STAT 482 Final Report: Traffic Data Analysis

Jonathan Chu, Mali Selman, Pranika Madhavan,
Gyumin Hwang, and Michael Lee

May 2021

1 Introduction

Auto collisions are steady causes of stress in our street frameworks. A lot of it is the consequence of our clogged streets tormented in rush hour gridlock. A noticeable area in California, with its exceptionally thick populace and clogged streets, ends up being the very illustration of this. Therefore, creating a tracker to screen the traffic in hot spots within California streets may fill in as an incredible instrument in understanding the relationship between the measure of vehicles passing by a specific stretch and the measure of car crashes that happen. With the assistance of sensors, we will actually want to distinguish the spaces of clog and reach inferences on occasions, for example, the normal time between successive car crashes on a bustling street and how much more probable a mishap will happen if there is two-fold the amount of traffic at a specific stretch. Additionally, when it comes to the focus of our datasets, we needed to analyze data points that coincide with holidays and understand their link to possible traffic patterns. This in turn, would solidify our outlier identification process and help in using our model in other datasets in the future.

2 Exploratory Data Analysis

Figure 1: Map displaying sensor locations

As shown in the map above, there are 10 sensors investigated in total with all of them located in northeast Sacramento, CA along a major highway. All of the

sensors are also located very close to each other, with the furthest between two sensors being only 2.3 miles. This allows for the population to provide incident data which remains relatively constant throughout the data collection. For the full calendar year of 2017, the sensors reported the traffic every five minutes, adding up to 288 data points recorded each day. This method was utilized with the sole exception of daylight savings day on Sunday, March 12 of that year.

Examples of graphs that were generated from this data are shown in Figure 2. The average for a day of the week was computed (with pre-decided holidays excluded) and mapped in red. The blue line displays the holiday's (respective to the graph title) traffic pattern. The goal in visualizing these differences was to get a more full scope understanding of how the daily traffic varied from the holiday traffic. For example, the graph labeled "Thanksgiving Day vs. Normal Day" compares the points from Thanksgiving Day of 2017 to the average "normal" Thursday of 2017. By doing this, we hoped to gather significant comparison information on how certain sensors may change during holidays and special events.

We can gather from the above graphs that the average days, regardless of the actual day of the week, seem to follow highly similar patterns (aside from the weekend where there is a noticeable peak in the morning). We can infer this differential stems from typical workday traffic vs. weekend traffic. It can also be noted that on weekdays, traffic incidents remain stagnant throughout the day until a steep drop in the evening, where significantly less people are out driving. These rush hour spikes appear throughout the week and slightly on Saturdays, however Sunday demonstrates its own unique traffic pattern. This can be noted on the example in Figure 2 displaying the Super Bowl Sunday traffic comparison.

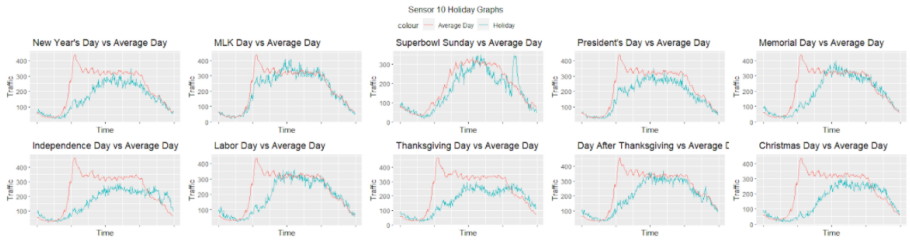

Figure 2: Example of EDA using sensor 10. The blue line displays the holiday traffic pattern and the red line displays the average for the respective day of the week which the holiday falls on.

After careful consideration, we decided that the most appropriate path forward in detecting holidays based on traffic incidents would arise from outlier detection. This method would allow for us to classify points as either potential holiday candidate points, or as within the threshold for a "normal" day that

year. Two methods were employed to perform the analysis to the fullest extent possible.

3 Statistical Models and Result

3.1 Method 1: Outlier Detection with Standard Deviation Envelope

The initial attempt in the formal analysis consisted of counting the number of outliers and determining if this number exceeded a predetermined threshold. If the number indeed exceeded this threshold, the curve would be classified as a holiday. A point was classified as an outlier if the difference between the count on the actual day and the count on the average corresponding day was greater than one standard deviation from the mean. Looking at the graphs produced, the days that represent holidays clearly displays visual differences that are quite noticeable.

