Using Origin-Destination Flow Graph and Public Transit Information to Enhance Short-Term Ridership Prediction in Bike-Sharing Systems

Kaifa Lu [1]

[1] International Center for Adaptation Planning and Design, College of Design, Construction and Planning, University of Florida, PO Box 115706, Gainesville, FL 32611-5706, USA

Abstract:
With the rapid growth of bike-sharing systems (BSS), accurate prediction of bike-sharing ridership is increasingly important to facilitate decision-makers in formulating bike-sharing dispatching and rebalancing plans towards maximizing efficiency and user satisfaction. In the literature, previous studies have developed lots of sophisticated bike-sharing ridership forecast models to improve the prediction accuracy while mostly ignoring model interpretability. There is a lack of interpretable prediction models that explicitly reflect the effects of different influential factors on bike-sharing ridership, which is important for better BSS design and planning to promote its use. To fill the gap, this project proposed an interpretable feature engineering approach to accurately predicting bike-sharing ridership by fully considering its spatiotemporal features, as well as the associations with meteorology, built environment, and public transit system. Specifically, we first employed a spatial vector autoregressive LASSO model to identify important variables that greatly influenced bike-sharing ridership. Based on the selected variables, we used machine learning approaches including Extreme Gradient Boosting (XGBoost) and Multi-Layer Preception (MLP) to predict future bike-sharing ridership and further reveal effects of different variables on the prediction accuracy. Using bike trip data in Washington D.C. as a case study, results suggested that public transit information and bike ridership in the last 24th,72nd,96th,120th,144th,168th hours were identified as key variables for ridership prediction and that XGBoost and MLP outperformed other baseline models. Further, public transit information and origin-destination flow graph attributes (i.e., degree and PageRank) certainly improved the prediction accoracy over non-zero bike ridership dataset than rush-hour and all datasets.

Keywords: Bike-sharing system; Ridership prediction; Variable selction; Machine learning

1. Introduction

Bike-sharing systems (BSS) are blooming globally as a sustainable, easily accessible, and affordable mode of trasnpostation over short distances, particularly over the first mile and last mile (Oeschger et al., 2020). As a result, bike-sharing market is becoming increasingly competitive in the first- and last-mile travel such that the presence of BSS can reduce people's dependencies on private vehicles and even encourage the use of public transit if they are well integrated (Tavassoli and Tamannaei, 2020). With the rapid growth of BSS in recent years, this type of modal substitution is becoming common practice for residents and commuters living in BSS-intensive areas, thus exhibiting huge potentials to reduce traffic congestion, air pollution, and greenhouse gas emissions mainly brought by private vehicles (Ferguson and Sanguinetti, 2021). However, there are still several tricky issues facing BSS, a typical one of which is imbalance between bike supply and demand in the context of time and space.

Specifically, as a flexible micro-mobility option, BSS efficiently extends geographical and demographic reach that the existing public transit network cannot cover, however, this flexibility also easily leads to the spatiotemporal supply-demand imbalance of sharing bikes (Chen et al., 2017). For instance, in the morning rush hours, an rapid increase in commuting trips may potentially result in a shortfall or even unavailable provision of sharing bikes in docked stations near residential areas (Yang et al., 2020), which does great harm to user satisfaction with service reliability of bike sharing systems. This similar situation may also happen in docked stations near commerical areas or workplace during the evening rush hours. To address the imbalance, a potential solution is to provide accurate spatiotemporal prediction of bike-sharing ridership that is crucial to assist decision-makers in formulating effective bike dispatching and rebalancing plans to improve system efficiency and user satisfaction and further promote the usage of BSS.

2. Background

In the literature, there are two typical methods of dealing with bike-sharing ridership forecasting problems (Yang et al., 2020): at an individual station level (Lin et al., 2018) and at an aggregated group level (Li et al., 2015). Due to a burst in the number of docked stations and high flexibility of sharing bikes, it is common that geographically adjacent bike stations usually serve for the same travel behaviors like trip origins and destinations, as well as trip purposes (Li et al., 2015). In this context, the bike-sharing ridership prediction at a station level is highly challenging and sometimes no use due to demand biases and uncertainties among adjacent stations. Instead, many studies (Li et al., 2015; Zhou et al., 2019) have attempted to predict bike-sharing ridership over a small spatial scale where bike riders almost have the same travel behaviours.

In terms of prediction models, most of related studies is based on either of parametric statistical models (Kaltenbrunner et al., 2010) or machine learning techniques (Cho et al., 2021). The former mainly includes using time-series prediction models (Jaber et al., 2022) such as ARIMA (Autoregressive Integrated Moving Average) and its variants to predict short-term or long-term bike-sharing ridership based on historical data of bike usage, however, they rarely consider the impact of spatial dependency among bike stations that may improve the prediction accuracy. To address this issue, machine learning techniques like random forest and support vector machine (Gao and Chen, 2022), convolutional neural networks (Li et al., 2023), and graph neural networks (Lin et al., 2018) are winning widespread research interests in predicting short-term bike-sharing ridership, owing to their strong capabilities of automatically capturing temporal dependency, spatial dependency, and semantic dependency in the BSS (Yang et al., 2020). Meanwhile, this powerful prediction capability of machine learning approaches demands for higher availability of big data and computing power, and furthermore, those complex machine learning techniques, particularly for neural networks, are usually difficult to interpret their predictive performances (Yang et al., 2020). To make a compromise among data need, computing power, prediction accuracy and model interpretability, it is vital to propose an interpretable feature engineering apporach, which is rarely discussed in related studies (Li et al., 2023; Xu et al., 2021; Yang et al., 2023), to informing model and feature selection through a combination of interpretable machine learning techniques like XGBoost and partially manual extraction of features from limited raw data.

