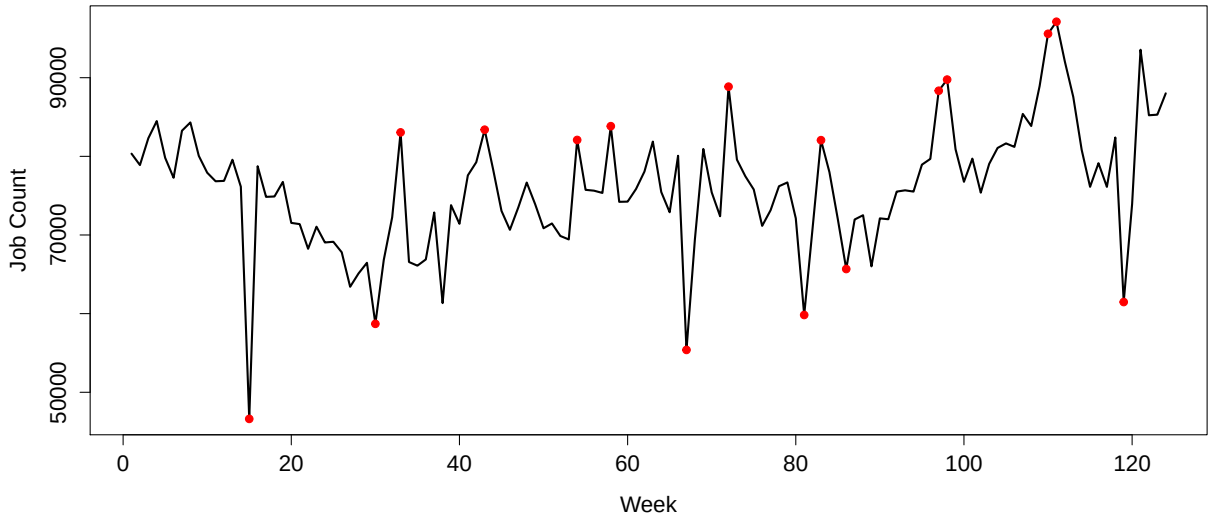
Explaining Changes in Aggregated Time-Series

