Functional Zone Based Hierarchical Demand Prediction For Bike System Expansion

Junming Liu
Rutgers University, USA
jl1433@rutgers.edu

Leilei Sun
Tsinghua University, China
Dalian University of Technology
leisun@mail.dlut.edu.cn

Qiao Li
Rutgers University, USA
qiao.li1218@rutgers.edu

Jingci Ming
Rutgers University, USA
jingci.ming@rutgers.edu

Yanchi Liu
Rutgers University, USA
yanchi.liu@rutgers.edu

Hui Xiong*
Rutgers University, USA
hxiong@rutgers.edu

ABSTRACT

Bike sharing systems, aiming at providing the missing links in public transportation systems, are becoming popular in urban cities. Many providers of bike sharing systems are ready to expand their bike stations from the existing service area to surrounding regions. A key to success for a bike sharing systems expansion is the bike demand prediction for expansion areas. There are two major challenges in this demand prediction problem: First. the bike transition records are not available for the expansion area and second. station level bike demand have big variances across the urban city. Previous research efforts mainly focus on discovering global features, assuming the station bike demands react equally to the global features, which brings large prediction error when the urban area is large and highly diversified. To address these challenges, in this paper, we develop a hierarchical station bike demand predictor which analyzes bike demands from functional zone level to station level. Specifically, we first divide the studied bike stations into functional zones by a novel Bi-clustering algorithm which is designed to cluster bike stations with similar POI characteristics and close geographical distances together. Then, the hourly bike check-ins and check-outs of functional zones are predicted by integrating three influential factors: distance preference, zone-to-zone preference, and zone characteristics. The station demand is estimated by studying the demand distributions among the stations within the same functional zone. Finally, the extensive experimental results on the NYC Citi Bike system with two expansion stages show the advantages of our approach on station demand and balance prediction for bike sharing system expansions.

CCS CONCEPTS

• **Applied computing** → **Transportation**; **Forecasting**; • **Theory of computation** → *Facility location and clustering*;

KEYWORDS

Bike sharing system; Demand prediction; Clustering

1 INTRODUCTION

Recent years have witnessed worldwide prevalence of public bike sharing systems [4, 16] which provide short-term bike rental services with many bike stations scattering over an urban city. Indeed, these bike sharing systems offer an environment-friendly solution for the first-and-last-mile connection and help bridge the gap between existing transportation modes such as subways and bus systems.

With the success of bike sharing system, most urban cities are planning or have been constructing bike sharing network expansion to attract more customers. For example, NYC has completed two bike sharing network expansions since its foundation in 2013. However, despite the significant benefits from bike sharing network expansion, it is very challenging to decide the expansion strategy which relies on an accurate bike demand prediction for expansion areas. An accurate bike demand prediction can help bike sharing system designers estimate how many new customers will be attracted and how much additional operation cost they need to spend on a larger system. To this end, in this paper, we study the bike demand prediction problem for bike sharing system expansion. There are two major challenges for this problem. First, there are no historical bike transition records available in the expansion areas. This challenge makes it impractical to conduct a direct supervised learning model on the station network after expansion. Second, the station level bike demand has big variances across the city, which can be impacted by multiple factors, such as time, location, surrounding environment (Point of Interest (POI) structure), transportation network, and human mobilities.

A number of recent researchers have studied the bike demand prediction problem. Most studies on bike demand prediction are based on single factor predictors like stochastic process [1, 15] without considering the impact of other influential factors. A promising way to improve bike demand

[*] Corresponding Author

KDD ’17, August 13-17, 2017, Halifax, NS, Canada
© 2017 Association for Computing Machinery.
ACM ISBN 978-1-4503-4887-4/17/08. . . $15.00
https://doi.org/10.1145/3097983.3098180

prediction accuracy is to leverage a variety of data that is directly or indirectly related to the public bike sharing service [7, 9]. However, these methods rely on the availability of historical bike transition records to train the proposed model and are not applicable for bike sharing system expansion. Our previous work [8] proposes a global station level bike demand predictor based on a set of fine grained global features which are used for current station network redesign, however, considering the complexity of urban city structures and station demand variances across the large area of urban city, the global features may not affect bike demand equally in different regions.

Indeed, the emergence of multi-source big data enables a new paradigm for enhancing bike demand predictions. Along this line, we exploit multi-source data related to bike sharing services, such as trip records, station status records, POI dataset, taxi trip records for developing station level bike demand solutions. Specifically, starting from the existing bike sharing system (which we call it principle bike system) with its historical trip records available, we build a hierarchical prediction model to analyze bike demand from zone level to station level. The station in service area is firstly divided into different functional zones through our Bi-Clustering algorithm which considers the POI structures and station locations simultaneously. Then, a zone level bike check-in and check-out predictor is studied based on the bike trip distance preference, zone-to-zone preference, zone characteristics and the historical transitions of the principle bike system. The check-ins and check-outs are then distributed to each station within the functional zone according to the POI structures and their links to other transportation networks. To predict the station bike demand after system expansion, we re-estimate the zone level features by considering the expanded zone-to-zone network and the inner-zone demand distribution for expansion area stations.

Finally, we carry out extensive experiments on a real-world dataset of three different time periods from the NYC Citi Bike system: Stage 1. the principle bike station system consisting of 329 stations from 07/01/2013 to 07/31/2015, Stage 2. the bike system after first expansion including 486 stations from 08/06/2015 to 07/18/2016 and Stage 3. the third stage starts from 07/23/2016 to 11/30/2016 with 617 stations in service after second expansion. Figure 1(a) presents the three stages of station distributions with each dot representing a bike station in New York City. The red dots represent the principle bike station distribution. The orange and the blue dots represent the second and third stages of station expansions respectively. In additional, a few stations represented by different symbols are closed in different stages.

2 PROBLEM FORMULATION

In this section, we first introduce some preliminaries used throughout this paper, and then formally define the problem of station bike demand prediction for bike system expansion.

2.1 Preliminaries

2.1.1 Station bike demand and unbalance. The station bike demand is defined as the pick-up (drop-off) frequency per unit time when the station is available. Station availability means the station is in service and there are bikes available for pick-up (drop-off). Station unavailability is usually due to maintenance, street block, empty dock (for pick-up) and full dock (for drop-off). We do not consider the station demand during its unavailable period.

