Accepted Manuscript

iBike: Intelligent public bicycle services assisted by data analytics

Yin Zhang, Haoyu Wen, Feier Qiu, Zie Wang, Haider Abbas

PII: S0167-739X(18)32278-7
DOI: https://doi.org/10.1016/j.future.2018.12.017
Reference: FUTURE 4638

To appear in: *Future Generation Computer Systems*

Received date : 24 September 2018
Revised date : 1 December 2018
Accepted date : 12 December 2018

iBike: Intelligent Public Bicycle Services Assisted by Data Analytics

Yin Zhang[a,*], Haoyu Wen[a], Feier Qiu[a], Zie Wang[a], Haider Abbas[b]

[a]*School of Information and Safety Engineering, Zhongnan University of Economics and Law, Wuhan, China*
[b]*Department of Information Security, National University of Sciences and Technology, Islamabad, Pakistan*

Abstract

Bicycle sharing systems are becoming increasingly prevalent in urban environments. These systems provide an environmentally friendly transportation alternative in cities. The management of these systems faces many optimization problems. The most important of these problems are the individual maintenance of bicycles, rebalancing and shared facilities, and the use of systems by creating requirements in asymmetrical patterns. A series of data mining tasks based on real data sets is performed to solve the problem of unbalanced bicycle use.

By analyzing the characteristics of each station, the stations are modeled from the perspective of individuals and clusters by means of different models. The evaluation indicators used to address the accuracy of the results provide an effective method for shared bicycle predictions.

Keywords: public bicycle services; data analytics; artificial intelligence; intelligent transportation.

1. Introduction

With the economic development in recent years, human activities have caused increasingly serious pollution and damage to the natural environment [1]. With the expansion of the Internet of Vehicles (IoV) in smart cities, the public bicycle system has developed quickly as a new green transportation mode [2, 3] and is substantially changing the travel habits of citizens around the world, especially in China.Therefore, sustainable development has become the consensus of the international community [4]. In this case, the public bicycle system (PBS) was developed as a substitute for short-distance vehicles due to its low pollution, low energy consumption and high flexibility. In addition to reducing travel needs for personal vehicles, public bicycle sharing systems not only help to extend the range of transit and walking travel and provide people with healthy transportation options but also generate greater interest in cycling and increase the number of people who ride bicycles [5]. As of the end of 2016, more than 1,100 cities actively operated an automated bicycle sharing system, with a total of 2,000,000 public bicycles deployed worldwide. With a bicycle sharing system, users can easily use a smart card to rent a bicycle at a nearby station, use the bicycle on short trips, and return the bicycle at another station [6].

The shared bicycle system can be considered to belong to the IoV [7]. Scanning a code to unlock a bicycle is a feature of shared bicycles. These smart locks are based on the Internet of Things.

The IoV principle of shared bicycles adopts the architecture of mobile terminal-cloud-bicycle terminal. PBSs use the more common Internet of Things (IoT) application architecture: cloud-user-terminal. Although the IoT architecture is not the only possibility, most of the current services use this architecture.

Due to the increasing importance of bicycle sharing programs and the operational difficulties in managing them, people have become concerned about various issues related to bicycle sharing [8]. Many problems need to be resolved in the rapidly rising industry of shared bicycles. 1. The uneven distribution of stations wastes resources: In some regions, there are fewer users with multiple stations, and supply is in short supply. In some regions, there are few users and the utilization rate is low. After some stations are put into operation, the use of bicycles is low, and the use of regional stations is low, resulting in a waste of resources. By contrast, other stations are extremely active, and the delivery of vehicles does not satisfy the demand. 2. The user experience is poor: In places where the number of bicycles borrowed and returned does not match, causing no bicycles to be available. 3. The frequency of use of bicycles is not balanced, resulting in high maintenance costs: The number of uses per bicycle varies greatly, and frequent use and frequent idleness can affect the useful life of bicycles and increase the cost of vehicle maintenance. Bicycles in short supply at popular stations are frequently. Thus, we must optimize these maintenance issues via data analytics.

To solve the above problems, this paper focuses on the research of station-based shared bicycles. The main con-

*Corresponding author
Email addresses: yin.zhang.cn@ieee.org (Yin Zhang), haoyuwen@zuel.edu.cn (Haoyu Wen), feierqiu@zuel.edu.cn (Feier Qiu), ziewang@zuel.edu.cn (Zie Wang), dr.h.abbas@ieee.org (Haider Abbas)

tribution of the article is discussing shared bicycle behavior from the perspective of individual stations and clusters by analyzing and processing the behavior of station-based shared bicycles in a city. A clustering model is used to predict the behavior of shared bicycles. The stations are allocated according to usage. The scheduling problem of vehicles is solved according to the borrowed and returned numbers for the predicted stations. Finally, bicycle usage is summarized to schedule proper maintenance of equipment. The existing shared bicycle system is improved by solving the above problems, so the user has a better experience and the manufacturer has a better plan for vehicle scheduling and maintenance.

2. Related works

Researchers have conducted substantial research to describe the nature of bicycle sharing systems, business models, optimization, and the combination of the Internet of Things. For example, Shaheen et al. reviewed the history, advantages and disadvantages of global bicycle sharing systems [9]. Martin et al., with the advent of public bicycle sharing, assessed the dynamics of transit mode transitions [10]. Parkes et al. used diffusion theory to compare the adoption of bicycle sharing in Europe and North America [11] and proposed the concept of the Internet of Shared Bicycle (IoSB) for the first time to find a feasible solution for the technical problems of shared bicycles.[2]. Although the research direction is different, it has inspired us to promote shared bicycle systems.

We know that shared bicycle systems and the IoV are inextricably linked, and IoV technology is an integral part of the Internet of Things. At present, most shared bicycles use traditional 2G networks for networking, but these systems will gradually transition to Narrow Band Internet of Things (NB-IoT) networks in the future [12]. In fact, many companies have begun to use NB-IoT technology [13] in shared bicycle network transmission. NB-IoT is an emerging technology in the IoT space that supports low-power devices via cellular data connection, also known as low-power wide area network [14]. NB-IoT supports efficient connection of devices with long standby time and high network connection requirements. Related research has been performed to make it easier to collect shared bicycle system data using IoT technology. Zguira.Y et al. proposed an efficient, IoB-DTN, a routing protocol based on data aggregation that applies the delay tolerant network (DTN) paradigm to IoT applications running data collection on urban bicycle sharing system based sensor networks [15]. Such a technology represents great convenience for shared bicycle systems. The data collection system mentioned in this paper is also supported by this technology.

The problem of bicycle sharing inspired researchers to study the optimization of sharing systems. The main focus was on the dynamic perception of bicycle sharing system data, and two topics were considered extensively, namely, clustering and prediction. Currently, most clustering methods can identify movement patterns [16] in bicycle use and divide these stations into clusters based on usage. For example, in [17], two clustering techniques use activity statistics derived from the evolution of station occupancy or the number of bicycles available on a given day [18]. Other researchers have focused on the user and motivated users to balance the bicycle sharing system. Singla A et al.[19] proposed a crowdsourcing mechanism that motivates a user to perform a bicycle repositioning process by providing an alternative to the user to select or return the bicycle in exchange for a monetary incentive. This idea is a subjective and innovative approach, but based on the consumption level of shared bicycle systems, the use of a monetary incentive mechanism to greatly optimize this method is challenging. However, most of these studies focused on studying vehicle scheduling problems through clustering and forecasting and were unable to implement the regional distribution of the stations and the vehicles themselves. In this paper, the demand for bicycles is predicted, and the vehicle scheduling and maintenance are experimentally elaborated.

