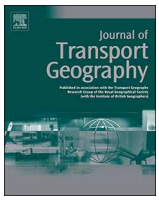
Journal of Transport Geography 83 (2020) 102661

Contents lists available at ScienceDirect

Journal of Transport Geography

journal homepage: www.elsevier.com/locate/jtrangeo



Using machine learning for direct demand modeling of ridesourcing services in Chicago

Xiang Yan[a], Xinyu Liu[b], Xilei Zhao[c,*]

[a] Department of Urban and Regional Planning, University of Florida, Gainesville, FL, USA
[b] H. Milton Stewart School of Industrial and Systems Engineering, Georgia Institute of Technology, Atlanta, GA, USA
[c] Department of Civil and Coastal Engineering, University of Florida, Gainesville, FL, USA

ARTICLE INFO

Keywords:
Ridesourcing
Travel demand
Random forest
Machine learning
Direct demand model

ABSTRACT

The exponential growth of ridesourcing services has been disrupting the transportation sector and changing how people travel. As ridesourcing continues to grow in popularity, being able to accurately predict the demand for it is essential for effective land-use and transportation planning and policymaking. Using recently released trip-level ridesourcing data in Chicago along with a range of variables obtained from publicly available data sources, we applied random forest, a widely-applied machine learning technique, to estimate a zone-to-zone (census tract) direct demand model for ridesourcing services. Compared to the traditional multiplicative models, the random forest model had a better model fit and achieved much higher predictive accuracy. We found that socioeconomic and demographic variables collectively contributed the most (about 50%) to the predictive power of the random forest model. Travel impedance, the built-environment characteristics, and the transit-supply-related variables are also indispensable in ridesourcing demand prediction.

1. Introduction

The emergence and rapid rise of app-based, on-demand ride services provided by transportation network companies (TNCs) such as Uber and Lyft, which is commonly called ridesourcing services, is disrupting the transportation sector and changing how people travel (Henao and Marshall, 2017). By introducing a new form of mobility featured with many advantages such as on-demand, low cost (compared to traditional taxi services), and reliability, major TNCs have quickly attracted millions of users and generated billions of trips. For example, Uber reported that as of December 2018, it had 91 million monthly active users and 14 million trips completed worldwide each day (Uber Newsroom, 2019). As TNCs become increasingly popular, they affect cities both in positive and negative ways. For example, while many cherish the convenience and economic benefits brought by TNCs to both riders and drivers (Chen et al., 2017; Wallsten, 2015), others are concerned that TNCs may worsen traffic congestion (Erhardt et al., 2019) and threaten the already struggling transit industry (Clewlow and Mishra, 2017). Recent evidence suggests that ridesourcing can act both as a substitute and supplement to public transit (Hall et al., 2018; Young et al., 2020). To maximize the potential benefits and to mitigate the harms brought by this new mobility option, it is important for policymakers to understand and forecast the demand for ridesourcing services so that they can plan accordingly and develop regulatory measures where necessary.

The conventional approach for travel-demand modeling is a "four-step" process, which includes trip generation, trip distribution, mode share, and traffic assignment (McNally, 2007). An alternative approach to this sequential approach is to develop a model subsuming trip generation, distribution, and mode choice, which is called a *direct demand model* (Lave, 1972; Talvitie, 1973). The direct demand model is attractive not only because it handles several submodels simultaneously, but also because it may overcome some drawbacks of the four-step approach; for example, it avoids the problem of accumulating step-by-step errors associated with the sequential four-step approach (Choi et al., 2012). However, the direct demand model has received limited application in the intraurban context (de Dios Ortuzar and Willumsen, 2011), possibly due to two reasons. First, recent advances in travel-demand prediction have focused on studying individual travel behavior (e.g., travel-behavior modeling informed by the microeconomic random-utility theory), and yet direct demand models are inflexible to accommodate these behavioral insights. Second, assembling a large amount of spatially accurate origin-destination (O-D) trip-matrix data is challenging. However, as the O-D transit trip matrix data (i.e., smart-card data) becomes available in recent years, the direct demand model has been applied to model station-to-station transit-use demand (Choi

* Corresponding author.
E-mail address: xilei.zhao@essie.ufl.edu (X. Zhao).

https://doi.org/10.1016/j.jtrangeo.2020.102661
Received 6 January 2020; Received in revised form 24 January 2020; Accepted 3 February 2020
Available online 29 February 2020
0966-6923/

et al., 2012; Ding et al., 2019; Iseki et al., 2018). The recent public release of TNC/ridesourcing trip records in the City of Chicago promises another application of the direct demand model.

The increasing popularity of machine learning may provide another boost to the direct demand model. Some recent studies have found that, by allowing a more flexible model structure to capture the complex interrelationships among variables (e.g., nonlinear relationships and interaction effects), machine learning algorithms often perform better than logit models in predicting travel mode choice (Cheng et al., 2019; Wang et al., 2019; Xie et al., 2003). Machine learning has also been shown effective in forecasting transit ridership, achieving better predictive performance compared to conventional statistical approaches (Ma et al., 2018). These findings suggest that if powered by machine learning algorithms, direct demand models may achieve higher predictive accuracy than traditional methods such as the multiplicative model (de Dios Ortuzar and Willumsen, 2011). To our knowledge, however, few previous work has applied machine learning algorithms to direct demand models.

In light of the above discussion, this study applies random forest, a widely-applied machine learning algorithm, to model and predict the O-D ridesourcing demand in the City of Chicago. Our model is based on a recently released dataset of ridesourcing-trip records, supplemented by a variety of other open-source datasets such as the census data, the General Transit Feed Specification (GTFS) data, and some geospatial datasets available on the Chicago Data Portal. We further compare the random forest model with a traditional direct demand model that takes a multiplicative function form in terms of their predictive power and goodness-of-fit. The unique contributions of this study are summarized as follows:

- This paper is one of the first studies to model and predict demand for ridesourcing services in a large US city. We extend the current literature on this topic by further examining how trip-cost-related variables (e.g., ridesourcing trip fare and travel time) shape ridesourcing demand.
- This paper presents a machine learning application to model zone-to-zone ridesourcing demand. The results show that the random forest model has significantly higher predictive accuracy and better model fit compared to the traditional multiplicative model.

The rest of the paper is organized as follows. In Section 2, we review the previous studies on ridesourcing demand modeling and machine learning applications in travel demand modeling. Section 3 describes the fundamentals of the direct demand model, its classic multiplicative form, and the random forest algorithm. It also describes the performance metrics used for model comparison. In Section 4, we introduce the Chicago ridesourcing data and the independent variables examined. Section 5 presents a descriptive analysis of ridesourcing trips and then compare the performance of four direct demand models (i.e., a benchmark model, two multiplicative models, and a random forest model). Lastly, in Section 6, we conclude the paper by summarizing findings, identifying limitations, and suggesting future work.

2. Literature review

2.1. Studies on ridesourcing demand modeling

Studies on demand prediction of ridesourcing services are limited in supply due to the short history of ridesourcing and the lack of publicly available data. Gerte et al. (2018), Lavieri et al. (2018) and Yu and Peng (2019) are among the first studies that applied ridesourcing-trip data to estimate demand models. The Gerte et al. (2018) study was based on the Uber trip data in the City of New York and the other two studies were based on the RideAustin (a non-profit TNC company in Austin, Texas) data. A main difference between the two datasets and the the Chicago ridesourcing-trip data used in this study is that the Chicago data recorded trip-cost-related information (e.g., ridesourcing trip fare and travel time) but the other two datasets did not.

Gerte et al. (2018) estimated a panel-based random effects model to estimate Uber-trip generation at the taxi zone level. Results showed that the following variables were statistically significant: time variables (e.g., season dummies), weather-related variables, built-environment characteristics, and demographic factors. To predict zone-to-zone ridesourcing trips, Lavieri et al. (2018) built two separate models: a spatially lagged multivariate count model for trip generation, and a fractional split model for trip distribution. Their list of independent variables included socioeconomic and demographic characteristics, built-environment variables, and transit-supply variables. They found that zones with higher proportions of white population had a weaker demand for ridesourcing services but zones with higher median household income were associated with more weekday ridesourcing trips (Lavieri et al., 2018). Yu and Peng (2019) explored a wider range of built-environment variables and fitted a trip-generation model using geographically weighted Poisson regression. A major finding was that higher population density is associated with greater ridesourcing demand (Yu and Peng, 2019). These studies informed the selection of independent variables examined in this study, and we further included some trip-cost-related variables (e.g., ridesourcing trip fare and travel time) that were unaccounted for in previous studies.

