Statistical Analysis of Surface Water Data

Trend tests and Cross correlations

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**: 1) trend tests; 2) cross correlations

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**: 1) trend tests; 2) cross correlations

Outline

Outline

Trend Tests of mean

- Question of interest: is there any general trend in means?

Means for Barium concentration in five time periods

Period	1963 – 1972	1973 – 1986	1987 – 1996	1997 – 2006	2007 – 2014
Mean	236.39	108.03	140.49	42.50	218.50
No. of observ.	18	671	856	3807	7770

Means for Sulphate concentration in five time periods

Period	1921 – 1940	1941 – 1960	1961 – 1982	1983 – 2001	2002 – 2015
Mean	77.79	110.10	114.98	229.07	143.55
No. of observ.	152	4256	20936	12598	34728

Proposed testing procedure:

- 1st Hypothesis Test H_0: $\mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5$
- If H_0 is NOT rejected, there is no general trend
- If H_0 is rejected, 2nd hypothesis test to determine the trend

Trend Tests of mean

- Question of interest: is there any general trend in means?

Means for Barium concentration in five time periods

Period	1963 – 1972	1973 – 1986	1987 – 1996	1997 – 2006	2007 – 2014
Mean	236.39	108.03	140.49	42.50	218.50
No. of observ.	18	671	856	3807	7770

Means for Sulphate concentration in five time periods

Period	1921 – 1940	1941 – 1960	1961 – 1982	1983 – 2001	2002 – 2015
Mean	77.79	110.10	114.98	229.07	143.55
No. of observ.	152	4256	20936	12598	34728

Proposed testing procedure:

- 1st Hypothesis Test H_0: $\mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5$
- If H_0 is NOT rejected, there is no general trend
- If H_0 is rejected, 2nd hypothesis test to determine the trend

Trend Tests of mean

- Question of interest: is there any general trend in means?

Means for Barium concentration in five time periods

Period	1963 – 1972	1973 – 1986	1987 – 1996	1997 – 2006	2007 – 2014
Mean	236.39	108.03	140.49	42.50	218.50
No. of observ.	18	671	856	3807	7770

Means for Sulphate concentration in five time periods

Period	1921 – 1940	1941 – 1960	1961 – 1982	1983 – 2001	2002 – 2015
Mean	77.79	110.10	114.98	229.07	143.55
No. of observ.	152	4256	20936	12598	34728

Proposed testing procedure:

- 1st Hypothesis Test H_0: $\mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5$
- If H_0 is NOT rejected, there is no general trend
- If H_0 is rejected, 2nd hypothesis test to determine the trend

Multiple Mean test for Barium concentration

Given the following means for Barium concentration

Period	1963 – 1972	1973 – 1986	1987 – 1996	1997 – 2006	2007 – 2014
Mean	236.39	108.03	140.49	42.50	218.50

Now, we consider H_0: $\mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5$

ANOVA F test for equality of all means

	Df	Sum Sq	Mean Sq	F value	p value
Period	4	8.12×10^7	2.03×10^7	0.16	0.9601
Residuals	13117	1.70×10^{12}	1.30×10^8		

- Our conclusion: *no general trend*

Multiple Mean test for Sulphate concentration

Given the following means for Sulphate concentration

Period	1921 – 1940	1941 – 1960	1961 – 1982	1983 – 2001	2002 – 2015
Mean	77.79	110.10	114.98	229.07	143.55

Now, we consider H_0: $\mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5$

ANOVA F test for equality of all means

	Df	Sum Sq	Mean Sq	F value	p value
Period	4	1.13×10^8	2.83×10^7	8.82	0.000***
Residuals	72665	2.33×10^{11}	3.21×10^6		

- Our conclusion: there exists some trend

Trend test of mean for Sulphate concentration

Given the following means for Sulphate concentration

Period	1921 – 1940	1941 – 1960	1961 – 1982	1983 – 2001	2002 – 2015
Mean	77.79	110.10	114.98	229.07	143.55

Next, we consider a sequence of two-sample t tests:

Welch's two sample t tests for equality of all means

H_0	H_a	p-value	Bonferroni correction
$\mu_1 = \mu_2$	$\mu_1 < \mu_2$	0.068	Don't reject H_0
$\mu_2 = \mu_3$	$\mu_2 < \mu_3$	0.053	Don't reject H_0
$\mu_3 = \mu_4$	$\mu_3 < \mu_4$	0.000^{***}	Reject H_0
$\mu_4 = \mu_5$	$\mu_4 > \mu_5$	0.000^{***}	Reject H_0

- Our conclusion: under Bonferroni correction $\mu_1 = \mu_2 = \mu_3$, but there is still a significant trend that $\mu_3 < \mu_4$ and $\mu_4 > \mu_5$

Trend test of mean for Sulphate concentration

Given the following means for Sulphate concentration

Period	1921 – 1940	1941 – 1960	1961 – 1982	1983 – 2001	2002 – 2015
Mean	77.79	110.10	114.98	229.07	143.55

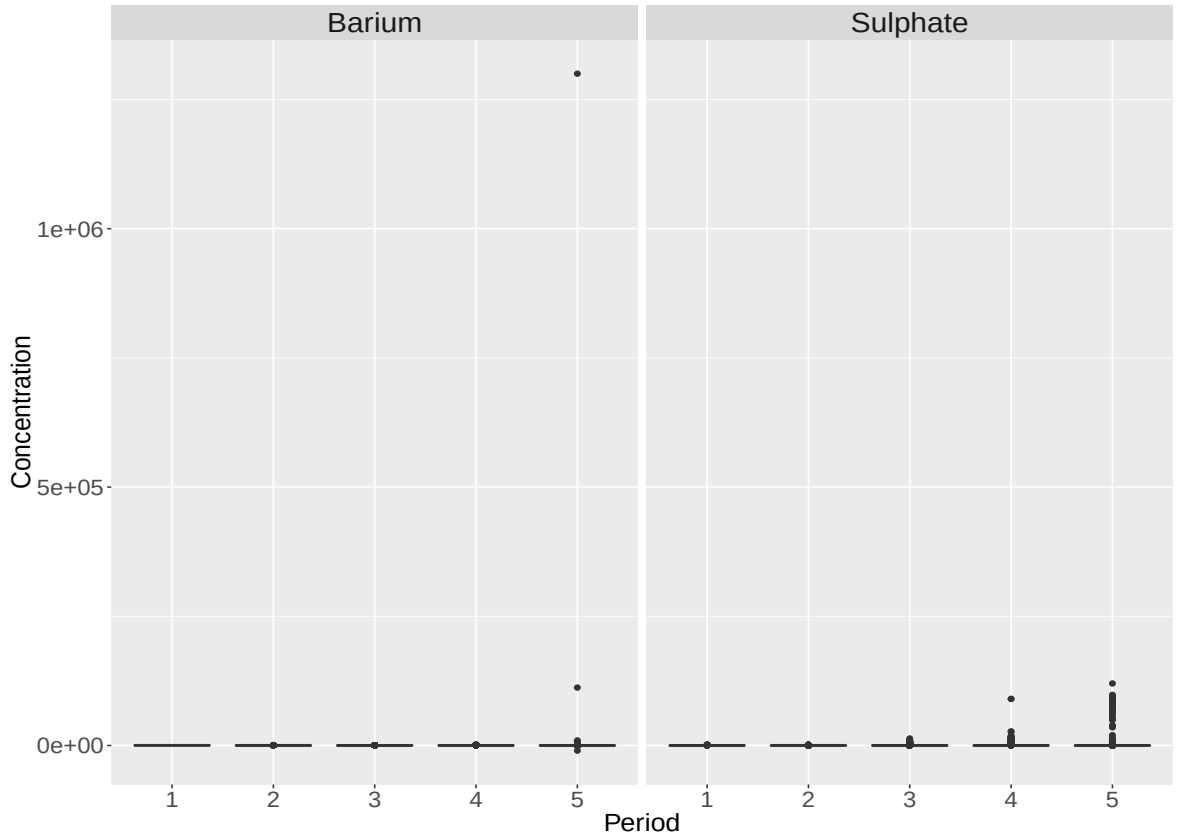
Next, we consider a sequence of two-sample t tests:

