Multi-source Data Analysis for Bike Sharing Systems

Nguyen Thi Hoai Thu, Le Trung Thanh, Chu Thi Phuong Dung, Nguyen Linh-Trung, Ha Vu Le
University of Engineering and Technology, Vietnam National University, E3, 144 Xuan Thuy, Cau Giay, Hanoi, Vietnam

***Abstract*—Bike sharing systems (BSSs) have become common in many cities worldwide, providing a new transportation mode for residents' commutes. However, the management of these systems gives rise to many problems. As the bike pick-up demands at different places are unbalanced at times, the systems have to be rebalanced frequently. Rebalancing the bike availability effectively, however, is very challenging as it demands accurate prediction for inventory target level determination. In this work, we propose two types of regression models using multi-source data to predict the hourly bike pick-up demand at cluster level: Similarity Weighted K-Nearest-Neighbor (SWK) based regression and Artificial Neural Network (ANN). SWK-based regression models learn the weights of several meteorological factors and/or taxi usage and use the correlation between consecutive time slots to predict the bike pick-up demand. The ANN is trained by using historical trip records of BSS, meteorological data, and taxi trip records. Our proposed methods are tested with real data from a New York City BSS: Citi Bike NYC. Performance comparison between SWK-based and ANN-based methods is provided. Experimental results indicate the high accuracy of ANN-based prediction for bike pick-up demand using multi-source data.**

***Index Terms*—bike sharing system, regression model.**

I. Introduction

Bike sharing is a service which provides available bikes as a shared use for individuals on a short-term basis, either free or at a reasonable price. A bike sharing system (BSS) allows users to rent a bike from one station and return it at any other station within the system. BSSs have been deployed in various cities around the world since the second half of 20th century and become more popular in recent years [1], [2]. These systems provide access to bicycles for short-distance trips as an alternative to private vehicles or motorized public transport such as bus or subway in an urban area. In addition, they help reduce the traffic congestion, air pollution and noise. Moreover, they have been considered as a way to solve the "last mile" problem [3]. Finally, they help bridge the gap between existing transportation modes such as subways and bus systems [4] and connect users to public transit networks.

Beside the benefits mentioned above, BSSs face many problems, one of which is the availability imbalance. Due to the fact that movements of customers are highly dynamic [5], the bike usage is non-stationary, changing markedly with time and location [6]. Therefore, some stations may be short of available bikes for rent while some are full and do not have enough docks for returned bikes. A general approach to solve this problem is that the system should monitor and redistribute bikes between stations frequently using trucks or bike-trailers. Real-time monitoring and redistribution, however, take too much time to execute, especially during rush hours, and therefore become unrealistic. It is desirable to make accurate prediction of the pick-up/drop-off demand and inventory level at each station. To address this issue, a solution must consist of two main stages: (i) bike pick-up/drop-off demand prediction, and (ii) rebalancing route optimization. Our work presented in this paper focuses only on the bike pick-up demand prediction. We can predict the number of bikes that will be picked up at each station in the near future, based on the historical trip records as well as meteorological data. However, apart from meteorological data, the bike traffic can be influenced by many other factors, such as time of day, day of week, events, demographic factors, and the correlation between stations. They make the problem become more challenging.

There have been a number of studies on bike demand prediction. Some methods are based on historical demand [7] or stochastic process [8], [9]. As we mentioned above, the bike pick-up demand is influenced by many factors. Thus, exploitation of multiple sources of data affecting BSSs is highly beneficial to improving bike demand prediction accuracy. It can be considered as a promising approach and attracts many studies. For instance, Liu *et al.* [4] proposed a Meteorological Similarity Weighted K-Nearest Neighbor (MSWK) model that combines meteorological data and the past bike demand to predict the hourly bike demand at the station level. Li *et al.* [6] first chose to cluster bike stations into groups using a bipartite clustering algorithm, and then used meteorological data with a multi-similarity-based inference model to predict the bike demand at the cluster level. Singhvi [10] applied a log-log regression model using taxi usage and spatial variables considered as covariates to predict bike demand at the neighborhood level.

