

# Transportation Letters

The International Journal of Transportation Research

ISSN: (Print) (Online) Journal homepage: <https://www.tandfonline.com/loi/ytrl20>

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

Mohsen Ramezani, Yue Yang, Jacob Elmasry & Porsiem Tang

To cite this article: Mohsen Ramezani, Yue Yang, Jacob Elmasry & Porsiem Tang (2022): An empirical study on characteristics of supply in e-hailing markets: a clustering approach, Transportation Letters, DOI: [10.1080/19427867.2022.2079869](https://doi.org/10.1080/19427867.2022.2079869)

To link to this article: <https://doi.org/10.1080/19427867.2022.2079869>



Published online: 26 May 2022.



Submit your article to this journal [↗](#)




View related articles [↗](#)



View Crossmark data [↗](#)

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

Mohsen Ramezani , Yue Yang, Jacob Elmasry and Porsiem Tang

School of Civil Engineering, The University of Sydney, Sydney, NSW, Australia

## 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 behavior 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. The results of the clustering method are demonstrated to be effective for prediction of supply.

## KEYWORDS

Mobility On-Demand; driver behavior; contractors; supply prediction

## Introduction and motivation

The recent rise in mobility on-demand (MOD) companies, such as Uber, Lyft, and Didi, has disrupted the existing transportation market. The representative ride-hailing services provided by these companies has expanded dramatically in their decade-long history (Shaheen et al. 2016). In September 2018, Uber reached a milestone of 10 billion trips worldwide, up from 140 million trips in 2014 (Uber 2019), Lyft accumulated more than 1 billion trips (Lyft 2019), and Didi also provided services for over 25 million trips on each day of 2018 in China (Xu et al. 2018). Meanwhile, self-driving technology and autonomous vehicles are also gradually deployed in these on-demand transportation services to achieve a higher trip requests' responsiveness (Chen, Valadkhani, and Ramezani 2021). It has been found that these app-exclusive ride-sourcing services (Nair et al. 2020) have significantly decreased the demand for taxis, specifically among younger, more affluent people (Contreras and Paz 2018; Young and Farber 2019; Tirachini and Del Río 2019). Additionally, ride-sourcing services have been found to both compete with (Tirachini and Del Río 2019) and complement (Shaaban and Kim 2016; Su, Nguyen-Phuoc, and Johnson 2021; Shen, Zhang, and Zhao 2018; Aghaabbasi et al. 2020; Chen and Nie 2017; Nocera et al., 2021) public transportation services (i.e. city buses, trams, and trains).

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) (Wang and Yang 2019). Research in this field has traditionally focused on the demand side of ride-sourcing services (Vij 2020; Su, Nguyen-Phuoc, and Johnson 2021; Lavieri and Bhat 2019; Aghaabbasi et al. 2020; Tirachini and Del Río 2019; Young and Farber 2019; Hamedmoghadam, Ramezani, and Saberi 2019) with it being stated that the supply side will rise to meet the demand because of monetary incentives in the market (Henao and Marshall 2019; Button 2020; Vij 2020). Part of this expectation of sufficient driver supply stems from the lack of regulation as there are frequently no limits on the number of drivers, no regulation on surge-pricing, and no requirements for drivers to be officially

trained (Ke et al. 2019; Harding 2016). However, the current trends of increasing supply may not be sustainable congestion-wise (Alisoltani, Leclercq, and Zargayouna 2021; Beojone and Geroliminis 2021), while it is revealed that ride-sourcing drivers are frequently earning less than originally advertised, occasionally even below minimum wage (Henao and Marshall 2019). Furthermore, the driver supply and passenger demand cannot be modeled as always being in equilibrium since the market needs time to adjust due to temporal and spatial dynamic natures of the travel demand and vehicle supply (Nourinejad and Ramezani 2020). On top of this, ride-sourcing drivers are independent contractors who could be working part-time or full-time and are under no obligation to be active in the market at any point in time. Thus, more research is required on the supply side of the market to identify influential drivers characteristics and behavior patterns, such as when they join, when they leave, the number of shifts they work, and whether they work part-time or full-time. This paper offers new empirical contributions to answer these questions.

Previous works modeled ride-sourcing driver behaviors based on empirical surveys (Ashkrof et al. 2020), economic analysis (Zha, Yin, and Du 2018), clustering methods (Ma et al. 2019), network flows (Riascos and Mateos 2020), and machine learning (Zhao et al. 2020). (Chen, Zahiri, and Zhang 2017) carried out descriptive statistics on various characteristics of ride-hailing demand and supply patterns. Based on a large-scale trajectory data from Shenzhen, (Nie 2017) investigated the impact of ride-hailing platforms on the traditional taxi drivers. According to the differences in the distributions of vehicle travel time periods and origin-destination (OD) points, (Dong et al. 2018) compared travel service patterns and driver behavior patterns of taxi and internet-based ride-sharing services. (Xu et al. 2020) conducted an empirical study on the working hours of ride-sharing drivers, with often exclusive focus on a driver's reaction to their income levels and whether they follow a neo-classical (Farber 2008, 2015) or income-targeting models (Köszegi and Rabin 2006; Crawford and Meng 2011). Considering the structural supply deficits/surpluses existing in the ride-sourcing market, (de Ruijter et al. 2022) investigated system-level factors may influence individual driver's labor decisions.

Although ride-sourcing driver behaviors 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 this paper is to develop a data-driven clustering method (i) to discover the characteristics and market-behavioral patterns of ride-sourcing drivers and (ii) to be incorporated for supply prediction. The data used in this paper has been provided by Didi Chuxing and includes anonymized trajectory records of drivers and all serviced orders in the city of Chengdu during November 2016. The data is cleaned by removing unrealistic data points. Six behavioral 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 behavior. 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 identified characteristics of the three clusters are used to predict the number of active drivers in the market within test days. The prediction method demonstrates a promising accuracy.

The remainder of the paper is structured as follows: Section 2 describes the data used in this work. Section 3 introduces the clustering method to categorize drivers into groups. The characteristics of each group of drivers are explored. Section 4 builds upon the results of drivers clustering and introduces a method to predict the number of active drivers in the network over a day. Section 5 offers some policy recommendations and discusses the limitations of the analysis. Finally, Section 6 summarizes the study and discusses potential extensions for future work.

## Data

The data used in this study is collected from the DiDi GAIA Initiative Project.<sup>1</sup> The project shares the complete ride trajectory and order data of DiDi Express and DiDi Premier, two of DiDi Chuxing's primary ride-sourcing services, in the city of Chengdu, China, from November 1 to November 30, 2016. The trajectory data recorded anonymized driver IDs and order IDs as well as the Unix timestamp, latitude and longitude of each driver approximately every 3 seconds (a sample of this data is shown in Table A1 in Sec: Appendix). The order data includes order ID, order start and stop times, pick-up latitude and longitude, and drop-off latitude and longitude for each order (a sample of this data is shown in Table A2 in Sec: Appendix).

In November 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. This means that there could be unserved order requests, which might be significant during specific times of the day (i.e. peak hours). 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. Specifically, the variable travel distance should fall in the range of 1.1 [km] – 29.7 [km], the average speed of a vehicle should be between 0.7 [km/h] – 54.5 [km/h], and the travel time must be within the range of 4.1 [min] – 72.4 [min]. After data cleaning, 4.7% of the data are marked as outliers and removed from the dataset.

Figure 1 presents the daily number of unique drivers and trip orders. There are approximately 36,000–44,000 unique drivers and 180,000–220,000 orders every day in Chengdu. The figure illustrates a weekly pattern of the number of drivers and orders; Fridays and Saturdays have the most drivers and orders, while Mondays have the least number of drivers and orders. Figures 2 and 3 show daily and hourly trends of average trip length and average travel time. The daily average trip length falls into the range of 8.1 [km] – 9.5 [km] and the daily average trip time is in the range of 21.0 [min] – 24.0 [min]. The hourly average trip length falls into the range of 7.6 [km] – 11.0 [km] and the hourly average trip time is in the range of 15.0 [min] – 27.0 [min]. Interestingly, these demonstrate a relatively stable day-to-day pattern with significant variations within day.

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. We define the gap between the sequential served orders as the time gap between the drop off and pick up of two successive serviced passengers. As shown in Figure 4, we can observe that over 95% of the gaps between two serviced orders are shorter than 120 [min]. We also observe a very spread tail in the distribution. To partition the activity timeline of drivers into multiple shifts, we chose the threshold of 2 [h]. That is once the time gap between serviced orders (between the drop off and pick up of two successive serviced passengers) is greater than 2 hours, this is considered as a break between two shifts. In other words, if the time gap to the next trip

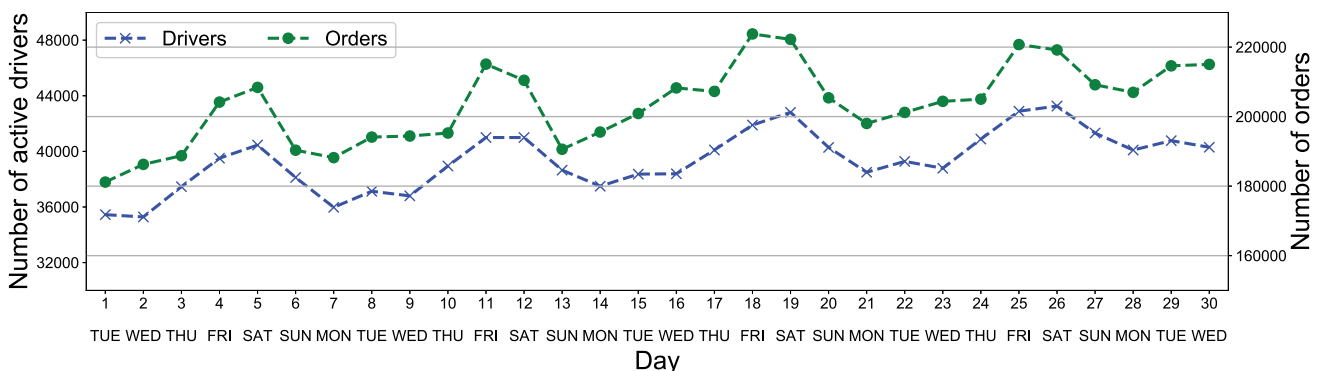


Figure 1. Daily number of active drivers and trip orders using Didi's services in Chengdu in November 2016.

