# Variable Selection for Bike-Sharing Demand Prediction: A Spatial Vector Autoregressive LASSO Approach

Kai-Fa Lu

***Abstract* - Accurate prediction of bicycle demand is pivotal to facilitate decision-makers in planning and designing bike-sharing systems towards efficiency and equity. Since bike-sharing demand fluctuates spatiotemporally, it is important to incorporate the two dimensions into the prediction models. Although previous studies have attempted to employ a spatial vector autoregressive (SpVAR) model to reveal and predict the spatiotemporal dynamics of bike-sharing demand, the SpVAR model assumes spatial homogeneity of the variables due to insufficient degrees of freedom, conflicting with the fact that bike-sharing demand is usually heterogeneous. To fill the gap, we proposed a spatial vector autoregressive LASSO (SpVAR-LASSO) approach without the homogeneity assumption where SpVAR modelled both temporal and spatial dynamics of bike-sharing demand and LASSO identified those most important predictors to avoid overfitting and improve predictive accuracy. Using bike trip data in Washington D.C. as a case study, results suggested that the spatial predictor (the number of nearby transit stops) and temporal predictors (i.e., the bike demand in the last 24 hours and 168 hours) both exerted a greater influence on bicycle demand prediction than other predictors. Moreover, the SpVAR-LASSO model outperformed other baseline models in predicting bicycle demand, in terms of both interpretability and accuracy.**

***Keywords***: Bike-Sharing System; Demand Prediction; Spatial Vector Autoregression; LASSO; Variable Selection

## I. Introduction

Bike-sharing system is ubiquitous globally due to its high flexibility and accessibility in providing an alternative mobility option, particularly over the first mile and last mile, that traditional public transit cannot cover. However, this flexibility also brings large uncertainty in bike-sharing demand, as well as serious challenges about how to design and plan bike-sharing system towards efficiency and equity. A key to address this issue is to accurately predict the bike-sharing demand in the context of time and space, thus providing essential insights into planning efforts seeking to implement bicycle facilities [1].

Driven by this, lots of researchers have attempted to develop different prediction models of bike-sharing demand. It is well-known that bike-sharing demand varies over time and space [2]. For instance, Wang et al. [2] observed that bike-sharing demand peaked at two periods on a weekday: 6:00-10:00 am and 3:00-7:00 pm in Montreal, Canada, and different spatial distributions in the two periods. In this context, most of the prediction models fall into three categories: Time-series models, spatial regression models, and spatiotemporal prediction models (**Table I**).

TABLE I
Summary of Temporal and Spatial Prediction Models of Bike-Sharing Demand

| Type | Model | Description |
|---|---|---|
| Time-series model | Hourly Average (HA)<br>Autoregression (AR)<br>Autoregressive Moving Average (ARMA) [3]<br>Vector Autoregression [4] | Analyze and forecast bike-sharing demand data that varies over time |
| Spatial regression model | Geographically Weighted Regression (GWR) [5]<br>Spatial Autoregression [6]<br>Spatially Varying Coefficient [2] | Analyze and model spatially dependent bike demand data |
| Spatial and temporal prediction model | Spatial Vector Autoregression (SVA) [7]<br>Machine Learning and Deep Learning Models [1, 8-10] | Analyze and model bike data that varies over space and time |

The idea of time-series models is to model and analyze the temporal patterns behind historical data of bike-sharing demand and further extrapolate and predict future demand. For instance, Toman et al. [4] used vector autoregression to account for trend and a weekly seasonal structure of bike demand data and further produce medium-term forecasts for sharing bike use. While for spatial regression models, the common practice is to predict the bike-sharing demand at a certain bike station based on its spatial relationships with neighboring bike stations. Guidon et al. [6] employed spatial autoregression model and booking data from an electric bike-sharing system to estimate the forecasts of bike-sharing demand. However, either of the two prediction models only considers temporal or spatial dependencies of bike-sharing demand while ignoring joint effects of spatiotemporal features on the accuracy of bike demand prediction.

