Journal of Applied Research and Technology

www.jart.icat.unam.mx

Journal of Applied Research and Technology 21 (2023) 886-898

Original

Analysis and prediction of New York City taxi and Uber demands

D. Correa[a,b]* • C. Moyano[a]

[a]*The University of Azuay, Cuenca, Ecuador*
[b]*The University of Cuenca, Cuenca, Ecuador*

Received 09 25 2022; accepted 03 01 2023
Available 10 31 2023

Abstract: Taxi and Uber are imperative transportation modes in New York City (NYC). This paper investigates the spatiotemporal distribution of pick-ups of medallion taxis (yellow), Street Hail Livery Service taxis (green), and Uber services in NYC, within the five boroughs: Brooklyn, the Bronx, Manhattan, Queens, and Staten Island. Regression models and machine learning algorithms such as XGboost and random forest are used to predict the ridership of taxis and Uber dataset combined in NYC, given a time window of one-hour and locations within zip-code areas. The dataset consists of over 90 million trips within the period April-September 2014, yellow with 86% the most used in the city, followed by green with 9%, and Uber with 5%. In the outer boroughs, the number of pick-ups is 12.9 million (14%), while 77.9 million (86%) were made in Manhattan only. Yellow is the predominant option in Manhattan and Queens, while green is preferred in Brooklyn and Bronx. In Staten Island, the market is shared between the three services. However, Uber presents a highly rising trend of 81% in Manhattan and 145% in outer boroughs during the analysis period. The regression model XGboost performed best because of its exceptional capacity to catch complex feature dependencies. The XGboost model accomplished an estimation of 38.51 for RMSE and 0.97 for R^2. This model could present valuable insights to taxi companies, decision-makers, and city planners in responding to questions, e.g., how to situate taxis where they are required, understand how ridership shifts over time, and the total number of taxis needed to dispatch to meet de the demand.

Keywords: Large scale data analysis, GPS-enabled taxi data, machine learning algorithms, taxi and Uber demand prediction, visual analytics, New York City

*Corresponding author.
E-mail address: dcorreab@nyu.edu (D. Correa).
Peer Review under the responsibility of Universidad Nacional Autónoma de México.

1. Introduction

The taxi and ride-sourcing industries are a critical component of the transportation infrastructure, especially in large urban areas such as New York City (NYC). The taxi and ride-sourcing services industry like Uber are managed and regulated by the NYC's Taxi & Limousine Commission (TLC) (De Bilasio & Joshin, 2016). NYC has the highest number of taxis of any city in the United States (U.S., 2019). They provide passengers with innovative, high-quality mobility, comfortable, convenient, and prompt trips (Jin et al., 2019); they can be used as a compliment or a substitute to mass transit systems (Lin, et al., 2012), especially in regions where transit is less accessible.

Predicting the demand is challenging due to the extensive number of factors involved in human decision-making. Thus, pick-ups count, which is an indicator of taxi companies' productivity, has been analyzed in many studies. Despite the previous works, in this study, we use a combination of taxi and Uber pick-ups in the analysis to demonstrate the potential for combining large-scale Spatio-temporal data with two broadly used machine learning (ML) algorithms, eXtreme Gradient boosting (XGboost), and random forest, to infer the spatiotemporal demand distribution of taxi and Uber pick-ups in the city.

Inferring demand and understanding how the transportation systems have been changed and the long-term impact of this change are essential questions for taxi dispatchers, planners, and policymakers to answer, especially to evaluate the accessibility and reliability of transportation systems' effects of adverse weather. Policymakers can use this model to anticipate the effects of future spatial-temporal policies related to Uber: expansion/reduction of service coverage, and increased congestion effects due to adverse weather, among others.

This research is divided into two parts; the first part investigates the spatiotemporal distribution of trip pick-ups of Medallion taxis (yellow), Street Hail Livery Service taxis (green), and Uber services in NYC, within its five boroughs: Brooklyn (BK), the Bronx (BX), Queens (QN), Staten Island (SI) and the busiest Manhattan (MN) (De Bilasio & Joshin, 2016), where these transportation companies share the road space throughout the day (Djavadian & Chow, 2017).

The second part develops regression models (RM) to predict the ridership of taxis and Uber combined in NYC, given a window of one hour time intervals and locations within zip-code areas (TAZ) of NYC, using the data taxis and Uber combined. Exogenous factors like weather conditions, which indirectly influence the demand for taxi ridership, are also considered. This model could present valuable insights to taxi companies, decision-makers, and city planners in responding to questions, e.g., how to situate taxis where they are required, understand how ridership shifts over time, and the total number of taxis needed to dispatch to meet de the demand.

The dataset consists of over 90 million trips, generated by 13,587 yellow taxis, 7,676 green taxis, and an unknown number of Ubers from April–September 2014 (De Bilasio & Joshin, 2016). The difference between yellow, green, and Ubers is that yellow cabs choose to operate in more dense areas of NYC, such as MN and the airports. On the other hand, green cabs are not allowed to pick up passengers in street hails from the most significant part of MN (On the West Side, below 110th St., and the East Side, below 96th St.), or either of JFK or LaGuardia airports, while Ubers cannot accept street hails in NYC. The following section reviews previous studies on taxi demand and modeling

2. Literature review

An empirical study of the impact of the emerging app-based for-hire vehicles is conducted using quantitative analyses of Uber and taxi demands for the neighborhoods of Chicago, developing forecasting models for the spatial dependence of Uber and taxi trips (CHICAGO, 2015).

A study of available data of taxi and Uber data in NYC (Correa et al., 2017) investigated the impact of app-based for-hire vehicles on the taxi industry through an empirical spatiotemporal analysis (between April-September 2014 and January-June 2015). They found a high spatial correlation between taxis and Uber pick-ups, especially in central areas of NYC.

As expanding volumes of urban information are captured, new opportunities for data, and information arise. Thus, a useful data visualization tool named TaxiVis (Ferreira et al., 2013) allows users to query taxi trips by considering spatial, temporal, visual, and other constraints throughout NYC. Another study conducted using data from Taipei City showed that in 60 to 73% of their operation hours, cab drivers, drove without customers because they did not know where potential clients were, leaving them with no choice other than wandering around the city (Chang et al., 2010). Past studies (Ma et al., 2016) have shown that applying Bayesian networks to model travel mode choice behavior, for the trips based on expert prior knowledge, can explicitly estimate the causality structure between variables.

Ride-sourcing service companies match passengers and drivers online and in real-time through the so-called e-hailing process. With the arrival of smartphones, an increasing number of e-hailing applications have emerged in recent years, making communication between drivers and passengers more efficient and convenient. Ride-sourcing is transforming urban mobility by providing more flexibility, especially in large cities in North America such as New York (Correa & Moyano, 2022; Correa et al., 2021).

As a result of the shortage of information, and the little portion of ride-hailing trips within large travel reviews, explicit mode decision displaying for ride-hailing trips has been hard to build up. Utilizing taxi and other shared mobility modes, however not explicitly ride-hailing trips, (Welch et al., 2018) discovered customers of shared mobility alternatives were especially cost- sensitive, and their utilization related to non-work trip purposes.

