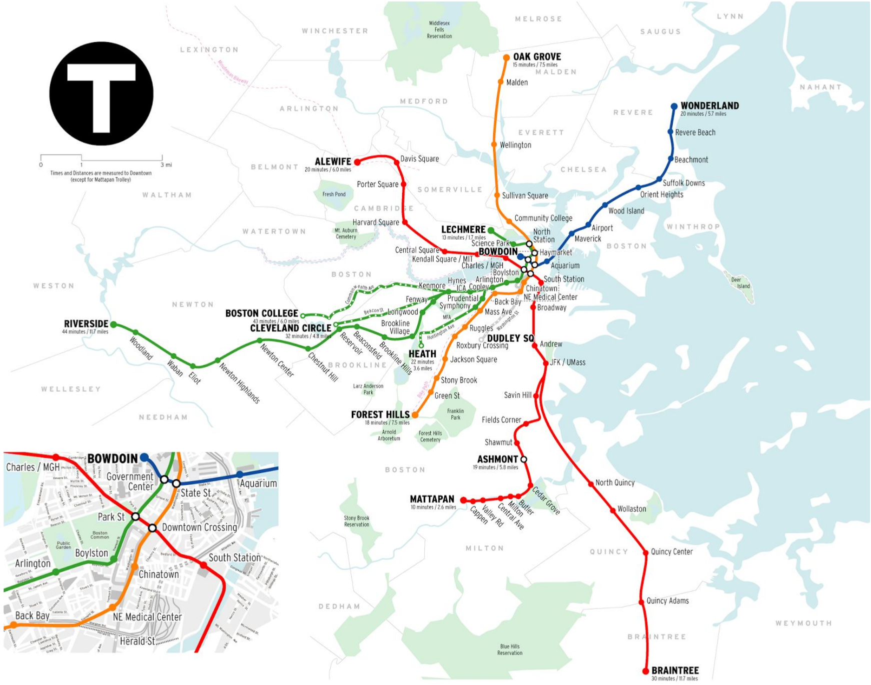
Relation of Boston Neighborhoods to MBTA Reliability and the Demographics of Surrounding Neighborhoods

Group Members: Sameena Bajwa, Shreya Pandit, Rudhra Raveendran, Nathan Weinberg

Introduction

In large cities all around the world, many residents rely on public transportation to get to and from work. The City of Boston is no different. For residents here, public transportation means the use of Massachusetts Bay Transit Authority (MBTA or "the T") services. The MBTA offers an incredibly diverse set of transportation options for its customers, including bus services, light-rail, heavy-rail, commuter rail, and even boating options.

The MBTA's transportation network services the entire Greater Boston Area, as far out as Providence, RI, but for our purposes, we will be focused specifically on neighborhoods within and surrounding the City of Boston itself and only on the MBTA's bus, rapid-transit, and commuter rail services.



Geographically-accurate map of MBTA rapid-transit services

The City of Boston hosts a number of very diverse and distinct neighborhoods. Due to a number of demographic factors such as socioeconomic status, some of these neighborhoods are more reliant on MBTA services and other. With this in mind our project does a two-fold analysis:

First, we focus on exactly "how" these neighborhoods contribute to the reliability of public transportation in Boston. Regression analysis on the reliability of public transport routes along with the neighborhoods they pass through, gives us a way to *quantify* how much each neighborhood contributes to the overall reliability metric.

Second, we attempt to see how tightly geographical proximity among neighborhoods is linked to the spread of this demographic across neighborhoods. For example, one way to think about this can be: Given a neighborhood A, how does geographical proximity influence the spread or trickling of poverty from neighborhood A to its surrounding neighborhoods? We do this by analyzing a distance matrix between neighborhoods and computing its correlation with various demographic data.

