Identifying Key Factors Associated with Ridesplitting Adoption Rate and Modeling Their Nonlinear Relationships

Yiming Xu[a], Xiang Yan[b], Xinyu Liu[c], Xilei Zhao[a,1]

[a] *Department of Civil and Coastal Engineering, University of Florida, Gainesville, 32611, FL*
[b] *Department of Urban and Regional Planning, University of Florida, Gainesville, 32611, FL*
[c] *H. Milton Stewart School of Industrial and Systems Engineering, Georgia Institute of Technology, Atlanta, 30332, GA*

Abstract

Ridesharing is critical for promoting transportation sustainability. As a new form of ridesharing services, ridesplitting has rarely been studied. Based on the Chicago ridesourcing trip data, this study explores how ridesplitting adoption rate (i.e., the proportion of ridesourcing trips with ridesharing authorization) varies across space and what factors are associated with these variations. We find large variations in ridesplitting adoption rates across neighborhoods (Census Tracts) and across origin-destination (Census-Tract-to-Census-Tract) pairs. Particularly, the ridesplitting adoption rate is low for airport rides. We further apply a random forest model to explore which factors are key determinants of ridesplitting adoption rate across O-D pairs and to explore their nonlinear associations. The results suggest that the socioeconomic and demographic variables collectively contribute to 68.60% of the predictive power of the model, but travel-cost variables and built-environment-related factors are also important. The most important variables associated with ridesplitting adoption are ethnic composition, median household income, education level, trip distance, and neighborhood density. We further identify the nonlinear association between neighborhood ridesplitting adoption rate and several key variables such as the percentage of white population, median household income, and neighborhood Walk Score. The revealed nonlinear patterns can help transportation professionals identify neighborhoods with the greatest potential to promote ridesplitting.

Keywords: ridesplitting; random forest; nonlinearity; variable importance; interpretable machine learning; ridesharing

1. Introduction

The shared version of app-based, on-demand ridesourcing service, usually provided by transportation network companies (TNCs) such as Uber and Lyft, is commonly termed as ridesplitting. Many transportation observers have suggested that promoting pooled rides such as ridesplitting is the key to transportation sustainability (e.g., Sperling, 2018; Merlin, 2017; Shaheen and Cohen, 2019). By pooling riders together, ridesplitting can bring a variety of environmental benefits as it can mitigate traffic congestion, reduce fuel consumption and greenhouse gas emissions, and curb parking demand (Feigon and Murphy, 2016; Amey et al., 2011; Ferguson, 1997; Kelly, 2007; Morency, 2007). The importance of ridesplitting is further undergirded by research on autonomous vehicles (AVs) that shows dramatic differences in

[1] Corresponding author.
Email address: xilei.zhao@essie.ufl.edu (Xilei Zhao)

simulation outputs as the proportion of shared trips is assumed to vary. For example, various studies have shown that ridesplitting is crucial for the introduction of AVs not to be accompanied by increases in vehicle miles traveled, traffic, and carbon emissions (Levin et al., 2017; Merlin, 2017; Wang et al., 2018).

However, the overall understanding of ridesplitting is lacking in critical aspects. Notably, we have little knowledge of how ridesplitting adoption rate (i.e., the proportion of ridesourcing trips with shared-trip authorization) varies across space (i.e., neighborhoods and origin-destination (O-D) pairs) and what factors are associated with these variations. An understanding of these issues can guide transportation professionals in developing planning strategies and policy measures to promote ridesplitting. For example, identifying neighborhoods with the low ridesplitting adoption rate can inform TNCs and transportation policymakers on where to target if they aim to promote shared rides. Moreover, while neighborhood characteristics such as population density and public safety may shape ridesplitting adoption, we barely know the extent of these associations. Such knowledge can help transportation planners develop place-based strategies (e.g., improving neighborhood street connectivity) that can facilitate ridesplitting.

The public release of the ridesourcing trip data in the city of Chicago provides a unique research opportunity to shed light on some of these issues. This trip-level dataset contains essential information about a trip such as its origin and destination, fare, distance, travel time, and if the trip is shared-trip authorized, but it contains no information regarding the riders. While this dataset is inappropriate for conducting in-depth behavioral analysis (for example, examining individual preferences for different trip attributes and various travel options) as researchers often do with individual-level travel data, it allows researchers to examine the spatial variation in the ridesplitting adoption rate (across neighborhoods and O-D pairs) and to explore how it is related to trip characteristics and neighborhood characteristics (at the trip origin and destination). This study presents such an analysis.

Specifically, we first conduct a descriptive analysis of the ridesplitting adoption rate across neighborhoods in the city of Chicago, noting several salient spatial patterns. We then construct a variety of variables, including trip characteristics and neighborhood-level variables at both the trip origin and destination, and apply random forest (a widely-adopted machine learning algorithm) to model the association between these variables and the ridesplitting adoption rate for each O-D pair. Interpreting the model outputs allows us to identify the important factors associated with the ridesplitting adoption rate. Since the random forest algorithm can readily reveal the nonlinear relationships between input variables and the outcome variable, we further explore the nonlinear associations between several important factors and the ridesplitting adoption rate. Our research contributes to the existing knowledge on ridesplitting, informs simulation studies on forming sensible assumptions on individuals' willingness to share rides, and guides transportation professionals to develop targeted policies that promote ridesplitting.

The remaining paper is structured as follows. Section 2 reviews the existing literature related to this study. Section 3 describes the dataset and methods used for analysis. Section 4 introduces the analysis and results of this study. Section 5 presents the discussions on results and potential applications of the findings. Section 6 concludes the paper by summarizing the strengths and limitations of the study and suggesting future research directions.

2. Literature review

2.1 Studies on ridesharing adoption

Ridesplitting is a new form of ridesharing, and traditional forms of ridesharing include carpooling and vanpooling. Investigating the determinants of ridesharing adoption has been a major research topic in the transportation literature. Among the various forms of ridesharing, carpooling has been studied most extensively. More recent research interests lie in ridesplitting and pooled AV rides (Delhomme and Gheorghiu, 2016; Lavieri and Bhat, 2019; Li et al., 2019; Shaheen and Cohen, 2019). The various ridesharing schemes have notable differences: people usually carpool with people of a certain degree of familiarity (e.g., acquaintances and co-workers), ridesplitting means sharing rides with strangers but involves a driver, and pooled AV rides operate in the absence of a driver who can potentially offer a sense of security.

Despite these differences, the literature shows that the facilitators and deterrents to various ridesharing schemes are largely the same. Studies consistently find that travel-time savings and monetary savings are the main motivation for both drivers and passengers to accept carpooling (Habib et al., 2011; Shaheen et al., 2016; Morris et al., 2019), ridesplitting (Sarriera et al., 2017), and pooled AV rides (Lavieri and Bhat, 2019). Also, researchers have shown that the socioeconomic and demographic characteristics of participants, such as age, gender, income, and education level, are important factors that affect ridesharing adoption (Sarriera et al., 2017; Delhomme and Gheorghiu, 2016; Rayle et al., 2016). In addition to the factors discussed above, some recent studies have investigated psychosocial factors associated with ridesplitting adoption, for example, personality types and concerns about privacy (Sarriera et al., 2017; Amirkiaee and Evangelopoulos, 2018; Moody, Middleton, and Zhao, 2019). Survey results from these studies suggest that travelers are hesitant about sharing a car with strangers due to a desire for personal space, aversion to social situations, distrust, and concerns about security and privacy (Shaheen and Cohen, 2019; Tahmasseby et al., 2016; Amirkiaee and Evangelopoulos, 2018; Morris et al., 2019).

Ultimately, the decision of whether to share rides with others is shaped both by individual travel preferences and by the context in which a trip takes place. Neighborhood environments can shape individual decisions to share rides or not in several ways. First, neighborhood characteristics such as population density and employment concentration can affect the price difference between an unshared ridesourcing trip and a shared one. This is because TNCs determine the discount for ridesplitting trips based on the likelihood of a trip being matched up, and trips originated from denser neighborhoods with more ridesourcing demand are more likely to be successfully matched (Perea, 2016). Second, individuals' willingness to share rides with strangers is likely affected by the type of people they are expected to encounter. Previous research has shown that individuals tend to prefer to share rides with people of similar socioeconomic class and the same race/ethnicity (Shaheen, Chan, and Garnor, 2016). Finally, neighborhoods vary in the degree of safety perceived by ridesourcing users, which in turn can affect their decision to adopt ridesplitting or not (Alemi et al., 2018).

2.2 Existing analysis of ridesourcing data

Most studies on ridesharing adoption are based on surveys, interviews, and focus groups. Recently, some researchers have analyzed ridesourcing trip data that are either directly provided by TNCs or made publicly available on government websites (Chen et al., 2017; Chen et al., 2018; Lavieri et al., 2018; Brown, 2019; Li et al., 2019; Yu and Peng, 2020). Compared

to conventional survey data, ridesourcing trip data have both pros and cons. A major advantage of the ridesourcing trip data is that all the trips made by all travelers are recorded, thus avoiding any sampling bias problems. Also, while survey research requires great monetary and labor investments in interacting with the respondents, the acquisition of large-volume ridesourcing trip data is cheap and efficient. On the other hand, since the ridesourcing trip data usually contain little information on the riders (largely due to privacy concerns), they cannot be used for deriving insights on individual preferences, attitudes, and behavior.

Researchers have applied large-scale ridesourcing trip data to examine various research topics. Chen et al. (2017) applied an ensemble learning approach to predict the ridesplitting choice of passengers in Hangzhou. In a different study, they further applied the same dataset to explore the impacts of ridesplitting on multimodal mobility (Chen et al., 2018). Li et al. (2019) used the DiDi data in Chengdu to compare the characteristics of ridesplitting trips with those of unshared ridesourcing trips. Lavieri et al. (2018) and Yu and Peng (2020) and applied the RideAustin data to explore factors associated with ridesourcing demand. Yan et al. (2020) applied random forest to model and forecast ridesourcing demand in the City of Chicago. Finally, Brown (2019) examined the association between Lyft travel, the built environment, and neighborhood socioeconomic characteristics. More recently, Brown (2020) further analyzed the factors that are associated with where ridesplitting occurs and who uses ridesplitting.

These studies have generated valuable insights regarding ridesplitting adoption and the characteristics of ridesplitting trips. For example, Chen et al. (2018) showed that trip travel time, trip costs, trip length, waiting time fee, and travel time reliability are the main factors that shape ridesplitting behavior. Brown (2019, 2020) found that people living in low-income neighborhoods make shorter, cheaper, and more shared trips than those living in higher-income neighborhoods. Nevertheless, none of these studies have examined how the ridesplitting adoption rate (i.e., the proportion of ridesourcing trips with shared-trip authorization) varies across space and what factors contribute to these variations. As we have discussed above, knowledge of these topics can guide the design of strategies to promote ridesplitting across neighborhoods and to facilitate the process of transportation planning in general. This study fills this knowledge gap by examining the Chicago ridesourcing data.

2.3 The application of machine learning to explore nonlinear relationships

Most of the studies on ridesharing adoption are based on linear models (Lavieri and Bhat, 2019; Sarriera et al., 2017), where the relationships between ridesharing adoption and the associated factors are assumed to be linear. However, the influence of a factor on traveler's ridesplitting adoption may differ in different value intervals and may have upper and lower thresholds. For example, the marginal effects of household income on ridesplitting adoption may become negligible when the household income level reaches a certain threshold; that is, upper-middle-income households may behave similarly as high-income households. At the neighborhood level, such nonlinearity may manifest as ridesplitting adoption rate decreasing with median household income increases initially, but this negative association becomes insignificant once a median household income level is reached. Galster (2018) provided an in-depth discussion on the behavioral mechanisms underlying the nonlinear and threshold relationships present in many neighborhoods, such as household mobility behaviors and property investment decisions.