Inspired by the above multi-source data approach, in this paper we propose two regression models for predicting hourly bike pick-up demand at the cluster level instead of the station level, namely the Similarity Weighted K-Nearest-Neighbor (SWK) regression model with the correlation among consecutive time slots (SWKcor), and the Artificial Neural Network (ANN) based model. Data utilized by these models in our work include historical trip records of the BSS, meteorological data, and taxi trip records. To our best knowledge, there have not been any studies combining these data sources in analyzing the BSS demand problem.

This paper is organized as follows. Section II provides background of the study, including some preliminaries, a

978-1-5386-2896-6/17/$31.00 ©2017 IEEE

clustering algorithm, the original MSWK regression model, and a brief introduction to ANN. Section III introduces our proposed method to improve the accuracy of bike pick-up demand prediction. Section IV shows the experimental results. Section V concludes the work with some notes about future directions.

II. Background

A. Preliminaries

In this section, we introduce some preliminaries, which will be used throughout this work. Let $\mathcal{G} = \{\mathcal{S}, \mathcal{E}\}$ be a directed graph representing a system (network) of bike stations. Each bike station $s \in S$ is called a node or a vertex, while the edge $e_{i,j} \in \mathcal{E}$ represents a directed path from node s_i to node s_j. Each node or edge has several attributes.

The BSS is constructed by tracking a set of trip records. A trip record $\mathbf{tr} = (s_0, s_d, \tau_0, \tau_d)$ is a bike usage record of a trip from an origin station s_0 to a destination station s_d, and τ_0 and τ_d are the pick-up time and the drop-off time, respectively. Only trips with duration $(\tau_d - \tau_0)$ larger than one minute are recorded. Each day is separated into 24 time slots of one hour duration, $t \in \{0, 1, \ldots, 23\}$.

Definition 1 (Station bike demand): Let $s_i(D^t)$ denote the pick-up demand of station s_i during time slot t of day D.

Definition 2 (Cluster bike demand): Let $c_i(D^t)$ denote the total pick-up demand of stations in cluster c_i during time slot t of day D.

B. Bipartite clustering

A number of recent studies have used clustering algorithms for bike pick-up demand prediction. There are two reasons why we need to predict bike pick-up demand at the cluster level instead of the station level. First, because the bike pickup demand is affected by multiple factors, such as the weather and the correlation between stations, traffic of a single station seems too chaotic to predict [6]. Thus, clustering makes the prediction easier. Second, it is not necessary to predict the pick-up demand of each individual station because people often pick the bikes up at a random station near their place. Therefore, for bike reallocation, knowing the pick-up demand of each cluster is sufficient. In fact, most of existing BSSs have a real-time status map which shows the number of available bikes and docks. If a station is empty or full of bikes, it is convenient for a user to find another nearby station. Besides, if there is some event happening which influences bike usage, it is common that the bike traffic of an area encompassing several stations will be affected.

We employ a clustering algorithm, proposed by Li *et al.* in [6], to group individual stations into clusters based on their geographical locations and transition patterns. It generates a set of matrices, each is called a t-matrix which describes the transition pattern of a station. Each entry of a t-matrix, $(\mathbf{A}_i)_{l,j}$, is the probability that a bike will be dropped off at cluster $C_{1,j}$ given that it was picked up from station s_i. The pseudo-code of this algorithm is described in Alg. 1.

Algorithm 1 Bipartite clustering algorithm

Input: Stations $\{s_i\}_{i=1}^n$, historical trips $\{\mathbf{tr}_i\}_{i=1}^n$; iteration threshold K, parameters $K_2 < K_1$;