Apart from studies that model ridesourcing demand at an aggregate (zonal) level, some researchers have examined individual preference for ridesourcing services using disaggregate data collected by travel surveys. By directly modeling individual travel behavior, these studies provide microeconomic and behavioral insights on ridesourcing demand. For example, Rayle et al. (2016) conducted a survey-based comparison of taxis, transit, and ridesourcing services in San Francisco and found that at least half of ridesourcing trips replaced modes other than taxi, including driving and public transit. Dias et al. (2017) estimated a bi-variate ordered Probit model using a survey data set derived from the 2014–2015 Puget Sound Regional Travel Study and showed that the users of car-sharing and ridesourcing services tend to be young, well-educated, higher-income, working individuals residing in higher-density areas. Alemi et al. (2019) estimated an ordered Probit model to evaluate the frequency of use of Uber/Lyft in California, and found that frequent long-distance travelers use ridesourcing more often. Lavieri and Bhat (2019) investigated objective and subjective factors that influence the adoption, frequency, and characteristics of ridesourcing trips using the survey data from the Dallas-Fort Worth Metropolitan Area and found that low residential location density and people's privacy concerns are the main deterrents to pooled ridesourcing adoption.

2.2. Direct demand modeling

The direct demand models, first proposed by Kraft and Wohl (1967), integrate trip generation, trip distribution, and mode split into a single equation. As Talvitie (1973) argued, the direct demand models can address some of the limitations associated with the four-step models: a) trip generation can be associated with accessibility measures between different zone-to-zone pairs; b) calibration of the direct demand model is accomplished by the zonal interchanges; and c) trip generation and mode split are determined in the same equation, meaning variables can affect mode choice and trip generation at the same time. Moreover, the direct demand models were believed to overcome the problem of step-by-step error accumulation associated with the sequential four-step approach (Choi et al., 2012).

However, the applications of direct demand models were limited in the last century. This is probably due to the difficulties encountered in calibrating such models, mainly the lack of O-D trip data (Dagenais et al., 1986). With the technological advancement in intelligent transportation systems, a large amount of spatially accurate O-D trip data for public transit systems have become available. Hence, in recent years,

there is a growing interest in applying direct demand models to forecast transit demand (e.g. Cervero et al., 2010; Choi et al., 2012; Kepaptsoglou et al., 2017; Zhao et al., 2014). For example, Cervero et al. (2010) constructed a direct demand model for predicting bus rapid transit in Los Angeles County, California. Zhao et al. (2014) adopted direct demand models to reveal associations between metro demand and the influencing factors in a case study of Nanjing, China. Kepaptsoglou et al. (2017) used a direct demand model approach to estimating the ridership of a new light rail transit in Cyprus.

Recently, some U.S. cities, e.g., Austin, Texas and Chicago, Illinois, have partnered with TNCs to publish the ridesourcing trip data, which promises another important application of direct demand models. However, to the best of our knowledge, there is no existing work that applies the direct demand modeling approach to predict ridesourcing demand.

2.3. Machine learning applications in travel demand modeling

Machine learning has gained increasing popularity in the field of travel demand modeling in recent years. For example, a number of recent studies have shown that machine learning can significantly improve the accuracy of individual-level travel mode choice predictions compared to the traditional logit models, e.g., Cheng et al. (2019); Hagenauer and Helbich (2017); Xie et al. (2003). More specifically, Xie et al. (2003) applied decision trees and artificial neural networks to model travel mode choice and showed improved predictive accuracy over the multinomial logit model. Hagenauer and Helbich (2017) compared multiple machine learning algorithms for modeling travel mode choice and found that the random forest model outperformed all the other models in terms of prediction. More recently, Cheng et al. (2019) used random forest to model travel mode choice and achieved high prediction accuracy, fast computation speed, and good interpretability. In addition, some researchers have tried to leverage machine learning to forecast the travel demand for different travel modes (the outcome variable is trip frequency rather than mode choice), such as public transit (Ma et al., 2018), ridesourcing (Geng et al., 2019), and bikesharing (Lin et al., 2018).

These studies have demonstrated the superior predictive capabilities of machine learning and suggested the potential of applying machine learning to facilitate travel demand modeling. Machine learning models such as random forest and deep neural networks have a more flexible modeling structure than conventional statistical methods, and they are capable of modeling nonlinear relationships between the independent and dependent variables and capturing complex interactions among the independent variables (Lhéritier et al., 2018). However, few existing work has applied machine learning algorithms in direct demand models. One exception is the Baek and Sohn (2016) study, which developed a artificial neural networks model to forecast bus ridership at the stop-to-stop level. Baek and Sohn (2016) verified that, with a sufficient sample size, the stop-to-stop direct demand model achieved good predictive performance. This study extends this line of work to apply machine learning to model travel demand for ridesourcing.

3. Methods

3.1. The direct demand model

Adapted from Choi et al. (2012); Talvitie (1973); Zhao et al. (2014), a direct demand model of ridesourcing trips can be defined to have the following functional form:

$$V_{ij} = F(L_{ij}, SED_i, SED_j, BE_i, BE_j, TS_i, TS_j),$$

where V_{ij} is the number of ridesourcing trips from i to j, L_{ij} is the set of travel-impedance variables (e.g., trip distance, ridesourcing trip cost, and ridesourcing travel time) from i to j, SED_i and SED_j are the sets of socioeconomic and demographic variables for origin zone i and destination zone j, respectively, BE_i and BE_j are the sets of built-environment variables for origin zone i and destination zone j, respectively, and TS_i and TS_j are the sets of transit-supply variables for origin zone i and destination zone j, respectively. All of these variables are commonly used in travel demand modeling. In the next subsection, we will introduce two different approaches, i.e., the traditional multiplicative model and the random forest model to forecast the demand for ridesourcing services.

3.2. Modeling approaches

3.2.1. The multiplicative model

The classic direct demand models took a multiplicative form (de Dios Ortuzar and Willumsen, 2011; Talvitie, 1973). Compared to other model classes, such as the Poisson regression model, the multiplicative model is easier to interpret and to draw inferences (Choi et al., 2012). The multiplicative model is defined as follows:

$$V_{ij} = \phi \Pi_{p=1}^{P} X_{ip}^{\alpha_p} \Pi_{p=1}^{P} X_{jp}^{\beta_p} \Pi_{q=1}^{Q} Z_{ijq}^{\gamma_q}, \quad (1)$$

where $X_{ip} \in SED_i \cup BE_i \cup TS_i$ is the p^{th} feature regarding origin zone i, $X_{jp} \in SED_j \cup BE_j \cup TS_j$ is the p^{th} feature regarding destination zone j, $Z_{ijq} \in L_{ij}$ is the q^{th} feature regarding the travel impedance from i to j, ϕ is the scale parameter, α_p, β_p, γ_q are the parameters to be estimated, P is the total number of variables that include socioeconomic and demographic, built-environment, and transit-supply variables, and Q is the total number of travel-impedance variables.

Then, by taking a natural logarithm on the both sides of Eq. (1), we can transform it into a linear form:

$$\ln(V_{ij}) = \ln(\phi) + \sum_{p=1}^{P} \alpha_p \ln(X_{ip}) + \sum_{p=1}^{P} \beta_p \ln(X_{jp}) + \sum_{q=1}^{Q} \gamma_q \ln(Z_{ijq}). \quad (2)$$

In this study, we use the log-transformed data to conduct multiple linear regression and then transform the predicted response variable, $\ln(\widehat{V_{ij}})$, back to its original domain, $\widehat{V_{ij}}$, in order to evaluate the model's predictive accuracy.

