




Article

Understanding and Predicting Ride-Hailing Fares in Madrid: A Combination of Supervised and Unsupervised Techniques

Tulio Silveira-Santos ^{1,*} , Anestis Papanikolaou ², Thais Rangel ^{1,3}  and Jose Manuel Vassallo ¹ 

¹ Transport Research Center (TRANSyT), Universidad Politécnica de Madrid, 28040 Madrid, Spain; thais.rangel@upm.es (T.R.); josemanuel.vassallo@upm.es (J.M.V.)

² Volkswagen Data:Lab, Volkswagen AG, 80805 Munich, Germany; anestis.papanikolaou@volkswagen.de

³ Department of Organizational Engineering, Business Administration and Statistics, Universidad Politécnica de Madrid, 28012 Madrid, Spain

* Correspondence: tulio.silveira@upm.es

Abstract: App-based ride-hailing mobility services are becoming increasingly popular in cities worldwide. However, key drivers explaining the balance between supply and demand to set final prices remain to a considerable extent unknown. This research intends to understand and predict the behavior of ride-hailing fares by employing statistical and supervised machine learning approaches (such as Linear Regression, Decision Tree, and Random Forest). The data used for model calibration correspond to a ten-month period and were downloaded from the Uber Application Programming Interface for the city of Madrid. The findings reveal that the Random Forest model is the most appropriate for this type of prediction, having the best performance metrics. To further understand the patterns of the prediction errors, the unsupervised technique of cluster analysis (using the k-means clustering method) was applied to explore the variation of the discrepancy between Uber fares predictions and observed values. The analysis identified a small share of observations with high prediction errors (only 1.96%), which are caused by unexpected surges due to imbalances between supply and demand (usually occurring at major events, peak times, weekends, holidays, or when there is a taxi strike). This study helps policymakers understand pricing, demand for services, and pricing schemes in the ride-hailing market.

Keywords: ride-hailing; dynamic pricing; machine learning; artificial intelligence; data analytics; prediction error; clustering analysis; decision-making process; transport policy



Citation: Silveira-Santos, T.; Papanikolaou, A.; Rangel, T.; Manuel Vassallo, J. Understanding and Predicting Ride-Hailing Fares in Madrid: A Combination of Supervised and Unsupervised Techniques. *Appl. Sci.* **2023**, *13*, 5147. <https://doi.org/10.3390/app13085147>

Academic Editors: Ahmad Al-Khasawneh and Abdalwali Lutfi

Received: 23 March 2023

Revised: 16 April 2023

Accepted: 18 April 2023

Published: 20 April 2023



Copyright: © 2023 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Urban transportation has changed significantly in recent years, given the fast development of innovative technologies. New app-based mobility services such as ride-hailing are becoming more and more popular because of the consumer behavior shift from ownership to accessibility [1]. Ride-hailing has recently exploded in popularity, as indicated by the business success of transportation network companies (TNCs) such as Uber and Lyft [2].

The popularity of these companies can be explained in part by the fact that they often provide cheap, comfortable, on-demand door-to-door transportation options in urban areas [3]. In many cities, this service has thus become an essential part of the transportation system. In tandem with the global rise of this mobility alternative, several scholars have investigated the impact of ride-hailing services on individuals' travel behavior and mode choice, pricing structure, etc. Nevertheless, the findings are still uncertain in several aspects, particularly regarding the pricing scheme.

Ride-hailing companies use real-time dynamic algorithms to adjust their fares at any moment, whereas taxi fares are usually fixed and regulated [4]. Dynamic pricing, also known as surge pricing, is an automated system based on demand and supply principles. While it is unclear how these rates are adjusted at any given time, understanding the behavior of the ride-hailing fares will be valuable to (i) users, to anticipate fares in advance;

(ii) drivers, to monitor and be prepared for fare rises and hence determine in advance potential opportunities to collect additional revenue; and (iii) policymakers, to apply regulatory measures to improve the overall transport system.

The main objective of this paper is to better understand, explain and predict the behavior of ride-hailing fares by combining statistical and supervised machine learning models for informing transport policymaking. To that end, data from Uber (one of the most popular TNCs in the world) were gathered in order to make predictions and explain its fares as a function of a range of explanatory variables. The Uber Application Programming Interface (API) was utilized to collect data on Uber ride supply in the city of Madrid over 10 months (from September 2018 to June 2019). This paper addresses the following research question: “Can existing open (big) data be combined with statistical and supervised machine learning techniques to help predict ride-hailing fares?”.

This paper also sets forth a conceptual and methodological framework for combining open (big) data of ride-hailing fares and additional information with predictive modeling for understanding the pricing mechanisms and incorporating them with the decision-making processes of agencies, stakeholders, and policymakers for the ride-hailing market.

In Spain, the ride-hailing market is limited, as is the number of ride-hailing licenses [5]. The city of Madrid was chosen for this paper because it is one of the most populous cities in the European Union and has a variety of transport modes. There has also been a big conflict between ride-hailing companies and conventional taxi services over the last few years [5,6]. In addition, the authors were able to build a large dataset in that city, which is not easy to find in other locations. Rangel et al. (2021) [5] used part of this dataset in a previous study in the city of Madrid considering only those origin-destination routes where fares experience notable variations over time (dynamic prices are effective). The authors explored ride-hailing fares using an econometric model—a Generalized Linear Model. The current paper goes further in the application of machine learning models (more robust models), applied in this city, and considers the complete dataset thus working with fixed and dynamic prices. The current dataset is 240% larger than the dataset used by Rangel et al. (2021) [5].

This research contributes to understanding and predicting ride-hailing fares, by combining supervised and unsupervised techniques. Three statistical and supervised machine learning models were used for short-term prediction, using scikit-learn’s open-source machine learning library (such as Linear Regression, Decision Tree, and Random Forest). Each model predicts Uber fares using different algorithms that can be compared in terms of performance metrics. The unsupervised technique of cluster analysis (using the *k*-means clustering method) was also applied to verify the difficulties of predicting Uber fares according to the prediction errors of the models.

Besides this introductory section, the paper has six additional sections. The background and literature review are presented in Section 2. The case study selection is described in Section 3, which is followed by data collection and analysis in Section 4. The methodology used to obtain the results of this research is detailed in Section 5. The results and discussion are presented in the Section 6, followed by conclusions and policy recommendations.

2. Background and Literature Review

Uber, Lyft, Cabify, and Didi are examples of ride-hailing services that use information and communication technologies (ICT). Ride-hailing companies typically use smartphone apps to provide their services, allowing the users to request a ride and receive information about the pick-up time, vehicle location, and the fare they will pay in advance.

The pricing strategy of ride-hailing companies is needed for their long-term success [7]. Dynamic pricing is a strategy by which products or services prices are adjusted in response to real-time supply–demand imbalances using a dynamic algorithm [8]. Up to about ten years ago, dynamic pricing was primarily limited to a few industries, such as airlines and hotels. Now dynamic pricing is used by companies in many other sectors, such as ride-hailing companies. The regular fare for a ride might, for example, increase during

rainy conditions, trip delays (congestion), morning and evening peak hours, and leisure days [5], which influence supply or demand.

The flexibility of dynamic pricing should increase the global welfare for society if the industry market is perfect, and thus externalities are internalized. However, the significance and distributions of welfare gains are unclear. Many critics suggest that dynamic pricing can decrease welfare gains for riders [9] or drivers [10].