The issue of forecasting bicycle sharing demand has attracted considerable attention. Based on the spatial granularity of focus, three sets of prediction models are considered in the current study: city level, cluster level, and station level. For city-level groups, the goal is to predict bicycle use throughout a city. In 2014, Kaggle, the world's largest predictive modeling and analysis competition platform, invited participants to predict the total hourly demand for the Capital Bikeshare system in Washington, D-C. Giot and Cherrier (2014) analyzed the demand forecast for the next 24 hours and provided the city granularity of the Capital Bikeshare system[20]. They tested a variety of machine learning algorithms, such as ridge regression, Adaboost regression, support vector regression, stochastic prediction trees, and gradient-enhanced regression trees, and showed that the former two performed better than the others [21]. Other researchers have used deep learning methods [22, 23] to estimate inventory rebalancing needs in shared bicycle systems. Mrazovic et al. [24] proposed a multi-input and multi-output deep learning model based on long short-term memory networks to predict users' long-term needs. Although this article is not intended to improve the algorithm, we compare various algorithms in the experiment, and finally choose one with the best results to provide an experimental demonstration.

In this demand forecasting problem, the ultimate goal of researchers is not to obtain specific values but to optimize the system to schedule the rebalancing of shared bicycles. Thus, cluster classification is applicable. Stations with the same characteristics are grouped into the same class and the same planning is applied to each station within a cluster. Zhou (2015) applied community detection algorithms and agglomerative hierarchical clustering to group similar bicycle traffic and stations in the Chicago

PBS. The paper confirms that the clustered bicycle usage patterns are different for daytime, user, orientation, and land use profiles [25]. Other studies have applied clustering algorithms and shared demand forecasts for cluster-level bicycle assessments. Li et al. (2015) proposed a bicycle sharing demand forecasting framework that introduces a dual-station clustering algorithm to group individual stations. The bicycle sharing needs of the entire city are predicted based on a gradient-elevation regression tree; then, based on multiple similarity reasoning, models are segmented between clusters [26]. Chen et al. (2016) noted that station clustering should be updated based on time and weather factors and social and traffic events. They proposed a geographically constrained station clustering method on a weighted correlation network to dynamically group stations with similar bicycle usage patterns. Then, they estimated the rent and return rate at the cluster level and adjusted the average value of the cluster through the inflation rate[27]. However, a small error in the cluster-level prediction will be multiplied several times, resulting in an unsatisfactory final result, which is accompanied by a large error risk. Therefore, this paper directly predicts these rates at the station level.

Station-level bicycle sharing demand forecasts are challenging and have attracted considerable interest from researchers. Faghih-Imani et al. (2014) built a similar linear hybrid model based on a two-day data set of the Montreal BIXI (bicycle+taxi) bicycle sharing system to predict the hourly bicycle sharing needs at the station. A 250 meter long buffer was set for each station to generate an explanatory variable in the model [28]. The two studies did not consider any potential correlation between stations. For example, if a bicycle station near a subway exit has higher demand during peak hours, another station nearby may also have high demand. Yang et al. (2016) proposed a probabilistic movement model that considers previous checkout records and journey duration to estimate the future registration number for each station [18]. For bicycle inspections and demand forecasts, they apply a random forest tree algorithm for each individual station, without considering spatial or temporal correlations between stations [29]. The research in this article is based on station-level bicycle sharing needs. The main issues discussed include station activity, vehicle scheduling, and vehicle maintenance. This article considers these issues from the perspective of a single station with clustering. Considering each station's arriving bicycle/returning bicycle and time dimension, the stations are divided into several categories to study separately. This distinguishes this study from previous research. Many researchers have considered only the perspective of clustering when studying sharing bicycles. However, these studies are not comprehensive, they are not considered from a single point, and they are not specific to the station or the vehicle itself. Finally, we use K-means clustering and the extreme gradient boosting (XBGOOST) algorithm to predict the features extracted from the data preprocessing stage. The results are evaluated using the root mean square error (RMSE) and mean absolute error (MAE). The prediction results are found to be satisfactory.

Table 1: 1 Data set properties

LEASEDATE	*CARDSN*	*RTLOCATIVEID*
LEASETIME	*RTSHEDID*	*VIPCARDSN*
USETIME	*RTDATE*	*OPTYPE*
SHEDID	*RTTIME*	*OPNAME*
LOCATIVEID	*OVERTIME*	

3. Framework

The leasing company collects user data through the public bicycle service system. After obtaining the shared bicycle data, we separate the important information in the data for separate analyses. For the purpose of forecasting, the first thing to discuss here is whether or not borrowing a bicycle and returning a bicycle need to be discussed. After analyzing the data, we find that the borrowing mode of the vehicle is essentially the same. Both have similar characteristics and can use a common model.

Then, we analyze the data from the dimension of time. Each month has the same loan repayment mode, the same weekly loan repayment mode, and the same daily loan repayment distribution. Therefore, the following time-related characteristics can be considered: month, date, hour, minutes, and whether it is a workday or a rest day. From the dimension analysis of the station, a liveness feature is defined for the station according to the number of borrowed and returned bicycles. From this aspect, it is possible to determine the location of a station and analyze it. Clearly, historical vehicle borrowing analysis must consider the past behavior, including data for every half-day, every month, and the average number of borrowed vehicles.

Then, the features discussed are used as input one after another, and the features that have a significant impact on the prediction result are analyzed in greater detail. Finally, the significant features are input to predict the bicycle behavior. The prediction and analysis results are then applied to solve the problem of station allocation, vehicle scheduling and vehicle maintenance.

The experimental process is illustrated in Figure 3.

4. Data analysis

4.1. Data Set

The data set for this article is from a shared bicycle rental company. The attributes of the data set are shown in Table 1.

We divide the data set and use the data from May to July 2015 as the training set and the data from August 2015 as the test set. In the experimental process, the time,

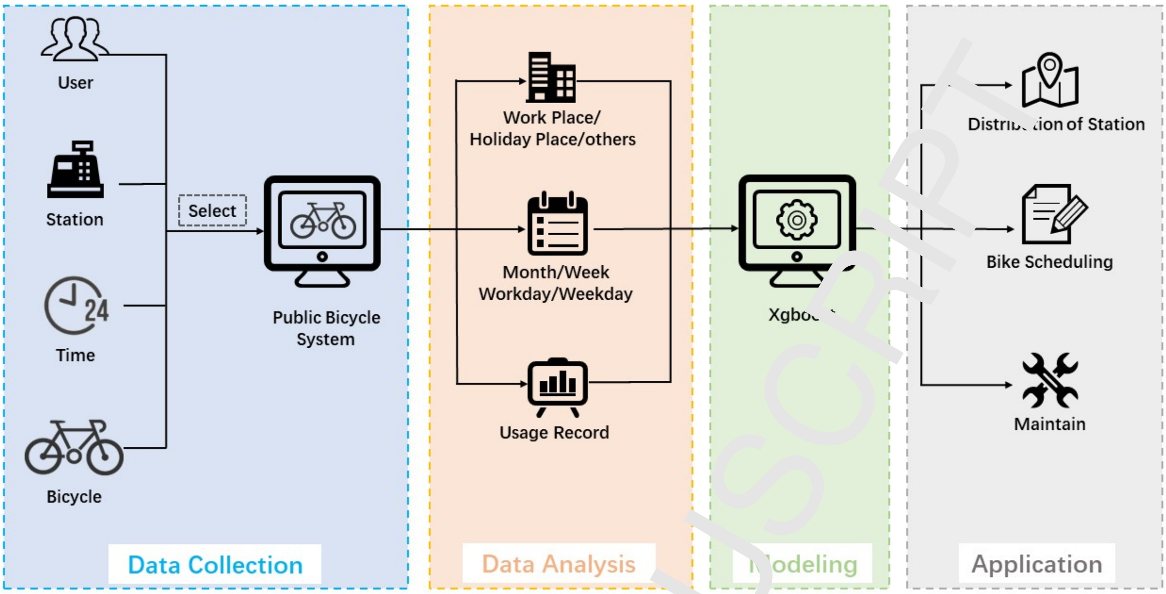

Figure 1: framework

station, bicycle, returning bicycle, and analysis process are taken as input, and the station, time, bicycle and return amount are the forecast results.