Output: K_1 clusters: $C_{1,1}, C_{1,2}, ..., C_{1,K_1}$

1: Cluster $\{s_i\}_{i=1}^n$ into K_1 clusters: $C_{1,1}, C_{1,2}, ..., C_{1,K_1}$ by K-mean based on locations
2: ***Initialization***:
3: k $\leftarrow 0$
4: ***Recursion***:
5: While $k < K$
6: **For** $i = 1$ to n **do**
7: Generate matrix $\{\mathbf{A}_i\}_{i=1}^n$ of station $\{s_i\}_{i=1}^n$.
8: Matrix $\mathbf{A}_i$ is the probability that a bike will be checked into cluster $\{C_j\}_{j=1}^k$.
9: $$P(\mathbf{A}_i \rightarrow C_j) = \frac{Trips(\mathbf{A}_i \rightarrow C_j)}{Trips(\mathbf{A}_i \rightarrow \{C_l\}_{l=1}^k)}$$
10: Cluster $\{s_i\}_{i=1}^n$ into K_2 clusters: $C_{2,1}, C_{2,2}, ..., C_{2,K_2}$ by K-mean based on $\{\mathbf{A}_i\}_{i=1}^n$
11: **For** $j = 1$ to K_2 **do**
12: Cluster stations in $C_{2,j}$ into $\left[\frac{N_j K_2}{n}\right]$ clusters with N_j are numbers stations of $C_{2,j}$
13: Obtain K_1 updated clusters
14: **If** $C_{1,1}, C_{1,2}, ..., C_{1,K_1}$ do not change **Then**
15: Break;
16: $k \leftarrow k + 1$;
17: Return K_1 clusters $C_{1,1}, C_{1,2}, ..., C_{1,K_1}$

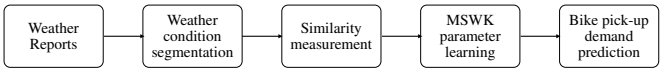


Fig. 1. Structure of original MSWK model.

C. Meteorological SWK regression model (MSWK)

We inherit a recently proposed MSWK regression model in [4] which predicts the station level pick-up demand and introduces some new methods making use of clustering and correlation between time slots to improve its performance.

A regressor is built to predict $s_i(D_q^t)$, the station level bike pick-up demand during time slot t of day D_q based on the demand during same time slot t of previous days and a meteorological multi-similarity function. The structure of this regression model is shown in Fig.1.

1) Weather condition segmentation: The weather report $R_{D_p^t}$ contains the weather condition $W_{D_p^t}$ (sunny/raining/etc.), temperature $F_{D_p^t}$, wind speed $S_{D_p^t}$, and visibility $V_{D_p^t}$ of time slot t on day D_p. In this step, weather conditions are divided into four groups according to their suitability for outdoor cycling: $G_1 = \{\text{sunny}, \text{cloudy}\}$; $G_2 = \{\text{fog, mist, haze}\}$; $G_3 = \{\text{snow, rain, light snow, light rain}\}$; $G_4 = \{\text{heavy snow, heavy rain}\}$.

Fig. 2 shows the effect of meteorological factors on bike usage in Citi Bike NYC. The bike pick-up demand is sensitive to temperature: the demand tends to decrease when the tem-

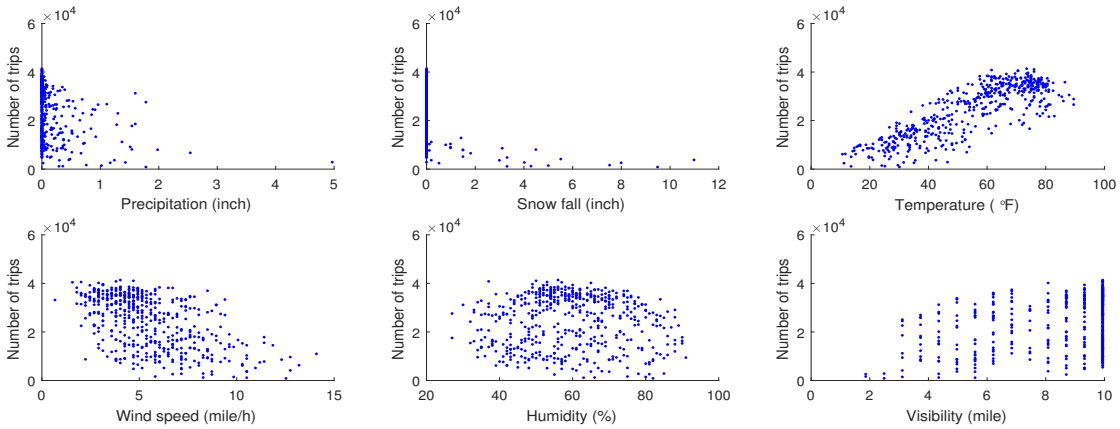


Fig. 2. The effect of meteorological factors on bike usage in Citi Bike NYC.

perature is too low (under 50^oF) or too high (over 85^oF). The wind speed and visibility show strong correlation in affecting the bike demand: the demand decreases with higher wind speed/lower visibility, and vice versa. The humidity seems to have no significant impact on bike usage.

