Australasian Transport Research Forum 2021 Proceedings
8-10 December, Brisbane, Australia
Publication website: http://www.atrf.info

An empirical study on characteristics of supply in e-hailing markets: A clustering approach

Yue Yang[1], Jacob Elmasry[1], Porsiem Tang[1], Mohsen Ramezani[1]

[1] University of Sydney, school of Civil Engineering

Email for correspondence: mohsen.ramezani@sydney.edu.au

Abstract

E-hailing services have disrupted how, when, and where people travel in cities. This paper characterizes the attributes of the supply of e-hailing markets that is reflective of the labor characteristics of the drivers (contractors). Based on a clustering analysis of the observed behaviour of an e-hailing company's drivers over a month, the analysis identifies three major groups of drivers: (i) part-time drivers working flexible hours, (ii) part-time drivers working in the evenings, and (iii) full-time drivers. The clustering results of the e-hailing market supply is verified to have consistent characteristics over different days. Ultimately, the analysis is used to predict the number of active drivers in the market during a day.

1. Introduction

The recent rise in mobility on-demand (MOD) companies such as Uber, Lyft, and Didi has disrupted the existing transport market. One approach to investigate the ride-sourcing services is through a two-sided market analysis with passengers exhibiting the desire to travel (the demand side) and drivers being willing to offer the service to transport them to their destination (the supply side). From the side of supply, ride-sourcing drivers are independent contractors who could work part-time or full-time and are under no obligation to be active in the market at any point in time. This paper offers insight on the supply side of the market to identify influential drivers characteristics and behaviour patterns such as when they join, when they leave, the number of shifts they work, and whether they work part-time or full-time.

Recent works modeled ride-sourcing driver behaviours based on empirical surveys (Ashkrof et al., 2020), economic analysis (Zha et al. 2018), network flows (Riascos and Mateos 2020), and machine learning (Zhao et al. 2020). Although ride-sourcing driver behaviours have received attentions in aforementioned existing works, few studies have offered refined analyses that consider the differences among various groups of ride-sourcing drivers.

The motivation of the paper is to develop a data-driven method to discover the characteristics and market-behavioural patterns of ride-sourcing drivers using a clustering approach. The data used in this paper has been provided by Didi Chuxing and includes anonymised trajectory records of drivers and all serviced orders in the city of Chengdu during November 2016. The data are cleaned by removing unrealistic data points. Six behavioural features are extracted for the clustering analysis. A k-means clustering method is then undertaken on two weeks of training data to discover different patterns of driver behaviour. The clustering result reveals that there exist three distinct driver groups: (i) part-time drivers working in flexible hours, (ii) part-time drivers working in evening hours, and (iii) full-time drivers. A detailed analysis of operational properties of the three clusters is provided. The characteristics of the three clusters

are then used to predict the number of active drivers available in the market within a day of testing data.

2. Data

The data used in this study is collected from the DiDi GAIA Initiative Project. The project shares the complete ride trajectory and order data of DiDi Express and DiDi Premier in the city of Chengdu, China, from November 1st to November 30th, 2016. There were approximately 1.2 million unique driver IDs and 6.1 million unique trip requests in the dataset. Driver IDs were re-anonymized each day, meaning that an individual driver cannot be tracked over multiple days. The orders in the GAIA dataset only represent trips that were successfully serviced. The data is cleaned by removing trajectories that show an outlier in travel distance, the average speed of vehicle, or travel time. An outlier is defined as a value that is not within the 99% range of the dataset for each of the variables.

Ride-hailing platforms provide the drivers with the freedom to choose their working hours due to the fact that they do not have direct employment relationships but are rather considered as independent contractors. To effectively capture the working duration of each driver, their operation period within a day can be segmented into one or more shifts. A shift starts once the driver serves a trip request, and it may contain one or more served orders. Once the time gap between serviced orders is greater than 2 hours, this is considered as a break between two shifts.

