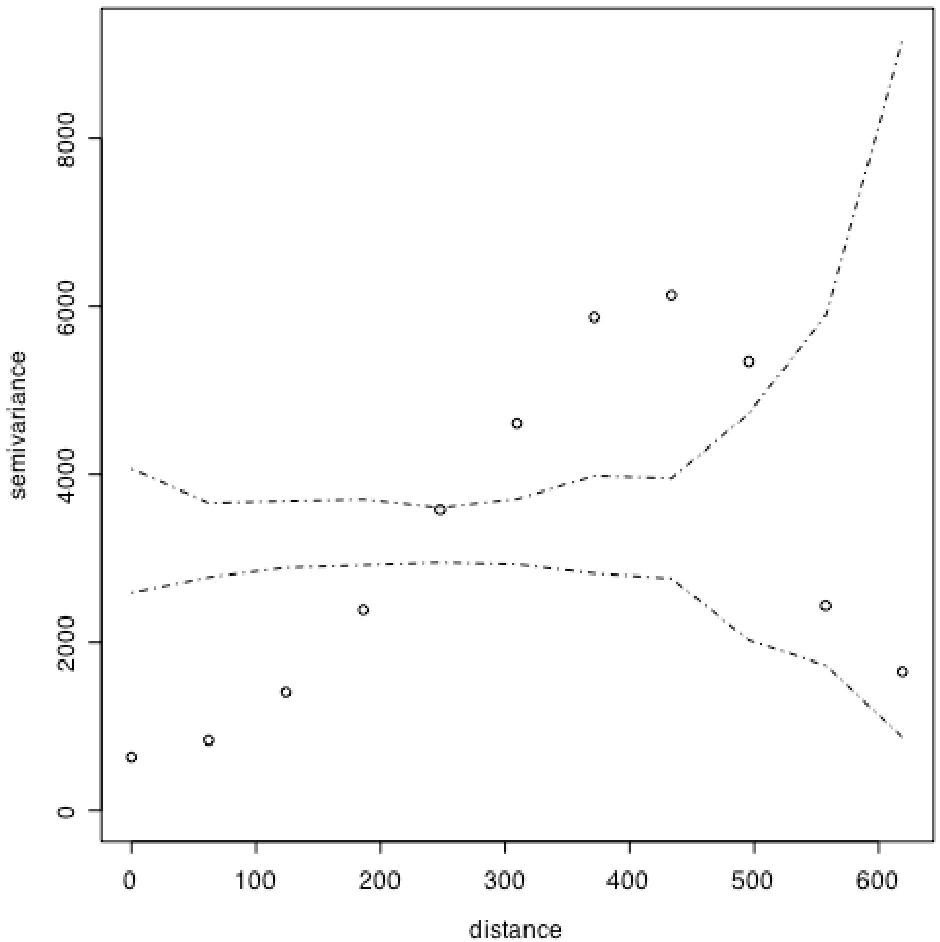
Parana precipitation



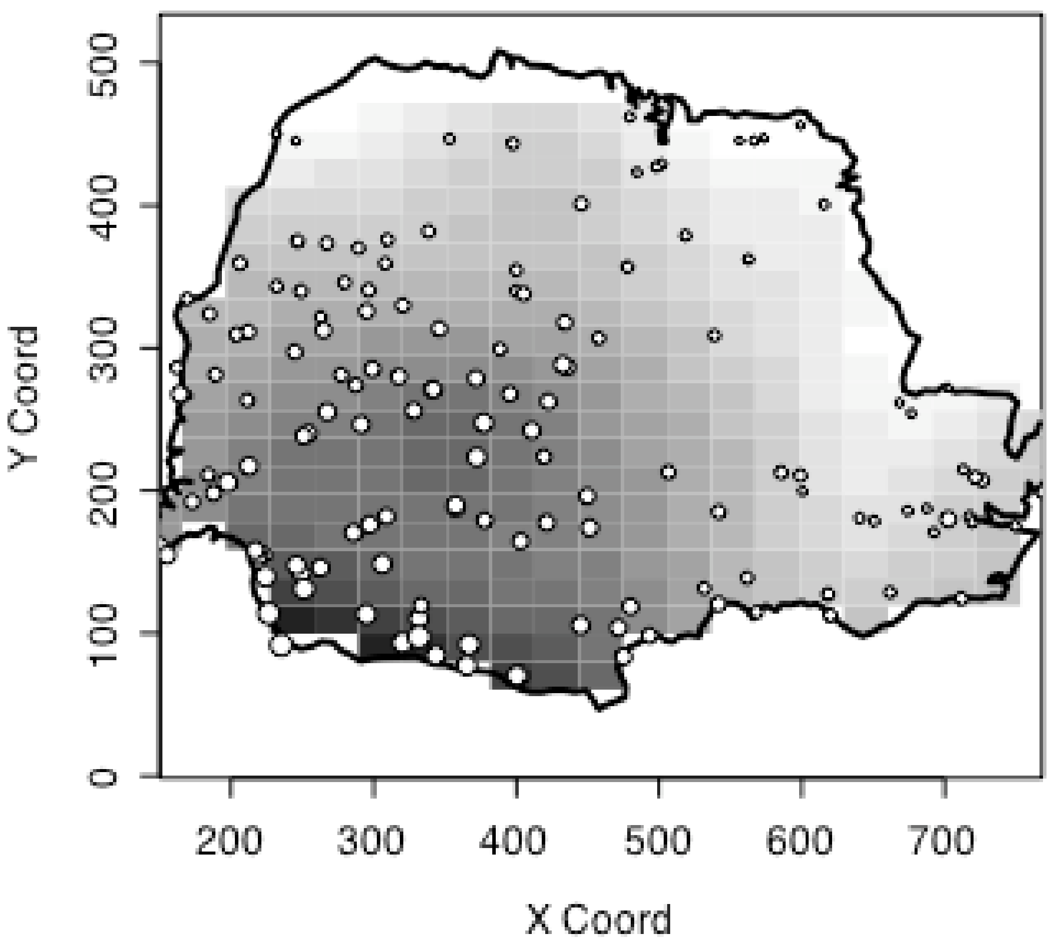
Fitted variogram



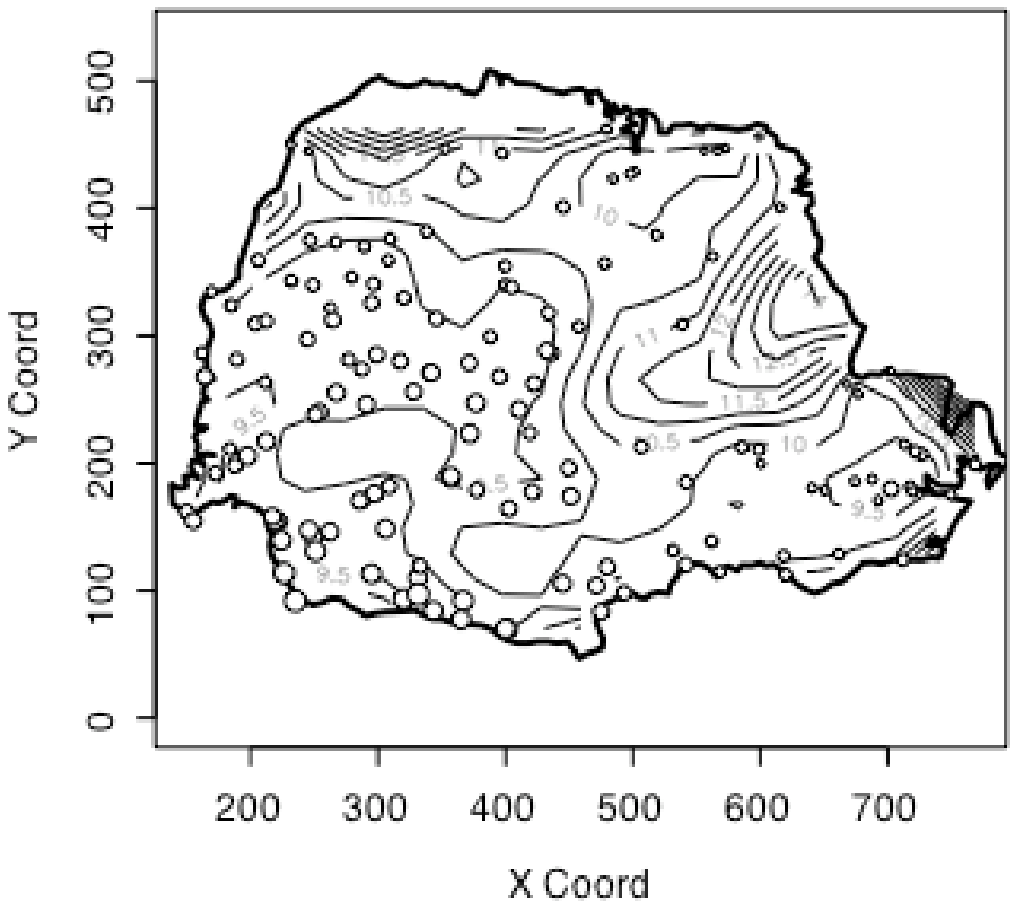
Is it significant?



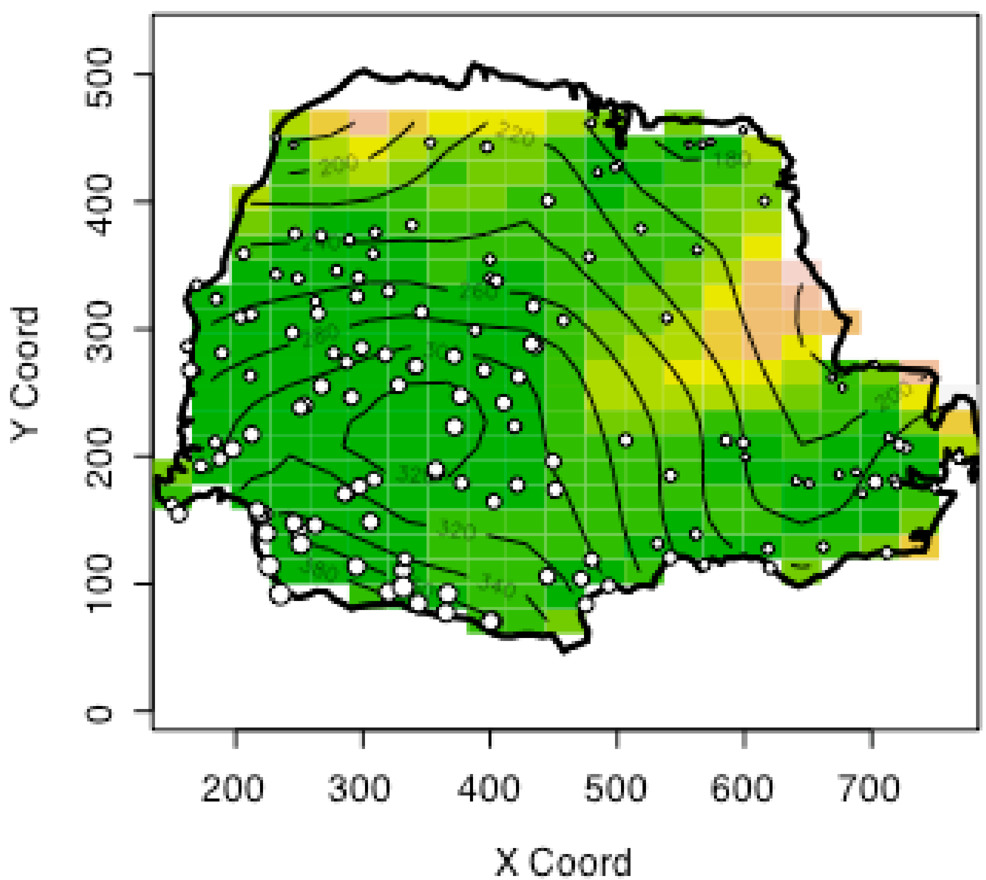
Kriging surface



Kriging standard error



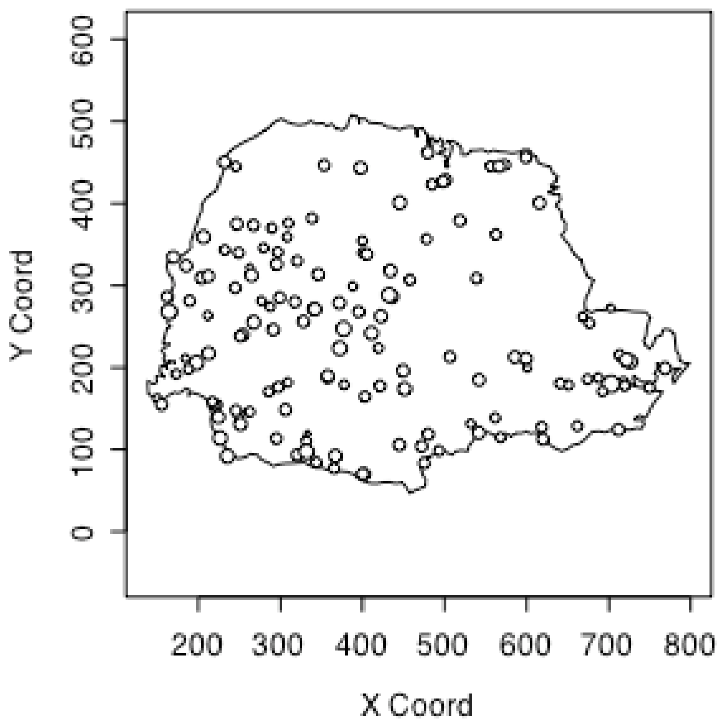
A better combination



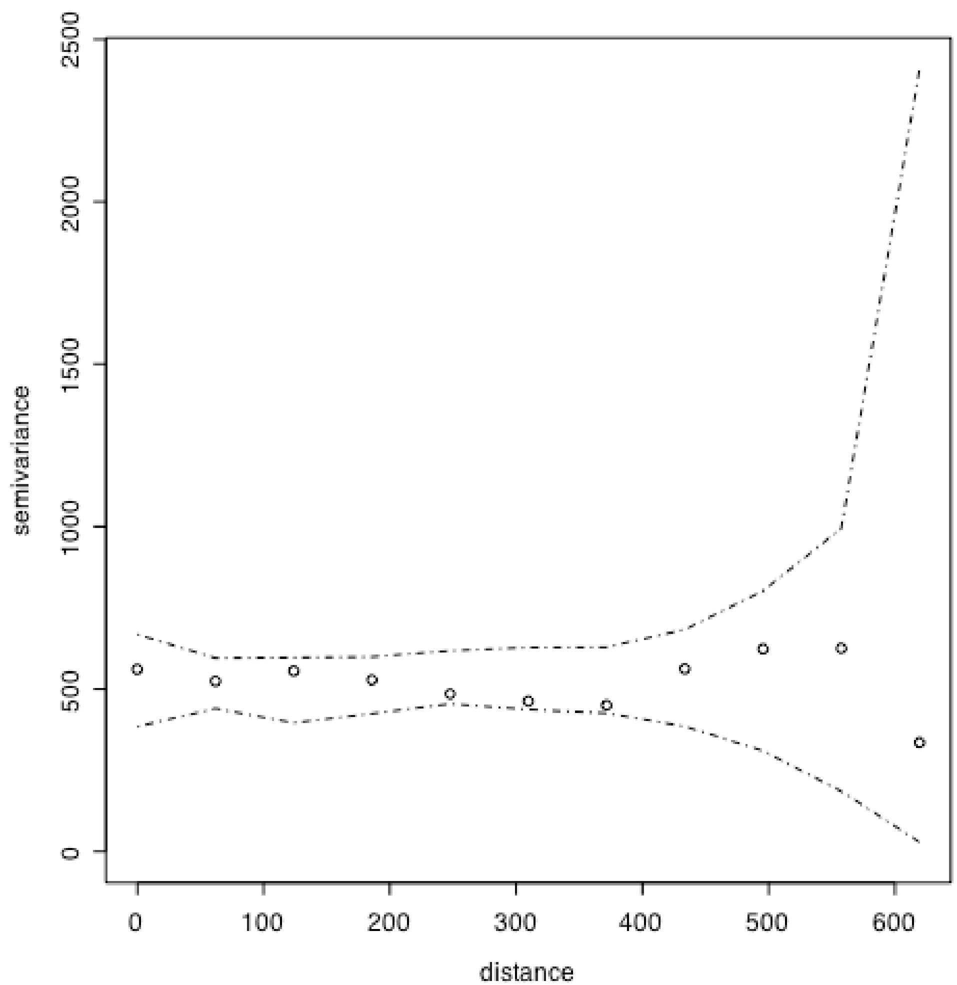
Spatial trend

Indication of spatial trend

Fit quadratic in coordinates



Residual variogram



Geometric anisotropy

If $C(x,y) = C(\|x - y\|)$ we have an *isotropic* covariance (circular isocorrelation curves).

If $C(x,y) = C(\|Ax - Ay\|)$ for a linear transformation A, we have *geometric anisotropy* (elliptical isocorrelation curves).

General nonstationary correlation structures are typically locally geometrically anisotropic.

The deformation idea

In the geometric anisotropic case, write

$$C(x,y) = C(\|f(x) - f(y)\|)$$

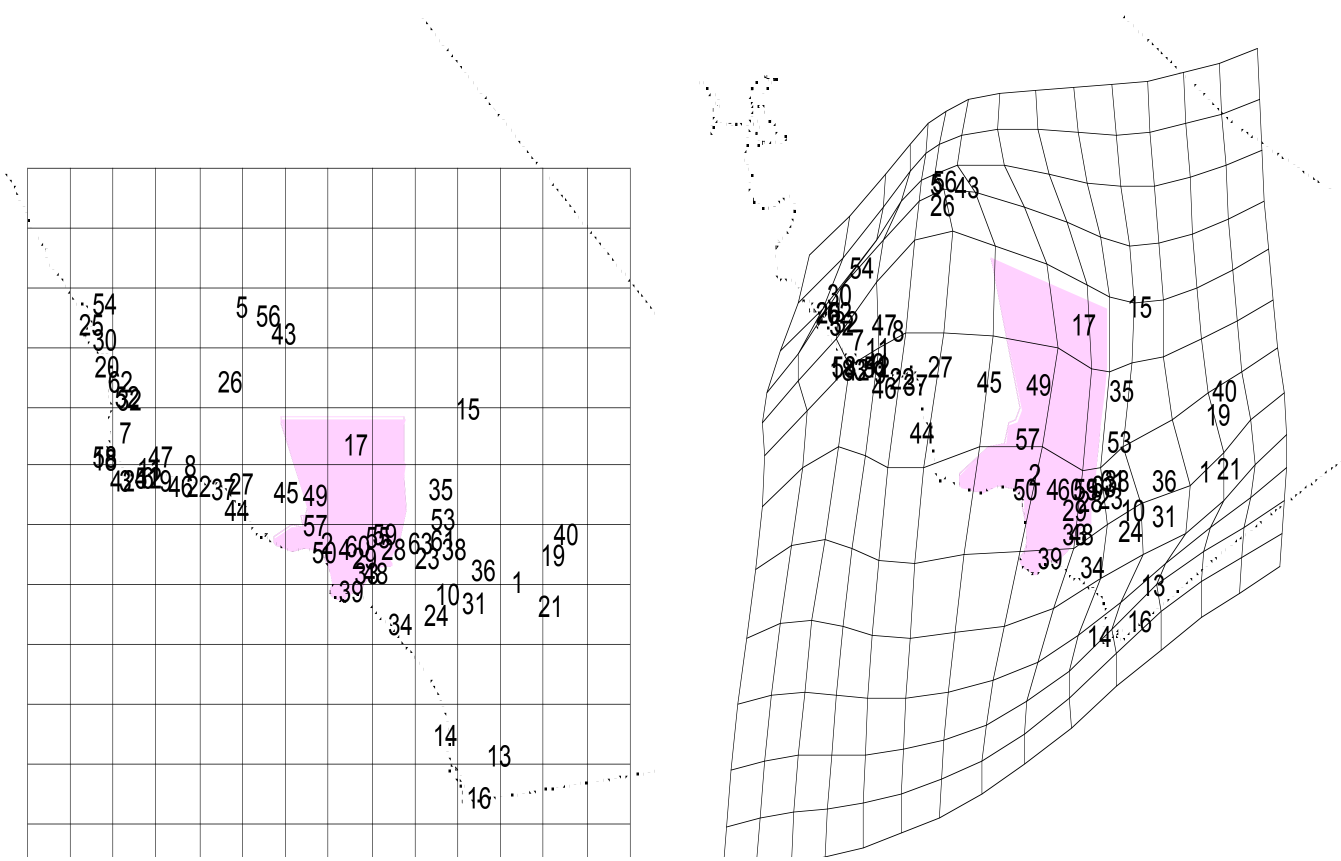
where f(x) = Ax. This suggests using a general nonlinear transformation
$f: R^2 \to R^d$. Usually d=2 or 3.

