NRCSE

Spatio-temporal modeling: Mapping mean fields

<u>A</u> common framework for spatio-temporal modeling:

$$\textit{Response}(x,t) = \textit{Trend}(x,t) + \textit{Residual}(x,t)$$

Some (obvious) observations on issues of scale: the practical definition and estimation of *Trend* and *Residual* depend on:

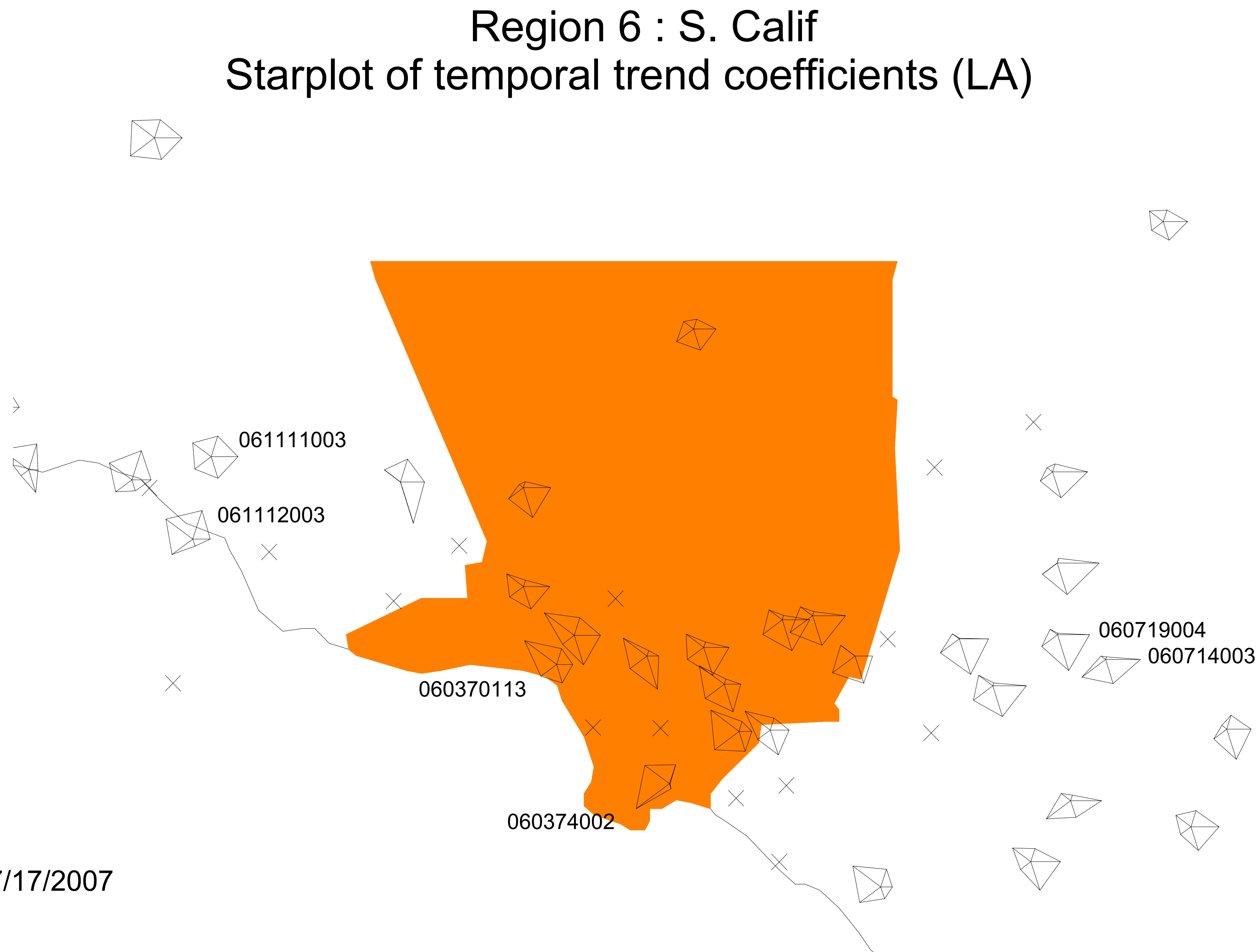
✓ Understanding of the spatio-temporal processes of interest

✓ Scientific questions to be addressed by analysis and modeling

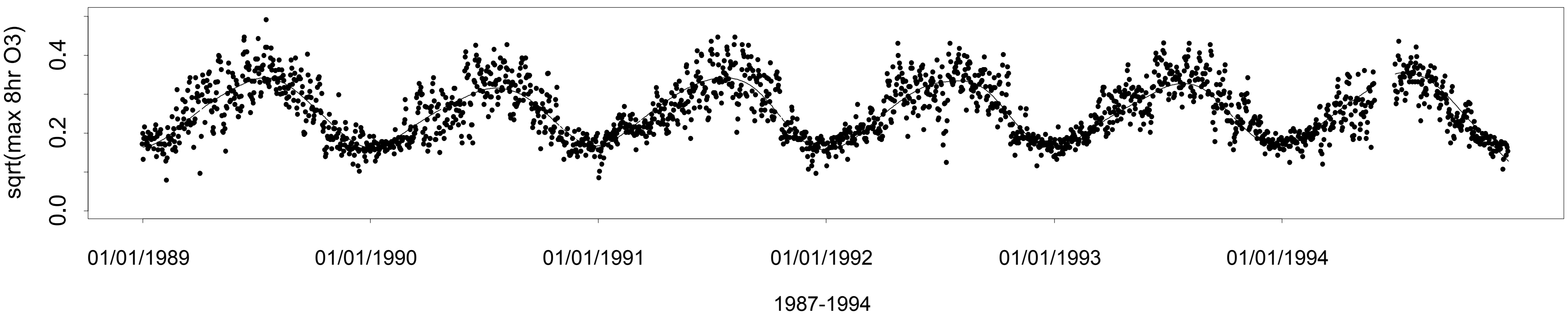
✓ Details of available spatio-temporal sampling data—inference at ``small'' spatial and/or temporal scales generally requires sampling and modeling at corresponding scales.

Focus today is *Trend*(x,t)

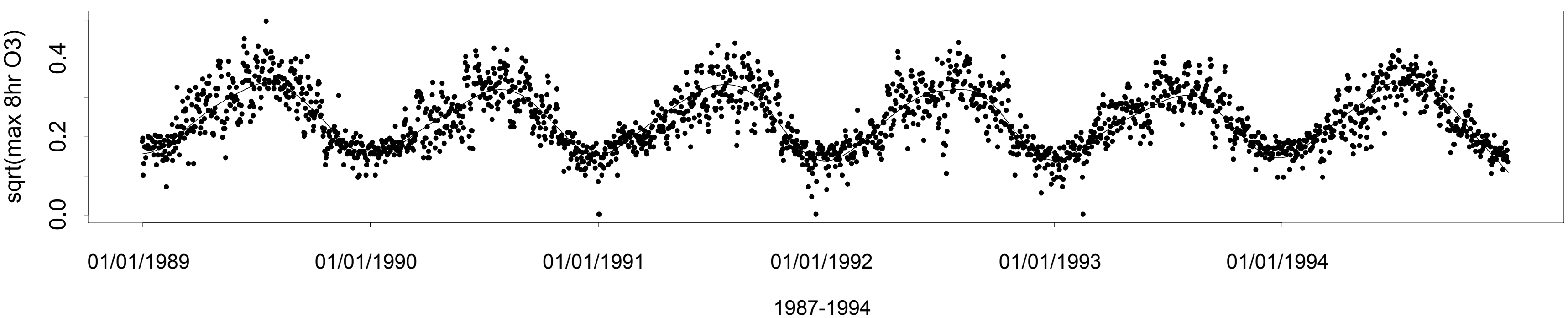
Example 1. Analyses of daily max 8hr ozone



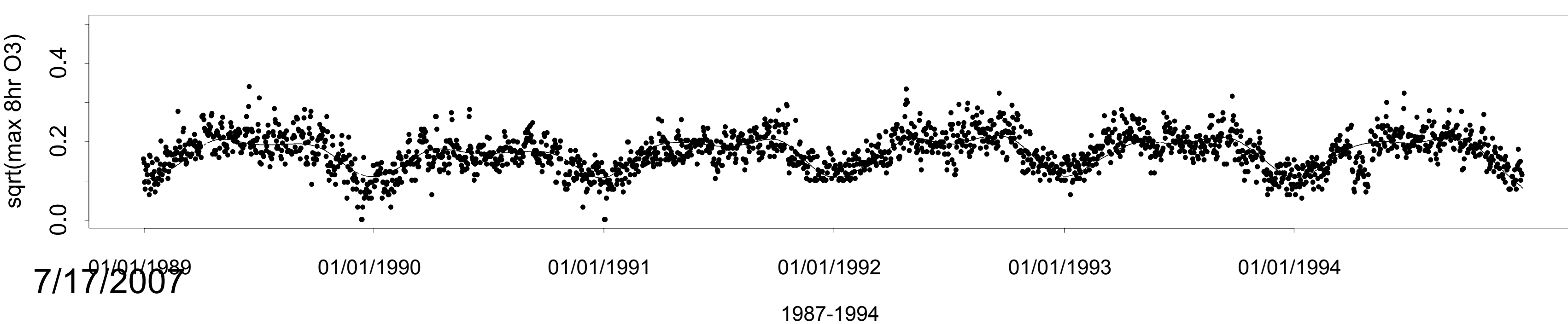
60714003



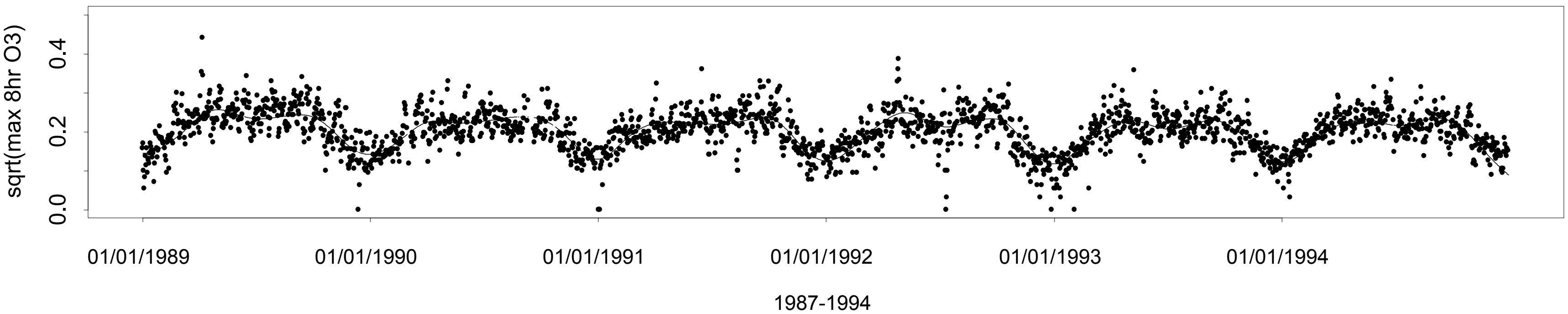
60719004



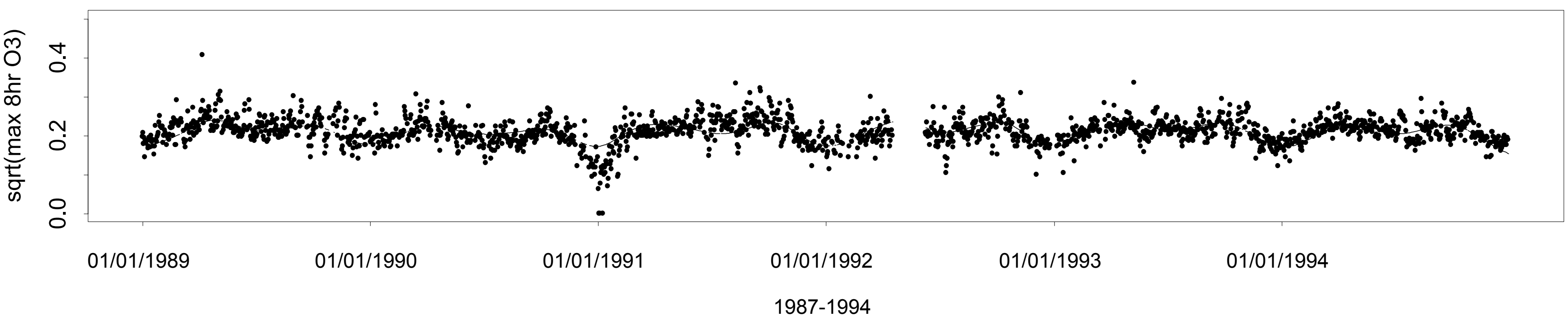
60374002



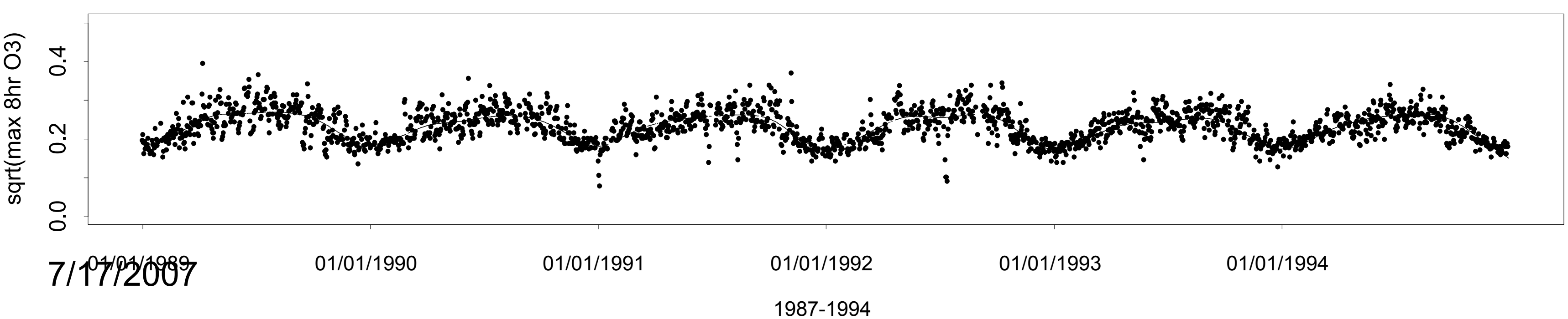
60370113



61112003



61111003



Zidek, slide #9, re hourly ozone data

Example (cont'd)

Footnotes:

- little randomly missing data
- marked daily cycles
- amplitude varies dynamically over season
- stations started up at varying times
- not all measure same species
- physical model data also available but at different resolution
- time variation similar from site-to-site