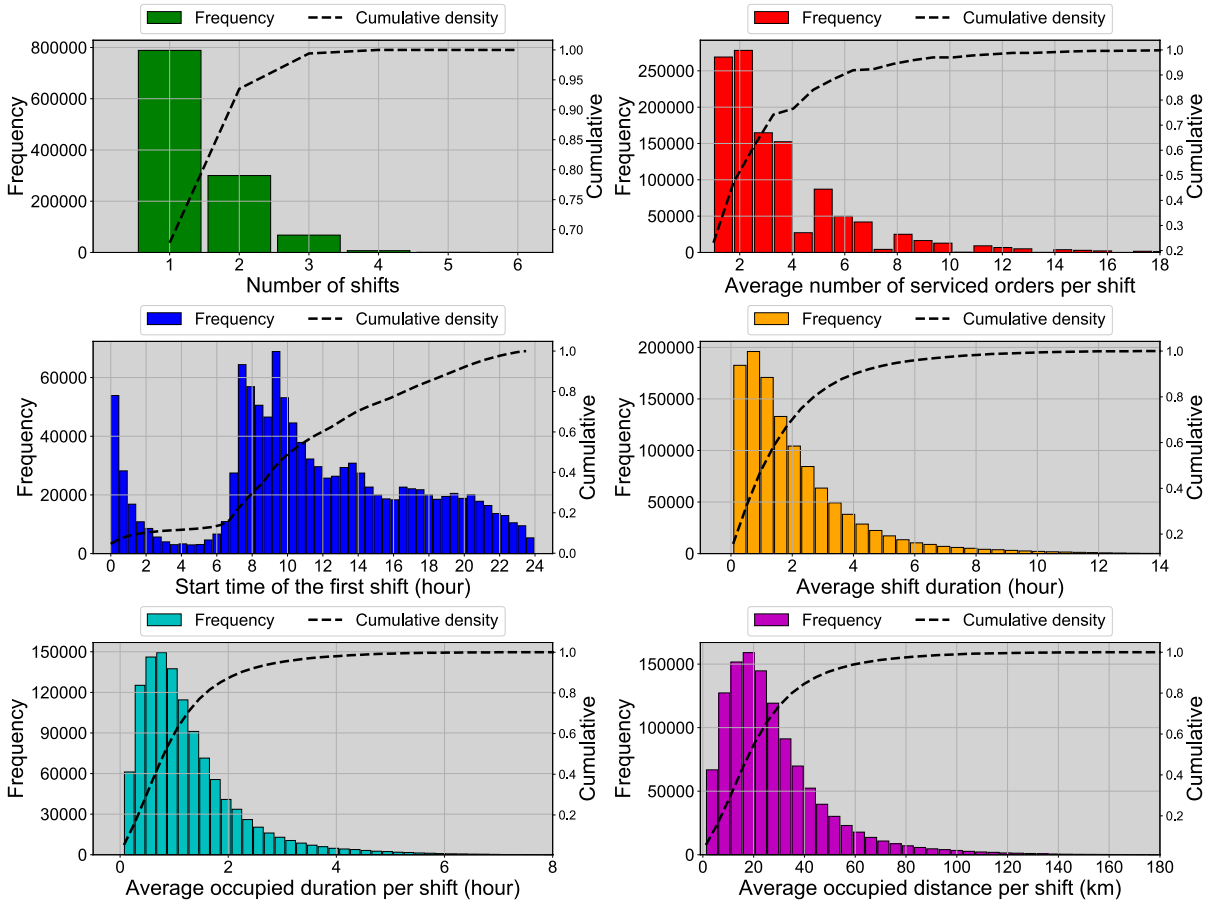

Figure 1: The double y-axes figures describe the distributions of six operation features. The first y-axis and histogram reflect the occurrence frequency of each feature, and the second y-axis and curve depict the cumulative density of each feature.