3.2.2. The random forest model

Random forest (RF) is among the most accurate general-purpose classifiers so far with the capability of handling high dimensional data (Biau, 2012). RF is also sufficiently robust: the input variables for RF can be of any type (numerical, categorical, continuous, or discrete) and RF is insensitive to skewed distributions, outliers, missing values, and the inclusion of irrelevant variables (Breiman, 2001). In addition, RF requires fairly minor efforts in tuning hyperparameters (i.e., two major hyperparameters) and is usually not very sensitive their values (Caruana et al., 2008; Liaw et al., 2002). It also needs relatively short training time (Liaw et al., 2002). More importantly, RF is able to model complex nonlinear relationships between the input variables and the response variable and capture high-order interactions among variables due to its flexible modeling structure (Breiman, 2001).

In recent years, RF has shown to be effective in various fields. To name a few use cases, the RF has been successfully applied to identify crash severity (Wang et al., 2019), predict construction injury (Tixier et al., 2016), forecast Alzheimer's disease (Lebedev et al., 2014), detect credit card fraud (Bhattacharyya et al., 2011), and classify earthquake building damage (Mangalathu et al., 2019). Random forest has also been applied to model people's travel behavior (e.g. Cheng et al., 2019; Hagenauer and Helbich, 2017; Lhéritier et al., 2018; Zhao et al., 2020). Recently, Cheng et al. (2019) provided a detailed summary of the recent studies applying random forest method in the field of transportation. Given its various strengths and widespread application, we choose to use RF to model the demand for ridesourcing services in Chicago.

Random forest is essentially a tree-based ensemble method: it trains multiple decision trees and combines the predictions of all the decision trees to generate a final prediction (Breiman, 2001). A decision tree model

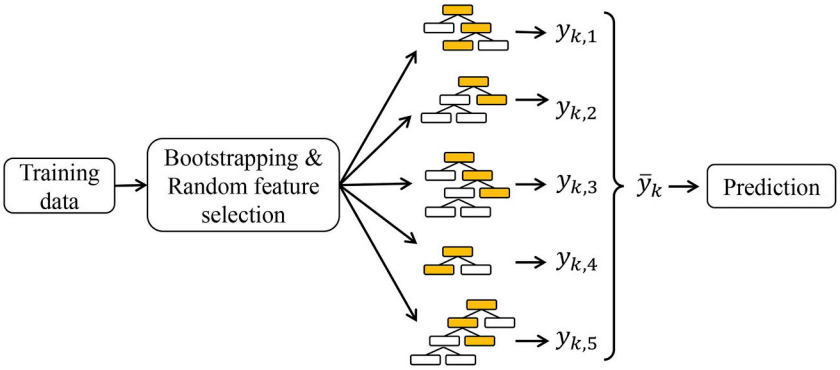


Fig. 1. Illustration of random forest model with soft prediction mechanism.

builds classification or regression trees to predict either a categorical or a continuous outcome variable. In this paper, the decision tree model builds regression trees where each internal node of the tree recursively partitions the data based on the value of a single feature and the terminal nodes of the tree represent the predicted value for the response variable, i.e., V_{ij}. The decision tree model is sensitive to noise and susceptible to overfit (Last et al., 2002; Quinlan, 2014). To address the overfitting issue, the tree-based ensemble techniques were proposed to form more robust, stable, and accurate models, and random forest is one of the most widely used tree-based ensemble methods (Breiman, 1996; Hastie et al., 2001). Random forest trains multiple decision trees in parallel by bootstrapping the training data, i.e., sampling with replacement (Breiman, 2001). It then selects a random subset of all the features to train the decision trees. More precisely, the trees in random forest use all the variables, but for each tree, it only chooses a random subset of variables for splitting. By doing so, random forest can overcome the overfitting problems of a single decision tree, and reduce variance between correlated trees. For regression problems (i.e., the response variable is continuous rather than categorical), random forest makes predictions by averaging over the predictions of all decision trees (see Fig. 1). More formally, the predicted number of ride-sourcing trips from i to j can be indicated as

$$f(\boldsymbol{X}_i, \boldsymbol{X}_j, \boldsymbol{Z}_{ij} \mid \hat{\theta}) = \frac{1}{B}\sum_{b=1}^{B} \hat{T}_b, \tag{3}$$

where $\boldsymbol{X}_i = [X_{i1}, X_{i2}, \ldots, X_{iP}]$, $\boldsymbol{X}_j = [X_{j1}, X_{j2}, \ldots, X_{jP}]$, $\boldsymbol{Z}_{ij} = [Z_{ij1}, Z_{ij2}, \ldots, Z_{ijQ}]$, $\hat{\theta}$ is the estimated hyperparameter vector for the random forest model, B is the total number of decision trees in random forest (one of the estimated hyperparameters), and $\hat{T}_b$ is the predicted value for decision tree b, $b = 1, \ldots, B$. In this paper, the random forest model is built using the Python library *scikit-learn* (Pedregosa et al., 2011). After applying grid search to tune the hyperparameters for the random forest model, we choose the candidate random forest model with 1000 decision trees and 58 variables considered at each split. For each decision tree, we choose 2 as the minimum number of samples for each leaf.

3.3. Model comparison: Cross validation and performance metrics

In this study, we apply 10-fold cross validation to compare and evaluate the predictive capability of different models. To conduct the 10-fold cross validation, the dataset is first randomly split into 10 disjoint subsets. Holding out one subset at a time, we use the remaining nine subsets to train the selected models. The trained models are then used to make predictions for the holdout set, which are compared with the observed (true) values to evaluate the models' out-of-sample predictions. After this process is repeated for each one of the 10 subsets, the validation results for each model are averaged to compute a mean estimate of the predictive performance metrics.

There are many types of performance metrics to measure the predictive capability of a model, and in this study, we choose two commonly-used metrics, including mean absolute error (MAE) and root mean square error (RMSE). To be specific, the MAE and RMSE are respectively defined as

$$MAE = \frac{1}{N}\sum_{k=1}^{N} |\hat{y}_k - y_k|, \tag{4}$$

$$RMSE = \sqrt{\frac{\sum_{k=1}^{N} (\hat{y}_k - y_k)^2}{N}}, \tag{5}$$

where N is the total of number of observations in the training set, y_k is the k^{th} observed value for the response variable, and $\hat{y}_k$ is the k^{th} predicted value for the response variable. The MAE has a clear interpretation as it quantifies the average absolute difference between the observed (true) and predicted values and each error contributes to MAE in proportion to the absolute value of the error (Pontius et al., 2008). On the other hand, the RMSE represents the square root of the second sample moment of differences between the predicted and observed values. The impact of each error on RMSE is proportional to the size of the squared error, so larger errors will have a disproportionately large impact on RMSE, making it more sensitive to outliers (Pontius et al., 2008).

We compute the out-of-sample MAE and RMSE to evaluate the predictive performance of each model, and we calculate the in-sample MAE and RMSE to assess their goodness-of-fit.

4. The data

Data on the Chicago ridesourcing trips came from the publicly accessible Chicago Data Portal.[1] The City of Chicago ordinance required TNCs to report all ridesourcing trips (starting November 2018) took place within the city boundary on a quarterly basis. This study collected data released until March 31, 2019, which includes a total of 45,338,599 trips. Each trip record contained a variety of attributes, among which the following were used in this study: trip fare,[2] distance (in miles), duration (in seconds), pickup (trip-origin) location, and drop-off (trip-destination) location. To protect privacy, the city of Chicago applied masking to several data fields. Namely, fares were rounded to the nearest $2.50, and pickup and drop-off locations were aggregated at the level of census tract.

To prepare the data for demand modeling, we processed the ride-sourcing-trip data with the following procedure. First, since a

[1] The data can be downloaded from the following link: https://data.cityofchicago.org/Transportation/Transportation-Network-Providers-Trips/m6dm-c72p.

