Spatio–temporal methods in environmental epidemiology:

Lectures 13 and 14

Overall outline

1. Temporal processes
2. Spatial processes or “fields”
   - Point referenced data
   - Area data
   - Point process data
3. Spatio–temporal processes

Review

Time - usually discrete index, $t = 1, \ldots, T$.
Spatial locations indexed by $s \in D$.

- **Point referenced data:** D = continuum or dense spatial grid; measurements made at irregular network of locations.
  E.g: ozone field
- **Lattice processes:** D = not necessarily regular grid of areal regions or specified locations D where measurements are made.
  E.g: death counts per county; centroids = lattice points
- **Point processes:** Measurements or "marks". made at randomly selected points in continuum D
  E.g: lightning strikes

Hierarchical modeling: Alternate formulation with

[X] = probability distribution of X

- $[parameters] = [\theta]$
- $[process|parameters] = [Y \mid \theta]$
- $[measurement|process, parameters] = [Z \mid Y, \theta]$

Spatio–temporal processes

Spatio–temporal modeling

Handling time.

- Depends on random response paradigm: point referenced; lattice; point process.
- Active area of current development

General approaches to incorporating time

Approach 1: Treat continuous time as like another spatial dimension with stationarity assumptions. Eg. Spatio–temporal Kriging[1]. **NOTE:** Constructing covariance models is more involved[2]. Note lose advantage of time ordering.

Approach 2: Integrate spatial fields over time. Eg. Given a spatial lattice let **Y(t)** : $m \times 1$ be vectors of spatial responses at lattice points. Eg. use multivariate autoregression.

Approach 3: Integrate times series across space. For a temporal lattice let **Y(s)** : $1 \times T$ be vector of temporal responses at - use multivariate spatial methods. Eg.co–Kriging; BSP.

[1] Bodnar and Schmid [2010]

[2] Fuentes et al. [2008]

Specialized approaches

Approach 4: Build a statistical framework on physical models that describe the evolution of physical processes over time

Features of a good spatio-temporal process theory

A good theory should:

- incorporate all sources of uncertainty[3]
- should come with a good theory of measurement - based inference
- admit a multivariate extension
- be computationally feasible to implement
- come with a theory about how to measure it - optimal design
- produce well calibrated predictive distribution error bands – 95% should be 95%!!!

[3] See Appendix B

Basic issues

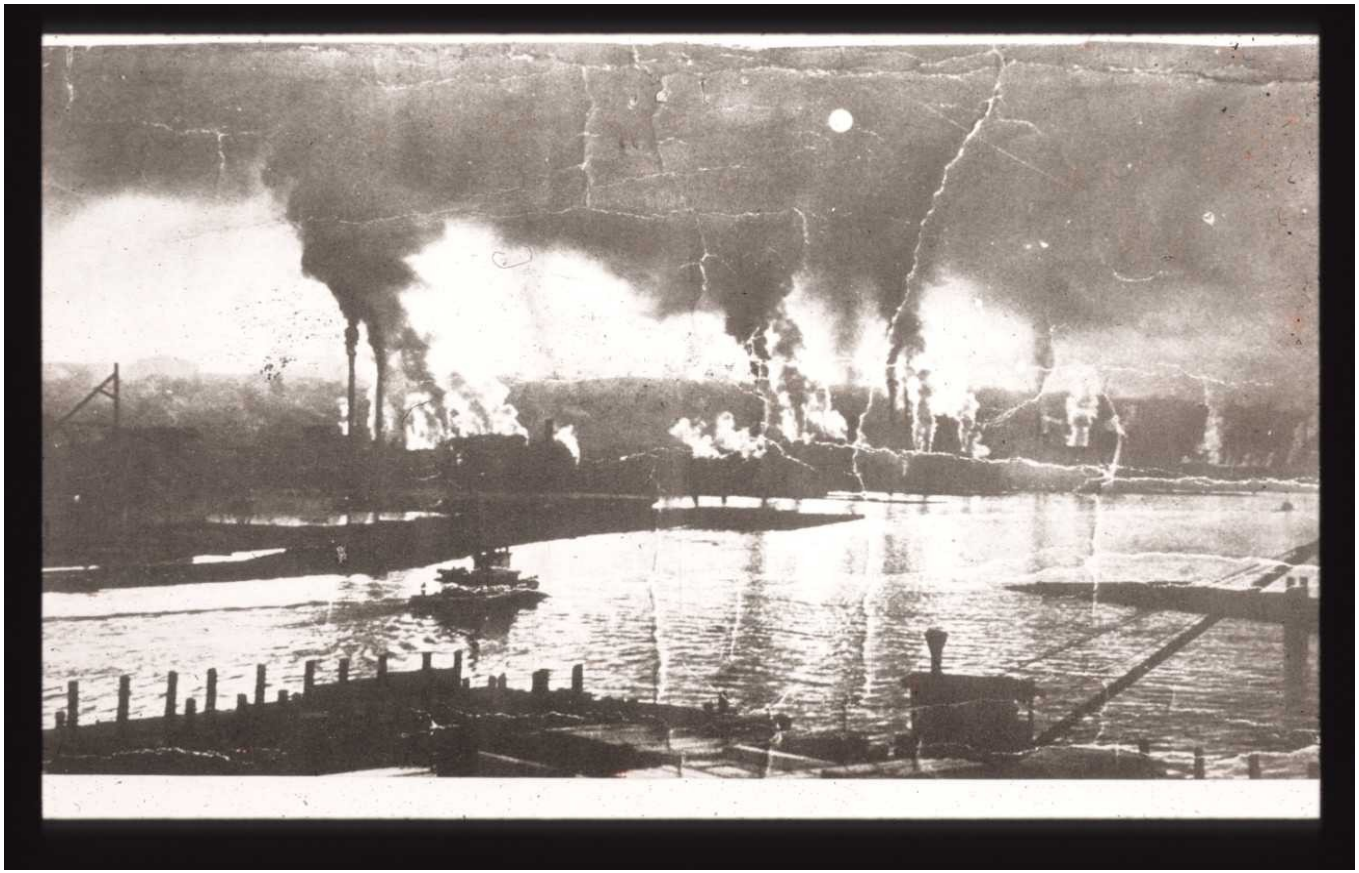
- Point referenced, lattice or point process?
- Temporal forecasting or spatial prediction?
- Association in space versus association in time?
  - Spatially linked temporal processes?
  - Temporally linked spatial processes?
- Continuous time vs. discretized time?
- Time $\neq$ space. Time order has advantages. Simplistic extensions of geostatistical models lose them.

Representing responses

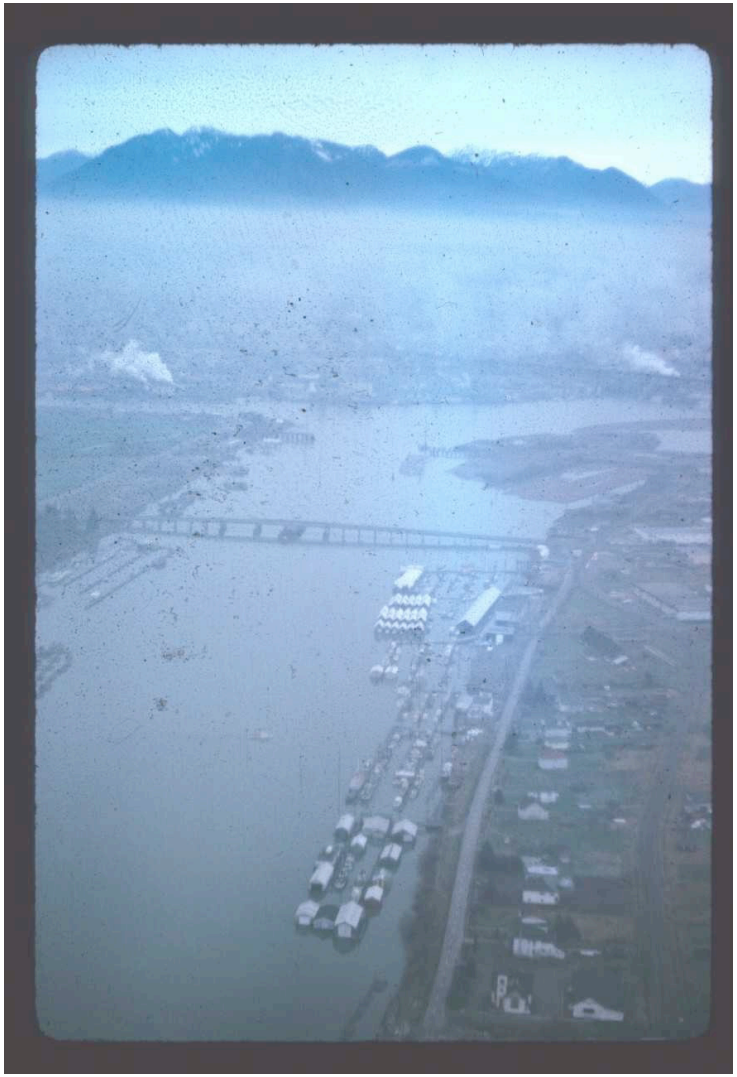
Processes fields are commonly situated on continuous temporal and spatial domains. In contrast measurements are restricted to a finite set of possible points in space and times. Time points will usually be equally spaced for administrative reasons. Potential spatial locations for measurements will often be restricted to a relatively small number of locations for the same reason. In this case, it may be technically simpler to work with a finite space time domain time $t = 1, \ldots, T$ and sites $s = 1, \ldots, p$ or $s = s_1, \ldots, s_p$ depending on context/convenience.

Example: Air pollution in Metro Vancouver

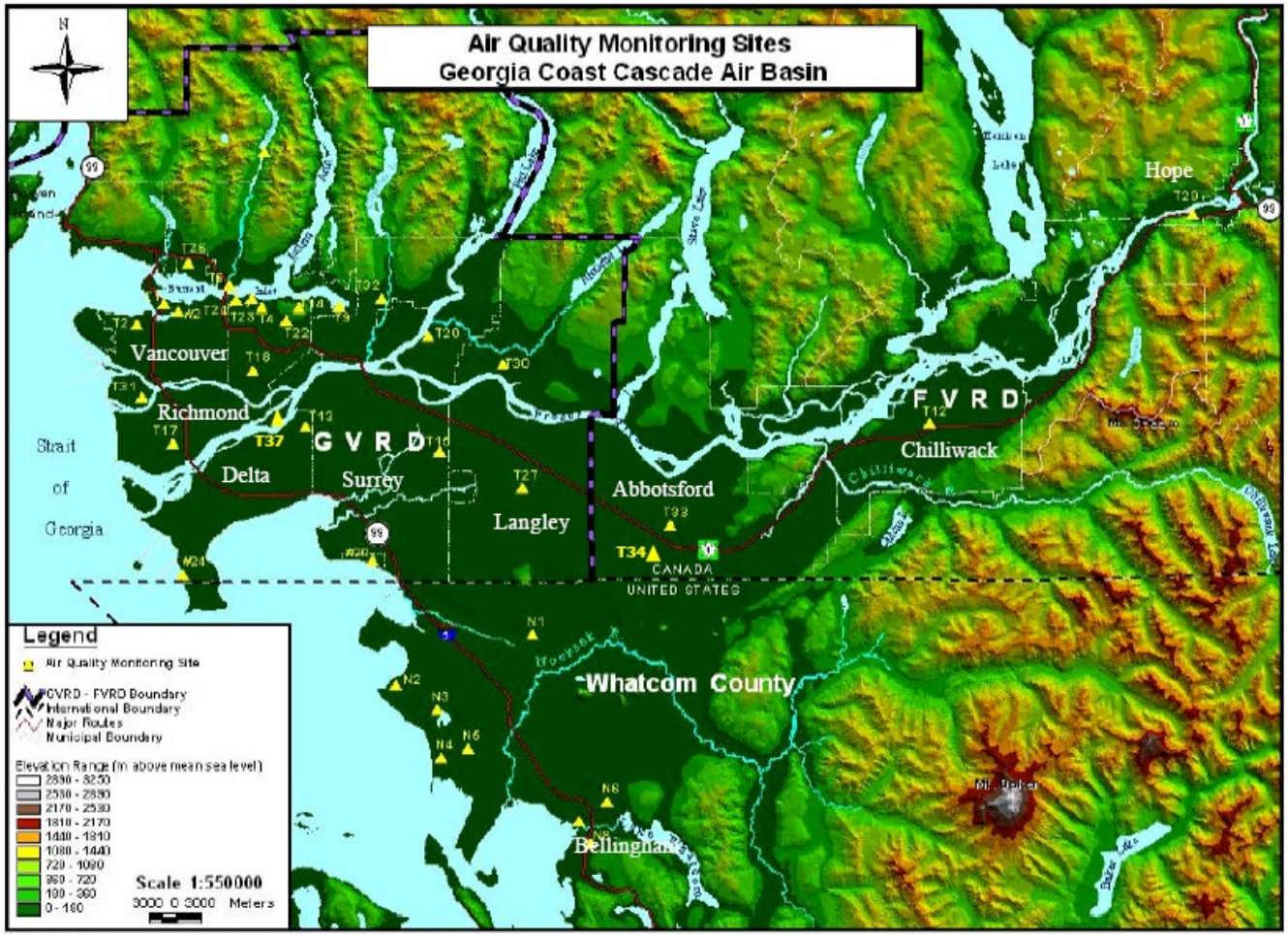
Coutesy of Professor Down Steyn, Earth & Ocean Sciences, UBC.
First slide shows False Creek in the early 1900s.



