Bayesian Methods Final Project: DC Metro Transit Times

Team Members: Chi Cheung, Samuel Kupfer, Neeraj Walia

I. Introduction

For years, the DC Metro system was plagued by deferred maintenance on its rails. In a recent report by the Washington Post, riders indicated that reliability, cost, safety, and cleanliness were among factors important to their daily commute. To see how well WMATA was doing a year after completing their Safe Track repairs program, our team put their reliability to test on a very personal level. First, we wanted to measure from our work locations, what was a given train's probability of arriving at Foggy Bottom-GWU station "on time" for class (as defined below)? Without a published train schedule, riders have a difficult time determining when they need to be at the station and are left guessing when they would have to leave work. For the commuter, this could be costly in loss of time, as riders of the New York City Subway have experienced - often arriving early, hours before their workplace opening its doors.

Similarly, this process can easily be applied to other Metro riders for their daily commute. Riders have a way to objectively decide for themselves, how much progress Metro is making in improving their reliability and timeliness. Equally applicable, riders can use the data to make informed decision, such as the point where it would be better to drive to work than take the Metro, the impact of a job change or household move on their commute, or arrange their working hours to minimize the time spent in transit.

II. About the Data Set

Our team used a live data feed from the WMATA Live Train Position API. We were unable to locate suitable historic data or query just the dataset we were interested in, therefore we had to use the API. The data collection for this project began June 10, 2018 and ended June 23, 2018. The API returned uniquely identifiable trains in service and station locations giving us the ability to determine the arrival and departure times. This amounted to 6 train lines serving 92 stations on average 18.8 hours a day. Real-time train position data was queried at a rate of once every 10 seconds. The size of the data for 13 days of data (before preprocessing and filtering) exceeded 1.5Gb of storage space. This period alone contained over 3 million rows of data. To handle the query, we collected the data on a storage drive hosted on Amazon Web Services.

For the data that we needed to work with our Bayesian model, we spent considerable time processing the initial dataset. We used Python to identify and keep only the subsets needed for our project. We examined the trains traveling from the L'Enfant Plaza station to the Foggy Bottom-GWU station between 3PM and 5PM. Trains arriving at Foggy Bottom-GWU between 4:50PM and 5:00PM were labelled as "on time", trains arriving before 4:50 or after 5:00 were labelled as "not on time". The same code could be modified to isolate the same type of data for any two stations. After filtering the dataset for only the L'Enfant Plaza to Foggy Bottom-GWU trips in the above timeframe, we were left with 538 rows (trips).

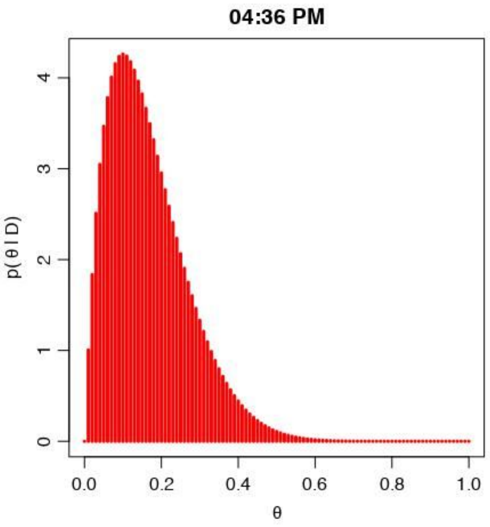
The dataset was further binned into 3 minute groups. This appeared to provide the most informative results with the wide range of possible arrival times and the the limited data available. With more data, we could possibly have binned the times into 1-minute intervals. We chose to use the transit direction between L'Enfant and Foggy Bottom-GWU because of the frequency of arrivals operating between the two stations using three different train lines, orange, silver, and blue, all on the same track.

