TRR
JOURNAL OF THE TRANSPORTATION RESEARCH BOARD

Research Article

Transportation Research Record
2020, Vol. 2674(9) 383–394

DOI: 10.1177/0361198120929338
journals.sagepub.com/home/trr
SAGE

K-Prototypes Segmentation Analysis on Large-Scale Ridesourcing Trip Data

Jason Soria[1], Ying Chen[2], and Amanda Stathopoulos[1]

Abstract

Shared mobility-on-demand services are expanding rapidly in cities around the world. As a prominent example, app-based ridesourcing is becoming an integral part of many urban transportation ecosystems. Despite the centrality, limited public availability of detailed temporal and spatial data on ridesourcing trips has limited research on how new services interact with traditional mobility options and how they affect travel in cities. Improving data-sharing agreements are opening unprecedented opportunities for research in this area. This study examined emerging patterns of mobility using recently released City of Chicago public ridesourcing data. The detailed spatio-temporal ridesourcing data were matched with weather, transit, and taxi data to gain a deeper understanding of ridesourcing's role in Chicago's mobility system. The goal was to investigate the systematic variations in patronage of ridehailing. K-prototypes was utilized to detect user segments owing to its ability to accept mixed variable data types. An extension of the K-means algorithm, its output was a classification of the data into several clusters called prototypes. Six ridesourcing prototypes were identified and discussed based on significant differences in relation to adverse weather conditions, competition with alternative modes, location and timing of use, and tendency for ridesplitting. The paper discusses the implications of the identified clusters related to affordability, equity, and competition with transit.

Transportation network companies (TNCs) are prominent in many urban transportation ecosystems. The growth of ridesourcing patronage is attributed to the convenience compared with traditional modes. Widespread adoption of smartphones embedded with GPS technology has enabled travelers to street-hail rides through mobile applications, get real-time information about waiting times, and make cashless e-payments, as well as rating driver performance. More recently, major operators have used algorithms to match passengers along similar routes in real time, providing a new generation of shared rides. However, the proliferation of ridesourcing services has been a disruptive force in the mobility landscape and many questions have been raised about negative externalities and the socio-spatial equity of supply. There is significant need for more insight on ridesourcing use patterns for cities to prepare policy, regulation, and infrastructure plans. Yet, this mobility transformation has not been widely studied, as many TNCs are reluctant to make their data publicly available. Recent data-sharing agreements with the City of Chicago, IL has enabled researchers to examine the role of ridesourcing in the transportation ecosystem using detailed temporal and spatial data on ridesourcing trips.

Several studies have characterized the adoption, frequency, and attitudes toward ridesourcing (*1–4*), but none have used trip data at the scale and scope provided by the City of Chicago. This study uses this newly released trip data, to develop insights about the role ridesourcing plays in the transportation ecosystem. The detailed spatio-temporal patronage data from operators Uber, Lyft, and Via is merged with local transit, taxi, and weather observations. The goal of this research was to investigate the variations in patronage of ridehailing. By studying the emerging mobility patterns present in the data and examining the uneven locational and

[1]Northwestern University, Department of Civil and Environmental Engineering, Technological Institute, Evanston, IL
[2]Northwestern University, Department of Civil and Environmental Engineering, Transportation Center, Evanston, IL

Corresponding Author:
Jason Soria, jason.soria@u.northwestern.edu

pricing patterns we developed insights about the motivations of users.

This study utilized an unsupervised learning algorithm to examine the underlying patterns in the data. Owing to the mixed data types (i.e., data containing both numeric and qualitative/categorical variables), the K-prototypes algorithm was selected (*5*). This is an extension of the K-means algorithm that deals with categorical data. The model produces a classification of the data into K number of prototypes, similar to K-means clusters.

This study contributes to the literature by providing a closer look at large, relatively disaggregate TNC data from a major metropolitan area. After tuning parameters for the best fit, the optimal number of prototypes to describe the data was six. The first group of users (i.e., prototype) contained trips that occurred in adverse weather conditions such as rainy weather. The second prototype involved trips that occurred in the evening. The third prototype represented trips that were typically longer in distance but not shared. The fourth prototype was defined by the trip origin and destination being to one of the two major airports in Chicago: O'Hare and Midway International. The fifth prototype was defined by short, solo trips occurring in areas that are well served by transit. The sixth prototype was defined by nearly all observations being shared rides.

This paper is structured with the following sections. A Literature Review that covers the state-of-the-art in TNC research follows. The Methodology section then explains model development and attribute selection. The subsequent section reviews the algorithm's output and leads into a policy discussion. Finally, the Conclusions section contains a review of what was achieved in this study, its limitations, and possible future works.

Literature Review

Since the inception of Uber in 2009, shortly followed by Lyft, ridehailing has already undergone significant service evolution, as discussed by Shaheen and Cohen (*6*). Among these, shared ridehailing, or ridesplitting, matches individuals in real time based on shared routes; and microtransit, or curb-to-curb options, match users into van-sized vehicles based on dynamic or planned routing. For the major TNCs, riders can now typically decide between solo travel or ridesplitting in the same application. Authorizing sharing typically results in lower fares but longer travel times, as the trip now includes several stops that may cause the vehicle to deviate from the optimal path for a single origin–destination pair.

Researchers have tried to develop a better understanding of ridesourcing travel, but data are scarce. Uber and Lyft do not generally share comprehensive trip level data in a several of its markets thereby impeding the progress of empirical studies in this field. Owing to empirical data scarcity, researchers have become creative to gain insights on usage and possible impacts. Henao and Marshall went so far as to become a TNC driver to collect trip information (*7*). Their research suggested extra vehicle miles were generated by deadheading, a form of inefficiency that is difficult to map using stated data. After accounting for deadheading mileage, the average occupancy of a TNC vehicle was less than one person. Such empirical approaches can offer new insights but some caveats limit the generality. The trips recorded in this study were a relatively small sample of total rides in the entire region and are biased because of all data coming from a single customer search strategy (*7*). Other groups have utilized their own ridehailing experiments to highlight competition with transit (*8*) and equity compared with taxis (*8*,*9*), further highlighting the need for a comprehensive dataset for analysts to access. Along with access to data, connections between stated and revealed preference data may also lead to more informed policy decisions (*10*).