The availability of machine learning (ML) methods allows researchers to explore nonlinear and threshold effects conveniently. ML is unlike conventional statistical methods that often assume a pre-determined functional form, ML allows the model structure to freely vary and

thus can readily capture the nonlinear patterns underlying the data (Lhéritier et al., 2019; Molnar, 2019; Zhao et al., 2020). Some empirical studies have applied ML to explore the nonlinear relationship between input factors and the outcome variable and proven the effectiveness of ML in these applications (Ding et al., 2018a; Ding et al., 2018b; Ding et al., 2019; Auret and Aldrich, 2012; Golshani et al., 2018). For example, Ding et al. (2018a) applied gradient boosting decision trees (GBDT) to model driving distance in Oslo and found that built environment characteristics have salient nonlinear effects on driving distance. Notably, the researchers found that when population density reached 3000 persons per square kilometer, additional density produced a trivial effect on driving distance reduction. In a different study context, Ding et al. (2018b) showed that built environment characteristics at residential locations and workplaces, such as population density, employment density, and bus stop density, have nonlinear impacts on the probability of choosing driving for commute in Washington, D.C. Moreover, Tao et al. (2020) found that spatial attributes such as population density and street-network density have strong nonlinear associations with walk distance to transit.

3. Data and Methods

3.1 The data

The main data source for this study came from the ridesourcing trip data published by the City of Chicago on the publicly available Chicago Data Portal[2]. The City of Chicago ordinance required TNCs to report all trips (starting from November 2018) that took place within the city boundary every quarter. This study collected data released until March 31, 2019, which includes a total of 45,338,599 trips. Every trip record noted whether the rider authorized ridesharing (i.e., whether the rider is willing to take a potential shared trip). The data also detailed trip attributes such as fare, distance, duration, start and end times, as well as pick-up and drop-off locations. To protect privacy, the data publisher, the City of Chicago, aggregated the pick-up and drop-off locations at the Census-Tract level, rounded the trip-start and trip-end times to the nearest 15 minutes, and rounded the fares and tips to the nearest $2.50 and $1.00, respectively.

To prepare the data for analysis, we need to summarize the TNC trips at the disaggregate level (i.e., individual trip records) into an aggregate level (i.e., origin-destination [O-D] pairs of Census Tracts). However, a significant proportion of ridesourcing trips lack the Census Tract ID. This is because the City of Chicago has applied de-identification and aggregation techniques to reduce the risk of linking individuals' trip location data to their identities. For instance, if there are no more than two trips in the same Census Tract and 15-min time window, the Census Tract IDs of these trips will be removed, and the city will report both ends of such trips at the Community Area level. Since our unit analysis is Census-Tract-to-Census-Tract, we would have to either removing these trips from the analysis or inferring their locations at the Census Tract level[3]. We found that a disproportionately large number of trips with such

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

[3] Note that the Community Areas are essentially groups of Census Tracts. Hence, based on the one-to-multiple correspondence between the Community Areas and the Census Tracts, we estimated the Census Tract ID of each trip with a stratified sampling procedure consisting of the following steps: 1) For each recorded Census Tract i, we counted its total occurrence for mode m as either pickup or drop-off, respectively as C_j^m. 2) For each Community Area j, let the set I_j contains all Census Tracts in Community Area j, which is a known input. We calculated the empirical distribution for each Community

characteristics were authorized for ridesharing, which means that excluding them would lead to an underestimate of the ridesplitting adoption rate.

To solve this issue, we developed an algorithm to infer trip origins/destinations from the Community Area level to the Census Tract level. The trips recovered from the above procedure were used to calculate the aggregate-level variables including ridesplitting adoption rate for each O-D pair, median trip fare and travel time for ridesharing authorized trips, and median trip fare and travel time for unshared ridesourcing trips. Note that before the calculation, we removed some uncommon trip records that we considered as outliers[4]. Finally, to minimize the impact of randomness, we excluded some O-D pairs with less than 50 trips throughout November 2018 and March 2019. These data processing steps prepares a total number of 92,683 O-D pairs for analysis, with 784 Census Tracts as trip origins and 787 as trip destinations (the total number of Census Tracts in the city of Chicago is 801).

The "Ridesplitting adoption rate" variable is the main variable of interest in this study, which is likely to be shaped by a variety of factors, including trip characteristics (e.g., the cost and travel-time differences between a shared and unshared ridesourcing trip), socioeconomic and demographic characteristics of the neighborhoods at which the trip starts and ends, and some neighborhood-environment variables (at both trip origins and destinations) such as crime rate and road-network density. We obtained a list of socioeconomic and demographic variables from the American Community Survey (ACS) 2013-2017 5-year estimates data, some employment- and worker-related variables from the 2015 Longitudinal Employer-Household Dynamics (LEHD) data, and crime rate data from the Chicago Data Portal. Furthermore, we used General Transit Feed Specification (GTFS) data to estimate some transit-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. Table 1 presents the description and the descriptive statistics of the outcome variable and the input variables examined this study. In addition, a descriptive profile of ridesplitting ridesourcing trips with ridesharing authorized and unauthorized trips is presented in Table 2 to provide a sense of their characteristics and how they compare. In general, we found that ridesplitting authorized trips have lower trip fare costs, longer trip distance, and longer trip duration.

Table 1. Descriptive profile of the outcome variable and input variables

Variable	Unit	Mean	St. dev.	Data source
Outcome Variable				
Ridesplitting adoption rate	-	32.34%	15.55%	Ridesourcing trips
Trip-cost-related variables[1]				
Median trip distance	mile	4.90	3.54	Ridesourcing trips
Median fare for unshared trips minus shared trips	US dollar	2.72	2.46	Ridesourcing trips

Area over all of its potential Census Tracts for each mode separately, denoted by $P_j^m(i) \coloneqq P$ (a trip of mode m in Community Area j is attributed to a Census Tract $i \in I_j$), as $P_j^m(i) = C_i^m / (\sum_{i' \in I_j} C_{i'}^m)$. 3) For each trip and mode combination with hidden Census Tract and known Community Area j, we sample from the set I_j according to the empirical distribution calculated in Step (2). This estimate is chosen such that, $\hat{\imath}_j^m = \{i \ w.p. \ P_j^m(i) \ \forall \ i \in I_j\}$, which only depends on the Community Area and whether the estimate is for a pickup or a drop-off location.

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

Median fare for shared trips divided by unshared trips	US dollar	0.78	0.22	Ridesourcing trips
Socioeconomic and demographic variables[2]				
Percentage of male population	-	48.29%	7.16%	ACS 2013-2017
Percentage of population with bachelor's degree and above	-	49.16%	28.35%	ACS 2013-2017
Percentage of population aged 18-44	-	50.01%	14.81%	ACS 2013-2017
Percentage of white population	-	55.25%	31.25%	ACS 2013-2017
Percentage of Hispanic population	-	21.23%	24.53%	ACS 2013-2017
Percentage of Asian population	-	7.72%	9.76%	ACS 2013-2017
Percentage of households with at least one car	-	70.38%	16.07%	ACS 2013-2017
Percentage of workers taking transit to work	-	32.74%	12.89%	ACS 2013-2017
Median household income	US dollar	65637.43	34916.86	ACS 2013-2017
Percentage of renter-occupied housing units	-	59.06%	17.14%	ACS 2013-2017
Percentage of single-family homes	-	21.88%	20.28%	ACS 2013-2017
Percentage of workers with earnings $3,333/month or less	-	68.28%	16.31%	2015 LEHD
Percentage of workers with bachelor's degree and above	-	20.29%	6.51%	2015 LEHD
Neighborhood environment variables[2]				
Population density	per square mile	22359.54	15674.83	ACS 2013-2017
Employment density	per square mile	21083.07	68482.70	ACS 2013-2017
Retail employment density	per square mile	172.16	117.57	ACS 2013-2017
Density of violent crime	per square mile	171.33	184.55	ACS 2013-2017
Road network density	miles per square mile	25.68	8.56	Smart Location Database
Intersection density	per square mile	125.92	104.93	Smart Location Database
Walk Score of centroid of Census Tract	-	79.28	17.48	Walkscore.com API
Aggregate service hours for rail routes	hours per day	363.75	427.14	GTFS
Bus stop density	per square mile	64.38	34.62	GTFS
Rail stop density	per square mile	1.58	3.40	GTFS
Percentage of tract within 1/4 mile of a bus stop	-	93.04%	15.98%	GTFS
Percentage of tract within 1/4 mile of a rail stop	-	23.54%	31.05%	GTFS

Note: 1. The trip-cost-related variables are at the zone-to-zone (i.e., Census-Tract-to-Census-Tract) level;
2. The socioeconomic, demographic, and neighborhood-environment variables are at the Census Tract level; each variable is included twice in the model, once at the trip origin and once at the trip destination.

Table 2. Descriptive profile of ridesplitting authorized and unauthorized trips

	Ridesplitting authorized trips	Ridesplitting unauthorized trips
Total number of trips	8,781,680	27,151,808
Median of trip fare (in US dollars)	7.50	10.05
Median of trip distance (in miles)	3.35	2.66
Median of trip duration (in minutes)	14.50	11.83

It should be noted that we initially examined more variables that are potentially related to the ridesplitting adoption rate. However, to reduce multicollinearity, we excluded most variables that had a variance inflation factor (VIF) score greater than 10, a common threshold applied to determine multicollinearity (Sheather, 2009).[5] The VIF values of the remaining 53 variables are shown in Appendix A.

[5] The following variables were examined but excluded: mean trip fare, mean trip distance, mean trip duration, median trip fare, median trip duration, average household size, percentage of black population, unemployment rate, income per capita, percentage of individuals below poverty, percentage of moderate-income households ($25-$50k), percentage of middle-

3.2 Methods

To understand how the ridesplitting adoption rate varies across neighborhoods, we first applied Geographic Information System techniques to examine the spatial variations of ridesplitting adoption rate across the city. To further explore what factors shape these variations and to explore their potential nonlinear relationships, we applied ML (i.e., random forest [RF]) to predict the ridesplitting adoption rate at the O-D pair (Census-Tract-to-Census-Tract) level. We also fit an Ordinary Least Squares (OLS) model and compared its results with those of the RF model. The purpose of this comparison is twofold. First, previous studies have shown that there are trade-offs between conventional statistical models and ML models regarding predictive accuracy and causal inference (Shmueli, 2010). ML models can perform better in prediction but not necessarily lead to causal inference. Thus, the results of the OLS model can corroborate the findings from the RF model. Also, as many researchers are less familiar with ML than conventional statistics, interpreting its results with those of an OLS model can improve the understanding of the RF model.

3.2.1 Modeling the ridesplitting adoption rate using random forest

To model the ridesplitting adoption rate, we use random forest (RF), one of the most popular supervised learning methods. RF is among the most accurate general-purpose method with the ability to deal with high dimensional data (Biau, 2012). RF is quite robust as it can directly model different data types and insensitive to skewed distributions, missing values, outliers, and inclusion of irrelevant variables (Breiman, 2001). In addition, it is relatively easy to tune the hyperparameters of RF as it has two major hyperparameters, i.e., the number of variables randomly sampled as candidates at each split and the number of trees in the forest. And RF is usually not very sensitive to the values of these hyperparameters (Liaw and Wiener, 2002). Most importantly, RF is able to model the complex nonlinear relationships between independent and dependent variables, thanks to its flexible modeling structure (Breiman, 2001).

