2018 4th International Conference on Big Data Computing and Communications

Demand Estimation of Public Bike-Sharing System Based on Temporal and Spatial Correlation

Xiawen Yao, Xingfa Shen*
School of Computer Science & Technology
Hangzhou Dianzi University, China
735649296@qq.com, shenxf@hdu.edu.cn

Tian He
Dept. of Computer Science & Engineering
University of Minnesota, Twin Cities
tianhe@cs.umn.edu

Sang Hyuk Son
Dept. of Information and Communication Engineering
Daegu Gyeongbuk Institude of Science and Technology (DGIST), Korea
son@dgist.ac.kr

***Abstract*—Nowadays, public Bike-Sharing Systems (BSSs) are broadly deployed in many cities around the world. It is important to obtain accurate user demand of BSS for better system planning and bicycle scheduling. The actual user demand includes not only the users who are served, but also those who are not served by BSS. In this study, we take into account the situations that users are not served for the first time. We propose a three-step demand estimation model to infer the situations that users are not served from both the temporal and spatial correlation, based on the two characteristics of station usage, long-term stability and short-term volatility. The demand estimation model proposed is evaluated based on Washington D.C. bike-sharing system and uses the comprehensive information of three datasets, user trip data, station status data, and station location data. Compared with the ground truth of user demand, the minimum relative error in the experimental results of the entire system is 45.5%.**

***Keywords*-Bike-sharing System, Demand Estimation, Temporal Correlation, Spatial Correlation**

I. Introduction

Recently, public Bike-Sharing Systems (BSSs) have developed rapidly in many cities. Demand estimation is critical to BSS. With accurate user demand, BSS can better plan the deployment of stations, the number of docks in each station and hence improve service quality. The actual user demand includes not only the users who are served, but also those who are not served by system. The amount of users served is the number of trips recorded by system. However, the situations the users are not served (failed to rent or return a bicycle) would not be recorded by system although they actually happen in real world.

There are many studies on BSS which predict the user amount in the future, taking only the amount of users served into consideration, ignoring those users who are not served. Considering the situations that users are not served, in this paper, the study of demand estimation is conducted for the first time.

The key of estimating the actual user demand is to be able to infer the situations that users are not served. We observed that the daily usage of a station has long-term stability and short-term volatility, reflected in temporal correlation and spatial correlation. Based on these two characteristics, we propose a three-step demand estimation model to infer the situations that users are not served and further to estimate the actual user demand. First, from temporal perspective, we make a Pearson correlation analysis between the historical usage data and the usage data of current day, looking for similar days with current day in the historical days. Based on the usage trend of these similar historical days, we infer the situations that users are not served in the current day and obtain an estimated demand. Second, from spatial perspective, we analyze the trip record data of neighboring stations close to the current station on the same day. Based on the usage trend of these adjacent stations, we infer the situations that users are not served in the current station and obtain an estimated demand. Third, we integrate the above two steps to get a final estimated demand.

The demand estimation model proposed is evaluated based on Washington D.C. bike-sharing system. There are a total of three datasets as inputs, user trip data, station status data, and station location data.

Specially, the key contributions of this paper are as follows:

- To the best of our knowledge, we put forward a demand estimation model of BSS for the first time, considering the situations that users are not served occur in real world.
- Our demand estimation model infers the situations that users are not served from both the temporal and spatial perspectives, based on the two characteristics of station usage, long-term stability and short-term

∗ Corresponding author, E-mail: shenxf@hdu.edu.cn

978-1-5386-8021-6/18/$31.00 ©2018 IEEE
DOI 10.1109/BIGCOM.2018.00016

volatility.

- Our demand estimation model is evaluated based on Washington D.C. bike-sharing system and leverages upon the comprehensive information of three datasets, user trip data, station status data, and station location data.

The rest of the paper is organized as follows. Section II introduces the related work. Section III describes the motivation of this work and section IV describes the design concept and the framework of the demand estimation model. Section V proposes the demand estimation model in details, followed by its evaluation in section VI. Section VII presents conclusive remarks and the future work.

