Project Progress Report: Analyzation of Traffic Patterns in Relation to Delays in Boston Public Transit

Cristian Mendivil
crme7282@colorado.edu

Lucas Lyon
luly2738@colorado.edu

Vamshi Arugonda
vaar2387@colorado.edu

Theodore Freeman
thfr5770@colorado.edu

ABSTRACT

This paper explains the process we will use to find interesting patterns related to traffic in Boston by using a dataset from Uber and from Massachusetts Bay Transit Authority.

ACM Reference Format:
Cristian Mendivil, Lucas Lyon, Vamshi Arugonda, and Theodore Freeman. 2018. Project Progress Report: Analyzation of Traffic Patterns in Relation to Delays in Boston Public Transit. In *Proceedings of ACM Conference (Conference'17).* ACM, New York, NY, USA, Article 4, 6 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

The authors of this paper will layout the methods and sources of data for an analysis of traffic, ride sharing, and public transportation in the Boston metro area. Data will be sources from 2016 to present, and most of the data will serve to provide a baseline for normal travel times between various sectors of the city. This paper will go into the application of knowledge gained, previous work that has been done on the data, details about the data sets and where to download them, methods of evaluation, data mining tools, temporal-based milestones, and a summary of peer review that was received in class.

2 PROBLEM STATEMENT/MOTIVATION

Ride-sharing on Uber and Lyft has exploded in popularity in the past decade. The growth of these companies has far surpassed legislators' abilities to regulate them and city planners' abilities to design cities around the services. Public transit has been a staple of Boston since 1947, with the creation of the Massachusetts Bay Transit Authority (MBTA) [1]. We aim to use Uber travel times in Boston as well as data provided by the MBTA on the reliability of rail, subway, and bus routes to establish a link between delays in MBTA service and an increase in traffic times. With this link, we then hope to train an algorithm to predict delays in traffic given delays in public transit.

The algorithm we train could be used for many things, including better city planning. Specifically, there may be certain routes that are resistant to delays in traffic. We can identify those routes and try to emulate characteristics of those routes in future roadways. We could also use our algorithm to help the city of Boston plan for and mitigate delays in public transit during natural disasters or major city-wide events.

The algorithm could further be applied to a commercial market in which the prediction of delays could help to improve route mapping for those who make deliveries or use the road for their main mode of business. While other traffic prediction algorithms/applications focus on time delays in general based on day-to-day patterns, our algorithm focuses on the correlation between public transport delays and traffic times.

3 LITERATURE SURVEY

There has been quite a bit of prior work about traffic prediction and analysis. One such study was done by a group of researchers at the University of Southern California with a goal of accurately predicting and quantifying impact of traffic incidents. This is a pretty good study as in their conclusion they claim that their model can increase "prediction accuracy of baseline approaches by up to 45% [1]" for the impact of traffic incidents on road networks. We have still yet to read through the whole paper, but get the feeling that it will be a valuable source for inspiration for where we can take our project, and has it is own references and prior work which we can also look through and possibly utilize. Another group did some work on developing a support system for using real time bus location data to accurately estimate arrival times. This study may be useful considering that all the data we intend on using is public transit data or Uber. Perhaps it can give us some ideas of how to use our public transit data in a cleverer way. A way in which our project will be different from the described research above is in a couple of ways. First, the most recent of these projects was done in 2016 so there is potential at least to have more currently relevant results. Second our work is going to try to learn how individual traffic delays affect city-wide transit rather than just providing time estimates for when the next bus will arrive or route prediction for obstruction avoidance. Both projects may be useful to us though by providing different ideas for how to use and view our data as well as what we might avoid. If we find that we are getting stuck in a corner though and neither of these are able to help get us out it seems there is plenty of other research out there which if we searched for we may be able to find our answers. [2][4]. [1] [2]

Conference'17, July 2017, Washington, DC, USA

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

[1] B. Pan, U. Demiryurek, C. Shahabi, and C. Gupta, âĂIJForecasting Spatiotemporal Impact of Traffic Incidents on Road Networks,âĂİ in 2013 IEEE 13th International Conference on Data Mining, 2013, pp. 587âĂŞ596.

[2] F. Sun, Y. Pan, J. White, and A. Dubey, âĂIJReal-Time and Predictive Analytics for Smart Public Transportation Decision Support System,âĂİ in 2016 IEEE International Conference on Smart Computing (SMARTCOMP), 2016, pp. 1âĂŞ8.