To identify the characteristics and patterns of different drivers, we introduce six operation features for each driver in a day: (i) Number of shifts, (ii) Average number of serviced orders

per shift, (iii) Start time of the first shift, (iv) Average shift duration, (v) Average occupied duration per shift, and (vi) Average occupied distance per shift. Fig. 1 presents the distributions of these six features in the whole dataset. The figure shows that over 60 % of the drivers only work one shift of 2-hour duration during which they serve 2-3 orders under 30 km cumulative distance for an hour (i.e. occupancy rate of 50 %). This indicates that most of the drivers are part-time contractors (at least for Didi).

3. Characteristics of Drivers: Clustering the Market Supply

K-means clustering is one of the most widely-used unsupervised learning methods for partitioning data into *k* clusters by minimizing the error function. Based on the training data (the first fourteen days of November 2016), the optimal number of clusters *k* is 3 by using the average silhouette width criterion (ASWC). Accordingly, the drivers are partitioned into three clusters by the k-means clustering method.

The mean values and standard deviations of the six features for each cluster are presented in Fig. 2. Drivers in Cluster A start their first shift early in the morning at an average of 8 AM, work nearly two separate short shifts, serving 2 to 3 orders in each shift. Drivers in Cluster B only work a short shift with an average of 2 to 3 orders, start their first shift late in the day with an average of around 6 PM. Drivers in Cluster C have one long shift with an average of 8 to 9 orders, start their first shift at late morning (around 10 AM). The three clusters represent three groups of drivers. In summary, Cluster A: Part-time drivers working flexible hours; Cluster B: Part-time drivers working in the evening; and Cluster C: Full-time drivers.

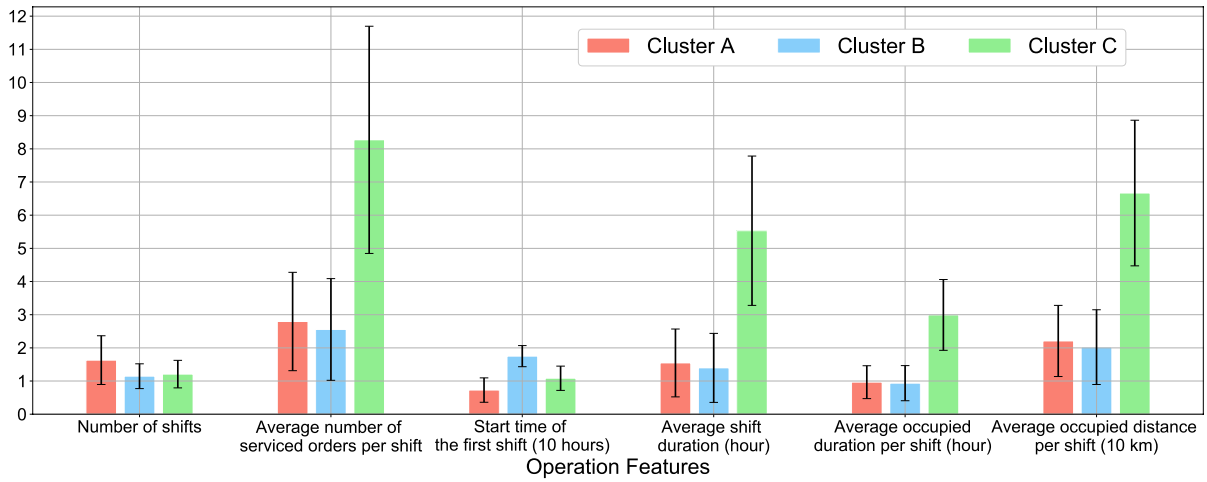

Figure 2: The mean value of the features of the three clusters; The whisker on the bar indicates the standard deviation of each feature in each cluster.

4. Supply Prediction: Forecasting the number of active drivers throughout the day

In this section, the number of active drivers in the ride-sourcing system during a day is predicted. The schematic of the overall procedure is summarized below:

(1) Firstly, the average number drivers for each group is generated and each driver is assigned to one of the three clusters.

(2) For a unique driver, the number of shifts and the start time of the first shift can be sampled from the estimated distributions.

(3) The subsequent procedure is a loop to sample the shift duration, update the shift end time, sample the gap to the next shift and update the start time of the next shift (if exists). The loop is terminated if the final shift has been updated.

