Predicting Traffic Flow using Bluetooth in Austin, Texas

Brandon Sams

Bellevue University

# Executive Summary

The City of Austin, Texas has a traffic problem. This is not unique to the city of Austin, of course, but they have implemented a solution that can, at least, keep track of the problem more easily. Dispersed around several intersections in the city are very small sensors that can keep track of Bluetooth signals that are present. This technology is used to monitor traffic flow throughout the city.

This happens because the Bluetooth sensors are able to uniquely identify a Bluetooth source, such as a smartphone from a vehicle. The sensors are smart enough to cross reference with each other, which gives the city a view into how long it takes to travel from one intersection to another, and how many vehicles travel down that route. Prior to the data being made public, the unique identifiers are obfuscated for security and privacy reasons.

This paper looks at a particular section of the road, which is 3.5 miles long and on a major roadway. If any section of road is going to give us results, it is this one. We were able to observe that the traffic volume is different on weekdays and weekends, but that is only true from about 10:00 am until 5:00 pm. Otherwise, the traffic volume was similar, regardless of the day of the week.

We also were able to identify that the travel times will increase by about 50% at the 2:00 pm peak. Yes, that also means that the peak time for this section of road is not at 5:00, as originally thought. This does not appear to be a time zone conversion error, as the time was reported in local time. This research was able to predict travel times and traffic volume for a particular section of road in Austin, Texas. I anticipate that similar techniques could be used around the remainder of the city.

# Intro/Background of the Problem

Traffic in the City of Austin, Texas is, like in most big cities, one of the most frustrating parts of living there. Estimates that are published in a local newspaper, The Statesman, state that Austin is ranked 14th in the country when it comes to congestion. This is a big sink for not only time, but money as well. Austin commuters lose an average of $1270 annually. This puts it roughly in the same ballpark as some other notable cities as San Diego, California or Portland, Oregon.

In order to deal with this, the city installed a series of bluetooth travel sensors to better keep track of how vehicles move about the city. Specifically, these travel sensors will simply record when they observe a bluetooth device in range, and an identifier for that bluetooth device. This makes it so that when the same bluetooth device appears in range of another sensor down the road, it knows it is from the same bluetooth device. Combining this data with information such as the distance between sensors, and when the bluetooth device was observed, measurements can be made for travel time and average speed over a given road segment at a particular time of day.

And if you are concerned about data privacy, the MAC addresses that the bluetooth device records are anonymized prior to being made public. The City of Austin's Github Repo for the project states "there are no real MAC addresses in the Bluetooth sensor data. Instead, MAC addresses are generated randomly on a daily basis. The MAC address mapping is retained for one day so that OD pairs can be identified for a 24-hour period."

This makes for a large repository of data that deserves to be turned into actionable insights for the city planners and traffic engineers.

The data that is available from the city is not available in a real-time format. However, it is available as a historical record. The raw data is available for each and every bluetooth signal that these sensors pick up on, but it is also available in two other formats.

There is a large table, titled "Individual Traffic Match Files (ITMF)", that keeps track of a single vehicle's journey between sensors for every record in the table. It is similar to the raw data, but is filtered to include only records that are very likely to be sourced from vehicles.

There is also a slightly smaller table, titled "Traffic Match Summary Records (TMSR)". This has summary statistics for how many cars travel down a stretch of road, what the average speed is during that time, and a collection of other attributes. This is broken up into 15 minute sections, going back to 2014.

These tables are available as a downloadable .csv file, but it can also be queried using a http GET request.

# Methods

Due to the scale of the dataset, it appears that it would be best to not analyze all road segments in the City of Austin. They all vary a lot in length, average speed, speed limit, and traffic flow, among other things. Any model that is generated for one road segment is not likely to be predictive of the behavior of another, due to the distinct nature of each segment. Instead, this paper will focus on a single segment of road.

But which road segment to choose? There are 470 options given in the original dataset, but the best one would have a long length and a high volume of traffic flow. For that reason, this paper will focus on the eastbound road segment between Ben White/FM 973 and Ben White/Riverside. It has a very long length, at 3.63 miles. It also is a main road, and is expected to have high traffic volume.

A picture containing text, outdoor, map, sitting

Description automatically generated

Because the data is available from both an online API and a .csv file, it before proceeding forward, it would be best to check and make sure the data is identical between the datasets. I plotted the following in python and found that the data was in fact identical.