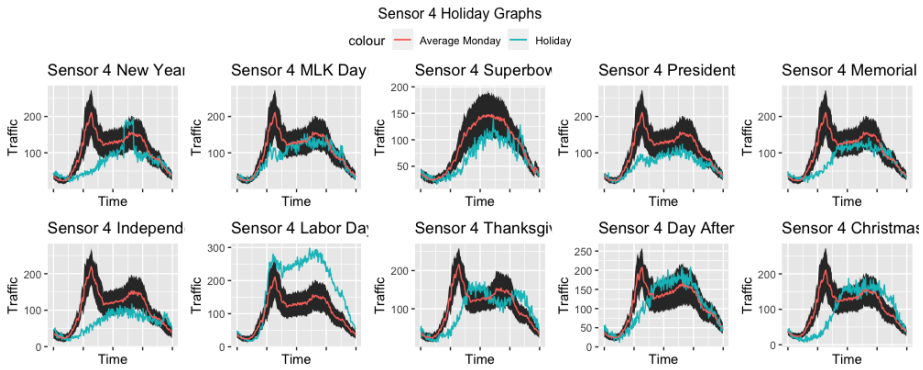

Figure 3: All holidays with standard deviation envelopes for Sensor 4

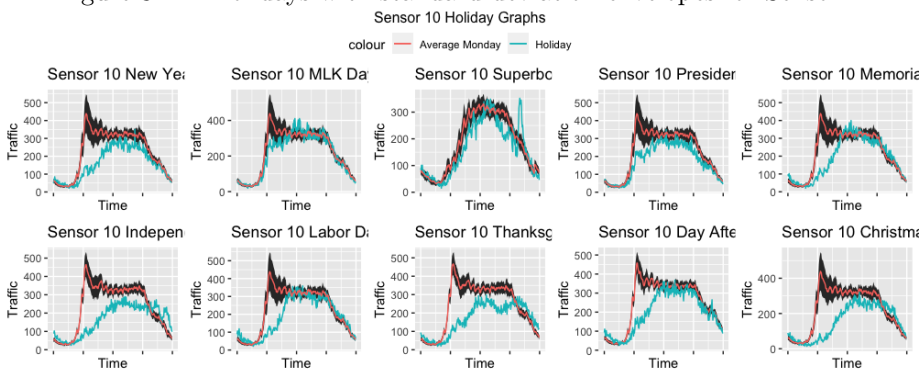

Figure 4: All holidays with standard deviation envelopes for Sensor 10

This led us to believe that this method could offer high accuracy while being a simple statistical method. Thus, the next step was to create a function to count the number of outliers in each holiday date at each sensor. Once these values were obtained, the total number of outliers for a given holiday was plotted for each sensor. As the graphs below show, the range of outliers in the sensors is large.

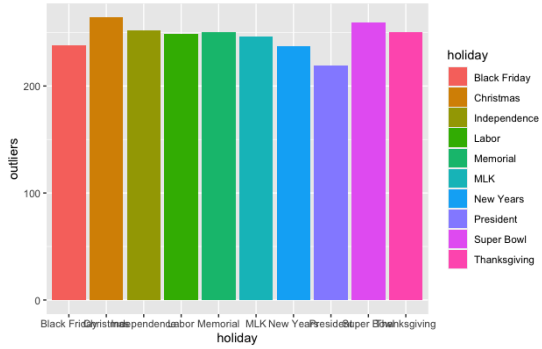

Figure 5: All holidays with standard deviation envelopes for Sensor 4

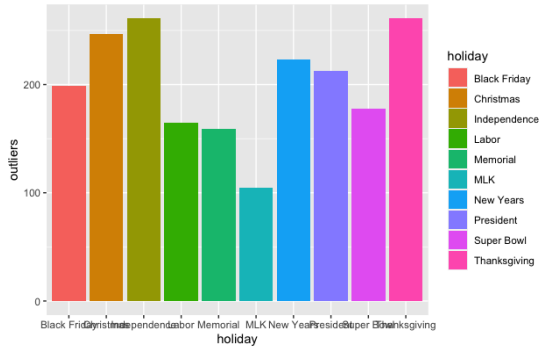

Figure 6: All holidays with standard deviation envelopes for Sensor 10

This highlights the flaw in this method as the wide range in the number of outliers creates difficulty in determining a universal threshold that fits every sensor. Furthermore, we do not want to set this threshold too low as that would cause serious overfitting that results in a powerless statistical model. However, the range is still too large to develop a threshold that is reliably accurate.