Example 2. PM2.5 data for MESA Air Study

- Multi-Ethnic Study of Atherosclerosis (MESA) Air Pollution Study
  - Ten year study funded by U.S. EPA
- Objective
  - Examine relationship between air pollution and cardiovascular disease (CVD) progression
- Approach
  - Prospective cohort study
    - 6000-7000 subjects
    - 6 metropolitan areas (Los Angeles, New York, Chicago, Winston-Salem, Minneapolis-St. Paul, Baltimore)
  - Predict long term exposure for each subject
  - Estimate effect of air pollution on progression of subclinical CVD and on clinical events

Sources of Concentration Data

- EPA monitors air pollution in U.S. cities
  - Variable number of monitors by pollutant, city, and year
    - 29 for NO_x in Los Angeles area in 2006
    - Fewer in other cities
  - Variable measurement periods
    - Hourly, daily, every third day, every sixth day

- MESA AIR collecting additional data to fill in gaps for prediction model
  - Design space emphasizes traffic effects, population density, and actual home locations

MESA Spatio-Temporal Monitoring Design

2-week average measurements

- Fixed sites
  - 3-7 sites in each city, continuous monitoring
  - Additional time series data (near road), calibration with AQS monitors
- Home outdoor
  - 50-100 sites in each city
  - Each site measured in two different seasons, 4 at a time
  - Capture realistic subject home data
- Community Co-pollutant (ComCo) [NOx only]
  - 100-150 sites in each city
    - Most arranged in 6 site gradients around major roads
  - All sites in each city measured simultaneously in three different seasons
  - Characterize traffic and population effects

Monitoring Data Structure (NO_x)

Spatial Locations ($N \approx 175$)

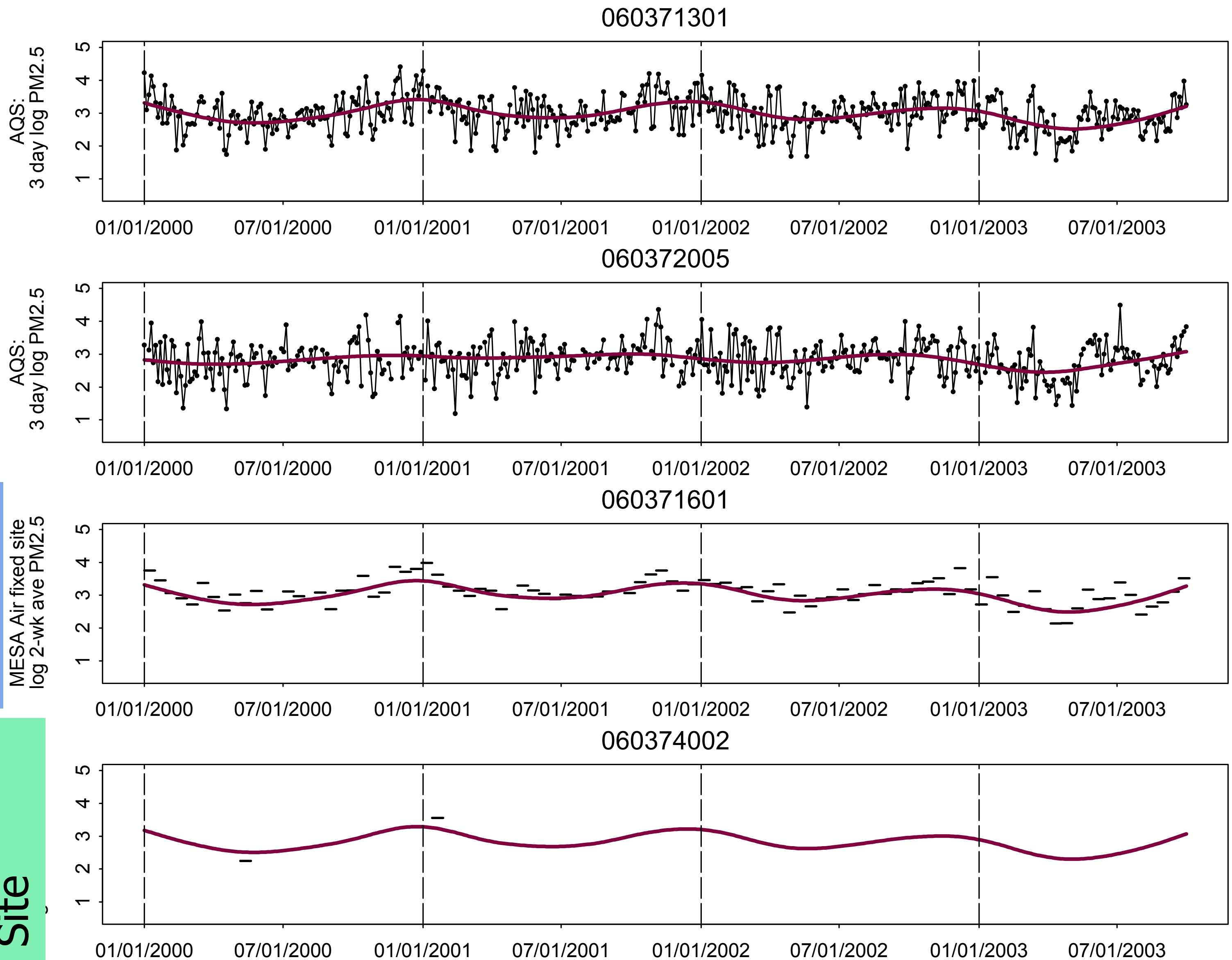
		1	2	3	4	5	6	7	...	24	25	26	27	28	29	30	31	32	33	...	45	46	47	48	49	50
Fixed (EPA) (number varies by location)	1	X	X	X	X	X	X	X	...	X	X	X	X	X	X	X	X	X	X	...	X	X	X	X	X	X
	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
	20	X	X	X	X	X	X	X	...	X	X	X	X	X	X	X	X	X	X	...	X	X	X	X	X	X
Fixed (MESA) (5 sites)	1	X	X	X	X	X	X	X	...	X	X	X	X	X	X	X	X	X	X	...	X	X	X	X	X	X
	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
	5	X	X	X	X	X	X	X	...	X	X	X	X	X	X	X	X	X	X	...	X	X	X	X	X	X
Home Outdoor (50 sites)	1	X										X														
	2	X										X														
	3	X										X														
	4	X										X														
	5		X										X													
	6		X										X													
	7		X										X													
	8		X										X													
	...								...											...						
	47										X															X
	48										X															X
	49										X															X
	50										X															X
Community Co-Pollutant (approx. 100 sites)	1					X											X						X			
	2					X											X						X			
	3					X											X						X			
	4					X											X						X			
	...					...											...						...			
	97					X											X						X			
	98					X											X						X			
	99					X											X						X			
	100					X											X						X			

Realized Data Examples ($PM_{2.5}$)

EPA/AQS Sites

MESA Air Site

Home Outdoor Site



$PM_{2.5}$ data for MESA Air Study

- Plots of
  - log transforms of 2-week averaged AQS PM2.5 mass concentration data
  - log transforms of 2-week integrated PM2.5 mass concentration data from MESA Air supplemental monitoring sites

[see pdf file, “Trendfits.lme4.pdf”]

- How would you model spatially varying temporal trend?
- What are feasible approaches to computing predictions at a point in space at time t?
  - assuming residuals uncorrelated in time
  - with (spatio-)temporally correlated residuals
  - how important is the temporal correlation structure?

Spatial Deformation Model

Damian et al., 2000 (*Environmetrics*), 2003 (*JGR*)

$$Z(x,t) = \mu(x,t) + \nu(x)^{1/2} H_t(x) + \varepsilon(x,t)$$

$\mu(x,t)$ spatio-temporal trend
parametric in time; mv spatial process

$\nu(x)$ temporal variance at x,
log-normal spatial process

$\varepsilon(x,t)$ msmt error and short-scale variation
$N(0,\sigma_\varepsilon^2)$, independent of $H_t(x)$

$H_t(x)$ mean 0, var 1, 2nd-order cont. spatial process
Cov$(H_t(x), H_t(y)) \xrightarrow[x \to y]{} 1.$

Decompose smoothly varying spatio-temporal trend:

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

- 1st term represents long-term mean concentration and will derive from a Bayesian analysis of a spatial regression model combining average concentration data from fixed-site ambient monitors and spatial covariate information encoded in a GIS.
- 2nd component represents mainly smooth seasonal temporal variation.

(1) Estimation of the long-term mean spatial field

Following Hoek and colleagues (2002 *Atmos Env*, 2003 *Epidemiology*), assume the long-term mean concentration can be expressed with a linear regression model:

$$\mu_1(s_i) = \mu_0 + \delta_1 v_{i1} + \delta_2 v_{i2} + \cdots + \delta_k v_{ik}$$

where $v_1, \ldots, v_k$ represent pop density, proximity to roads, traffic density, and possibly local topographic and climatic wind patterns. In the field of spatial epidemiology this is now called a "land use regression".

(2) Smooth, spatially varying, seasonal temporal variation.

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

The spatial index in the 2nd component allows for the magnitude and details of the seasonal variation to vary over the spatial scale of the regional target communities.

Over sufficiently small regions one may find homogeneous seasonal variation, permitting an additive (separable) decomposition of the spatio-temporal trend.

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

Characterize and estimate the seasonal structure of air pollutant concentrations in terms of a model written as:

$$C(s_i,t) = \mu_1(s_i) + \mu_2(s_i,t) + \nu(s_i,t)$$

$$\mu_2(s_i,t) = \sum_{j=1}^{J} \beta_j(s_i) f_j(t)$$

Where the $f_j(t)$ are temporal basis functions describing possible seasonal trend patterns, and the $\beta_j(s_i)$ represent spatially varying coefficients of these trend patterns.

Choice of basis functions: fourier components, splines, … ?

Approach here is consistent with methods for analysis of "functional data" where data to be modeled are curves.

$$C(s_i,t) = \mu_1(s_i) + \sum_{j=1}^{J} \beta_j(s_i) f_j(t) + \nu(s_i,t)$$

$$= \sum_{j=0}^{J} \beta_{ij} f_j(t) + \nu_{it}$$

or

$$C = F\mathrm{B} + \mathrm{N}$$

where we are writing C as an $T \times S$ (time-space) matrix of observations, B is an $(J+1) \times S$ matrix of coefficients multiplying the matrix F, $T \times (J+1)$, with columns containing values of the basis functions evaluated at the T observation times (j=1,…,T).

“Obvious” calculation is an SVD of the concentration matrix *C*, providing the best (least squares) rank-(J+1) approximation.

$$C = F\mathrm{B} + \mathrm{N} = UDV^{T} = \sum_{j=1}^{S} d_j u_j v_j{}'$$

$$= \left(\sum_{j=1}^{J+1} d_j u_j v_j{}' \right) + \left(\sum_{j=J+2}^{S} d_j u_j v_j{}' \right)$$

$$= U^{(J+1)} \left(D^{(J+1)} V^{(J+1)\prime} \right) + N$$

where the columns of the (truncated) matrix of left singular vectors are considered to represent the matrix of values of the *J*+1 temporal basis functions:

$$F = U^{(J+1)}$$

Issues: Smoothness of the singular vectors as components of trend; computation with missing data.

Compute trend components empirically as smoothed versions of the temporal singular vectors of the $T \times S$ data matrix (rather than assuming parametric forms such as trigonometric functions). Arbitrary amounts of missing data are accommodated in an EM-like iterative calculation of the SVD – or Peter Hoff's recent Bayesian model-based approach to the SVD.

In sufficiently small spatial domains the number of temporal trend components required will likely be J=1 and the assumption of separability would lead to coefficients of these components not varying in space. $\beta_j(s_i) = \beta_j$.

However, a Bayesian spatial regression model can incorporate these parameters as spatial fields, and thus provide the basis for estimation of $\mu_2(s_i, t)$ at target homes.

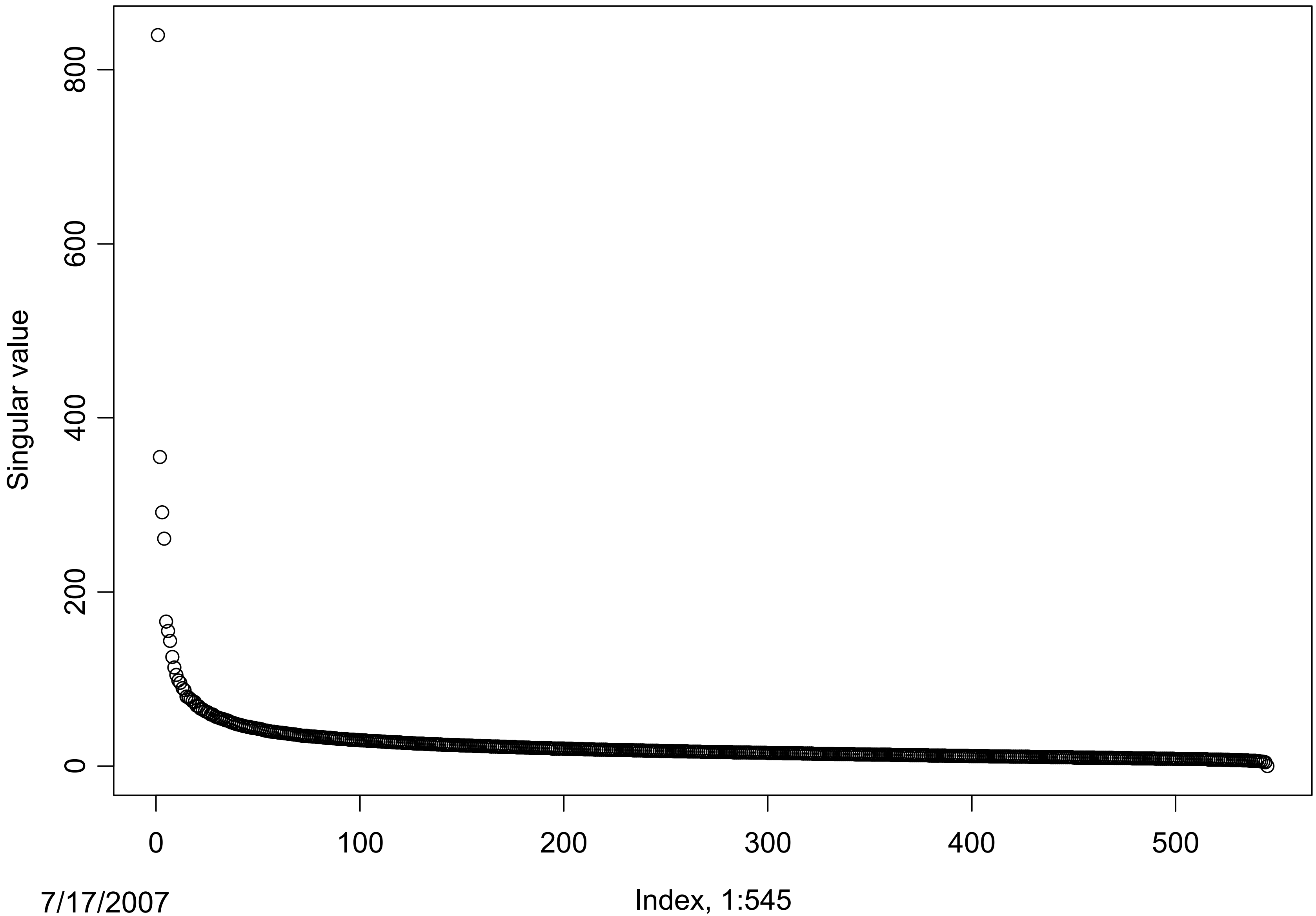
- The SVD could be computed using an iterative algorithm in which the left singular vectors are computed by regressions in which smoothness is imposed directly. However, it has proven adequate to simply compute smoothed versions of the left singular vectors and then compute the coefficient matrix by ordinary least squares regressions of the columns of the response matrix on the set of smoothed singular vectors .