II. Related Work

Nowadays, BSSs attract increasing attention and are widely adopted in many major cities around the world. BSSs will generate large amounts of data. Analysis and mining these data will find some special rules and phenomena, and can do some interesting experiments. Yao *et al.* [1] analysed the usage load-unbalance phenomenon in the existing bike-sharing system and proposed a hybrid bicycle allocation strategy to reduce the degree of usage load-unbalance. Shen *et al.* [2] used the LDA model and K-means clustering algorithm to find the functional areas of BSSs. BSS as a public transportation mode, accurate user demand is crucial to better serve the user.

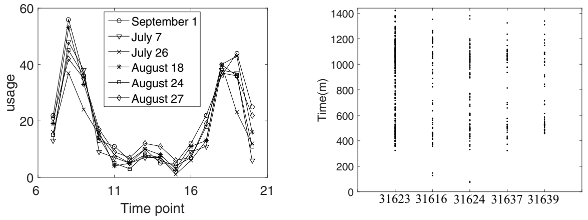
Bicycle usage prediction: Kaltenbrunner *et al.* [3] proposed a statistical model to predict the number of available bicycles at any station. Yoon *et al.* [4] built a spatio-temporal prediction system based on a modified ARIMA model, extended to include seasonal trends and regression-based spatial correlations. Singhvi *et al.* [5] used taxi data, weather and space factors as covariates to predict the demand for public bicycles and further analyze the impact of precipitation and week on public bicycle demand.

Demand estimation: There are also a large number of researches on demand estimation issues for other transport modes in the field of transportation. Zhang *et al.* [6] designed and evaluated a taxicab passenger model Dmodel based on big data, which employs roving taxicabs as real-time mobile sensors to infer passenger demand by customized online training with both historical and real-time data. Kang *et al.* [7] proposed a prediction model of combining the conditional transition distribution and the neighboring information on taxi passenger demands. Akiyama *et al.* [8] conducted a long-term estimation of traffic demand on urban expressway using neural networks.

III. Motivation

A. User Demand in Different Station Status

User demand can be divided into two types, the demand of users to rent bicycles and those to return bicycles. Without loss of generality, only the demand of users to rent bicycles are taken as the example to illustrate our demand estimation model. When a user comes to a BSS station to rent a bicycle, there are two cases, one is the station has available bicycles and the user successfully rents a bicycle, while another case is the station does not have available bicycle and the user fails to rent a bicycle.In the second case, although it is happened in real world, we could not get this information from the system data, hence we need to infer this situation and then estimate the actual user demand.



(a) Usage of different days

(b) Usage of different stations on a same day

Figure 1. Data feature analysis

Previous researches on BSS did not consider the situations that users are not served. In order to improve service quality for users, it is necessary to estimate the user demand considering the situations that users are not served.

B. Temporal Correlation of User Demand

For a specific station, the amount of daily bicycle usage has a certain pattern, has a long-term stability, which is reflected in a temporal correlation. Fig.1(a) shows the usage of station whose terminal ID is 31623 for a total of 14 hours on Sept. 1^{st}, 2015, compared with the large amount of records trips generated in the past two months to find that there are five days whose usage per hour are very close to that of Sept. 1^{st}. As can be seen from the Fig.1(a), a total of six days including Sept. 1^{st}, these usages are very close to each other, and the trends of usage for the whole day are also very similar.

C. Spatial Correlation of User Demand

In BSS, adjacent stations in a region have similarities, and their usage behaviors will be related to each other, although their usage amount will be different (limited by the stations' capacity). Therefore, in a region, the daily usage of a station has a short-term volatility, which is reflected in a spatial correlation.

Fig.1(b) shows the usage behavior distribution of station whose terminal ID is 31623 on the day of Sept. 1^{st}, 2015, with another four stations closed to it. The vertical axis in Fig.1(b) is the time of a day, and the unit is minute. If there are users successfully renting bicycles in a minute, there is a corresponding point in Fig.1(b). There is no corresponding point in Fig.1(b) means that there is no user renting bicycle in this minute. As shown in Fig.1(b), compared with station 31623, its four adjacent stations, although their usage behaviors are less and the points are sparser, the usage behavior distributions have a clear correlation. The peak time period and the bottom time period of five stations during the day are very close.

IV. Design Concept

A. Model Design

In order to estimate the actual user demand, we need to infer the situations that users are not served, which is not recorded by the system. We want to be able to infer this part of the non-existent data through the existing available data, and the key of that is to find the correlation between the existing available data and the non-existent data.