2) Similarity measurement: The meteorological similarity between two different days, D_p^t and D_q^t, are calculated by using a linear combination of three units: weather condition similarity, temperature similarity, and wind-speed-visibility similarity. Each unit is associated with a weighted coefficient that is learned to improve the accuracy.

Weather similarity At first, we manually segment the weather conditions into different levels according to their effects on bike pick-up demand. Let (G_1, G_2, G_3, G_4) = (0.25, 0.5, 0.75, 1), the weather condition similarity is defined as:

$$\lambda_1(W_{D_p^t}, W_{D_q^t}) = \frac{1}{2\pi\sigma_1} \exp\left\{-\frac{\left(W_{D_p^t} - W_{D_q^t}\right)^2}{\sigma_1^2}\right\} \quad (1)$$

Temperature similarity:

$$\lambda_2(F_{D_p^t}, F_{D_q^t}) = \frac{1}{2\pi\sigma_2} \exp\left\{-\frac{\left(F_{D_p^t} - F_{D_q^t}\right)^2}{\sigma_2^2}\right\} \quad (2)$$

Wind-speed-visibility similarity:

$$\lambda_3 = \frac{1}{2\pi\sigma} \exp\left\{-\left(\frac{(S_{D_p^t} - S_{D_q^t})^2}{\sigma_3^2} + \frac{(V_{D_p^t} - VS_{D_q^t})^2}{\sigma_4^2}\right)\right\} \quad (3)$$

Similarity function Before calculating similarity measures, we normalize the temperature, wind speed, and visibility values into range [0, 1] and set $\sigma_k = 1$ $(k = 1, 2, 3, 4)$ to simplify the equations (1)-(3). The similarity function is defined by the following linear combination:

$$M(D_p^t, D_q^t; a) = \sum_{i=1}^{3} a_i \lambda_i \quad (4)$$

3) Bike pick-up demand prediction: Given predefined values of H and a, we select the top H days $\{D_1^t, D_2^t, ..., D_H^t\}$ having the highest similarity scores, calculated by the similarity function, to our target day D_q^t. Then $s_i(D_q^t)$ is predicted by a similarity-weighted KNN:

$$s_i(D_q^t; a) = \frac{\sum_{p=1}^{H} M(D_p^t, D_q^t; a) * s_i(D_p^t)}{\sum_{p=1}^{H} M(D_p^t, D_q^t; a)} \quad (5)$$

4) MSWK parameter learning: The weight of a given meteorological similarity a in equation (5) is obtained via a training process such that the minimum prediction absolute error of predicted value $\hat{s_i}(D_q^t; a)$ against the ground truth $s_i(D_q^t)$ is reached by brute force searching:

$$\underset{a}{\text{minimize}} \frac{1}{N} \sum_{i-1}^{N} |\hat{s_i}(D_q^t; a) - s_i(D_q^t; a)| \quad (6)$$

D. Artificial neural networks

Artificial neural networks (ANNs) are a bio-inspired learning models aimed at simulating the behavior of biological neural networks [11]. ANNs are indeed connectionist systems which are composed of a number of interconnected artificial neurons. They are typically organized into layers. Objects are presented to an ANN through the input layer, which communicates with one or more hidden layers via weighted connections. The hidden layers connect to the output layer in the same manner. As an useful analytical tool, ANN is widely used to solve prediction problems in various research fields.