Ride-hailing services have provided new travel options to urban residents. Therefore, efficient ridesharing solutions could improve congestion. To examine how the optimal routes, change as a function of incentives for ridesharing (Wang et al., 2016) study the effect of travel time and the toll on optimal routes by using a pick up and drop-off problem with time windows.

Recently, a study of operational performance of a shared-use automated vehicle (AV) mobility service (SAMS) fleet (Hyland et al., 2019), analyze, evaluate, and quantify the effect of demand forecast spatial resolution on SAMS in New York City. While fleet performance improved with better resolution, the forecast quality declines.

A predictive model of the number of vacant taxis in each area based on the time of day, day of the week, and weather conditions in Lisbon is presented (Phithakkitnukoon et al., 2010). Another study applied time series forecasting techniques to real-time vehicle location systems for taxis to make short-term predictions of passenger demand in the city of Porto, Portugal *(*Moreira-Matias et al., 2012*)*.

Research on modeling the variation of taxi pick-ups was developed using Poisson (Austin & Zegras, 2012) and negative binomial (Yang & Gonzales, 2017) models. These models suggest that adjacent census tracts have correlated residuals, meaning that spatial autocorrelation exists. Another paper focusing on green taxis and Uber (Korsholm et al., 2016) uses the TLC public taxi data in the outer boroughs of NYC.

From the spatial viewpoint, ridesourcing may reduce the first mile/last-mile problem, appearing as a feeder for transit. For instance, (Jin et al., 2019) studied the case of Uber in NYC, where Uber trips are higher in zones with low transit coverage, like Queens, and still high in zones with high public transit coverage like Manhattan.

A machine learning approach capturing the effects of meteorology, time of day, driving behavior, and driver experience on trip-level emissions is a study in (Xu et al., 2020). Gradient boosting, as a group of prediction models, is a machine learning technique for regression and classification (Bühlmann & Hothorn, 2007). It sets up the model in a stage-wise manner by permitting optimization of an arbitrary differentiable loss function (Svetnik et al., 2005). XGboost is a practical implementation of the gradient tree boosting model, formalized to regulate over-fitting (Chen & He, 2023). It has been shown to beat other machine learning methods significantly and consistently in terms of precision for regulated and tabular data (Robinson et al., 2017).

Travel mode detection using smartphone GPS data, comparing between random forest and wide-and-deep learning study on (Yang el at., 2019). A random forest is a group of decision trees (Ho, 1995). The training packages for random forest are typically included in most existing software for artificial intelligence (e.g., random forest classifier in Python) (Toddwschneider, 2017). Several previous studies have used the random forest for prediction such as (Tang et al., 2018; Xiong & Zhang,2013), and the model of random forest is selected as the benchmark for this research. The effect of climate conditions on travel demand has been well explored with conditions such as wind, fog, rain, snow, and weather intensity levels, influencing safety, traffic demand, and flow relationship (Maze et al., 2006).

Urban traffic patterns embody the intricate and dynamic interplay between vehicular movements and human activities within road networks. Understanding and accurately estimating these patterns can enhance traffic efficiency, safety, and sustainability in urban areas (Correa & Ozbay, 2022).

Precipitation, low temperature, and winter months have caused a mode shift from bicycle to transit and car. The opposite behavior was found to be the case for elevated temperatures (Sabir et al., 2008). Inclement weather conditions such as low temperatures and precipitation have led to reduced bicycle usage (Flynn et al., 2012; Miranda-Moreno & Nosal, 2011). Therefore, the direct response of transit ridership to variations in weather conditions is potentially overlooked given that passengers could postpone or prepone their journey in poor weather conditions are studied in (Singhal et al., 2014), another study where cabdrivers were daily income targeting (Brodeur & Nield, 2018). Using Tableau, an interactive data visualization software, an exploratory data analysis of NYC taxi demands, and the impact of weather are well described in (Gong et al., 2016).

In recent times, several studies have researched the correlation between Uber and public transport, (Tirachini & Gomez-Lobo, 2020) found that the ride-hailing services mostly increased vehicle kilometers traveled (VMT), resulted from Monte Carlo simulation based on an online data survey.

3. dataset

The study area is NYC, where most of the taxi and Uber trips are made Figure 1. For this study, geographic filter rules are established to clean the outliers from the data; therefore, trips entering and leaving NYC are ignored. For the analysis, two different datasets were combined into one. The first one consists of trip record data that was obtained from the NYCTLC, via a Freedom of Information Act request (Korsholm et al., 2016). The second one was one consisting of trip record

data of Uber pick-ups in NYC, obtained from Todd W. Schneider's by the website fivethirtyeight.com (Toddwschneider, 2017) because of a Freedom of Information Law request on July 20, 2015. Both datasets were provided as CSV files. Each row represents a ride, and each column is an attribute specific to that ride.

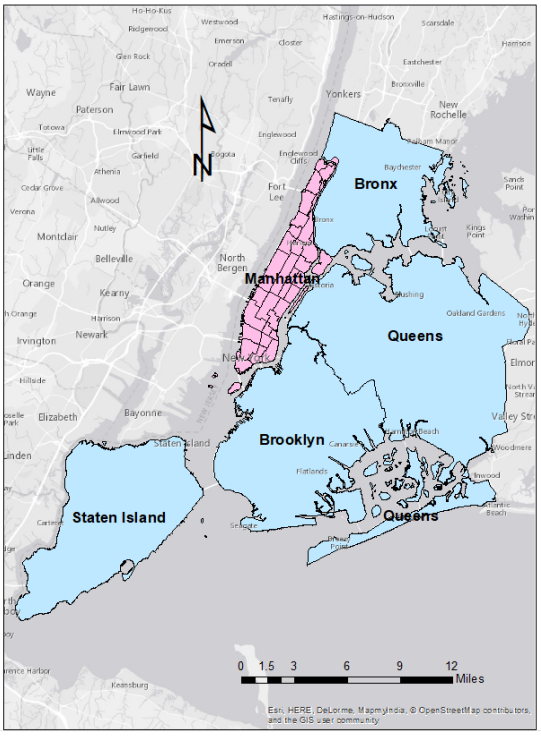


Figure 1. Study area in New York City.