Iterative algorithm for "SVD" with missing data

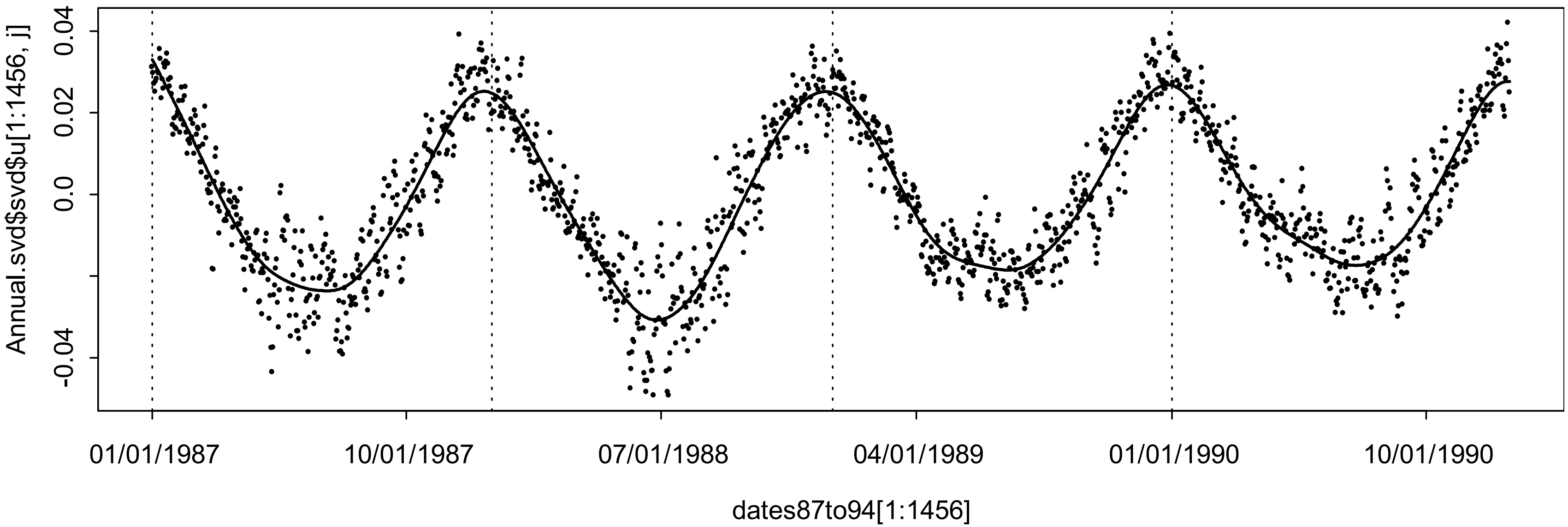
A simple, "EM-like" iterative algorithm for the SVD:

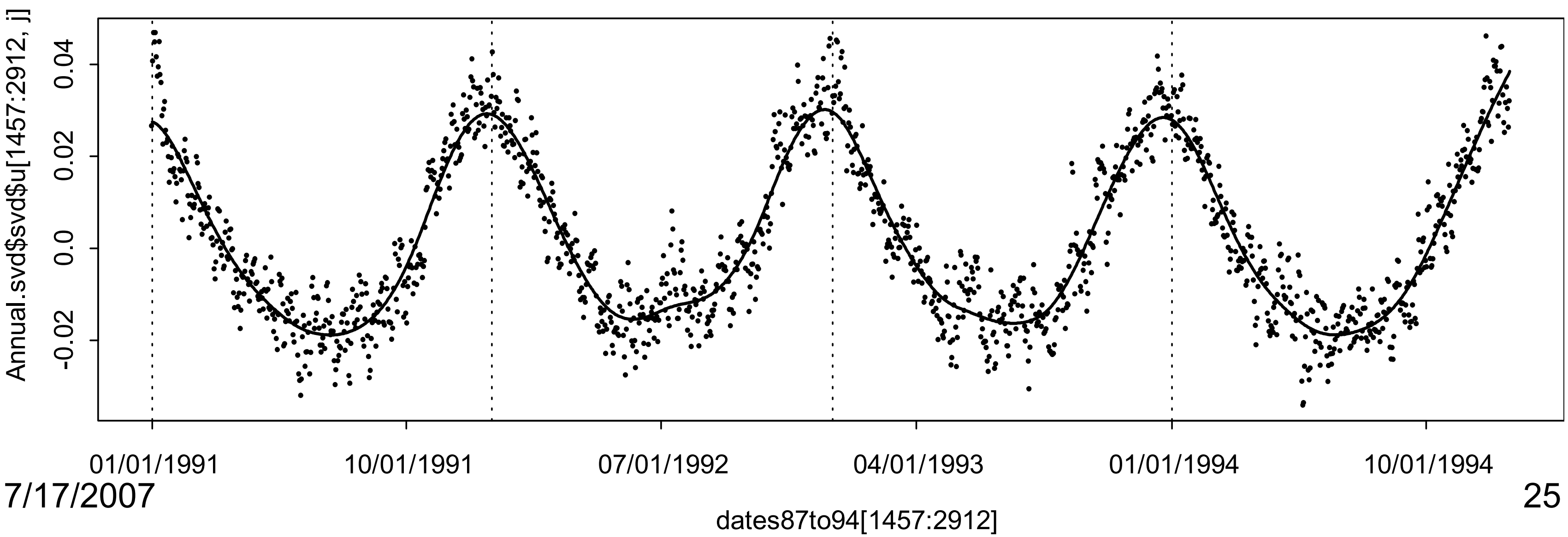
1. Specify a dimension (rank), J, for the SVD model.
2. Scale the observations at each monitoring site (columns of C) to norm (variance) one.
3. Fill in the missing observations in the data matrix using elements of an initial rank-one approximation provided by a regression through the origin of each column of C on the vector computed as the average over sites of the columns of C.
4. Compute the rank-J SVD-approximation of the now complete data matrix
5. Replace the missing values in C by the elements of the rank-J SVD approximation.
6. Return to step 4 and iterate to convergence.