(4) After computing all the shifts of a driver, the above process was repeated until all of the drivers have been considered.

Consequently, each day was split into 240 time slots (each slot is 6 minutes) as $T = \{1,\ldots,240\}$. If drivers are working a shift during a time slot, they are considered to be active drivers within that time slot. Four weekdays and two weekends are selected as the testing data. Fig. 3 compares the empirical results with the prediction results, which indicates that the behavior patterns of the drivers can be well characterized by the proposed method.

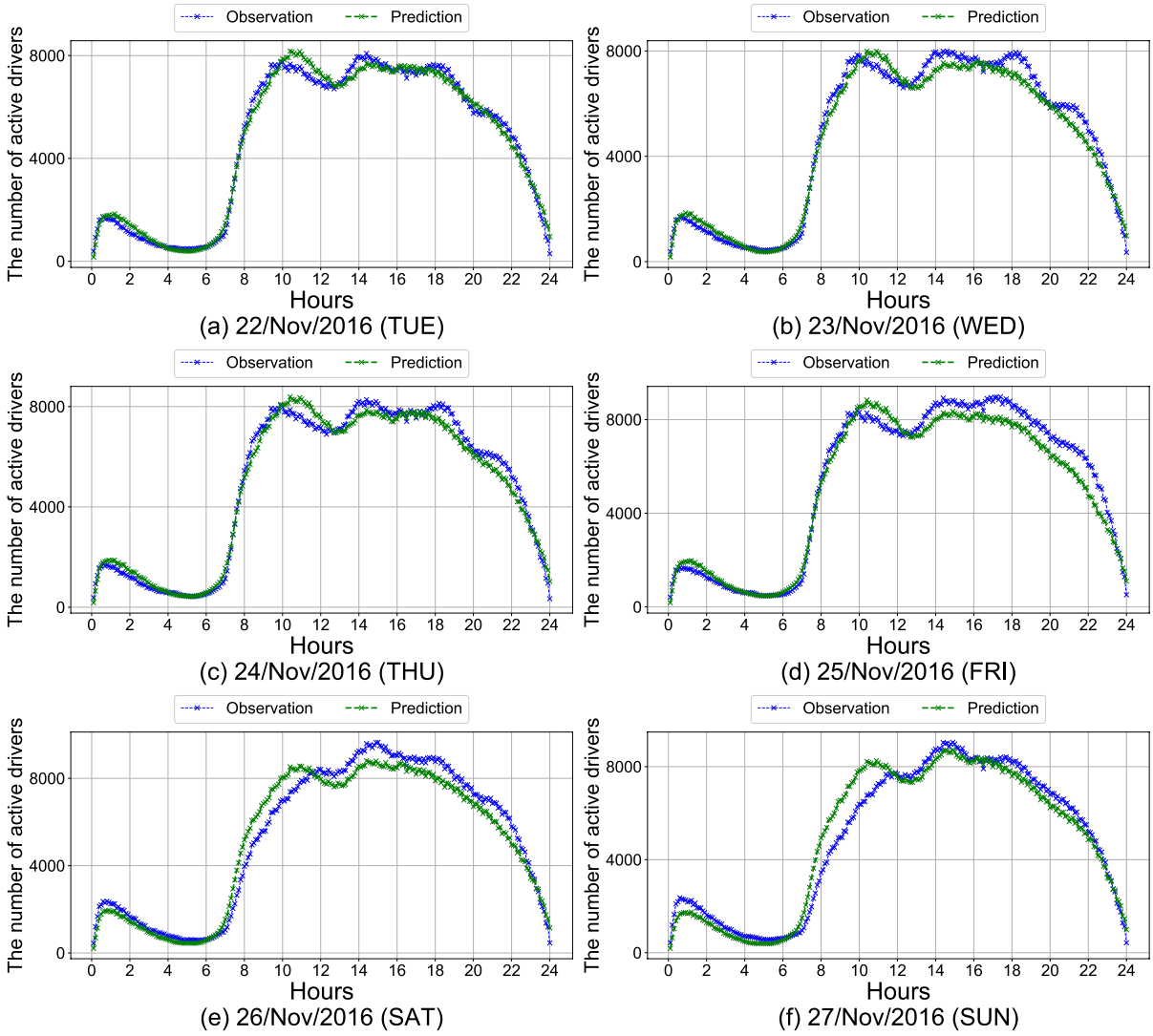

Figure 3: A comparison of the predictions of time-varying numbers of active drivers in the market with observed values.