To solve this issue, Beenstock and Felsenstein [11] proposed a spatial vector autoregression (SpVAR) model that integrated both temporal and spatial dynamics through modelling with spatiotemporal lags of variables [7]. However, this method has not yet been broadly applied to bike demand prediction. Instead, researchers are more likely to adopt machine learning and deep learning methods, i.e., convolution neural networks [8], Long-Short Term Memory [9], and graph neural networks [10], to model both temporal and spatial characteristics of bike-sharing data for demand prediction with a higher accuracy but usually with low interpretability. Motivated by this dilemma, this study aims to explore the potential of the SpVAR model in predicting spatiotemporal demand of sharing bike usage, in terms of both accuracy and interpretability. However, a naïve SpVAR cannot capture local variations of bike-sharing demand that is spatially heterogeneous due to insufficient degrees of freedom [7]. Since the least absolute shrinkage and selection operator (Lasso) can resolve the insufficient issue of degree of freedom by shrinking

Submitted for grading on April 15, 2023.

K.-F. Lu is with International Center for Adaptation Planning and Design, College of Design, Construction and Planning, University of Florida, PO Box 115706, Gainesville, FL 32611-5706, USA (email: kaifa.lu@ufl.edu).

some coefficients to zero [12], the combination of SpVAR and LASSO may be a solution to further improve model accuracy by reducing the variance of the predicted values. Also, Jin and Lee [7] have applied this to explore spatiotemporal dynamics in a housing market, but rare work is done to explore its potentials in bike-sharing demand prediction.

In this project, our goal is to develop a novel spatial vector autoregressive LASSO (SpVAR-LASSO) model to accurately predict bike demand. In this approach, SpVAR is used to model the spatiotemporal dynamics of bicycle demand, as well as the impacts of different influential factors while LASSO is used for variable selection to identify important variables contributing to the prediction accuracy and improve model interpretability and generalizability. Finally, bike trip data of 2019 in Washington D.C. is employed to further demonstrate the prediction model.

The rest of this study is organized below: Section II presents the trip data of bike demand and its influencing factors. Section III presents a SpVAR-LASSO approach to predict bike-sharing demand. Section IV describes the predictive performance of the proposed approach. Section V summarizes the major findings.

## II. Data

### A. Bike-Sharing Demand

Using bike trip data from Capital Bikeshare program during 2019 in Washington D.C., we illustrated the spatial distribution of hourly ridership at 352 bike stations, as well as the variations of hourly average ridership in 12 months and a day in **Figure 1**. The hourly ridership at 352 bike stations displayed significant spatiotemporal distribution patterns summarized as follows:

(1) The downtown areas in Washington D.C. showed higher hourly ridership than the surroundings.
(2) The peak of the hourly ridership happened during 6-9 am and 4-7 pm in April-May and September-October due to increased short-trip commuting activities.

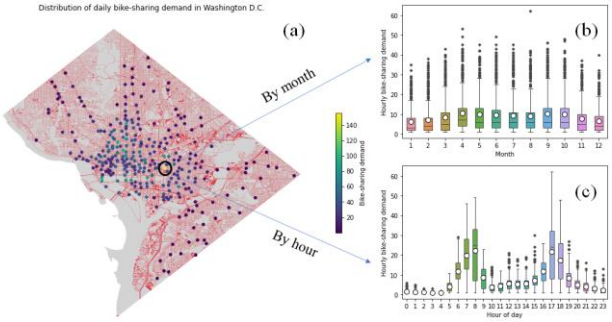


**Figure 1.** Spatiotemporal distribution of bike-sharing demand: (a) Spatial distribution of hourly demand, (b) Hourly demand in 12 months, and (c) Hourly demand in a day.

### B. Influential Factors

Distribution of bike-sharing demand is shaped by numerous factors, for instance, hour of a day, day of week, demographics, built environment, and weather [5]. Since we model the general spatiotemporal dynamics of bike-sharing demand, the weather variables with high uncertainty are not considered in this project [2]. In this context, we divide variables influencing the forecasts of bike-sharing demand into three types: 1) Temporal predictors including the bike trip data in the past several hours; (2) Spatial predictors, i.e., trip data in neighboring bike stations, population density, land use diversity, bike lane density, housing unit density, and the number of nearby transit stops; (3) Graph-based predictors, i.e., degree, PageRank, and betweenness [13].

