# How could the station-based bike sharing system and the free-floating bike sharing system be coordinated?

**Cheng, L., Yang, J., Chen, X., Cao, M., Zhou, H. and Sun, Y.**

# How could the station-based bike sharing system and the free-floating bike sharing system be coordinated?

Long Cheng[a,b], Junjian Yang[a,1], Xuewu Chen[a,*], Mengqiu Cao[c,d,e], Hang Zhou[a], Yu Sun[a]

[a] *School of Transportation, Southeast University, China*

[b] *Department of Geography, Ghent University, Belgium*

[c] *School of Architecture and Cities, University of Westminster, United Kingdom*

[d] *Bartlett School of Planning, University College London, United Kingdom*

[e] *Department of Statistics, London School of Economics and Political Science, United Kingdom*

* *Corresponding authors: Si Pai Lou #2, Nanjing, 210096, China; Email: chenxuewu@seu.edu.cn (X. Chen)*

[1] *Equally contributed first author.*

## Abstract

The station-based bike sharing system (SBBSS) and the free-floating bike sharing system (FFBSS) have been adopted on a large scale in China. However, the overlap between the services provided by these two systems often makes bike sharing inefficient. By comparing the factors that affect the usage of the two systems, this paper aims to propose appropriate strategies to promote their coordinated development. Using data collected in Nanjing, a predictive model is built to determine which system is more suitable at a given location. The influences of infrastructure, demand distribution, and land use attributes at the station level are examined via the support vector machine (SVM) approach. Our results show that the SBBSS tends to be favored in areas where there is a high concentration of travel demand, and proximity to metro stations and commercial properties, whereas locations with a higher density of major roads and residential properties are associated with more frequent use of the FFBSS. With regard to the methods used, a comparison of several machine learning approaches shows that the SVM has the best predictive performance. Our findings could be used to help policymakers and transportation planners to optimize the deployment and redistribution of docked and dockless bikes.

**Keywords:** Station-based bike sharing system; Free-floating bike sharing system; Support vector machine; Coordinated development; Land use

## 1. Introduction

The bike sharing system was first launched in Amsterdam, in the Netherlands, in 1965 in order to meet the 'last-mile' travel demand (DeMaio, 2009; Parkes et al., 2013). The system provides a new mode of transportation for citizens, which is recognized to have both traffic and health benefits, such as flexible mobility, support for multimodal transportation connections, and engagement in physical activity (Shaheen et al., 2010). Another advantage it offers is that traveling by bike is relatively quick and convenient, and therefore saves time (Faghih-Imani et al., 2014). In general, bike sharing systems can be divided into two categories according to whether they have docking stations or not. A station-based bike sharing system (SBBSS) has fixed docking stations, and users have to rent and return bikes at a station (see Figure 1a). The emerging free-floating bike sharing system (FFBSS) has no docking stations, and bikes can be parked anywhere that is appropriate (see Figure 1b). The built-in GPS tracking module allows users to locate and unlock bikes that are nearby via smartphone applications (Xu et al., 2018). The FFBSS system has attracted numerous users due to the freedom that it offers, as well as convenience of payment and parking (Pan et al., 2019).

(a) Station-based bike sharing system (b) Free-floating bike sharing system

Figure 1. Station-based bike sharing system and free-floating bike sharing system

(Source: (a) Nanjing Institute of City and Transport Planning (2018); (b) Li et al. (2018))

The FFBSS differs from the SBBSS in terms of users' travel demands, socio-economic characteristics, and operating modes (Chen et al., 2020a; Du and Cheng, 2018; Hua et al., 2020; Lyu et al., 2020). The FFBSS is often supported by venture capital funding. With the purpose of making profits, the majority of bikes are placed in 'hot' zones of a city where users are concentrated and the demand is high. In this case, inhabitants in 'cold' zones, such as suburbs and peripheral areas, are better served by the SBBSS (a non-profit making scheme). The wider spatial distribution of SBBSS docking stations satisfies the travel demand in lower-density neighborhoods (particularly suburban areas). In addition, consistent evidence suggests that these two systems show different patterns of use by people with different socio-demographic characteristics (Chen et al., 2020a; Fishman et al., 2015; Li et al., 2019). According to observations in Hangzhou and Kunming, the SBBSS is used more by older population groups and those with a relatively low level of education, whereas younger people and those with a higher level of education are more likely to favor the FFBSS (Chen et al., 2020a; Li et al., 2019). This may be due to a greater willingness to use and trust new technologies among younger and more educated people. In addition, with regards to the purpose for which people use the schemes, one recent study showed that the SBBSS is primarily used for commuting between homes and offices, while FFBSS services are used for various purposes, including commuting, recreation, and tourism, etc. (Chen et al., 2020a).

Provided by the government as a basic public transport service, most SBBSS schemes in China are subsidized. Therefore, they can offer good-quality services at relatively low costs, which benefits lower-income cohorts. The first 30 or 45 minutes of use is free in most cities for every journey made. In addition, the SBBSS has advantages in terms of controlling and redistributing the bicycle fleets between designated stations. However, due to the limited amount of docks at each station, it is often difficult to meet the demand for rent and return turnovers during peak hours. Having to walk the first and last mile to and from a docking station also reduces the attractiveness of the scheme (Link et al., 2020). By contrast, the FFBSS allows users to easily locate and unlock a bike via a smartphone and return it almost anywhere (as long as parking is permitted) once they have completed a trip. Consequently, it avoids the problem of there being no available docks nearby and provides seamless travel with a door-to-door service. Despite its strengths, the FFBSS also has some drawbacks, such as its unsustainable business model, over-sized fleets, disruption of public spaces, and vandalism (Shen et al., 2018; Xu et al., 2019). For instance, in China, competition for market share between different companies has led to an excess of bikes. In combination with inadequate redistribution and maintenance, this has caused a large amount of abandoned and damaged bikes to remain on the streets and in public spaces. As the redistribution of FFBSS bikes occurs over a larger geographical area rather than just between designated docking stations, it is more difficult to control and redistribute shared bikes than with the SBBSS.

