

Addressing Temporal Correlation, Incomplete Source Profile Information, and Varying Source Profiles in the Source Apportionment of Particulate Matter

William F. Christensen, C. Shane Reese, Matthew J. Heaton, Basil Williams,
& Jeff Lingwall
Department of Statistics
Brigham Young University
william@stat.byu.edu

EPA STAR PM Source Apportionment
Progress Review Workshop
Research Triangle Park
June 21, 2007



Although the research described in the article has been funded wholly or in part by the U.S. Environmental Protection Agency's STAR program through grant RD-83216001-0, it has not been subjected to any EPA review and therefore does not necessarily reflect the views of the Agency, and no official endorsement should be inferred.

Outline

- I. Pollution source apportionment and Bayesian methods
- II. Dirichlet based Bayesian multivariate receptor modeling
- III. Dirichlet Process (DP) model for temporally-evolving source profiles
- IV. Bayesian approach for the identification of pollution source directions
- V. Conclusions and additional research directions

Pollution source apportionment and Bayesian methods

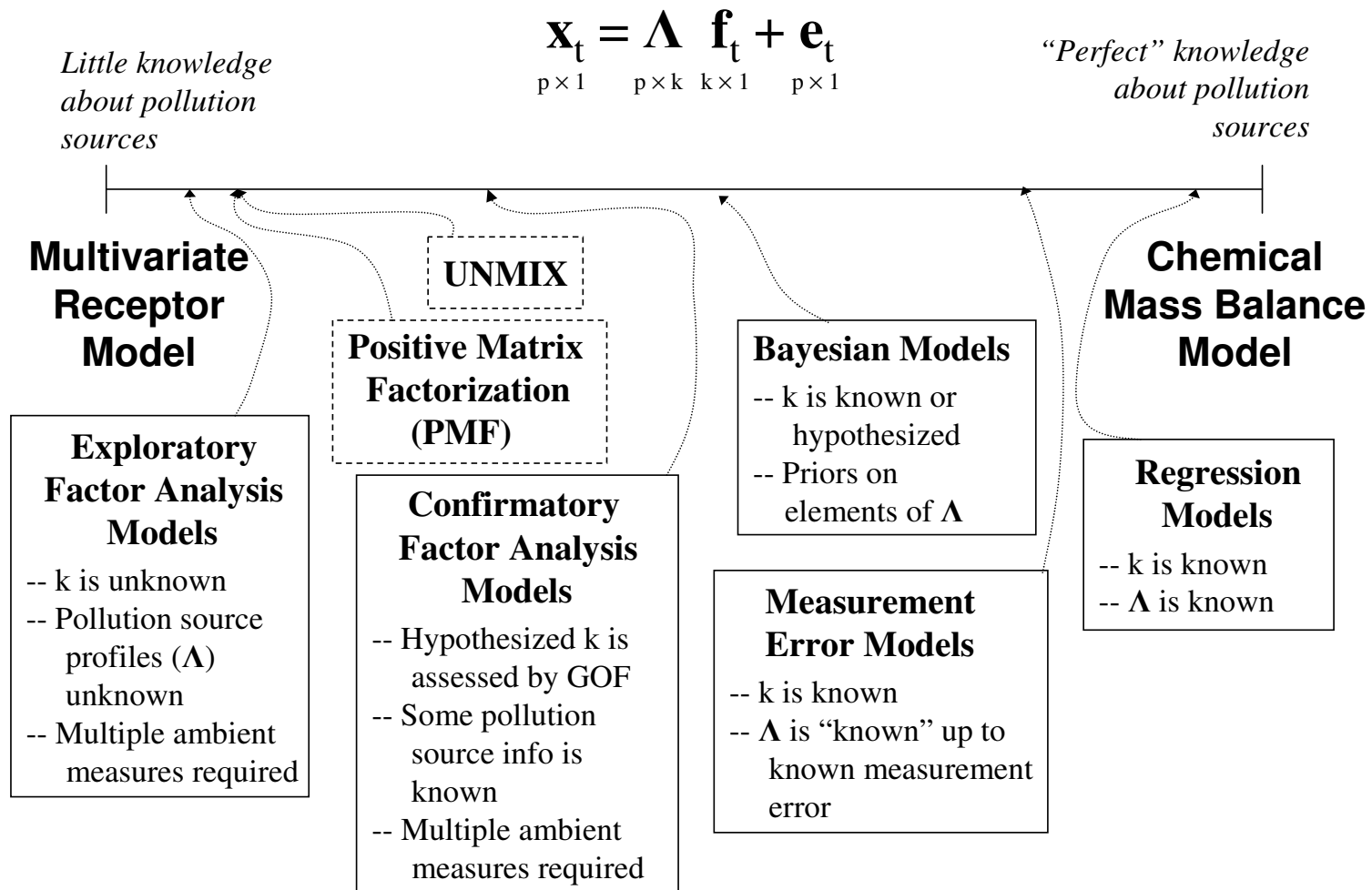
$$\underbrace{\mathbf{x}_t}_{p \times 1} = \underbrace{\Lambda}_{p \times k} \underbrace{\mathbf{f}_t}_{k \times 1} + \underbrace{\mathbf{e}_t}_{p \times 1}$$

For example, the abundance of EC particulates at time t :

$$\begin{aligned} x_{1t} = & [\% \text{ EC in auto exhaust}] \times \\ & [\text{concentration of auto exhaust in atmosphere } (\mu\text{g}/\text{m}^3)] \\ & + [\% \text{ EC in zinc smelter emissions}] \times \\ & [\text{concentration of zinc smelter emissions } (\mu\text{g}/\text{m}^3)] \\ & + \dots + e_{1t} \end{aligned}$$

- Λ unknown \Rightarrow model is called *multivariate receptor model* and is fit using factor analytic methods
- Λ known \Rightarrow model is called *chemical mass balance model* and is fit using regression methods

“Receptor Models”



**IDEAS AND
PERSPECTIVES**

Why environmental scientists are becoming Bayesians

James S. Clark
*Nicholas School of the
Environment and Department
of Biology, Duke University,
Durham, NC 27708, USA*
*Correspondence: E-mail:
jimclark@duke.edu*

Abstract

Advances in computational statistics provide a general framework for the high-dimensional models typically needed for ecological inference and prediction. Hierarchical Bayes (HB) represents a modelling structure with capacity to exploit diverse sources of information, to accommodate influences that are unknown (or unknowable), and to draw inference on large numbers of latent variables and parameters that describe complex relationships. Here I summarize the structure of HB and provide examples for common spatiotemporal problems. The flexible framework means that parameters