The BSSs have the peculiar characteristics. The first characteristic is long-term stability. Peoples lives are regular. People go to work/school from home in the morning and go home from work/school in the evening. So, for a specific station, the daily usage has a certain pattern. In a large amount of historical usage data, the daily usage of a station has a long-term stability, which is reflected in a temporal correlation. In the previous section, we make an analysis of a large amount of historical usage data for a specific station, and prove that there has a strong temporal correlation. Based on the above analysis of the first characteristic, when a station status is no available bicycles at some time period of current day, we can find similar days with current day in the historical days, and infer the usage trend of current day through the usage trend of these similar days.

The second characteristic is short-term volatility. Usually, the stations of BSS are located at fixed positions in the city. Stations within a region are affected by the same short-term external factors, e.g., sudden rain and sudden traffic accident. So, for these stations within a region, the usage has a same short-term volatility, which is reflected in a spatial correlation. We also prove that there has a spatial correlation between the usage of a station and its adjacent stations in the previous section. Based on the above analysis of the second characteristic, when current station status is no available bicycles at some time period of current day, we can look for adjacent stations, and infer the usage trend of current station through the usage trend of these adjacent stations.

B. Framework

Fig.2 shows the framework of the demand estimation model. We need to judge the station status. If the station status is no available bicycles, it needs to use demand estimation model to obtain the actual user demand. There are two operational branches for demand estimation based on temporal and spatial correlation, and finally, integrate these two branches to obtain an actual user demand.

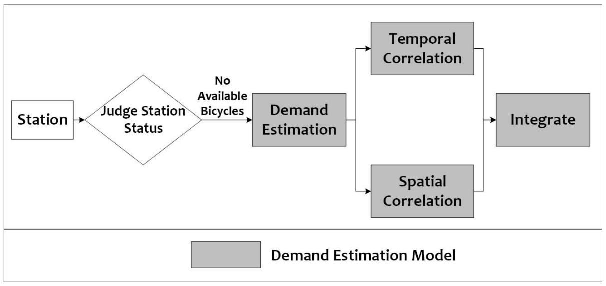


Figure 2. Framework of the demand estimation model for BSS

Table I
Notations

Variables	Definitions
S	The reference station
S_i	The i^{th} station
d	The reference day
d_j	The j^{th} historical day
t	Current time
T	Length of Time period
TR	Trip record generated by the system
$U_{S_i,d_j,t}$	Usage amount of the station S_i in the time t of the day d_j
$D_{S_i,d_j,t}$	Demand of the station S_i in the time t of the day d_j
$CV_{S_i,d_j,t}$	Usage change value of the station S_i in the time t the day d_j
$CR_{S_i,d_j,t}$	Rate of usage change of the station S_i in the time t the day d_j
$F_{S_i,d_j,t}$	Distribution of the usage amount of the station S_i in the time t of the day d_j
$corr_{d_j}$	Pearson correlation coefficient of the historical day d_j
$corr_{S_i}$	Pearson correlation coefficient of the adjacent station S_i
w_{d_j}	Weight value of the historical day d_j
w_{S_i}	Weight value of the adjacent station S_i

V. Estimation Model

A. Notations

This subsection defines the notations (Table I).

B. Models

1) Demand Estimation Based on Temporal Correlation: Considering three months that July, August, and September, whose external factors, temperature, humidity and so on, keep almost constant, we conduct a Pearson correlation analysis of the usage on the day of Sept. 1^{st} with the historical usage data of past two months (July and August) to find out five days with the highest correlation, and the comparison of their usages are shown in Fig.1(a).

From Fig.1(a), we can see that among these five days with the highest correlation, two days are in July and three days in August. Although not all are very close to the day of Sept. 1^{st}, their usage data per hour are very close to those of Sept. 1^{st}. Their trends in a day as a whole are very close, increasing or decreasing in a consistent way.