III. Methods

A. SWK based models

We improve the MSWK model proposed in [4] by i) using the bipartite clustering algorithm in [6], then predicting the pick-up demand based on the correlation between consecutive time slots, ii) using SWKcor with the meteorological similarity measure (MSWKcor), and iii) using SWKcor with a taxi usage similarity measure (TSWKcor). Our proposed modification is motivated by the observation that, when the system is expanded, the demand increases significantly, or if there is an event happening in the target day or the previous days,

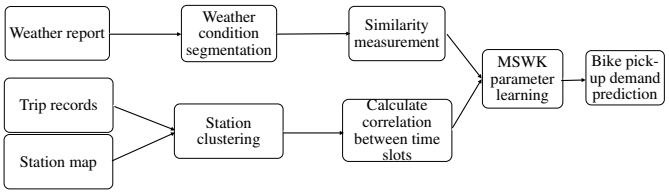


Fig. 3. Structure of the MSWKcor regression model.

the demand at the same time slot of the target day is highly different from past days. Because the bike pick-up demand of one cluster is supposedly a time series, the demand at time slot t may relate to the demands at previous time slots. We calculate the covariances $\mathrm{Cov}_i(t, t-1)$ and $\mathrm{Cov}_i(t, t-2)$ between pick-up demand of cluster c_i at time slot t with time slots $t-1$ and $t-2$, respectively.

1) SWKcor with meteorological data (MSWKcor): The structure of the modified MSWKcor model is shown in Fig. 3.

After calculating the similarity by (4), given H and a, we can select the top H days $D = \{D_1^t, D_2^t, , D_H^t\}$ having the highest meteorological similarity scores to our target day D_q^t. With each $D_p^t \in D$, we obtain the coefficient $x_i(D_p^t)$ by the following formula:

$$c_i(D_p^t) = \frac{\mathrm{Cov}_i(t, t-1)c_i(D_p^{t-1}) + \mathrm{Cov}_i(t, t-2)c_i(D_p^{t-2})}{x_i(D_p^t)}. \quad (7)$$

With each $x_i(D_p^t)$, we estimate $\hat{c}_i(D_q^t, p)$

$$\hat{c}_i(D_q^t, p) = \frac{\mathrm{Cov}_i(t, t-1)c_i(D_q^{t-1}) + \mathrm{Cov}_i(t, t-2)c_i(D_q^{t-2})}{x_i(D_p^t)}. \quad (8)$$

Then $c_i(D_p^t)$ is predicted by a similarity-weighted KNN by

$$\hat{c}_i(D_q^t; a) = \frac{\sum_{p=1}^{H} M(D_p^t, D_q^t; a)\hat{c}_i(D_q^t, p)}{\sum_{p=1}^{H} M(D_p^t, D_q^t; a)}. \quad (9)$$

2) SWKcor with taxi usage data (TSWKcor): Bike and taxi are two types of transportation which are very popular and convenient in cities. The bike pick-up demand of BSS and the number of taxi trips seem to have no relationship when one is a simple vehicle with low speed and the other is a motor vehicle with high speed. However, they do have something in common, for example, most of the bike trips and taxi trips travel short distances, between 0.5 miles and 10 miles (98% bike trips and 95% taxi trip in NYC). Besides, users (passengers) may switch from taxi to bike when there is a traffic jam, or switch from bike to taxi when it rains or when it is too cold. Therefore, taxi usage data may give hints to predict the bike pick-up demand. In order to realize that idea, we first collect data about taxi trips picked up in the bike cluster, which have the properties suitable for bike trips.

The structure of the TSWKcor model is similar to the MSWKcor model, it uses taxi usage data for similarity measurement instead of meteorological data. In this model, to predict the demand $c_i(D_q^t)$ of cluster c_i at time slot t of day D_q , we measure the similarity of taxi usage between two different days D_q and D_p. Given taxi usage data of two previous time slots $(t-1)$ and $(t-2)$ of day D_q at cluster c_i, denoted $T_i(D_q^{t-1})$ and $T_i(D_q^{t-2})$ respectively, the similarity scores are calculated as follows:

$$\lambda_1 = \frac{1}{2\pi\sigma_1} \exp\left\{-\left(\frac{(T_i(D_p^{t-1}) - T_i(D_q^{t-1})}{\sigma_1}\right)^2\right\}. \quad (10)$$

$$\lambda_2 = \frac{1}{2\pi\sigma_2} \exp\left\{-\left(\frac{(T_i(D_p^{t-2}) - T_i(D_q^{t-2})}{\sigma_2}\right)^2\right\}. \quad (11)$$

The taxi usage similarity are also calculated by using the following linear combination:

$$M(D_p^t, D_q^t; i; a) = \sum_{j=1}^{2} a_j \lambda_j \quad (12)$$

After finding the top H days having the highest taxi usage similarity scores, the TSWK parameter learning step and the bike pick-up demand prediction step of the TSWKcor model will be carried out in a similar manner to the MSWKcor model.

