AGILE: GIScience Series, 4, 39, 2023. https://doi.org/10.5194/agile-giss-4-39-2023
Proceedings of the 26th AGILE Conference on Geographic Information Science, 2023.
Editors: P. van Oosterom, H. Ploeger, A. Mansourian, S. Scheider, R. Lemmens, and B. van Loenen.
This contribution underwent peer review based on a full paper submission.

Modeling the choice of shared micro-mobility services using XGBoost machine learning algorithm

Qilin Ren[1], Pengxiang Zhao[1], and Ali Mansourian[1]

[1]Department of Physical Geography and Ecosystem Science, Lund University, Lund, Sweden

Correspondence: Pengxiang Zhao (pengxiang.zhao@nateko.lu.se)

Abstract. In recent years, shared micro-mobility services (e.g., bikes, e-bikes, and e-scooters) have been popularized at a rapid pace worldwide, which provide more choices for people's short and medium-distance travel. Accurately modeling the choice of these shared micro-mobility services is important for their regulation and management. However, little attention has been paid to modeling their choice, especially with machine learning. In this paper, we explore the potential of the XGBoost model to model the three types of shared micro-mobility services, including docked bike, docked e-bike, and dockless e-scooter, in Zurich, Switzerland. The model achieves an accuracy of 72.6%. Moreover, the permutation feature importance is implemented to interpret the model prediction. It is found that trip duration, trip distance, and difference in elevation present higher feature importance in the prediction. The findings are beneficial for urban planners and operators to further improve the shared micro-mobility services toward sustainable urban mobility.

Keywords. Shared micro-mobility, Machine learning, Vehicle availability data, Feature importance, Mode choice

1 Introduction

Shared micro-mobility services have been proliferating worldwide in the past few years, including dockless e-scooters, dockless and docked bikes and e-bikes, etc. Due to their environmentally-friendly and flexible characteristics, shared micro-mobility services have been widely used to supplement public transport by dealing with first and last-mile problems (Qin et al., 2018; Romm et al., 2022). However, there are also many problems that are brought by them, such as disorderly placement of vehicles, and overproduction of facilities leading to resource wasting. Thus, it is crucial to understand shared micro-mobility usage and its influencing factors for better urban planning and management (Li et al., 2020; Abduljabbar et al., 2021; Mangold et al., 2022).

In addition, various shared micro-mobility services are provided in many cities, which provide more travel mode choices to users, especially for short and medium-distance travel. Understanding how users adopt and utilize each type of service is beneficial for urban planners and policymakers to develop pertinent regulations on their usage. Hence, it is necessary to model the choice of shared micro-mobility services. The previous studies on shared micro-mobility usage are mainly concentrated in bike-sharing (Faghih-Imani et al., 2017; Li et al., 2021), e-scooter sharing (Huo et al., 2021; Li et al., 2022), comparison between bike-sharing and e-scooter sharing (McKenzie, 2019; Zhao et al., 2021; Blazanin et al., 2022), and integration of bike-sharing / e-scooter sharing and public transport (Campbell and Brakewood, 2017; Cao et al., 2021), etc. However, the usage of various shared micro-mobility services among users is not yet well investigated. Reck et al. (2021) firstly examined the mode choice of four different micro-mobility services (i.e. dockless e-scooters, dockless e-bikes, docked e-bikes and docked bikes) by developing a multinomial logit (MNL) model.

Although MNL model and its variants have been demonstrated to be effective in travel mode choice analysis, it is still difficult to deal with a high degree of complexity in a dataset (Kim, 2021). With the advent of the geospatial big data era, new methods are required to analyze travel behavior and travel mode choice based on various sources of geospatial big data. To fill the above-mentioned research gaps, this study aims to systematically model the choice of shared micro-mobility services with interpretable machine learning. First, the machine learning model XGBoost is developed to model the choice. Second, the model prediction is interpreted based on permutation feature importance. Three types of shared micro-mobility services in Zurich, Switzerland are explored, including docked-bike, docked e-bike, and dockless e-scooter.