Therefore, the SBBSS and FFBSS are likely to serve different groups of travelers and to supplement each other because they each have their own merits and demerits. As Gu et al. (2019) and Chen et al. (2020b) also concluded in their recent review papers, it is difficult to judge which one is better. The two systems could coexist in a coordinated and complementary way that enriches the urban transportation systems. However, currently, the two types of bike sharing system overall operate independently of each other, and there is a lack of research into their coordinated development. There is always some overlap between the services provided by the two systems, resulting in fierce competition. In some regions, the bicycle fleet exceeds travel demand, resulting in inefficient utilization of the bike sharing system as a whole. This study takes Nanjing as a case study with which to analyze how the development of the two systems could be properly coordinated. There are currently more than 96,000 SBBSS bikes and 2,656 stations operating in Nanjing. Meanwhile, the FFBSS has rapidly expanded since 2017 and now accounts for a sizeable proportion of the bike sharing market. There were more than 500,000 dockless bikes in Nanjing at the end of 2017 (Nanjing Institute of City and Transport Planning, 2018).

In this study, we use the SBBSS smart card data and the FFBSS order data to compare the factors affecting demand for the two bike sharing systems. The database is augmented with information on infrastructure, demand distribution, and land use attributes. The main objectives are to evaluate which system is more suitable at a given location, as well as to quantify the importance of the influencing factors.

The rest of the paper is organized as follows. Section 2 provides a literature review of prior studies and positions our research within the existing literature. Section 3 explains the case study and the multi-source data used. Section 4 describes the methods. Section 5 presents the model estimation results, and finally, Section 6 concludes the paper by summarizing the main findings.

## 2. Literature review

In order to improve the performance of the conventional bike sharing system with docking stations, numerous studies have been conducted to understand the factors that affect the demand for SBBSS. Buck and Buehler (2012) explored the influence of various factors – including the number of households without a car, cycle lanes, population density, and retail destinations around the stations – on bike sharing trips in Washington D.C. Rixey (2013) investigated the effects of demographic and built environment characteristics on monthly bicycle usage in three different cities in the US at the station level. He concluded that population density, job density, income levels, and the proportion of alternative commuters are the critical factors affecting bike sharing ridership. A station level analysis was also conducted in Toronto to examine the effects of the built environment and weather on bike sharing demand (El-Assi et al., 2017). Faghih-Imani et al. (2014) examined the influence of meteorological data, temporal characteristics, land use and built environment attributes on arrival and departure flows. Zhao et al. (2014) investigated the effects of urban features and bike sharing system characteristics on daily trip frequency. Wang et al. (2016) conducted analyses which considered annual rates for each station and examined the effects of socio-demographics, nearby business and job densities, the built environment, and transportation infrastructure variables on annual usage flows.

Several recent studies have discussed the factors that influence travel demand for the FFBSS. For example, using data collected from Hangzhou, Chen et al. (2020a) compared the factors influencing the use frequency of the SBBSS and FFBSS. They found that gender and type of monthly mobile phone data package purchased were the two most significant factors affecting the usage of the FFBSS; while education, household car ownership, travel purpose and travel distance were the key factors influencing SBBSS usage. In addition, Li et al. (2018) explored the factors affecting FFBSS users' behaviors, using data collected in Jiangsu Province. According to their analyses, a higher level of education, higher daily transportation costs, the convenience of picking up and parking/dropping off, and the health benefits for users have significant effects on FFBSS usage. Furthermore, the built environment is also related to travel demand for the FFBSS, through factors such as whether an area is residential and commercial, land use mix, public transit accessibility, and population density (Du et al., 2019; Shen et al., 2018; Xu et al., 2019). In Singapore, Xu et al. (2019) found that locations with a higher density of public housing are associated with more shared bicycle trips in the evening and fewer in the morning. Du and Cheng (2018) studied the factors that affect FFBSS travel patterns and observed that the number of malfunctioning bicycles is an important factor which restricts the development of the FFBBS. Weather is another important factor that plays a role in FFBSS usage. It was found that fewer trips were made on cold and rainy days, in a case study of Munich (Reiss and Bogenberger, 2015), while a similar trend was also observed with respect to hot weather in Singapore (Shen et al., 2018).

Different modeling techniques have been used to identify which factors contribute to bike sharing demand. Campbell et al. (2016) used a multinomial logit model to investigate which factors contributed to users' choice of shared bikes, via a stated preference survey conducted in Beijing. Zhao et al. (2014) employed a partial least squares regression model to explore the relationship between daily trip frequency and urban population, government expenditure, the number of registered scheme members, the number of shared bikes, and the number of docking stations. Faghih-Imani et al. (2014) employed a linear mixed model to study how land use and urban form impact bicycle flows. El-Assi et al. (2017) evaluated the effects of the built environment and weather on bike sharing demand in Toronto using a distributed lag model. Meanwhile, Fournier et al. (2017) developed a sinusoidal model to predict the pattern of seasonal demand for bike sharing.

In general, previous studies have provided valuable insights into understanding the factors that influence the demand for SBBSS and FFBSS. However, few studies have drawn comparisons between the two systems in terms of these influencing factors. To fill this gap, this paper examines the differences in the relative importance of these factors, and thereby proposes strategies that could be used to develop effective cooperation between the bike sharing systems.

## 3. Study area and data

3.1 Study area

The data used were collected from an area totalling 782.45 km$^2$, comprising the central part of the city of Nanjing (Figure 2), and consisting of Xuanwu (XW) District, Gulou (GL) District, Jianye (JY) District, Qixia (QX) District, Qinhuai (QH) District, and Yuhuatai (YHT) District.[1] As the capital of Jiangsu Province, Nanjing is a large city located on the east coast of China. In September 2017, the study area contained 1,212 SBBSS stations (Figure 2). A buffer area with a radius of 250 meters is set as the service area for each SBBSS station, which is in line with the study carried by Faghih-Imani et al. (2014). This is also considered to be the overlapping service area for the SBBSS and FFBSS.