The key to an interpretable feature engineering approach is about how to accurately extract features that significantly contribute to the accuracy of bike-sharing ridership prediction. As revealed in related research, temporal features (i.e., hour of a day, day of week and holidays) and meterological factors (e.g., temperature, wind, and precipitation) both exert a great influence on bike-sharing ridership (Cantelmo et al., 2020; Jaber et al., 2022). Also, research done by (Liu and Lin, 2019; Madapur et al., 1970) investigated the associations between bike-sharing ridership and built environment attributes including density and diversity. In addition, the operation of public transit systems including metro and bus schedules & ridership could also affect the variations in bike-sharing ridership (Bigazzi and Wong, 2020; Fan et al., 2019; Reck et al., 2021; Saberi et al., 2018). However, there is a lack of studies to explicitly incorporate the effects of these influential factors on short-term bike-sharing ridership into the prediction model, particularly an absence of considerations for the impacts of public transit information on bike-sharing ridership. Furthermore, there are more unexplored features present in time-series data of bike trips such as origin-destination (O-D) relationship among bike stations that may improve the accuracy of bike-sharing ridership forecast. Thus, this project aims to fill the research gap: whether bike trip O-D flow and public transit information can further improve the predictive accuracy of bike-sharing ridership in the BSS, in addition to temporal features, meteorological factors, and built environment attributes.

To sum up, we will propose a data-driven feature engineering approach that fully incorporates temporal and spatial features of bike trip data, meteorological factors, built environment attributes, O-D flow patterns, and public transit schedule information into an easily interpretable machine learning method - XGBoost to predict future bike-sharing ridership and further reveal their effects on the prediction accuracy. Finally, we will use the bike trip data during March 2020 from Captial Bikeshare program in Washington D.C. to further demonstrate the proposed interpretable feature enginering approach for bike-sharing ridership prediction.

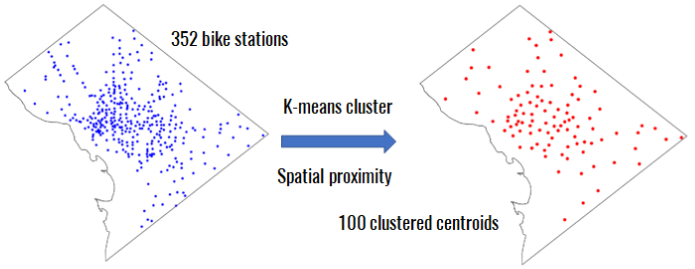
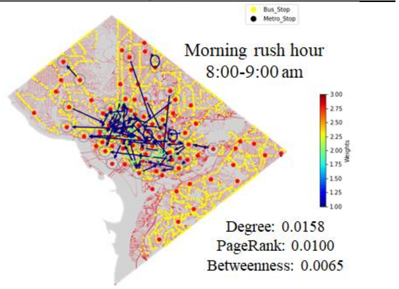
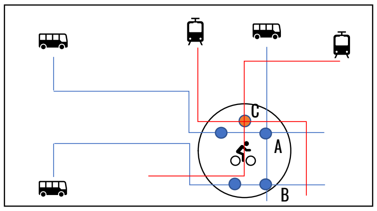
3. Approaches: Data and Methods

3.1. Data Preprocessing and Aggregation

3.1.1. Bike Station Clustering

Based on the spatial proximity, we used a hierarchical clustering method like K-means clustering (Hartigan and Wong, 1979) to group 352 bike stations in Washington D.C. into 100 clusters for bike-sharing ridership prediction (Table 1). As a result, each centroid of 100 clusters is a good representative of 3-5 geographically adjacent bike stations with the similar travel behaviors like trip origins and destinations (Yang et al., 2020).

Table 1 Illustration of data preprocessing

<table>
<tr><th colspan="2">1.1 Bike station clustering</th><th>1.2 OD flow graph construction</th></tr>
<tr><td colspan="2">
</td><td>
</td></tr>
<tr><td>1.3 Public transit information</td><td>
</td><td><i>Concept</i>: The number of transit (bus or metro) stops around the 100 clustered centroids characterizes the effect of public transit on bike-sharing system
Dynamic number of metro stops around centroids: 2
Static number of metro stops around centroids: 1
Dynamic number of bus stops around centroids: 6
Static number of bus stops around centroids: 4</td></tr>
</table>

3.1.2. Origin-Destination (O-D) Flow Graph Construction

Based on OD trip data of sharing bikes, we applied graph theory to construct spatiotemporal directed graphs for each hour over the BSS network (Saberi et al., 2016; Yang et al., 2019) where the 100 clustered centroids were nodes of each graph, the OD trips between two nodes were used to generate edges, and the volume of OD bike trips between the two nodes represented weights over edges. Then we extracted three typical graph attributes including degree, betweenness, and PageRank from each directed graph to represent hourly O-D flow patterns (Yang et al., 2020). We selected morning rush hour (8-9 am) as a time window to demonstrate the O-D flow graph construction and feature extraction process (Table 1).