Haze over Burnaby in early 1900s.



Modern Metro Vancouver monitoring site locations. They yield hourly measurements.



NOTE: Detailed elevation data for the area north of the FVRD was unavailable at the time of map preparation. This gives the topography a flat appearance.

Monitor at Robson Square.

Another at Kitsilano High School.

Representing the random process responses

When $t = 1, \dot{s}, T$ and sites $s = 1, \ldots, p$ we may take

$$Y = \begin{pmatrix} Y_{11} & \cdots & Y_{1p} \\ \vdots & \vdots & \vdots \\ Y_{T1} & \cdots & Y_{Tp} \end{pmatrix}$$

Separability of time & space

Time & space[4] may be separable for Gaussian processes. In covariance form this means:

$$Cov(Y_{ts}, Y_{t's'}) = \sigma^2 \rho_1(i, j')\rho_2(s, s')$$

with $s = 1, \ldots, p$ & $t = 1, \ldots, T$. Then in matrix form, we get Kronecker product form:

$$\boldsymbol{\Sigma}^{Tp \times Tp} = \sigma^2 \boldsymbol{\rho}_1^{T \times T} \otimes \boldsymbol{\rho}_2^{p \times p}$$

This condition simplifies things a lot. Non-separable processes are difficult to understand and model. Example later of “correlation leakage”.

[4] Gneiting et al. [2006]

Using the Kronecker product

In general for any two matrices and $A : R \times T$ and $B : S \times p$ their Kronecker product $A \otimes B$ is defined as a linear operator acting on the space of response matrices, $\{Y : T \times p\}$ as follows:

$$(A \otimes B)Y = AYB'$$

It then follows that $(A \otimes B)' = A' \otimes B'$ and when A and B are nonsingular that $(A \bigotimes B)^{-1} = A^{-1} \otimes B^{-1}$. (**Exercise**)

Thus if[5]

$$Y : T \times p \sim N(\mu, \sigma^2 \rho_1 \otimes \rho_2)$$

$$(\rho_1^{-1/2} \otimes I) Y \sim N((\rho_1^{-1/2} \otimes I)\mu, \sigma^2 (I \otimes \rho_2)).$$

So if ρ_1 were known or well estimated (say when T is large or this covariance depends on a small number of parameters like having an AR(1) structure) , we could assume the between row correlations are approximately zero. The the temporal process of successive spatial fields could be viewed as iid.

[5] See Appendix A

Spatio–temporal process models - a simple version

- Gaussian point-referenced data, t continuous

$$Y_t(s) = \mu_t(s) + w_t(s) + \epsilon_t(s)$$

- Non-Gaussian data, instead use appropriate likelihood with link

$$g(E(Y_t(s))) = \mu_t(s) + \omega_t(s)$$

- Modeling: $Y_t(s) = \mu_t(s) + \omega_t(s) + \epsilon_t(s)$, or $g(E(Y_t(s)) = \mu_t(s) + \omega_t(s)$
- $\epsilon_t(s) \sim indN(0,' \tau_t^2)$
- For $\omega_t(s)$
  - $\omega_t(s) = \alpha_t + \omega(s)$
  - $\omega_t(s)$ independent for each t
  - $\omega_t(s) = \omega_{t-1}(s) + \eta_t(s)$ independent spatial process innovations

Dynamic spatio–temporal model

Measurement model:

$$Y_t(s) = \mu_t(s) + \epsilon_t(s),\ \epsilon_t(s) \sim indN(0, \sigma^2_\epsilon)$$
$$\mu_t(s) = x'_t(s)\tilde{\beta}_t(s)$$
$$\tilde{\beta}_t(s) = \beta_t + \beta_t(s)$$

Process model:

$$\beta_t = \beta_{t-1} + \eta_t,\ \eta_t \sim indN(0, \Sigma_\eta)$$
$$\beta_t(s) = \beta_{t-1}(s) + w_t(s)$$

where $w_t(s) = Av_t(s)$, $v_t(s) = v_{(1:n)t}(s)$.

The $v_{lt}(s)$ are replications of a Gaussian processes with unit variance and correlation function $\rho_l((\rho_l)$

DLM: General version

- Background:
  - Goes back to Kalman-Bucy filter
  - Developed for statistics by Harrison & Stevens[6]
  - Much extended since. (See book by Harrison and West, 1997)
  - made practical by modern computational tools
- Very general/flexible
- We use variation of Stroud, Muller, Sanso[7] & Huerta, Sanso, Stroud[8]

[6] Harrison and Stevens [1971]
[7] Stroud et al. [2001]
[8] Huerta et al. [2004]

Data: $\mathbf{Z_t}$: $n \times 1$ $t = 1, 2, \ldots$. The "observation" and "evolution equations":

- $$\mathbf{Z_t} = F_t' \mathbf{Y_t} + \nu_\mathbf{t}, \quad \nu_\mathbf{t} \sim N[\mathbf{0}, V_t], \tag{1}$$

- $$\mathbf{Y_t} = G_t \mathbf{Y_{t-1}} + \omega_\mathbf{t}, \quad \omega_\mathbf{t} \sim N[\mathbf{0}, W_t], \tag{2}$$

  $F_t : p \times n$, $G_t : p \times p$, $V_t : n \times n$, & $W_t : p \times p$ being known matrices.

Notes: F_t= "design matrix"; $\mathbf{Y_t}$ = process (or state) vector; $\nu_\mathbf{t}$ = observational error; G_t = state matrix; $\omega_\mathbf{t}$=evolution error with evolution matrix W_t.

DLM Continued

DLM's specification completed by specifying initial information:

$$(\mathbf{Y_0}|\mathbf{Z_0}) \quad \sim \quad N[\mathbf{m_0}, C_0].$$

Forward filtering - backward sampling

"Forward filtering":

$$
\begin{aligned}
(\mathbf{Y_{t-1}}|Z_{1:t-1},\theta) &\sim N[\mathbf{m_{t-1}}, C_{t-1}] \\
(\mathbf{Y_t}|Z_{1:t-1},\theta) &\sim N[\mathbf{a_t}, R_t] \\
(\mathbf{Z_t}|Z_{1:t-1},\theta) &\sim N[\mathbf{f_t}, Q_t] \\
(\mathbf{Y_t}|Z_{1:t},\theta) &\sim N[\mathbf{m_t}, C_t], \text{ where}
\end{aligned}
$$

$$
\begin{aligned}
\mathbf{a_t} &= G_t \mathbf{m_{t-1}} & R_t &= G_t C_{t-1} G_t' + W_t \\
\mathbf{f_t} &= F_t' \mathbf{a_t} & Q_t &= F_t' R_t F_t + V_t \\
\mathbf{e_t} &= \mathbf{Z_t} - \mathbf{f_t} & A_t &= R_t F_t Q_t^{-1} \\
\mathbf{m_t} &= \mathbf{a_t} + A_t \mathbf{e_t} & C_t &= R_t - A_t Q_t A_t'.
\end{aligned}
$$

Forward filtering - backward sampling

"Backward sampling": Let $B_t = C_t G'_{t+1} R^{-1}_{t+1}$.

- For $0 \leq k \leq T-1$,

$$(\mathbf{Y}_{T-k} | Z_{1:T}, \theta) \sim N[\mathbf{a}_T(-k), R_T(-k)], \tag{3}$$

- where

$$\begin{aligned} \mathbf{a_T}(-k) &= \mathbf{m}_{T-k} + B_{T-k}[\mathbf{a_T}(-k+1) - \mathbf{a}_{T-k+1}] \\ R_T(-k) &= C_{T-k} + B_{T-k}[R_T(-k+1) - R_{T-k+1}]B'_{T-k} \text{ with} \end{aligned}$$

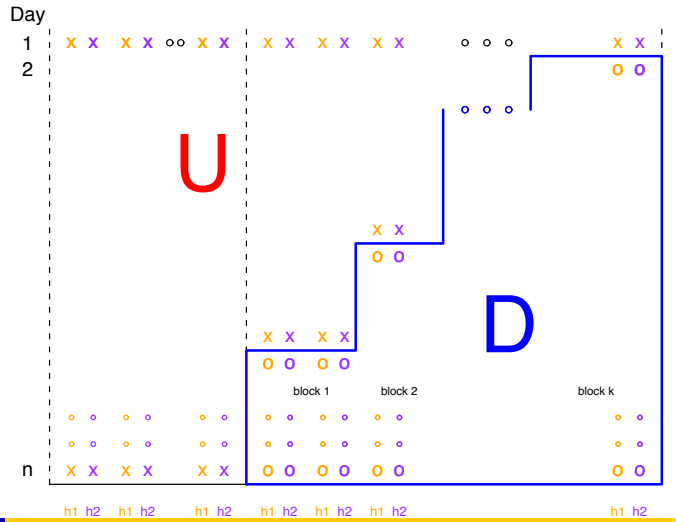
$\mathbf{a}_T(0) = \mathbf{m}_T$, $R_T(0) = C_T$; $\mathbf{a}_{T-k}(1) = \mathbf{a}_{T-k+1}$, & $R_{T-k}(1) = R_{T-k+1}$.

Bayesian Spatial Prediction (BSP) approach

Multivariate response: Can be used for vectors of chemical species. But we use it for hourly ozone by stacking hourly ozone concentrations into a vector for each day.

$$Y = \left[Y_{ij(p)} \right] = \left[Y^{[u]}, \begin{pmatrix} Y^{[g_1^m]} \\ Y^{[g_1^o]} \end{pmatrix}, \cdots, \begin{pmatrix} Y^{[g_k^m]} \\ Y^{[g_k^o]} \end{pmatrix} \right]$$

The dataset is assumed to have a staircase design, where “U” means ungauged site or missing data while “G” means gauged sites. Graphical display.



Why multivariate?

After all, ozone in the application below is a univariate response

Answers:

- reduces need to model fine scale autocorrelation within the day
- partially side-steps correlation leakage problem described below.
  - Occurs:
    - when time & space are not separable
    - even when separable due to imperfect correlation modeling + sampling error
  - More later

Multivariate BSP - Model

Model

$$\begin{cases} Y \mid \beta, \mathbf{\Sigma} \sim N(Z\beta, A \otimes \mathbf{\Sigma}) \\ \beta \mid \mathbf{\Sigma}, \beta_0, F \sim N(\beta_0, F^{-1} \otimes \mathbf{\Sigma}) \\ \mathbf{\Sigma} \sim GIW(\Theta, \delta) \end{cases}$$

A: Is assumed to be known

Example: For many pollutant monthly averages: $A = I_n$

Special case - no staircase

$$\mathbf{\Sigma} \sim IW(\Psi, \delta)$$

Le & Zidek (1992-2002, 2006)

BSP: Predictive distribution

- $(Y_U \mid D, \mathcal{H}) \sim \left(Y^{[u]} \mid Y^{[g_1^m,\ldots,g_k^m]}, D, \mathcal{H}\right) \times$

$$\prod_{j=1}^{k-1} \left(Y^{[g_j^m]} \mid Y^{[g_{j+1}^m,\ldots,g_k^m]}, D, \mathcal{H}\right) \times \left(Y^{[g_k^m]} \mid D, \mathcal{H}\right)$$

  - Each component follows a matric-t distribution
    Mean, covariance, and df: functions of $\mathcal{H}$ and D
  - Completely characterized given $\mathcal{H}$ (all hyper's)

- $\mathcal{H}$: Empirical Bayes
  Computation simpler
  Separability: $\Psi = \Lambda \otimes \Omega$
  Non-stationarity (Sampson & Guttorp 1992)

Application: spatial prediction of the ozone field

Model for space - time field of hourly ozone concentrations

Why???

