Cluster Analysis of Endogeneous Taxi Driver Schedule Patterns

Pradeep Pendem*
Jessica Leete†

Abstract

In this paper, we utilize detailed trip and fare data of more than 30,000 New York yellow taxi drivers to understand similar patterns in their chosen schedules within a day. We use in-depth trip by trip information along with driver id to create a dataset at the intra-day level and cluster the observations using a K-means clustering algorithm. Based on a threshold of error measure, we identified and choose four clusters. Two clusters of schedule patterns were found to be of drivers who provide service during daylight hours. Remaining two clusters were found to be drivers choosing to work during night hours. This study has important policy implications for the parent organization on incentivizing drivers to participate in taxi service thereby providing an opportunity to match supply and demand.

1 Introduction

The taxi industry is an important component of city transport to provide safe, quality, and efficient access to car services. Typically, taxis are mostly owned and operated by a large number of individuals while licensing and fare is regulated by a larger corporation. With no pressure on service intensity, drivers choose their working schedules or number of service hours in a day based on profitability and operating cost [1]. This leads to a variation in working hours and schedules. As a result, taxi supply and demand vary in a non-homogeneous pattern over time and space. However, taxi service systems are commonly modeled to have uniform supply and demand by aggregating different time periods into one static period. The objective of this project is to better understand the dynamic supply of taxis. We will accomplish this in three steps. First, identify intra-day patterns of taxi driver schedules. Second, group similar schedule patterns using K-means clustering algorithm. Third, identify and report sources of fare variation among clusters (if any). In the following sections, we explain our dataset, provide basic summary, method, results, and conclusion.

2 Dataset and Summary

To address our objective in this paper, we utilized publicly available data of New York yellow taxis at NYC Taxi and limousine commission. We selected first week taxi trip and fare data which are approx 4 million records and has more than 30,000 drivers. We merged trip and fare data by driver id and pickup date time to create a unified dataset. All data analysis in this paper was carried out in R-software, and charts were created with ggplot2 package. Table 1 displays the variables present in our sample dataset.

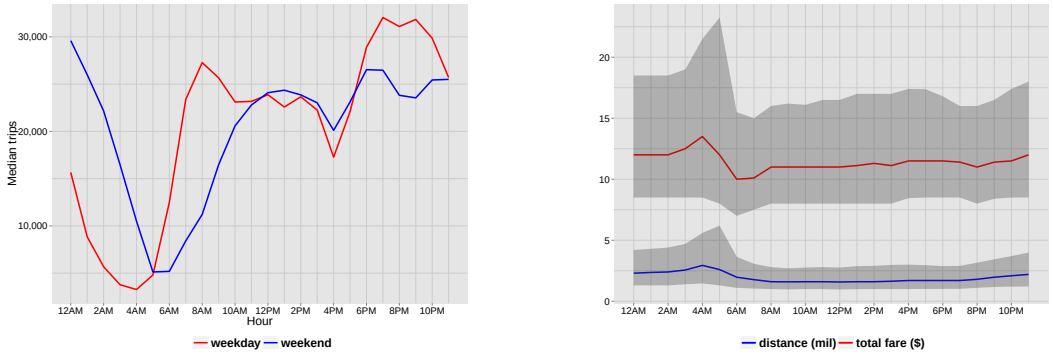
During a detailed data analysis, we identified records with abnormal dist mil, dur secs, speed (miles per hour) and total fare. Hence we choose observations with dist mil more than 0 and less than 35 miles, dur secs more than 0 and less than 6300 secs, speed more than 0 and less than 70 miles per

*UNC Chapel Hill, pradeep_pendem@kenan-flagler.unc.edu
†Duke University, jessica.leete@duke.edu

Table 1: Dataset variables and description

Variable	Description
driver id	unique id for each driver
pickup datetime	date and time stamp of the pickup
dropoff datetime	date and time stamp of the dropoff
pickup longitude	pickup longitude coordinate
pickup latitude	pickup latitude coordinate
dropoff longitude	drop off longitude coordinate
dropoff latitude	drop off latitude coordinate
passenger count	number of passengers who availed the trip
dist mil	distance of the trip in miles
dur secs	duration of the trip in seconds
pay type	card or cash
fare	actual fare of the trip
surcharge	surcharge based on the demand for trip
mta tax	metropolitan commuter transportation mobility tax
tip	tip to the driver
toll	toll fare
tot fare	total fare = fare + surcharge + mta tax + toll

hour, and total fare more than 0 and less than $150.
Though we can provide extensive summary with our data, we limit our charts to trips, and fare and distance per trip in our sample dataset. Figure 1a shows the median total trips for weekdays and weekend for each hour. It is visual evidence of a dynamic supply of taxis through driver endogenous schedules. Keen observation of the plot shows drivers tend to participate more in providing service during mid part of day and until late night, and less in early morning. Figure 1b shows median and variation of distance and fare per trip by each hour. It is interesting to see drivers tend to choose longer distance trips during 2AM-4AM and shorter trips during 6AM-8AM compared to average values through out other part of the day. Hence similar trends are observed in fare as well.