[2] The Chicago ridesourcing-trip dataset contains four trip-cost-related variables, including fare, tips, additional charges (taxes, tolls, and any other charges such as wait time fee), and total trip costs. This study analyzes trip fare only because tips and additional charges are often individual-specific variable costs, which may not be important predictors of travel demand.

significant proportion of trips recorded the geographic coordinates of the pickup/drop-off census tract's centroid but not the census tract ID, we recovered these IDs by applying the Python "census geocode" package. We found that a disproportionately large number of trips with such characteristics were authorized rideshare trips, and so without this step authorized rideshare trips would be underestimated. Second, we summarized the ridesourcing-trip data from a disaggregate level (i.e., individual trip records) into an aggregate level (i.e., origin-destination O-D pairs). We calculated the following new attributes: total number of trips, median trip fare, median trip distance, median trip duration, standard deviation of trip fare, standard deviation of trip distance, and standard deviation of trip duration. Before the calculation, we removed some trip records that we considered as outliers.[3] The "median" variables meant to measure the characteristics of a UberX or regular Lyft trip (the most popular ridesourcing trip type), and the "standard deviation" variables meant to capture the variations in rider characteristics (riders who select different types of ridesourcing service, i.e., UberPool versus UberX), price fluctuations (e.g., surge pricing), traffic condition, route choice, etc. Finally, to minimize the impact of randomness on demand prediction, we excluded some O-D pairs with less than 50 trips over the period of November 2018 and March 2019. After these data processing steps, the total number of O-D pairs to be analyzed was 62,943, which included 713 census tracts as trip origins and 705 as trip destinations (the total number of census tracts in the City of Chicago is 801).

TotalRidesourcingTrips (i.e., the total number of ridesourcing trips) is the outcome variable modeled in this paper. Together with a variable measuring the driving distance of an O-D pair (data obtained using the GraphHopper Directions API), the ridesourcing-trip-related variables were used to represent the travel-impedance characteristics between an origin census tract and a destination census tract. We further supplemented these data with a list of socioeconomic and demographic variables obtained from the American Community Survey 2013–2017 5-year estimates data, some employment and worker characteristics obtained from the 2015 Longitudinal Employer-Household Dynamics data, and the crime rate data obtained from Chicago Data Portal. Furthermore, we used GTFS data to estimate some transit-supply-related variables, applied geographic information system (GIS) techniques to calculate several built environment variables, and used the Walkscore.com API to obtain the Walk Score of a census tract's centroid. All these variables were calculated at the census tract level. Tables 1 and 2 present the variable description and the descriptive statistics, respectively, of the variables examined in this study. All of the independent variables examined here have a variance inflation factor (VIF) value less than 10, which suggests that multicollinearity should be of little concern in our models (Sheather, 2009).[4]

Table 1
Variable description.

Variable	Description
TotalRidesourcingTrips[a]	Total number of ridesourcing trips
Travel-impedance variables	
Distance_median	Median trip distance (mile)
Time_median	Median trip duration (second)
Fare_sd	The standard deviation of trip fare ($)
Distance_sd	The standard deviation of trip distance (mile)
Time_sd	The standard deviation of trip duration (second)
Socioeconomic and demographic variables	
PctMale[b]	Percentage of male population
PctYoung	Percentage of population aged 18–44
PctWhite	Percentage of White population
PctHisp	Percentage of Hispanic population
PctAsian	Percentage of Asian population
PctCar	Percentage of households with at least one car
PctTransit	Percentage of workers taking transit to work
PctLowInc	Percentage of low-income households ($25 k less)
PctModInc	Percentage of moderate-income households ($25-$50 k)
PctMidInc	Percentage of middle-income households ($50 k to $75 k)
PctRentUnit	Percentage of renter-occupied housing units
PctSinFam	Percentage of single-family homes
PctWacWorker54[c]	Percentage of workers (workplace) workers aged 54 or younger
PctWacLowMidWage[c]	Percentage of workers (workplace) with earnings $3333/month or less
PctWacBachelor[c]	Percentage of workers (workplace) with bachelor's degree and above
CrimeDen	Density of violent crime
Commuters	Total number of commuters (from origin census tract to destination census tract)
Built-environment variables	
PopDen	Population density
EmpDen	Employment density
EmpRetail	Retail employment density
RdNetwkDen	Percentage of commuters with bachelor's degree and above
Walkscore	Walkscore of centroid of census tract
InterstDen	Intersection density
Transit-supply variables	
SerHourBusRoutes	Aggregate service hours for bus routes
SerHourRailRoutes	Aggregate service hours for rail routes
BusStopDen	Number of bus stops per square mile
RailStopDen	Number of rail stops per square mile
PctBusBuf	Percentage of tract within 1/4 mile of a bus stop
PctRailBuf	Percentage of tract within 1/4 mile of a rail stop

[a] This is the dependent variable.
[b] The variables listed before PctMale are at the zone-to-zone level, and the remaining variables (including PctMale) are at the zonal (census tract) level.
[c] Variable codes with "Wac" refers to variables extracted from workplace area characteristics data files of the Longitudinal Employer-Household Dynamics data. These files describe the characteristics of workers aggregated at the area of their workplace.

5. Results

5.1. Exploratory analysis of ridesourcing data in Chicago

The three maps shown in Fig. 2 show the spatial distributions of trip generation (origin), trip attraction (destination), and travel flows (O-D pairs) of ridesourcing services in Chicago over the five-month period. From Fig. 2(a) and (b), we infer two important observations. First, there is no obvious imbalance (i.e., places with much more pickups than drop-offs or the other way around) in the spatial distribution of trip origins and destinations. Second, a majority of the trips took place at areas surrounding the downtown and the "North Side",[5] and trips

[3] We first removed observations with trip fare equal to 0, trip duration less than 1 min, or trip distance less than 0.25 miles. Moreover, for trips sharing an O-D pair, one would expect their trip distance and duration to be reasonably close. We thus removed trips that are identified as outliers, which were defined as trips whose distance or duration was more than 6 interquartile ranges away from either the upper or the lower quartile. For each data point x_{ij} denoting the j^{th} feature of the i^{th} observation, we use the notation X_j for the j^{th} feature vector, $Q1(\cdot)$ for the lower quartile (25% quantile) and $Q3(\cdot)$ for the upper quartile (75% quantile), then $IQR(\cdot) \coloneqq Q3(\cdot) - Q1(\cdot)$ is the interquartile range. The data point x_{ij} is considered an outlier in the vector X_j if $x_{ij} > Q3(X_j) + 6 \times IQR(X_j)$ or $x_{ij} < Q1(X_j) - 6 \times IQR(X_j)$. This outlier analysis was not performed on trip fare because this variable had zero interquartile range on some O-D pairs due to rounding. In the end, we removed about 3.6% of the trip records.

[4] We initially examined more variables that are potentially related to ridesourcing demand, but to reduce model complexity we excluded some of them (e.g., household size and job accessibility) based on preliminary model results. We also excluded variables that had a VIF score greater than 10, which includes median fare, median household income, percentage of black population, and percentage of adults with a bachelor degree.

[5] See a map and description of Chicago community areas at https://en.wikipedia.org/wiki/Community_areas_in_Chicago.

Table 2
Descriptive statistics of the variables used for modeling.