Basic probability:

$$\Pr\{A,B,C\} = \Pr\{A|B,C\} \times \Pr\{B|C\} \times \Pr\{C\}$$

For complex problems (Berliner, 1996):

$$p\{\text{data,process,parameters}\} = p\{\text{data|process,params}\} \times p\{\text{process|params}\} \times p\{\text{params}\}$$

“data model”

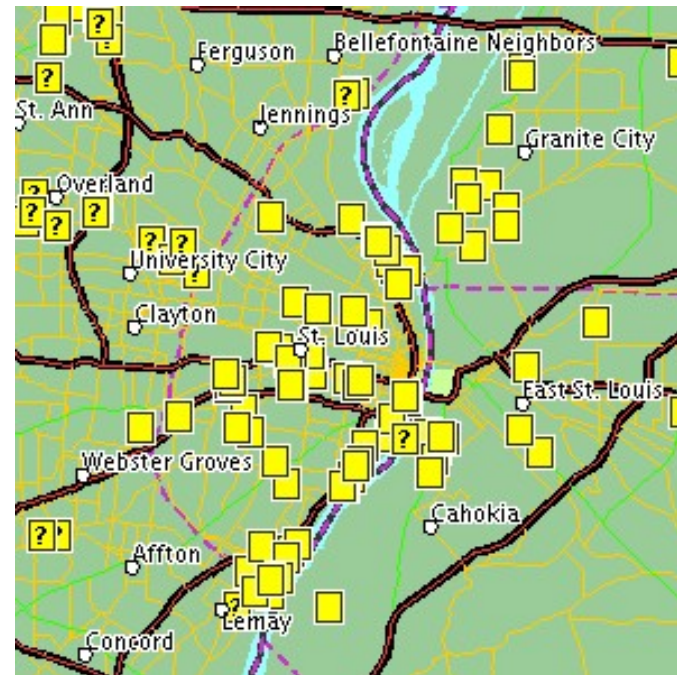
“process model”

“parameter
model”

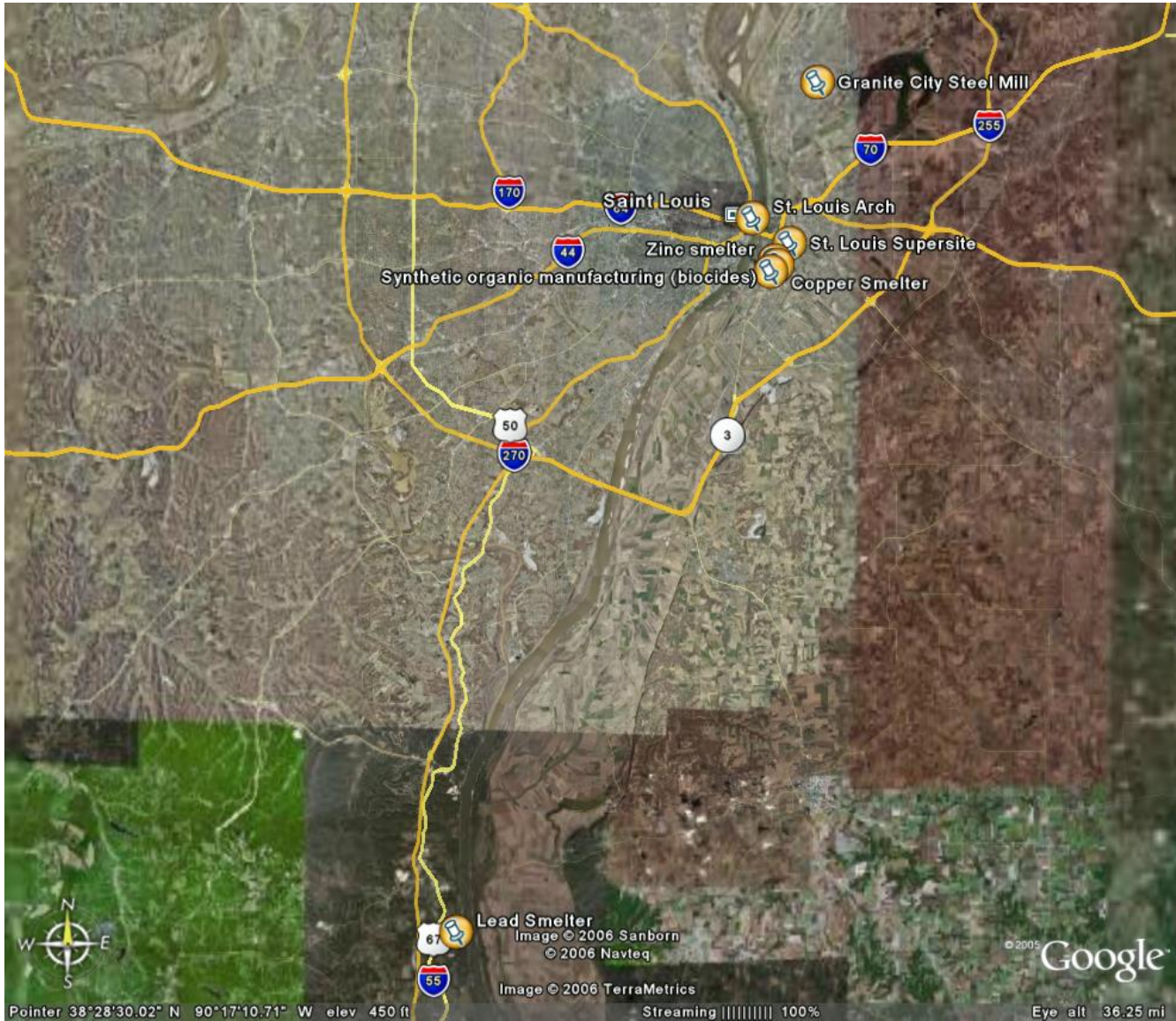
1. “Data”: ambient PM concentrations, meteorological data
2. “Process”: transport/dispersion, meteorology, seasonality, atmospheric chemistry, etc.
3. “Parameters”: daily source contribution values, source profile values

Interest in $p\{\text{parameters|data,process}\}$

- Auxiliary information for enhancing source apportionment



- Toxic release inventories →
- Wind direction & other meteorological data
- Dispersion models (e.g., EPA's AERMOD)



II. Dirichlet based Bayesian multivariate receptor modeling

(Lingwall, Christensen, and Reese, submitted)

- Data from St. Louis EPA Supersite includes two years of daily measurements of metals, carbon, and ions. Also...
 - Particle size data
 - Weekly organics measurements (extremely important for wood/agricultural burning, auto/diesel split, etc.)
- Model for ambient PM data, \mathbf{X}

$$\mathbf{x}_t = \Lambda \mathbf{f}_t + \mathbf{e}_t$$

- Likelihood and Priors:

- Source profiles (columns of Λ) \sim Generalized Dirichlet (Rogers and Young, 1973)