Specifically, the RF algorithm generates a set of decision trees where each decision tree is trained on a bootstrapped sample from the original data set, and the optimal node splitting variable is selected from a random subset of all the independent variables. Both bootstrapping and random selection of variables could reduce the correlation between the generated decision trees and thus the average prediction of all the decision trees is expected to overcome the overfitting problems and has lower variance than individual decision trees.

RF has two major hyperparameters to tune: the number of variables randomly sampled as candidates at each split, and the number of trees in the forest. In this study, the hyperparameters of the RF are tuned by using 10-fold cross-validated grid-search. Three commonly-used measures (i.e., square root of the total number of variables, base two logarithm of the total number of variables, and the total number of variables) were tested to determine the number of variables in the random subset at each node. For the number of trees in the forest, we examined values from 10 to 200 at an interval of 10. In total, there were 60 possible combinations. The optimal number of variables in the random subset is selected as 7 (i.e., the square root of the total number of variables), and the optimal number of trees in the forest is 110.

After selecting the best RF model, we further compare its model fit and predictive capability with the benchmark OLS model, and two other popular machine learning models, including decision tree (DT) and multi-layer perceptron (MLP). Both DT and MLP are widely used

---

income households ($50k to $75k), job accessibility by auto, worker accessibility by auto, total number of commuters, percentage of commuters aged 54 or younger, total number of jobs, percentage of jobs taken by workers aged 54 or younger, percentage of jobs with earnings $3,333/month or less, total number of workers, percentage of jobs taken by workers with bachelor's degree and above, percentage of workers aged 54 or younger, and aggregate service hours for bus routes.

nonparametric ML algorithms that are famous for modeling nonlinear relationships (Molnar, 2019). The four models are evaluated by 10-fold cross-validation. The model performance is evaluated by in-sample and out-of-sample root mean squared error (RMSE) and mean absolute error (MAE).

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

$$MAE = \frac{1}{N}\sum_{k=1}^{N}|\hat{y}_k - y_k| \quad (2)$$

where N is the total number of observations, y_k is the kth observed value for the dependent variable, and $\hat{y}_k$ is the kth predicted value for the dependent variable.

The model comparison results are presented in Table 3.

Table 3 Model comparison results

Model	In-Sample Performance		Out-of-Sample Performance	
	RMSE	MAE	RMSE	MAE
OLS	0.1011 ± 0.0001	0.0781 ± 0.0001	0.1012 ± 0.0013	0.0782 ± 0.0010
DT	0.0994 ± 0.0004	0.0756 ± 0.0003	0.0997 ± 0.0011	0.0759 ± 0.0008
MLP	0.0829 ± 0.0012	0.0622 ± 0.0010	0.0842 ± 0.0011	0.0632 ± 0.0009
RF	0.0285 ± 0.0001	0.0207 ± 0.0000	0.0761 ± 0.0011	0.0557 ± 0.0007

For the in-sample performance, RF is significantly better than the other three models: the RF's RMSE and MAE are around 30% of the other three models' results. For the out-of-sample performance, RF is also better than the others: the RF's RMSE and MAE are around 10%-20% better than the others' outputs. The better model fit and higher predictability of RF are probably due to its flexible modeling structure that can automatically capture nonlinearities and variable interactions. Moreover, RF can effectively reduce the negative influences of outliers on model performance by binning them.

3.2.2 Model interpretation methods

To identify key factors associated with the ridesplitting adoption rate and examine their relationships, we further interpret the "black-box" RF model with three ML interpretation methods, including variable importance (Breiman, 2001), partial dependence plots (PDP) (Hastie et al., 2009) and accumulated local effects (ALE) plots (Apley, 2016). We used variable importance to identify key determinants of ridesplitting adoption rate, and we used PDP and ALE plots to reveal the relationships between input variables and the dependent variable.

The most commonly used variable importance measure for RF is *Mean Decrease Impurity*. It evaluates the importance of variable x_i by computing the mean decrease in node impurities (measured by variance) from splitting on this variable. We will report the relative importance of each variable, with the total relative importance of all variables scaled to 100%. Note that variable importance represents the relative contribution of a variable to the predictive power of a model, and it does not indicate the direction to which a variable is associated with the outcome variable.

The PDP shows the marginal effect that a variable has on the predicted outcome of a ML model (Friedman, 2001). PDP works by marginalizing the ML model output over the distribution of the variables in the complement set of the selected variable(s), so PDP shows

the relationship between the selected variable(s) we want to evaluate and the predicted outcome (Molnar, 2019). However, PDP assumes that the variable(s) under evaluation is independent of the other variables. If they are highly correlated, PDP creates new data points in the areas of the variable distribution where the actual probability is very low, often leading to biases in results. One solution to tackle this problem is applying ALE plots (Apley, 2016; Molnar, 2019).

ALE plot is an alternative approach for visualizing the effects between the selected variables and the predicted outcome for ML models. ALE plot averages the changes in the predictions and accumulates them over the grid. Note that the ALE plot is centered at zero and thus the mean effect is zero. The value of the ALE can be interpreted as the main effect of the variable at a given value compared to the average prediction of the data; hence, ALE plot is an unbiased alternative to PDP (Apley, 2016; Molnar, 2019). The most important advantage of ALE plot is that it can generate valid interpretations when variables are correlated. ALE plot is also less computationally expensive (Apley, 2016). However, the implementation of ALE plot is much more complex compared to PDP. It is also tricky to set the number of intervals when constructing ALE plots: if the number is too small, the ALE plot might not be very accurate because of too few observations per interval; if the number is too high, the ALE curve may become very bumpy. As PDP and ALE plots have their pros and cons, we apply both to help interpret the RF model.

4. Results

4.1 Spatial patterns of ridesplitting adoption rate

As shown in Table 1, the mean ridesplitting adoption rate across O-D pairs was 32.34%, with a standard deviation of 15.55%. In other words, on average, about a third of trips for each O-D pair are authorized for ridesharing. Since it is difficult to identify meaningful spatial patterns from visualizing 92,683 O-D pairs, we adopted the following visualization strategy. First, we computed ridesplitting adoption rates for each Census Tract as trip destinations and origins, respectively, and developed two corresponding Choropleth maps (Figure 1). These maps allow us to learn the variation of ridesplitting adoption rates across neighborhoods. Second, we mapped the O-D pairs with the highest (90th percentile) and lowest (10th percentile) ridesplitting adoption rates, respectively, in two separate maps (Figure 2(a) and Figure 2(b)). To facilitate the interpretation of these two maps, we further created a map that visualizes the ridesourcing trip flows (Figure 2(c)). For all of these maps, we drew a line only for O-D pairs with more than 300 trips.

Several notable spatial patterns are presented in Figure 1. First, the two Choropleth maps are very similar to each other, suggesting that there is little difference in the ridesplitting adoption rate for each Census Tract being the trip origin versus being the trip destination. Second, the city can be roughly divided into three distinctive zones according to the levels of ridesplitting adoption rate: Northern and Central (low), West and Southwest (high), and South and Far South (medium). These patterns appear to be closely related to the income levels and ethnic composition of Census Tracts. Third, trips to and from the two airports had a relatively low ridesplitting adoption rate.

We inferred some additional insights from Figure 2. Noticeably, O-D pairs with a high ridesplitting adoption rate and O-D pairs with a low ridesplitting adoption rate are taking up almost completely different and concentrated spaces. The ridesplitting adoption rate is the highest for trips happening in the West Side and the Southwest Side and the lowest for trips connecting to the two airports and those happening in the North Side and in Central Chicago.

Airport travel should thus be a main target if the city wishes to promote ridesplitting. Also, considering that large volumes of ridesourcing trips concentrate at the North Side and Central Chicago but the ridesplitting adoption rate is not particularly high, transportation planners should consider developing spatially concentrated strategies to facilitate ridesplitting in these areas. In the following sections, we apply multivariate modeling to further explore which factors shape ridesplitting adoption across different neighborhoods.

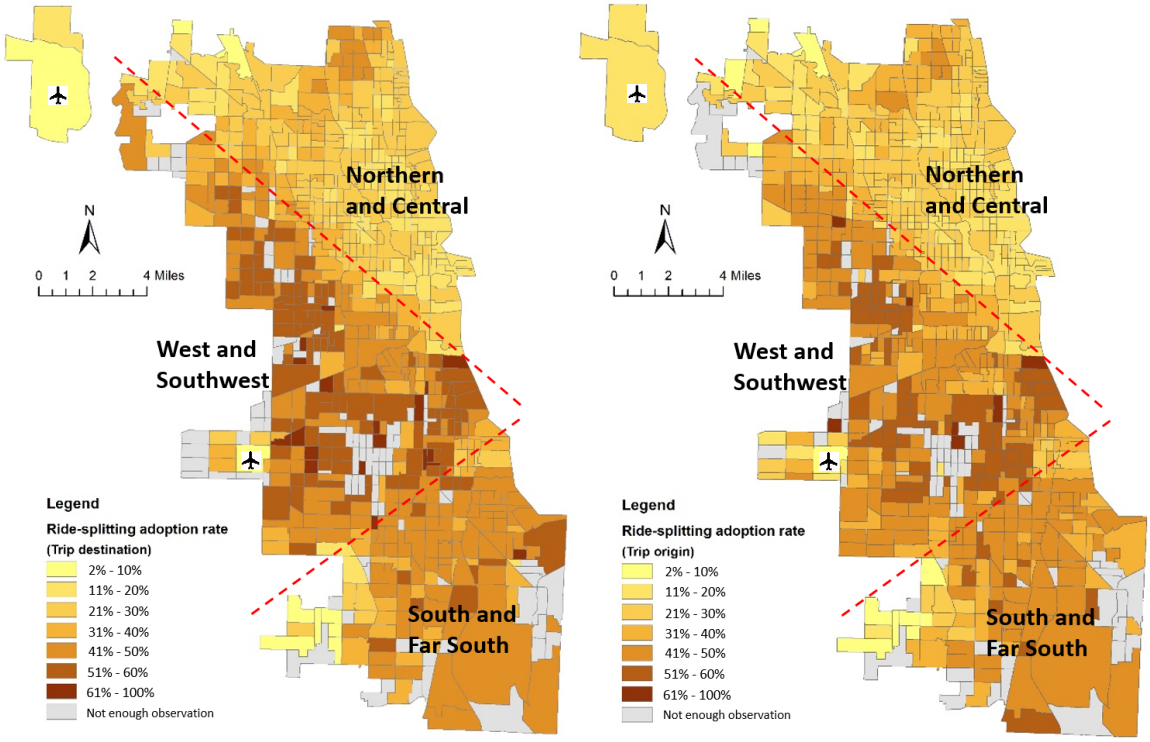

Figure 1. Ridesplitting adoption rate for each Census Tract as trip destination (left) and as trip origin (right)

