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

This research is funded by
U.S. EPA - Science To Achieve Results (STAR) Program
Grant # RD-83216001-0

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

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

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

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

- $\boldsymbol{\Lambda}$ unknown $\Rightarrow$ model is called *multivariate receptor model* and is fit using factor analytic methods

- $\boldsymbol{\Lambda}$ known $\Rightarrow$ model is called *chemical mass balance model* and is fit using regression methods

"Receptor Models"

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

Little knowledge about pollution sources

"Perfect" knowledge about pollution sources

Multivariate Receptor Model

Chemical Mass Balance Model

UNMIX

Positive Matrix Factorization (PMF)

Exploratory Factor Analysis Models

- -- k is unknown
- -- Pollution source profiles (Λ) unknown
- -- Multiple ambient measures required

Confirmatory Factor Analysis Models

- -- Hypothesized k is assessed by GOF
- -- Some pollution source info is known
- -- Multiple ambient measures required

Bayesian Models

- -- k is known or hypothesized
- -- Priors on elements of Λ

Measurement Error Models

- -- k is known
- -- Λ is "known" up to known measurement error

Regression Models

- -- k is known
- -- Λ is known

Ecology Letters, (2005) **8**: 2–14 doi: 10.1111/j.1461-0248.2004.00702.x

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):

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

"data model" "process model" "parameter model"

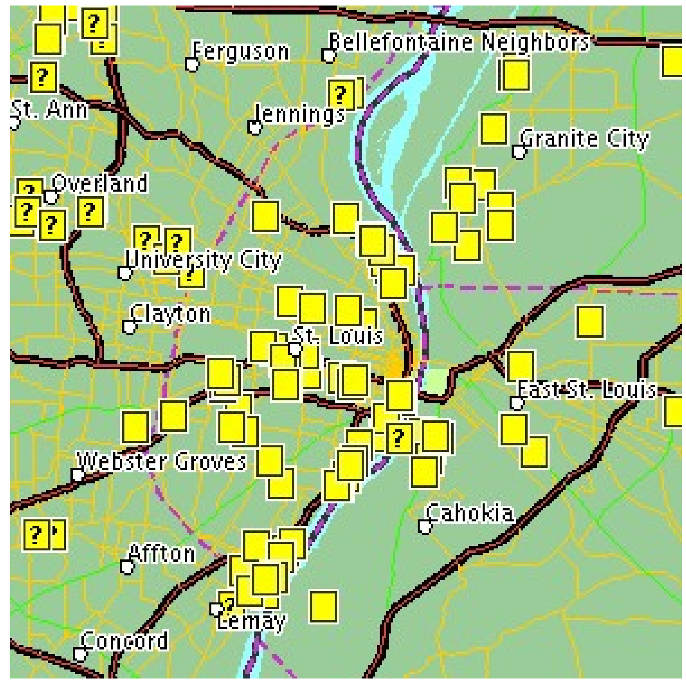
1. "Data": ambient PM concentrations, meteorological data

2. "Process": transport/dispersion, meteorology, seasonality, atmospheric chemistry, etc.

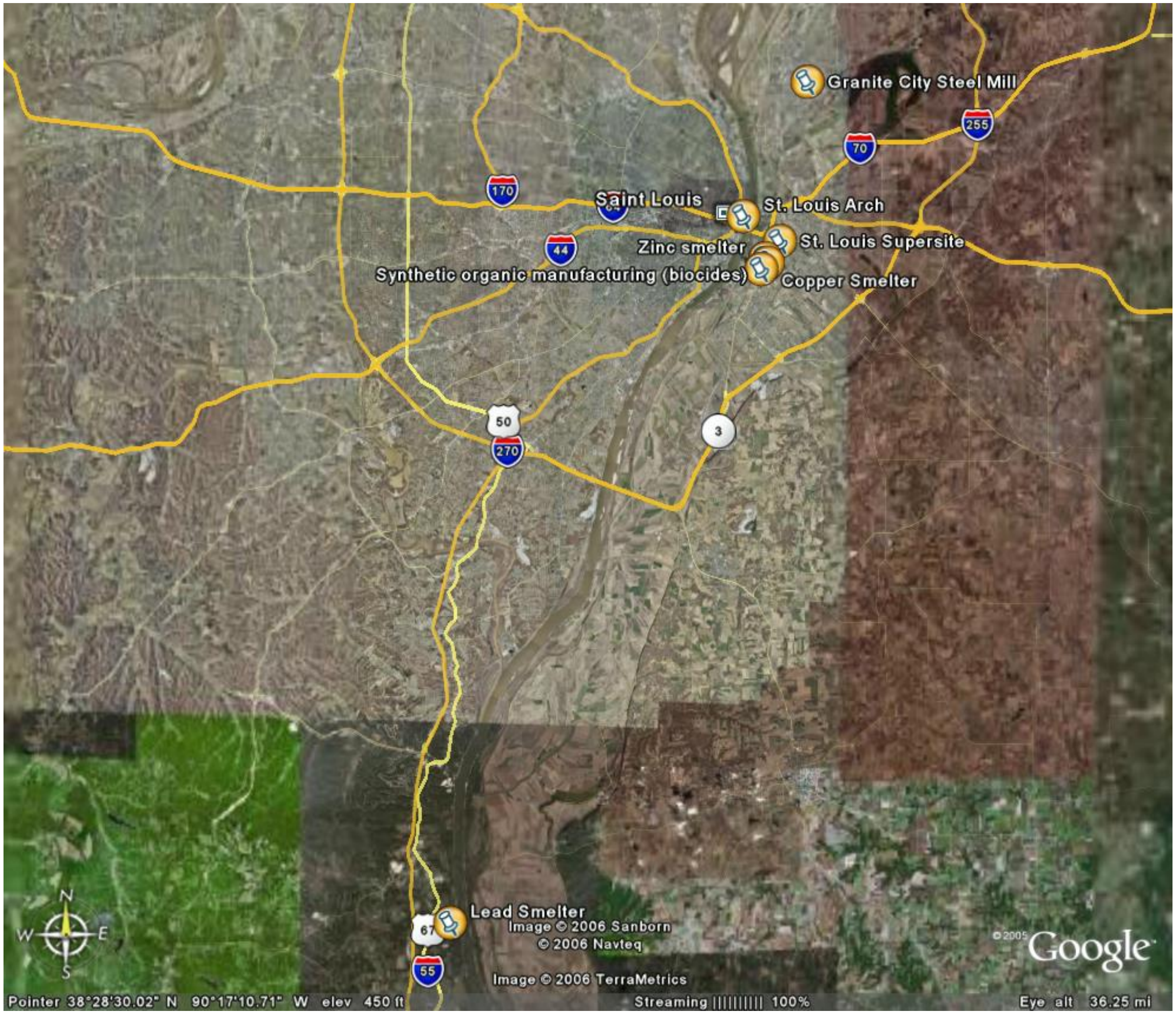
3. "Parameters": daily source contribution values, source profile values