- * Individual elements of an *a priori* source profile ($\tilde{\lambda}_k$) are associated with different degrees of certainty, but variances of elements of Dirichlet vector cannot be individually tuned

$$\lambda_k \sim \text{Dirichlet}(\eta_k \tilde{\lambda}_k)$$

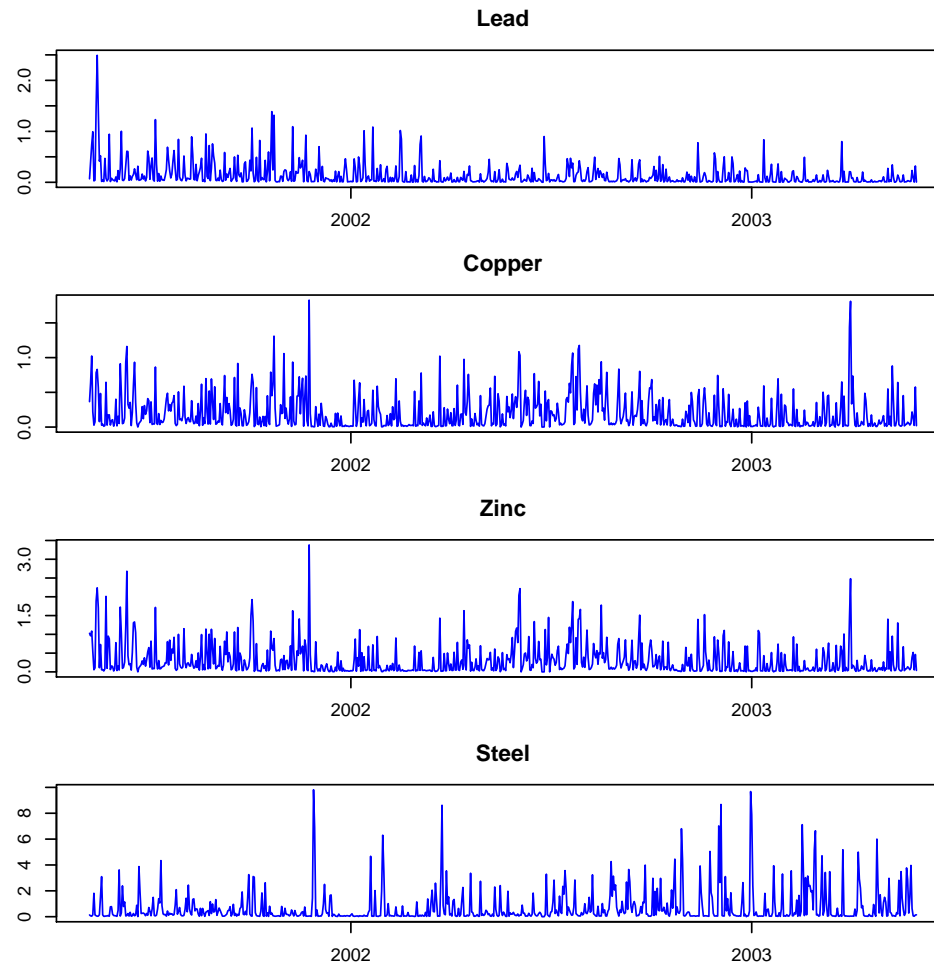
- * Generalized Dirichlet is sum of Gamma random variables with *differing* scale parameters, so individual variances can be at least partially tuned to desired degree of uncertainty (e.g., with genetic algorithm)

- * Priors for profile parameters informed by:

- Available profiles
- Past studies

- Likelihood and Priors:
 - Source contributions (elements of \mathbf{f}_t) \sim Lognormal
 - * Priors for contribution parameters informed by:
 - Toxic release inventories
 - Wind data
 - Particle size distributions
 - Daily, weekly, yearly cycles (e.g., seasonal patterns in secondary formation and traffic flow)

EPA's AERMOD dispersion model: fate of pollutants emitted from point source locations



Simulation Studies

- Generate pseudo-data based on source apportionment analysis of Washington DC PM_{2.5} data
- Use approximate profiles as *a priori* information in Bayesian model (via prior distributions) and PMF (via “source profile targeting”)
- No *a priori* information for contribution matrix in this simulation
- Calculate Total Median Absolute Error (TMAE) for estimating source contributions and source profiles:
 - PMF _{$\tilde{\Lambda}$} (uses *a priori* information on Λ)
 - PMF (does *not* use *a priori* information on Λ)
 - Bayesian _{$\tilde{\Lambda}$} (uses *a priori* information on Λ)
 - Bayesian (does *not* use *a priori* information on Λ)

TMAE for estimating source contributions and source profiles

Parameters	CV_Y	CV_Λ	$PMF_{\tilde{\Lambda}}$	Bayesian $_{\tilde{\Lambda}}$	PMF	Bayesian
F	0.3	0.2	4.22	4.02	6.84	4.77
Λ	0.3	0.2	0.0048	0.0016	0.0136	0.0031
F	0.3	0.4	5.17	4.36	6.84	4.77
Λ	0.3	0.4	0.0065	0.0019	0.0136	0.0031
F	0.3	0.6	5.01	4.42	6.84	4.77
Λ	0.3	0.6	0.0069	0.0023	0.0136	0.0031
F	0.6	0.2	7.24	7.05	10.21	7.87
Λ	0.6	0.2	0.0019	0.0022	0.0283	0.0056
F	0.6	0.4	9.13	7.67	10.21	7.87
Λ	0.6	0.4	0.0265	0.0035	0.0283	0.0056
F	0.6	0.6	9.80	8.05	10.21	7.87
Λ	0.6	0.6	0.0296	0.0051	0.0283	0.0056
Average Relative TMAE{ F }			114%	100%	144%	107%
Average Relative TMAE{ Λ }			459%	100%	757%	157%

III. Dirichlet Process (DP) model for temporally-evolving source profiles

(Heaton, Reese, and Christensen, in preparation)