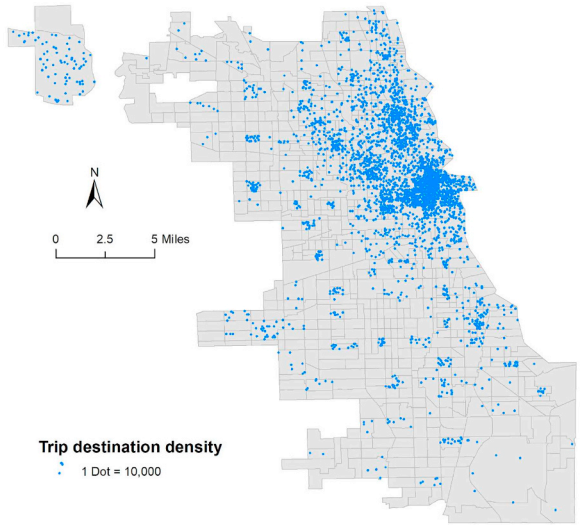
		Mean	SD	Min	Max
TotalRidesourcingTrips		570.889	2,111.720	50	77,661
Distance_median		5.082	3.887	0.356	38.153
Time_median		988.192	498.412	145	4,321
Fare_sd		3.019	1.460	0.461	20.198
Distance_sd		0.694	0.466	0.019	8.750
Time_sd		292.097	157.660	48.873	1,236.964
PctMale	Origin	0.487	0.074	0	0.884
	Destination	0.487	0.075	0	0.884
PctYoung	Origin	0.528	0.148	0	0.897
	Destination	0.529	0.149	0	0.897
PctWhite	Origin	0.622	0.280	0	0.973
	Destination	0.626	0.276	0	0.973
PctHisp	Origin	0.195	0.217	0	0.996
	Destination	0.191	0.212	0	0.996
PctAsian	Origin	0.086	0.099	0	0.898
	Destination	0.087	0.098	0	0.898
PctCar	Origin	0.700	0.162	0	0.995
	Destination	0.699	0.164	0	0.995
PctTransit	Origin	0.343	0.130	0	0.706
	Destination	0.343	0.131	0	0.706
PctLowInc	Origin	0.220	0.143	0	0.840
	Destination	0.218	0.142	0	0.840
PctModInc	Origin	0.169	0.085	0	0.481
	Destination	0.167	0.085	0	0.476
PctMidInc	Origin	0.259	0.076	0	0.556
	Destination	0.259	0.076	0	0.556
PctRentUnit	Origin	0.593	0.163	0	1
	Destination	0.592	0.164	0	1
PctSinFam	Origin	0.184	0.167	0	1
	Destination	0.182	0.167	0	1
PctWacAge54	Origin	0.796	0.067	0	1
	Destination	0.796	0.066	0	1
PctWacLowMidWage	Origin	0.673	0.164	0	1
	Destination	0.668	0.166	0	1
PctWacBachelor	Origin	0.208	0.064	0	0.474
	Destination	0.209	0.065	0	0.500
CrimeDen	Origin	163.943	184.639	0	1,560.648
	Destination	161.982	183.698	0	1,560.648
Commuters	Work to Home	6.551	33.513	0	2,213
	Home to Work	6.725	33.456	0	2,213
Popden	Origin	24,192.546	16,386.369	0	306,466.800
	Destination	24,165.331	16,618.945	0	306,466.800
EmpDen	Origin	49,465.901	479,572.778	0	96,915,379.783
	Destination	51,052.207	479,378.450	0	96,915,379.783
EmpRetail	Origin	1,914.562	5,783.325	0	55,862.618
	Destination	1,874.870	5,703.093	0	55,862.618
RdNetwkDen	Origin	26.446	8.997	0	64.169
	Destination	26.520	9.162	0	64.169
Walkscore	Origin	81.732	17.232	0	100.000
	Destination	81.601	17.460	0	100.000
InterstDen	Origin	135.643	111.832	0	743.911
	Destination	136.988	113.667	0	743.911
SerHourBusRoutes	Origin	1,311.829	870.711	0	4,852.068
	Destination	1,326.514	885.268	0	4,852.068
SerHourRailRoutes	Origin	416.461	439.134	0	1,681.712
	Destination	424.616	443.161	0	1,681.712
BusStopDen	Origin	66.916	36.458	0	216.449
	Destination	66.909	36.715	0	216.449
RailStopDen	Origin	1.829	3.669	0	23.336
	Destination	1.854	3.709	0	23.336
PctBusBuf	Origin	0.933	0.162	0	0.998
	Destination	0.931	0.164	0	0.998
PctRailBuf	Origin	0.269	0.325	0	0.998
	Destination	0.271	0.325	0	0.998

happening in other areas were also highly clustered. Fig. 2(c) further shows that ridesourcing travel flows concentrated at the central and north part of Chicago and that the quantity of airport rides was large.

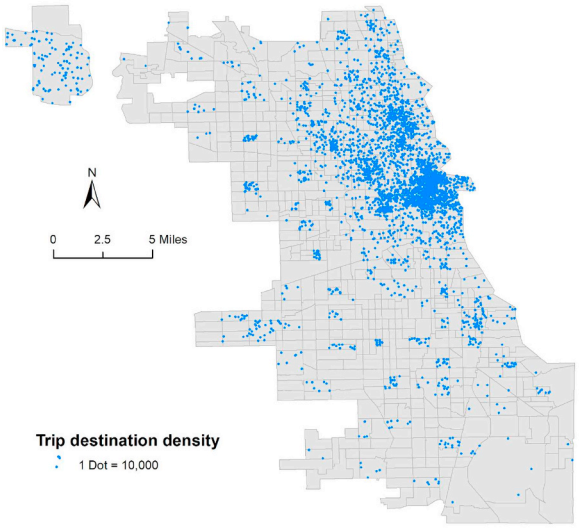
Fig. 3 presents the histograms of trip fare, distance and duration. Unsurprisingly, the trip fare and trip duration distributions had a log-normal-like shape, and the plot of trip distance was close to an exponential shape. In addition, these figures showed that that most trips had a trip fare between $5 and $15, a trip distance less than 7.5 miles, and a trip duration less than 15,000 s (25 min).

5.2. Direct demand model comparison

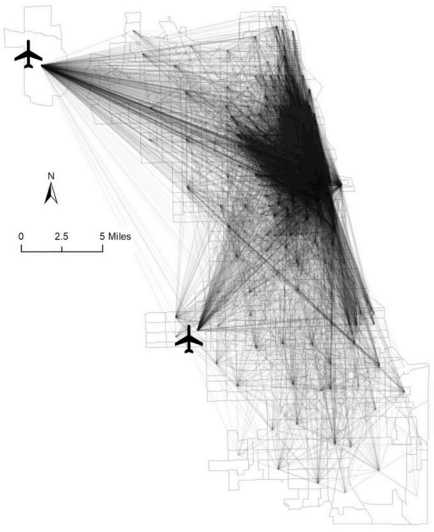
We implemented four different direct demand models in order to have a comprehensive comparison. To be specific, we built the



(a) Trip generation

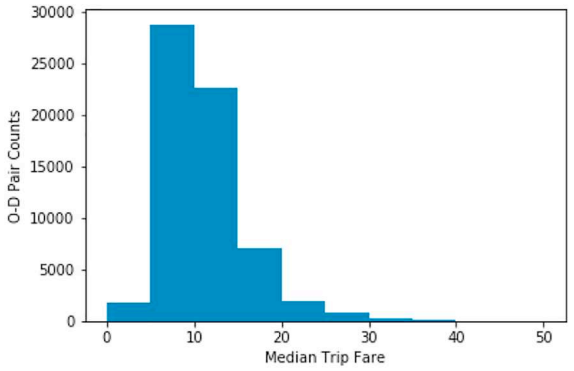


(b) Trip attraction

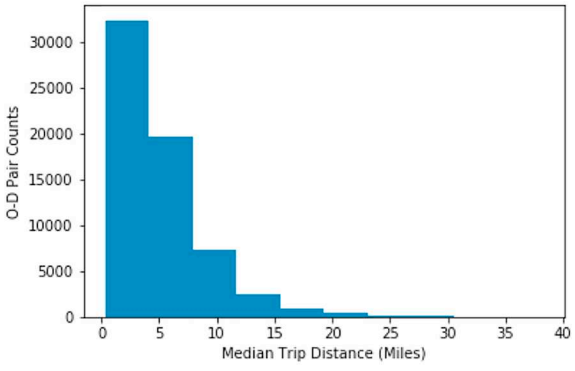


(c) Ridesourcing travel flows

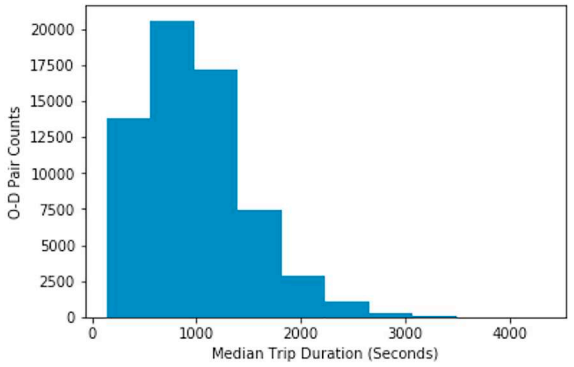
Fig. 2. Spatial distribution of ridesourcing trips.
Note: 1. Map (a) and (b) are point density maps, which means that the dots in them are symbolic and do not indicate actual pickup/drop-off locations. 2. For Map (c), we have only shown O-D lines with a minimal flow of 300 ridesourcing trips.