Singular values of T=2912 x S=545 observation matrix

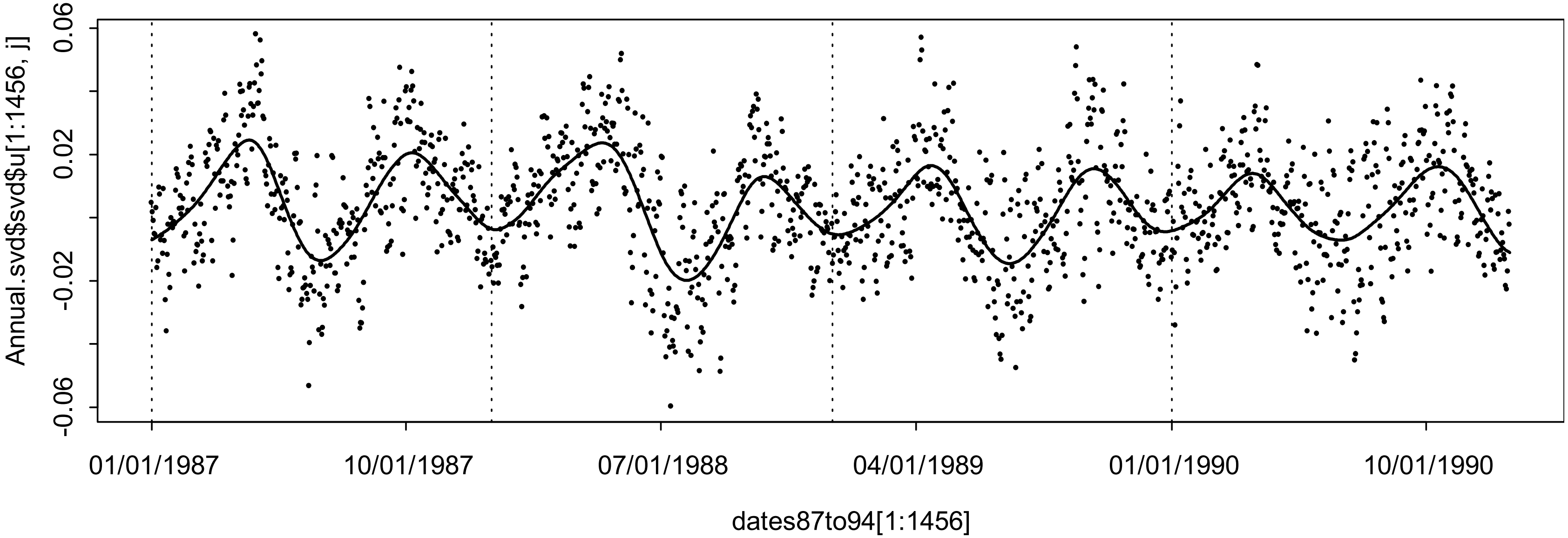


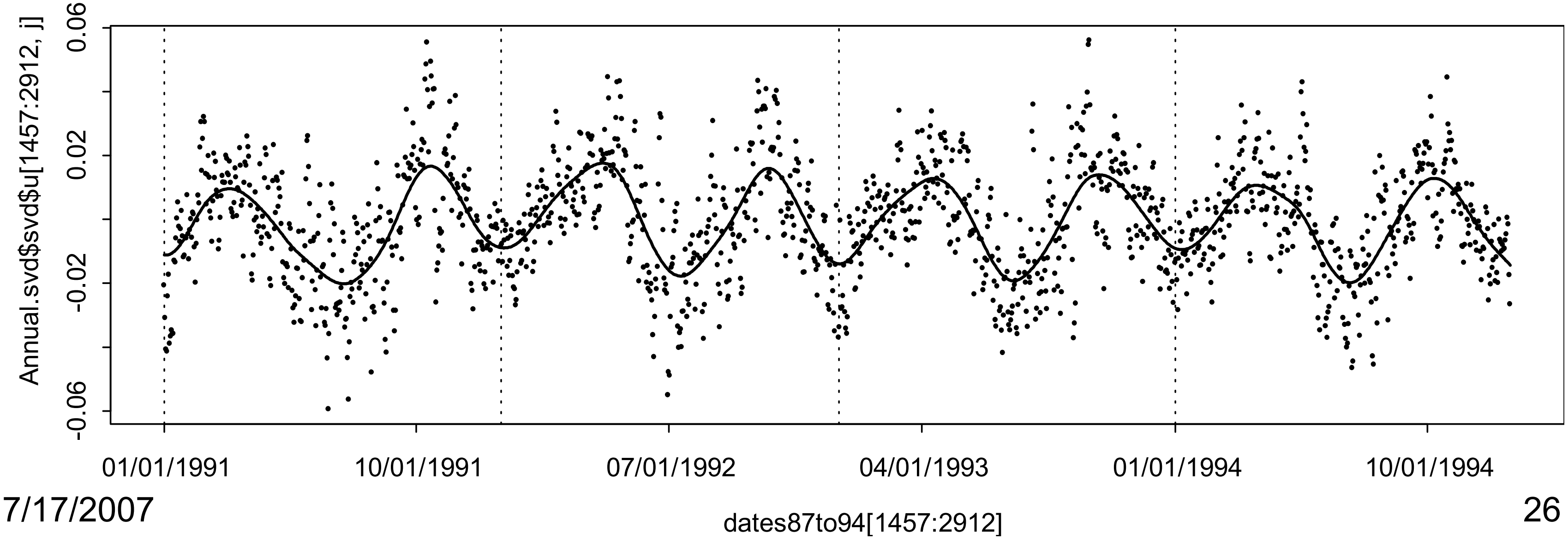
Annual Trend Component 1



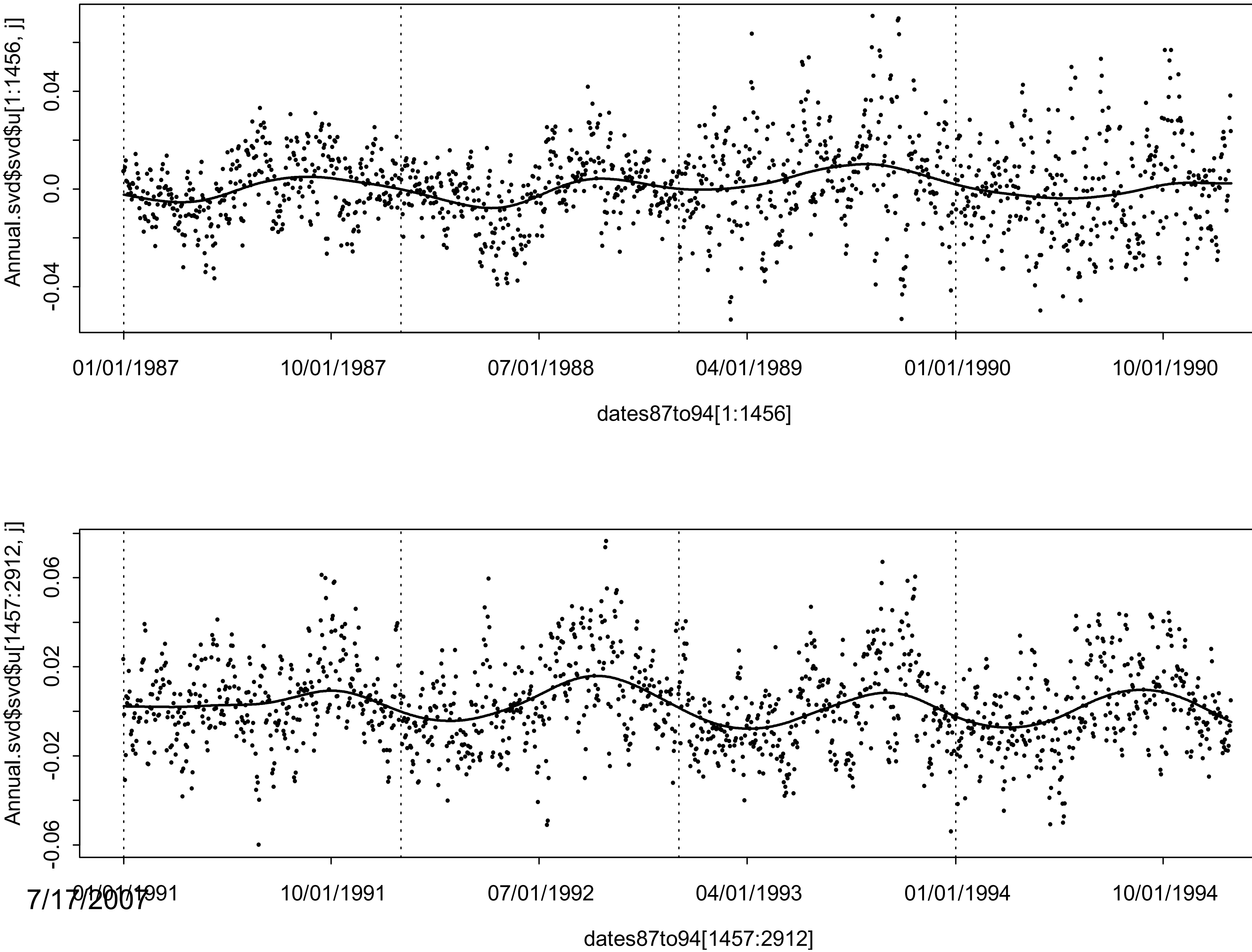


Annual Trend Component 2

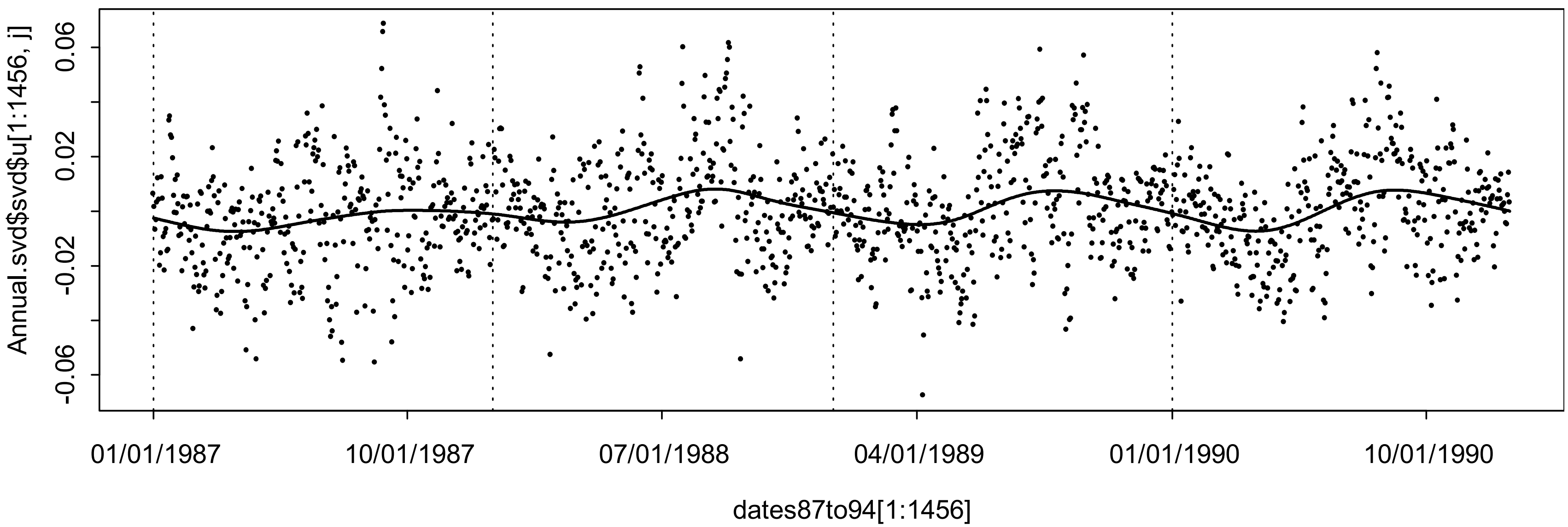


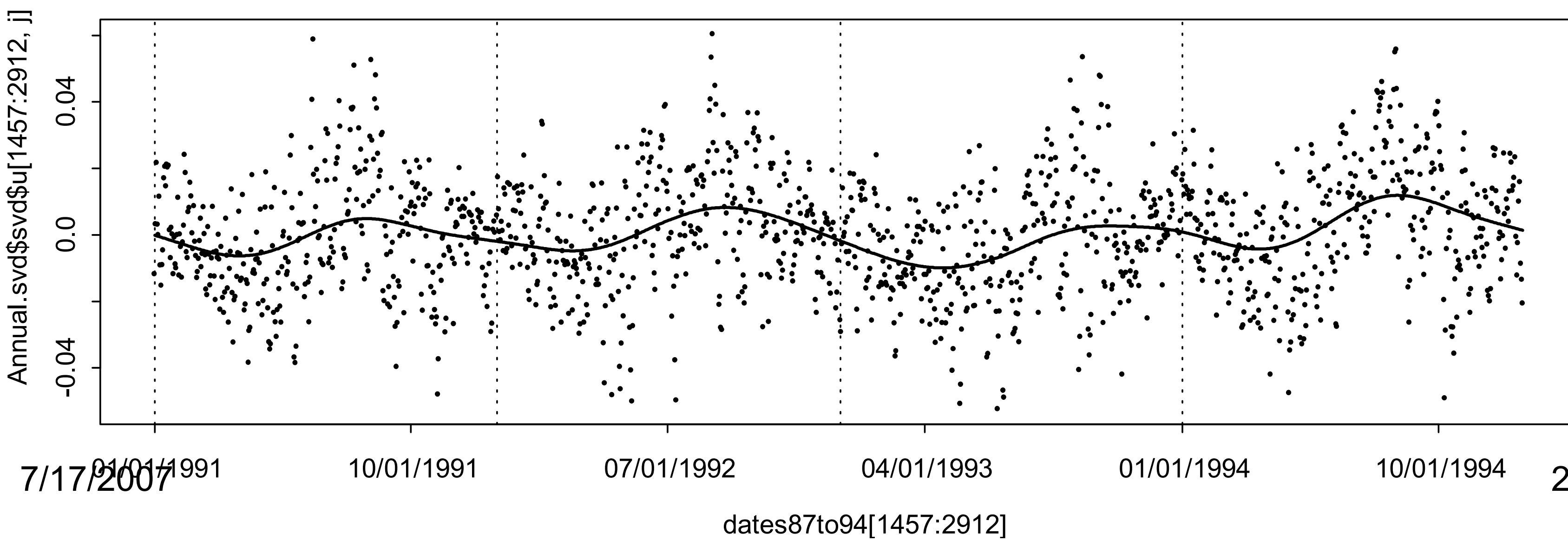


Annual Trend Component 3



Annual Trend Component 4





Note: this is essentially the “orthogonal series” or “principal components” computation Jim explained in lect 3.1, slides 25-27. But the perspective and purpose is different.

Modelling X’s joint distribution

APPROACH 6. Orthogonal series representations

$X_{it} = \sum_{k=1}^{K} a_k(t)\phi_k(i)$

are non-random orthonormal basis functions.

- the ϕs can be eigenvectors from decomposition of spatial covariance.
- incomplete - represents spatial dispersion not temporal
- long history, used in many ways (e.g. EOFs = empirical orthogonal functions.

Modelling X's joint distribution

EOFs: Suppose for n spatial sites $\mathbf{X}_t : n \times 1,\ t = 1, \ldots, T$ are independent copies of a r.vector $\mathbf{X}_0$. Let $\mathbf{X} = [\mathbf{X}_1, \ldots, \mathbf{X}_T]$ and estimate the spatial covariance by

$$\hat{\Sigma}_{\mathbf{X}} = \frac{1}{T-n}\mathbf{X}\mathbf{X}' = \frac{1}{T-n}\sum \mathbf{X}_t\mathbf{X}_t'$$

Diagonalize it

$$\hat{\Sigma}_{\mathbf{X}} = \mathbf{Q}'\mathbf{D}^2\mathbf{Q},\ \text{with } D_1 > \cdots > D_p$$

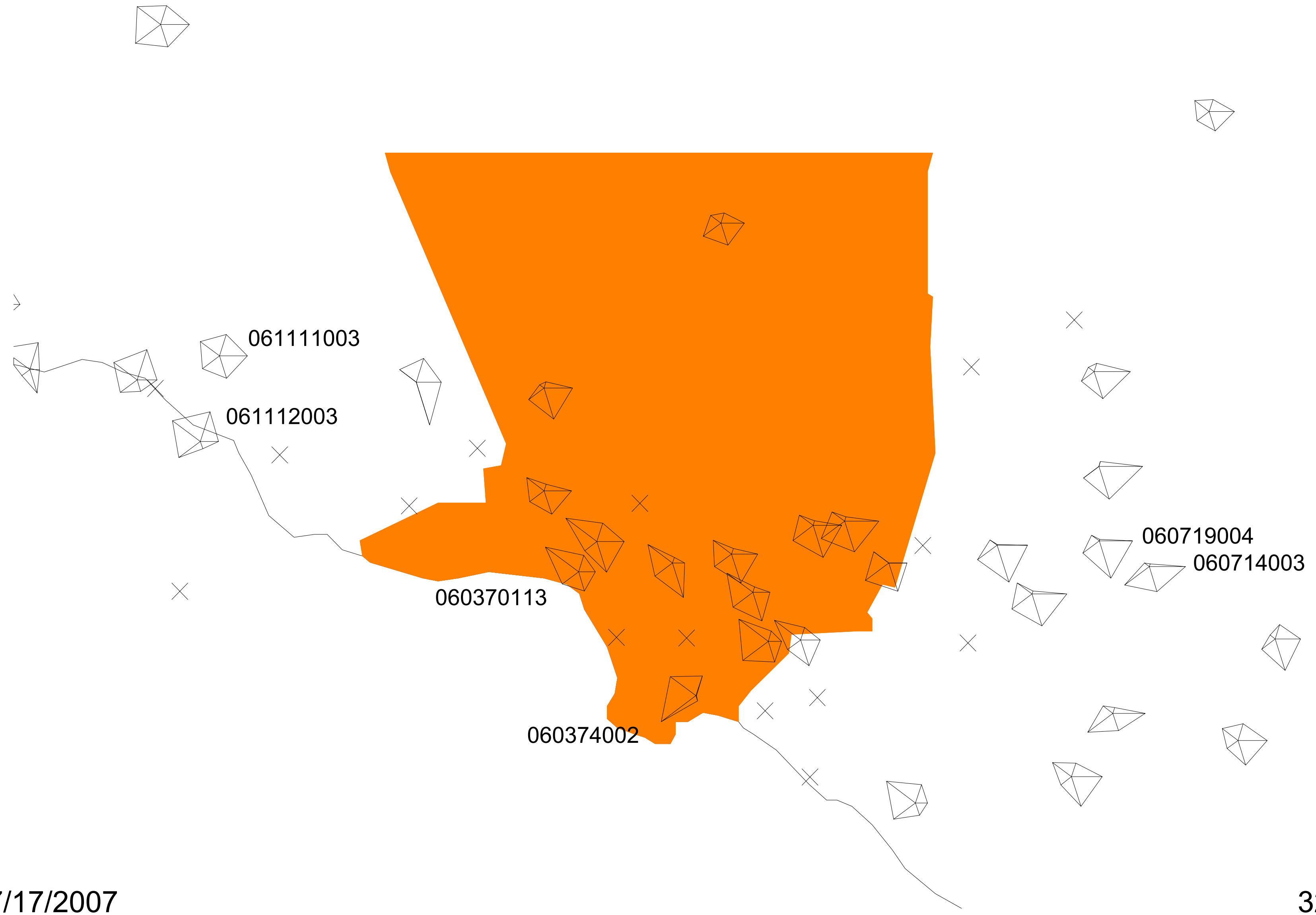
Then $\mathbf{W}' \doteq \mathbf{D}^{-1}\mathbf{Q}$ is orthogonal ($\mathbf{W}'\mathbf{W} = I_n$),
$\mathbf{P} \doteq \mathbf{W}\mathbf{X}/(T-n) : n \times T$ is orthonormal ($\mathbf{P}\mathbf{P}' = I_n$).
Furthermore,

Modelling X's joint distribution

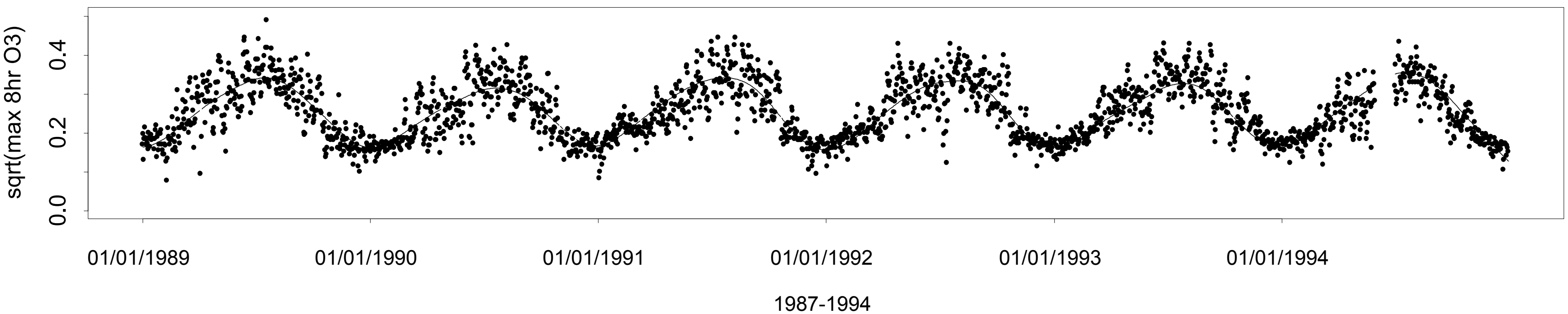
$$\begin{aligned} \mathbf{X} &\equiv \mathbf{WP} \\ &= [\mathbf{W}_1, \ldots, \mathbf{W}_n][\mathbf{P}'_1, \ldots, \mathbf{P}'_n]' \\ &= \sum_1^n \mathbf{W}_i \mathbf{P}_i \sim \mathbf{W}_1\mathbf{P}_1 + \mathbf{W}_2\mathbf{P}_2 \end{aligned}$$

The $\{\mathbf{W}_i\}$ are called the **EOFs** as they capture the spatial variation. For example $\mathbf{W}_1$, the most important component of spatial variation might represent the northern - southern hemisphere component of temperature variation across space. The $\{\mathbf{P}_i\}$ are called the "**principal components**" altho this means something different to statisticians. The approximation using just two EOFs would mean a big reduction in the data file.