4. Discussions

Based on the results of clustering and prediction methods, this section discusses policy recommendations of this study:

(1) **Driver incentivization**: In our result, nearly 85% of drivers are part-time from the perspective of Didi, but if they are contractors for multiple platforms driving might still be

their full-time occupation. These multi-homing ride-sourcing drivers could become loyal to one specific platform if they were offered wage and incentive programs in various formats. For instance, a time-varying wage (increasing the wage rate for drivers in peak hours) and a commission (decreasing the platform commission percentage during peak hours) could attract part-time drivers to stay in the platform and help curb the supply shortage during peak demand periods. Another scheme can be a joint spatial-temporal monetary incentive. This could be achieved by allocating a bonus for drivers who complete a repositioning instruction (vacant trip) to imbalanced supply-demand hotspots. This would target the over-supply in parts of the city and the supply shortage in other parts.

(2) **Congestion Management**: To curb congestion caused by ride-hailing services, distance-based taxes for ride-hailing vehicles and platforms can be evaluated by using the Clusters' characteristics determined by this study. Once vehicle-based distance-based taxes are imposed for ride-hailing drivers, drivers of Cluster C are most affected and may avoid a long cruising distance or shorten their working hours in over supply periods. For drivers of Clusters A who are full-time multi-homing drivers among multiple platforms, distance-based taxes might stimulate them to be full-time drivers in one platform to avoid getting lower priority (longer cruising distances) from switching platforms. As for Cluster B drivers, who have the least occupied distance per shift, will experience a negligible impact triggered by distance-based taxes.

(3) **Ridesharing**: To guarantee effective ridesharing matching, full time drivers of Cluster C play a considerable role to be matched to ridesharing requests as they spend more time in the market and offer the flexibility in their working hours to accommodate the longer trip and detours associated with ridesharing. This, in return, would increase Cluster C's occupation duration percentage.

5. Summary

The paper has analysed the behaviour of contractor drivers in a ride-hailing platform from the data provided by Didi Chuxing. After cleaning outlier data points, six operation features: (i) number of shifts, (ii) average number of serviced orders per shift, (iii) start time of the first shift, (iv) average shift duration, (v) Average occupied duration per shift, and (vi) Average occupied distance per shift were extracted and fed for clustering. Employing a k-means clustering method, three representative clusters of drivers are identified: (i) part-time drivers working in flexible hours, (ii) part-time drivers working in evening hours, and (iii) full-time drivers. This analysis provides a better understanding of the characteristics and market-behavioural patterns of ride-hailing drivers.

6. Reference

[1] Ashkrof, P., de Almeida Correia, G. H., Cats, O., & van Arem, B. (2020). Understanding ride-sourcing drivers' behaviour and preferences: Insights from focus groups analysis. Research in Transportation Business & Management, 37, 100516.
[2] Zha, L., Yin, Y., & Du, Y. (2017). Surge pricing and labor supply in the ride-sourcing market. Transportation Research Procedia, 23, 2-21.
[3] Riascos, A. P., & Mateos, J. L. (2020). Networks and long-range mobility in cities: A study of more than one billion taxi trips in New York City. Scientific reports, 10(1), 1-14.
[4] Zhao, P., Liu, X., Kwan, M. P., & Shi, W. (2020). Unveiling cabdrivers' dining behavior patterns for site selection of 'taxi canteen'using taxi trajectory data. Transportmetrica A: Transport Science, 16(1), 137-160.