Definition 1: Station Bike demand. Let $s_i.pf(t)(s_i.df(t))$ and $s_i.pa(t)(s_i.da(t))$ represent the pick-up (drop-off) frequency and the pick-up (drop-off) available time of station i during time slot t. Each time slot t represents a 60 minutes time duration. The bike demand during the day is split into 24 time slots: $t \in \{0, 1, ..., 23\}$. The station pick-up (drop-off) demand $s_i.pd(t)$ $((s_i.dd(t)))$ is defined as follows:

$$s_i.pd(t) = \frac{s_i.pf(t)}{s_i.pa(t)} \quad (s_i.dd(t) = \frac{s_i.df(t)}{s_i.da(t)}) \tag{1}$$

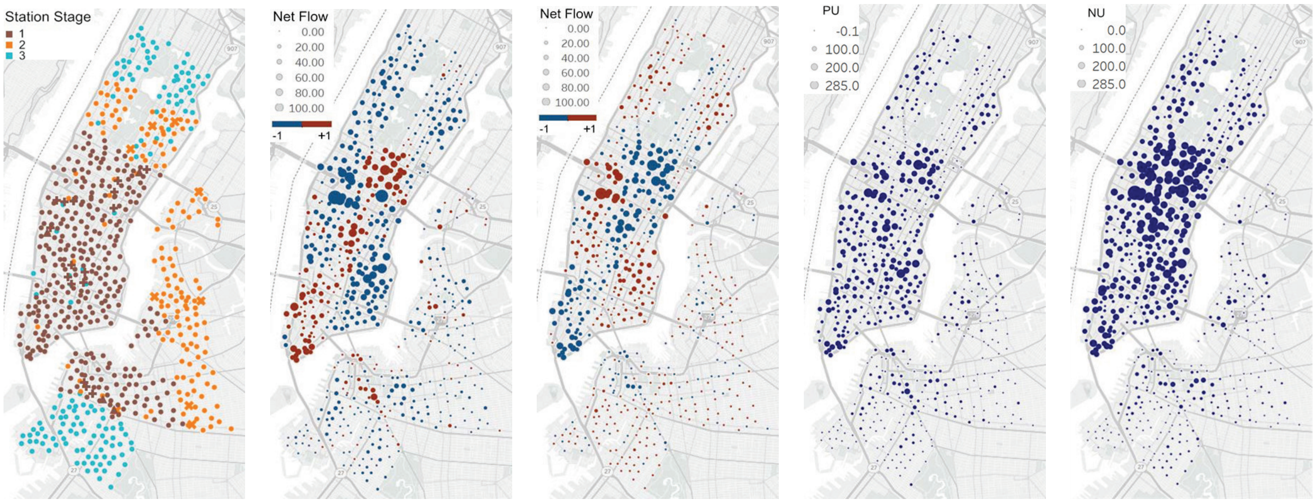
Due to unbalanced bike demand distribution, some bike stations can have continuous large positive bike flows (drop-off demand is much larger than pick-ups) or negative bike flows. The station bike net flow distributions during AM and PM rush hours are presented in Figure 1(b) and Figure 1(c) as an example. In Figure 1(b) and Figure 1(c), each dot represents a bike station in stage 3 with its size representing the absolute value of net flow. The red color represents a positive net flow (drop-off frequency is larger than pick-up frequency) and the blue color represents a negative net flow. As can be seen, the station demand distribution is unbalanced both geographically and temporally. The unbalanced bike flows will cause full station or empty station status and require more operation cost to rebalance the inventory level. We investigate the station balance problem by first introducing the concept of station positive unbalance $s_i.pu$ and negative unbalance $s_i.nu$ based on accumulated station net flow.

Definition 2: Station unbalance. Let $s_i.pd(t)$ and $s_i.dd(t)$ represent the pick-ups and drop-offs at station s_i during time slot t, the station unbalance is defined as the maximum accumulate bike net flow during the day. The positive unbalance and negative unbalance are defined as the contiguous subarray of series $\{s_i.dd(t) - s_i.pd(t)\}$ whose values have the largest positive sum and smallest negative sum. The station positive unbalance $s_i.pu$ and negative unbalance $s_i.nu$ are formally defined as follows:

$$s_i.pu = \max_{t_i,t_j}\{\sum_{t=t_i}^{t_j} s_i.dd(t) - s_i.pd(t), \quad t_i < t_j\} \tag{2}$$

$$s_i.nu = \max_{t_i,t_j}\{\sum_{t=t_i}^{t_j} s_i.pd(t) - s_i.dd(t), \quad t_i < t_j\} \tag{3}$$

Ideally, a self-balanced station with balanced bike flow will make $s_i.pu$ and $s_i.nu$ close to 0. However, most bike stations in NYC are far from balanced status. The NYC Citi Bike station daily averaged positive unbalance and negative unbalance distributions are presented in Figure 1(d) and Figure 1(e) as an example.



(a) Station Distribution (b) Net Flow AM (c) Net Flow PM (d) Positive Unbalance (e) Negative Unbalance

Figure 1: Station Distribution of NYC Citi Bike System.

2.1.2 Functional Zone. Since need-based customers will choose the station closest to their current locations or final destinations, we partition the bike station in service area using a Voronoi-based gridding method [2], from which the map is partitioned into regions based on walking distance to bike stations. Each grid is centered by one bike station and the points within one region is closest to its center. As a result, pick-up/drop-off points for taxi trips and POIs are mapped to the nearest bike station.

Definition 3: Voronoi Region. Let X be a space coordinate endowed with a walking distance wd extracted from Google Maps Distance Matrix API. The Voronoi region R_{s_i} associated with station s_i is the set of points in X whose distance to s_i is no greater than that to other stations:

$$R_{s_i} = \{x \in X | wd(x, s_i) \leq wd(x, s_j), \forall j \neq i\} \qquad (4)$$

Definition 4: Functional Zone. A functional zone Z_K is comprised of a group of regions $\{R_{s_i}^K\}$ with similar urban functions identified by the distribution of socioeconomic activities [21]. Each functional zone has its major category characterized by its POI structure. For example, the commercial zones have a lot of shopping centers while the transportation junctions have many transportation centers compared to other functional zones.

2.2 Problem Definition

Expansion Station Bike Demand Prediction. Given a set of existing principle bike station locations S_l^p and a set of expansion station locations S_l^e, the problem of expansion station bike demand prediction is to predict the hourly pick-up (drop-off) demand $s_i.pd(t)$ ($s_i.dd(t)$) of the expanded station (including principle stations along the edge of expansion area and coverage expansion stations) during a day.

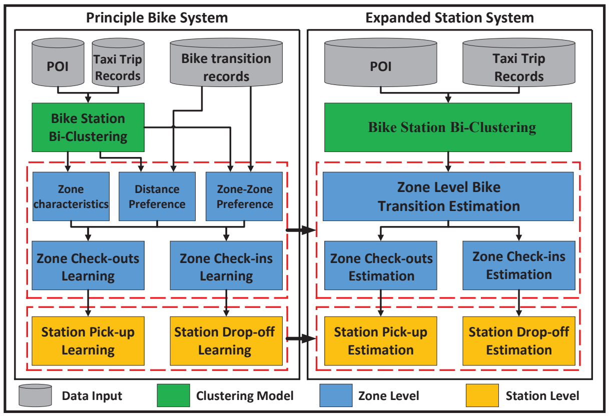


Figure 2: Framework Overview

Station unbalance Prediction. Once the hourly bike demand is estimated, we can further estimate the station unbalance characteristic according to definition 2.

2.3 Framework Overview

Figure 2 shows the framework overview of our proposed method which consists of three major sections: functional zone based bike station Bi-Clustering, Zone level bike transition prediction and station level bike demand prediction.