4 PROPOSED WORK

The data sets that we are using are not in perfect form to be used during the project. We have to modify some of its data to start our project. The Uber dataset have some noisy values which we will have to clean. The data set also have some rows with incomplete data that does not give us any good information because it is information from trips that were canceled. We will delete all the rows with incomplete information which will help us to reduce our data to speed up the process of finding patterns. We will use the data to create a baseline to predict if a trip in Uber will have some delay. We intend on creating this baseline by taking averages of transit times for each pair of source and destination ids in our dataset.

The MBTA dataset also have some data that we will have to clean. For example, there are some public transit rides that does not have a route. We will remove entries without routes as they will not be useful for our work. Another issue we discovered is that the data comprised of rail line entries have metrics that are percent on time or passenger wait time. These measures donâĂŹt tell us anything useful for our study, so we will remove all the rail entries, and attributes which were necessary before to differentiate between rail and bus entries that are no longer needed.

We will need to find a map of the us census tracts for Boston which include area ids in order to have a method for comparing MTBA and Uber data. If this is unavailable we may need to instead reach out to Uber about finding what areas are which. After this, the MBTA datasets will be combined with the Uber baselines to see if there is any correlation between delays in public transit rides and delays in Uber.

See figure 3 for a visual representation of the methods that we will use during the prepossessing part for our project

5 DATA SET

Our team will utilize two data sets: one from Uber, and one from MBTA. The Uber dataset must be downloaded in quarter-year increments from https://movement.uber.com/. The MBTA dataset can be downloaded in one file from their dashboard at *http://www.mbtabackontrack.com/performance/index.html#/download*. You must select the radio box "Reliability". Our team had difficulty downloading the entire dataset at once, and had to split the download into three time frames: January 1st 2016 - January 1st 2017, January 2nd 2017 - January 1st 2018, and January 2nd 2018 - March 5th 2018.

The MBTA Reliability dataset has 395,130 rows and 9 attributes. The attributes include service date and time, whether the row is for Off-Peak service or Peak service times, the type of transport (including rail, commuter, and bus), the route line, stop station, metric measured (including Passenger Wait Time and Schedule Adherence), and varying numerators and denominators for those metrics.

We will primarily use the service date and time, type of transport, stop station, metric type, and numerator and denominator attributes. The Peak vs Off-Peak hours attribute is not very helpful for establishing a link between delays in public transit and traffic, but will be helpful to establish two baselines for normal wait times, one during peak hours and one during off-peak hours.

The Uber Movement dataset is split into 7 distinct .csv files, each containing 3 months worth of travel times between every Uber-defined source and destination in Boston. Each file has 7 attributes: sourceid, dstid, hod, meanTravelTime, standardDeviationTravelTime, geometricMeanTravelTime, and geometricStandardDeviationTravelTime. Every attribute in the file will be useful for building a model of baseline travel times between different sectors of the city.

6 EVALUATION METHODS

The data sets will be combined to see any interesting pattern between public transit rides and Uber delays. We will use correlation analysis and multivariate linear regression to show whether or not there is a correlation between the datasets.

For correlation analysis, we will use the delay time for public transit rides and Uber to see if the delays have some correlation.

We will use multivariate linear regression to check public transit rides and Uber trips are delay depending on different factors such as time, or day of the week.

7 TOOLS

(i) **Python:** We will conduct the majority of our programming in Python 3, through Python we will be able to organize and utilize our data to aid our research.
(ii) **Spyder/Jupyter NoteBook:** Our team will use one (or both) of these tools to implement our code and work with the data. Spyder is an excellent tool because it acts as a text editor and interpretor, also Spyder has numerous built in libraries that will help us organize our data. Jupyter has the benefit of running blocks of code at a time in order to debug more efficiently.
(iii) **Pandas/NumPy:** Both of these libraries will act as data organization and analysis tools. Pandas will be our primary data frame which we will clean and organize to create visuals and compare our datasets. NumPy will act as our primary tool for performing numerical computations.
(iv) **Matplotlib:** A Python library that our team will use to generate graphs/tables.
(v) **SciPy:** A machine learning python tool which we will use to perform computations and regression analyses.
(vi) **SciKit Learn:** A python machine learning library which we may use to build a model for predicting traffic delay impacts.

8 MILESTONES

See Figure 2

8.1 Milestones Completed

Research attributes in MBTA dataset:

The MBTA dataset had many different attributes related to the different types of transportation. We did some research to know more about how the attributes could be use in our projects. We found that most of the attributes were related to only one type of transportation, and we use that information to help us clean and reduce the data set. The research was done in the MBTA website where we found a file containing details for all the attributes. We

concluded that some of the attributes in the dataset were not important in our project, and that we would delete rows in the datasets containing those attributes.