- Ozone linked to:
  - decrements in lung function (reduced levels of FEV
  - mortality (e.g. asthma, bronchitis)
  - possibly mortality (not certain)
- Thus ozone one of 6 criteria pollutants:
  - particulate matter, carbon monooxide, nitrogen dioxide, sulfur dioxide and lead
  - must be regulated to protect human health and human welfare (US Clean Air Act, 1970)

- To characterize outdoor human exposure [interpolate between center (ambient) monitoring site measurements]
- For input in computer models that account for indoor exposures such as:
  - **APEX**: developed by US Environmental Protection Agency (EPA) [interpolated values may yet be used?]
  - **pCNEM**: UBC model used to set air quality standards in Canada [interpolated values not used]
  - **SHEDS**: Another EPA model for particulate matter [interpolated values used]

DLM Approach

Model: $Y_t(s) = Z_t(s)'\beta_t + S_{1t}\alpha_{1t}(s) + S_{2t}\alpha_{2t}(s) + \epsilon_t(s)$

$$\beta_t = \beta_{t-1} + \omega_t, \qquad \alpha_{jt}(s) = \alpha_{j,t-1}(s) + \omega_{jt}(s)$$

- S_{jt} sine's + cosine's for 12hr and 24hr cycles
- α_{jt} for amplitudes
- $\text{Cov}(\epsilon_t) = \sigma_y^2 \exp(-D/\lambda_y)$; $D =$ intersite distance matrix (spatial smoothness)
- Parameters change dynamically - random walk

- $\text{Cov}(\omega_{jt}) = \sigma_y^2 \tau_j^2 \exp(-D/\lambda_j)$ spatial smoothness
- $\omega_t \sim N(0, \sigma_y^2 \tau_y^2)$
- Very flexible: incorporate trend, spatial, temporal correlation, etc directly via model parameters
- Build sub-models for meteorology, etc
- Implements via MCMC
- $\tau_j^2, \tau_y^2, \lambda_j$ fixed in advance - trial & error (not easy!)
  results sensitive to these.

DLM's Computational issues:

- **DLM very computational intensive**
  - Huerta et al (2004) apply it only to 10 Mexico City sites for 7 days (168)
  - We need it for about 300 sites over 120 days (2880 hours), yielding about **1.7 mi parameters**. Not feasible! Not scalable.
  - Our runs restricted to 10 sites to make them comparable to Huerta et al. but various speed ups make somewhat larger numbers feasible.
- **DLM hyperparameter modeling difficult**
  - sensitive to some of the hyper-parameter specs (bad)
  - random walk parameter model unsatisfactory

The BSP approach

Bayesian spatial prediction with prefiltering (**BSP**)

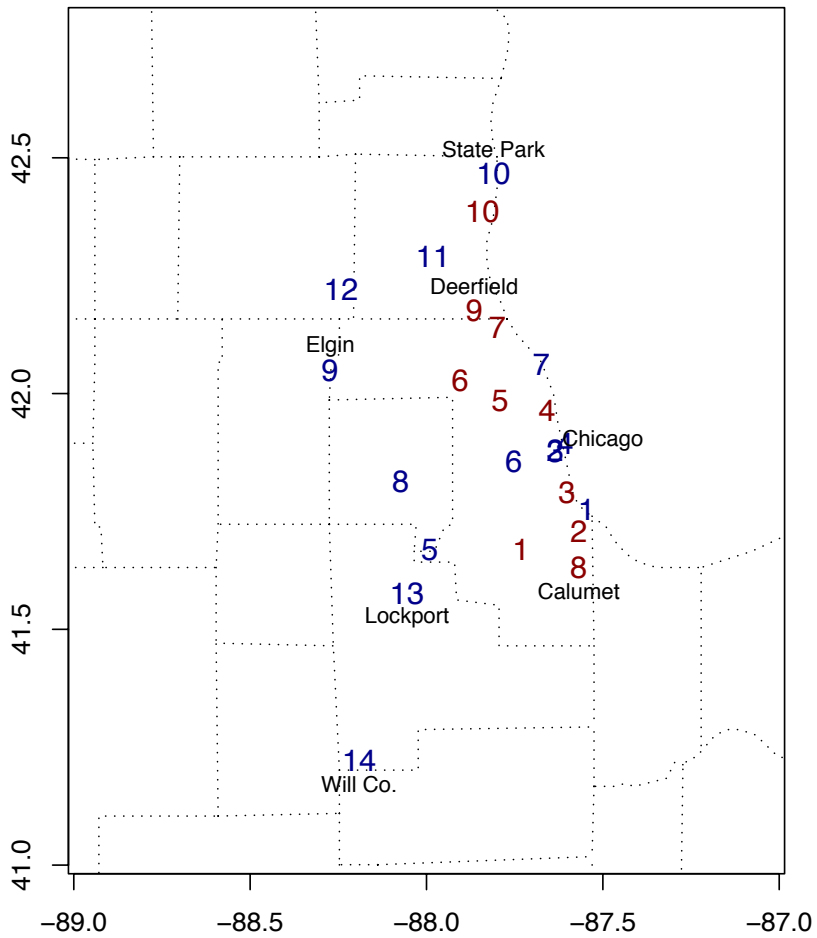
- Background:
  - developed by Le, Zidek + co-investigators (1992-2002)
  - can require prefiltering to remove
    - systematic components
    - autocorrelation
- Featured in Le & Zidek (2006)
- Software available at http://enviro.stat.ubc.ca
- Computationally efficient versus DLM

Application

Model the hourly ozone concentration field over Chicago by both DLM & BSP (Prefiltering) approaches

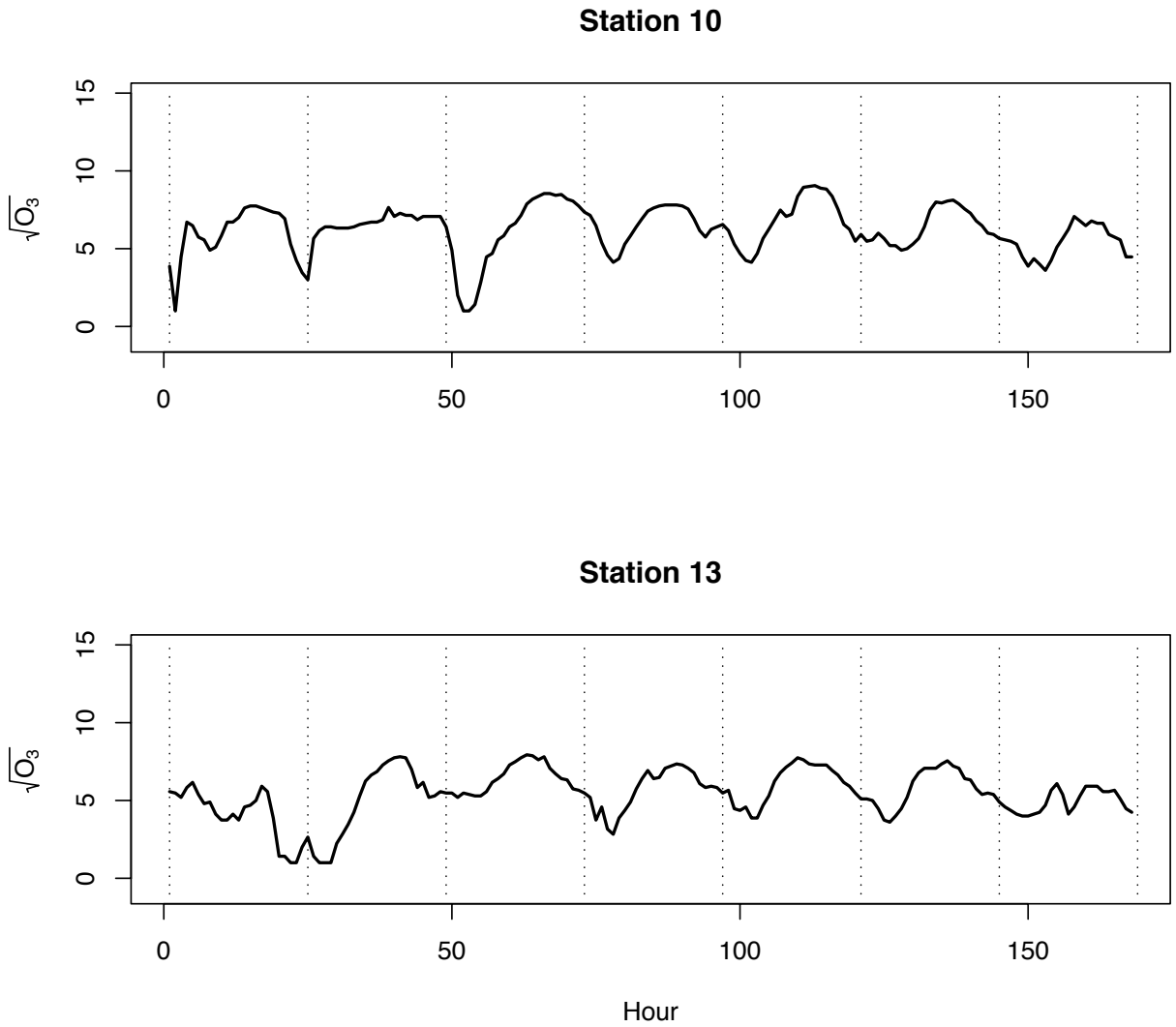
- Ozone data comes from the AIRS database
  - widely dispersed network of monitoring sites in the US
  - urban sites used in regulation for compliance
  - covers many species - **only ozone considered here**
- Requires $\sqrt{\text{transformation}}$ for distribution symmetry
- Includes periodic components - need removing for BSP (**detrending!**)

