A Cluster Analysis of Uber Request Data via the Transit App in New York City

Jing Guo
Department of Civil and Environmental Engineering
University of Tennessee, Knoxville, Knoxville, TN
Email: jguo31@vols.utk.edu

Antora Mohsena Haque
Department of Civil and Environmental Engineering
University of Tennessee, Knoxville, Knoxville, TN
Email: ahaque1@vols.utk.edu

Cassidy Crossland
Department of Civil and Environmental Engineering
University of Tennessee, Knoxville, Knoxville, TN
Email: ccrossl3@vols.utk.edu

Candace Brakewood, PhD
Department of Civil and Environmental Engineering
University of Tennessee, Knoxville, Knoxville, TN
Email: cbrakewo@utk.edu

Word count: 5599 + 3 tables (250 * 3 = 750) = 6,349

Submission Date: August 1, 2020

ABSTRACT

As ridehailing services like Uber become increasingly common in urban transportation systems, it is necessary to understand their usage patterns. Since private ridehailing companies do not publicly disclose their ridership data, usage patterns can be analyzed using other data sources, such as the Transit app's Uber request data. The objectives of this research are three-fold: (1) explore the temporal characteristics of Uber requests through data visualization, (2) identify groups of users through cluster analysis, and (3) compare Transit app Uber request data with overall Uber usage data and transit data. The exploratory analysis results suggest that requests occurred most frequently during AM and PM peak periods. K-means clustering identified eight groups of Uber users: long duration and frequent users, off-peak users, PM peak users, AM peak users, party goers, long duration and infrequent users, holiday users, and weekend users. The main trip purposes determined by the clustering analysis were going to social events and to and from the workplace or home. Comparing the Transit app data to the overall ridehailing usage data and transit data suggest that the time distribution pattern of Transit app Uber requests is a combination of transit and Uber usage while the time usage features are more similar to those of Uber users. These results will help transportation departments to better coordinate ridehailing services and public transportation to meet users' travel needs.

Keywords: Uber request, Transit app, K-means, Cluster, Data Mining

INTRODUCTION

Ridehailing is a pre-arranged and on-demand transportation service where drivers and passengers are connected through a digital application (*1*). There has been significant growth in the adoption of ridehailing services since the introduction of Uber in 2009 (*2*). Ridehailing, which is also known as ridesharing, ridesourcing, or Transportation Network Companies (TNCs), has become an increasingly popular travel mode for passengers in recent years. Many researchers have become interested in different aspects of the development of ridehailing, such as the frequency and use of ridehailing, the ridehailing trip purpose, and the relationship between ridehailing and other travel modes.

Analyzing ridehailing services usage not only helps to infer and understand the travel patterns of travelers, but also helps transportation departments and researchers coordinate the transportation network system. However, limitations on accessing trip data from ridehailing companies, such as Uber and Lyft, create difficulties for planners and researchers to explore and analyze ridehailing usage. The dataset used in this study, which will be called "Uber requests" through the remainder of this paper, was obtained from the Transit app. Transit app is a free digital application that collects and maps data on public transit and shared mobility services, such as ridehailing and bikesharing. The Transit app allows users to request Uber by clicking on an icon within the application. During the study period, the Uber request would then be handed off to Uber through its smartphone application for fulfillment (*3*). Therefore, the dataset used in this study is likely an indicator of demand for Uber but does not reflect actual trips. It should be noted that this feature has since been updated and now allows users to book an Uber directly in the app. The Uber requests data was obtained from the Transit app in New York City between October 2016 and October 2017 and analyzed using a data-driven approach and data mining technique. A K-means unsupervised machine learning algorithm was used to understand the temporal usage pattern of Uber requests from the Transit app users by identifying different clusters.

The paper proceeds in the following way. First, the literature review section describes the prior studies related to temporal characteristics of ridehailing usage, trip purpose of ridehailing users, the relationship between public transportation and ridehailing, and Uber requests from the Transit app. Then, objectives of this paper are set forth. The methodology is described next. The next section discusses the results of the study, followed by the comparison with Uber usage and transit usage. The conclusions and future research are described in the last section.

LITERATURE REVIEW

With the development of smartphones, online-based ridehailing developed rapidly and sparked many questions relating to this new travel mode. Relevant studies analyzing the characteristics of ridehailing focused on three major areas: the time of day and day of week distribution of ridehailing usage, the trip purpose of ridehailing, and the relationship between ridehailing and public transportation. These topics make up the first three sections of the literature review. In the final section of the literature review, prior studies using similar Uber request datasets from the Transit app are described.

Time of day and day of week distribution of ridehailing usage

At least seven prior studies have analyzed the temporal characteristics of ridehailing usage, as listed in **Table 1**. Exploration of hourly and daily usage is necessary to understand the temporal distribution of ridehailing usage. Researchers found ridehailing trip data, including trip time information, to be the most useful type of data for temporal data analysis. a small number of prior studies were able to use ridehailing trip data, such as Uber trips data (*4*), New York City Taxi and Limousine Commission (TLC) trips data (*5*), Uber and Lyft APIs via Northeastern University (*6*) and trip data obtained from the TNC (*7*). Several studies used surveys to obtain information related to subjective questions like a person's income, education level, travel mode preferences, and travel purpose (*8; 9*). These surveys can be used to supplement the trip data and to analyze the temporal characteristics of ridehailing usage.

Key findings from the seven studies shown in Table 1 are described here, first for hour of the day and then for day of the week. Ridehailing trip records from New York City showed that ridehailing demand was concentrated during the morning and evening peak periods and late evenings (*5*). An in-vehicle intercept survey conducted in Massachusetts suggested that 40% of weekday ridehailing rides occurred during the four-hour morning and evening commute periods (*10*). An additional study found that the evening peak was higher and longer than the morning peak on weekdays (*6*). Another research study noted that the highest ridehailing usage volume hour fell on Saturday nights at 9 or 10 p.m. (*7*). Many other studies indicated that ridehailing trips occurred most often later in the evening. An online survey in California suggested that the majority of ridehailing trips were taken between 10 p.m. and 4 a.m., and an in-vehicle intercept survey in Massachusetts suggested that 42% of weekend ridehailing rides happened between 7 p.m. and midnight (*10; 11*). After combining research results with the opening hours of local bars, researchers found that there was a large spike in ridehailing demands during bar closing time in Pittsburgh (*4*).

The temporal characteristics of ridehailing services were also seen through the differences in demand depending on the day of the week. According to prior studies, most ridehailing trips occurred on Fridays and Saturdays (*6; 12*) while the fewest ridehailing requests occurred on Sundays (*6*).

Trip purpose of ridehailing

There were seven studies that analyzed the trip purpose of ridehailing, and those are organized chronologically in **Table 2**. Five of the seven studies found that going out or social events were the most common trip purpose for using ridehailing (*4; 12-15*). Three studies suggested that home was usually the origin or destination among ridehailing trips (*10; 14; 16*). Another three studies found that some ridehailing trips were taken for work or commuting purposes (*12-14*). A final purpose of using ridehailing was to get to or from the airport (*14*).