Based on the above analysis, there are high degrees of temporal correlation in daily usages of a station. There will be many days with a high degree of relevance can be found out from the large amount of historical data. In these days, the higher the relevance to the reference day is, the higher weight value we assign to it in the station user demand estimation. After we get the historical days' weight values, we also need to analyze the specific trend of usage of each day to comprehensively infer the number of users who are not recorded by system for the specific station. We analyze the specific trends of each day based on the change values between the usage of each hour and that of its precedent hour for each day (as shown in Eq.1).

$$CV_{S_i,d_j,t} = U_{S_i,d_j,t} - U_{S_i,d_j,t-1} \tag{1}$$

Based on the above analysis, we obtain the weight values by the Pearson correlation between the usages of historical days and that of the reference day (as shown in Eq.2). Combined with the usage change values of the historical days at the same hour, a weighted average can be obtained as the estimated usage change values at the same hour of the reference day (as shown in Eq.3). Finally, the user demand at the same hour is estimated by Eq.4 by combining the actual usage of the station at the precedent hour and the estimated usage change values at the moment.

$$w_{d_j} = \frac{corr_{d_j}}{\sum corr_{d_j}} \tag{2}$$

$$CV_{S_i,d,t} \approx \sum w_{d_j} CV_{S_i,d_j,t} \tag{3}$$

$$D_{S_i,d,t} \approx D_{S_i,d,t-1} + \sum w_{d_j} CV_{S_i,d_j,t} \tag{4}$$

2) Demand Estimation Based on Spatial Correlation: As shown in Fig.1(b), based on the usage behaviors of these stations on Sept. 1^{st}, 2015, there are five stations with high correlation with this reference station whose information is shown in Table II. Because the different capacity of each station will lead to the different number of usages, we analyze the detailed station usage on minute-scale.

In Fig.1(b), station 31623 is the station whose user traffic is the largest, and the other four stations are close to it. The time of renting bicycle started is almost the same in the morning, while the time of renting bicycle is in the adjacent time period at night. Additionally, the dense time period and the sparse time period are also similar during the day. The different durations of time periods of dense bicycle renting are caused by the different capacities of the stations themselves.

Based on the above analysis, the adjacent stations in a region have similarities, and their usage behavior distributions have a great spatial correlation. Among these adjacent stations around a station, the bigger the similarity with the reference station is, the higher weight value we assign to it in the user demand estimation. After obtaining the weight value of each adjacent station, we also need to analyze the specific trend of usage of each adjacent station. In the previous section, when examining the trend of usage for a specific station, we use the usage change value as shown in Eq.1. However, in this section, it is not appropriate to adopt the usage change value when examining the specific trend of usage due to the fact that different stations have different capacities. Instead, we adopt the rate of usage change of each station between each hour and its precedent hour (as shown in Eq.5).

$$CR_{S_i,d_j,t} = \frac{U_{S_i,d_j,t} - U_{S_i,d_j,t-1}}{U_{S_i,d_j,t-1}} \tag{5}$$

Based on the above analysis, we obtain the weight values by the similarities between the adjacent stations and the reference station (by Eq.6). Combined with the rates of usage change of the adjacent stations at the moment, a weighted average can be computed as the estimated rate of usage change of the reference station at the moment (as shown in Eq.7). As a result, the user demand at the moment is estimated by Eq.8 by combining the actual usage of the station at the precedent hour and the estimated rate of usage change at the moment.

$$w_{S_i} = \frac{corr_{S_i}}{\sum corr_{S_i}} \tag{6}$$

$$CR_{S,d_j,t} \approx \sum w_{S_i} CR_{S_i,d_j,t} \tag{7}$$

$$D_{S,d_j,t} \approx D_{S,d_j,t-1} + D_{S,d_j,t-1} \sum w_{S_i} CR_{S_i,d_j,t} \tag{8}$$

3) User Demand Integrated: In the last two subsections, we analyze the temporal correlation of the station with itself and the spatial correlation with its adjacent stations, and estimate the user demand for a station based on temporal correlation and spatial correlation, respectively.

Different types of stations will result in different proportions of temporal and spatial correlation in demand estimation. Some stations have more stable traffic patterns, which are located in residential area, near the school, etc. These stations have stronger long-term stability than short-term volatility. Therefore, the proportions of temporal correlation is larger than that of spatial correlation in conducting user demand estimation for these stations. As a contrast, stations near tourist attractions are the opposite, which have stronger short-term volatility than long-term

Table II
STATION INFORMATION

Terminal ID	Station Name	Longitude	Latitude
31623	Columbus Circle / Union Station	-77.00493	38.89696
31616	3rd & H St NE	-77.001949	38.900412
31624	North Capitol St & F St NW	-77.009888	38.897446
31637	North Capitol St & G Pl NE	-77.008911	38.899703
31639	2nd & G St NE	-77.003666	38.89967

stability. For simplicity, we take the same proportion of temporal correlation and spatial correlation to conduct experiments.