The advent of ride-hailing services has significantly impacted the taxi market. Its real impact is difficult to measure due to the limited data. According to Chang (2017) [11], Uber reduced regular taxi drivers' income by 12% in Taiwan, and up to 18% after three years. Willis and Tranos (2021) [12] conclude that traditional taxi trips in New York have decreased after the entry of Uber. Akimova et al. (2020) [13] showed that ride-hailing services have had a significant negative impact on the profitability of taxi companies in Madrid and Barcelona. Besides, the taxi supply is limited by the municipalities in many countries such as Spain.

Ride-hailing and taxi services operate under different legal and regulatory frameworks in many countries [5,7]. The ride-hailing platforms retain a percentage of the total fare as commissions once the ride is completed and paid, and the rest is transferred to the driver. The Chinese Department of Transportation started to regulate in August 2021 the commissions charged by the ride-hailing platforms to standardize the business of those companies and reduce their excessive commissions. In addition, some cities in China, such as Beijing, Shanghai, and Hangzhou, have adopted regulatory measures to restrict the number of drivers registered in Didi [5,7].

In recent years, the competition between Uber, Cabify, and taxis has caused strong opposition from taxi drivers, who have organized protests and strikes in countries such as Spain [5] and Chile [7]. Currently, the Chilean government is discussing the definition of a basic framework to regulate the system in Congress [7]. In New York, the government has taken regulatory measures to limit the number of Uber drivers in 2019 [7]. In Spain, Denmark, Italy, and Sweden, Uber services have already been declared illegal at some point [14].

Despite the increasing interest in ride-hailing topics, there are still some gaps in the literature that have motivated this research. Statistical models are the most common methodological tools used in ride-hailing studies. From the demand perspective, Faghih et al. (2019) [15] used time series to predict ride-hailing demand. From the supply perspective, Rangel et al. (2021) [5] used an econometric model to explore Uber fares in Madrid, based only on time-varying fare data. In recent years, there has been growing interest in applying machine learning methods in ride-hailing studies. For instance, Battifarano and Qian (2019) [16] proposed a general real-time framework for predicting surge multipliers. Their approach was based on a log-linear model, and their model was able to predict Uber surge multipliers in Pittsburgh up to two hours in advance. Yan et al. (2020) [17] applied Random Forest to model and predict the demand for ride-hailing services in Chicago. Chen et al. (2021) [18] adopted deep learning networks for short-time prediction of demand for ride-hailing services. Silveira-Santos et al. (2022) [19] analyzed Lyft fares in Atlanta and Boston before and during the COVID-19 pandemic, with a focus on applications of time series forecasting and machine learning models. However, the scientific literature on the prediction of ride-hailing fares is still limited. Short-term ride-hailing fares forecasting has room to be improved using machine learning models.

To sum up, this paper departs from previous studies and contributes to the prediction of ride-hailing fares in the following ways:

- Total fares were predicted, not just the surge multiplier, as noted by Battifarano and Qian (2019) [16]. This also made it possible to identify cases of low demand, in addition to cases of high demand.
- Data from a European city were analyzed (most research had previously focused on American cities) and data were collected over an extended period, a total of ten months.

- Statistical and machine learning models were applied and compared considering the complete dataset, thus working with fixed and dynamic prices, not only dynamic prices, as noted by Rangel et al. (2021) [5].
- A conceptual framework was described that can be adopted by interesting parties to better understand the pricing dynamics of the ride-hailing market.
- Valuable information and policy recommendations for the ride-hailing market are provided.

3. Case Study Selection

The study was conducted in the city of Madrid, which is one of the most populated cities in the European Union and has a variety of transport modes, both public and private. For a detailed description of the case study, the reader is referred to reports such as Ayuntamiento de Madrid (2021) [20] and Consorcio Regional de Transportes de Madrid (2019) [21].

There has also been a major conflict between ride-hailing companies and conventional taxi services in recent years. Taxi drivers complained that ride-hailing companies did not pay taxes in Spain, did not follow labor laws, and benefited from the freedom to change their fares whenever they wanted. As a result, taxi sector protests and strikes have become common in Spain in recent years, particularly in the city of Madrid [13].

In Spain, Uber and Cabify are the two largest ride-hailing companies [6,22]. Nevertheless, this paper only focuses on Uber services due to the lack of data available from its main competitor Cabify.

To better understand the cost of the ride, it is essential to know how Uber fares are estimated. The service fee p (total fare for a ride) for Uber is split into two parts (see Equation (1)).

$$p_{\text{service fee}} = p_{\text{base cost}} + p_{\text{dynamic pricing}} \quad (1)$$

The first component (base cost) includes regular fees, such as one-off fees, and trip fees proportional to the trip's duration and distance. The second component (dynamic pricing) reflects the result of Uber's surge pricing algorithm depending on supply and demand (S&D) imbalances [8,19].

Uber provides three different services in Spain (UberX, Uber Black, and Uber Van). UberX provides rides in regular vehicles for up to four people, while Uber Black is the premium service and Uber Van is a service for groups of up to 6 people. This paper primarily focuses on UberX rides because it is the most popular Uber product [23,24]. Table 1 describes the factors that influence UberX fares in Madrid during the period analyzed, based on multiple factors.

Table 1. Factors that influence UberX fares in Madrid (Adapted from El Confidencial, 2019 [25]).

Service Fee	Variable	UberX (EUR)
Base cost	One-off fee	0.40
	Cost per minute	0.15
	Cost per kilometer	1.22
Dynamic pricing *		S&D
Minimum fare		3.50

* Reflects the time evolution of supply and demand (S&D) imbalances.

Uber fares are determined by the company's policy, and the base cost considers three factors: (i) the one-off fee, which remains constant regardless of the length and duration of the ride; (ii) the cost per minute; (iii) the cost per kilometer. Dynamic pricing is applied depending on supply and demand through a real-time dynamic algorithm [4,26]. In addition, the minimum fare is also included, which is a minimum fare to compensate drivers for short rides.

It is worth mentioning that the real-time dynamic algorithm is not 'open sourced' from Uber (as well as other TNCs), and in most cases, information on fares is also not available. As a result, the underlying pricing mechanism is not known to transport professionals or policymakers. Given that fact, this research aims to better understand the behavior of

ride-hailing fares, using applications of machine learning models that may be useful for transport policy purposes.

4. Data Collection and Analysis

Data collection was obtained using Uber's Application Programming Interface (API). It was not possible to obtain information from other ride-hailing companies operating in the city (such as Cabify), since their APIs did not provide the availability of that information. Given the lack of up-to-date official empirical data on ride-hailing demand in Spain, ride-hailing prices can serve as a good proxy for estimating the level of demand, as the supply of ride-hailing services in Spain is very steady since car licenses are limited and most of the drivers work full time [22,27].

Using the web-scraping technique, a script was created in which the computer was taught to find the data that were deemed appropriate [28]. These tools allow for the real-time collection of requested ride information while controlling for the latitude and longitude coordinates of the chosen OD points (as was conducted by Rangel et al., 2021 [5], and Silveira-Santos et al., 2022 [19]).

This study is not intended to compete with existing open tools for fare prediction (such as Uber's Fare Estimator (<https://www.uber.com/global/en/price-estimate/>, accessed on 10 January 2022) and UberFareFinder (<https://uberfarefinder.com/>, accessed on 10 January 2022)), but rather to define a framework to identify the issues of ride-hailing fare prediction and the errors associated with it.