Luke Rhodes-Leader

5^{th} September 2014

Supervisor: Lawrence Bardwell

Aggregated Time-Series

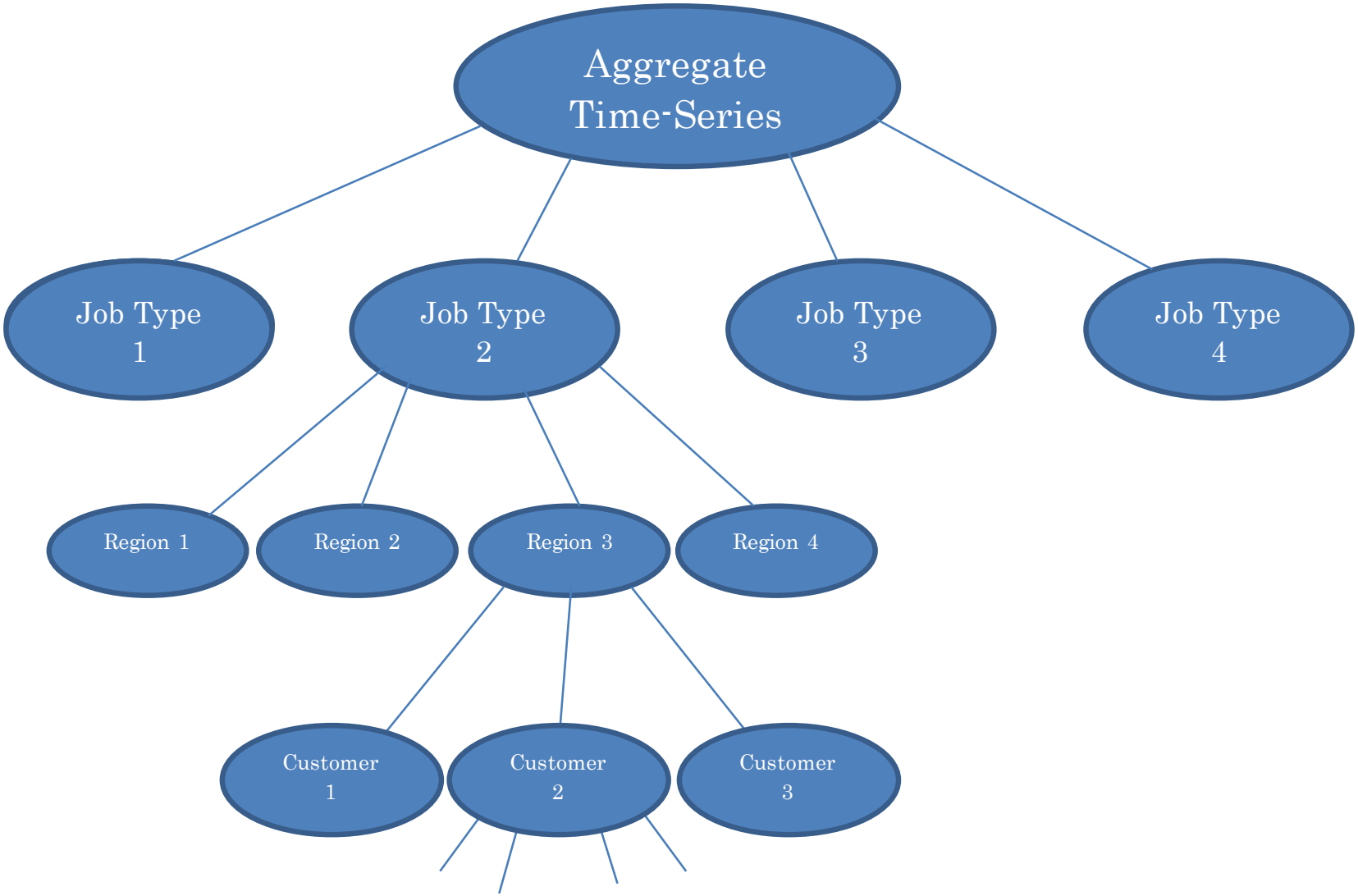


- Time-series, X, is of the number of maintenance jobs done per week on a communications network.
- It is an aggregated time-series.
- The series contains outliers and trends, for which we want to offer an explanation.
- This study focuses on:
  - Trying to find the attributes that explain the outlier at week 33.
  - Explanations of trends.

Problem

- Data represents the number of faults in a network.
- Aggregated time-series allows us to search for the set of attributes that "most caused" an outlier.
- What does "most caused" mean?
- We use two different measures of contribution.
- Explanations would help in preventing such faults reoccurring.