Data cleaning and streamlining for both Uber and MBTA:

The original MTBA data was broken up into smaller spans of time so that it could be downloaded. In order to get a complete data set again we concatenated the pandas data frames together from what we read in from the individual sets. Then, we looked at what kind of attributes we had in our data set to evaluate which would be useful and what was excess. We came to the conclusion that the data from rail lines were not going to be useful for us as the method provided for evaluating rail traffic wouldn't work.

We also were able to download all of the Uber data which had to similarly be put together. This data was also cleaned and reduced to get rid of what we felt were un-needed attributes and useless entries.

Establishing baselines:

As we have planned, work has been done to begin forming a baseline from our Uber data. This has been a very resource intensive step in our project as we have needed to find average transit times between all pairs of census tracts zones which have any data. Although running the code for this took hours to complete, it is the first half of the work we need to produce the initial results we're looking for.

8.2 Milestones Todo

- Figure out what census tract zones have which id's in the uber dataset so we can map them with the bus route information.
- Come up with an average travel time for bus routes that have the same start and stop location ids as the Uber set.
- Perform initial correlation analysis.
- Possibly revise data set with outlires in mind.
- Revise initial results with respect to time of day.
- Perform further correlation analysis.
- Perform Multivariate Linear Regression.
- Potentially build a model for predicting impact of delays for citywide traffic.

9 RESULTS SO FAR

Currently we have no results to report. A lot of our time has so far gone into getting our data organized and getting our baseline from the Uber data. Calculating the baseline has been an exorbitantly resource-consuming endeavor in both time and space complexity, once this is computed we will begin comparisons with the MTBA dataset.

10 GRAPHICS

***Figure 1: Correlation Coefficient Graphs**[3]*

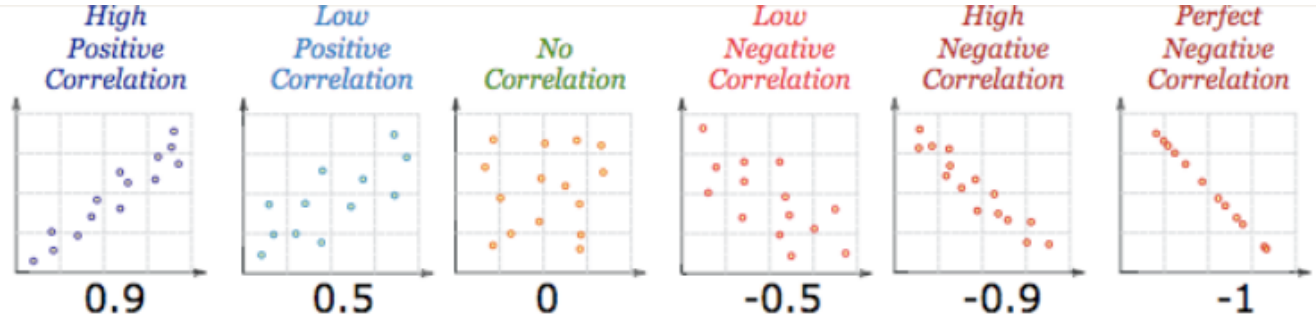


Figure 2: Milestones and Plan

Task	Date								
	6-Mar	13-Mar	20-Mar	27-Mar	3-Apr	10-Apr	17-Apr	24-Apr	1-May
Proposal Submission									
Data cleaning and streamlining for both Uber and MBTA									
Analyze data for patterns and correlations between public transit delays									
Progress Report									
Create final analyses, visualizations, and begin final writeup									
Final Project Submission									