Additionally, there are two key issues when calculating these influential factors: 1) Graph representation of bicycle network; 2) Aggregation of influential factors. For the former, we used hourly bike trip data to construct a set of time-dependent graphs where bike stations were considered as nodes of a graph and the volume of bike trips between two nodes were used to generate the edges and weights over these edges [13]. Then we extracted graph attributes including degree, PageRank, and betweenness to represent origin-destination (OD) flow patterns, as illustrated in **Figure 2**. For the latter, we aggregated temporal, spatial, and graph predictors into corresponding bike stations, specifically, aggregation of spatial predictors was based on the geographical proximity between census tracts and bike stations, to facilitate station-level bike-sharing demand predictions.

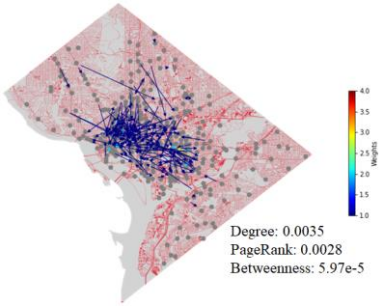


**Figure 2.** Graph representation of bike-sharing networks.

## III. Methodology

### A. Spatial Vector Autoregression (SpVAR)

The SpVAR model extends the time-series model spatially to deal explicitly with spatiotemporal autocorrelation in variables influencing bike demand [7], which is formalized as follows:

$$y_{i,t} = \beta_0 + \beta_1 \boldsymbol{x}_i + \beta_2 \sum_j w_{ij} y_{j,t} + \beta_3 \boldsymbol{g}_{i,t-1} + \beta_4 \boldsymbol{y}_{i,t-1\sim t-n} + \varepsilon_{i,t} \quad (1)$$

Where $y_{i,t}$ means the bike demand of bike station $i$ at time $t$; $\boldsymbol{x}_i$ denotes a $5 \times 1$ vector affiliated to bike station $i$ including population density, land use diversity, bike lane length, housing unit density, and the number of nearby transit stops; $\sum_j w_{ij} y_{j,t}$ is a geographically weighted sum of the bike-sharing demand of neighboring bike stations at time $t$ where $w_{ij}$ is computed based on the geographical proximity (the inverse of geographic distance between bike station $i$ and $j$); $\boldsymbol{g}_{i,t-1}$ represents a $3 \times 1$ vector affiliated to bike station $i$ at time $t-1$ including degree, PageRank, and betweenness; $\boldsymbol{y}_{i,t-1\sim t-n}$ denotes a $n \times 1$ vector describing the bike demand of bike station $i$ in the last $n$ hours; $\varepsilon_{i,t}$ is the error item of bike station $i$ at time $t$; $\beta_0$ represents station-specific effect; $\beta_0 \sim \beta_4$ are the coefficients of parameter.

For the whole bike network consisting of 352 bike stations, the SpVAR model can be written into its matrix form:

$$\boldsymbol{Y_t} = \boldsymbol{\beta_0} + \boldsymbol{\beta_1 X} + \boldsymbol{\beta_2 W Y_t^*} + \boldsymbol{\beta_3 G_{t-1}} + \boldsymbol{\beta_4 Y_{t-1\sim t-n}} + \boldsymbol{\varepsilon_t} \quad (2)$$

Since the general matrix form is nonlinear due to the spatial autoregressive item $\boldsymbol{Y_t^*}$, Beenstock and Felsenstein [11] used a reduced form as an equivalent alternative as follows:

$$(\boldsymbol{I}-\boldsymbol{\beta}_2\boldsymbol{W})\boldsymbol{Y}_t=\boldsymbol{\beta}_0+\boldsymbol{\beta}_1\boldsymbol{X}+\boldsymbol{\beta}_3\boldsymbol{G}_{t-1}+\boldsymbol{\beta}_4\boldsymbol{Y}_{t-1\sim t-n}+\boldsymbol{\varepsilon}_t \quad (3)$$

$$\boldsymbol{Y}_t=\boldsymbol{\Gamma}_0+\boldsymbol{\Gamma}_1\boldsymbol{X}+\boldsymbol{\Gamma}_3\boldsymbol{G}_{t-1}+\boldsymbol{\Gamma}_4\boldsymbol{Y}_{t-1\sim t-n}+\boldsymbol{u}_t \quad (4)$$