The current understanding of ridesourcing travel is mostly informed by survey research. In the following we briefly provide an overview of relevant ridesourcing work and relate findings to the current analysis of real large-scale data. Several studies delve into the trip purposes of ridesourcing trips. Defined by its utilization of large-capacity vehicles, microtransit (also known as demand-responsive transit or on-demand transit) can serve as a tool to address public transit overcrowding and the first–last mile problem (*11*). It is mainly utilized for commuting (*11*,*12*). Instead, trips made by the more taxi-like TNCs are mostly for social/recreational trips (*7*,*13*–*15*). Trip purpose is not included in the current analysis because of data anonymization. However, in future works, spatial examination of locations of interest combined with other trip attributes could be used to infer trip types.

The effects of TNCs on the transportation system, specifically via the competition or complementarity with other modes, is a core area of research. In particular, owing to the similarity of the services, the impact on taxis has been widely studied. TNCs have significantly reduced the demand for traditional taxi services such that taxi drivers have altered their strategies to remain profitable (*16*–*21*). Focusing on transit competition, Schwieterman and Smith found that TNCs are preferred over public transit, especially when origin–destination pairs are not well served by transit (*8*). Along similar lines, ridesourcing's relationship with public transit was found to be complementary within large cities with small transit agencies (*22*). Further determinants of ridesourcing use relate to the travel environment. Frei et al.

found that weather affects TNC usage (*23*). Though their study focused on microtransit, other TNC services may also be affected by adverse weather.

Owing to these observations, the ridesourcing trip data used in this project were matched in space and time with data on weather, equivalent origin–destination transit travel times, and peak taxi demand.

Model Development

The data analysis used to examine patterns of ridesourcing use in this project was an unsupervised learning technique called K-prototypes. K-prototypes is similar to K-means, since both aim to cluster several observations together according to their attributes. The advantage K-prototypes has in this situation is its ability to also accept categorical variables. More details on K-prototypes development can be found in research by Huang (*5*).

The challenge of dealing with categorical variables has been considered for segmentation analysis. The problem is that the K-means algorithm relies on all variables being numerical. Specifically, in the K-means algorithm, for a continuous variable such as travel time, the distance between an observation's travel time and the proposed cluster's mean travel time is the key element for identifying clusters among observations. With a categorical variable such as vehicle type, distance is no longer applicable. One strategy to include categorical variables in the K-means algorithm is to code each category as a dummy variable (0 or 1). The distance calculated by K-means algorithm for a categorical variable is then 0 or 1, which is not informative. With the K-prototypes algorithm the mode of the category is used as is a measure of a matching coefficient. The formulation from Huang of the K-prototypes algorithm is summarized in Equations 1 to 4 (*5*).

The matching of observations to prototypes involves reducing the error or cost function. This cost function represents the distance between observation data and the assigned prototype center. Equation 1 shows that the error, E, is the sum of distances from the prototype center. X_i is the attributes of trip i, Q_l is the center of Prototype l, and y_{il} is a dummy variable that is equal to 0 when trip i is assigned to Prototype l. This is then the sum of squared distances for n TNC trips across k number of prototypes. Equation 2 breaks down $d(X_i, Q_l)$ into numerical and categorical components, where the first term is the squared numerical distance of attribute j of trip i from the center for attribute j of Prototype l; the second term includes a component to determine the weight, γ_l, of the categorical variables to the total error E. The error of Prototype l is then calculated in Equation 3, where E_l^c is further explained by Equation 4. C_j is the set of all unique values of categorical attribute j, and $p(c_j \in C_j \mid l)$ is then the probability of unique value q_j from set C_j being in Prototype l.

$$E = \sum_{l=1}^{k} \sum_{i=1}^{n} y_{il} d(X_i, Q_l) \tag{1}$$

$$d(X_i, Q_l) = \sum_{j=1}^{m_r} \left(x_{ij}^r - q_{lj}^r\right)^2 + \gamma_l \sum_{j=1}^{m_c} \delta\left(x_{ij}^c, q_{lj}^c\right) \tag{2}$$

$$E_l = \sum_{i=1}^{n} y_{il} \sum_{j=1}^{m_r} \left(x_{ij}^r - q_{lj}^r\right)^2 + \gamma_l \sum_{i=1}^{n} y_{il} \sum_{j=1}^{m_c} \delta\left(x_{ij}^c, q_{lj}^c\right) = E_l^r + E_l^c \tag{3}$$

$$E_l^c = \gamma_l \sum_{j=1}^{m_c} n_l \left(1 - p\left(q_{lj}^c \in C_j | l\right)\right) \tag{4}$$

The advantage of using the K-prototypes algorithm over other clustering algorithms is highlighted by Equation 4. A common way to code categorical variables for other data-driven methods is to use one-hot encoding. Using this method, the unique values of a category are coded as a dummy variable equal to 1 when denoting the variable of interest and 0 otherwise. Algorithms using one-hot encoded data fail to recognize that these unique values belong to a categorical variable because categories are reduced to 0 or 1. The advantage of the K-prototypes algorithm lies in the recognition that these values denote categorical variables and using the probability of a unique value from a set C_j being in Prototype l.

This model was implemented and tuned with the R programming language using the "clustMixType" package (*24*,*25*). Using this package, the error is minimized and the weighting of the categorical error is optimized. Much like other clustering methods, the number of prototypes is a tunable parameter. The final tunable parameters are discussed in the Results section.

Data Description

The data used in this project were drawn from TNC trip data provided by the City of Chicago (*26*). The trip data began on November 1, 2018 and was updated monthly. For the purpose of obtaining lower optimization times and being able to match the equivalent transit travel times, the data was partitioned to weekdays in November 2018. Holidays were not included. This left a total of 3,085,070 trips in the dataset. The trips were grouped at the census tract level and include variables such as travel time, travel distance, fare, whether it was a shared trip (and if it was, how many other passengers were included), census tract origin–destination pairs, and timestamp of pickup and drop-off rounded to the nearest 15-min increment.

The weather data were collected from OpenWeatherMap, specifically for the City of Chicago in November 2018 (*27*). The data were at the hourly level and include amount of rain and snow in the previous hour, a qualitative description of the weather (such as raining, hazy, sunny, etc.), and temperature. The station collecting the data was located at O'Hare International Airport at the northwest tip of the city limits. The supplementary transit travel times dataset was created for each unique origin–destination–time–day tuple. Transit travel time estimates were obtained using the Google Distance Matrix (Advanced) API by providing the census tract of origin, the census tract of destination, travel mode (transit), and departure time (*28*). From the API, approximate transit travel times between point to point origin–destination pairs were collected. This data included the expected access (walking to transit stop) and expected wait time just as one would view them from a navigation assistant device/app. Since the data were only available from 6:00 a.m. to 10:00 p.m. for much of the network, the TNC trips data, which are publicly available and collected by the City of Chicago, were also restricted to these hours. Even when available, transit typically operates at reduced capacity after 10:00 p.m., so to enable a fair comparison we restricted the analysis to these regular travel hours. The taxi trip data used in this research corresponded to peak taxi demand in 2014 and are further described by Chen et al. (*29*). The data are the monthly taxi trips between census tract origin–destination pairs, which are referred to as monthly taxi frequency later in the analysis. These data were included to characterize and compare the spatial relationship of taxi usage by matching each ridesourcing trip with the total taxi flow between the same origin and destination. Table 1 contains descriptive statistics for the numerical analysis data.