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

- Auxiliary information for enhancing source apportionment



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


Granite City Steel Mill
255
70
170
Saint Louis
St. Louis Arch
44
St. Louis Supersite
Zinc smelter
Synthetic organic manufacturing (biocides)
Copper Smelter
50
270
3
Lead Smelter
Image © 2006 Sanborn
© 2006 Navteq
Image © 2006 TerraMetrics
67
55
Google
Pointer 38°28'30.02" N 90°17'10.71" W elev 450 ft
Streaming |||||||||| 100%
Eye alt 36.25 mi

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, $\mathbf{X}$

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

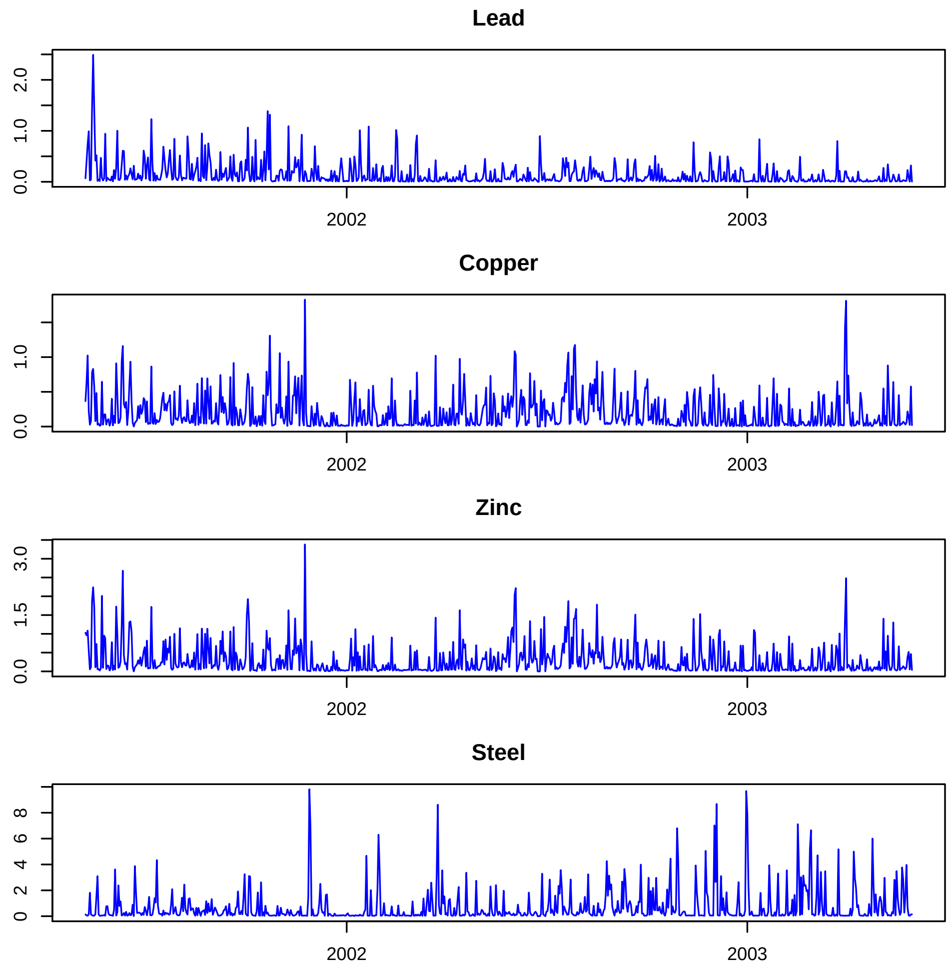
- Likelihood and Priors:
  - Source profiles (columns of $\mathbf{\Lambda}$) $\sim$ Generalized Dirichlet (Rogers and Young, 1973)
    - Individual elements of an *a priori* source profile ($\tilde{\boldsymbol{\lambda}}_k$) are associated with different degrees of certainty, but variances of elements of Dirichlet vector cannot be individually tuned

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

    - 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:
  - $\text{PMF}_{\tilde{\Lambda}}$ (uses *a priori* information on $\mathbf{\Lambda}$)
  - PMF (does *not* use *a priori* information on $\mathbf{\Lambda}$)
  - $\text{Bayesian}_{\tilde{\Lambda}}$ (uses *a priori* information on $\mathbf{\Lambda}$)
  - Bayesian (does *not* use *a priori* information on $\mathbf{\Lambda}$)

TMAE for estimating source contributions and source profiles

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

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

(Heaton, Reese, and Christensen, in preparation)