Functional zone based station clustering. We first propose a functional zone based Bi-clustering algorithm to cluster stations into different groups based on their Voronoi region POI structures [2, 8] and station locations. The stations within one functional zone are close to each other and designed to serve for the same functional zone customers.

Zone level bike transition prediction. The zone level bike transition prediction integrates the bike trip distance preference, zone-to-zone transition preference and zone characteristics to predict zone check-ins and check-outs based on the Random Forest predictor.
Station level bike demand prediction. The station level bike demand prediction is to distribute the inner zone bike check-ins and check-outs to individual stations based on their covered resources (POI densities).

3 METHODOLOGY

3.1 Principle station network learning

The principle station network learning studies the station level bike demand prediction model by analyzing the historical bike transition records of the principle station network in 3 steps: 1. functional zone identification; 2. zone-to-zone bike transition learning and 3. inner zone station level bike demand prediction.
Functional zone identification. We first discuss how to divide the whole bike sharing system in service area into many functional zones, where a functional zone is a subregion contains several bike stations and their associated Voronoi regions, the stations in the same functional zone have similar POI distribution and close geographical locations.

Assume the POI matrix of bike stations is $\mathbf{P} = \{p_{ij}\}$, where p_{ij} is an indicator of j-th type of POIs of bike station i. POI matrix $\mathbf{P}$ is derived from POI counts in Voronoi regions of stations, which is essentially a POI heat matrix of the bike stations. This paper considers many types of POIs. Some types of POIs have similar geographical distribution, for example, pharmacy and convenience store, they may have a symbiotic relationship; while the other types of POIs show quite different geographical distributions, e.g., financial service or car rental service. So it is very important to study the relationships between different types of POIs before partition the studied region into functional zones.

Algorithm 1 presents a novel Bi-clustering algorithm which clusters the bike stations and POI features alternatively. In step 1, we get initial station clustering result and POI feature clustering result at the same time. Then we construct virtual bike stations in step 4∼6, and get new clustering result of POI features (or POI category) in step 7. Step 8 is a break condition, where NMI is a popular information-based evaluation metric of clustering results, which describes the coherence of two clustering results[17]. In step 12∼14, we use new POI features to represent each station, where new POI features are generated according to current clustering result of POI categories. Step 15 gets new station clustering result according to stations in new feature space. K^f and K^s are the number of station clusters and that of feature clusters respectively. From a theoretical view, the setting of the number of clusters is essentially a balance of fitting error and model complexity; while in practical applications, we set the parameters according to data volume. In this problem, K^s is set as approximately 10% number of stations, K^s is about 10% number of POI types. Table 1 shows the Bi-clustering result of the studied POI categories, where two POI categories are assigned into the same cluster means the two POI types have similar geographical distribution and symbiotic relationship. Take 4-th cluster, for example, pharmacy, grocery, and store, often locate together, while the 3-rd cluster indicates the geographical distribution of taxi pickups and taxi dropoffs are almost the same.

Algorithm 1 $BiC\text{-}POIs(\mathbf{P},K^f,K^s,ItrMax)$

```
Require: Input: $\mathbf{P},K^f,K^s,ItrMax$;
 1: $Itr = 0$, $\mathbf{c}^s = kmeans(\mathbf{P}, K^s)$, $\mathbf{c}^f_0 = kmeans(\mathbf{P}^T, K^f)$;
 2: while $Itr < ItrMax$ do
 3:     $Itr = Itr + 1$;
 4:     for $i = 1 : K^s$ do
 5:         $Idxs = find(\mathbf{c}^s = i)$;
 6:         $\mathbf{x}^f_{i\cdot} = mean(\mathbf{P}_{Idxs\cdot}, row)$;
 7:     end for
 8:     $\mathbf{c}^f = kmeans(\mathbf{X}^f, K^f)$; % $\mathbf{X}^f = \{\mathbf{x}^f_{i\cdot}\}^T$.
 9:     if $NMI(\mathbf{c}^f, \mathbf{c}^f_0) = NMI(\mathbf{c}^f, \mathbf{c}^f)$ then
10:         $Break$;
11:     else
12:         $\mathbf{c}^f_0 = \mathbf{c}^f$.
13:     end if
14:     for $j = 1 : K^f$ do
15:         $Idxf = find(\mathbf{c}^f = j)$;
16:         $\mathbf{x}^s_{\cdot j} = mean(\mathbf{P}_{\cdot Idxf}, col)$;
17:     end for
18:     $\mathbf{c}^s = kmeans(\mathbf{X}^s, K^s)$; % $\mathbf{X}^s = \{\mathbf{x}^s_{\cdot j}\}$.
19: end while
```

Table 1: Clustering result of POI categories

Cluster	POI categories
1^{st}	subway, transit station, train station, finance, ...
2^{nd}	park, museum, bus station, amusement park, ...
3^{rd}	taxi pickup, taxi dropoff, ...
4^{th}	pharmacy, grocery, supermarket, store, hair care, school, shopping mall, florist, lodging, doctor, ...
5^{th}	food, cafe, bar, night club, church, spa, ATM, ...
6^{th}	parking, car rental, car wash, repair, car dealer, ...

After clustering the original POI categories into several groups, we generate new POI features based on the clustering result. The new POI heat matrix is represented by $\mathbf{H} = \{h_{ij}\}$, h_{ij} represents the heat of bike station i w.r.t. j-th category of POIs. We will discover functional zones by clustering bike stations according to their geographical locations and POI heats. However, it is a very challenging clustering problem, most of the existing clustering algorithms cannot be employed in this task because: 1) this problem requires the consideration of both geographical locations and POI features of bike stations; 2) similarities between objects are required to be predefined in most of the previous methods, and the definitions of similarities usually mix all the features of objects together. Therefore, the POI characteristics of a bike station will fade away as a result of average effect. For example, if the POI heat of a bike station is $\mathbf{h}_i = [1, 0, \cdots, 0]$, it should become a representative station in a functional zone as it has the highest heat value of the first POI type. However, if we cluster stations according to similarity defined by station feature vectors like $\|\mathbf{h}_i - \mathbf{h}_j\|$, it will be very difficult to identify representative stations with distinguished POI characteristics. That is why we develop a novel Heat Peaks based Clustering (HPC) algorithm in this paper. Algorithm 2 presents the proposed HPC method.