Where $\boldsymbol{\Gamma}_0=(\boldsymbol{I}-\boldsymbol{\beta}_2\boldsymbol{W})^{-1}\boldsymbol{\beta}_0$ , $\boldsymbol{\Gamma}_1=(\boldsymbol{I}-\boldsymbol{\beta}_2\boldsymbol{W})^{-1}\boldsymbol{\beta}_1$ , $\boldsymbol{\Gamma}_3=(\boldsymbol{I}-\boldsymbol{\beta}_2\boldsymbol{W})^{-1}\boldsymbol{\beta}_3$, $\boldsymbol{\Gamma}_4=(\boldsymbol{I}-\boldsymbol{\beta}_2\boldsymbol{W})^{-1}\boldsymbol{\beta}_4$. Hence, the statistical significance of $\boldsymbol{\Gamma}_1,\boldsymbol{\Gamma}_3,\boldsymbol{\Gamma}_4$ denotes the effects of spatial, graph, and temporal predictors, respectively.

To estimate the coefficients of parameters, we introduce the seemingly unrelated regression (SUR) [14] to solve $Y=X\beta+u$ whose disturbances are correlated contemporaneously. Then the coefficients of the equation are derived as follows:

$$\hat{\beta}=(X'\hat{\Omega}^{-1}X)^{-1}X'\hat{\Omega}^{-1}Y \quad (5)$$

Where $\hat{\Omega}$ denotes the covariance matrix of disturbance.

### *B. Least Absolute Shrinkage and Selection Operator (LASSO)*

The LASSO regression applies to model cases whose number of variables is larger than or closer to the number of data points. The main idea of LASSO is to regulate the coefficients of the regression model by imposing a penalty on its residual sum of squares with a parameter $\gamma$ to minimize mean square error [7]:

$$\beta=\underset{\beta}{\operatorname{argmin}}\left[(Y-X\beta)'(Y-X\beta)+\gamma\textstyle\sum_j|\beta_j|\right] \quad (6)$$

The estimate of coefficients is:

$$\hat{\beta}=(X'X+\gamma W^{-1})^{-1}X'Y \quad (7)$$

Where $W^{-1}$ is a diagonal matrix with diagonal elements $|\hat{\beta}_j|$ and those less significant or small coefficients are forced by the constraints term of the LASSO to shrink to zero. As a result, the LASSO renders both coefficient estimate and variable selection available to keep most important variables whose coefficients are non-zero. Therefore, the introduction of LASSO may solve the issue of the insufficient degrees of freedom facing SpVAR.

### *C. SpVAR-LASSO Model for Bike-Sharing Demand Prediction*

Due to the simplicity and efficiency in model selection and estimation, the LASSO has been widely adopted to improve the performance of generalized linear model [15] and vector autoregressive model [16]. The project extends the applicability of the LASSO into the SpVAR to develop a SpVAR-LASSO approach to predicting the bike-sharing demand:

$$\boldsymbol{\Gamma}=\underset{\boldsymbol{\Gamma}}{\operatorname{argmin}}\left[(\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\Gamma})'(\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\Gamma})+\gamma\textstyle\sum_j|\boldsymbol{\Gamma}_j|\right] \quad (8)$$

By combining SUR (**Eq**. 5) and LASSO estimations (**Eq**. 8), the coefficients of the SpVAR-LASSO model can be derived as:

$$\hat{\boldsymbol{\Gamma}}=(\boldsymbol{X}'\hat{\boldsymbol{\Omega}}^{-1}\boldsymbol{X}+\gamma\boldsymbol{W}^{-1})^{-1}\boldsymbol{X}'\hat{\boldsymbol{\Omega}}^{-1}\boldsymbol{Y} \quad (9)$$

The SpVAR-LASSO model does not require variable blocks or homogeneity assumptions in contrast to the SpVAR model, instead, it excels in variable selection because the Lasso is free from lack of degrees of freedom. Thus, the SpVAR-LASSO model is useful to explore heterogeneous relationships between bike-sharing demand and its influential variables in the context of space and time [7].