3.1.3. Public Transit Information Extraction

As revealed in the literature (Hasselwander et al., 2022; Zuo et al., 2020), the intergration of BSS as a feeder mode of public transit system could efficiently enhance public transit accessibility and ridership, as well as transportation equity. Meanwhile, network planning and schedules of public transit system also affect bike-sharing ridership nearby (Cho et al., 2021). Due to unavailability of public transit ridership in Washington D.C., we only extracted the static and dynamic number of transit stops around each clustered centroid from transit network and schedules as metrics to characterize the effects of public transit system on bike-sharing ridership. Specifically, as presented in Table 1.3, within the buffer area of a clustered centroid, there are 1 metro stop and 4 bus stops over the transit network and therefore the static number of metro and bus stops around the centroid is 1 and 4; While based on transit schedules, the metro stop C and bus stops A&B are separately passed twice in an hour and thus the dynamic number of metro and bus stops around the centroid is 2 and 6. In fact, the dynamic number of transit stops around the clustered centroids is a better and more realistic representative of the relationship between public transit and BSS than the static one.

3.1.4. Data Aggregation

We aggregated hourly bike-sharing ridership data, meteorological factors (i.e., visibility, temperature, dew point temperature, relative humidity, wind speed & direction, and precipitation), built environment factors such as population density, housing unit density, land use diversity (Lin et al., 2020), and bike lane density, graph attributes including degree, betweenness, and PageRank (Yang et al., 2020), and public transit metrics (i.e., the static and dynamic number of transit stops around the centroids) into the same spatiotemporal scale to be ready for bike-sharing ridership prediction (Table 2).

Table 2 Descriptive statistics affiliated to the 100 clustered centroids in Washington D.C.

Metric	Statistical indicator					Variable description
	Mean	Std	Min	Median	Max	
Hourly ridership	1.7971	3.9791	0	0	72	Hourly volume of bike trips
Graph attributes						
Degree	0.0208	0.0406	0	0	2	No. of neighboring nodes
Betweenness	0.0081	0.0215	0	0	0.5	No. of links passing a node
PageRank	0.0098	0.0250	0	0.0024	1	Node importance
Public transit metrics						
Dynamic number of metro stops	9.2286	38.76	0	0	383	Around each clustered centroid
Static number of metro stops	0.0682	0.2521	0	0	1	
Dynamic number of bus stops	55.78	104.86	0	0	1119	
Static number of bus stops	1.9117	2.7488	0	0	15	
Built environment attributes						
Population density	0.0052	0.0037	2e-6	0.0040	0.0160	Continuous (people/100m^2)
Housing unit density	0.0034	0.0029	0	0.0024	0.0131	Continuous (HU/m^2)
Land use diversity	1.9413	0.3822	0.6107	1.9573	2.6330	Continuous
Bike lane density	0.2761	0.2025	0.0088	0.2427	1.0000	Continuous (mile/m^2)
Meteorological factors						
Visibility	9.5221	1.4178	0.1633	10	10	Continuous
Temperature	11.4	4.7469	-2.2	11.1	27.8	Continuous (C)
Dew point temperature	3.4911	6.1396	-11.4	3.9	16.7	Continuous (C)
Relative humidity	62.30	20.80	15	62	100	Continuous (%)
Wind speed	9.2003	4.7373	0	9	29	Continuous (mph)
Wind direction	174.97	109.70	0	180	360	Continuous
Precipitation	0.0021	0.0090	0	0	0.1025	Continuous (inch)

3.2. Spatial Vector Autoregressive LASSO Model for Variable Selection

Since bike-sharing ridership fluctuates spatiotemporally and also exhibits strong associations with a bundle of factors, i.e., time (historical ridership), space (hourly ridership in other bike stations), built environment, public transit, and meteorology, we introduced a spatial vector autoregression LASSO (SpVAR-LASSO) method for both temporal and spatial modelling of bike-sharing ridership (Jin and Lee, 2019) to investigate the effects of these influential factors on bike-sharing ridership and further identify important variables for bike-sharing ridership prediction. The SpVAR-LASSO model is formalized as follows:

$$y_{i,t} = \beta_0 + \boldsymbol{\beta_1} \boldsymbol{x}_i + \beta_2 \sum_j w_{ij} y_{j,t} + \boldsymbol{\beta_3} \boldsymbol{y}_{i,t-1\sim t-n} + \gamma \sum_j \left|\boldsymbol{\beta}_j\right| + \varepsilon_{i,t} \tag{1}$$

Where $y_{i,t}$ is the bike-sharing ridership of centroid i at time t; $\boldsymbol{x}_i$ means a 8×1 vector affiliated to centroid i such as population density, housing unit density, land use diversity, bike lane density, static and dynamic number of nearby metro & bus stops; $\sum_j w_{ij} y_{j,t}$ denotes a geographically weighted sum of the bike-sharing ridership of neighboring centroids at time t where w_{ij} is computed based on their geographical proximity (the inverse of geographic distance between centroid i and j); $\boldsymbol{y}_{i,t-1\sim t-n}$ refers to a $n \times 1$ vector describing the bike-sharing ridership of centroid i in the last n hours; $\gamma \sum_j \left|\beta_j\right|$ is a LASSO term used to regulate the coefficients of the SpVAR model by imposing a penalty on the regression coefficients with a parameter γ; $\varepsilon_{i,t}$ is the error item of centroid i at time t; β_0 represents centroid-specific effect; $\beta_0 \sim \beta_4$ are the coefficients of parameter.