4.2. Feature Analysis

Many features, such as riding time, are included in the initial data analysis phase to predict the distance between stations or the links between stations, the links between active stations, the impact of a user's activity, the number of vehicles to use in consideration of damage and repairs, station activity to balance delivery, station liveness to predict locations, etc.

4.2.1. Intuition/Assumption

The experiments showed that not every feature impacts the predictions. Only some features play a role in the prediction result. The following most effective features are selected through experimental screening.

1. Daily peak hours of bicycle use: Daily commuting hours (8:00-9:00, 11:00-12:00, 13:00-14:00 and 17:00-18:00). Bicycle use is higher during peak bicycle use, and the other times are ordinary times.
2. Weekdays/days off: Mondays to Fridays are classified as working days, and Saturdays, Sundays, and holidays are classified as holidays. Workdays are compared with the peak hours of holidays.
3. Station activity: Researchers previously used clustering tools to classify traffic based on stations that showed similar usage, not the stations themselves. The difference in this article is to classifying the activity of stations at 3 levels. The classification is based on the following: the quarter of stations with the lowest average daily activity of borrowing/transfer are the low-active stations; the quarter of stations with the highest activity are the active stations; and the remaining are general stations.
4. Fixed users of a station: Near each station, there are users who borrow bicycles frequently and users who borrow bicycles from only one or two stations. These users are assumed to have fixed bicycles at a certain station. There are several ways to identify a fixed user: a user borrows a bicycle from a certain station and also returns the bicycle to this station, or based on a threshold of the total number of times a certain user borrows a bicycle at a specific station (candidate).
5. Whether the station is at a work place/tourist location: To determine whether a station is in a work area, the active stations are considered. Suppose a station is a tourist attraction or work area. If a user arrives to the location and returns a bicycle here, when they are leaving, the user is very likely to borrow a bicycle from the same location. If there are many similar users, then on a daily basis, there will not be a considerable change in the number of bicycles borrowed and returned. Therefore, if a station is in equilibrium from Monday to Friday, we classify it as a place of work; if a station is in equilibrium on holidays, we will classify it as a tourist attraction.
6. Monthly/weekly/daily lending information: There is an overall trend in the daily borrowing/returning information for each week, which will have an impact on our forecast.

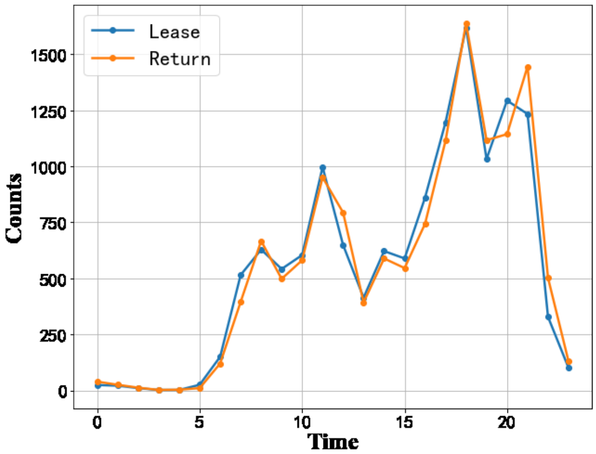

Figure 2: LEASE/RETURN Information

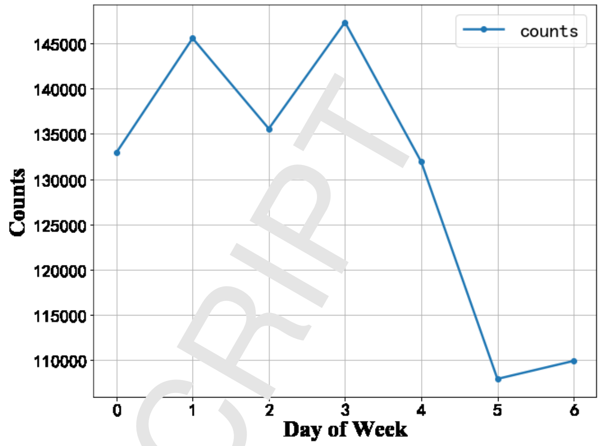

Figure 3: Borrowing from Monday to Sunday (0 means Monday, 6 means Sunday)

The above analysis showed that the characteristics that have a significant impact on the forecast results are station activity level, forecast time period, total monthly/weekly vehicle returns, working days/off days, and the distribution of borrowed vehicles. We collected relevant information from the data set to illustrate these features.

4.2.2. Feature Explanations

The features proposed in 4.2.1 can be further explained by considering the corresponding data from the data set. For example, for the vehicle borrowing information, we randomly selected and compared stations in terms of bicycle and bicycle returning information. The results are shown in Figure 2. The blue broken line in the figure represents the number of returned bicycles, and the orange curve represents the number of borrowed bicycles. The curves for borrowing and returning are very similar, and the curves are basically consistent. Therefore, there is no significant difference between the number of bicycles borrowed from and returned to each station.

The data from Monday to Sunday show a downward trend in the number of bicycles borrowed on the weekend (Figure 3). In this situation, we consider the weekend as a vacation. The number of commuters on weekends is less than the number on working days, which leads to the reduction in volume. Therefore, when using the model predictions, holidays and workdays must be studied separately.

To more clearly illustrate the use of vehicles at different times of the day, the bicycle usage data for each time period from 0 to 24 hours is extracted, as shown in Figure 4.

The abscissa of Figure 4 represents time, and the ordinate represents number of rentals. The total number of bicycles borrowed at 7:00 from Monday to Friday is approximately 14,000-15,000, the total number of bicycles borrowed at 11:00 is approximately 9,000-1,0000, the total number of bicycles borrowed at 14:00 is approximately 6,000-7,000, and the total number of bicycles borrowed at 18:00 is approximately 15,000-16,000.

The number of bicycles borrowed shows several fluctuations throughout the day, and the trends of the number of bicycles borrowed from Monday to Friday are essentially the same. Although the number of leases on weekdays is significantly less than the number on weekdays, the hourly trends are similar: peaks are observed at 7:00, 11:00, 14:00 and 18:00. Therefore, we conclude that the peak hours of workdays are 7:00, 11:00, 14:00 and 18:00. The peak period of bicycle use is a key feature for predicting vehicle scheduling.

Because the sharing system is oriented towards society, if the system is widely distributed in cities and even the whole country, the issue of cost will have to be considered. From the perspective of the equipment itself, we can effectively reduce the reinvestment in bicycles and avoid large amounts of waste if we maintain the vulnerable vehicles in time and recycle them. The damage to various vehicles is assessed in terms of the following aspects: (1) vehicle is used ¡= 1 min, (2) a vehicle is no longer used after the last time the vehicle is used, (3) the vehicle has not been used during the peak period of the active, (4) the last time the bicycle was used, the bicycle was abandoned. This information can be used only as a reference; the data set does not show that a bicycle was not returned, but the rental company will have corresponding information.