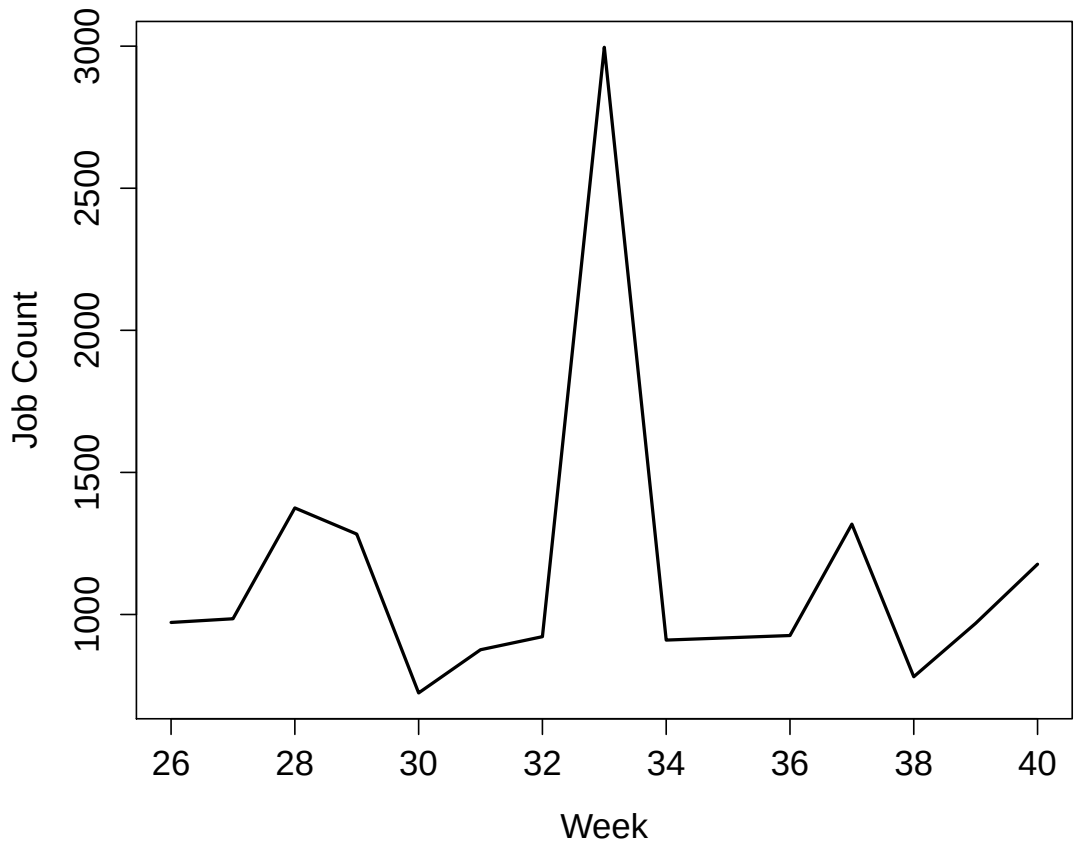
Attributes


Aggregate Time-Series
Job Type 1
Job Type 2
Job Type 3
Job Type 4
Region 1
Region 2
Region 3
Region 4
Customer 1
Customer 2
Customer 3