(a) Median trips by hour

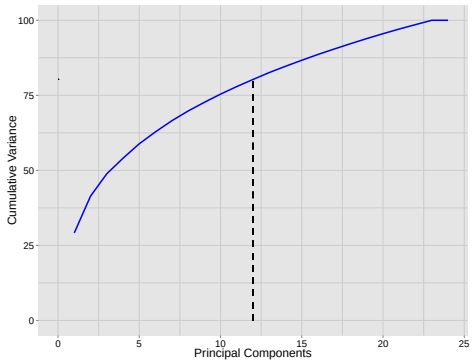
(b) Distance and fare per trip by hour

Figure 1

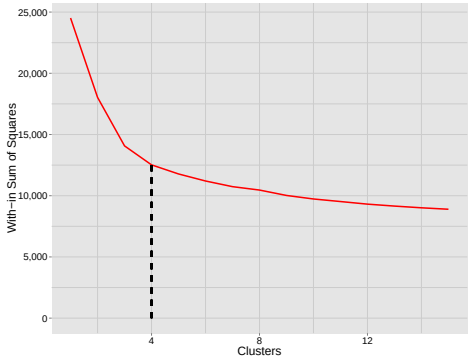
3 Methods:

The main strength of our dataset is availability of driver id which helps our to track path taken by a driver in real time and hence number of trips taken in each hour. Using the valuable information of driver id, we create a dataset to represent their schedules. Given a day, driver id we calculate the number of trips taken during each hour and total value. We then divide number of trips in each hour to total number of trips which gives proprotion of trips. Hence for each driver id, day we have 24

columns each representing proportion of trips take in the particular hour of the day.
As we have 24 variables, we attempt to reduce the dimensionality of the data using Principal Component Analysis (PCA). After analyzing the principal components, we realize it requires a minimum of 12 principal components to capture 80% of variation the data (Figure 2a). While we could reduce the dimensions from 24 to 12, it was not a drastic enough change to warrant loosing 20% of the information. Thus, we clustered on the unreduced data set, but did utilize information of the principal component breakdown to later visualize the data.



(a) Cumulative variance of principal components



(b) Within Sum of Squares for each choice of clusters

Figure 2

K-means clustering divides observations into K clusters by iteratively calculating the mean of each cluster, then labeling each data point as belonging to the nearest cluster mean. Our K-means model is found in Equation 1 where K is the number of clusters, N_l is the set of observations in cluster l and $N_1 \cup N_2 \cup \cdots \cup N_K = \{1, 2....n\}$ where n is the number of observations, and $prop\_trips_{ij}$ is the proportion of trips in observation i in the jth hour. Equation 1 minimizes the distance between observations in the cluster, which can be shown as equivalent to minimizing the distance between the observations and the center of the cluster [2].

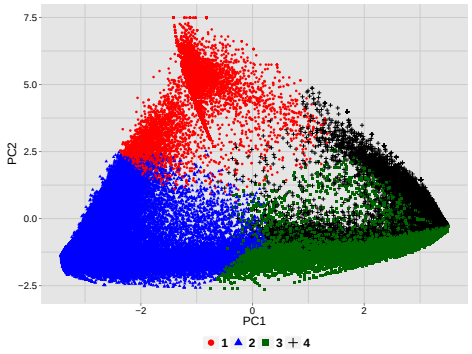
$$\min_{N_1,N_2,...,N_k} \sum_{k=1}^{K} \frac{1}{|N_k|} \sum_{i,i' \in N_k} \sum_{j=1}^{24} (prop\_trips_{ij} - prop\_trips_{i'j})^2 \tag{1}$$

We performed K-means clustering with $K = 1, 2, 3...15$ clusters to determine the number of clusters that most closely represents the natural groupings in the data. We measured the fit of the clusters using Within Sum of Squares (WSS). The WSS is calculated by summing the distance between each point in each cluster to the center of the cluster. A larger WSS would mean that larger variation occurs between observations in a cluster, while a smaller WSS indicates denser clusters. Four clusters is where adding another cluster starts only marginally decreasing the WSS (Figure 2b). Thus, we decided to use 4 clusters.