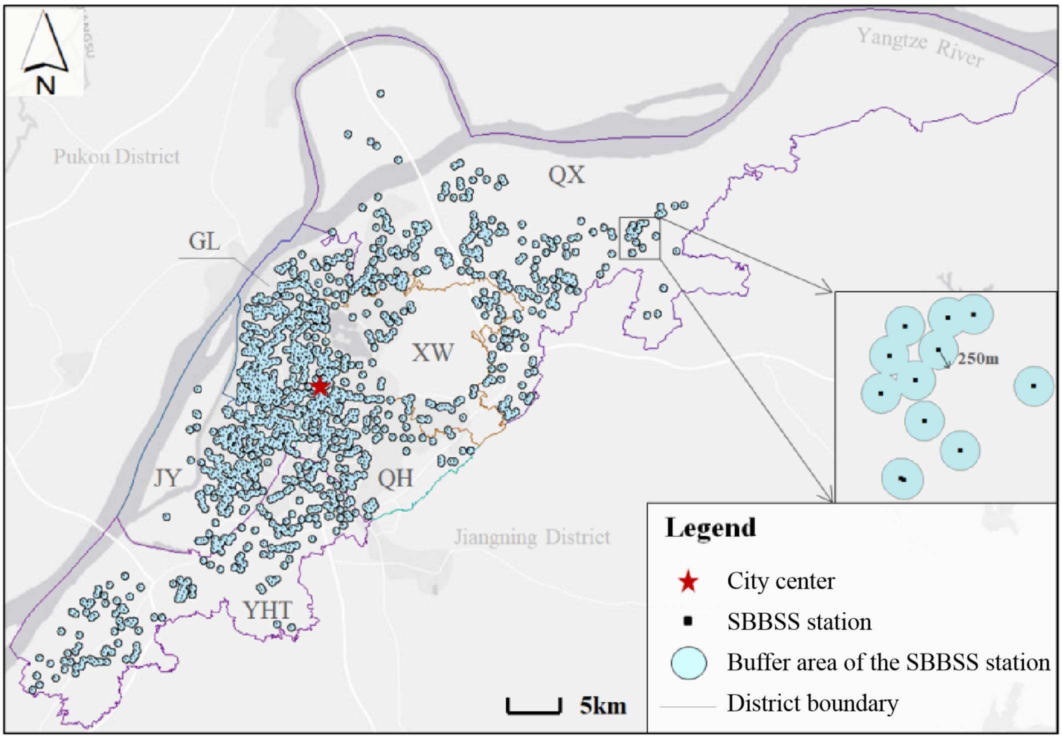



Figure 2. Spatial distribution of SBBSS stations in the study area
(Sources: Authors' elaboration)

3.2 Data source

The multi-source data used in this study consists of SBBSS smart card data, FFBSS order data, and, urban spatial data. The SBBSS data were provided by the Nanjing Public Bicycle Company, and the FFBSS data were obtained from one of the operating companies in Nanjing. To inform our analysis, we examined the SBBSS smart card data and the FFBSS order data from 11 to 24

[1] Nanjing consists of 11 administrative districts in total. However, the study area only covers 6 of these districts (i.e. the central part of the city).

September 2017. The smart card data consists of the anonymized user ID, bike ID, leasing station name, leasing time, returning station name, and returning time. The information relating to the order data is consistent with the SBBSS smart card data, except for the station location. The station location information contained in the FFBSS order data includes the latitude and longitude of pick-up and destination points. We removed the records of trips that occurred outside the study area. As a result, the SBBSS smart card dataset is composed of 1.64 million trips, 42.46 thousand docked bikes, and 0.12 million users. Meanwhile, the FFBSS dataset comprises a total of 5.63 million trips made by 0.51 million users using 187.81 thousand dockless bikes. The SBBSS docking station data were also provided by the Nanjing Public Bicycle Company, and include the name, longitude, and latitude of each station.

The urban spatial data within the buffer area were extracted from the web mapping service developed by AutoNavi on Amap. AutoNavi (also known as 'Gaode' in Chinese) is a Chinese web mapping, navigation and location-based services provider. The data are divided into three categories: (a) spatial vector surface data of buildings, which include function type (business/office, residential property, etc.), number of floors, and floor area; (b) spatial vector line data of roads, which include road grade (major road, minor road, etc.) and road length; (c) spatial location data of points of interest (POI), which include the latitude, longitude and function type of each POI (e.g. residential property, business/office, or commercial property).

### 3.3 Variable generation

In this study, we examine whether or not a service area is SBBSS dominant. More specifically, if SBBSS usage exceeds that of FFBSS in the overlapping service area, it is regarded as SBBSS dominant and denoted by a one; otherwise, it is denoted by a zero. The independent variables considered in our analysis are divided into three categories: (a) infrastructure, (b) demand distribution, and (c) land use. The selection of variables is based on theoretical considerations (e.g. a wider spatially distributed demand favors the use of the FFBSS) and a review of other relevant studies (e.g. Buck and Buehler, 2012; Chen et al., 2020a; Du et al., 2019; Faghih-Imani et al., 2014; Shen et al., 2018). The infrastructure data consist of major road density, minor road density, and number of neighboring docking stations. These variables are able to capture the effect of cycling facilities on bike sharing demand. Moreover, the influences of minor roads and major roads can help to identify cyclists' preferences for different routes. The number of neighboring stations within the service area is computed to capture the spatial dependence effect. The number of metro stations is also taken into consideration.