Dirichlet Process (DP) Model

$$\mathbf{y}_t | \Lambda_t, \mathbf{f}_t, \Sigma \sim \text{LN}[\Lambda_t \mathbf{f}_t, \Sigma]$$

$$\lambda_{kt} \sim \text{DIR} [g_k \lambda_{k(t-1)}]$$

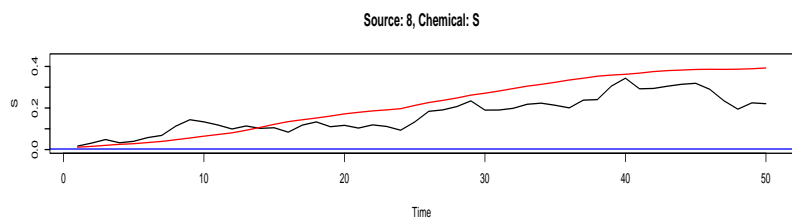
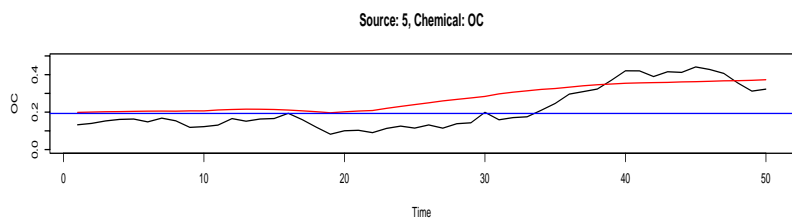
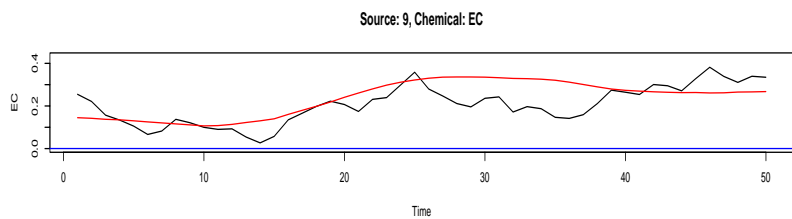
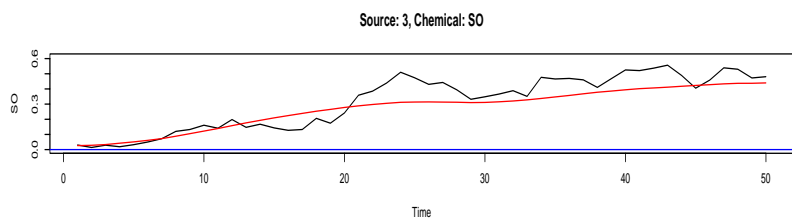
Assumptions

1. Source emission compositions vary through time.
2. Errors are log-normally distributed.
3. Concentrations are time dependent.

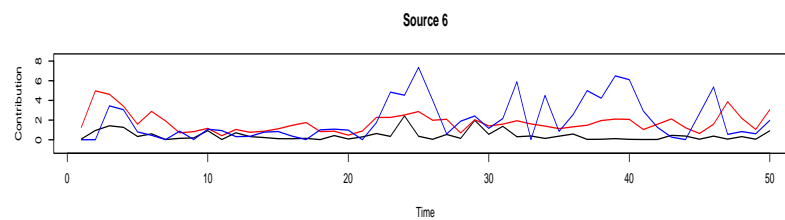
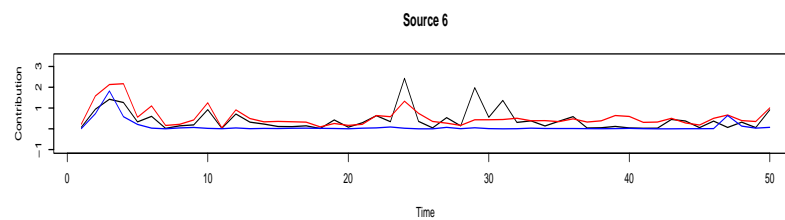
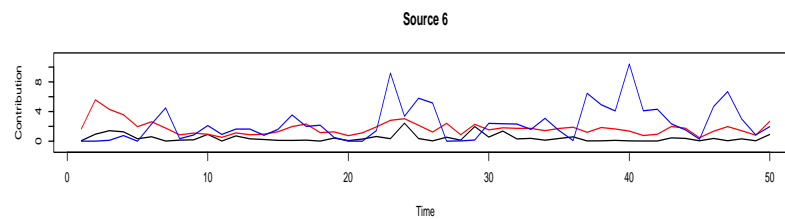
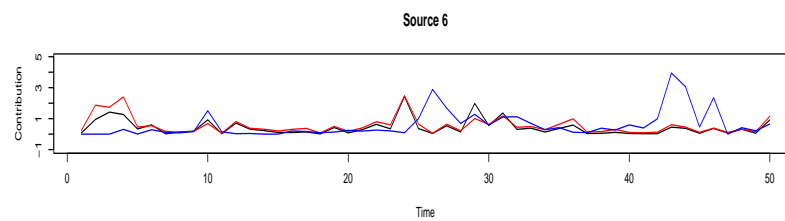
Goal: Compare DP Model to PMF by simulating data sets under varying degrees of variability in \mathbf{y}_t and Λ_t .

Comparison under time-varying profiles (True, Bayesian, PMF)

Profile Plots



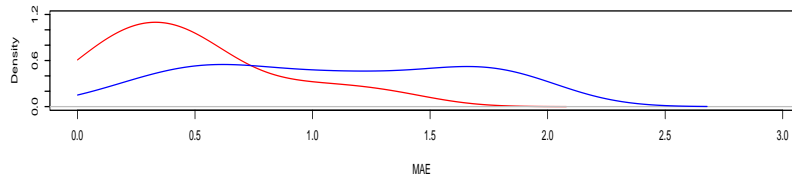
Contribution Plots



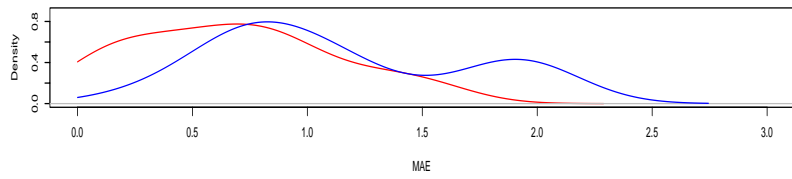
Comparison under time-varying profiles (Bayesian, PMF)