Welch's two sample t tests for equality of all means

H_0	H_a	p-value	Bonferroni correction
$\mu_1 = \mu_2$	$\mu_1 < \mu_2$	0.068	Don't reject H_0
$\mu_2 = \mu_3$	$\mu_2 < \mu_3$	0.053	Don't reject H_0
$\mu_3 = \mu_4$	$\mu_3 < \mu_4$	0.000^{***}	Reject H_0
$\mu_4 = \mu_5$	$\mu_4 > \mu_5$	0.000^{***}	Reject H_0

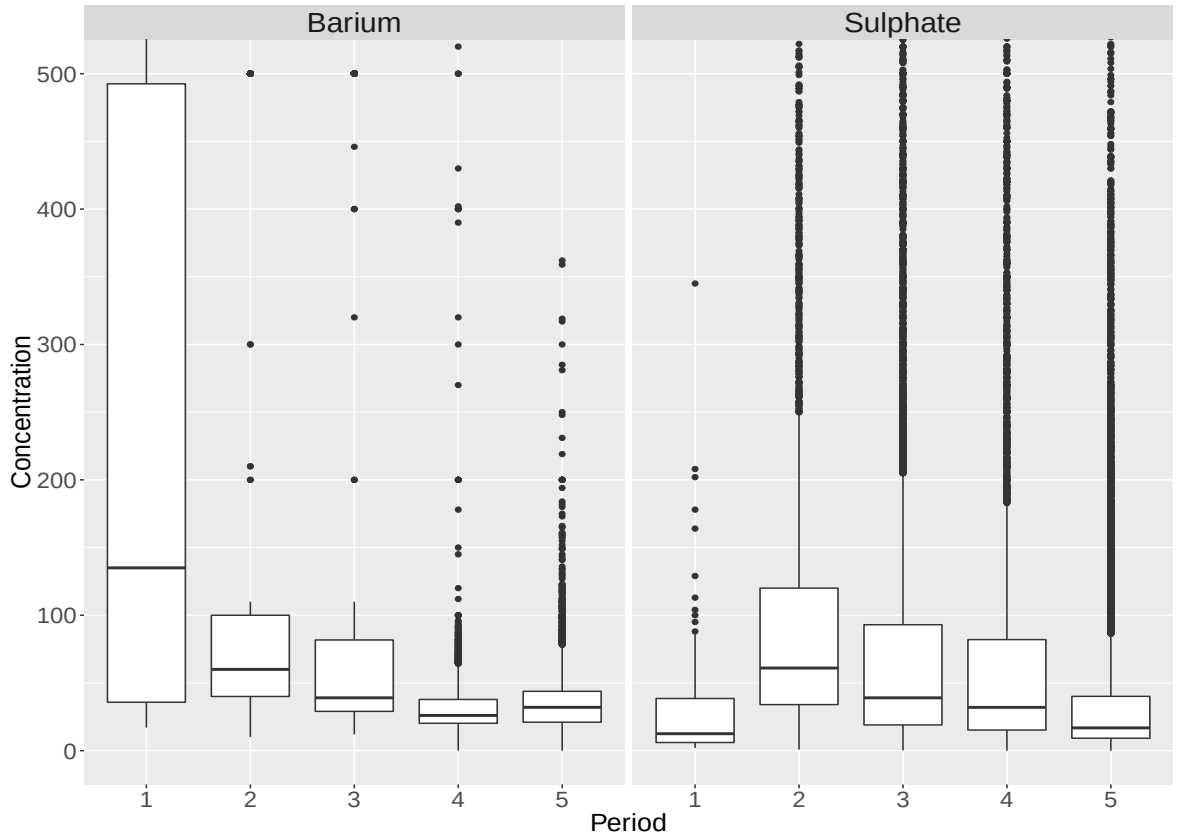
- Our conclusion: under Bonferroni correction $\mu_1 = \mu_2 = \mu_3$, but there is still a significant trend that $\mu_3 < \mu_4$ and $\mu_4 > \mu_5$

Visualization of the skewness



Boxplots for concentration against different periods for Ba and Su

Visualization of the median trend in a limited range



Boxplots for concentration against different periods for Ba and Su

Trend Tests of median

- Question of interest: is there any general trend in median?

Medians for Barium concentration in five time periods

Period	1963 – 1972	1973 – 1986	1987 – 1996	1997 – 2006	2007 – 2014
Median	135	60	39	26	32

Medians for Sulphate concentration in five time periods

Period	1921 – 1940	1941 – 1960	1961 – 1982	1983 – 2001	2002 – 2015
Median	12.5	61	39	31.95	16.8

Proposed testing procedure:

- 1st Hypothesis Test H_0: $m_1 = m_2 = m_3 = m_4 = m_5$
- If H_0 is NOT rejected, there is no general trend
- If H_0 is rejected, 2nd hypothesis test to determine the trend

Trend Tests of median

- Question of interest: is there any general trend in median?

Medians for Barium concentration in five time periods

Period	1963 – 1972	1973 – 1986	1987 – 1996	1997 – 2006	2007 – 2014
Median	135	60	39	26	32

Medians for Sulphate concentration in five time periods

Period	1921 – 1940	1941 – 1960	1961 – 1982	1983 – 2001	2002 – 2015
Median	12.5	61	39	31.95	16.8

Proposed testing procedure:

- 1st Hypothesis Test H_0: $m_1 = m_2 = m_3 = m_4 = m_5$
- If H_0 is NOT rejected, there is no general trend
- If H_0 is rejected, 2nd hypothesis test to determine the trend

Trend Tests of median

- Question of interest: is there any general trend in median?

Medians for Barium concentration in five time periods

Period	1963 – 1972	1973 – 1986	1987 – 1996	1997 – 2006	2007 – 2014
Median	135	60	39	26	32

Medians for Sulphate concentration in five time periods

Period	1921 – 1940	1941 – 1960	1961 – 1982	1983 – 2001	2002 – 2015
Median	12.5	61	39	31.95	16.8

Proposed testing procedure:

- 1st Hypothesis Test H_0: $m_1 = m_2 = m_3 = m_4 = m_5$
- If H_0 is NOT rejected, there is no general trend
- If H_0 is rejected, 2nd hypothesis test to determine the trend

Multiple Median tests for Barium concentration

Given the following medians for Barium concentration

Period	1963 – 1972	1973 – 1986	1987 – 1996	1997 – 2006	2007 – 2014
Median	135	60	39	26	32

Now, we consider H_0: $m_1 = m_2 = m_3 = m_4 = m_5$

Non-parametric tests for equality of all medians

	Test statistic	Df	p value
Mood	737.51	4	0.000***
Kruskal-Wallis	1204.50	4	0.000***

- Our conclusion: *there exists some trend*

Trend test of median for Barium concentration

Given the following medians for Barium concentration

Period	1963 – 1972	1973 – 1986	1987 – 1996	1997 – 2006	2007 – 2014
Median	135	60	39	26	32

Next, we consider a sequence of two-sample median tests:

Two sample Wilcoxon tests

H_0	H_a	p-value	Bonferroni correction
$m_1 = m_2$	$m_1 > m_2$	0.035	Don't reject H_0
$m_2 = m_3$	$m_2 > m_3$	0.000^{***}	Reject H_0
$m_3 = m_4$	$m_3 > m_4$	0.000^{***}	Reject H_0
$m_4 = m_5$	$m_4 < m_5$	0.000^{***}	Reject H_0

- Our conclusion: under Bonferroni correction, $m_1 = m_2$ but there is still a significant trend that $m_2 > m_3 > m_4$ and $m_4 < m_5$

Trend test of median for Barium concentration

Given the following medians for Barium concentration

Period	1963 – 1972	1973 – 1986	1987 – 1996	1997 – 2006	2007 – 2014
Median	135	60	39	26	32

Next, we consider a sequence of two-sample median tests:

Two sample Wilcoxon tests

H_0	H_a	p-value	Bonferroni correction
$m_1 = m_2$	$m_1 > m_2$	0.035	Don't reject H_0
$m_2 = m_3$	$m_2 > m_3$	0.000^{***}	Reject H_0
$m_3 = m_4$	$m_3 > m_4$	0.000^{***}	Reject H_0
$m_4 = m_5$	$m_4 < m_5$	0.000^{***}	Reject H_0