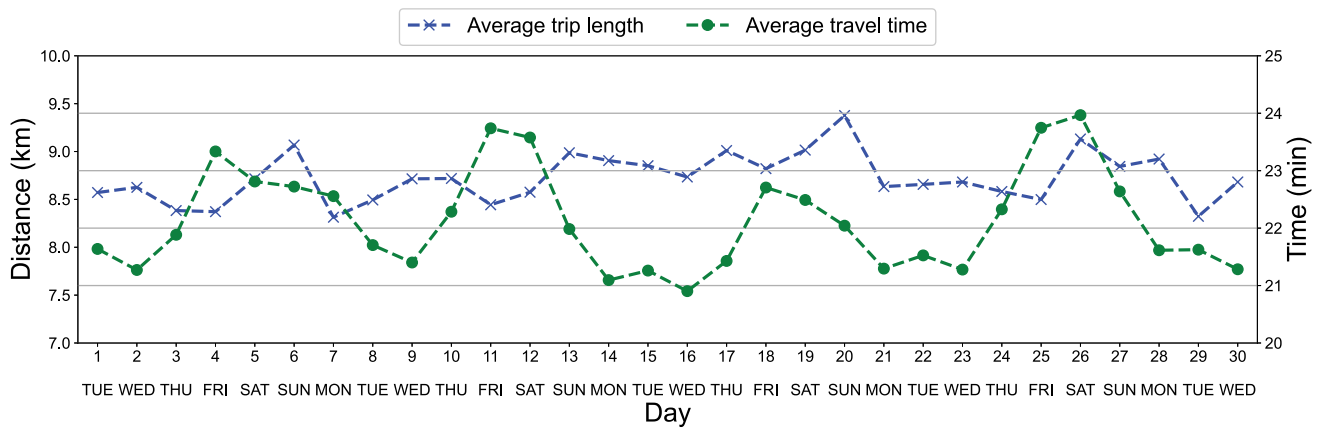


Figure 2. Daily average trip length and travel distance in Chengdu in November 2016.

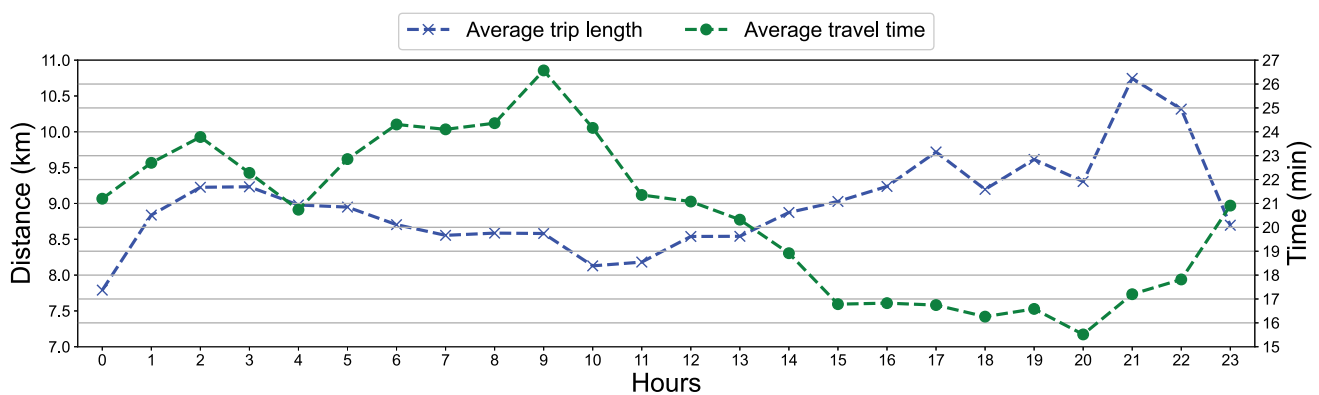


Figure 3. Hourly average trip length and travel distance in Chengdu in November 2016.

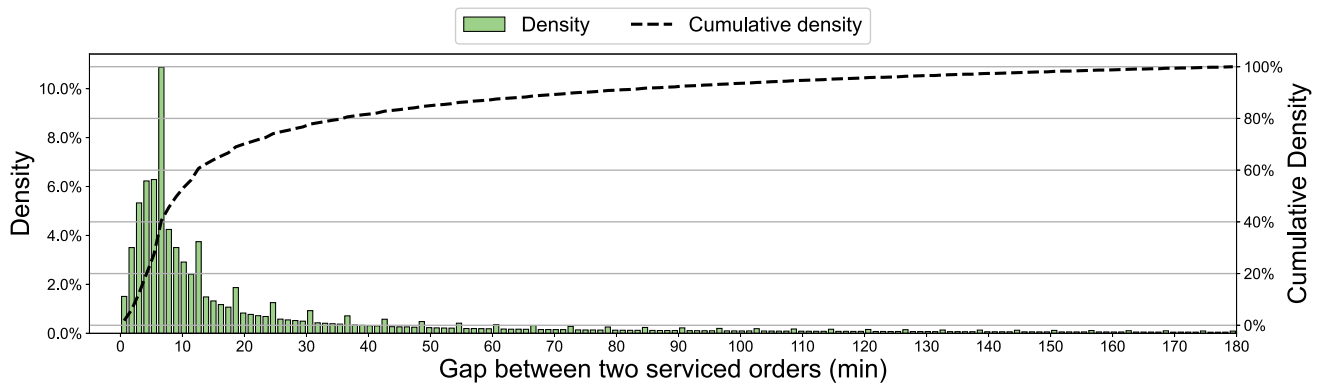


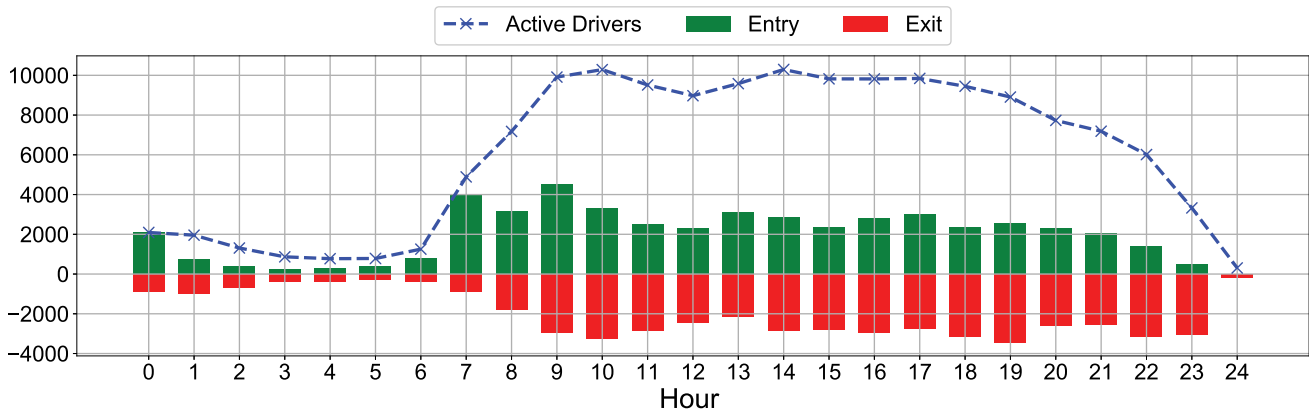
Figure 4. The observed distributions of the gap between two sequential orders.

order is greater than 2 hours, the drop-off time of the order is regarded as the end time of the current shift. Further, if the drivers end their shift, they will be considered to leave the market until their next shift starts.

Figure 5 illustrates the hourly numbers of active drivers in the market, the number of entries to the market (drivers starting their shift), and the number of exits from the market (drivers ending their shift) on November 8 2016 (Wednesday). A (relatively) stable trend of the number of active drivers is observed during 09:00 AM – 06:00 PM. However, the hourly number of entries and exits from the market changes more significantly. This should be considered in the equilibrium analysis of the market. Another noteworthy observation is the sharp increase and decrease in the number of active

drivers in the morning (06:00 AM – 09:00 AM) and evening periods (09:00 PM – 12:00 AM). This shows considerable diurnal change in the market supply. Thus, an accurate prediction of the supply is a challenging task. Section 4 introduces a method to predict the number of active drivers in the market that can be a foundation for tackling many issues in ride-hailing systems (e.g. supply shortage, vehicle dispatching, wage and fare management, etc.).

Flexibility, freedom, and independence were acknowledged by all drivers as the main motivations for joining ride-hailing platforms (Ashkrof et al. 2020). In general, drivers can independently decide when and where to start and finish their shifts. As a consequence, there exist different types of drivers in the platform, namely full-time drivers and part-time drivers. To identify the

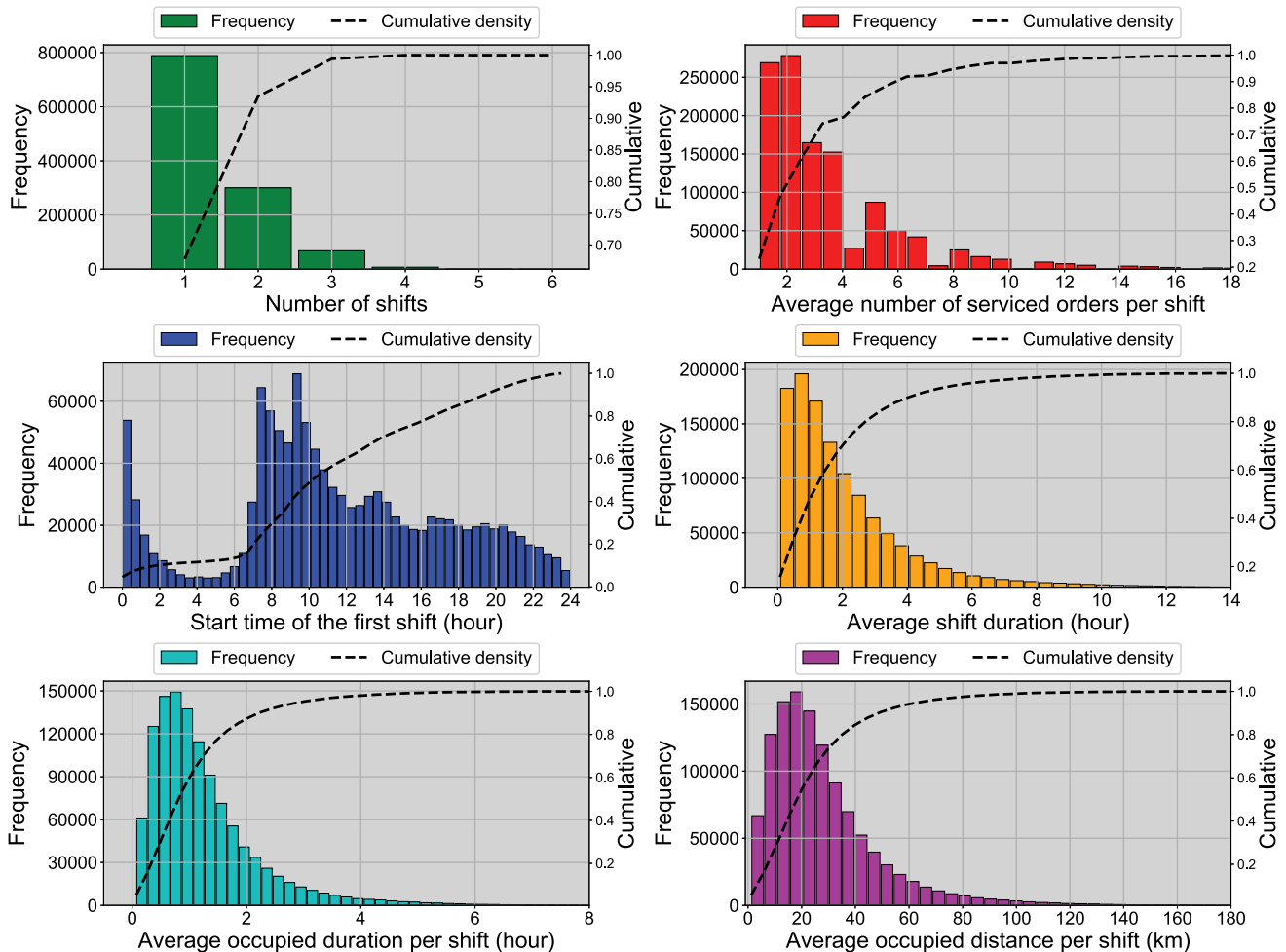


**Figure 5.** The hourly number of active drivers in the market (blue), the number of drivers entering to the market (green), and the number of drivers exiting from the market (red) on November 8 2016 (Wednesday).