III. The Models

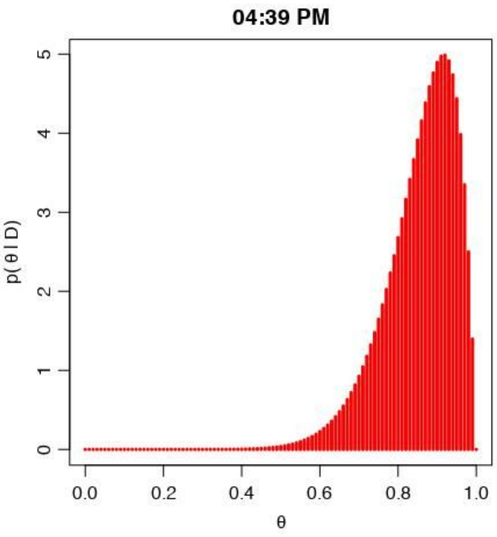
The project entailed demonstration of two models based on Bayes rule. First was the use of formal analysis to determine the probability of arriving on time between two stations for each departure time interval. Using the formal analysis method with Bayes' rule, we treated the dataset as a binary problem with a beta distribution as the prior to determine the posterior probability. This dataset was relatively small (538 rows of data) allowing formal analysis to work well without compromising processing speed. We applied formal analysis, first using a uniform beta distribution for each time interval. Then, we repeated this analysis using more informed priors (also beta distributions) based on our anecdotal experience on Metro. We compared the results (posteriors) generated by these two sets of priors. After completing our formal analysis, we used Markov Chain Monte Carlo (MCMC) to determine the difference between this probability for two time intervals (4:36-4:39 PM and 4:39-4:42 PM) and the statistical significance of this difference.

III. Results

We used Bayes' rule to perform a separate formal analysis on each 3-minute time interval. We were able to do this because we used beta distributions for our priors, which led to beta posterior distributions. First, we tried using uniform prior distributions (beta distributions with α=1 and ϐ=1) for each 3-minute interval. The posteriors for two time intervals (4:36-4:39 PM and 4:39-4:42 PM) are shown below. Also shown is the number of on-time trips (z) and total trips (N) in our data for each interval.

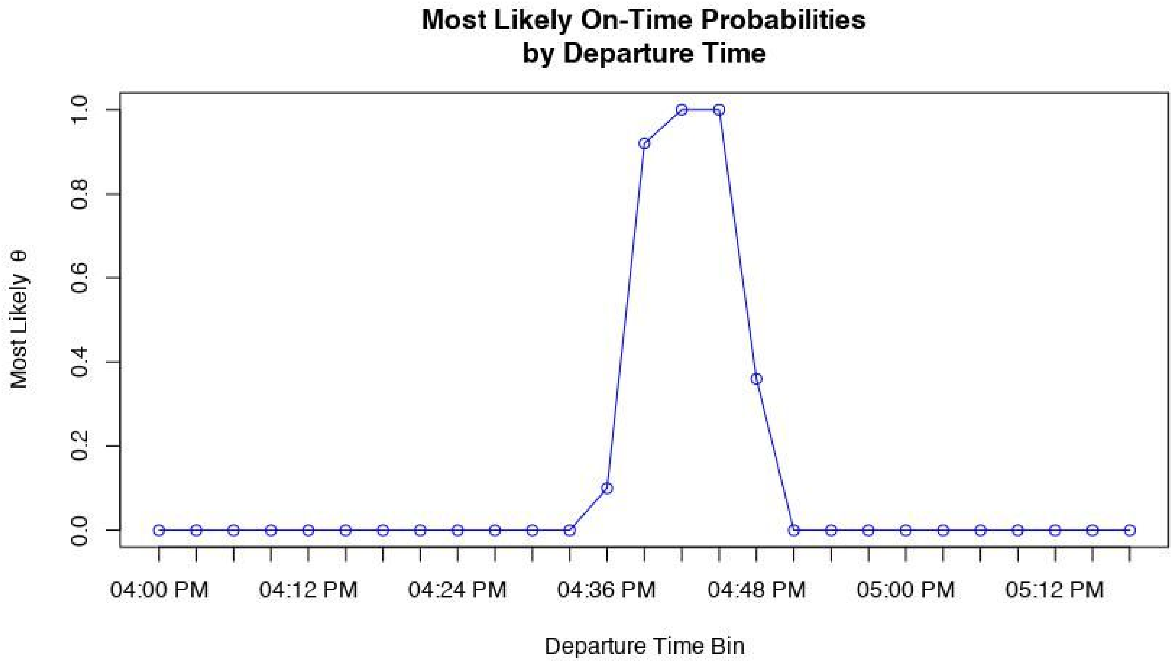


(z=1, N=10)