- Our conclusion: under Bonferroni correction, $m_1 = m_2$ but there is still a significant trend that $m_2 > m_3 > m_4$ and $m_4 < m_5$

Multiple Median tests for Sulphate concentration

Given the following medians for Sulphate concentration

Period	1921 – 1940	1941 – 1960	1961 – 1982	1983 – 2001	2002 – 2015
Median	12.5	61	39	31.95	16.8

Now, we consider H_0: $m_1 = m_2 = m_3 = m_4 = m_5$

Non-parametric tests for equality of all medians

	Test statistic	Df	p value
Mood	6697.60	4	0.000***
Kruskal-Wallis	8920.90	4	0.000***

- Our conclusion: *there exists some trend*

Trend test of median for Sulphate concentration

Given the following medians for Sulphate concentration

Period	1921 – 1940	1941 – 1960	1961 – 1982	1983 – 2001	2002 – 2015
Median	12.5	61	39	31.95	16.8

Next, we consider a sequence of two-sample median tests:

Two sample Wilcoxon tests

H_0	H_a	p-value	Bonferroni correction
$m_1 = m_2$	$m_1 < m_2$	0.000^{***}	Reject H_0
$m_2 = m_3$	$m_2 > m_3$	0.000^{***}	Reject H_0
$m_3 = m_4$	$m_3 > m_4$	0.000^{***}	Reject H_0
$m_4 = m_5$	$m_4 > m_5$	0.000^{***}	Reject H_0

- Our conclusion: under Bonferroni correction, there is still a significant trend that $m_1 < m_2$ but $m_2 > m_3 > m_4 > m_5$

Trend test of median for Sulphate concentration

Given the following medians for Sulphate concentration

Period	1921 – 1940	1941 – 1960	1961 – 1982	1983 – 2001	2002 – 2015
Median	12.5	61	39	31.95	16.8

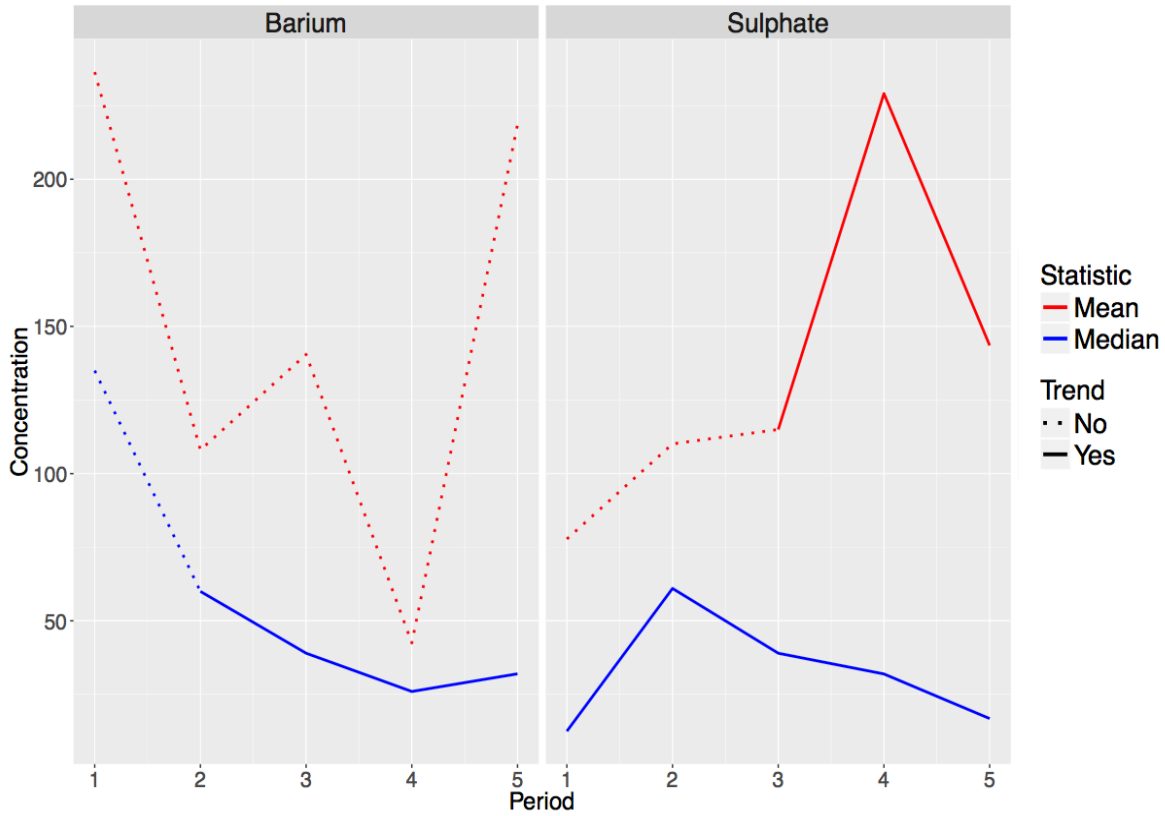
Next, we consider a sequence of two-sample median tests:

Two sample Wilcoxon tests

H_0	H_a	p-value	Bonferroni correction
$m_1 = m_2$	$m_1 < m_2$	0.000^{***}	Reject H_0
$m_2 = m_3$	$m_2 > m_3$	0.000^{***}	Reject H_0
$m_3 = m_4$	$m_3 > m_4$	0.000^{***}	Reject H_0
$m_4 = m_5$	$m_4 > m_5$	0.000^{***}	Reject H_0

- Our conclusion: under Bonferroni correction, there is still a significant trend that $m_1 < m_2$ but $m_2 > m_3 > m_4 > m_5$

Summary of trend tests for full dataset



Plot of Mean and Median concentration against different periods showing trend significance for Ba and Su

Outline

Trend Tests of mean

- Question of interest: is there any general trend in means?

Means for Barium concentration in five time periods

Period	1963 – 1972	1973 – 1986	1987 – 1996	1997 – 2006	2007 – 2014
Mean	150.38	43.80	38.34	26.09	32.33
No. of observ.	13	321	651	1080	1178

Means for Sulphate concentration in five time periods

Period	1921 – 1940	1941 – 1960	1961 – 1982	1983 – 2001	2002 – 2015
Mean	79.31	110.10	115.34	165.01	42.76
No. of observ.	149	4256	20807	7823	16332

Proposed testing procedure:

- 1st Hypothesis Test H_0: $\mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5$
- If H_0 is NOT rejected, there is no general trend
- If H_0 is rejected, 2nd hypothesis test to determine the trend

Trend Tests of mean

- Question of interest: is there any general trend in means?

Means for Barium concentration in five time periods

Period	1963 – 1972	1973 – 1986	1987 – 1996	1997 – 2006	2007 – 2014
Mean	150.38	43.80	38.34	26.09	32.33
No. of observ.	13	321	651	1080	1178

Means for Sulphate concentration in five time periods

Period	1921 – 1940	1941 – 1960	1961 – 1982	1983 – 2001	2002 – 2015
Mean	79.31	110.10	115.34	165.01	42.76
No. of observ.	149	4256	20807	7823	16332

Proposed testing procedure:

- 1st Hypothesis Test H_0: $\mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5$
- If H_0 is NOT rejected, there is no general trend
- If H_0 is rejected, 2nd hypothesis test to determine the trend

Trend Tests of mean

- Question of interest: is there any general trend in means?