Relationship between ridehailing and public transportation

The complementary or substitutionary relationship between ridehailing and public transit has been a controversial topic since the emergence of ridehailing services. Some researchers found that ridehailing and public transit have a complementary relationship, especially for those transit systems with low ridership (*17*). However, other researchers found that ridehailing services compete with public transportation based on speed of travel, convenience and comfort (*5*). A final study using both ridehailing trip data and survey data concluded that there is no relationship between the level of peak-hour ridehailing use and longer-term changes in the study regions' public transit usage (*7*).

Prior studies of Uber requests from the Transit app

Some prior studies specifically focused on the Uber request feature within the Transit app. Davidson et al. used the Uber request data from the Transit app to visualize the spatial distribution in New York City and concluded that Transit app Uber requests were located closer to subway stations than those trips requested directly from the Uber application (*18*). In another study, the number of Uber requests from the Transit app during an extreme weather event were also analyzed; the results suggested that ridehailing services may have played a role in recovery after the transportation system shutdown due to extreme weather events (*19*).

However, these studies did not focus on the temporal distribution and trip purpose of Uber requests via the Transit app. The Uber requests data generated by the Transit app are not the same as the data generated from ridehailing applications or the data gathered by surveys due to the users' transit-planning tendency. This paper aims to fill this gap from the studies described above related to analysis of ridehailing usage. It is also a good opportunity to assess the relationship between public transit and ridehailing usage since it is assumed that those who send Uber requests from the Transit app were initially considering using public transportation.

TABLE 1 Studies Related to the Time of Day and Day of Week Exploration of Ridehailing Usage

Studies	Location	Data Sources	Methodology	Related Findings
Uber and MA DD; 2015 (*4*)	Seattle; Miami; Pittsburgh; Chicago; Austin; San Francisco; California	• Uber Trip Data • Online Survey Conducted by Benenson Group (N = 807)	• Graph Descriptions • Summary Statistics	• Large spike in Uber requests during bar closing time in Pittsburgh
Rayle et al.; 2016 (*12*)	San Francisco	• Intercept Survey (May and June 2014, N = 380) • San Francisco Municipal Transportation Authority Taxi User Survey • GPS Trip Logs for a taxi company • 2013 American Community Survey	• Descriptive Statistics	• 48% of ridehailing trips were taken on Friday or Saturday
Schaller; 2017 (*5*)	New York City	• New York City Taxi and Limousine Commission Trip Data (June 2013, 2015, and 2016, N = 43,014,512)	• Descriptive Statistic	• TNC trip growth in Manhattan has been concentrated during the morning and evening peak periods, late evenings, and weekends
Circella et al; 2018 (*11*)	California	• Online Survey (Fall 2015, N = 2,400)	• Descriptive Statistics	• Majority of ridehailing trips were taken between 10p.m. and 4 a.m.
Cooper et al; 2018 (*6*)	San Francisco International Airport	• Uber and Lyft APIs via Northeastern University (November 12 to December 21, 2016)	• Iterative Proportional Fitting Algorithm	• TNC trips increase throughout the week (130,000 on Monday to 220,000 on Friday and Saturday) • Number of TNC trips are lowest on Sunday • Evening peak is higher and longer than the morning peak on weekdays • TNCs have evening peaks that extend further into the nights with a second peak around 11 p.m. on Thursdays and Fridays
Feigon and Murphy; 2018 (*7*)	Chicago; Los Angeles; Nashville; San Francisco; Seattle; Washington D.C.	• TNC data source for 5 cities (May 2016) • San Francisco County Transportation Authority modeled information • Shared Mobility Survey in 8 cities (2015, N = 10,000 transit and shared mobility users) • Four Agency Survey in 4 cities (2015, N = 357 transit riders administered by public transit agencies)	• Descriptive Statistics	• Highest TNC usage volume hour falls on Saturday night at 9 or 10 p.m. • Lowest TNC usage volume hours uniformly fall on early weekday mornings
Gehrke; 2019 (*10*)	Massachusetts	• In-vehicle Intercept Survey (October-November 2017, N = 944)	• Descriptive Statistics	• 42% of weekend ridehailing trips occurred between 7p.m. and midnight • 40% of weekday ridehailing trips occurred during the 4-hour morning and evening commutes

TABLE 2 Studies Related to the Trip Purpose of Ridehailing

Studies	Location	Data Sources	Methodology	Related Findings
Uber and MA DD; 2015 (*4*)	Seattle; Miami; Pittsburgh; Chicago; Austin; San Francisco; California	• Uber Trip Data • Online Survey Conducted by Benenson Group (Date Unknown, N = 807)	• Graphs Description • Summary Statistics	• Going Out/Social: In Chicago, late night origins were within 50 meters of an establishment that serves alcohol
Rayle et al.; 2016 (*12*)	San Francisco	• Intercept Survey (May and June 2014) • San Francisco Municipal Transportation Authority Taxi User Survey • GPS Trip Logs for a taxi company • 2013 American Community Survey	• Descriptive Statistics	• Going Out/Social: 67% of ridehailing trips are for social/leisure • Work/Commuting: 16% of ridehailing trips are for work
Feigon and Murphy; 2016 (*13*)	Austin; Boston; Chicago; Los Angeles and San Francisco; Seattle; Washington, DC	• User survey (September and October 2015, N=4,551)	• Descriptive Statistics	• Going Out/Social: 54% of trips were taken for recreational or social events • Work/Commuting: 21% of trips were taken for commuting purposes
Henao; 2017 (*14*)	Denver	• Survey administered during trip (Fall 2016, N = 416 ridehailing trips and 311 passenger surveys) • Driver Trip Data	• Descriptive Statistics • Travel Demand Model	• Going Out/Social: 16% of trips originated at a social outing while 18% of trips in this study had a destination at a social outing • Work/Commuting: 13% of trips originated at work while 17% of trips in this study had a destination at work • To/From Airport 12% of trips had a destination at an airport • To/From Home 41% of trips originated at home; 29% of trips in this study had a destination at home
Mahmoudifard et al.; 2017 (*15*)	Chicago	• Online Survey recruited by Uber drivers	• Descriptive Statistics	• Going Out/Social: 53.84% of Uber trips were for a social/leisure activity
Gehrke; 2019 (*10*)	Massachusetts	• In-vehicle Intercept Survey (October-November 2017, N = 944)	• Descriptive Statistics	• To/From Home: 58% of ridehailing trips that started somewhere other than the rider's home went home
Erhardt; 2019 (*16*)	San Francisco	• San Francisco's Travel Demand Model • 2 TNCs Application Programming Interface • INRIX	• Fixed-Effect Panel Regression Model	• To/From Home: Uber is more likely to be chosen for the return home from an activity rather than going to the activity

OBJECTIVES

The objective of this paper is to understand patterns of Uber request data via the Transit app data. Three specific objectives are set forth to achieve this goal, which are:

1. Explore the temporal Uber requests pattern based on the time of day and day of week;
2. Infer trip purpose and identify distinct groups of users by K-means unsupervised machine learning algorithm; and
3. Compare temporal characteristics of Uber requests from the Transit app with Uber pickup data released publicly through the TLC and General Transit Feed Specification (GTFS) service information for the subway system in New York City.

METHODOLOGY

The methodology of this study was broken down into five steps as shown in **Figure 1**. The steps are described below.