B. ANNs

Four ANNs are used for four datasets: 1) BSS data, 2) BSS data with meteorological data, 3) BSS data with taxi usage data, and 4) BSS data with meteorological and taxi usage data.

1) ANN for BSS data: Because the bike pick-up demand varies from cluster to cluster, time slot to time slot, and day to day; and the demands at previous time slots also influence the target time slot; our proposed ANN for the BSS data uses the following features as inputs: time of day (time slot), day of week, cluster ID, and the pick-up demands of previous time slots.

2) ANN for BSS data with meteorological data: In addition to the features used by the first ANN, the following meteorological factors also used by this ANN: weather condition, temperature, wind speed, and visibility.

3) ANN for BSS data with taxi usage data: In addition to the features used by the first ANN, inputs to this ANN also include taxi usage data of the previous time slots.

4) ANN for BSS data with meteorological and taxi usage data: This ANN's inputs include all features used by the first three ANNs, as listed in Table I.

IV. Experimental results

A. Experimental data

We conduct our experiments on bike trip history data, taxi usage data, and meteorological data from NYC, focusing on rush hours (from 7 am to 22 pm) on workdays, from July 1^{st}, 2013 to June 30^{th}, 2014. The dataset is divided into a training set (July 1^{st} - April 30^{th}) and a testing set (May 1^{st} - June 30^{th}).

1) Citi Bike Data: Citi Bike transactions are generated by NYC BSS which is publicly available from Citi Bike official website. This data set contains the following information: station id, bicycle pick-up station, bicycle pick-up time, bicycle drop-off station and bicycle drop-off time.

TABLE I
FEATURES OF BBS, METEOROLOGICAL, AND TAXI USAGE DATA.

Data	Feature
Time	Time of day (Time slot)
	Day of week
Location	Cluster ID
Bike trip records	Bike demand of previous time slots
Meteorology	Weather condition
	Temperature
	Wind speed
	Visibility
Taxi	Taxi usage of previous time slots

2) Taxi Data: The taxi usage data used in this work are for the Yellow Taxi service and made public on the website of NYC Taxi and Limousine Commission (TLC). The Yellow Taxi is one of the main taxi service of New York which can pick up passengers anywhere in New York City. The trip records include: pick-up and drop-off times, pick-up and drop-off locations, trip distances, itemized fares, rate types, payment types, and driver-reported passenger counts.

3) Hourly Weather Report: The NYC weather report data contains the hourly weather report with the format: time, temperature, humidity, wind speed, visibility, and weather condition.

B. Data preprocessing

1) BSS data: We filter out the bike trips that have the duration trip larger than 1 hour. After dividing a day into 24 time slot with duration of 1 hourly, we classify bike trips according to stations, time slot and day. Next, we calculate the total demand of a station in one time slot.

2) Weather data: The missing meteorological data is estimated by its previous and next weather report.

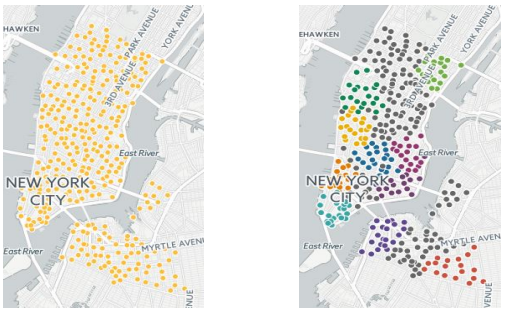
3) Taxi usage data: We collect the taxi trips that are suitable for a bike trip. Because we just collect the bike trips that have trip duration less than one hour and according to Google maps cycling time estimate in New York City the average cycling speed is about 10 miles per hour, we remove the taxi trips which have the distance more than 10 miles. After that, we map the pick-up and drop-off location of a taxi trip to the nearest bike station in the radius of 0.25 mile. Only the taxi trips have both pick-up and drop-off location belong to a bike station are collected. These taxi trips are then grouped by pick-up and drop-off bike station to get a count of taxi trips for each station pair.

