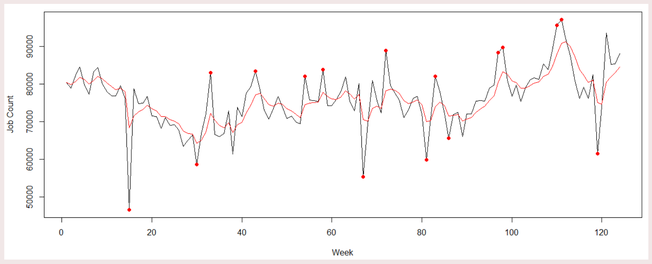
Explaining Changes in Aggregated Time-Series

Luke Rhodes-Leader[1] Lawrence Bardwell[1]

[1]STOR-i, Maths & Statistics Department, Lancaster University.

Background

An *aggregated time-series*, X, is one created from the sum of the time-series from various different attributes. Any outliers that occur within the aggregated time-series may not be a shared property of all the time-series, but may actually belong to specific attributes. In this case, it is possible to find the set of attributes with this outlier in its time-series, and thus offer an explanation for the cause of the outlier.



The data set considered contained the number of maintenance jobs done on a communications network per week, and so the number of faults occurring each week. Each time point is the sum over each attribute with 8 different properties including Job Type and Region. Identifying the attributes that caused the outliers is crucial in helping to identify problems and allocating a work force to specific tasks to prevent such events reoccurring.

The data can also be viewed as a rate of jobs per week. Therefore, as well as looking at high additional increases in job rate, the percentage increase over the median is also important, as this indicates an abnormal fault in that part of the system. These are also important in prevention of future problems.

Standard Influence

Let $\mathcal{A}$ be the set of all attributes contributing to a time-series X, let $\mathcal{S} \subset \mathcal{A}$, and let $agg(\mathcal{S})_t$ be the value of the time-series associated with set $\mathcal{S}$ at time t. Thus, $X_t = agg(\mathcal{A})_t \ \forall t$. For each outlier, also define a set of points which are considered "normal", $\mathcal{N}$.

- The first measure of influence used is the *standard influence*, defined as

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

- Measures the absolute difference between $agg(\mathcal{S})_t$ and $\max_{i \in \mathcal{N}}\{agg(\mathcal{S})_i\}$.
- Ranking sets allows removal of outliers as it favours large changes.
- Does not allow much detail, as generally, the more restrictions imposed on the set, the fewer attributes it contains, and so the peak size reduces.

Multiplicative Influence

- The other measure used is the *multiplicative influence*:

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

- A measure of the factor that $agg(\mathcal{S})_t$ is above median$(\{agg(\mathcal{S})_i : i \in \mathcal{N}\})$.
- Favours sets with low median and relatively large peak.
- Add 1 to top and bottom to prevent dividing by 0. (All time-series points are non-negative).
- This is useful if you view the time-series as a rate of jobs done.
- Fails to pick out some large peaks due to division.

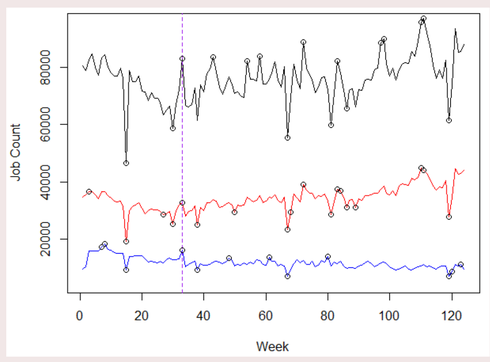
We will refer to the number of restrictions on a set as its level.

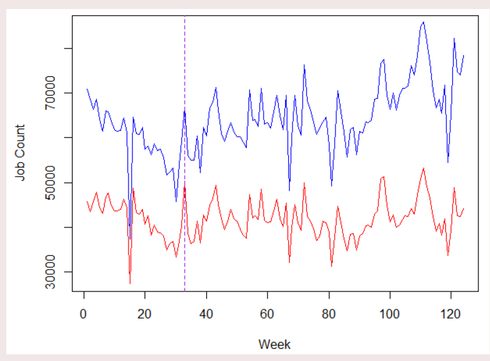
Outlier 33 - Standard Influence

The main focus of this study was on the outlier at $t = 33$ in the time-series.

- No single set of exclusively restricted attributes could remove this outlier.
- The highest value of $\mathcal{I}(\mathcal{S})$ was 8915 for the set of attributes with Product 2 type 8, $(\mathcal{S}_{p8})$.
- Another very influential set was that of attributes with Job Type 2, $(\mathcal{S}_{J2})$, with $\mathcal{I}(\mathcal{S}) = 6498$.