3.1. Yellow and green taxis dataset

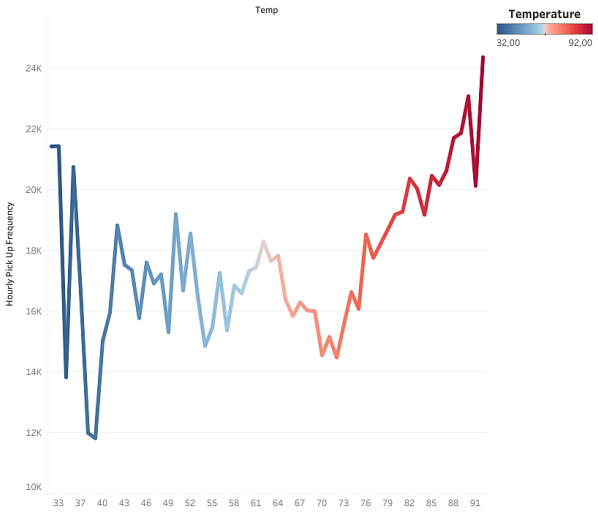
Green taxis were introduced with the proposal of providing more services to the residents of BK, QN, BX, SI, and Upper MN, where the availability of taxis tends to be minimal. Yellow cabs prefer to operate in the densest areas of the city: The Central Business District of MN (CBD), and the two airports, JFK, and LaGuardia. The data are collected using meters and GPS devices installed in all licensed taxis in the city. The trip record includes fields capturing pick up and drop-off dates/times, pick up and drop-off locations, trip distances, itemized fares, rate types, and payment types (e.g., cash, credit card).

3.2. Uber dataset

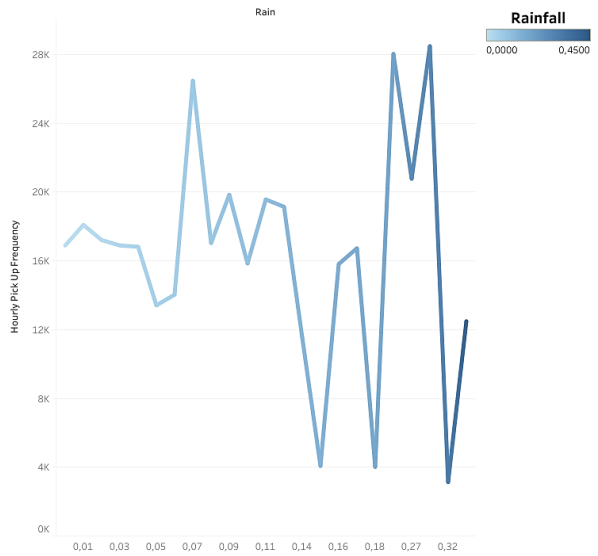
Uber's data was collected from FiveThirtyEight GitHub repo, from Todd W. Schneider's GitHub repository (Toddwschneider, 2017). Less detailed than the taxi data, the times and locations are available only for Uber pick-ups; there is also some publicly available data covering 4.4 million Uber rides in NYC from April–September 2014, that was incorporated into the dataset. The Uber data is not as detailed as the taxi data; Uber provides information related to a single Uber ride, exact GPS coordinates, time and date of pick up, no customer service information, no fare, and no drop-off information.

3.3. Weather dataset

Hourly weather information was collected from the National Centers for Environmental Information (NOAA) climate data website (NCDC, n.d.). Weather data is incorporated in the big taxi dataset. We investigate how the demand for taxis is influenced by the weather, using two features: rainfall and temperature Figure 2a, b. The temperature is measured in Fahrenheit (°F) and Hourly rain in an inch (in).



(a



(b)

Figure 2. Weather in NYC: (a) hourly taxi demand by temperature; (b) hourly taxi demand by rainfall.

The temperature has a significant impact on taxi demand, as Figure 2a shows fluctuations as the temperature becomes colder or hotter. The graph also shows the need for taxi services when the temperature goes below 36°F. On the other side, when the heat is rising, more people are willing to go to the city, which increases the demand for taxis.

From Figure 2b, the taxi demand (hourly) has a high fluctuation; as the rainfall increases, the taxi demand (hourly) may be affected, but not significantly.

4. Data processing

4.1. Visualize pick up locations

Heatmaps of taxi and Uber pick-ups are shown in Figure 3 to visualize the distribution of demand over space. These maps are made up of dots, each of them representing a single pick up location. The bright color is caused by concentrated dots and indicates higher demand activity. Figure 3a represents activities by yellow taxis, most of which are heavily concentrated in Manhattan, as well as airports and other boroughs. Although inhabitants in Manhattan account for less than 20 percent of the total population (U.S., 2019), Manhattan still presents significantly higher demand than other boroughs.

As mentioned above, green cabs are not allowed to pick up passengers in most parts of MN, as is displayed in Figure 3b. However, they can pick up passengers anywhere outside of these areas, as shown in Figure 3c, showing that Uber users are primarily interested in getting around from "Transportation Hubs" and shopping areas. However, Uber usage is spreading within MN as well as in the surrounding areas.

In Figure 3d, e, f, the order of colors indicates different hours of the day, e.g., red (midnight), yellow (4 a.m.), green (8 a.m.), cyan (noon), blue (4 p.m.), purple (8 p.m.), and back to red (since hours and colors are both cyclic). There are clearly hotspots by the hour that should be investigated. The data shows that Uber is busiest around 5 and 6 p.m., while yellow and green taxi ridership dip around that time. The CBD is active during the evening rush hour and the evening. This data could be telling that people are using Uber as a last-mile and first-mile connection to transit.

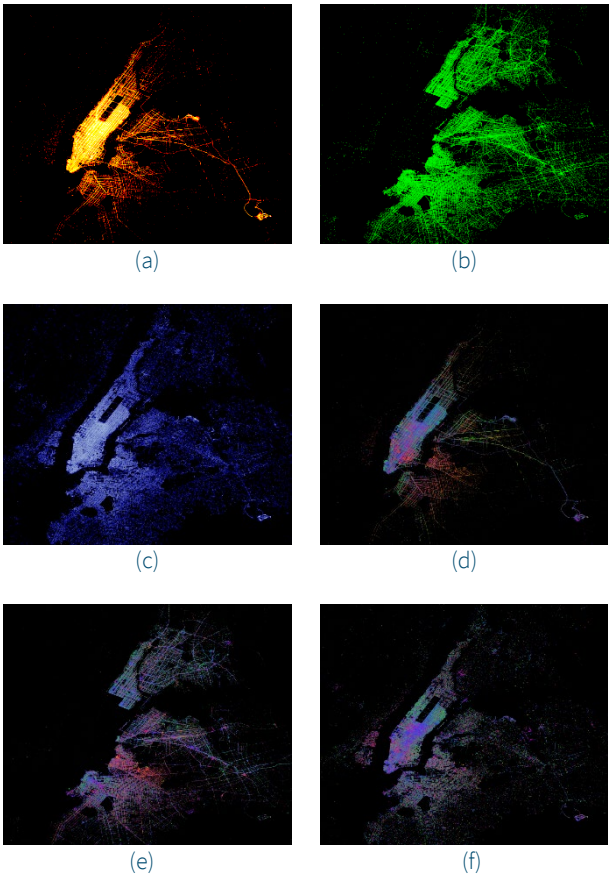


Figure 3. Spatial visualization of pick-ups in NYC, (a) yellow, (b) green,
(c) Uber and spatial visualization of pick-ups by the time-of-day, d) yellow, (e) green, (f) Uber.

4.2. Processing analysis and modeling workflow

Open-source tools such as Python and R were used to process and visualize extensive taxi data. The overall data processing sequence is shown in Figure 4.