Means for Barium concentration in five time periods

Period	1963 – 1972	1973 – 1986	1987 – 1996	1997 – 2006	2007 – 2014
Mean	150.38	43.80	38.34	26.09	32.33
No. of observ.	13	321	651	1080	1178

Means for Sulphate concentration in five time periods

Period	1921 – 1940	1941 – 1960	1961 – 1982	1983 – 2001	2002 – 2015
Mean	79.31	110.10	115.34	165.01	42.76
No. of observ.	149	4256	20807	7823	16332

Proposed testing procedure:

- 1st Hypothesis Test H_0: $\mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5$
- If H_0 is NOT rejected, there is no general trend
- If H_0 is rejected, 2nd hypothesis test to determine the trend

Multiple Mean test for Barium concentration

Given the following means for Barium concentration from filtered dataset

Period	1963 – 1972	1973 – 1986	1987 – 1996	1997 – 2006	2007 – 2014
Mean	150.38	43.80	38.34	26.09	32.33

Now, we consider H_0: $\mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5$

ANOVA F test for equality of all means

	Df	Sum Sq	Mean Sq	F value	p-value
Period	4	2.87×10^5	7.18×10^5	150.49	0.000***
Residuals	3238	1.54×10^6	477.22		

- Our conclusion: *there exists some trend*

Trend test of mean for Barium concentration

Given the following means for Barium concentration from filtered dataset

Period	1963 – 1972	1973 – 1986	1987 – 1996	1997 – 2006	2007 – 2014
Mean	150.38	43.80	38.34	26.09	32.33

Next, we consider a sequence of two-sample t tests:

Welch's two sample t tests for equality of all means

H_0	H_a	p-value	Bonferroni correction
$\mu_1 = \mu_2$	$\mu_1 > \mu_2$	0.025	Don't reject H_0
$\mu_2 = \mu_3$	$\mu_2 > \mu_3$	0.000^{***}	Reject H_0
$\mu_3 = \mu_4$	$\mu_3 > \mu_4$	0.000^{***}	Reject H_0
$\mu_4 = \mu_5$	$\mu_4 < \mu_5$	0.000^{***}	Reject H_0

- Our conclusion: under Bonferroni correction, $\mu_1 = \mu_2$ but there is still a significant trend that $\mu_2 > \mu_3 > \mu_4$ and $\mu_4 < \mu_5$

Trend test of mean for Barium concentration

Given the following means for Barium concentration from filtered dataset

Period	1963 – 1972	1973 – 1986	1987 – 1996	1997 – 2006	2007 – 2014
Mean	150.38	43.80	38.34	26.09	32.33

Next, we consider a sequence of two-sample t tests:

Welch's two sample t tests for equality of all means

H_0	H_a	p-value	Bonferroni correction
$\mu_1 = \mu_2$	$\mu_1 > \mu_2$	0.025	Don't reject H_0
$\mu_2 = \mu_3$	$\mu_2 > \mu_3$	0.000^{***}	Reject H_0
$\mu_3 = \mu_4$	$\mu_3 > \mu_4$	0.000^{***}	Reject H_0
$\mu_4 = \mu_5$	$\mu_4 < \mu_5$	0.000^{***}	Reject H_0

- Our conclusion: under Bonferroni correction, $\mu_1 = \mu_2$ but there is still a significant trend that $\mu_2 > \mu_3 > \mu_4$ and $\mu_4 < \mu_5$

Multiple Mean test for Sulphate concentration

Given the following means for Sulphate concentration from filtered dataset

Period	1921 – 1940	1941 – 1960	1961 – 1982	1983 – 2001	2002 – 2015
Mean	79.31	110.10	115.34	165.01	42.76

Now, we consider H_0: $\mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5$

ANOVA F test for equality of all means

	Df	Sum Sq	Mean Sq	F value	p-value
Period	4	9.19×10^7	2.30×10^7	53.16	0.000***
Residuals	49362	2.13×10^{10}	4.32×10^5		

- Our conclusion: *there exists some trend*

Trend test of mean for Sulphate concentration

Given the following means for Sulphate concentration from filtered dataset

Period	1921 – 1940	1941 – 1960	1961 – 1982	1983 – 2001	2002 – 2015
Mean	79.31	110.10	115.34	165.01	42.76

Next, we consider a sequence of two-sample t tests:

Welch's two sample t tests for equality of all means

H_0	H_a	p-value	Bonferroni correction
$\mu_1 = \mu_2$	$\mu_1 < \mu_2$	0.082	Don't reject H_0
$\mu_2 = \mu_3$	$\mu_2 < \mu_3$	0.042	Don't reject H_0
$\mu_3 = \mu_4$	$\mu_3 < \mu_4$	0.003^{**}	Reject H_0
$\mu_4 = \mu_5$	$\mu_4 > \mu_5$	0.000^{***}	Reject H_0

- Our conclusion: under Bonferroni correction, $\mu_1 = \mu_2 = \mu_3$ but there is still a significant trend that $\mu_3 < \mu_4$ and $\mu_4 > \mu_5$

Trend test of mean for Sulphate concentration

Given the following means for Sulphate concentration from filtered dataset

Period	1921 – 1940	1941 – 1960	1961 – 1982	1983 – 2001	2002 – 2015
Mean	79.31	110.10	115.34	165.01	42.76

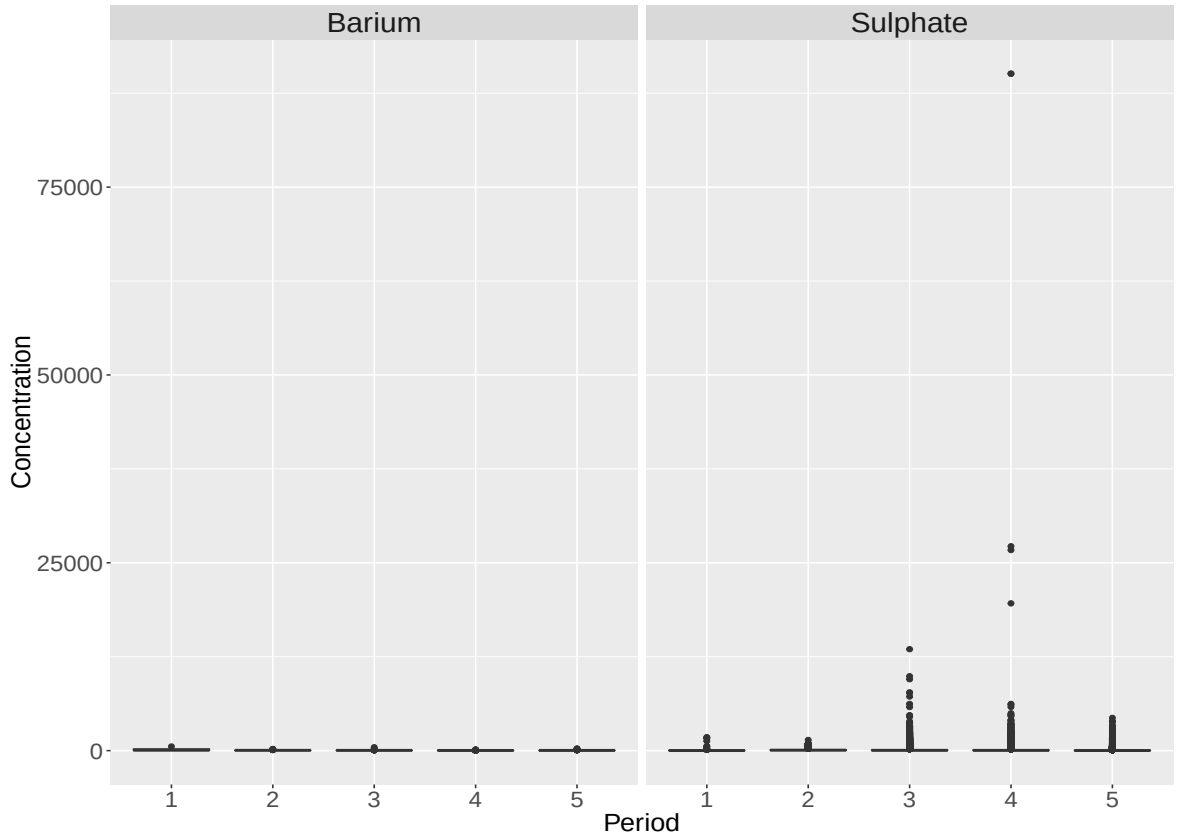
Next, we consider a sequence of two-sample t tests:

Welch's two sample t tests for equality of all means

H_0	H_a	p-value	Bonferroni correction
$\mu_1 = \mu_2$	$\mu_1 < \mu_2$	0.082	Don't reject H_0
$\mu_2 = \mu_3$	$\mu_2 < \mu_3$	0.042	Don't reject H_0
$\mu_3 = \mu_4$	$\mu_3 < \mu_4$	0.003^{**}	Reject H_0
$\mu_4 = \mu_5$	$\mu_4 > \mu_5$	0.000^{***}	Reject H_0