2 Literature review

2.1 Travel mode choice

Travel mode choice is roughly categorized as walking, cycling, public transport, and private car. Most studies are conducted among these mode choices (Hu et al., 2018; Narayan et al., 2020; Bucher et al., 2020; Liu et al., 2022). Over the past decades, a number of models have been developed to conduct transport mode choice analysis based on the above-mentioned travel models and the related influencing factors. Traditionally, the logit models such as the MNL model, the nested logit model, and the mixed logit model, are probably one of the most commonly-used travel mode choice models (Zhao et al., 2020). In recent years, machine learning has been popularized and pervasive in many fields, including but not limited to transportation, such as transportation mode recognition (Jahangiri and Rakha, 2015), traffic flow prediction (Pun et al., 2019), road extraction (Jiao et al., 2022), etc. A series of recent studies have indicated that machine learning can outperform logit models in travel mode choice modeling. For instance, Lee et al. (2018) compared four Types of artificial neural networks (ANN) with an MNL model for travel mode choice modeling, which showed that the ANN models are superior to the MNL model. Zhao et al. (2020) conducted the prediction and behavioral analysis of travel mode choice by comparing machine learning with logit models. It was found that the random forest model achieves much higher predictive accuracy compared to MNL model and mixed logit model.

2.2 Influence factors of shared micro-mobility

There are plenty of empirical studies that have been conducted for examining the influencing factors of shared micro-mobility services, including both bike-sharing and e-scooter sharing. Taking bike-sharing services as an example, the influencing factors can be categorized into the following aspects (Eren and Uz, 2020), namely weather conditions (e.g., temperature, precipitation, wind speed), urban built environment (e.g., bicycle infrastructure, access to urban facilities, land use), public transportation, and socio-demographic factors (e.g., age, gender, education, income), temporal factors, and safety (e.g., the use of helmet). For example, Li et al. (2020) conducted an empirical study on dockless bike-sharing utilization and its explanatory factors. It is found that factors such as the proximity to public transport, and population density are significantly related to the utilization. Some other studies also explored how various influencing factors impact e-scooter sharing services. For instance, the study by Huo et al. (2021) examined the influence of the built environment on e-scooter sharing ridership in five cities. It was reported that the e-scooter sharing ridership is positively correlated with population density, employment density, intersection density, land use mixed entropy, and bus stop density.

Overall, machine learning has been widely used for travel mode choice modeling based on various influencing factors. However, little attention has been paid to the choice of multiple shared micro-mobility services, especially with machine learning techniques. This study will model the choice of shared micro-mobility services with machine learning based on the influencing factors.

3 Methodology

3.1 Data and software availability

The data is collected in Zurich, which is the largest city in Switzerland. The base map and location of the study area are presented in Figure 1. The main transport modes in Zurich are comprised of public transportation (46.9%) and private motorized transport (40.6%) in 2019 (Federal statistical office, 2021). There are 340 km of cycling lanes and tracks in Zurich, which provides shared micro-mobility a large application and development possibility.

The records from three types of shared micro-mobility services are collected from open-accessible APIs of a service provider in Switzerland from Feb 1 to Feb 29, 2020. The raw dataset includes the vehicle location data which are offered by the shared micro-mobility operator. After the data preprocessing, the dataset contains 58,048 trip records. Each record represents a trip, including the information of id and type of the vehicle, the start/end location and time, trip length, trip duration, and average speed of the trip. The sample size of each service for analysis is shown in Table 1.

Table 1. Samples of vehicle availability data records

services	sample size	percentage
Docked bike	9286	16.0%
Docked e-bike	25808	44.5%
E-scooter	22954	39.5%

In addition, some geographic and climate data are used to calculate the influencing factors, including DEM, road network, points of interest (POI), and weather conditions.