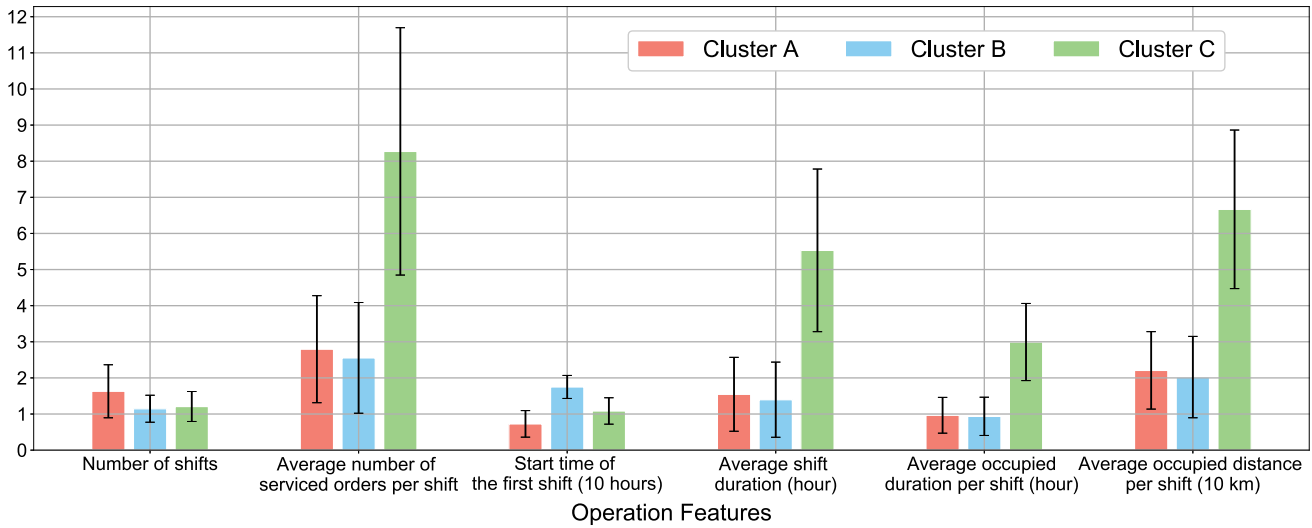
characteristics and patterns of different drivers, we introduce six operation features for each driver in a day: i) Number of shifts, ii) Average number of serviced orders per shift, iii) Start time of the first shift, iv) Average shift duration, v) Average occupied duration per shift, and vi) Average occupied distance per shift. Figure 6 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-h 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). According to the above features, Section 3 proposes a clustering method to identify different behavioral patterns of drivers and categorize them into different groups.

Although the distributions of the six features are shown independently in Figure 6, these features are correlated to various extents. To investigate their combined relationship with the



**Figure 6.** 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.



**Figure 7.** The mean value of the features of the three clusters; Cluster a: part-time drivers working flexible hours; Cluster b: part-time drivers working evening hours; Cluster c: full-time drivers. The whisker on the bar indicates the standard deviation of each feature in each cluster. Apart from the above six features, the mean values of occupied duration percentage (defined as the percentage of occupied duration in operated duration) are also calculated: Clusters a, b, and c have occupied duration percentages of 72%, 78%, and 56%, respectively.

characteristics and patterns of the drivers' behavior and predict the number of active drivers, the first 14 days of November 2016 of GAIA dataset are selected as the training data for clustering analysis in Section 3.

### Characteristics of drivers: clustering the market supply

In this section, the main objective is to identify and cluster the characteristics and market-behavioral patterns of the drivers. Drivers' behavior patterns stem from their characteristics (i.e. being full-time or part-time) and are coupled with their day-to-day and within day decisions. These decisions are mainly shaped around when they enter and exit the market and more, specifically, about the number of shifts within a day, the start time of a shift, and the duration of a shift. Based on the six features mentioned in Section 2 and the training data, a clustering method is employed to partition the drivers into a non-predefined number of groups. Considering the computational efficiency, a k-means clustering method (Hartigan and Wong 1979) is selected to categorize drivers and analyze the characteristics of each cluster.

K-means clustering is one of the most widely used unsupervised learning methods for partitioning data into  $k$  clusters by minimizing the error function. The number of clusters,  $k$ , has to be determined exogenously before performing the clustering of the dataset. In this paper, the average silhouette width criterion (ASWC) is used for the selection of the optimal number of clusters (Rousseeuw 1987). In general, ranging from  $-1$  to  $+1$ , ASWC demonstrates how well the objects (drivers) are grouped into the clusters. Higher ASWC values indicate a higher quality of the clustering process in terms of within-cluster homogeneity and between-cluster separation. Assuming that the data has been clustered in  $k$  clusters, for

each data point  $x_i$  in cluster  $C_i$ , the silhouette coefficient of data point  $x_i$ ,  $S_{x_i}$ , is calculated as follows (Kaufman and Rousseeuw 2009):

$$S_{x_i} = \frac{b_{x_i} - a_{x_i}}{\max(b_{x_i}, a_{x_i})} \quad (1)$$

where  $a_{x_i}$  is the average distance between  $x_i$  and other data points in the same cluster and  $b_{x_i}$  is the minimum average distance between data point  $x_i$  and data points in any other clusters. Then, the average silhouette value of Cluster  $j$  and the average silhouette value of all clusters can be calculated using equations:

$$SWC_j = \frac{1}{n} \sum_{i=1}^{n_j} S_{x_i} \quad (2)$$

$$ASWC = \frac{1}{k} \sum_{j=1}^k SWC_j \quad (3)$$

where  $n_j$  is the number of data points in Cluster  $j$  and  $k$  is the number of total clusters.

Based on the training data, the ASWC scores of different numbers of clusters are provided in Table 1. It can be seen that the ASWC score of  $k = 3$  provides the best k-means clustering results, and therefore the data indicates a natural categorization of the drivers into three clusters.

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 Figure 7. 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, and have nearly a 72% occupied duration percentage. 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, and have nearly a 78% occupied duration percentage. 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), and have nearly a 56% occupied duration percentage. 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.

**Table 1.** Average silhouette width criterion (ASWC) of k-means clustering with different numbers of clusters. Higher values represent better clustering performance. The optimal ASWC value indicates that there are three groups of drivers, which characterize the supply of ride-hailing markets.

$k$	2	3	4	5	6	7	8	9	10
ASWC	0.479	0.573	0.451	0.415	0.396	0.379	0.356	0.347	0.338



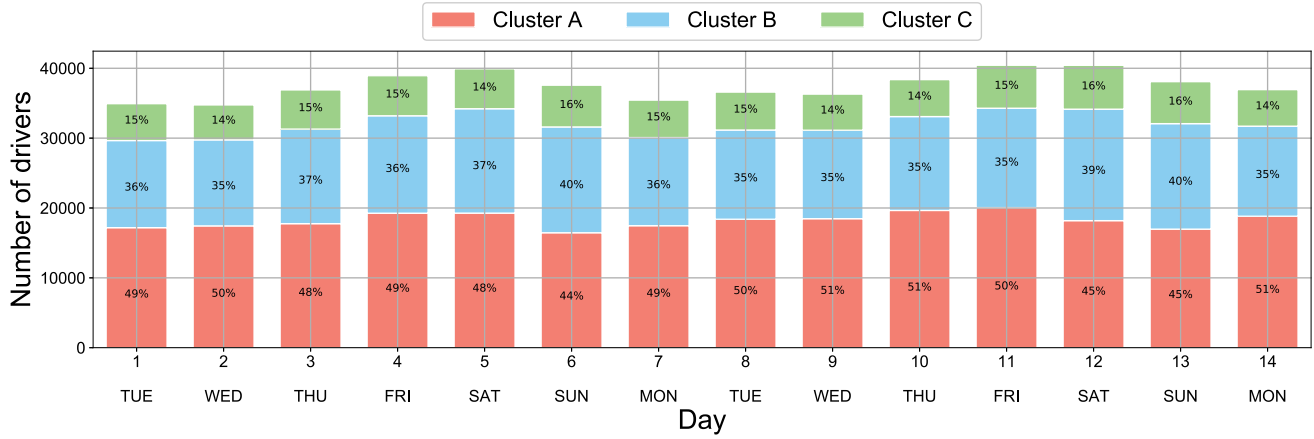


Figure 8. The absolute number and percentage of the identified drivers in each of the three clusters from the training data. Clusters a and b indicate the part-time drivers.

Figure 8 presents the percentage and absolute number of drivers in each cluster for the training days. The percentages show a very similar and consistent trend across the 14 days including the week-ends, indicating that the clustering result is representative. Drivers of Clusters A, B, and C are approximately 50%, 35%, and 15% of the supply in each day, respectively. Therefore, part-time drivers make up 85% of total supply of contractors on average in a day.

### Supply prediction: forecasting the number of active drivers throughout the day

In this section, we present a method based on the proposed clustering analysis to predict the time-varying number of active drivers during a day. Using the training data, the statistical estimations of the following features are first presented: (i) The number of shifts, (ii) Start time of the first shift, (iii) Shift duration, and (iv) Gap (time) between two shifts. The first two items have been introduced in the previous section. Note that each cluster of drivers exhibits different patterns of these four features. After estimating the probability distribution of the four features, a method to generate a prediction of the number of active drivers in the ride-sourcing market during a day is proposed. Finally, the predicted results are compared with the empirical observations from six testing days (four weekdays and two weekends).