Table 1. Variables and descriptive analysis

| Variables | Mean | Std. | Min. | Max. |
|---|---|---|---|---|
| Dependent variable | | | | |
| Whether a service area is SBBSS dominant (yes = 1; no = 0) | 0.34 | 0.47 | 0 | 1 |
| Independent variable | | | | |
| *Infrastructure* | | | | |
| Major road density (km/km$^2$) | 0.83 | 0.86 | 0 | 5.64 |
| Minor road density (km/km$^2$) | 0.89 | 0.63 | 0 | 3.97 |
| Number of neighboring stations | 0.70 | 0.93 | 0 | 5 |
| Number of metro stations | 0.09 | 0.29 | 0 | 1 |
| *Demand distribution* | | | | |
| Dispersion of demand points | 44.96 | 12.40 | 0 | 74.68 |
| Average station distance to demand points (m) | 105.85 | 29.78 | 10.18 | 244.61 |
| Std. of station distance to demand points | 53.22 | 15.13 | 0 | 103.98 |
| *Land use* | | | | |
| Business/office area (km$^2$) | 161,699 | 423,497 | 0 | 4,329,883 |
| Residential area (km$^2$) | 505,063 | 571,677 | 0 | 4,879,651 |
| Number of residential properties | 5.17 | 4.97 | 0 | 35 |
| Number of commercial properties | 5.05 | 6.26 | 0 | 53 |
| Number of businesses/offices | 17.19 | 28.79 | 0 | 252 |
| Number of Edu. & Cul. | 7.72 | 11.22 | 0 | 120 |
| Dispersion of residential properties | 30.32 | 25.84 | 0 | 91.08 |
| Dispersion of commercial properties | 24.59 | 26.01 | 0 | 98.11 |
| Dispersion of businesses/offices | 30.87 | 26.17 | 0 | 95.72 |
| Dispersion of Edu. & Cul. | 71.81 | 53.08 | 0 | 195.18 |
| Average station distance to residential properties (m) | 124.91 | 51.30 | 0 | 210.99 |
| Average station distance to commercial properties (m) | 103.64 | 62.56 | 0 | 210.94 |
| Average station distance to businesses/offices (m) | 117.14 | 55.01 | 0 | 211.41 |
| Average station distance to Edu. & Cul. (m) | 113.76 | 58.20 | 0 | 210.44 |
| Std. of station distance to residential properties | 33.93 | 24.16 | 0 | 118.65 |
| Std. of station distance to commercial properties | 29.22 | 26.42 | 0 | 112.11 |
| Std. of station distance to businesses/offices | 32.13 | 23.98 | 0 | 111.28 |
| Std. of station distance to Edu. & Cul. | 32.42 | 25.95 | 0 | 131.82 |

Note: Std. = standard deviation; Edu. & Cul. = educational and cultural amenities; Network distance is used in the calculation.

Because the FFBSS has no fixed stations, the demand points are spread haphazardly over space. We therefore use the dispersion of demand points and average distance to the nearest docking station to represent the demand distribution of the FFBSS in the overlapping service area. Three variables relating to demand distribution are generated accordingly: the dispersion of demand points, the average and standard deviation of the distances between demand points and the nearest docking station. It should be noted that the dispersion referred to here is described by the standard deviation of the distance between demand points and the geometric center of the buffer area (i.e. the position corresponding to the average latitude and longitude of all points in the buffer area).

Land use variables aim to capture the influence of different types of land use within the overlapping service area. Residential and business/office areas are calculated using the spatial vector surface data. We also consider two types of POI-related variables: (a) the number and dispersion of POIs associated with residential property, commercial property, business/office, educational and cultural amenities; and (b) the average and standard deviation of the distance between an SBBSS station and POIs. A descriptive summary of the data is provided in Table 1.

## 4. Methodology

### 4.1 Model comparisons

Predicting the dominant type of bike sharing system is essentially a binary classification problem. Several measures are used to evaluate the effectiveness of a method: accuracy, precision, sensitivity, and the receiver operating characteristic (ROC) curve. These measures can be readily calculated based on a confusion matrix which contains information about actual and predicted classifications (Provost and Kohavi, 1998). Table 2 shows the confusion matrix for a two-class classifier.

Table 2. Confusion matrix

| Actual | Predicted | |
|---|---|---|
| | Positive | Negative |
| Positive | True positive (TP) | False negative (FN) |
| Negative | False positive (FP) | True negative (TN) |

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{FN} + \text{TN}} \tag{1}$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{2}$$

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{3}$$

The ROC curve is a reliable technique based on the values of the true positive rate and the false positive rate. Therefore, it offers a trade-off between precision and sensitivity (Akay, 2009). The true positive rate is also known as the sensitivity, and the false positive rate is defined as:

$$\text{False positive rate} = \frac{FP}{FP + TN} \tag{4}$$

In order to find the optimal method, we compared the predictive abilities of different approaches, i.e. support vector machine (SVM), logistic regression (Logistic), K-nearest neighbors (KNN), random forest (RF), and Gaussian process regression (GPR)). These approaches are commonly used in the literature to address classification problems (Cheng et al., 2019; Sze and Wong, 2007; Xu et al., 2020). They were run using Python with normalized raw data. The grid search and ten-fold cross-validation technique were used to determine the optimal parameter setting for the SVM. The penalty parameter was set as one and the kernel was set as 'linear'. To maintain fairness when making comparisons, each method was run multiple times with varying parameter settings, and the one with the optimal performance was chosen to carry out the comparison. Finally, the penalty parameter of the logistic regression was set as one. The K value of the KNN was set as five, and the output was weighted by the reciprocals of the Euclidean distances between five points and its neighbors. For the RF, the number of estimators was set as 30, and the maximum number of features was set as 19. The performance of the GPR relies on the selection of the covariance function (Yu et al., 2016). In this study, the squared exponential function expressed in Equation (5) was chosen to calculate the covariance (Rasmussen and Williams, 2003):

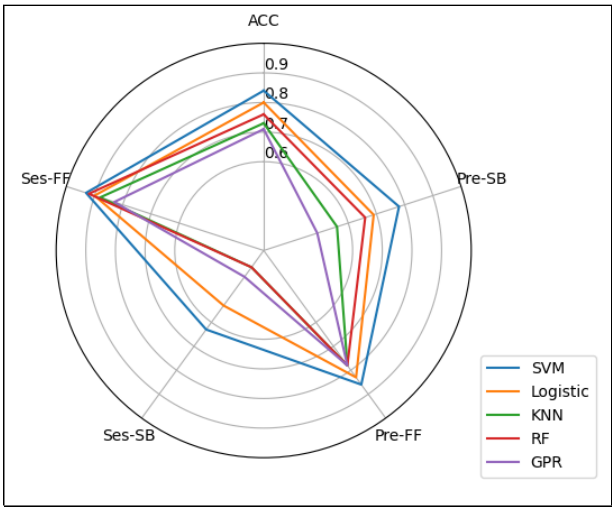
$$K_{SE}(x_i,x_j) = \sigma_s^2 \exp\left( - \frac{1}{2l^2}(x_i - x_j)^2 \right) + \sigma_n^2 \delta_{ij} \tag{5}$$

where $\sigma_s^2$ is the signal variance, $l$ is the length scale and $\sigma_n^2$ is the noise variance. The optimal setting of these three parameters can be achieved by maximizing the marginal log-likelihood function with hyperparameters (Yu et al., 2016). All five of the methods (i.e. SVM, Logistic, KNN, RF, and GPR) were undertaken using the same training sample. Table 3 and Figure 3 compare their predictive performances for the validation sample.