To collect information about ride-hailing fares, 10 locations in the city of Madrid were selected as the origin and destination (OD) of the requested rides (see Figure 1). These locations were chosen to cover the city uniformly, including spots of high demand (e.g., airports, public transport stations, etc.). Ride-hailing demand is high at three special locations in the city of Madrid (i.e., Madrid-Barajas Airport, Atocha Train Station, and Chamartín Train Station). Then, using a GIS tool, another seven points were chosen to uniformly cover the city. In the end, 10 locations in the city of Madrid were defined (making up a network with 90 potential routes).

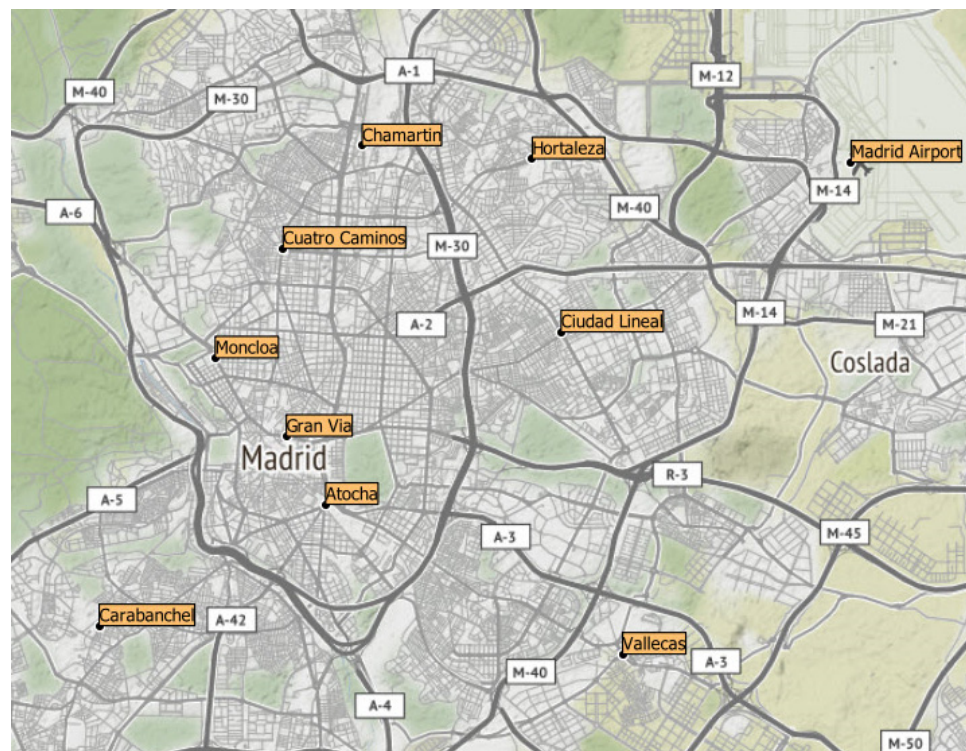


Figure 1. Madrid city and selection of the ODs of the requested rides.

For each ride requested through the Uber API, the following data were gathered: (i) fare; (ii) trip distance; (iii) trip duration; and (iv) trip request time (with year, month, day, and hour information). The Uber fare represents the cost of the ride as displayed by the app. The trip distance and duration indicate the distance and travel time required to travel to a specific OD.

Data were collected through the Uber API and stored at 1 h intervals over 10 months (from September 2018 to June 2019), and 667,051 entries were collected. Data cleaning was required in cases where the fare, distance, and travel time variables had zero values. After the data cleaning process, the final dataset ended up containing 665,977 entries (99.84% of the original sample).

Additional exogenous variables were added using feature engineering techniques (the process of extracting characteristics, properties, and attributes from raw data using domain knowledge) and queries to other information sources in addition to the variables obtained via the Uber API. Specifically, information on delay, month, and hour (time of day) was extracted. Other sources of information were consulted to obtain additional data on variables that may influence demand, such as rain precipitation, business days and holidays, peak hours, and taxi strike periods. All these new variables were incorporated into the predictive models to better explain the results.

Since the Uber API's travel time is estimated based on current traffic conditions, the delay variable was calculated as the difference between the travel time of an OD pair and the shortest travel time for that OD pair. This variable is a good indicator of the expected road congestion of each ride.

Categorical variables that control time-related features for requested trips, such as month and hour (time of day), were included to capture changes in ride-hailing fares at different times. In addition, holidays and peak hours in Madrid were verified, obtaining the variables of the business day (which does not include weekends and holidays) and peak hours (on business days from 07:00 to 09:00 and 18:00 to 20:00, according to EMESA, 2019 [29]).

Weather conditions are included in the analysis because they can affect Uber demand and thus influence ride fares [30]. Data on rain precipitation (measured in millimeters) was collected over 1 h. The State Meteorological Agency (AEMET) of Spain provided that information.

To account for special events that affect its main competitor's transportation supply, the analysis also considers taxi strikes that occurred during the analysis period. A categorical variable was included to capture Madrid's strike days from September 2018 to June 2019. Taxi strikes occurred for 20 days during the period studied.

Table 2 shows the descriptive statistics for the final data sample.

Table 2. Summary statistics of explanatory variables.

Variable	Typology	Unit	Summary Statistics	
UberX fare (FARE)	Continuous	Euro (EUR)	Mean	18.91
			Median	18.00
			Max.	149.00
			Min.	3.50
			SD	8.14
Trip distance (DIST)	Continuous	Kilometers (Km)	Mean	11.16
			Median	10.80
			Max.	36.97
			Min.	2.03
			SD	5.05
Travel time (TTIME)	Continuous	Minutes (min)	Mean	18.32
			Median	18.00
			Max.	53.00
			Min.	5.00
			SD	5.51

Table 2. Cont.

Variable	Typology	Unit	Summary Statistics	
Delay (DELAY)	Continuous	Minutes (min)	Mean	4.57
			Median	4.00
			Max.	32.00
			Min.	0.00
			SD	3.29
Rain precipitation (PREC)	Continuous	Millimeter (mm)	Mean	0.14
			Median	0.00
			Max.	91.00
			Min.	0.00
			SD	1.64
Business day (BUSINESS_DAY)	Categorical	-	Business day	448,300
Peak hour (PEAK_HOUR)	Categorical	-	Not business day	217,677
			Peak hour	110,709
Taxi strike (STRIKE)	Categorical	-	Not peak hour	555,268
			Strike	42,333
			No strike	623,644

Note: The Month (MONTH) and Hour (HOUR) variables were also considered in this research, being included as dummy variables.

Figure 2 shows the pair plot of continuous variables.