### Statistical estimation

#### The number of shifts

The number of shifts performed by the drivers follows a discrete probability distribution. Let  $s$  represent the number of shifts. We approximate the probability mass function (PMF) as below:

$$\begin{aligned} p_A(s=i) &= \frac{n_A(s=i)}{n_A}, \quad i \in \{1, 2, 3, 4\} \\ p_B(s=i) &= \frac{n_B(s=i)}{n_B}, \quad i \in \{1, 2, 3, 4\} \\ p_C(s=i) &= \frac{n_C(s=i)}{n_C}, \quad i \in \{1, 2, 3, 4\} \end{aligned} \quad (4)$$

where  $p_A(s=i)$ ,  $p_B(s=i)$ , and  $p_C(s=i)$  are the probabilities that drivers of Clusters A, B, and C have  $i$  shifts in a day, respectively. Particularly,  $n_A$ ,  $n_B$ , and  $n_C$  are the number of drivers of Clusters A, B, and C, respectively, and  $n_A(s=i)$ ,  $n_B(s=i)$ , and  $n_C(s=i)$  are the number of drivers of Cluster A, B, and C with  $i$  shifts, respectively. According to the clustering results, the estimation of the PMF of the number of shifts is presented in Table 2.

Table 2. The estimation of the probability mass function (PMF) for the number of shifts.

$s$	Cluster A	Cluster B	Cluster C
1	0.51	0.86	0.80
2	0.37	0.13	0.20
3	0.11	0.01	0.00
4	0.01	0.00	0.00

#### Start time of the first shift

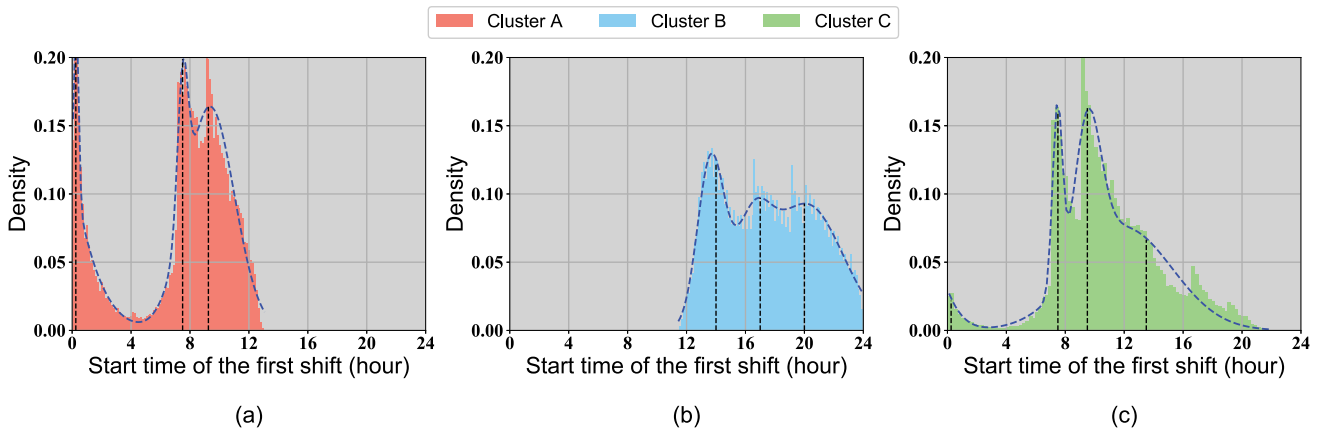
Figure 9 presents the density distribution of start time of the first shift from the training data for the three identified clusters of drivers. It intuitively suggests that histograms of all three clusters are asymmetrical and have distinct peaks during the day. Cluster A shows three peaks. The centers of three peaks are near 00:30 AM, 07:30 AM, and 09:30 AM. Cluster B also shows 3 peaks with centers roughly at 02:00 PM, 04:30 PM, and 08:00 PM. Cluster C displays four peaks corresponding to four centers at 00:30 AM, 07:30 AM, 09:30 AM, and 02:00 PM. Therefore, it can be assumed that the start time of the first shift follows a multi-modal distribution.

A one sample Kolmogorov–Smirnov (K-S) test is a powerful tool for testing if a random variable (the start time of the first shift in this case) follows a given distribution (Goodman 1954). The null hypothesis ( $H_0$ ) of the K-S test was conjectured as the observed distribution of the start time of the first shift follows a multi-modal Gaussian distribution. The fitting parameters are estimated by the trust-region reflective least-squares algorithm (Vogel 2002). Table 3 shows the result of K-S test for a different number of modes with a significance level of 95%. All of the p-values in Table 3 are significantly greater than 0.05, which indicates that  $H_0$  is accepted for all of the clusters. According to the minimum test statistic and the maximum p-value, three, three, and four are selected as the number of modes for Clusters A, B, and C, respectively. The fitted multi-modal distributions of three clusters are shown as the blue dashed lines in Figure 9.

Table 3. One sample K-S test of start time for the first shift for a multi-modal Gaussian distribution.

Number of modes	Cluster A		Cluster B		Cluster C	
	D	p-value	D	p-value	D	p-value
2	0.14	0.26	0.11	0.54	0.18	0.17
3	0.08 (**)	0.89 (**)	0.04 (**)	0.95 (**)	0.09	0.73
4	0.09	0.74	0.09	0.76	0.08 (**)	0.89 (**)

\*\* : The minimum test statistic  $D$  and the maximum p-value in the column.



**Figure 9.** The observed distributions of the start time of the first shift of drivers in the three clusters. Blue dash lines represent the fitted multi-modal distributions for each cluster. Black dash lines indicate the peaks of the modes in each cluster.

### Shift duration

Figure 10 presents the joint distribution of shift duration and shift start time for the three identified clusters. This figure suggests that shift duration is strongly correlated to the start time of the shift. Taking Cluster A as an example, if drivers start a shift in the morning period (i.e. 08:00 AM – 12:00 PM), the shift often lasts for a long time (up to a maximum of 6 hours), while if they start a shift in late evening (i.e. 08:00 PM – 12:00 AM), the shift is often short (less than 1 h).

To estimate the distribution of shift duration as a function of the shift start time, a discrete time horizon  $ST = \{st_1, \dots, st_n\}$  is considered to represent the range of the start times of the shifts.

**Table 4.** Average results of one sample K-S test for the shift duration distribution estimations.

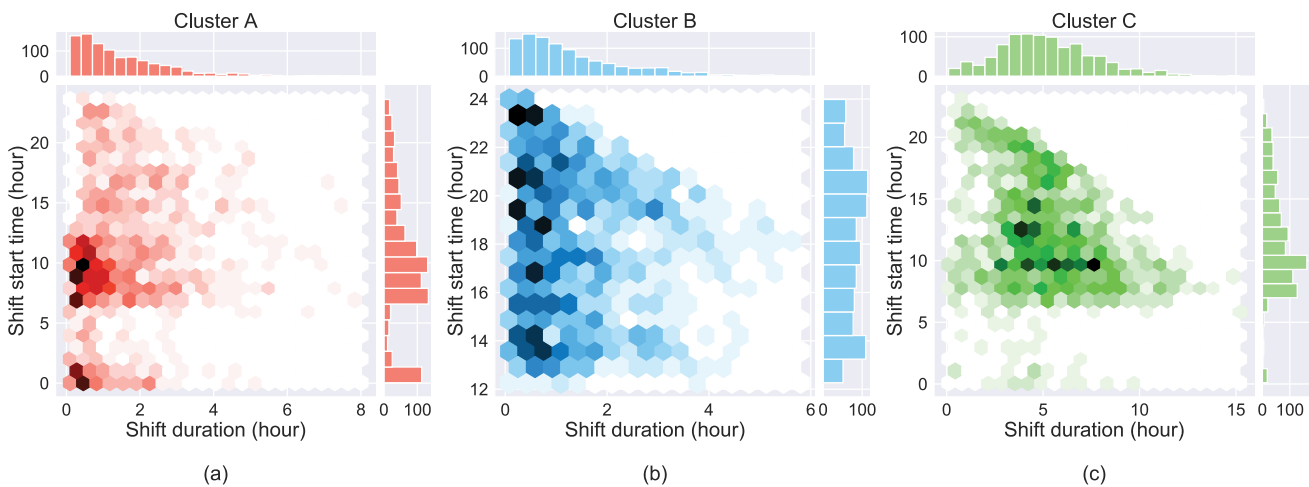
Distribution	Cluster A		Cluster B		Cluster C	
	D	p-value	D	p-value	D	p-value
Gaussian	0.27	0.81	0.26	0.82	0.23 (**)	0.84 (**)
Gamma	0.24 (**)	0.89 (**)	0.19 (**)	0.85 (**)	0.29	0.76
Lognormal	0.23	0.86	0.31	0.84	0.25	0.81
Weibull	0.26	0.84	0.28	0.84	0.31	0.74

\*\* : The minimum test statistic  $D$  and the maximum p-value in the column.

Without loss of generality, we split  $ST$  into 240 segments ( $n = 240$ ), that is, for each shift start time segment (every 6 minutes) a shift duration distribution will be estimated. Assuming these 240 distributions follow the same distribution type, the parameters of each one of them can be estimated from the training data. Table 4 presents the average results of one sample K-S tests over the 240 distributions. According to the minimum test statistic  $D$  and the maximum p-value, Gamma, Gamma, and Gaussian are selected as the probability distributions representing the shift durations of drivers in Clusters A, B, and C, respectively.

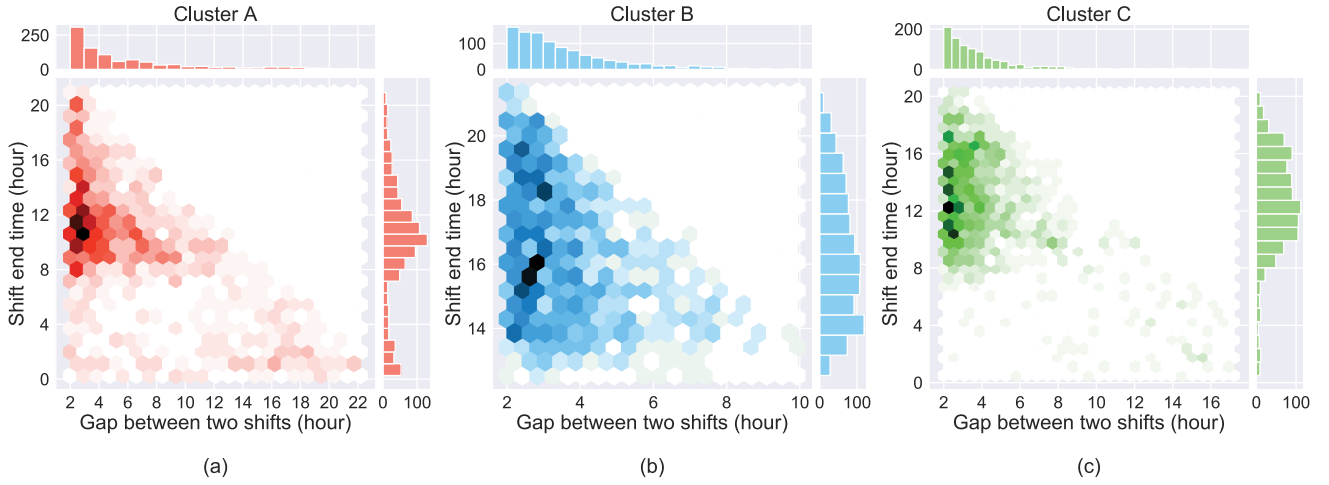
### Gap between two shifts

Figure 11 depicts the joint distribution of the gap between each pair of two consecutive shifts and the end time of the first shift for the three identified clusters. This figure illustrates that the gap between two shifts is strongly correlated to end time of the first shift. Taking cluster A as an example, if a driver ends a shift in the morning (i.e. 08:00 AM – 12:00 PM), the gap to the next shift can often be a long time (up to a maximum of 15 hours). However, since the drivers' activities of the next day cannot be tracked, if drivers end a shift in evening hours (i.e. 08:00 PM –



**Figure 10.** The joint distributions of shift duration and shift start time for each cluster. The darker hexagons indicate a higher density. The marginal distributions of shift duration and shift start time are shown at the top and the right of each figure respectively. Cluster a drivers start their shift in the range of 00:00–24:00 and work 0–8 hours per shift. Because cluster a represents the group of drivers who have a higher chance of working 2 shifts, the distributions in Figure 9(a) and Figure 10(a) show a substantial difference. Cluster b drivers start their shift in the range of 12:00–24:00 and work 0–6 hours per shift. Cluster c drivers start their shift in the range of 00:00–24:00 and work 0–15 hours per shift.