Data Sets - Background and Origins

Means of Commuting, Poverty Rates, and Household Income

Names: nathansw_sbajwa.commuting, nathansw_sbajwa.povertyrates, nathansw_sbajwa.householdincome
In order to analyze the experiences and challenges of residents in Boston in regards to their use of public transportation, we had to collect some background information on the different demographic areas. This data was found in a publicly accessible PDF online created by the Research division of the Boston Redevelopment Authority. The BRA created this PDF, entitled "Boston in Context: Neighborhoods," by collecting data from the U.S. Census Bureau's 5-Year American Community Survey, ranging from 2010-2014. It was very difficult to find data sets pertaining to the Greater Boston area that also show the breakdowns by neighborhood, which is why these data sets were especially important to us.

For our final project, we used three data sets from this PDF: Means of Commuting, Poverty Rates, and Household Income. All three data sets were converted to CSV files with the help of online conversion tools, before being converted to JSON. The reason for the intermediate step was so that we could manually edit the data where necessary. The manual edits consisted of ensuring that column headers were unique. Once the CSV files were ready, they were converted to JSON and added to the Data Mechanics file uploader, where our scripts could access them. The PDF can be obtained at this link: https://www.bostonplans.org/getattachment/7b9b1201-8b4f-4fa9-b0f2-4acbbe083198.

MBTA Performance Data

Name: nathansw_rooday_sbajwa_shreyap.MBTAPerformance
Historic performance data for the MBTA is provided online by the MBTA Performance Dashboard website. The site offers different metrics for analyzing performance data, such as reliability, ridership, financials, and customer satisfaction. For this project, we found reliability

to be the most useful metric for us. We downloaded historic data ranging from 11/1/16 - 6/3/17 as a large Excel file with 129,413 rows. The performance data can be accessed here: http://www.mbtabackontrack.com/performance/index.html#/download

Changes:
Because of the large nature of the data set, a Python script needed to be written in order to restructure the data and convert the file to JSON. The Excel file was organized in such a way where every row included a date and then information regarding the performance of a route on that particular date. Assigning the dates to be keys would not work, as there were hundreds of observations for each date. Therefore, the script created a nested JSON file; each key was the date and its value was a nested JSON object, where the key within those objects were the row numbers (from the Excel file) and the corresponding values came from that row. This way, there was no chance of overwriting any data, as every row had its own unique place within the JSON file.

On-Time Performance by Line

Name: nathansw_rooday_sbajwa_shreyap.OTP_by_line
This data set was created in OTP_by_line.py by transforming and aggregating the previously collected MBTA Performance data. The script parses the performance data and keeps track of the on time performance values for each line. Furthermore, it also keeps track of whether the OTP values are being reported during peak or off-peak service times. The final JSON file is formatted such that each key is the name/ID of a line (ex: 57 or Green Line B) and the values are the OTP averages calculated for both peak and off-peak service times.

Trickling

Name: nathansw_rooday_sbajwa_shreyap.trickling
The trickling data set was manually created by using each neighborhood's latitude and longitude coordinates that are available online. With this information, we were able to calculate the distances between each neighborhood. Each key in this data set is a neighborhood and the values are a nested dictionary, where the keys are every neighborhood and the corresponding values are the distance values.

Neighborhood Map

Name: nathansw_rooday_sbajwa_shreyap.neighborhoodMap
We manually created this data set based off of the information in the trickling data set file. The keys are the neighborhoods in Boston and the values are the neighborhoods that are closest to the key. The values are chosen based off of the distance values calculated in trickling.

Stops vs. Lines

Name: nathansw_rooday_sbajwa_shreyap.stopsVslines

This data set was created using OTP_by_line (described above) and the MBTA real-time data API. The script parses OTP_by_line to grab all the keys (i.e. lines) and their OTP peak-service values. If necessary, the line names are changed so that they follow the syntax necessary for the MBTA API. The MBTA API is then used to generate all the IDs for the stops that each line passes through (using their stopsbyroute query). The stop IDs returned were appended to the values of each key in the final JSON file.

