Applied Data Science - Project Proposal

**Estimating Time and Cost Sensitivity in New Yorker's Transport Decisions:**

**Evidence from the Second Avenue Subway**

*Guobing Chen, Fangshu Lin, Hao Xi, Nick Jones*

|  |
| --- |
| **Summary:**  In January 2017, three subway stations opened on the Upper East Side as part of the long-awaited Second Avenue Subway. This study will use high-resolution transport and weather data to understand how the subways affected New Yorkers' decisions on what mode of transport to take. Using the new stations as a 'natural experiment' should allow us to get new estimates of customer's demand functions for different modes of transportation, which will be helpful for transport planning such as the siting of new stations in future. |

**Research question**

We are interested in how the mobility of the neighborhood was affected by the opening of the new subway stations (72 St, 86 St, 96 St). To understand this we investigate taxi pick-ups and drop-offs in the surrounding area. Using the geographic and time-series data available from the TLC, we are able to examine change in:

* number of outbound/inbound taxi drips by time of day
* number outbound taxi trips by destination ZIP code
* number of inbound taxi trips by origin ZIP code
* average taxi trip length
* effects of distance from subway station

The team is particularly interested in how distance from subway station affects the number of pick-ups. We hypothesize that customers are more likely to switch from subway to taxi if the station is nearby (eg. 1 block away instead of 4 blocks away). We will test for the existence of this distance to subway effect. If we confirm that a significant effect exists, we will estimate its size. Estimating the propensity of customers to substitute between taxi and subway will have useful applications for transport planning (see below).

**Why is it important**

New York, like other cities, has a transport system that is increasingly multi-modal. People choose between subway, bus, taxi, Uber, walking and cycling. The city aims to promote quick and easy mobility, at low cost.

It is particularly important to use data and evidence for transport planning because subways are extremely expensive: estimated at $4.5 billion for the three Second Avenue stations. Without evidence on new subways actually affect people's transport decisions, it is hard to make a strong case for new investments.

Our estimate of the distance sensitivity of demand for subway trips vis-a-vis taxi trips will be helpful for MTA planning decisions. For example, if an area between stations shows little drop-off in taxi pick-ups for equivalent routes (eg. where trips to midtown are possible via both modes), the subway stations may be too distant.

**Data**

We will use the data listed below:

|  |  |
| --- | --- |
| **Data type** | **Accessibility and description** |
| **TLC Trip Record Data** | Variables of interest include pick-up time, drop-off time, trip distance, longitude, and latitude. Data is high spatial and temporal resolution.  <http://www.nyc.gov/html/tlc/html/about/trip_record_data.shtml> |
| **MTA data** | MTA turnstile data for each device are recorded every 4 hours. Timestamps are accurate to seconds.  <http://web.mta.info/developers/turnstile.html> |
| Real time MTA data is available using API.  <http://web.mta.info/developers/turnstile.html> |
| A new dashboard has been released recently, which displays information such as wait time, incidents, and travel times, etc.  <http://dashboard.mta.info> |
| **Weather data** | Customized historical and real-time weather datasets are available.  <https://www.weather.gov/mdl/ndfd_home>  <http://w2.weather.gov/climate/index.php?wfo=okx>  <https://www.wunderground.com/history/> |
| **Maps** | Subway entrances:  <https://data.cityofnewyork.us/Transportation/Subway-Entrances/drex-xx56>  Subway stations:  <https://data.cityofnewyork.us/Transportation/Subway-Stations/arq3-7z49>  Subway lines:  <https://data.cityofnewyork.us/Transportation/Subway-Lines/3qz8-muuu> |

**Methods**

The analysis for the project is in two parts: (i) outbound trips; (ii) inbound trips.

First, outbound trips will be analyzed with respect to variations in destination, duration, pick-up time, pick-up location, and weather condition. By comparing taxi trips' pick-up locations, time and destination before and after the Second Avenue Subway was opened, the project will evaluate the influence of the new subway stations on the taxi journeys.

The study will look at two time periods. We will select a period from before the new stations were opened (tentatively January 1 – July 1 2016), and name this P0. We will then study the six months after subway opening, naming this P1. We will adjust for underlying trend changes in taxi volumes by using a control group (a neighborhood of similar income, tentatively the Upper East Side between 60th – 70th St). We will then conduct regressions to test for statistically significant changes in taxi pick-ups (including a distance-to-subway effect) in P1 with respect to P0.

Using data from before the new stations were opened, a taxi demand model will be built. This will relate number of taxi pick-ups to distance from subway station, time, weather condition, and destination. Another taxi demand model will be built with the same attributes but using data from P1. By comparing the parameters in the two models, the influence of the new stations on taxi pick-ups will be revealed.

Figure : Second Avenue Subway

Second part of the analysis focuses on the inbound trips. Two models predicting taxi drop-off numbers will be built. One will be based on the taxi data before the Second Avenue subway was opened, the other will be based on the data after the stations are completed.

**Analytical issues**

The team will bear in mind several methodological issues while doing the project:

1. Making valid comparisons. To estimate the sensitivity of taxi demand with respect to subway station proximity, we need to compare journeys that are feasible and of similar duration on both modes of transport.

2. Ability to bring in new data. The study will develop a method that could incorporate Uber and Lyft data, once these become available for 2017.

3. Potential for complex changes outside the model parameters. For example, if commuters save 1 hour travel time per day using the subway, they may make additional taxi journeys in the evening, to go out for dinner. Bearing in mind the potentially complex changes taking place, the study will therefore focus on the specific question of willingness to substitute taxi and subway journeys based on distance.

**Anticipated roles**

The key tasks include: Literature review, data collection, data wrangling, data analysis, modeling, visualization, reporting. The team members expect to share the tasks:

**Hao**: GIS lead, data collection, data wrangling

**Fangshu**: Literature review, data collection, data analysis, model building

**Guobing**: Data collection, data analysis, model building, visualization

**Nick**: Literature review, data collection and analysis, model building,