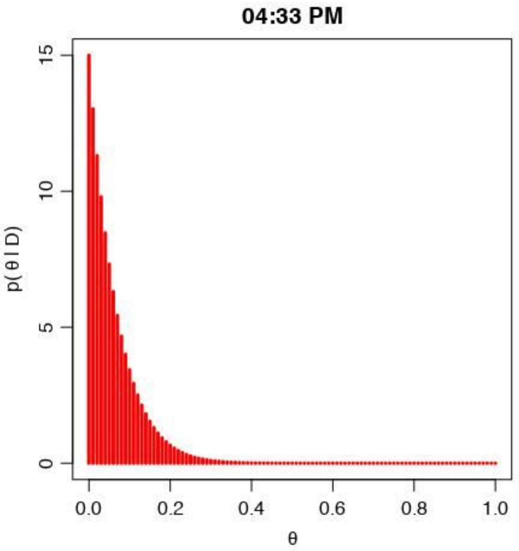


(z=11, N=12)

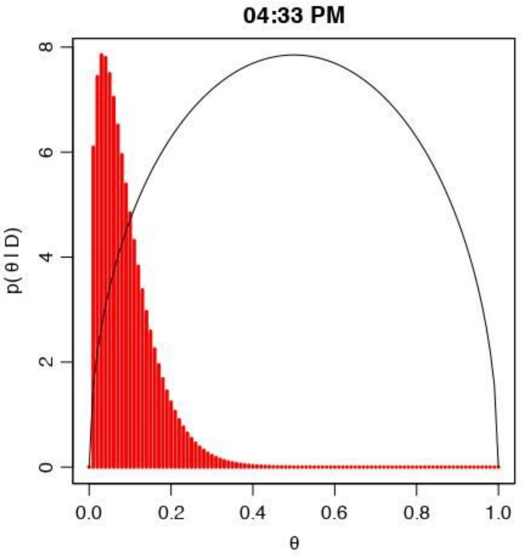
We summarized these results by taking the Ө of the highest probability density in each time interval's posterior and plotting them, showing the relationship between departure time and chances of on-time arrival.



Next, we wanted to see how these results would be different with more informed prior distributions - a few different priors for the different time intervals - rather than simply using uniform priors for every time interval. We used a few different beta distributions for different time intervals, based only on our anecdotal experience taking the orange/blue/silver Metro lines. See our code for more details on which beta distributions we used for each interval. Shown below are two posterior distributions for the same time interval - the first used the uniform priors, and the second used our more “informed” prior distribution (this prior is shown in black). In this time interval, there were 14 trips, none of which were on time. Note the different shapes and peaks of the two posterior distributions.