C. Metric

The metric we adopt to measure the performance is the error and the error rate (*ER*) of bike demand:

$$error_i(t) = \frac{|\hat{c}_i(t) - c_i(t)|}{c_i(t)} \tag{13}$$

$$ER = \frac{\sum |\hat{c}_i(t) - c_i(t)|}{\sum c_i(t)} \tag{14}$$



(a) NYC's map (b) Clusters

Fig. 4. Clustering result

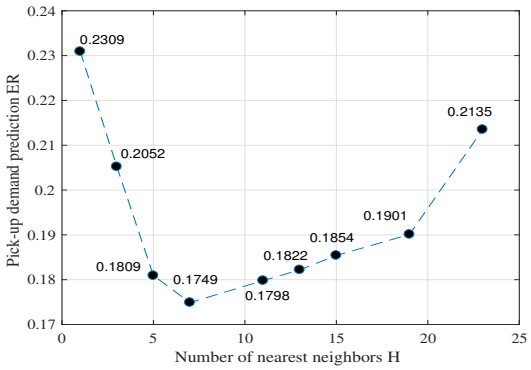


Fig. 5. Performance comparison of SWK based model with different *H*.

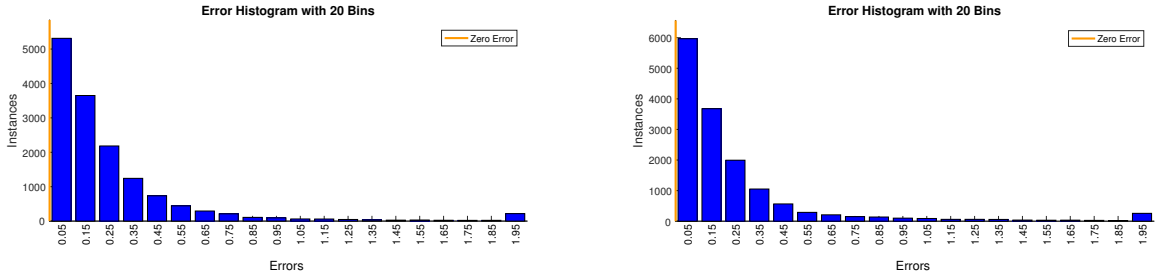
Here, $c_i(t)$ is the ground truth of the pick-up demand of cluster c_i during time slot t. $\hat{c}_i(t)$ is the corresponding prediction value.

D. Results

1) Clustering result: It is clear that the small number of clusters will get higher accuracy than the large number of clusters. However, if the number of clusters is too small, the cluster will be too large. It will be not convenient to the user to pick-up or drop-off bikes at clusters. Thus, the number of clusters needs to be chosen carefully by knowledge and experience. In our experiments, we use the same number of clusters as in [6], all the stations in the NYC Citi Bike system are clustered into 23 clusters (Fig. 4). The stations in each cluster are close to each other. However, there are still some outlier stations which are in the same cluster but quite far from each other.

2) SWK based regression model: First, after finding out the optimal value of the weight of different similarity function a, several values of H are tested to get the best performance. Fig. 5 shows the average *ER* of three SWK based regression models: original MSWK, MSWKcor and TSWKcor with different values of H. We can see that with the value of $H = 7$, the models get the lowest error rate.

3) ANNs: We apply a simple ANN with one hidden layer because it reduces the complexity and the result is good. After



(a) MSWKcor regression model.

(b) ANN with BSS, meteorological and taxi usage data

Fig. 6. Error histogram of 2 regression models.

.

TABLE II
THE RESULTS OF 7 MODELS.

Model	ER	ME
Original MSWK	0.2011	0.4035
MSWKcor	0.1749	0.2995
TSWKcor	0.1836	0.3216
ANN - BSS data	0.1711	0.3298
ANN - BSS and Meteorological data	0.1642	0.2874
ANN - BSS and Taxi data	0.1672	0.3183
ANN - BSS, Meteorological and Taxi data	0.1355	0.2594

that, we change the number of neurons in the hidden layer and then decide the best structure of ANN with the best trade-off between accuracy and complexity for bike demand prediction problem. The configuration that we used for 4 ANN models is a feedforward neural network, with 1 hidden layer which contains 20 neurons. The hyperbolic tangent sigmoid function is used as the active function of the ANNs.