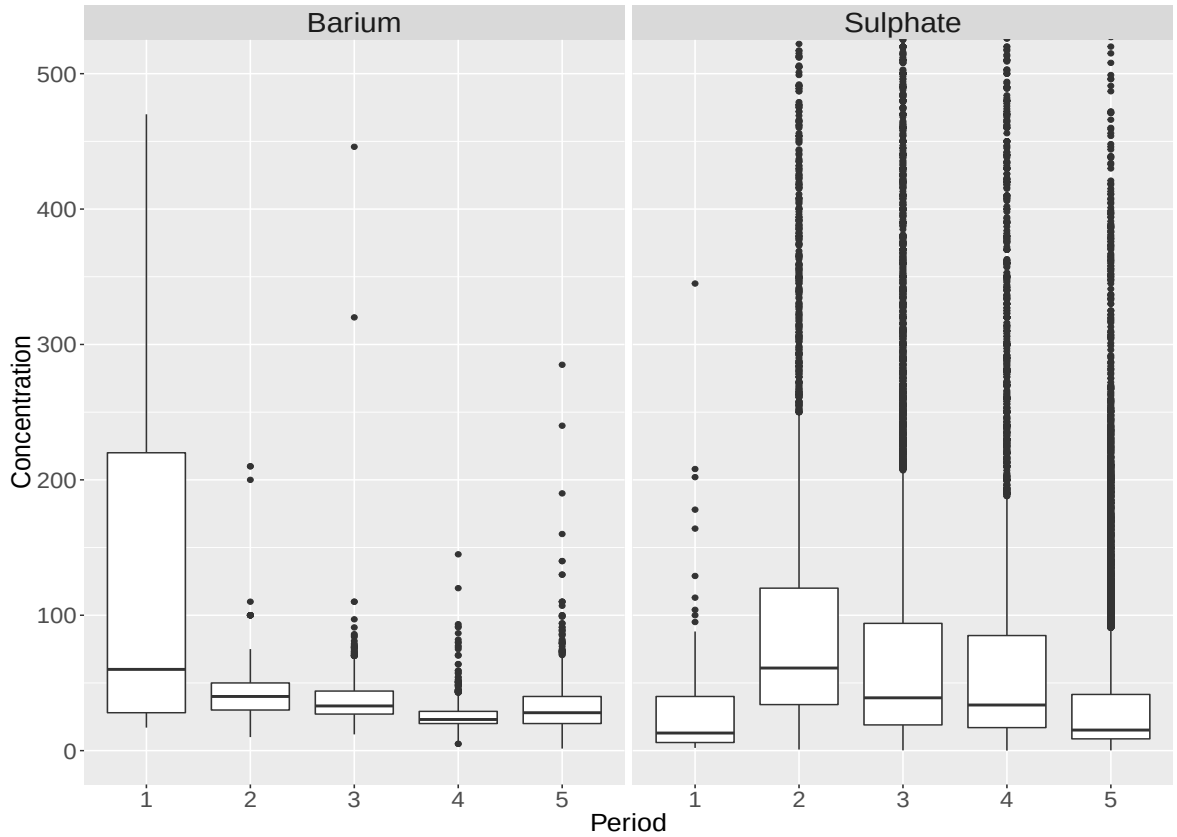
- Our conclusion: under Bonferroni correction, $\mu_1 = \mu_2 = \mu_3$ but there is still a significant trend that $\mu_3 < \mu_4$ and $\mu_4 > \mu_5$

Visualization of the skewness



Boxplots for concentration against different periods for Ba and Su

Visualization of the median trend in a limited range



Boxplots for concentration against different periods for Ba and Su

Trend Tests of median

- Question of interest: is there any general trend in median?

Medians for Barium concentration in five time periods

Period	1963 – 1972	1973 – 1986	1987 – 1996	1997 – 2006	2007 – 2014
Median	60	40	33	23	28

Medians for Sulphate concentration in five time periods

Period	1921 – 1940	1941 – 1960	1961 – 1982	1983 – 2001	2002 – 2015
Median	13	61	39	33.7	15.2

Proposed testing procedure:

- 1st Hypothesis Test H_0: $m_1 = m_2 = m_3 = m_4 = m_5$
- If H_0 is NOT rejected, there is no general trend
- If H_0 is rejected, 2nd hypothesis test to determine the trend

Trend Tests of median

- Question of interest: is there any general trend in median?

Medians for Barium concentration in five time periods

Period	1963 – 1972	1973 – 1986	1987 – 1996	1997 – 2006	2007 – 2014
Median	60	40	33	23	28

Medians for Sulphate concentration in five time periods

Period	1921 – 1940	1941 – 1960	1961 – 1982	1983 – 2001	2002 – 2015
Median	13	61	39	33.7	15.2

Proposed testing procedure:

- 1st Hypothesis Test H_0: $m_1 = m_2 = m_3 = m_4 = m_5$
- If H_0 is NOT rejected, there is no general trend
- If H_0 is rejected, 2nd hypothesis test to determine the trend

Trend Tests of median

- Question of interest: is there any general trend in median?

Medians for Barium concentration in five time periods

Period	1963 – 1972	1973 – 1986	1987 – 1996	1997 – 2006	2007 – 2014
Median	60	40	33	23	28

Medians for Sulphate concentration in five time periods

Period	1921 – 1940	1941 – 1960	1961 – 1982	1983 – 2001	2002 – 2015
Median	13	61	39	33.7	15.2

Proposed testing procedure:

- 1st Hypothesis Test H_0: $m_1 = m_2 = m_3 = m_4 = m_5$
- If H_0 is NOT rejected, there is no general trend
- If H_0 is rejected, 2nd hypothesis test to determine the trend

Multiple Median tests for Barium concentration

Given the following medians for Barium concentration from filtered dataset

Period	1963 – 1972	1973 – 1986	1987 – 1996	1997 – 2006	2007 – 2014
Median	60	40	33	23	28

Now, we consider H_0: $m_1 = m_2 = m_3 = m_4 = m_5$

Non-parametric tests for equality of all medians

	Test statistic	Df	p value
Mood	486.86	4	0.000***
Kruskal-Wallis	531.03	4	0.000***

- Our conclusion: *there exists some trend*

Trend test of median for Barium concentration

Given the following medians for Barium concentration from filtered dataset

Period	1963 – 1972	1973 – 1986	1987 – 1996	1997 – 2006	2007 – 2014
Median	60	40	33	23	28

Next, we consider a sequence of two-sample median tests:

Two sample Wilcoxon tests

H_0	H_a	p-value	Bonferroni correction
$m_1 = m_2$	$m_1 > m_2$	0.071	Don't reject H_0
$m_2 = m_3$	$m_2 > m_3$	0.000^{***}	Reject H_0
$m_3 = m_4$	$m_3 > m_4$	0.000^{***}	Reject H_0
$m_4 = m_5$	$m_4 < m_5$	0.000^{***}	Reject H_0

- Our conclusion: under Bonferroni correction, $m_1 = m_2$ but there is still a significant trend that $m_2 > m_3 > m_4$ but $m_4 < m_5$

Trend test of median for Barium concentration

Given the following medians for Barium concentration from filtered dataset

Period	1963 – 1972	1973 – 1986	1987 – 1996	1997 – 2006	2007 – 2014
Median	60	40	33	23	28

Next, we consider a sequence of two-sample median tests:

Two sample Wilcoxon tests

H_0	H_a	p-value	Bonferroni correction
$m_1 = m_2$	$m_1 > m_2$	0.071	Don't reject H_0
$m_2 = m_3$	$m_2 > m_3$	0.000^{***}	Reject H_0
$m_3 = m_4$	$m_3 > m_4$	0.000^{***}	Reject H_0
$m_4 = m_5$	$m_4 < m_5$	0.000^{***}	Reject H_0

- Our conclusion: under Bonferroni correction, $m_1 = m_2$ but there is still a significant trend that $m_2 > m_3 > m_4$ but $m_4 < m_5$

Multiple Median tests for Sulphate concentration

Given the following medians for Sulphate concentration from filtered dataset

Period	1921 – 1940	1941 – 1960	1961 – 1982	1983 – 2001	2002 – 2015
Median	13	61	39	33.7	15.2