5. Modeling

5.1. Problem Definition

Definition 1. *Trip.Travel $Tr = (So, Sd, o, d)$ is a bicycle usage record, where So represents the starting station, represented by RTSHEDID; Sd represents the destination station, represented by LEASESHEDID; and o and d are the bicycle Check-out and Check-in time [26].*

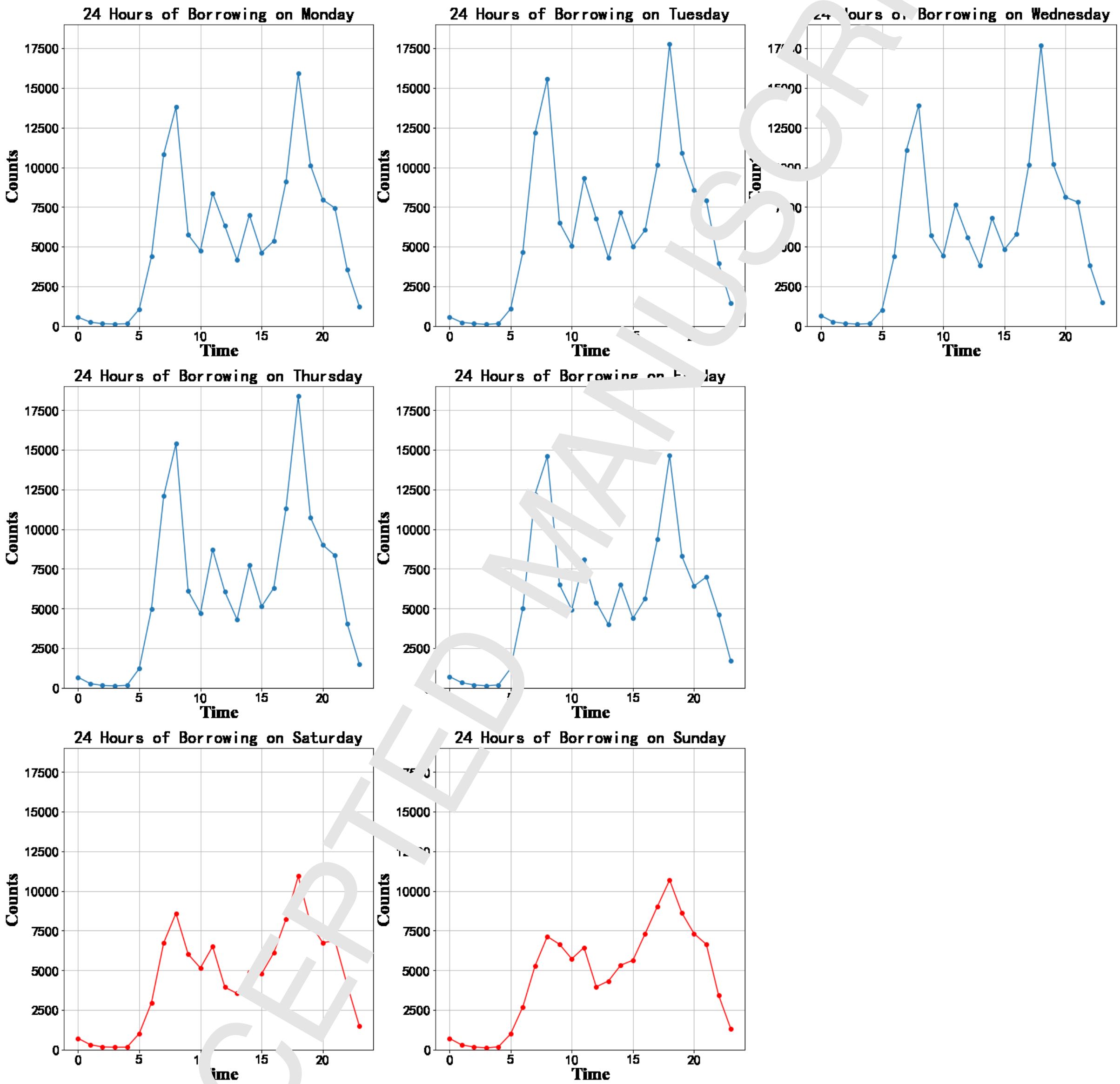

Figure 4: 0-24 hours borrowing information

Definition 2. *TimeInterval. The bicycle use time period $Time = \{Tv, Tw, Tr\}$, where Tv and Tw indicate WorkDay and WeekDay, respectively, and Tr is the peak daily bicycle use.*

Definition 3. *Check−out/Check−in. Check-out/Check-in refers to the number of bicycles that are loaned out and returned by Time.*

Definition 4. *UsageCounter. UsageCounter refers to the number of uses per bicycle.*

Definition 5. *Problem definition: Check-in/Check-out prediction problem. Given a set of historical trips $TH = \{Tr1, Tr2, Tr3...TrH\}$, we want to predict the Check-in/Check-out of each station within a certain time period.*

Modeling from a single station: Each station's information is assessed in terms of monthly, weekly, daily, work day and off-station considerations, and the peak period of station usage is compared to the average period. These factors can have a substantial impact on the station's Check-in/Check-out.

Modeling from a station cluster aggregation perspective: Stations are divided into three levels with different levels of activity. The one-quarter of stations with the lowest average daily number of borrowed/returned bicycles are low-active stations; the one-quarter of stations with the highest number of borrowed/returned bicycles are the active stations; and the others are general stations.

5.2. Algorithms

This article uses the *XGBoost* algorithm to solve the shared bicycle prediction problem. Three aspects must be considered when selecting an algorithms. First, station-based and time-based feature selection and some specific classifications based on these two major points, the algorithm of this paper must support an approximate classification. The method requires high classification accuracy. Second, high flexibility in the optimization and evaluation of results. Third, a high-performance and fast algorithm. Taking into account the above points, we select the *XGBoost* algorithm, which implements a generic tree boosting algorithm. One representative of this algorithm is the gradient boosting decision tree (*GBDT*), which was proposed in February 2014 and focuses on gradient learning algorithms. The library has received extensive attention due to its excellent learning effect and efficient training. It combines much of the previous work on the gradient lifting algorithm, has been applied to many optimization problems in project implementations, and is one of the most successful machine learning algorithms.

The objective function of *XGBoost* is as follows:

$$o_{bj} = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k) \tag{1}$$

where $l(y_i, \hat{y}_i)$ represents the training loss value, y_i is the real number of bicycles leased and returned, and $\hat{y}_i$ is the predicted number of bicycles leased and returned. The regularization term $\Omega(f_k)$ controls the complexity of the model, including the number of leaf nodes T and the L2 modulus square of the leaf score. γ represents the minimum loss reduction required to further partition a leaf node of the tree: the larger the value is, the more conservative the algorithm will be. λ represents the L2 regularization penalty coefficient.

$$\Omega(f_t) = \gamma \mathrm{T} + \frac{1}{2}\lambda \sum_{j=1}^{\mathrm{T}} w_j^2 \tag{2}$$

The details of the formula derivation are provided in the references [30].

This formula agrees with the ID3 algorithm (using entropy to calculate the gain) and the CART algorithm [31] (which uses the Gini index to calculate the gain). The formula uses a certain value after splitting minus a certain value before splitting to determine the gain. To limit the growth of the tree, we can add a threshold and allow a node to split when the gain is greater than the threshold. Gamma in the above equation is the threshold, which is the coefficient of the number of leaf nodes T in the regularization term. Therefore, *XGBoost* is equivalent to the objective function.

Pre-pruning can then be performed. In addition, the coefficient lambda in the above formula is the coefficient of the L2 norm square of the leaf score in the regularization term. Smoothing the leaf score also plays a role in preventing overfitting. This feature is not included in the conventional gradient boosting decision tree (*GBDT*). Moreover, regularization is used in *XGBoost*. The standard gradient boosting machine (*GBM*) implementation does not have regularization steps, in contrast to *XGBoost*. Regularization is also helpful in reducing overfitting. This method has a high degree of flexibility, enabling users to define custom optimization goals and evaluation criteria. Additionally, *XGBoost* has built-in rules for handling missing values.

After selecting the algorithm, the features are entered from different perspectives, including the following:

Modeling from a single station: Each station's information is assessed in terms of monthly, weekly, daily, work day and off-station considerations, and the peak period of station usage is compared to the average period. These factors can have a substantial impact on the station's Check-in/Check-out.