Then we adopted a classic method - seemingly unrelated regression (SUR) (Zellner, 1962) to solve the SpVAR-LASSO model: $\boldsymbol{Y}_t = \beta_0 + \boldsymbol{\beta_1} \boldsymbol{X} + \beta_2 \boldsymbol{W} \boldsymbol{Y}_t + \boldsymbol{\beta_3} \boldsymbol{Y}_{t-1\sim t-n} + \gamma \sum_j \left|\boldsymbol{\beta}_j\right| + \varepsilon_{i,t}$ as follows:

$$\boldsymbol{\beta} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \left[(\boldsymbol{Y}_t - (\boldsymbol{X}, \boldsymbol{Y}_{t-1\sim t-n}) \boldsymbol{\beta})' (\boldsymbol{Y_t} - (\boldsymbol{X}, \boldsymbol{Y}_{t-1\sim t-n}) \boldsymbol{\beta}) + \gamma \sum_j \left|\boldsymbol{\beta}_j\right| \right] \tag{2}$$

$$\widehat{\boldsymbol{\beta}} = \left((\boldsymbol{X}, \boldsymbol{Y}_{t-1\sim t-n})' \widehat{\boldsymbol{\Omega}}^{-1} (\boldsymbol{X}, \boldsymbol{Y}_{t-1\sim t-n}) + \gamma \boldsymbol{W}^{-1} \right)^{-1} (\boldsymbol{X}, \boldsymbol{Y}_{t-1\sim t-n})' \widehat{\boldsymbol{\Omega}}^{-1} \boldsymbol{Y}_t \tag{3}$$

Where $\widehat{\boldsymbol{\Omega}}$ denotes the covariance matrix of disturbance; $\boldsymbol{W}^{-1}$ is a diagonal matrix with diagonal elements $\left|\widehat{\boldsymbol{\beta}}_j\right|$ and smaller coefficient estimates are forced by the constraints term of the LASSO to shrink into zero. As a result, the variables with non-zero coefficient estimates are considered as important variables affecting bike-sharing ridership and further used for subsequent ridership prediction.

3.3. A Machine Learning Approach for Bike-Sharing Ridership Prediction

Based on the selected variables from the SpVAR-LASSO model, we incorporated the corresponding dataset into an interpretable maching learning approach - Extreme Gradient Boosting (XGBoost) (Yang et al., 2020) to predict the bike-sharing ridership with inputs presented in Table 3 and quantify their contributions to the prediction accuracy. XGBoost is an ensemble learning algorithm of implementing multiple gradient boosted decision trees where each decision tree in this model works as a weak learner and these weak learners learn sequentially and adaptively to improve model predictions of decision trees (Yang et al., 2020). As a result, XGBoost has been frequently confimed to be capable of obtaining better or at least equivalent performances to complex neural networks (i.e., convolution neural networks and recurrent neural networks) in predicting traffic demand (Lin et al., 2018; Yang et al., 2020). Motivated by this merit and its stronger interpretability than neural networks, we used XGBoost as a feature engineering modelling framework to systematically study the potentials of different features in improving the predictive performance of bike-sharing ridership.

Table 3 Variables in the XGBoost model

Y: Dependent variable	Hourly ridership of sharing bikes in the next hour (+ 1 h)
X: Independent variable	
Temporal features (T)	Time-lagged bike-sharing ridership (-24, -72, -96, -120, -144, -168 h)
Meteorological factors (M)	Visibility, temperature, dew point temperature, relative humidity, wind speed & direction, precipitation
Spatial features (S)	Geographically weighted bike-sharing ridership
Time-lagged graphs (G)	Time-lagged degree, PageRank, and betweenness (-24, -72, -96, -120, -144, -168 h)
Public transit (PT)	Dynamic number of metro & bus stops, static number of bus stops around centroids

In addition, we selected mean sqaure error (MSE) and r^2 as two indicators to characterize the predictive performance of bike-sharing ridership as follows:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{4}$$

$$r^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y}_i)^2} \tag{5}$$

Where $y_i, \bar{y}_i$ denote the ground truth observations and their mean values, and $\hat{y}_i$ refer to the predictions.

4. Results

4.1. Spatiotemporal Distribution of Bike-Sharing Ridership

To accurately predict the bike-sharing ridership, it is important to acquire a basic understanding of its spatial and temporal distribution patterns. Here, we aggregated the hourly ridership by day of week, hour of a day, and 100 centroids to separately characterize its weekly, daily, and spatial variations, as illustrated in Fig. 1.

Temporal distribution: The bike-sharing ridership hardly displayed any significant weekly variation. There was no explicit evidence showing any difference of bike ridership between weekdays and weekends in a week. Instead, there were lots of zero bike ridership at certain hourly periods, regardless of weekdays or weekends. More specifically, these zero bike ridership frequently occurred during 11 pm-6 am in a day. While for other time periods, bike ridership exhibited significant daily variation with two peaks separately happening during morning (8-9 am) and evening (5-6 pm) rush hours.

Spatial distribution: We selected 8-9 am and 5-6 pm as two typical time windows to describe the spatial distribution patterns of bike-sharing ridership in the morning and evening rush hours. Fig. 1 (c)-(d) showed that the hourly bike ridership in the downtown area was higher than that in the surroundings and that spatial usage patterns of sharing bikes also reflected the land use patterns. Specifically, the areas A and B displayed a higher bike-sharing ridership separately in the morning and evening rush hours. This just complied with the land use patterns that the area A was more likely to be residential areas while the area B was commerical areas and workplace. Also, there were lots of zero bike ridership in the surrounding clustered centroids even in the morning and evening rush hours, which provided important hints for subsequent ridership prediction.