AIRS locations: Gauged & Ungauged

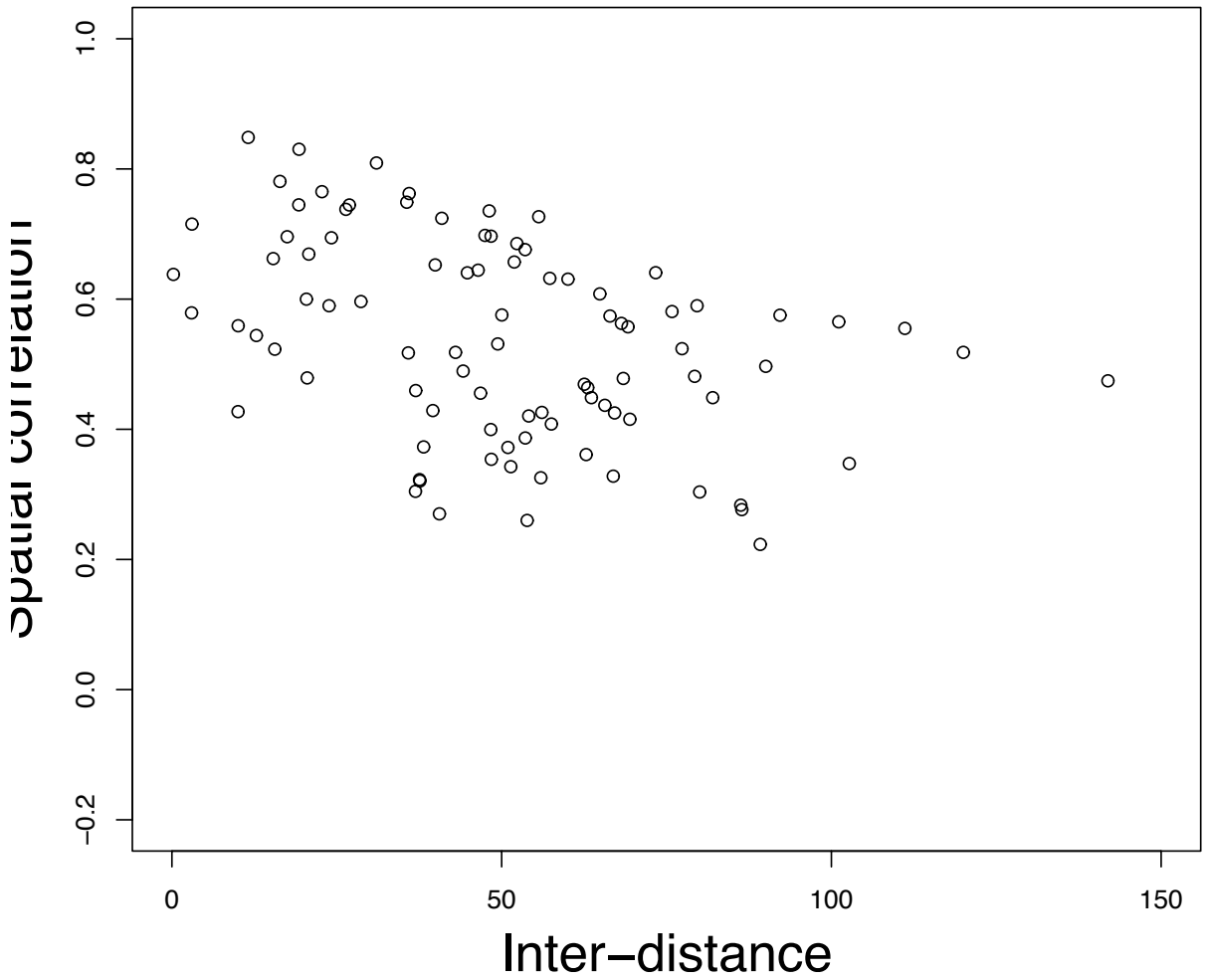


Hourly Observations: May 1-Aug 31, 2000

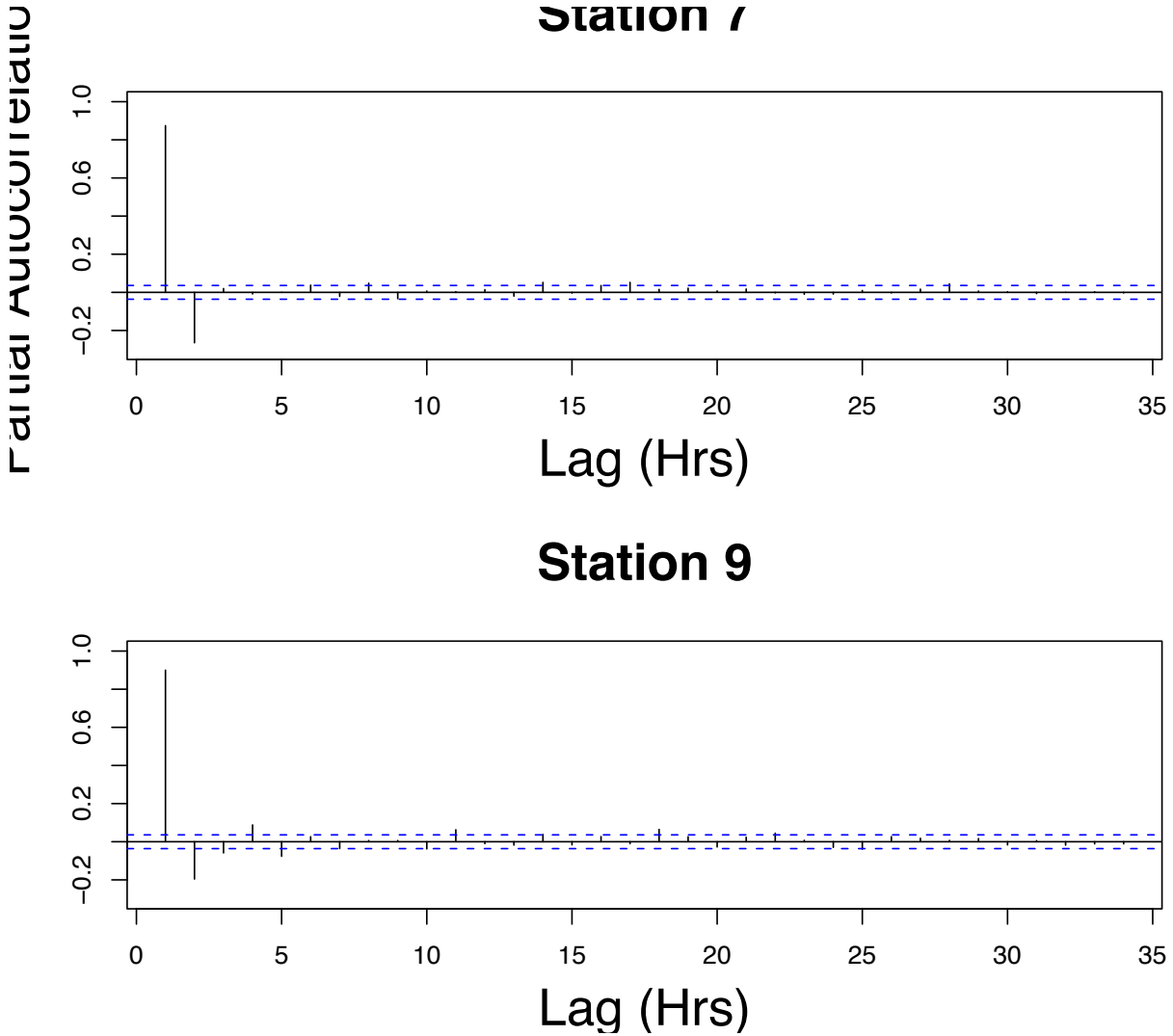
Week 1: Ozone levels (ppb)



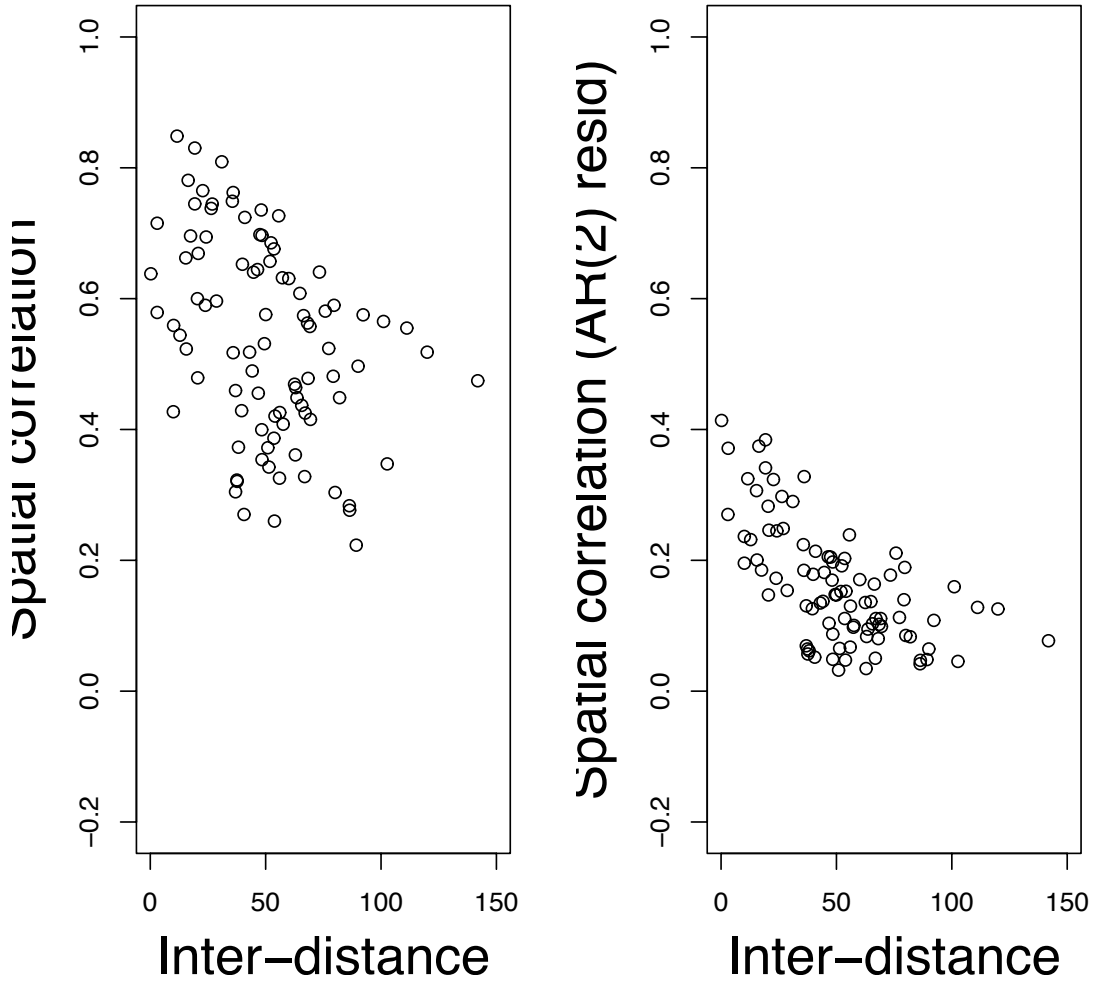
Spatial Correlation - detrended series



Partial autocorrelation - detrended series



Spatial Correlation Leakage



Spatial Correlation Leakage - Simple Case

AR(1) Model: $Z_t(s) = \alpha Z_{t-1}(s) + \epsilon_t(s)$

$\epsilon_t(s)-$ time independent with spatial correlation

Spatial correlation:

$$cor(\epsilon_t(s), \epsilon_t(s')) = cor(Z_t(s), Z_t(s')) \; -$$

$$\frac{\alpha}{\sqrt{(1-\alpha^2)}}[cor(Z_{t-1}(s), \epsilon_t(s') + cor(Z_{t-1}(s'), \epsilon_t(s)]$$

Cross-corr $= 0 \rightarrow cor(Z_t(s), Z_t(s')) = cor(\epsilon_t(s), \epsilon_t(s'))$

Correlation leakage occurs since sample ones $\neq 0$

- substantial when α is

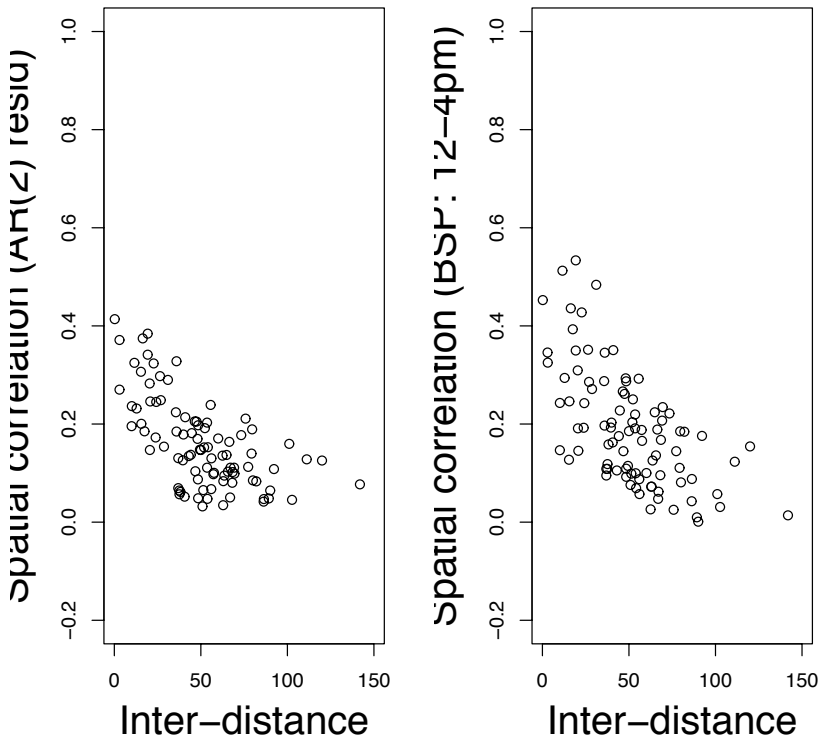
Implementing BSP

- Need to deal with A
  - assumed known in theory $Y \mid \beta, \mathbf{\Sigma} \sim N(Z\beta, A \otimes \mathbf{\Sigma})$
  - is estimated from data (prefiltering!) to get (approx) $A = I_n$ - empirical Bayes step
- multivariate responses: treat block of 5 hourly concentrations in single vector, day-by-day, then move to next block and repeat **24 times**
  - helps reduce correlation leakage
  - 5×1 vectors unauto-correlated over days

	Day			
	1	**2**	...	**123**
5 AM	$\begin{pmatrix} X \\ X \\ X \\ X \\ X \\ O \\ O \\ \vdots \\ O \end{pmatrix}$	$\begin{pmatrix} X \\ X \\ X \\ X \\ X \\ O \\ O \\ \vdots \\ O \end{pmatrix}$	... ...	$\begin{pmatrix} X \\ X \\ X \\ X \\ X \\ O \\ O \\ \vdots \\ O \end{pmatrix}$

- Need to deal with A
  - assumed known in theory $Y \mid \beta, \mathbf{\Sigma} \sim N(Z\beta, A \otimes \mathbf{\Sigma})$
  - is estimated from data (prefiltering!) to get (approx) $A = I_n$ - empirical Bayes step
- multivariate responses: treat block of 5 hourly concentrations in single vector, day-by-day, then move to next block and repeat **24 times**
  - helps reduce correlation leakage
  - 5×1 vectors unauto-correlated over days
- borrows strength across space & time: 5^{th} hour at unmonitored (ungauged) sites predicted from 4 previous hours at gauged sites as well as hour 5 there