Table 1. Descriptive Statistics of Ridehailing, Transit, Taxi, and Weather Data

Numerical variable	Median	Mean (SD)
Travel time (min)	13.32	15.47 (9.98)
Distance (mi)	2.70	3.79 (3.19)
Total fare ($)	10.00	11.24 (6.27)
Parties joined in trip	1	1.32 (0.77)
Humidity (%)	71.00	73.58 (11.89)
Wind speed (mph)	3.00	3.82 (2.33)
Rain last hour (in.)	0.00	0.061 (0.26)
Minute after midnight	930.00	887.5 (268.20)
Transit travel time (min)	17.95	21.10 (15.40)
Monthly taxi frequency	1004	14,976 (36,875.57)

Note: SD = standard deviation.

Analysis of Results

During the estimation phase, the K-prototypes algorithm was tuned to select the optimal number of prototypes. This was determined by developing models including several prototypes ranging from 2 to 14 and calculating the total cost across all observations. The final number of prototypes chosen was six, based on interpretability of segmentation variables and guidance from the plot, which in Figure 1 shows a clear "elbow" at six prototypes (*30*). An elbow occurs when adding more clusters does not sufficiently improve the objective function. γ is the tradeoff between numerical cost and categorical cost optimized by the "kproto" function in the clustMixType package and was estimated to be 1.33 for all prototypes as per Equations 2 and 4 (*25*). There is no intuitive meaning to this value except that it can be user-specified, and higher values mean that the categorical variables receive a higher weight. Figure 1 shows how many observations belong in each prototype cluster. A summary of the top six origin and destinations, respectively, are given in Table 2. The clustering results are shown in Table 3 along with mean values of the explanatory attributes in each prototype.

An important observation related to variable selection in the presence of potential correlation needs to be made. In practice, transportation modeling often deals with concerns surrounding the correlation among time, distance, and cost, either by interacting or dropping variables. Yet ridesourcing represents a special case because of the dynamic demand-responsive pricing that relaxes this typical correlation. While we cannot separate out instances of surge pricing from this data we note that some interesting relationships were discovered when

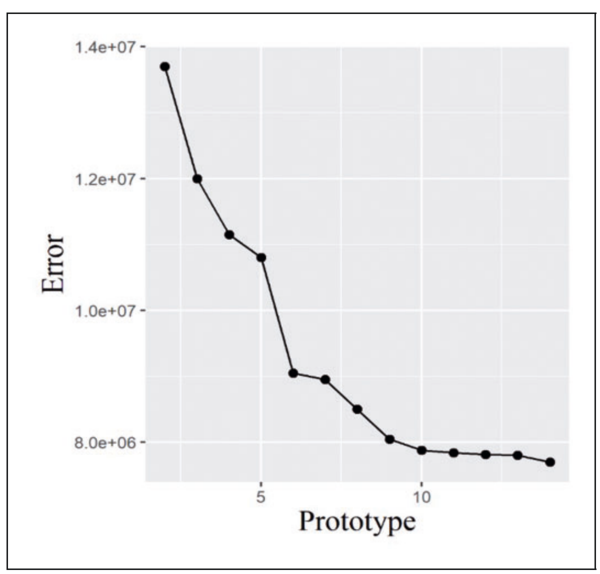

Figure 1. Selection K number of prototypes.

Table 2. Community Area Characteristics

Community area	Per capita income ($)	Bar and tavern density (per mi^2)	Transit access time* (min)
Chicago average	32,534	4.78	19.75
Near North Side	91,948	32.66	13.00
Near West Side	50,394	10.51	10.57
West Town	54,429	11.86	11.03
Loop	77,722	46.84	9.53
Lincoln Park	73,965	13.43	12.94
Lake View	67,066	19.15	11.17
Midway	28,925	3.27	33.79
O'Hare	27,212	0.17	84.64

comparing prototypes. Notably, though the variables are correlated, *on average*, within the specific clusters the relationship revealed vast differences in per mile costs. Table 4 and related discussions highlight these insights.

We now turn to summarize the contours of the six user clusters. On the whole, the analysis did not produce prototypes that were heavily differentiated by temperature or snow fall in the past hour. Yet weather effects were evident in the first segment of users (Prototype 1 or P1_weather). P1_weather was the second largest prototype and was characterized by its relatively low total fares and short travel times and distances. This short-distance travel, averaging 4 mi, was coupled with the strongest weather impacts observed, namely the presence of adverse weather seen with rain, humidity, and wind speed. The distinct nature of Prototype 1 suggests the use of ridesourcing for short-distance travel to cope with adverse weather in the early part of the day.

Prototype 2 (P2_late-night) was the largest segment with 30.2% of users. Though still representing shorter trips, it was distinct from P1 owing to the trip timing in the evening (average was 1,080 min after midnight or 6:00 p.m.) and the lack of relationship to weather conditions. Inspecting Table 5, these trips were most heavily focused in the wealthy downtown and near north areas. Furthermore, Table 2 illustrates that trips in this cluster originated from areas with the highest bar and tavern densities. This sizable cluster suggests a strong tendency to use ridesourcing for evening travel, which is in line with findings from Lavieri and Bhat (*31*).