VI. EXPERIMENTS

This study presents a three-step model for station demand estimation. For the station status is no available bicycles, we use the demand estimation model proposed to estimate the actual user demand of the station, which needs to be verified in terms of the feasibility and effectiveness.

A. Settings

1) Datasets: We conduct experiments on the user trip dataset of Capital Bike-share system, which is located in Washington D.C.. There are also two other datasets taken as inputs, one is the station location data, which contains the geometric information of each station in Capital Bike-share system, e.g., latitudes and longitudes, and the other is station status data, which contains real-time station status of each stations in Capital Bike-share system, e.g., no available bicycles or no available docks. The detailed information of the datasets is shown in Table III.

Table III
DETAILS OF THE DATASETS

Data Source	Washington D.C.
Time Span	July 1^{st} - Sept. 30^{th}, 2015
Valid Days	92
Number of Stations	355
Geometric Information	Latitudes and Longitudes
Number of User Trip Records	1056269
Number of Station Status Records	112844

2) Metrics: The metric we adopt to evaluate the results is as follow:

***RE* (Relative Error)**: $\delta = \frac{|D_{S_i,d_j,t} - \overline{D_{S_i,d_j,t}}|}{D_{S_i,d_j,t}} \times 100\%$

The variable $D_{S_i,d_j,t}$ represents the user demand estimated by our demand estimation model, and the variable $\overline{D_{S_i,d_j,t}}$ represents the actual user demand.

B. Data Filtering

In this study, we mainly have three parts of the data filtering.

Firstly, for some remote stations, the user demand is relative small and the station will not appear in the status of no available bicycles. These stations do not need to use the demand estimation model proposed, so we do not carry out the experiments of the demand estimation model.

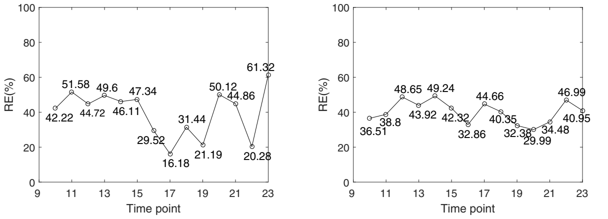
Secondly, the usage we used to estimate the user demand should be the user trip data when the station status is having available bicycles. Once the station status is no available bicycles, the usage is not the actual user demand, and we need to filter this part of station data. The station status data is mainly used to deal with this part of data filtering.

Thirdly, after we use the above two kinds of data filtering and the demand estimation model, we adopt relative error metric *RE* to measure the advantages and disadvantages of the estimation model used. When the actual user demand is a very small value, the estimated value obtained by our demand estimation model is very close to real value, whose difference is very small. However, it still will lead to a large *RE*, which will affect the correct evaluation of the model. Therefore,in the final evaluation, we filter out the station moments whose actual user demand is smaller than 10, which can reflect the advantages and disadvantages of the model more accurately.

C. Experimental Results

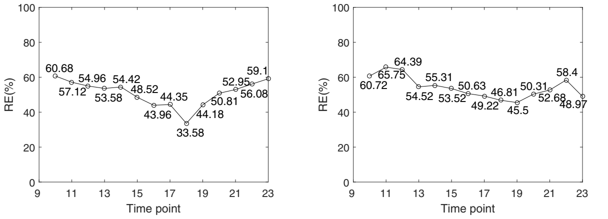
We select a station whose terminal ID is 31623 (specific information can be found in Table II) to conduct experiment for a period of one-month (Sept. 1^{st}, 2015-Sept. 30^{th}, 2015). The selected experimental time period is 9:00 am-23:00 pm. For each hour, we estimate the station user demand, and compute the average *RE* of a month. When we estimate user demand based on temporal correlation, the historical usage data we used is ranging from July 1^{st} to the day before the reference day.

Fig.3(a) shows the experimental results of user demand estimating for station 31623 based on temporal correlation. As shown in Fig.3(a),from day to night,the RE of the experiment decreases with the increase of human traffic and the minimum RE of the experiment is only 16.18%.



