Statistical Analysis of Surface Water Data

Time series segmentation and Outlier visualization

Outline

1. Motivation
2. Time series segmentation
   - Barium Dataset
   - Sulphate Dataset
3. Outlier Visualization for Barium
4. Other project ideas
5. Discussion

Outline

1. Motivation
2. Time series segmentation
    - Barium Dataset
    - Sulphate Dataset
3. Outlier Visualization for Barium
4. Other project ideas
5. Discussion

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

1. Motivation
2. Time series segmentation
   - Barium Dataset
   - Sulphate Dataset
3. Outlier Visualization for Barium
4. Other project ideas
5. Discussion

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

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

- 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):

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

  - 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

1. Motivation
2. Time series segmentation
   - Barium Dataset
   - Sulphate Dataset
3. Outlier Visualization for Barium
4. Other project ideas
5. Discussion

Full time series

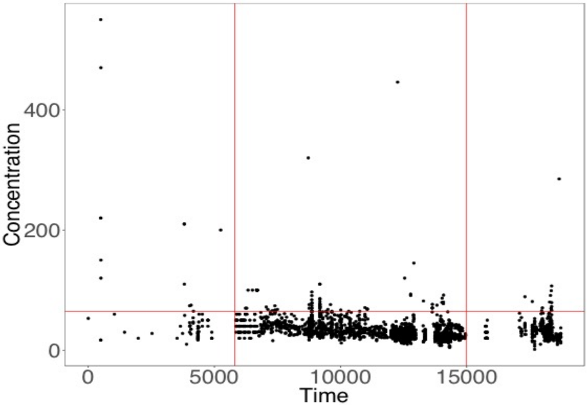

Figure: Full time series for Barium dataset

Dense time series

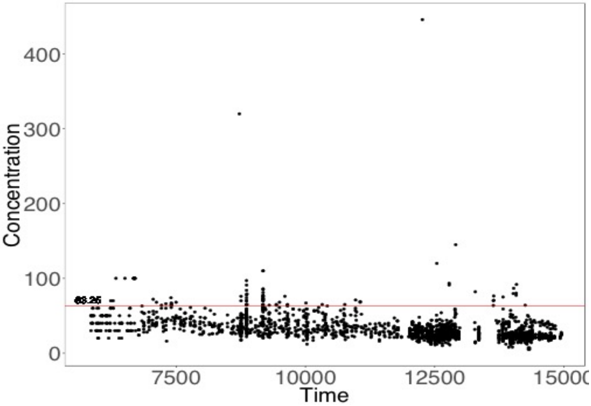

Figure: Dense time series for Barium dataset

BIC and RSS

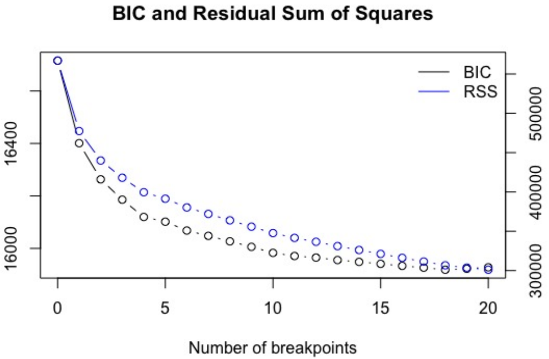

Figure: BIC and RSS plot for dense part of Barium dataset

Time segmentation for dense part

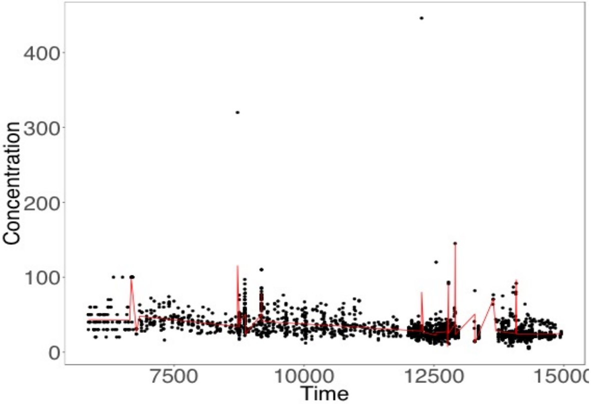

Figure: Time segmentation for dense part of Barium dataset with 18 breakpoints

Outline

1. Motivation
2. Time series segmentation
   - Barium Dataset
   - Sulphate Dataset
3. Outlier Visualization for Barium
4. Other project ideas
5. Discussion