**Figure 11.** The joint distribution of shift gap and shift end time for each cluster. The darker hexagons indicate higher densities. The marginal distributions of shift gap and end time of the first shift are shown at the top and the right of each figure, respectively. Drivers in cluster A end their shift between 12 AM–09 PM and have a 2–22 hour gap to the next shift. Drivers in cluster B end their shift between 12 PM–10 PM and have 2–10 hour gaps to the next shift. Drivers in cluster C end their shift between 12 AM–08 PM and have a 2–17 hour gap to the next shift.

**Table 5.** Average results of a one sample K-S test for the gap between two shifts.

Distribution	Cluster A		Cluster B		Cluster C	
	D	p-value	D	p-value	D	p-value
Gaussian	0.31 (**)	0.79 (**)	0.36	0.74	0.39	0.69
Gamma	0.38	0.73	0.22 (**)	0.86 (**)	0.21 (**)	0.89 (**)
Lognormal	0.37	0.75	0.27	0.81	0.25	0.81
Weibull	0.39	0.73	0.21	0.83	0.22	0.86

\*\* : The minimum test statistic  $D$  and the maximum p-value in the column.

12:00 AM) and have a subsequent shift, the gap to the next shift is often very short because the start time of the next shift should be earlier than 12:00 AM.

Subsequently, we introduce a discrete time set  $ET = \{et_1, \dots, et_m\}$  to represent the range of end times of the first shift. After splitting  $ET$  into 240 segments ( $m = 240$ ), a distribution of the gap between two shifts is estimated for each of the 240 segments of the end time of the first shift. Assuming these 240 distributions of gaps between two shifts follow the same type of distribution, the parameters of each distribution can be estimated. Table 5 shows the average result of a one sample K-S test over the 240 distributions. According to the minimum test statistic  $D$  and the maximum p-value, the Gaussian, Gamma, and Gamma distributions are selected as the probability distribution of the gap between two shifts of the drivers in clusters A, B, and C, respectively.

### Simulation-based prediction

In this section, the number of active drivers in the ride-sourcing system during a day is predicted. The proposed method first generates three clusters of drivers and then simulates their operation in a 24-h time frame based on the statistical analysis in Section 4.1. Each individual driver is represented as an agent. Simulated drivers start and end multiple shifts based on clustering analysis and the statistical estimation of their features. After a full-day simulation of all drivers, the number of active drivers can be readily estimated.

The schematic of the overall procedure is outlined in Figure 12. Firstly, the average number of drivers for each group in Figure 8 is generated and each driver is assigned to one of the three clusters. For

a unique driver, the number of shifts and the start time of the first shift can be sampled from the estimated distributions in Tables 2 and 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. 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  $T = \{1, \dots, 240\}$ . If drivers are working a shift during a time slot, they are considered to be active drivers within that time slot.

To evaluate the performance of the proposed method in comparison with empirical observations, the Mean Absolute Percentage Error (MAPE), Mean Absolute Error (MAE), and Root Mean Square Error (RMSE) are considered as performance metrics:

$$\begin{aligned}
 MAPE &= \frac{1}{T} \sum_{t=1}^T \frac{|Y_t - \hat{Y}_t|}{Y_t} \cdot 100 \\
 MAE &= \frac{1}{T} \sum_{t=1}^T |Y_t - \hat{Y}_t| \\
 RMSE &= \sqrt{\frac{1}{T} \sum_{t=1}^T (Y_t - \hat{Y}_t)^2}
 \end{aligned} \tag{5}$$

where  $Y_t$  and  $\hat{Y}_t$  are the observed and predicted number of active drivers at time slot  $t$ .

Four weekdays (Nov 22nd, 23rd, 24th, and 25th) and two weekends (Nov 26th and 27th) are selected as the testing data. Figure 13 compares the empirical results with the prediction results. The distributions of the number of active drivers within a day express different patterns between weekdays and weekends. There are three supply peaks (10:00 AM, 02:00 PM, and 06:00 PM) in weekdays and 3 peaks in weekends (12:00 PM, 02:30 PM, and 06:00 PM). The comparison also demonstrates that the proposed method can effectively predict the empirical data on different days. Table 6 summarizes the evaluation metrics for the six testing days. The prediction method based on clustering analysis achieves 8.9–10.3 MAPE (%), 232.5–422.2 MAE, and 284.1–526.8 RMSE for the four weekdays. The prediction errors for the two weekends are 16.0 MAPE (%), 479.6–615.1 MAE, and 649.5–715.8 RMSE. This indicates that the behavior patterns of the

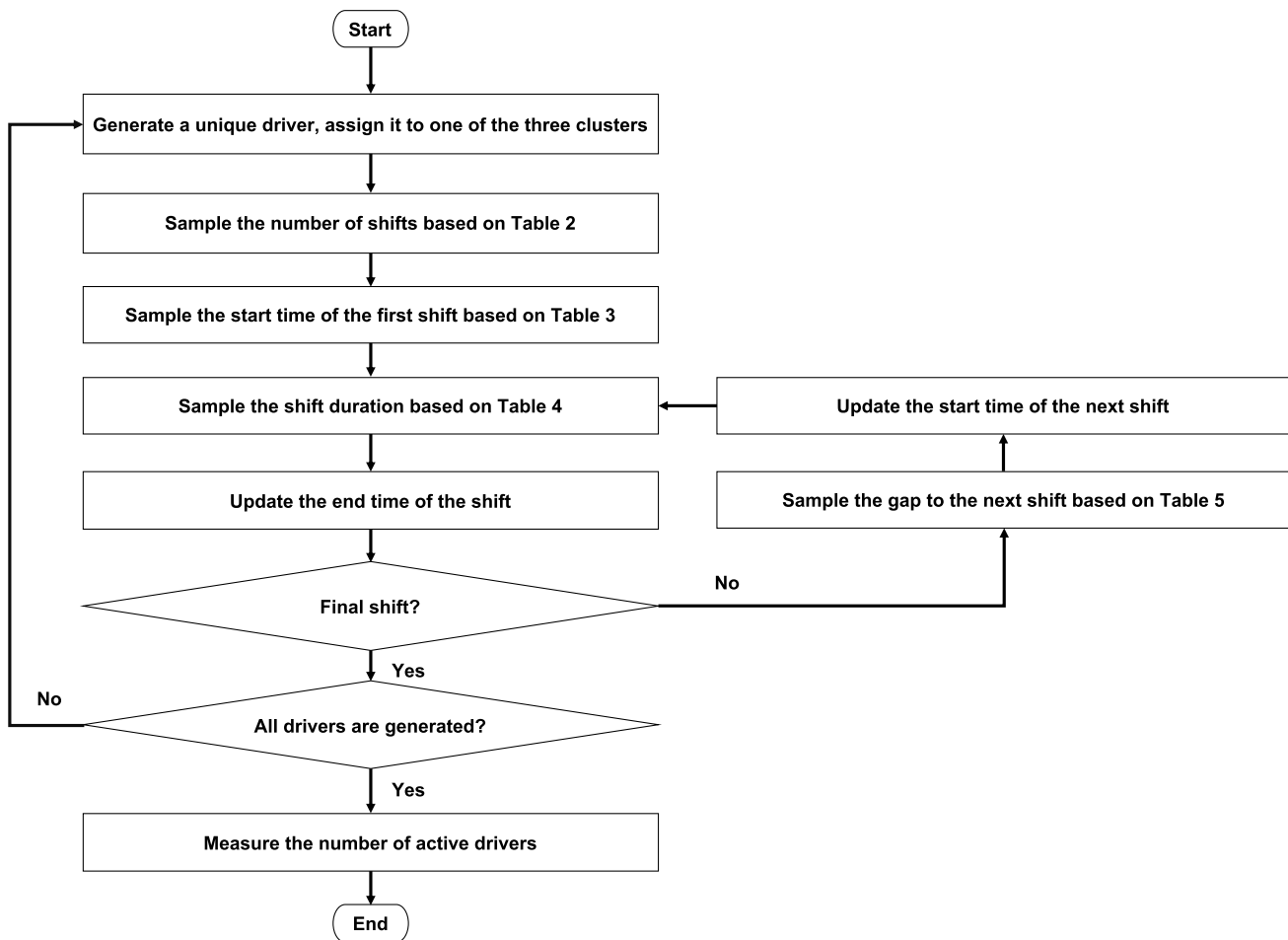


Figure 12. The overall procedure of the simulation-based prediction of the number of active drivers in the market.