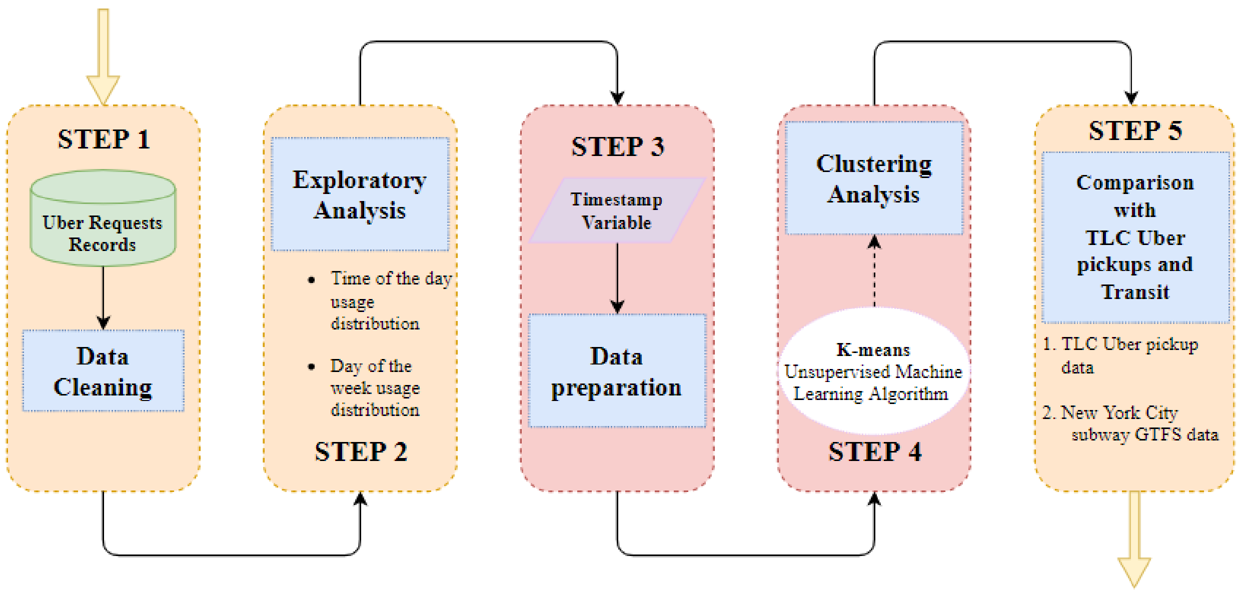

Figure 1 Flowchart showing the methodology of this study

Step 1: Data Cleaning

The Transit app dataset used in this paper contained a sample of over 200,000 Uber requests in New York City for over 100,000 users (by unique device ID). First, all the Uber requests made by each unique device ID were sorted. Some Uber requests occurred within a very short time for some of the devices. A ten-minute threshold period was used to check the interval between two consecutive Uber requests by an individual device ID. If the interval between two consecutive requests was no more than ten minutes, the former request was discarded since it was assumed that the user did not complete an Uber trip. This procedure was followed for all requests made for each unique device ID. There were more than 100,000 Uber requests left after this data cleaning process.

Step 2: Exploratory Analysis

Exploratory analysis of the Uber requests was performed next. Trend line graphs were created to explore the characteristics by time of day and day of week to identify temporal usage patterns of Uber requests via the Transit app.

Step 3: Data Preparation

Next, those users with only one Uber request were removed from the dataset since it is hard to recognize patterns with only one request. Based on prior studies that focused on the temporal usage pattern of ridehailing services and the relationship between ridehailing and public transportation, several variables were designed to explain the temporal usage patterns of Uber requests. The variables used in the clustering analysis were created from the timestamp variable. A brief description of each variable used in the K-means clustering analysis is described below:

(1) Count of Days

The count of days an individual user (unique device ID) requested an Uber in the Transit app during the one-year period was calculated. This variable showed the request frequency.

(2) Duration

The duration was defined as the time interval between the first and last day an Uber was requested in the Transit app by an individual user during the one-year study period.

(3) AM Peak Usage Rate

For an individual user, the AM peak usage rate was calculated as follows: the count of the total number of Uber requests made between 6:00 a.m. and 8:59 a.m. on weekdays divided by the total number of Uber requests made in the one-year period.

(4) PM Peak Usage Rate

For an individual user, the PM peak usage rate was calculated as follows: the count of the total number of Uber requests made between 4:00 p.m. and 6:59 p.m. on weekdays divided by the total number of Uber requests made in the one-year period.

(5) Party-time Usage Rate

For an individual user, the party-time usage rate was calculated as follows: the count of the total number of Uber requests made between 10:00 p.m. on Friday and 2:59 a.m. on Saturday and between 10:00 p.m. on Saturday and 2:59 a.m. on Sunday divided by the total number of Uber requests made in the one-year period.

(6) Weekend Usage Rate

For an individual user, the weekend usage rate was calculated as follows: the count of the total number of Uber requests made during the weekends excluding party-time (Saturday 3:00 a.m. to 9:59 p.m. and Sunday 3:00 a.m. to 11:59 p.m.) divided by the total number of Uber requests made in the one-year period.

(7) Holiday Usage Rate

For an individual user, the holiday usage rate was calculated as follows: the count of the total number of Uber requests made during holidays in this period divided by the total number of Uber requests made in study time period. The holidays designated for this variable included: Halloween, Thanksgiving, Black Friday, Christmas Eve, Christmas Day, New Year’s Eve, New Year’s Day, Superbowl Sunday, Valentine's Day, Saint Patrick’s Day, and Independence Day.

Step 4: Clustering Analysis

K-means is a well-known unsupervised classification method that helps to identify groups based on their characteristics. Hence, the K-means clustering method was selected to identify different user groups based on the time variables of the Uber requests via the Transit app.

The K-means algorithm partitions the dataset into a predefined number (*k* groups) of clusters by minimizing the distance of observations within clusters and maximizing the distance of observations in different clusters. These two criteria are known as within cluster sum of squares (WCSS) and between cluster sum of squares (BCSS). Since the total variance is constant, minimizing WCSS is equivalent to maximizing the BCSS (*20*). The mathematical formula is shown in **Equation 1**:

$$WCSS\ (k) = \sum_{j=1}^{k} \sum_{i=1}^{n} \|x_i - u_j\|^2 \quad (1)$$

Where:
$\|x - u_i\|$ = the Euclidian distance between the observation x_i *and* centroid u_j
u_j = the sample mean in cluster j
k = number of clusters
n = number of cases
The K-means clustering algorithm typically follows four steps (*21*):

1. Initialization: set an initial cluster by picking a random value as k.
2. Assignment: partition all the observations to the nearest centroid.
3. Update: recalculate centroids for observations assigned to each cluster.
4. Repeat the assignment step and update step until the assignments do not change.

Step 5: Comparison with TLC Uber Pickups and Transit

To further understand the temporal difference between ridehailing usage and transit, a comparison of Uber request data via the Transit app with Uber pickups data and subway GTFS data in New York City was conducted.

The Uber pickup data from January 1 to June 30, 2015 was obtained from the TLC. As one of the for-hire-vehicles in New York City, Uber has been required to report pickups data to the TLC from the dispatch bases under New York law since 2014. On July 20, 2015, FiveThirtyEight obtained Uber pickup data for two time periods (April to September 2014 and January to June 2015) by submitting a Freedom of Information Law (FOIL) request and made the data public with FOIL documentation on a GitHub repository (*22*). In this study, the Uber pickup data from January to June 2015 was selected to compare with the Uber requests data via the Transit app.