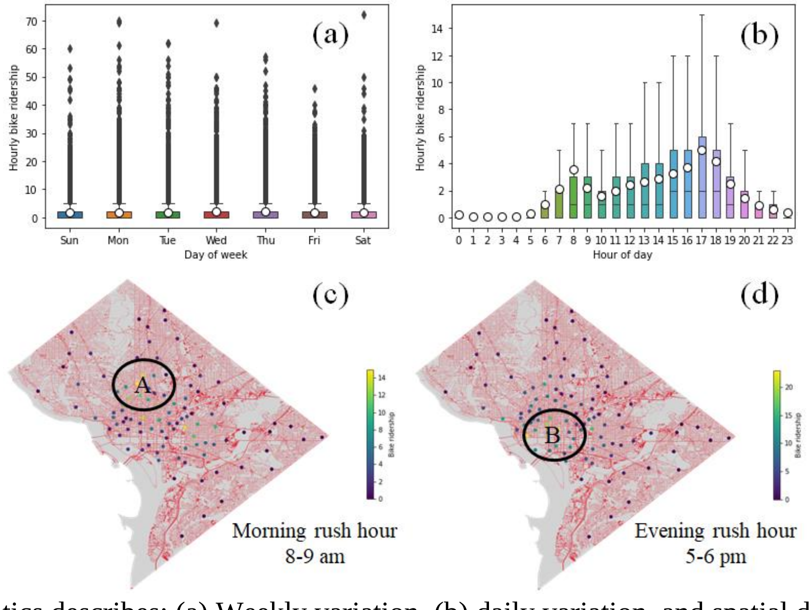


Fig. 1 The schematics describes: (a) Weekly variation, (b) daily variation, and spatial distribution of bike-sharing ridership in the moring (c) and (d) evening rush hours.

4.2. Variable Selection for Spatiotemporal Prediction of Bike-Sharing Ridership

Since the bike-sharing ridership varies over time and space, there must be certain close associations between its distribution patterns and potential temporal & spatial influential factors (shown in Table 2). To determine which variables that are important for bike ridership prediction, we input the corresponding data sets (shown in Eq. 2 and Table 2) into the SpVAR-LASSO model for variable selection.

Hyperparameter settings: In the SpVAR-LASSO model, there were two hyperparameters significantly affecting the model performance. (1) The time lag n of bike-sharing ridership: To determine the optimal n, we computed and visualized the autocorrelation function (ACF) of bike ridership in Fig. 2(a) and the ACF curve suggested that the bike-sharing ridership in the last week still exhibited strong correlations with that at the current moment. Thus, we set the optimal n as 168 (7 days * 24 hours). (2) The LASSO regularization parameter γ: To enhance the model accuracy and generalizability, we introduced a 10-fold cross-validation to decide the optimal γ by setting γ at the range of 1e-4 to 100 (Baumanis et al., 2023). As displayed in Fig. 2(b), when γ ranged from 1e-4 to 100, the corresponding MSE of the SpVAR-LASSO model first remained almost unchanged and then increased rapidly and that the MSE reached the minimum when $\gamma = 0.0001$.

Variable selection: We run the SpVAR-LASSO model for each hour in March 2020 with a total of 576 times and derived corresponding coefficient estimates of each independent variable at different time periods. Then we selected the variables whose number of non-zero coefficient estimates were over 90 based on 576 regression models and illustrated the distribution of coefficient estimates of these variables in Fig. 2(c). We found that 3 public transit factors such as the dynamic number of metro & bus stops and the static number of bus stops around centroids and 6 time-lagged variables including the bike ridership data in the last 24th, 72nd, 96th, 120th, 144th, and 168th hours were identified to be more important variables than others (i.e., built environment factors and other time-lagged ridership variables) when modelling bike-sharing ridership data. In particular, the dynamic number of bus and metro stops around 100 clustered centroids exerted a greater influence on the spatiotemporal variations of bike-sharing ridership in Washington D.C. Furthermore, these findings were reliable and significant due to a lower MSE of the SpVAR-LASSO model (Fig. 2(d)).

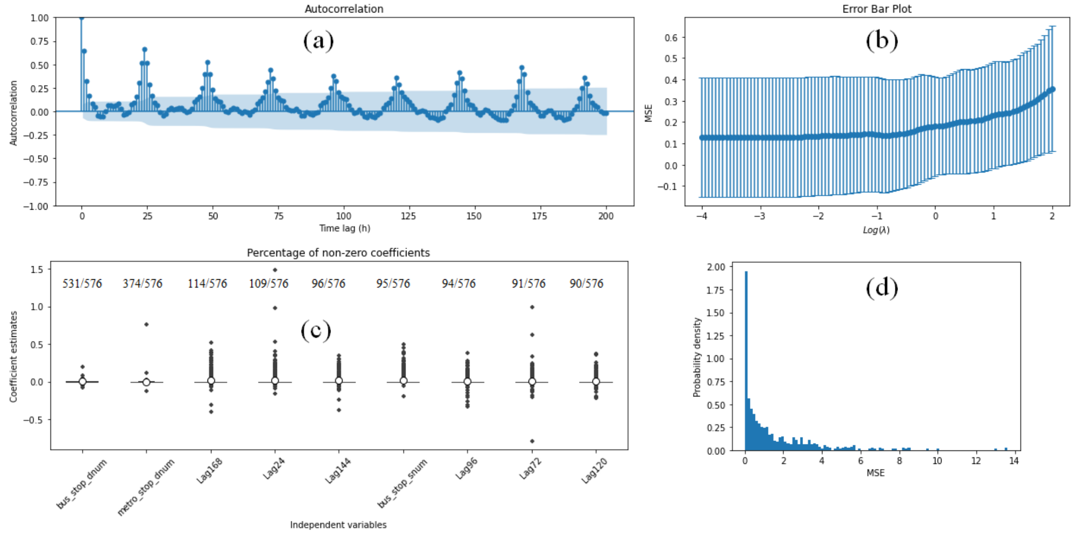


Fig. 2 The schematics describes: (1) Autocorrelation function of bike ridership, (b) error bar plot of 10-fold cross validation for γ, (c) distribution of coefficient estimates of variables with the largest number of non-zero regression coefficients, and (d) probability density function of MSE from the SpVAR-LASSO model.