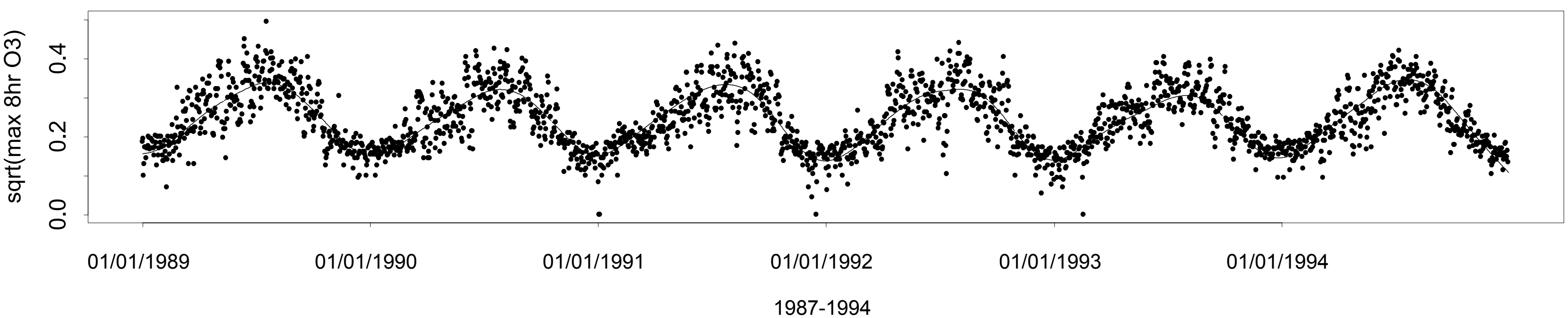
Region 6 : S. Calif
Starplot of temporal trend coefficients (LA)



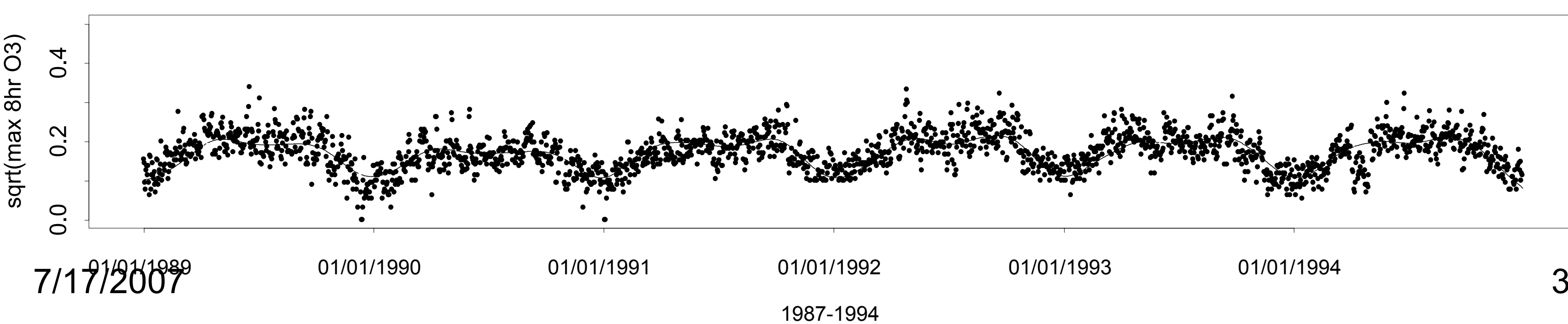
60714003



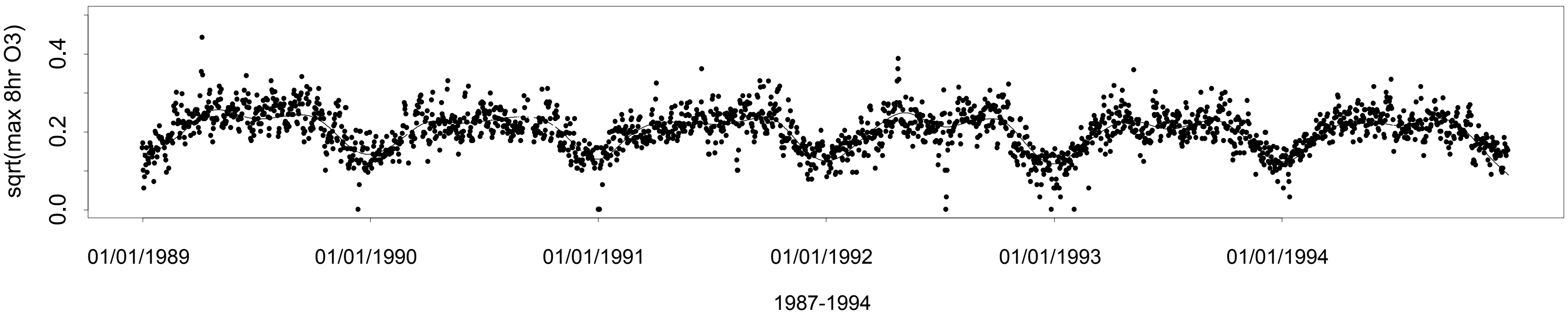
60719004



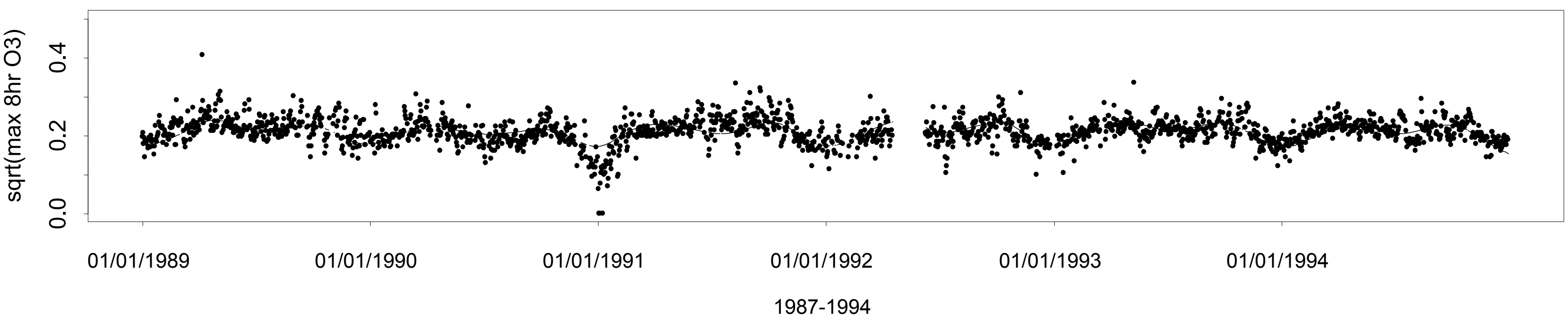
60374002



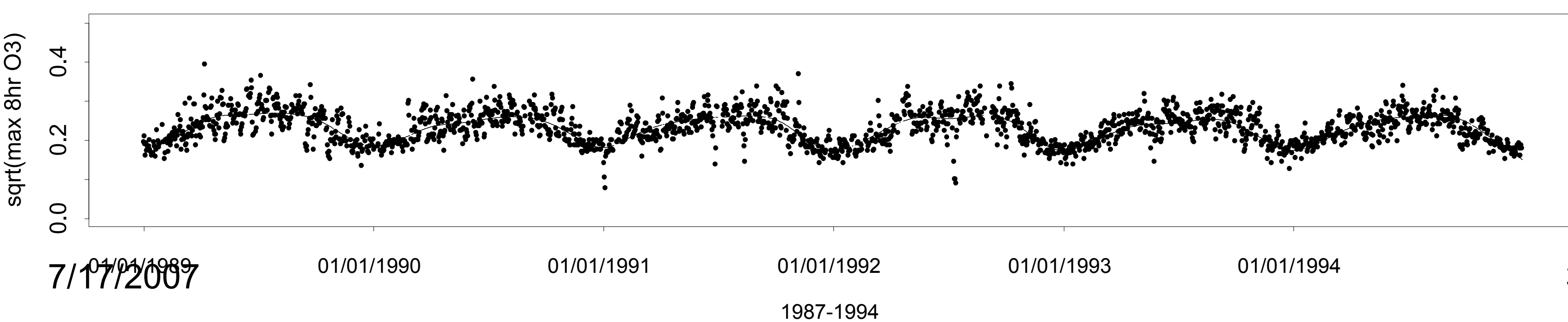
60370113



61112003



61111003



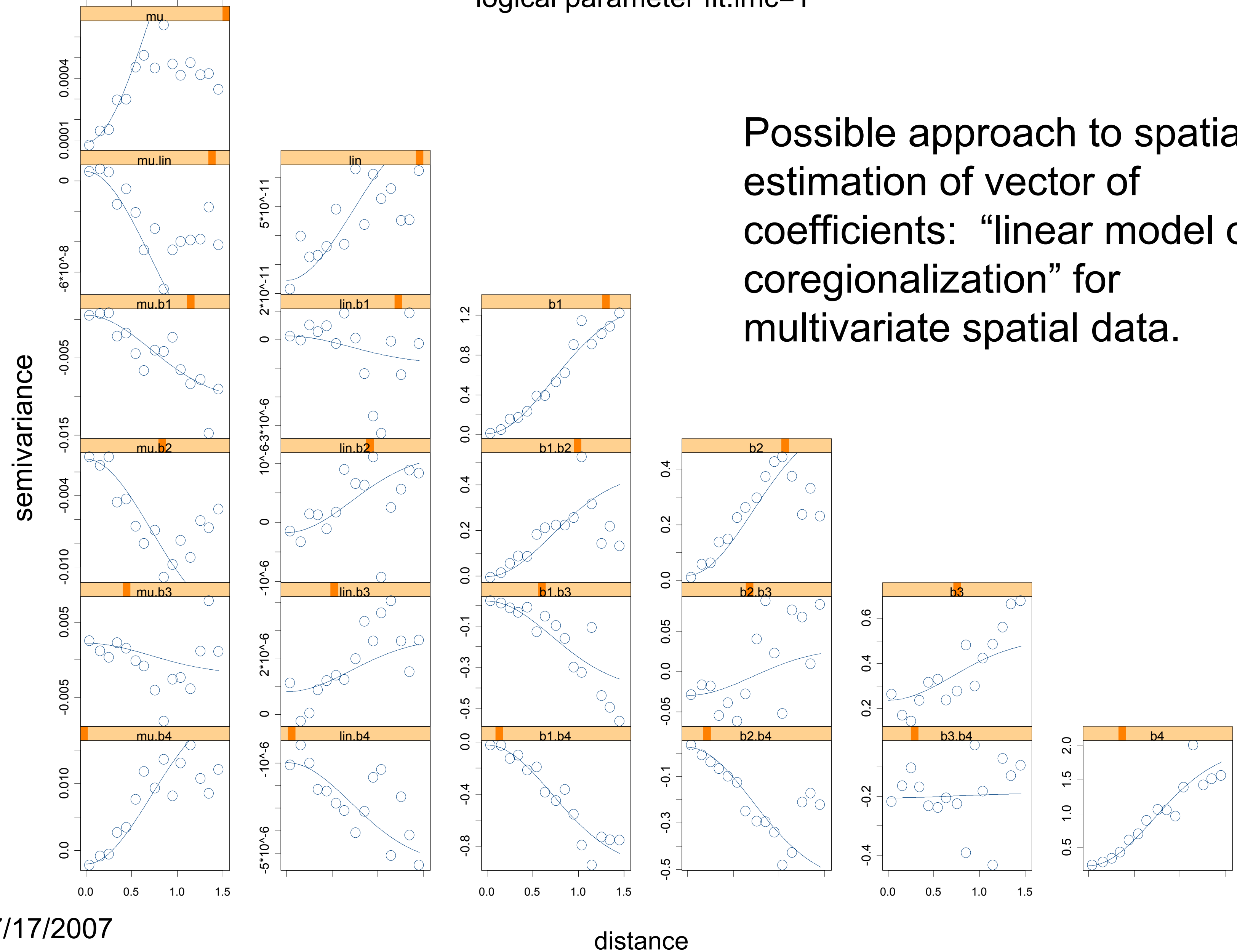
Prediction at an unmonitored site

$$C(s_0,t) = \hat{\mu}_1(s_0) + \hat{\mu}_2(s_0,t) + \hat{\nu}(s_0,t)$$

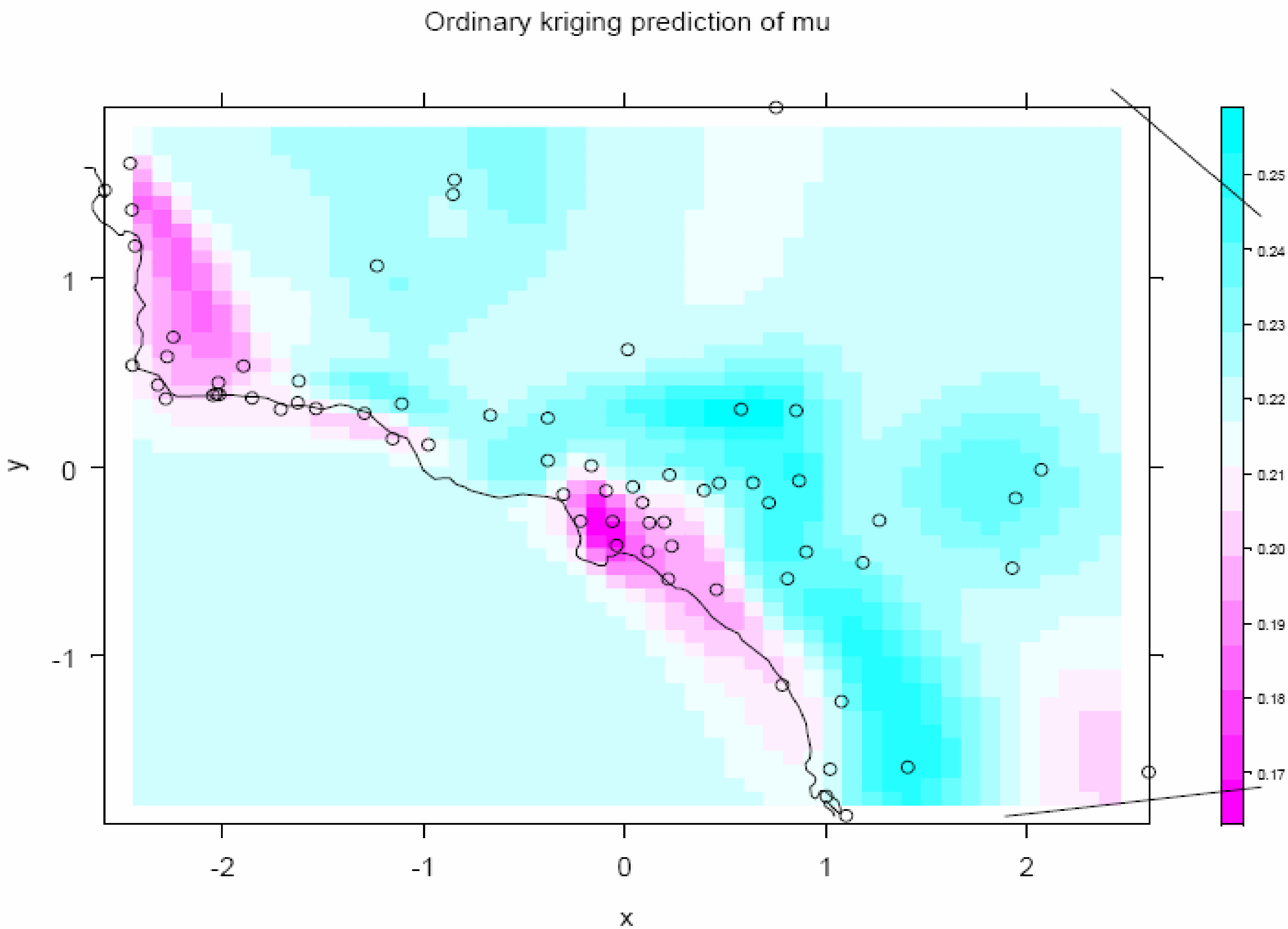
$$\hat{\mu}_2(s_0,t) = \sum_{j=1}^{J} \hat{\beta}_j(s_0) f_j(t)$$

where $\hat{\beta}_j(s_0)$ is a kriged estimate of $\beta_j(s_0)$, considered as a random field