MAE for $\hat{\Lambda}_t$

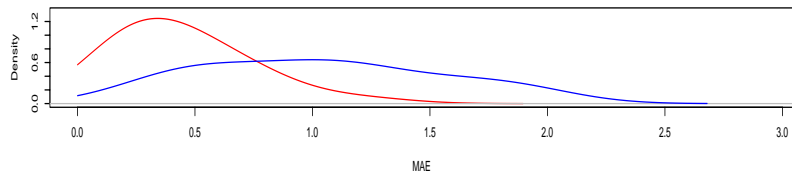
MAE_i



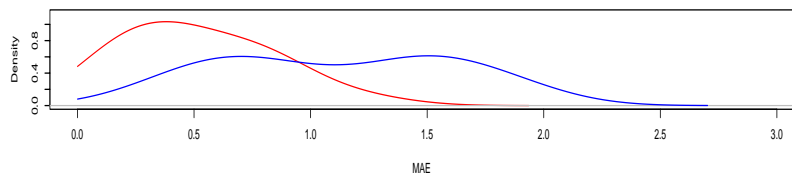
MAE_i



MAE_i

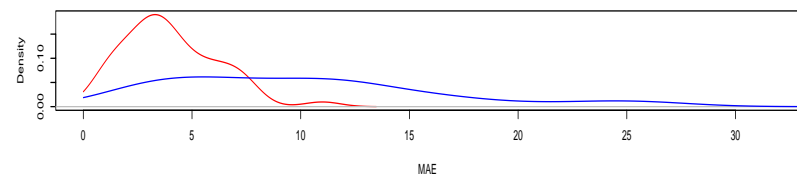


MAE_i

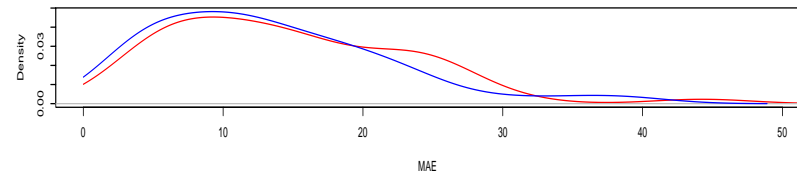


MAE for \hat{f}_t

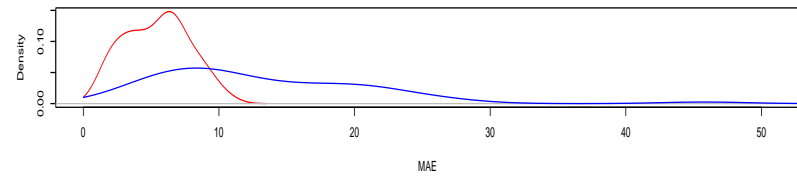
MAE_F



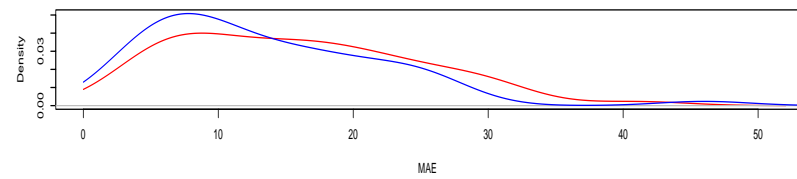
MAE_F



MAE_F



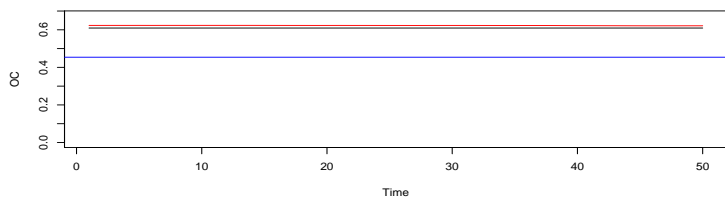
MAE_F



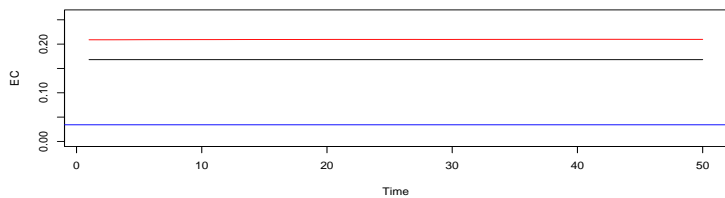
Comparison under time-constant profiles (True, Bayesian, PMF)