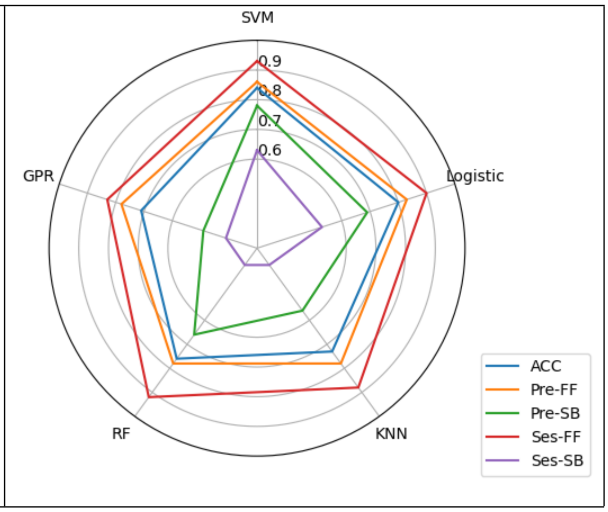
Table 3. Performance comparison of different methods

| Method | Accuracy | Precision | | Sensitivity | |
|---|---|---|---|---|---|
| | | FFBSS | SBBSS | FFBSS | SBBSS |
| SVM | 0.8418 | 0.8603 | 0.7805 | 0.9286 | 0.6275 |
| Logistic | 0.7966 | 0.8261 | 0.6923 | 0.9048 | 0.5294 |
| KNN | 0.7345 | 0.7762 | 0.5588 | 0.8810 | 0.3725 |
| RF | 0.7627 | 0.7838 | 0.6552 | 0.9206 | 0.3725 |

| | | | | | |
|---|---|---|---|---|---|
| GPR | 0.7062 | 0.7761 | 0.4884 | 0.8254 | 0.4118 |




(a) Predictive performance for different indices




(b) Predictive performance of different methods

Figure 3. Comparison of predictive performances for different methods (ACC = Accuracy, Pre = Precision, Ses = Sensitivity)

There are twice as many FFBSS dominant areas in our sample as there are SBBSS dominant areas (Table 1). Due to potential concerns about class imbalance, it is necessary to consider precision and sensitivity as well as accuracy. The SVM has significantly higher predictive accuracy than other methods. Logistic regression is the second most accurate approach, and the difference between the SVM and Logistic regression is 4.52%. The SVM also performs best in terms of precision and sensitivity. The overall performance of the Logistic regression is nearly as good as the SVM (Figure 3a). This is reasonable because the optimization problem of the Logistic regression method is similar to that of the SVM. Figure 3b shows that the precision and sensitivity of the FFBSS predictions are higher than those of the SBBSS, which is due to the likelihood of a FFBSS dominant prediction caused by class imbalance. It is worth noting that these indices for the SBBSS predictions are more important for policy makers than those for the FFBSS, because the locations of SBBSS stations are regulated by the government, while the FFBSS bikes are controlled by market forces. Furthermore, the cost of constructing a new SBBSS station is higher than that of repositioning FFBSS bikes. In terms of precision and sensitivity, the SVM method reveals a powerful predictive capability for the SBBSS (Figure 3a). The differences between the SVM and the Logistic regression method (i.e. the second-best) in terms of precision and sensitivity are 8.82% and 9.81%, respectively.

An ROC curve is a graphical plot that illustrates the diagnostic ability of a binary classifier system. The area under the curve (AUC) provides a tool for selecting potentially optimal methods in general circumstances. Table 4 and Figure 4 compare the AUC and ROC curves for different methods. The GPR performs the worst, followed by the KNN and the RF. GPR is a kernel-based regression technique used to model nonlinear multivariate relationships, and it assumes that the random variables follow Gaussian distributions. This rule may not be applicable to our dataset. The SVM outperforms other methods according to this index.

Table 4. AUC for different methods

| Method | SVM | Logistic | KNN | RF | GPR |
|---|---|---|---|---|---|
| AUC | 0.8604 | 0.8409 | 0.7399 | 0.8268 | 0.5962 |

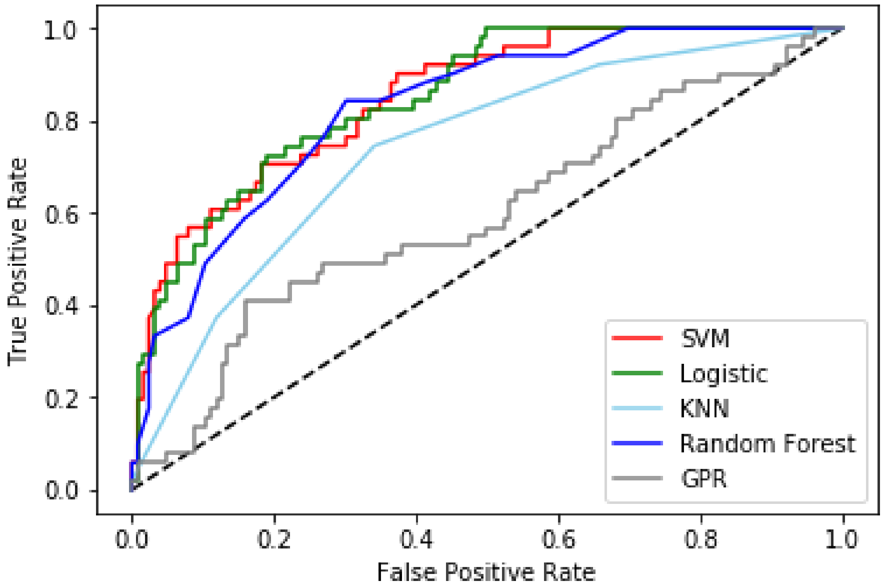



Figure 4. ROC curves for different methods

4.2 Support vector machine method

In addition to the tests results for the predictive abilities of various methods, described above, in which the SVM method ranked highest, many prior studies have also shown that the SVM is very useful for classification problems because of its flexibility, lower risks of overfitting, computational efficiency and capacity to handle high-dimensional data (Nguyen and De la Torre, 2010; Plakandaras et al., 2019; Yu and Abdel-Aty, 2013). However, the SVM was not originally developed as a feature selection tool (Bradley and Mangasarian, 1998). A conventional SVM approach extracts features and learns parameters independently. However, performing these two steps independently might result in a loss of information in relation to the classification process (Nguyen and De la Torre, 2010). To address this problem, we used a method of feature selection based on SVM recursive feature elimination (SVM-RFE) in our analyses. The method has demonstrated good generalization performance and is able to overcome the overfitting problem (Guyon et al., 2002). In addition, it is capable of recognizing key factors that influence the target variable.