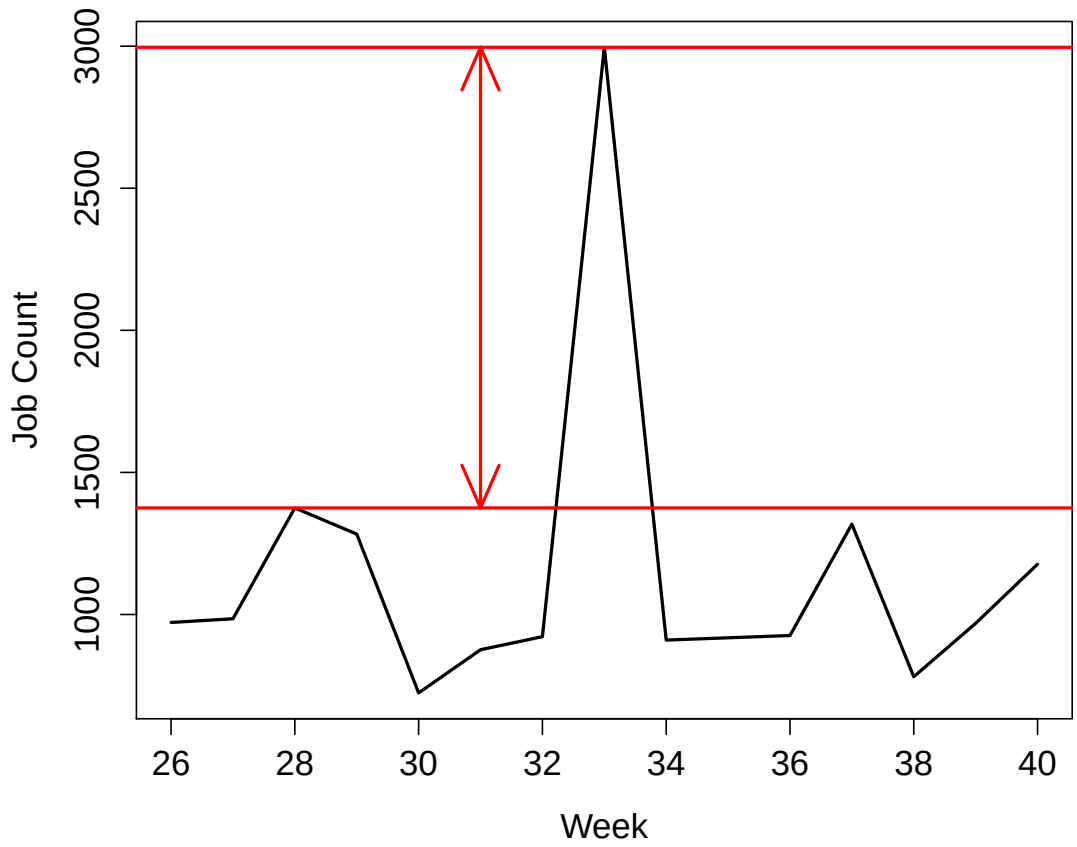
Attributes

- The different attributes are Job Type, Customer, Region, Subregion, Product 1, Product 2, Loc 1 and Loc 2.
- This study imposes restrictions on a set of attributes, $\mathcal{S}$, and measures its influence on the outlier or trend.
- Restrictions decrease the number of attributes in the set, so generally decreases the job count of that set.
- More restrictions imply a smaller impact on an outlier.
- Balance this with wanting to know as much detail as possible.
- Due to computational power limitations, at most three restrictions were made.

Standard Influence



Standard Influence



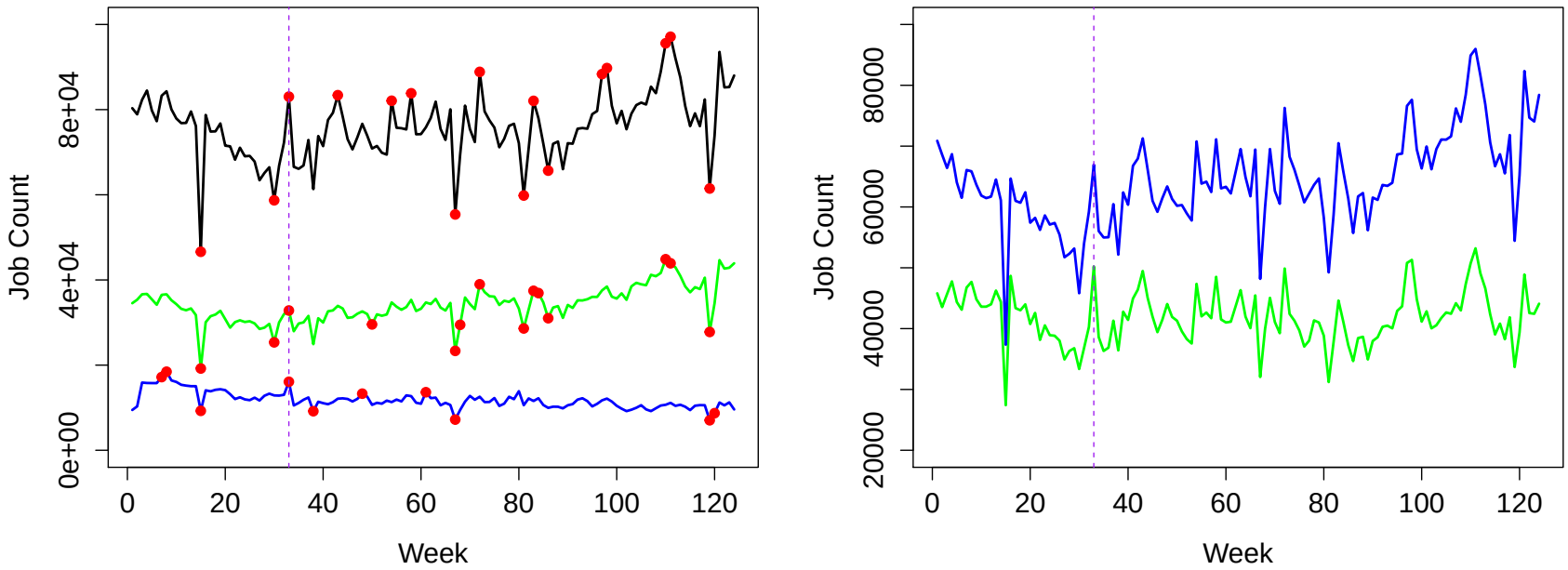
Standard Influence

- Let $\mathcal{A}$ be the set of all attributes.
- Define $agg(\mathcal{S})_t$ to be the value of the time-series of set $\mathcal{S} \subseteq \mathcal{A}$ at time t.
- Let $\mathcal{N}$ be the Normal set.
- Define standard influence of a set $\mathcal{S} \subset \mathcal{A}$ as

$$\mathcal{I}(\mathcal{S}) = agg(\mathcal{S})_t - \max_{i \in \mathcal{N}}\{agg(\mathcal{S})_i\}. \quad (1)$$

- Penalises a set if it contributes a lot to the normal set.
- Favours sets with a large additive difference.