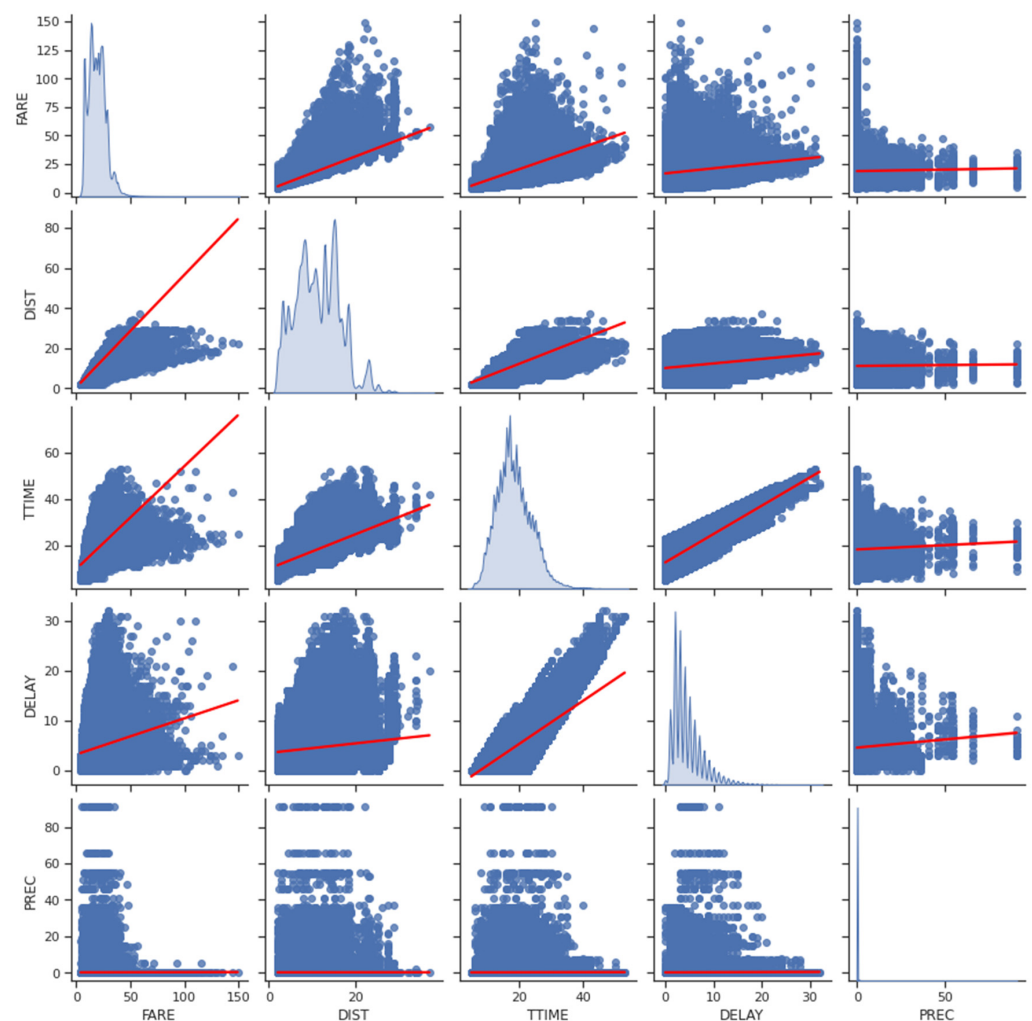


Figure 2. Pair plot of continuous variables. Notes: The points are in blue and linear regressions are in red, and the main diagonal represents kernel density estimates (KDEs).

The rain precipitation variable was highly skewed ($\gamma_1 = 23.59$) and has 653,048 (98.1%) zeros. The data out of the main diagonal represent scatter plots of continuous variables, in which the correlation between them can also be observed. Figure 3 shows the boxplot of continuous variables.

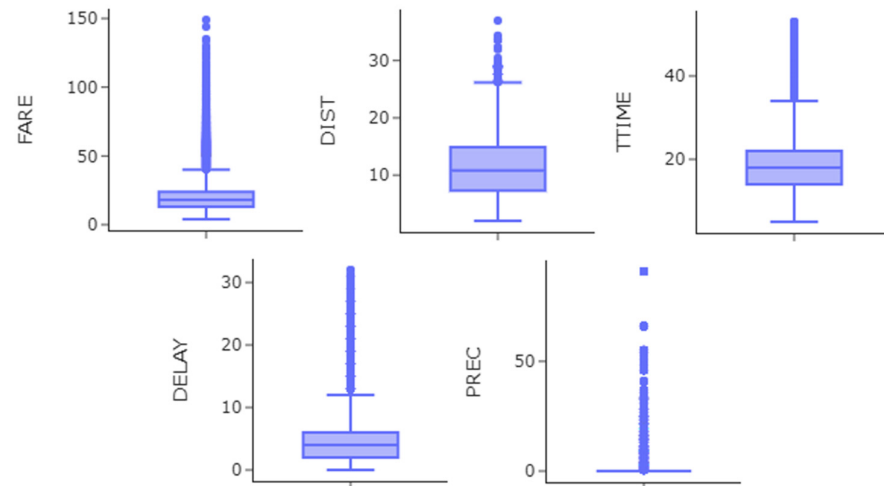


Figure 3. Boxplot of continuous variables.

It is worth noting that no outlier data were removed during the data cleaning process. “Extreme” ride-hailing fares were kept in the dataset because it is believed they represented some type of instant “market irregularity”, due to either low supply in cases of high demand or the opposite. Figure 4 shows the average fare per hour and type of day (business, weekends, and holidays) for a better understanding of the fare trend throughout the day.

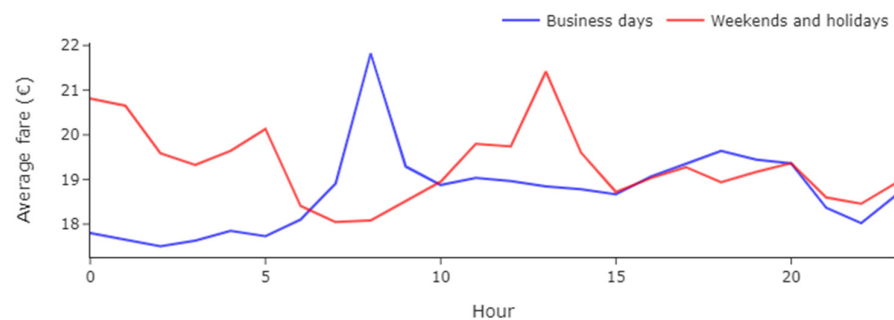


Figure 4. Average fare per hour and type of day.

Uber fares on business days appear to have a morning peak at 08 h and a smoother and wider afternoon peak at around 18 h. Weekend and holiday fares, on the other hand, are more expensive late at night (which can be linked to leisure trips back home) and in the early afternoon (which can be linked to leisure activities such as dining, shopping, etc.). Uber fares are on average EUR 18.72 on business days and EUR 19.29 on weekends and holidays.

5. Methodology

The conceptual methodological framework was developed in five sequential steps (see Figure 5).

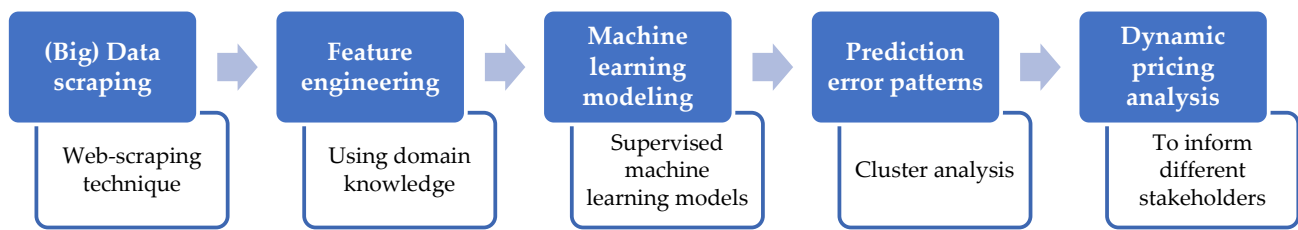


Figure 5. Methodological framework.

The first step includes (big) data scraping, which involves extracting information from a website and putting it into a database. The second stage includes the feature engineering technique, which involves the process of extracting characteristics, properties, and attributes from raw data using domain knowledge. These first two steps have already been presented in data collection and analysis (see Section 4).

