Cluster Analysis of Endogeneous Taxi (NY Yellow Taxi) Driver Schedule Patterns

Pradeep Pendem and Jessica Leete

Kenan-Flagler Business School, University of North Carolina

Computational Biology and Bioinformatics, Duke University

Objectives

Understand with in a day dynamic supply of taxis through driver endogenous schedules:

- Identify intra-day patterns of taxi driver schedules
- Group similar schedule patterns using K-means clustering algorithm
- Identify and report (if any) sources of fare variation among clusters

Introduction

- Taxi drivers typically choose their working schedules or number of service hours in a day based on profitability and operating cost[1]
- Dynamics in the service system results in supply of taxis varying in a non-homogeneous pattern over time and space
- Traditionally taxi service systems are modeled to have uniform supply and demand by aggregating different time periods into one static period

Dataset

- First week of New York Yellow Taxi trip and fare data ; approx - 4 million records
- Merged trip and fare data by Driver ID, pickup date time
- Key variables - Driver ID, Pickup and Dropoff date & time, Distance (miles), Total fare etc
- Aggregated number of trips for each driver and day ; Normalized to create proportion of trips
- Interested in driver schedule patterns against to total trips
- Figure 1 and 2 show total trips (median) for weekday and weekend, variation in fare and distance (mil) by 1 hour
- Analysis was carried out in R-software, Charts with ggplot2 package

Data Visualization, Methods & Results

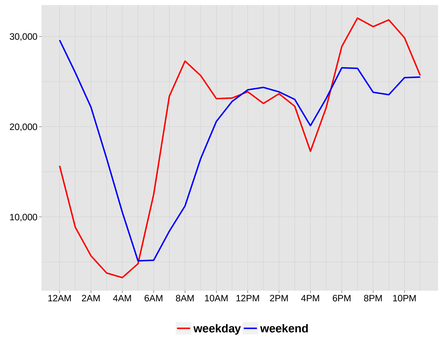

Figure 1: Median number of trips by each hour

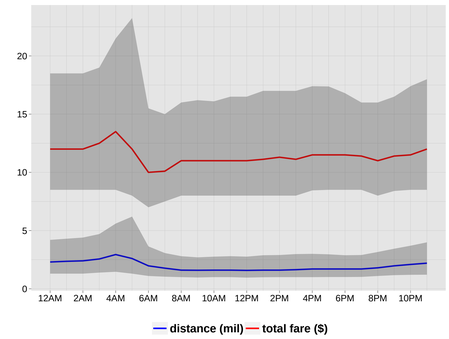

Figure 2: Median total fare and distance per trip (mil) by each hour

Important Result

Identified four unique Taxi driver working schedules patterns. Excluding one pattern, others show similar structure separated by a time lag.

- Each observation - Driver ID, day, proportion of trips in each 1 hour (total-24)
- PCA - More than 12 Principal components explain at least 80% variation ; Clustering on original data
- Optimization model for K-means ;

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

- K-means algorithm repeated with 1 to 15 clusters ; cut-off at 4; Error - Within Sum of Squares (WSS)

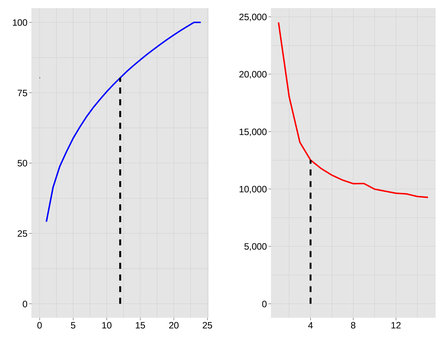

Figure 3: Cumulative variance of Principal components, WSS by cluster choice

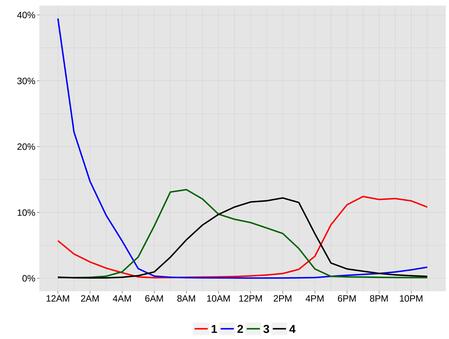

Figure 4: Mean proportion of trips by cluster in each hour

Results

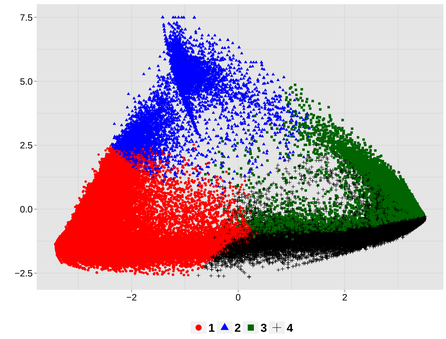

Figure 5: Clusters representation on first two principal components

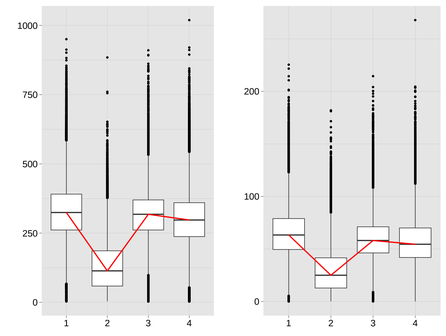

Figure 6: Total fare and distance boxplots by cluster

Conclusion

- Identified four different scheduling pattern of driver schedules
- Two clusters emulate mirror image pattern

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):501–515, 2005.