Prototype 3 (P3_solo-non-transit) had longer travel times, which tend to be associated with longer distances (albeit not associated with airport travel), and higher total charges. This large user segment (20.4% of usage) suggests some transit gap-filling capacity of ridesourcing in Chicago, whereas the origin–destination and time-matched potentially available transit trip would take 30% longer on average with transit travel-time taken as base. Notably, considering the fixed transit pricing of $2.25, the ridesourcing trips were on average six

Table 3. Prototype Attribute Results and Percentiles

Prototype	Travel time (min)	Distance (mi)	Total fare ($)	Parties joining trip**	Humidity (%)	Wind speed (mph)**	Rain last hour (in.)**	Minute after midnight	Transit travel time (min)	Monthly taxi frequency	Percent ridesplitting (%)
P1_weather (Percentile)	637.40* (36th)	2.16 (40th)	9.05 (41th)	1.07	**82.96 (79th)**	**4.32**	**0.15**	702.5 (37th)	844 (39th)	11,695 (76th)	18.77
P2_late-night	600.9 (33rd)	2.07 (38th)	8.85 (41st)	1.07	66.07 (31st)	3.57	0.01	**1,080 (72nd)**	804 (37th)	10,031 (75th)	17.65
P3_solo-non-transit	1,284.0 (78th)	5.74 (80th)	15.56 (85th)	1.04	72.29 (54th)	3.66	0.03	878 (45th)	**1,838 (78th)**	**3,558 (64th)**	**10.48**
P4_airport	**2,014.0 (94th)**	**12.25 (97th)**	27.64 (97th)	1.17	75.09 (61st)	3.96	0.08	826.4(40th)	**3,392 (97th)**	**3,840 (65th)**	16.58
P5_transit-competitive	572.2 (31st)	1.39 (21st)	8.40 (40th)	1.12	73.64 (56th)	3.70	0.05	815.9(39th)	**501 (19th)**	**14,4687 (98th)**	13.76
P6_ridespli-tting	1,320.0 (79th)	4.83 (74th)	7.43 (15th)	**3**	74.12 (59th)	3.66	0.05	870.7(45th)	**1,545 (69th)**	**5,286 (68th)**	**100**

Note: Bold type indicates important feature.
**Non-continuous variables with low range do not have percentiles included.

Table 4. Prototype Specific Average Costs and Speed

Prototype	Average $ per mile traveled	Average $ per minute travel time	Average speed (mph)	% transit travel time above ridesourcing equivalent trip*
All trips	2.97	0.73	12.16	36.39
P1_weather	4.19	0.85	12.20	32.41
P2_late-night	4.28	0.88	12.40	33.80
P3_solo-non-transit	2.71	0.73	16.09	43.15
P4_airport	2.26	0.82	21.90	68.42
P5_transit_competitive	6.04	0.88	8.75	−12.44
P6_ridesplitting	1.54	0.34	13.17	17.05

*(Transit travel time – Ridesourcing travel time)/Ridesourcing travel time

Table 5. Prominent Prototype Origins and Destinations

	Origins		Destinations	
Prototype	Community	% in prototype	Prototype	% in prototype
P1_weather	Near North Side	22.62	Near North Side	24.22
	Near West Side	13.63	Loop	15.40
	West Town	9.638	Near West Side	14.15
	Loop	9.537	West Town	5.280
	Lincoln Park	5.878	Lincoln Park	5.012
	Lake View	5.169	Lake View	4.561
P2_late-night	Near North Side	24.96	Near North Side	23.78
	Near West Side	12.84	Near West Side	13.16
	Loop	12.11	West Town	8.932
	West Town	7.502	Lincoln Park	8.162
	Lincoln Park	7.224	Loop	8.085
	Lake View	7.103	Lake View	7.664
P3_solo-non-transit	Near North Side	16.77	Loop	18.21
	Loop	10.47	Near North Side	12.07
	Lake View	9.795	Near West Side	11.25
	Near West Side	8.263	Lake View	7.916
	Lincoln Park	7.144	West Town	4.967
	West Town	6.115	Lincoln Park	4.723
P4_airport	**Midway**	**13.80**	**O'Hare**	**16.17**
	O'Hare	**9.523**	**Midway**	**15.08**
	Near North Side	7.606	Near North Side	9.966
	Loop	6.306	Loop	7.152
	Near West Side	5.624	Near West Side	6.974
	Lake View	4.607	Lake View	3.531
P5_transit-competitive	**Loop**	**45.57**	**Loop**	**54.35**
	Near North Side	**32.15**	**Near North Side**	**21.88**
	Near West Side	8.719	Near West Side	8.013
	Lake View	5.986	Lake View	5.982
	West Town	3.156	West Town	4.160
	Lincoln Park	2.688	Lincoln Park	3.112
P6_ridesplitting	Near West Side	13.81	Near North Side	14.90
	Near North Side	11.54	Near West Side	13.20
	Loop	10.60	Loop	13.18
	West Town	7.686	West Town	6.060
	Lake View	6.510	Lake View	6.058
	Lincoln Park	5.489	Lincoln Park	4.873

Note: Bold type denotes important prototype features.

times more costly. Trips in this prototype were also typically not shared and concentrated in wealthier areas. This finding mirrors observations by Schwieterman and Smith that ridesourcing is used even in areas with a wealth of transit options, although our analysis suggests that transit speeds were relatively low (Table 4) a factor that is easily tracked by travelers using real-time smartphone navigation tools (*8*).

Prototype 4 (P4_airport) represented a small group of users with long travel times dominated by trips to and from the main airports, O'Hare or Midway International (Table 5). This prototype also had trips where the origins and destinations were not served well by transit, as seen with the average transit travel time being more than 70% longer. Along with poor transit connectivity, this cluster featured relatively low taxi frequency. The low taxi frequency showed low demand for taxis between similar airport-based trips, likely because airport trips were relatively infrequent and could be completed by carpooling with known associates such as a family member or friend. These trips' fares were more expensive than in other prototypes, but relative to the cost of traditional taxis, were still affordable. Given that it is also more convenient to utilize ridehailing than to ask a family member to drive, the strong connection between airport travel and ridehailing is unsurprising. Taxi pickups at airports are declining and other revenue streams such as parking and rental cars have also been negatively affected (*32*,*33*). This prototype highlighted the strong competitive position against both transit and taxi for airport access, albeit it did not account for the issue of waiting time that might change this assessment in particular considering departures from Chicago airport where TNCs have limited access.

Interestingly, Prototype 5 (P5_transit-competitive) was a small cluster that stood out as representing the shortest trips and for being the only case for which trips could have been served better by transit. Notably, the average transit travel times would have been 12.44% lower than the observed TNC travel times. This is in stark contrast to other prototypes, as Table 4 shows that most other prototypes' transit travel times were at least 30% longer than the ridesourcing equivalent ride. Most of these trips were in the Chicago Loop or just north of it where transit is highly concentrated in the core commercial area.

Prototype 6 (P6_ridesplitting) with 12.8% of users was defined by representing nearly all shared authorized trips. This segment appears to reflect a more cost-conscious user group given that the ridesourcing price per mile was the lowest, and the competition in relation to price and time was closer to the potentially available transit trip.