The whole study is conducted on a computer with Intel(R) Core(TM) i7-4930K CPU 3.40GHz and 32.0 GB RAM, and the program is coded with Python language

3.2 Influencing factors

Due to the privacy protection agreement, the profiles of user groups are unavailable, such as age, gender, and income, etc. Therefore, the influencing factors mainly consist of trip attributes and the built environment at the origin and destination of each trip. The built environment factors include elevation, and point of interest (POI). Some types

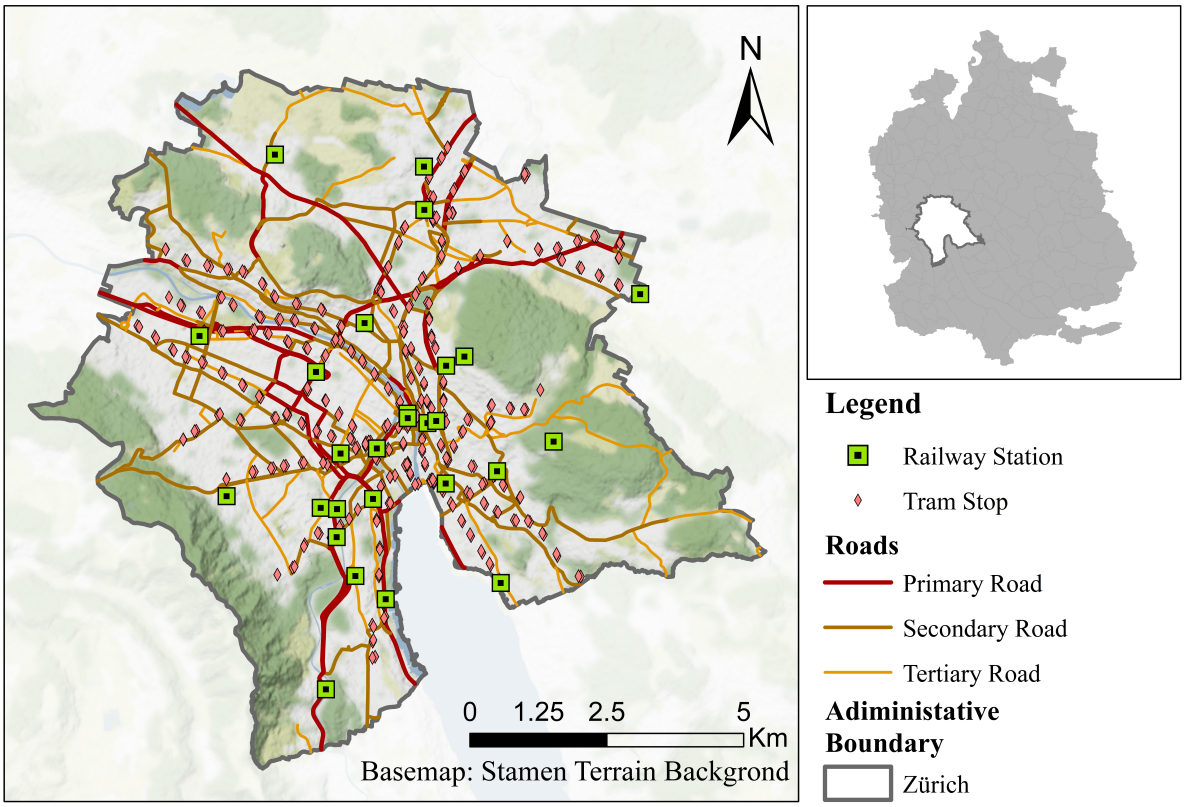

Figure 1. Study area.

of POIs that are closely related to cycling activities are selected, including public transport, education institutions, tourist attractions, and public facilities (e.g., theater, museum, post office, etc).

The influencing factors are described as follows: (1) The trip attributes, such as trip duration, trip distance, and trip speed. (2) The start time to each hour of the day and whether the trip happened on a weekday or weekend are aggregated. (3) Weather-related factors are considered, including temperature (°C) and wind speed (km/h). (4) The differences in elevation at the start and end are calculated and represented as $Elevation\_difference$. (5) The number of each type of POIs surrounding the trip end is counted by defining a 100-meter buffer. The following table shows the descriptions of the selected influencing factors or features.

Table 2. Type and value of each feature.