While the approach of this statistical model is effective in theory, it fails to be universal. Even though outlier detection is effective and simple, the next step, involving the determination of a flexible threshold, shows the futility of this model. Although the detection of individual outlying points is practical, there is too much sacrifice in terms of overall accuracy for this model to be implemented. Therefore, we decided to try FDA as our statistical model in hopes

to improve the overall accuracy and functionality of the model even at the cost of efficient and simple outlier detection.

3.2 Method 2: Functional Data Analysis (FDA)

The FDA method uses techniques to identify ways curves vary from one another. Therefore, we were able to use this approach to understand how incidents of traffic on a particular day vary from the average day of the week. This provides the necessary information to answer our problem on whether holidays are a good indicator of whether or not we choose to monitor and accommodate for these days in the future.

In our data, every day of the week was plotted for each sensor. Then a functional boxplot was obtained to identify any days which were outliers. Figures 7 and 8 show these plots for Mondays of the year.

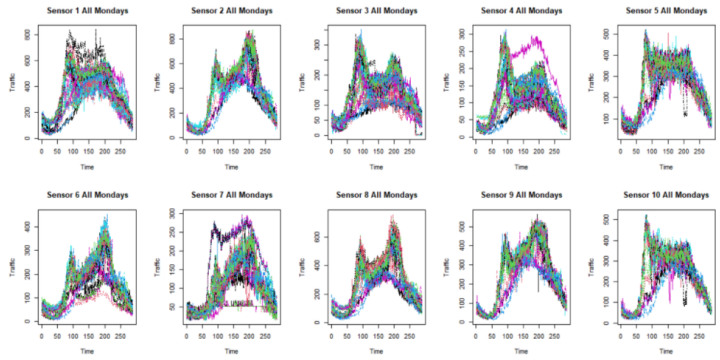

Figure 7: All traffic incidents for each Monday of the year graphed for each of the 10 sensors

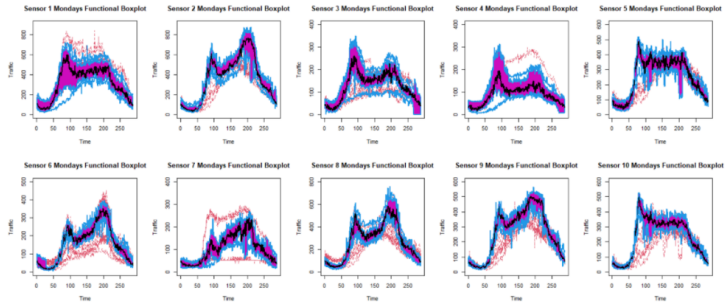

Figure 8: Functional boxplots of Mondays for each of the 10 sensors

Figure 8 shows the curves that are identified as outliers as red lines. The black line is the median curve while the purple and blue lines represent the curves

that lie within the margins of a boxplot. When using the functional boxplot, a list of outpoints were obtained that identified which curves were outliers. Since the curves were ordered chronologically, we were able to cross verify whether a holiday was correctly identified as an outpoint. Figure 9 shows the Monday holidays alongside their corresponding order in the year. A list of outpoints for sensor 10 Mondays is also shown below.

Holiday	nth Monday of the year
New Years	1
MLK Day	3
President's Day	6
Memorial Day	22
Labor Day	36
Christmas	52

```
$outpoint
 [1]  1  3  5  6  8 22 25 26 36 49 52
```

Figure 9: Table showing a correspondence table for the nth Monday of the year and the related holiday. The output contains the Mondays selected for outlier analysis.

One can see that every holiday was identified amongst every Monday. Functional boxplots oftentimes also included days which were not listed holidays as well. Figure 9 shows that the 5th, 8th, 25th, 26th, and 49th were identified as outliers among all the curves of sensor 10 mondays. It is possible to include future analysis to classify and identify these days as well. Looking at a sensor by sensor analysis however, we observed varying results on the efficiency of using FDA. For sensor 10, every single holiday we considered was correctly identified. However in some sensors such as sensor 1 and 4, FDA missed a large proportion of the holidays. The plots in Figure 10 show the functional boxplots for each day of the week of sensor 1.