G-plane D-space

We do not want f to fold.

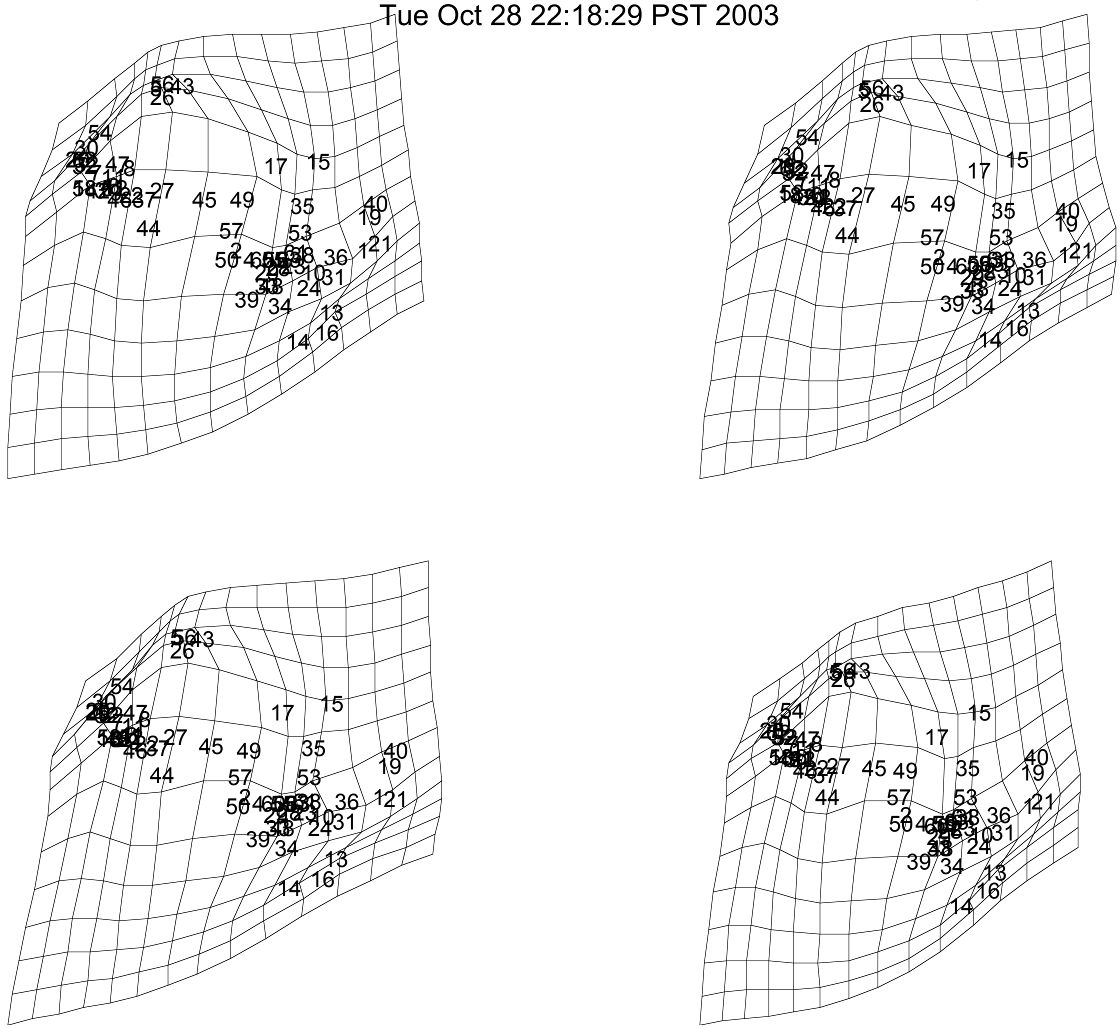
Do a Bayesian implementation using thin plate splines

California ozone

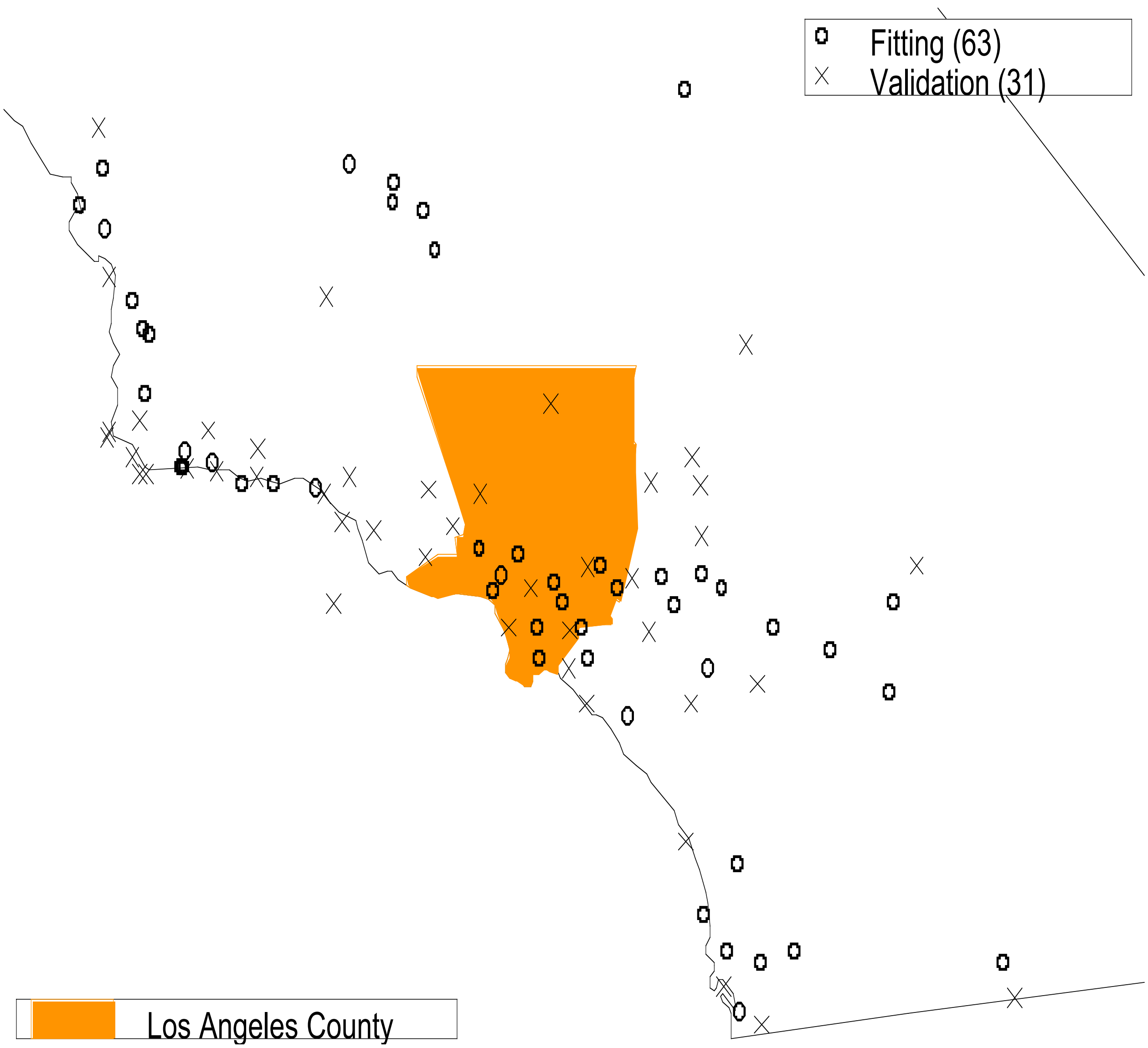


Posterior samples

N=63, S. Calif: 4 samples from the posterior distribution of deformations reflecting spatial covariance
Tue Oct 28 22:18:29 PST 2003



Region 6: S Calif, all 94 sites, fitting and validation



Trend model

$$\mu(s_i,t) = \mu_1(s_i) + \mu_2(s_i,t)$$

$$\mu_1(s_i) = \mu_0(s_i) + \sum \delta_k V_{ik}$$

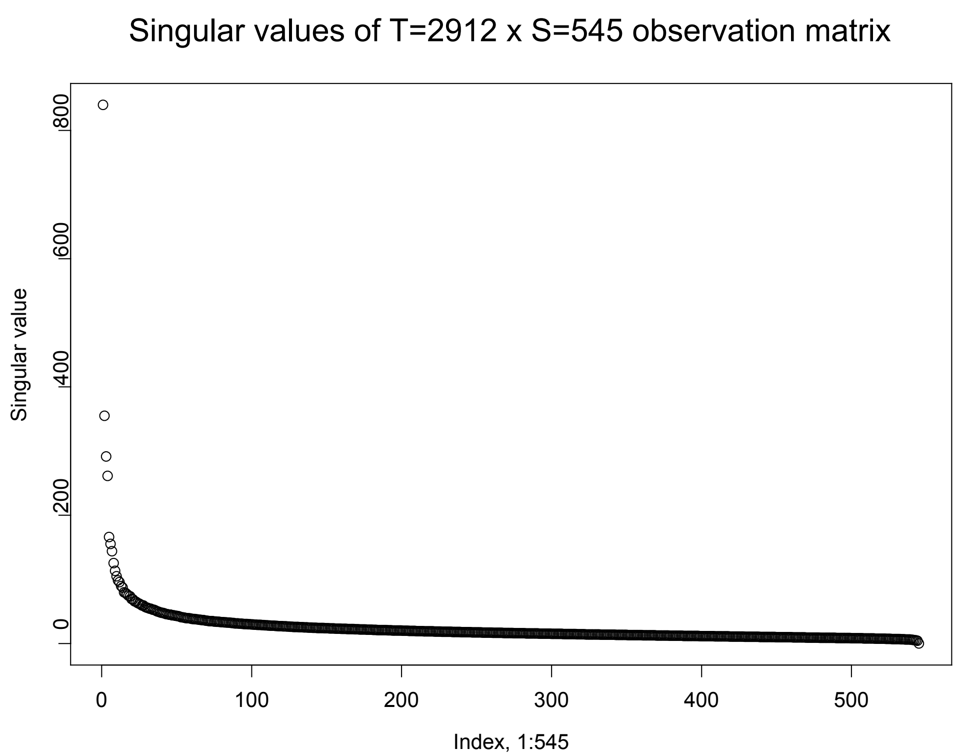
where V_{ik} are covariates, such as population density, proximity to roads, local topography, etc.

$$\mu_2(s_i,t) = \sum \rho_j(s_i) f_j(t)$$

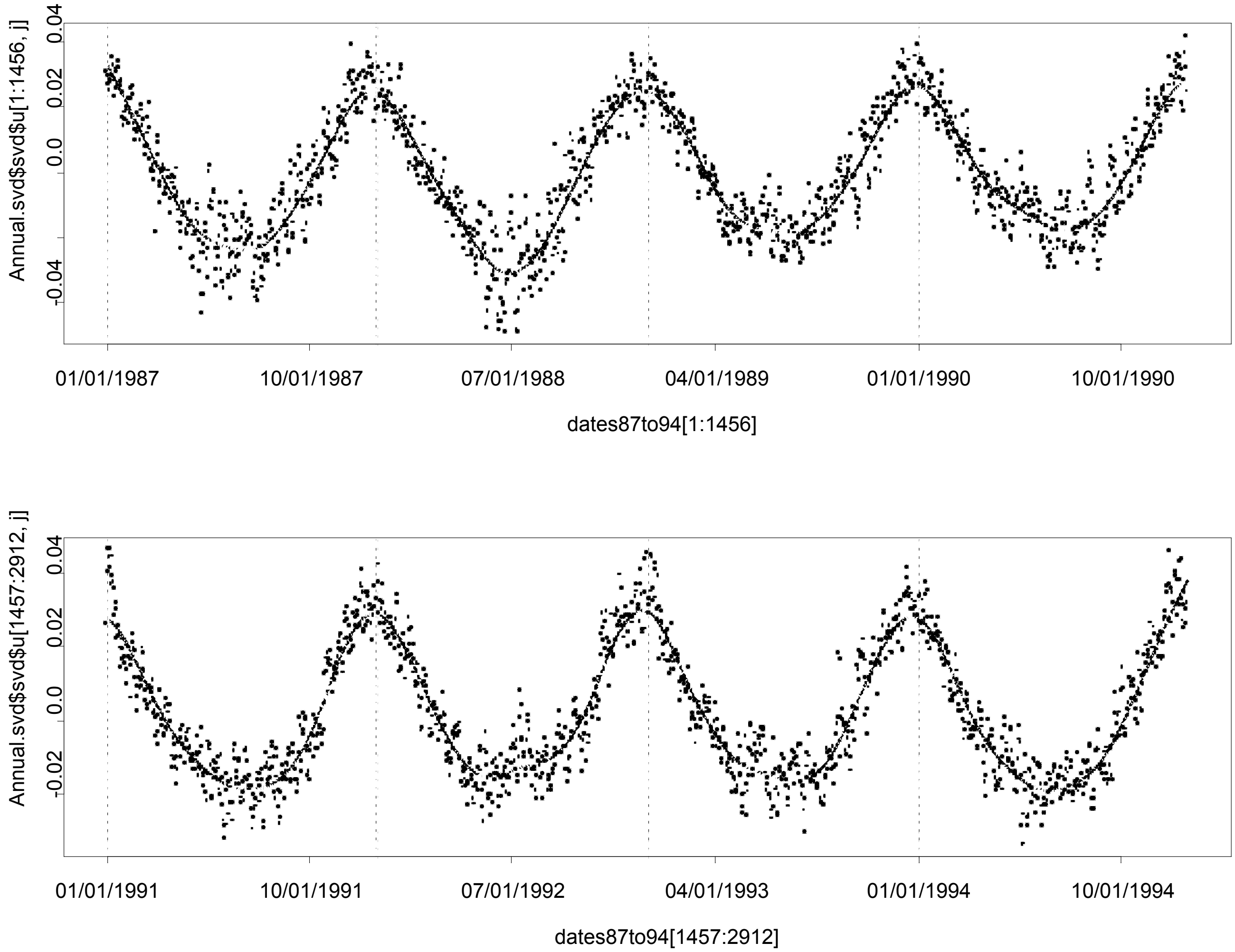
where the f_j are smoothed versions of temporal singular vectors (EOFs) of the TxN data matrix.

We will set $\mu_1(s_i) = \mu_0(s_i)$ for now.

SVD computation

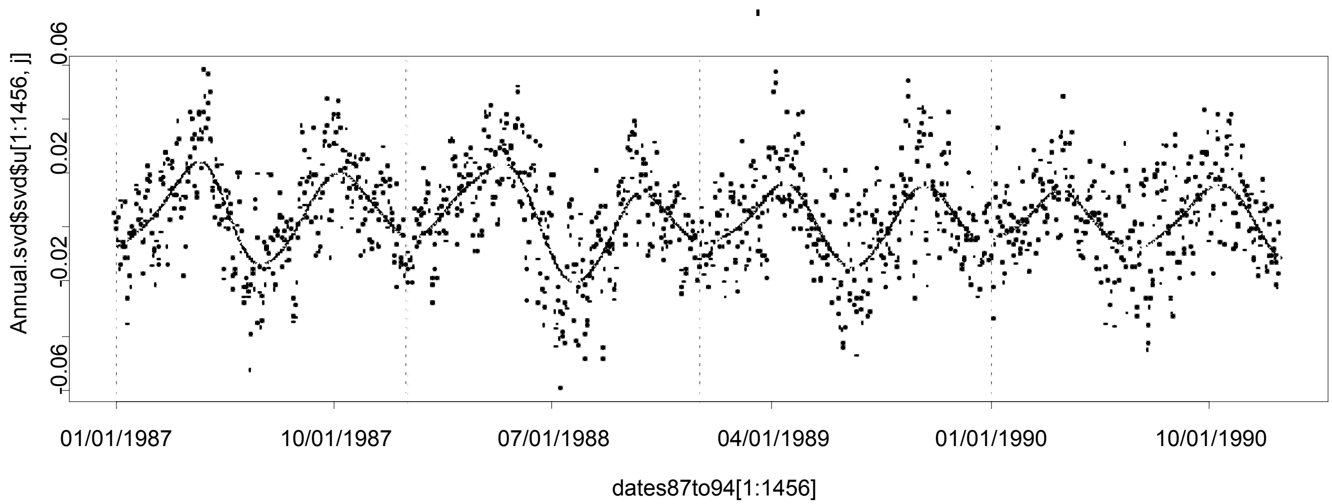


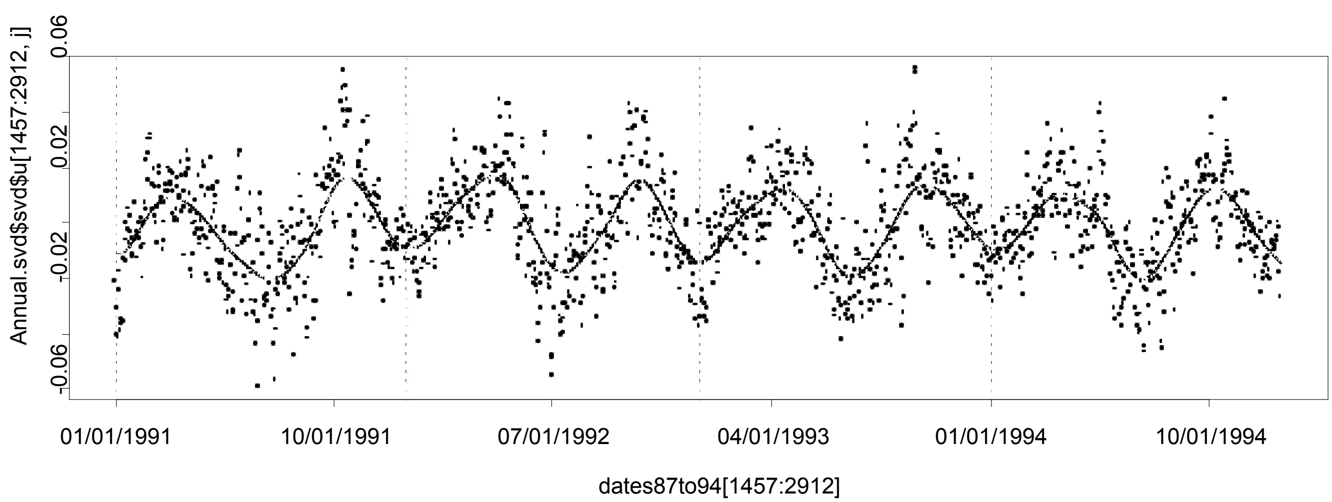
EOF 1



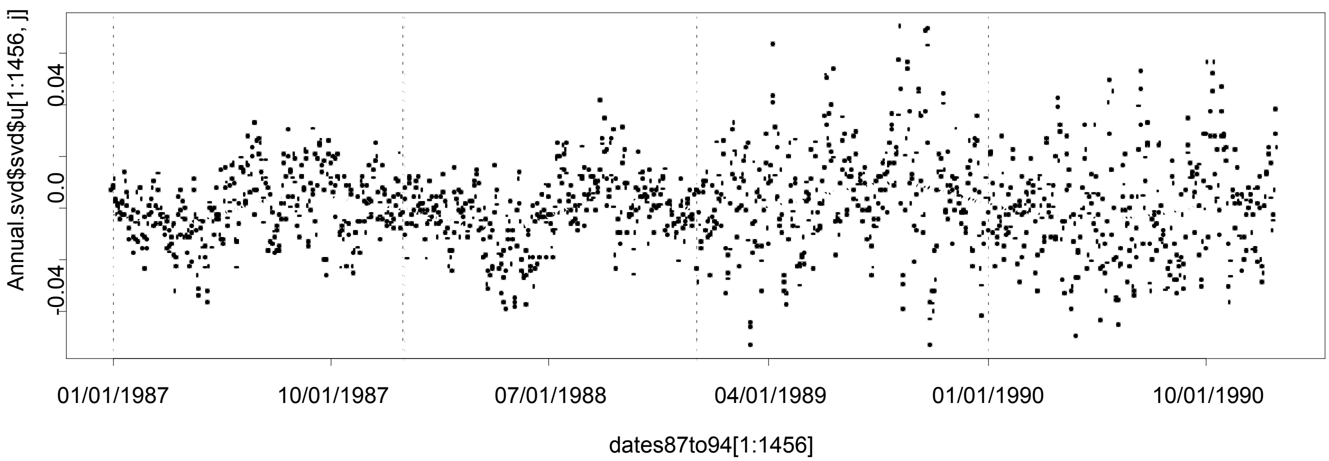
Annual.svdsvdu[1:1456, j]
0.04
0.02
0.0
-0.04
01/01/1987
10/01/1987
07/01/1988
04/01/1989
01/01/1990
10/01/1990
dates87to94[1:1456]
Annual.svdsvdu[1457:2912, j]
0.04
0.02
0.0
-0.02
01/01/1991
10/01/1991
07/01/1992
04/01/1993
01/01/1994
10/01/1994
dates87to94[1457:2912]