## IV. Experiments And Results

### *A. Experimental Design and Parameter Settings*

The station-level aggregated bike trip data and its influential variables during 2019 in Washington D.C. were used to solve the SpVAR-LASSO approach for estimating the coefficients of parameters and predicting bike-sharing demand. To enhance its model accuracy and generalizability, we introduced the 10-fold cross-validation to decide the optimal hyperparameter $\gamma$ where $\gamma$ ranged from 1e-4 to 100 [1]. **Figure 3** showed when $\gamma$ ranged from 1e-4 to 100, the mean square error (MSE) displayed first slightly declining and then gradually increasing trend. The MSE reached the minimum when $\gamma=0.02656$.

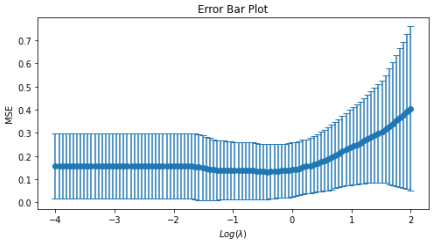


**Figure 3**. Plotted $\gamma$ cross validation curve.

The other hyperparameter is the time lag $n$ of bike-sharing demand. To determine the optimal $n$, we calculated and plotted the autocorrelation function (ACF) curve in **Figure 4**. It can be easily seen that the bike-sharing demand in the last week still exerted an influence on the current bike-sharing demand and therefore, we set the optimal $n$ as 168 (7*24).

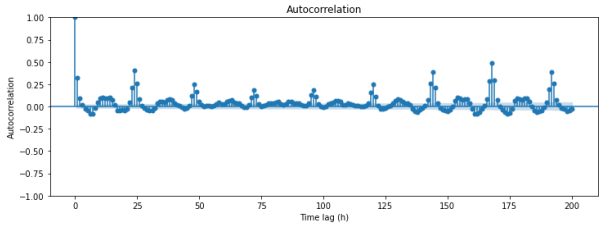


**Figure 4.** Autocorrelation function of bike-sharing demand.

### *B. Model Estimate and Variable Selection*

We run the SpVAR-LASSO model for each hour in 2019 and derived corresponding coefficient estimates of each predictor at different time periods. Then we identified those most important variables whose coefficients were nonzero with a probability of $> 0.2$ based on 8,592 regression models (an hour corresponded to one regression model) and illustrated them in **Figure 5**. We found that spatial predictor such as the number of nearby transit stops and 10 temporal predictors including the bike trip data in the last 1, 23, 24, 25, 48, 96, 120, 144, 167, and 168 hours were more important than other predictors such as those graph-based predictors to the accurate forecasts of the bike-sharing demand. Furthermore, the number of nearby transit stops, bike trip data in the last 24 hours (one day) and 168 hours (one week) had the most contributions to forecasting bike demand, as revealed by higher percentage, larger mean, and lower variance of nonzero coefficients across the 8,592 regression equations in 2019.

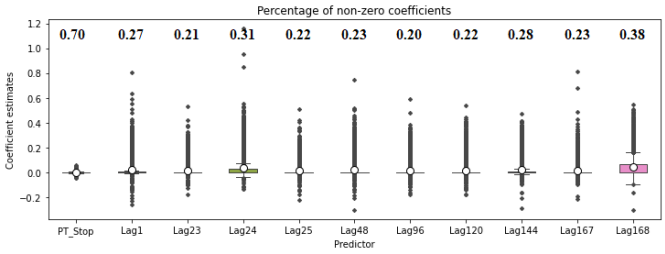


**Figure 5**. Boxplot of coefficient estimates of most important predictors to the bike-sharing demand prediction.

Then we plotted hourly variations of coefficients of the three most important predictors in **Figure 6** to further explore if and how their effects on the prediction accuracy of the bike demand varied by time. Results suggested that the coefficient estimates of the bike trip data in the last 24 hours and 168 hours exhibited significant temporal patterns that two peaks separately occurred during the morning and evening rush hours while the coefficient estimates of the number of nearby transit stops hardly varied by time, showing higher stability and lower variance.