Profile Plots

Source: 6, Chemical: OC

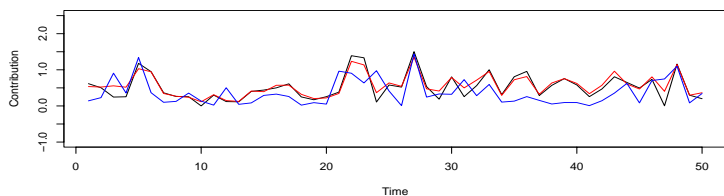


Source: 3, Chemical: EC

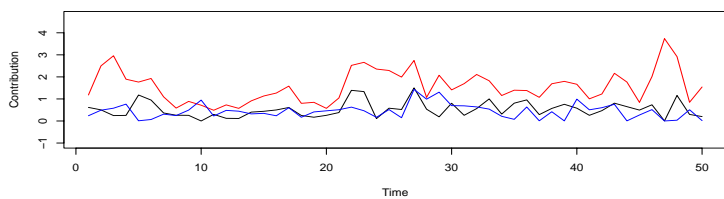


Contribution Plots

Source 4

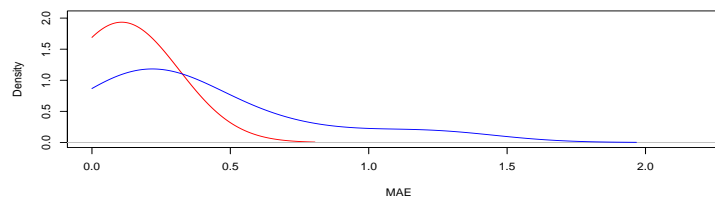


Source 4

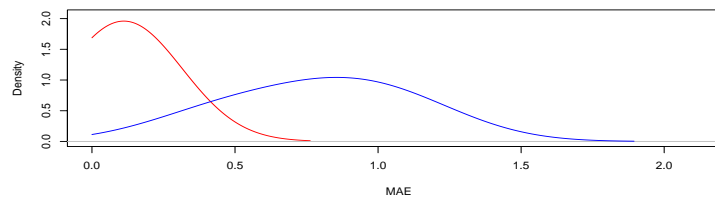


MAE Comparison

MAE_I

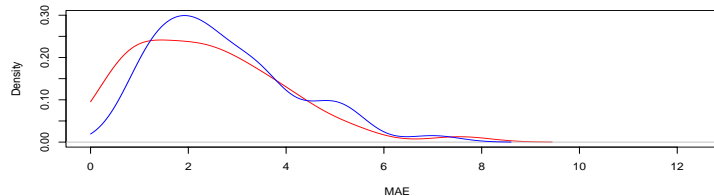


MAE_L

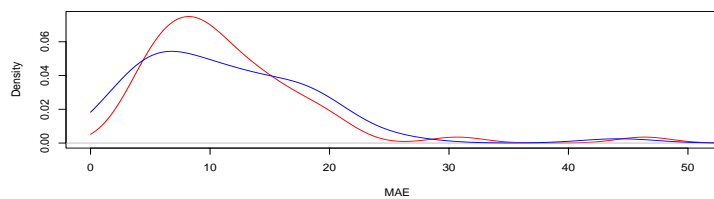


MAE Comparison

MAE_F



MAE_F



Summary of DP Model Performance

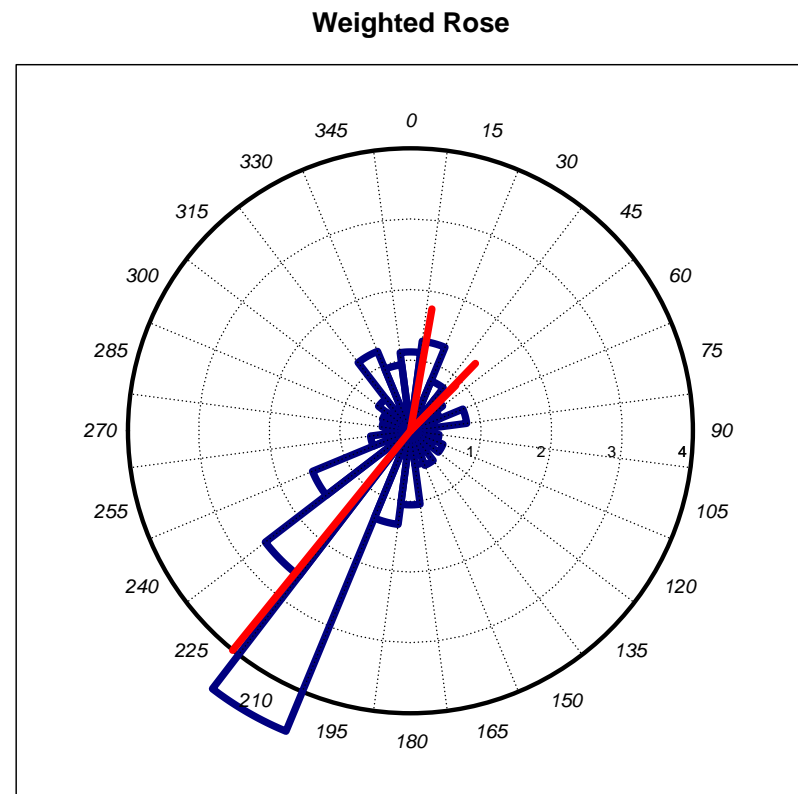
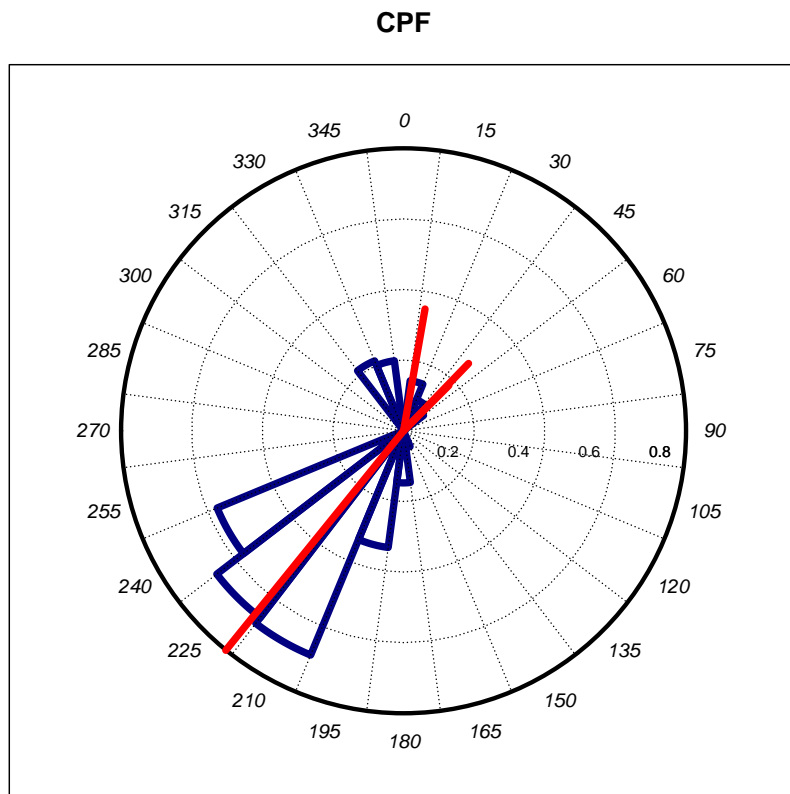
Profile Smoothness	Uncertainty	Source Profiles		Source Contributions	
		DP Model	PMF	DP Model	PMF
low	low (CV=0.2)	✓		✓	
low	high (CV=0.8)	✓		✓	✓
high	low	✓		✓	
high	high	✓		✓	✓
flat	low	✓		✓	
flat	high	✓		✓	

In the majority of circumstances, the DP model out performs PMF.

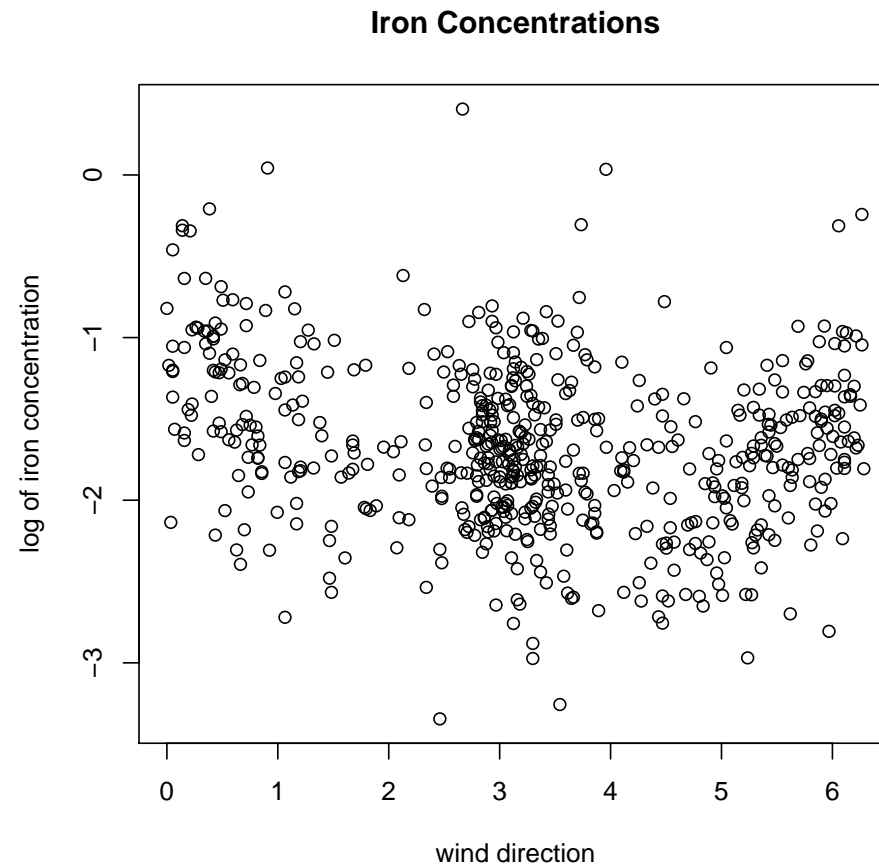
IV. Bayesian approach for the identification of pollution source directions