EOF 2

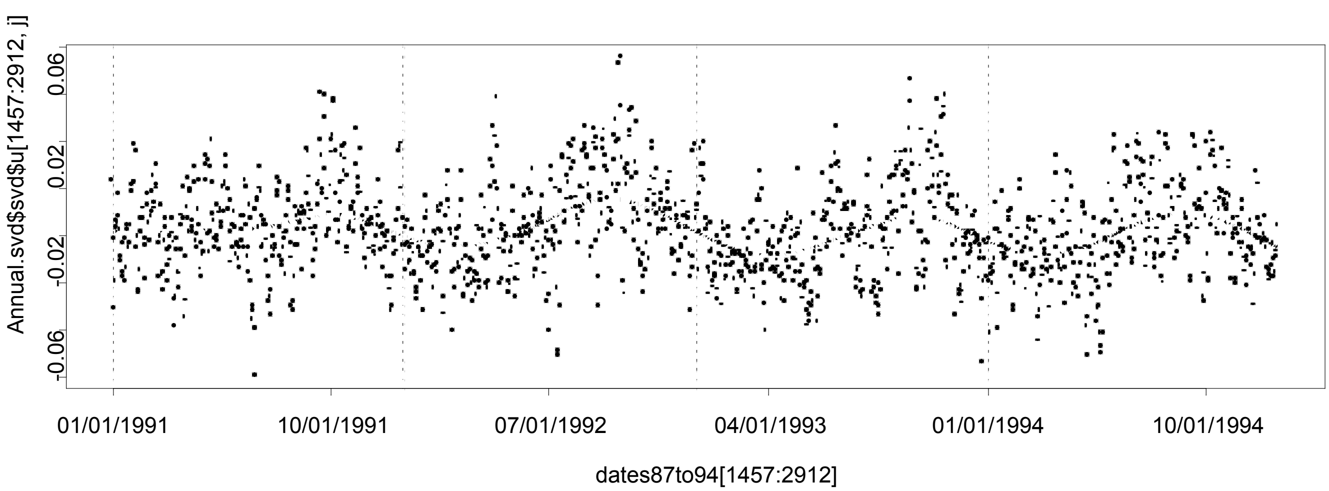

Annual.svdsvdu[1:1456, j]
0.06
0.02
-0.02
-0.06
01/01/1987
10/01/1987
07/01/1988
04/01/1989
01/01/1990
10/01/1990
dates87to94[1:1456]


Annual.svdsvdu[1457:2912, j]
0.06
0.02
-0.02
-0.06
01/01/1991
10/01/1991
07/01/1992
04/01/1993
01/01/1994
10/01/1994
dates87to94[1457:2912]

EOF 3

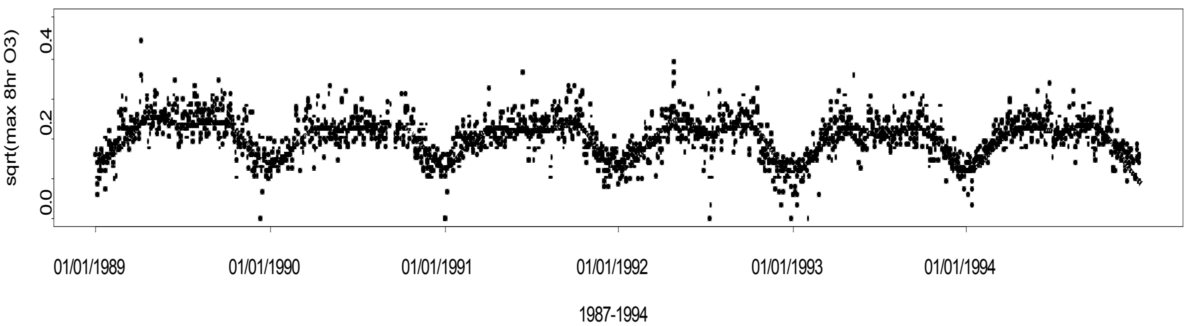


Annual.svdsvdu[1:1456, j]
0.04
0.0
-0.04
01/01/1987
10/01/1987
07/01/1988
04/01/1989
01/01/1990
10/01/1990
dates87to94[1:1456]

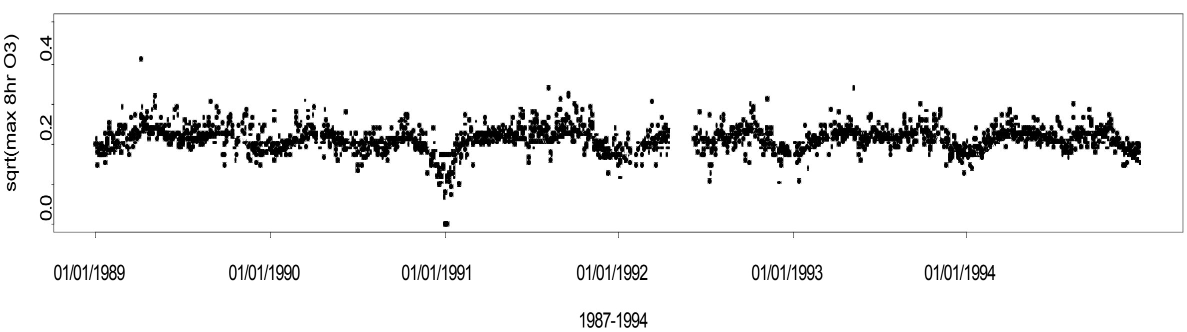


Annual.svdsvdu[1457:2912, j]
0.06
0.02
-0.02
-0.06
01/01/1991
10/01/1991
07/01/1992
04/01/1993
01/01/1994
10/01/1994
dates87to94[1457:2912]

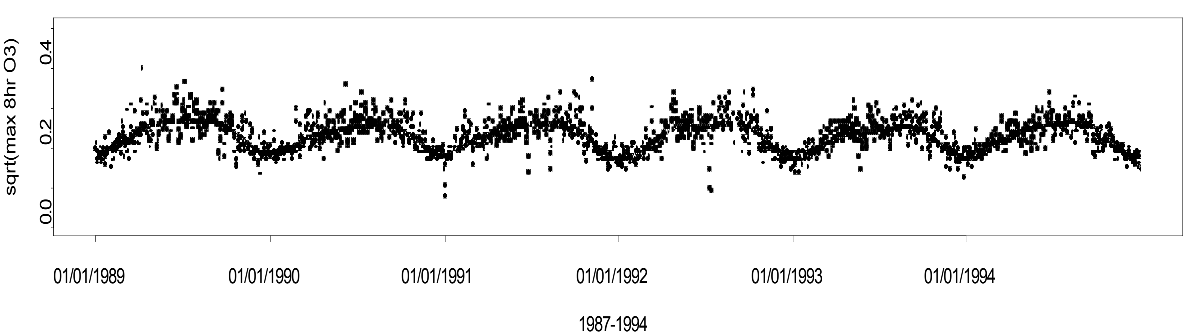
60370113



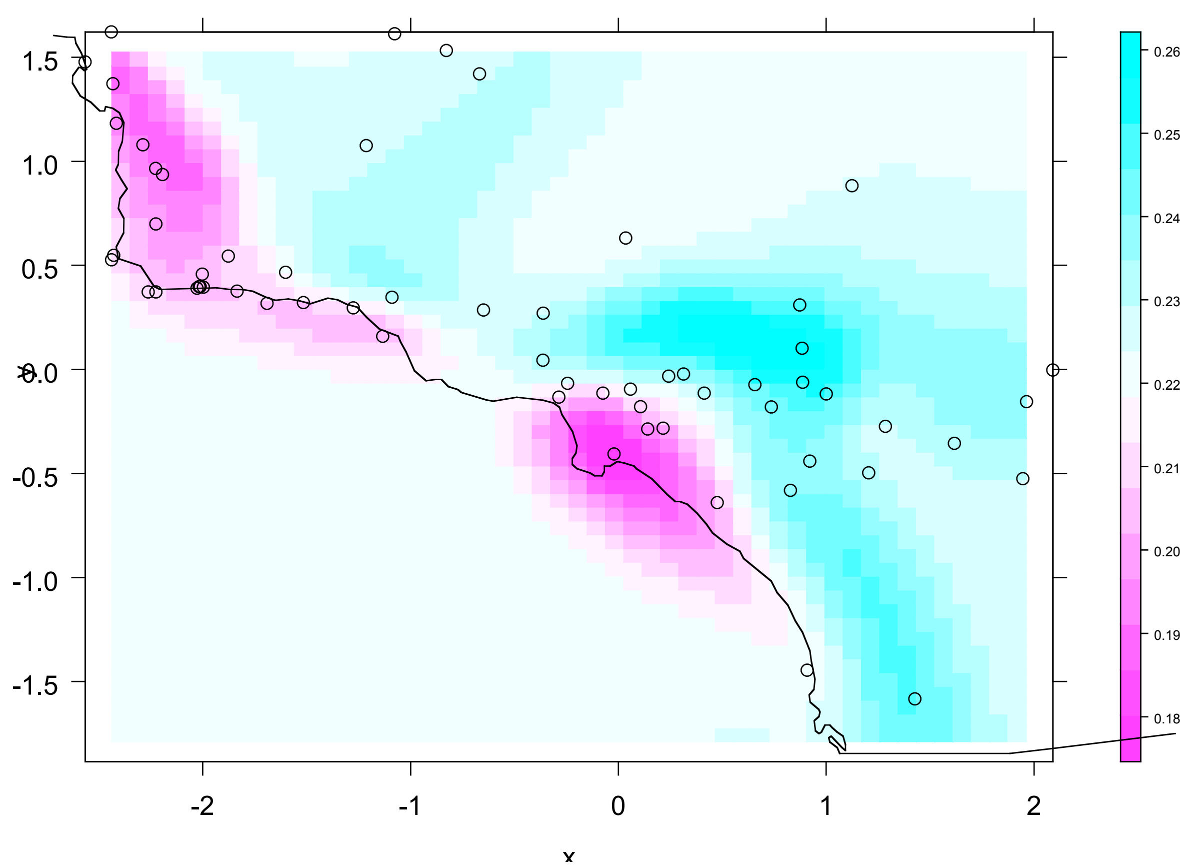
61112003



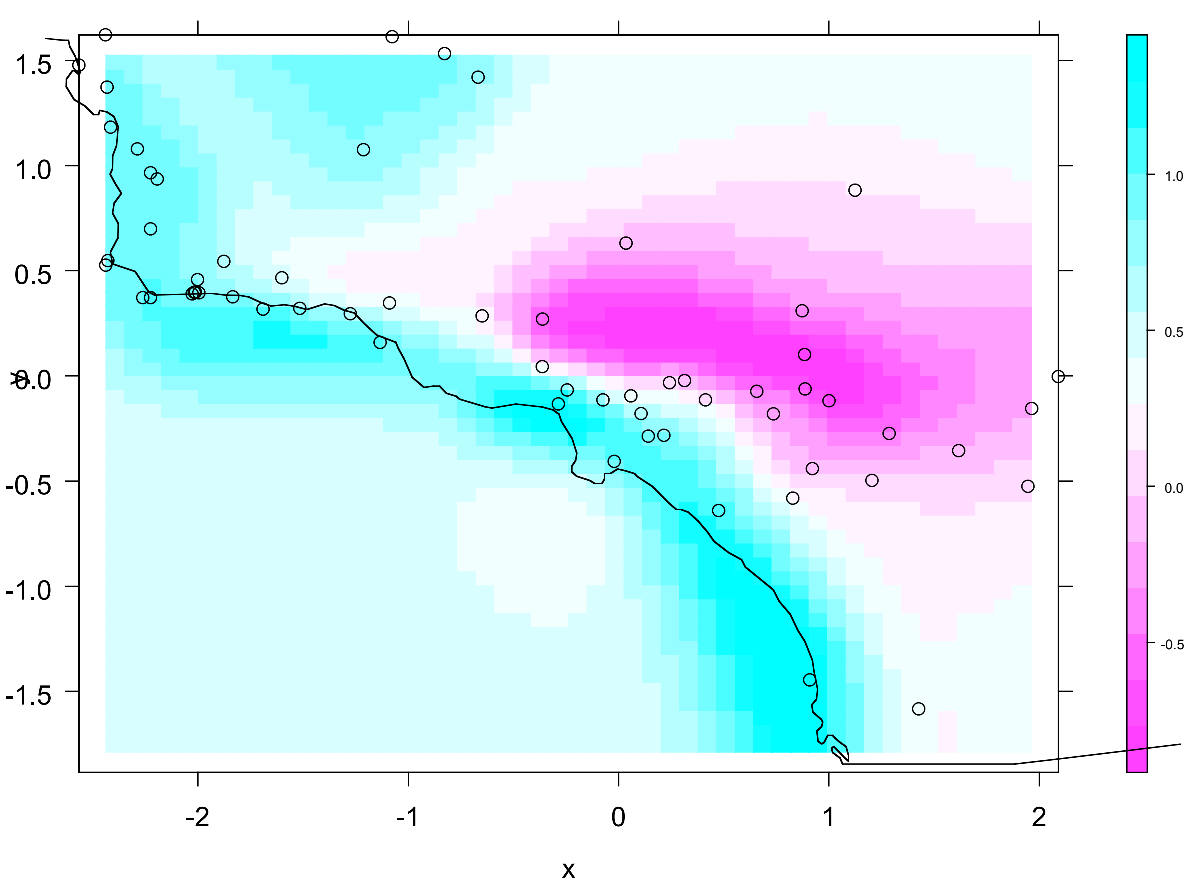
61111003



Kriging of μ_0

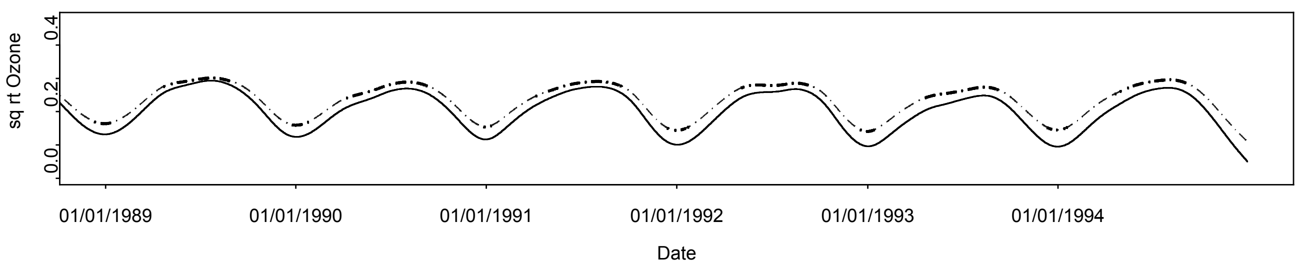


Kriging of ρ_2

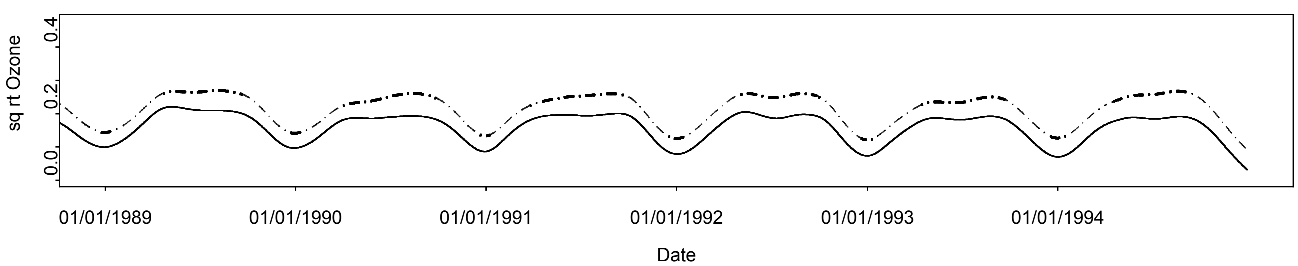


Quality of trend fits

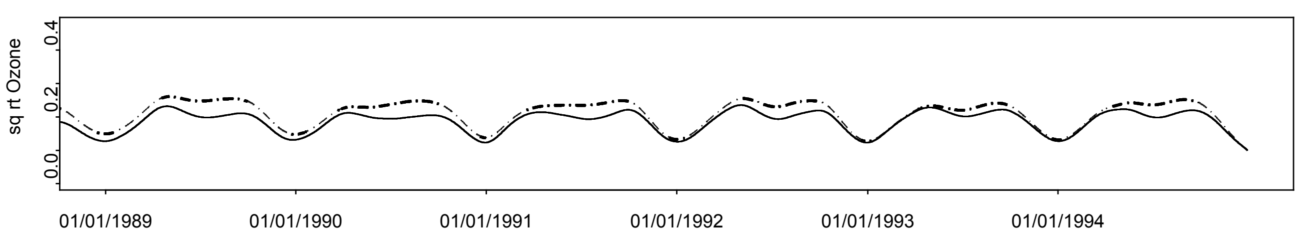
Fitted trend (solid) vs Predicted (dashed): 060371002



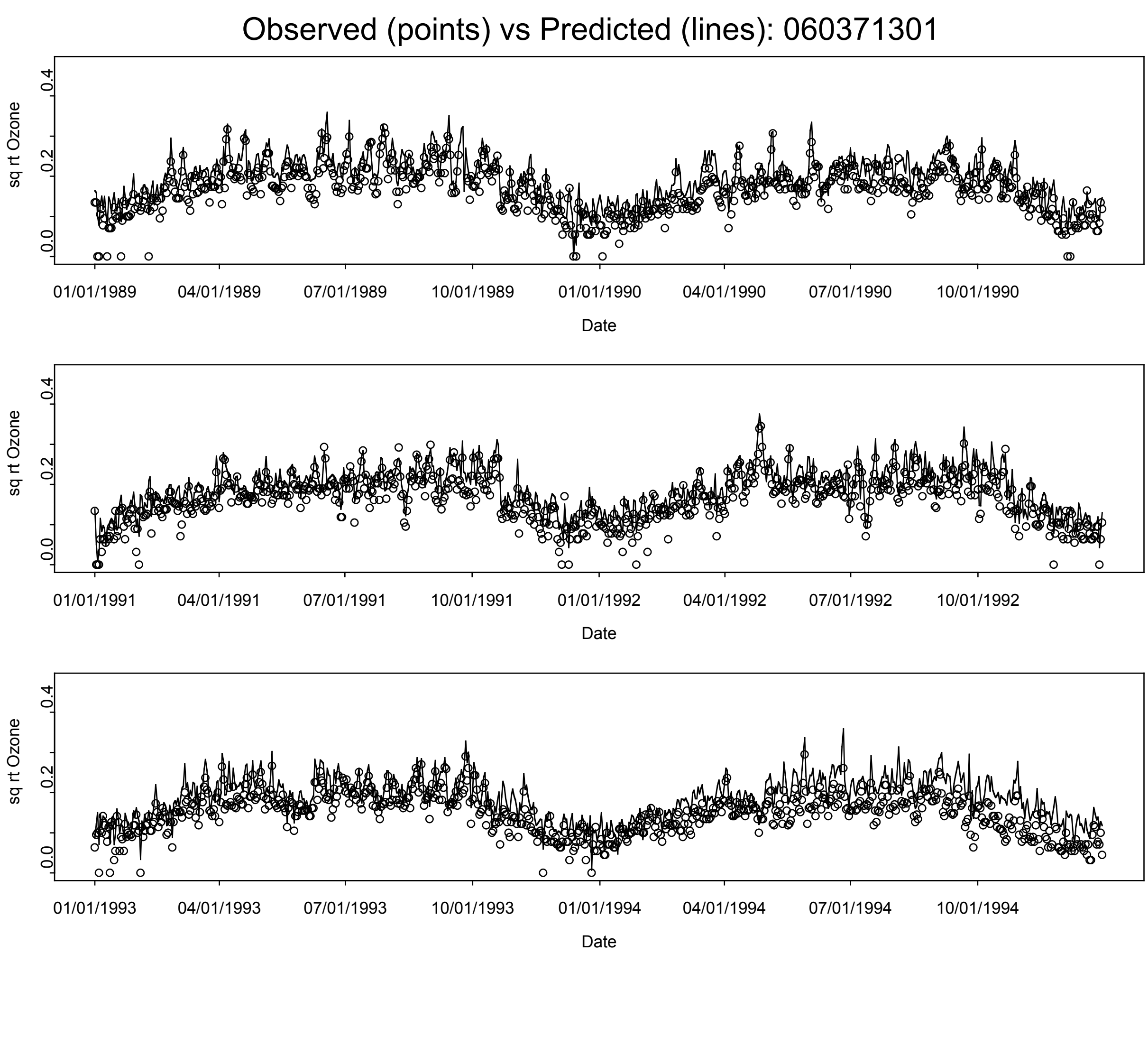
Fitted trend (solid) vs Predicted (dashed): 060371301