To further understand motivations of different users, Table 4 highlights the insights from comparing tradeoffs within clusters, namely fare per mile, fare per minute, and average speed to the average reference of all ridesourcing trips. Table 4 shows that P1_weather, P2_late-night, and P5_transit-competitive prototypes have a more premium fare point with higher fare per mile and fare per minute than their counterparts. The results also show steep discounts for P6_ridesplitting as it had the lowest fare per mile and fare per minute. These results confirmed the prototype interpretations, as premiums were expected (through surge pricing or similar dynamics) for rides in bad weather, late at night when drivers may be few and far between, when potential-riders are unable to drive because of inebriation, and transit-competitive trips mostly occurring in the Loop community area, which is the core commercial area. Discounts were also expected to appear with the ridesplitting prototype, as reduced fares were expected with delays incurred by the detours when picking up a different party.

Discussion

The K-prototypes analysis was geared to finding relationships in the ridesourcing data by grouping similar observations together. The merging of multiple datasets further enabled the prototypes search to identify the main ridesourcing profiles with regards to trip attributes (e.g., travel time, fare, origin and destinations, being private or shared), and competing mobility services (transit and taxi), along with weather conditions. This discussion section focuses on how the results relate to current research and can inform future research directions. Four areas of investigation are highlighted, centering on weather impacts, competition with transit and taxi, ridesplitting patterns, and spatial distribution of ridesourcing.

Weather Dependence

We found that while weather did not have a pervasive impact on ridesourcing across clusters, it did strongly determine the choices in P1_weather highlighted by higher average windspeed, humidity, and rainfall in the last hour. The identification of this prototype provided evidence that weather can have a significant impact on TNC usage for as many as 25% of trips. Taken together with results from Frei et al. demonstrating weather impacts in a microtransit choice experiment, this illustrates the importance of including weather as an explanatory variable in future TNC analyses (*23*). Inclusion of weather variables in TNC analyses could further explain the interactions between ridesourcing and other modes. For example, weather was shown to affect active modes of transport, so including weather as an explanatory variable between the relationship of ridesourcing and active mobility could inform demand in the future (*34*). This is especially useful for understanding how TNCs might relate to bikeshare, as adverse weather has been shown to decrease its demand and contribute to increased ridership of other modes (*35*). Brodeur and Nield (*36*) find that ridesourcing demand increases during adverse weather conditions and compared the supply of TNC drivers to taxis. Their results illustrate the benefit of TNCs, specifically, its dynamic pricing over taxis as a tool to increase the supply of drivers and meet consumer demand (*36*).

Mode Substitution with Transit and Taxi

The importance of understanding the relationship TNCs have with other modes was further highlighted by P4_airport and P5_transit-competitive prototypes. The airport prototype showed that airport trips were a major source of demand for ridesourcing because it provides more effective service than current transit options for many users.

The P5_transit-competitive prototype illustrated the competitive nature beyond travel time of TNCs. Though Figure 2 shows that this is a smaller portion of the trips, representing only 5.1% of the data, this is still an interesting prototype because it emphasizes how TNCs offer several advantages that go beyond shorter travel times. As discussed by Lavieri and Bhat, this is troubling because ridesourcing's relationship with transit is complex, as solo rides do not necessarily substitute transit trips (*31*). With shorter transit travel times and some demand previously met effectively by taxis, there is a need to map out the difficult to measure variables such as comfort, safety, and convenience that must be considered in conjunction with travel time. These insights may

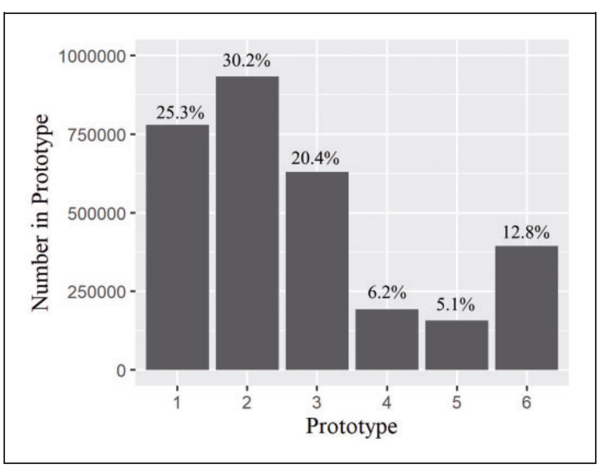

Figure 2. Prototype shares among total ridesourcing trips.

be critical to understanding the differing user perspective toward solo and shared ridehailing.

The relationship between taxis and ridehailing was more straightforward as the services are more comparable. Although the literature review section briefly discussed changes in the taxi industry, a thorough investigation of the interaction between these modes was completed by Nie (*16*). Ridesourcing was found to be an attractive alternative to taxis, however, there still remains a role for taxis in the transportation system as they remain competitive in highly dense areas during peak commuting hours. The substitution of taxis for ridesourcing also (though unintended) led to improved mobility equity in struggling communities as it is an option for those who do not possess bank accounts, credit cards, or smartphones (*37*).

Ridesplitting Patterns

Another major area of the literature focuses on the potential for TNCs to be more efficient people movers than privately driven vehicles. The dynamic ridesharing literature examines the efficiency gains of ridesplitting over private modes (*38*,*39*). Despite theoretical findings on the advantages of ridesplitting, there has been limited exploration of how this functions in real systems. A notable result from this study was the low share of split rides despite a relatively high share of riders indicating that they would be willing to share their ride. For the complete dataset, 26.7% of all trips were authorized to be shared but of these only 68.5% were actually shared. That implies that only 18.3% of the overall rides were truly pooled, likely reflecting a lack of matching travel itineraries that were close enough in space and time for the matching to occur. The percentage of authorized shared trips of all prototypes except for P6_ridesplitting was well below the 26.7% figure.

When compared with the other prototypes, the ridesplitting prototype showed that pooled trip making can be seen as a separate profile of use. To further examine the patterns of ridesplitting, Figure 3 shows the number of trips by separate trip-makers within a pooled trip for each prototype. P6_ridesplitting had a much higher share of pooled trips including more than three riders. However, this prototype only constituted 12.8% of the data. With such a small share of trips being shared, decision-makers that support TNCs should consider strategies that increase the number of pooled trips.