The third step describes the methods used to better understand and predict the behavior of ride-hailing fares, focusing on statistical and machine learning modeling. Supervised machine learning models were used for short-term prediction. All models were used to predict the expected ride-hailing fare for the case study of Madrid with a one-hour forecast horizon.

Machine learning is used as a technique to “learn” from data [31,32]. These techniques were proposed in this paper as a computational alternative to solve the problem of interest using standard prediction error metrics and model evaluation techniques. In total, three predictive models were trained on Uber data (see Table 3), using scikit-learn’s open-source machine learning library (such as Linear Regression, Decision Tree, and Random Forest).

Table 3. Overview of each model used.

Model	Definition	Main Purpose	Main Advantages
Linear Regression (LR)	Ordinary least squares are used for regression problems.	Minimize the residual sum of squares between the observed targets in the dataset and the targets predicted by the linear approximation by fitting a linear model with coefficients.	Simple and most used statistical model. It is also one of the most basic machine learning algorithms.
Decision Tree (DT)	Non-parametric supervised learning method, used for classification and regression.	Create a model that uses simple decision rules inferred from data features to estimate the value of a target variable. A tree approximates a piecewise constant.	Simple to learn and analyze; minimum data preparation required; capable of handling both numerical and categorical data; capable of dealing with multi-output problems; employs a white box model; etc.
Random Forest (RF)	Ensemble method based on randomized decision trees, used for classification and regression.	A meta estimator that employs averaging to increase predicted accuracy and control overfitting by fitting several classifying decision trees on various sub-samples of the dataset.	One of the most accurate general-purpose machine learning methods; is robust; ability to minimize overfitting without increasing error related to bias; minor hyper-parameter tuning; quick training time; etc.

Source: Developed by authors and based on scikit-learn (<https://scikit-learn.org/>, accessed on 10 January 2022), towards data science (<https://towardsdatascience.com/>, accessed on 10 January 2022) and several authors [33–38].

The predictive ability of the models can be verified in terms of some performance metrics. The Root Mean Squared Error (RMSE) and the Mean Absolute Percentage Error (MAPE) are two metrics commonly used to assess the accuracy of prediction models [39].

The fourth step includes the analysis of error patterns, which involves analyzing the differences between the observed and predicted values of the models. To better understand the prediction results of the models, the unsupervised technique of cluster analysis was

employed (using the models' MAPE prediction errors as variables) to compare the profiles of each group with the other variables.

Capitalizing on the prediction errors analysis, the last step of the framework aims to highlight lessons learned from the behavior of dynamic pricing analysis and the behavior of misprediction (error) patterns caused by various supply–demand irregularities. Finally, the results of applying statistical and machine learning models to better understand the behavior of ride-hailing fares are linked to transport policies, highlighting the benefits of knowing/predicting fares for different stakeholders.

6. Results and Discussion

6.1. Statistical and Machine Learning Models

This subsection shows how statistical and machine learning models were used to predict the fare of the Uber service in Madrid with a one-hour prediction horizon (using scikit-learn's open-source machine learning library). Several input features were used, which include Uber API data and the new exogenous variables imposed on the model (as shown in Section 4).

The Uber fare variable (FARE) was used as the target variable explained by all the other variables listed in Table 2 as features (e.g., trip distance, travel time, delay, etc.). In this study, the default hyperparameters for the three machine learning models were used. It is noteworthy that the same random data split of training and testing sets was used for all models (i.e., Train/Test equal to 80/20) for comparison purposes. Predictive accuracy is evaluated and compared across all models (see Table 4).

Table 4. Comparison of performance metrics of the models.

Performance Metrics	Linear Regression	Decision Tree	Random Forest
RMSE (EUR)	3.41	3.85	3.40
MAPE (%)	8.01	6.60	6.32

The results show that the Random Forest model has the best average performance according to the RMSE (EUR3.40) and MAPE (6.32%) metrics. Although the RMSE results do not fluctuate as much, the MAPE results show that the Random Forest model performs better than the Decision Tree and Linear Regression models. Figure 6 shows the scatterplots of real fare values versus predicted ones.

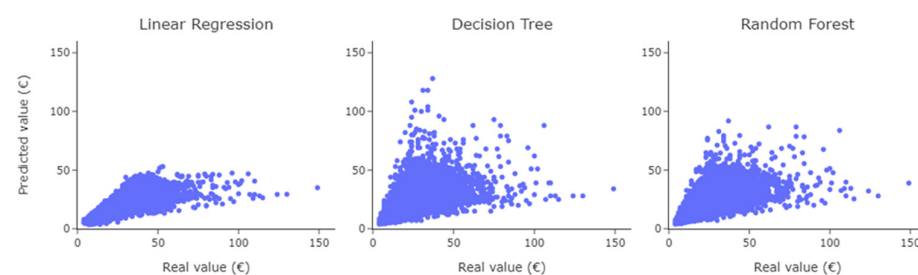


Figure 6. Scatterplots of real fare values versus predicted values.

The Linear Regression model has difficulty in predicting higher fare values (mainly values above EUR 50), while the Decision Tree and Random Forest models show similar trends, and both predict better high fares (but it is the Random Forest that best predicts fares overall). However, it is still possible to observe some fare peaks in which the predicted values differ greatly. Thus, the next subsection presents the application of cluster analysis to highlight the differences in predicting Uber fares across the alternative models.

6.2. Cluster Analysis of the Prediction Errors

To inform policymaking, it is necessary to have an in-depth understanding of how Uber fares react concerning demand or supply shortages. To this end, in this subsection

cluster analysis is applied using the prediction errors of the models as variables (namely, the distribution of MAPE prediction errors, which are mostly used for comparison purposes), identifying, for example, cases in which models overpredict or underpredict the observed fares.

Cluster analysis is a powerful technique for examining group features. Many studies based on clustering approaches have been conducted on feature recognition and analysis. Each observation belonging to one cluster is like the other ones belonging to it and different from all the other ones belonging to other clusters. The *k*-means algorithm is one of the most frequently employed techniques in group division and feature analysis [40].

This study applied the *k*-means clustering method for which the number of clusters is one of the most critical decisions. In this analysis, the *k* cluster number was set to three, as was conducted by other authors, such as Kumar et al. (2016) [41]. This study considers the number of clusters based on the distribution of MAPE prediction errors of the three models (i.e., Linear Regression, Decision Tree, and Random Forest), namely: (i) Low error; (ii) Medium error; and (iii) High error.

Fifty-nine iterations were necessary to achieve stability in the cluster centers. Table 5 shows the average distance of each variable (i.e., the MAPE prediction errors of the three machine learning models) to every cluster center and Figure 7 shows the number and percentage of observations per cluster.

Table 5. Final cluster centers.

MAPE Prediction Errors	Cluster		
	#1 (Low Error)	#2 (Medium Error)	#3 (High Error)
Linear Regression	6.030	21.458	22.544
Decision Tree	2.555	26.611	76.960
Random Forest	2.917	25.045	55.860

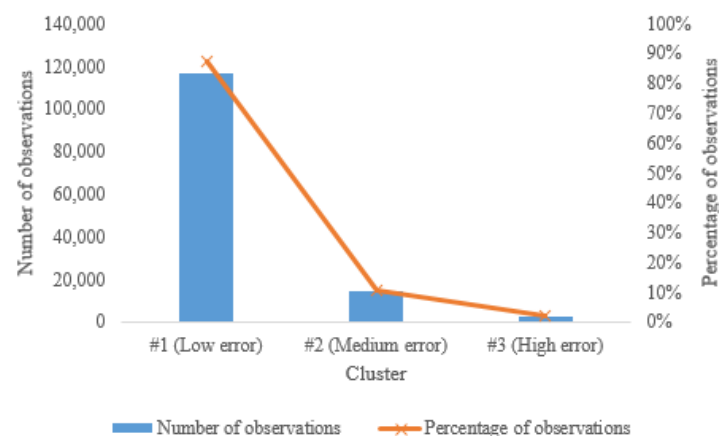


Figure 7. Number and percentage of observations per cluster.