The format is as follows: {line1: [Peak Average OTP, stop1, stop2, stop3,...]}

Stops

Name: nathansw_rooday_sbajwa_shreyap.stops

This data set was created by utilizing both the MBTA real-time data API and online resources. The API was called to grab every stop that the bus, light rail, and commuter rail passes through, and the stop's associated ID number, latitude and longitude coordinates, and parent station name. The location coordinates were used to determine the city, neighborhood, and zip code that the stop resides in.

Methods

Regression Analysis

Data sets - On-Time Performance by Line (otp_by_line), Stops, and Stops vs. Lines (stops_vs_lines)

Description and Usability:

Regression analysis was used to predict the reliability of various MBTA routes, depending on which neighborhoods the routes travel through. In regression analysis, the focus is to determine how much of an impact the independent variables have on the outcome of the dependent variables. The independent variables in this situation were the different neighborhoods that routes travel within, and the dependent variables were the average off-peak on-time performance values for the routes. We the algorithm return coefficients that showed us the size of the effect each independent variable was having on our dependent variable.

Data Transformations and Process:

Our data had to be in a specific format in order to carry out regression analysis. The first step was to join "stops_vs_lines" with "stops." In doing this, we now had all information regarding location (ex: neighborhood, location coordinates, zip code) and routes for each MBTA stop. We filtered the results so that only stops with neighborhood information was included, since this was necessary information for our algorithm. The resulting data set was named "stop_route_neighborhood."

The next data set created to perform regression analysis was merged_stop, built by merging “otp_by_line” with “stop_route_neighborhood” to add the average off-peak and peak on-time performance values for each MBTA route. We devised a boolean matrix with the information at our disposal, entitled “stop_dummy_city.” The rows represented all stops and the columns represented all cities/neighborhoods. For a given stop, if it was located in a city/neighborhood, the entry was represented by a “1” within the matrix.

“Merged_stop” was joined with “stop_dummy_city” to gather all of the necessary information we had either retrieved from previously constructed data sets or programatically created. The results, initially organized by stops, was grouped by routes, which effectively summarized all of our statistics. At this point, the boolean matrix represented the neighborhoods that routes travel within, rather than the neighborhood that an individual stop is located in. Regression analysis was performed for each route in our data set, and its corresponding off-peak on-time performance values and neighborhoods.

Limitations:
Although the results returned from regression analysis were quite powerful and important to our findings, this statistical method suffers from some serious limitations. First, if we use our discoveries to predict MBTA reliability in the future, we are also assuming that the relationship we found between variables won’t change drastically. Second, it’s important to note that the data we were basing our findings off of was limited. The MBTA Performance data’s on-time performance values were based off of different metrics, depending on if the mode of transportation was bus, light rail, or commuter rail. This is especially important when it comes to the light rail’s data; the performance metric was calculated based on the estimated number of passengers waiting longer than the scheduled time and the estimated number of passengers waiting at the station. It’s clear from their labels that these are merely estimates, and there is no way of collecting accurate, comprehensive data for this metric. Finally, the process of computing regression analysis is lengthy and complicated compared to other statistical methods.

Correlation Analysis
Data sets - Trickling, Neighborhood Map (neighborhood_map), Means of Commuting, Poverty Rates, and Household Income

Description and Usability:
Correlation analysis is a method of statistical evaluation which focuses on studying the strength of the relationship between a set of variables. For our project, we were interested in not only observing the demographic factors among neighborhoods in Boston, but to also go one step further and analyze the relationships between a neighborhood’s demographics and those of its

surrounding neighborhoods. This way, we would be able to see this size of a neighborhood's "trickle effect."

Data Transformations and Process:

Using the data from "trickling," we constructed a data set entitled "neighborhood_pair," where the first two columns were every pairing combination of neighborhoods (except for the case when neighborhood == neighborhood) and the third column held the distance values between the two neighborhoods. Similarly, a new data set entitled "neighborhood" was built by transforming neighborhood_map, where the first two columns were pairs of neighborhoods and the third was a boolean value indicating whether or not the pairs were considered neighbors. Finally, "neighborhood_pair" and "neighborhood" were merged together to combine all pairs of neighborhoods, their distances to each other, and the boolean column to indicate their neighbor status.