I also decided that I would be using the Traffic Match Summary Records (TMSR) dataset. This has been filtered to include data from actual vehicles, but it is also nicely split into 15-minute increments for each day.

A screenshot of a cell phone

Description automatically generated

After deciding which data source I would be using, along with which table to use, I saved the .csv file into a database on my local machine. This made it easier to query, without needing to load the entire dataset (47,385,655 records) into a dataframe.

From a Jupyter Notebook, I then ran the following query, and saved the results into a dataframe.

SELECT \* FROM atx\_traffic.tmsr WHERE origin\_reader\_identifier = 'fm973\_tx71' AND destination\_reader\_identifier = 'benwhite\_riverside';

This returned just the travel information that is pertinent to the segment that I am interested in. I then decided to drop fields from this dataframe that were not going to be useful to me.

df = data[['segment\_length\_miles','timestamp','average\_travel\_time\_seconds','average\_speed\_mph','number\_samples','standard\_deviation']]

The timestamp column was then converted from a string to a datetime. I also saved the time and weekday as separate fields, since I expect that traffic will behave differently at different times of day, and on different days of the week.

df['timestamp'] = pd.to\_datetime(df['timestamp'],infer\_datetime\_format=True)  
df['time'] = df['timestamp'].dt.time  
df['weekday'] = df['timestamp'].dt.weekday

With the columns how I wanted them, I then observed that many of the values for average\_speed\_mph were -1. I then removed these records from the dataframe, as they were indicative of an invalid entry.

df = df[df.average\_speed\_mph != -1]

To make some calculations easier down the line, I also computed some intermediate fields.

df['totalspeed'] = df.average\_speed\_mph \* df.number\_samples  
df['totaltraveltime'] = df.average\_travel\_time\_seconds \* df.number\_samples  
df['pooled\_variance'] = (df.number\_samples - 1) \* (df.standard\_deviation) \*\* 2

With that, the data has been cleaned, but I also found myself wanting to create a summary table that would better suit my purposes. This resulting dataframe still had more than 15,000 records, which is a bit much for analysis in my opinion. Also, it likely will be imbalanced in some ways. For example, there will likely be more records present from 4:15 pm than those at 3:00 am. Therefore, I decided to create a summary table that contained aggregate data for a given time of day.

Therefore, I needed to loop through each possible time of day and generate summary data for that given time of day. I was also careful to not mix the summary statistics for weekdays with weekends, since I wanted to compare these two scenarios.

This summary table contained:

* Time of Day
* Weekday Mean Number of Samples
* Weekend Mean Number of Samples
* Weekday Mean Speed
* Weekend Mean Speed
* Weekday Mean Travel Time
* Weekend Mean Travel Time
* Weekday Speed Pooled Standard Deviation
* Weekend Speed Pooled Standard Deviation

This summary table serves as the backbone for the analysis. I exported it as a .csv file, so that it could be analyzed using R. I find that data cleaning and preparation is far easier in Python 3, but the analysis portion is better suited to being performed in R.

# Results

Once I got the data into R, I decided to try and visualize some of the data that is present in the summary table. Given that the data was split into different columns depending on if the data was from a weekend or a weekday, I decided to include both of them on each graph, so that we could compare both of the datasets. So let's start by comparing the traffic volume on this given segment of road.

A close up of a map

Description automatically generated

This plot tells us quite a bit about the number of cars that are traveling down the road. There are three things that are worth noting.

The most significant observation here is that weekends and weekdays have very similar traffic volumes from about 5:00 in the afternoon until 10:00 the following morning. The lines here seem to change in parallel, with about 7.5 vehicles being observed in each 15 minute chunk, starting at 5:00 pm and lasting until midnight. The traffic flow then will decrease gradually the following day, until about 10:00 am, where we see that only one vehicle was being observed every 15 minutes.

The second thing to note here is the divergence that takes place between weekend and weekday traffic flow from 10:00 am to 5:00. On weekdays, we can see a drastic increase in traffic flow, which appears to be a linear increase. It will then taper off, peaking at 9 observed vehicles in each 15 minute window. This peak takes place at about 2:00 pm. During this time, the weekend traffic flow is also increasing, but at a much slower rate. From the hours of 2:00 pm to 5:00 pm, we see that weekday traffic flow is decreasing, while weekend traffic flow is still increasing, and has actually accelerated. During this time is the only time where we can see that traffic flow is truly different.