Modeling from a station cluster/aggregation perspective: Stations are divided into three levels with different levels of activity. The one-quarter of stations with the lowest average daily number of borrowed/returned bicycles are low-active stations; the one-quarter of stations with the highest number of borrowed/returned bicycles are the active stations; and the others are general stations. From the perspective of clustering, this paper uses K-means clustering and specifies the number of categories as three. K-means clustering is performed using the average number of borrowed and returned bicycles from Monday to Sunday at each station. The results of the clustering method are not ideal, and the deviation is large.

5.3. Evaluation indicators

The evaluation metrics used in this paper are RMSE [32] and MAE. The RMSE, also called the standard error, is the square root of the square of the observed deviation from the true value and the ratio of the observed value times n. The average absolute error is the average of the absolute error. The average absolute error accurately reflects the actual prediction error. This article uses two types of evaluation indicators to display and compare the evaluation results to make the prediction results more rigorous.

The *XGBoost* algorithm is used to make predictions, taking $\{Time, Station, RT, LEASE, F1, ...Fn\}$ as input, where $Time$ represents time, $Station$ represents the station, RT and $LEASE$ indicate borrowed and returned, respectively, and $F1$ through Fn represent the available features. The output is $\{Time, Station, RT, LEASE\}$, that is, the number of bicycles borrowed and returned at each station at a certain time.

6. Experiments

6.1. Experimental design

This paper now begins to analyze different angles, different features, and combinations of features. Modeling is performed using the *XGBoost* algorithm, which has high accuracy compared to other algorithms, and then the algorithm is used for prediction. The influence of different features on the results are compared to identify the important features. The goal is to find the combination of features with the highest accuracy. The shared bicycle system, including bicycle rebalancing, bicycle scheduling, and bicycle maintenance, is then optimized based on these characteristics. We find that the time dimension and the spatial dimension features have the strongest effects on the demand for bicycles at a certain station. The time dimension is the time when the bicycle is borrowed and returned. The spatial dimension includes the distribution of the station itself, the geographical location, and the degree of activity. We compare the impact of each feature on the prediction results to identify the most important features to improve the accuracy of our predictions and help us to better analyze and optimize the problems raised above.

6.2. Station distribution

Public bicycle stations have obvious problems with irrational distribution, which results in a lack of access to stations in small-use areas and a waste of resources. In some regions, the use of bicycles is high, and the number of stations is small. This issue has led to the use of vehicles in such regions being concentrated on a small number of stations, and the pressure on these stations is excessive.

To solve the problem of irrational station distribution, this paper analyzes the Check-in/Check-out data during peak hours and regular hours for each station on May 2 (Saturday) and May 4 (Monday). The data were visualized, and the results are shown in Figure 5 to Figure 8.

The above visualizations are force layout diagrams drawn with the EChart tool[33]. Figures 5 and 6 show the following. 1. Count: the total number of vehicles at the considered stations at certain times of the day, which provides a representation of size. 2. Statistics of traffic between all involved stations during a certain period of time each day, which is represented on the diagram by a line between circles, in which the duplicated links (same as borrowing stations) are represented by a single curve. 3. The borrowing information of the station itself is shown as a separate circular representation on the map.

The station-related information is provided in the visualizations shown in Figure 5 and Figure 6. There is a significant increase in the peak period with the normal period. We believe that the number of users during the commuting period is greater than that in other periods. Furthermore, the stations with different degrees of activity in Figure 5 and Figure 6 are clustered together and have close relationships. We speculate that many active stations are also interconnected.

The visualization also provides an important piece of information about the distribution of stations. Although the specific locations are not shown in the figure, the relative locations of the stations are clearly visible, such as the distribution of stations in the two red circles in Figure 5.

The stations in the red circle on the left side are far from most of the clustered stations and can be considered to be far from the city center. At location 220, the stations are relatively aggregated, and the usage rate is not high. Therefore, the number of stations in the area should be reduced. The stations in the red circle area on the right side of the figure are dense. There is a shortage of bicycles and stations.

For example, the areas labeled 182, 168, and 181 in the upper-right corner have a large number of stations, but there are few available bicycles. New stations should be added in nearby areas to reduce the pressure on the individual stations. There are still a lot of these two types of stations, such as 222 and 318 with low bicycle usage and 181 and 301 with high usage. It would be reasonable to reduce the number of with lower usage, and in high usage areas, new stations can alleviate the pressure on individual stations to ensure the rational and efficient distribution of public bicycle stations.

6.3. Vehicle scheduling

6.3.1. Method Comparison

In the course of the experiment, some features of data analysis are proposed, and the time-related Check-in/Check-out and station-related Check-in/Check-out data are input