(a) Trip fare



(b) Trip distance



(c) Trip duration

Fig. 3. Histograms of key trip characteristics.

multiplicative model 1 using all explanatory variables and the multiplicative model 2[6] with all statistically significant variables with a *p*-value that is no greater than 0.05 in multiplicative model 1. In addition, we trained a random forest model with all the variables included and the hyperparameters set to the tuned values as described in the method section. Lastly, a mean model was built as a baseline for comparison, which used the sample mean of the training data to forecast the future values in the testing set. The four models, including the mean model, multiplicative model 1, multiplicative model 2, and the random forest model, were compared using 10-fold cross validation. The results of several performance metrics, including in-sample and out-of-sample MAEs, and in-sample and out-of-sample RMSEs, for the four models were visualized using boxplot, as shown in Fig. 4.

Fig. 4 clearly shows that the random forest model had much lower

[6] The following variables were removed for multiplicative model 2: popden_Ori, pctlowinc_Ori, pctsinfam2_Ori, popden_Des, pctlowinc_Des, Walkscore_Des, and InterstDen_Des.

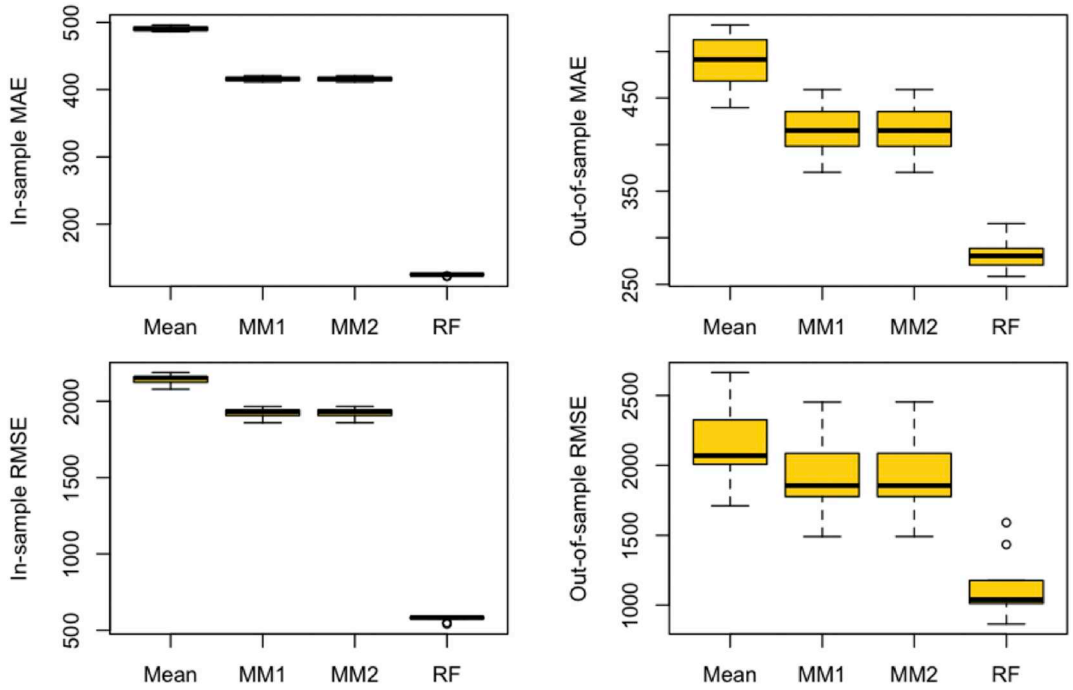


Fig. 4. Boxplots of model performance metrics: Mean–mean model; MM1–multiplicative model 1; MM2–multiplicative model 2; RF–random forest.

Table 3
Variable importance by category.

Variable category	Sum of importance	Count	Average importance
Travel impedance	15.30%	5	3.06%
Socioeconomic and demographic	49.87%	40	1.47%
Built environment	17.48%	6	1.46%
Transit supply	17.35%	12	1.45%
Total:	1	63	

prediction error than the other three models. Both the in-sample and out-of-sample errors were much smaller for the random forest model than for the multiplicative models. Notably, the values of in-sample MAE and RMSE for the random forest model were only about one fourth of those for the multiplicative models. The higher predictability of random forest is probably due to its flexible modeling structure, which allows it to capture the nonlinear relationships and variable interactions and to cope with variable correlations (Lhéritier et al., 2018; Molnar, 2019). Furthermore, random forest can effectively reduce the negative influences of outliers on model performance by binning them. Inside a random forest, each decision tree tends to grow very deep, so the trees have low bias and high variance. By averaging all the uncorrelated decision trees, a random forecast model maximizes the reduction in variance, which consequently leads to relatively low bias and low variance (both of which contribute to the overall prediction error). By contrast, the multiplicative model assumes a log-linear relationship between the independent variables and the outcome variable and predetermines an error distribution that assumes independence among independent variables. These assumptions reduces the model's flexibility to adapt to the particularities of a particular dataset and consequently lowers its predictive performance.

5.3. Interpreting the random forest model

To shed light on how the random forest model makes its prediction, we computed the variable importance for the random forest model. Variable importance measures how much each feature contributes to the prediction of the outcome variable. The more a model depends on a variable to make predictions, the more important it is. For the random forest model, we generated the variable importance of a particular variable by computing the mean decrease in node impurities (measured by variance) from splitting on this variable. Given the large number of variables examined in this study, we decided not to present the results for all variables. Instead, we presented the variable importance aggregated by each variable category (Table 3) and results for the three most important variables in each category (Table 4).

As shown in Table 3, among the four variable categories, the socioeconomic and demographic variables contributed the most to the prediction of ridesourcing demand. These variables explained almost half of the model's predictive power. The sums of variable importance for the other three variable categories were quite close (15.30%, 17.48%, and 17.35%). These results are consistent with decades of travel-behavior research which suggests that the socioeconomic and demographic factors have the largest influence on travel behavior but the built-environment characteristics and trip prices also matter (Boarnet et al., 2001; Ewing and Cervero, 2001, 2010).[7] We further averaged the variable importance within each variable category to assess the contribution of individual variables to prediction. The results showed that travel-impedance variables (i.e., factors associated with travel costs), on average, have the greatest impact on predicting ridesourcing demand. Therefore, the results presented in Table 3 highlight an important contribution of our study to the literature on ridesourcing demand modeling: in addition to the variables commonly considered in the existing literature, we have further examined the travel-cost-related factors and assessed their significance in predicting ridesourcing demand.

In Table 4, we also present the three most important variables in each category and their overall ranking among all independent variables. As shown in Table 4, ridesourcing trip distance and trip time were the most important variables in the travel-impedance category. Note that the median fare of ridesourcing trips for an O-D pair was excluded from our models due to multicollinearity concerns (trip fare is

[7] Note that in the travel behavior research, transit-supply-related variables are often considered as built-environment factors.

Table 4
Three most important variables in each category.

Variable	Importance	Direction of association	Category	Overall ranking
Distance_median	4.24%	−	Travel impedance	5
Distance_sd	3.36%	+	Travel impedance	7
Time_median	3.32%	−	Travel impedance	8
CommutersHomeToWork	20.51%	+	Socioeconomic and demographic	1
CommutersWorkToHome	10.76%	+	Socioeconomic and demographic	2
PctWacLowMidWage_Ori	1.44%	−	Socioeconomic and demographic	13
EmpDen_Ori	8.60%	+	Built environment	3
EmpDen_Des	3.16%	+	Built environment	9
Walkscore_Ori	0.99%	+	Built environment	17
SerHourBusRoutes_Ori	4.55%	+	Transit supply	4
SerHourRailRoutes_Ori	4.08%	+	Transit supply	6
SerHourBusRoutes_Des	2.94%	+	Transit supply	11

Note: Variable codes ending with "_Ori" and "_Des" refers to variables measured at trip origin and trip destination, respectively.

mainly determined by trip distance and trip time); otherwise, we would expect it to be a variable of top importance. For the socioeconomic and demographic category, the most important variables were the total number of commuters (home-to-work and work-to-home) and the percentage of workers (at workplace) with a wage of less than $3333 per month at the origin zone. It is intuitive that O-D pairs with large commuting travel flows have great demand for ridesourcing. Among the built-environment variables, employment density (at both trip origin and destination) and Walk Score at trip origin had the greatest impact on predicting ridesourcing demand. Finally, among the transit-supply factors, the service frequencies of bus and rail services had the strongest correlation with ridesourcing demand.