BSP: Improvement in spatial correlation



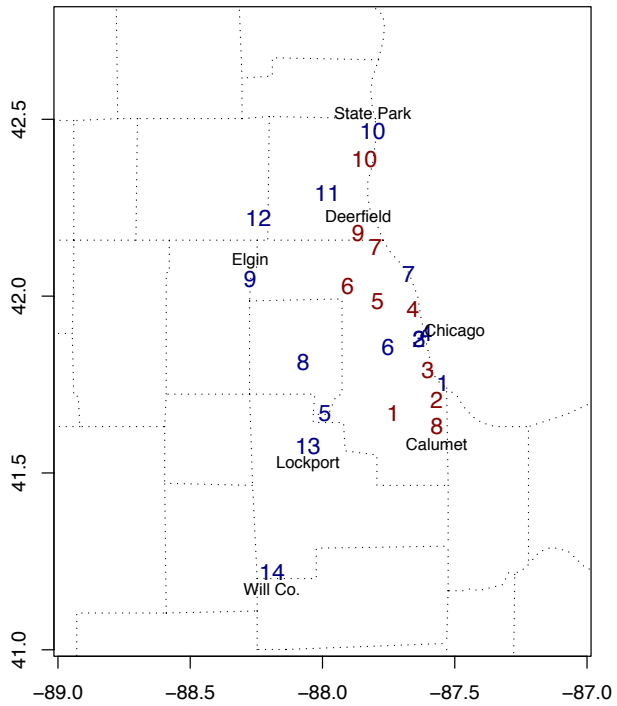
BSP byproduct:between hour correlations

- **5** $\times$ **5** autocorrelation matrices (24 of them that could be stitched together)
- $\Psi = \Lambda \otimes \Omega$ (between stations and hours)

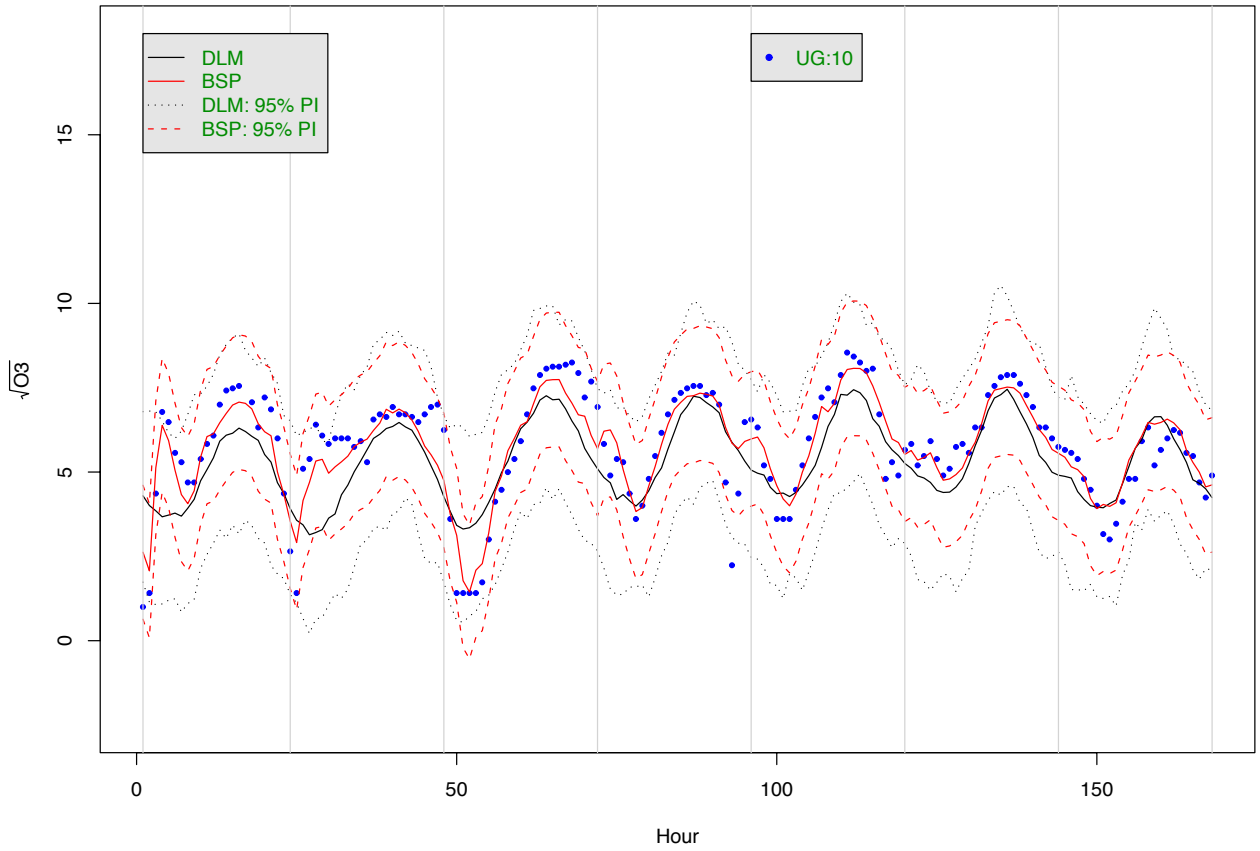
Example:

$$\Omega = \begin{array}{c} \begin{array}{ccccc} \text{12pm} & \text{1pm} & \text{2pm} & \text{3pm} & \text{4pm} \end{array} \\ \begin{bmatrix} 1.00 & 0.76 & 0.54 & 0.44 & 0.37 \\ 0.76 & 1.00 & 0.76 & 0.58 & 0.49 \\ 0.54 & 0.76 & 1.00 & 0.80 & 0.66 \\ 0.44 & 0.58 & 0.80 & 1.00 & 0.81 \\ 0.37 & 0.49 & 0.66 & 0.81 & 1.00 \end{bmatrix} \end{array}$$

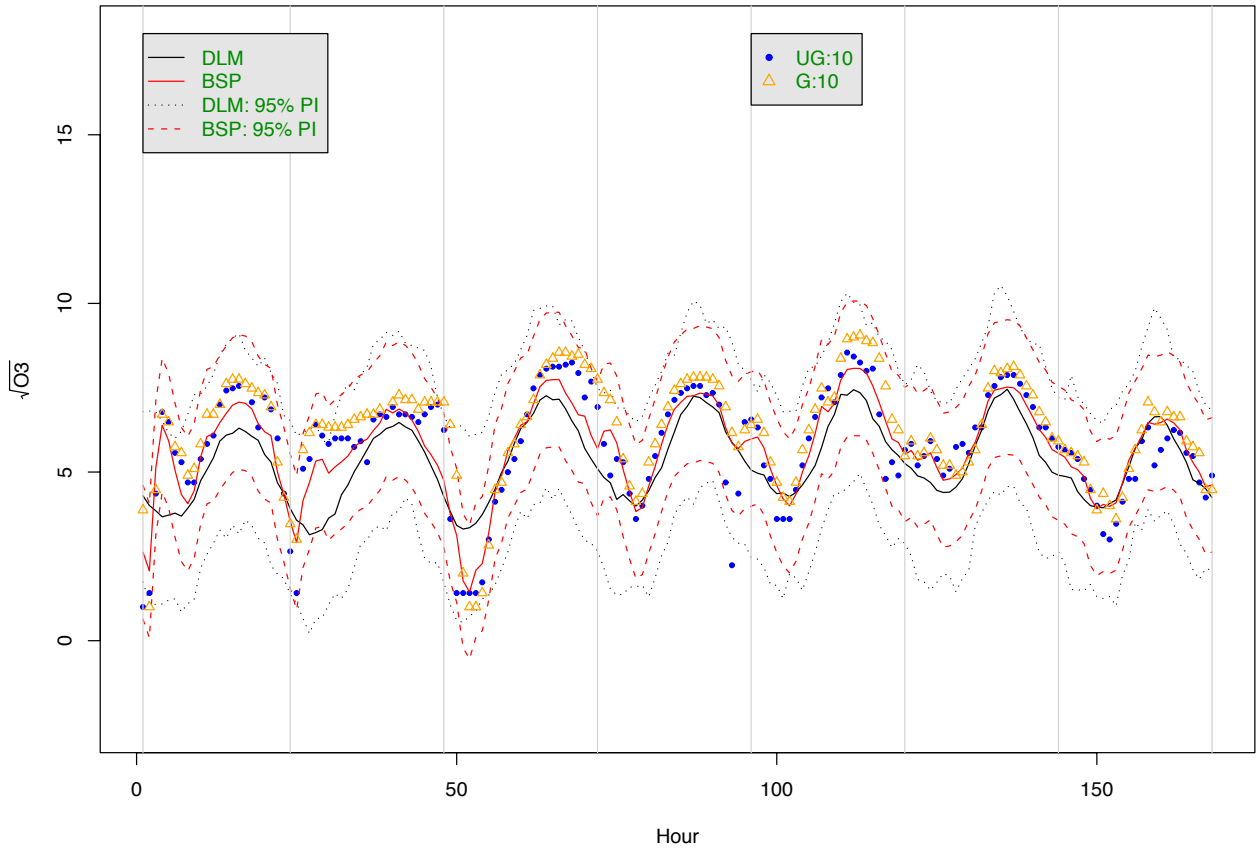
Results


State Park
Deerfield
Elgin
Chicago
Calumet
Lockport
Will Co.
42.5
42.0
41.5
41.0
−89.0
−88.5
−88.0
−87.5
−87.0

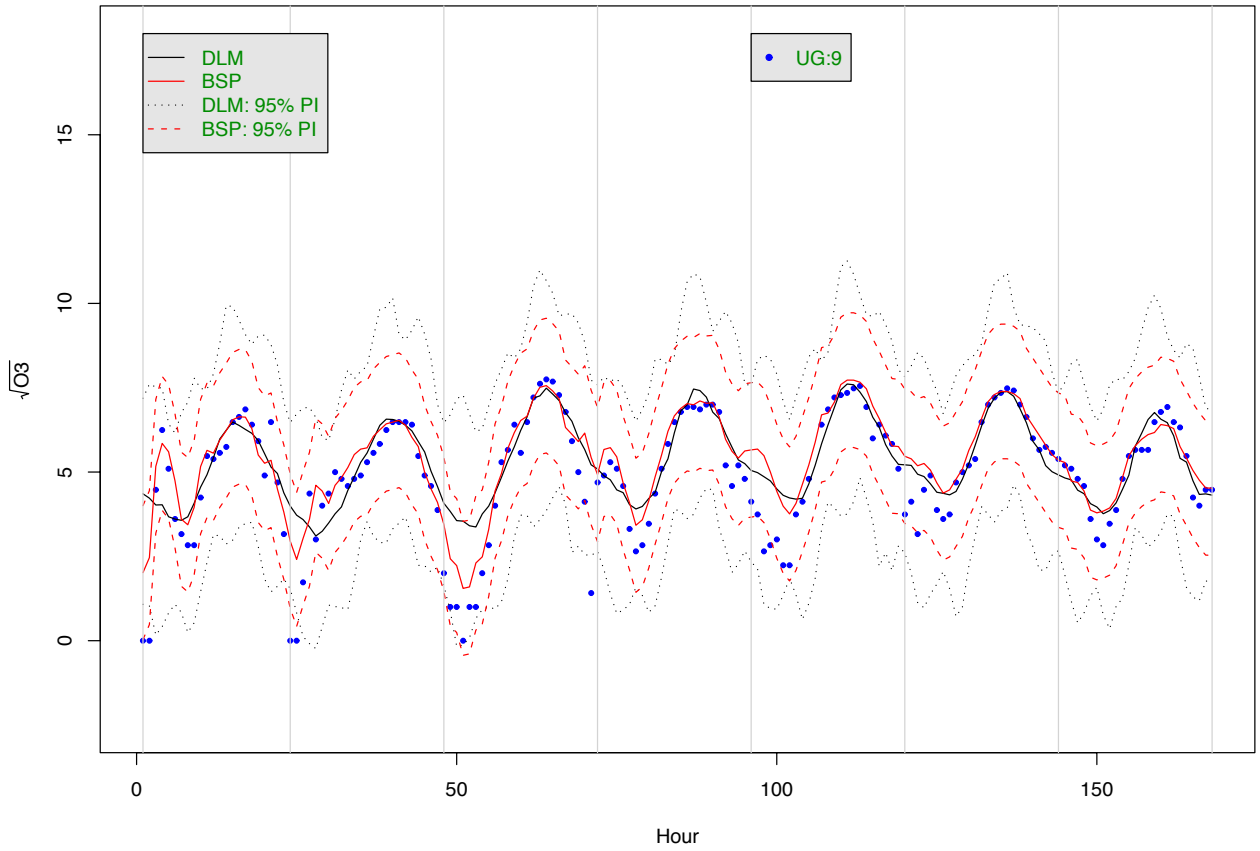
Pred mean + 95%CI for ungauged site 10 - week 1