The New York City subway GTFS schedule data from November 2016 to December 2017 was used to build estimates of transit demand across the day and week. The New York City subway GTFS data can be easily accessed and downloaded from *OpenMobilityData (23)*, a free service collecting open data about mobility from around the world. GTFS represents transit service availability, which is assumed to reflect transit demand patterns.

RESULTS AND DISCUSSION

Exploratory Analysis

Figure 2 presents the temporal distribution from three perspectives from top to bottom: daily Uber requests over the entire study period, the average number of Uber requests for each day of the week, and the average number of Uber requests for the hour of the day. The top graph indicates that there were more Uber requests during the spring and summer months than in the winter months. From the distribution of the average number of Uber requests per hour in a week, which is shown as the middle graph, there was a spike late at night, from 9 p.m. to 1 a.m., on Friday and Saturday when people typically go out for social and leisure events, such as going to bars. As shown in the bottom graph, the average number of Uber requests per hour in a day by weekday and weekend shows clear a.m. and p.m. peaking patterns on weekdays, when commuters travel to or from the workplace. In addition, this graph also suggests that the number of late night Uber requests was always high, no matter the day of the week.

Clustering Analysis

This section describes the results of K-means clustering algorithm on the Uber request data. There were 28,636 users (unique device IDs) left for clustering analysis after the data preparation step.

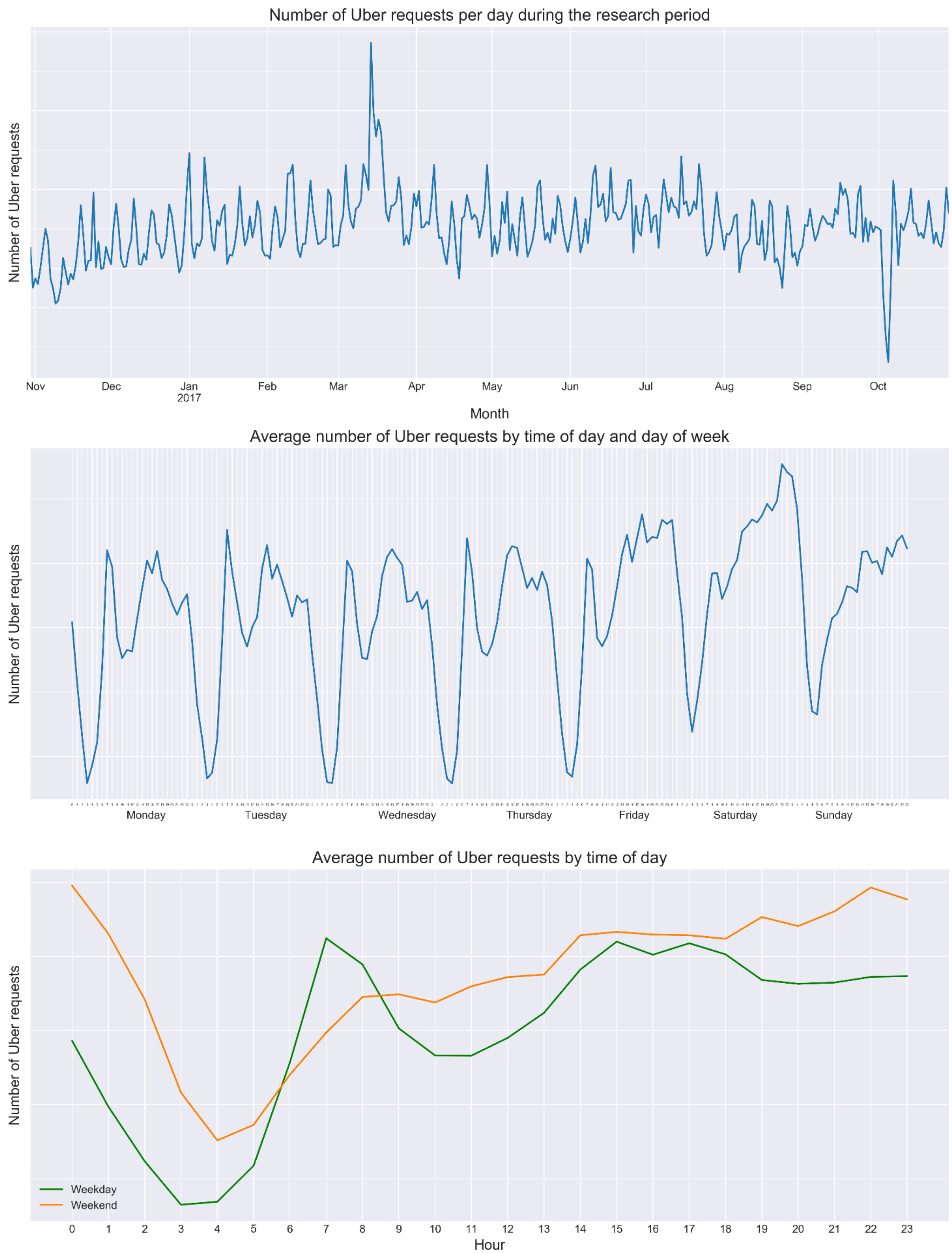

Figure 2 Temporal Distribution of Uber Requests

Selection of K with BCSS and WCSS

The value of the evaluation indices, BCSS and WCSS, of the K-means algorithm with k values ranging from 1 to 15 were plotted and are shown below in **Figure 3**. There is an elbow when the number of clusters (x-axis) is equal to eight. Therefore, eight was chosen as the number of clusters for the K-means clustering and interpretation of this study.