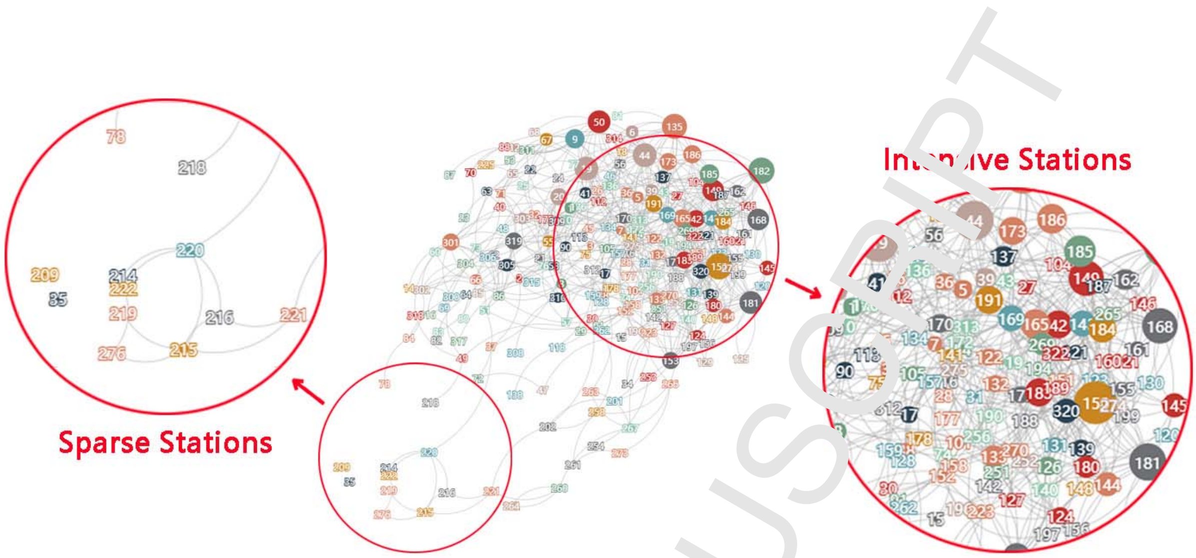

Figure 5: May 2, 8:00-8:30

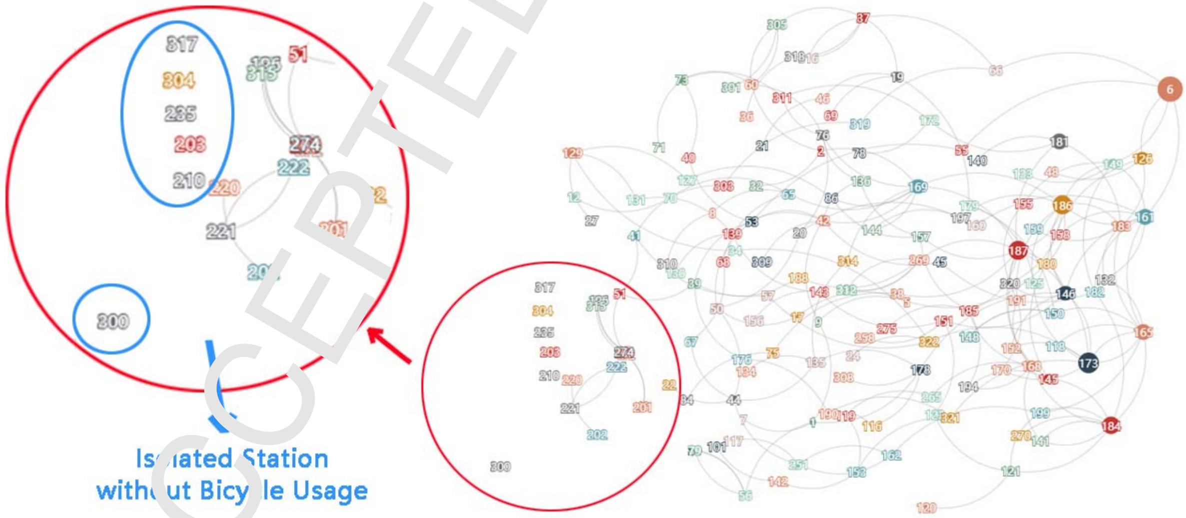

Figure 6: May 4, 15:00-15:30

Table 2: Analysis of the results

Feature	RMSE	MAE
Every day	9.3282	6.0824
Working days/days off	10.1171	6.5500
Time features	10.0124	6.8103
Station repayments	9.9555	6.3723
Station activity	9.9641	6.3999
Station features	9.9595	6.3751
All features	10.4201	6.7787

into the *XGBoost* prediction. The RMSE and MAE are used to compare the results. By comparing the deviations of different feature prediction results, we can determine the influence of the corresponding features.

Then, the K-means clustering results are compared with the prediction results of the *XGBoost* algorithm to determine which method is more suitable for shared bicycle predictions.

6.3.2. Results Comparison

In the experiment, some features were extracted, and the daily and weekly time characteristics and all time characteristics were input. The station's return quantity and liveness were input, and the results were assessed in terms of RMSE and MAE. The comparison results are shown in Table 2.

6.3.3. Analysis of the results

Comparison of the results in the previous section shows that the features have different influences on the prediction results, and the characteristics of the time dimension have a greater influence than those of the station dimension. However, when the complete set of valid features is used as input, the results are not good because other features discussed in the data analysis are also included in the experiment. Some of these features, such as time, play a positive role in the prediction results. The relevant characteristics include some redundant features, although these redundant features do not significantly improve the prediction results. Some of the features bias the prediction results but provide important information for bicycle prediction and are therefore necessary to include.

The experimental results show that when the daily borrowing amount of a specific station is included in the feature analysis, the prediction results are significantly improved. Additionally, when the time characteristics are expanded to a weekly or monthly time point or the station characteristics are expanded to all stations, the effect on forecasting is reduced. In this case, we believe that the more specific and detailed the features are, the greater the degree of influence on the forecast, and the more abstract the features are, the less the impact on the forecast results. This article also used K-means clustering for prediction, but the results were not ideal and are not considered further. Of course, this conclusion is based on the data sets and experiments in this article.

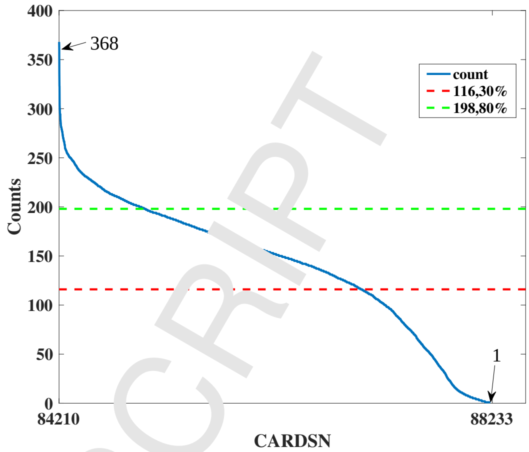

Figure 7: Bicycle usage

This article uses the above experimental results to predict the number of vehicles to be retired at each station, to implement effective vehicle scheduling, and to distribute public bicycles to each station. Proper management will relieve pressure on the peak usage at a station while balancing the use of bicycles and avoiding the status quo where no bicycles ares available or no space is available to return a bicycle.

6.4. Vehicle maintenance

To realize the long-term development of public bicycles, another problem that must be solved is the maintenance of bicycle equipment. The cost of each bicycle is high, and damage to or abandonment of bicycles often occurs. Additionally, when some bicycles are used too frequently, they cannot be repaired in time. This problem requires an effective solution.

The number of uses of each bicycle in the data set for three months is shown in Figure 7.

In Figure 7, we can see that the maximum number of times a bicycle was used in the three months was 368, and the minimum number of uses was 1. Frequent bicycles are used more than 2-3 times per day, whereas other bicycles are left untouched for months. We propose a corresponding countermeasure, which divides the bicycles into those that are used too frequently and those that are used to infrequently. Considering that there are relatively few vehicles with extremely frequent use in data sets, the number of uses for the 30th percentile (116) is the lower limit, and vehicles with fewer than 116 uses are considered to be used too infrequently. The number of uses of the quintile (198 times) is the upper limit. Vehicles used more than 198 times are classified as being used too frequently. We believe there are two situations in which bicycles are used too infrequently. First, the vehicle is at an active station, but the use is abnormal (less often). Second, the vehicle

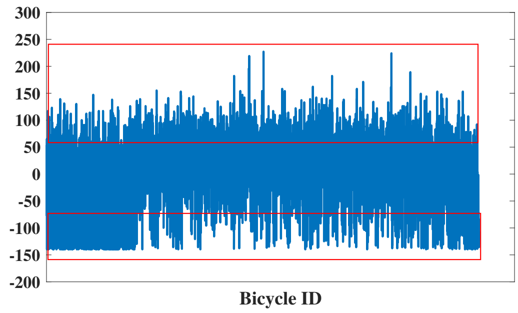

Figure 8: Use frequency fluctuation diagram

is used only once or several times. Bicycles that fall into these conditions are in question (to be maintained). Then, we need to inform the manufacturer to perform timely recovery or maintenance to put the bicycle back into use. For vehicles that are used too frequently, vehicle wear is more serious. If it is not repaired in time, the bicycle may become unusable. The manufacturer should be informed to ensure prompt recovery for maintenance. For such unbalanced vehicles, effective maintenance methods can effectively reduce the cost of waste and save production materials, and vehicle redistribution strategies can be implemented.

The frequency of bicycle use shown in the figure fluctuates considerably. The abscissa represents the distribution of all bicycle IDs. On the vertical axis, 0 represents the average number of bicycle use times, negative values are below the average, and positive values are above the average. In only three months, the highest value differs from the lowest value by more than 360. The red box shows that more bicycles are used too frequently than too infrequently. This figure accurately shows that bicycle usage is very unbalanced.

After applying our analysis and solution methods, the number of uses of infrequently used bicycles will gradually increase, whereas the number of uses of frequently used bicycles will gradually decrease. Finally, the two will gradually approach a mean value, reaching a relatively balanced state.

7. Discussion

7.1. Scenarios

Public bicycles are used by a wide range of people, and stations are spread all across the country. In the case of short-distance travel and temporary hurdles, the masses tend to use public bicycles. Bicycles provide energy saving and environmental protection, and users can exercise and use them conveniently. However, as public bicycles have become more popular, the number of input stations has increased, and more and more vehicles have been put into use. The management and maintenance of these bicycles will continue to cause many problems.

Scenario 1: Dense/Sparse Residential Areas

In urban centers, the population is dense, and short-distance travel is common. Public bicycle stations distributed around dense residential areas are used more frequently. However, the stations in theses area are limited. The use of public bicycles is concentrated on nearby stations and is insufficient to satisfy the demand. There is a huge demand for use in these regions. In the relatively sparsely populated areas of urban suburbs, there are more long-distance trips. Although there are many public bicycle stations, more people choose public or other means of transportation. Therefore, many of the stations in these region are not active (in some cases, no one visits them), causing a waste of resources.

