Fare Share: Flow and Efficiency in NYC's Taxi System

Abraham Neuwirth
Touro College
abraham.neuwirth@student.touro.edu

Fatima Chebchoub
NYC College of Technology
fatima.chebchoub@mail.citytech.cuny.edu

Jai Punjwani
Adelphi University
jaipunjwani@mail.adelphi.edu

Marieme Toure
NYC College of Technology
marieme.toure@mail.citytech.cuny.edu

Introduction

Millions of people move throughout New York City each day and yet relatively little is understood about where and when people travel, both at the individual and aggregate levels. Better insights around these travel patterns could play crucial roles in everything from simply understanding people's habits to improving traffic flow and optimizing taxi provisioning. In this paper, we address several such questions using highly detailed data for 13 million NYC taxi trips in 2013 to better understand the flow and efficiency of the taxi system and the people it services.

First, we use the patterns of pickups and dropoffs across different neighborhoods to get an overview of the entire city, showing how people move between neighborhoods during a typical week. Next, we look at the role of drivers in the taxi system, specifically investigating how earnings vary across drivers and quantifying how much of this variation is due to skill versus chance. Somewhat surprisingly, we find that while factors such as time of day and weather have a large impact on efficiency of the taxi system, skill plays a sizeable role in determining driver efficiency with some drivers consistently earning up to 30% more than average. Finally, we use the highly granular nature of this data to identify opportunities to improve the efficiency of the taxi system through a simple carpooling strategy. Specifically, we identify locations throughout the city with consistently redundant trips, where two or more taxis leave from the same place at the same time, traveling to the same destination. We show that a taxi stand policy requiring people to wait no more than five minutes to carpool with another rider at these locations could improve the system by upwards of 5%, eliminating more than 650,000 trips and saving consumers $8.5 million each month.

In the remainder of the paper we discuss more details of the data and methods used to obtain these results.

Data and Methods

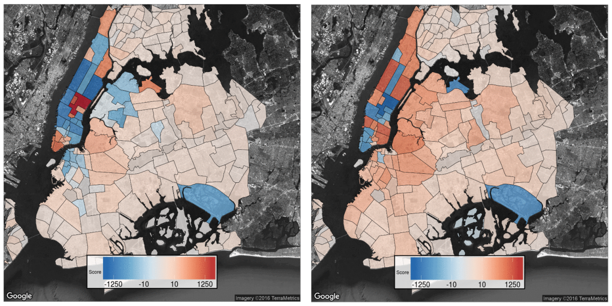
Our data set [1] had records of every taxi trip in a yellow cab in 2013, with detailed information such as the start and end time of each trip, pickup and drop off coordinates for the trip, the total fare paid, and an anonymized driver's license, among other fields. The anonymized driver's licenses, which were unavailable in more recent data sets (2014 & 2015), allow us to associate trips with drivers, which aid in the understanding of driver behavior. We chose to work with the month of July 2013 for our analysis, which contains more than 13 million rides, for an average of 420,000 trips per day driven by over 32,000 different drivers. We joined this data with daily weather measured in Cental Park by the National Oceanic and Atmospheric Association [2] to aid in our analysis.

Flow

First, we looked at the overall flow of taxis in NYC to find the general trends that define the movement of people within the city. To understand flow, we used shapefiles [3] to map pickup and dropoff coordinates into one of 266 New York City neighborhoods. We then grouped trips by neighborhood and looked at the change in population (number of passengers that entered minus number of passengers that left) at every hour, averaged over all weekdays in July. Finally, we used the median change in population of each hour to represent the net flow.

Figure 1, which shows the net flow for 7 AM and 7 PM for weekdays in July, helps us visualize the movement of people and understand the flow in NYC. We can see, for example, that in the morning Midtown and the Financial District have a high inflow while most of the residential areas in Manhattan and Downtown Brooklyn have negative flow scores, indicating a high level of outflow. In the late evening this trend reverses, with neighborhoods in Greenwich Village, Uptown, and Downtown Brooklyn receiving the highest inflow, while Midtown and the Financial District have high outflow scores. Another interesting observation is that most neighborhoods in the outer boroughs are either stable (they have the same inflow as outflow) or have a slightly positive inflow. The only notable exceptions are the airports (LaGuardia Airport and John F. Kennedy International Airport) which have relatively high inflows in the early morning hours but high outflows for the rest of the day, and neighborhoods close to to Manhattan such as Williamsburg, Greenpoint, and Long Island City which have relatively high outflow throughout the day.