Fitted trend (solid) vs Predicted (dashed): 060375001



Observed vs. predicted



2. Air quality standards, extremes and climate

Peter Guttorp
University of Washington

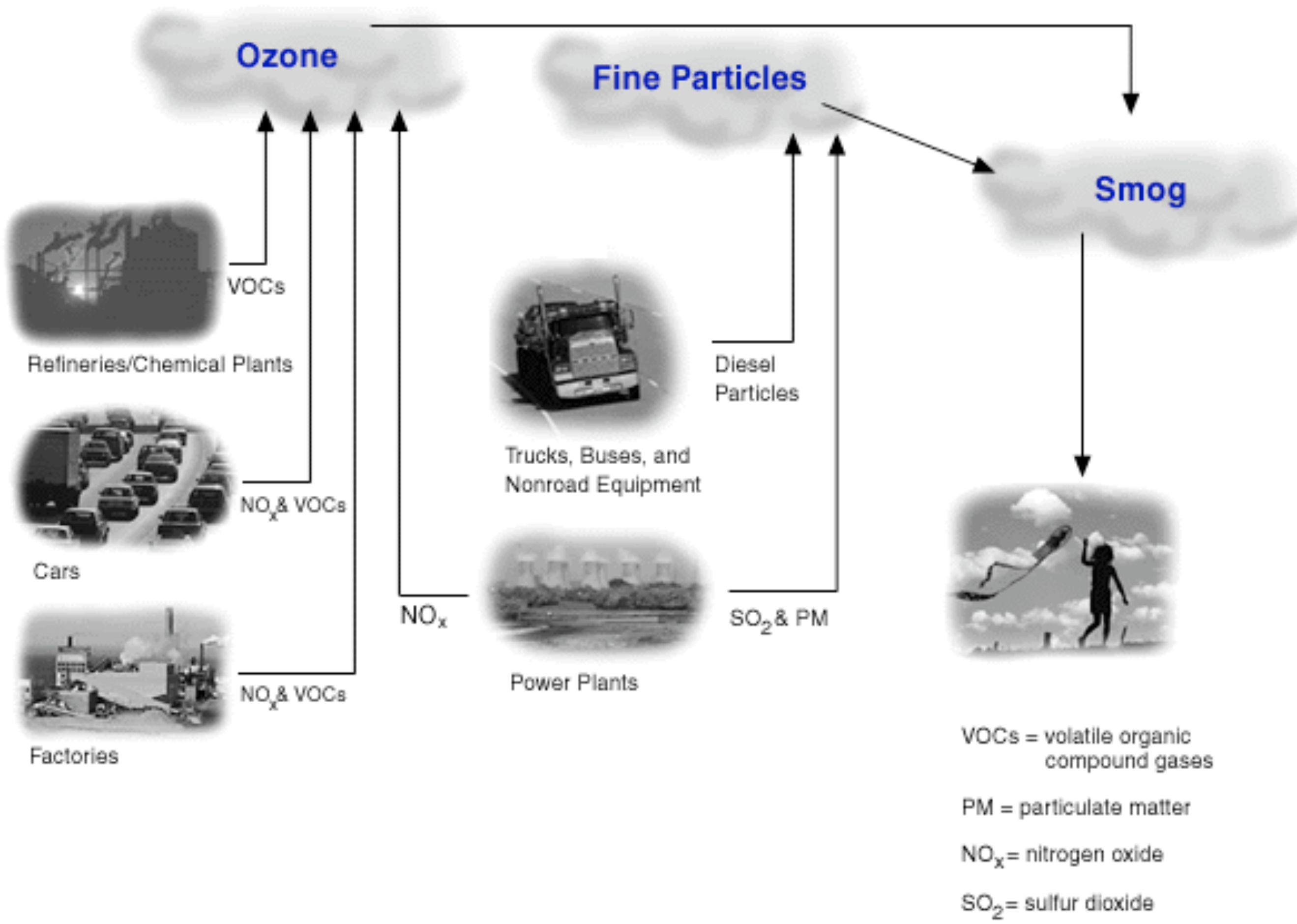
peter@stat.washington.edu
www.stat.washington.edu/peter

Acknowledgements

Joint work with
Sofia Åberg
David Caccia
Laura Knudsen
Nicola Loperfido
Paul Sampson
Mary Lou Thompson
Larry Cox
Anders Grimvall
Georg Lindgren
Lars Bärring
Erik Kjellström
Jan Heffernan

Outline

Health effects
Regulations
Implementation
Statistical quality considerations
Trends in extremes


Ozone
Fine Particles
Smog
VOCs
Refineries/Chemical Plants
NO_x & VOCs
Cars
NO_x & VOCs
Factories
Diesel Particles
Trucks, Buses, and Nonroad Equipment
NO_x
SO_2 & PM
Power Plants
VOCs = volatile organic compound gases
PM = particulate matter
NO_x = nitrogen oxide
SO_2 = sulfur dioxide

Health effect studies

Often opportunistic
Rarely yield clear cutoff values
Uncertainty associated with dose-response curve
What are important health outcomes for policy setting?

Network bias

- Many health effects studies use
  - air quality data from compliance networks
  - health outcome data from hospital records
- Compliance networks aim at finding large values of pollution
- Actual exposure may be lower than network values

A calculation

$$\begin{pmatrix} X_{1,t} \\ X_{1,t-1} \\ X_{2,t-1} \end{pmatrix} \sim N_3 \left[\begin{pmatrix} \mu_1 \\ \mu_1 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} 1 & \rho & \alpha\rho \\ \rho & 1 & \rho \\ \alpha\rho & \rho & 1 \end{pmatrix} \right]$$

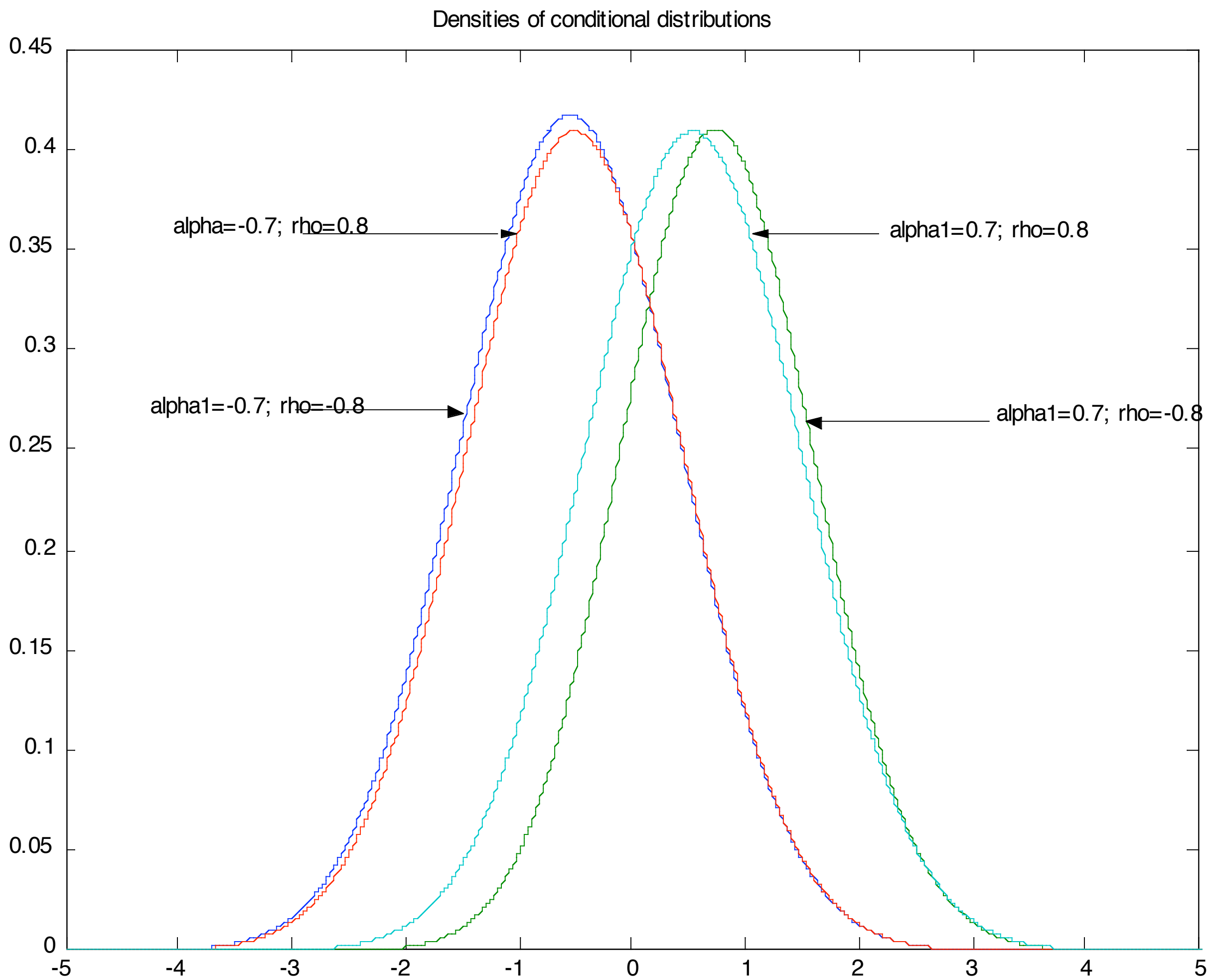
$$0 < |\rho| < 1 \qquad \frac{2\rho^2 - 1}{\rho} \le \alpha \le \frac{1}{\rho}$$

$$E\left(X_{1,t} \middle| X_{1,t-1} > X_{2,t-1}\right) = \mu_1 + \alpha\sqrt{\frac{1-\rho}{2}}\,\xi_1\left(\frac{\mu_1 - \mu_2}{\sqrt{2-2\rho}}\right)$$

$$\xi_1(t) = \frac{d}{dt}\log\Phi(t)$$

Special cases

Case	Bias
$\mu_1 >> \mu_2$	negligible
$\mu_1 = \mu_2$	$\approx \alpha\sqrt{(1-\rho)/\pi}$
$\mu_1 << \mu_2$	$\approx \alpha(\mu_2 - \mu_1)/2$


Densities of conditional distributions
alpha=-0.7; rho=0.8
alpha1=0.7; rho=0.8
alpha1=-0.7; rho=-0.8
alpha1=0.7; rho=-0.8
0.45
0.4
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
-5
-4
-3
-2
-1
0
1
2
3
4
5

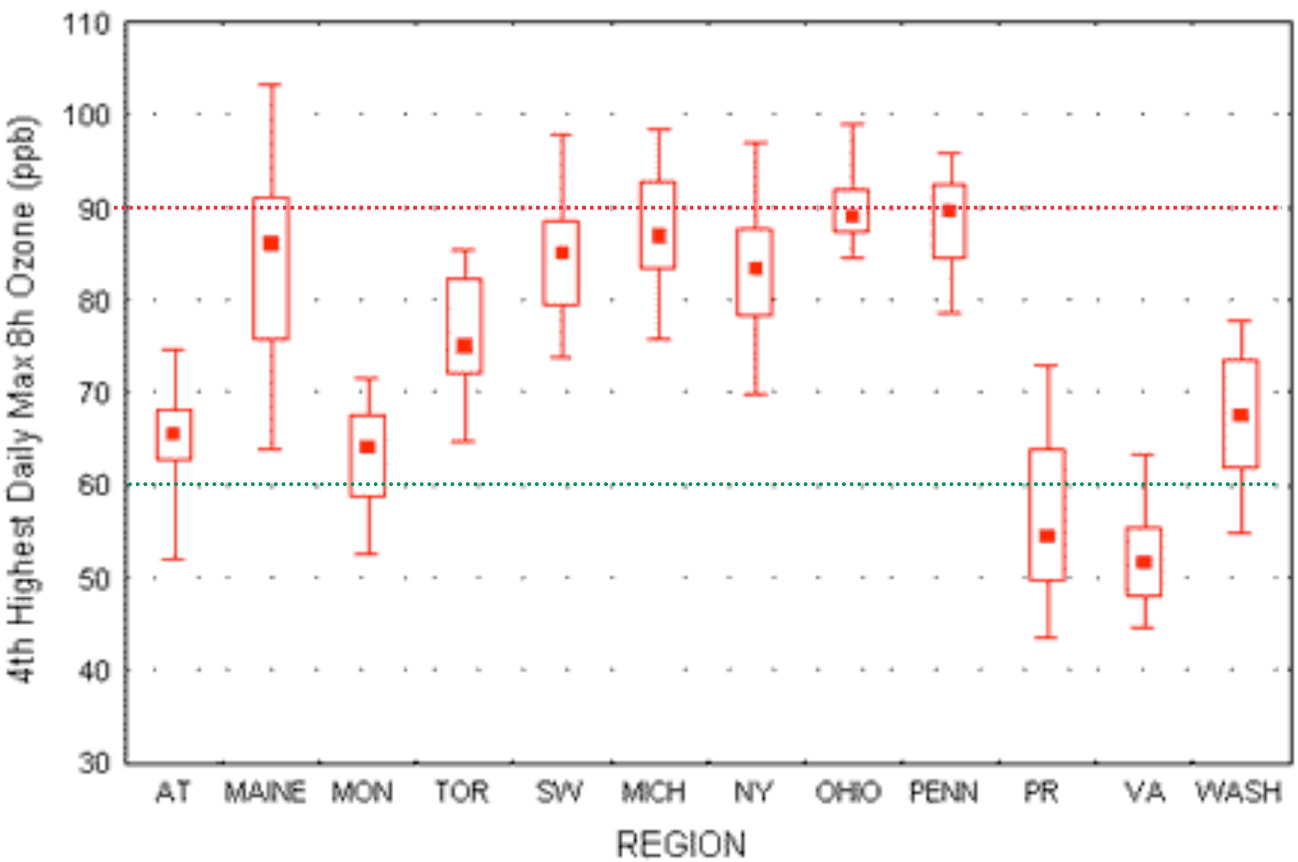
WHO health effects estimates for ozone

10% most sensitive healthy children get 5% reduction in lung capacity at .125 ppm hourly average

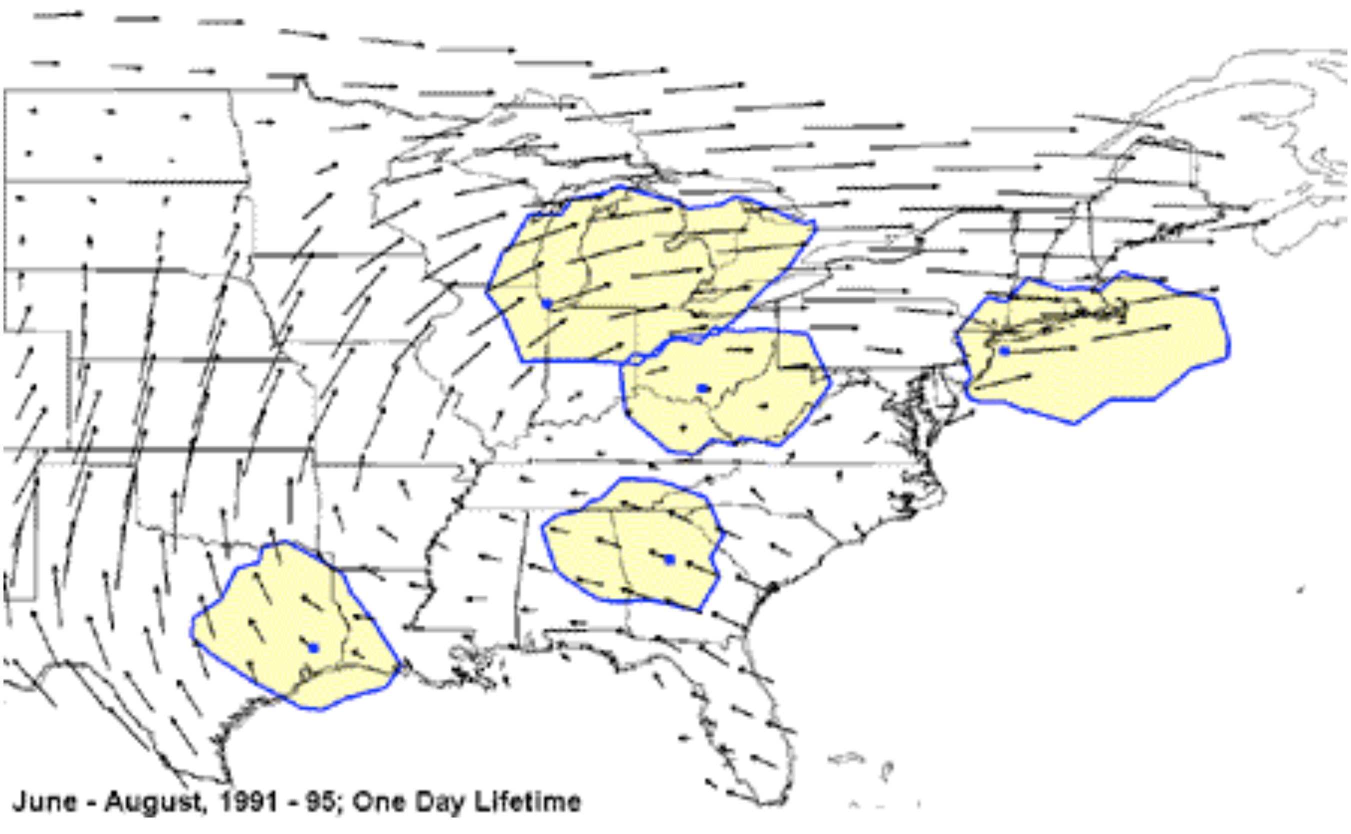
Double inflammatory response for healthy children at .09 ppm 8-hr average

Minimal public health effect at .06 ppm 8-hr average

North American ozone measurements 94-96



Transport wind vectors high regional O_3 days



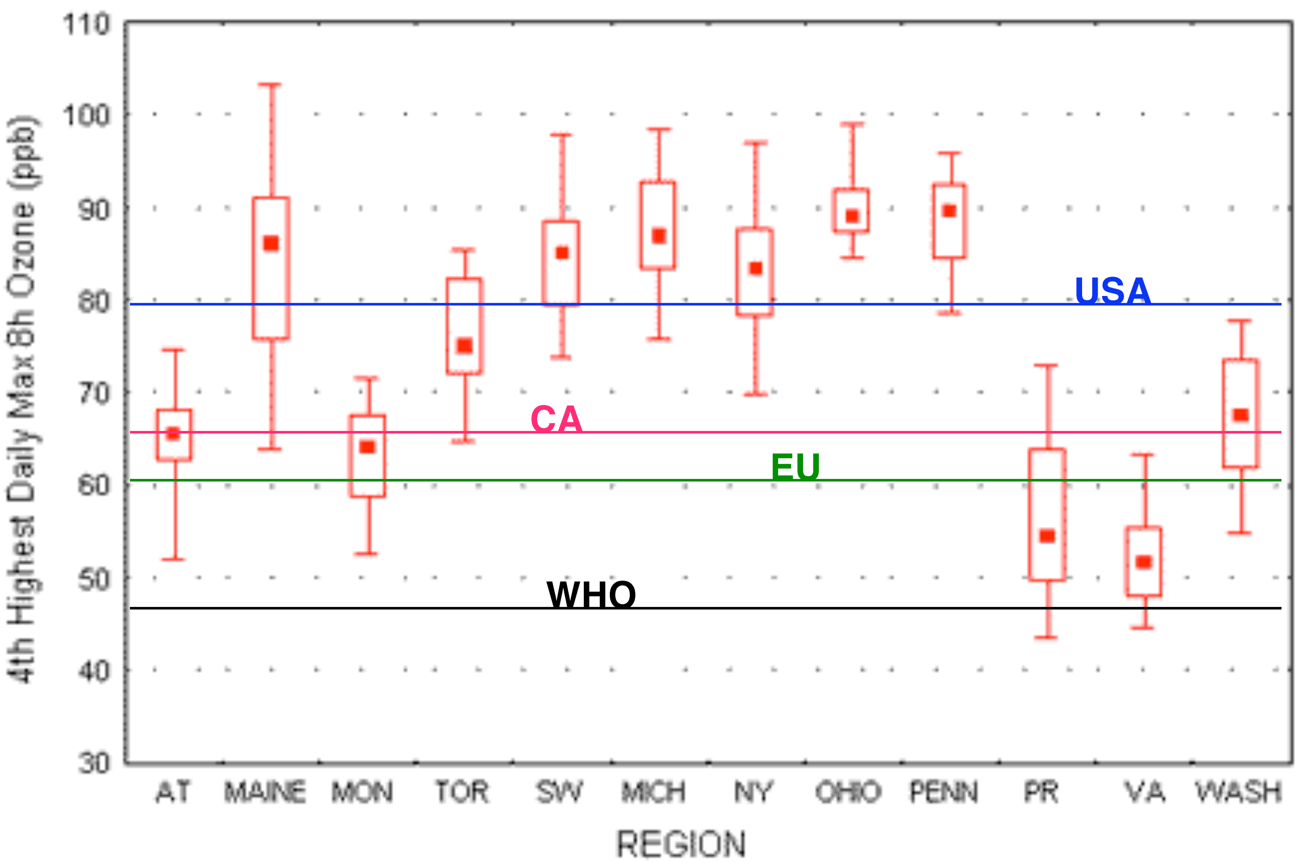
Task for authorities

- Translate health effects into limit values for standard
- Determine implementation rules for standard
- Devise strategies for ozone reduction
  - Need to limit emissions of primary pollutants in summertime

Some standards

	Ozone	$PM_{2.5}$
WHO	100 μg/m^3 (46.7 ppb)	25 μg/m^3
USA	80 ppb	35 μg/m^3
EU	60 ppb	50 μg/m^3
Canada	65 ppb	30 μg/m^3
	Max 8 hr average	24 hr ave

North American ozone measurements 94-96



US 1-hr ozone standard

In each region the expected number of daily maximum 1-hr ozone concentrations in excess of 0.12 ppm shall be no higher than one per year

Implementation: A region is in violation if 0.12 ppm is exceeded at any approved monitoring site in the region more than 3 times in 3 years

A hypothesis testing framework

The US EPA is required to protect human health. Hence the more serious error is to declare a region in compliance when it is not.