Spatial Patterns of Use

Lastly, we discuss the spatial distribution of travel. Notably, the majority of trips occurred in or around the Chicago Loop or airports with standouts Near North Side and Near West Side where there are typically

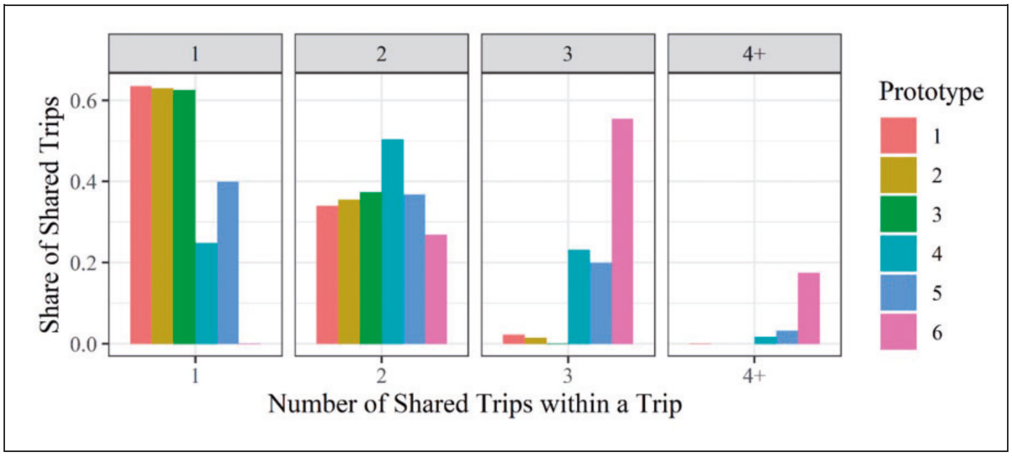

Figure 3. Number of travelers pooling a ride for actual shared trips.

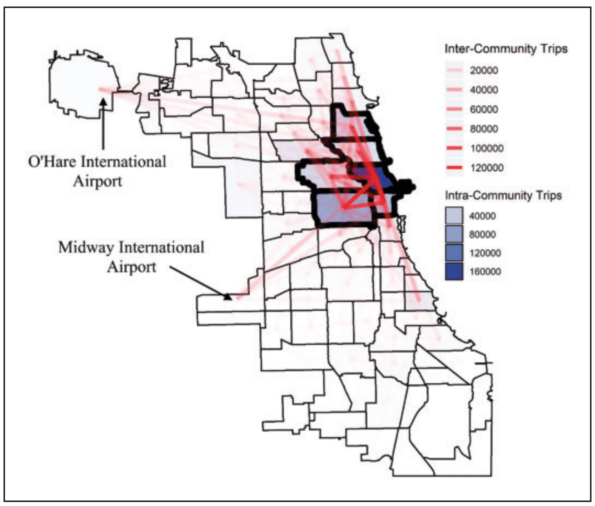

Figure 4. Ridesourcing flows in the city of Chicago, with bolded boundaries of prominent community areas.

more residential units than in the Loop and overall higher density compared with the rest of the city. Table 5 confirms that the top six origins and destinations hardly differ across prototypes. The strong concentration of flows is further illustrated in Figure 4, which shows the location of the top origin–destination pairs distinguished by bold borders. These areas tend to have a higher influx of visitors, along with more leisure landmarks such as restaurants and night clubs. The residents of these community areas tend to have higher average incomes and possess higher educational attainments than the average Chicagoan. These results are in line with findings from Clewlow and Mishra who found those who are college-educated, younger, and living in denser areas are more likely to adopt ridehailing (*40*).

Policy Implications

This study identified several patterns of ridehailing usage across Chicago that highlight the need for careful policy implementation. The discussion of policy implications will focus on modal interactions and ridesplitting owing to the need for insights to guide ongoing efforts to tweak fares, promote partnerships, and regulate ridehailing to better serve the comprehensive mobility needs of Chicago residents. The core questions that need to be explored relate to a) the challenge of providing effective service in areas with poor (or strong) transit options and b) advocating equity in hailing-access by understanding and promoting more affordable ridesplitting. Because ridehailing has been a disruptive innovation and there has been a lack of access to a comprehensive dataset on TNC activity, there is limited understanding of its relationship with other transport modes and the variation in ridesplitting adoption.

Much of the policy debate has focused on determining whether ridehailing is complementary or a substitute for other modes; this section discusses strategies that may facilitate synergy in the transport ecosystem.

Ridesourcing and Air Mobility Accessibility

Given the identification of an airport prototype with strong connections to the core commercial areas of Chicago, one major policy trend has been to control ridesourcing's effect on airport infrastructure. Examples of this include extra fees to ride into airports and curbside management of drop-offs and pickups. This prototype serves as evidence for continued development of policies that will better manage the relationship between airports and urban mobility including prominent use of ridesourcing. With this prototype showing a strong connection between the commercial core of

Chicago, policies should focus on connections that will appeal to business travelers. This remains a challenging area of research, as new options, including vertical urban air mobility, are being tested in initiatives such as UberElevate with electric vertical takeoff and landing vehicles (*41*,*42*). This highlights the need to craft regulations and partnership arrangements such as security checkpoints and luggage drop-offs (*43*). The rise of new services also highlights renewed equity and affordability concerns as they might give rise to further erosion of transit options.

Ridesourcing and Transit Performance

Conversely, the airport prototype also suggests the need for policies to improve transit connections between downtown Chicago and the airports. The segmentation analysis revealed some intriguing patterns of competition. Ridesourcing appears to be used by a small group of users even when transit is seemingly the better option (P5_transit-competitive: 5.1%), and at the same time, a sizable segment will turn to their mobility-apps in areas where transit is in abundant supply but time-performances is poor (P3_solo-non-transit: 20.3%). This opens a debate about perception and motivations of users, communicating options to travelers, and developing new partnerships.

With the transit-competitiveness prototype showing that there are real possibilities for transit to be faster than ridesourcing, a practical policy effort would be to improve the dissemination of transit information. Local transit agencies could develop advanced traveler information systems that highlight cases for which transit is competitive to increase their ridership (*44*). Other strategies could be used in conjunction with MaaS (mobility as a service) in multi-modal systems to nudge riders toward transit. Studies have shown that travel behavior can be influenced using soft strategies (*45*). These strategies such as making transit the default option or highlighting the broader benefits of supporting transit through patronage can be facilitated through a navigation application. Although this type of policy improves transit competitiveness, ridesourcing may still be dominant in many areas and promotion of sharing is vital in this situation.

More Ridesplitting?

Promoting ridesourcing naively may worsen traffic conditions, however, promoting shared rides to increase the demand for ridesplitting may be a reasonable solution. Policies that incentivize shared rides such as a tax that increases fees for exclusive rides could lead to higher demand for sharing and increased transit ridership (*46*). The tradeoff between delays and lower fares could be used to promote sharing and even increase mobility for disadvantaged groups where high fares turn them away. Policies providing travel support for unemployed and low-income residents via vouchers or further lowering fares may increases travel and opportunities when other modes are not feasible. The ongoing debate in Chicago and cities around the United States has focused on the lack of broader coverage, outside transit rich areas, of ridesourcing. Figure 4 highlights the lower share of rides occurring in and between historically underserved communities on the south and west sides of Chicago. Policies geared to promoting shared ridesourcing between underserved areas represent an opportunity to both reduce vehicle miles traveled and support disadvantaged communities.