Results from the Standard Influence



- Most influential set is $\mathcal{S}_{p8}$ with $\mathcal{I}(\mathcal{S}_{p8}) = 8915$.
- Other influential sets include $\mathcal{S}_{p8,C5}$, $\mathcal{S}_{C5}$ and $\mathcal{S}_{J2}$.
- Each of these reduces the outlier considerably.
- No sets could remove the outlier.
- This method of ranking did not give much detail.

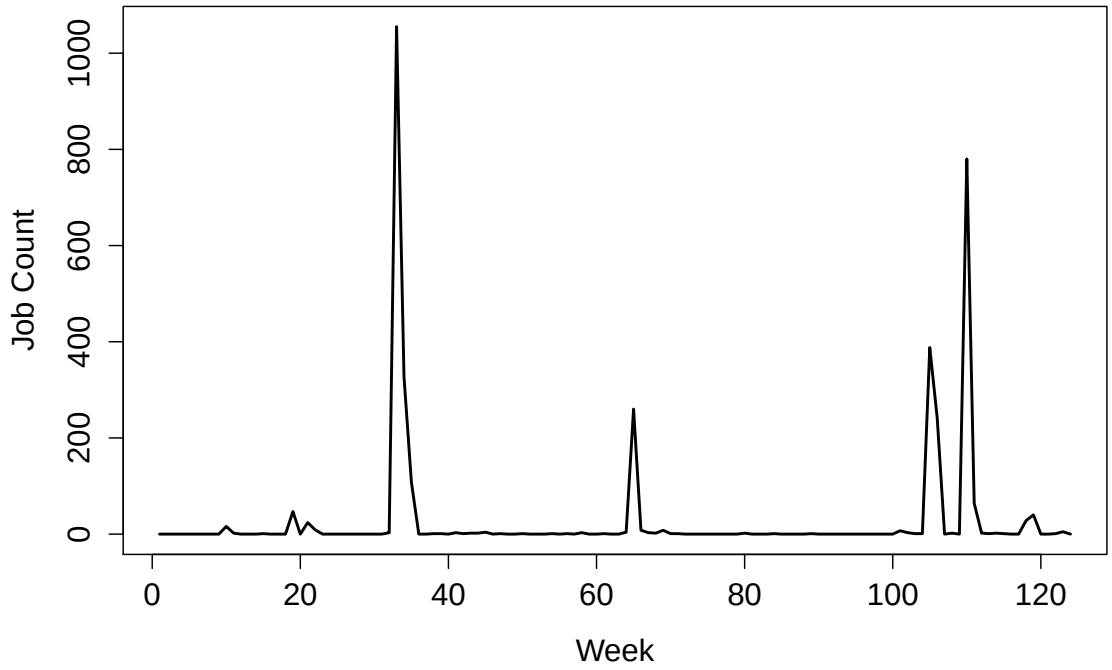
Multiplicative Influence

- Now consider the data as the rate of jobs per week.
- Then we look at the percentage change of rate, rather than additive change.
- This helps to pick out attributes that are normally stable, but undergo abnormal fault levels.
- Define multiplicative influence as

$$\mathcal{M}(\mathcal{S}) = \frac{agg(\mathcal{S})_t + 1}{Median(\{agg(\mathcal{S})_i : i \in \mathcal{N}\}) + 1}. \tag{2}$$

- This is a measures of how far above the median of the normal set the outlier is, proportional to the median.
- It favours sets with a low median and a large peak.

Results from Multiplicative Influence



- Most influential set is $\mathcal{S}_{SR23,J3,PONL4}$, with $\mathcal{M}(\mathcal{S}_{SR23,J3,PONL4}) = 1056$.
- This set has median of 0, and a large spike of 1055.
- Clearly, something went very wrong at $t = 33$.