The SVM-RFE method was proposed by Guyon et al. (2002), and has been widely used in different fields (Akay, 2009; Huang et al., 2014; Xue et al., 2018). It is a sequential backward feature elimination method based on SVM. By removing an irrelevant feature from the feature set in each iteration, a less important or irrelevant feature is eliminated in advance or sorted to the back of a ranked feature list (Huang et al., 2014). The method can be described starting from a conventional SVM algorithm, as follows. Consider *n* training data pairs $\{(x_i, y_i)\}_{i=1}^{n}$, where $x_i \in R^m$ is a feature vector representing the *i*th sample, and $y_i$ is the class label of $x_i$. The decision function of SVM is $f(x) = w^T x + b$, where $w = [w_1, w_2, \ldots, w_m]^T$ is the weight vector and $b$ is a scalar. By using a kernel trick, the dual optimization problem associated with SVM can be expressed as follows (Akay, 2009; Xue et al., 2018):

$$\begin{cases} \min \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j y_i y_j K(\boldsymbol{x}_i x_j) - \sum_{i=1}^{n}\alpha_i \\ \quad s.t. \sum_{i=1}^{n}\alpha_i y_i = 0 \\ \quad 0 \le \alpha_i \le C, i = 1, \cdots, n \end{cases} \tag{6}$$

where the predefined parameter $C$ represents a trade-off between training accuracy and model complexity. If $\alpha_i^*$ is the non-zero optimal solution, the classifier function can be expressed as follows:

$$f(x) = \mathrm{sgn}\left(\sum_{i=1}^{n}\alpha_i^* y_i K(x_i, x_j) + b^*\right) \tag{7}$$

The linear SVM uses the linear kernel $K(x_i, x_j) = x_i^T x_j$. In the nonlinear SVM, a range of different options exist and the Gaussian kernel is widely used:

$$K(x_i, x_j) = \exp\left(-\frac{\|x_i - x_j\|^2}{2\sigma^2}\right) \tag{8}$$

where $\| \bullet , \bullet \|$ denotes the distance between two vectors (usually defined as the Euclidean distance), $\sigma$ is a constant and the value can be obtained through a cross-validation process.

The SVM-RFE method is used to remove irrelevant features from the feature set. In each iteration, the importance of various features depends on their weight coefficients. If the kernel is a linear kernel, the weight is $w = \sum_{i=1}^{n}\alpha_i y_i x_i$, where $\alpha_i$ is obtained by solving Equation (6). The ranking score of the *p*th feature can be computed as $c_{p\_linear} = w_p^2$. If the kernel is a

Gaussian kernel, the ranking score of feature *p* can be written as follows (Xue et al., 2018):

$$c_{p\_gaussian} = \frac{1}{2}\left|\alpha^T H\alpha - \alpha^T H^{(-p)}\alpha\right|, p = 1,\ldots,m \tag{9}$$

where $K(x_i,x_j)$ is the Gaussian kernel, $H \in R^{n \times n}$ with $H_{i,j} = y_i y_j K(x_i,x_j)$, $H^{(-p)} \in R^{n \times n}$ with $H_{i,j}^{(-p)} = y_i y_j K(x_i^{(-p)},x_j^{(-p)})$. The notation $(-p)$ means that feature $p$ has been removed. The feature with the smallest ranking score was removed in each iteration.

## 5. Results

In this section, the results of the SVM-RFE are discussed in order to gain insights into the effects of different variables. It should be noted that only statistically significant coefficients at a 95% confidence level are presented in Table 5. The positive value of a variable means it has a positive correlation with the probability that a service area is classified as SBBSS dominant.

Table 5. Model estimation results

| Variables | Coef. | Variables | Coef. |
|---|---|---|---|
| Major road density | -0.324 | Dispersion of residential properties | 0.097 |
| Minor road density | 0.039 | Dispersion of commercial properties | -0.101 |
| Number of neighboring stations | 0.180 | Dispersion of businesses/offices | 0.018 |
| Number of metro stations | 0.135 | Dispersion of Edu. & Cul. | -0.037 |
| Dispersion of demand points | -0.449 | Average station distance to residential properties | -0.027 |
| Average station distance to demand points | -0.616 | Average station distance to commercial properties | -0.363 |
| Std. of station distance to demand points | 0.220 | Average station distance to businesses/offices | -0.027 |
| Business/office area | -0.014 | Average station distance to Edu. & Cul. | 0.062 |
| Residential area | -0.012 | Std. of station distance to residential properties | -0.026 |
| Number of residential properties | -0.578 | Std. of station distance to commercial properties | -0.208 |
| Number of commercial properties | -0.283 | Std. of station distance to businesses/offices | -0.085 |
| Number of businesses/offices | 0.437 | Std. of station distance to Edu. & Cul. | 0.011 |
| Number of Edu. & Cul. | -0.288 | | |

Note: Std. = standard deviation; Edu. & Cul. = educational and cultural amenities

### 5.1 Infrastructure

As expected, there is a positive correlation between the number of neighboring docking stations and SBBSS dominant areas. The more docking stations there are, the more likely an area is to be SBBSS dominant. The density of major roads has a negative impact, indicating that this variable contributes to the use of the FFBSS. However, the density of minor roads is positively correlated with SBBSS usage. SBBSS users often combine their bicycle trips with travelling by metro, which is reflected by the positive impact of the presence of metro stations. This implies that the proximity to a metro station is likely to increase SBBSS usage. This is plausible because the SBBSS provides more reliable services near metro stations than the FFBSS due to the existence of fixed docking stations. In reality, there is at least one SBBSS station near every metro station in Nanjing.