(a) Based on temporal correlation (b) Based on temporal and spatial correlation

Figure 3. Result of single station



(a) Based on temporal correlation (b) Based on temporal and spatial correlation

Figure 4. Result of entire system

Fig.3(b) shows the experimental results for station 31623 based on both temporal and spatial correlation. Compared with Fig.3(a), the result in Fig.3(b) is much more stable and almost under 50%. In the evening peak time, the *RE* values of the experiment are distributed among 30%-35%.

After examining the experimental results of the demand estimation model for a single station, we further use the demand estimation model to conduct experiments for the entire BSS. There are 96 stations left by filtering out the stations that are remote. Other experimental conditions are the same as above, and for each hour, we obtain the average *RE* of a month for each station as above, and then get the station-average across the 96 stations.

Fig.4(a) shows the average *RE* of the experiment for the entire BSS based on temporal correlation. As shown in Fig.4(a), the *RE* of the experiment for the entire BSS is relatively stable compared with a single station. It can be seen more clearly that the RE of the experiment decreases with the increase of human traffic,from day to night. At the evening peak time, the *RE* of the experiment is reduced to 33.58%.

Fig.4(b) shows the average *RE* of the experimental for the entire BSS based on both temporal and spatial correlation. Considering the spatial correlation, the *RE* of the experiment is much more stable and less affected by changes of user traffic, where the minimum *RE* of the experiment is reduced to 45.5%.

VII. Conclusion

In this paper, we put forward a three-step demand estimation model to estimate the actual user demand taking into account the users who are not served. We analyze the two characteristics of the daily station usage, long-term stability and short-term volatility. Based on these two characteristics, our demand estimation model infers the situations that users are not served from both the temporal and spatial perspectives, and then obtains the user demand integrated. We evaluate our demand estimation model on Washington D.C. Bike-share system. Moreover, our demand estimation model is evaluated leveraged upon the comprehensive information of three datasets, user trip data, station status data, and station location data. In the future, we would like to consider how to better integrate the temporal correlation and spatial correlation, and use more datasets to evaluate our demand estimation model.

Acknowledgment

This work was supported by the National Science Foundation of China (NSFC) under Grants No.61672198, 61772551 and Zhejiang Provincial Natural Science Foundation under Grant No.LY14F020047, LS17G03001, and DGIST Global Research Laboratory Program (2013K1A1A2A02078326), Funded by the Korean MSIP (No.B0101-15-0557, Resilient Cyber-Physical Systems Research).

References

[1] X. Yao, X. Shen, L. Wang, and T. He. Hybrid Bicycle Allocation for Usage Load Balancing and Lifetime Optimization in Bike-Sharing Systems. In *IEEE International Conference on Mobile Data Management*, pages 112–117, 2017.

[2] X. Shen and L. Wang. Rental points clustering and function identification of public bicycle system. *Computer Engineering*, 2017(in Chinese).

[3] A. Kaltenbrunner. Urban cycles and mobility patterns: Exploring and predicting trends in a bicycle-based public transport system. *Pervasive and Mobile Computing*, 6(4):455–466, 2010.

[4] J. W. Yoon, F. Pinelli, and F. Calabrese. Cityride: A Predictive Bike Sharing Journey Advisor. In *Proc. of the 13th IEEE ICMDM*, pages 306–311, 2012.

[5] D. Singhvi, S. Singhvi, and P. I. Frazier. Predicting Bike Usage for New York Citys Bike Sharing System. In *AAAI 2015 Workshops on Computational Sustainability*, 2015.

[6] D. Zhang, T. He, S. Lin, S. Munir, and J. A. Stankovic. Taxi-Passenger-Demand Modeling Based on Big Data from a Roving Sensor Network. *IEEE Trans. Big Data*, 3(3):362–374, 2017.

[7] S. H. Kang, H. B. Bae, R. M. Kil, and H. Y. Youn. Predicting Taxi Passenger Demands Based on the Temporal and Spatial Information. In *ICONIP*, number 5, pages 267–274, 2017.

[8] T. Akiyama and H. Inokuchi. Long term estimation of traffic demand on urban expressway by neural networks. In *International Symposium on Soft Computing & Intelligent Systems*, pages 185–189, 2014.