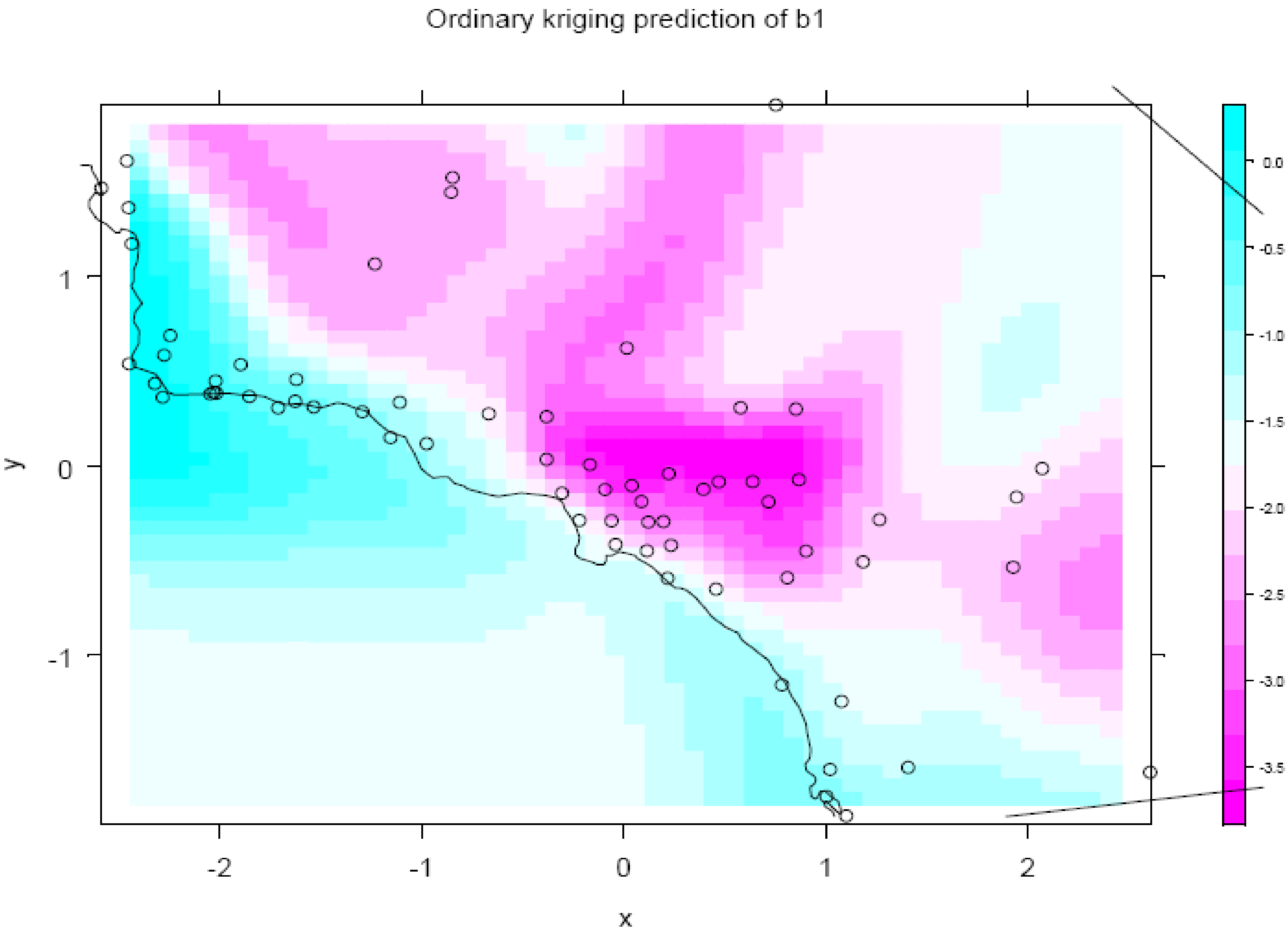
Best fit achieved with Gaussian variogram, logical parameter fit.lmc=T



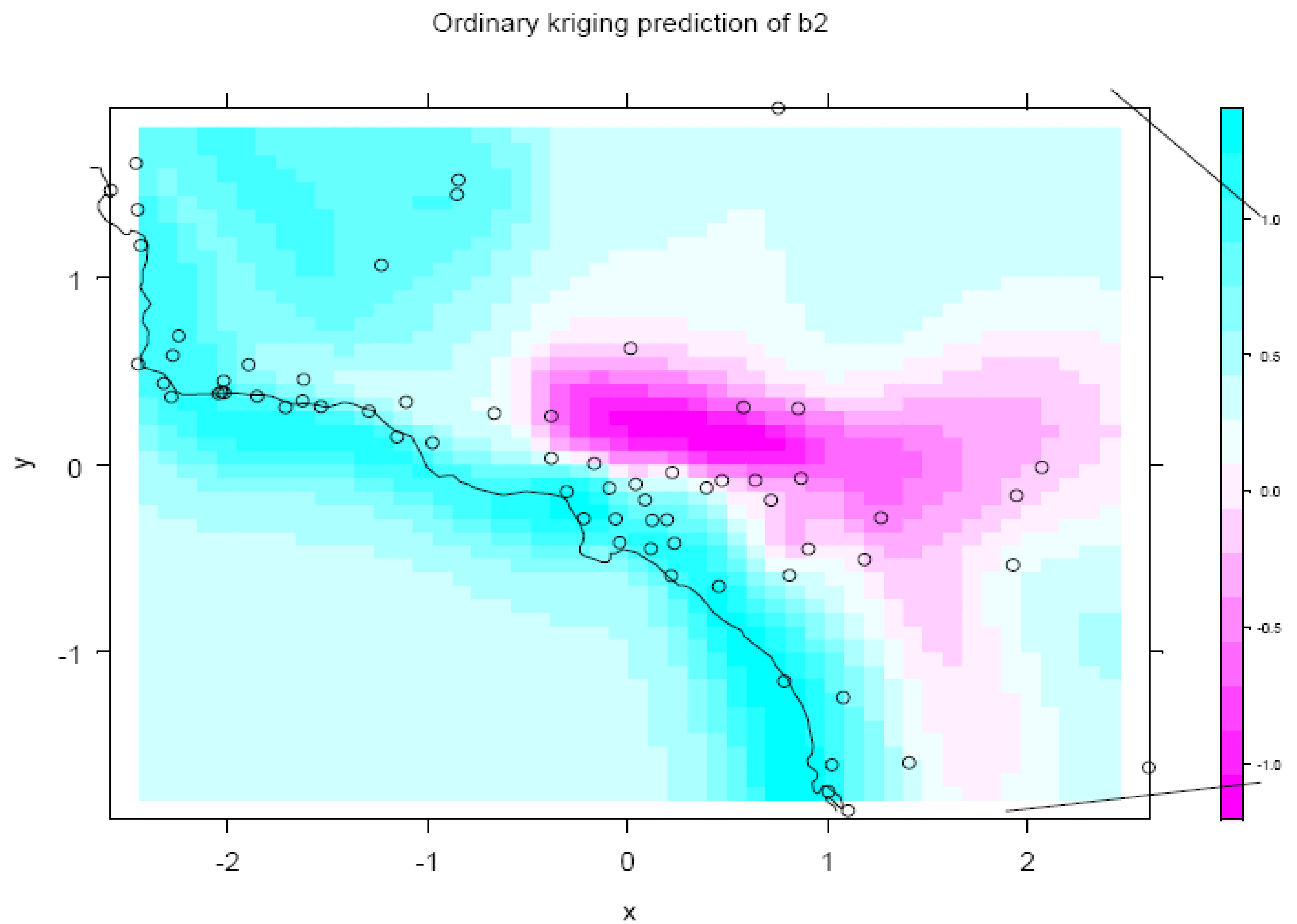
Possible approach to spatial estimation of vector of coefficients: “linear model of coregionalization” for multivariate spatial data.

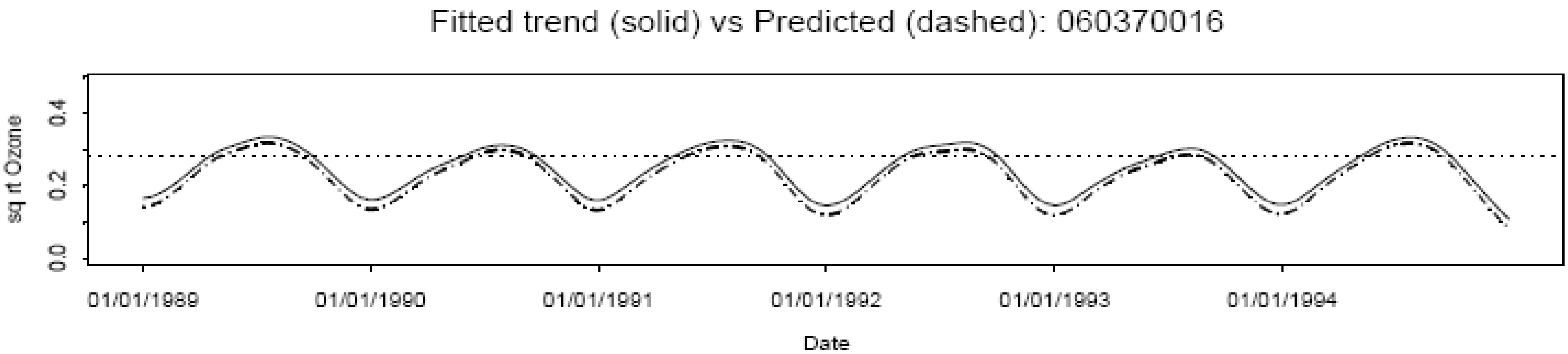


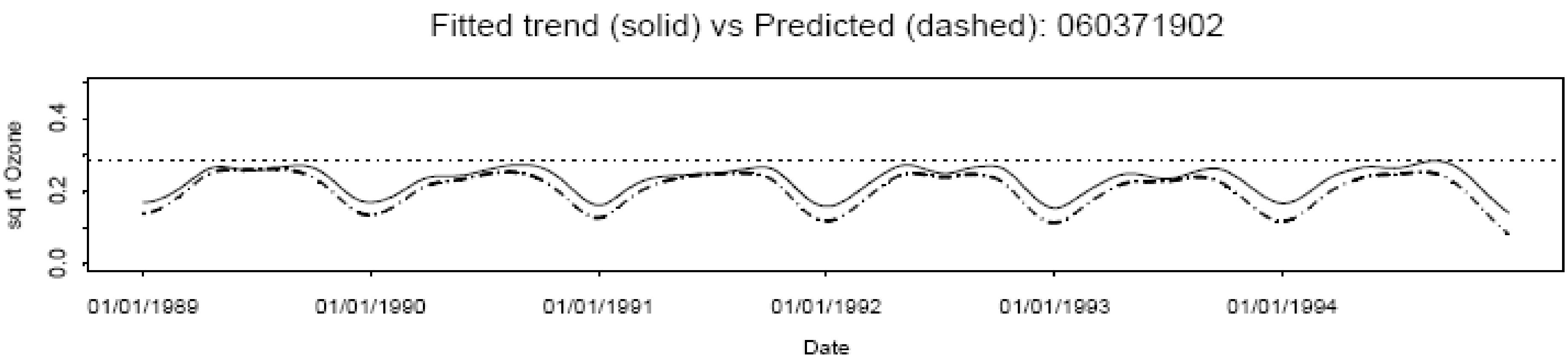
Ordinary kriging prediction of mu
y
1
0
-1
-2
-1
0
1
2
x
0.25
0.24
0.23
0.22
0.21
0.20
0.19
0.18
0.17

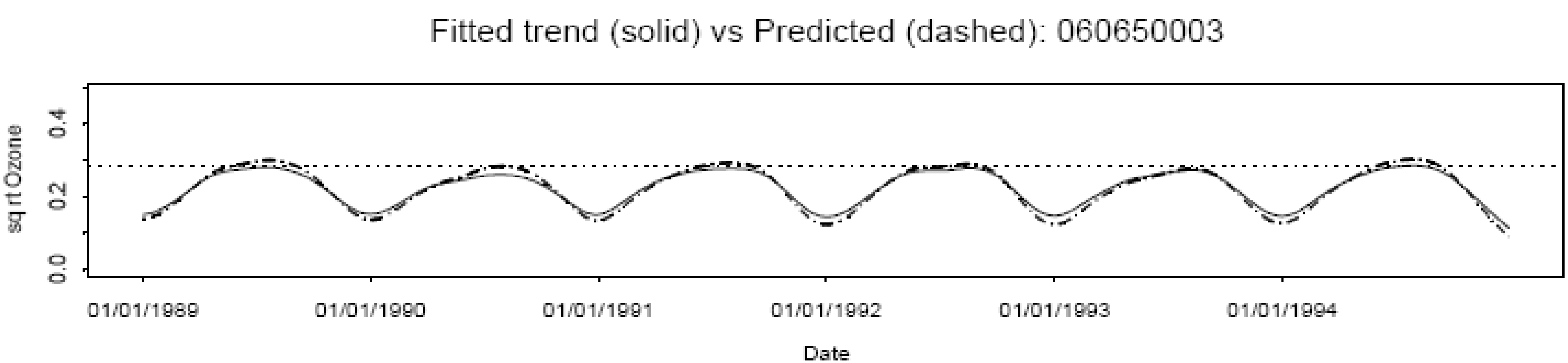


Ordinary kriging prediction of b1
y
x
1
0
-1
-2
-1
0
1
2
0.0
-0.5
-1.0
-1.5
-2.0
-2.5
-3.0
-3.5