Now, we consider H_0: $m_1 = m_2 = m_3 = m_4 = m_5$

Non-parametric tests for equality of all medians

	Test statistic	Df	p value
Mood	4339.60	4	0.000***
Kruskal-Wallis	6546	4	0.000***

- Our conclusion: *there exists some trend*

Trend test of median for Sulphate concentration

Given the following medians for Sulphate concentration from filtered dataset

Period	1921 – 1940	1941 – 1960	1961 – 1982	1983 – 2001	2002 – 2015
Median	13	61	39	33.7	15.2

Next, we consider a sequence of two-sample median tests:

Two sample Wilcoxon tests

H_0	H_a	p-value	Bonferroni correction
$m_1 = m_2$	$m_1 < m_2$	0.000^{***}	Reject H_0
$m_2 = m_3$	$m_2 > m_3$	0.000^{***}	Reject H_0
$m_3 = m_4$	$m_3 > m_4$	0.000^{***}	Reject H_0
$m_4 = m_5$	$m_4 > m_5$	0.000^{***}	Reject H_0

- Our conclusion: under Bonferroni correction, there is still a significant trend that $m_1 < m_2$ but $m_2 > m_3 > m_4 > m_5$

Trend test of median for Sulphate concentration

Given the following medians for Sulphate concentration from filtered dataset

Period	1921 – 1940	1941 – 1960	1961 – 1982	1983 – 2001	2002 – 2015
Median	13	61	39	33.7	15.2

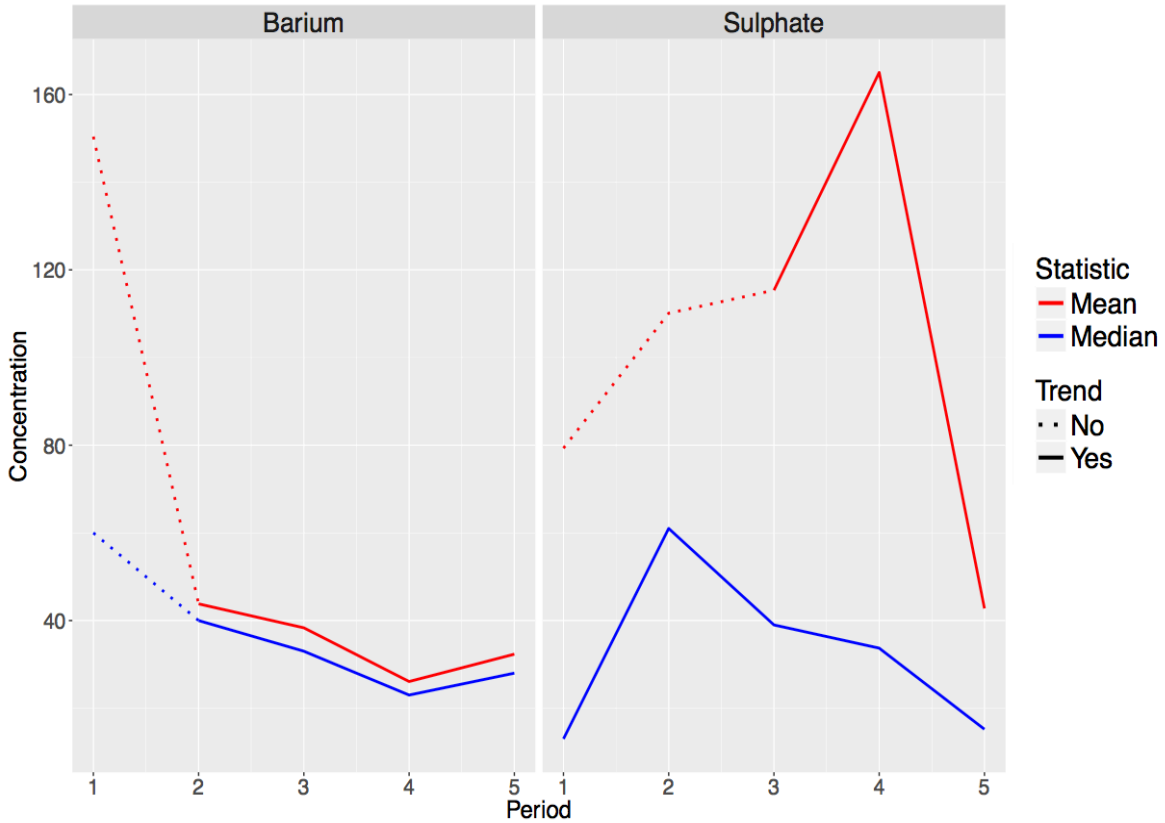
Next, we consider a sequence of two-sample median tests:

Two sample Wilcoxon tests

H_0	H_a	p-value	Bonferroni correction
$m_1 = m_2$	$m_1 < m_2$	0.000^{***}	Reject H_0
$m_2 = m_3$	$m_2 > m_3$	0.000^{***}	Reject H_0
$m_3 = m_4$	$m_3 > m_4$	0.000^{***}	Reject H_0
$m_4 = m_5$	$m_4 > m_5$	0.000^{***}	Reject H_0

- Our conclusion: under Bonferroni correction, there is still a significant trend that $m_1 < m_2$ but $m_2 > m_3 > m_4 > m_5$

Summary of trend tests for filtered dataset



Plot of Mean and Median concentration against different periods showing trend significance for Ba and Su

Outline

Motivation

- Barium and Sulphate conc. for 11 PA counties from '02-'09

County	Armstrong	Blair	Centre	Chester	Crawford	Cumberland
Barium	75	65	79	95	88	95
Sulphate	85	81	88	96	91	96

County	Lycoming	Pike	Somerset	Sullivan	Warren
Barium	93	93	87	89	80
Sulphate	95	94	91	90	92

Note: shown are the numbers of monthly observations from 2002-2009

- Observation: *irregularly* sampled time series!

Motivation

- Barium and Sulphate conc. for 11 PA counties from '02-'09

County	Armstrong	Blair	Centre	Chester	Crawford	Cumberland
Barium	75	65	79	95	88	95
Sulphate	85	81	88	96	91	96

County	Lycoming	Pike	Somerset	Sullivan	Warren
Barium	93	93	87	89	80
Sulphate	95	94	91	90	92

Note: shown are the numbers of monthly observations from 2002-2009

- Observation: *irregularly* sampled time series!

Motivation
Cont'd

Time Series Scatter Plots for Barium and Sulphate conc.

Cross correlation function (CCF)

CCF studies linear association between x_t and y_t at lag k

- CCF for two regularly sampled time series x_t and y_t

$$\hat{\rho}(k) = \frac{1}{\hat{\sigma}_x \hat{\sigma}_y (N-k)} \sum_{t=1}^{N-k} (x_t - \overline{x})(y_{t+k} - \overline{y})$$

- CCF for two irregularly sampled time series x_t and y_t

$$\hat{\rho}(k) = \frac{\sum_{i=1}^{N_x} \sum_{j=1}^{N_y} (x_i - \overline{x})(y_j - \overline{y}) K_h(|t_j^y - t_i^x|)}{\hat{\sigma}_x \hat{\sigma}_y \sum_{i=1}^{N_x} \sum_{j=1}^{N_y} K_h(|t_j^y - t_i^x|)},$$

Alternative: Lomb-Scargle estimator (Scargle, 1981) based on Fourier transform

Cross correlation function (CCF)