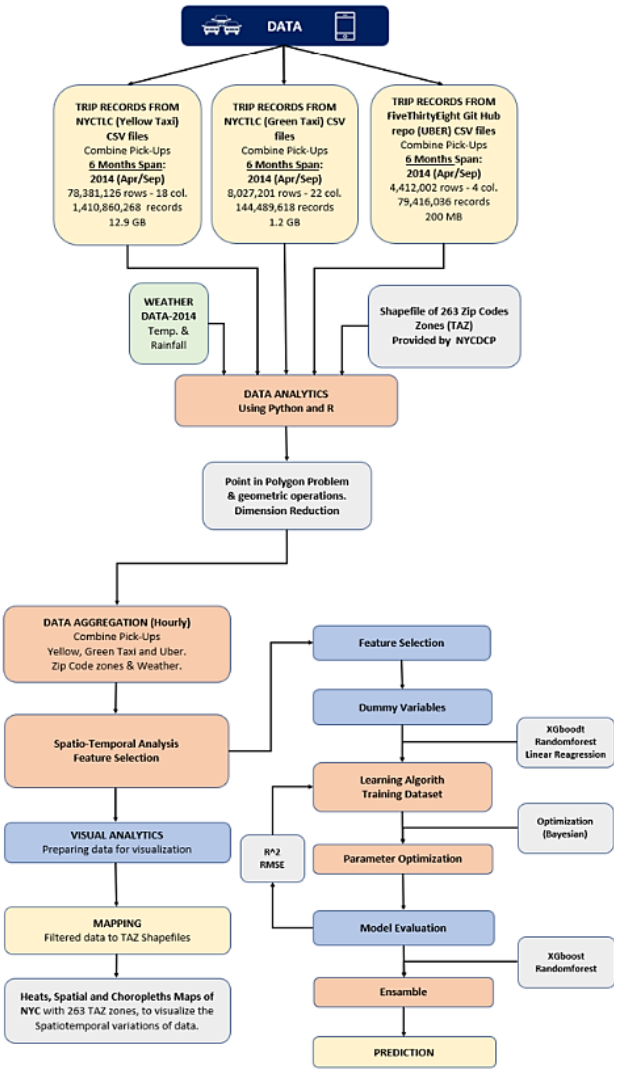


Figure 4. Flowchart of data analysis and modeling workflow.

The following steps were performed to analyze the dataset. Step 1, clean the data, remove outliers, and erroneously recorded trips, selecting only the features for the analysis. Step 2 geocode pick up and drop-off coordinates from trip record data using Python. Step 3, to constrain the problem to NYC, for this study, we only consider trips that start and end within the NYC area, as shown in Figure 1, which represents a trip. Step 4 removes excessively short trips that skew the data, i.e., shorter than a threshold, i.e., trips that were < 10 seconds long.

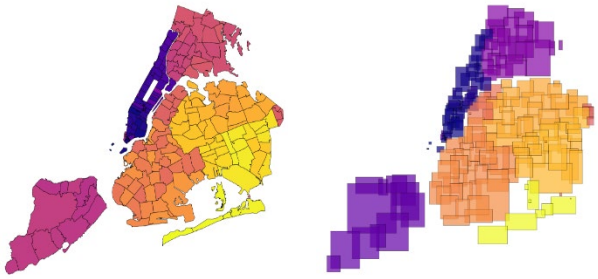


Figure 5. (a) TAZ zones Choropleth map, b) bounding boxes for each TAZ.

Figure 5a is a map of Zip Codes (TAZ) with 263 zones, colored by the id. To find the number of pick-ups (PU) and drop-offs (DO) within a TAZs, we compute and solve the point in the polygon problem. The point in a polygon is computationally expensive; therefore, to speed this up, we calculate the bounding boxes of each Zip Code. The boxes are "building block" classes, which are computed from max/min Lon/Lat points for each TAZ zone, as shown in Figure 5b, using the Python's shapely library TAZ zones are assigned to Lon/Lat pairs. Now, given a (Lon/Lat) coordinate pair, bounding boxes that contain that pair can be efficiently calculated with an R-tree. Only the polygons (TAZs) that have bounding boxes that contain the coordinate pair need to be examined; this process reduces computation time drastically.

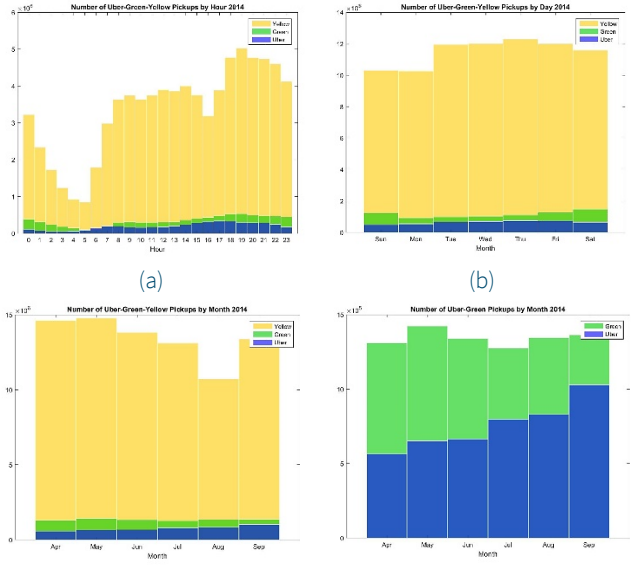
Table 1. Yellow, green, and Uber pick-ups by borough.