Dirichlet Process (DP) Model

$$\mathbf{y}_t|\mathbf{\Lambda}_t, \mathbf{f}_t, \mathbf{\Sigma} \sim \mathsf{LN}\left[\mathbf{\Lambda}_t \mathbf{f}_t, \mathbf{\Sigma}\right]$$

$$\boldsymbol{\lambda}_{kt} \sim \mathsf{DIR}\left[g_k \boldsymbol{\lambda}_{k(t-1)}\right]$$

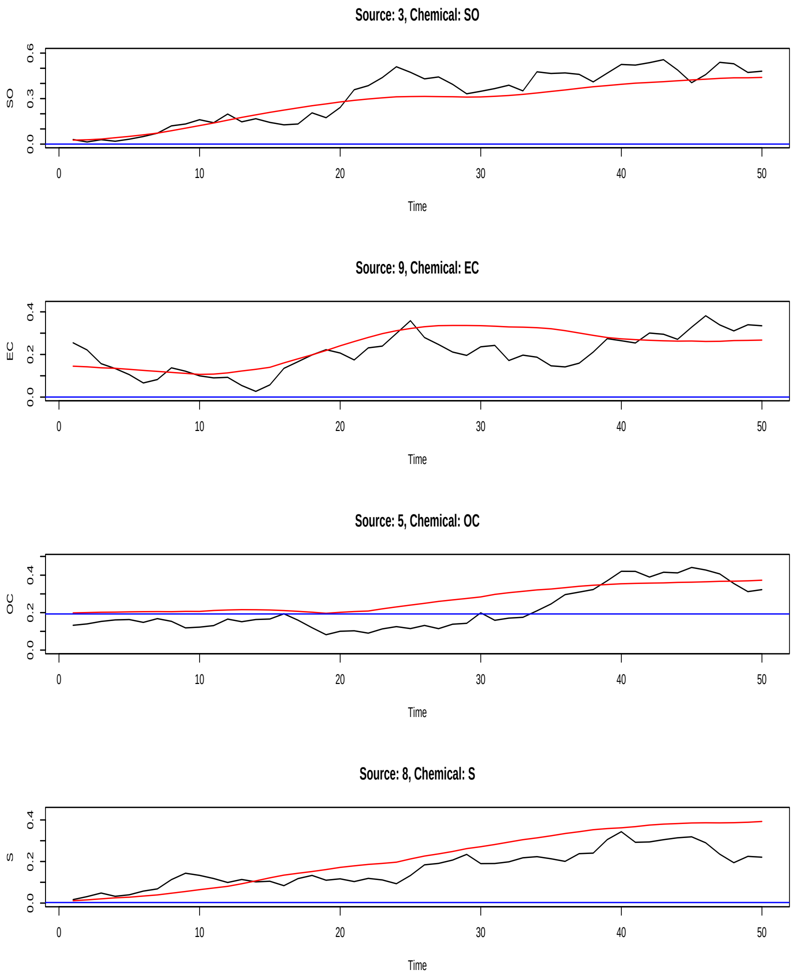
Assumptions

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

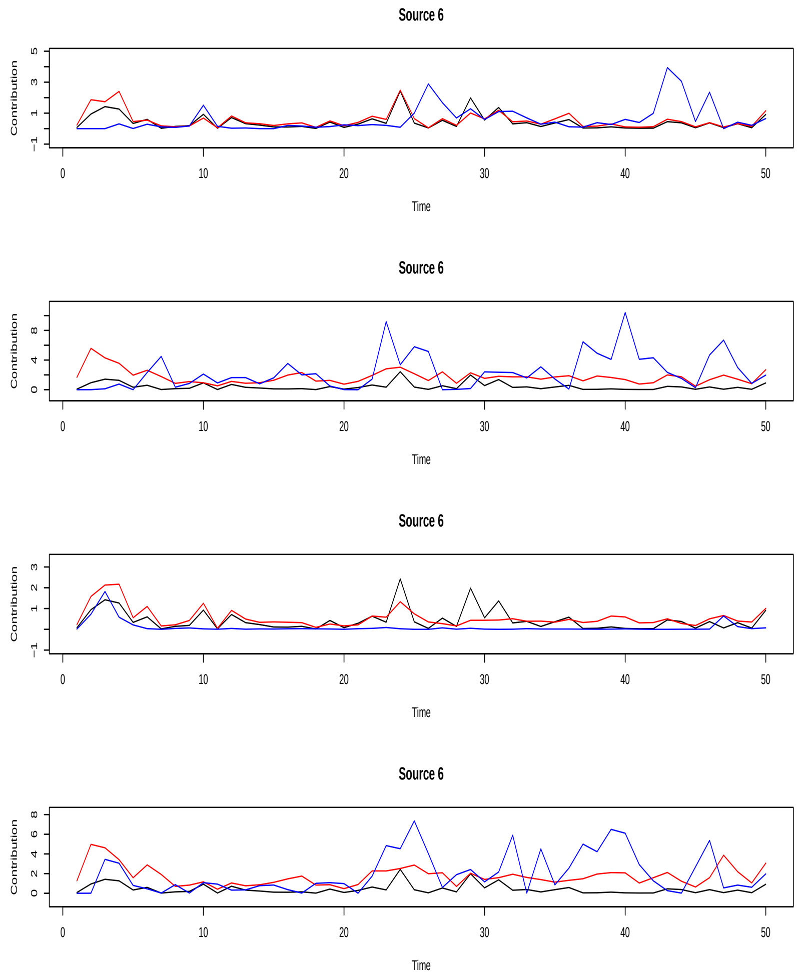
Goal: Compare DP Model to PMF by simulating data sets under varying degrees of variability in $\mathbf{y}_t$ and $\mathbf{\Lambda}_t$.