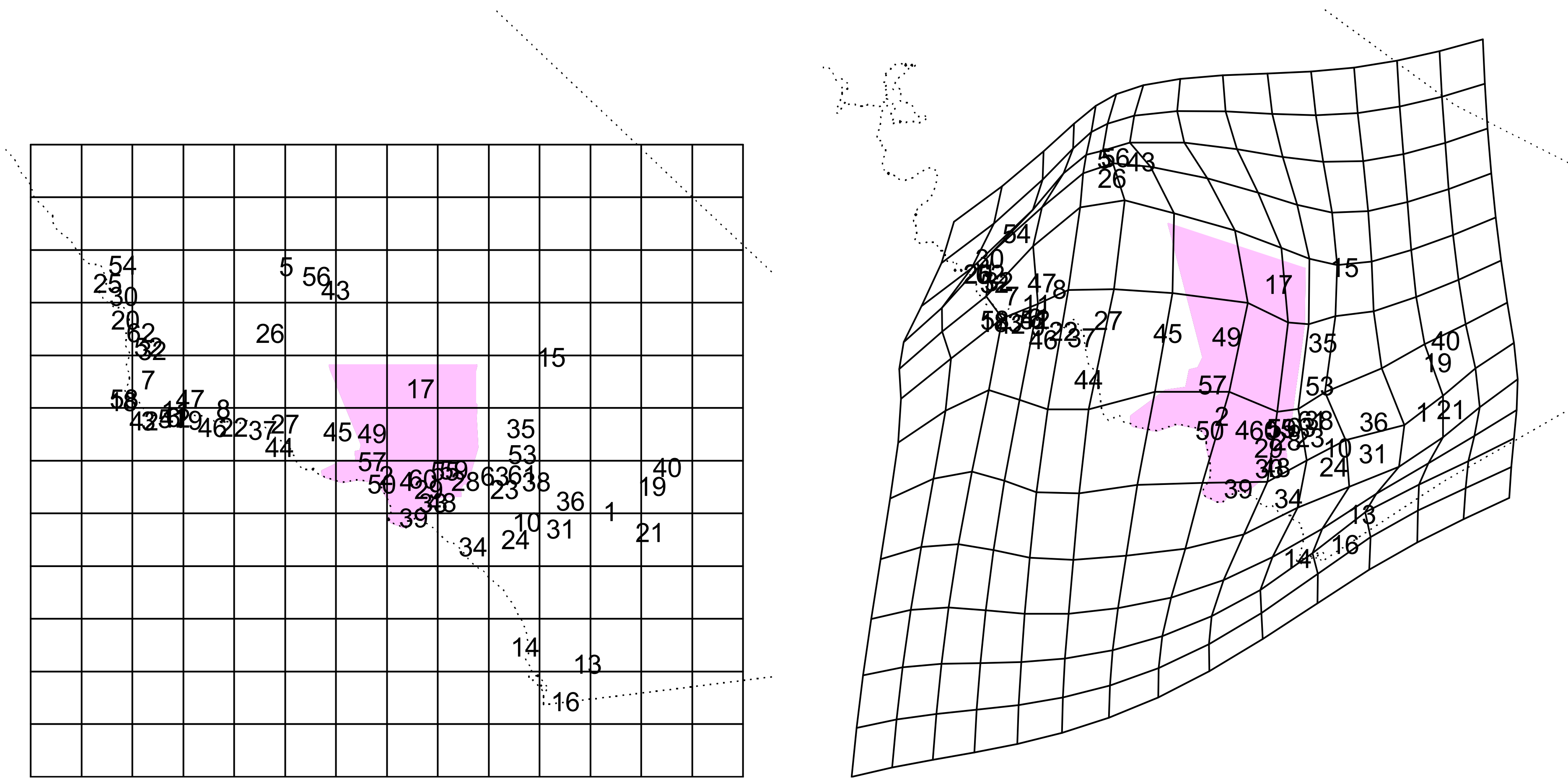

Ordinary kriging prediction of b2
x
y
-2
-1
0
1
2
1
0
-1
1.0
0.5
0.0
-0.5
-1.0


Fitted trend (solid) vs Predicted (dashed): 060370016
sq rt Ozone
0.0
0.2
0.4
01/01/1989
01/01/1990
01/01/1991
01/01/1992
01/01/1993
01/01/1994
Date

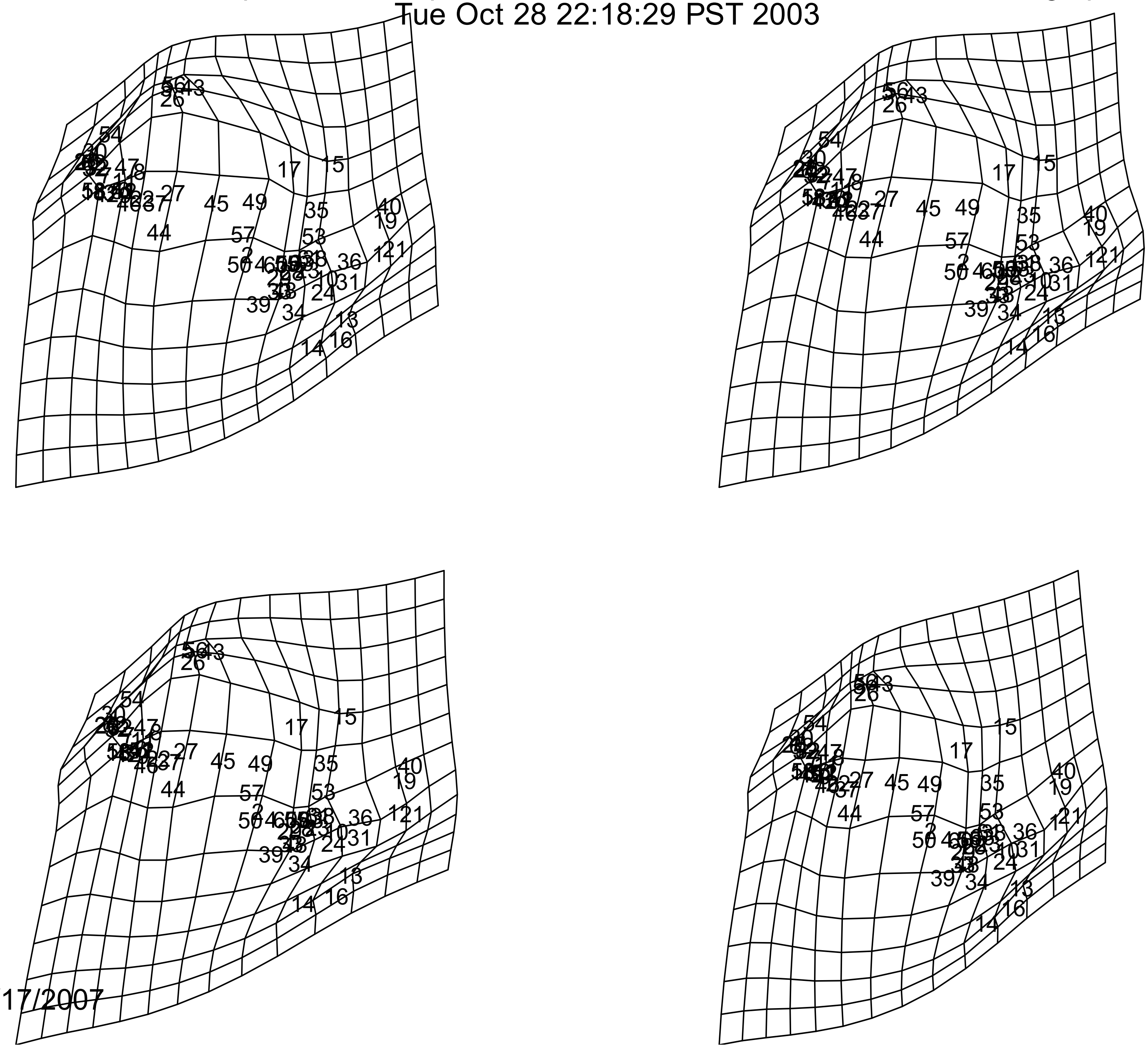

Fitted trend (solid) vs Predicted (dashed): 060371902
sq rt Ozone
0.0
0.2
0.4
01/01/1989
01/01/1990
01/01/1991
01/01/1992
01/01/1993
01/01/1994
Date

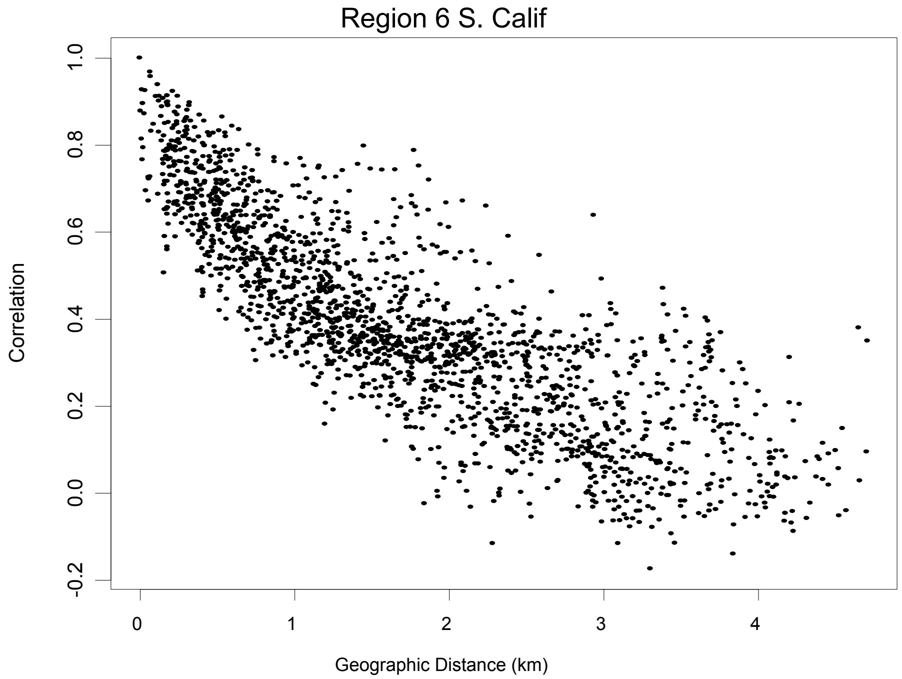

Fitted trend (solid) vs Predicted (dashed): 060650003
sq rt Ozone
0.0
0.2
0.4
01/01/1989
01/01/1990
01/01/1991
01/01/1992
01/01/1993
01/01/1994
Date

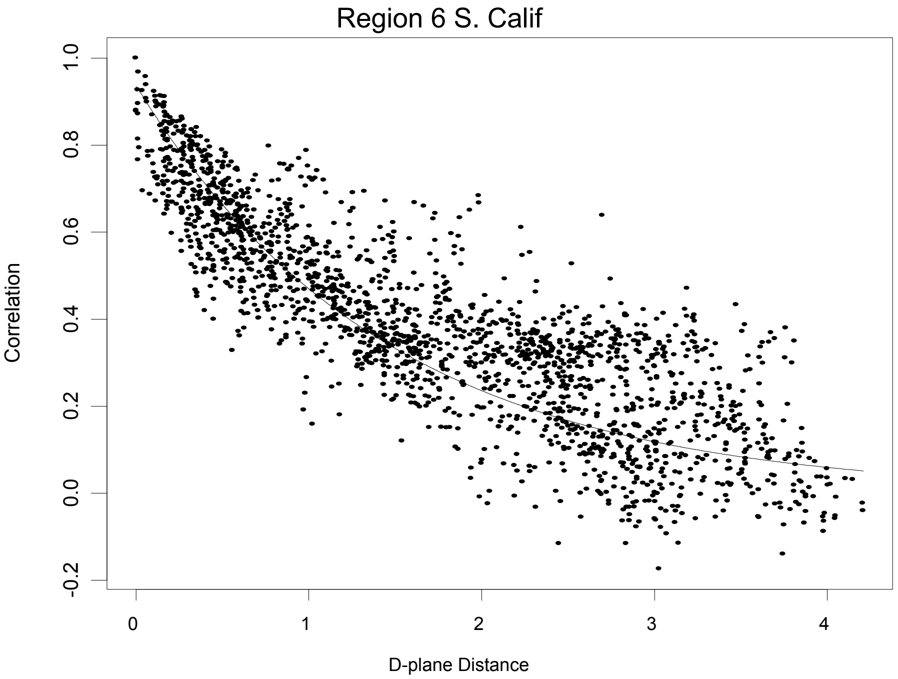
63 Region 6 monitoring sites and their representation in a deformed coordinate system reflecting spatial covariance
Thu Oct 30 00:12:36 PST 2003