Boro	Month	Yellow	Green	Uber	Total	Graph
Brooklyn (BK)	Apr.	350,947	397,158	61,686	809,791	
	May	382,365	459,633	73,453	915,451	
	Jun	339,574	449,618	77,627	866,819	
	Jul	295,308	428,379	105,127	828,814	
	Aug	226,568	472,851	129,314	828,733	
	Sep	300,762	474,118	146,390	921,270	
	Total	**1,895,524**	**2,681,757**	**593,597**	**5,170,878**	
Bronx (BX)	Apr.	9,976	128,219	3,023	141,218	
	May	9,805	121,579	3,407	134,791	
	Jun	9,239	107,305	3,955	120,499	
	Jul	9,073	99,335	5,452	113,860	
	Aug	8,128	99,220	7,108	114,456	
	Sep	8,191	92,601	8,639	109,431	
	Total	**54,412**	**648,259**	**31,584**	**734,255**	
Manhattan (MN)	Apr.	13,251,420	409,175	453,567	14,114,162	
	May	13,205,943	431,358	516,642	14,153,943	
	Jun	12,324,240	396,190	516,840	13,237,270	
	Jul	11,717,859	379,980	602,003	12,699,842	
	Aug	9,525,078	388,261	594,322	10,507,661	
	Sep	12,043,609	406,970	760,188	13,210,767	
	Total	**72,068,149**	**2,411,934**	**3,443,562**	**77,923,645**	
Queens (QN)	Apr.	691,675	370,696	32,881	1,095,252	
	May	803,707	404,634	43,105	1,251,446	
	Jun	760,206	380,518	47,660	1,188,384	
	Jul	718,121	362,335	62,194	1,142,650	
	Aug	644,795	381,087	72,775	1,098,657	
	Sep	744,902	384,876	83,610	1,213,388	
	Total	**4,363,406**	**2,284,146**	**342,225**	**6,989,777**	
Staten Island (SI)	Apr.	159	302	121	582	
	May	158	106	102	366	
	Jun	147	111	116	374	
	Jul	196	71	171	438	
	Aug	138	240	229	607	
	Sep	134	275	295	704	
	Total	**932**	**1,105**	**1,034**	**3,071**	
New York City (NYC)	Apr.	14,304,177	1,305,550	551,278	16,161,005	
	May	14,401,978	1,417,310	636,709	16,455,997	
	Jun	13,433,406	1,333,742	646,198	15,413,346	
	Jul	12,740,557	1,270,100	774,947	14,785,604	
	Aug	10,404,707	1,341,659	803,748	12,550,114	
	Sep	13,097,598	1,358,840	999,122	15,455,560	
	Total	**78,382,423**	**8,027,201**	**4,412,002**	**90,821,626**	
Outer Boroughs (BK, BX, QN, SI)	Apr.	1,052,757	896,375	97,711	2,046,843	
	May	1,196,035	985,952	120,067	2,302,054	
	Jun	1,109,166	937,552	129,358	2,176,076	
	Jul	1,022,698	890,120	172,944	2,085,762	
	Aug	879,629	953,398	209,426	2,042,453	
	Sep	1,053,989	951,870	238,934	2,244,793	
	Total	**6,314,274**	**5,615,267**	**968,440**	**12,897,981**	
	% of NYC	**8.1%**	**70.0%**	**22.0%**	**14.2%**	

For each trip record, latitude and longitude were used to indicate the pick up location in the attributes. A summary of taxis and Ubers pick-ups is shown in Table 1. With more than 90 million trips, taxis and Uber are clearly imperative transportation modes in NYC, yellow with 86% the most used, followed by green taxis with 9%, and finally Uber with 5%.

In the outer boroughs of NYC, the number of taxis plus Uber pick-ups was 12.9 million, which represents 14% of the total, while most of the trips: 77.9 million (86%), were made in MN, as shown in data in Table 1Table 1. Yellow taxis are the predominant option in MN and QN, while green taxi is preferred in BK and BX. On the other hand, in SI, the market is shared between all three services from the six-month analysis period.

Uber is growing in NYC, as Figure 6d shows; there is a rising trend in the volume of pick-ups every month. Looking at the data by day of the week, people used it more during the weekdays Figure 6b shows. However, New Yorkers still do not use Uber regularly in the early morning hours and even use it heavily during the evening rush hour, as is shown in **Figure 6**a yellow and green taxis and Uber follow a similar pattern. The taxi demand tops around Thursdays and Fridays, while a decline around Sundays and Mondays. Then, it peaks between 6–8 p.m. and decreases between 4–5 a.m.



(a) (b)

(c) (d)

Figure 6. Histogram of Pick-ups in NYC in 2014:
(a) number of Uber-green-yellow pick-ups by hour; (b) number of Uber-green-yellow pick-ups by day; (c) number of Uber-green-yellow pick-ups by month; (d) number of Uber -green pick-ups by month.

Compared with taxis, the demand for Uber tends to be distributed more evenly throughout the city as shown in Figure 6d, and it is rising in the outer boroughs of NYC, a trend continued as shown in Table 1. A spatial heat map of pick-ups in NYC is shown in a) further details about the data NYC taxi GPS data can be found in (Chang et al.,2010).

Once the data have been processed and ready to be used for training and testing in the model, each row represents a combination of TAZ, weather, and the total amount of pick-ups aggregated by hour, e.g., "2014-06-01 07:00:00, 69.0, 0, 00083, 975" represents that, on June 6 of 2014, there were 975 trips taken from the TAZ 00083 (Central Park in MN) zone from 7 a.m. to 8 a.m. 69°F, and no rain.

5. Methodology

5.1. Feature selection

In the following section, we describe the procedure of feature extraction from each data point. First, TAZ, we presume the location would be a good predictor of taxi activity. Second, the hour of the day, $\in$ [0, 23] we use the entire day and are expected to be high during peak hours. Third, the day of the week, $\in$ [0, 6] traffic is expected to be correlated according to each day of the week. Four, temperature (°F), high and low temperatures are expected to increase the demand for taxis.

Fifth, precipitation (in), precipitation is expected to increase taxi ridership. Finally, the TAZ area code but treated as a categorical variable. We create 263 dummy variables according to the number of TAZ. A dummy variable is one that takes the values 0 or 1 to indicate the absence or presence of some categorical effect that may be expected to shift the outcome. If a record takes one of these values, then the corresponding dummy variable is assigned a value of 1, while all the rest dummies are assigned a 0 value.

5.2. Estimation

For proposing of evaluation, the dataset is divided into training and test sets, the first one containing 80% of the data, and the second including the remaining 20%. Before usage, both datasets are ordered chronologically.

To analyze differences between values forecasted by the model or an estimator and the values observed, we use a frequently used measure of the differences, the root mean square error (RMSE), which measures how the spread out these residuals are. RMSE favors uniformity and penalizes predictions with a high deviation from the correct number of pick-ups.

Another used estimator for comparing the results between models is the coefficient of determination (R^2) value. While R^2 is a relative measure of fit, RMSE is an absolute measure of fit. Both parameters are used to evaluate how well the models perform.

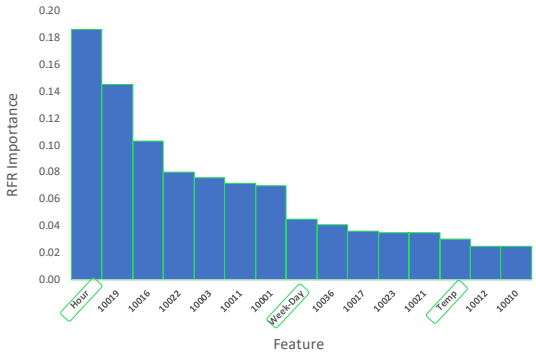
5.3. Multiple-linear regression

Multiple linear regression (MLR) attempts to model the relationship between two or more explanatory variables and a response variable by fitting a linear equation and exploiting linear patterns in the data set. For its easy implementation and efficiency on large datasets, this method is often the first choice. The results are shown in Table 1a.

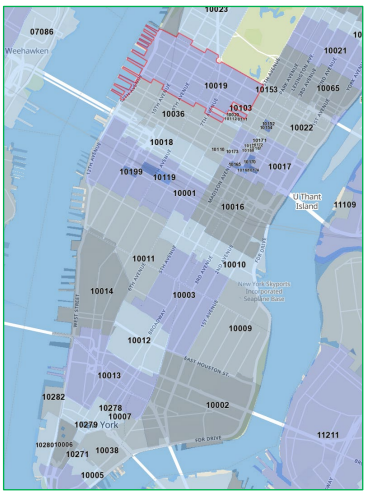
5.4. The random forest regression model

Random forest is a simple and flexible machine learning algorithm that, most of the time, produces excellent results with minimum time spent on parameter tuning. It can be used for both regression and classification. Random forest prevents overfitting and is robust against outliers.