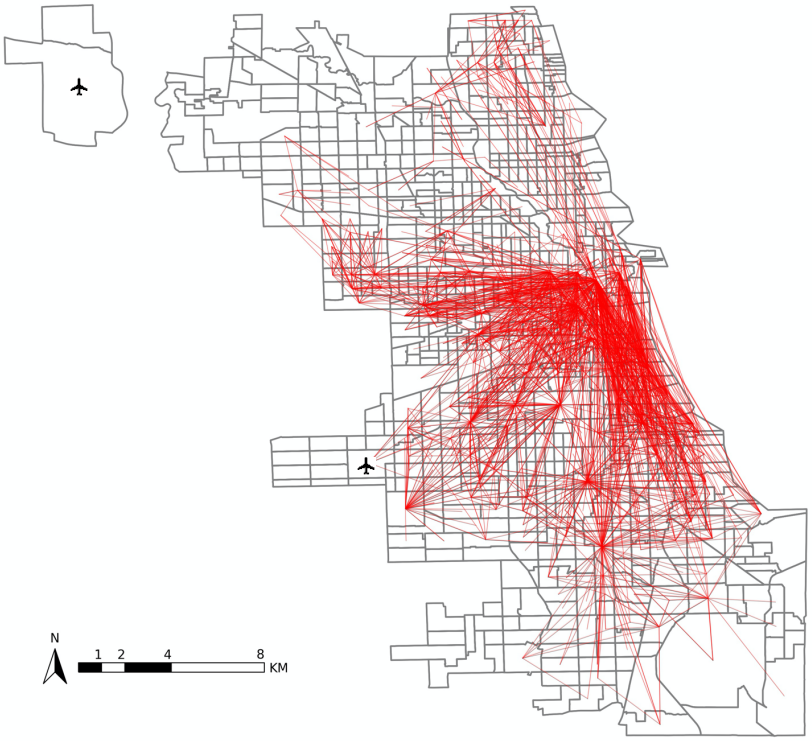

Figure 2(a) O-D pairs with highest (90th percentile) ridesplitting adoption rates

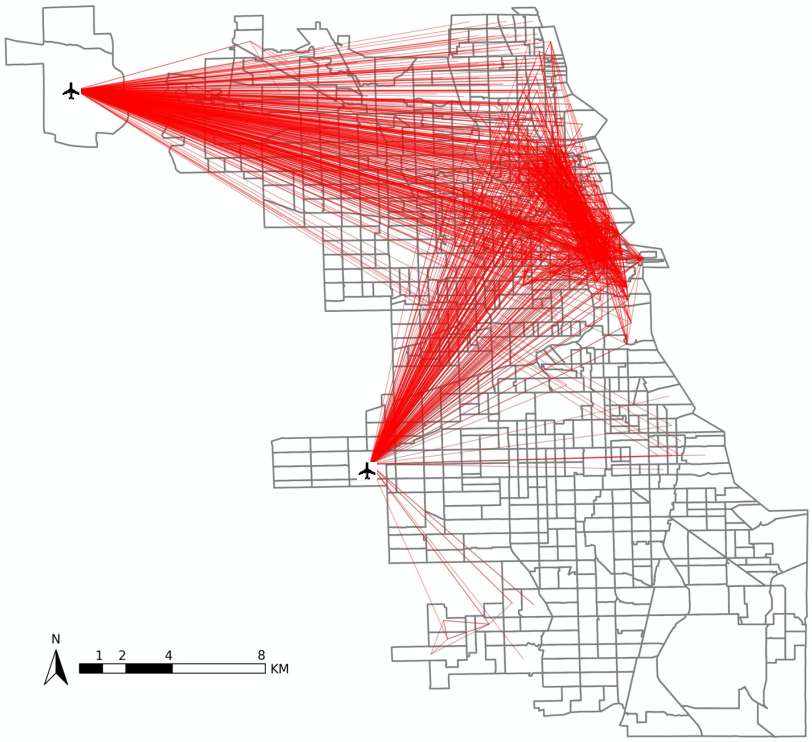

Figure 2(b) O-D pairs with lowest (10th percentile) ridesplitting adoption rates

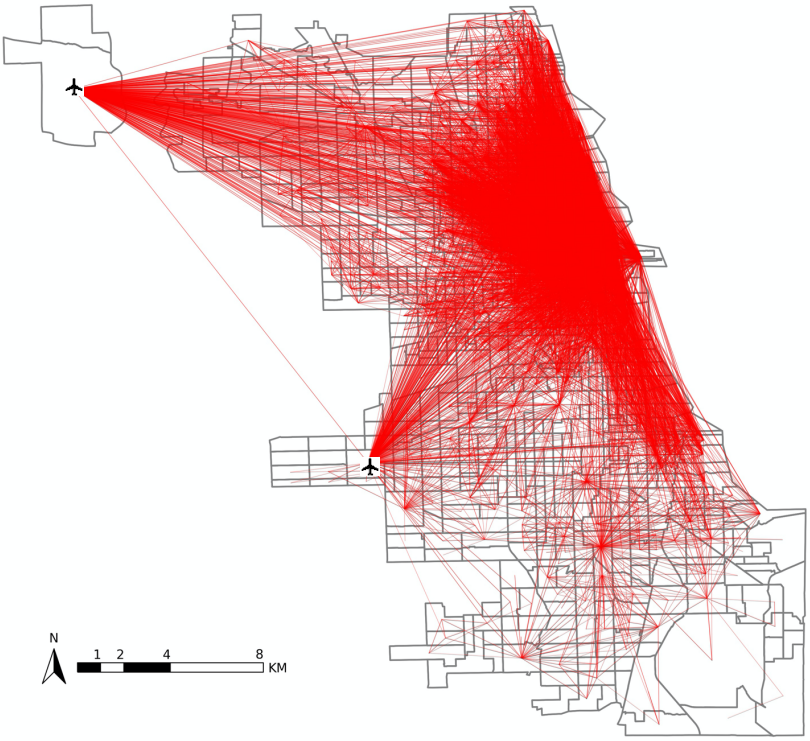

Figure 2(c) Ridesourcing trip flows

Figure 2. O-D pairs with highest (90th percentile) and lowest (10th percentile) ridesplitting adoption rate and ridesourcing trip flows

4.2 Factors associated with ridesplitting adoption rate

In this section, we present the outputs and interpretations of the RF model and compare the results with those obtained from the OLS model. As we discussed above, the comparison allows us to both substantiate the findings of the RF model and to demonstrate the additional insights (e.g., nonlinear relationships) that can be extracted from the RF model. We first discuss the variable importance, which shows the strength of association between each independent variable and the dependent variable. We then interpret the PDP and ALE plots generated from the RF model, which can reveal the direction of these associations, potential nonlinear relationships, and threshold effects.

4.2.1 Variable importance

Figure 3 shows the top 20 most important variables (variables with the highest relative variable importance values) in the RF model. We computed a similar variable importance measure for the OLS model and presented the results in the same figure as a comparison.[6] In the graph, we also show the direction of the association between each variable and ridesplitting adoption rate but leave the discussions to the next subsection.

[6] For OLS model, standardized Beta coefficients can represent the impact of each independent variable on the outcome. The relative importance of each variable in OLS model is the relative scale of standardized Beta coefficients. $F_i = \beta_i / \sum_{k=1}^{n} \beta_k$, where F_i is relative importance of variable i, β_i is standardized Beta coefficients of variable i, n is the number of variables.

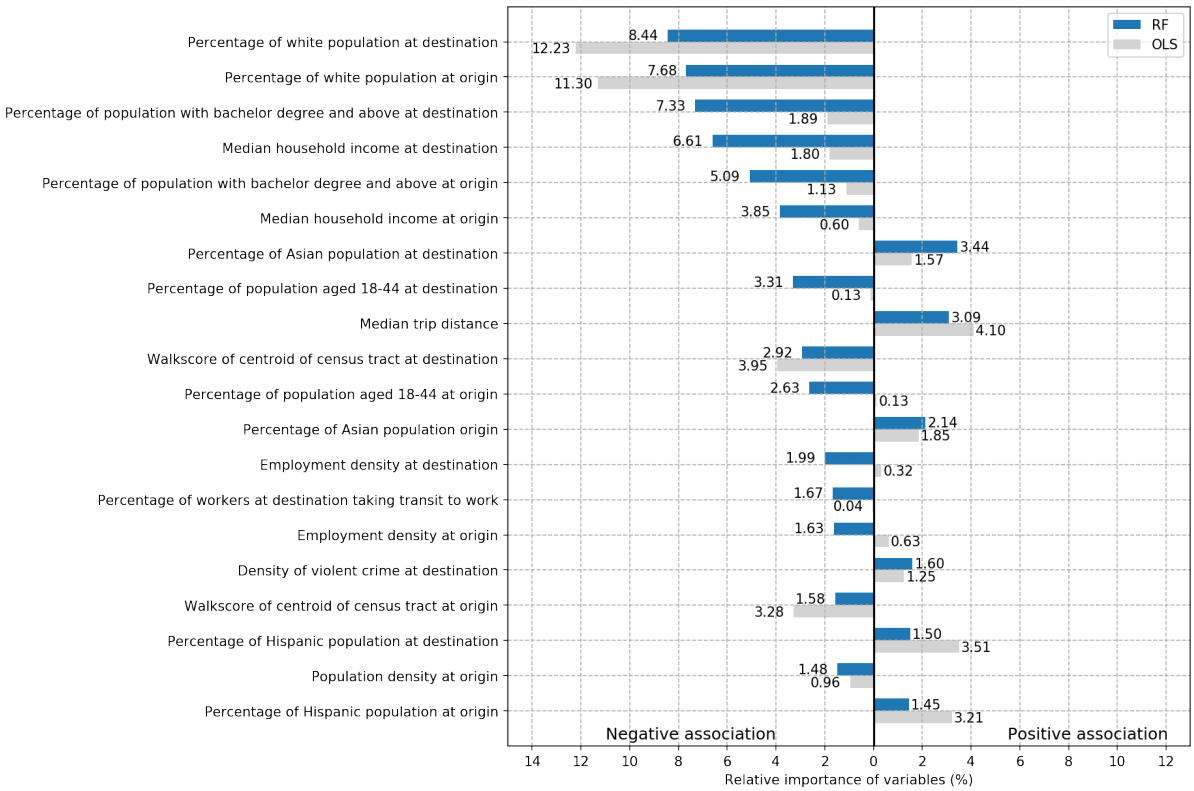

Figure 3. Relative importance of variables and direction of associations

For the RF model, the most important variable to predict ridesplitting adoption rate is percentage of white population. Other racial/ethnic variables, including percentage of Asian population and percentage of Hispanic population, also rank quite high. These results suggest that the racial/ethnic composition of a neighborhood is strongly associated with the ridesplitting adoption rate, which may be because ridesplitting is a more popular concept among certain racial/ethnic groups than others. For example, previous research has suggested that individuals living in non-white-majority communities, especially immigrant neighborhoods, use carpool as a frequent mode of travel (Blumenberg and Smart, 2014). The variables that followed are percentage of population with bachelor's degree and above, median household income, and percentage of population aged 18-44, suggesting that education level, household income, and age are important factors that are related to individuals' willingness to share rides. A trip-cost-related variable, median trip distance, ranks 9th with a relative importance value of 3.09%. Previous studies on carpooling studies have also shown that trip distance matters: long-distance commuters are more likely to carpool (e.g., Tahmasseby et al., 2016). Finally, some built-environment variables such as Walk Score, employment density, and population density are also of great importance. We believe that these variables shape ridesplitting adoption by influencing the money and time costs of a ridesourcing trip. As we mentioned above, TNCs determine the discount for ridesplitting trips based on the likelihood of a trip being matched, and trips happen in denser neighborhoods are more likely to be successfully matched (Perea, 2016). On the other hand, dense environments are often more congested, which can decrease people's willingness to share rides as congestion causes greater uncertainty in travel times.