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

Profile Plots

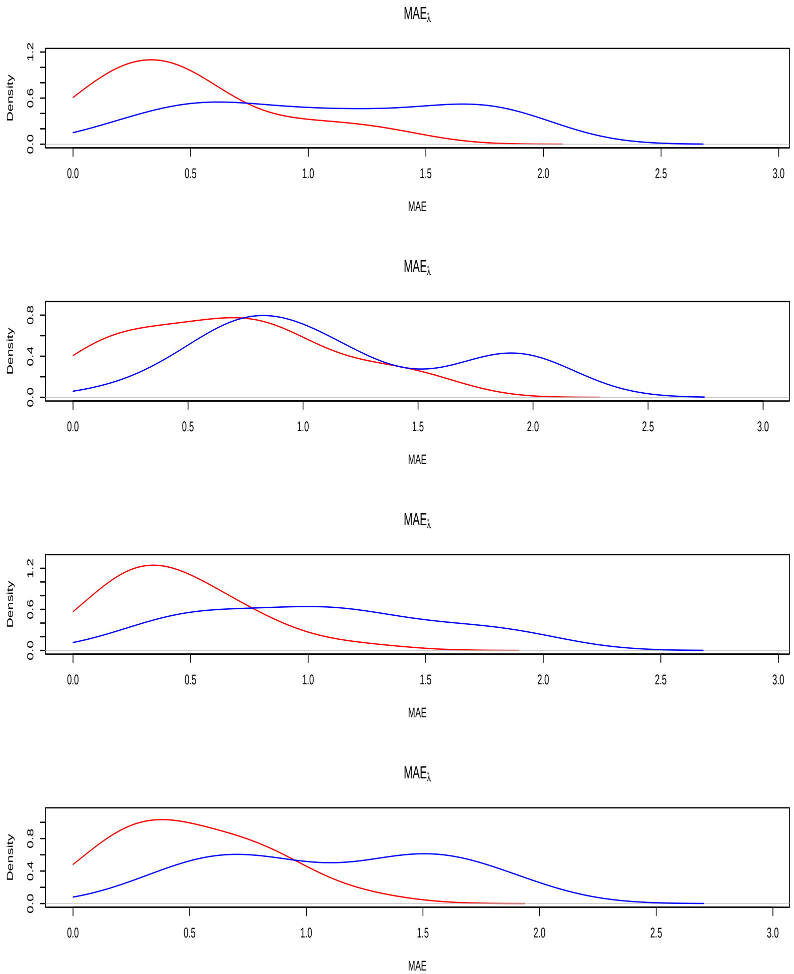


Contribution Plots

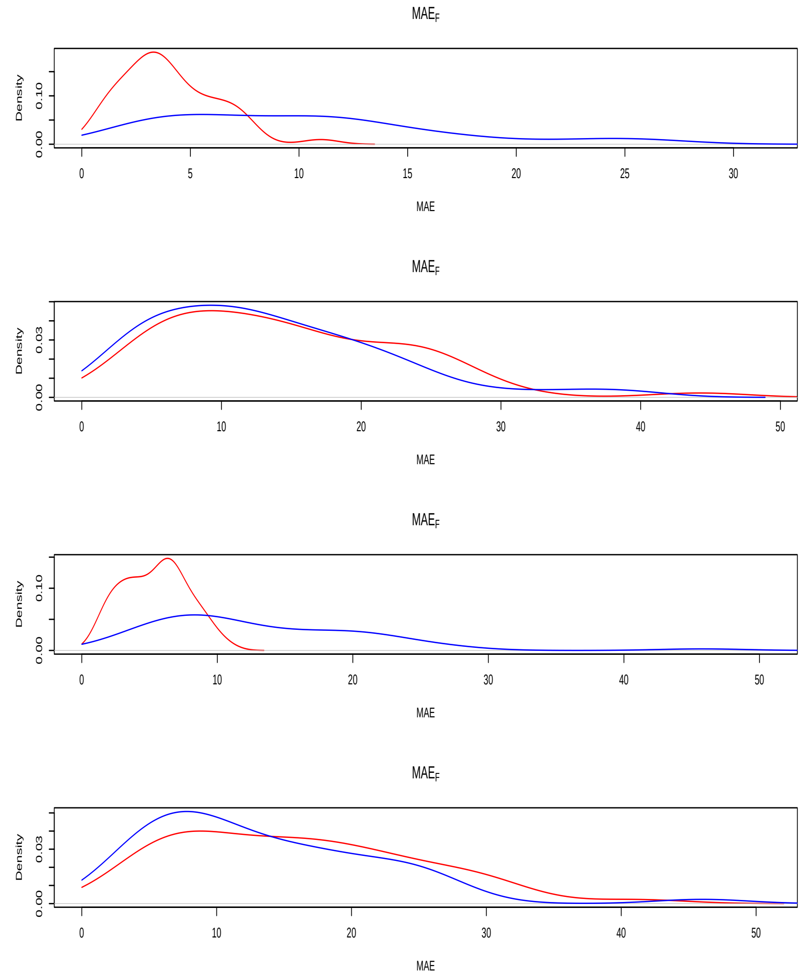


Comparison under time-varying profiles (Bayesian,PMF)