From the error histogram (Fig. 6), we can see that most of the errors are at the low value such as MSWKcor has about 5000 out of 14000 instances at the error smaller than 0.1 and ANN with three datasets has about 6000 instances which has error smaller than 0.1.

Table II shows the results of 7 models: the original MSWK model, the MSWKcor model, the TSWKcor and 4 ANNs with different sets of data. Where *ME* is the mean of errors. It can be clearly seen that in 3 MSWK based models, our proposed models: MSWKcor and TSWKcor have the better performance than the original MSWK. In 4 ANNs with different dataset, the ANN with BSS data sources have the worst performance while the ANN with three data sources has the best performance with the lowest $ER = 0.1355$, lowest mean(error). Thus, these results show that the more data sources that affect the bike demand are used, the higher accuracy we can get. In comparison with SWK based models, for the same dataset, the ANN has better performance than the SWK based models although we just use the feedforward neural network with one hidden layer.

V. CONCLUSION

In this paper, we developed a multi-source data analysis approach for addressing the rebalancing problem of BSSs by using BSS historical trip records, meteorological data and taxi usage data. Specifically, we used a station clustering algorithm base on geographical locations and transition patterns and two methods for predicting pick-up demand: Similarity Weighted KNN (SWK) with the correlation between time slots and Artificial Neural Networks (ANNs). The experimental results show that the ANNs have better performance than the SWK based models. These results also indicate that it is possible to use multi-source data for predicting the bike pick-up demand with high accuracy. For future work, we want to use more data source to understand the insights of the bike sharing system for better prediction, such as the housing and demographic data for station clustering algorithm, bus data, and subway data to predict the bike demand and to expand the system.

REFERENCES

[1] P. DeMaio, "Bike-sharing: History, impacts, models of provision, and future," *Journal of public transportation*, vol. 12, no. 4, p. 3, 2009.
[2] S. Shaheen, S. Guzman, and H. Zhang, "Bikesharing in europe, the americas, and asia: past, present, and future," *Transportation Research Record: Journal of the Transportation Research Board*, no. 2143, pp. 159–167, 2010.
[3] I. Focus, "The last mile and transit ridership," *Institute for Local Government. http://www. cailg. org/node/3216*, 2011.
[4] J. Liu, L. Sun, W. Chen, and H. Xiong, "Rebalancing bike sharing systems: A multi-source data smart optimization," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 2016, pp. 1005–1014.
[5] P. Vogel, T. Greiser, and D. C. Mattfeld, "Understanding bike-sharing systems using data mining: Exploring activity patterns," *Procedia-Social and Behavioral Sciences*, vol. 20, pp. 514–523, 2011.
[6] Y. Li, Y. Zheng, H. Zhang, and L. Chen, "Traffic prediction in a bike-sharing system," in *Proceedings of the 23rd SIGSPATIAL International Conference on Advances in Geographic Information Systems*. ACM, 2015, p. 33.
[7] J. Froehlich, J. Neumann, N. Oliver *et al.*, "Sensing and predicting the pulse of the city through shared bicycling." in *IJCAI*, vol. 9, 2009, pp. 1420–1426.
[8] R. Alvarez-Valdes, J. M. Belenguer, E. Benavent, J. D. Bermudez, F. Muñoz, E. Vercher, and F. Verdejo, "Optimizing the level of service quality of a bike-sharing system," *Omega*, vol. 62, pp. 163–175, 2016.
[9] J. Schuijbroek, R. Hampshire, and W.-J. van Hoeve, "Inventory rebalancing and vehicle routing in bike sharing systems," 2013.
[10] D. Singhvi, S. Singhvi, P. I. Frazier, S. G. Henderson, E. O'Mahony, D. B. Shmoys, and D. B. Woodard, "Predicting bike usage for new york city's bike sharing system." in *AAAI Workshop: Computational Sustainability*, 2015.
[11] D. J. Livingstone, *Artificial Neural Networks: Methods and Applications (Methods in Molecular Biology)*. Humana Press, 2008.