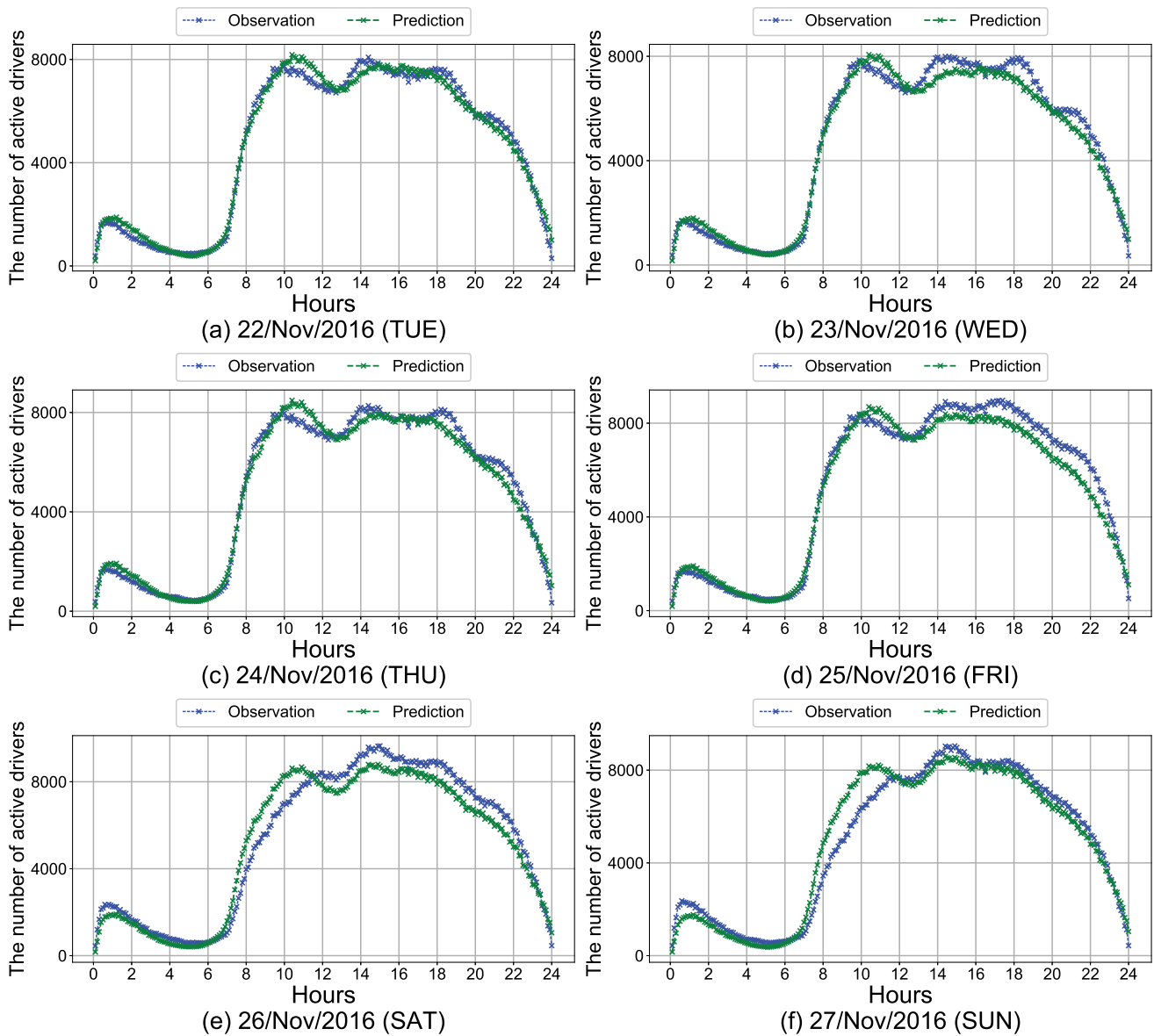
drivers can be well characterized by the proposed method. In addition, it can be observed that the proposed method performs better on weekdays than weekends. This is expected since the proposed method is trained on data from weekdays, which may bias the predictions to follow weekdays patterns. More specifically, it can also be observed that the MAPE of Cluster B is significantly higher than that of the other clusters on weekdays, which reflects the limited prediction ability of the proposed method at the cluster level.

### Discussions and limitations

The clustering and prediction analysis discussed in this paper provide key insights into the characteristics and market-behavioral patterns of ride-hailing drivers. Based on the results of clustering and prediction methods, this section discusses policy recommendations and limitations of this study. Section 5.1 offers driver incentive mechanisms to ride-sourcing operators for the avoidance of supply shortage or oversupply. Section 5.2 presents a discussion on how congestion may be alleviated and managed by imposing distance-based tax policy. Section 5.3 considers a potential solution on implementing and promoting ride-sharing services by leveraging the behavioral pattern of drivers. Finally, Section 5.4 discusses the limitations of the analysis.

### Driver incentivization

Designing appropriate wage and incentive schemes for drivers plays a crucial role in competitive ride-sourcing markets. To the best of our knowledge, there are multiple ride-sourcing platforms like Didi, Kuaidi, and Gaode in Chengdu city, and the majority of drivers may be multi-homing who simultaneously work for more than one platform and subsequently provide services on multiple ride-sourcing platforms. In our result, nearly 85% of drivers are part-time from the perspective of Didi, but if they are a contractor 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 (Leng et al. 2015; Henao and Marshall 2019; Fang, Huang, and Wierman 2020; Chen et al. 2020; Xu et al. 2020; Yu et al. 2021; Sun and Ertz 2021). 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.



**Figure 13.** A comparison of the predictions of time-varying numbers of active drivers in the market with observed values. Blue curves show that there are roughly 3 peaks (10:00 AM, 02:00 PM, and 06:00 PM) on weekdays and 3 peaks (12:00 PM, 02:30 PM and 06:00 PM) on weekends. Green curves demonstrate that the prediction results are reasonably accurate in reflecting the empirical number of active drivers on different days.

Figure 14 shows hourly numbers of active drivers of Clusters A, B, and C, hourly numbers of active drivers and orders, and hourly supply–demand (number of active drivers/number of new (serviced) orders) ratio on November 8 2016. Note that the set of active drivers includes occupied and idle drivers. Thus, the supply–demand ratio is overestimated. Considering the possible matching friction in ride-sourcing markets (Zha, Yin, and Yang 2016), a threshold of 5.0 is considered to delineate the supply shortage. Accordingly, two supply shortage periods during 04:00 AM – 08:00 AM and 10:00 PM – 12:00 AM can be observed. For the first supply shortage period (04:00 AM – 08:00 AM), discounting the platform commission rate or offering monetary bonus may incentivize more drivers in Clusters A and C to start joining the market earlier in the morning. For the second supply shortage period (10:00 PM – 12:00 AM), the platform can provide a completion-target monetary bonus (i.e. offering drivers a bonus once they complete a specific

number of consecutive orders), so drivers of all three clusters may prolong their working hours in the late evening. Another point is that to avoid a possible over supply and over competition during 10:00 AM – 06:00 PM, the platform can set higher commission rates during the day hours to encourage drivers of Clusters A and B to finish their shift and start their shift later.

The clustering analysis identifies full-time drivers in Cluster C who offer their services in the system for an extended time. Similarly, drivers in Cluster A are part-time drivers but with flexible work hours (usually with two shifts per day). They could be full-time multi-homing drivers among multiple platforms. In contrast, drivers in Cluster B only work in the evening, which suggests that they probably have another full-time job throughout the day. Those drivers are unlikely to adjust their working hours unless the wage and incentives exceed their full-time income. As shown in Figure 8, Cluster B includes 35–40% of all the drivers. This means that the wage and incentive programs may not be

**Table 6.** The evaluation metrics of the proposed prediction method compared to empirical observations. The units of MAE and RMSE are vehicles.

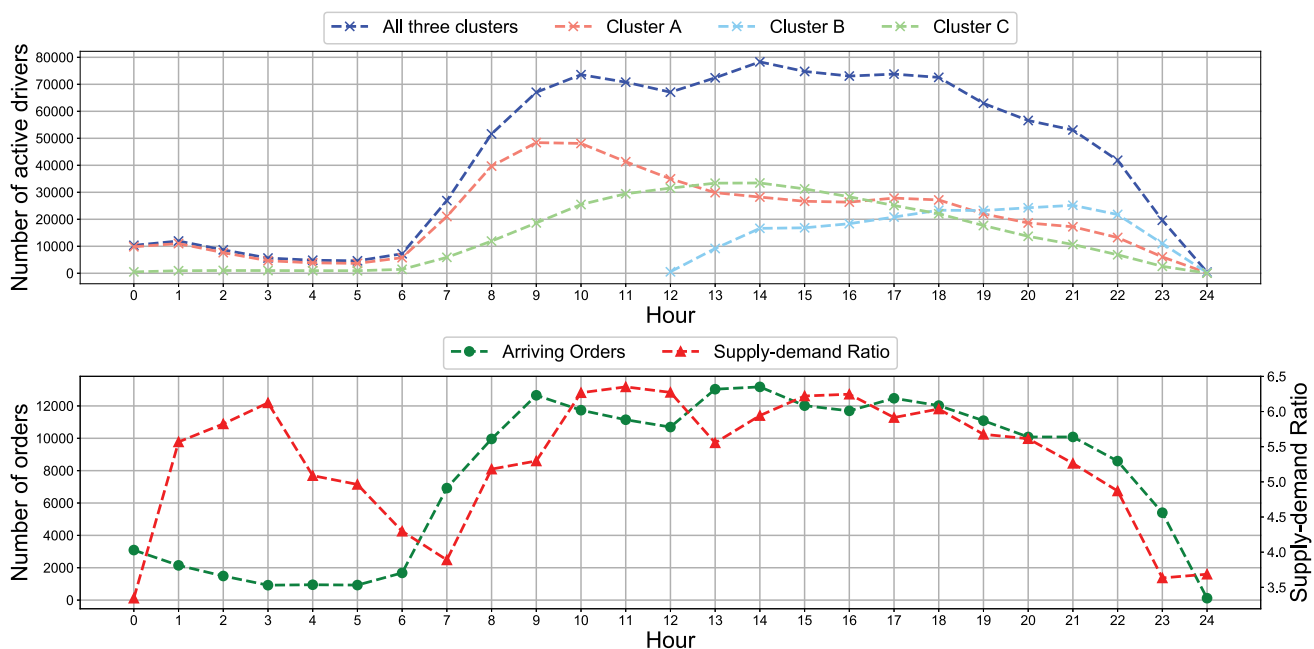
Date	Metrics	Overall	Cluster A	Cluster B	Cluster C
22/Nov/2016	MAPE (%)	9.3	15.3	69.5	11.8
	MAE (veh)	232.5	237.0	154.6	114.0
	RMSE (veh)	284.1	296.7	182.5	161.2
23/Nov/2016	MAPE (%)	9.2	12.3	35.9	14.8
	MAE (veh)	292.6	209.1	181.3	131.1
	RMSE (veh)	373.0	283.6	211.5	189.1
24/Nov/2016	MAPE (%)	8.9	13.5	59.7	14.3
	MAE (veh)	266.8	235.9	148.3	138.5
	RMSE (veh)	344.6	290.0	178.3	191.1
25/Nov/2016	MAPE (%)	10.3	13.1	42.3	18.4
	MAE (veh)	422.2	244.2	254.6	175.3
	RMSE (veh)	526.8	319.2	279.4	231.4
26/Nov/2016	MAPE (%)	16.0	16.8	18.3	25.4
	MAE (veh)	615.1	337.5	287.1	306.1
	RMSE (veh)	715.8	423.3	330.0	386.7
27/Nov/2016	MAPE (%)	16.0	18.3	17.6	25.3
	MAE (veh)	479.6	330.4	207.4	271.9
	RMSE (veh)	649.5	424.4	254.2	364.7

effective for slightly less than half of all the drivers. Although drivers in Cluster C are most likely to respond to the wage and incentive programs, they only make up 15% of all the drivers. Drivers in Cluster A are around 45–50% of all the drivers. Given proper wage and incentive programs, those drivers are able to extend their work hours and/or only serve for one platform. Therefore, the platform can mainly target drivers in Cluster A by motivating them to be full-time drivers or to prolong their working hours to mitigate the supply shortage.