The correct null hypothesis therefore is that the region is *violating* the standard.

Optimal test

One station, observe

Y_3 = # exceedances in 3 years

Let $\theta = E(Y_1)$

H_0: $\theta > 1$ vs. H_A: $\theta \leq 1$

When $\theta = 1$, approximately

$Y_3 \sim Bin(3 \cdot 365, 1/365) \approx Po(3)$

and the best test rejects for small Y_3.

For $Y_3 = 0$ $\alpha = 0.05$.

In other words, no exceedances should be allowed.

How did the US EPA perform the test?

EPA wants $Y_3 \le 3$, so $\alpha = 0.647$

The argument is that $\theta \approx Y_3 / 3$

(Law of large numbers applied to n=3)

Using $Y_3 / 3$ as test statistic, equate the critical value to the boundary between the hypotheses (!).

This implementation of the standard does not offer adequate protection for the health of individuals.

An example

Let $\sqrt{Z_i} \sim N(\mu, \sigma^2)$. For Houston, TX, μ=0.235 (0.059 ppm) and σ=0.064.

The station exceeds 0.12 ppm with probability 0.041, for an expected number of exceedances of 15 (18 were observed in 1999)

At level 0.18 ppm (severe violation) the exceedance probability is 0.0016, corresponding to 0.6 violations per year (1 observed in 1999)

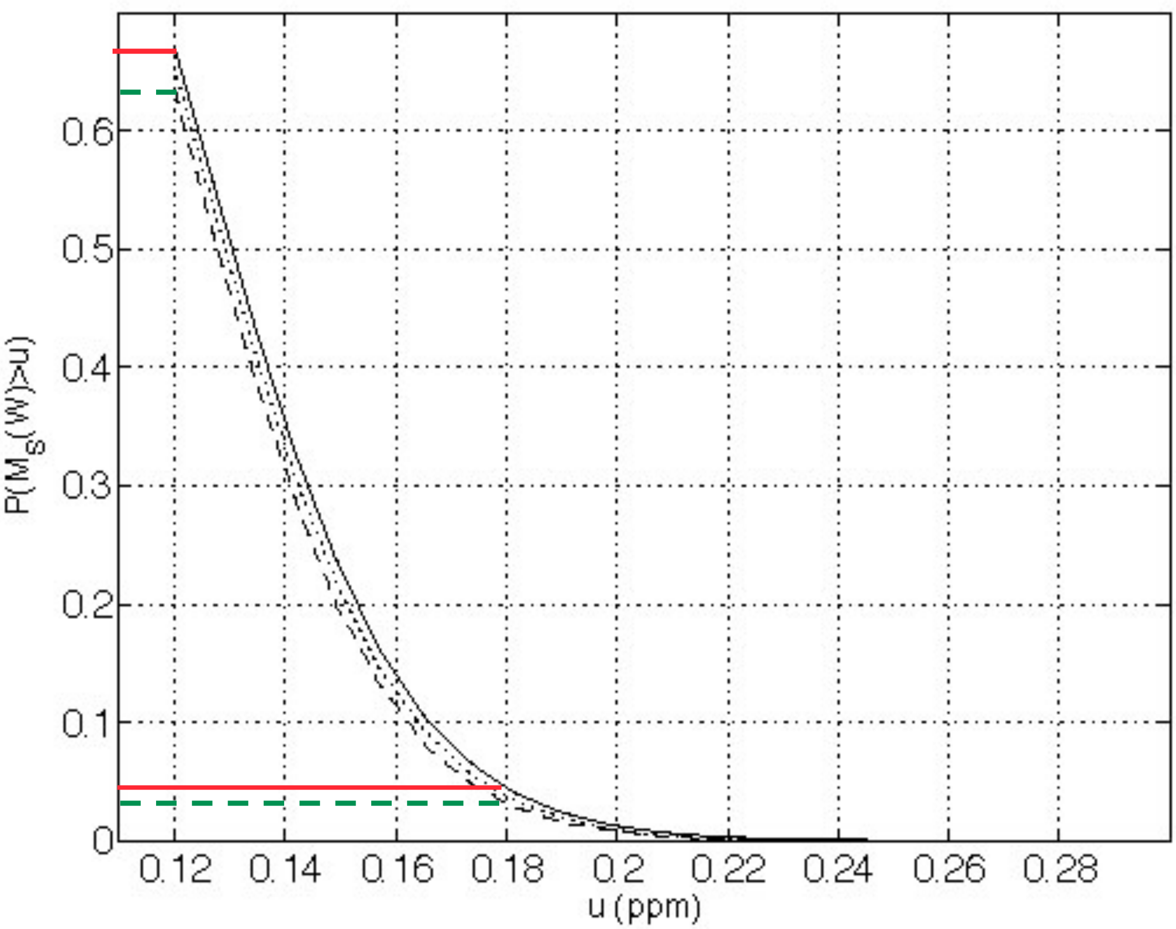
In order to have an exceedance probability of 1/365=.0027 we need the mean reduced to 0.182 (0.033 ppm)

A conditional calculation

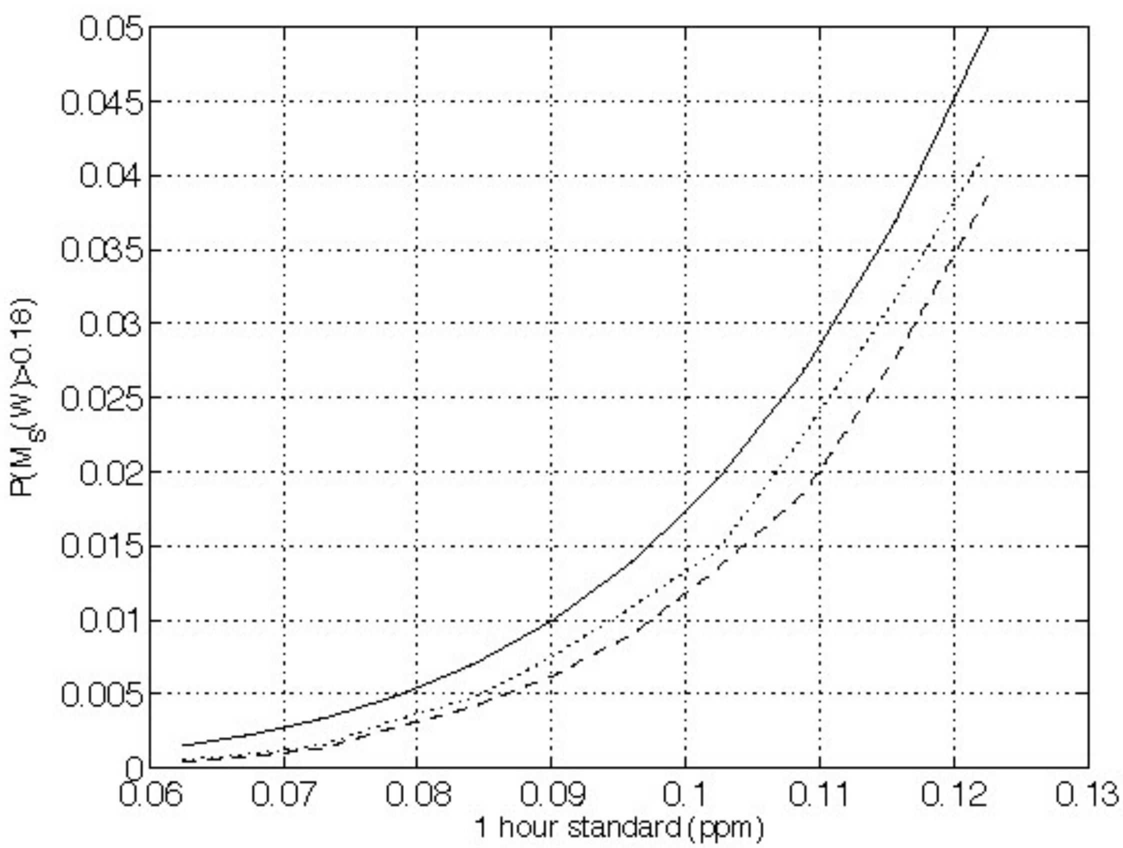
Given an observation of .120 ppm in the Houston region, what is the probability that an individual in the region is subjected to more that .120 ppm?

Need to calculate maximum of Gaussian process (after transformation) over a region that is highly correlated with measurement site, taking into account measurement error.

Probability of exceeding level u



Level of standard to protect against 0.18 ppm



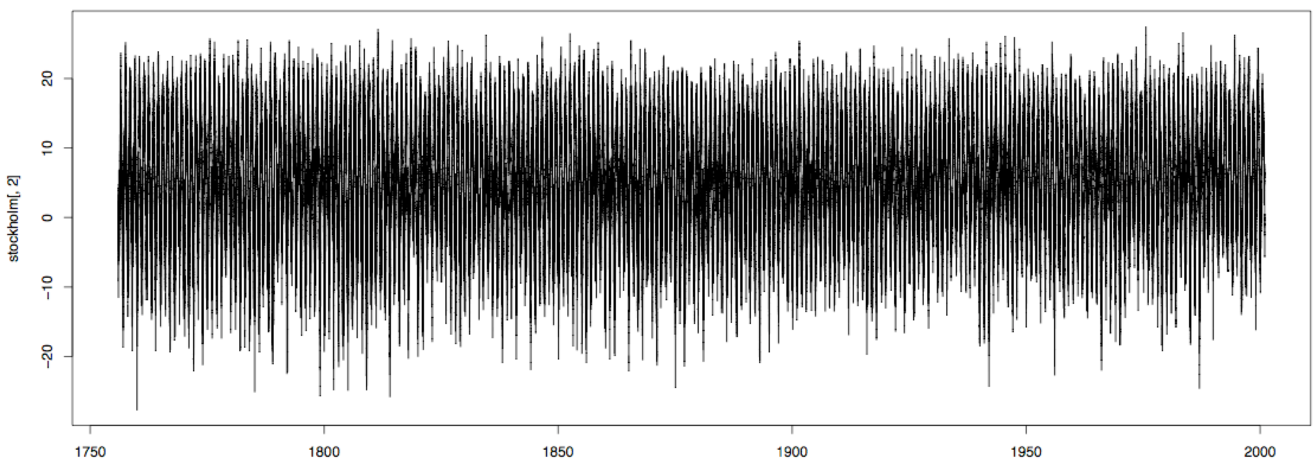
General setup

Given measurements $X(s_i, t_j)$ of a Gaussian field $\xi(s,t)$ observed with error, find $c_{[t]}$ such that

$$P(\sup_{\{v:\rho(u,v)\geq\beta\}} \xi(u,t) > c_{[t]}) \leq \alpha$$

where [t] denotes season and the mean of $\xi(u,t)$ equals the γ-quantile of the estimated health effects distribution.

Stockholm daily temperatures 1756-2000



Is there a trend?

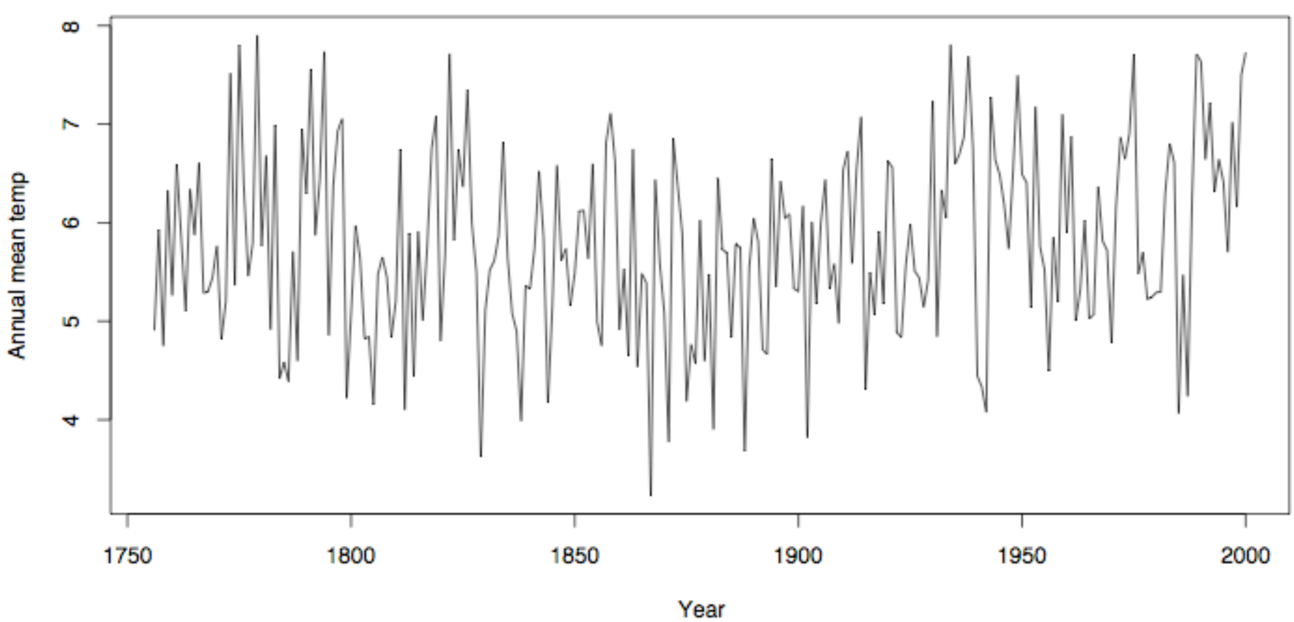
What does climate models predict?

Increasing global mean annual temperature

Decreasing annual temperature range

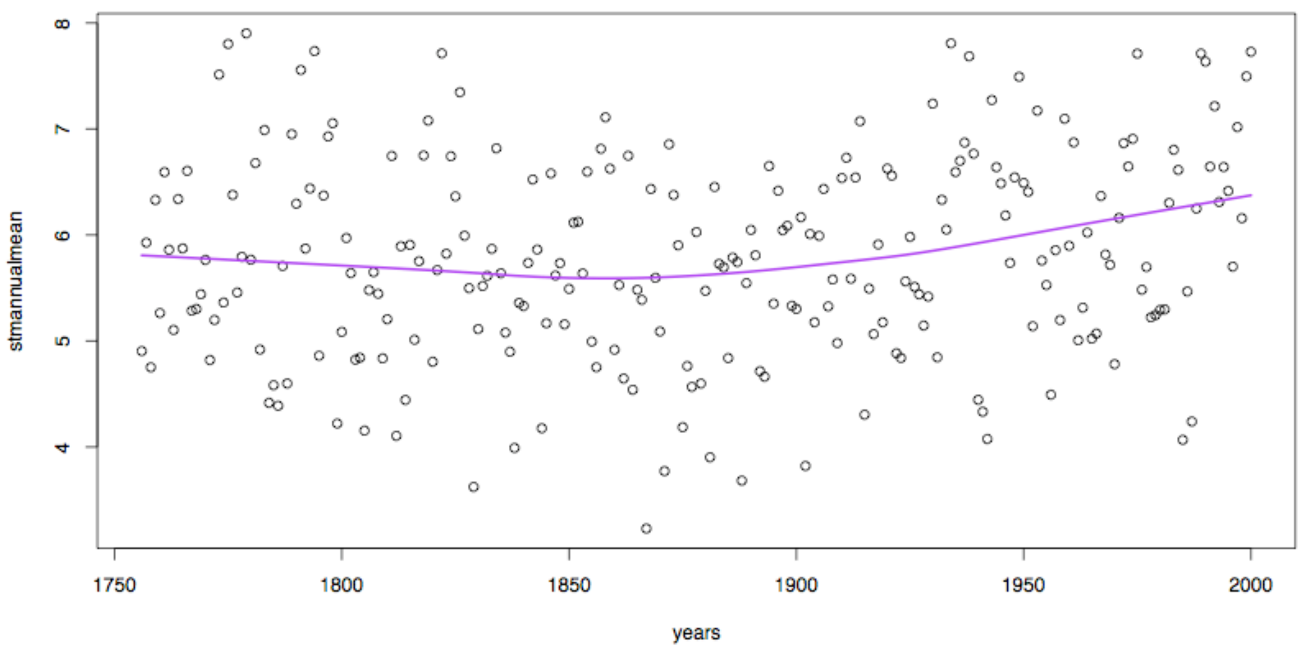
Increasing minimum temperatures

Looking at annual averages

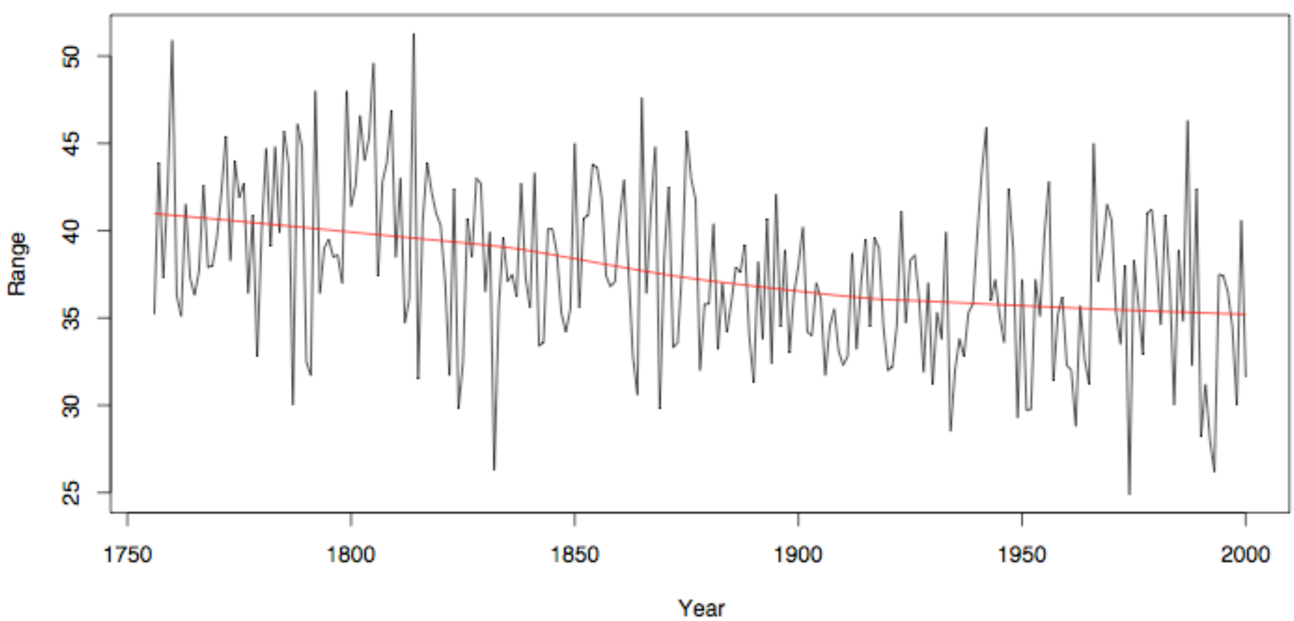


Is there a trend now?