Algorithm 2 $HPC(\mathbf{H}, \mathbf{D}, NP^0, NP^I, \delta, NCMin)$

```
Require: Input: $\mathbf{H}, \mathbf{D}, NP^0, NP^I, \delta, NCMin$;
 1: $PeaksAll = \emptyset$; $\mathbf{c}^s = \mathbf{0}$;
 2: for $f = 1 : N^f$ do
 3:     $PeaksI = PeakDiscovery(\mathbf{H}_{\cdot f}, D, NP^0)$;
 4:     $PeaksAll = PeaksAll \cup PeaksI$;
 5: end for
 6: $PeaksAll0 = PeaksAll$;
 7: while $\exists\ c_i^s = 0$ do
 8:     for $i = 1 : N$ do
 9:         if $i \in PeaksAll$ then
10:             $c_i^s = i$;
11:         else
12:             $PeaksNI = \{j | D(i,j) \leq \delta, j \in PeaksAll\}$;
13:             if $PeaksNI \neq \emptyset$ then
14:                 $c_i^s = \underset{k \in PeaksNI}{\arg\max}\ S^f(i,k)$
15:             end if
16:         end if
17:     end for
18:     for $k = 1 : NP^t$ do
19:         $Cluster^s(k) = \left\{ i | c_i^s = PeaksAll(k), i = 1, 2, \cdots, N \right\}$
20:         if $|Cluster^s(k)| \leq NCMin$ then
21:             $c_i^s = 0,\ i \in Cluster^s(k)$;
22:             $PeaksAll = PeaksAll \backslash PeaksAll(k)$;
23:         end if
24:     end for
25:     if $PeaksAll = PeaksAll0$ then
26:         $Break$;
27:     else
28:         $PeaksAll0 = PeaksAll$;
29:     end if
30:     $I = find(\mathbf{c}^s = 0)$, $\mathbf{H}^u = \mathbf{H}_{(I,\cdot)}$, $\mathbf{D}^u = \mathbf{D}_{(I,I)}$;
31:     for $f = 1 : N^f$ do
32:         $PeaksI = PeakDiscovery(\mathbf{H}^u_{\cdot f}, D^u, NP^I)$;
33:         $PeaksAll = PeaksAll \cup PeaksI$;
34:     end for
35: end while
36: for $i = 1 : N$ do
37:     if $c_i^s = 0$ then
38:         $c_i^s = \underset{k \in PeaksAll}{\arg\max}\ S^f(i,k)$;
39:     end if
40: end for
```

The core of Algorithm 2 is $PeakDiscovery(\mathbf{h}, \mathbf{D}, K)$, which is used in step 3 and 25. This function finds K heat-peak stations according to distribution of POI heat $\mathbf{h}$ and geographical distances between stations $\mathbf{D}$, where a heat-peak station satisfies two conditions: 1) it has relatively larger POI heat value, and 2) there is no station with even higher heat value in its neighborhood. Besides heat value h_i, the other indicator of *i-th* station is defined as

$$\gamma_i = \min_{j, h_j \geq h_i} D_{ij}. \tag{5}$$

$PeakDiscovery(\mathbf{h}, \mathbf{D}, K)$ is to pick out K stations with largest η values, where $\eta_i = h_i \cdot \gamma_i$. It can be known that a heat-peak station selected by our method is a station with highest heat value in a relative large region.

In Algorithm 2, step 2~4 discover the first batch of heat-peak stations, which select NP^0 heat-peak stations from each of N^f POI categories. In the following, step 7~13 assign each bike station a cluster label by first finding heat-peak stations in its δ-neighborhood, then assigning the stations to a heat-peak station with the most similar POI mode. If there are no heat-peak stations in its δ-neighborhood, the cluster label of the station is set 0. Step 14~18 reset the cluster label of a station as 0 if the scale of the cluster it belongs to is less than $NCMin$. In step 23~26, we find new heat peaks from unlabeled bike stations. Repeat 7~26 till there are no new heat-peak stations added in an iteration. The residual unlabeled bike station is finally assigned in step 27~29.

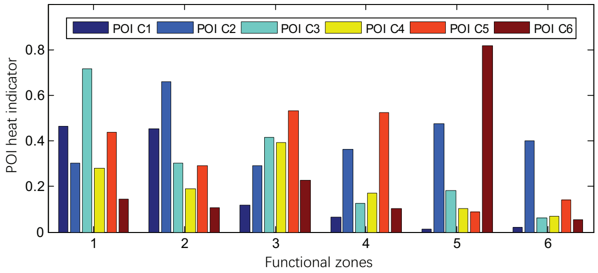


Figure 3: POI characteristics of 6 FZ categories.

We partition the current Citi Bike in service area into 6 functional zone categories. Figure 3 shows POI characteristics of the 6 functional zones categories. The first category can be defined as a mixed business zone as it contains a balanced high density of POIs from the first, third, and fifth POI category, while the third category is the residential area, which contains a large POIs density like grocery, pharmacy, and food. The other categories also have distinguished POI characteristics that can be categorized as transportation area (second functional zone category), scenic spots (fourth functional zone category), car services area (5th functional category) and education area and Park areas (the park zones that have few POIs).

It can be known that the proposed HPC algorithm can discover functional zones with the consideration of both geographical locations and POI characteristics of bike stations. The identified functional zones consist of bike stations with similar POI mode and close geographical distance. Demand prediction of bike station can not only benefit from POI features of functional zones, but also from zone-to-zone transition patterns.

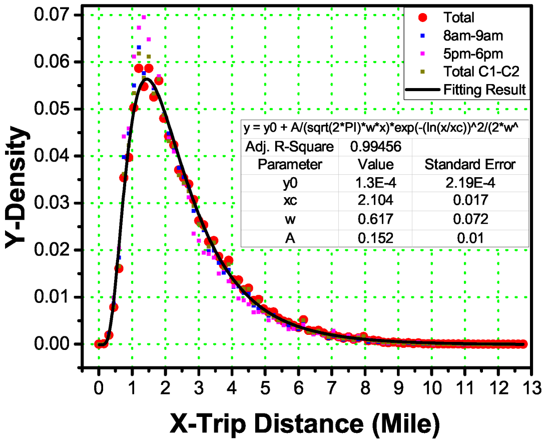


Figure 4: Bike Transition Distance Preference

	C1	C2	C3	C4	C5	C6
C1	0.11	0.18	0.16	0.23	0.18	0.15
C2	0.14	0.10	0.17	0.24	0.23	0.12
C3	0.18	0.18	0.12	0.18	0.16	0.18
C4	0.22	0.22	0.18	0.10	0.18	0.10
C5	0.12	0.21	0.15	0.21	0.07	0.24
C6	0.18	0.15	0.17	0.09	0.33	0.09

(a) Weekday 8am to 9am

	C1	C2	C3	C4	C5	C6
C1	0.08	0.11	0.15	0.21	0.16	0.29
C2	0.13	0.08	0.15	0.23	0.22	0.19
C3	0.15	0.22	0.12	0.17	0.14	0.20
C4	0.18	0.25	0.16	0.11	0.18	0.11
C5	0.13	0.21	0.12	0.18	0.11	0.26
C6	0.16	0.10	0.19	0.10	0.32	0.13

(b) Weekday 5pm to 6pm

	C1	C2	C3	C4	C5	C6
C1	0.11	0.24	0.13	0.19	0.12	0.22
C2	0.16	0.08	0.21	0.20	0.22	0.13
C3	0.17	0.22	0.12	0.15	0.15	0.19
C4	0.21	0.22	0.13	0.10	0.23	0.11
C5	0.14	0.20	0.10	0.19	0.11	0.26
C6	0.31	0.07	0.12	0.07	0.35	0.08

(c) Weekend 8am to 9am

	C1	C2	C3	C4	C5	C6
C1	0.07	0.16	0.12	0.18	0.16	0.30
C2	0.15	0.10	0.16	0.23	0.20	0.17
C3	0.17	0.29	0.10	0.16	0.13	0.16
C4	0.18	0.26	0.14	0.11	0.20	0.12
C5	0.16	0.22	0.12	0.20	0.12	0.18
C6	0.31	0.14	0.13	0.10	0.19	0.13

(d) Weekend 5pm to 6pm

Figure 5: Zone-to-Zone Transition Matrix

Zone-to-Zone transition learning. The zone-to-zone transition learning focuses on the three major factors that would affect the check-outs and check-ins of each functional zone: trip distance preference, zone-to-zone preference, and zone characteristics.

