Statistical Analysis of Surface Water Data

Time series segmentation

Outline

Outline

Surface Water Data

- It includes *irregularly* spaced time series of Barium and Sulphate concentrations at approx. 80 PA counties from 1921-2015
- Deciphering spatial and temporal correlations may provide important insights about water quality deterioration due to energy extraction processes
- **Challenges**: spatial and temporal "sparsity"
- **Our current work**:
  - Naive spatial clustering using k-means
  - Time series segmentation

Surface Water Data

- It includes *irregularly* spaced time series of Barium and Sulphate concentrations at approx. 80 PA counties from 1921-2015
- Deciphering spatial and temporal correlations may provide important insights about water quality deterioration due to energy extraction processes
- **Challenges**: spatial and temporal "sparsity"
- **Our current work**:
  - Naive spatial clustering using k-means
  - Time series segmentation

Outline

Spatial K means clustering

- K means clustering over latitude and longitude.
- Criteria: Spatial density of points;
  - Dense measurements over one region form a cluster.
- Clustering selection:
  - Number of clusters chosen using **scree plot**.
  - Sum of squares within clusters vs. number of clusters.

Outline

Model and Goals
Time series segmentation

- Assumption: There are m breakpoints.
- Model[1]:

$$\begin{aligned} \text{Ba}_i(\text{Su}_i) &= \text{Time}_i\beta_1 + u_i \quad (i = 1, ..., T_1) \\ &\vdots \\ \text{Ba}_i(\text{Su}_i) &= \text{Time}_i\beta_m + u_i \quad (i = T_{m+1}, ..., T) \end{aligned}$$

- Goals:
  - Point estimate of breakpoints.
  - Interval estimate of breakpoints.

[1] Bai, Jushan, and Pierre Perron. "Computation and analysis of multiple structural change models." Journal of applied econometrics 18.1 (2003): 1-22.

Methodology
Time series segmentation

- Notation:
  - h: Minimum segment length.
  - $\text{SSR}^{r}_{t_1:t_2}$: Sum of squared residuals for the time segment t_1-t_2 with r breakpoints.
- Recursive problem:

$$\text{SSR}_{1:T} = \min_{mh \leq j \leq T-h} \left[\text{SSR}^{m-1}_{1:j} + \text{SSR}^{0}_{(j+1):T} \right]$$

- Construction of triangular matrix of sums of squared residuals.
- Time Complexity: $O(T^2)$

Confidence intervals for breakpoints

- Asymptotic distribution of breakpoint estimate:

$$A_T(\hat{T}_i - T_i) \xrightarrow{d} \text{some Wiener process}$$

  where A_T is a normalization constant.

- Using above asymptotic distribution function of the breakpoint.[2], 95% confidence intervals for point estimates $\hat{T}_1$,..., $\hat{T}_m$ can be created.

[2]Bai, Jushan. "Estimation of a change point in multiple regression models." Review of Economics and Statistics 79.4 (1997): 551-563.

Choosing the number of segments

- Bayesian Information Criterion (BIC):

$$\text{BIC} = -2\log(\hat{L}) + k\log(n)$$

  - Limitation in our case: $n >> k$ condition not satisfied by Barium data for some clusters
- Residual sum of squares:
  - Always decreases and so can't be used for choosing segmentation.

Outline

Outline

Scree plot for K Means clustering

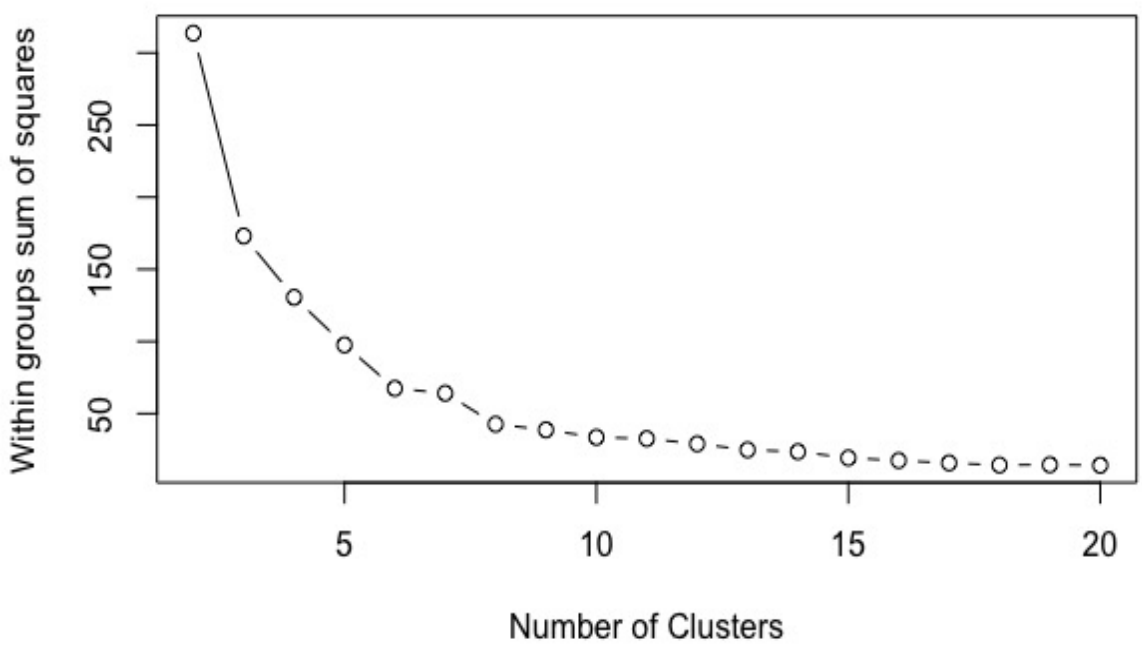

Figure: Scree plot for K Means clustering for Barium; K chosen as **8** based on elbow