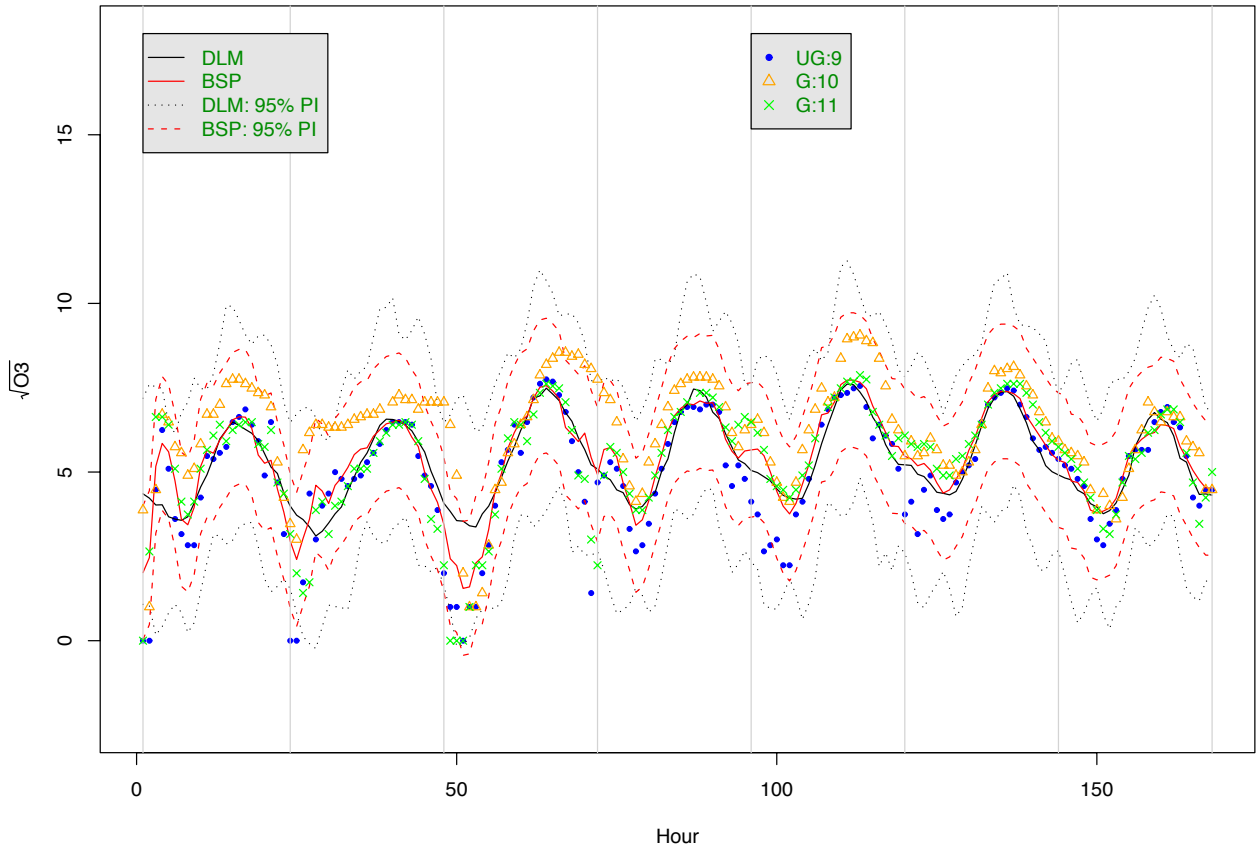
Pred mean + 95%CI for ungauged site 10 - week 1



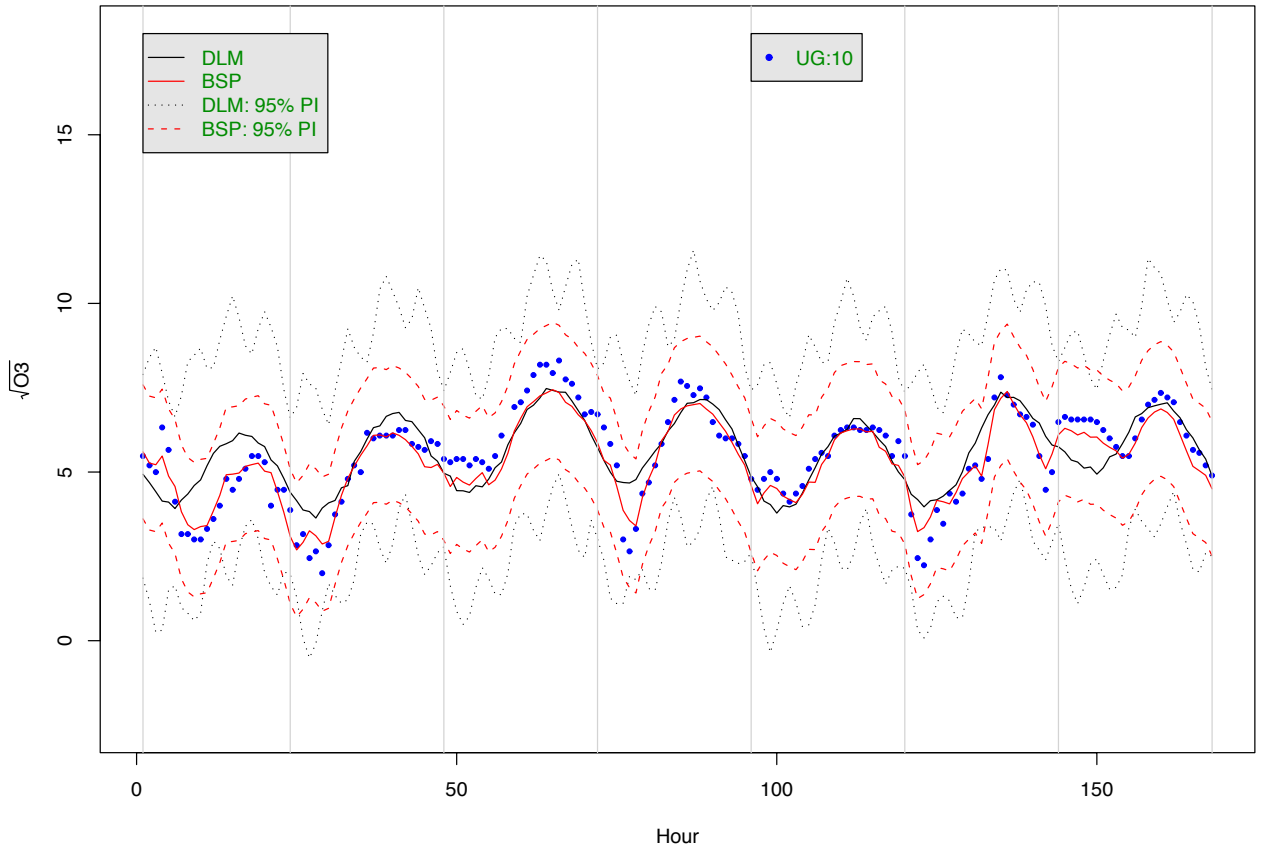
Pred mean + 95%CI for ungauged site 9 - week 1



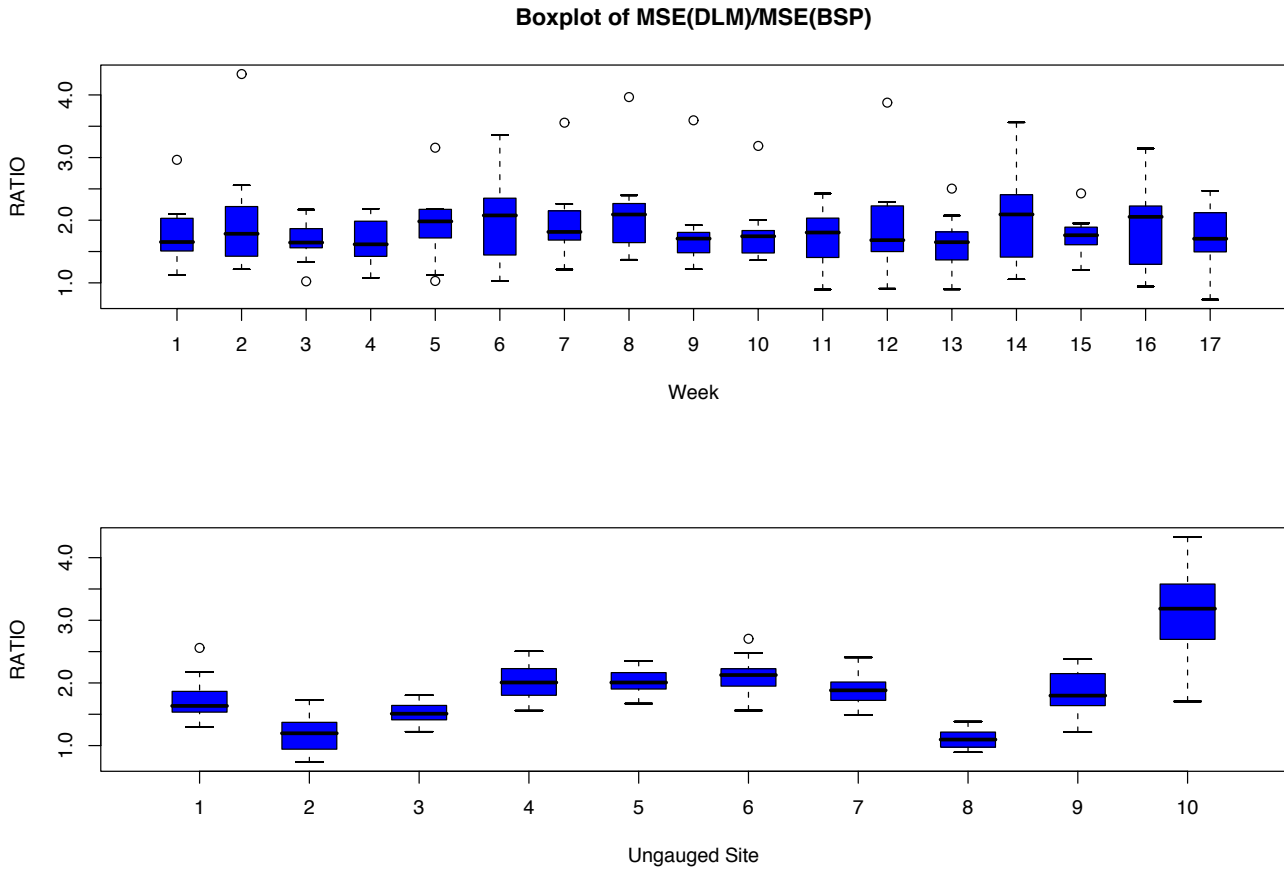
Pred mean + 95%CI for ungauged site 9 - week 1



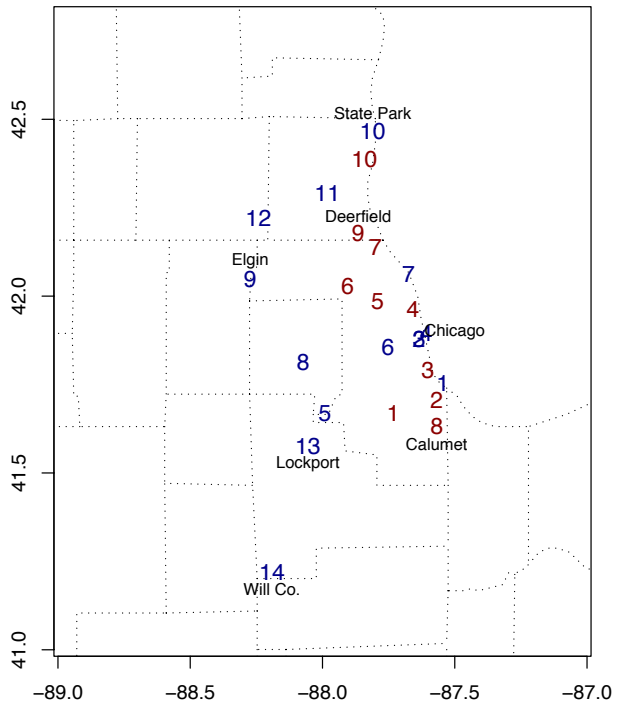
Pred mean + 95%CI for ungauged site 10 - week 10