The process of calculating correlation analysis was conducted three times, once for each demographic factor we were looking into. As an example, poverty rate will be used to walk through the course of the algorithm. For each neighborhood, we used our poverty rates data set and grabbed the poverty rate percentage, the percentage being of the population of that neighborhood, not the population of all of Boston. We then calculated the difference between each pair of neighborhood's poverty percentage, only if those neighborhoods were deemed neighbors, and appended the results as an extra column to "neighborhood_pairs."

Once the differences were found, we processed "neighborhood_pairs" to find the correlation coefficients for each neighborhood. These coefficients indicated the size of their trickling effect, or how much the status of their demographics impacts those of surrounding neighborhoods.

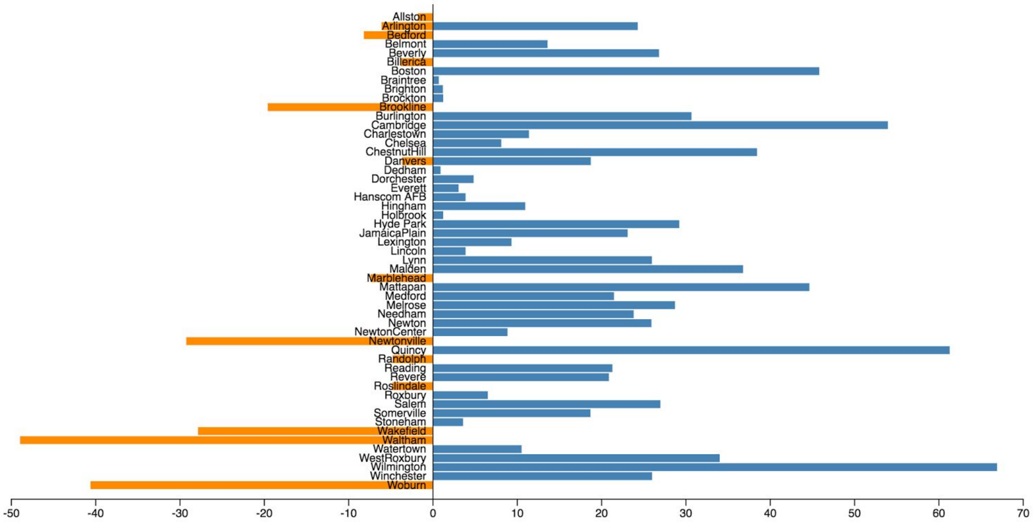
Limitations

An important limitation of correlation analysis that is often overlooked is that correlation doesn't imply causation. If we find that two variables, A and B, are positively correlated, all we can conclude is that when A increases in value, then B will as well, and vice versa. We cannot go so far as to assume that the performance of A directly influences the performance of B.

Results

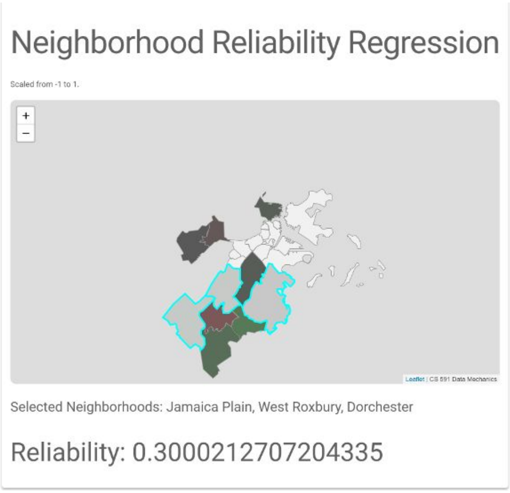
Regression Analysis

The implementation of our Regression Analysis produced the chart seen below:



The orange bars on the left-hand side of the chart indicate a neighborhood with a negative correlation coefficient and the blue bars on the right-hand side of the chart indicate a neighborhood with a positive correlation coefficient. More generally speaking, let's assume we are given an MBTA route, whether it be bus, rapid-transit, or commuter rail. If a neighborhood has an orange bar on the left-hand side of the chart, our analysis shows that the neighborhood has a negative effect on that route, i.e. routes that go through this neighborhood have their reliability (in terms of on-time performance) negatively affected. Conversely, if a neighborhood has a blue bar on the right-hand side of the chart, our analysis shows that the neighborhood has a positive effect on that route, i.e. routes that go through this neighborhood have their reliability positively affected.