(Williams, Christensen, and Reese, in preparation)

Exploratory Graphical Methods



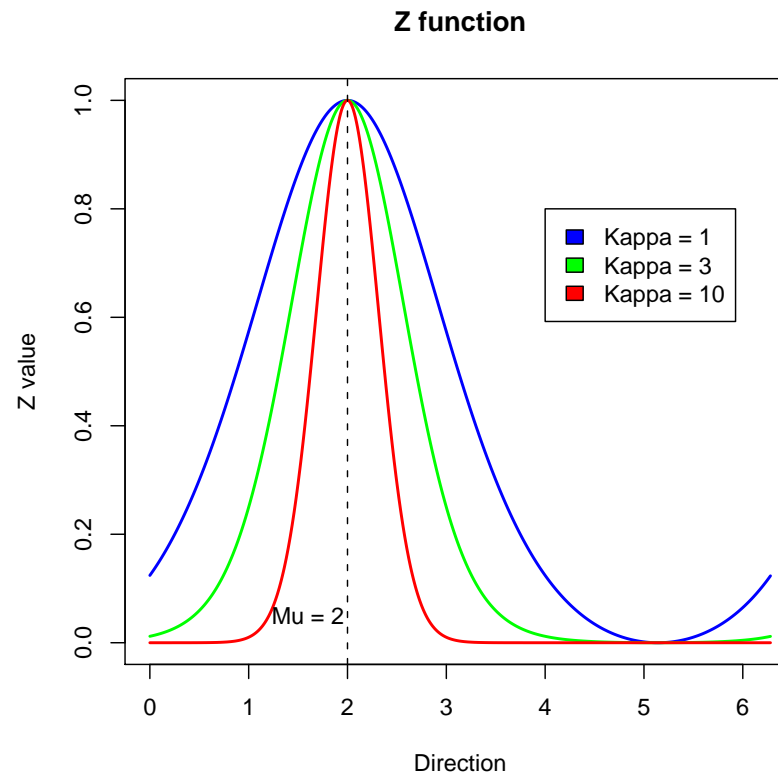
- Need method amenable to statistical inference
- Must account for the circular nature of the data



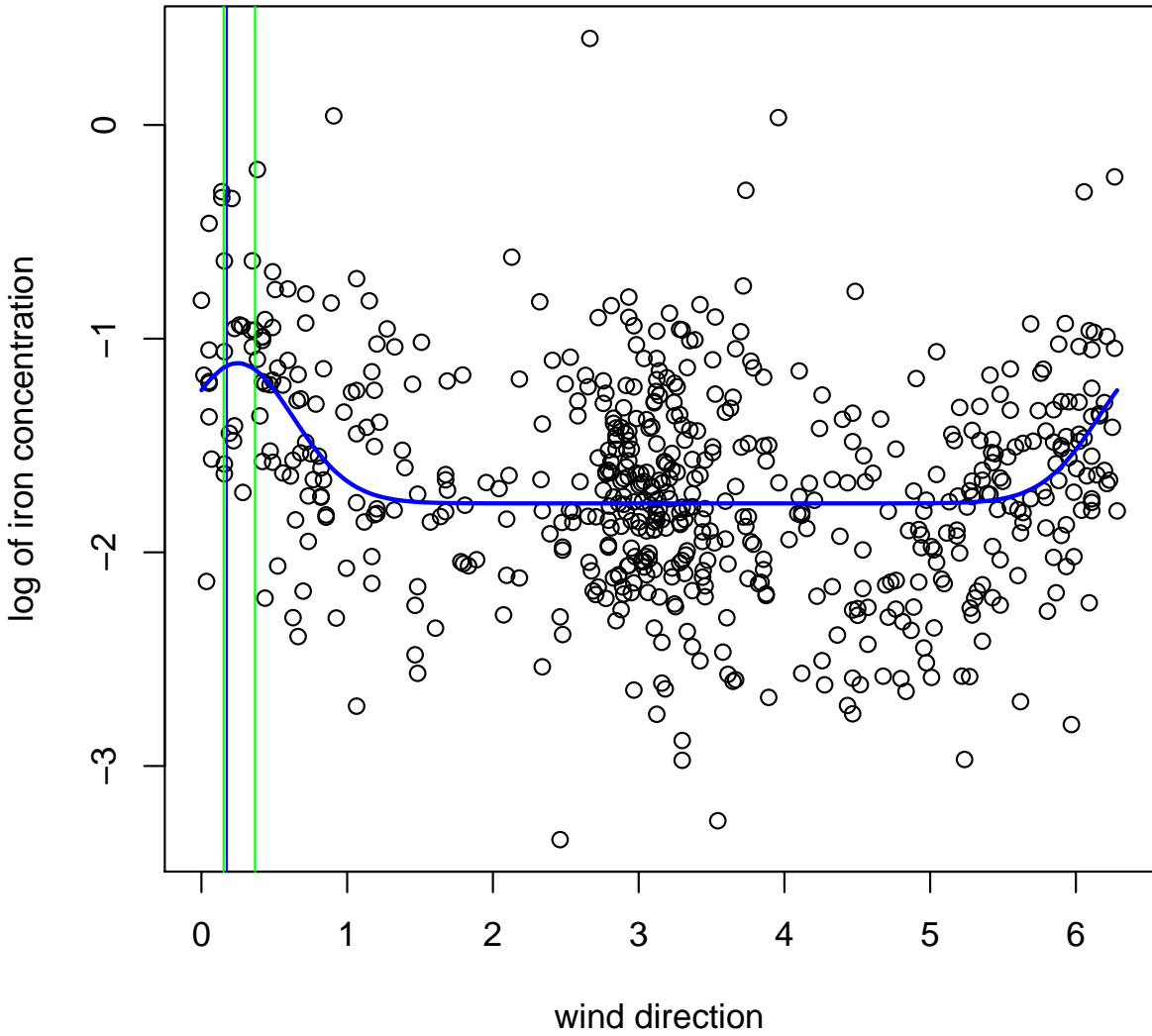
Model

$$y \sim LN(\beta_0 + \beta_1 Z(\theta, \mu, \kappa) + \beta_3 s, \sigma)$$

$$Z(\theta, \mu, \kappa) = \frac{e^{\kappa \cos(\theta - \mu)} - e^{-\kappa}}{e^{\kappa} - e^{-\kappa}}$$