Distance Preference Learning. The distance preference refers to the distance range that a person prefers to taking a bike other than other transportation methods such as subways or taxis. Mathematically, the pick-up frequency density versus transition distance forms a log-normal distribution (see the blue fitting line in Figure 4). As can be seen, the bike transition distance distributions are identical during different time periods and between different functional zones. Therefore, given the locations of origin $o.c$ and destination $d.c$, associated with their distance $x \equiv \|o.c - d.c\|$, we can estimate the users' distance preference of taking bicycles, which is defined by a Distance Preference Score (DPS):

$$DPS(x) = y_0 + \frac{A}{\sqrt{2\pi}wx}exp(-\frac{(ln(x/x_c))^2}{2w^2}) \tag{6}$$

Where y_0, A, w, x_c are fitting parameters (see fitting results in inserted table of Figure 4). The formula of DPS indicates that people would not like to take bikes for long term trip (larger than 4 miles) or within walking distances. People who have an origin-to-destination distance in the range of $FWHM$ (full width at half maximum) of DPS (1.5 miles $\sim$ 2.7 miles) are more willing to take bicycles.

Zone-to-zone preference learning. Besides the distance preference, customers have their functional zone preference in different periods. For example, during AM rush hour, customers will have a high preference to take bikes from subways exits to business areas even though a nearby parking lot has a higher DPS. Here we define a functional zone transition matrix $T(C_i, C_j)$ to describe the transition preference from functional zones of class C_i to functional zones of class C_j that satisfies:

$$T(C_i, C_j)\sum_{n=1}^{N} DPS(x; o_n.c \in C_i, d_n.c \in C_j) = N \tag{7}$$

Where N represents the total transitions from functional zone of class C_i to C_j. Table 5 presents the normalized FZ transition preference matrix of 4 time periods: weekday 8 am-9 am 5(a), weekday 5 pm-6 pm 5(b), weekend 8 am-9 am 5(c) and weekend 5 pm-6 pm 5(d). As can be seen, bike users are least likely to move between the functional zones of the same class (small diagonal value of T), which indicates that in order to motivate more bike users, the functional zones should be diversified. The transition matrix varies in different time periods and some classes have high links in different time periods.

Therefore, given two functional zone Z_m, Z_n with its location center $Z_m.l, Z_n.l$ and class $Z_m.C, Z_n.C$, we define the Zone Transition Score ZTS of time period t as follows:

$$ZTS(Z_m, Z_n; t) = T(Z_m.C, Z_n.C; t)DPS(|Z_m.l - Z_n.l|) \tag{8}$$

The functional zone check-out preference score ZPS_{out} and check-in preference score ZPS_{in} are then defined by considering the transition scores to and from all surrounding functional zones:

$$ZPS_{out}(m; t) = \sum_{n \neq m} ZTS(Z_m, Z_n; t) \tag{9}$$

$$ZPS_{in}(m, in; t) = \sum_{n \neq m} ZTS(Z_n, Z_m; t) \tag{10}$$

Zone Characteristics. The last factor considered in our prediction model that could affect the bike demands is the characteristics of each functional zone, including the densities of 6 major POI categories, historical taxi check-outs and check-ins, and the number of available docks.

Entire traffic prediction. After we extract the most influential factors from the zone characteristics, zone check-out and check-in preference scores, the entire check-outs $Z_{out}(m; t)$ and check-ins $Z_{in}(m; t)$ are predicted by feeding the factors into the Random Forest Regressor (RF) for different time periods.

Station level bike demand prediction. After we predict the check-ins and check-outs of each Functional zone, the station bike demand in the functional zone is predicted by ridge regression $f^C(s_i.F)$, indicating the number of check-ins or check-outs distributed to each station based on the station level feature vector F (Voronoi area POI densities and their distance to nearest transportation entrances, such as parking lots, subway entrances and bus stops). The logistic regressor for pick-up demand and drop-off demand are trained by the historical transition records of stations within the same functional zone of category C. For each station s_i located in Functional Zone Z_m of category C, the station level pick-up $s_i.pd$ and drop-off demand $s_i.dd$ are formally predicted as follows:

$$s_i.pd(t) = Z_{out}(m;t)\frac{f_p^C(s_i.F)}{\sum_{s_j \in Z_m} f_p^C(s_j.F)} \tag{11}$$

$$s_i.dd(t) = Z_{in}(m;t)\frac{f_d^C(s_i.F)}{\sum_{s_j \in Z_m} f_d^C(s_j.F)} \tag{12}$$

3.2 Demand prediction after expansion

There are two kinds of stations for expansion: station coverage expansion and complementary station expansion. Different kinds of expansion strategy have different effects on the demand related factors.

The station setup for coverage expansion is to set up new stations in the area that has no stations before. As a result, the new functional zones of the expansion areas will affect the zone-to-zone connection preference and transition distance preference of the principle system. By the definition of Zone Transition Score ZTS, which decreases fast for long distance transportation, the coverage expansion stations will have fewer effects on the functional zones located far away compared to the functional zones located near the expansion edges. The complementary station expansion aims at reducing the station workload by adding one or more stations to existing bike sharing system covered functional zones. The complementary stations have fewer effects on zone level bike check-ins/check-outs predictions but will redistribute the bike pick-ups and drop-offs within the functional zones.

Although different expansion strategies have different effects, these effects are reflected in the changes of our predictor input feature vectors. To predict the station demand prediction after expansion, we reconstruct the input features at zone level and station level after expansion and implement the predictors we have trained in the principle station network learning.

4 EXPERIMENTAL RESULTS

To validate the efficiency and effectiveness of our proposed method, extensive experiments are performed on real world NYC CitiBike trip data of three different time periods. The first stage is the principle bike station system consisting of 329 stations from 07/01/2013 to 07/31/2015. The second stage has 486 stations from the completion of first expansion on 08/06/2015 to 07/18/2016. And the third stage starts from 07/23/2016 to 11/30/2016, with 617 stations in service after the second expansion. All experiments are conducted on a PC 7 with an Intel(R) Core i7-4790 CPU, 3.6 GHz, and 16 GB RAM running 64-bit Windows 10 system.

4.1 Experimental Data

We conduct our experiments with bike sharing system data, Google Place API[1], taxi trip records from NYC with their statistics presented in Table 2. Citibike transition records are generated by NYC Bike Sharing System which is public available from Citibike official website [2]. 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. In addition, the station status is crawled every 10 minutes from station status feed site [3] which contains the information of station in service status, currently available bikes and station capacity.