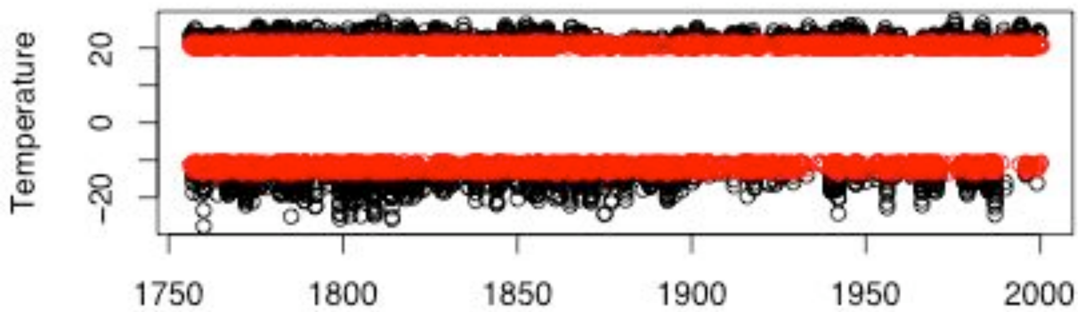
Trendline

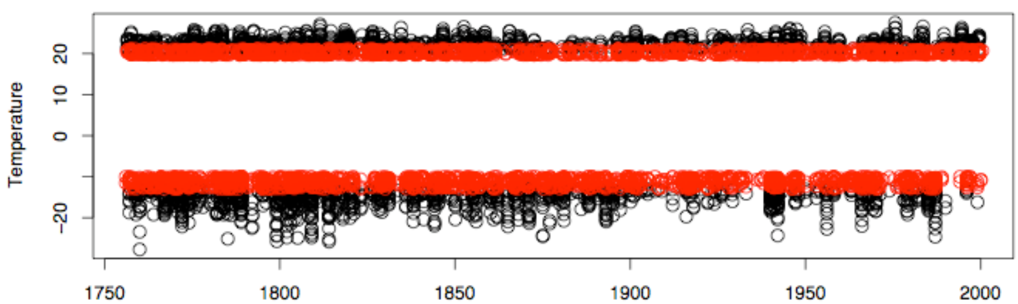


What about the range?



Are the extremes changing?





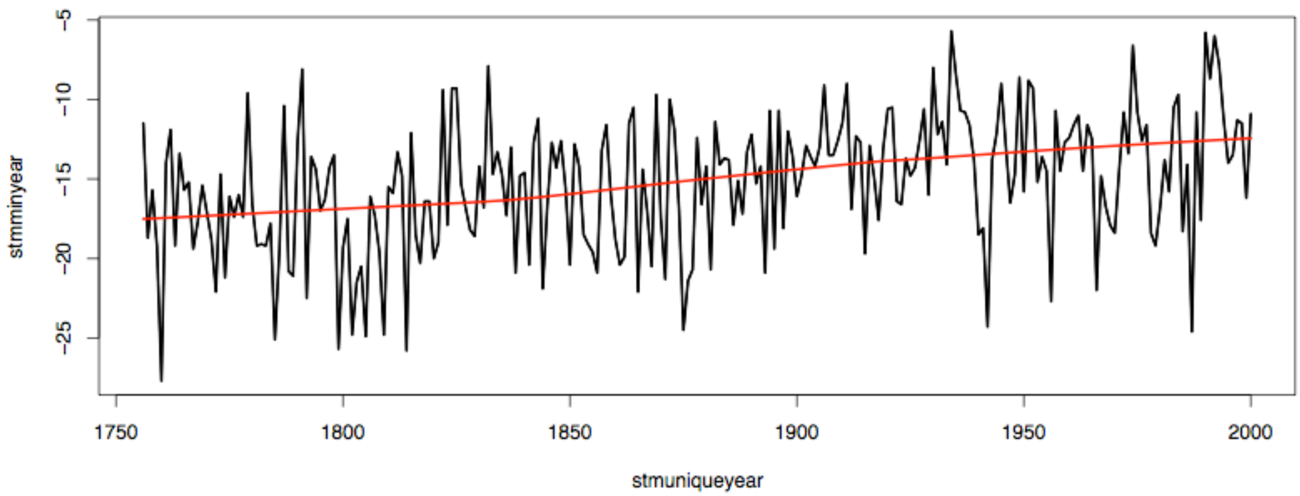
It matters what you base the quantiles on:

(min ,1%,2.5%.97.5%,99%,max)

all data (-27.7,-13.5,-10.5,20.0,21.6,27.5)

late data (-24.6,-13.0,-10.0,19.7,21.2,27.5)

Annual minimum



Is the trend due to climate change?

Multiple variables

Extreme in one, not extreme in others?

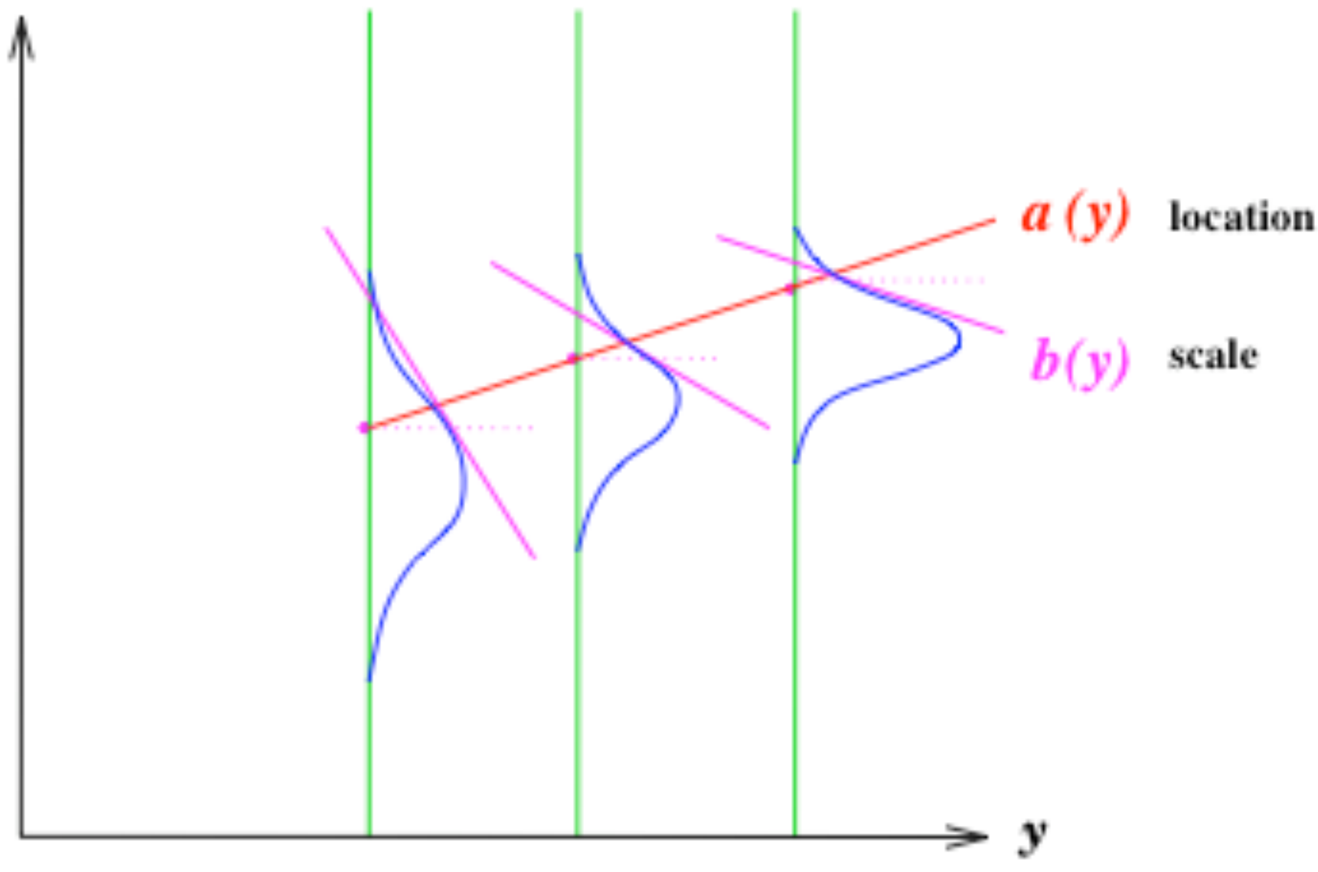
Interesting scenario:

Medium temperature, about 0C

Large snowfall

Extreme winds

What do we mean by trends in extreme values?



3. Modeling compositional data

Peter Guttorp
University of Washington

peter@stat.washington.edu
www.stat.washington.edu/peter

Outline

Compositional data

An algebra for compositions

Examples:

- **air quality**
- **ecology**
- **water quality**

Background

NAPAP, 1980's

Workshop on biological monitoring, 1986

Dirichlet process: Gary Grunwald, 1987

Current framework: Dean Billheimer, 1995

Other co-workers: Adrian Raftery, Mariabeth Silkey, Eun-Sug Park

Compositional data

Vector of proportions

$$\mathbf{z} = (z_1, \ldots, z_k)^T \quad z_i > 0 \quad \sum_1^k z_i = 1 \quad \mathbf{z} \in \nabla^{k-1}$$

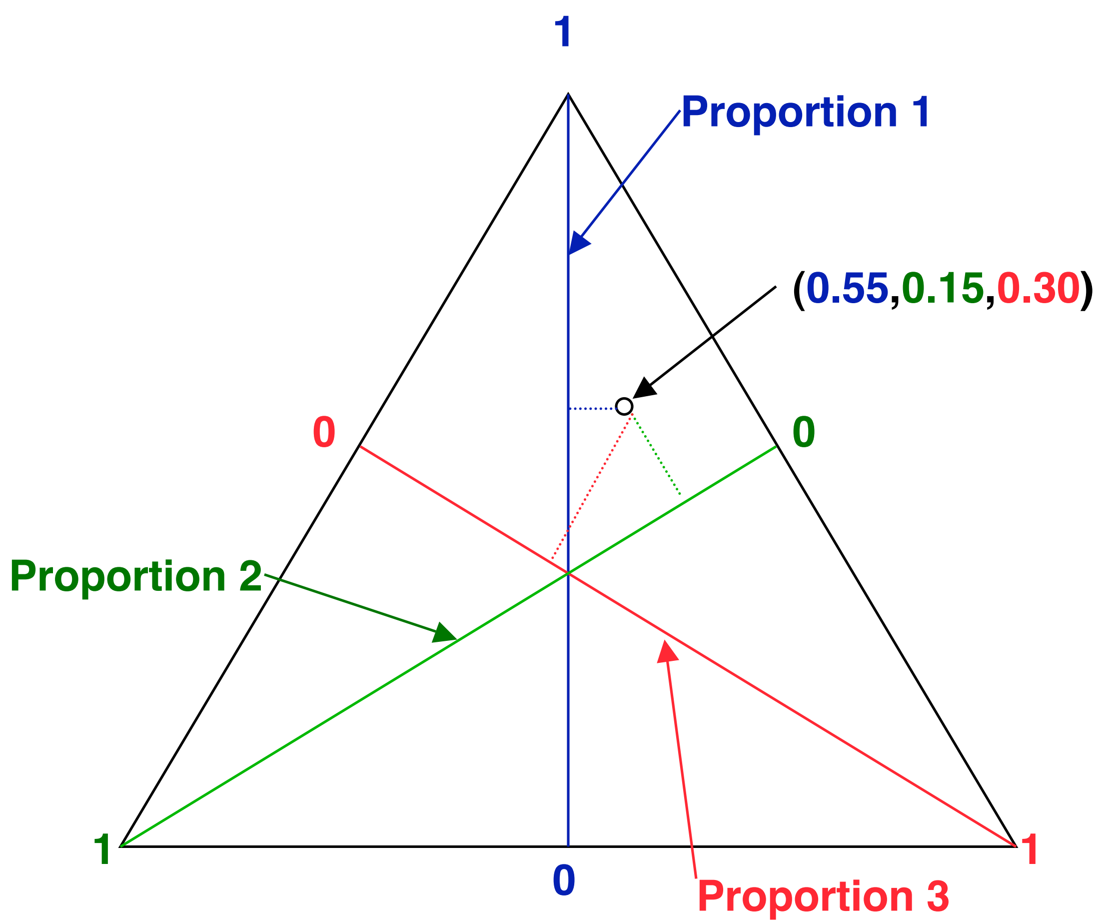
Proportion of taxes in different categories

Composition of rock samples

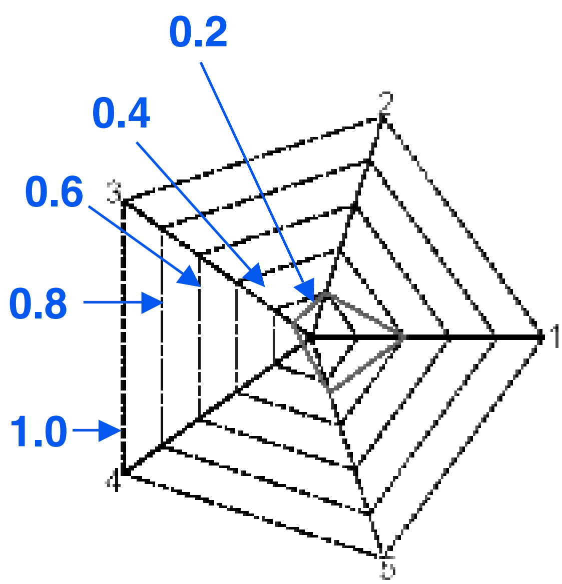
Composition of biological populations

Composition of air pollution

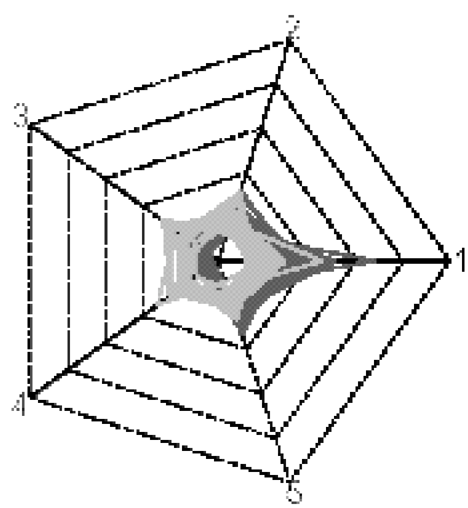
The triangle plot



The spider plot



(0.40,0.20,0.10,0.05,0.25)



An algebra for compositions

Perturbation: For $\xi, \alpha \in \nabla^{k-1}$ define

$$\xi \oplus \alpha = \left(\frac{\xi_1 \alpha_1}{\sum_1^k \xi_i \alpha_i}, \ldots, \frac{\xi_k \alpha_k}{\sum_1^k \xi_i \alpha_i} \right) \in \nabla^{k-1}$$

The composition $\iota = \left(\frac{1}{k}, \ldots, \frac{1}{k} \right)$ acts as a zero, so $\xi \oplus \iota = \xi$.

Set $\xi^{-1} = \left(\frac{1}{\xi_1}, \ldots, \frac{1}{\xi_k} \right)$ so $\xi \oplus \xi^{-1} = \iota$.

Finally define $\xi - \eta = \xi \oplus \eta^{-1}$.

The logistic normal

If $\text{alr}(z) = \left(\log\frac{z_1}{z_k}, \ldots, \log\frac{z_{k-1}}{z_k} \right)^T \sim MVN(\mu, \Sigma)$

we say that z is logistic normal, in short Z ~ LN(μ,Σ).

Other distributions on the simplex:

Dirichlet — ratios of independent gammas

“Danish” — ratios of independent inverse Gaussian

Both have very limited correlation structure.

Scalar multiplication

Let a be a scalar. Define

$$\xi \otimes a = \left(\frac{\xi_1^a}{\sum \xi_i^a}, \ldots, \frac{\xi_k^a}{\sum \xi_i^a} \right)$$

$\left(\nabla^{k-1}, \oplus, \otimes\right)$ is a complete inner product space, with inner product given, e.g., by

$$\langle \xi, \eta \rangle = alr(\xi)^T N^{-1} alr(\eta)$$

N is the multinomial covariance $N = I + jj^T$

j is a vector of k-1 ones.

$\|\xi\| = \langle \xi, \xi \rangle$ is a norm on the simplex.

The inner product and norm are invariant to permutations of the components of the composition.

Some models

Measurement error:

$z_j = \xi \oplus \varepsilon_j$ **where** $\varepsilon_j \sim LN(0,\Sigma)$.

Regression:

$\xi_j = \xi \oplus \gamma \otimes u_j$ ← **centered covariate**

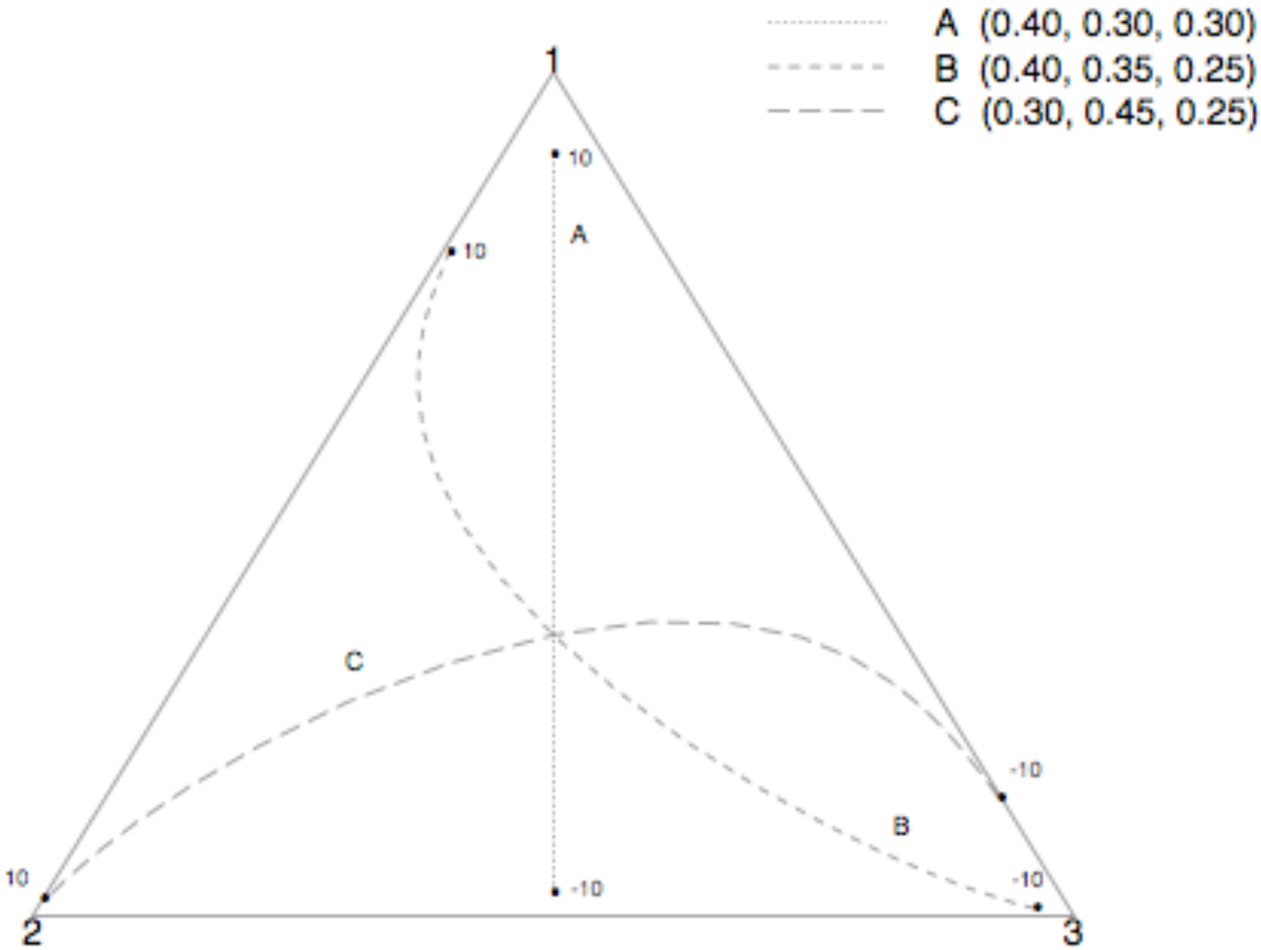
compositions (ξ_j, ξ, γ)