4.3. Short-Term Bike-Sharing Ridership Prediction and Model Comparison

4.3.1. Experimental Design

Based on the selected variables from the SpVAR-LASSO model, we conducted the following experimental design within the XGBoost framework to investigate the effects of different feature groups and data sets on the prediction accuracy of bike-sharing ridership, as well as its performance comparison with other models:

Feature input: We built a set of scenarios with different feature groups as model inputs. First, we set a basic scenario with temporal features and meteorological factors as model inputs. Then we incrementally added more feature groups into the XGBoost prediction models until all feature groups were incorporated: (1) Adding spatial features; (2) Adding O-D flow graph attributes; (3) Adding public transit factors.

Dataset selection and splitting: Since the bike ridership dataset contained lots of zero ridership and also exhibited significant discrepancy in the rush hours (6-10 am & 4-8 pm) and in the non-rush hours, we built three types of datasets: all dataset, non-zero dataset, and rush-hour dataset as model inputs to study whether different datasets could affect the prediction accuracy. Also, we randomly split the three datasets into the training set (80%) and the testing set (20%), separately ready for bike-sharing ridership prediction.

Model comparison: In addition to the easily interpretable XGBoost model, we constructed another four baseline models: (1) Hourly Average (HA), (2) Autoregressive Intergated Moving Average (ARIMA), (3) Ordinary Least Squares (OLS), and (4) Multi-Layer Perception (MLP) (Yang et al., 2020) to compare their predictive performances, in terms of prediction accuracy, generalizability, and interpretability.

4.3.2. Model Evaluation and Interpretation

According to the above experimental design, we incorporated the corresponding datasets and feature groups into the five models for bike-sharing ridership prediction, and their predictive performances under different scenario settings were presented in Table 4.

Table 4 Comparison of prediction accuracy

Model	Dataset	All dataset (N=57,600)		Non-zero dataset (N=23,162)		Rush-hour dataset (N=19,200)	
	Scenario	MSE	r^2	MSE	r^2	MSE	r^2
HA	T	15.57	-2.93	24.79	-5.22	29.27	-1.64
ARIMA		19.97	-0.58	N/A			
OLS	T+M	4.54/4.33	0.65/0.64	9.45/9.36	0.58/0.59	7.43/7.61	0.68/0.66
	T+M+S	4.14/3.90	0.68/0.68	8.71/8.41	0.61/0.63	6.80/6.25	0.70/0.73
	T+M+S+G	4.10/3.89	0.68/0.68	8.62/8.31	0.62/0.64	6.57/6.79	0.72/0.69
	T+M+S+PT	4.12/3.88	0.68/0.68	8.65/8.40	0.61/0.63	6.63/6.84	0.71/0.69
	T+M+S+G+PT	4.08/3.87	0.68/0.68	8.57/8.30	0.62/0.64	6.55/6.77	0.72/0.69
MLP	T+M	3.31/3.88	0.74/0.68	7.53/7.38	0.66/0.68	5.48/6.14	0.76/0.72
	T+M+S	3.10/3.57	0.76/0.71	6.74/6.62	0.70/0.71	4.97/4.86	0.78/0.79
	T+M+S+G	2.90/3.47	0.77/0.71	6.34/6.81	0.72/0.70	4.34/5.49	0.81/0.75
	T+M+S+PT	2.97/3.46	0.77/0.71	6.53/6.68	0.71/0.71	4.55/5.57	0.80/0.75
	T+M+S+G+PT	2.88/3.49	0.78/0.71	6.22/6.86	0.72/0.70	4.68/5.84	0.80/0.74
XGBoost	T+M	3.07/3.70	0.76/0.70	6.46/7.14	0.71/0.69	4.26/5.45	0.82/0.75
	T+M+S	2.83/3.41	0.78/0.72	5.98/6.56	0.73/0.71	4.16/4.96	0.82/0.79
	T+M+S+G	2.82/3.42	0.78/0.72	5.95/6.63	0.73/0.71	4.02/5.46	0.83/0.75
	T+M+S+PT	2.68/3.27	0.79/0.73	5.80/6.37	0.74/0.72	3.96/5.19	0.83/0.77
	T+M+S+G+PT	**2.67/3.25**	**0.79/0.73**	**5.71/6.44**	**0.75/0.72**	**3.93/5.30**	**0.83/0.76**

Note: a/b in MSE and R^2 columns represent training and testing MSE and R^2, respectively.

Model evaluation: Regardless of the types of datasets and feature groups as model inputs, the XGBoost model achieved the best prediction preformance of bike-sharing ridership with the lowest MSE and highest r^2 in both training and testing sets, followed by the MLP, OLS, HA, and ARIMA models. Therefore, the XGBoost model demonstrated a higher predictive accuracy, as well as a stronger generalizability than other models in predicting the bike-sharing ridership.