Visualization of spatial clusters

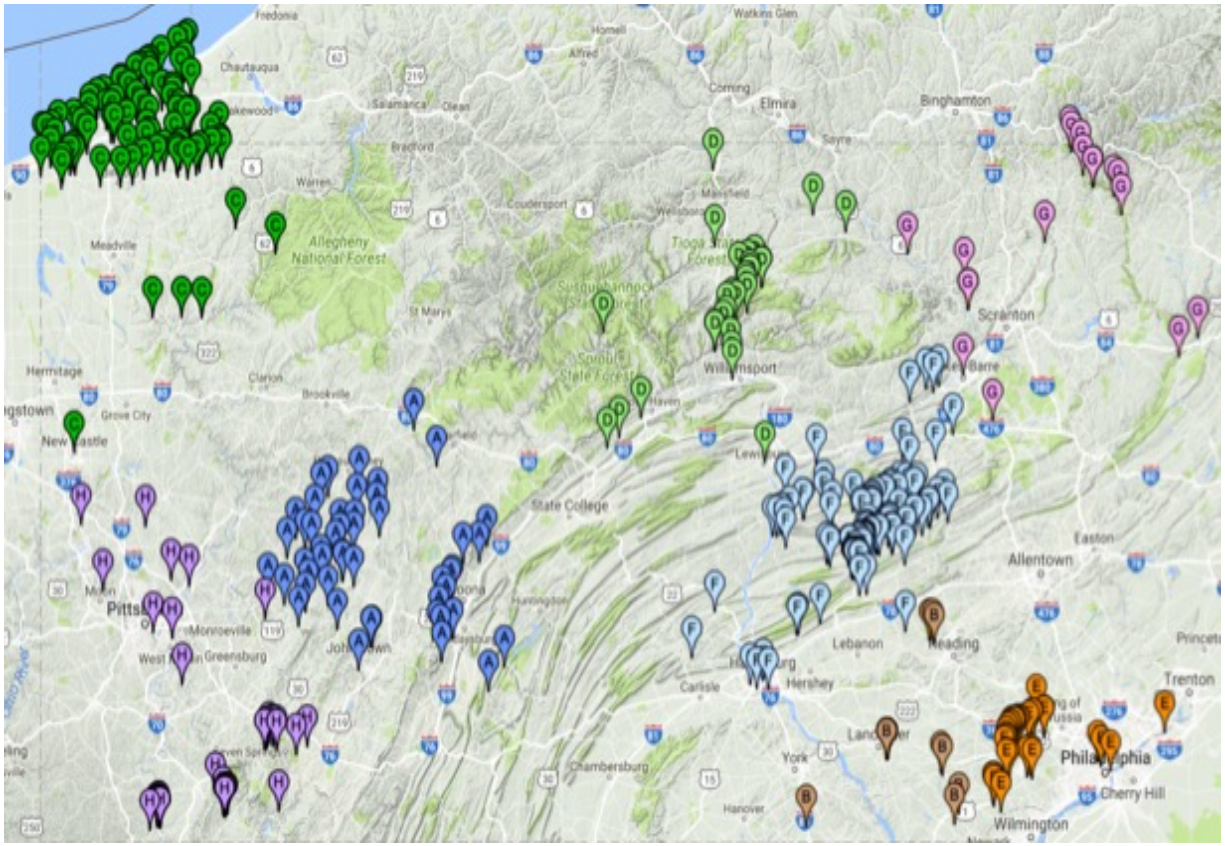

Figure: Visualization of spatial clusters on map

Outline

Sample BIC and RSS plots for cluster A

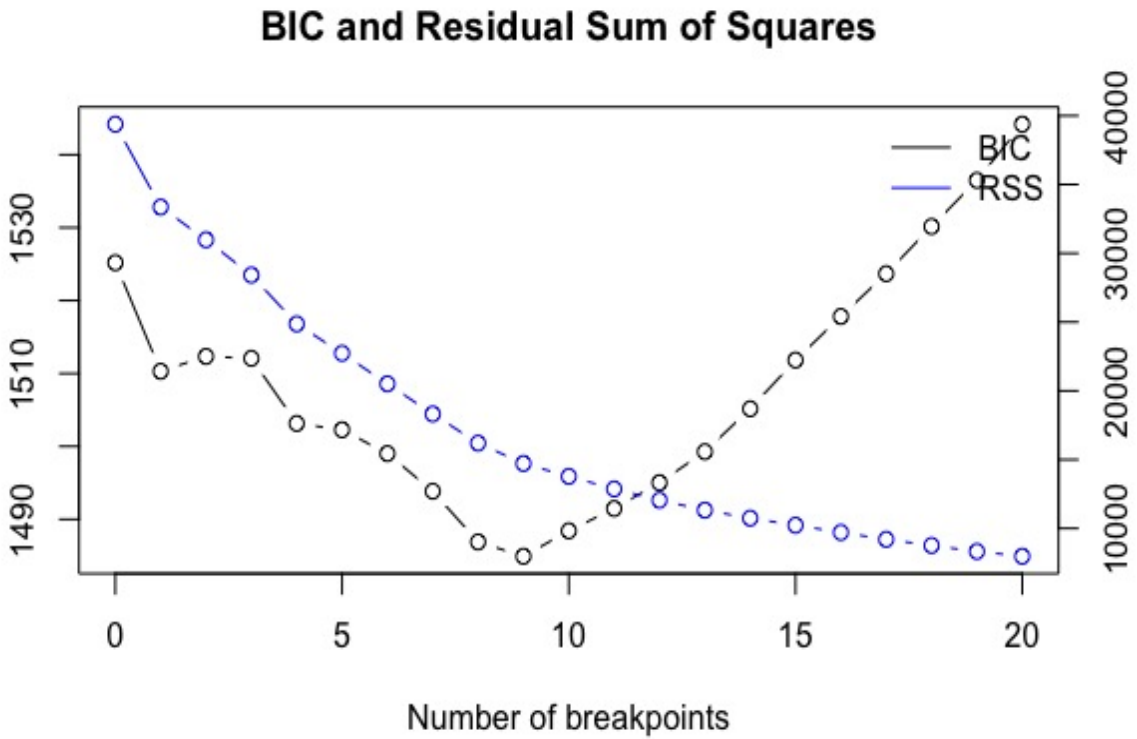

Figure: BIC and RSS for cluster A

Cluster A

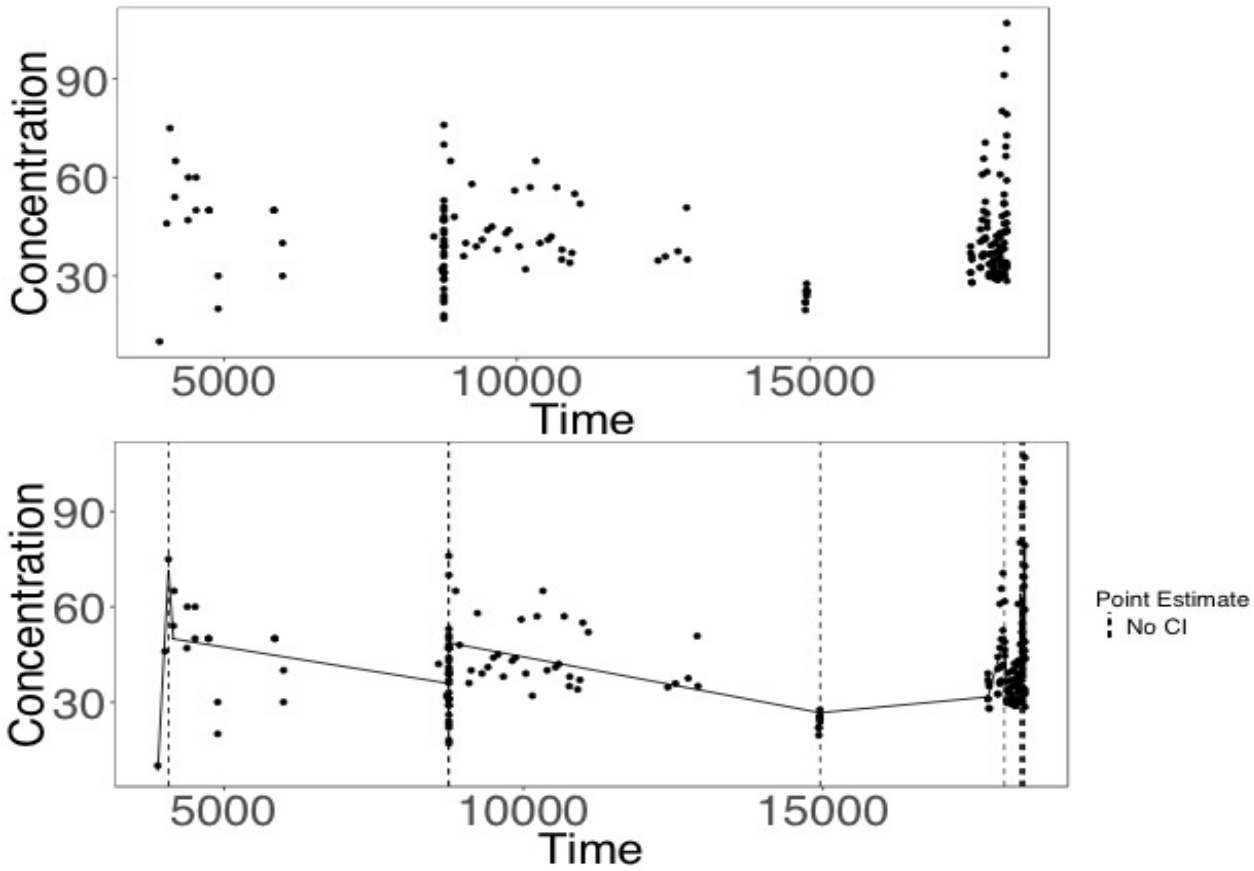