CCF studies linear association between x_t and y_t at lag k

- CCF for two regularly sampled time series x_t and y_t

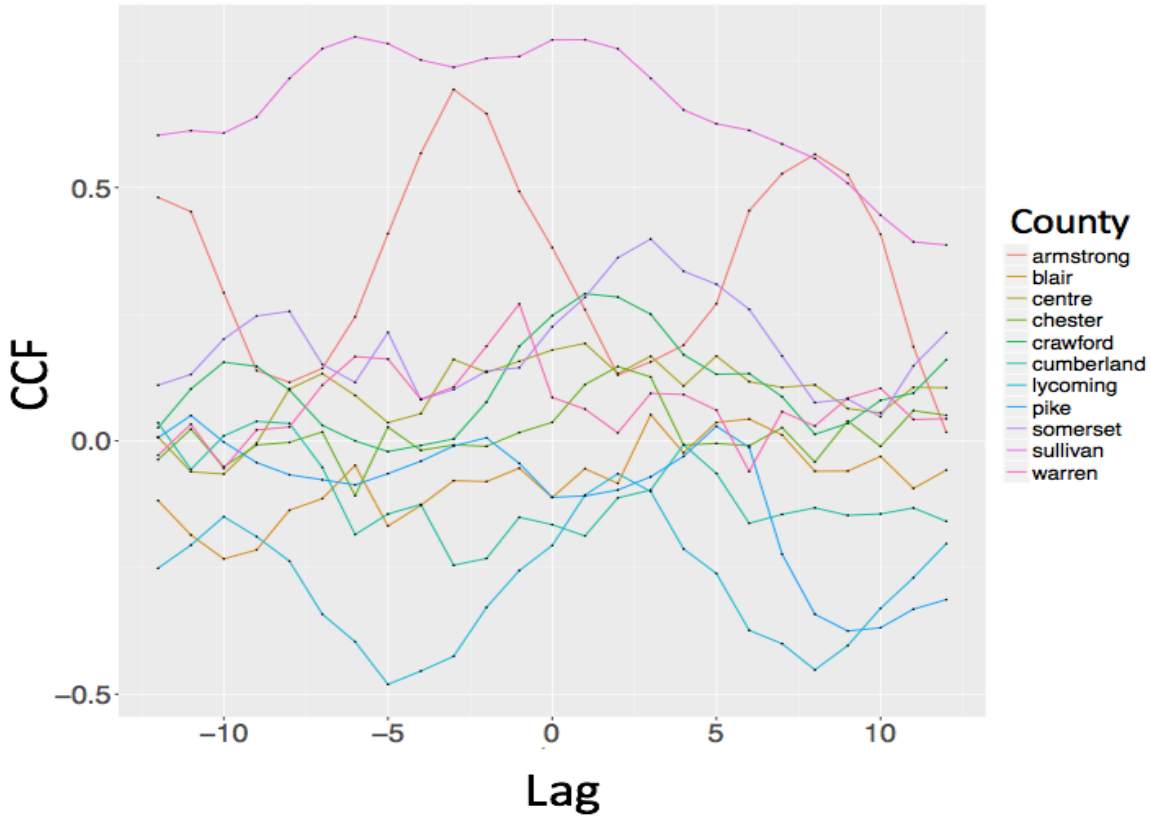
$$\hat{\rho}(k) = \frac{1}{\hat{\sigma}_x \hat{\sigma}_y (N-k)} \sum_{t=1}^{N-k} (x_t - \overline{x})(y_{t+k} - \overline{y})$$

- CCF for two irregularly sampled time series x_t and y_t

$$\hat{\rho}(k) = \frac{\sum_{i=1}^{N_x} \sum_{j=1}^{N_y} (x_i - \overline{x})(y_j - \overline{y}) K_h(|t_j^y - t_i^x|)}{\hat{\sigma}_x \hat{\sigma}_y \sum_{i=1}^{N_x} \sum_{j=1}^{N_y} K_h(|t_j^y - t_i^x|)},$$

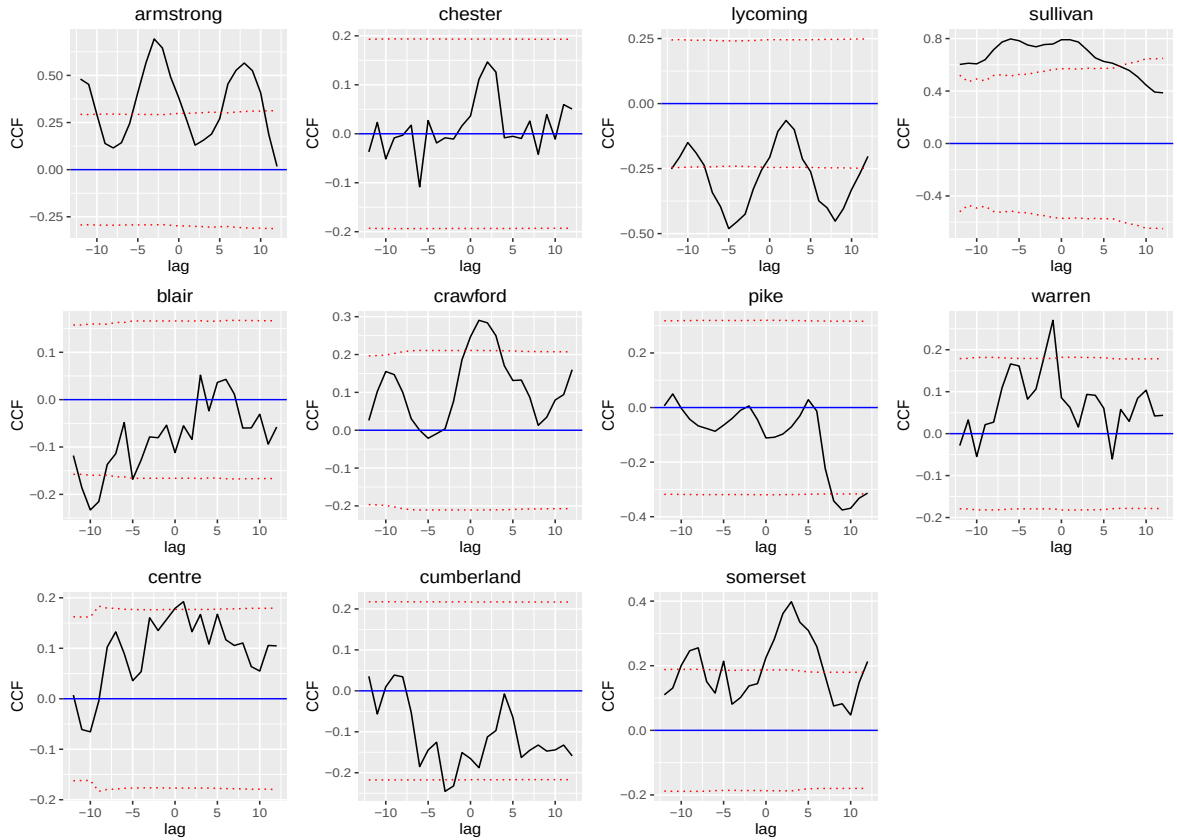
Alternative: Lomb-Scargle estimator (Scargle, 1981) based on Fourier transform

Cross correlations for Barium and Sulphate conc.



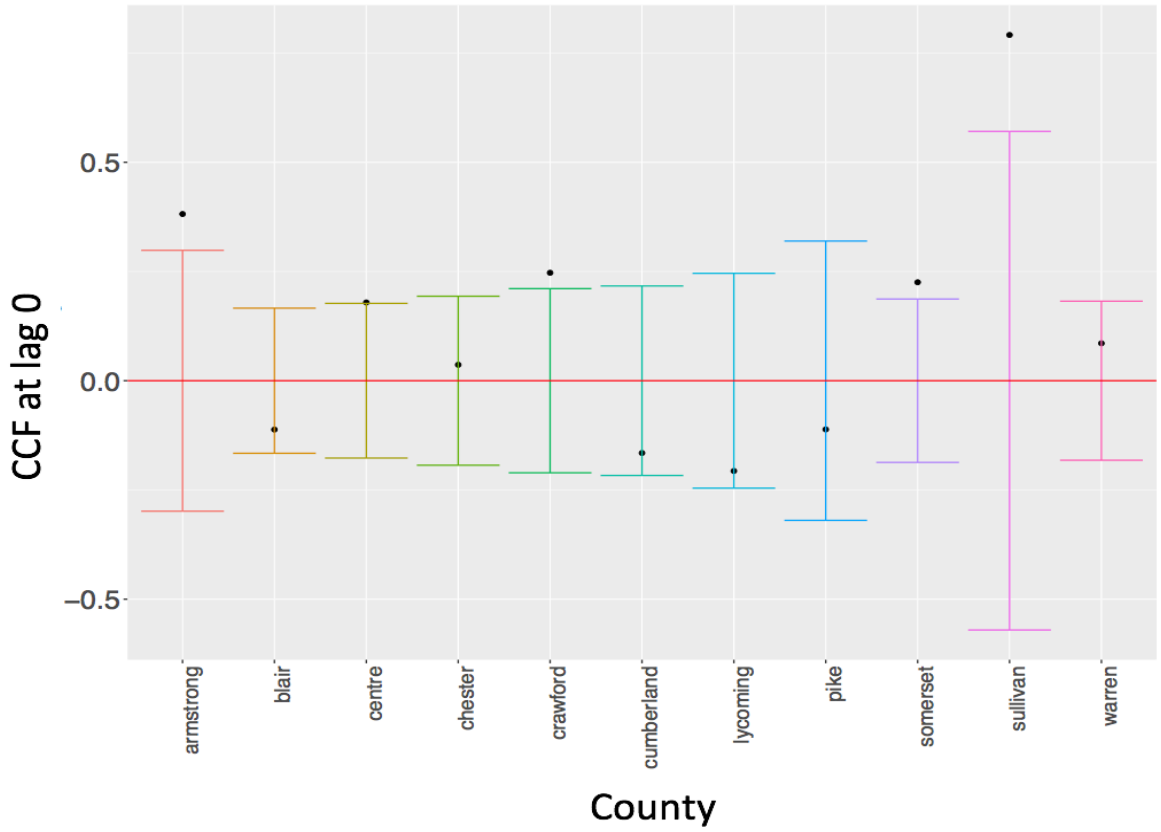
Plot of estimated CCFs with lag $k = -12, -11, \ldots, 11, 12$