Effect of feature groups: With the temporal & spatial features and meteorlogy as model inputs, the OLS, MLP, and XGBoost models all obtained a satisfactory prediction results of the bike-sharing ridership with the MSE less than 4.2 and r^2 higher than 0.65. This implied significant spatiotemporal distribution patterns of bike-sharing ridership in Washington D.C. Regarding all dataset, simply adding graph attributes into the prediction models could not greatly improve the prediction accuracy of the OLS and XGBoost models while the introduction of graph attributes worked well for the MLP model. By contrast, only adding public transit factors into the prediction models can improve the prediction accuracy of the XGBoost model while it did not work well for the OLS and MLP models. When we added graph attributes and public transit factors into the prediction models simultaneously, the predictive accuracy of the OLS, MLP, and XGBoost models did not improve a lot, as compared with the scenario where graph attributes and public transit were separately input into the prediction models. One of the main reasons was that all dataset contained too many zero data such that the corresponding graph attributes and public transit factors were no use in predicting ridership.

Effect of dataset selection: Inspired by this, we further compared the model performances under various datasets. We found that although the models' predictive performances over non-zero dataset and rush-hour dataset were worse than that over all dataset due to the fact that the zero ridership was easier to predict, an addition of graph attributes and public transit factors into the models, whether separately or simultaneously, both improved the prediction accuracy with a lower MSE, particularly for the MLP and XGBoost models over the non-zero dataset. This suggested that when the hourly ridership was non-zero (often excluding the time period from 11 pm to 5 am), the corresponding graph attributes and public transit factors were valuable and helpful to predict the bike-sharing ridership. More specifically, the XGBoost and MLP models with all feature groups as inputs achieved the best prediction accuacy over the same dataset, especially for the non-zero dataset. However, at the same time, we also still found that public transit factors were more useful for improving the prediction accuracy of the XGBoost model while graph attributes were more valuable for enhancing the prediction accuracy of the MLP model, which would not vary over any of the three datasets.

5. Conclusion and Discussions

The accurate prediction of bike-sharing ridership is essential for bike management and fleet rebalancing to mitigate the spatial and temporal imbalance of sharing bikes and improve user satisfaction with bike-sharing systems (BSS). However, most of the literature emphasized more on the optimizaiton and improvement of model itself while almost ignoring the effect of domain knowledge and feature engineering on the prediction accuracy. In this project, we proposed an easily interpretable feature engineering approach to predicting the short-term bike-sharing ridership by incrementally adding more feature groups into the prediction model to see their impacts on the prediction accuracy. Specifically, using bike trip data in Washington, D.C., we first employed a K-means clustering method to group 352 bike stations into 100 clusters based on their spatial proximity to alleviate the effects of ridership biases and uncertainties among adjacent bike stations on the prediction accuracy. Then we constructed a set of time-dependent origin-destination (O-D) flow graphs and further extracted corresponding graph attributes to characterize the houly mobility patterns among the 100 clustered centroids. Also, we extracted four public transit factors that may affect the bike-sharing ridership from transit networks and schedules.

By aggregating the potential temporal and spatial influential factors, we first established a spatial vector autoregressive LASSO model for variable selection to identify those important variables significantly contributing to the bike ridership prediction. The results suggested that public transit factors like the dynamic and static number of bus & metro stops and the static number of bus stops around centroids and temporal features such as the ridership data in the last 24th,72nd,96th,120th,144th,168th hours were identifed to be more important variables than others. Then we integrated these important features and meteorology, as well as the corresponding time-lagged graph attributes, into the XGBoost model to explore the effects of differrent feature groups on the prediction accuracy. In addition, we examined the impacts of different data sets on the model predictive performance. The results indicated that the XGBoost model outperformed other models in both prediction accuracy, generalizability, and interpretability, with a lower MSE and higher r^2 in both training and testing sets. Furthermore, the introduction of OD flow graph attributes and public transit information could greatly improve the prediction performance, particularly for public transit information over all, non-zero, & rush-hour datasets and graph attributes over rush-hour dataset.

The findings of this project not only demonstrated an easily interpretable feature engineering approach to predicting future bike-sharing ridership in a high accuracy, but also provided important insights into the associations between bike-sharing ridership and different influential factors. In particular, the enhancement of public transit information on the prediction accuracy at least partially implied the significant interactions between bike-sharing systems and public transit networks. However, this project has some limitations, for example, unavailability of transit ridership restricting in-deeper analysis of effects of public transit on bike-sharing system, as well as only a study area lacking generalizability. In addition, there are some possibilities of combining graph neural networks (Xu et al., 2021), Bayesian methods (Peng et al., 2023), with XGBoost to solve deeper prediction tasks (i.e., uncertainty quantification) in future research.

References

Baumanis, C., Hall, J., Machemehl, R., 2023. A machine learning approach to predicting bicycle demand during the COVID-19 pandemic. Research in Transportation Economics.

Bigazzi, A., Wong, K., 2020. Electric bicycle mode substitution for driving, public transit, conventional cycling, and walking. Transportation Research Part D: Transport and Environment 85.

Cantelmo, G., Kucharski, R., Antoniou, C., 2020. Low-dimensional model for bike-sharing demand forecasting that explicitly accounts for weather data. Transportation Research Record: Journal of the Transportation Research Board 2674, 132-144.

Chen, L., Ma, X., Nguyen, T.-M.-T., Pan, G., Jakubowicz, J., 2017. Understanding bike trip patterns leveraging bike sharing system open data. Frontiers of Computer Science 11, 38-48.

Cho, J.-H., Ham, S.W., Kim, D.-K., 2021. Enhancing the accuracy of peak hourly demand in bike-sharing systems using a graph convolutional network with public transit usage data. Transportation Research Record: Journal of the Transportation Research Board 2675, 554-565.