The results of the two models (RF and OLS) have similarities and notable differences. Consistent with the RF model, several race/ethnicity-related variables (percentage of white population and percentage of Hispanic population), median trip distance, and Walk Score of centroid of Census Tract are also of top importance in the OLS model. But the relative

importance values of these variables are much greater in the OLS model than those in the RF model. Moreover, compared to RF, the contribution of median household income and percentage of population with bachelor's degree and above to the predictive power of the OLS model is much smaller. Finally, the relative importance values appear to be more evenly distributed among the variables for the OLS model than for the RF model.

In Table 4, we present the relative importance of variables by variable category. For the RF model, the sum of relative importance values for trip-cost variables, socioeconomic and demographic characteristics, and neighborhood-environment factors is 4.56%, 68.60%, and 26.84%, respectively. These values are reasonably close to those of the OLS model. We further average the relative importance of variables within each category to assess the contribution of individual variables to outcome prediction. Both models suggest that the socioeconomic and demographic variables contributed the most to predicting the neighborhood ridesplitting adoption rate.

Table 4. Relative importance of variables by category in the RF and OLS model

	RF model			OLS model		
Category	Importance Sum (%)	Variable Count	Importance Avg. (%)	Importance Sum (%)	Variable Count	Importance Avg. (%)
Travel cost	4.56	3	1.52	6.78	3	2.26
Socioeconomic and demographic	68.60	27	2.54	63.94	27	2.37
Neighborhood environment	26.84	24	1.12	29.28	24	1.22
Total	100.00	54	1.85	100.00	54	1.85

4.3.2 Direction of association, nonlinear relationships, and threshold effects

In Figure 4, we show the PDPs and ALE plots developed from the RF model and the PDPs for the OLS model. Note that for the OLS model, the slope of the PDPs is equivalent to the magnitude of the OLS coefficient estimates. Also note that the relationships revealed by these plots are correlations, and one should not readily conclude causality from them.

For variables that were measured at both the trip origin and destination, the results are quite similar. In other words, there is no significant difference between the association of these variables with the ridesplitting adoption rate at trip starts and trip ends, and so we present the results at the trip origin only. Figure 4 shows that the ALE plots largely overlap with the PDPs. As we have discussed above, the ALE plot is an unbiased alternative to PDPs, as ALE plots are less vulnerable to variable correlations. These results suggest that if there is any bias introduced to the PDPs by variable correlation, the bias is minimal. Accordingly, the discussion below focuses on the PDPs only. Finally, we observe that the directions of association between the independent variables and the outcome variable estimated by the RF model and by the OLS model are largely consistent. On the other hand, the PDPs for the RF models often display nonlinear patterns (the PDPs for the OLS model are all linear lines), which suggest that the RF model has a much more complex and flexible modeling structure than the OLS model. This can be the main reason that the RF model has achieved much higher predictive accuracy and better model fit than the OLS model.

We now interpret the PDPs for the RF model. We first infer the direction of associations by examining how the outcome variable (i.e., ridesourcing adoption rate) changes with increases in an independent variable. At the bottom of each plot, we show the distribution of each independent variable: a higher density indicates more data points; and in the ranges where the

data points are sparse, the results are less robust and thus should be interpreted with caution. We further examine the PDPs to see if there are salient nonlinear patterns and threshold effects.

The following variables are positively associated with the ridesplitting adoption rate: percentage of Asian population, percentage of Hispanic population, and density of violent crimes. Some previous studies have also shown that people traveling from or to neighborhoods with a larger share of minority population are more willing to share rides with others (e.g., Brown, 2020). However, it is surprising that the density of violent crimes is positively associated ridesplitting rate, but this association is rather weak. This unexpected result may be caused by omitted variable bias. If anything, it suggests that higher rates of violent crime in some areas don't seem to deter ridesplitting.

Ridesplitting adoption rate is negatively associated with percentage of white population, median household income, and percentage of population with bachelor's degree and above. In other words, everything else being equal, neighborhoods with a lower proportion of white population, a lower median household income, and a lower proportion of college graduates are associated with a higher level of ridesplitting adoption rate. We believe that individual income level is a major intervening factor underlying these associations: people with a higher income tend to be less willing to share rides with others (Lavieri and Bhat, 2019), and white people, college-educated individuals, and people living in neighborhoods with a higher median household income tend to have higher income. In other words, we hypothesize that even though median household income (for a Census Tract) is included as an independent variable in the models, it does not completely capture the effects of riders' income level on ridesplitting adoption. Median household income is an aggregate measure, which has much less variation than ridesourcing users' income levels. Moreover, the median household income variable only accounts for residents of a neighborhood, but ridesourcing users often come from other neighborhoods.

The associations between the built-environment variables (i.e., Walk Score, population density, and employment) and ridesplitting adoption rate appear to be rather weak. As shown in Figures 4(h), 4(i), and 4(j), the direction of these associations differs at different value ranges of each variable. At value ranges, the curves are flat, which indicates negligible impacts of each independent variable on the outcome variable. These results are somewhat consistent with those of Brown (2020), which showed that population density and employment density had opposite signs at the trip origin and the trip destination when these variables are used to predict if a ridesourcing trip is shared.

We further observe some nonlinear relationships between the independent variables and the ridesplitting adoption rate. The percentage of white population had a negative correlation with ridesplitting adoption rate in general, and this negative relationship was particularly salient at the value range between 50% and 70% (the two curves in Figure 4(a) had a sharp downward slope in this value range). This means that ridesplitting adoption rate increased rapidly as trips changed from happening in white-majority neighborhoods to being in non-white-majority neighborhoods. A possible explanation for this finding is that people who embrace a more racially diverse environment (e.g., whites choosing to live in a mixed-race rather than a white-majority neighborhood) are also more willing to share rides. The variable median household income also has a nonlinear association with the ridesplitting adoption rate (Figure 4(c)). Notably, the ridesplitting adoption rate decreases significantly when median household income increases from $50,000 to $70,000, but barely decreases after median household income reaches $90,000. This may be attributable to a threshold effect: income is negatively associated with ridesplitting adoption up to a threshold; that is, individuals living in high-income and very-high-income neighborhoods are equally reluctant to share rides with others.

According to Figure 4(h), Walk Score appears to be unrelated to ridesplitting adoption rate until it reached a score of 80, and ridesplitting adoption rate decreased as Walk Score increases from 80 to 100. We explored the Walk Score value of neighborhoods in Chicago and found that the most walkable neighborhoods (i.e., neighborhoods with a Walk Score above 80) concentrated in the North Side and Central Chicago. A plausible reason underlying the negative association between Walk Score and ridesplitting adoption rate is the high levels of congestion in these areas, which increases the uncertainty in travel times; as previous research has shown, uncertain travel time is a major deterrent to ridesplitting (Sarriera et al., 2017). Moreover, pick-ups and drop-offs are likely to be more difficult in these dense areas, which may also make riders be less willing to use ridesplitting (Morris et al., 2019). In addition, Figure 4(e) reveals a nonlinear relationship between median trip distance and ridesplitting adoption rate: the ridesplitting adoption rate increases significantly when the median trip distance is below 12 miles, but the increases become insignificant afterward. These results suggest that riders in long-distance trips are more likely to use ridesplitting, but there is a distance threshold: when the trip distance exceeds 12 miles, the effects of trip distance on ridesplitting adoption rate are negligible.