### 5.2 Demand distribution

Although the existence of docking stations makes the SBBSS more reliable, it also restricts the service coverage. If the demand distribution is dispersed, users will have to walk farther, on average, to use the SBBSS. Consequently, people may be more likely to use the FFBSS if the demand distribution is dispersed. This assumption is supported by the negative coefficients of two variables: dispersion of demand points; and average station distance to demand points. This finding can assist in understanding the relationship between the SBBSS and FFBSS. When a new SBBSS station is to be constructed, demand distribution is a significant factor that should be considered. If the distribution is not dispersed, the SBBSS station should be located at the geographical center of the distribution. However, if the demand distribution is more dispersed, this location may not be appropriate for an SBBSS station.

### 5.3 Land use

Four types of POIs were considered in our analyses: residential property, commercial property, business/office, and educational and cultural amenities. It was expected that people may be more likely to use the FFBSS if an SBBSS station is located farther away from these places. This is supported by the negative coefficients for distance to residential properties, commercial properties, and businesses/offices. Therefore, it is important to locate SBBSS stations as close as possible to crowded places. It is noteworthy that the coefficients for the number of residential properties and the number of commercial properties are negative. In fact, the capability of an SBBSS station is limited but there may be any number of FFBSS bikes in an area. SBBSS users may therefore encounter a problem if they ‘pick up’ a bike at an empty docking station or ‘return’ it to a fully occupied one. A service area will become FFBSS dominant if it contains a lot of residential properties or shopping centers, even though the SBBSS utilization rate could be high. Unlike residential and commercial properties, the coefficient for the number of companies is positive. The explanation for this is that the trip pattern for people traveling to work is regular and has fixed origins and destinations. Commuters are more likely to choose to use the SBBSS if there are SBBSS stations located near their homes and workplaces.

## 6. Conclusions

This study analyzed the differences between the SBBSS and FFBSS by examining the factors influencing the usage of the two bike sharing systems. The impacts of infrastructure, demand distribution, and land use attributes were examined at the station level using multi-source data.

A predictive model was built to assess whether a specific location was suitable for the SBBSS or FFBSS. The SVM-RFE method was used to determine the relative importance of the variables. The findings will be helpful for improving the coordinated development of the bike sharing system as a whole.

To coordinate the two bike sharing systems more effectively, we compared the relative importance of influencing factors, represented by the magnitude of the coefficients, in order to propose appropriate strategies. First, the demand distribution should be a primary consideration when planning a new SBBSS station. If the demand distribution is dispersed, the location would not be suitable for the SBBSS. In an SBBSS dominant area, docking stations should be located near the geographical center of demand distribution. FFBSS bikes should be repositioned so that they are away from this area. In this way the area covered by the two systems will expand. Second, SBBSS usage decreases when the station is located farther away from commercial properties. Therefore, the SBBSS station should be located as close as possible to commercial properties (if permitted). This finding could help to determine suitable areas for planning new SBBSS stations to ensure high utilization efficiency. Third, the proximity to metro stations promotes increased SBBSS usage relative to FFBSS. Thus, it would be good practice to construct SBBSS stations near metro stations. Fourth, it is important to reposition FFBSS bikes to locations with a high major road density and residential density.

Our research contributes to the existing literature in three ways. First, it captures the effect of infrastructure, demand distribution, and land use attributes on the usage of SBBSS and FFBSS, whereas most previous studies have only focused on one of the two bike sharing systems (Du et al., 2019; Faghih-Imani et al., 2014; Shen et al., 2018). Second, it uses an SVM-RFE method to strike a balance between classification performance and interpretability, something which has been overlooked in previous research (Akay, 2009; Nguyen et al., 2010; Yu and Abdel-Aty, 2013). Third, this paper provides policy implications for SBBSS planning and FFBSS repositioning from the perspective of coordinated development.

In terms of limitations, a city’s cycling infrastructure can also play a significant role in affecting bike sharing demand, through aspects such as the density and length of the cycle lanes (Lazarus et al., 2020). However, due to the restriction of data availability, the aforementioned factors have not been taken into consideration in this study, but this is something which could be addressed in further research.

**Acknowledgments**

This research is funded by the National Key R&D Program of China (Project No.:

2018YFB1601300), the National Natural Science Foundation of China (Project No.: 71801041 and 51808392), the EPSRC (EPSRC Reference: EP/R035148/1), the SCUE Research Fund, and School Funding from the University of Westminster. Thanks to the extensive comments from the anonymous reviewers, which have significantly improved the paper.

## References

Akay, M.F., 2009. Support vector machines combined with feature selection for breast cancer diagnosis. Expert Systems with Applications, 36(2), 3240-3247.

Bradley, P.S., Mangasarian, O.L., 1998. Feature selection via concave minimization and support vector machines. In Proceedings of the 13th International Conference on Machine Learning, San Francisco, USA.

Buck, D., Buehler, R., 2012. Bike lanes and other determinants of capital bikeshare trips. In: Paper presented at the 91st Transportation Research Board Annual Meeting, Washington D.C., USA.

Campbell, A.A., Cherry, C.R., Ryerson, M.S., Yang, X., 2016. Factors influencing the choice of shared bicycles and shared electric bikes in Beijing. Transportation Research Part C, 67, 399-414.

Chen, M., Wang, D., Sun, Y., Waygood, E.O.D., Yang, W., 2020a. A comparison of users' characteristics between station-based bikesharing system and free-floating bikesharing system: Case study in Hangzhou, China. Transportation, 47(2), 689-704.

Chen, Z., van Lierop, D., Ettema, D., 2020b. Dockless bike-sharing systems: what are the implications? Transport Reviews, 40(3), 333-353.

Cheng, L., Chen, X., De Vos, J., Lai, X., Witlox, F., 2019. Applying a random forest method approach to model travel mode choice behavior. Travel Behaviour and Society, 14, 1-10.

DeMaio, P., 2009. Bike-sharing: History, impacts, models of provision, and future. Journal of Public Transportation, 12(4), 41-56.

Du, M., Cheng, L., 2018. Better understanding the characteristics and influential factors of different travel patterns in free-floating bike sharing: Evidence from Nanjing, China. Sustainability, 10(4), 1244.

Du, Y., Deng, F., Liao, F., 2019. A model framework for discovering the spatio-temporal usage patterns of public free-floating bike-sharing system. Transportation Research Part C, 103, 39-55.

El-Assi, W., Mahmoud, M.S., Habib, K.N., 2017. Effects of built environment and weather on bike sharing demand: a station level analysis of commercial bike sharing in Toronto. Transportation, 44(3), 589-613.