MAE for $\widehat{\mathbf{\Lambda}}_t$

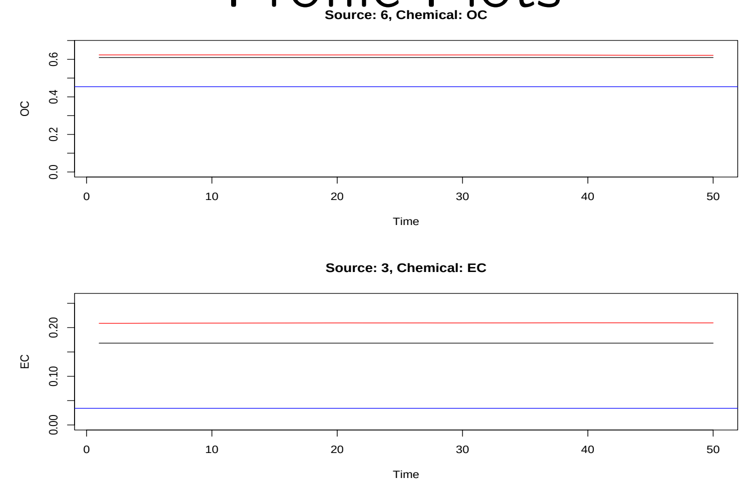


MAE for $\widehat{\mathbf{f}}_t$

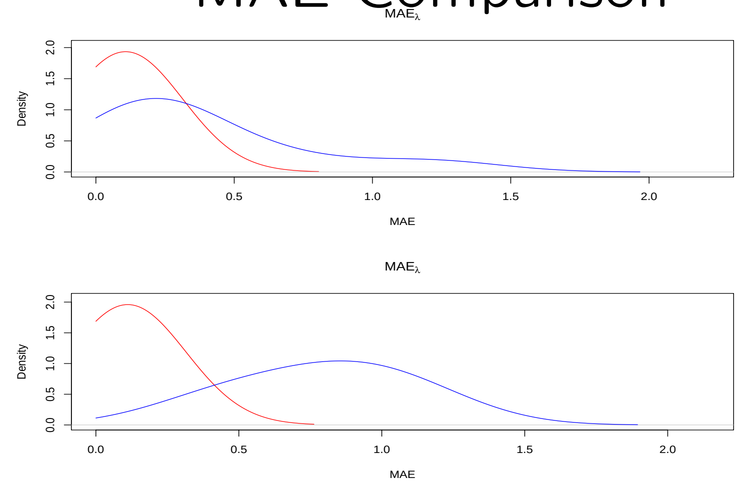


Comparison under time-constant profiles (True,Bayesian,PMF)

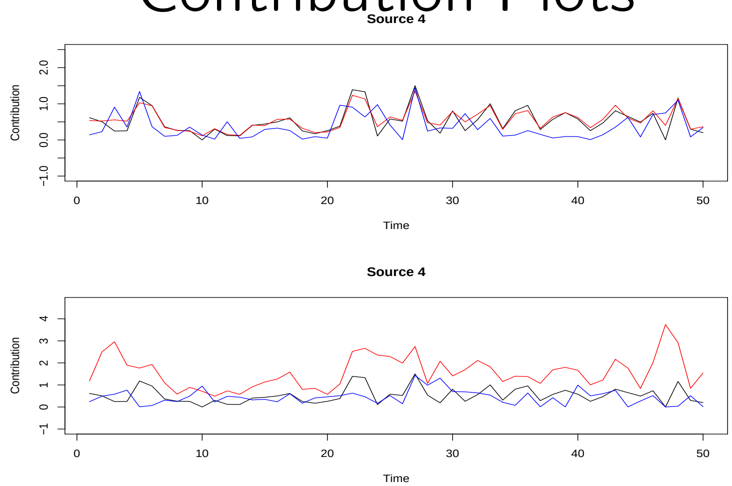
Profile Plots



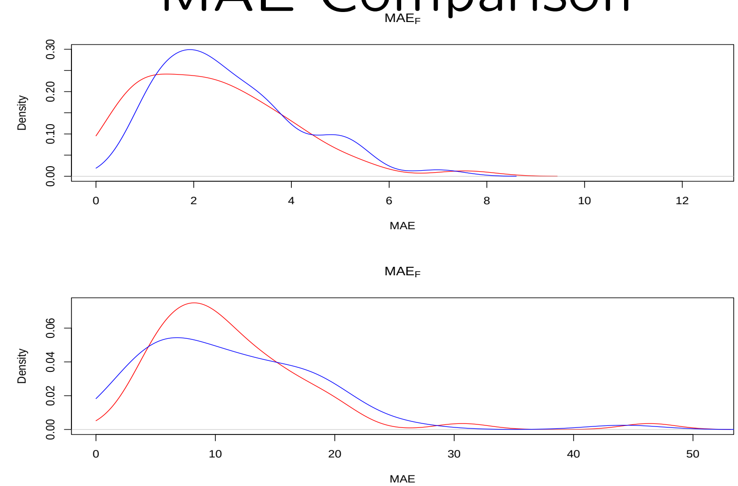
MAE Comparison



Contribution Plots



MAE Comparison



Summary of DP Model Performance

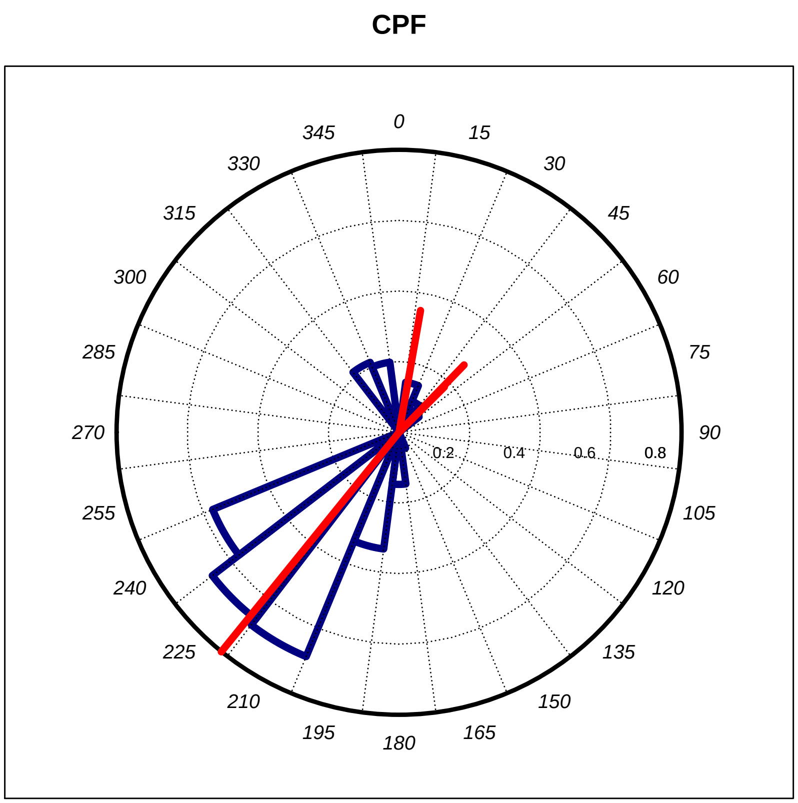
		Source Profiles		Source Contributions	
Profile Smooth-ness	Uncertainty	DP Model	PMF	DP Model	PMF
low	low (CV=0.2)	✓		✓	
low	high (CV=0.8)	✓		✓	✓
high	low	✓		✓	
high	high	✓		✓	✓
flat	low	✓		✓	
flat	high	✓		✓	