Figure: Cluster A with 184 observations and 9 breakpoints

Cluster B

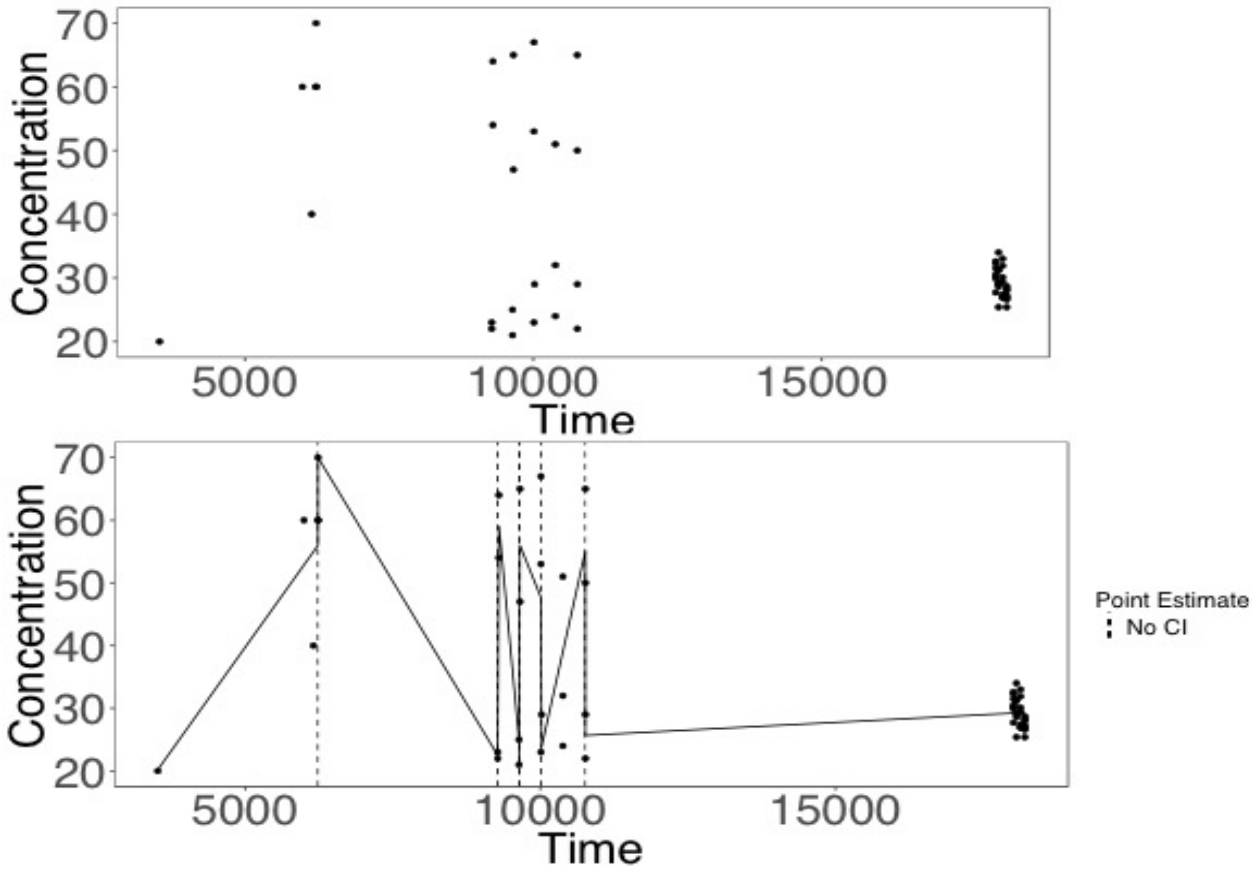

Figure: Cluster B with 53 observations and 6 breakpoints

Cluster C

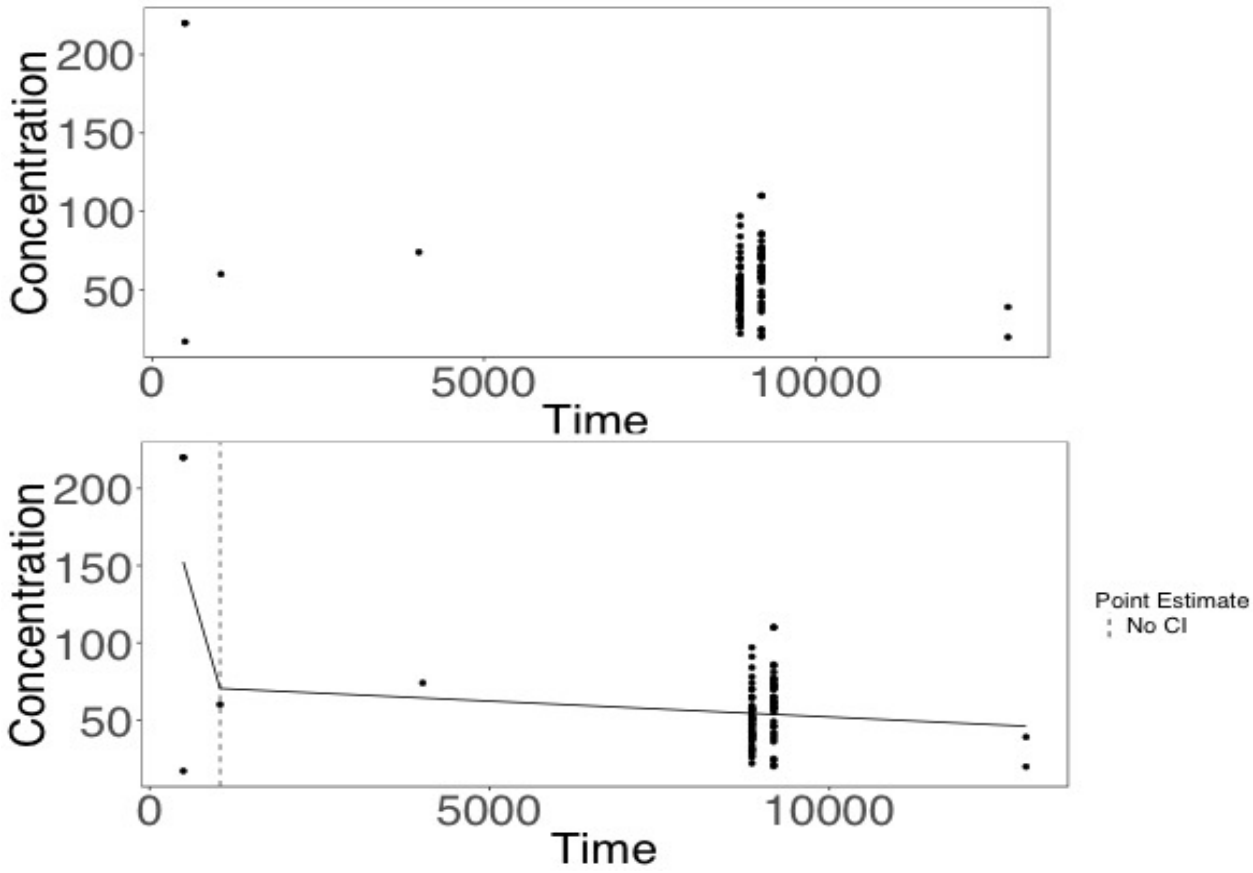

Figure: Cluster C with 105 observations and 1 breakpoint