These graphs show both of these and the time-series with these removed, the white dots are outliers in the data. ($\mathcal{S}_{p8}$ in red, $\mathcal{S}_{J2}$ in blue.)

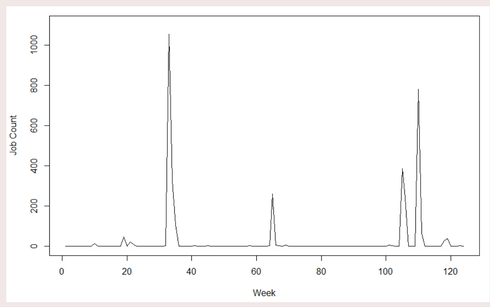




- The clear peaks at $t = 33$ is common of many sets, suggesting that the problem was widespread.
- The graphs also show that Job Type 2 carries much of the graph's shape, and removes the other outliers.

Outlier 33 - Multiplicative Influence

- This influence gave quite different results.
- Most influential set is $\mathcal{S}_{SR23,J3,PONL4}$, with $\mathcal{M}(\mathcal{S}_{SR23,J3,PONL4}) = 1056$.
- This is a level 3 set, with median 0 and spike of 1055 at $t = 33$.
- Given that the expected job rate for this set is 0, this is a sign of something serious going wrong.
- The great increase of rate suggests an abnormal fault occurred.

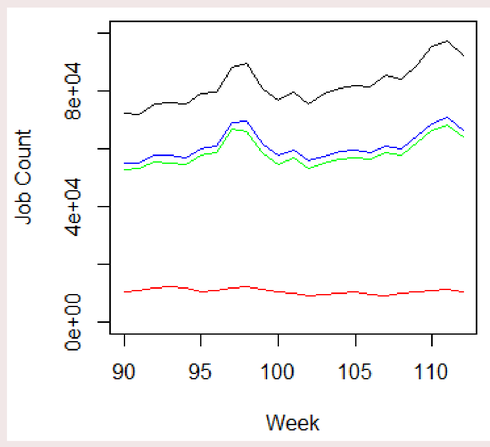


Weather of Late April 2012

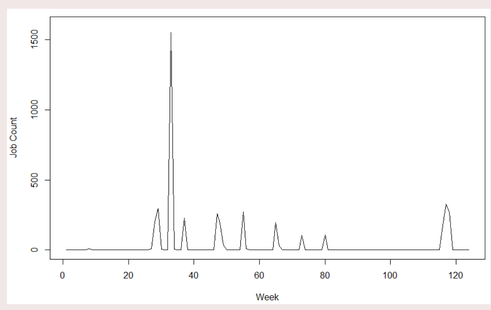
- The outlier at $t = 33$ corresponds to the week beginning 28^{th} April 2012.
- This was the end of the wettest April in a century, with some parts of the UK seeing up to 300% of their normal April rainfall and widespread flooding.
- The serious weather may have caused the large increase in jobs, and would explain how widespread the problem was.

Trends

- There are sections where the number of jobs is increasing or decreasing.
- For example, the section from week 8 to week 30 shows a clear decrease, whereas between weeks 90 and 112, there is a strong increase.
- By removing different sets, the gradient of the section's linear fit can be changed. The flatter the resulting gradient, the more influence a set has on the trend at that point.
- Level 1 sets were always the most influential, particularly $\mathcal{S}_{J2}$ and $\mathcal{S}_{p8}$.
- Between weeks 90 and 112, removing $\mathcal{S}_{J2}$ (shown in red) reverses the gradient from 794.3 to -63.6.
- For more detail, the most influential level 3 set is $\mathcal{S}_{P1,p1,J2}$ (blue), with gradient 373.8.
- This section is the only one in which Products $P1$ and $p1$ (green) were more influential than $P6$ and $p8$. This could either suggest an increasing fault in $P1$ and $p1$, or that they are products that are being phased in over this period.



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:
  - Its peak is 1553, suggesting a serious problem.
  - It is level 3, giving more detail than other sets.
  - However, the peaks on either side reduce its $\mathcal{I}(\mathcal{S})$, so it is not picked out in comparison to the level 1 sets.
  - Also, its median is 2.5 which reduces its $\mathcal{M}(\mathcal{S})$, lower than a spike of 500 and median 0.
- This example demonstrates that it is possible for both measures to miss an attribute that contains a significant fault.
- Further work could include finding a measure that would pick out such sets of attributes.
- This study also considered an algorithm to detect which attributes contain the outliers. Methods included use of a Gibbs Sampler, and further development of this process could be done.

References

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

l.rhodes-leader@lancaster.ac.uk