While this snapshot is certainly informative in its



(a) *7 AM* **(b)** *7 PM*

Figure 1: *Net flow of people for weekdays in July 2013*

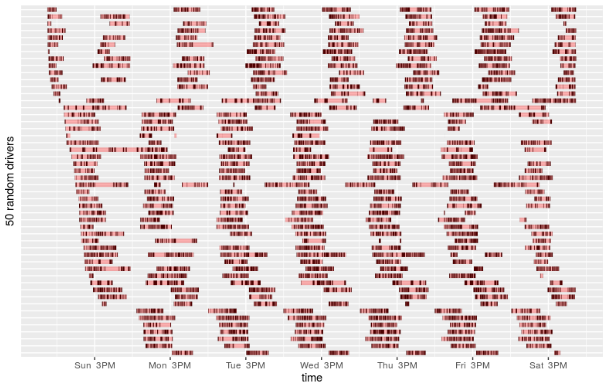


Figure 2: *Rides (in black) and shifts (in red) for 50 random drivers over a week's time in July.*

own right, it betrays the level of detail and intricacy of the data at hand. Different people live in different neighborhoods, and each of these neighborhoods have their own story to tell. To tease out these stories, we grouped all the rides by source neighborhood, time of the week (weekends and weekdays) and hour of day, and for each group we computed the distribution of probability for each possible destination. We then built a tool[1] which allows one to explore this data to see trends of popular destinations for individual neighborhoods at different times.

Driver and Shift Efficiency

Given this high-level understanding of how people move throughout the city, next we investigated the role of drivers in the taxi system. In particular, we looked at two questions: first, how do driver earnings vary, and second, how much of this variation is due to skill versus chance?

We began by computing the efficiency of a driver in a work period, defined as the ratio of the total metered fare earned in that period to the total time worked. The total metered fare earned on each trip was available directly in our data set. Unfortunately, however, the data does not contain the time each driver spent working in a shift; rather, it only logs the times that a driver was with a passenger, making it difficult to identify the length of a shift. Previous work used the distance between a dropoff and the next pickup to approximate total shift time [4]. We took a more refined approach to identify driver shifts by looking at the downtime, or time between a drop off and the next pickup, for all trips. Consistent with past work on driver earnings [5], we found that very few drivers have downtimes of six hours, with the time between most trips either ranging from a few minutes to 12 hours. Accordingly, we defined any dropoff where there is no driver activity for at least six hours as the end of a shift, with the following pickup marking the beginning of a new shift. With this information we were able to group each driver's rides into shifts and compute driver efficiency for each shift as defined above. Figure 2 shows rides grouped into shifts across one week for 50 randomly selected drivers. Our six-hour rule of thumb seems to work well as many drivers have consistent shifts throughout the week, as a typical working adult would have. For example, we see drivers with shifts that start at a similar time and span a similar duration throughout the week.

Our new shift data revealed that drivers have passengers for approximately half of their shift time and earn an average of $30 per shift, with a reasonable amount of variation in earnings across drivers. That said, it is unclear what drives this variation—is it simply due to the string of pickup and dropoff locations a driver happens to be assigned, or can it be attributed to some inherent difference in skill across drivers? To better understand variation in efficiency, we fit a linear regression to predict shift efficiency using the following model:

$$\beta_{\text{driver id}} + \beta_{\text{hour}} + \beta_{\text{weekend}} + \beta_{\text{hour:weekend}} + \\ \gamma_p x_p + \sum_{n=1}^{N} \rho_n p_n + \sum_{n=1}^{N} \delta_n d_n,$$

where each β represents the effect of its corresponding subscript—whether it be the driver's ID, the start hour of a shift, or whether it is a weekend of the weekday—x_p is precipitation in inches, and γ_p is its coefficient. The two summations represent the percentage of pickups (p_n) and drop-offs (d_n), respectively, in each neighborhood, with ρ and δ as their coefficients.

After running our model, we observed a significant variance in earnings, with some drivers consistently making ±$10/hour from the average. However, this model does not necessarily confirm the relationship between drivers and skill, as the high variance could have been due to chance. To see what we would expect from chance, we randomized the assignment of driver ids to shifts and re-ran our regression. In the randomized data set, the variance was cut in half to ±$5/hour (Figure 3). The difference between the distributions of earnings signifies that the actual variation in driver earnings is, in fact, much larger than we would expect from chance: there are drivers who consistently make a great deal more than the average

[1] http://bit.ly/nyc_taxi (accessed August 11, 2016)