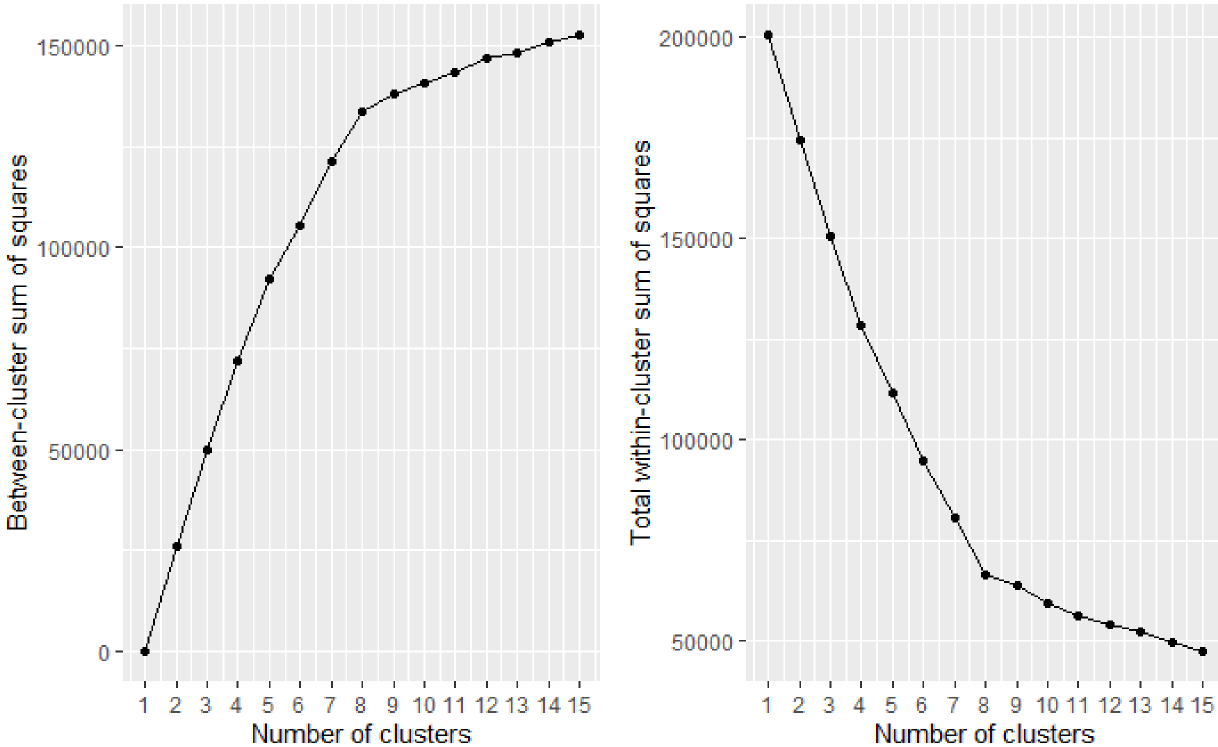

Figure 3 Cluster Evaluation Indices

Clusters Interpretation

K-means clustering identified eight types of Uber requests user groups based on seven variables, which are the count of days, duration, AM peak usage rate, PM peak usage rate, party time usage rate, weekend usage rate, and holiday usage rate. The values of these variables were visualized in two types of charts, a bubble chart and a radar chart, for each cluster. **Figure 4** represents the count of days and duration for each cluster where the size of the bubble represents the number of users in the cluster. The radar chart shown in **Figure 5** describes how the clusters are related to the different time period usage rate variables. Based on the visualizations of the clusters shown in **Figure 4** and **Figure 5**, the clusters were labeled as long duration and frequent users, off-peak users, PM peak users, AM peak users, party goers, long duration and infrequent users, holidays users, and weekend users. **Table 3** summarizes the mathematical results for each cluster. It shows that the highest number of users are in the weekend users' group (21.3%) followed by the off-peak users' group (21.2%).

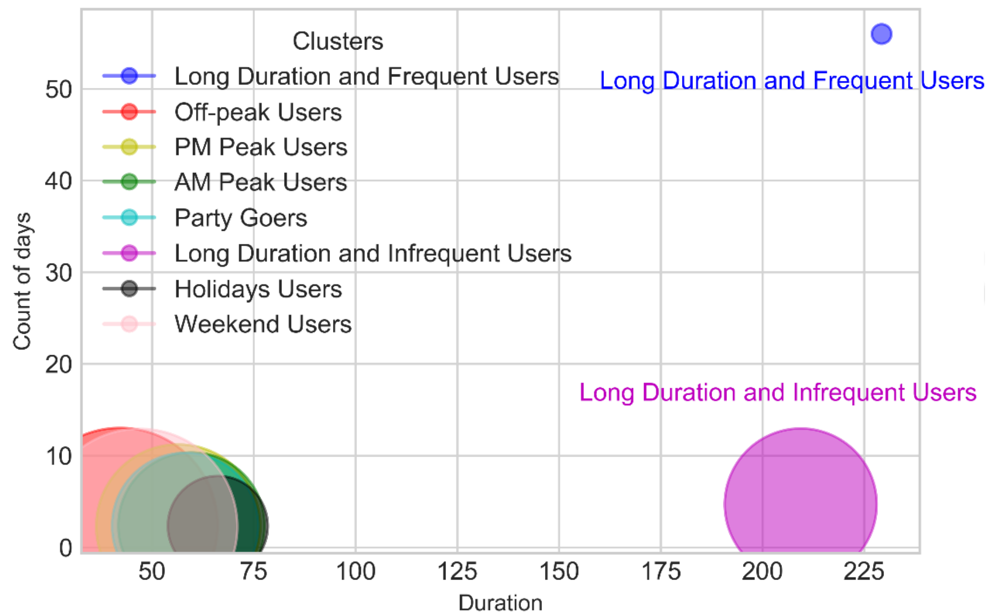

Figure 4 Bubble Chart of Duration and Count of Days for Each Cluster

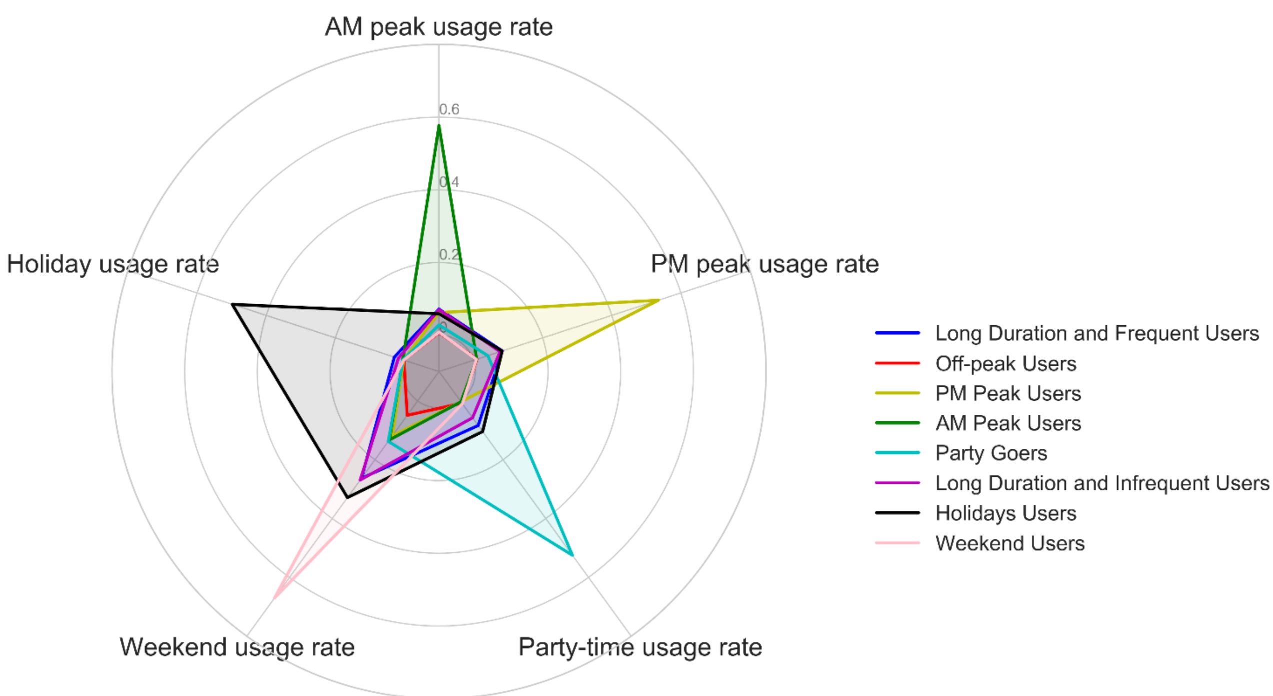

Figure 5 Radar Chart of Usage Time Rate for Each Cluster

TABLE 3 Clusters of users and mean value of attributes for each cluster

	Long Duration and Frequent Users	Off-peak Users	PM Peak Users	AM Peak Users	Party Goers	Long Duration and Infrequent Users	Holidays Users	Weekend Users
AM peak usage rate	7.2%	0.7%	6.2%	57.7%	2.8%	7.1%	5.9%	0.9%
PM peak usage rate	8.3%	1.0%	53.5%	0.9%	4.1%	7.7%	8.3%	0.8%
Party-time usage rate	8.4%	0.8%	0.3%	0.3%	52.5%	5.7%	10.4%	1.0%
Weekend usage rate	26.3%	4.9%	11.9%	13.3%	13.8%	26.9%	32.9%	67.0%
Holiday usage rate	2.9%	0.2%	0.3%	0.5%	0.4%	1.8%	49.8%	0.2%
Percent of Users	0.2%	21.2%	15.0%	12.0%	11.9%	12.9%	5.6%	21.3%

Each cluster is briefly described one by one below. Additionally, **Figure 6** visualizes the time usage distribution of each cluster by the hour of the day and day of the week was created to better understand the temporal usage difference among the eight clusters.