Weather of Late April 2012

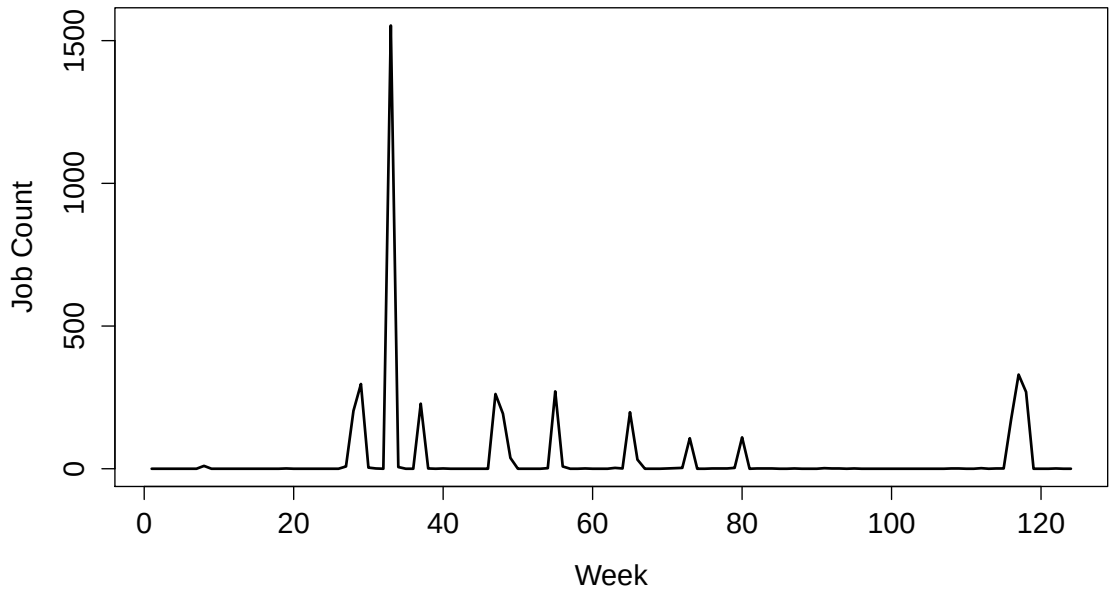
- The outlier at $t = 33$ corresponds to the week beginning 28^{th} April 2012.
- This was the end of the wettest April on record:
  - Some parts of the UK saw up to 300% of average April rainfall.
  - Widespread flooding.
- The serious weather may have caused the large increase in jobs, and would explain how widespread the problem was.

Trends

- The time-series also appears to have some trends.
- Compared gradient of linear fit of time-series and time-series without a set.
- The greater the reduction in slope, the greater the influence of a set.
- It was found that for most sections, $\mathcal{S}_{J2}$ was the most influential.
- $\mathcal{S}_{p8}$ and $\mathcal{S}_{P6}$ were also highly influential in certain sections.

Set	Gradient of Section				
	8:30	30:44	44:53	73:90	90:112
$\mathcal{A}$	-737.0	980.6	-663.0	-393.7	794.3
$\mathcal{A} \setminus \mathcal{S}_{J2}$	-198.6	-140.6	-115.2	93.4	-63.6
$\mathcal{A} \setminus \mathcal{S}_{p8}$	-272.9	307.5	-84.6	-183.1	430.7
$\mathcal{A} \setminus \mathcal{S}_{P6}$	-180.2	617.7	-183.9	-163.1	600.5

Further Work



- This time-series is of the set $\mathcal{S}_{SR17,J3,PONL4}$.
- It displays a problem with both the standard and the multiplicative influence measure.
- This example demonstrates that it is possible for both measures to miss something that could be very important.
- Further work would include finding a measure that would pick out such sets of attributes.

Questions

Thank you for listening.
Any Questions?

References

1. Wu, E. and Madden, S. (2013). Scorpion: Explaining away outliers in aggregate queries. PVLDB, 6(8):553-564.
2. http://www.metoffice.gov.uk/news/releases/archive/2011/wettest-april-on-record
3. http://www.theguardian.com/uk/2012/may/02/uk-may-need-standpipes-drought