We created a web visualization to represent the same results, but on an interactive map so that users can click and choose the areas they wish to see data on. Unlike the chart above, the coefficients for this map are scaled to be from -1 to 1. Additionally, users can select more than one neighborhood to see the average on-time performance correlation coefficient amongst all that are chosen, as seen in this screenshot.


Neighborhood Reliability Regression

Scaled from -1 to 1.

Selected Neighborhoods: Jamaica Plain, West Roxbury, Dorchester

Reliability: 0.3000212707204335

Conclusions:

We can see from our chart that neighborhoods like Cambridge, Quincy, and Wilmington have very positive effects on MBTA routes that pass through them, given their high correlation coefficients. Conversely, we also see that neighborhoods such as Brookline, Waltham, and Woburn seem to have a negative effect on MBTA routes that pass through them, given their high negative correlation coefficients.

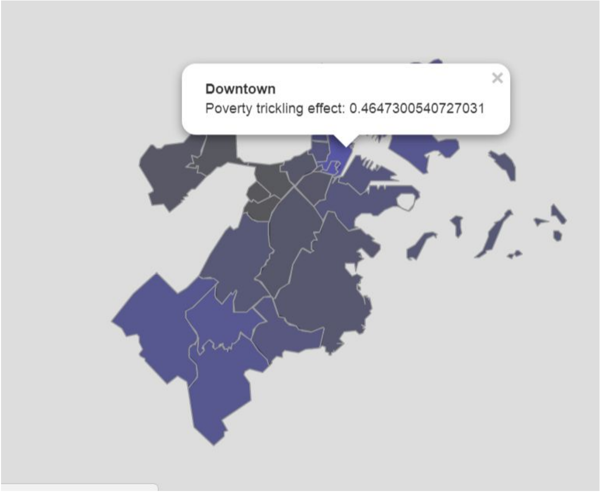
Correlation Analysis

The implementation of our Correlation Analysis produces several visualizations seen below, along with the conclusions we can draw from each one of them. All maps follow the same color range, where blue represents a large trickling effect, dark blue represents a moderate trickling effect, dark gray represents a very small trickling effect, and light gray represent non-existent trickling.

All maps can be accessed here: https://shreyapandit.github.io/data-science-boston/

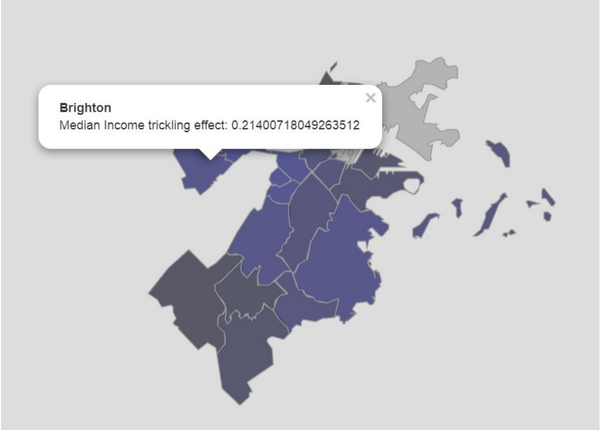
Poverty Rate

This map shows the trickling effect of each neighborhood in Boston for poverty, i.e. how strongly the poverty rate in a given neighborhood affects the neighborhoods directly surrounding it. We can see from our visualization that areas such as Downtown, Roslindale, West Roxbury, and Hyde Park have a high trickling effect, and that the poverty in those areas has a high effect on the poverty in surrounding areas. Conversely, we can also see that areas such as Allston, Fenway/Kenmore, and Mission Hill have a lower trickling effect, and that poverty in these areas has a low effect on poverty in the surrounding areas.