Feature	Type	Value
Hour_day	Categorical	[0, 1, ..., 23]
Day_week	Categorical	[Weekday, Weekend]
Trip_duration	Continuous	$[0, \infty]$
Trip_distance	Continuous	$[0, \infty]$
Trip_speed	Continuous	$[0, \infty]$
Elevation_difference	Continuous	[-203,211]
Temperature	Continuous	[-3, 18]
Wind_speed	Continuous	[0, 56]
Education	Continuous	$[0, \infty]$
Tourist_attraction	Continuous	$[0, \infty]$
Transport	Continuous	$[0, \infty]$
Public_facility	Continuous	$[0, \infty]$

3.3 Machine learning model

In this study, the machine learning model XGBoost is applied to model the choice of shared micro-mobility services. XGBoost is shorthand for extreme gradient boosting, which is a widely applicable implementation of the gradient boosting framework Chen and Guestrin (2016). Gradient boosting is also a boosting algorithm, which likewise attempts to build a strong classifier by an ensemble of weak classifiers Friedman (2001). In the framework of gradient boosting, the decision tree is the most applicable basis estimator. In contrast to adaptive boosting, the core of gradient boosting is that each classifier is trained with the residuals of all previous classifiers. Here, the residual is a numerical value that can be used to obtain the true value by adding it to the predicted value. The training process is iteratively conducted until the residuals approach zero. In addition, compared with random forest building an ensemble of independent trees, gradient boosting decision trees construct an ensemble of successive trees, in which each tree is trained based on the previous one.

The superiority of XGBoost lies in several important innovations compared to gradient boosting, including a regularized learning objective, shrinkage and column subsampling, optimization in storage and computation, etc. The tuning hyperparameters in XGBoost contain the number of trees (n_trees), learning rate, maximum tree depth, the fraction of observations to be randomly sampled for each tree, and the fraction of features for each tree.

3.4 Model interpretability

Machine learning models have demonstrated high performance in learning complex patterns from massive data through increased model complexity (e.g., deep learning), which are often referred to as black boxes (Murdoch et al., 2019). As a consequence, the rationale behind their predictions is difficult to understand and interpret. In order for humans to trust black-box methods, the interpretability of results is required. Hence, machine learning interpretability has attracted an increasing amount of attention in both academia and industry.

Permutation feature importance is a commonly-used technique in interpretable machine learning, which can be used for calculating the importance of each feature in a machine learning model. It measures the increase in the prediction error of the model after one feature's values are permuted or shuffled since the permutation breaks the relationship between the feature and the true outcome (Altmann et al., 2010). The idea behind this method is that if a feature is important, then permuting its values should have a significant impact on the performance of the model.

4 Results

4.1 Prediction evaluation

In this study, the prediction performance of XGBoost model is evaluated based on the typical evaluation metrics, including accuracy, F1 score, precision, and recall. First, since the numbers of trips for docked bike and e-scooter are more than that for docked e-bike, we randomly select 10,000 trips for docked bike and e-scooter respectively to guarantee the data balance. Second, the selected trips of three types of shared micro-mobility services and the influencing factors at the trip level are randomly split into 21,964 training data instances (75%) and 7,322 test data instances (25%). Finally, the grid search with 5-fold cross-validation is conducted to tune the hyperparameters of the XGBoost model using the training data. The best hyperparameter settings are used to train the model.

Figure 2 shows the confusion matrix from the XGBoost model. It indicates that the e-scooter trips can be predicted accurately, while the trips from docked bike and docked e-bike can not be distinguished very well. The four evaluation metrics are calculated based on the confusion matrix, as shown in Table 3. The XGBoost model yields an accuracy of 72.6%. The F1-score, precision, and recall results show how the XGBoost model performs on the three types of micro-mobility services individually. Overall, the trained XGBoost model can predict the choice of e-scooter sharing service, while not able to model the choices of bike-sharing and e-bike sharing services very well.

