Police Data Challenge

Simon Fraser University

Barinder Thind, Matthew Reyers
Brad Smallwood, Ryan Sheehan

Fall 2017

Methodology Description

We decided to focus almost exclusively on the Baltimore Emergency Call data which provided information on calls made within each district and the priority of that call. The goal we set for ourselves was to create a model that can look at proportions of the various priority levels by week and intelligently filter out the number of weeks prior that had the highest influence in the proportions present in the current week.

One of the main issues to deal with was that of seasonality trends within the data. We viewed this data through a time series lens and so, we looked at all possible number of lags for some given week to see when the effect of seasonality would become apparent. By looking at the autocorrelation coefficients, we were able to determine that the white noise effect came into play at either the 4th, 5th, or 6th week depending on the district we were looking at. This finding caused us to build our model with respect to the x_i[1] most recent weeks.

The model is based on functional data analysis as we treat each week of data as a function[2]. Every week prior (in some combination depending on the number of lags) is compared with the current week by using the weighted fit values for the corresponding lags (provided by the autocorrelation function). As an example, assume that we are looking at the high priority of some district; then the following formula will give us the predicted result for the current week:

$$high_n = w_{n-1} \cdot f(high_{n-1}) + w_{n-2} \cdot f(high_{n-2}) + w_{n-3} \cdot f(high_{n-3}) + w_{n-4} \cdot f(high_{n-4})$$

The formula is extended to a 5^{th} or 6^{th} lag term in districts that require five or six lags and is calculated identically to the terms above. Each weight is determined as a normalized correlational coefficient and must satisfy the following equation:

$$\sum_{i=1}^{\mathit{max(lag)}} w_{n-i} = \sum_{i=1}^{\mathit{max(lag)}} \frac{abs(r_{n-i})}{\sum_{i=1}^{\mathit{max(lag)}} abs(r_{n-i})} = 1$$

The correlational components are calculated by running historical data through the autocorrelation function. The current week, as it does not have results yet, has to be independent of the data we are using. The result generated by the $high_n$ function above represents our prediction for the proportion of high priority calls in a given district. The model naturally extends to all other priorities by using the corresponding weights for those given priorities.

As a way of cross-validating this methodology, we looked to show that the lag for which our autocorrelation function hits its boundary for insignificance (i.e white noise) is the same lag at which our error for the weighted spline predictions is minimized. For example, we can consider the district *WD* error plot and it's autocorrelation function for some randomly picked priority (the other priorities will display similar results):

[1] Where $x = [4 \cup 5 \cup 6]$ and i refers to a particular district

[2] Predicted by a smoothing spline

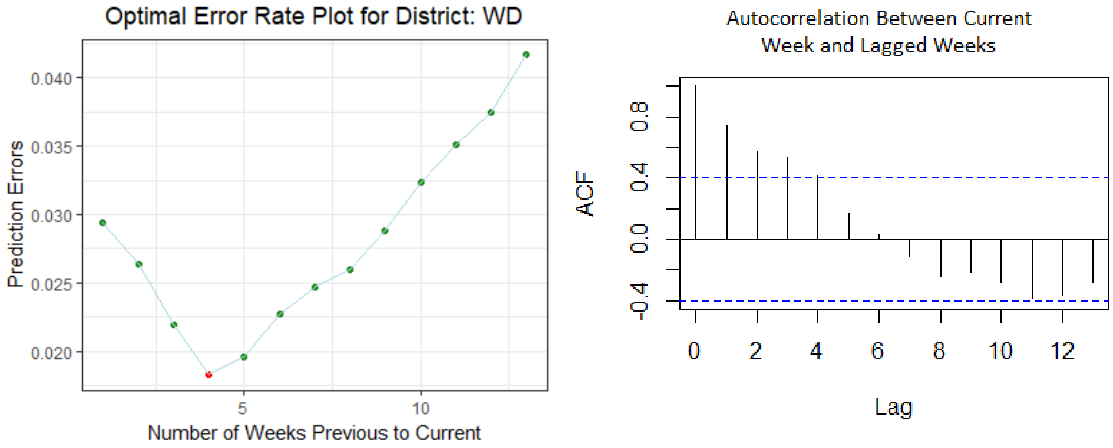

Figure 1: An example (District: WD, Priority: Medium) to showcase the optimization using our cross-validation method. The red dot on the left plot is the minimum prediction error whereas the blue dashed lines in the right plot indicate the significance cut-off. Both occur at $lag = 4$.

The equivalence in Figure 1 isn't always exactly the case for all priority levels of a given district however, on average, and more often than not, this result is shown. Also, the places at which its not seen is usually for district and priority combinations with very few observations. A more detailed explanation can be made available if needed![3]

[3] Also, links to some interactive tableau visualizations are included in the presentation aspect of the competition