Conclusions and Future Work

This study examined a unique TNC dataset from Chicago, IL by utilizing the unsupervised learning K-prototypes algorithm that accepts categorical data. The goal of this study was to identify patterns of TNC patronage with regard to service attributes, weather, transit, taxis, characteristics of origins and destinations, and ridesplitting. The analysis revealed six distinct ridehailing user segments. The segments were identified in relation to adverse weather conditions, evening trips, longer trips, trips to the airport, trips that would be better served by transit, and trips that are pooled. The segments were discussed in the context of the relative performance of ridesourcing as well as examining the origin and destination of flows to better interpret the spatial and performance variation.

The identification of these distinct trip types has shown where future research is warranted. The discussion in this study focused on how future research should consider factors such as weather and other external factors when estimating the demand for TNCs and other modes, airport-based mobility options in the future, understanding why TNCs have competitive advantages besides faster travel times, and why more trips are not shared. The last point made in the discussion emphasized how most of the trips were completed in and surrounding the central business district of Chicago. In summary, the concentration of trips in the downtown area where mobility options and amenities are abundant, along with notable variation in performance of ridehailing across user clusters, prompted a deeper discussion of *where* and *for whom* ridehailing enables mobility.

The main limitations of this study come from the constraints of the merged datasets. Firstly, the weather data was collected at only one location. Considering the size of Chicago and the location of the station, the data may not be representative of local weather. Secondly, the

TNC, taxi, and transit data were aggregated at the census tract level. This aggregation was needed to jointly analyze mode performance and supply but might have led to less precise findings about competing transit service. To increase the accuracy of these comparisons, more data with smaller sizes of spatial aggregation and trip details such as trip purpose are needed. Lastly, future research should expand the analysis to a longer panel of observations, thereby capturing more variation in weather and other seasonal factors that determine demand for mobility.

Acknowledgments

We thank the reviewers for their thoughtful comments that contributed toward improving our manuscript.

Author Contributions

The authors confirm contribution to the paper as follows: study conception and design: J.S. and Y.C.; data collection: J.S.; analysis and interpretation of results: J.S., Y.C., and A.S.; draft manuscript preparation: J.S., Y.C., and A.S. All authors reviewed the results and approved the final version of the manuscript.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: Amanda Stathopoulos was supported in part by the US National Science Foundation CAREER (Award no. 1847537).

References

1. Circella, G., K. Tiedeman, S. Handy, F. Alemi, and P. Mokhtarian. *What Affects US Passenger Travel? Current Trends and Future Perspectives*. UC Davis, National Center for Sustainable Transportation, CA, 2016. https://escholarship.org/uc/item/2w16b8bf. Accessed April 8, 2019.
2. Dias, F. F., P. S. Lavieri, V. M. Garikapati, S. Astroza, R. M. Pendyala, and C. R. Bhat. A Behavioral Choice Model of the Use of Car-Sharing and Ride-Sourcing Services. *Transportation*, Vol. 44, No. 6, 2017, pp. 1307–1323.
3. Alemi, F., G. Circella, S. Handy, and P. Mokhtarian. What Influences Travelers to Use Uber? Exploring the Factors Affecting the Adoption of On-Demand Ride Services in California. *Travel Behaviour and Society*, Vol. 13, 2018, pp. 88–104.
4. Alemi, F., G. Circella, P. Mokhtarian, and S. Handy. Exploring the Latent Constructs Behind the Use of Ridehailing in California. *Journal of Choice Modelling*, Vol. 29, 2018, pp. 47–62.
5. Huang, Z. Extensions to the K-Means Algorithm for Clustering Large Data Sets with Categorical Values. *Data Mining And Knowledge Discovery*, Vol. 2, No. 3, 1998, pp. 283–304.
6. Shaheen, S., and A. Cohen. Shared Ride Services in North America: Definitions, Impacts, and the Future of Pooling. *Transport Reviews*, Vol. 39, No. 4, 2018, pp. 1–16.
7. Henao, A., and W. E. Marshall. The Impact of Ride-Hailing on Vehicle Miles Traveled. *Transportation*, Vol. 46, 2018, pp. 2173–2194.
8. Schwieterman, J., and C. S. Smith. Sharing the Ride: A Paired-Trip Analysis of UberPool and Chicago Transit Authority Services in Chicago, Illinois. *Research in Transportation Economics*, Vol. 71, 2018, pp. 9–16.
9. Brown, A. E. *Ridehail Revolution: Ridehail Travel and Equity in Los Angeles*. PhD dissertation. University of California, Los Angeles, CA, 2018.
10. Chen, X., H. Zheng, Z. Wang, and X. Chen. Exploring Impacts of On-Demand Ridesplitting on Mobility via Real-World Ridesourcing Data and Questionnaires. *Transportation*, 2018.
11. Shaheen, S., and N. Chan. Mobility and the Sharing Economy: Potential to Facilitate the First- and Last-Mile Public Transit Connections. *Built Environment*, Vol. 42, No. 4, 2016, pp. 573–588.
12. Lewis, E. O. C., and D. MacKenzie. UberHOP in Seattle: Who, Why, and How? *Transportation Research Record: Journal of the Transportation Research Board*, 2017. 2650: 101–111.
13. Rayle, L., S. Shaheen, N. Chan, D. Dai, and R. Cervero. *App-Based, On-Demand Ride Services: Comparing Taxi and Ridesourcing Trips and User Characteristics in San Francisco. University of California Transportation Center*, Berkeley, CA, 2014.
14. Zhen, C. *Impact of Ride-Sourcing Services on Travel Habits and Transportation Planning*. Master's thesis, University of Pittsburgh, PA, 2015.
15. Mahmoudifard, S. M., A. Kermanshah, R. Shabanpour, and A. Mohammadian. Assessing Public Opinions on Uber and Ridesharing Transportation Systems: Exploratory Analysis and Results in a Survey in Chicago. Presented at 96th Annual Meeting of the Transportation Research Board, Washington, D.C., 2017.
16. Nie, Y. How Can the Taxi Industry Survive the Tide of Ridesourcing? Evidence from Shenzhen, China. *Transportation Research Part C: Emerging Technologies*, Vol. 79, 2017, pp. 242–256.
17. Kim, K., C. Baek, and J-D. Lee. Creative Destruction of the Sharing Economy in Action: The Case of Uber. *Transportation Research Part A: Policy and Practice*, Vol. 110, 2018, pp. 118–127.
18. Contreras, S. D., and A. Paz. The Effects of Ride-Hailing Companies on the Taxicab Industry in Las Vegas, Nevada. *Transportation Research Part A: Policy and Practice*, Vol. 115, 2018, pp. 63–70.
19. Jiang, W., and L. Zhang. The Impact of the Transportation Network Companies on the Taxi Industry: Evidence from

Beijing's GPS Taxi Trajectory Data. *IEEE Access*, Vol. 6, 2018, pp. 12438–12450.