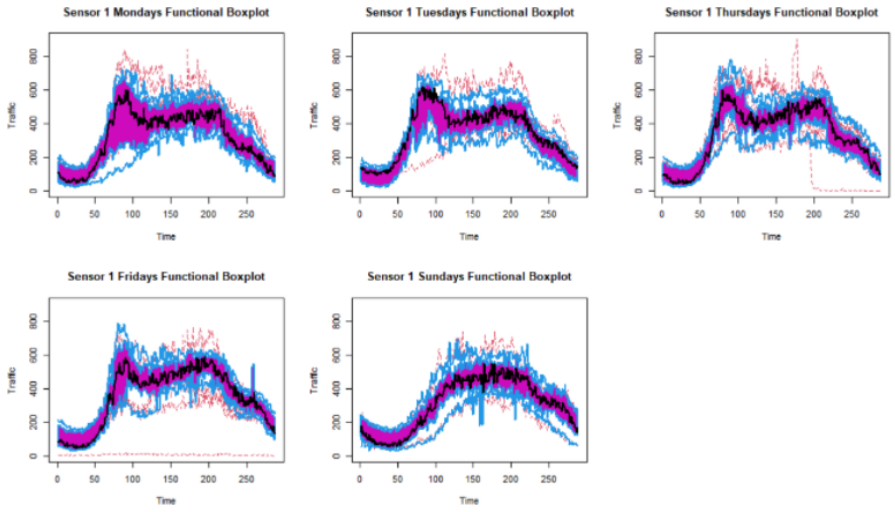

Figure 10: Functional boxplots of Tuesdays for each of the 10 sensors

Visually analyzing these plots, we can see that there are clearly curves which could be considered outliers that were not highlighted. This is especially apparent in Sensor 1 Sundays Functional Boxplot. When matching with the outpoints provided by this method, we were unable to identify every single holiday as an outlier. Thus, to better understand the efficiency of our FDA model, we observed the frequency of missed detections. Figure 11 tabulates the frequency of a holiday being missed, as well as the number of missed detections for a certain sensor.

Holidays and corresponding frequency of failed detections:

Mondays	Missed	Tuesdays	Missed	Thursdays	Missed	Fridays	Missed	Sundays	Missed
New Years	1	Independence	1	Thanksgiving	4	Black Friday	5	Superbowl	1
MLK	6								
Presidents	6								
Memorial	4								
Labor	2								
Christmas	3								

Sensor	Proportion of Holidays Identified
1	0.2
2	0.8
3	0.6
4	0.1
5	0.7
6	0.8
7	0.5
8	0.7
9	0.9
10	1.0

Figure 11: Tables of frequency of missed detections

We can see that certain sensors had more success with FDA than other sensors. Notably, sensors 9 and 10 were able to identify most if not all of the sensors while sensors 1 and 4 were unsuccessful in almost all of them. Reasoning behind the sensors that were not so successful has to do mostly with the outliers that were not holidays. Such outliers were oftentimes much more extreme than the holiday curves. In terms of preparing traffic conditions for future years, we take a look at the frequency of correctly identifying a holiday. Take note in figure 11 that independence day was the most identified holiday, with it being detected by 90 percent of all sensors. This shows that when comparing all Tuesdays, Independence day in particular would require the most monitoring or change to increase the efficiency of traffic. Meanwhile for holidays that were frequently undetected, they were oftentimes overshadowed by other outlier days when using the FDA method and does not necessarily mean that they are "unimportant" days that we can ignore.

Overall when comparing the model in method 1 with our model in the FDA approach, we can see that FDA is less flexible when determining outlier points. We have problems with underfitting in the FDA method. In this analysis of traffic, the seasonality was based on the day of the week. However, there may be different methods of grouping days which might have returned better results for FDA since some days of the week had considerable amounts of outpoints that were not holidays.

3.3 Conclusion

California served as an appropriate choice to implement the processes used for the analysis as it houses an exceptionally dense population which links to its high traffic congestion. Detecting outliers using standard deviation envelope worked well for determining differences as it overfit the data but on the other hand, the FDA method tended to underfit the data so it only worked on some sensors that were prone to being identified more easily such as Sensor 10 and weaker for others such as Sensor 4. Both methods were successful for certain cases in determining these differences and could serve as outlier detection methods to use for future or other datasets regarding traffic.

4 References

Chang, G.; Xiang, H. (2003). The Relationship Between Congestion Levels and Accidents. Maryland State Highway Administration, 1-86. doi:MD-03-SP 208B46

Manolakis, D. ; Truslow, E. ; Tsitsopoulos, G. (2018).Event Detection In Time Series:A Traffic Data Challenge[PowerPoint slides]. MIT Lincoln Laboratory

5 Appendix

Data Analysis Code: https://github.com/ghwang0307/stat_capstone