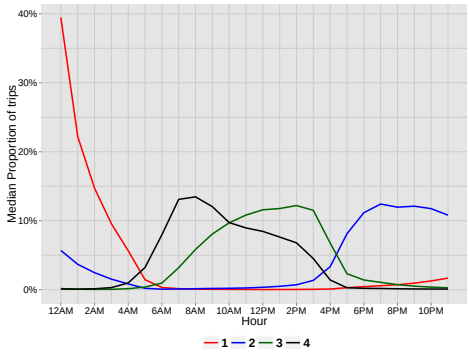
4 Results:

K-means clustering identified 4 different types of schedules, some more similar than others. Figure 3a shows that clusters 3 and 4 are more similar to each other than to the other clusters. They even overlap when plotted on the first two principal components because only some of the information is actually shown.

We can see that drivers tend to follow four different schedules: (1) only providing trips in the morning, (2) providing trips in the evening, with a few trips in the early morning, (3) starting later in the workday and carrying into the evening, and (4) starting in the early part of the workday, but ending earlier (Figure 3b). Also, each schedule cluster has a similar proportion of trips at their peak hours, with group 1 being the exception. However, as we can see in Figure 1a, less trips happen in



(a) The 4 clusters shown on the first two principal components.

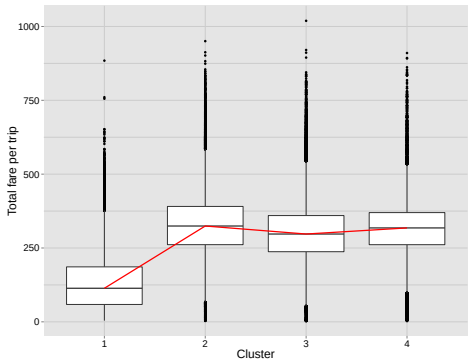


(b) The median proportion of trips in each cluster over the day.

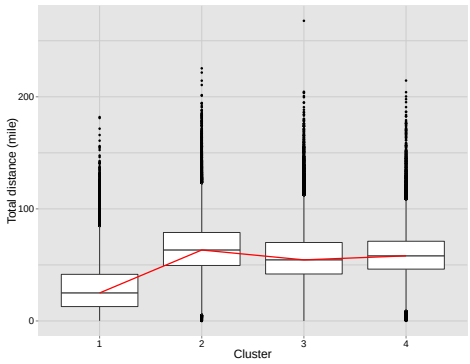
Figure 3

the early morning, so while a higher proportion of trips is happening in the early morning, there is not a higher number of trips.

Next, we compared the total fare and distances of trips in each cluster to determine if some schedules were more profitable than others. While clusters 2, 3, and 4 were indistinguishable, cluster 1 had a lower distribution of total fare (Figure 4a), but this is probably because less trips are occurring in the early morning (Figure 1a) and shorter trips were grouped into cluster 1 (Figure 4b).



(a) Total fare per trip by cluster



(b) The total distance driven by cluster.

Figure 4

5 Conclusion:

While our method did identify different driver schedules, we did face a few limitations. First, we could only utilize one week of data based on computing resources. Exploration and analysis of the whole data set would lead to more rigorous and accurate answers. It would also allow for comparisons between weekdays and weekends, which we noticed have slightly different trends, but were not able to treat differently. Second, we worked under the assumption that a shift began and ended in the same day. While this was true for clusters 3 and 4, and did not drastically affect group 2, it could have led to group 1 being identified wrongly. For example, for a driver that started their shift Monday night and ended Tuesday morning, their Tuesday hours would have been grouped into cluster 2 if they worked again Tuesday night, but would be in cluster 1 if they did not work again on Tuesday. Thus, we do not know if the drivers in cluster 1 actually only work in the early morning, or if they have the night shift and simply took a day off.

Better understanding of the dynamic daily supply of taxis has implications for drivers, the larger parent organization, and future modelers. Drivers can use this information to pick a schedule that best fits thier needs for distance and number of trips. Drivers might have a preference for short or long trips, which can be easily influenced by what shift they are driving. The parent organization can influence the supply of taxis at anytime by using this information to incentive certain schedules. Future modelers can use this to more accurately describe the supply of taxis. Currently, the supply of taxis is often modeled as a uniform distribution [1], however we have found that it is not (Figure 1a). A possibly more accurate model could be a mixture model, where the number of available taxis is draw from a mixture of the different distributions we have identified as driver schedules. As such, understanding endogenous taxi driver schedules is a complex and open problem requiring more study and analysis with potential benefits for all those involved.

Acknowledgments

References

[1] Hai Yang, Min Ye, Wilson Hon-Chung Tang, and Sze Chun Wong. A multiperiod dynamic model of taxi services with endogenous service intensity. Operations research, 53(3):501515, 2005

[2] James, Gareth, Daniela Witten, and Trevor Hastie. "An Introduction to Statistical Learning: With Applications in R." (2014).