Table 2: Details of the datasets

Data Source	**New York City Bike System**		
Time from	7/1/13	8/6/15	7/29/16
to	7/31/15	7/18/16	11/30/16
Weekdays	524	238	85
(Weekends)	(237)	(110)	(40)
#Stations	329	486	617
#Records	17.58 million	11.76 million	6.07 million

Data Source	**Google Place API**		
POI type	number	POI type	number
establishment	70335	car service	1088
education	2784	supermarket	4077
shopping mall	206	entertainment	996
store	28418	bus station	1981
lodging	1262	railway station	1142
home service	1166	finance	8103
convenience	9914	estate agency	5693
health center	42164	restaurant	11825
night life	4115	travel agency	1595
fitness	1357	···	···

Data Source	**New York City Taxi Trip Records**	
effective days	time Period	# of trip records
31	08/2013	12.6 million

4.2 Baselines & Metric

The methods proposed in our work to predict the station level pick-up demand and drop-off demand are denoted as Functional Zone based Random Forest Regressor (**FZ+RF**). In order to confirm the effectiveness of our models, we conduct experiments to compare our methods with the following baselines:

Station Level Predictor [7]: The station level predictors estimate the bike demand based on a set of global feature elements. The baselines of station level predictors we use in this paper include Random Forest **(RF)**, K-Nearest Neighbor Regressor **(KNN)**, Neural Network (NN) and Gradient

[1] https://developers.google.com/places/

[2] https://www.citibikenyc.com/system-data

[3] https://feeds.citibikenyc.com/stations/stations.json

Boosting Regressor **(GBR)**. The features used for station level predictors include the 19 fine grained POI densities, Voronoi region taxi check-ins and check-outs.

Hierarchical Demand Predictor: Considering our Functional Zone based station clustering is the first attempt, we use the **FZ+GBR** as a baseline. The only difference between the **FZ+GBR** and our method is that it uses GBR for zone level bike transitions prediction.

Metric: The metrics we adopt to measure the performance are the Error Rate *ER* and Root Mean Squared Logarithmic Error *RMLSE*, which are formally defined as follows:

$$ER(t) = \frac{\sum_{i=1}^{N} |\hat{s}_i.d(t) - s_i.d(t)|}{\sum_{i=1}^{N} s_i.d(t)}$$

$$RMLSE(t) = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(log(\hat{s}_i.d(t)+1) - log(s_i.d(t)+1))^2}$$

Here $s_i.d(t)$ is our ground truth of bike pick-up or drop-off demand of station i during time slot t and $\hat{s}_i.d(t)$ is the corresponding prediction value. The principle stations in Stage 1 are used as training set. For the expansion analysis of Stage 2 and Stage 3 bike sharing systems, only the bike stations in the expansion areas, the stations located within the functional zones taht are adjacent to the expansion boundaries, and the complementary stations are included in the testing set.

4.3 Demand Prediction

Hourly station bike demand prediction after first expansion. The performance comparison for first expansion bike demand prediction (including weekday pick-up demand, weekday drop-off demand, weekend pick-up demand and weekend drop-off demand) between our proposed FZ+RF and baselines is summarized in Figure 6. From Figure 6, we can see that for all time periods, both of the Error Rate (ER) and the Root Mean Squared Logarithmic Error (RMSLE) obtained from our proposed method are much lower than all the baselines with a significant margin. Moreover, the hierarchical demand predictor based on functional zone station clustering (FZ+GBR and FZ+RF represented by dot lines) achieve a better performance than station level predictors based on global features (represented by star symbol lines). The high ER of early morning predictions is mainly due to the few transition records which amplify the ER but leads to a very small RMSLE.

Hourly station bike demand prediction after second expansion. Figure 7 presents the performance comparison for bike demand prediction after the second system expansion. As can be seen, our proposed method lower the ER and RMSLE of bike demand predictions of different time periods. However, compared to the bike demand prediction performance after the first expansion, the ER and RMSLE increase. It might be due to the reason that the functional zones in the second expansion areas have a larger difference compared to that in the principle area.

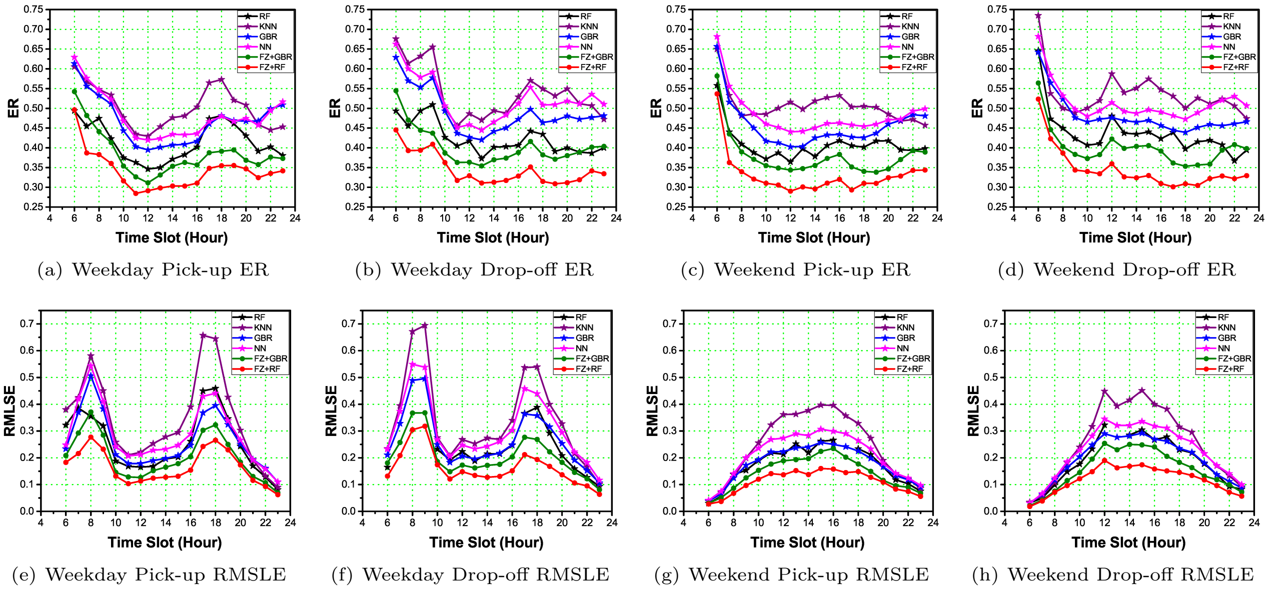
Overall Performance Comparison. The daily averaged pick-up demand, drop-off demand, positive balance and negative balance prediction accuracy comparisons are represented in Figure 8. For the first stage expansion, our method achieves an overall pick-up ER of 0.3118 which is 0.0482 lower than the other hierarchical demand predictor (stage 1) and 0.0863 lower than the most competitive station level predictor RF. The overall drop-off demand ER of our method is 0.3295 which is much lower than other baselines. In terms of RMLSE, our method achieves an overall RMLSE of 0.1096, 0.1184, 0.1509 and 0.157 for stage-1 pick-up demand prediction, stage-1 drop-off demand prediction, stage-2 pick-up demand prediction and stage-2 drop-off demand prediction respectively. Moreover, an accurate hourly demand prediction can also benefit the station unbalance prediction. The Figure 8(c) and Figure 8(d) summarize the performance of the positive unbalance and negative unbalance prediction. As can be seen, our proposed method can provide a more accurate unbalance status prediction which can further help bike sharing system designers estimate the rebalancing operation cost after bike sharing network expansion.