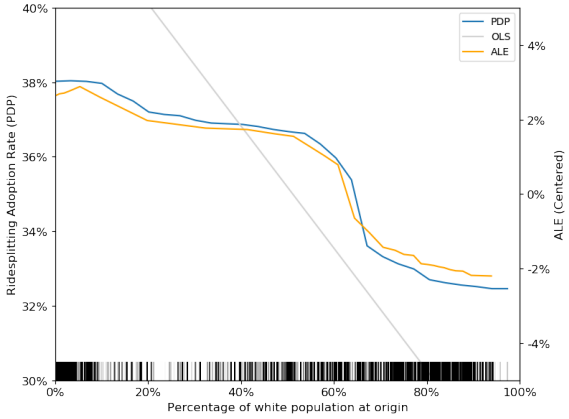

Figure 4(a) Percentage of white population

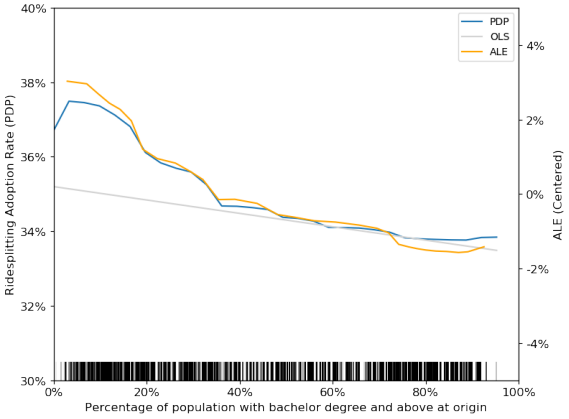

Figure 4(b) Percentage of population with bachelor's degree and above

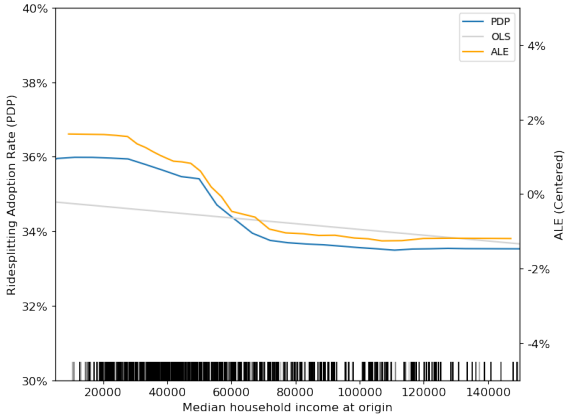

Figure 4(c) Median household income

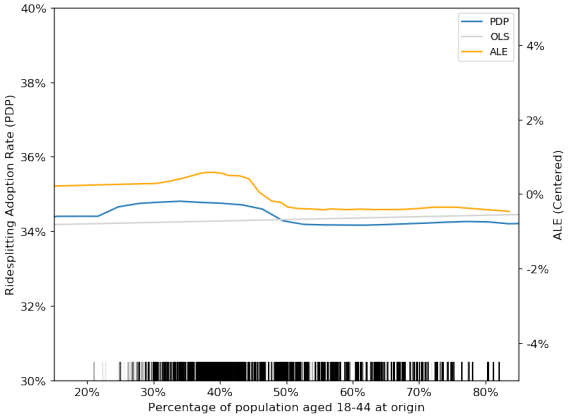

Figure 4(d) Percentage of population aged 18-44

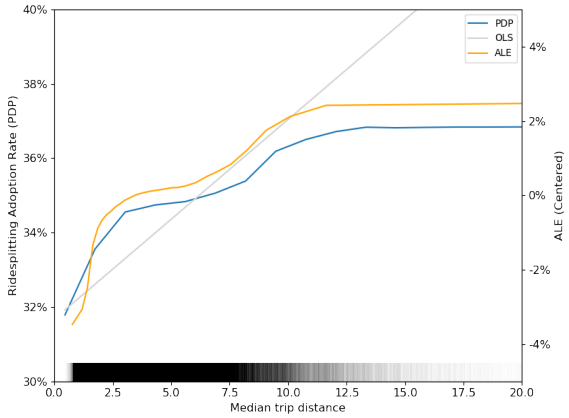

Figure 4(e) Median trip distance

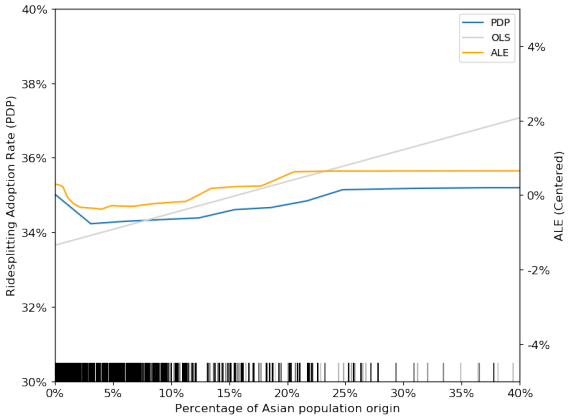

Figure 4(f) Percentage of Asian population

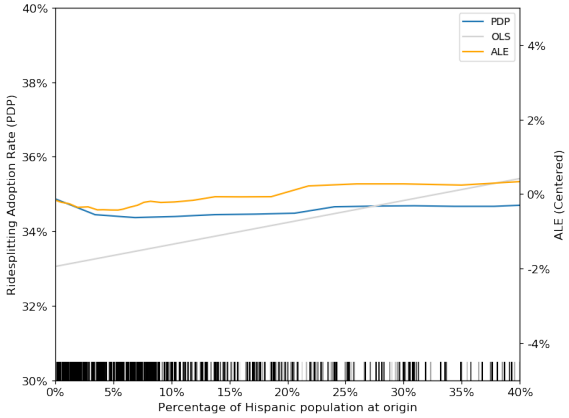

Figure 4(g) Percentage of Hispanic population

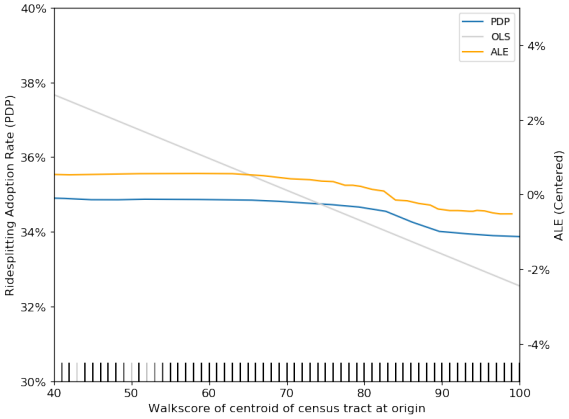

Figure 4(h) Walk Score of centroid of Census Tract

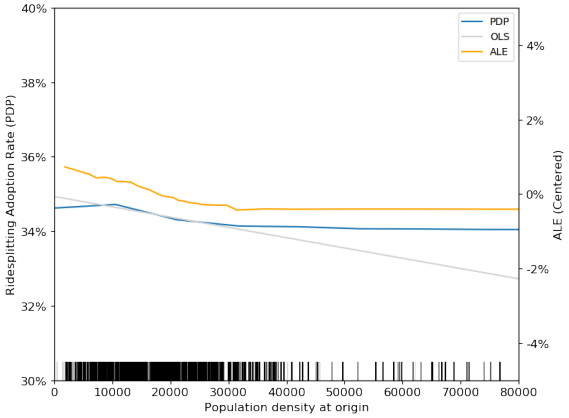

Figure 4(i) Population density

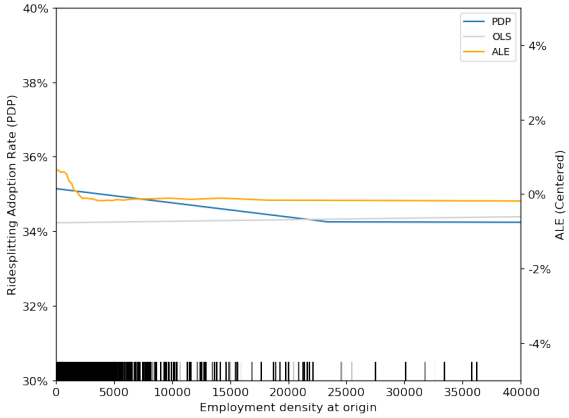

Figure 4(j) Employment density

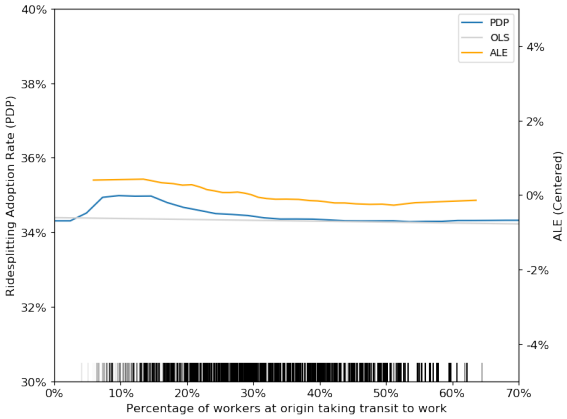

Figure 4(k) Percentage of workers taking transit to work

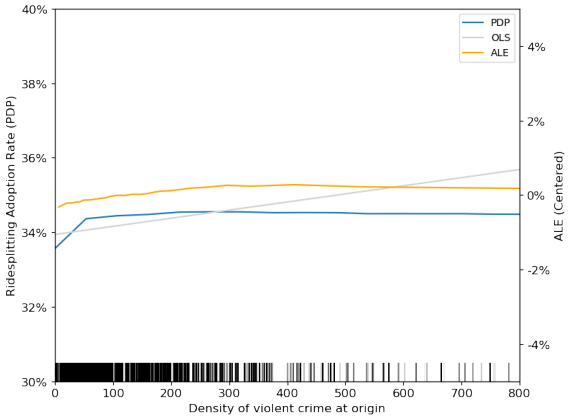

Figure 4(l) Density of violent crime

Figure 4. PDPs and ALEs plots of selected key variables

5. Discussions

Ridesplitting adoption rate (i.e., the proportion of trips that are authorized for ridesharing) varies considerably across neighborhoods in the City of Chicago. Nearly half of the trips happening in the West Side and the Southwest Side are authorized for ridesplitting, compared to less than a quarter of trips in the North Side and Central Chicago. The North Side and Central Chicago, however, generate a much large volume of trips than many other neighborhoods in the city. Therefore, if policymakers in the city wish to promote greater adoption of ridesplitting, they should target at trips happening in these areas. Pricing is a possible strategy to incentivize riders to choose ridesplitting over unshared ridesourcing services. As the City of Chicago has already implemented congestion pricing policies in the downtown area that impose a higher tax on unshared trips than on shared trips during peak times ($1.25 versus $3.00), future research may examine the impacts of these taxes on ridesplitting adoption. If this pricing strategy is proven effective, the city may consider implementing similar measures at other locations such as the North Side and the two airports. Notably, while airport trips account for 2,346,299 of all ridesourcing trips, only 9% of them are ridesplitting trips. Currently, a $5 Special Zone charge is imposed on all airport trips; to promote ridesplitting, the city may consider charging more on unshared trips and less on shared trips.

In addition to pricing schemes, transportation officials may consider other measures that can facilitate ridesplitting. For instance, the city may consider improving street design and better manage the curbside to make pick-ups and drop-offs easier. As previous research has shown, difficult pick-ups and drop-offs are a major deterrent for ridesplitting (Morris et al., 2020). We have partially attributed the low ridesplitting adoption rate in the North Side and in Central Chicago to the difficulties of pick-ups and drop-offs, and future research should verify this hypothesis.

In general, we find that Census Tracts with a higher white population, a higher median household income, and a higher proportion of people with a college degree are associated with a lower ridesplitting adoption rate. Previous studies on early adopters of ridesourcing services have found the opposite direction of associations to be true (Clewlow and Mishra, 2017). Considered together and despite potential omitted variable bias, these findings imply that a large proportion of current ridesourcing users only take unshared trips. As Brown (2020) suggests, in Los Angeles, the top 10 percent of Lyft users who adopt ridesplitting made 94

percent of Lyft Shared trips. Therefore, to promote ridesplitting, a primary focus should be on the ridesourcing users who do not currently share rides (Brown, 2020).

The preliminary results on nonlinear relationships and threshold effects can help policymakers identify which neighborhoods to target to better promote ridesplitting. For example, the ridesplitting adoption rate barely decreases after median household income reaches an upper threshold, and the rate of decrease is greatest in the value between $50,000 and $70,000. Similarly, we find that the ridesplitting rate rises rapidly as trips changed from happening in white-majority neighborhoods to non-white-majority neighborhoods. If future studies can confirm this relationship to be causal, it suggests that strategies that target middle-income and racially diverse neighborhoods would be more effective to promote ridesplitting. These strategies can be further strengthened by prioritizing the efforts on ridesourcing users who do not share rides.

6. Conclusion

This study presents a study of the ridesplitting adoption rate across O-D (Census-Tract-to-Census-Tract) pairs in the city of Chicago based on the recently released ridesourcing-trip data. By aggregating the data at the Census Tract level (as trip origins and trip destinations), we find that ridesplitting adoption rate varies greatly across Census Tracts. Based on the level of ridesplitting adoption rate, the city can be roughly divided into three distinct zones: Northern and Central (low), West and Southwest (high), and South and Far South (medium). Also, a low percentage of airport rides, which is in large volume, have adopted ridesplitting. An analysis of the ridesplitting adoption rate across O-D pairs further suggests that trips from/to Central Chicago (the Loop and the Midtown) have the highest ridesplitting adoption rate.

A RF model is applied to further explore which factors are key determinants of ridesplitting adoption rate across O-D pairs and to explore their nonlinear associations. The results suggest that the socioeconomic and demographic variables collectively contribute to 68.60% of the predictive power of the model, but travel-cost variables and neighborhood environment (i.e., built environment, safety, and transit services) are also important. We find that the most important variables associated with ridesplitting adoption are ethnic composition, income and education level, trip distance, and neighborhood density. The RF model further reveals a nonlinear association between ridesplitting adoption rate and several key variables, including percentage of white population, median household income, and neighborhood Walk Score. These nonlinear patterns allow transportation professionals to identify neighborhoods of the greatest potential impact if they are to develop strategies to promote ridesplitting.

It is worth noting that, given the nature of the data used, this study analyzes ridesplitting adoption at an aggregate level (at the Census-Tract level and an O-D pair level). The study of travel behavior based on aggregate-level data is less preferable than based on disaggregate travel behavior/preference data because the causal inference is usually more convincing at the disaggregate level. However, we shall note that an aggregate-level analysis, such as the one presented in this study, has its advantages. Notably, the cost of data acquisition is much lower, and there is little concern for the issue of sampling bias (due to the insufficient sample size or the lack of representativeness). In addition, the Chicago ridesourcing-trip data, while inappropriate for disaggregate-level travel-behavior analysis, allows researchers to readily examine the spatial patterns associated with ridesplitting adoption.

Several issues require future research. Firstly, we mainly focused on applying and interpreting RF model in this paper, and future research could explore other ML methods such as gradient boosting trees and support vector machine. For example, future research could interpret other ML models to validate the nonlinearities captured by the RF model. Secondly,

more variables such as congestion levels may be needed to develop a more comprehensive model and to generate richer insights. Thirdly, we have not examined if and to what degree ridesplitting adoption differs between peak hours and non-peak hours and between weekdays and weekends. As traveler needs and traffic conditions differ significantly across these time periods, future research should pay attention to these differences. Finally, the results and insights regarding the ridesplitting adoption rate found in Chicago may not be directly transferable to other cities with distinctive characteristics. Therefore, transferability requires further research in the future.