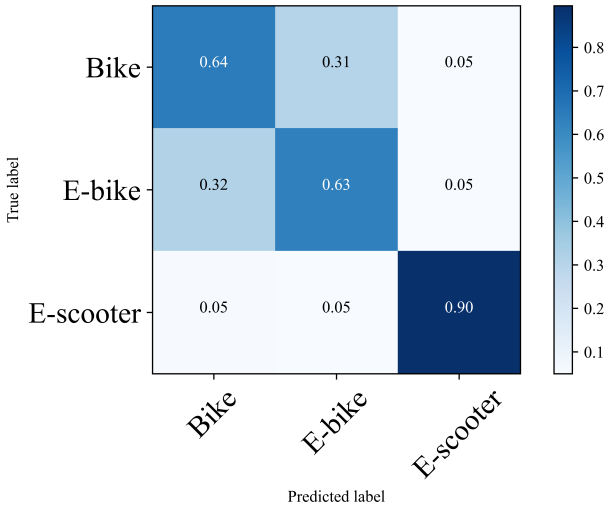

Figure 2. Confusion matrix from the XGBoost model.

Table 3. Evaluation metrics for the XGBoost model

Accuracy	72.6%		
	F1	Precision	Recall
Docked bike	63.6%	62.7%	64.5%
Docked e-bike	63.9%	64.6%	63.2%
E-scooter	89.8%	90.1%	89.6%

4.2 Feature importance analysis

Next, the permutation feature importance is calculated to measure the importance of the 12 selected features in the choice of modeling. In particular, each feature is permuted 50 times in the test data, and the reduction in accuracy is regarded as the feature importance. Figure 3 presents the importance of each feature with boxplot. It can be observed that trip duration and trip distance have higher feature importance in the XGBoost model. The study by McKenzie (2020) shows that e-bike sharing services have a higher average trip duration and distance than those of e-scooter sharing services, which can be used to distinguish the two types of services. The need for e-scooter is more affected by the long trip duration and distance. $Elevation_difference$ is displayed as the third important feature. The difference in elevations at the start and end delineates whether the trip is flat or hilly. The study by Li et al. (2021) reports that some areas with a high average elevation can be 200 m higher than those with a low elevation, and docked e-bikes are more attractive than

docked bikes traveling in those hilly regions. The study by Reck et al. (2021) also shows that e-scooters are more likely to be used on flat terrain, and e-bikes are preferable on tortuous terrains (both uphill and downhill). In addition, public facility and transport are followed in terms of the feature importance, which indicates that trip purpose also has an influence on the choice of shared micro-mobility services to some extent. For example, previous studies have confirmed that e-bikes are used for commuting, while e-scooters services are more often used for recreation (Bieliński and Ważna, 2020). Note that the feature $Hour\_day$ also shows the importance in the prediction. Its influence on the choice of shared micro-mobility services is also confirmed by the study (Reck et al., 2021). It is found that docked bike and docked e-bike have similar usage patterns while different from the usage pattern of shared e-scooters. Compared with the built environment and trip characteristics, two weather-related features do not display higher feature importance.

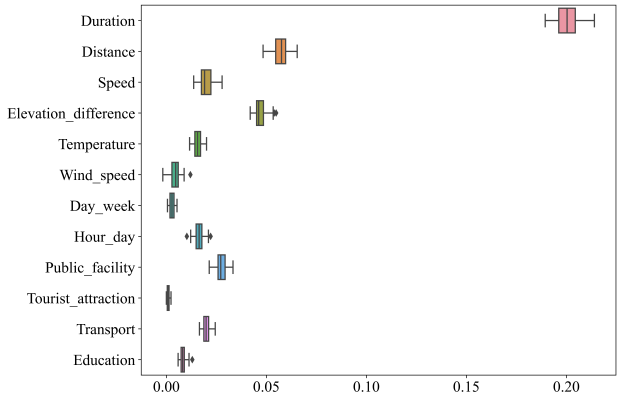

Figure 3. Feature importance in the XGBoost model.