Full time series

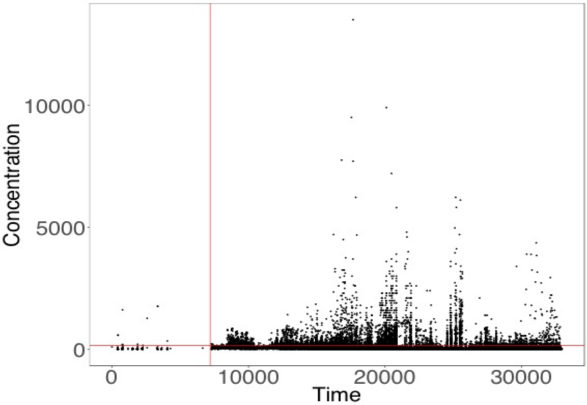

Figure: Full time series for Sulphate dataset

Dense time series

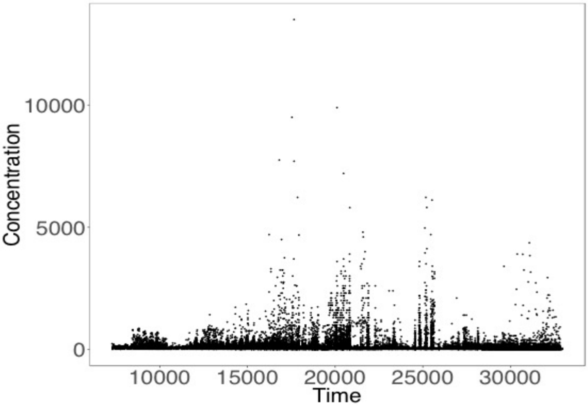

Figure: Dense time series for Sulphate dataset

Outline

1. Motivation
2. Time series segmentation
   - Barium Dataset
   - Sulphate Dataset
3. Outlier Visualization for Barium
4. Other project ideas
5. Discussion

Outlier map for Barium dataset

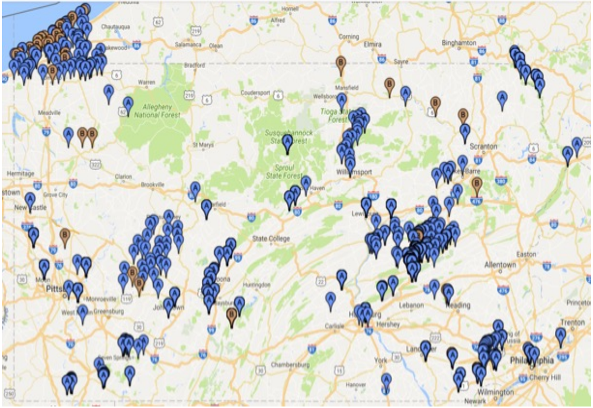

Figure: Outlier map for Barium dataset

Outline

1. Motivation
2. Time series segmentation
   - Barium Dataset
   - Sulphate Dataset
3. Outlier Visualization for Barium
4. Other project ideas
5. Discussion

Other project ideas

- Bedrock zone spatial clustering and time series segmentation.
- Analyzing river system networks:
    - Yearly correlation networks.
    - Identifying the community of outliers.
- High dimensional tests
    - Simultaneous testing for methane concentrations in 1000 wells before and after drilling.
    - False discovery rate to control the type I error.

Outline

1. Motivation
2. Time series segmentation
   - Barium Dataset
   - Sulphate Dataset
3. Outlier Visualization for Barium
4. Other project ideas
5. Discussion

Discussion

Summary of current works:

- Time series segmentation for dense regions
- Interactive Shiny app

Future works:

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

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

- Agglomerative clustering and Spatial segmentation based on trend

Discussion

Summary of current works:

- Time series segmentation for dense regions
- Interactive Shiny app

Future works:

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

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

- Agglomerative clustering and Spatial segmentation based on trend