Acknowledgments

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 No. 69A3551747104).

Appendix A. Final variable list with VIF values

Variable	VIF
Median trip distance	3.27
Median fare for unshared trips minus shared trips	6.32
Median fare for shared trips divided by unshared trips	3.64
Percentage of male population at origin	2.96
Percentage of male population at destination	3.04
Percentage of population with bachelor's degree and above at origin*	13.61
Percentage of population with bachelor's degree and above at destination*	13.21
Percentage of population aged 18-44 at origin	4.76
Percentage of population aged 18-44 at destination	4.71
Percentage of white population at origin	6.10
Percentage of white population at destination	5.92
Percentage of Hispanic population at origin	3.68
Percentage of Hispanic population at destination	3.51
Percentage of Asian population at origin	1.67
Percentage of Asian population at destination	1.69
Percentage of households with at least one car at origin	5.15
Percentage of households with at least one car at destination	5.22
Percentage of workers taking transit to work at origin	2.56
Percentage of workers taking transit to work at destination	2.58
Median household income at origin	7.89
Median household income at destination	7.81
Percentage of renter-occupied housing units at origin	3.51
Percentage of renter-occupied housing units at destination	3.48
Percentage of single-family homes at origin	3.24
Percentage of single-family homes at destination	3.17
Percentage of workers with earnings $3,333/month or less at origin	2.64

Percentage of workers with earnings $3,333/month or less at destination	2.66
Percentage of workers with bachelor's degree and above at origin	2.35
Percentage of workers with bachelor's degree and above at destination	2.36
Population density at origin	2.29
Population density at destination	2.28
Employment density at origin	2.91
Employment density at destination	2.97
Retail employment density at origin	1.44
Retail employment density at destination	1.43
Density of violent crime at origin	2.28
Density of violent crime at destination	2.28
Road network density at origin	3.22
Road network density at destination	3.31
Intersection density at origin	3.11
Intersection density at destination	3.16
Walk Score of centroid of Census Tract at origin	3.40
Walk Score of centroid of Census Tract at destination	3.39
Aggregate service hours for rail routes at origin	3.02
Aggregate service hours for rail routes at destination	3.05
Bus stop densityat origin	2.67
Bus stop densityat destination	2.73
Rail stop densityat origin	4.21
Rail stop densityat destination	4.22
Percentage of tract within 1/4 mile of a bus stop at origin	3.14
Percentage of tract within 1/4 mile of a bus stop at destination	3.18
Percentage of tract within 1/4 mile of a rail stop at origin	4.19
Percentage of tract within 1/4 mile of a rail stop at destination	4.16

Note: *Although the VIF values of these two variables exceed the threshold of 10, we included them in the model due to their theoretical importance. We have fit a separate model that excluded these two variables, and we found that the results did not change.

Appendix B. Distributions of ridesplitting authorized and unauthorized trips

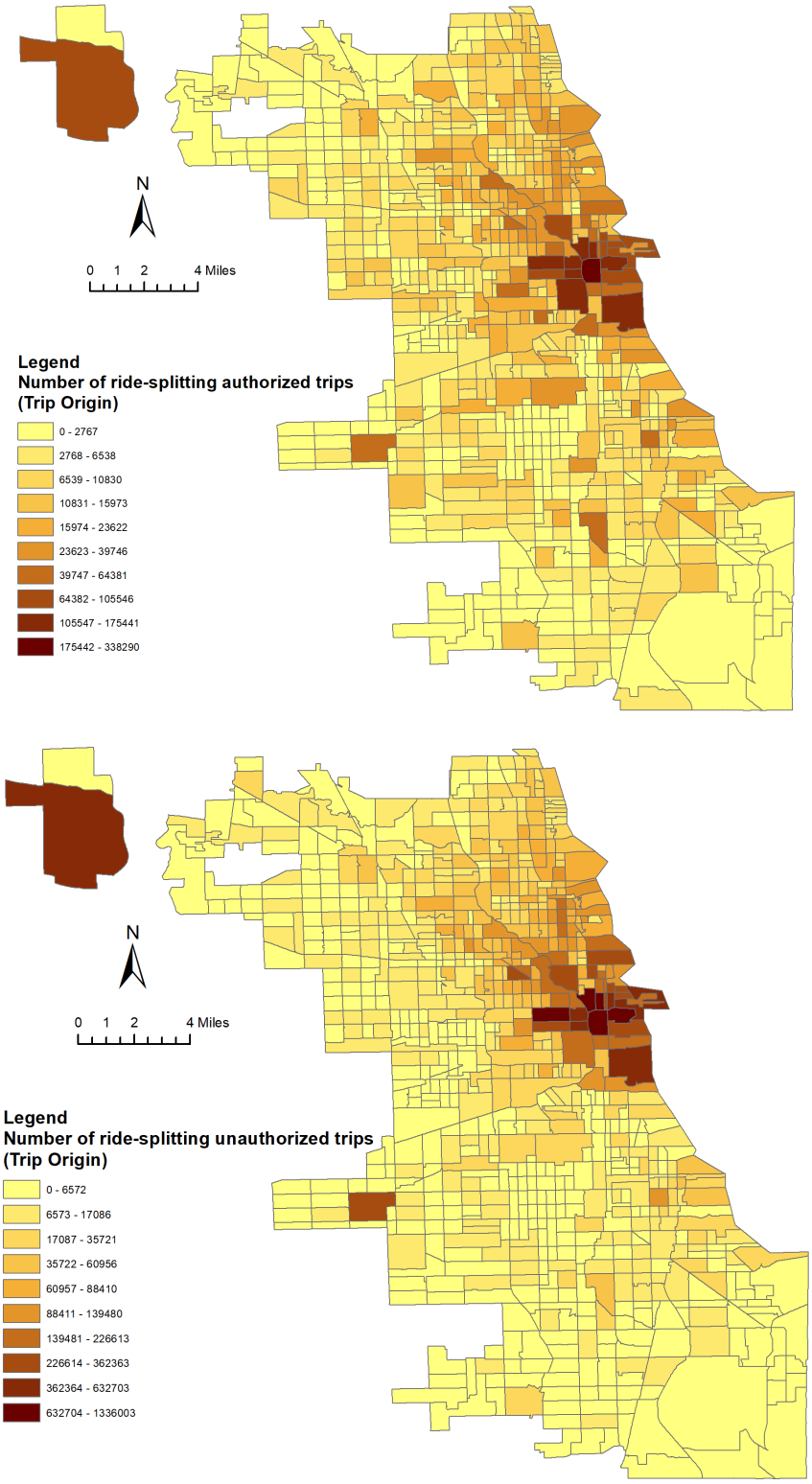

Figure B1. Number of ridesplitting authorized (top) and unauthorized (bottom) trips for each Census Tract as trip origin

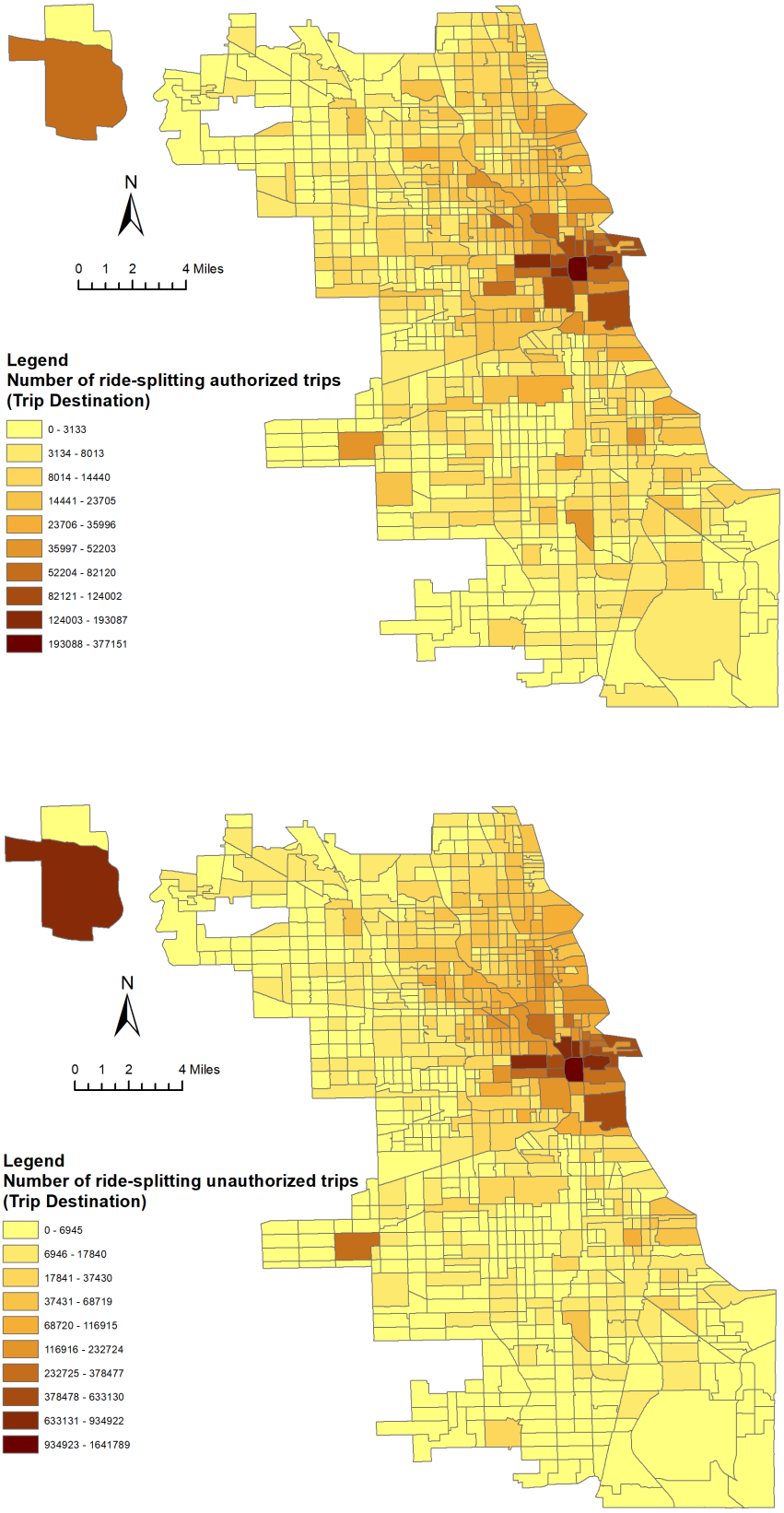

Figure B2. Number of ridesplitting authorized (top) and unauthorized (bottom) trips for each Census Tract as trip destination

References

Alemi, F., Circella, G., Handy, S. and Mokhtarian, P., 2018. What influences travelers to use Uber? Exploring the factors affecting the adoption of on-demand ride services in California. Travel Behaviour and Society, 13, pp.88-104.

Alin, A., 2010. Multicollinearity. Wiley Interdisciplinary Reviews: Computational Statistics, 2(3), pp.370-374.

Amey, A., Attanucci, J. and Mishalani, R., 2011. Real-time ridesharing: opportunities and challenges in using mobile phone technology to improve rideshare services. Transportation Research Record, 2217(1), pp.103-110.

Amirkiaee, S.Y. and Evangelopoulos, N., 2018. Why do people rideshare? An experimental study. Transportation research part F: traffic psychology and behaviour, 55, pp.9-24.