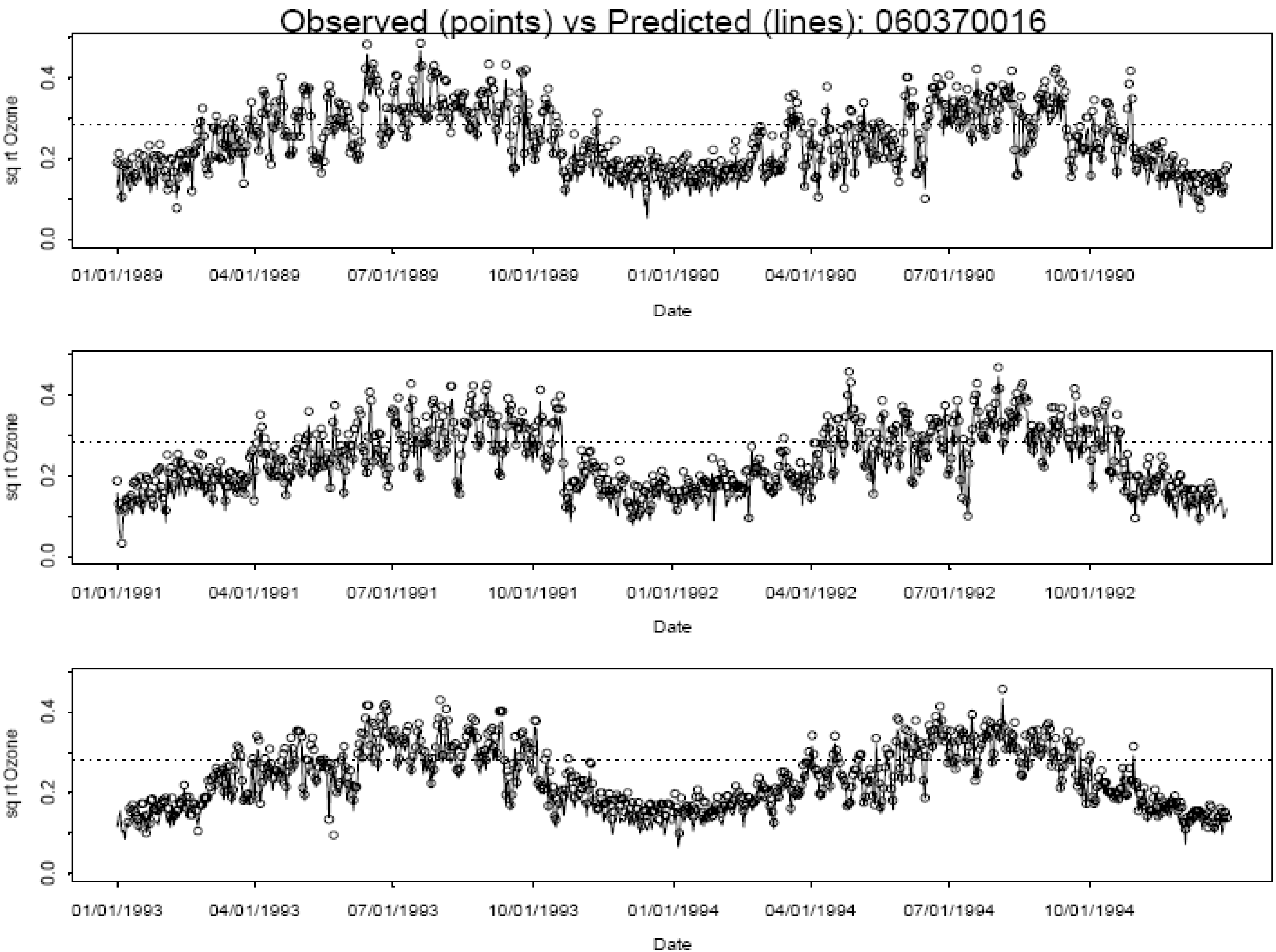


N=63, S. Calif: 4 samples from the posterior distribution of deformations reflecting spatial covariance
Tue Oct 28 22:18:29 PST 2003

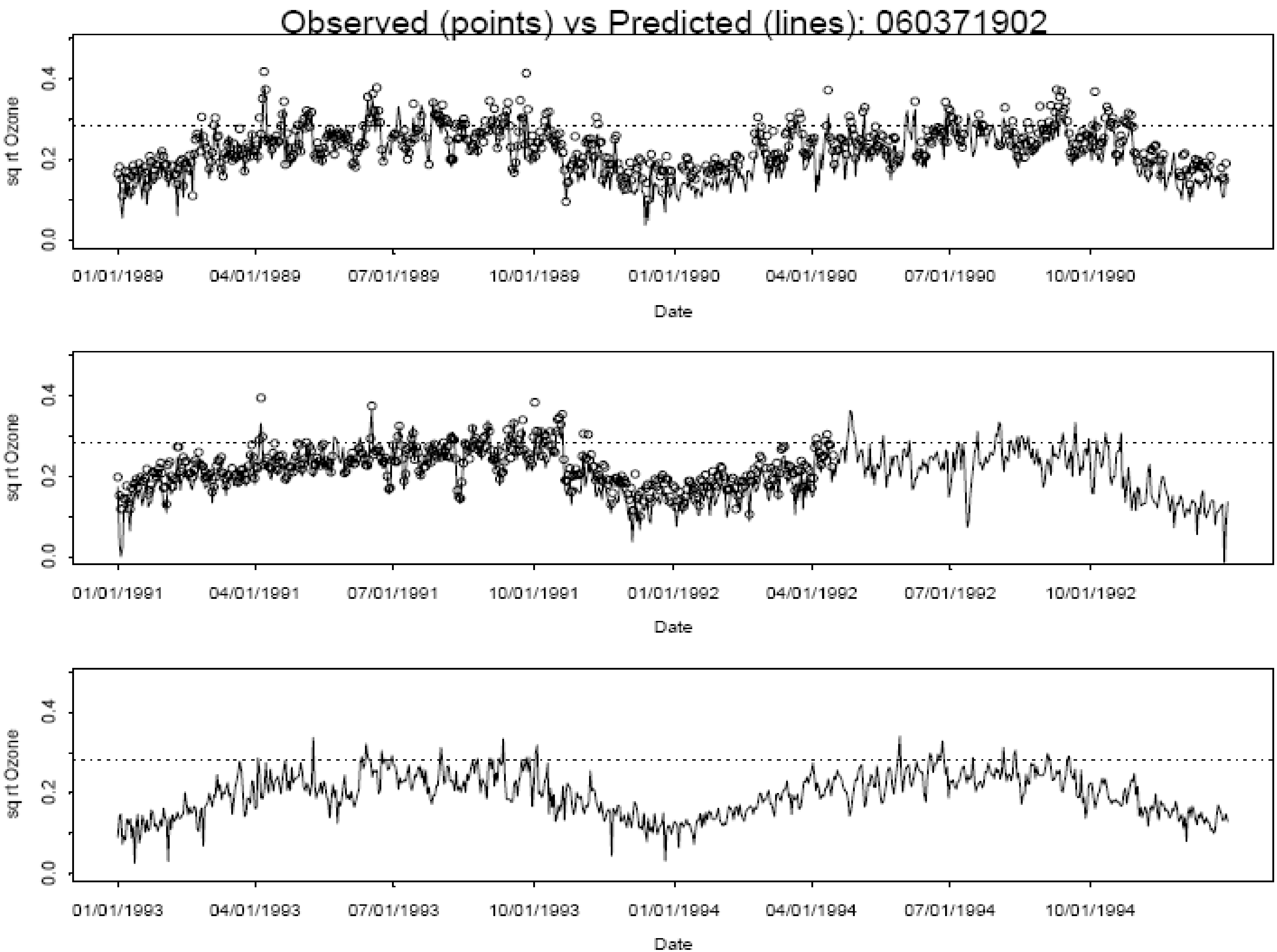



Region 6 S. Calif
Correlation
Geographic Distance (km)
-0.2
0.0
0.2
0.4
0.6
0.8
1.0
0
1
2
3
4

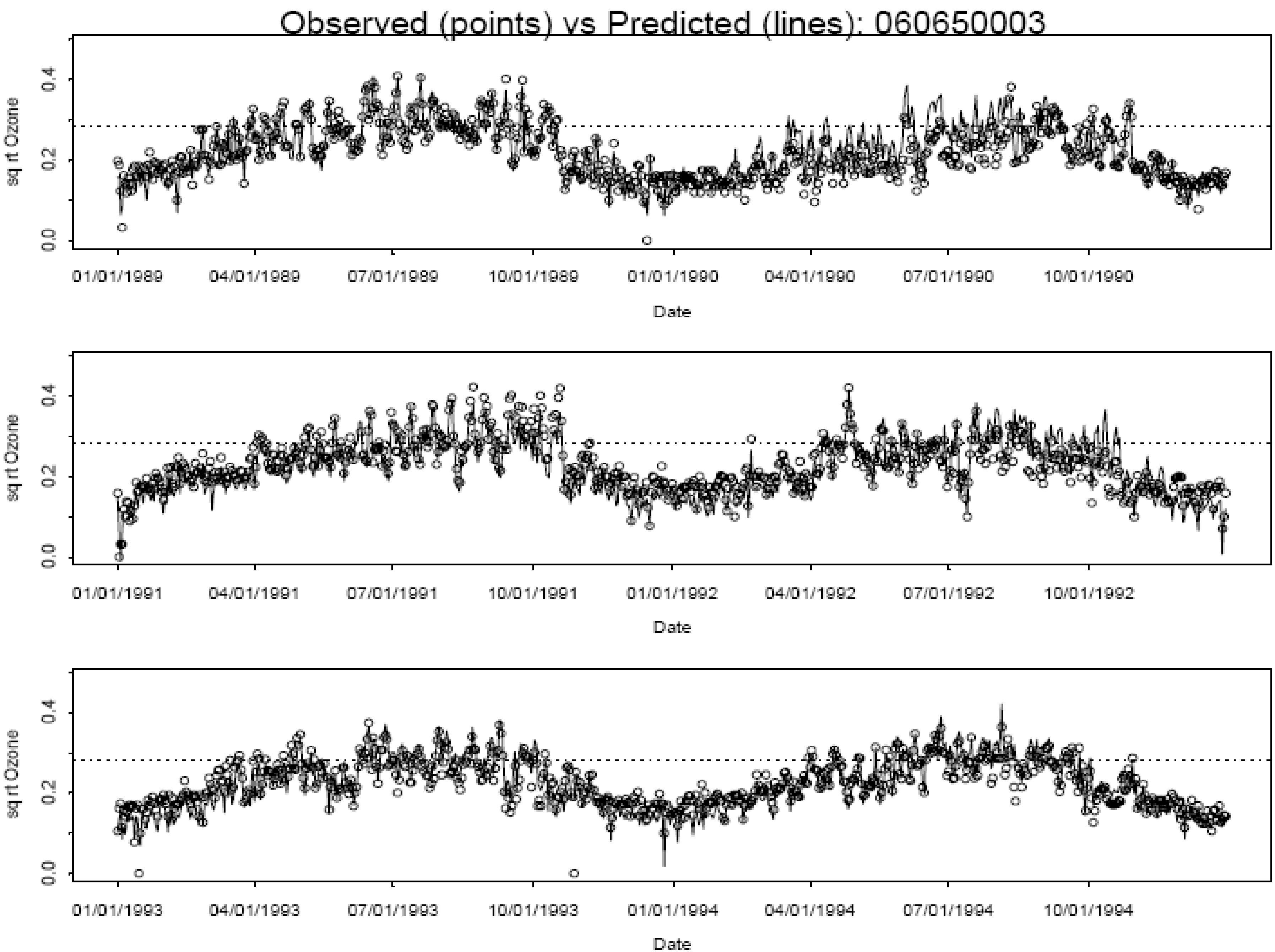

Region 6 S. Calif
Correlation
D-plane Distance
-0.2
0.0
0.2
0.4
0.6
0.8
1.0
0
1
2
3
4



Observed (points) vs Predicted (lines): 060370016
sq rt Ozone
0.0
0.2
0.4
01/01/1989
04/01/1989
07/01/1989
10/01/1989
01/01/1990
04/01/1990
07/01/1990
10/01/1990
Date
01/01/1991
04/01/1991
07/01/1991
10/01/1991
01/01/1992
04/01/1992
07/01/1992
10/01/1992
01/01/1993
04/01/1993
07/01/1993
10/01/1993
01/01/1994
04/01/1994
07/01/1994
10/01/1994



Observed (points) vs Predicted (lines): 060371902
sq rt Ozone
0.0
0.2
0.4
01/01/1989
04/01/1989
07/01/1989
10/01/1989
01/01/1990
04/01/1990
07/01/1990
10/01/1990
Date
sq rt Ozone
0.0
0.2
0.4
01/01/1991
04/01/1991
07/01/1991
10/01/1991
01/01/1992
04/01/1992
07/01/1992
10/01/1992
Date
sq rt Ozone
0.0
0.2
0.4
01/01/1993
04/01/1993
07/01/1993
10/01/1993
01/01/1994
04/01/1994
07/01/1994
10/01/1994
Date


Observed (points) vs Predicted (lines): 060650003
sq rt Ozone
0.0
0.2
0.4
01/01/1989
04/01/1989
07/01/1989
10/01/1989
01/01/1990
04/01/1990
07/01/1990
10/01/1990
Date
sq rt Ozone
0.0
0.2
0.4
01/01/1991
04/01/1991
07/01/1991
10/01/1991
01/01/1992
04/01/1992
07/01/1992
10/01/1992
Date
sq rt Ozone
0.0
0.2
0.4
01/01/1993
04/01/1993
07/01/1993
10/01/1993
01/01/1994
04/01/1994
07/01/1994
10/01/1994
Date

MESA Air PM2.5 preliminary analyses

- Trend basis functions in `TrendBases.all.ncomp4.pdf`
- Monitoring site fits in `TrendFits.lme4.pdf`

Recall Zidek, lect 3.2, slide #73

4.5 Hierarchical Bayesian Kriging

Le & Zidek (1992): A Bayesian alternative to kriging

Consider a simple setting - Same operational period

- Model: $Y_t \mid z_t, B, \Sigma \quad \sim \quad N_p(z_t B, \Sigma)$
- Prior: Conjugate

$$B \mid B_o, \Sigma, F \quad \sim \quad N_{kp}\left(B_o, F^{-1} \otimes \Sigma\right)$$

$$\Sigma \mid \Psi, \delta \quad \sim \quad W_p^{-1}(\Psi, \delta) \qquad \text{(inverted Wishart)}$$

- Predictive distribution - D observed data

$$Y_m^{(g)} \mid D \quad \sim \quad t_g\left(\mu_{gg}, \hat{\Psi}_{gg}, \delta + n - u - g + 1\right)$$

$$Y_m^{(u)} \mid Y_m^{(g)}, D \quad \sim \quad t_u\left(\mu_{u|g}, \hat{\Psi}_{u|g}, \delta - u + 1\right)$$

Comments on the LZ Hierarchical Bayesian Model

- Combine the response vectors Y_t as rows of a $t \times p$ matrix Y and do the same with the covariate vectors z_t Then the model

$$Y_t \mid z_t, B, \Sigma \quad \sim \quad N_p(z_t B, \Sigma)$$

can be written

$$Y = Z\mathrm{B} + E$$

- This is the same as our model written

$$C = F\mathrm{B} + \mathrm{N}$$

LZ Comments (cont)

- The “conventional” geostatistical approach fits a model to the “gauged” sites (my “S” sites = LZ’s “g” sites) and then in a 2nd step evaluates predictions at arbitrary new locations whereas LZ incorporate the target sites for prediction, the “u” ungauged sites, into the model from the outset.
- The spatial covariance structure of the regression parameters B in the LZ model is specified in the prior and is the <u>same</u> spatial covariance structure as the spatial covariance structure for the model responses, whereas my approach models the spatial covariance structure of the regression parameters B separately.
- In principal, the LZ hierarchical model could be extended to include a “land use regression” in the prior mean, B0, for the regression parameters, but the elegant analytical computations might be lost?