This paper provides a practical and reasonable solution for these two scenarios. The visualization graph (station-to-station relationship diagram) shows the use of regional stations and the links between stations. We can use the public bicycle service system to collect relevant information in the region, perform an aggregate analysis, and create a relationship map to increase the number of stations in regions with a particularly large number of visits to a single station or to several stations to satisfy the needs of users in the region. For less-visited areas, the number of stations can be reduced so users are concentrated in several nearby stations, thereby reducing the waste of station resources.

Scenario 2: City Living Area

The use of modern public bicycles is highest among people who are 20 to 35 years old. In the residential areas, office areas and other areas where public bicycles are frequently used, there are some challenging, repeated situations. In some time periods, such as 8:00 to 21:00, i.e., peak work hours, there are too many users for the available number of bicycles, resulting in some areas where no bicycles are available. At large shopping malls or markets, there are large crowds between 9:00 and 10:00. In this case, the nearby stations are full, and there is nowhere to return a bicycle. A similar situation occurs at tourist attractions on the holidays. When there is a large number of users at a station and no bicycles are available, the user experience suffer. A lack of parking will lead to randomly abandoning bicycles, which will not only prevent additional fares from being collected but may also affect traffic, causing negative social impact.

To solve these problems, we need to start from the dimensions of time and space to predict the active periods of active stations, place staff at these locations at specified times, conduct vehicle scheduling, and transfer vehicles to areas with excessive vehicle volume or insufficient parking places. The experiment in this paper uses relevant data to predict each station's borrowing information. The prediction results can help to balance borrowing to improve the user experience and achieve public recognition and acceptance of public bicycles.

Scenario 3: Damaged bicycles still occupy parking spaces

The widespread use of public bicycles and the lack of effective maintenance often result in vehicle damage or abandonment. Some vehicles are left unattended at active stations, and some vehicles only have their borrowed time recorded (abandoned). In addition, the use of each bicycle is different. Some vehicles in active areas are used hundreds of times in a short period of time, and some vehicles are used only a few times. Vehicles with few uses at an active station can be inspected, and if they are damaged, they can be promptly recovered. For bicycles that are used too frequently, the degree of wear may be considerable, and if they are not repaired in time, it may lead to situations in which they cannot be used or must be completely scrapped. Moreover, if a bicycle fails during use, it may cause an accident.

For the long-term development of public bicycle systems, it is necessary to solve the problem of uneven bicycle load. The use frequency of bicycles can be balanced, and vehicles that are used too frequently can be recovered and rested to solve potential safety risks. By analyzing the use of each bicycle, this experiment accurately identifies the vehicles that are used too frequently and too infrequently and promptly notify the manufacturers so they can perform recovery and maintenance, implement a reinvestment strategy, and put the maintained vehicles back into use. These steps will effectively reduce costs.

7.2. Open issues

The use of public bicycles is becoming increasingly widespread, and the problems faced will continue to increase. For instance, the repatriation of station-style public bicycles needs to be completed at a fixed station, and the bicycles are rented via magnetic cards. This limits the use of public bicycles and does not allow citizens to borrow bicycles any time and anywhere. Moreover, the current social citizenship is not sufficient to achieve a normal loan for each bicycle. Lost vehicles cannot be recovered in many cases. Therefore, anew public bicycle service system is required that can solve the current dilemma of public bicycles while maintaining the input cost within a controllable range.

Although current flexible and parked bicycles (such as OfO and Mobike) are popular, the current state of the system has a major problem, that is, too many vehicles affect traffic and hinders people's normal means of travel. Thus, we need a convenient, low-cost, lightweight and convenient travel tool. A new type of shared device may soon be introduced - folding shared bicycles that are light and fast, take up little space, and avoid parking problems.

8. Conclusion

This paper discusses the convenience of and problems faced by urban bicycles from the perspectives of users, bicycles, and businesses by studying the characteristics of bicycle sharing in cities. On the basis of real data from a public bicycle company for 15 years, this paper conducts a series of data mining tasks. Data visualization and comparative analysis methods illustrate the current problems faced by public bicycle systems, give appropriate recommendations, and promote the sharing of bicycle equipment. Positive development of shared equipment will better serve our society.

[1] M. Chen, J. Yang, X. Zhu, X. Wang, M. Liu, and J. Song, "Smart home 2.0: Innovative smart home system powered by botanical iot and emotion detection," *Mobile Networks and Applications*, vol. 22, no. 6, pp. 1159–1169, 2017.

[2] S. Shen, Z.-Q. Wei, L.-J. Sun, Y.-Q. Su, R.-C. Wang, and H.-M. Jiang, "The shared bicycle and its networkinternet of shared bicycle (iosb): A review and survey," *Sensors*, vol. 18, no. 8, 2018.

[3] M. Chen, Y. Tian, G. Fortino, J. Zhang, and I. Humar, "Cognitive internet of vehicles," *Computer Communications*, vol. 120, pp. 58–70, 2018.

[4] P. Demaio, "Bike-sharing: History, impacts, models of provision, and future," *Journal of Public Transportation*, vol. 12, no. 4, 2014.

[5] M. Chen, W. Li, Y. Hao, Y. Qian, and I. Humar, "Edge cognitive computing based smart healthcare system," *Future Generation Computer Systems*, vol. 86, pp. 403 – 411, 2018. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0167739X17326766

[6] J. Hu, Z. Yang, Y. Shu, P. Cheng, and J. Chen, "Data-driven utilization-aware trip advisor for bike-sharing systems," in *2017 IEEE International Conference on Data Mining (ICDM)*, vol. 00, Nov. 2018, pp. 167–176. [Online]. Available: doi.ieeecomputersociety.org/10.1109/\ICDM.2017.26

[7] M. Razzaque and S. Clarke, "Smart management of next generation bike sharing systems using internet of things," in *Smart Cities Conference*, 2015, pp. 1–8.

[8] E. O'Mahony and D. B. Shmoys, "Data analysis and optimization for (citi)bike sharing," in *Twenty-Ninth AAAI Conference on Artificial Intelligence*, 2015, pp. 687–694.

[9] S. Shaheen, S. Guzman, and H. Zhang, "Bikesharing in europe, the americas, and asia," *Transportation Research Record Journal of the Transportation Research Board*, vol. 2143, no. 1316350, pp. 159–167, 2010.

[10] E. W. Martin and S. A. Shaheen, "Evaluating public transit modal shift dynamics in response to bikesharing: a tale of two u.s. cities," *Journal of Transport Geography*, vol. 41, pp. 315–324, 2014.

[11] S. D. Parkes, G. Marsden, S. A. Shaheen, and A. P. Cohen, "Understanding the diffusion of public bikesharing systems: evidence from europe and north america," *Journal of Transport Geography*, vol. 31, no. 7, pp. 94–103, 2013.

[12] M. Chen, Y. Miao, Y. Hao, and K. Hwang, "Narrow band internet of things," *IEEE Access*, vol. 5, pp. 20 557–20 577, 2017.

[13] R. S. Sinha, Y. Wei, and S. H. Hwang, "A survey on lpwa technology: Lora and nb-iot," *Ict Express*, vol. 3, no. 1, 2017.

[14] R. Ratasuk, B. Vejlgaard, N. Mangalvedhe, and A. Ghosh, "Nb-iot system for m2m communication," in *Wireless Communications and NETWORKING Conference Workshops*, 2016, pp. 428–432.

[15] Y. Zguira, H. Rivano, and A. Meddeb, "Internet of bikes: A dtn protocol with data aggregation for urban data collection," *Sensors*, vol. 18, no. 9, 2018.

[16] M. Chen, Y. Hao, M. Qiu, J. Song, D. Wu, and I. Humar, "Mobility-aware caching and computation offloading in 5g ultra-dense cellular networks," *Sensors*, vol. 16, no. 7, p. 974, 2016.