## Congestion management

The previous section discussed how the e-hailing supply shortage and over supply can be tackled. However, a survey done by (Tirachini and Del Río 2019) in Santiago shows that ride-hailing may increase congestion due to its substitution of public transportation systems. This aligns with the findings from another study, which employed a Monte Carlo simulation model to predict the increase in traffic volume due to ride-hailing (Tirachini and Gomez-Lobo 2020). Therefore, regulators must intervene to balance the negative externalities of increased traffic, caused by the excessive empty trips of ride-hailing vehicles. (Note that these excessive empty trips benefit riders by reducing their waiting time.) Policies have been introduced in the past to control congestion caused by ride-hailing in other cities. This includes distance-based taxes in Sao Paulo (the distance-based tax was applied on the accumulated distance for the platform not for each driver, see Tirachini and Del Río 2019) and the integration of ride hailing with public transport (Young and Farber 2019).

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. As depicted in Figure 7, drivers of Cluster C occupy the highest average distance per shift while they have the lowest occupied duration percentage. 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



**Figure 14.** The hourly numbers of total and clustered active drivers are shown in the top figure. The hourly numbers of new (serviced) orders and the hourly supply–demand ratio are shown in the bottom figure (November 8 2016). Note that the total number of orders may be higher because unserviced canceled orders are not included in the dataset. In other words, the supply–demand ratio is an overestimation.

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.

The effectiveness of integrating e-hailing with public transport by subsidizing trips to a nearby public transport station (Young and Farber 2019) can be also evaluated by the Clusters' characteristics. Figure 6 shows that more than 80% of the shifts in the dataset occupied a distance of less than 40 km. If the total cost of a ride-hailing trip far exceeds the total cost of taking the public transport integrated with the ride-hailing (total cost includes monetary, waiting and in-vehicle travel time cost), the integration policy will significantly shorten the ride-hailing trip length distribution. Drivers of Clusters A and B who have time constraints in staying in the system may prefer driving shorter distance unlike drivers of Cluster C. Hence, the integration will be highly attractive for the part-time drivers of Clusters A and B since they can drive shorter distances and service more trips.

### Ridesharing

Ridesharing is a type of service that encourages passengers with similar trips and time schedules to share the same ride-hailing vehicle. This can contribute to tackling the supply shortage in systems and reducing congestion for society (Ferguson 1997; Chan and Shaheen 2012; Furuhashi et al. 2013). Both passengers and drivers can benefit from these services. Passengers with loose travel schedules usually receive a fare discount as compensation for their increase in travel time, while ride-hailing drivers can serve more passengers, increase their occupation rate, and earn more profit during their working shift. Aside from participants, ridesharing is also beneficial to society by mitigating traffic congestion and diminishing its environmental externalities (Agatz et al. 2012; Jin et al. 2018). A study has found that the implementation of Uber-sharing service in the United State has significantly reduced the congestion level (Li, Hong, and Zhang 2016) in 2016. Also, data from 'DiDi Hitch' (the ridesharing service launched by Didi Chuxing in China) confirmed the positive environmental effects of ridesharing. The data demonstrated that ride-sharing can improve vehicle utilization by 24%, and in 2016, DiDi users saved 1.44 million tons of carbon emissions by carpooling (Wang et al. 2019).

To attract more riders to opt-in for ridesharing services as opposed to solo ride-sourcing travel, a sufficient and spatially proportionate number of active drivers is required in the network to reduce waiting time of riders and limit detours (Chiabaut and Veve 2019; Veve and Chiabaut 2020). 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.

### Limitations of analysis

The dataset used has limitations. For example, the quality of the clustering result can be improved and a new cluster may be investigated if the information on late night drivers who start their shift before midnight and end after is available. In addition, individual drivers cannot be tracked across different days so their behavior cannot be specifically characterized. A unique driver can belong to Cluster A and Cluster C on different days if the driver works full-time in several different platforms. As driver IDs are re-anonymized, we cannot analyse/identify *individual-level* day-to-day decision-making patterns based on this dataset. A dataset with unique driver ID over

the days could shed more light on that. However, the collective behavior of drivers is analyzed day-to-day. The three clusters determined in this study have distinct characteristics despite the absence of the individual driver tracking as shown in Figure 7. In addition, the percentage of drivers in each cluster across different days depicted in Figure 8 is fairly constant. This shows that the results are meaningful.

Moreover, the information of the income of the drivers for each shift or any rush hour surcharge for Didi in Chengdu is also missing in the dataset. The income could have provided a better description of each cluster and determined whether drivers behave as neoclassical or income-targeting individuals (Xu et al. 2020). Similarly, the information on weather would also add a different angle to the analysis. Lastly, there is no data regarding passenger drop off and next passenger pick up. Hence, it is unknown whether the driver is actively searching for customers (driving or parking) or is taking a short break during the gaps. This information could provide a better definition of a shift for each individual driver and determine the efficiency of each shift.

### Summary

The paper has analyzed the behavior 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-behavioral patterns of ride-hailing drivers.

Based on the clustering analysis, the statistical estimation of four features for each cluster was investigated to build a prediction method for the number of active drivers within a day. The numerical results demonstrate that the proposed prediction method can accurately capture the within day fluctuation of supply on both weekdays and weekends compared with empirical observations. Based on the results of clustering and prediction methods, various policy, and operational recommendations including driver incentivization, congestion management, and ride-sharing are provided to tackle the mismatch of supply and demand (i.e. supply shortage and surplus), improve the operating efficiency, and address the competition.

### Notes

1. <https://outreach.didichuxing.com/research/opendata/en/>.

### Acknowledgments

This research was partially funded by the Australian Research Council (ARC) Discovery Early Career Researcher Award (DECRA) DE210100602.

### Disclosure statement

No potential conflict of interest was reported by the author(s).