Scikit-learn package (Toddwschneider, 2017) is used, and the number of splits at each tree (parameters max-features) and the number of trees (n-estimators) are determined using Bayesian optimization (3 iterations needed to find the best max-features and n-estimators). This study performs the best with max-features 15 and n-estimators of 520. For this study, we listed the best of 15 features by importance, produced by random forest, as shown in Figure 7a.



(a)



(b)

Figure 7. (a) Random Forest top 15 features, (b) top TAZ (Zip Codes).

The Featured selection results demonstrate that the hour, weekday, and temperature are essential features that directly impact the model, as well as some TAZ. As the featured selection results show that, hour, weekday, and temperature are prominent features that directly impact the model, as well as some TAZ as 10019 (Time Square-MN). Results are shown in Figure 7b and Table 3.

5.5. XGboost

XGboost is a robust machine learning algorithm (which provides an advanced gradient boosting algorithm), especially where speed and accuracy are concerned. XGboost is a sophisticated and powerful algorithm sufficient to deal with all types of abnormalities of data, allowing them to reach the optimal solution. The XGboost model requires parameter tuning to improve and fully leverage its advantages over other algorithms.

The overall parameters of the model can be divided into three categories: boosting parameters (these affect the boosting operation in the model), tree-specific parameters (these affect each tree in the model, and miscellaneous parameters (other parameters for overall functioning).

The XGboost model implementation in this study is based on the scikit-learn and XGboost Python libraries (Pedregosa et al., 2012). For this method, adjustment is made by Bayesian optimization to find the best combination of parameters. The max-depth is the maximum depth of a tree, n_estimators is the number of trees, min_child_weight is the smallest summation of weights of all deviations required at each split node, gamma is the minimum loss reduction at each split, learning rate controls the impact of each tree on the result to avoid overfitting, subsample is the percentage of samples used per tree. A low value can lead to underfitting, and colsample_bytree is the percentage of features used per tree. High value can lead to overfitting. Parameters are listed in Table 2 and Table 3.

We use both RMSE and R^2 values to the performance of the model. RMSE is obtained using Equation 1.

$$RMSE = \sqrt{\frac{\sum_{t=1}^{T}(\hat{y}_t - y_t)^2}{T}} \quad (1)$$

Where:

- T is the sample size.
- $\hat{y}_t$: The predicted value., and yt its observed value
- y_t : The observed value.

The coefficient of determination R^2 value is obtained using Equation 2.

$$R^2 = 1 - \frac{First\ Sum\ of\ Errors}{Second\ Sum\ of\ Errors} \quad (2)$$

The results of all three models are displayed in Table 3. Overall, XGboost performs best.

Table 2. Model parameters.

Model Parameters	Default	Typical values	Results
max depth	6.0	3 - 14	12
n estimators	-	50 - 1000	523
min child weight	1.0	1 - 10	7.71
gamma	0.0	0.01 - 1.0	0.94
learning rate	0.1	0.01 - 0.2	0.16
subsample	0.8	0.5 - 1.0	0.93
colsample bytree	0.8	0.5 - 1.0	0.83
objective	reg: linear	binary: logistic	binary: logistic

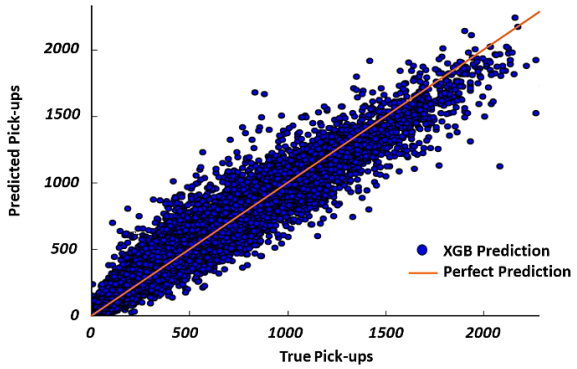
Table 3. Results of models.

Model	Train		Test	
	R^2	RMSE	R^2	RMSE
Multiple Linear Regression	0.74	122.26	0.75	120.54
Random Forest	0.97	33.71	0.95	40.46
XGBoost	0.98	32.78	0.97	38.51

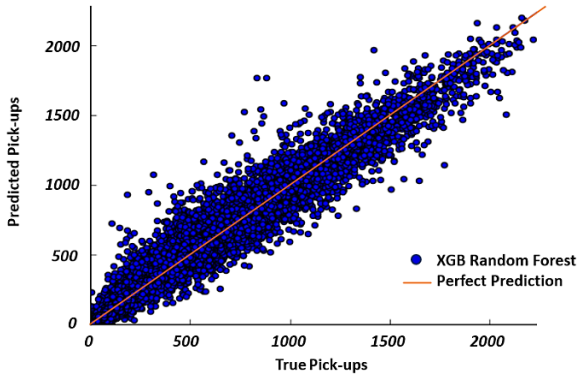
6. Model analysis

To visualize how well the models' random forest and XGboost perform, we plot the expected versus the actual number of pick-ups for each point. Results are shown in Figure 8a, b. The plots indicate that test sets of both models perform well, and overall predictions are close to the real values along the line. Hence, models do not analytically overestimate or underestimate the real number of taxi pick-ups.

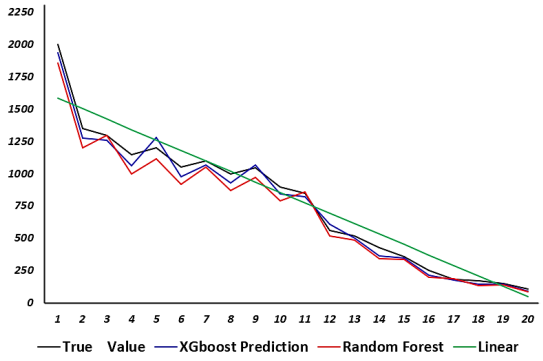
According to the plots, in both model's total prediction error increases as the real number of pick-ups increases as expected. Figure 8c shows the comparison of randomly picked ten samples between the real value and the prediction of two models.



(a)



(b)



(c)

Figure 8. (a) Predicted number of pick-ups. (a) Random forest, (b) XGboost, (c) comparison of models.

7. Discussion and future work

This paper investigates the spatiotemporal distribution of yellow & green taxis and Uber pick-ups in New York City, within the five boroughs: Brooklyn, the Bronx, Manhattan, Queens, and Staten Island: and develops a Zip Code zone (TAZ) based forecasting analysis of ride-hailing, app-based services Uber,

and taxi services, in NYC using large-scale Uber and taxi pick-ups data. Major contributions of this paper compared with previous studies are the use of Uber data and the consideration of XGboost and random forest, two popular methods for various machine learning tasks to forecast the readership of taxi and Uber in NYC., understanding the travel preferences, especially in outer boroughs.