5 RELATED WORK

There is an increasing interest in optimization problems arising in bike sharing systems. Below we describe some related studies that have been accomplished on demand prediction for bike sharing systems.

Station Clustering. The clustering algorithms have been proposed to discovery bike transition patterns, reduce the station demand variance and improve prediction accuracy. Patrick, etc [18] explored bike activity patterns based on temporal and spatial validation of clusters, and revealed imbalances in the distribution of bikes. Lin, etc [7] proposed a Bipartite station clustering algorithm consisting of Geo-clustering and Bike-Transit-Clustering according to the similarities of bike usage patterns and station locations. Similarly, Chen, etc. [3] proposed a Geographically-Constrained Station Clustering to group stations. However, their station clustering algorithms are based on historical bike transition records. Motivated by the Functional Zone discovery analysis [10, 21], our Bi-Clustering algorithm is based on both of POI structures and station geographical constraints that could be applicable for the expansion areas where no historical bike transition records are available. The identification of heat-peak bike stations is motivated by a recently proposed clustering method that is published in *Science* in 2014 [14], however, the cluster centers in our work are POI heat peaks rather than density peaks. In order to discover functional zones with distinguished POI characteristics, we proposed the HPC clustering algorithm, which can find representative stations via different POI categories. Different from most existing clustering algorithms based on predefined similarities of objects[11, 20], the POI characteristics of stations fadeaway in the process of computing similarities.

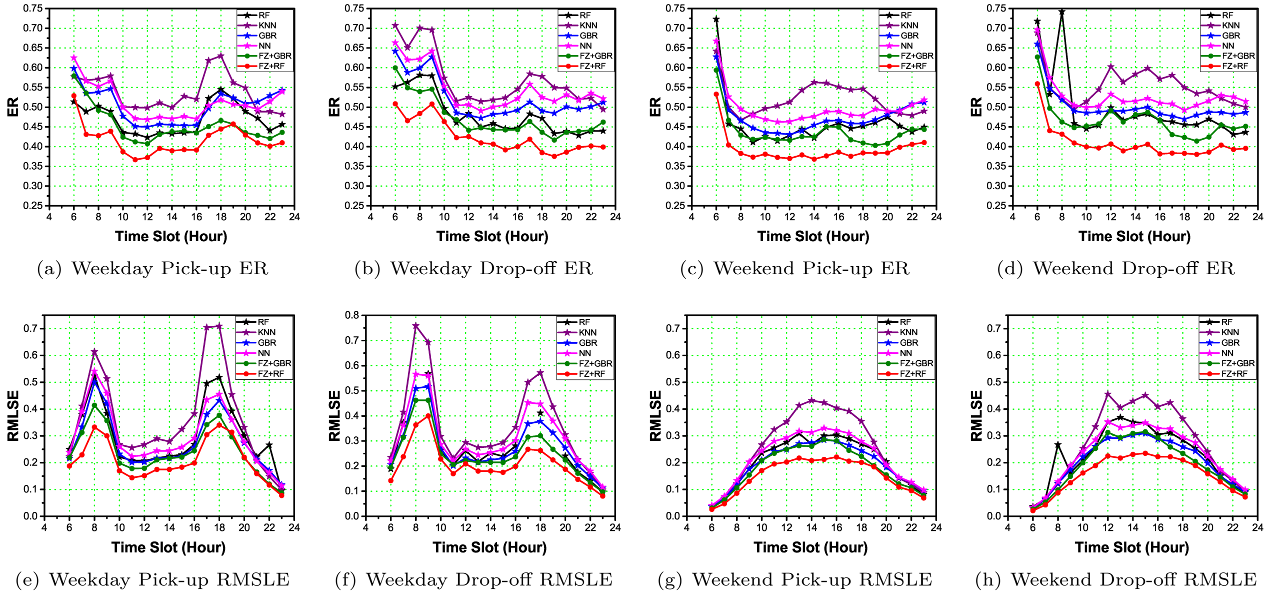
Bike Demand Prediction. The early research on bike sharing system focused on the studies of bike activity patterns discovery [6, 13, 23] or daily bike demand forecasting using data mining techniques and classical empirical statistical methods.



(a) Weekday Pick-up ER (b) Weekday Drop-off ER (c) Weekend Pick-up ER (d) Weekend Drop-off ER

(e) Weekday Pick-up RMSLE (f) Weekday Drop-off RMSLE (g) Weekend Pick-up RMSLE (h) Weekend Drop-off RMSLE

Figure 6: Performance comparison of station bike demand prediction after first expansion.

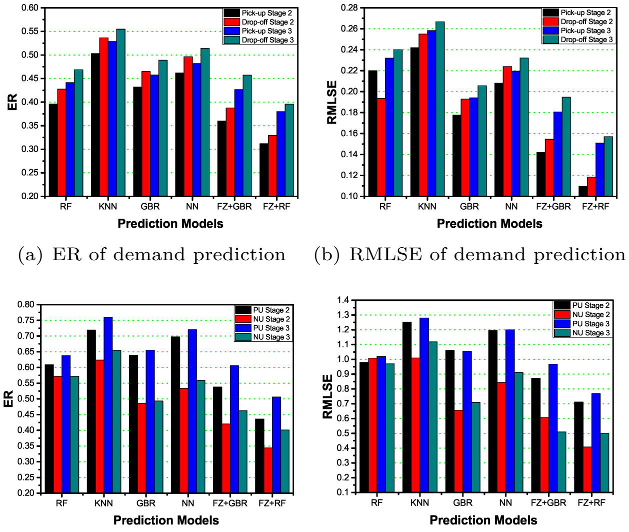


(a) Weekday Pick-up ER (b) Weekday Drop-off ER (c) Weekend Pick-up ER (d) Weekend Drop-off ER

(e) Weekday Pick-up RMSLE (f) Weekday Drop-off RMSLE (g) Weekend Pick-up RMSLE (h) Weekend Drop-off RMSLE

Figure 7: Performance comparison of station bike demand prediction after second expansion.