### ORCID

Mohsen Ramezani  <http://orcid.org/0000-0001-6839-6839>



## References

- Agatz, N., A. Erera, M. Savelsbergh, and X. Wang. 2012. "Optimization for Dynamic ride-sharing: A Review." *European Journal of Operational Research* 223 (2): 295–303. doi:10.1016/j.ejor.2012.05.028.
- Aghaabbasi, M., Z. A. Shekari, M. Z. Shah, O. Olakunle, D. J. Armaghani, and M. Moeinaddini. 2020. "Predicting the Use Frequency of ride-sourcing by off-campus University Students through Random Forest and Bayesian Network Techniques." *Transportation Research Part A: Policy and Practice* 136 (Jun): 262–281.
- Alisoltani, N., L. Leclercq, and M. Zargayouna. 2021. "Can Dynamic ride-sharing Reduce Traffic Congestion?" *Transportation Research Part B: Methodological* 145: 212–246. doi:10.1016/j.trb.2021.01.004.
- Ashkrof, P., G. H. de Almeida Correia, O. Cats, and B. van Arem. 2020. "Understanding ride-sourcing Drivers' Behaviour and Preferences: Insights from Focus Groups Analysis." *Research in Transportation Business & Management* 37: 100516. doi:10.1016/j.rtbm.2020.100516.
- Beojone, C. V., and N. Geroliminis. 2021. "On the Inefficiency of ride-sourcing Services Towards Urban Congestion." *Transportation Research Part C: Emerging Technologies* 124: 102890. doi:10.1016/j.trc.2020.102890.
- Button, K. January 2020. "The 'Ubernomics' of Ridesourcing: The Myths and the Reality." *Transport Reviews* 40 (1): 76–94. doi:10.1080/01441647.2019.1687605.
- Chan, N. D., and S. A. Shaheen. 2012. "Ridesharing in North America: Past, Present, and Future." *Transport Reviews* 32 (1): 93–112. doi:10.1080/01441647.2011.621557.
- Chen, P. W., and Y. M. Nie. 2017. "Connecting e-hailing to Mass Transit Platform: Analysis of Relative Spatial Position." *Transportation Research Part C: Emerging Technologies* 77: 444–461. doi:10.1016/j.trc.2017.02.013.
- Chen, X. M., M. Zahir, and S. Zhang. 2017. "Understanding Ridesplitting Behavior of on-demand Ride Services: An Ensemble Learning Approach." *Transportation Research Part C: Emerging Technologies* 76: 51–70. doi:10.1016/j.trc.2016.12.018.
- Chen, X. M., H. Zheng, J. Ke, and H. Yang. 2020. "Dynamic Optimization Strategies for on-demand Ride Services Platform: Surge Pricing, Commission Rate, and Incentives." *Transportation Research Part B: Methodological* 138: 23–45. doi:10.1016/j.trb.2020.05.005.
- Chen, L., A. H. Valadkhani, and M. Ramezani. 2021. "Decentralised Cooperative Cruising of Autonomous ride-sourcing Fleets." *Transportation Research Part C: Emerging Technologies* 131: 103336. doi:10.1016/j.trc.2021.103336.
- Chiabaut, N., and C. Veve. 2019. "Identifying Twin Travelers Using Ridesourcing Trip Data." *Transport Findings* 7. <https://findingspress.org/article/9223-identifying-twin-travelers-using-ridesourcing-trip-data>
- Contreras, S. D., and A. Paz. 2018. "The Effects of ride-hailing Companies on the Taxicab Industry in Las Vegas, Nevada." *Transportation Research Part A: Policy and Practice* 115 (Sep): 63–70.
- Crawford, V. P., and J. Meng. 2011. "New York City Cab Drivers' Labor Supply Revisited: Reference-dependent Preferences with rational-expectations Targets for Hours and Income." *American Economic Review* 101 (5): 1912–1932. doi:10.1257/aer.101.5.1912.
- de Ruijter, A., O. Cats, R. Kucharski, and H. van Lint. 2022. "Evolution of Labour Supply in Ridesourcing." *Transportmetrica B: Transport Dynamics* 10: 1–28.
- Dong, Y., S. Wang, L. Li, and Z. Zhang. 2018. "An Empirical Study on Travel Patterns of Internet Based ride-sharing." *Transportation Research Part C: Emerging Technologies* 86: 1–22. doi:10.1016/j.trc.2017.10.022.
- Fang, Z., L. Huang, and A. Wierman. 2020. "Loyalty Programs in the Sharing Economy: Optimality and Competition." *Performance Evaluation* 143: 102105. doi:10.1016/j.peva.2020.102105.
- Farber, H. S. 2008. "Reference-dependent Preferences and Labor Supply: The Case of New York City Taxi Drivers." *American Economic Review* 98 (3): 1069–1082. doi:10.1257/aer.98.3.1069.
- Farber, H. S. 2015. "Why You Can't Find a Taxi in the Rain and Other Labor Supply Lessons from Cab Drivers." *The Quarterly Journal of Economics* 130 (4): 1975–2026. doi:10.1093/qje/qjv026.
- Ferguson, E. 1997. "The Rise and Fall of the American Carpool: 1970–1990." *Transportation* 24 (4): 349–376. doi:10.1023/A:1004928012320.
- Furuhata, M., M. Dessouky, F. Ordóñez, M.-E. Brunet, X. Wang, and S. Koenig. 2013. "Ridesharing: The state-of-the-art and Future Directions." *Transportation Research Part B: Methodological* 57: 28–46. doi:10.1016/j.trb.2013.08.012.
- Goodman, L. A. 1954. "Kolmogorov-smirnov Tests for Psychological Research." *Psychological Bulletin* 51 (2): 160. doi:10.1037/h0060275.
- Hamedmoghdam, H., M. Ramezani, and M. Saberi. May 2019. "Revealing Latent Characteristics of Mobility Networks with coarse-graining." *Scientific Reports* 9 (1): 7545. number: 1 Publisher: Nature Publishing Group. doi:10.1038/s41598-019-44005-9.
- Harding, S. 2016. "Taxi Apps, Regulation, and the Market for Taxi Journeys." *Transportation Research Part A: Policy and Practice* 88: 15–25.
- Hartigan, J. A., and M. A. Wong. 1979. "Ak-means Clustering Algorithm." *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 28 (1): 100–108.
- Henao, A., and W. E. Marshall. 2019. "An Analysis of the Individual Economics of ride-hailing Drivers." *Transportation Research Part A: Policy and Practice* 130: 440–451.
- Jin, S. T., H. Kong, R. Wu, and D. Z. Sui. 2018. "Ridesourcing, the Sharing Economy, and the Future of Cities." *Cities* 76: 96–104. doi:10.1016/j.cities.2018.01.012.
- Kaufman, L., and P. J. Rousseeuw. 2009. *Finding Groups in Data: An Introduction to Cluster Analysis*. Vol. 344. Hoboken, New Jersey, USA: John Wiley & Sons.
- Ke, J., H. Yang, H. Zheng, X. Chen, Y. Jia, P. Gong, and J. Ye. November 2019. "Hexagon-Based Convolutional Neural Network for Supply-Demand Forecasting of Ride-Sourcing Services." *IEEE Transactions on Intelligent Transportation Systems* 20 (11): 4160–4173. doi:10.1109/TITS.2018.2882861.
- Köszegi, B., and M. Rabin. 2006. "A Model of reference-dependent Preferences." *The Quarterly Journal of Economics* 121 (4): 1133–1165.
- Lavieri, P. S., and C. R. Bhat. 2019. "Modeling Individuals' Willingness to Share Trips with Strangers in an Autonomous Vehicle Future." *Transportation Research Part A: Policy and Practice* 124: 242–261.
- Leng, B., H. Du, J. Wang, L. Li, and Z. Xiong. 2015. "Analysis of Taxi Drivers' Behaviors within a Battle between Two Taxi Apps." *IEEE Transactions on Intelligent Transportation Systems* 17 (1): 296–300. doi:10.1109/TITS.2015.2461000.
- Li, Z., K. Hong, and Z. Zhang. 2016. "An Empirical Analysis of on-demand Ride Sharing and Traffic Congestion." In: *Proc. International Conference on Information Systems*.
- Lyft. March 2019. "Form S-1, at the United States Securities and Exchange Commission." <https://www.sec.gov/Archives/edgar/data/1759509/000119312519059849/d633517ds1.htm>
- Ma, Q., H. Yang, H. Zhang, K. Xie, and Z. Wang. October 2019. "Modeling and Analysis of Daily Driving Patterns of Taxis in Reshuffled Ride-Hailing Service Market." *Journal of Transportation Engineering, Part A: Systems* 145 (10): 04019045. doi:10.1061/JTEPBS.0000266.
- Nair, G. S., C. R. Bhat, I. Batur, R. M. Pendyala, and W. H. Lam. 2020. "A Model of Deadheading Trips and pick-up Locations for ride-hailing Service Vehicles." *Transportation Research Part A: Policy and Practice* 135 (May): 289–308.
- Nie, Y. M. 2017. "How Can the Taxi Industry Survive the Tide of Ridesourcing? Evidence from Shenzhen, China." *Transportation Research Part C: Emerging Technologies* 79: 242–256. doi:10.1016/j.trc.2017.03.017.
- Nocera, S., Pungillo, G., & Bruzzone, F. 2021. How to evaluate and plan the freight-passengers first-last mile. *Transport Policy* 113: 56–66.
- Nourinejad, M., and M. Ramezani. 2020. "Ride-Sourcing Modeling and Pricing in non-equilibrium two-sided Markets." *Transportation Research Part B: Methodological* 132 (Feb): 340–357. doi:10.1016/j.trb.2019.05.019.
- Riascos, A. P., and J. L. Mateos. December 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): 4022. doi:10.1038/s41598-020-60875-w.
- Rousseeuw, P. J. 1987. "Silhouettes: A Graphical Aid to the Interpretation and Validation of Cluster Analysis." *Journal of Computational and Applied Mathematics* 20: 53–65. doi:10.1016/0377-0427(87)90125-7.
- Shaaban, K., and I. Kim. 2016. "Assessment of the Taxi Service in Doha." *Transportation Research Part A: Policy and Practice* 88 (Jun): 223–235.
- Shaheen, S. A., A. P. Cohen, I. H. Zohdy, B. Kock. 2016. "Smartphone Applications to Influence Travel Choices: Practices and Policies." Institute of Transportation Studies, Research Reports, Working Papers, Proceedings.

- Shen, Y., H. Zhang, and J. Zhao. 2018. "Integrating Shared Autonomous Vehicle in Public Transportation System: A supply-side Simulation of the first-mile Service in Singapore." *Transportation Research Part A: Policy and Practice* 113 (Jul): 125–136.
- Su, D. N., D. Q. Nguyen-Phuoc, and L. W. Johnson. February 2021. "Effects of Perceived Safety, Involvement and Perceived Service Quality on Loyalty Intention among ride-sourcing Passengers." *Transportation* 48 (1): 369–393. doi:10.1007/s11116-019-10058-y.
- Sun, S., and M. Ertz. 2021. "Dynamic Evolution of ride-hailing Platforms from a Systemic Perspective: Forecasting Financial Sustainability." *Transportation Research Part C: Emerging Technologies* 125: 103003. doi:10.1016/j.trc.2021.103003.
- Tirachini, A., and M. Del Río. 2019. "Ride-hailing in Santiago de Chile: Users' Characterisation and Effects on Travel Behaviour." *Transport Policy* 82 (Oct): 46–57. doi:10.1016/j.tranpol.2019.07.008.
- Tirachini, A., and A. Gomez-Lobo. 2020. "Does ride-hailing Increase or Decrease Vehicle Kilometers Traveled (Vkt)? A Simulation Approach for Santiago de Chile." *International Journal of Sustainable Transportation* 14 (3): 187–204. doi:10.1080/15568318.2018.1539146.
- Uber. March 2019. "Form S-1. at the United States Securities and Exchange Commission." <https://www.sec.gov/Archives/edgar/data/1543151/000119312519103850/d647752ds1.htm>
- Veve, C., and N. Chiabaut. 2020. "Estimation of the Shared Mobility Demand Based on the Daily Regularity of the Urban Mobility and the Similarity of Individual Trips." *PLoS one* 15 (9): e0238143. doi:10.1371/journal.pone.0238143.
- Vij, A., 2020. "Consumer Preferences for on-demand Transport in Australia, 17."
- Vogel, C. R. 2002. *Computational Methods for Inverse Problems*. Philadelphia, PA, USA: SIAM.
- Wang, H., and H. Yang. 2019. "Ridesourcing Systems: A Framework and Review." *Transportation Research Part B: Methodological* 129: 122–155. doi:10.1016/j.trb.2019.07.009.
- Wang, Y., J. Gu, S. Wang, and J. Wang. 2019. "Understanding Consumers' Willingness to Use ride-sharing Services: The Roles of Perceived Value and Perceived Risk." *Transportation Research Part C: Emerging Technologies* 105: 504–519. doi:10.1016/j.trc.2019.05.044.
- Xu, Z., Z. Li, Q. Guan, D. Zhang, Q. Li, J. Nan, C. Liu, W. Bian, and J. Ye. 2018. "Large-scale Order Dispatch in on-demand ride-hailing Platforms: A Learning and Planning Approach." In: *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, London. pp. 905–913.
- Xu, Z., D. AMC Vignon, Y. Yin, and J. Ye. 2020. "An Empirical Study of the Labor Supply of ride-sourcing Drivers." *Transportation Letters* 1–4. doi:10.1080/19427867.2020.1788761.
- Young, M., and S. Farber. 2019. "The Who, Why, and When of Uber and Other ride-hailing Trips: An Examination of a Large Sample Household Travel Survey." *Transportation Research Part A: Policy and Practice* 119 (Jan): 383–392.
- Yu, J., D. Mo, N. Xie, S. Hu, and X. M. Chen. 2021. "Exploring multi-homing Behavior of ride-sourcing Drivers via real-world Multiple Platforms Data." *Transportation Research Part F: Traffic Psychology and Behaviour* 80: 61–78. doi:10.1016/j.trf.2021.03.017.
- Zha, L., Y. Yin, and H. Yang. 2016. "Economic Analysis of ride-sourcing Markets." *Transportation Research Part C: Emerging Technologies* 71: 249–266. doi:10.1016/j.trc.2016.07.010.
- Zha, L., Y. Yin, and Y. Du. 2018. "Surge Pricing and Labor Supply in the ride-sourcing Market." *Transportation Research Part B: Methodological* 117: 708–722. doi:10.1016/j.trb.2017.09.010.
- Zhao, P., X. Liu, M.-P. Kwan, and W. Shi. December 2020. "Unveiling Cabdrivers' Dining Behavior Patterns for Site Selection of 'Taxi Canteen' Using Taxi Trajectory Data." *Transportmetrica A: Transport Science* 16 (1): 137–160. doi:10.1080/23249935.2018.1505972.

## Appendix

The data structure of trajectories and orders are provided in the following Tables.

**Table A1.** Trajectory data structure.

Field	Type	Sample	Comment
Driver ID	String	glox.jrrlltBMvCh8nxqktdr2dtopmlH	Anonymized
Order ID	String	jkkt8kxniovlFuns9qrrlvst@iqnpkwz	Anonymized
Time Stamp	String	1501584540	Unix timestamp
Longitude	String	104.0439	GJ-02 Coordinate System
Latitude	String	30.6673	GJ-02 Coordinate System

**Table A2.** Order data structure.

Field	Type	Sample	Comment
Order ID	String	jkkt8kxniovlFuns9qrrlvst@iqnpkwz	Anonymized
Ride Start Time	String	1501584540	Unix timestamp
Ride Stop Time	String	1501584540	Unix timestamp
Pick-up Longitude	String	104.0439	GJ-02 Coordinate System
Pick-up Latitude	String	30.6673	GJ-02 Coordinate System
Drop-off Longitude	String	104.0439	GJ-02 Coordinate System
Drop-off Latitude	String	30.6673	GJ-02 Coordinate System