Cluster D

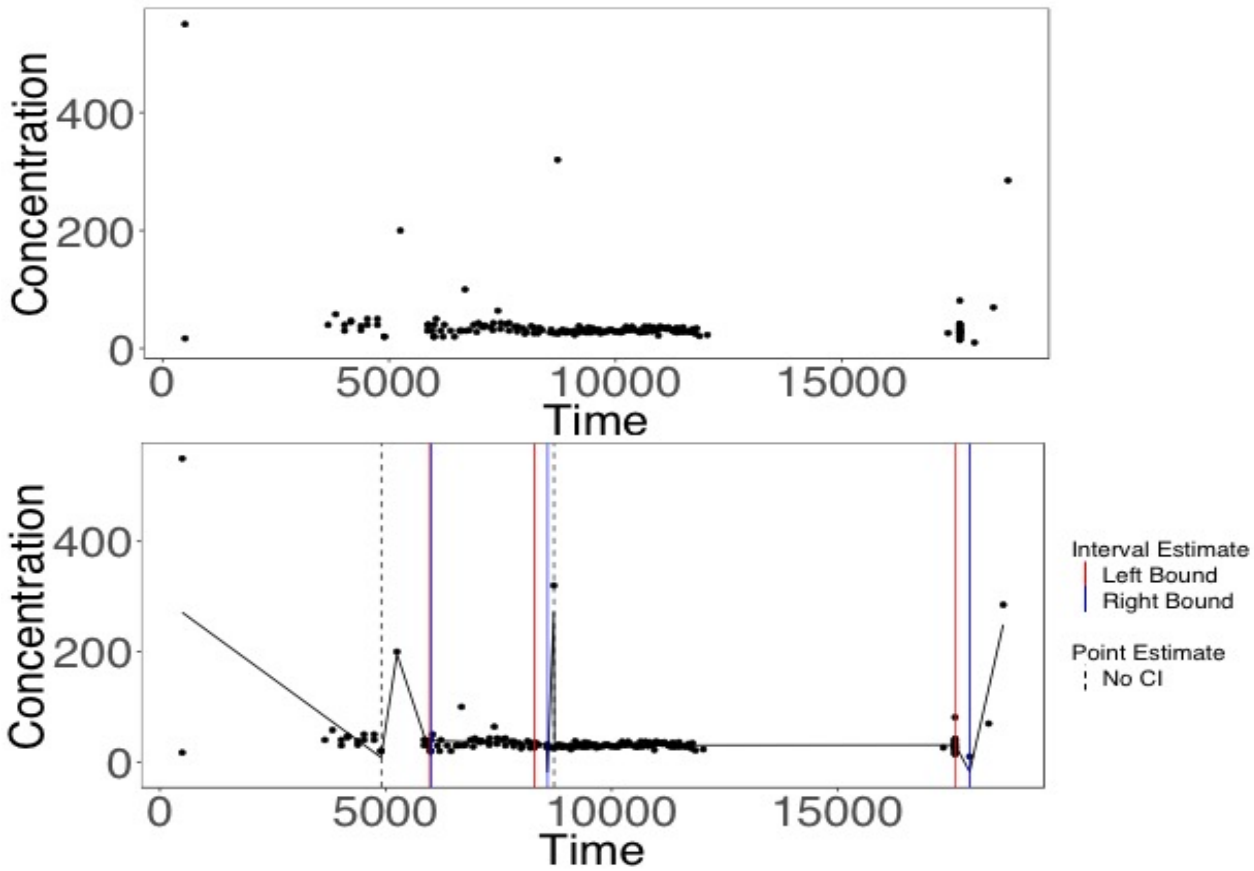

Figure: Cluster D with 162 observations and 5 breakpoints

Cluster E

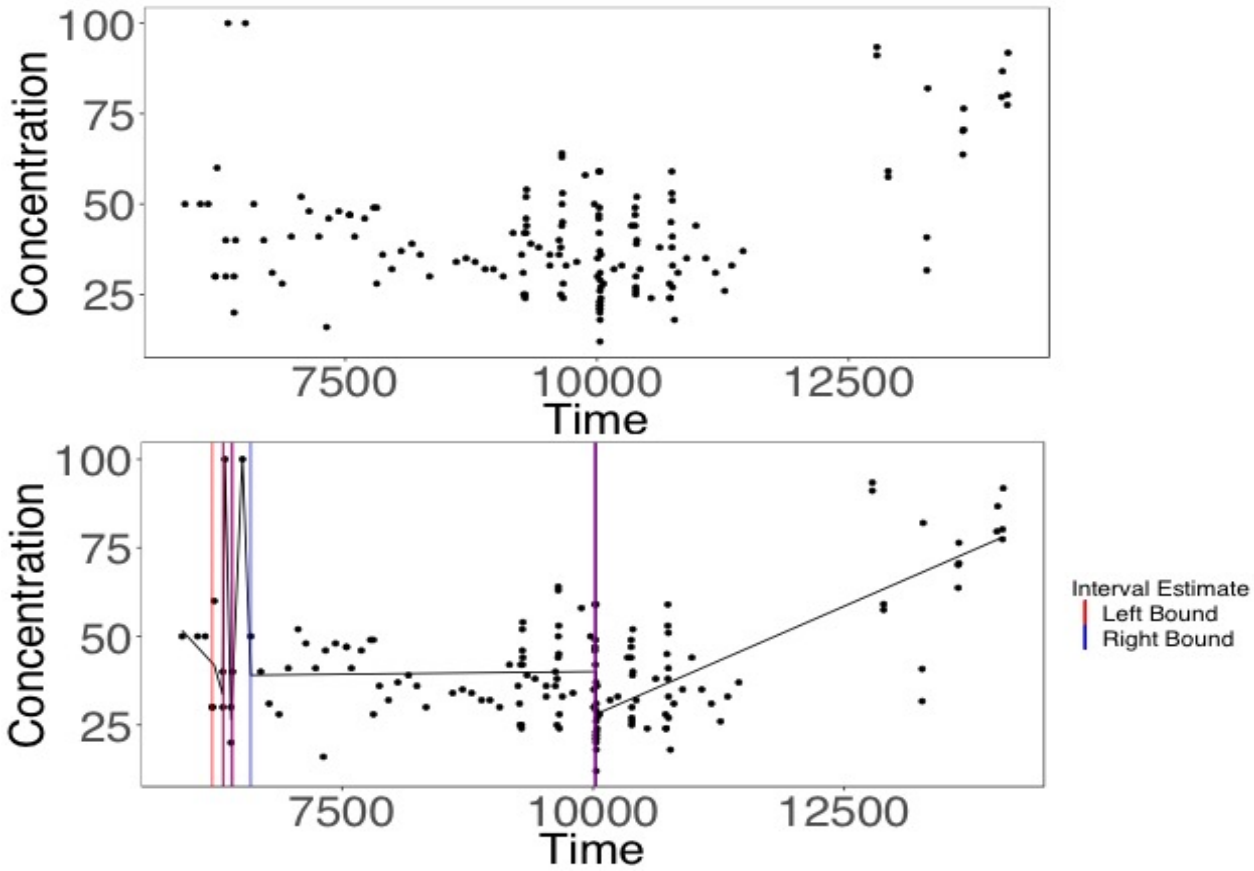

Figure: Cluster E with 152 observations and 4 breakpoints

Cluster F

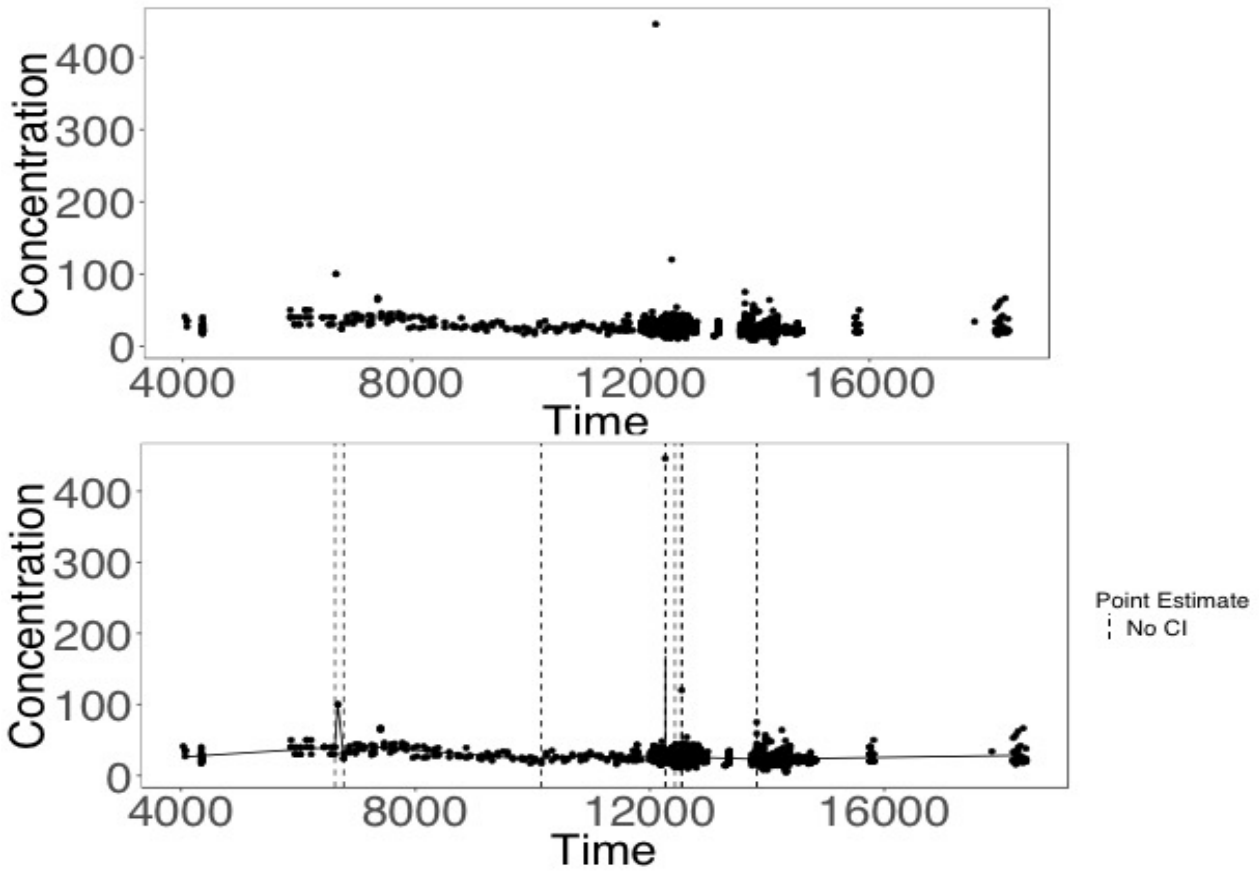