Correspondence in Euclidean space:

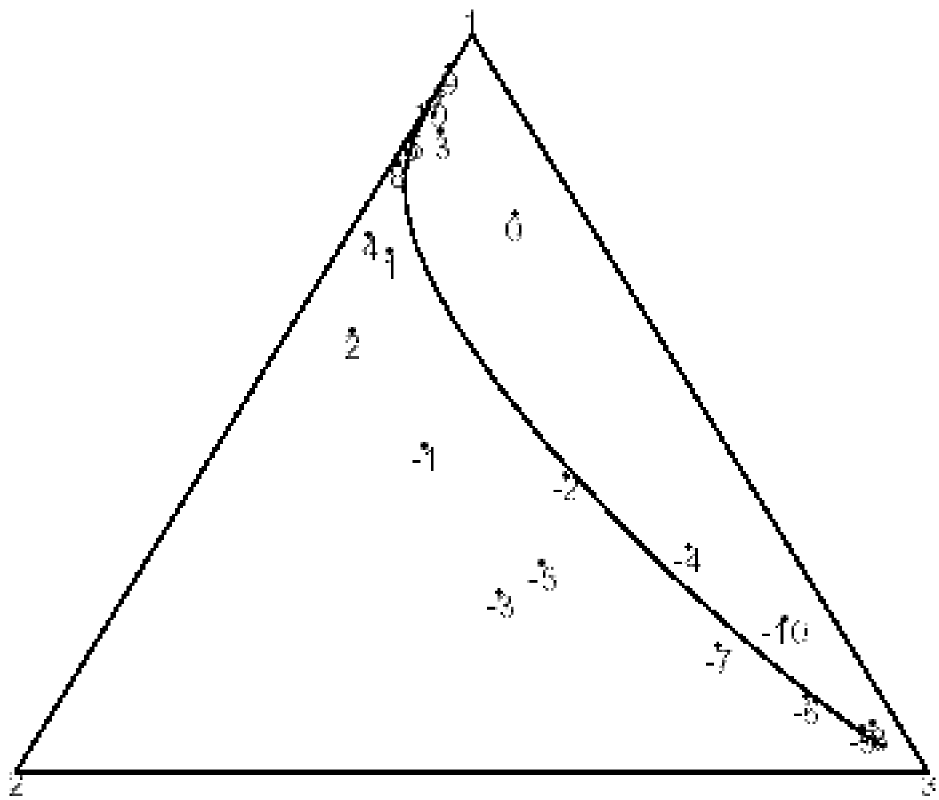
$$\mu_j \quad = \quad \beta_0 \quad + \quad \beta_1 \quad (x_j - \bar{x})$$

$$\underbrace{\mathrm{alr}^{-1}(\mu_j)}_{\xi_j} = \underbrace{\mathrm{alr}^{-1}(\beta_0)}_{\xi} \oplus \underbrace{\mathrm{alr}^{-1}(\beta_1)}_{\gamma} \otimes \underbrace{(x_j - \bar{x})}_{u_j}$$

Some regression lines



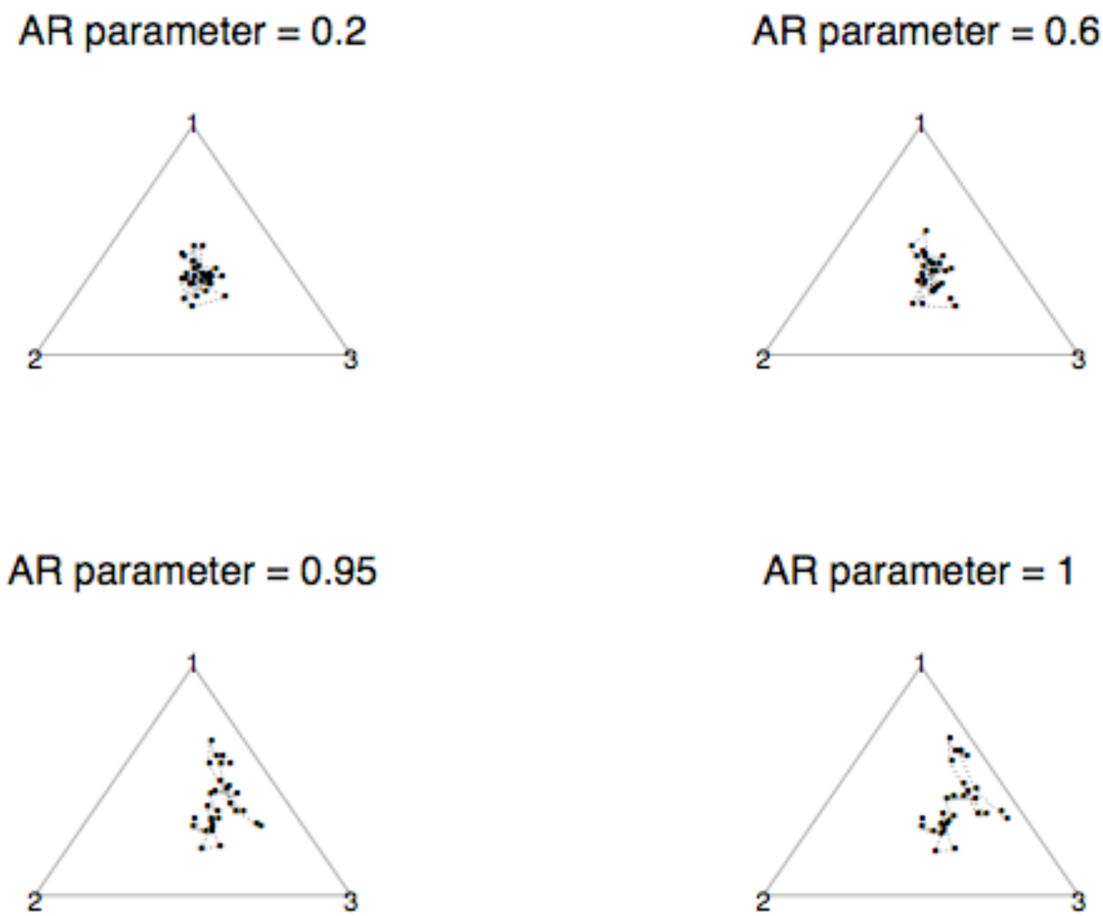
A regression example



γ=(0.40,0.35,0.25)

Time series (AR 1)

$$\mathbf{z}_{k+1} = \phi \otimes \mathbf{z}_k \oplus \varepsilon_k$$



A source receptor model

Observe relative concentration Y_i of k species at a location over time.

Consider p sources with chemical profiles θ_j. Let α_i be the vector of mixing proportions of the different sources at the receptor on day i.

$$EY_i = \sum_{i=1}^{p} \alpha_{ij}\theta_j = \Theta\alpha_i$$

$$Y = \Theta\alpha_i \oplus \varepsilon_i$$

Θ ~ LN, α_i ~ indep LN, ε_i ~ zero mean LN

Juneau air quality

50 observations of relative mass of 5 chemical species. Goal: determine the contribution of wood smoke to local pollution load.

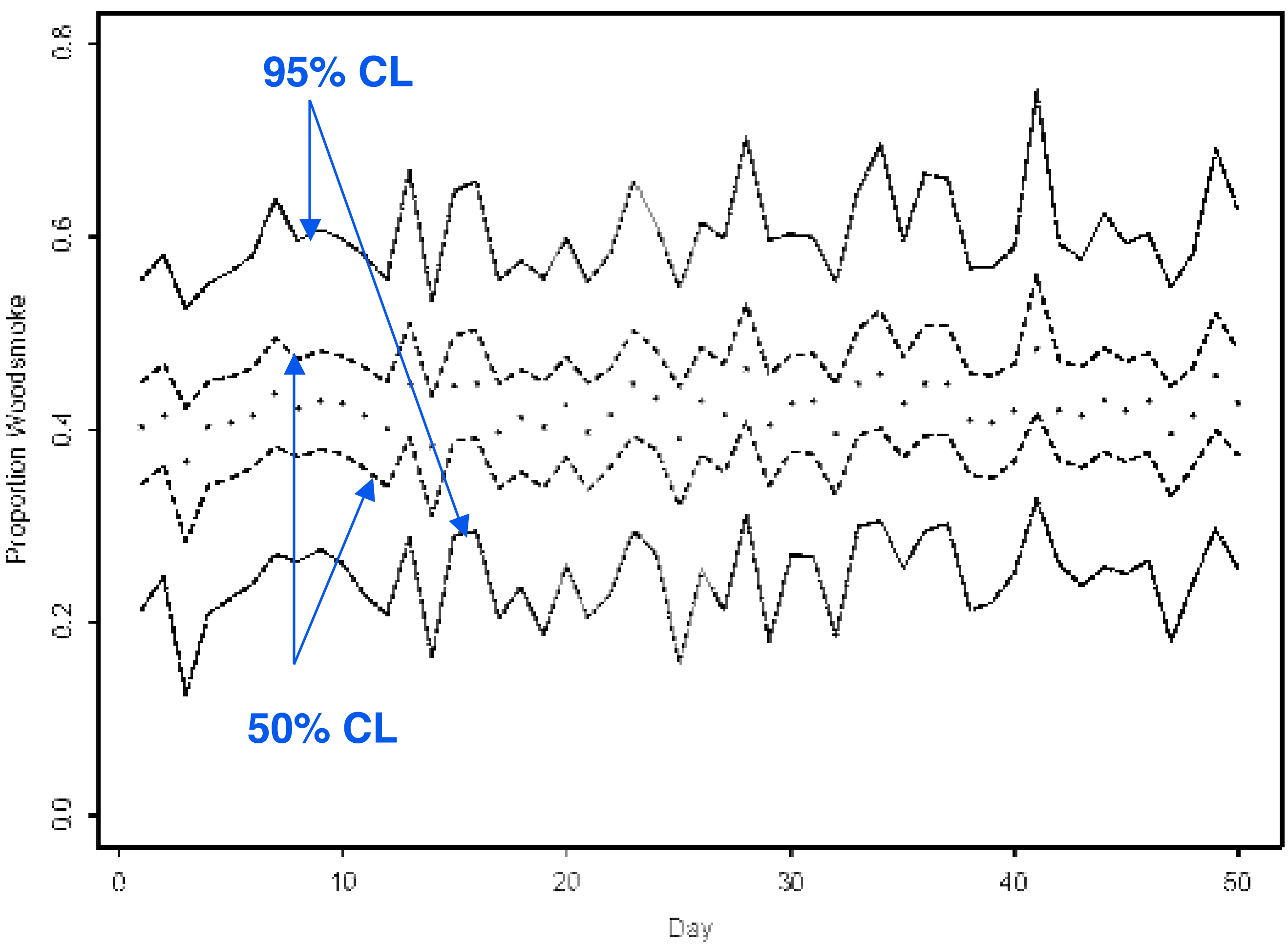
Prior specification:

$$f(\Theta, \alpha_i, \varepsilon_i, \mu_\alpha, \Gamma, \Sigma_\varepsilon) =$$

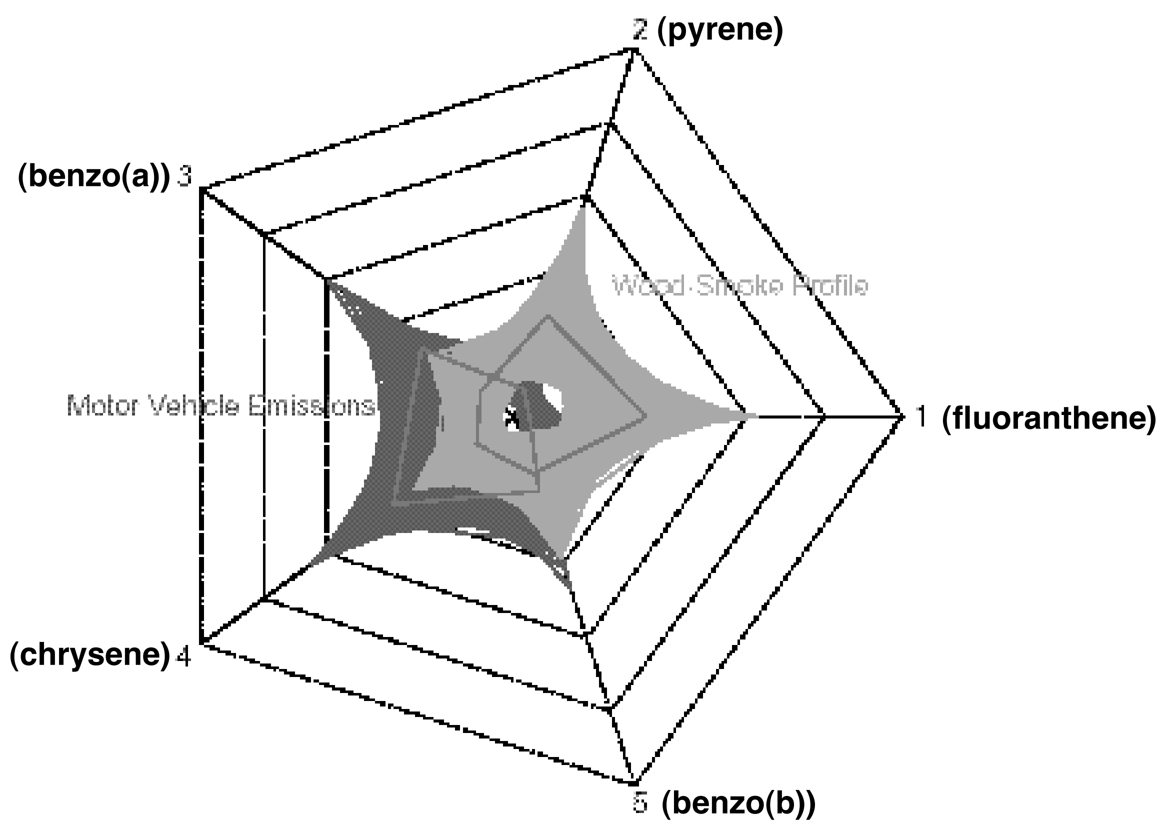
$$f(\alpha_i \mid \mu_\alpha, \Gamma) f(\varepsilon_i \mid \Sigma_\varepsilon) f(\mu_\alpha) f(\Gamma) f(\Sigma_\varepsilon)$$

Inference by MCMC.

Wood smoke contribution



Source profiles



State-space model

Space-time model of proportions

State-space model:

z_j unobservable composition $\sim LN(\mu_j, \Sigma_j)$

y_j k-vector of counts $\sim Mult(\sum_{i=1}^{k} [y_j]_i, z_j)$

Inference using MCMC again

Stability of arthropod food webs

Omnivory thought to destabilize ecological communities

Stability: Capacity to recover from shock (relative abundance in trophic classes)

Mount St. Helens experiment: 6 treat-ments in 2-way factorial design; 5 reps.

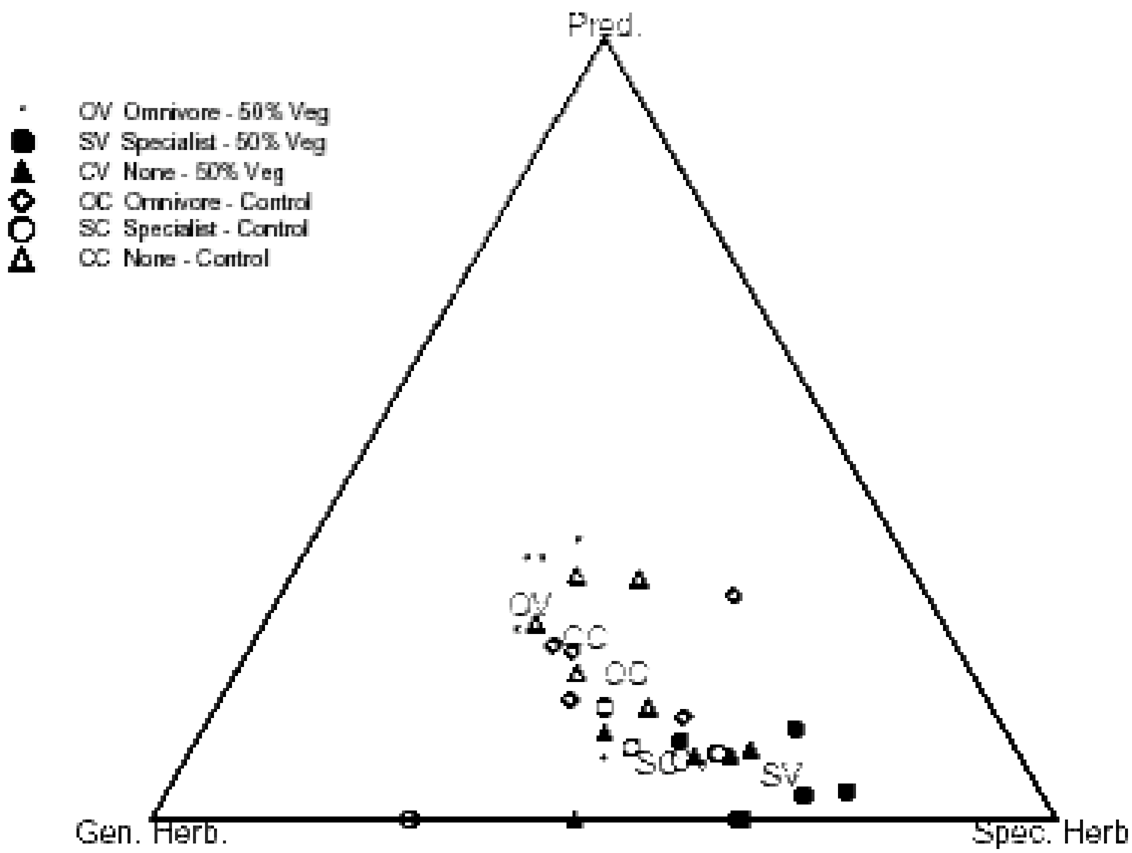
- **Predator manipulation (3 levels)**
- **Vegetation disturbance (2 levels)**

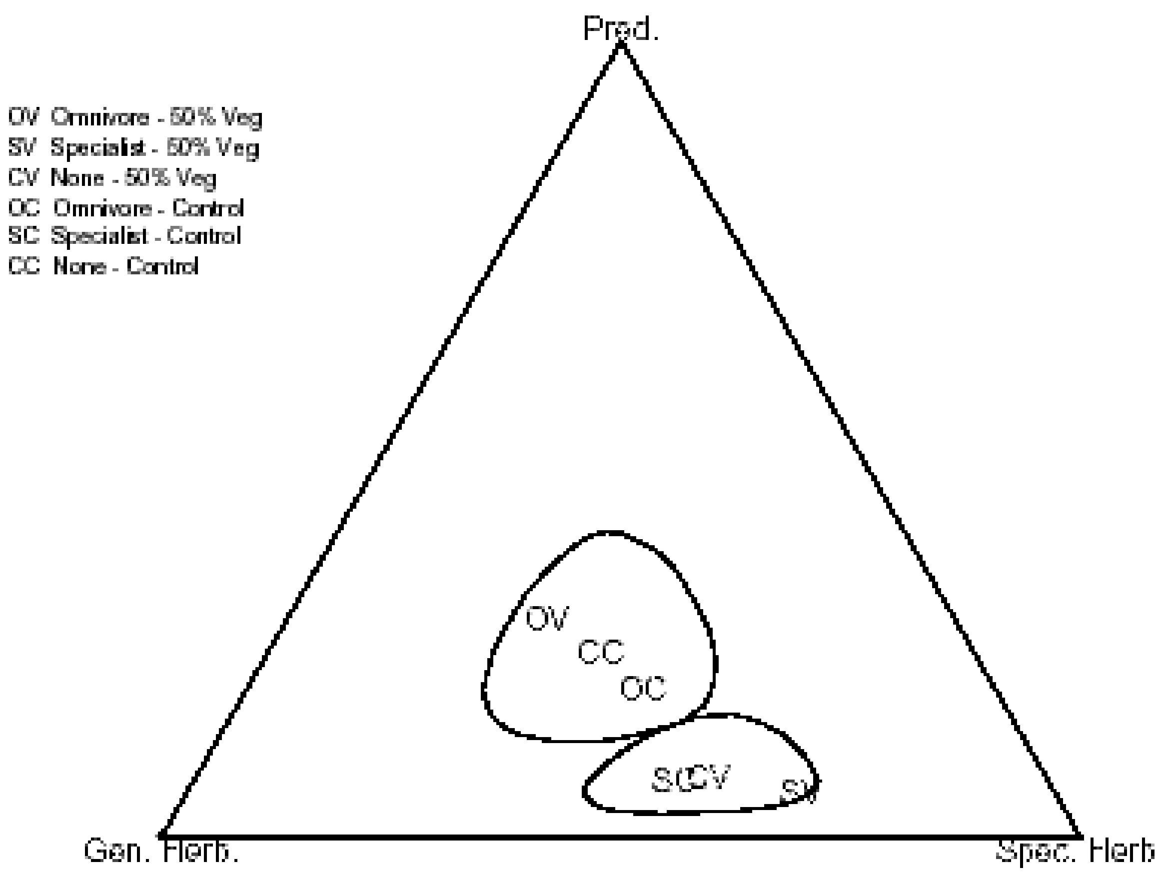
Count anthropods, 6 wks after treatment. Divide into specialized herbivores, general herbivores, predators.