*1. Long Duration and Frequent Users (shown in the 2nd column of **Table 3** and the 1st row of **Figure 6**)*
As shown in **Table 3**, the long duration and frequent users cluster only has 0.2% of all users. The count of days in this cluster is around 56 on average (results not shown), which is considered frequent users compared to other clusters. This cluster was also known as long duration users since the average duration between first and last Uber request in the Transit app is 229 days (results not shown). Its time usage distribution shown in the 1st row of **Figure 6** indicates that there was no clear time usage pattern among long duration and frequent users.

*2. Off-peak Users (shown in the 3rd column of **Table 3** and the 2nd row of **Figure 6**)*
As shown in **Table 3**, the users in this cluster have a small value for Uber request rates in all the variables related to time usage rate. Which indicates that Uber requests made by these users occurred on weekdays except during peak period and party-time. The time distribution shown in the 2nd row of **Figure 6** suggests the same conclusion. Therefore, this cluster was labeled as off-peak users.

*3. PM Peak Users (shown in the 4th column of **Table 3** and the 3rd row of **Figure 6**)*
As shown in **Table 3**, although the users in this cluster only requested twice on average, there is a clear pattern in which most of the cluster's Uber requests were made during the PM peak period on weekdays. Uber requests made by these users during other time periods are small. The PM peak users accounted for a relatively large proportion of all users (15.0%). This group might consist of commuters who take Uber for returning home from work or school. Its time usage distribution shown in the 3rd row of **Figure 6** indicates that there was a large spike during the PM peak period.

*4. AM Peak Users (shown in the 5th column of **Table 3** and the 4th row of **Figure 6**)*
As shown in **Table 3**, the user in this cluster has AM peak usage rate of 57.7%. Uber requests made during other time periods by these users are small. This cluster might consist of commuters who take Uber for going to work or school in the morning. Its time usage distribution shown in the 4th row of **Figure 6** indicates that there was a large spike during the AM peak period.

*5. Party Goers (shown in the 6th column of **Table 3** and the 5th row of **Figure 6**)*
As shown in **Table 3**, the user in this cluster has party-time usage rate of 52.5% and weekend usage rate of 13.3%. Those users might use Uber when the transit system is running less frequently late at night or they do not want to drive after drinking. Its time usage distribution shown in the 5th row of **Figure 6** indicates that there was a large spike during 10:00 p.m. on Friday to 2:59 a.m. on Saturday and 10:00 p.m. on Saturday

to 2:59 a.m. on Sunday.

*6. Long Duration and Infrequent Users (shown in the 7^{th} column of **Table 3** and the 6^{th} row of **Figure 6**)*
As shown in **Table 3**, there are 209 days on average from the first Uber request to the last Uber request in this cluster (results not shown). However, the number of day count is much smaller. This cluster was named it the long duration and infrequent users to distinguish it from others. Its time usage distribution shown in the 6^{th} row of **Figure 6** indicates that there was no clear time usage pattern among long duration and infrequent users.

*7. Holidays Users (shown in the 8^{th} column of **Table 3** and the 7^{th} row of **Figure 6**)*
As shown in **Table 4**, the users in this cluster have a holiday usage rate of 50.0% and weekend usage rate of 32.9%. This could be explained by the overlap of the day of holidays and weekends. The users in this cluster might use Uber because the transit system has less availability due to the schedule changes for the holiday.

*8. Weekend Users (shown in the 9^{th} column of **Table 3** and the 8^{th} row of **Figure 6**)*
As shown in **Table 3,** this cluster accounts for 21.3% of all users. Uber requests made during other time periods by these users are small. Its time usage distribution shown in the 8^{th} row of **Figure 6** indicates that there was a large spike in Uber requests on the weekend.