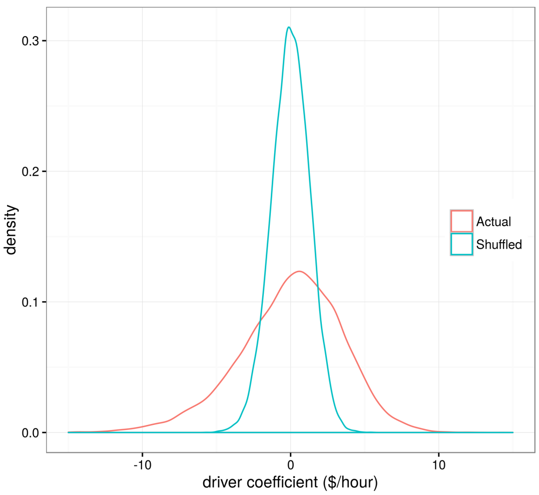


Figure 3: *Effect of Driver ID on Efficiency*

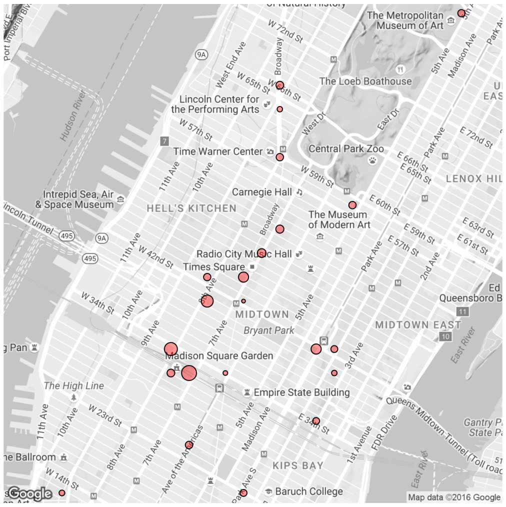


Figure 4: *Top carpooling hotspots in Manhattan. Points are sized by how frequent a carpooling potential occurs at a location.*

driver, as well as drivers who make a great deal less. Thus, while anyone can become a cab driver, there are certain skills that distinguish good drivers from less effective ones.

Carpooling

After looking at flow and driver behavior, we wanted to find ways to improve the overall taxi system. Past work has focused on recommender systems [6] or information systems [7], which would require taxi drivers to be notified via a mobile application or a similar solution to re-route drivers and improve efficiency. We take a different approach and look at a policy that can be implemented at existing taxi stands, with little overhead. We noticed that there were lots of trips occurring between similar locations. As a result, we considered a scenario in which people would carpool. Our thought experiment made the following assumptions: Customers would be willing to (1) share a cab with strangers, (2) wait up to five minutes to find someone to carpool with, (3) walk up to one block, (4) share a destination within 1 kilometer of their own destination. We also assumed that customers would carpool only for trips between two distinct neighborhoods, because customers would probably not want to wait for trips that are relatively short.

To look for carpooling potential, we rounded the start time of trips to the nearest 5 minutes, pickup latitude and longitude to the nearest two thousandth of a degree, and dropoff coordinates to the nearest hundredth of a degree. We then counted the number of trips and passengers within each "carpooling potential" bin. Somewhat surprisingly, we found that a significant number of trips left from the same place at the same time, going to the same destination. After ranking these trips and plotting their pick-up points on a map, we identified the top carpooling hotspots in NYC (figure 4). Unsurprisingly, many of these places are either major transportation hubs (JFK airport, LaGuardia airport, Penn Station, Port Authority Bus Terminal, and Grand Central Terminal) or popular cultural attractions (the Metropolitan Museum of Art, the Lincoln Center, the Theater District, etc.). For instance, we found that on weekday mornings around 7am, there are roughly 25 redundant trips from Port Authority to Rockefeller center that take place every five minutes for the duration of rush hour.

Next, we used these redundant trips to calculate the potential savings that a simply carpooling strategy could produce. To do so, we assumed that up to four passengers would fit in one cab, since 92% of taxis fit four passengers [8]. The minimum number of trips needed per each five minute bin is given by ⌈number of passengers/4⌉. The number of "unecessary" trips per bin are thus actual number of trips minus minimum number of trips needed, and the potential fare savings are average fare per trip × "unnecessary" trips. Our potential savings over the month of July were over 650,000 thousands trips, around 5% of total trips and over $8.5 million, around 6% of total money spent by consumers.

We also found that with less restrictive assumptions, i.e. widening the waiting period to 6 minutes, and including rides taking place within the same neighborhood, we can improve the savings up to 14% for the number of rides and fare paid. In addition to saving money, carpooling would also help the environment and reduce traffic.

Acknowledgements

We thank our mentors Ashton Anderson (Microsoft Research), Christopher Riederer (Columbia University), Jake Hofman (Microsoft Research), and Siddhartha Sen (Microsoft Research) for their mentorship as a part of the 2016 Microsoft Research Data Science Summer School. Maps were produced using the ggmap package in R [9].

References

[1] Chris Whong. *FOILing NYCs Taxi Trip Data*, March 2014 (accessed August 11, 2016).

[2] National Centers For Enviornmental Information. *Climate Data Online*, (accessed August 11, 2016).

[3] Pediacities. *Pediacities NYC Neighborhoods*, June 2015.

[4] Samuel K. Lee. *7 Habits of Highly Effective Hacks*, March 2015 (accessed August 11, 2016).

[5] Henry S Farber. Why You Can't Find a Taxi in the Rain and Other Labor Supply Lessons from Cab Drivers. October 2014.

[6] Xianyuan Zhan, Xinwu Qian, and Satish V. Ukkusuri. *Measuring the Efficiency of Urban Taxi Service System*, August 2014 (accessed August 11, 2016).

[7] Hyunmyung Kim, Jun-Seok Oh, and R. Jayakrishnan. *Effect of Taxi Information System on Efficiency and Quality of Taxi Services*, 2005.

[8] New York City Taxi & Limousine Commission. *Taxi 07: Roads Forward*, December 2007 (accessed August 11, 2016).

[9] David Kahle and Hadley Wickham. ggmap: Spatial Visualization with ggplot2. January 2013.