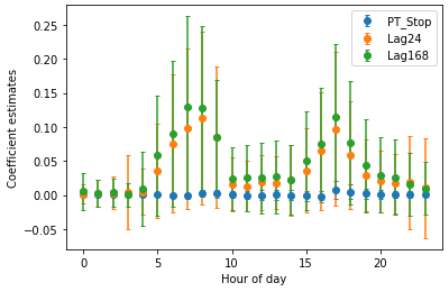


**Figure 6**. Hourly variations of coefficient estimates of most important predictors to the bike-sharing demand prediction.

### *C. Model Evaluation and Performance Comparison*

Based on these coefficient estimates, we predicted the hourly bike-sharing demand in 2019 and computed the MSE between ground truth observations and predictions. Then we plotted the probability density function and temporal distribution of hourly MSE in 2019, as shown in **Figure 7**.

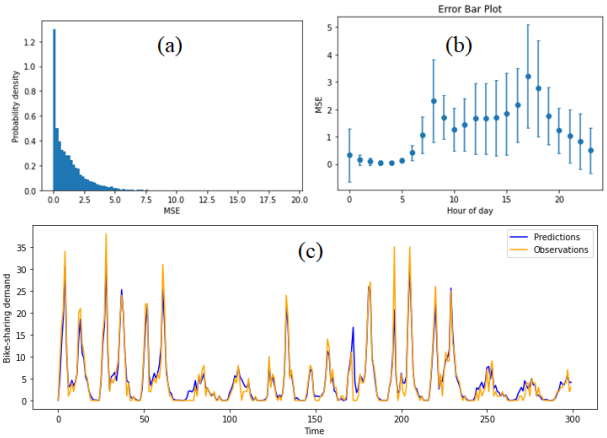


**Figure 7**. Predictive performance of bike-sharing demand.

**Figure 7**(a) suggested that the majority of hourly MSE fell into the range of 0-2.5 and **Figure 7**(b) displayed that the hourly MSE exhibited similar time-varying patterns to the bike-sharing demand (**Figure 1**(c)) due to higher flexibility in bike-sharing demand at morning and evening rush hours. Then we randomly selected one bike station to further plot the fitting curve between observations and predictions of bicycle demand in **Figure 7**(c). The fitting curve suggested that the SpVAR-LASSO achieved good predictive performances. Then we further compared the predictive performance of SpVAR-LASSO with other baseline models (**Table II**). The results showed that the SpVAR-LASSO outperformed other baseline models such as HA, AR (24), and ARMA (24,1), in terms of both accuracy and interpretability (less inputs were required). Although the SpVAR had a higher accuracy than the SpVAR-LASSO model, the former required significantly more (ten times) variables as model inputs than the latter, implying that the SpVAR-LASSO approach had superior model interpretability, as well as a high prediction accuracy.

TABLE II
COMPARISON OF PREDICTIVE PERFORMANCE BETWEEN THE SPVAR-LASSO APPROACH AND BASELINE MODELS

| Model | MSE | No. of input variables |
|---|---|---|
| Hourly average | 4.6387 | 24 |
| AR (24) | 1.5341 | 24 |
| ARMA (24, 1) | 1.4657 | 24 |
| SpVAR | 0.3274 | 176 |
| SpVAR-LASSO | 1.2212 | Avg: 18 (Range: 0-119) |