[17] J. Froehlich, J. Neumann, and N. Oliver, "Sensing and predicting the pulse of the city through shared bicycling." in *IJCAI 2009, Proceedings of the International Joint Conference on Artificial Intelligence, Pasadena, California, Usa, July*, 2009, pp. 1420–1426.

[18] Z. Yang, J. Hu, Y. Shu, P. Cheng, J. Chen, and T. Moscibroda, "Mobility modeling and prediction in bike-sharing systems," in *the 14th Annual International Conference*.

[19] M. Santoni, M. Santoni, M. Meenen, M. Meenen, and A. Krause, "Incentivizing users for balancing bike sharing systems," in *Twenty-Ninth AAAI Conference on Artificial Intelligence*, 2015, pp. 723–729.

[20] M. Wozniak, "Fitness function for evolutionary computation applied in dynamic object simulation and positioning," in *2014 IEEE Symposium on Computational Intelligence in Vehicles and Transportation Systems (CIVTS)*, Dec 2014, pp. 108–114.

[21] P. BORGNAT, P. ABRY, P. FLANDRIN, J.-B. ROUQUIER, and E. FLEURY, "Shared bicycles in a city: A signal processing and data analysis perspective," *Advances in Complex Systems*, vol. 14, no. 03, pp. 1 100 295–, 2011.

[22] M. Chen, Y. Hao, K. Lin, Z. Yuan, and L. Hu, "Label-less learning for traffic control in an edge network," *arXiv preprint arXiv:1809.04525*, 2018.

[23] M. Chen, X. Shi, Y. Zhang, D. Wu, and M. Guizani, "Deep features learning for medical image analysis with convolutional autoencoder neural network," *IEEE Transactions on Big Data*, p. 1. [Online]. Available: doi.ieeecomputersociety.org/10.1109/TBDATA.2017.2717439

[24] P. Mrazovic, J. L. Larriba-Pey, and M. Matskin, "A deep learning approach for estimating inventory rebalancing demand in bicycle sharing systems," in *IEEE Computer Software and Applications Conference*, 2018, pp. 110–115.

[25] X. Zhou, "Understanding spatiotemporal patterns of biking behavior by analyzing massive bike sharing data in chicago." *Plos One*, vol. 10, no. 10, p. e0137922, 2015.

[26] Y. Li, Y. Zheng, H. Zhang, and L. Chen, "Traffic prediction in a bike-sharing system," in *Sigspatial International Conference on Advances in Geographic Information Systems*, 2015, pp. 1–10.

[27] L. Chen, D. Zhang, L. Wang, D. Yang, X. Ma, S. Li, Z. Wu, G. Pan, T. M. T. Nguyen, and J. Jakubowicz, "Dynamic cluster-based over-demand prediction in bike sharing systems," in *ACM International Joint Conference on Pervasive and Ubiquitous Computing*, 2016, pp. 841–852.

[28] A. Faghih-Imani, N. Eluru, A. M. El-Geneidy, M. Rabbat, and U. Haq, "How land-use and urban form impact bicycle flows: evidence from the bicycle-sharing system (bixi) in montreal," *Journal of Transport Geography*, vol. 41, pp. 306–314, 2014.

[29] L. Lin, Z. He, S. Peeta, and X. Wen, "Predicting station-level hourly demands in a large-scale bike-sharing network: A graph convolutional neural network approach," *Transportation Research Part C: Emerging Technologies*, 2017.

[30] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," *CoRR*, vol. abs/1603.02754, 2016. [Online]. Available: http://arxiv.org/abs/1603.02754

[31] A. M. Prasad, L. R. Iverson, and A. Liaw, "Newer classification and regression tree techniques: Bagging and random forests for ecological prediction," *Ecosystems*, vol. 9, no. 2, pp. 181–199, 2006.

[32] T. Chai and R. R. Draxler, "Root mean square error (rmse) or mean absolute error (mae)? - arguments against avoiding rmse in the literature," *Geoscientific Model Development Discussions*, vol. 7, no. 3, pp. 1247–1250, 2014.

[33] L. Zhang, S. Tang, Z. Yang, J. Hu, Y. Shu, P. Cheng, and J. Chen, "Demo: Data analysis and visualization in bike-sharing systems," in *International Conference on Mobile Systems, Applications, and Services Companion*, 2016, pp. 128–128.

Yin Zhang is an Associate Professor of the School of Information and Safety Engineering, Zhongnan University of Economics and Law (ZUEL), China. He is a Wenlan Distinguished Scholar at ZUEL and a Chutian Distinguished Scholar, Hubei, China. He is Vice-chair of IEEE Computer Society Big Data STC. He serves as editor or associate editor for IEEE Network, IEEE Access, Journal of Information Processing Systems, etc. He is a Guest Editor for Future Generation Computer Systems, IEEE IoT Journal, ACM/Springer Mobile Networks & Applications, Sensors, Neural Computing and Applications, Multimedia Tools and Applications, Wireless Communications and Mobile Computing, Electronic Markets, Journal of Medical Systems, New Review of Hypermedia and Multimedia, etc. He also served as Track Chair of IEEE CSCN 2017, TPC Co-Chair of CloudComp 2015 and TRIDENTCOM 2017, etc. He has published more than 80 prestigious conference and journal papers, including 8 ESI Highly Cited Papers. He is an IEEE Senior Member since 2016. He got the Systems Journal Best Paper Award of the IEEE Systems Council in 2018. His research interests include intelligent service computing, big data, social network, etc.

Haoyu Wen received his Bachelor of Engineering degree at Zhongnan University of Economics and Law in 2017. He is currently pursuing a Master's degree at the School of Information and Safety Engineering, Zhongnan University of Economics and Law. His research interests include intelligent service computing, machine learning, etc.

Feier Qiu received her Bachelor of Management degree at Zhongnan University of Economics and Law in 2018. She is currently pursuing a Master's degree at the School of Information and Safety Engineering, Zhongnan University of Economics and Law.
Her research interests include intelligent service computing, big data, social network, etc.

Zie Wang received his Bachelor of Hubei Engineering University in 2017. He is currently pursuing a Master's degree at the School of Information and Safety Engineering, Zhongnan University of Economics and Law. His research interests include data mining, machine learning, big data etc.

Haider Abbas is a Cyber Security Professional who took professional trainings and certifications from Massachusetts Institute of Technology (MIT), United States, Stockholm University, Sweden, IBM and EC-Council. He received his MS in Engineering and Management of Information Systems (2006) and PhD in Information Security (2010) from KTH- Royal Institute of Technology, Sweden. He is an associate editor or on the editorial board of a number of well-established scientific international journals including IEEE Communication Magazine, IEEE Transactions on Information Technology in Biomedicine/IEEE Journal of Biomedical and Health Informatics, Journal of Network and Computer Applications, Neural Computing and

Applications, Electronic Commerce Research (Springer), IEEE Access, and Cluster Computing (Springer). Dr. Abbas also won many awards and received several research grants for ICT related projects from various research funding authorities and working on scientific projects in US, EU, KSA and Pakistan. He is the principal advisor for several graduate and doctoral students at King Saud University, KSA, National University of Sciences and Technology, Pakistan and Florida Institute of Technology, United States. In recognition of his services to the international research community and excellence in professional standing, he has been awarded one of the youngest Fellows of the Institution of Engineering and Technology, U.K.; a fellow of the British Computer Society, U.K.; and a fellow of the Institute of Science and Technology, U.K.

Yin Zhang

Haoyu Wen

Feier Qiu

Zie Wang

Haider Abbas

- It analyzes and deals with the behavior of quay-shared bicycles in cities, and discusses the behavior of shared bicycles from the perspective of individual sites and clusters.
- Clustering model is proposed to predict the behavior of shared bicycles, according to the use of different regional sites to reasonably allocate the site.
- Based on the predicted site borrowing and lending volume, it solves the vehicle scheduling problems, and achieve a considerable bicycle maintenance.