The characteristics of the clusters are described below:

- Cluster #1 (Low error): This cluster has the smallest average distance values across all MAPE prediction error variables. The two smallest distances were found in the Decision Tree and Random Forest models, indicating that these models predict observations with smaller MAPE prediction errors than Linear Regression.
- Cluster #2 (Medium error): This cluster has intermediate average distance values on all MAPE prediction error variables. The smallest distance was found in the Linear Regression model.
- Cluster #3 (High error): This cluster has the highest average distance values across all MAPE prediction error variables. The smallest distance was found in the Linear

Regression model, indicating that this model predicts the observations in the group with larger MAPE prediction errors better.

The number of observations in each cluster also shows that cluster #1 (Low error) contains 87.32% of the observations, followed by cluster #2 (Medium error) with 10.72% and cluster #3 (High error) with 1.96%. All valid observations are included in the clusters. The results show that a high share of the observations have small MAPE prediction errors (87.32%), which shows good accuracy of the machine learning models used to predict the UberX fare (the Random Forest is the machine learning model with the best performance metrics and is also one of the models contributing the most to the group with small MAPE prediction errors). Likewise, it shows a very small share of the observations with high MAPE prediction errors (only 1.96%), which can be caused by unexpected surges due to imbalances between supply and demand, as well as being related to outliers that were not removed and/or other variables (see Section 4).

Table 6 presents a summary of the statistics of the clusters in terms of key continuous variables (e.g., UberX fare, trip distance, travel time, etc.).

Table 6. Relationship between the continuous variables and the formed clusters.

Cluster	Summary Statistics	FARE (EUR)	DIST (Km)	TTIME (min)	DELAY (min)	PREC (mm)
#1 (Low error)	Mean	18.57	11.35	18.47	4.60	0.14
	SD	7.29	5.01	5.51	3.30	1.66
#2 (Medium error)	Mean	20.90	9.74	17.02	4.31	0.14
	SD	11.52	5.08	5.17	2.95	1.56
#3 (High error)	Mean	23.38	10.24	17.93	4.62	0.21
	SD	16.38	4.83	5.35	3.51	2.07

Forecasting ride-hailing fares appears more difficult (high MAPE prediction errors) when the supply value of the fare variable (FARE) and when rain precipitation (PREC) is higher, but the high standard deviations of the errors overshadow this effect (due to the existence of outliers in all groups). It is also noteworthy there is no clear trend for the variables trip distance (DIST), travel time (TTIME), and delay (DELAY) concerning the errors across clusters, which means the models have already captured the statistical signal from these variables. Table 7 shows the percentage of frequency of observations of the categorical variables within the three clusters.

Table 7. Percentage of frequency of observations of the categorical variables within the formed clusters.

Cluster	STRIKE (Strike Is True)	BUSINESS_DAY (Business Day Is True)	PEAK_HOUR (Peak Hour Is True)
#1 (Low error)	6.25%	71.06%	16.02%
#2 (Medium error)	6.77%	42.79%	20.77%
#3 (High error)	10.20%	31.79%	20.44%

The results of the percentages of frequency of observations of the categorical variables per cluster show how much they interfere with the MAPE prediction errors, mainly in cluster #3 (High error). The prediction of ride-hailing fares becomes slightly more complex when there is a taxi strike, as well as during peak hours (from 07:00 to 09:00 and from 18:00 to 20:00). The results also show that the forecasts are more accurate on business days, being thus less accurate on weekends and holidays, in which there is a high percentage of frequency of observations in clusters #2 (57.21%) and #3 (68.21%). The previous reasons appear to be related to potential demand peaks that cause an imbalance between supply and demand. Figure 8 shows the percentage of frequency of observations per hour and cluster.

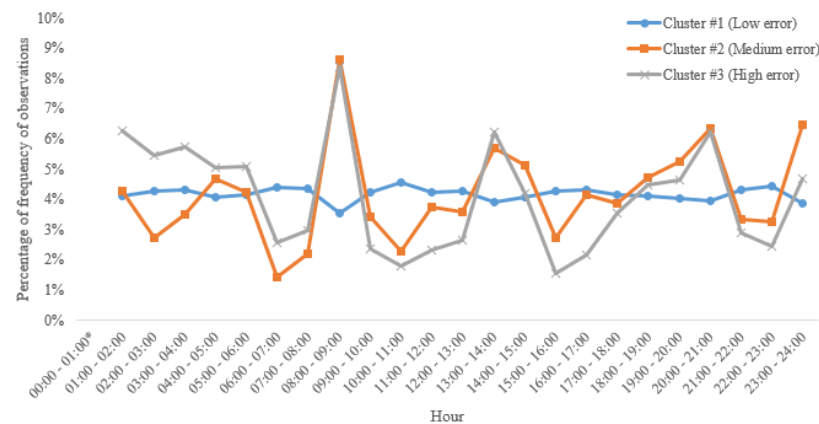


Figure 8. Percentage of frequency of observations per hour and cluster.

Cluster #1 (Low Error) is the one with the highest number of observations, being the frequency of observations almost constant across hours of the day, ranging from 3.52% to 4.56%. Clusters #2 (Medium error) and #3 (High error) behave in quite different ways. Despite having fewer observations, there is a greater share of high MAPE prediction errors in peak hours, the early afternoon, and late at night, which is also in line with what was presented in Figure 4.

To verify the cases in which the models overestimate or underestimate the observed fares, Table 8 presents the over-prediction and under-prediction by cluster and model.

Table 8. Over-prediction and under-prediction per cluster and model.

Percentage of Observations	#1 (Low Error)			#2 (Medium Error)			#3 (High Error)		
	LR	DT	RF	LR	DT	RF	LR	DT	RF
Over-prediction	75%	78%	71%	44%	53%	52%	67%	74%	74%
Under-prediction	25%	22%	29%	56%	47%	48%	33%	26%	26%
Total	100%	100%	100%	100%	100%	100%	100%	100%	100%

The models show more over-prediction than under-prediction in all clusters, especially in clusters #1 (Low error) and #3 (High error). Cluster #2 (Medium error) presents similar percentages of over-prediction and under-prediction and is the only one where the Linear Regression model differs more from the tree-based models.

6.3. Discussion

The results show differences in prediction using different models, especially in MAPE prediction errors (see Table 4) and in the scatterplots of real fare values versus predicted values (see Figure 6).

To better understand and explore the reasons for ‘highly mispredicted’ fares, cluster analysis was applied, using only the MAPE prediction errors from the three models as variables (see Section 6.2). The results of this analysis show that the prediction of ride-hailing fares become more difficult in the following cases: (i) the supply value of the fare variable is very high; (ii) there is higher rain precipitation; (iii) there is a taxi strike; (iv) during peak hours; and (v) during weekends and holidays. All the previous cases appear to be related to unexpected demand rises that produce an imbalance between supply and demand. The analysis thus helps identify in which circumstances there are imbalances between supply and demand.