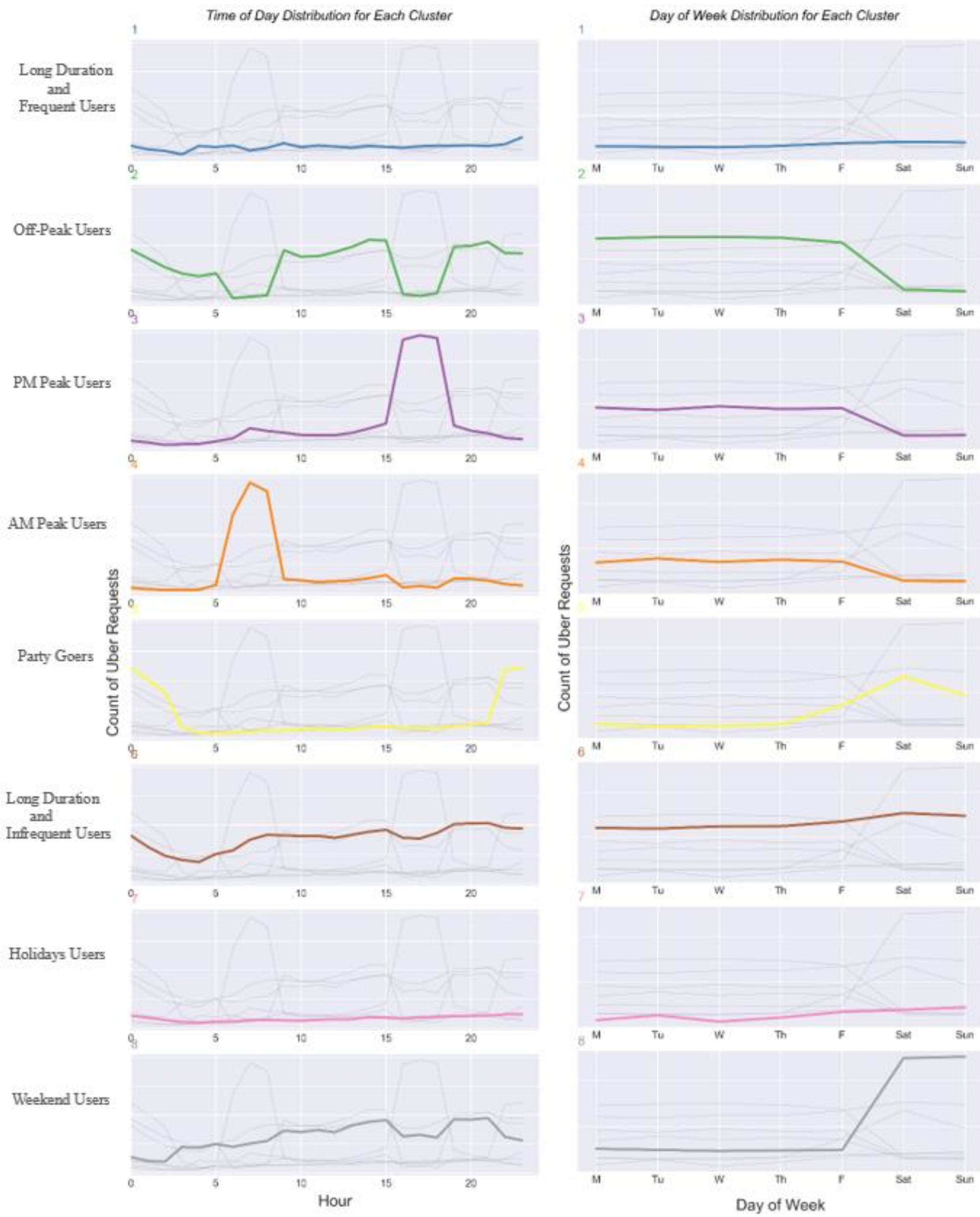

Figure 6 Temporal Features of Each Cluster

Comparison with TLC Uber pickup data and transit

The Uber requests data via the Transit app from January to June in 2017 was selected to compare with the TLC Uber pickup data for the same months in 2015 and subway vehicle trip data for November 2016 to

December 2017. The subway vehicle trip data was grouped by weekday, Saturday, and Sunday since the subway schedule is the same throughout the week and changes for weekend service.

Figure 7 shows the temporal distribution of TLC Uber pickups, Transit app Uber requests, and subway vehicle trips for each day of the week. As **Figure 7** shows, the AM peak period is different between TLC Uber pickups and Transit app Uber requests. The TLC's peak period is around 8 a.m. while the Transit app's peak period is concentrated around 7 a.m. Another difference between the two datasets is the comparison of AM and PM peak period usage for each dataset. The TLC Uber pickups shows higher usage in the PM peak period than in the AM peak period while the Transit app Uber requests data have similar usage in the AM peak period and the PM peak period. However, there are some similarities between the two datasets. This is reflected in late night demand and weekend demand. The time of day with the highest usage for both TLC Uber pickups and Transit app Uber requests occurred late at night, which is from approximately 10 p.m. to 1 a.m., especially on Fridays and Saturdays. This temporal usage feature is consistent with previous research.

Compared to Transit app Uber requests, subway trips were highly concentrated on weekdays and during the AM and PM peak time periods. Subway vehicle trips showed similar AM and PM peak spikes to the Transit app Uber requests. The subway had a much lower demand during the night compared to Transit app Uber requests.

Transit app users had similar late night and weekend patterns as Uber users. However, in terms of the difference in the peak usage of these three datasets, the usage of users who sent Uber requests via the Transit app were more similar to the subway during the AM peak. In short, the time distribution characteristics of Uber requests via the Transit app are a combination of transit usage and Uber usage, but the temporal features are more like those of Uber users.

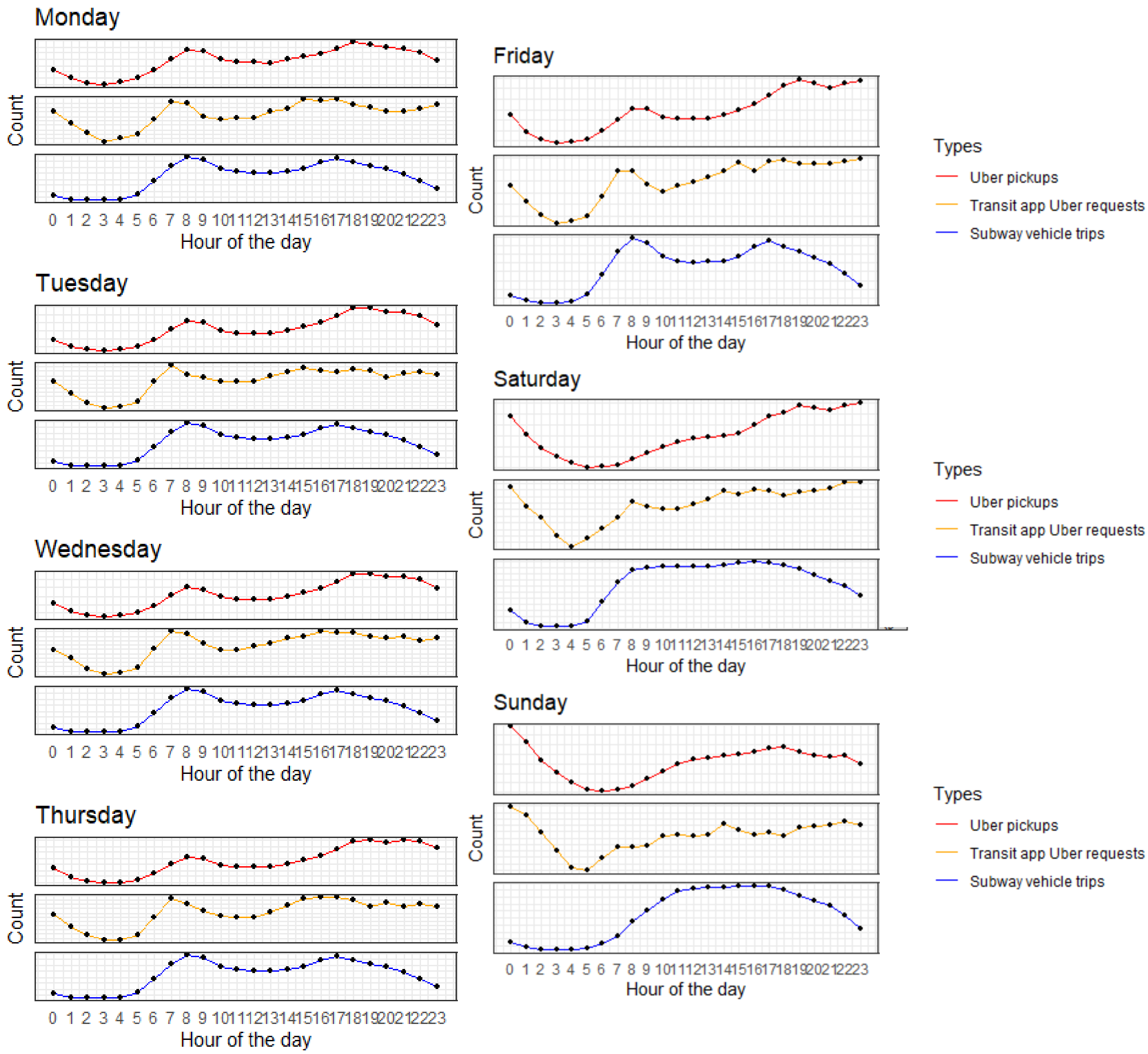

Figure 7 Uber Requests, Uber Pickups, and Subway Vehicle Trips by Time of Day and Day of Week

CONCLUSIONS AND FUTURE RESEARCH

The temporal characteristics of Uber requests via Transit App in New York City were analyzed to find patterns in Uber requests. Exploratory results suggest that the time trend line peaks of Transit App users' Uber requests were mostly concentrated in the AM and PM peak periods during weekdays and the late night periods on Fridays and Saturdays. These results are consistent with the findings from prior studies related to temporal analysis of ridehailing usage. Based on Uber requests data, K-means clustering showed eight distinguishable clusters of users: long duration and frequent users, off-peak users, PM peak users, AM peak users, party goers, long duration and infrequent users, holidays users, and weekend users. The clustering resulted in different trip purposes among users who sent Uber requests from the Transit app. This supports the main trip purposes of using ridehailing services from prior research, which were going to social events and going to or from work or home. The comparison of Uber request data via the Transit app with TLC Uber pickup data and subway GTFS data in New York City was conducted to further understand the temporal difference between ridehailing and transit. The results of this comparison suggest that the time distribution characteristics of Uber requests via the Transit app are likely a combination of transit usage and Uber usage, but the temporal features are more like those of Uber users.