Faghih-Imani, A., Eluru, N., El-Geneidy, A.M., Rabbat, M., Haq, U., 2014. How land-use and urban form impact bicycle flows: evidence from the bicycle-sharing system (BIXI) in Montreal. Journal of Transport Geography, 41, 306-314.

Fishman, E., Washington, S., Haworth, N., Watson, A., 2015. Factors influencing bike share membership: An analysis of Melbourne and Brisbane. Transportation Research Part A, 71, 17-30.

Fournier, N., Christofa, E., Knodler Jr, M.A., 2017. A sinusoidal model for seasonal bicycle demand estimation. Transportation Research Part D, 50, 154-169.

Gu, T., Kim, I., Currie, G., 2019. To be or not to be dockless: Empirical analysis of dockless bikeshare development in China. Transportation Research Part A, 119, 122-147.

Guyon, I., Weston, J., Barnhill, S., Vapnik, V., 2002. Gene selection for cancer classification using support vector machines. Machine Learning, 46(1-3), 389-422.

Hua, M., Chen, X., Zheng, S., Cheng, L., Chen, J., 2020. Estimating the parking demand of free-floating bike sharing: A journey-data-based study of Nanjing, China. Journal of Cleaner Production, 244, 118764.

Huang, M.L., Hung, Y.H., Lee, W.M., Li, R.K., Jiang, B.R., 2014. SVM-RFE based feature selection and Taguchi parameters optimization for multiclass SVM classifier. The Scientific World Journal, 2014.

Lazarus, J., Pourquier, J.C., Feng, F., Hammel, H., Shaheen, S., 2020. Micromobility evolution and expansion: Understanding how docked and dockless bikesharing models complement and compete – A case study of San Francisco. Journal of Transport Geography, 84, 102620.

Li, X., Zhang, Y., Sun, L., Liu, Q., 2018. Free-floating bike sharing in Jiangsu: Users' behaviors and influencing factors. Energies, 11(7), 1664.

Link, C., Strasser, C., Hinterreiter, M., 2020. Free-floating bikesharing in Vienna – A user behaviour analysis. Transportation Research Part A, 135, 168-182.

Lyu, Y., Cao, M., Zhang, Y., Yang, T. and Shi, C. 2020. Investigating users' perspectives on the development of bike-sharing in Shanghai. Research in Transportation Business and Management, 100543.

Nanjing Institute of City and Transport Planning, 2018. Annual Report of Nanjing Traffic Development 2018. Nanjing, China.

Nguyen, M.H., De la Torre, F., 2010. Optimal feature selection for support vector machines. Pattern Recognition, 43(3), 584-591.

Pan, L., Cai, Q., Fang, Z., Tang, P., Huang, L., 2019. A deep reinforcement learning framework for rebalancing dockless bike sharing systems. In Proceedings of the 33rd AAAI Conference on Artificial Intelligence, Honolulu, USA.

Parkes, S.D., Marsden, G., Shaheen, S.A., Cohen, A.P., 2013. Understanding the diffusion of public bikesharing systems: Evidence from Europe and North America. Journal of Transport Geography, 31, 94-103.

Plakandaras, V., Papadimitriou, T., Gogas, P., 2019. Forecasting transportation demand for the US market. Transportation Research Part A, 126, 195-214.

Provost, F., Kohavi, R., 1998. Glossary of terms. Journal of Machine Learning, 30(2-3), 271-274.

Rasmussen, C.E., Williams, C.K.I., 2003. Gaussian process for machine learning. In Summer School on Machine Learning, Berlin, Germany.

Reiss, S., Bogenberger, K., 2015. GPS-data analysis of Munich's free-floating bike sharing system and application of an operator-based relocation strategy. In Proceedings of the 2015 IEEE 18th International Conference on Intelligent Transportation Systems, Washington D.C., USA.

Rixey, R., 2013. Station-level forecasting of bikesharing ridership: station network effects in three US systems. Transportation Research Record, 2387, 46-55.

Shaheen, S.A., Guzman, S., Zhang, H., 2010. Bikesharing in Europe, the Americas, and Asia: past, present, and future. Transportation Research Record, 2143, 159-167.

Shen, Y., Zhang, X., Zhao, J., 2018. Understanding the usage of dockless bike sharing in Singapore. International Journal of Sustainable Transportation, 12(9), 686-700.

Sze, N.N., Wong, S.C., 2007. Diagnostic analysis of the logistic model for pedestrian injury severity in traffic crashes. Accident Analysis & Prevention, 39(6), 1267-1278.

Wang, X., Lindsey, G., Schoner, J.E., Harrison, A., 2016. Modeling bike share station activity: Effects of nearby businesses and jobs on trips to and from stations. Journal of Urban Planning and Development, 142(1), 04015001.

Xu, C., Ji, J., Liu, P., 2018. The station-free sharing bike demand forecasting with a deep learning approach and large-scale datasets. Transportation Research Part C, 95, 47-60.

Xu, D., Wang, Y., Peng, P., Beilun, S., Deng, Z., Guo, H., 2020. Real-time road traffic state prediction based on kernel-KNN. Transportmetrica A: Transport Science, 16(1), 104-118.

Xu, Y., Chen, D., Zhang, X., Tu, W., Chen, Y., Shen, Y., Ratti, C., 2019. Unravel the landscape and pulses of cycling activities from a dockless bike-sharing system. Computers, Environment and Urban Systems, 75, 184-203.

Xue, Y., Zhang, L., Wang, B., Zhang, Z., Li, F., 2018. Nonlinear feature selection using Gaussian kernel SVM-RFE for fault diagnosis. Applied Intelligence, 48(10), 3306-3331.

Yu, H., Chen, D., Wu, Z., Ma, X., Wang, Y., 2016. Headway-based bus bunching prediction using transit smart card data. Transportation Research Part C, 72, 45-59.

Yu, R., Abdel-Aty, M., 2013. Utilizing support vector machine in real-time crash risk evaluation. Accident Analysis & Prevention, 51, 252-259.

Zhao, J., Deng, W., Song, Y., 2014. Ridership and effectiveness of bikesharing: The effects of urban features and system characteristics on daily use and turnover rate of public bikes in China. Transport Policy, 35, 253-264.