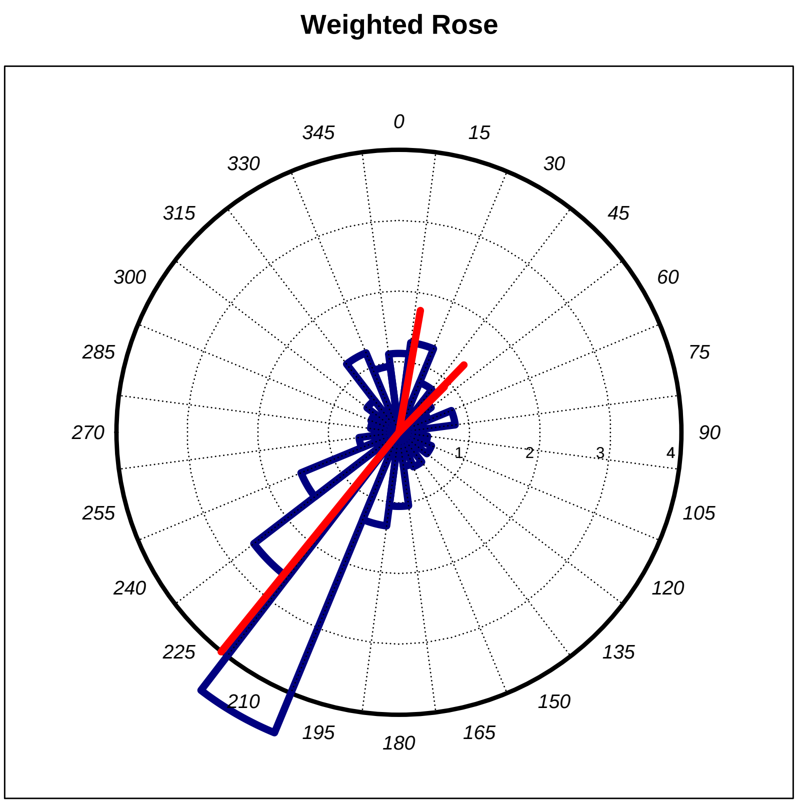
In the majority of circumstances, the DP model out performs PMF.

IV. Bayesian approach for the identification of pollution source directions

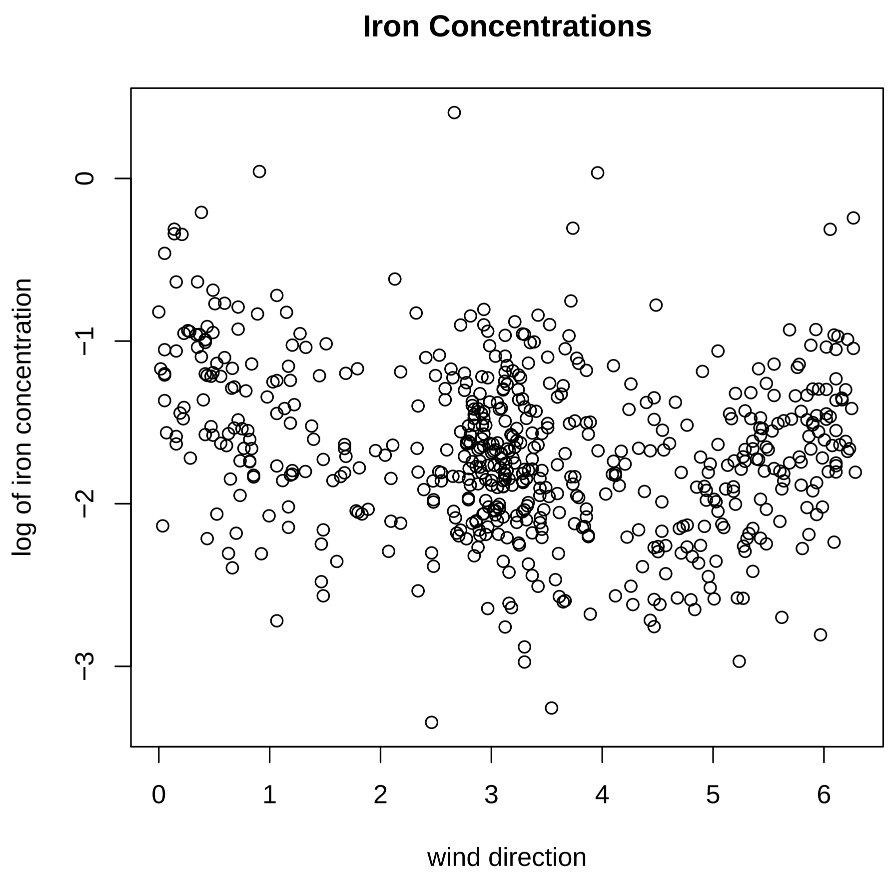
(Williams, Christensen, and Reese, in preparation)

