Access our youtube video here! (https://youtu.be/zv2PxOoe01U)

Metro Mania

Background

Wes has always loved trains, and in high school, he had an idea to redesign train schedules to optimize efficiency (which he entered into the Connecticut science fair, and again into another competition held by the NYC metro authority). In college, Wes got a job at the Volpe Center in Cambridge doing transportation crashworthiness work, which is not the same as his project but is still under the topic of locomotives. He reached out to the WMATA because he knew that they have much better data collection than NYC does, and they sent him this dataset to work with. This CS 109 project provided the opportunity for Wes to (finally) get going on it, and he plans to continue working on it as an independent project in the winter. Erika doesn't know much about trains but is interested in large data sets! Erika is working in the Stanford School of Medicine on an algorithm to identify COVID-19 before symptom onset using smart watch data, which would allow people to self-isolate before becoming symptomatic and contagious and infecting others. As engineering majors, Wes (ME) and Erika (EE) met in ENGR-14 last year and became instant friends after building a bridge together!

Source Data

In the original data, each row represents data about one passenger's journey; their starting station, their ending station, and the start and end time of their trip (down to the second resolution). WMATA is able to track this because passengers must swipe out at their destination instead of freely leaving the station. This data was collected for the entire month of October 2018.

Sample:

START_STATION_NAME	*END_STATION_NAME*	*START_TIME*	*END_TIME*
Van Dorn Street	*Pentagon*	*10/1/2018 6:34:02*	*10/1/2018 6:53:55*
Court House	*Smithsonian*	*10/1/2018 7:24:30*	*10/1/2018 7:44:21*

Processed Data

For this project, we used almost entirely frequentist probability. Although that may not be the best way of modeling ever-changing passenger flows, it was a simple way of getting to some really interesting observations.

This data set is gigantic (15 million rows), so we summarized it by splitting the day into 96 discrete 15-minute chunks. Using RStudio, we used these to calculate different probabilities corresponding to each time chunk. Each row of our new data set contains data characterizing types of journeys, with "type" being origin station, destination station, and journey start time. Then we recorded the number of people traveling each line during each "chunk," which is the word we use to refer to unique trip types. Each row represents an origin-destination pair within a 15 minute time window e.g. passengers going from Dupont Circle to Federal Triangle between 8:45 and 9:00. Our "**n**" column represents the total number of passengers making the above type of journey over the course of the entirety of October 2018. **N** is the basis for almost every probability calculation.

In order to calculate conditional probabilities, we need to find the total number of passengers in all chunks which meet certain conditions, then divide **n** by that number. For example, the

$$P(\textit{passenger is going to Federal Triangle} \mid \textit{they entered the Smithsonian station at } 6:12\,PM) = \frac{\textit{\# of passengers who entered Smithsonian between } 6PM \textit{ and } 6{:}15PM \textit{ en route to Federal Triangle}}{\textit{\#of passengers who entered Smithsonian between } 6PM \textit{ and } 6{:}15PM}$$

For these conditional probabilities, the numerator is **n** and the denominator is any of three sums "**stationSum**," "**timeSum**," and "**entrySum**;" representing the sum of passengers who entered at this station, the sum of passengers who entered at this time, and the number of passengers who entered at this at station at this time. For example, the stationSum column for any row describing a type of journey from Dupont Circle to Federal Triangle represents the sum of passengers going from Dupont Circle to Federal Triangle at any time for the entirety of October 2018 (i.e. any origin-destination pair going from Dupont Circle to Federal Triangle will have the same stationSum, but Federal Triangle to Dupont Circle will have a different stationSum). Using this information we were able to calculate (for each row of the processed data) the

*"**probabilityGivenTime**" = P(origin & destination | time window),*

*"**probabilityGivenStations**" = P(time window | origin & destination),*

*"**probabilityGivenEntry**" = P(destination | origin & time window).*

We also calculated (for each row)

*"**averageLengthForTime**" = E[trip time | entry time],*

*"**averageLengthForStations**" = E[trip time | origin & destination],*

*"**averageLengthForEntry**" = E[trip time | origin & entry time],*

*"**averageLengthForTimeAndStation**" = E[trip time | origin & destination & entry time]*

The Code

Wes used R studio to make numerous plots describing this data. See the video to learn more about each plot and interesting observations we can draw from each one.

Plot A is a heatmap of passenger traffic, with the axes being origin and destination station. The axes are ordered based on entryRatio (see plot E).

Plot B is a heatmap of total trip time, with the axes being origin and destination station. The axes are ordered based on entryRatio (see plot E).

Plot C is a map showing the probabilities of different destinations from the origin (in this case, DuPont Circle) over the 24-hour day.

Plot D is a different representation of the same data as in plot C, showing the expectation of passengers traveling to different destinations from Dupont Circle over time. It highlights the morning and evening commute spike.

Plot E is a map showing the entryRatio variable for each station, signifying whether more passengers enter in the morning or the evening. From this, we can approximate how residential a neighborhood is.

To build the map visualizations, Erika found the GPS coordinates of the stations, found the centroid, then subtracted that off from each point so that the GPS coordinates would be centered around (0,0) for plotting. This allowed an easy transposing of longitude/latitude coordinates onto the XY plane. Wes then drew the map of the system over these plotted points.

Significance

Why is it important to see these passenger flows? Perhaps it will be able to help us understand where people work given their neighborhood and their expected commute time (E[trip time | origin & time window]). Seeing what times certain lines are busy can also help people avoid busy times if they are old or disabled and at risk for falling or have a compromised immune system and are at risk for infection. As a continuation of this project (perhaps for Wes in the Winter), it would be interesting to see how ridership has changed during the pandemic. Perhaps this knowledge could expand to other cities, or help D.C. decide which lines to keep lines running when, and change the schedule to save fuel and cost. Perhaps everyone with tech jobs are now working from home, so people who formerly traveled on the lines going to techy work spots are not as crowded anymore, and D.C. would not need as many trains visiting the techy workstations or neighborhoods. Additionally, Wes calculated “entryRatioStart,” which is the ratio of 8:00 AM entries to 5:15 entries for the start station. The higher this value, the more residential this station is, as 8:00 and 5:15 are the two peak entry periods in the day, one for the morning commute and one for the evening commute. Using this ratio, Wes can determine which neighborhoods are residential and how residential they are, and then use P(destination | origin & entry time) to see where people from each neighborhood works. This may reveal interesting socio economic connections between different work districts and, perhaps, zip code prices. If he is able to get ridership data during the pandemic, he might be able to see in which neighborhoods people are still working in person, how that correlates to socioeconomic status, and also how that correlates to COVID-19 cases.

Fundamentally, being able to model passenger traffic probabilistically will be crucial to designing fair and efficient transportation systems in the future. As self-driving technology and scheduling technology improve, it is likely that our passenger rail systems become smarter, adapting to ever-changing passenger flow patterns and conditions. To enable that process, we need a robust and reliable way of predicting where and when passengers will be and where they need to go. While this project doesn’t find any solutions, it’s an important first step in understanding how people use existing transportation systems.

Note: October_2018 (source data) and passengersByChunk6 (processed data) can be found here for anyone who is interested! (https://drive.google.com/file/d/1d0wzDGAHKmmVJxdviMMCDmnq5-UdxOhW/view?usp=sharing)