The final thing to notice about this graph is variance of the data from 5:00 pm until midnight. While traffic flow for the remainder of the day is fairly predictable, we can see that, at least on this section of the road, the traffic volume is far messier.

Now that we have analyzed the traffic volume, we can move onto the traffic speed analysis. This is how the traffic volume is impacting travel times. Here is the graph of that information. Again, the weekday information is shown in red, and the weekend information is shown in blue. I have plotted the data and fit a line to both weekday and weekend data.

A close up of a map

Description automatically generatedThis graph tells a story that is more descriptive of what a trip down this road segment will be like. We can make models of traffic volume, but unless that increase in volume leads to decreased speed and increased travel times, a driver may not care.

There are two points of interest on this chart. The first of which is that from 10:00 am until midnight, the travel times for a weekday are *consistently* higher than those for a weekend. Travel times on a weekend are typically around 300 seconds, regardless of the time of day. Pas 10:00 am, however, the weekday travel times hover around 400 seconds.

The second point of note here is that starting at 10:00 am, however, we can see that the weekday travel times increase, peaking at about 450 seconds at 2:00 pm. Travel times then drop back down to 350, before finally rising around midnight.

A close up of a map

Description automatically generatedA close up of a map

Description automatically generatedI also did some analysis for the speed and pooled standard deviation, but I was unable to find any insights about how traffic flow changes over the course of a day that were not already present on the previous two charts. This makes sense because speed would be a scalar multiple of travel time, given that the length of this road segment is unchanging. For reference, however, here are the plots for speed and pooled standard deviation of speed as well.

# Discussion/Conclusion

Coming into this project, I expected to see a difference in traffic flow between weekdays and weekends. I was unsurprised, yet delighted when I was able to observe this difference. Additionally, I expected to see a spike in traffic volume at about 5:00 pm, but I was surprised to see that the spike came much earlier, at about 2:00 pm. I do not think that the data was mislabeled, but maybe there is time difference. Perhaps the data was not in local time, or there really is a spike in traffic volume at 2:00 pm.

I was excited to see that there are distinct sections of a day where traffic flow will be different, which seems to lend itself to a more complex model. Perhaps this would have been a good use case for a piecewise model.

I like the approach I took because it could, in theory, be reapplied to each and every road segment in the area. Perhaps some future research could be done to extend the modeling to a more general flow of traffic through the city.

# Acknowledgments

Thank you to Bellevue University for the education opportunity. I also extend a gracious "Thank you" to my professor, Dr. Fadi Alsaleem. The content this term was very helpful in steering me in (what I belive to be) the correct direction. Finally, I would like to thank my peers for their incredibly valuable feedback. It was invaluable for the learning process.

# References

“Bluetooth Travel Sensors - Individual Address Files (IAFs) | Open Data | City of Austin Texas.” Austin, https://data.austintexas.gov/Transportation-and-Mobility/Bluetooth-Travel-Sensors-Individual-Address-Files-/qnpj-zrb9/data. Accessed 7 Aug. 2020.

“Bluetooth Travel Sensors - Individual Traffic Match Files (ITMF) | Open Data | City of Austin Texas.” Austin, https://data.austintexas.gov/Transportation-and-Mobility/Bluetooth-Travel-Sensors-Individual-Traffic-Match-/x44q-icha/data. Accessed 7 Aug. 2020.

Bluetooth Travel Sensors -Traffic Match Summary Records (TMSR) | Open Data | City of Austin Texas. https://data.austintexas.gov/Transportation-and-Mobility/Bluetooth-Travel-Sensors-Traffic-Match-Summary-Rec/v7zg-5jg9. Accessed 7 Aug. 2020.

“Cityofaustin/Hack-the-Traffic.” GitHub, https://github.com/cityofaustin/hack-the-traffic. Accessed 7 Aug. 2020.

Hall, Katie. “Austin Traffic Worsens, Now Ranking 14th Most Congested City in Nation.” Austin American-Statesman, https://www.statesman.com/news/20190822/austin-traffic-worsens-now-ranking-14th-most-congested-city-in-nation. Accessed 7 Aug. 2020.