Exploratory Graphical Methods





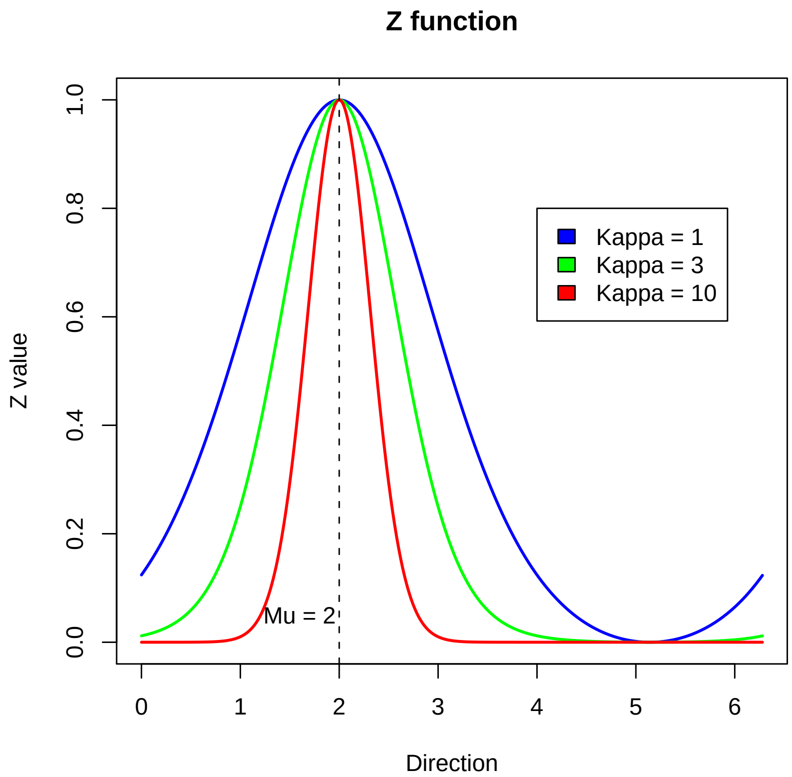
- Need method amenable to statistical inference
- Must account for the circular nature of the data