Apley, D.W. and Zhu, J., 2016. Visualizing the effects of predictor variables in black box supervised learning models. arXiv preprint arXiv:1612.08468.

Auret, L. and Aldrich, C., 2012. Interpretation of nonlinear relationships between process variables by use of random forests. Minerals Engineering, 35, pp.27-42.

Bagley, M.N. and Mokhtarian, P.L., 2002. The impact of residential neighborhood type on travel behavior: A structural equations modeling approach. The Annals of regional science, 36(2), pp.279-297.

Biau, G., 2012. Analysis of a random forests model. The Journal of Machine Learning Research, 13(1), pp.1063-1095.

Blumenberg, E. and Smart, M., 2014. Brother can you spare a ride? Carpooling in immigrant neighbourhoods. Urban Studies, 51(9), pp.1871-1890.

Breiman, L., 2001. Random forests. Machine learning, 45(1), pp.5-32.

Breiman, L., 2017. Classification and regression trees. Routledge.

Brown, A.E., 2019. Redefining Car Access: Ride-Hail Travel and Use in Los Angeles. Journal of the American Planning Association, pp.1-13.

Brown, A.E., 2020. Who and where rideshares? Rideshare travel and use in Los Angeles. Transportation Research Part A: Policy and Practice, 136, pp.120-134.

Chan, N.D. and Shaheen, S.A., 2012. Ridesharing in North America: Past, present, and future. Transport Reviews, 32(1), pp.93-112.

Chen, X.M., Zahiri, M. and Zhang, S., 2017. Understanding ridesplitting behavior of on-demand ride services: An ensemble learning approach. Transportation Research Part C: Emerging Technologies, 76, pp.51-70.

Chen, X., Zheng, H., Wang, Z. and Chen, X., 2018. Exploring impacts of on-demand ridesplitting on mobility via real-world ridesourcing data and questionnaires. Transportation, pp.1-21.

Cheng, L., Chen, X., De Vos, J., Lai, X. and Witlox, F., 2019. Applying a random forest method approach to model travel mode choice behavior. Travel behaviour and society, 14, pp.1-10.

Clewlow, R.R. and Mishra, G.S., 2017. Disruptive transportation: the adoption, utilization, and impacts of ride-hailing in the United States. Institute of Transportation Studies, University of California. Davis, Research Report UCD-ITS-RR-17-07.

de Souza Silva, L.A., de Andrade, M.O. and Maia, M.L.A., 2018. How does the ride-hailing systems demand affect individual transport regulation?. Research in Transportation Economics, 69, pp.600-606.

Delhomme, P. and Gheorghiu, A., 2016. Comparing French carpoolers and non-carpoolers: which factors contribute the most to carpooling?. Transportation Research Part D: Transport and Environment, 42, pp.1-15.

Ding, C., Cao, X. and Liu, C., 2019. How does the station-area built environment influence Metrorail ridership? Using gradient boosting decision trees to identify nonlinear thresholds. Journal of Transport Geography, 77, pp.70-78.

Ding, C., Cao, X. and Næss, P., 2018. Applying gradient boosting decision trees to examine nonlinear effects of the built environment on driving distance in Oslo. Transportation Research Part A: Policy and Practice, 110, pp.107-117.

Ding, C., Cao, X. and Wang, Y., 2018. Synergistic effects of the built environment and commuting programs on commute mode choice. Transportation Research Part A: Policy and Practice, 118, pp.104-118.

Dunnett, C.W., 1955. A multiple comparison procedure for comparing several treatments with a control. Journal of the American Statistical Association, 50(272), pp.1096-1121.

Farrar, D.E. and Glauber, R.R., 1967. Multicollinearity in regression analysis: the problem revisited. The Review of Economic and Statistics, pp.92-107.

Feigon, S. and Murphy, C., 2016. Shared mobility and the transformation of public transit (No. Project J-11, Task 21).

Ferguson, E., 1997. The rise and fall of the American carpool: 1970–1990. Transportation, 24(4), pp.349-376.

Friedman, J.H., 2001. Greedy function approximation: a gradient boosting machine. Annals of statistics, pp.1189-1232.

Galster, G.C., 2018. Nonlinear and Threshold Effects Related to Neighborhood: Implications for Planning and Policy. Journal of Planning Literature, 33(4), pp.492-508.

Golshani, N., Shabanpour, R., Mahmoudifard, S.M., Derrible, S. and Mohammadian, A., 2018. Modeling travel mode and timing decisions: Comparison of artificial neural networks and copula-based joint model. Travel Behaviour and Society, 10, pp.21-32.

Habib, K.M.N., Tian, Y. and Zaman, H., 2011. Modelling commuting mode choice with explicit consideration of carpool in the choice set formation. Transportation, 38(4), pp.587-604.

Hastie, T., Tibshirani, R. and Friedman, J., 2009. The elements of statistical learning: data mining, inference, and prediction. Springer Science & Business Media.

Ke, J., Zheng, H., Yang, H. and Chen, X.M., 2017. Short-term forecasting of passenger demand under on-demand ride services: A spatio-temporal deep learning approach. Transportation Research Part C: Emerging Technologies, 85, pp.591-608.

Kelly, K.L., 2007. Casual carpooling-enhanced. Journal of Public Transportation, 10(4), p.6.

Lavieri, P.S. and Bhat, C.R., 2019. Modeling individuals' willingness to share trips with strangers in an autonomous vehicle future. Transportation Research Part A: Policy and Practice, 124, pp.242-261.

Levin, M.W., Kockelman, K.M., Boyles, S.D. and Li, T., 2017. A general framework for modeling shared autonomous vehicles with dynamic network-loading and dynamic ride-sharing application. Computers, Environment and Urban Systems, 64, pp.373-383.

Lhéritier, A., Bocamazo, M., Delahaye, T. and Acuna-Agost, R., 2019. Airline itinerary choice modeling using machine learning. Journal of Choice Modelling, 31, pp.198-209.

Li, W., Pu, Z., Li, Y. and Ban, X.J., 2019. Characterization of ridesplitting based on observed data: A case study of Chengdu, China. Transportation Research Part C: Emerging Technologies, 100, pp.330-353.

Liaw, A. and Wiener, M., 2002. Classification and regression by randomForest. R news, 2(3), pp.18-22.

Lin, J. and Long, L., 2008. What neighborhood are you in? Empirical findings of relationships between household travel and neighborhood characteristics. Transportation, 35(6), p.739.

Mansfield, E.R. and Helms, B.P., 1982. Detecting multicollinearity. The American Statistician, 36(3a), pp.158-160.

Merlin, L.A., 2017. Comparing automated shared taxis and conventional bus transit for a small city. Journal of Public Transportation, 20(2), p.2.

Molnar, C., 2019. Interpretable machine learning. https://christophm.github.io/interpretable-ml-book/.

Morency, C., 2007. The ambivalence of ridesharing. Transportation, 34(2), pp.239-253.

Morris, E.A., Pratt, A.N., Zhou, Y., Brown, A., Khan, S.M., Derochers, J.L., Campbell, H. and Chowdhury, M., 2019. Assessing the Experience of Providers and Users of Transportation Network Company Ridesharing Services.

Neter, J., Kutner, M.H., Nachtsheim, C.J. and Wasserman, W., 1996. Applied linear statistical models (Vol. 4, p. 318). Chicago: Irwin.

O'brien, R.M., 2007. A caution regarding rules of thumb for variance inflation factors. Quality & quantity, 41(5), pp.673-690.

Perea, C. 2016. How does UberPool pricing really work? The Rideshare Guy: A Blog and Podcast for Rideshare Drivers. Retrieved from https://therideshareguy.com/how-does-uberpool-pricing-work/.

Rayle, L., Dai, D., Chan, N., Cervero, R. and Shaheen, S., 2016. Just a better taxi? A survey-based comparison of taxis, transit, and ridesourcing services in San Francisco. Transport Policy, 45, pp.168-178.

Rice, J.A., 2006. Mathematical statistics and data analysis. Cengage Learning.

Sarriera, J.M., Álvarez, G.E., Blynn, K., Alesbury, A., Scully, T. and Zhao, J., 2017. To share or not to share: Investigating the social aspects of dynamic ridesharing. Transportation Research Record, 2605(1), pp.109-117.

Shaheen, S.A., Chan, N.D. and Gaynor, T., 2016. Casual carpooling in the San Francisco Bay Area: Understanding user characteristics, behaviors, and motivations. Transport Policy, 51, pp.165-173.

Shaheen, S. and Cohen, A., 2019. Shared ride services in North America: definitions, impacts, and the future of pooling. Transport reviews, 39(4), pp.427-442.

Shaheen, S.A., Cohen, A.P., Zohdy, I.H. and Kock, B., 2016. Smartphone applications to influence travel choices: practices and policies (No. FHWA-HOP-16-023). United States. Federal Highway Administration.

Sheather, S., 2009. A modern approach to regression with R. Springer Science & Business Media.

Shi, X., Wong, Y.D., Li, M.Z.F., Palanisamy, C. and Chai, C., 2019. A variable learning approach based on XGBoost for driving assessment and risk prediction. Accident Analysis & Prevention, 129, pp.170-179.

Shieh, Y.Y. and Fouladi, R.T., 2003. The effect of multicollinearity on multilevel modeling parameter estimates and standard errors. Educational and psychological measurement, 63(6), pp.951-985.

Shmueli, G., 2010. To explain or to predict?. Statistical science, 25(3), pp.289-310.

Sperling, D., 2018. Three revolutions: steering automated, shared, and electric vehicles to a better future. Island Press, Washington, DC.

Tahmasseby, S., Kattan, L. and Barbour, B., 2016. Propensity to participate in a peer-to-peer social-network-based carpooling system. Journal of Advanced Transportation, 50(2), pp.240-254.

Tao, T., Wang, J. and Cao, X., 2020. Exploring the nonlinear associations between spatial attributes and walking distance to transit. Journal of Transport Geography, 82, p.102560.

Wager, S. and Athey, S., 2018. Estimation and inference of heterogeneous treatment effects using random forests. Journal of the American Statistical Association, 113(523), pp.1228-1242.

Wang, Y., Zheng, B. and Lim, E.P., 2018. Understanding the effects of taxi ride-sharing—A case study of Singapore. Computers, Environment and Urban Systems, 69, pp.124-132.

Yan, X., Levine, J. and Zhao, X., 2019. Integrating ridesourcing services with public transit: An evaluation of traveler responses combining revealed and stated preference data. Transportation Research Part C: Emerging Technologies, 105, pp.683-696.

Yan, X., Liu, X. and Zhao, X., 2020. Using machine learning for direct demand modeling of ridesourcing services in Chicago. Journal of Transport Geography, 83, p.102661.

Yu, H. and Peng, Z.R., 2020. The impacts of built environment on ridesourcing demand: A neighbourhood level analysis in Austin, Texas. Urban Studies, 57(1), pp.152-175.

Zhang, Y. and Haghani, A., 2015. A gradient boosting method to improve travel time prediction. Transportation Research Part C: Emerging Technologies, 58, pp.308-324.

Zhao, Q. and Hastie, T., 2019. Causal interpretations of black-box models. Journal of Business & Economic Statistics, pp.1-19.

Zhao, X., Yan, X., Yu, A. and Van Hentenryck, P., 2020. Prediction and behavioral analysis of travel mode choice: A comparison of machine learning and logit models. Travel Behaviour and Society, 20, pp.22-35.