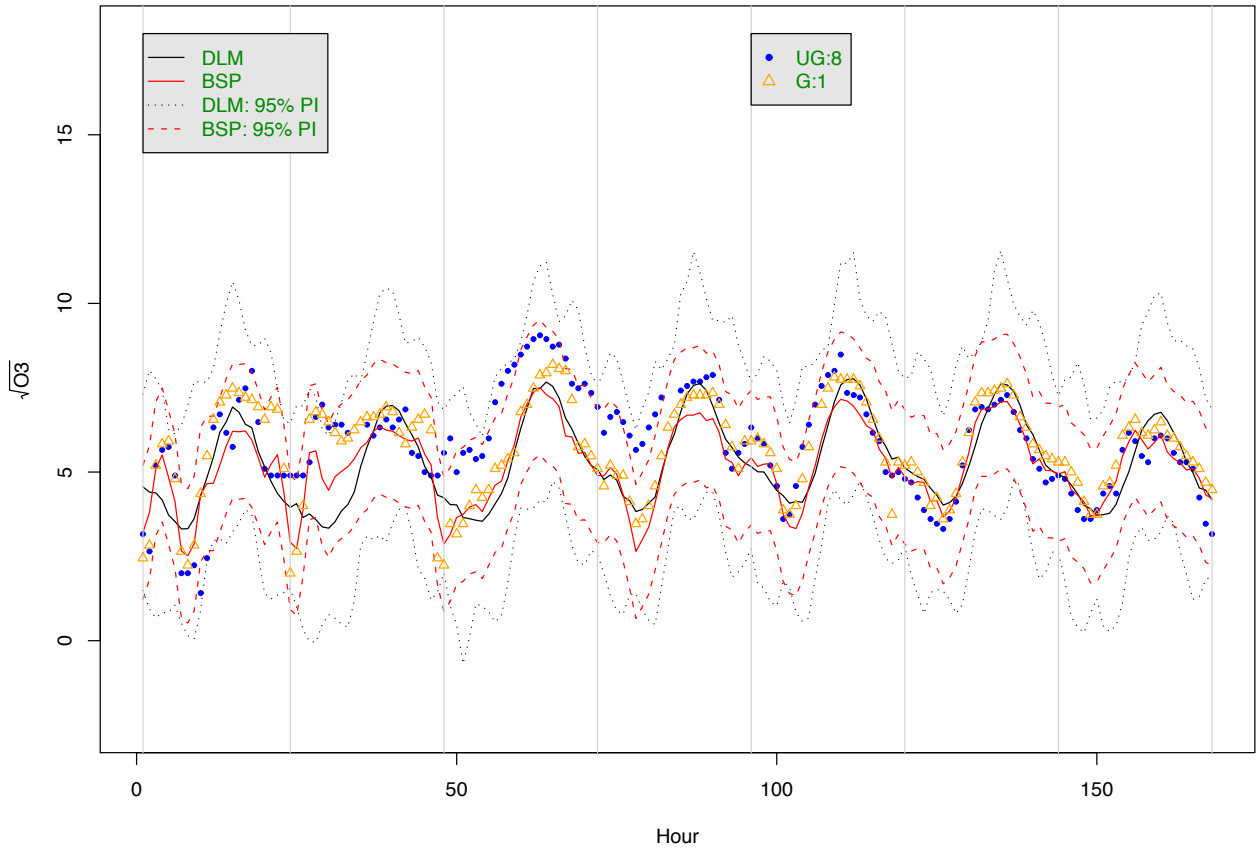
Ratio of MSE's



Ungauged Location 8



Pred mean + 95%CI for ungauged site 8 - week 1



Confidence Interval Coverage

Nominal Level	DLM	BSP
95%	95	90
80%	88	79
60%	77	66

DLM:

- $\tau_j^2, \tau_y^2, \lambda_j$: trial & error to get good coverage 95%CI

Ones - good for other levels - quite different

Discussion

BSP:

- Performs reasonably well
- Can extend to multiple pollutants + hourly data
- Limitations:
  - need for time independence and prefiltering:
    - could add prior on A and use MCMC but loss of computational efficiency
  - assumption of separable covariance

Discussion

DLM:

- Powerful, flexible, intuitive
- Limitations:
  - good hyperparameter estimates hard to find
  - big computation burden:10 days for 3000 MCMC iterations-C code-dual processors. [BSP took about 4 hours].
  - variance increases over time due to random walk model for coefficients
  - wobbly, not well calibrated prediction intervals
  - spatial correlation leakage (evidence of-hard to assess)

Conclusions from empirical study

- Each method has strengths and weaknesses
- BSP (with some additional modeling), but not, DLM could be used to model full field over about 300 eastern US (the real goal of the project)
- DLM fairly easy to implement - but hard to tune the hyperparameters
- DLM more difficult to program but that has been done
- Separability and correlation leakage deserve much more attention in future work

Physical statistical modeling: dynamic processes

- physical models needed for background
  - prior knowledge often expressed by differential equations (de's)
  - can lead to big computer models
  - yield deterministic response predictions
  - can encounter difficulties:
    - butterfly effect
    - nonlinear dynamics
    - lack of relevant background knowledge
    - lack of sufficient computing power

- statistical models also desirable
  - prior knowledge expressed by statistical models
  - often lead to big computer models
  - yield predictive distributions
  - can encounter difficulty:
    - off-the-shelf-models too simplistic
    - lack of relevant background knowledge
    - lack of sufficient computing power

May be strength in unity but:

- big gulf between two cultures
- communication between camps difficult
- approaches different
- route to reconciliation unclear

Approach to reconciliation - depends on: purpose; context; # of (differential) equations; etc.

With many equations (e.g. 100):

- build a better predictive response density for [field response — deterministic model outputs]
  eg. input model value as prior mean
- view model output as response and create joint density for
  $[\text{field response}, \text{model output}] =$
  $\int [\text{field response}|\lambda][\text{model output}|\lambda] \times \pi(\lambda|\text{data}) d\lambda$

References: Fuentes and Raftery [2005], Liu et al. [2011]

With a few differential equations (de's)

Example: $dX(t)/dt = \lambda X(t)$.

Option 1: solve it and make known or unknown constants uncertain (i.e. random):
$X(t) = \beta_1 \exp \lambda t + \beta_0$

Option 2: discretize the de and add noise to get a state space model: $X(t+1) = (1+\lambda)X(t) + \epsilon(t)$

Option 3: use functional data analytic approach - incorporate de through a penalty term as in splines
$\sum_t (Y_t - X_t)^2 + (\textit{smoothing parameter}) \int (DX - \lambda X)^2 dt$

Downscaling physical models

Regression – like approaches may be used:

$$X(s,t) = \alpha_{st} + \beta_{Mst} M(S,T) + \beta_{st} Z^{\text{covariates}}(s,t)\delta(s,t)$$

where M is physical model output, $s \in S^{\text{grid cell}}$ & $t \in T^{\text{Time Interval}}$.

References: Berrocal et al. [2010a], Zidek et al. [2012]

References

- Software with demo available at http://enviro.stat.ubc.ca
- Tutorial for it in Chapter 14 of Le & Zidek (2006). Statistical Analysis of environmental space-time processes (Springer)

V.J. Berrocal, A.E. Gelfand, and D.M. Holland. A spatio-temporal downscaler for output from numerical models. *Journal of Agricultural, Biological, and Environmental Statistics*, 15(2): 176–197, 2010a. ISSN 1085-7117.

O. Bodnar and W. Schmid. Nonlinear locally weighted kriging prediction for spatio-temporal environmental processes. *Environmetrics*, 21:365–381, 2010.

M. Fuentes and A.E. Raftery. Model evaluation and spatial interpolation by bayesian combination of observations with outputs from numerical models. *Biometrics*, 61:36–45, 2005.

M. Fuentes, L. Chen, and J.M. Davis. A class of nonseparable and nonstationary spatial temporal covariance functions. *Environmetrics*, 19(5):487–507, 2008.

T. Gneiting, M.G. Genton, and P. Guttorp. Geostatistical space-time models, stationarity, separability, and full symmetry. *MONOGRAPHS ON STATISTICS AND APPLIED PROBABILITY*, 107:151, 2006.

PJ Harrison and CF Stevens. A bayesian approach to short-term forecasting. *Operational Research Quarterly*, pages 341–362, 1971.

G. Huerta, B. Sansó, and J.R. Stroud. A spatiotemporal model for mexico city ozone levels. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 53(2):231–248, 2004.

Zhong Liu, Nhu Le, and James Zidek. An empirical assessment of bayesian melding for mapping ozone pollution. *Environmetrics*, 22 (3):340–353, 2011. doi: 10.1002/env.1054.

J.R. Stroud, P. Muller, and B. Sanso. Dynamic models for spatio-temporal data. *Journal of the Royal Statistical Society, Series B*, 63:673–689, 2001.

James Zidek, Nhu Le, and Zhong Liu. Combining data and simulated data for space–time fields: application to ozone. *Environmental and Ecological Statistics*, 19(1):37–56, 2012. ISSN 1352-8505. doi: 10.1007/s10651-011-0172-1.

APPENDICES

Summary of Appendices

These appendices provide basic building blocks for spatial prediction theory and design of monitoring networks. First in Appendix A we see important distributions:

- multinormal distribution
- the Wishart & Inverted Wishart distributions
- the Generalized Inverted Wishart distribution
- the multivariate & matric t distributions

Then we looked at uncertainty & saw:

- how Bayesian statistics includes all the uncertainty, even about parameters
- how to quantify it through entropy

That paves the way to the theory of spatial prediction.
Then in Appendix B we see some theory for characterizing uncertainty, including discussion of the entropy which can be used to develop optimal networks for measuring spatio–temporal processes.

APPENDIX A. Basics: Probability theory

Basics: Probability theory

Explores basic probability theory for the normal distribution & its cousins. Important class since often space-time fields can be transformed to be approximately normal.
Some of the distributions are on the normal's uncertain parameters to enable us to use Bayesian methods described in Appendix B.

Multivariate normal distribution

$\mathbf{X} : p \times 1 \sim N_p(\mu, \Sigma)$ means for any $\mathbf{a} : p \times 1$,

$$\mathbf{a}^T\mathbf{X} \sim N(a'\mu, \mathbf{a}^T\Sigma\mathbf{a}).$$

If $\Sigma : p \times p > 0$ (ie positive definite, hence invertible)

$$f(x) = (2\pi)^{-p/2}|\mathbf{\Sigma}|^{-1/2} \exp\left\{-(x-\mu)^T\mathbf{\Sigma}^{-1}(x-\mu)/2\right\}$$

Multinormal properties

- $E(X) = \mu,\ Cov(X) = \Sigma = (\Sigma_{ij}), \Sigma_{ij} = E[(X_i - \mu_i)(X_j - \mu_j)]$

.

For example, Σ_{ij} could be the covariance between responses at sites i and j.

Matric normal distribution

$\mathbf{X} : n \times m \sim N(\mu, A \otimes B)$ means

$$f(x) = \frac{1}{(2\pi)^{nm/2}} |A|^{-m/2} |B|^{-n/2} \text{etr}\left\{ -\frac{1}{2}[A^{-1}(x-\mu)][(x-\mu)B^{-1}]^T \right\}$$

$A = a_{ij} : n \times n > 0$ & $B = (b_{ij}) : m \times m > 0$ where $etr = e^{tr}$ &
$tr(A) = trace(A) = \sum a_{ii}$ for any matrix A

Matric normal properties

Properties:

- $E(X) = \mu$
- $\mathrm{var}[vec(X)] = A \otimes B$ and $\mathrm{var}[vec(X^T)] = B \otimes A$.
- $X \sim N(\mu, A \otimes B)$ if and only if $X^T \sim N(\mu^T, B \otimes A)$.
- $\mathrm{cov}(x_i, x_j) = a_{ij}B, \quad \mathrm{cov}(x^{(i)}, x^{(j)}) = b_{ij}A$.
- For any matrix $C : c \times n$ and matrix $D : m \times d$

$$CXD \sim N(C\mu D, CAC^T \otimes D^T BD)$$

Matric norm Properties continued

- $$E[\mathbf{X}F\mathbf{X}^T] = \mu E \mu^T + A\mathrm{tr}(FB) \text{ for any } F : m \times m$$

  &

  $$E[\mathbf{X}^T G\mathbf{X}] = \mu D \mu^T + \mathrm{tr}(AG)B \text{ for any } G : n \times n.$$

  Thus

  $$E[\mathbf{X}B^{-1}\mathbf{X}^T] = \mu B^{-1}\mu^T + mA$$

  &

  $$E[\mathbf{X}^T A^{-1}\mathbf{X}] = \mu^T A^{-1}\mu + nB$$

Multivariate t-distribution

$\mathbf{X} : p \times 1 \sim t_p(\mu, A, \nu)$ means

$$f(x) = \frac{\Gamma\left(\frac{p+\nu}{2}\right)\sqrt{|A|}}{\Gamma(\nu/2)\sqrt{2\pi p}} \times \left[1 + \frac{1}{\nu}(x-\mu)^T A(x-\mu)\right]^{-(p+\mu)/2}.$$