5 Conclusion

The prevalence of shared micro-mobility services provides more choices for people's short and medium-distance travel. Studies on modeling the choice of shared micro-mobility services, especially at the trip level, are still scarce. The emergence of GPS-based vehicle availability data provides the possibility to investigate the research topic. In this paper, we model the choice of three types of shared micro-mobility services (i.e. docked bike, docked e-bike, and dockless e-scooter) with machine learning. To do this, we first determine the 12 influencing factors for the choice modeling by literature review, including trip attributes, the built environment, and weather conditions. Next, the XGBoost model is developed to model the choice of services based on the selected factors. Last, the prediction performance of the model is evaluated, and the interpretability of the model is analyzed in terms of permutation feature importance. The main findings of this study are summarized as follows.

First, the XGBoost model yields a prediction accuracy of 72.6%. By calculating the evaluation metrics F1, precision, and recall individually for each type of service, it is found that the choice of e-scooter sharing service can be modeled accurately, while docked bike and docked e-bike usage can not be distinguished very well. Next, the model prediction is interpreted by calculating permutation feature importance. The results show that trip duration, trip distance, and difference in elevation present higher feature importance in the prediction. This study has important implications with regard to the regulation and management of shared micro-mobility services.

There are also some limitations in the current study that call for future work. First, although the trained XGBoost model is capable of modeling the choice of e-scooter sharing service accurately, the prediction on the choices of the shared bike and e-bike services can be further improved. One solution could be to include more features (e.g., the urban built environment surrounding the start point of the trip) into the model. Second, this study only attempts the XGBoost model for choice modeling, other advanced machine learning / deep learning models deserve to be studied. Third, only the permutation feature importance analysis is implemented for the interpretability of the machine learning prediction. Some other interpretable machine learning techniques, such as partial dependence plots (PDP), and SHAP (SHapley Additive exPlanations), can be considered in future work.

References

Abduljabbar, R. L., Liyanage, S., and Dia, H.: The role of micro-mobility in shaping sustainable cities: A systematic literature review, Transportation research part D: transport and environment, 92, 102 734, 2021.

Altmann, A., Toloşi, L., Sander, O., and Lengauer, T.: Permutation importance: a corrected feature importance measure, Bioinformatics, 26, 1340–1347, 2010.

Bieliński, T. and Ważna, A.: Electric scooter sharing and bike sharing user behaviour and characteristics, Sustainability, 12, 9640, 2020.

Blazanin, G., Mondal, A., Asmussen, K. E., and Bhat, C. R.: E-scooter sharing and bikesharing systems: An individual-level analysis of factors affecting first-use and use frequency, Transportation research part C: emerging technologies, 135, 103 515, 2022.

Bucher, D., Martin, H., Hamper, J., Jaleh, A., Becker, H., Zhao, P., and Raubal, M.: Exploring Factors that Influence Individuals' Choice Between Internal Combustion Engine Cars and Electric Vehicles, GIScience, 1, 1–23, https://agile-giss.copernicus.org/articles/1/2/2020/, 2020.

Campbell, K. B. and Brakewood, C.: Sharing riders: How bike-sharing impacts bus ridership in New York City, Transportation Research Part A: Policy and Practice, 100, 264–282, 2017.

Cao, Z., Zhang, X., Chua, K., Yu, H., and Zhao, J.: E-scooter sharing to serve short-distance transit trips: A Singapore case,

Transportation research part A: policy and practice, 147, 177–196, 2021.

Chen, T. and Guestrin, C.: Xgboost: A scalable tree boosting system, in: Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining, pp. 785–794, ACM, 2016.

Eren, E. and Uz, V. E.: A review on bike-sharing: The factors affecting bike-sharing demand, Sustainable cities and society, 54, 101 882, 2020.

Faghih-Imani, A., Hampshire, R., Marla, L., and Eluru, N.: An empirical analysis of bike sharing usage and rebalancing: Evidence from Barcelona and Seville, Transportation Research Part A: Policy and Practice, 97, 177–191, 2017.

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

Hu, H., Xu, J., Shen, Q., Shi, F., and Chen, Y.: Travel mode choices in small cities of China: A case study of Changting, Transportation research part D: transport and environment, 59, 361–374, 2018.

Huo, J., Yang, H., Li, C., Zheng, R., Yang, L., and Wen, Y.: Influence of the built environment on E-scooter sharing ridership: A tale of five cities, Journal of Transport Geography, 93, 103 084, 2021.