As mentioned before, machine learning models were applied considering the complete dataset, thus working with fixed and dynamic prices, not only dynamic prices—as noted by Rangel et al. (2021) [5]. It is also noteworthy that these authors explored ride-hailing fares using an econometric model (adopting the Generalized Linear Model—GLM) but

decided to reduce the number of combinations to just 40 OD pairs (representing 42% of the entire sample of this research) by looking only at fare data with notable variations over time (i.e., dynamic prices). All Uber minimum fare data (fixed price equivalent to EUR 3.50) was excluded because the econometric model adopted did not work well with fixed prices. Although the sample is smaller than that of this research, these authors found similar prediction difficulties in the econometric model (such as the occurrence of taxi strikes and peak hours).

This research shows that machine learning methods can also handle the analysis of more data (which skewed the results of econometric models). The Random Forest model is the machine learning model with the best performance metrics (see Table 4) and is also one of the models that most contribute to the group with small MAPE prediction errors (see Table 5). The models also show more over-prediction than under-prediction in all clusters (see Table 8).

7. Conclusions and Policy Recommendations

Three models were applied to better understand the behavior of ride-hailing fares, using scikit-learn's open-source machine learning library (such as Linear Regression, Decision Tree, and Random Forest). The Random Forest was the one with the best performance metrics and is also one of the models that most contribute to the group with small MAPE prediction errors. The authors hence recommend combining statistical and supervised machine learning models with unsupervised techniques on the errors analysis to predict ride-hailing fares and better understand the conditions under which market imbalances occur, which lead to lower- or higher-than-expected ride-hailing fares.

From a transport policy point of view, the authors highlight several benefits of knowing/predicting the behavior of ride-hailing fares for different stakeholders: public authorities, regulatory authorities, users, and drivers.

Public authorities can take advantage of knowing and predicting ride-hailing fares to define and adopt policy measures to set a rational competition and coordination with the taxi industry and across ride-hailing companies. In Spain, for example, taxi services claim that they should be able to establish their prices with the same freedom as ride-hailing services to compete fairly with them [5]. Knowing the fares can also help them understand imbalances between mobility supply and demand and promote greater coordination with other mobility options (e.g., through Mobility as a Service package).

Regulatory authorities can use the methodology and result coming out from this paper to safeguard fair competition among different ride-hailing operators. The results can also help them identify bad practices from operators aimed to obtain larger earnings through a dominant position, as happened in China, according to the literature review.

The findings can help users know the ride's price in advance, thus facilitating them in choosing the most favorable option for their trips according to their priorities. They can also help drivers keep track of fare rises to secure higher pricing and hence more potential earnings.

Future research directions include: (i) adopting these methods to perform ride-hailing fare prediction using data from different stages of the COVID-19 pandemic; (ii) extending the research methods to other cities (as the ride-hailing market in Spain is restricted and the number of ride-hailing licenses is also limited); (iii) using other robust models for predicting ride-hailing fares, especially to better estimate unexpected surges; and (iv) collecting data from shorter intervals to more accurately capture peak fares.

Author Contributions: Conceptualization, T.S.-S., T.R. and J.M.V.; Data curation, T.S.-S. and T.R.; Formal analysis, T.S.-S. and A.P.; Investigation, T.S.-S. and T.R.; Methodology, T.S.-S., A.P. and T.R.; Resources, T.R. and J.M.V.; Software, T.S.-S. and A.P.; Supervision, J.M.V.; Validation, T.S.-S. and A.P.; Writing—original draft, T.S.-S. and T.R.; Writing—review and editing, T.S.-S., A.P., T.R. and J.M.V. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported by the Spanish Ministry of Science and Innovation, which has funded Project RTI2018-095501-B-I00 (MOBISHARING). This project has also been co-funded by the European Social Fund and the State Research Agency. Tulio Silveira-Santos is also grateful for his research grant (PRE2019-088587) funded by the Spanish Ministry of Science and Innovation and co-financed by the European Social Fund and the State Research Agency.

Data Availability Statement: The data presented in this study are available on request from the corresponding author. The data are not publicly available due to privacy.

Acknowledgments: The corresponding author thanks Natalia Sobrino for recommending the special issue “Algorithms and Applications regarding Big Data Analytics and Machine Learning” and ‘Stack Tecnologías’ for their support in the field of data science.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Gomez, J.; Aguilera-García, Á.; Dias, F.F.; Bhat, C.R.; Vassallo, J.M. Adoption and frequency of use of ride-hailing services in a European city: The case of Madrid. *Transp. Res. Part C Emerg. Technol.* **2021**, *131*, 103359. [CrossRef]
2. Dong, X.; Guerra, E.; Daziano, R.A. Impact of TNC on travel behavior and mode choice: A comparative analysis of Boston and Philadelphia. *Transportation* **2021**, *49*, 1577–1597. [CrossRef]
3. Dias, F.F.; Lavieri, P.S.; Garikapati, V.M.; Astroza, S.; Pendyala, R.M.; Bhat, C.R. A behavioral choice model of the use of car-sharing and ride-sourcing services. *Transportation* **2017**, *44*, 1307–1323. [CrossRef]
4. Chen, M.K.; Sheldon, M. Dynamic Pricing in a Labor Market: Surge Pricing and Flexible Work on the Uber Platform. In Proceedings of the 2016 ACM Conference on Economics and Computation, Maastricht, The Netherlands, 24–28 July 2016; pp. 1–19.
5. Rangel, T.; Gonzalez, J.N.; Gomez, J.; Romero, F.; Vassallo, J.M. Exploring ride-hailing fares: An empirical analysis of the case of Madrid. *Transportation* **2021**, *49*, 373–393. [CrossRef]
6. Vega-Gonzalo, M.; Aguilera-García, Á.; Gomez, J.; Vassallo, J.M. Traditional taxi, e-hailing or ride-hailing? A GSEM approach to exploring service adoption patterns. *Transportation* **2023**, 1–40. [CrossRef]
7. Zhong, Y.; Yang, T.; Cao, B.; Cheng, T.C.E. On-demand ride-hailing platforms in competition with the taxi industry: Pricing strategies and government supervision. *Int. J. Prod. Econ.* **2022**, *243*, 108301. [CrossRef]
8. Schröder, M.; Storch, D.M.; Marszal, P.; Timme, M. Anomalous supply shortages from dynamic pricing in on-demand mobility. *Nat. Commun.* **2020**, *11*, 4831. [CrossRef] [PubMed]
9. Dholakia, U.M. Everyone Hates Uber’s Surge Pricing—Here’s How to Fix It. Harvard Business Review. 2015. Available online: <https://hbr.org/2015/12/everyone-hates-ubers-surge-pricing-heres-how-to-fix-it> (accessed on 10 January 2022).
10. Goncharova, M. Ride-Hailing Drivers Are Slaves to the Surge. The New York Times. 2017. Available online: <https://www.nytimes.com/2017/01/12/nyregion/uber-lyft-juno-ride-hailing.html> (accessed on 10 January 2022).
11. Chang, H.H. The economic effects of uber on taxi drivers in Taiwan. *J. Compet. Law Econ.* **2017**, *13*, 475–500. [CrossRef]
12. Willis, G.; Tranos, E. Using ‘Big Data’ to understand the impacts of Uber on taxis in New York City. *Travel. Behav. Soc.* **2020**, *22*, 94–107. [CrossRef]
13. Akimova, T.; Arana-Landín, G.; Heras-Saizarbitoria, I. The economic impact of Transportation Network companies on the traditional taxi Sector: An empirical study in Spain. *Case Stud. Transp. Policy* **2020**, *8*, 612–619. [CrossRef]
14. OECD. Taxi, Ride-Sourcing and Ride-Sharing Services—Background Note by the Secretariat. *SSRN Electron. J.* **2018**, *2*, 1–38. [CrossRef]
15. Faghih, S.S.; Safikhani, A.; Moghimi, B.; Kamga, C. Predicting Short-Term Uber Demand in New York City Using Spatiotemporal Modeling. *J. Comput. Civ. Eng.* **2019**, *33*, 05019002. [CrossRef]
16. Battifarano, M.; Qian, Z.S. Predicting real-time surge pricing of ride-sourcing companies. *Transp. Res. Part C Emerg. Technol.* **2019**, *107*, 444–462. [CrossRef]
17. Yan, X.; Liu, X.; Zhao, X. Using machine learning for direct demand modeling of ridesourcing services in Chicago. *J. Transp. Geogr.* **2020**, *83*, 102661. [CrossRef]
18. Chen, L.; Thakuriah, P.V.; Ampountolas, K. Short-Term Prediction of Demand for Ride-Hailing Services: A Deep Learning Approach. *J. Big Data Anal. Transp.* **2021**, *3*, 175–195. [CrossRef]
19. Silveira-Santos, T.; González, A.B.R.; Rangel, T.; Pozo, R.F.; Vassallo, J.M.; Díaz, J.J.V. Were ride-hailing fares affected by the COVID-19 pandemic? Empirical analyses in Atlanta and Boston. *Transportation* **2022**, 1–32. [CrossRef] [PubMed]
20. Ayuntamiento de Madrid. Ayuntamiento de Madrid: Población por Distrito y Secciones Censales. 2021. Available online: <http://www-2.munimadrid.es/TSE6/control/seleccionDatosSeccion> (accessed on 9 July 2021).
21. Consorcio Regional de Transportes de Madrid. Encuesta Domiciliaria de Movilidad de la COMUNIDAD de Madrid 2018. 2019. Available online: <https://www.crtm.es/conocenos/planificacion-estudios-y-proyectos/encuesta-domiciliaria/edm2018.aspx> (accessed on 15 June 2021).
22. Aguilera-García, Á.; Gomez, J.; Velázquez, G.; Vassallo, J.M. Ridesourcing vs. traditional taxi services: Understanding users’ choices and preferences in Spain. *Transp. Res. Part A Policy Pract.* **2021**, *155*, 161–178. [CrossRef]