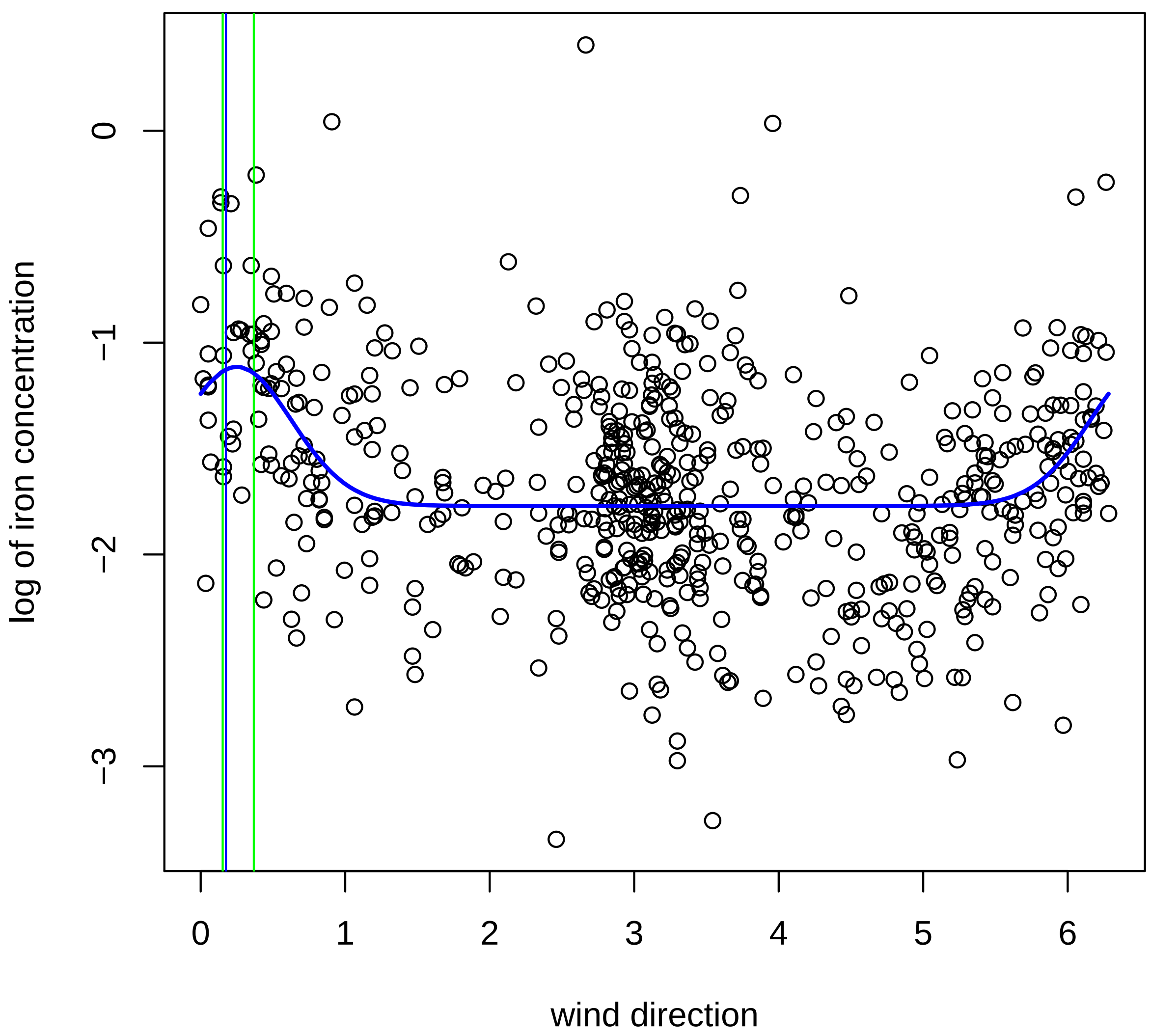
Model

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

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



MCMC Results for Iron Analysis

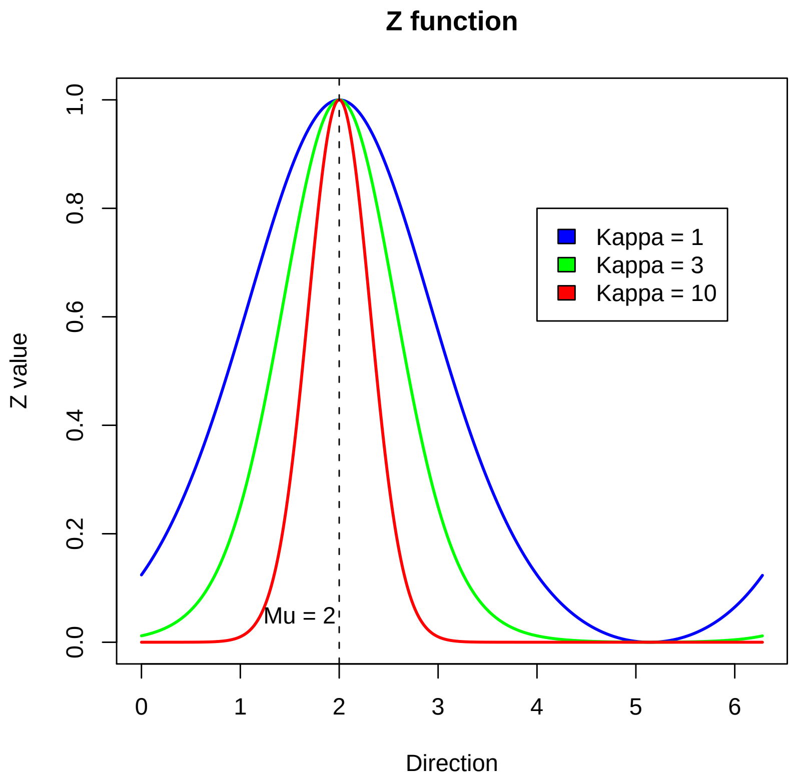


US EPA ARCHIVE DOCUMENT

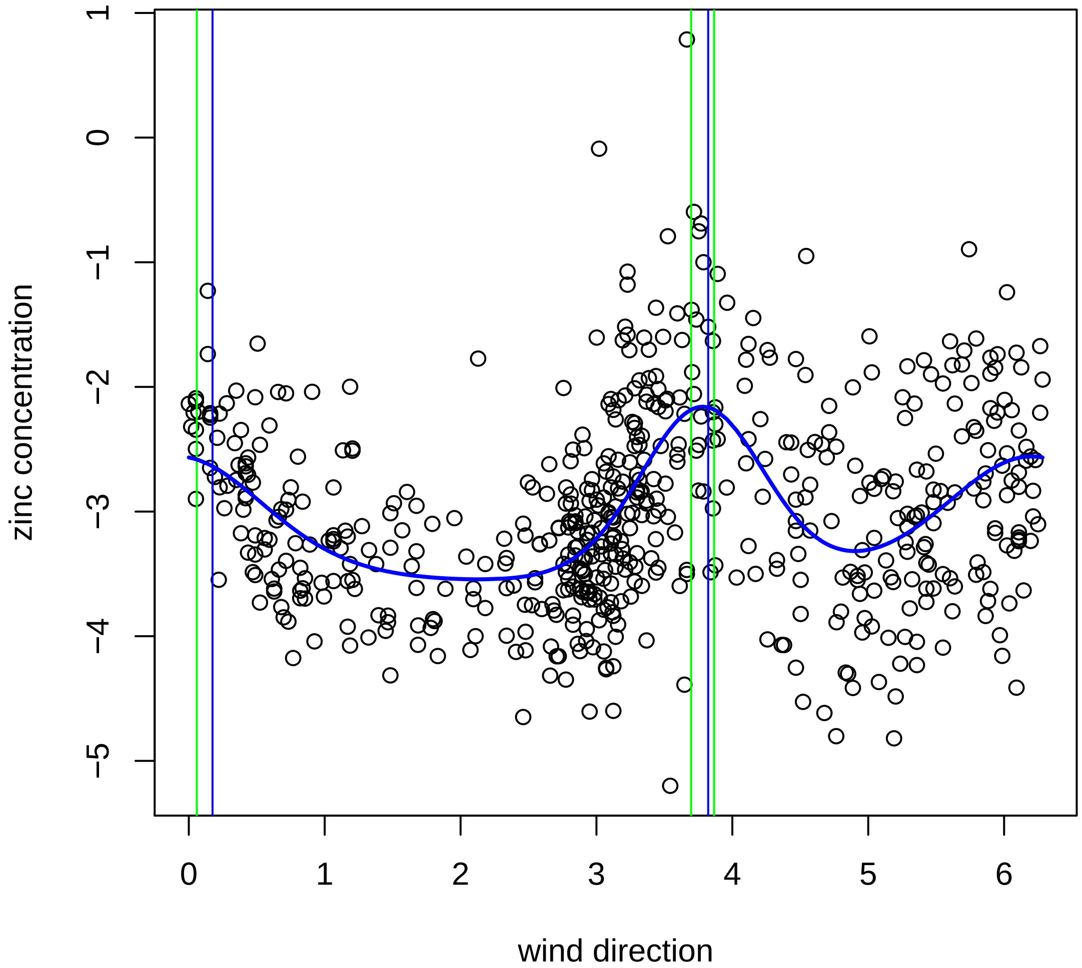
Two Source Model

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

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



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, phenomenonological 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)