To reveal the direction to which each variable is associated with ridesourcing demand, we have further generated partial independent plots for the 12 important features presented in Table 4. The partial dependence plot shows the marginal effect that a feature has on the predicted outcome of a machine learning model (Friedman, 2001). Due to the lack of space, however, we present these plots as supplementary materials and do not examine the results in detail. We simply discuss the direction of association between the features and ridesourcing demand as revealed by these plots. We found that the following variables were positively associated with ridesourcing trip frequency: *Distance_sd*, *CommutersHomeToWork*, *CommutersWorkToHome*, *EmpDen_Ori*, *EmpDen_Des*, *Walkscore_Ori*, *SerHourBusRoutes_Ori*, *SerHourRailRoutes_Ori*, and *SerHourBusRoutes_Des*. The following variables were negatively associated with ridesouricng demand: *Distance_median*, *Time_median*, and *PctWacLowMidWage_Ori*. Note that the "sign" of these variables derived from the random forest model are consistent with those provided by the multiplicative models.

It is quite intuitive that shorter trip distance, shorter travel time, greater commuter flows, and greater employment density were associated with greater volumes of ridesourcing trips. Also, *Distance_sd* was positively associated with ridesourcing trip frequency, and we believe that road congestion is the main mediating factor underlying this association: the existence of road congestion often makes ridesourcing take distinctive route choices, which in turn leads to a greater variation in trip distance; road congestion tends to happen in areas where travel demand is greater. Moreover, *PctWacLowMidWage_Ori* had a negative association with ridesourcing demand, which is likely because a lower proportion of low- and middle-wage ($3333/month or less) individuals use ridesourcing than that of higher-wage individuals. This finding is consistent with previous research which shows that a larger share of ridesourcing users come from higher-income neighborhoods (Brown, 2019). Finally, that a higher Walk Score and better transit services were associated with greater ridesourcing demand is probably due to the co-location of concentrated supply and demand: areas with a greater concentration of people and activities (e.g., downtown Chicago) generate more travel demand, and they are usually more walkable and better served by transit. The positive association between transit services and ridesourcing demand, by itself, does not tell us if ridesourcing is substituting or complementing transit.

6. Conclusions

This paper models travel demand for ridesourcing services in the City of Chicago. Using the ridesourcing trip data recently released by the City of Chicago along with over 63 variables collected from publicly available data sources to quantify the zonal characteristics, we develop direct demand models to predict the number of ridesourcing trips for each O-D (census-tract-to-census-tract) pair. We compare the model performance of a random forest model with that of the classic multiplicative model, and we find that the random forest model is superior in terms of predictive accuracy and model fit. These findings show the potential of leveraging machine learning techniques to improve travel demand modeling. We further compute variable importance for the random forest model to reveal which factors are strong predictors of ridesourcing demand. We find that socioeconomic and demographic variables contributed the most to the predictive power of random forest. However, travel impedance, built environment characteristics, and transit-related factors are also strongly correlated to ridesourcing demand.

It should be noted that the main purpose of this paper is to introduce the application of a machine-learning-based direct demand model to predict the travel demand for ridesourcing, an increasingly popular mobility option in cities. In other words, the intended application of the modeling approach demonstrated here is for empirical prediction rather than for causal explanation (see Shmueli et al. (2010) for an in-depth discussion on the difference between prediction and explanation). These interpretations can be facilitated by carefully examining partial dependence plots and accumulated local effects plots generated from machine learning models. These plots not only show the direction of association between each feature and the outcome variable but also reveal their nonlinear relationships (see an application by Ding et al. (2019)). Moreover, visualizing the partial dependence of two features at once allows one to explore the interactions between independent variables (Ding et al., 2018). Future work may explore the applications of these tools to generate insights into *how* different factors shape ridesourcing demand.

Moreover, the demand models built here intend to inform long-term transportation and land-use planning, and they are less appropriate for guiding travel-demand management. Notably, surge pricing is a strategy that TNCs heavily rely on to redistribute travel demand and supply across time and space, but it is unaccounted for in this study. When data becomes available, researchers should consider examining how pricing mechanisms such as surge pricing and fare discounts shape ridesourcing demand. Future work may also consider modeling travel demand for pooled and unpooled ridesourcing rides separately. Notably, the two types of ridesourcing services may attract distinctive

groups of travelers (e.g., travelers with different levels of income) and serve distinctive geographical areas (pooling rides is more feasible in higher-density areas). Nevertheless, this study did not do so because doing it would add great complexity to the discussion of model results. The main focus of this paper is to demonstrate the use of machine learning for demand modeling, and fitting separate models for different types of ridesourcing services adds little value to it. That being said, one can expect to build demand models of higher predictive capability by splitting ridesourcing trips into different service types, into weekdays versus weekends, and by time of day.

Another direction of future research can explore other machine learning algorithms and compare their predictive performance with the random forest model examined here. Future research may also compare the performance of the direct demand model with that of a sequential modeling approach (i.e., separate models on trip generation, trip attraction, and mode choice).

Acknowledgement

This research was partially supported by the U.S. Department of Transportation through the Southeastern Transportation Research, Innovation, Development and Education (STRIDE) Region 4 University Transportation Center (Grant P0147836 - Project B3).

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jtrangeo.2020.102661.