Median Household Income

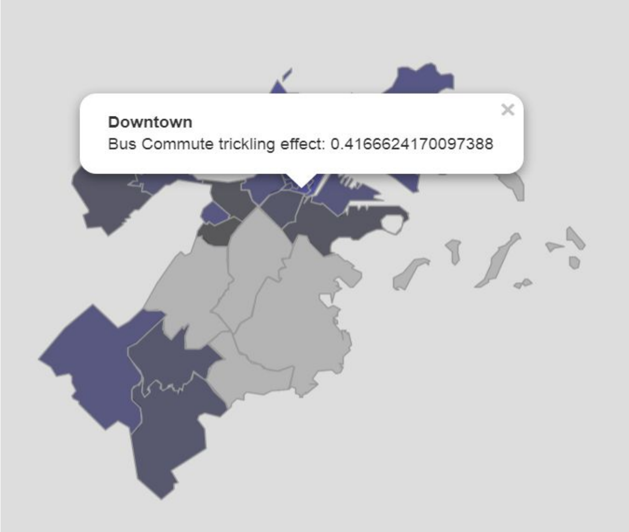
This map shows the trickling effect of household income for each neighborhood in Boston. We can conclude from this map that areas such as Brighton and Dorchester have a large positive trickling effect. If there were to be a massive round of layoffs that primarily affected residents of these neighborhoods, the drastic change in household income would have a high effect on that of surrounding areas, which in this



case would not be a fortunate circumstance. If the layoffs affected residents of East Boston, we would not see nearly the same effect.

Means of Commuting

This map shows the trickling effect for how each neighborhood of Boston commutes to work, specifically in regards to the residents' usage of the MBTA bus. Other maps regarding the residents' usage of cars, bikes, or their own legs (i.e. walking) can be found at the link posted. From this map, one can reason that if bus routes going through Downtown were discontinued, this would affect the surrounding neighborhoods as the residents probably rely on similar bus routes to get to work. On the other hand, areas such as Roxbury, Dorchester, and Mattapan show almost no trickling effect. One can reason that residents are so far from their workplace that they will continue to be reliant on their current means of commuting no matter what the circumstances are in surrounding neighborhoods. It's important to note that these are just theories (i.e. correlation doesn't imply causation), as we would not know the exact reason behind these trickling effects without further research and more extensive data.



Future Work

The results from both algorithms can be used in the future by policy makers in Boston, helping them to make informed decisions that would affect residents of the city and its neighbors.

Future MBTA Routes:

As the City of Boston and the surrounding area grows, expansions to the MBTA will become necessary. If policy makers ever want to consider adding another MBTA route, our algorithm would be able to predict the reliability of that route, given the neighborhoods that it passes through.

For example, the Green Line Extension project will extend the Green Line through Somerville and Medford. We can use our regression analysis to predict the reliability of this new route and other such projects based off the neighborhoods that they pass through.

Reformative Neighborhood Policies:

If any kind of reformative policies were written for a neighborhood, it is important to know if this change will trickle to surrounding neighborhoods, or if the policies will have to be explicitly implemented in surrounding neighborhoods.

For example, our data shows Downtown has a high trickling effect on surrounding neighborhoods regarding poverty. Therefore, if the City of Boston were to dedicate resources towards alleviating poverty Downtown, our data shows there would be greater effect on surrounding area than if they were to put a focus on other areas, making it a more valuable investment.

Correlation between Data and Potential Use:

There are also many potentially interesting and useful benefits to drawing conclusions using both datasets. By knowing the effect neighborhoods have on MBTA routes as well as a plethora of demographic information regarding said neighborhoods, including their “trickling” effect, we can make several inferences about the present and future.

For example, if several neighborhoods had certain demographic similarities as well as similar effects on MBTA performance, we could assume that there was a correlation between the two pieces of data, and could replicate or mitigate such scenarios depending on our desired outcome.