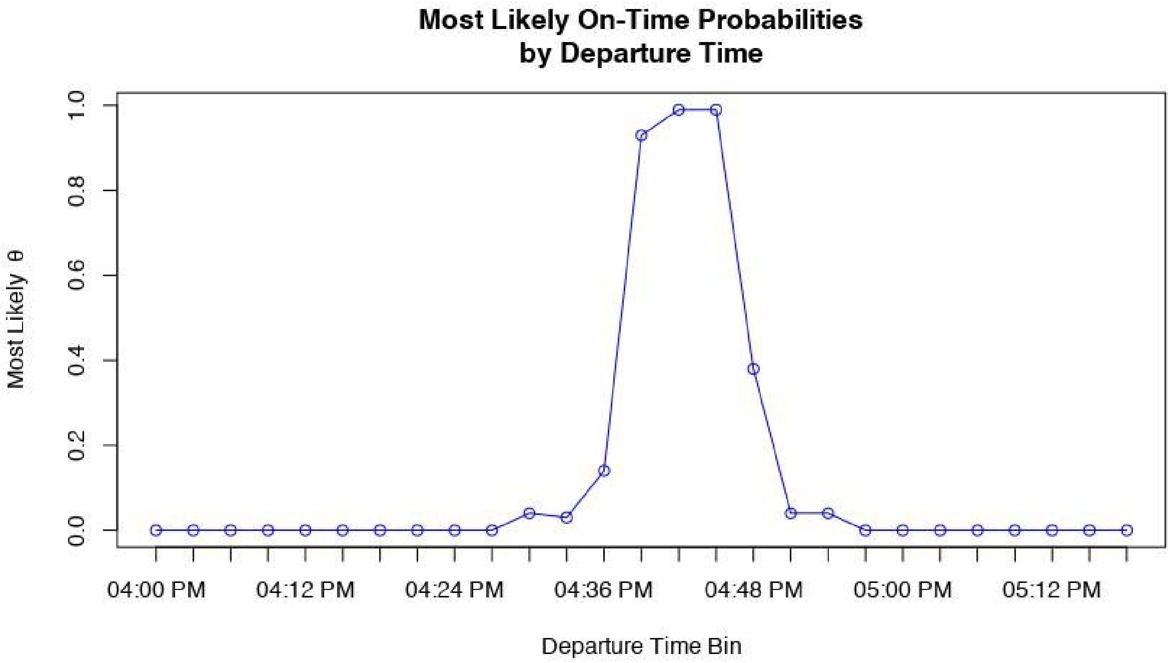


Posterior using uniform prior

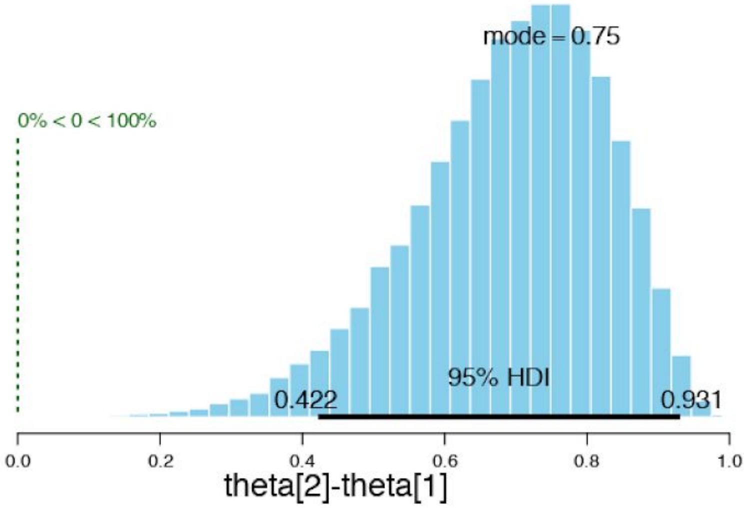


Posterior using “informed” prior of beta(α=1.5, ϐ=1.5)

And, the same plot of the most likely Θs for each time interval, for the analysis using our more informed priors. Note how this plot is “wider” than the one above because there are some intervals, like the one starting at 4:33 PM shown above, that had 0 on-time trips but were “pulled up” by their priors.



Finally, we used MCMC to compare the posterior distributions of two different time intervals (here, theta[2] is the 4:36-4:39 PM interval and theta[1] is 4:39-4:42 PM). This shows that by departing 3 minutes later, chances of arriving “on time” increase by 75%. Also, note that 0.0 is well outside of the 95% HDI, showing that the difference in on time probability is statistically significant.



IV. Conclusion

Our models had some inherent limitations that made the results interesting. While we attempted to come up with informed prior distributions, we could only base them on our own anecdotal experience. Our dataset was too small and it did not cover various scenarios like train breakdowns, track issues, maintenance, etc.

Using Bayes' rule, we were able to determine the departure times from L'Enfant Metro Plaza that are best suited to arrive on time for class. And, we were able to use MCMC to show that, despite our relatively small dataset, the differences in probabilities between the different time intervals was statistically significant. We believe with more data and better-informed priors we could get a wider graph of probability vs. departure time that is closer to the ground truth.

Our data did not have any outliers (other than some that were so far out that they were clearly data errors - these were filtered out in preprocessing). This could be because of the small data set for two weeks. We cannot judge reliability of the trains based on this data. If we had several months of data that could cover different variations, we would be able to better analyze the reliability.