Yutaka [12] and Juan[5] built multi-factor statistical models for bicycle demand prediction with the consideration of weather and geography. The hourly bike demand prediction was investigated by implementing statistical models or machine learning techniques on multi-source data [1, 9, 15]. A hierarchical bike traffic prediction model that integrating station clustering algorithm and meteorology reports were also studied [7]. However, all of these prediction models require the availability of historical transition records of target stations and thus are not applicable for expansion demand prediction problem. Liu etc.[8], Wang [19] and Zeng etc.[22] built station demand prediction models by extracting global features from multiple static factors of surrounding environment and public transportation networks. However, among these multi-factor prediction models, the feature bias among stations across the urban areas are neglected with the global features and predictors. Moreover, the direct analysis of station-to-station bike transition may suffer from insufficient transition records at station level.



(a) ER of demand prediction

(b) RMLSE of demand prediction

(c) ER of balance prediction

(d) RMLSE of balance prediction

Figure 8: Overall Performance Comparison

6 CONCLUSION

In this paper, we developed a hierarchical bike demand prediction models for expansion area station level bike demand prediction. Specifically, we first partitioned the station in service area into different functional zones based on our Bi-Clustering algorithm. Then based on the functional zones, we implemented Random Forest Regressor to estimate the functional zone bike transitions by integrating the bike trip distance preference, zone-to-zone preference, and zone characteristics. The station level bike demand was predicted by distributing the zone level check-ins and check-outs to each station with the consideration of their Voronoi region POI structures. Finally, the extensive experiments on real-world data from the 3-stage NYC Citi Bike System showed the advantages of our hierarchical strategy of bike demand prediction for bike sharing system expansion.

7 ACKNOWLEDGMENTS

This research was partially supported by the National Science Foundation (NSF) via the grant number IIS-1648664. Also, it was supported in part by the Natural Science Foundation of China (71329201).

REFERENCES

[1] Ramon Alvarez-Valdes, Jose M. Belenguer, Enrique Benavent, Jose D. Bermudez, Facundo Muñoz, Enriqueta Vercher, and Francisco Verdejo. 2016. Optimizing the level of service quality of a bike-sharing system. *Omega* 62 (2016), 163 – 175.

[2] Franz Aurenhammer. 1991. Voronoi diagrams, a survey of a fundamental geometric data structure. *ACM Computing Surveys (CSUR)* 23, 3 (1991), 345–405.

[3] Longbiao Chen, Daqing Zhang, Leye Wang, Dingqi Yang, Xiaojuan Ma, Shijian Li, Zhaohui Wu, Gang Pan, Thi-Mai-Trang Nguyen, and Jérémie Jakubowicz. 2016. Dynamic Cluster-based Over-demand Prediction in Bike Sharing Systems. In *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp '16)*. ACM, New York, NY, USA, 841–852.

[4] Paul DeMaio. 2009. Bike-sharing: History, impacts, models of provision, and future. *Journal of Public Transportation* 12, 4 (2009), 3.

[5] Juan Carlos GarcÍła-Palomares, Javier GutiÍęrrez, and Marta Latorre. 2012. Optimizing the location of stations in bike-sharing programs: A {GIS} approach. *Applied Geography* 35, 1ÍC2 (2012), 235 – 246.

[6] Andreas Kaltenbrunner, Rodrigo Meza, Jens Grivolla, Joan Codina, and Rafael Banchs. 2010. Urban cycles and mobility patterns: Exploring and predicting trends in a bicycle-based public transport system. *Pervasive and Mobile Computing* 6, 4 (2010), 455 – 466. Human Behavior in Ubiquitous Environments: Modeling of Human Mobility Patterns.

[7] Yexin Li, Yu Zheng, Huichu Zhang, and Lei Chen. 2015. Traffic Prediction in a Bike-sharing System. In *Proceedings of the 23rd SIGSPATIAL International Conference on Advances in Geographic Information Systems (SIGSPATIAL '15)*. ACM, New York, NY, USA, Article 33, 10 pages.

[8] J. Liu, Q. Li, M. Qu, W. Chen, J. Yang, H. Xiong, H. Zhong, and Y. Fu. 2015. Station Site Optimization in Bike Sharing Systems. In *2015 IEEE International Conference on Data Mining*. 883–888.

[9] Junming Liu, Leilei Sun, Weiwei Chen, and Hui Xiong. 2016. 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 (KDD '16)*. ACM, New York, NY, USA, 1005–1014.

[10] Ying Long and Zhenjiang Shen. 2015. Discovering functional zones using bus smart card data and points of interest in Beijing. In *Geospatial analysis to support urban planning in Beijing*. Springer, 193–217.

[11] Ulrike Luxburg. 2007. A tutorial on spectral clustering. *Statistics and Computing* 17, 4 (2007), 395–416.

[12] Yutaka Motoaki and Ricardo A. Daziano. 2015. A hybrid-choice latent-class model for the analysis of the effects of weather on cycling demand. *Transportation Research Part A: Policy and Practice* 75 (2015), 217 – 230.

[13] Oliver O?Brien, James Cheshire, and Michael Batty. 2014. Mining bicycle sharing data for generating insights into sustainable transport systems. *Journal of Transport Geography* 34 (2014), 262 – 273.

[14] Alex Rodriguez and Alessandro Laio. 2014. Clustering by fast search and find of density peaks. *Science* 344, 6191 (2014), 1492–1496.

[15] Jasper Schuijbroek, Robert Hampshire, and Willem-Jan van Hoeve. 2013. Inventory rebalancing and vehicle routing in bike sharing systems. (2013).

[16] Susan A Shaheen, Stacey Guzman, and Hua Zhang. 2010. Bike-sharing in Europe, the Americas, and Asia. *Transportation Research Record: Journal of the Transportation Research Board* 2143, 1 (2010), 159–167.

[17] Nguyen Xuan Vinh, Julien Epps, and James Bailey. 2009. Information Theoretic Measures for Clusterings Comparison: Is a Correction for Chance Necessary. In *Proceedings of the 26th ICML*. 1073–1080.

[18] Patrick Vogel, Torsten Greiser, and Dirk Christian Mattfeld. 2011. Understanding bike-sharing systems using data mining: Exploring activity patterns. *Procedia-Social and Behavioral Sciences* 20 (2011), 514–523.

[19] Wen Wang. 2016. Forecasting Bike Rental Demand Using New York Citi Bike Data. (2016).

[20] Rui Xu and D. Wunsch, II. 2005. Survey of Clustering Algorithms. *IEEE Transactions on Neural Networks* 16, 3 (2005), 645–678.

[21] Nicholas Jing Yuan, Yu Zheng, Xing Xie, Yingzi Wang, Kai Zheng, and Hui Xiong. 2015. Discovering urban functional zones using latent activity trajectories. *IEEE Transactions on Knowledge and Data Engineering* 27, 3 (2015), 712–725.

[22] Ming Zeng, Tong Yu, Xiao Wang, Vincent Su, Le T Nguyen, and Ole J Mengshoel. 2016. Improving Demand Prediction in Bike Sharing System by Learning Global Features. *Machine Learning for Large Scale Transportation Systems (LSTS)@ KDD-16* (2016).

[23] Xiaolu Zhou. 2015. Understanding Spatiotemporal Patterns of Biking Behavior by Analyzing Massive Bike Sharing Data in Chicago. *PLOS ONE* 10, 10 (10 2015), 1–20.