References

Alemi, F., Circella, G., Mokhtarian, P., Handy, S., 2019. What drives the use of ridehailing in California? Ordered probit models of the usage frequency of uber and lyft. Transp. Res. Part C: Emerg. Technol. 102, 233–248.
Baek, J., Sohn, K., 2016. Deep-learning architectures to forecast bus ridership at the stop and stop-to-stop levels for dense and crowded bus networks. Appl. Artif. Intell. 30, 861–885.
Bhattacharyya, S., Jha, S., Tharakunnel, K., Westland, J.C., 2011. Data mining for credit card fraud: a comparative study. Decis. Support. Syst. 50, 602–613. https://doi.org/10.1016/j.dss.2010.08.008.
Biau, G., 2012. Analysis of a random forests model. J. Mach. Learn. Res. 13, 1063–1095. URL. http://www.jmlr.org/papers/volume13/biau12a/biau12a.pdf.
Boarnet, M.G., Crane, R., et al., 2001. Travel by Design: The Influence of Urban Form on Travel. Oxford University Press on Demand.
Breiman, L., 1996. Bagging predictors. Mach. Learn. 24, 123–140.
Breiman, L., 2001. Random forests. Mach. Learn. 45, 5–32.
Brown, A., 2019. Redefining car access: ride-hail travel and use in Los Angeles. J. Am. Plan. Assoc. 85, 83–95.
Caruana, R., Karampatziakis, N., Yessenalina, A., 2008. An empirical evaluation of supervised learning in high dimensions. In: Proceedings of the 25th International Conference on Machine Learning. ACM, pp. 96–103.
Cervero, R., Murakami, J., Miller, M., 2010. Direct ridership model of bus rapid transit in los Angeles county, California. Transp. Res. Rec. 2145, 1–7.
Chen, M.K., Chevalier, J.A., Rossi, P.E., Oehlsen, E., 2017. The value of flexible work: Evidence from uber drivers. In: Technical Report. National Bureau of Economic Research.
Cheng, L., Chen, X., De Vos, J., Lai, X., Witlox, F., 2019. Applying a random forest method approach to model travel mode choice behavior. Travel Behav. Soc. 14, 1–10.
Choi, J., Lee, Y.J., Kim, T., Sohn, K., 2012. An analysis of metro ridership at the station-to-station level in Seoul. Transportation 39, 705–722.
Clewlow, R.R., Mishra, G.S., 2017. Disruptive transportation: The adoption, utilization, and impacts of ride-hailing in the united states. In: Institute of Transportation Studies. University of California, Davis.
Dagenais, M.G., Gaudry, M.J., et al., 1986. Can aggregate direct travel demand models work? In: Technical Report.
de Dios Ortuzar, J., Willumsen, L.G., 2011. Modelling Transport. John Wiley & Sons.
Dias, F.F., Lavieri, P.S., Garikapati, V.M., Astroza, S., Pendyala, R.M., Bhat, C.R., 2017. A behavioral choice model of the use of car-sharing and ride-sourcing services. Transportation 44, 1307–1323.
Ding, C., Cao, X., Wang, Y., 2018. Synergistic effects of the built environment and commuting programs on commute mode choice. Transp. Res. A Policy Pract. 118, 104–118.
Ding, C., Cao, X., Liu, C., 2019. How does the station-area built environment influence metrorail ridership? Using gradient boosting decision trees to identify non-linear thresholds. J. Transp. Geogr. 77, 70–78.
Erhardt, G.D., Roy, S., Cooper, D., Sana, B., Chen, M., Castiglione, J., 2019. Do transportation network companies decrease or increase congestion? Sci. Adv. 5 eaau2670.
Ewing, R., Cervero, R., 2001. Travel and the built environment: a synthesis. Transp. Res. Rec. 1780, 87–114.
Ewing, R., Cervero, R., 2010. Travel and the built environment: a meta-analysis. J. Am. Plan. Assoc. 76, 265–294.
Friedman, J.H., 2001. Greedy function approximation: a gradient boosting machine. Ann. Stat. 1189–1232.
Geng, X., Li, Y., Wang, L., Zhang, L., Yang, Q., Ye, J., Liu, Y., 2019. Spatiotemporal multi-graph convolution network for ride-hailing demand forecasting. In: 2019 AAAI Conference on Artificial Intelligence (AAAI'19).
Gerte, R., Konduri, K.C., Eluru, N., 2018. Is there a limit to adoption of dynamic ride-sharing systems? Evidence from analysis of uber demand data from New York city. Transp. Res. Rec. 2672, 127–136.
Hagenauer, J., Helbich, M., 2017. A comparative study of machine learning classifiers for modeling travel mode choice. Expert Syst. Appl. 78, 273–282.
Hall, J.D., Palsson, C., Price, J., 2018. Is uber a substitute or complement for public transit? J. Urban Econ. 108, 36–50.
Hastie, T., Tibshirani, R., Friedman, J., 2001. The Elements of Statistical Learning. vol. 1 Springer series in statistics, New York, NY, USA.
Henao, A., Marshall, W., 2017. A framework for understanding the impacts of ridesourcing on transportation. In: Disrupting Mobility. Springer, pp. 197–209.
Iseki, H., Liu, C., Knaap, G., 2018. The determinants of travel demand between rail stations: a direct transit demand model using multilevel analysis for the Washington dc metrorail system. Transp. Res. A Policy Pract. 116, 635–649.
Kepaptsoglou, K., Stathopoulos, A., Karlaftis, M.G., 2017. Ridership estimation of a new lrt system: direct demand model approach. J. Transp. Geogr. 58, 146–156.
Kraft, G., Wohl, M., 1967. New Directions for Passenger Demand Analysis and Forecasting. (Transportation Research/UK/).
Last, M., Maimon, O., Minkov, E., 2002. Improving stability of decision trees. Int. J. Pattern Recognit. Artif. Intell. 16, 145–159.
Lave, L.B., 1972. The demand for intercity passenger transportation. Reg. Sci. J. 12.
Lavieri, P.S., Bhat, C.R., 2019. Investigating objective and subjective factors influencing the adoption, frequency, and characteristics of ride-hailing trips. Transp. Res. Part C: Emerg. Technol. 105, 100–125.
Lavieri, P.S., Dias, F.F., Juri, N.R., Kuhr, J., Bhat, C.R., 2018. A model of ridesourcing demand generation and distribution. Transp. Res. Rec. 2672, 31–40.
Lebedev, A., Westman, E., Van Westen, G., Kramberger, M., Lundervold, A., Aarsland, D., Soininen, H.K., Loszewska, I., Mecocci, P., Tsolaki, M., et al., 2014. Random forest ensembles for detection and prediction of alzheimer's disease with a good between-cohort robustness. Neuro.Image: Clin. 6, 115–125. https://doi.org/10.1016/j.nicl.2014.08.023.
Lhéritier, A., Bocamazo, M., Delahaye, T., Acuna-Agost, R., 2018. Airline itinerary choice modeling using machine learning. J. Choice Model. 31, 198–209.
Liaw, A., Wiener, M., et al., 2002. Classification and regression by randomforest. In: R News. 2. pp. 18–22.
Lin, L., He, Z., Peeta, S., 2018. Predicting station-level hourly demand in a large-scale bike-sharing network: a graph convolutional neural network approach. Transp. Res. Part C: Emerg. Technol. 97, 258–276.
Ma, X., Zhang, J., Du, B., Ding, C., Sun, L., 2018. Parallel architecture of convolutional bi-directional lstm neural networks for network-wide metro ridership prediction. IEEE Trans. Intell. Transp. Syst. 20, 2278–2288.
Mangalathu, S., Sun, H., Nweke, C.C., Yi, Z., Burton, H.V., 2019. Classifying earthquake damage to buildings using machine learning. In: Earthquake Spectra, https://doi.org/10.1177/8755293019878137. 0000–0000.
McNally, M.G., 2007. The four-step model. In: Handbook of Transport Modelling, 2nd edition. Emerald Group Publishing Limited, pp. 35–53.
Molnar, C., 2019. Interpretable Machine Learning. https://christophm.github.io/interpretable-ml-book/.
Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al., 2011. Scikit-learn: machine learning in python. J. Mach. Learn. Res. 12, 2825–2830.
Pontius, R.G., Thontteh, O., Chen, H., 2008. Components of information for multiple resolution comparison between maps that share a real variable. Environ. Ecol. Stat. 15, 111–142.
Quinlan, J.R., 2014. C4. 5: Programs for Machine Learning. Elsevier.
Rayle, L., Dai, D., Chan, N., Cervero, R., Shaheen, S., 2016. Just a better taxi? A survey-based comparison of taxis, transit, and ridesourcing services in San Francisco. Transp. Policy 45, 168–178.
Sheather, S., 2009. A Modern Approach to Regression with R. (Springer Science & Business Media).
Shmueli, G., et al., 2010. To explain or to predict? Stat. Sci. 25, 289–310.
Talvitie, A., 1973. A direct demand model for downtown work trips. Transportation 2, 121–152.
Tixier, A.J.P., Hallowell, M.R., Rajagopalan, B., Bowman, D., 2016. Application of machine learning to construction injury prediction. Autom. Constr. 69, 102–114. https://doi.org/10.1016/j.autcon.2016.05.016.
Uber Newsroom, 2019. Company Info: Facts and Figures as of December 2018. Uber URL. https://www.uber.com/en-GB/newsroom/company-info.

Wallsten, S., 2015. The competitive effects of the sharing economy: how is uber changing taxis. In: Technology Policy Institute. 22. pp. 1–21.
Wang, K., Shi, X., Goh, A.P.X., Qian, S., 2019. A machine learning based study on pedestrian movement dynamics under emergency evacuation. Fire Saf. J. 106, 163–176.
Xie, C., Lu, J., Parkany, E., 2003. Work travel mode choice modeling with data mining: decision trees and neural networks. Transp. Res. Record: J. Transp. Res. Board 50–61.
Young, M., Allen, J., Farber, S., 2020. Measuring when uber behaves as a substitute or supplement to transit: an examination of travel-time differences in Toronto. J. Transp. Geogr. 82, 102629.
Yu, H., Peng, Z.R., 2019. Exploring the spatial variation of ridesourcing demand and its relationship to built environment and socioeconomic factors with the geographically weighted poisson regression. J. Transp. Geogr. 75, 147–163.
Zhao, J., Deng, W., Song, Y., Zhu, Y., 2014. Analysis of metro ridership at station level and station-to-station level in Nanjing: an approach based on direct demand models. Transportation 41, 133–155.
Zhao, X., Yan, X., Yu, A., Van Hentenryck, P., 2020. Prediction and behavioral analysis of travel mode choice: a comparison of machine learning and logit models. Travel Behav. and Soc (Forthcoming).