## V. CONCLUSIONS

In this study, we developed a spatial vector autoregressive LASSO (SpVAR-LASSO) approach to predicting bike-sharing demand. Specifically, we used SpVAR to model both temporal and spatial dynamics of bike-sharing demand and used LASSO to select those important predictors with non-zero coefficients. Then we used bike trip data in Washington D.C. to demonstrate the SpVAR-LASSO approach and the results suggested that the combination of SpVAR and LASSO can avoid overfitting and improve prediction accuracy (lower MSE) and interpretability (less variables were required). Temporal predictors and spatial predictors both exerted a greater influence on bicycle demand than graph predictors. Also, one spatial predictor (the number of nearby transit stops) and two temporal predictors (the bike-sharing demand in the last 24 and 168 hours) were identified as the most important variables in accurately predicting the bike-sharing demand. Moreover, the SpVAR-LASSO outperformed other baseline models in predicting bicycle demand, in terms of both accuracy and model interpretability.

## REFERENCES

1 Baumanis, C., Hall, J., and Machemehl, R.: 'A machine learning approach to predicting bicycle demand during the COVID-19 pandemic', Research in Transportation Economics, 2023
2 Wang, X., Cheng, Z., Trépanier, M., and Sun, L.: 'Modeling bike-sharing demand using a regression model with spatially varying coefficients', Journal of Transport Geography, 2021, 93
3 Holan, S.H., Lund, R., and Davis, G.: 'The ARMA alphabet soup: A tour of ARMA model variants', Statistics Surveys, 2010, 4, (none)
4 Toman, P., Zhang, J., Ravishanker, N., and Konduri, K.C.: 'Dynamic predictive models for ridesourcing services in New York City using daily compositional data', Transportation Research Part C: Emerging Technologies, 2020, 121
5 Munira, S., and Sener, I.N.: 'A geographically weighted regression model to examine the spatial variation of the socioeconomic and land-use factors associated with Strava bike activity in Austin, Texas', Journal of Transport Geography, 2020, 88
6 Guidon, S., Reck, D.J., and Axhausen, K.: 'Expanding a(n) (electric) bicycle-sharing system to a new city: Prediction of demand with spatial regression and random forests', Journal of Transport Geography, 2020, 84
7 Jin, C., and Lee, G.: 'Exploring spatiotemporal dynamics in a housing market using the spatial vector autoregressive Lasso: A case study of Seoul, Korea', Transactions in GIS, 2019, 24, (1), pp. 27-43
8 Li, X., Xu, Y., Zhang, X., Shi, W., Yue, Y., and Li, Q.: 'Improving short-term bike sharing demand forecast through an irregular convolutional neural network', Transportation Research Part C: Emerging Technologies, 2023, 147
9 Liu, X., Gherbi, A., Li, W., and Cheriet, M.: 'Multi features and multi-time steps LSTM based methodology for bike sharing availability prediction', Procedia Computer Science, 2019, 155, pp. 394-401
10 Xiao, G., Wang, R., Zhang, C., and Ni, A.: 'Demand prediction for a public bike sharing program based on spatio-temporal graph convolutional networks', Multimedia Tools and Applications, 2020, 80, (15), pp. 22907-22925
11 Beenstock, M., and Felsenstein, D.: 'Spatial vector autoregressions', The Econometric Analysis of Non-Stationary Spatial Panel Data, 2019, pp. 129-161
12 Tibshirani, R.: 'Regression shrinkage and selection via the Lasso', Journal of the Royal Statistical Society: Series B (Methodological), 1996, 58, (1), pp. 267-288
13 Yang, Y., Heppenstall, A., Turner, A., and Comber, A.: 'Using graph structural information about flows to enhance short-term demand prediction in bike-sharing systems', Computers, Environment and Urban Systems, 2020, 83
14 Zellner, A.: 'An efficient method of estimating seemingly unrelated regressions and tests for aggregation bias', Journal of the American Statistical Association,, 1962, 57, (298), pp. 348–368
15 Nardi, Y., and Rinaldo, A.: 'Autoregressive process modeling via the Lasso procedure', Journal of Multivariate Analysis, 2011, 102, (3), pp. 528-549
16 Hsu, N.-J., Hung, H.-L., and Chang, Y.-M.: 'Subset selection for vector autoregressive processes using Lasso', Computational Statistics & Data Analysis, 2008, 52, (7), pp. 3645-3657