23. Hughes, R.; MacKenzie, D. Transportation network company wait times in Greater Seattle, and relationship to socioeconomic indicators. *J. Transp. Geogr.* **2016**, *56*, 36–44. [\[CrossRef\]](#)
24. Jiao, J. Investigating Uber price surges during a special event in Austin, TX. *Res. Transp. Bus. Manag.* **2018**, *29*, 101–107. [\[CrossRef\]](#)
25. El Confidencial. Uber Cambia Sus Precios en Madrid: Estas Serán Ahora Sus Nuevas Tarifas. El Confidencial. 2019. Available online: https://www.elconfidencial.com/tecnologia/2019-04-12/uber-madrid-cabify-vtc-taxi_1940338/ (accessed on 13 July 2021).
26. Ngo, V. *Transportation Network Companies and the Ridesourcing Industry: A Review of Impacts and Emerging Regulatory Frameworks for Uber*; The University of British Columbia: Vancouver, BC, Canada, 2015.
27. De Miguel-Molina, M.; de Miguel-Molina, B.; Catalá-Pérez, D. The collaborative economy and taxi services: Moving towards new business models in Spain. *Res. Transp. Bus. Manag.* **2021**, *39*, 100503. [\[CrossRef\]](#)
28. Glez-Peña, D.; Lourenço, A.; López-Fernández, H.; Reboiro-Jato, M.; Fdez-Riverola, F. Web scraping technologies in an API world. *Brief. Bioinform.* **2013**, *15*, 788–797. [\[CrossRef\]](#)
29. EMESA. ¿Cuáles Son Las Horas Punta del Tráfico en Madrid? EMESA. 2019. Available online: <https://www.emesa-m30.es/hora-punta-de-los-atascos-en-madrid/> (accessed on 16 July 2021).
30. Shokoohyar, S.; Sobhani, A.; Sobhani, A. Impacts of trip characteristics and weather condition on ride-sourcing network: Evidence from Uber and Lyft. *Res. Transp. Econ.* **2020**, *80*, 100820. [\[CrossRef\]](#)
31. Basu, R.; Ferreira, J. Understanding household vehicle ownership in Singapore through a comparison of econometric and machine learning models. *Transp. Res. Procedia* **2020**, *48*, 1674–1693. [\[CrossRef\]](#)
32. Koushik, A.N.P.; Manoj, M.; Nezamuddin, N. Machine learning applications in activity-travel behaviour research: A review. *Transp. Rev.* **2020**, *40*, 288–311. [\[CrossRef\]](#)
33. Dogru, N.; Subasi, A. Traffic accident detection using random forest classifier. In Proceedings of the 2018 15th Learning and Technology Conference (L&T), Jeddah, Saudi Arabia, 25–26 February 2018; pp. 40–45. [\[CrossRef\]](#)
34. Maulud, D.; Abdulazeez, A.M. A Review on Linear Regression Comprehensive in Machine Learning. *J. Appl. Sci. Technol. Trends* **2020**, *1*, 140–147. [\[CrossRef\]](#)
35. Pekel, E. Estimation of soil moisture using decision tree regression. *Theor. Appl. Climatol.* **2020**, *139*, 1111–1119. [\[CrossRef\]](#)
36. Wu, J.; Liu, C.; Cui, W.; Zhang, Y. Personalized Collaborative Filtering Recommendation Algorithm based on Linear Regression. In Proceedings of the 2019 IEEE International Conference on Power Data Science (ICPDS), Taizhou, China, 22–24 November 2019; pp. 139–142. [\[CrossRef\]](#)
37. Zantalis, F.; Koulouras, G.; Karabetsos, S.; Kandris, D. A review of machine learning and IoT in smart transportation. *Futur. Internet* **2019**, *11*, 94. [\[CrossRef\]](#)
38. Patange, A.D.; Pardeshi, S.S.; Jegadeeshwaran, R.; Zarkar, A.; Verma, K. Augmentation of Decision Tree Model Through Hyper-Parameters Tuning for Monitoring of Cutting Tool Faults Based on Vibration Signatures. *J. Vib. Eng. Technol.* **2022**, 0123456789. [\[CrossRef\]](#)
39. Washington, S.P.; Karlaftis, M.G.; Mannering, F.L. *Statistical and Econometric Methods for Transportation Data Analysis*, 2nd ed.; Taylor & Francis Group: London, UK, 2011.
40. Kim, K. Exploring the difference between ridership patterns of subway and taxi: Case study in Seoul. *J. Transp. Geogr.* **2018**, *66*, 213–223. [\[CrossRef\]](#)
41. Kumar, P.; Gupta, S.; Agarwal, M.; Singh, U. Categorization and standardization of accidental risk-criticality levels of human error to develop risk and safety management policy. *Saf. Sci.* **2016**, *85*, 88–98. [\[CrossRef\]](#)

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.