$A > 0$= precision matrix &

- $E(\mathbf{X}) = \mu, Cov(\mathbf{X}) = \frac{\nu}{\nu-2}A^{-1}$

Matric t-distribution

$X : n \times m \sim t_{n \times m}(\mu,\ A \otimes B,\ \delta)$, δ being the degrees of freedom means

$$f(x) \propto |A|^{-m/2}|B|^{-n/2}|I_n + \delta^{-1}[A^{-1}(X-\mu)][(x-\mu)B^{-1}]^T|^{-\frac{\delta+n+m-1}{2}}$$

for matrices $A : n \times n > 0$, $B : m \times m > 0$ and $\mu : n \times m$. Normalizing constant:

$$K = (\delta\pi^2)^{-\frac{nm}{2}} \frac{\Gamma_{n+m}[(\delta+n+m-1)/2]}{\Gamma_n[(\delta+n-1)/2]\Gamma_m[(\delta+m-1)/2]}$$

$$\Gamma_p(t) = \pi^{\frac{p(p-1)}{4}} \prod_{i=1}^{p} \Gamma[t-(i-1)/2]$$

More Matric-t properties

- $\mathbf{X} \sim t_{n\times m}(\mu,\ A \otimes B,\ \delta)$, if & only if $\mathbf{X}^T \sim t_{m\times n}(\mu^T,\ B \otimes A,\ \delta)$.
- If $n = 1$ & $A = 1$, $\mathbf{X}$ has an m-variate t-distribution, *i.e.*,

$$\mathbf{X} \sim t_m(\mu,\ B,\ \delta).$$

- If $m = 1$ & $B = 1$, $\mathbf{X}$ has an n-variate t-distribution, *i.e.*,

$$\mathbf{X} \sim t_n(\mu,\ A,\ \delta).$$

- If $\mathbf{X} \sim t_{n\times m}(\mu,\ A \otimes B,\ \delta)$, & $C_{c\times n}$ & $D_{m\times d}$ are of full rank (ie. rank c & d respectively), then

$$\mathbf{Y} = C\mathbf{X}D \sim t_{c\times d}(C\mu D, CAC^T \otimes D^T BD,\ \delta)$$

Matric-t properties

- $E(X) = \mu$
- When $\delta > 2$,

$$\text{var}[vec(X)] = \delta(\delta - 2)^{-1} A \otimes B.$$

and

$$\text{cov}(x_i, x_j) = \delta(\delta - 2)^{-1} a_{ij} B, \quad \text{cov}(x^{(i)}, x^{(j)}) = \delta(\delta - 2)^{-1} b_{ij} A.$$

Wishart distribution

$\mathbf{S} : p \times p \sim W_p(A, m)$ means

$$f(\mathbf{s}) = \left[2^{mp/2}\Gamma_p(m/2)\right]^{-1} |A|^{-m/2}|\mathbf{s}|^{(m-p-1)/2} \exp^{-\frac{1}{2}tr(A^{-1}\mathbf{s})}$$

for any $A > 0$.

The Wishart generalizes the chi-squared distribution to $p \times p$ positive definite random matrices

Wishart & Inverted Wishart properties

- $\mathbf{Y} \sim W_p^{-1}(\Psi, \delta)$ if & only if $\mathbf{Z} = \mathbf{Y}^{-1} \sim W_p(\Psi^{-1}, \delta)$.
- If $\mathbf{Z} \sim W_p(\mathbf{\Sigma}, \delta)$ then $E(\mathbf{Z}) = \delta\mathbf{\Sigma}$ & $E(\mathbf{Z}^{-1}) = \mathbf{\Sigma}^{-1}/(\delta - p - 1)$ provided $\delta - p - 1 > 0$.
- If $\mathbf{Y} \sim W_p^{-1}(\Psi, \delta)$, then $E(\mathbf{Y}) = \Psi/(\delta - p - 1)$ & $E(\mathbf{Y}^{-1}) = \delta\Psi^{-1}$.
- If $\mathbf{Y} \sim W_p^{-1}(\Psi, \delta)$, then

$$E \log |\mathbf{Y}| = -p \log 2 - \sum_{i=1}^{p} \eta \left[\frac{1}{2}(\delta - i + 1)\right] + \log |\Psi|$$

where $\eta = \text{digamma function} = d[\log \Gamma(x)]/dx$

Inverted Wishart distribution

$\mathbf{\Sigma} : p \times p \sim W_p^{-1}(\Psi, \delta)$ means

$$f(\mathbf{\Sigma}) = 2^{mp/2}\Gamma_p(m/2)|\Psi|^{-\frac{1}{2}(\delta+p+1)} \exp\{-\frac{1}{2}\Sigma^{-1}\Psi\}$$

for some $\Psi > 0$.
This one generalizes the inverse gamma distribution.

Bartlett decomposition

(Needed for next distribution.) Let

$$\Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}, \Delta = \begin{pmatrix} \Sigma_{1|2} & 0 \\ 0 & \Sigma_{22} \end{pmatrix} \& T = \begin{pmatrix} I & \tau \\ 0 & I \end{pmatrix}.$$

$$\textbf{with } \Sigma_{1|2} \equiv \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}, \tau \equiv \Sigma_{12}\Sigma_{22}^{-1}.$$

$$\textbf{Then } \Sigma = T\Delta T^T$$

Hence

$$\boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{1|2} + \tau\boldsymbol{\Sigma}_{22}\tau^T & \tau\boldsymbol{\Sigma}_{22} \\ \boldsymbol{\Sigma}_{22}\tau^T & \boldsymbol{\Sigma}_{22} \end{pmatrix}.$$

Extendable recursively to more blocks

Generalized Inverted Wishart

The 2-block version extendable to k by recursion.

$$\textbf{Apply Bartlett to } \Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}$$

Then $\boldsymbol{\Sigma} \sim GIW(\Psi, \delta)$ means

$$\boldsymbol{\Sigma}_{22} \sim IW(\Psi_{22}, \delta_2), \Sigma_{1|2} \sim IW(\Psi_{1|2}, \delta_1)$$

$$\tau \mid \Sigma_{1|2} \sim N\left(\tau_{01}, H_1 \otimes \Sigma_{1|2}\right),$$

where

$$\Psi_{1|2} = \Psi_{11} - \Psi_{12}(\Psi_{22})^{-1}\Psi_{21}$$

NOTE: Different δs for each block - much richer than IW

APPENDIX A. Basics: Uncertainty

Basics: Uncertainty

- Gives brief intro to Bayesian theory wherein uncertainty is represented by probability distributions.
- Also discusses entropy, a general method of quantifying such

Elements of Bayesianity

- **Likeihood:** the likelihood of the data x given a hypothesized value of a model parameter $f(x \mid \theta)$. This is the voice of the data whose favored choice is $\hat{\theta} = mle = \arg\max_{\theta'} f(y \mid \theta')$
- **prior distribution:** $\pi(\theta \mid \alpha)$ reflects experimenter's uncertainty about θ, $\theta^{opt} = \arg\max_{\theta'} \pi(\theta' \mid \alpha)$ being preferred in the absence of data. α is a **hyperparameter**, which if uncertain can either be
  - estimated from the data **(empirical Bayes)**or
  - assigned another prior distribution **(hierarchical Bayes)**

Elements continued

- **Marginal likelihood:**

$$f(y \mid \alpha) = \int f(y \mid \theta)\pi(\theta \mid \alpha)d\theta$$

- Type II mle= $\hat{\alpha} = \arg\max_{\alpha'} f(y \mid \alpha)$ for the empirical Bayes approach

NOTES: θ can be of high dimension eg 1.7million coordinates. There ad hoc methods are commonly used, eg choosing **vague** (ie flat) priors. However, Bayesianity provides a coherent framework for analysis even there.

Entropy

Used in monitoring network design: select sites where uncertainty (entropy) highest.

Let $\mathbf{X} = E$ or $\overline{E}$ - unknown

- $p = \Pr(E)$: prob of event E
- $\phi(.)$: reduction in uncertainty with $\phi(p)$ if E occurs & $\phi(1-p)$ if $\overline{E}$ occurs

Thus expected reduction in uncertainty for observing $\mathbf{X}$

$$H(\mathbf{X}) = p\phi(p) + (1-p)\phi(1-p)$$

Axioms:

- $\phi \in [0, 1]$
- $H(\mathbf{X}, \mathbf{Y}) = H(\mathbf{X}) + H(\mathbf{Y})$ for $\mathbf{X}, \mathbf{Y}$ independent $\Longrightarrow \phi(p) = -log(p)$

General discrete case

For a general discrete case $\mathbf{X} = \{x_i, \ldots, x_n\}$
ENTROPY($\mathbf{X}$) denoted by $H(\mathbf{X})$ defined as

$$H(\mathbf{X}) = \sum_{i=1}^{n} p_i[-\log p_i]$$

is the average reduction of uncertainty for observing $\mathbf{X}$

Continuous case

For $X \sim f$

$$H(X) \equiv -E[\log f(X)] = -\int \log f(x) f(x) dx$$

by analogy is unsatisfactory: not invariant

- **Jaynes (1963)** introduces reference measure $h(x)$

$$H(X) \equiv -E\left[\log \frac{f(X)}{h(X)}\right]$$

  is invariant under $X \longrightarrow g(X)$
- $h(X)$ represents complete ignorance, eg Gaussian-IW, $h(X, \mu, \Sigma) = |\Sigma|^{-(p+1)/2}$

Entropy decomposition

$$\mathbf{X} \sim f(. \mid \theta)$$

Given data D

$$H(\mathbf{X}, \theta) = H(\mathbf{X} \mid \theta) + H(\theta)$$

where

$$\begin{aligned} H(\mathbf{X} \mid \theta) &= E[-\log(f(\mathbf{X} \mid \theta, D)/h_1(\mathbf{X})) \mid D] \\ H(\theta) &= E[-\log(f(\theta \mid D)/h_2(\theta)) \mid D] \\ h(\mathbf{X}, \theta) &= h_1(\mathbf{X})h_2(\theta) \end{aligned}$$

[Caselton,Kan,Zidek (1992);Le,Zidek (1994)]

Entropy: multivariate t -dist

$$\begin{aligned} \mathbf{Y} \mid \Sigma &\sim N_g(0, \Sigma) \\ \Sigma \mid \Psi, \delta &\sim IW(\Psi, \delta) \end{aligned}$$

Then

$$\mathbf{Y} \mid \Psi, \delta \sim t_g(0, s^{-1}\Psi, s)$$

with $s = \delta - g + 1$
Can be shown:

$$\Sigma \mid \mathbf{X}, \Psi, \delta \sim IW(\Psi + \mathbf{X}\mathbf{X}^T, \delta + 1)$$

Conditional on Ψ & δ

$$\begin{aligned} H(X,\Sigma) &= H(X \mid \Sigma) + H(\Sigma) \\ H(X,\Sigma) &= H(\Sigma \mid X) + H(X) \end{aligned}$$

Hence

$$H(X) = H(X \mid \Sigma) + H(\Sigma) - H(\Sigma \mid Y)$$

With

$$\begin{aligned} H(X \mid \Sigma) &= \frac{1}{2}E(\log|\Sigma| \mid \Psi) + \frac{g}{2}(\log(2\pi)+1) \\ &= \frac{1}{2}E(\log|\Psi|) + \frac{1}{2}E(\log|\Sigma\Psi^{-1}|) + \frac{g}{2}(\log(2\pi)+1) \\ &= \frac{1}{2}E(\log|\Psi|) + c(g,\delta) \end{aligned}$$

Here $c(g,\delta)$: constant depending on g and δ

Note $\Psi\Sigma^{-1} \sim W(I_g, \delta)$
Using $h(\mathbf{Y}, \Sigma) = h(\mathbf{Y})h(\Sigma) = |\Sigma|^{-(g+1)/2}$

$$
\begin{aligned}
H(\Sigma) &= E[\log f(\Sigma)/h(\Sigma)] \\
&= \frac{1}{2}\delta \log|\Psi| - \frac{1}{2}\delta E(\log|\Sigma|) \\
&\quad - \frac{1}{2}E(tr\Psi\Sigma^{-1}) + c_1(g,\delta) \\
&= -\frac{1}{2}\delta E(\log|\Sigma\Psi^{-1}|) - \frac{1}{2}E(tr\Psi\Sigma^{-1}) + c_1(g,\delta) \\
&= \frac{1}{2}\delta E(\log|\Sigma^{-1}\Psi|) - \frac{1}{2}E(tr\Psi\Sigma^{-1}) + c_1(g,\delta) \\
&= c_2(g,\delta)
\end{aligned}
$$

Similarly

$$\begin{aligned} H(\Sigma \mid X) &= \frac{1}{2}(\delta+1)\log|\Psi| - \frac{1}{2}(\delta+1)E(\log|\Psi + YY')|) \\ &\quad + c_3(g,\delta) \\ &= -\frac{1}{2}(\delta+1)E(\log|1 + X'\Psi^{-1}X)|) + c_3(g,\delta) \\ &= c_4(g,\delta) \end{aligned}$$

Note $|\Psi|(1 + X^T\Psi^{-1}X) = |\Psi + XX^T)|$
and $X^T\Psi^{-1}X \sim F$ with df's depending only on g, δ
Thus $H(X) = \frac{1}{2}\log|\Psi| + c_5(g,\delta)$