Figure 3: Preprocessing Data

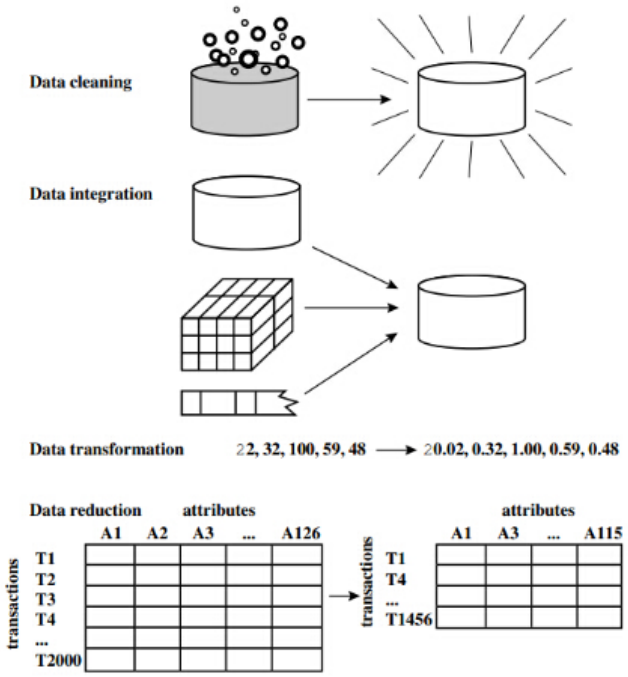
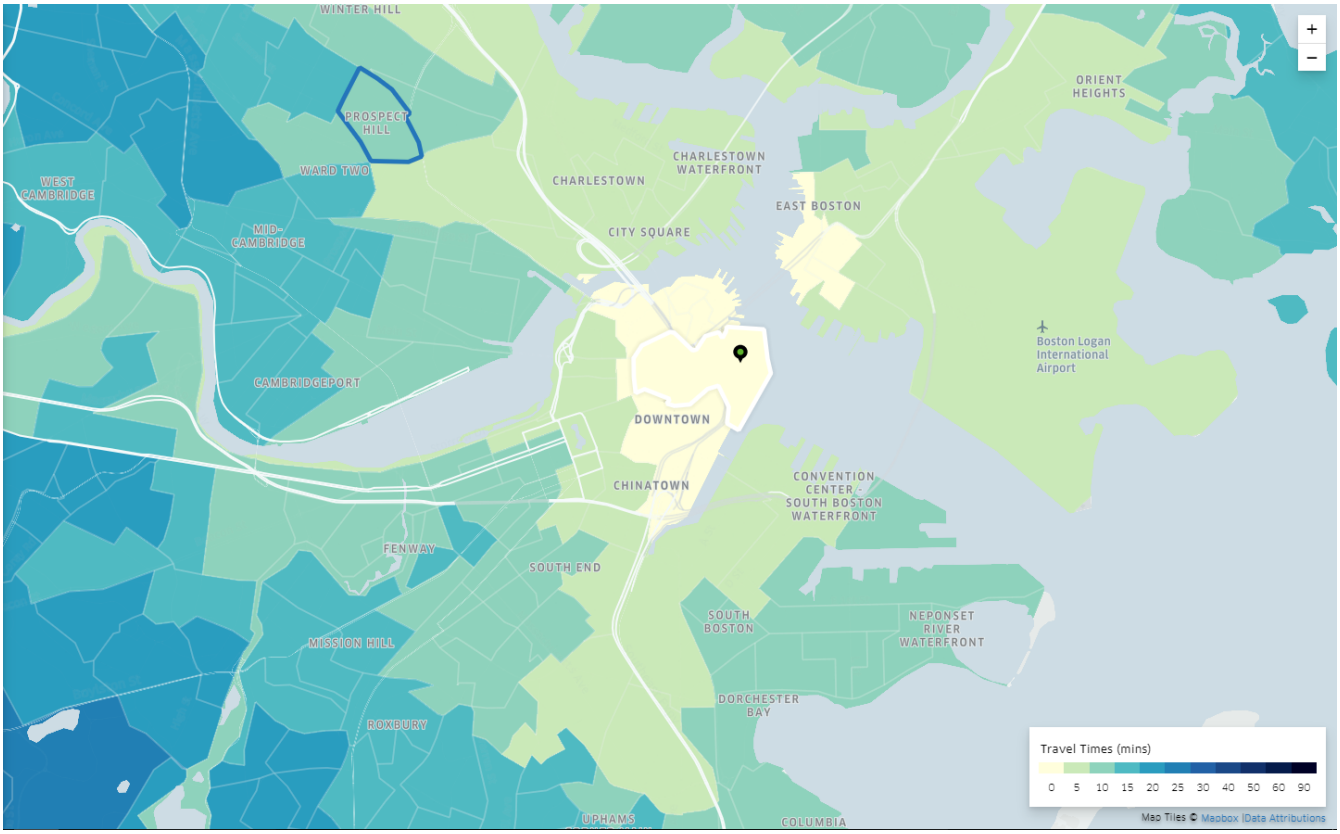


Figure 4: Uber data as shown on their website



A SECTIONS

A.1 Introduction

A.2 Problem Statement/Motivation

A.3 Literature Survey

A.4 Proposed Work

A.5 Data Set

A.6 Evaluation Methods

A.7 Tools

A.8 Milestones

A.9 Results So Far

A.10 Graphics

A.11 References

REFERENCES

[1] 2018. Massachusetts Bay Transportation Authority. (Mar 2018). https://en.wikipedia.org/wiki/Massachusetts_Bay_Transportation_Authority
[2] C. Shahabi B. Pan, U. Demiryurek and C. Gupta. [n. d.]. .
[3] Boese Elizabeth. 2018. Chapter 3 - Preprocessing. (2018).
[4] J. White F. Sun, Y. Pan and A. Dubey. [n. d.]. .