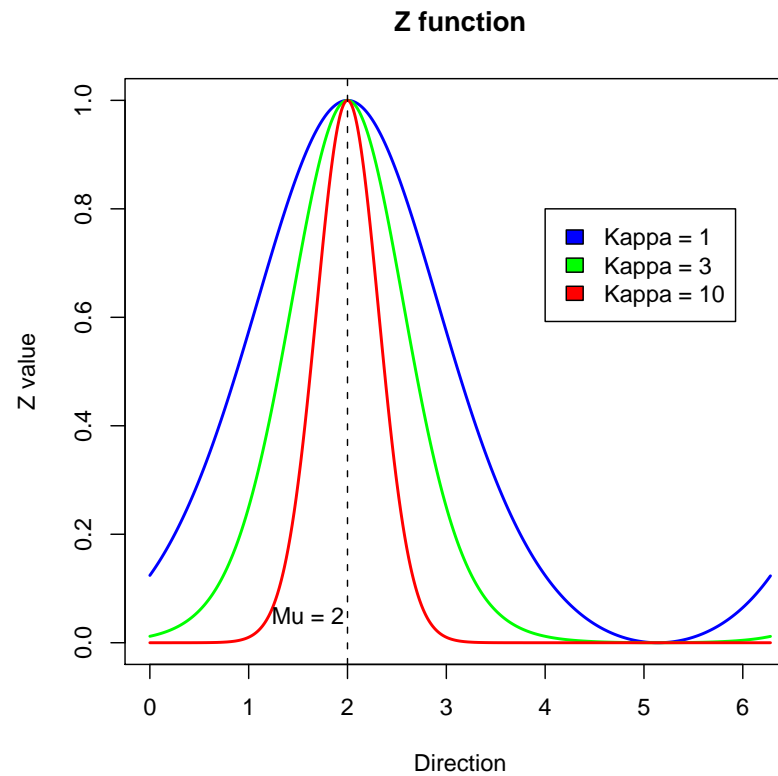
MCMC Results for Iron Analysis



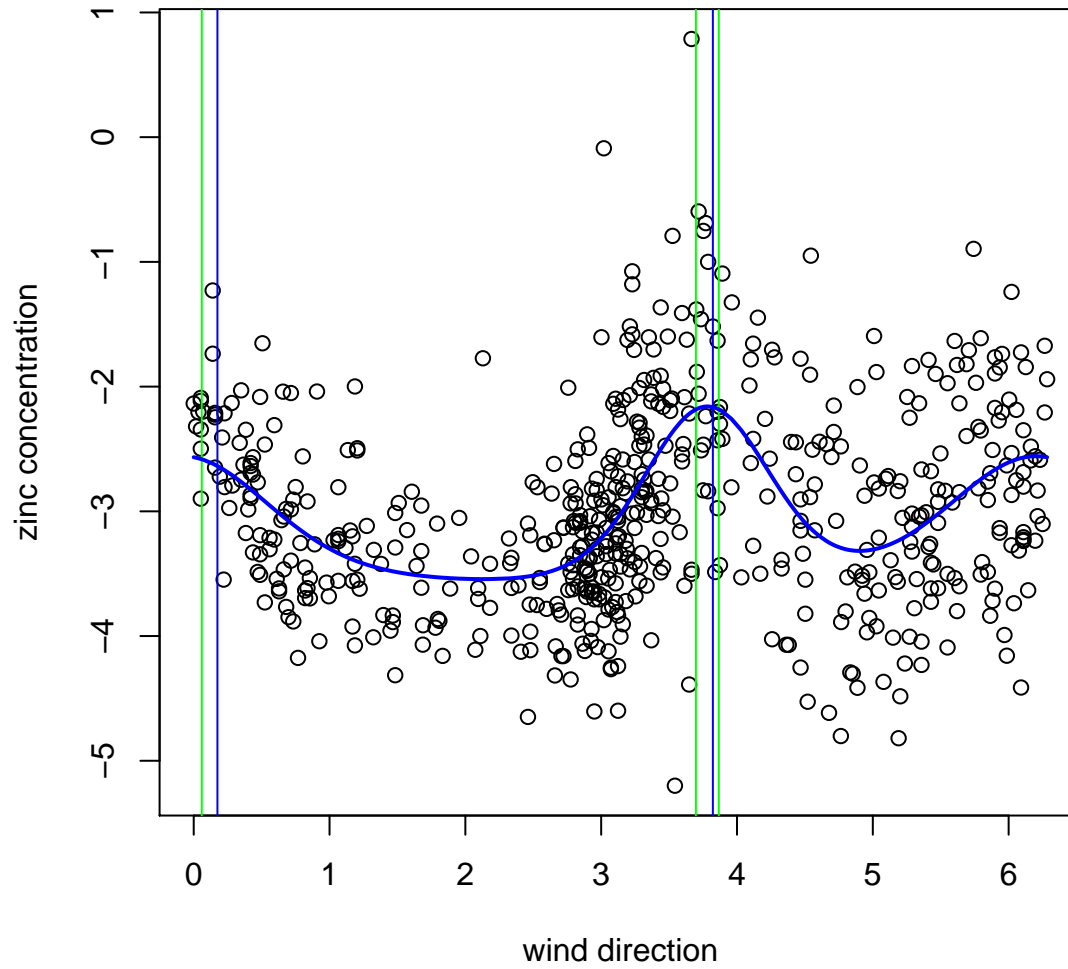
Two Source Model

$$y \sim LN(\beta_0 + \beta_1 Z(\theta, \mu_1, \kappa_1) + \beta_2 Z(\theta, \mu_2, \kappa_2) + \beta_3 s, \sigma)$$

$$Z(\theta, \mu, \kappa) = \frac{e^{\kappa \cos(\theta - \mu)} - e^{-\kappa}}{e^{\kappa} - e^{-\kappa}}$$



MCMC Result



V. Conclusions and additional research directions

- Bayesian approach has several advantages:
 - Efficient use of auxiliary information (in construction of priors)
 - * Partial source profile information
 - * Seasonal, meteorological, phenomenological effects on sources
 - Potential for incorporating partial information synthesizing data measured with differing temporal resolution (e.g., OC & EC available hourly while organics only measured weekly or monthly)
 - Potential for time varying source profiles along with time varying source contributions
 - In simulation, compares well with other source apportionment methods

- Current and future research directions
 - PSA using a priori information and PMF (Lingwall and Christensen, 2007)
 - Clustering species using size distribution data (Christensen, Dillner, Schauer, and Reese, 2007)
 - Species influence in PSA using PMF (Christensen and Schauer, in preparation)
 - Embedding deterministic dispersion model (AERMOD) into a Bayesian hierarchical model for identifying sources (current work)
 - Integrating meteorological information in PSA (current work)
 - Application to St. Louis Supersite data (current work)