Figure: Cluster F with 1349 observations and 10 breakpoints

Cluster G

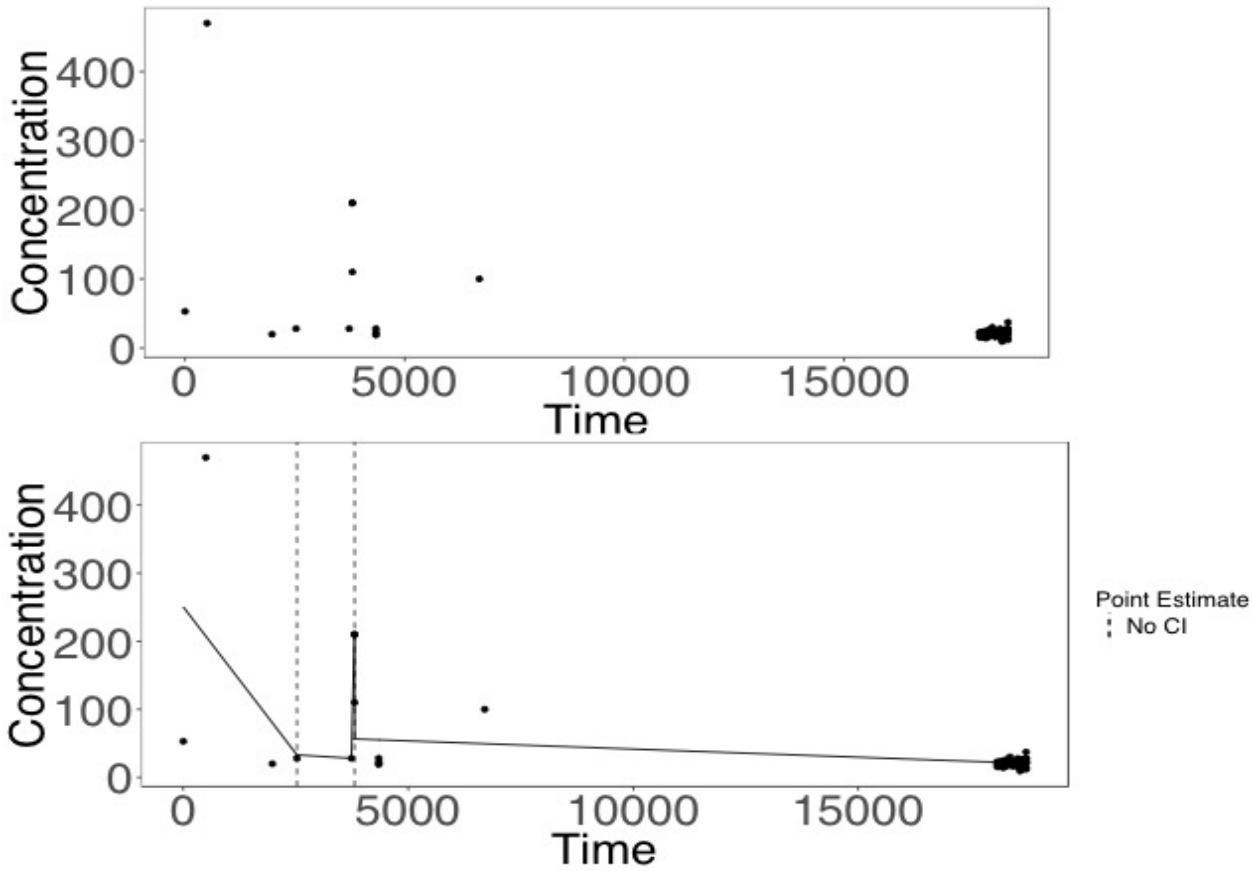

Figure: Cluster G with 79 observations and 2 breakpoints

Cluster H

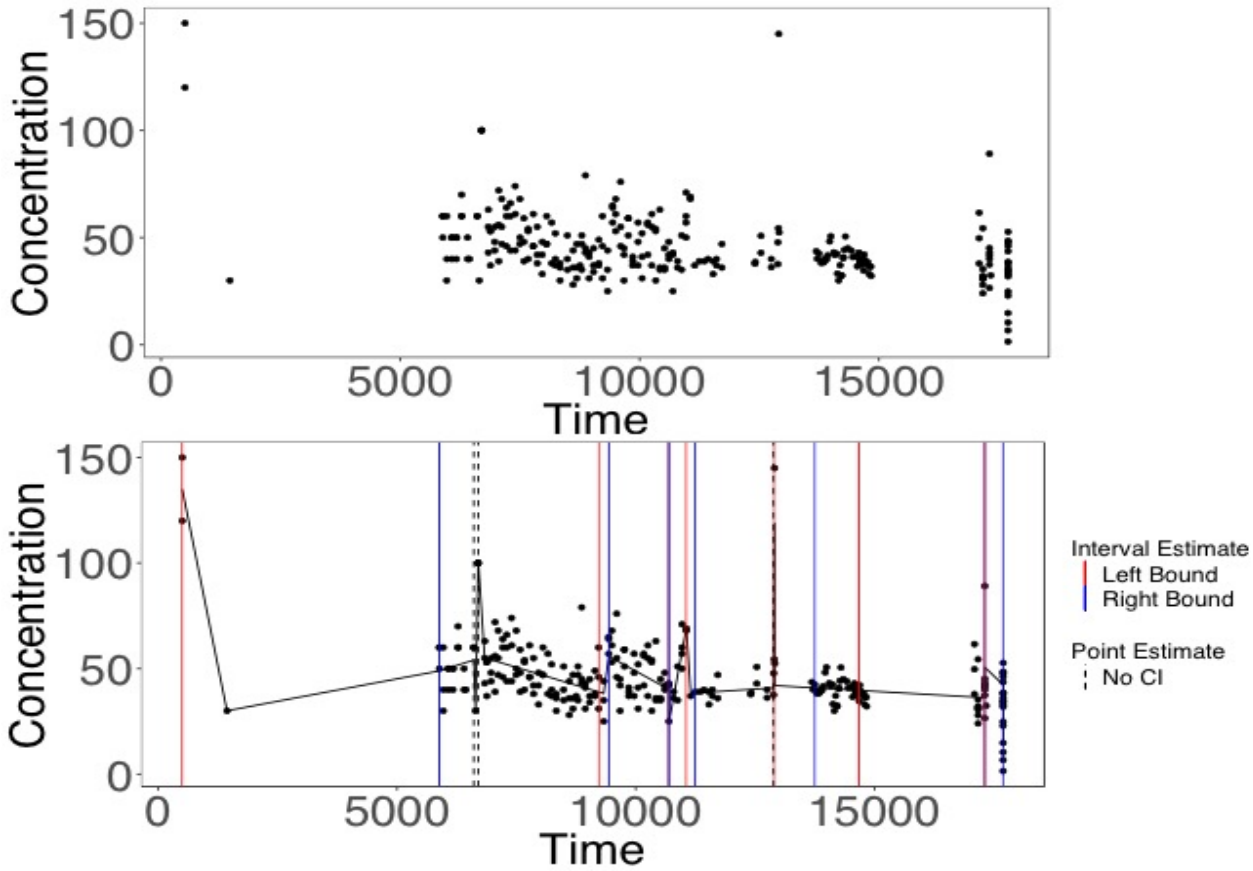

Figure: Cluster H with 269 observations and 10 breakpoints

Outline

Discussion

Summary of current works:

- Spatial K-Means Clustering
- Time series segmentation
- Interactive Shiny app

Future works:

- Robust estimation of breakpoints.
  - $SAR^{r}_{t_1:t_2}$: Sum of absolute residuals for the time segment t_1-t_2 with r breakpoints.

$$SAR_{1:T} = \min_{mh \leq j \leq T-h} \left[SAR^{m-1}_{1:j} + SAR^{0}_{(j+1):T} \right]$$

- Agglomerative clustering and Spatial segmentation based on trend

Discussion

Summary of current works:

- Spatial K-Means Clustering
- Time series segmentation
- Interactive Shiny app

Future works:

- Robust estimation of breakpoints.
  - $SAR^{r}_{t_1:t_2}$: Sum of absolute residuals for the time segment t_1-t_2 with r breakpoints.

$$SAR_{1:T} = \min_{mh \leq j \leq T-h} \left[SAR^{m-1}_{1:j} + SAR^{0}_{(j+1):T} \right]$$

- Agglomerative clustering and Spatial segmentation based on trend