Specification of structure

Σ is generated from independent observations at each treatment

mean depends only on treatment


Pred.
OV Omnivore - 50% Veg
SV Specialist - 50% Veg
CV None - 50% Veg
OC Omnivore - Control
SC Specialist - Control
CC None - Control
OV
CC
OC
SC
SV
Gen. Herb.
Spec. Herb.



Pred.
OV Omnivore - 50% Veg
SV Specialist - 50% Veg
CV None - 50% Veg
OC Omnivore - Control
SC Specialist - Control
CC None - Control
OV
CC
OC
SC
CV
SV
Gen. Herb.
Spec. Herb.

Benthic invertebrates in estuary

EMAP estuaries monitoring program: Delaware Bay 1990. 25 locations, 3 grab samples of bottom sediment during summer

Invertebrates in samples classified into

- –pollution tolerant
- –pollution intolerant
- –suspension feeders (control group; mainly palp worms)

Site j, subsample t

$$z_{jt} \sim LN(\theta_j + \beta x_j, \Psi)$$

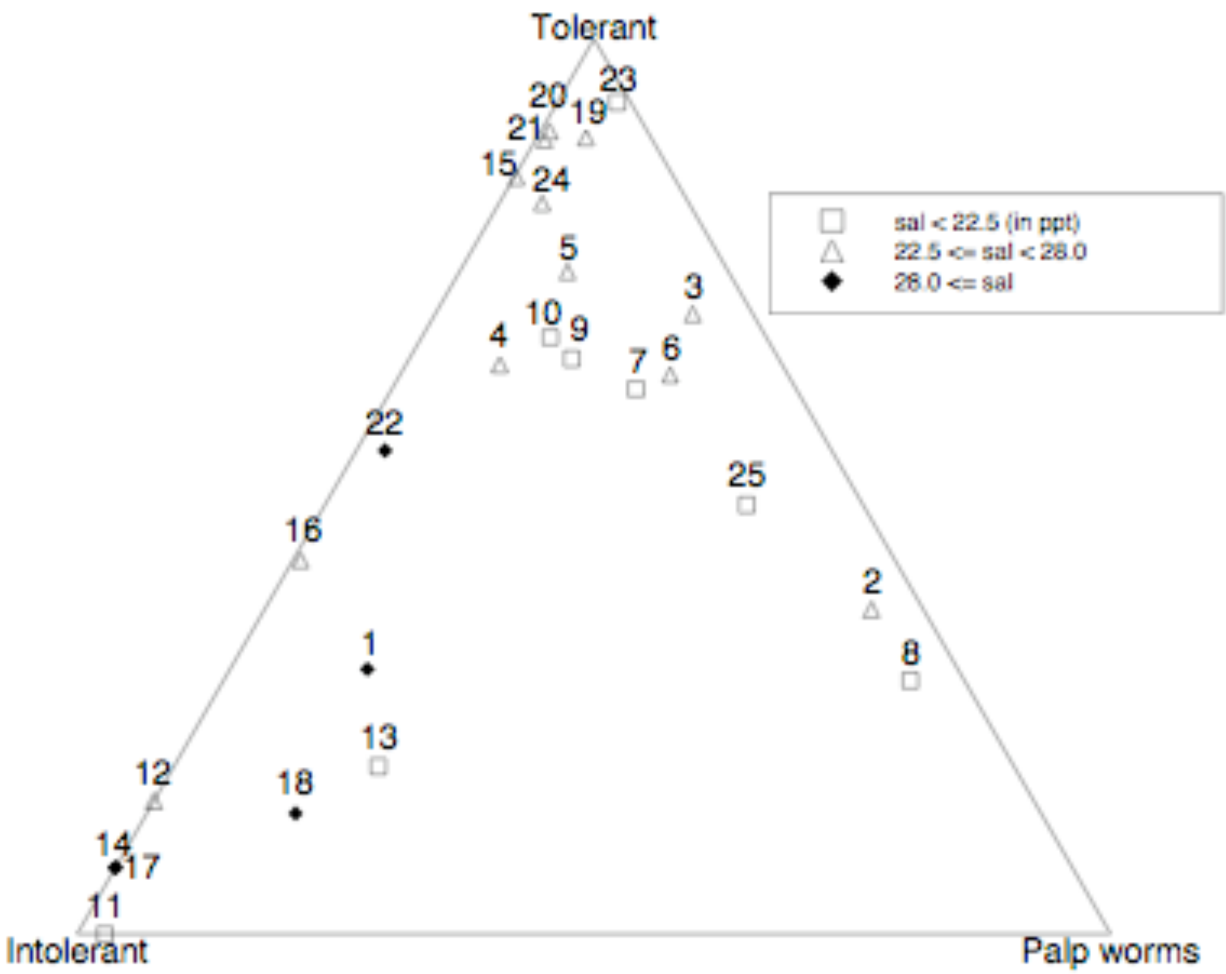
θ_j **~ CAR process**

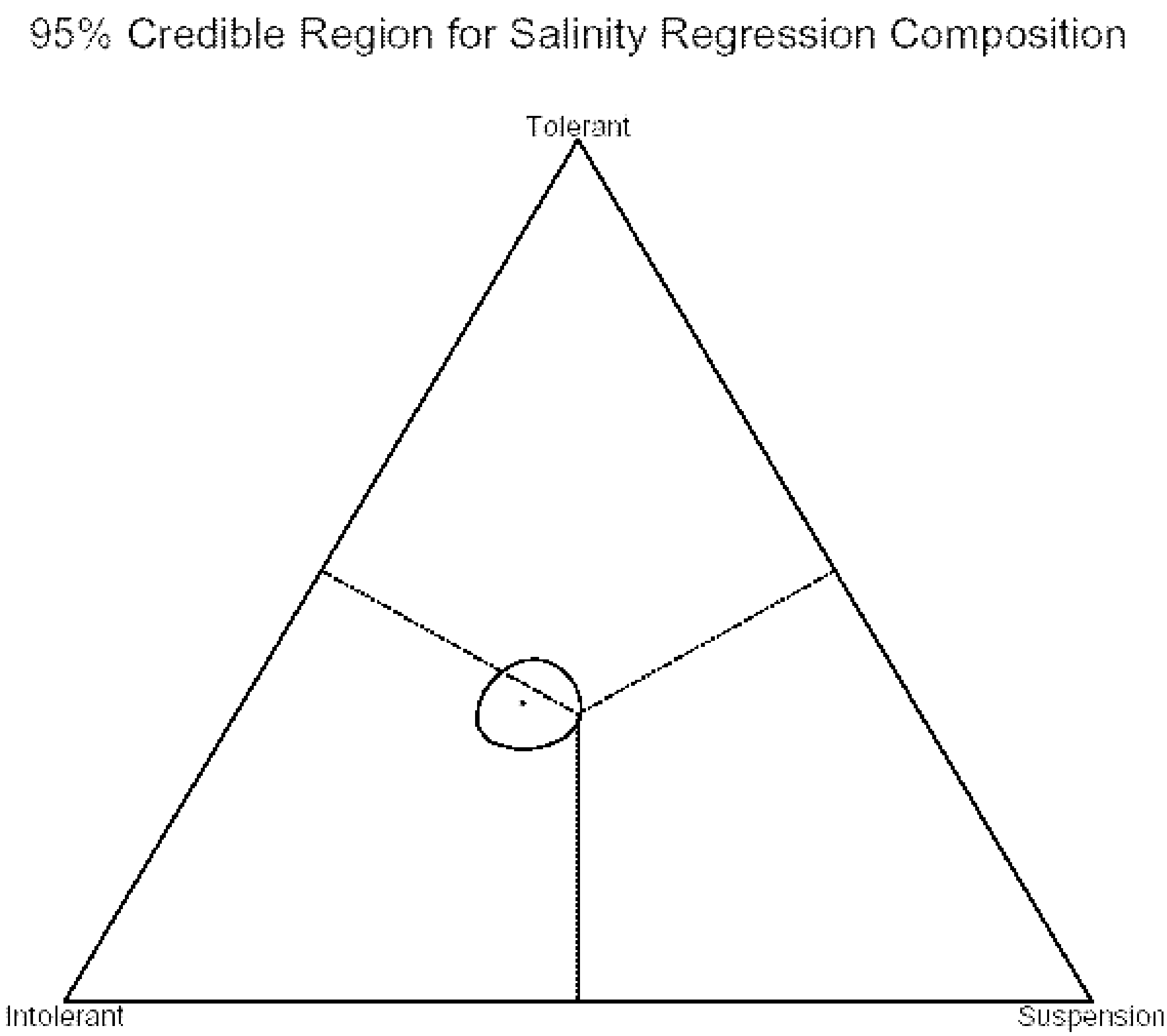
$$E(\theta_j \mid \theta_{-j}) = \mu + \sum_{k \in N(j)} \frac{\lambda}{n_j}(\theta_k - \mu)$$

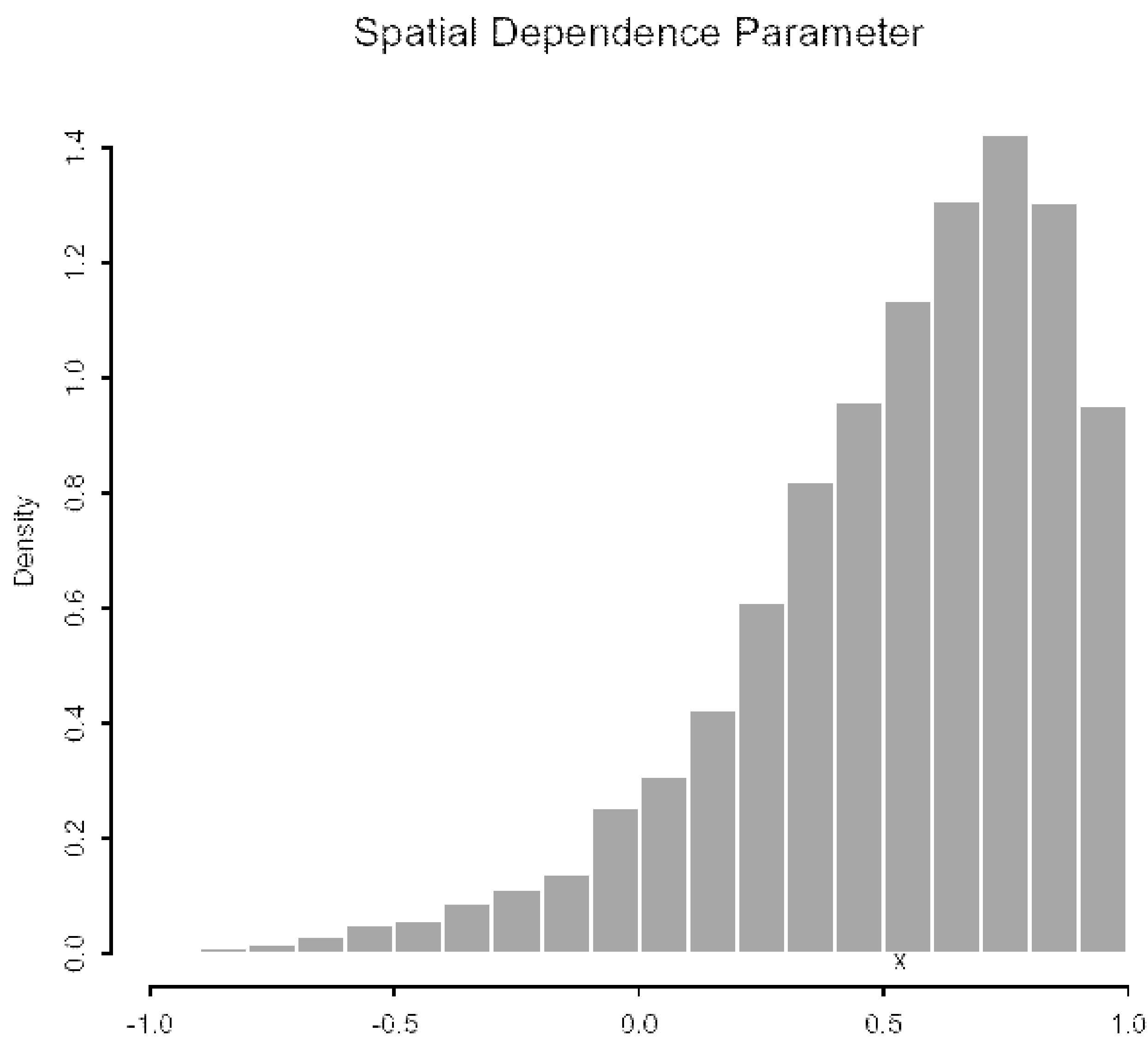
$$Var(\theta_j \mid \theta_{-j}) = \frac{\Gamma}{n_j}$$


11
13
10
25
9
8
7
23
2
6
5
4
3
24
20
21
19
17
16
18
14
15
22
12
1

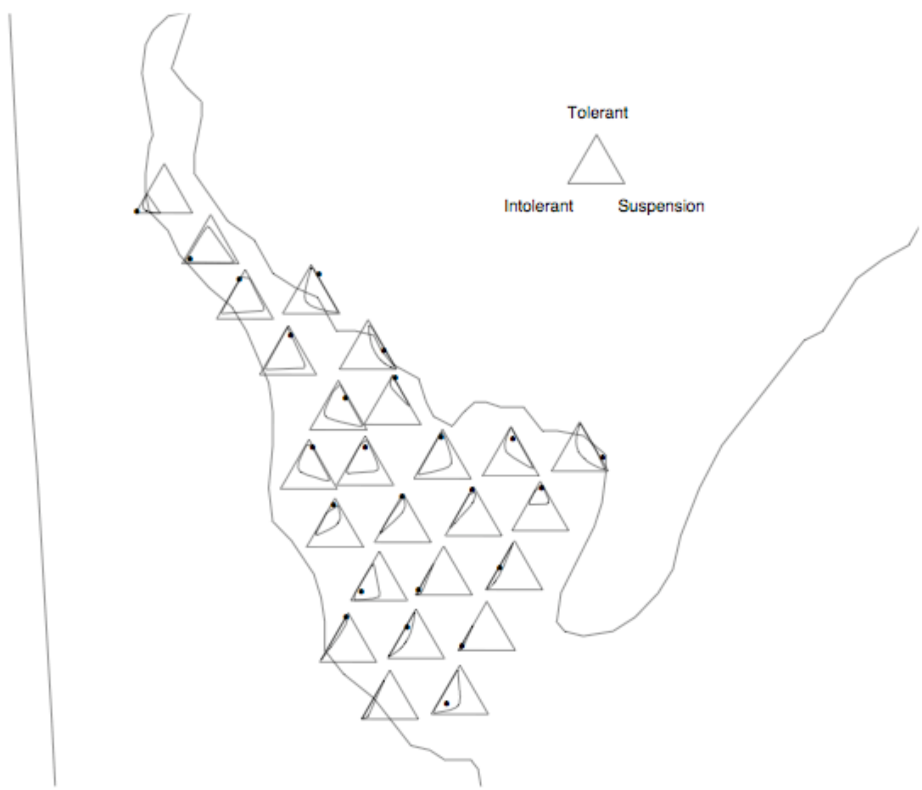
Effect of salinity

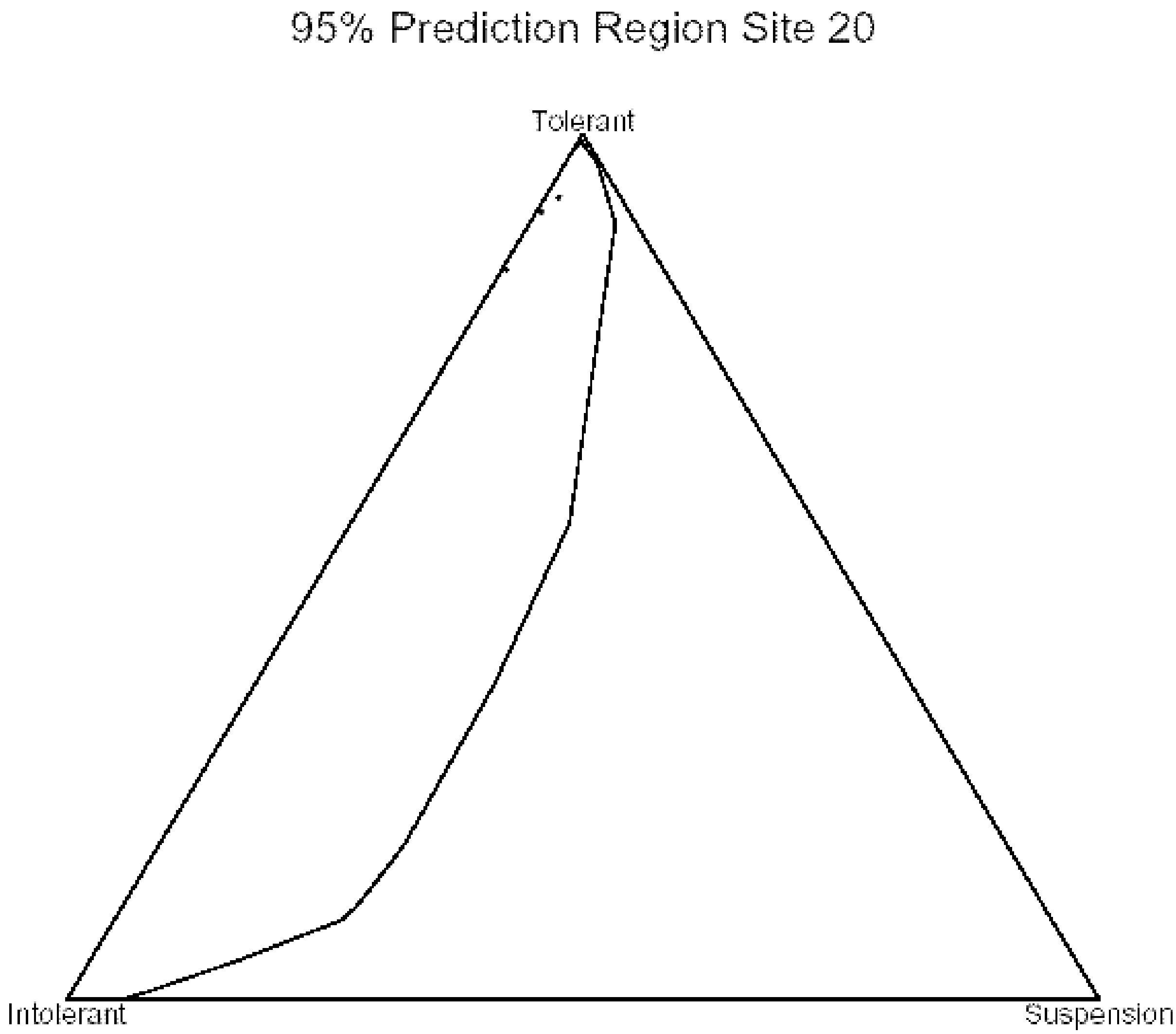



95% Credible Region for Salinity Regression Composition
Tolerant
Intolerant
Suspension


Spatial Dependence Parameter
Density
0.0
0.2
0.4
0.6
0.8
1.0
1.2
1.4
-1.0
-0.5
0.0
0.5
1.0
x

95% Prediction Regions for Hold-out Sub-Sample Compositions




95% Prediction Region Site 20
Tolerant
Intolerant
Suspension

95% Prediction Region Site 23