20. Dong, Y., S. Wang, L. Li, and Z. Zhang. An Empirical Study on Travel Patterns of Internet Based Ride-Sharing. *Transportation Research Part C: Emerging Technologies*, Vol. 86, 2018, pp. 1–22.
21. Berger, T., C. Chen, and C. B. Frey. Drivers of Disruption? Estimating the Uber Effect. *European Economic Review*, Vol. 110, 2018, pp. 197–210.
22. 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.
23. Frei, C., M. Hyland, and H. S. Mahmassani. Flexing Service Schedules: Assessing the Potential for Demand-Adaptive Hybrid Transit via a Stated Preference Approach. *Transportation Research Part C: Emerging Technologies*, Vol. 76, 2017, pp. 71–89.
24. R Development Core Team. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, 2008.
25. Szepannek, G., and R. Ascenbruck. https://cran.r-project.org/web/packages/clustMixType/index.html. Accessed January 28, 2019.
26. City of Chicago. *Transportation Network Providers – Trips*. https://data.cityofchicago.org/Transportation/Transportation-Network-Providers-Trips/m6dm-c72p. Accessed July 30, 2019
27. Open Weather Map. *Open Weather Map 2019*. https://openweathermap.org/. Accessed July 30, 2019.
28. Google. *Google Distance Matrix API*. https://developers.google.com/maps/. Accessed February 5th, 2019.
29. Chen, Y., M. Hyland, M. P. Wilbur, and H. S. Mahmassani. Characterization of Taxi Fleet Operational Networks and Vehicle Efficiency: Chicago Case Study. *Transportation Research Record: Journal of the Transportation Research Board*, 2018. 2672: 127–138.
30. Madhulatha, T. S. An Overview on Clustering Methods. *IOSR Journal of Engineering*, Vol. 2, No. 4, 2012, pp. 719–725.
31. Lavieri, P. S., and C. R. Bhat. Investigating Objective and Subjective Factors Influencing the Adoption, Frequency, and Characteristics of Ride-Hailing Trips. *Transportation Research Part C: Emerging Technologies*, Vol. 105, 2019, pp. 100–125.
32. Bergal, J. Airport Parking Takes Hit from Uber, Lyft. *Pew*. https://www.pewtrusts.org/en/research-and-analysis/blogs/stateline/2017/07/18/airport-parking-takes-hit-from-uber-lyft. Accessed March 23, 2020.
33. Wadud, Z. An Examination of the Effects of Ride-Hailing Services on Airport Parking Demand. *Journal of Air Transport Management*, Vol. 84, 2020, pp. 101783.
34. Saneinejad, S., M. J. Roorda, and C. Kennedy. Modelling the Impact of Weather Conditions on Active Transportation Travel Behaviour. *Transportation Research Part D: Transport and Environment*, Vol. 17, No. 2, 2012, pp. 129–137.
35. Gebhart, K., and R. B. Noland. The Impact of Weather Conditions on Bikeshare Trips in Washington, DC. *Transportation*, Vol. 41, No. 6, 2014, pp. 1205–1225.
36. Brodeur, A., and K. Nield. An Empirical Analysis of Taxi, Lyft and Uber Rides: Evidence from Weather Shocks in NYC. *Journal of Economic Behavior & Organization*, Vol. 152, 2018, pp. 1–16.
37. Young, M., and S. Farber. Ride-Hailing Platforms are Shaping the Future of Mobility, but for Whom? https://osf.io/pz7fk/. Accessed March 23, 2020.
38. Xue, M., B. Yu, Y. Du, B. Wang, B. Tang, and Y-M. Wei. Possible Emission Reductions from Ride-Sourcing Travel in a Global Megacity: The Case of Beijing. *The Journal of Environment & Development*, Vol. 27, No. 2, 2018, pp. 156–185.
39. Alonso-Mora, J., S. Samaranayake, A. Wallar, E. Frazzoli, and D. Rus. On-Demand High-Capacity Ride-Sharing via Dynamic Trip-Vehicle Assignment. *Proceedings of the National Academy of Sciences*, Vol. 114, No. 3, 2017, pp. 462–467.
40. Clewlow, R. R., and G. S. Mishra. Disruptive Transportation: The Adoption, Utilization, and Impacts of Ride-Hailing in the United States. Research Report UCD-ITS-RR-17-07. UC Davis: National Center for Sustainable Transportation, CA, 2017.
41. Kasliwal, A., N. J. Furbush, J. H. Gawron, J. R. McBride, T. J. Wallington, R. D. De Kleine, H. C. Kim, and G. A. Keoleian. Role of Flying Cars in Sustainable Mobility. *Nature Communications*, Vol. 10, No. 1, 2019, pp. 1–9.
42. Al Haddad, C., E. Chaniotakis, A. Straubinger, K. Plötner, and C. Antoniou. Factors Affecting the Adoption and Use of Urban Air Mobility. *Transportation Research Part A: Policy and Practice*, Vol. 132, 2020, pp. 696–712.
43. Merkert, R., and M. J. Beck. Can a Strategy of Integrated Air-Bus Services Create a Value Proposition for Regional Aviation Management? *Transportation Research Part A: Policy and Practice*, Vol. 132, 2020, pp. 527–539.
44. Shaheen, S., and A. Cohen. Is It Time for a Public Transit Renaissance?: Navigating Travel Behavior, Technology, and Business Model Shifts in a Brave New World. *Journal of Public Transportation*, Vol. 21, No. 1, 2018, pp. 67–81.
45. Gaker, D., Y. Zheng, and J. Walker. Experimental Economics in Transportation: Focus on Social Influences and Provision of Information. *Transportation Research Record: Journal of the Transportation Research Board*, 2010. 2156: 47–55.
46. Zhu, Z., X. Qin, J. Ke, Z. Zheng, and H. Yang. Analysis of Multi-Modal Commute Behavior with Feeding and Competing Ridesplitting Services. *Transportation Research Part A: Policy and Practice*, Vol. 132, 2020, pp. 713–727.