There are many potential areas for improvement and future research that emerged from this analysis. First, this study only considered temporal variables. In future research, the spatial information of data could be included in further exploration of users' trip patterns and trip purposes. In addition, the accessibility of transit facilities around the user's location might be a good measure to better analyze the relationship between Uber requests and transit usage. Future research can further help transportation departments to better coordinate the distribution of ridehailing services and public transportation to meet the travel needs of users.

ACKNOWLEDGMENTS

The authors would like to thank the Transit app for providing the data; they are also grateful to Jake Sion and David Block-Schachter for reviewing the manuscript. The authors would like to thank the Writing Center at University of Tennessee, Knoxville who provided proofreading for this paper.

AUTHOR CONTRIBUTIONS

The authors confirm contribution to the paper as follows: data collection: Brakewood; data preparation: Guo, Haque; study conception and design: Guo, Haque, Brakewood; literature review: Crossland, Guo; analysis and interpretation of the results: Guo; draft manuscript preparation: Guo, Brakewood, Crossland; All authors reviewed the results and approved the final version of the manuscript.

REFERENCES

[1] *Taxonomy and Definitions for Terms Related to Shared Mobility and Enabling Technologies.* SAE International. https://saemobilus.sae.org/content/J3163_201809/. Accessed July 27, 2020.

[2] Clewlow, R. R., and G. S. Mishra. *Disruptive Transportation: The Adoption, Ytilization, and Impacts of Ride-hailing in the United States*. Research Reports UCDITS-RR-17-07, Institute of Transportation Studies, University of California, Davis CA, 2017.

[3] Brakewood, C., N. Ghahramani, J. Peters, E. Kwak, and J. Sion. Real-time Riders: A First Look at User Interaction Data from the Back End of A Transit and Shared Mobility Smartphone App. *Transportation Research Record,* Vol. 2658, No. 1, 2017, pp. 56-63.

[4] *More Options. Shifting Mindsets. Driving Better Choices.*, Uber and Mothers Against Drunk Driving. https://newsroom.uber.com/wp-content/uploads/2015/01/UberMADD-Report.pdf. Accessed July 27, 2020.

[5] Schaller, B. *Unsustainable? The Growth of App-Based Ride Services and Traffic, Travel and the Future of New York City*, Schaller Consulting. http://www.schallerconsult.com/rideservices/unsustainable.htm. Accessed July 27, 2020.

[6] Cooper, D., J. Castiglione, A. Mislove, and C. Wilson. Profiling Transport Network Company Activity Using Big Data. *Transportation Research Record,* Vol. 2672, No. 42, 2018, pp. 192-202.

[7] Board, T. R., E. National Academies of Sciences, and Medicine. *Broadening Understanding of the Interplay Among Public Transit, Shared Mobility, and Personal Automobiles*. The National Academies Press, Washington, DC, 2018.

[8] Henao, A., and W. E. Marshall. The Impact of Ride-hailing on Vehicle Miles Traveled. *Transportation,* Vol. 46, No. 6, 2019, pp. 2173-2194.

[9] Gehrke, S. R., A. Felix, and T. G. Reardon. Substitution of Ride-Hailing Services for More Sustainable Travel Options in the Greater Boston Region. *Transportation Research Record,* Vol. 2673, No. 1, 2019, pp. 438-446.

[10] Gehrke, S. *A Survey of Ride-Hailing Passengers*. TREC Friday Seminar Series. 152, 2018.

[11] Circella, G., F. Alemi, K. Tiedeman, S. Handy, and P. Mokhtarian. *The Adoption of Shared Mobility in California And Its Relationship with Other Components of Travel Behavior*. National Center for Sustainable Transportation, Davis, CA, 2018.

[12] Rayle, L., D. Dai, N. Chan, R. Cervero, and S. Shaheen. Just A Better Taxi? A Survey-based Comparison of Taxis, Transit, and Ridesourcing Services in San Francisco. *Transport Policy,* Vol. 45, 2016, pp. 168-178.

[13] Board, T. R., E. National Academies of Sciences, and Medicine. *Shared Mobility and the Transformation of Public Transit*. The National Academies Press, Washington, D.C., 2016.

[14] Henao, A. *Impacts of Ridesourcing-Lyft and Uber-on Transportation Including VMT, Mode Replacement, Parking, and Travel Behavior*. University of Colorado at Denver, 2017.

[15] Mahmoudifard, S., A. Kermanshah, R. Shabanpour, and A. Mohammadian. Assessing Public Opinions on Uber as a Ridesharing Transportation System: Explanatory Analysis and Results of a Survey in Chicago Area. Presented at 96th Annual Meeting of the Transportation Research Board, Washington, D.C., 2017.

[16] Erhardt, G. D., S. Roy, D. Cooper, B. Sana, M. Chen, and J. Castiglione. Do Transportation Network Companies Decrease or Increase Congestion? *Science advances*, Vol. 5, No. 5, 2019, p. eaau2670.

[17] Hall, J. D., C. Palsson, and J. Price. Is Uber A Substitute or Complement for Public Transit? *Journal of Urban Economics*, Vol. 108, 2018, pp. 36-50.

[18] Davidson, A., J. Peters, and C. Brakewood. Interactive Travel Modes: Uber, Transit, and Mobility in New York City. Presented at 96th Annual Meeting of the Transportation Research Board, Washington, D.C., 2017.

[19] Ghahramani, N., and C. Brakewood. Requests for Ridehailing During an Extreme Weather Event: Exploratory Analysis of New York City. *Journal of Urban Planning Development,* Vol. 146, No. 2, 2020, p. 04020006.

[20] Kriegel, H.-P., E. Schubert, and A. Zimek. The (Black) Art of Runtime Evaluation: Are We Comparing Algorithms or Implementations? *Knowledge and Information Systems*, Vol. 52, No. 2, 2017, pp. 341-378.

[21] MacKay, D. J., and D. J. Mac Kay. *Information Theory, Inference and Learning Algorithms*. Cambridge university press, 2003.
[22] FiveThirtyEight.com. *uber-tlc-foil-response*, GitHub. https://github.com/fivethirtyeight/uber-tlc-foil-response. Accessed July 27, 2020.
[23] OpenMobilityData. *NYC Subway GTFS*. https://transitfeeds.com/p/mta/79?p=3. Accessed July 27, 2020.