In the empirical analysis, we explore the spatiotemporal patterns of Uber and taxi pick-ups. The demand for Uber tends to be distributed more evenly throughout the city and is preferred on Thursdays and Fridays. Customer preferences have also changed with respect to time of day, now the duration of PM Uber's peak demand is more significant than taxis. Uber is the busiest during the morning hours, while taxi rides dip around that time. Uber also is the preferred transportation mode during late nights, especially in Queens. In addition, regarding night rides, Uber's rides start inclining at 3 a.m. while green taxi rides decline from midnight to 5 a.m. Thus, Uber is more popular as a taxi service during nighttime, correspondingly.

Data for Staten Island is deficient compared to the other boroughs; however, it shows the consistent rise of using Uber over taxis. One of the most important aspects to be analyzed in depth is the rising trend Uber from April to September. Uber overpasses taxis, becoming the preferred option if the actual condition continues. Demand for green taxis is still growing in all the outer boroughs analyzed. Nevertheless, Uber rides in the same area are growing even more rapidly, especially in low-income boroughs, Ubers are performing better. Compared with taxis, the demand for Uber tends to be distributed more evenly throughout the city.

In general, the models performed well for predicting taxi pick-ups in NYC, being the best performed XGboost, because of its capacity to capture complex feature dependencies, achieving a value of 38.51 for RMSE and 0.97 for R^2. On the other hand, the Random Forest model attained a value of 40.46 for RMSE and 0.95 for R^2.

Although there are no fixed threshold values for RMSE, the smaller, the better. Comparing the RMSE of both test and train datasets, we can say that the model performs well because the RMSE of test data is like the training dataset. If the RMSE for the test set is much higher or lower than the RMSE of the training set, it is likely that the model is overfitting or underfitting the data, respectively. The RMSE of the training data is calculated using 80% of observations, and the RMSE of the test data is calculated using only 20% of observations. The test dataset RMSE gives an idea of how well the model will perform on test data.

Building a model using XGBoost is easy. However, improving the model using XGBoost is more difficult due to this algorithm using multiple parameters. To improve the model, the selection of parameters to tune and the ideal values to obtain optimal output becomes a challenging task.

From the feature selection, the analysis of the results shows that the time of the day is the most crucial factor to be considered. Zip Codes, especially around Time Square and Central Park, influence more than other features such as the day of the week and temperature. In contrast, the rainfall feature is not critical for model implementation.

The model could present valuable insights to taxi companies, decision-makers, and city planners in determining patterns in ridership and defining where to position taxicabs throughout the day. For future studies, the spatial lag model will be integrated with other techniques such as Bayesian networks. Improved versions of this spatial model will provide alternative ways of calculating.

To improve the model, in future studies, other techniques should be implemented like Neural network regression (NN) and K-means Clustering (K-means). The learning algorithm can automatically determine and model feature connections, rather than manually determining which features to combine. K-means could be implemented to find hidden patterns across the spatial distribution of data points. After the creation of some clusters, it could then fill in as an extra feature for the regression models.

Since Uber and taxi services are competing on travel demands in the real world, the inter-relationship between Uber and taxi demands should be considered in separate models. It is also possible to model the interaction between taxi demand and the demand for all other alternative transportation modes such as bike-sharing and other sharing services and new ride-sharing services. However, obtaining data for other ride-sharing services remains to be a major challenge for studies such as this one. Considerations like if Uber is being used for first/last mile problems as well as pick-ups/drop-offs clustered near subway stations in the outer boroughs will be the focus of future research.

Conflict of interest

The authors have no conflict of interest to declare.

Funding

The Secretariat of Higher Education, Science, Technology, and Innovation of Ecuador SENESCYT partially supported this research.

References

Austin, D., & Zegras, P. C. (2012). Taxicabs as public transportation in Boston, Massachusetts. *Transportation research record*, *2277*(1), 65-74. https://doi.org/10.3141/2277-0

Brodeur, A., & Nield, K. (2018). An empirical analysis of taxi, Lyft and Uber rides: Evidence from weather shocks in NYC. *Journal of Economic Behavior & Organization*, *152*, 1-16. https://doi.org/10.1016/j.jebo.2018.06.004

Bühlmann, P. & Hothorn, T. (2007). Boosting Algorithms: Regularization, Prediction and Model Fitting. Statistical Science, Vol. 22, No. 4, 2007, pp. 477–505. https://doi.org/10.1214/07-STS242

Chang, H. W., Tai, Y. C., & Hsu, J. Y. J. (2010). Context-aware taxi demand hotspots prediction. *International Journal of Business Intelligence and Data Mining*, *5*(1), 3-18. https://doi.org/10.1504/IJBIDM.2010.030296

Chen, T., & He, T. (2023). Xgboost: EXtreme Gradient Boosting. https://cran.r-project.org/web/packages/xgboost/vignettes/xgboost.pdf

CHICAGO: An Uber Case Study. (2015). Uber, Chicago case study. https://uber-static.s3.amazonaws.com/web-fresh/legal/Uber_Chicago_CaseStudy.pdf

Correa, D., & Ozbay, K. (2022). Urban path travel time estimation using GPS trajectories from high-sampling-rate ridesourcing services. *Journal of Intelligent Transportation Systems*, 1-16. https://doi.org/10.1080/15472450.2022.2124867

Correa, D., & Moyano, C. (2022). Dynamics of the Growth of UBRER vs Green Taxis in Outer Boroughs in New York City. *Available at SSRN 4229008*. http://dx.doi.org/10.2139/ssrn.4229008

Correa, D., Chow, J. Y., & Ozbay, K. (2021). Spatial-dynamic matching equilibrium models of New York City Taxi and Uber markets. *Journal of Transportation Engineering, Part A: Systems*, *147*(9), 04021048. https://doi.org/10.1061/JTEPBS.0000550

Correa, D., Xie, K. & Ozbay, K. (2017). Exploring the taxi and Uber demands in New York City: An empirical analysis and spatial modeling. *Available at SSRN 4229042*. http://dx.doi.org/10.2139/ssrn.4229042

De Bilasio, B. & Joshi, M. (2016). TLC Factbook. TLC trip record data. NYC. https://www.nyc.gov/assets/tlc/downloads/pdf/2016_tlc_factbook.pdf

Djavadian, S., & Chow, J. Y. (2017). An agent-based day-to-day adjustment process for modeling 'Mobility as a Service'with a two-sided flexible transport market. *Transportation research part B: methodological*, *104*, 36-57. https://doi.org/10.1016/j.trb.2017.06.015

Ferreira, N., Poco, J., Vo, H. T., Freire, J., & Silva, C. T. (2013). Visual exploration of big spatio-temporal urban data: A study of new york city taxi trips. *IEEE transactions on visualization and computer graphics*, *19*(12), 2149-2158. https://doi.org/10.1109/TVCG.2013.226