Fan, A., Chen, X., Wan, T., 2019. How have travelers changed mode choices for first/last mile trips after the introduction of bicycle-sharing systems: An empirical study in Beijing, China. Journal of Advanced Transportation 2019, 1-16.

Ferguson, B., Sanguinetti, A., 2021. Facilitating micromobility for first and last mile connection with public transit through environmental design: A case study of California Bay Area rapid transit stations. Proceedings of the Design Society 1, 1577-1586.

Gao, C., Chen, Y., 2022. Using machine learning methods to predict demand for bike sharing, in: Stienmetz, J.L., Ferrer-Rosell, B., Massimo, D. (Eds.), Information and Communication Technologies in Tourism 2022. Springer.

Hartigan, J.A., Wong, M.A., 1979. Algorithm AS 136: A K-Means Clustering Algorithm. Applied Statistics 28.

Hasselwander, M., Nieland, S., Dematera-Contreras, K., Goletz, M., 2022. MaaS for the masses: Potential transit accessibility gains and required policies under mobility-as-a-service.

Jaber, A., Csonka, B., Juhasz, J., 2022. Long term time series prediction of bike sharing trips: A cast study of Budapest City, 2022 Smart City Symposium Prague (SCSP), pp. 1-5.

Jin, C., Lee, G., 2019. Exploring spatiotemporal dynamics in a housing market using the spatial vector autoregressive Lasso: A case study of Seoul, Korea. Transactions in GIS 24, 27-43.

Kaltenbrunner, A., Meza, R., Grivolla, J., Codina, J., Banchs, R., 2010. Urban cycles and mobility patterns: Exploring and predicting trends in a bicycle-based public transport system. Pervasive and Mobile Computing 6, 455-466.

Li, X., Xu, Y., Zhang, X., Shi, W., Yue, Y., Li, Q., 2023. Improving short-term bike sharing demand forecast through an irregular convolutional neural network. Transportation Research Part C: Emerging Technologies 147.

Li, Y., Zheng, Y., Zhang, H., Chen, L., 2015. Traffic prediction in a bike-sharing system, Proceedings of the 23rd SIGSPATIAL International Conference on Advances in Geographic Information Systems, pp. 1-10.

Lin, L., He, Z., Peeta, S., 2018. Predicting station-level hourly demand in a large-scale bike-sharing network: A graph convolutional neural network approach. Transportation Research Part C: Emerging Technologies 97, 258-276.

Lin, P., Weng, J., Hu, S., Alivanistos, D., Li, X., Yin, B., 2020. Revealing spatio-temporal patterns and influencing factors of dockless bike sharing demand. IEEE Access 8, 66139-66149.

Liu, H.-C., Lin, J.-J., 2019. Associations of built environments with spatiotemporal patterns of public bicycle use. Journal of Transport Geography 74, 299-312.

Madapur, B., Madangopal, S., Chandrashekar, M.N., 1970. Micro-mobility infrastructure for redefining urban mobility. European Journal of Engineering Science and Technology 3, 71-85.

Oeschger, G., Carroll, P., Caulfield, B., 2020. Micromobility and public transport integration: The current state of knowledge. Transportation Research Part D: Transport and Environment 89.

Peng, B., Yang, K., Dong, X., 2023. Variable selection for quantile autoregressive model: Bayesian methods versus classical methods. Journal of Applied Statistics, 1-33.

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

Saberi, M., Ghamami, M., Gu, Y., Shojaei, M.H., Fishman, E., 2018. Understanding the impacts of a public transit disruption on bicycle sharing mobility patterns: A case of Tube strike in London. Journal of Transport Geography 66, 154-166.

Saberi, M., Mahmassani, H.S., Brockmann, D., Hosseini, A., 2016. A complex network perspective for characterizing urban travel demand patterns: graph theoretical analysis of large-scale origin–destination demand networks. Transportation 44, 1383-1402.

Tavassoli, K., Tamannaei, M., 2020. Hub network design for integrated Bike-and-Ride services: A competitive approach to reducing automobile dependence. Journal of Cleaner Production 248.

Xu, Y., Paliwal, M., Zhao, X., 2021. Real-time forecasting of dockless scooter-sharing demand: A context-aware spatio-temporal multi-graph convolutional network approach. arXiv preprint arXiv:2111.01355.

Yang, Y., Heppenstall, A., Turner, A., Comber, A., 2019. A spatiotemporal and graph-based analysis of dockless bike sharing patterns to understand urban flows over the last mile. Computers, Environment and Urban Systems 77.

Yang, Y., Heppenstall, A., Turner, A., Comber, A., 2020. Using graph structural information about flows to enhance short-term demand prediction in bike-sharing systems. Computers, Environment and Urban Systems 83.

Yang, Y., Shao, X., Zhu, Y., Yao, E., Liu, D., Zhao, F., Zou, Y., 2023. Short-term forecasting of dockless bike-sharing demand with the built environment and weather. Journal of Advanced Transportation 2023, 1-13.

Zellner, A., 1962. An efficient method of estimating seemingly unrelated regressions and tests for aggregation bias. Journal of the American statistical Association 57, 348-368.

Zhou, Y., Li, Y., Zhu, Q., Chen, F., Shao, J., Luo, Y., Zhang, Y., Zhang, P., Yang, W., 2019. A reliable traffic prediction approach for bike-sharing system by exploiting rich information with temporal link prediction strategy. Transactions in GIS 23, 1125-1151.

Zuo, T., Wei, H., Chen, N., Zhang, C., 2020. First-and-last mile solution via bicycling to improving transit accessibility and advancing transportation equity. Cities 99.