Jahangiri, A. and Rakha, H. A.: Applying machine learning techniques to transportation mode recognition using mobile phone sensor data, IEEE transactions on intelligent transportation systems, 16, 2406–2417, 2015.

Jiao, C., Heitzler, M., and Hurni, L.: A fast and effective deep learning approach for road extraction from historical maps by automatically generating training data with symbol reconstruction, International Journal of Applied Earth Observation and Geoinformation, 113, 102 980, 2022.

Kim, E.-J.: Analysis of travel mode choice in seoul using an interpretable machine learning approach, Journal of Advanced Transportation, 2021, 1–13, 2021.

Lee, D., Derrible, S., and Pereira, F. C.: Comparison of four types of artificial neural network and a multinomial logit model for travel mode choice modeling, Transportation Research Record, 2672, 101–112, 2018.

Li, A., Zhao, P., Huang, Y., Gao, K., and Axhausen, K. W.: An empirical analysis of dockless bike-sharing utilization and its explanatory factors: case study from Shanghai, China, Journal of Transport Geography, 88, 102 828, 2020.

Li, A., Zhao, P., Haitao, H., Mansourian, A., and Axhausen, K. W.: How did micro-mobility change in response to COVID-19 pandemic? A case study based on spatial-temporal-semantic analytics, Computers, environment and urban systems, 90, 101 703, 2021.

Li, A., Zhao, P., Liu, X., Mansourian, A., Axhausen, K. W., and Qu, X.: Comprehensive comparison of e-scooter sharing mobility: Evidence from 30 European cities, Transportation Research Part D: Transport and Environment, 105, 103 229, 2022.

Liu, L., Kong, H., Liu, T., and Ma, X.: Mode choice between bus and bike-sharing for the last-mile connection to urban rail transit, Journal of Transportation Engineering, Part A: Systems, 148, 04022 017, 2022.

Mangold, M., Zhao, P., Haitao, H., and Mansourian, A.: Geofence planning for dockless bike-sharing systems: a GIS-based multi-criteria decision analysis framework, Urban Informatics, 1, 17, 2022.

McKenzie, G.: Spatiotemporal comparative analysis of scooter-share and bike-share usage patterns in Washington, DC, Journal of transport geography, 78, 19–28, 2019.

McKenzie, G.: Urban mobility in the sharing economy: A spatiotemporal comparison of shared mobility services, Computers, Environment and Urban Systems, 79, 101 418, 2020.

Murdoch, W. J., Singh, C., Kumbier, K., Abbasi-Asl, R., and Yu, B.: Definitions, methods, and applications in interpretable machine learning, Proceedings of the National Academy of Sciences, 116, 22 071–22 080, 2019.

Narayan, J., Cats, O., van Oort, N., and Hoogendoorn, S.: Integrated route choice and assignment model for fixed and flexible public transport systems, Transportation Research Part C: Emerging Technologies, 115, 102 631, 2020.

Pun, L., Zhao, P., and Liu, X.: A multiple regression approach for traffic flow estimation, IEEE access, 7, 35 998–36 009, 2019.

Qin, J., Lee, S., Yan, X., and Tan, Y.: Beyond solving the last mile problem: the substitution effects of bike-sharing on a ride-sharing platform, Journal of Business Analytics, 1, 13–28, 2018.

Reck, D. J., Haitao, H., Guidon, S., and Axhausen, K. W.: Explaining shared micromobility usage, competition and mode choice by modelling empirical data from Zurich, Switzerland, Transportation Research Part C: Emerging Technologies, 124, 102 947, 2021.

Romm, D., Verma, P., Karpinski, E., Sanders, T. L., and McKenzie, G.: Differences in first-mile and last-mile behaviour in candidate multi-modal Boston bike-share micromobility trips, Journal of transport geography, 102, 103 370, 2022.

Zhao, P., Haitao, H., Li, A., and Mansourian, A.: Impact of data processing on deriving micro-mobility patterns from vehicle availability data, Transportation Research Part D: Transport and Environment, 97, 102 913, 2021.

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