Flynn, B. S., Dana, G. S., Sears, J., & Aultman-Hall, L. (2012). Weather factor impacts on commuting to work by bicycle. *Preventive medicine*, *54*(2), 122-124. https://doi.org/10.1016/j.ypmed.2011.11.002

Gong, Y., Fang, B. Zhang, S. & Zhang, J. (2016). Data Study to Predict New York City Taxi Demand. *NYC DATA SCIENCE ACADEMY*. https://nycdatascience.com/blog/student-works/predict-new-york-city-taxi-demand

Ho, T. K. (1995). Random decision forests. In *Proceedings of 3rd international conference on document analysis and recognition* (Vol. 1, pp. 278-282). IEEE. https://doi.org/10.1109/ICDAR.1995.598994

Hyland, M., Dandl, F., Bogenberger, K., & Mahmassani, H. (2019). Integrating demand forecasts into the operational strategies of shared automated vehicle mobility services: spatial resolution impacts. *Transportation Letters*, *12*(10), 671-676. https://doi.org/10.1080/19427867.2019.1691297

Jin, S. T., Kong, H., & Sui, D. Z. (2019). Uber, public transit, and urban transportation equity: A case study in new york city. *The Professional Geographer*, *71*(2), 315-330. https://doi.org/10.1080/00330124.2018.1531038

Korsholm, L., Dekkers, D., Wagenaar, N., Snijders, W., Lewinsky, B., Mukkamala, R. & Vatrapu, R. (2016). Green taxis vs. Uber in New York City. *IEEE International Congress on Big Data*.

Lin, Y., Li, W., Qiu, F., & Xu, H. (2012). Research on optimization of vehicle routing problem for ride-sharing taxi. *Procedia-Social and Behavioral Sciences*, *43*, 494-502. https://doi.org/10.1016/j.sbspro.2012.04.122

Maze, T. H., Agarwal, M., & Burchett, G. (2006). Whether weather matters to traffic demand, traffic safety, and traffic operations and flow. *Transportation research record*, *1948*(1), 170-176. https://doi.org/10.1177/036119810619480011

Miranda-Moreno, L. F., & Nosal, T. (2011). Weather or not to cycle: Temporal trends and impact of weather on cycling in an urban environment. *Transportation research record*, *2247*(1), 42-52. https://doi.org/10.3141/2247-06

Moreira-Matias, L., Gama, J., Ferreira, M., & Damas, L. (2012). A predictive model for the passenger demand on a taxi network. In *2012 15th International IEEE Conference on Intelligent Transportation Systems* (pp. 1014-1019). IEEE. https://doi.org/10.1109/ITSC.2012.6338680

NOAA. National Centers for Environmental Information. (n. d.). Climate Data Online (CDO). NOAA. National Oceanic and Atmospheric Administration. https://www.ncdc.noaa.gov/cdo-web/

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay, É. (2012). Scikit-learn: Machine learning in Python. *the Journal of machine Learning research*, *12*, 2825-2830. https://doi.org/10.48550/arXiv.1201.0490

Phithakkitnukoon, S., Veloso, M., Bento, C., Biderman, A., & Ratti, C. (2010). Taxi-aware map: Identifying and predicting vacant taxis in the city. In *Ambient Intelligence: First International Joint Conference, AmI 2010, Malaga, Spain, November 10-12, 2010. Proceedings 1* (pp. 86-95). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-16917-5_9

Robinson, C., Dilkina, B., Hubbs, J., Zhang, W., Guhathakurta, S., Brown, M. A., & Pendyala, R. M. (2017). Machine learning approaches for estimating commercial building energy consumption. *Applied energy*, *208*, 889-904. https://doi.org/10.1016/j.apenergy.2017.09.060

Sabir, M., Koetse, M. J. J., & Rietveld, P. (2008). The impact of weather conditions on mode choice: empirical evidence for the Netherlands. In *Proceedings of the BIVEC-GIBET Transport Research Day 2007* (pp. 512-527).

Toddwschneider. (2017). *GitHub - toddwschneider/NYC-taxi-data: Import public NYC taxi and for-hire vehicle (Uber, Lyft) trip data into a PostgreSQL or ClickHouse database*. GitHub.https://github.com/toddwschneider/nyc-taxi-data

Singhal, A., Kamga, C., & Yazici, A. (2014). Impact of weather on urban transit ridership. *Transportation research part A: policy and practice*, *69*, 379-391. https://doi.org/10.1016/j.tra.2014.09.008

Svetnik, V., Wang, T., Tong, C., Liaw, A., Sheridan, R. P., & Song, Q. (2005). Boosting: An ensemble learning tool for compound classification and QSAR modeling. *Journal of chemical information and modeling*, *45*(3), 786-799. https://doi.org/10.1021/ci0500379

Ma, T. Y., Chow, J. Y., & Xu, J. (2016). Causal structure learning for travel mode choice using structural restrictions and model averaging algorithm. *Transportmetrica A: Transport Science*, *13*(4), 299-325. https://doi.org/10.1080/23249935.2016.1265019

Tang, L., Pan, Y., & Zhang, L. (2018). *Trip Purpose Imputation Based on Long Term GPS Data* (No. 18-05010).

Welch, T. F., Gehrke, S. R., & Widita, A. (2018). Shared-use mobility competition: a trip-level analysis of taxi, bikeshare, and transit mode choice in Washington, DC. *Transportmetrica A: transport science*, *16*(1), 43-55. https://doi.org/10.1080/23249935.2018.1523250

Tirachini, A., & Gomez-Lobo, A. (2020). Does ride-hailing increase or decrease vehicle kilometers traveled (VKT)? A simulation approach for Santiago de Chile. *International journal of sustainable transportation*, *14*(3), 187-204. https://doi.org/10.1080/15568318.2018.1539146

U.S. Census Bureau. (2019). American Fact Finder. https://www.census.gov/data.html

Wang, X., Dessouky, M., & Ordonez, F. (2016). A pickup and delivery problem for ridesharing considering congestion. *Transportation letters*, *8*(5), 259-269. https://doi.org/10.1179/1942787515Y.0000000023

Xiong, C., & Zhang, L. (2013). A Descriptive Bayesian Approach to Modeling and Calibrating Drivers' En Route Diversion Behavior. *IEEE Transactions on Intelligent Transportation Systems*, *14*(4), 1817-1824. https://doi.org/10.1109/TITS.2013.2270974

Xu, J., Saleh, M., & Hatzopoulou, M. (2020). *A machine learning approach capturing the effects of meteorology, time of day, driving behaviour, and driver experience on trip-level emissions*. *Atmospheric Environment.* https://doi.org/10.1016/j.atmosenv.2020.117311

Yang, C., & Gonzales, E. J. (2017). Modeling taxi demand and supply in New York city using large-scale taxi GPS data. *Seeing Cities Through Big Data: Research, Methods and Applications in Urban Informatics*, 405-425. https://doi.org/10.1007/978-3-319-40902-3_22

Yang, D., Xiong, C., Tang, L., & Zhang, L. (2019). *Travel mode detection using smartphone GPS data: A comparison between random forest and wide-and-deep learning*.