Significance tests of cross correlations



Plot of estimated CCFs with confidence bounds around zero to evaluate significance at each lag

Significance tests of cross correlations
Cont'd



Plot of estimated CCFs with confidence bounds around zero to evaluate significance at lag $k = 0$

Cross correlation networks

Correlation networks

- a widely used data mining method for studying networks based on pairwise correlations between variables
- the nodes correspond to variables, and edges correspond to significant pairwise correlations

Barium or Sulphate cross correlation networks of 11 counties

- nodes correspond to 11 PA counties
- edges correspond to significant cross correlations at lag 0

Cross correlation networks

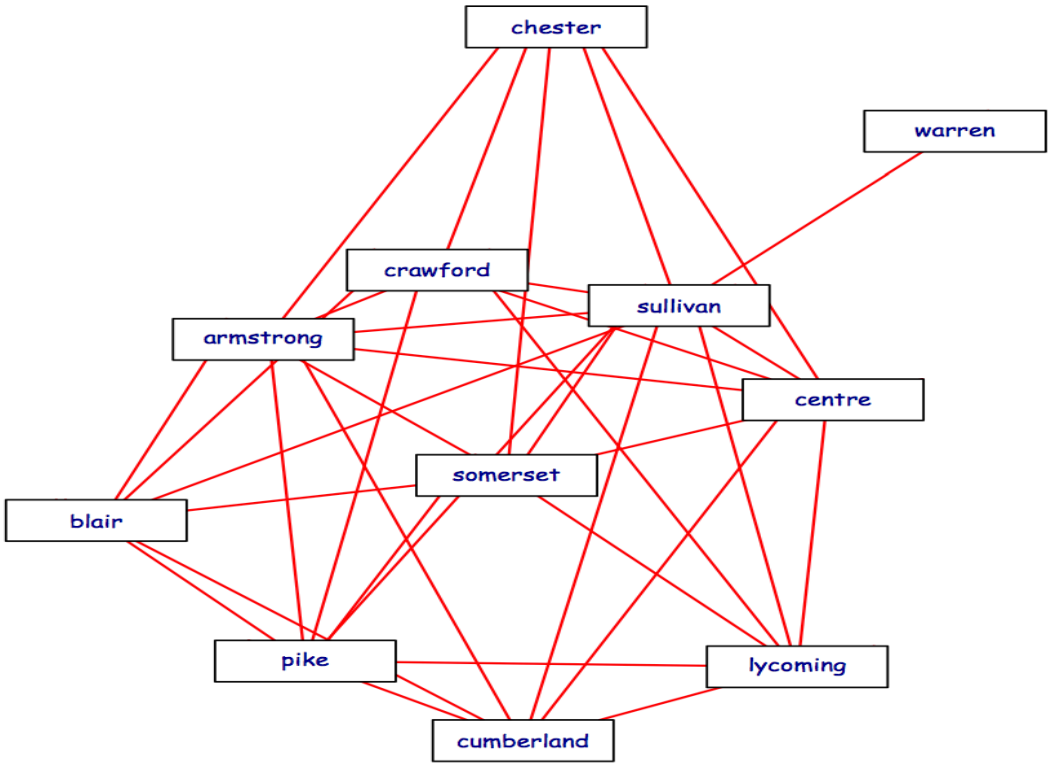
Correlation networks

- a widely used data mining method for studying networks based on pairwise correlations between variables
- the nodes correspond to variables, and edges correspond to significant pairwise correlations

Barium or Sulphate cross correlation networks of 11 counties

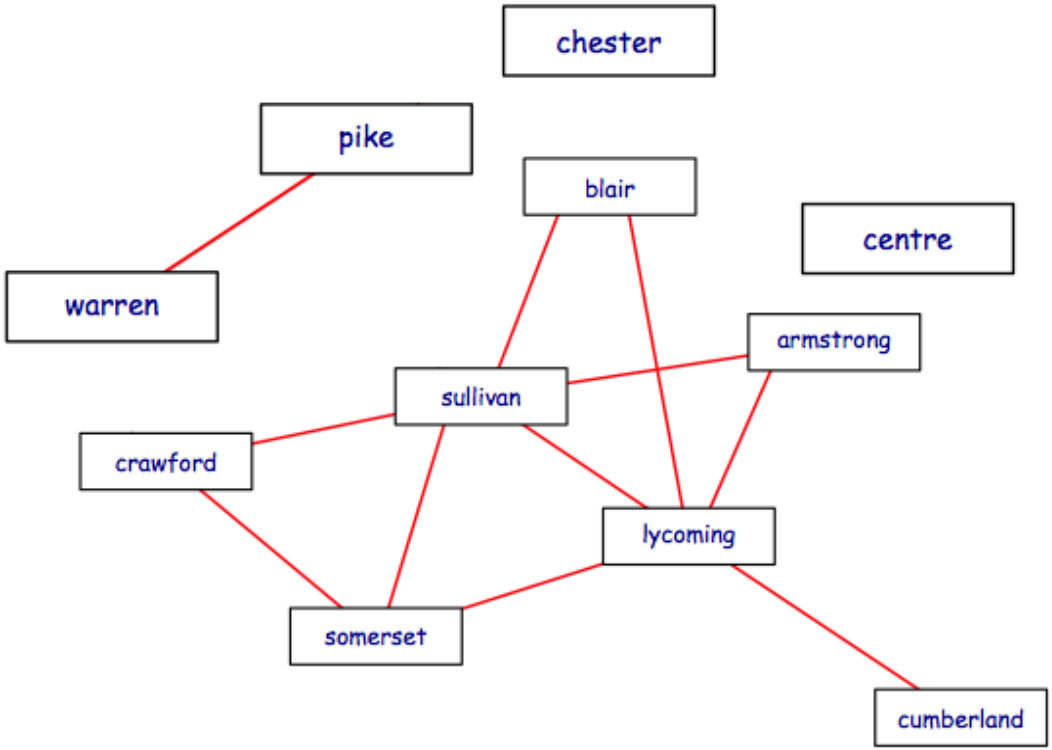
- nodes correspond to 11 PA counties
- edges correspond to significant cross correlations at lag 0

Barium cross correlation network



Network plot of 11 PA counties with significant cross correlations in Barium concentration at lag 0 from 2002-2009

Sulphate cross correlation network



Network plot of 11 PA counties with significant cross correlations in Sulphate concentration at lag 0 from 2002-2009

Outline

Discussion

Summary of current works

- Trend tests
- Cross correlation functions
- Cross correlation networks

Future works

- Estimation of large-scale cross correlation networks
- Analysis of cross correlation networks, such as clustering, causality, and many others

Discussion

Summary of current works

- Trend tests
- Cross correlation functions
- Cross correlation networks

Future works

- Estimation of large-scale cross correlation networks
- Analysis of cross correlation networks, such as clustering, causality, and many others