2016 17th IEEE International Conference on Mobile Data Management

Bicycle-Sharing System Analysis and Trip Prediction

Jiawei Zhang⋆, Xiao Pan†, Moyin Li⋆, Philip S. Yu⋆
⋆University of Illinois at Chicago, Chicago, IL, USA
†Shijiazhuang Tiedao University, China
jzhan9@uic.edu, smallpx@stdu.edu.cn, mli60@uic.edu, psyu@cs.uic.edu

***Abstract*—Bicycle-sharing systems, which can provide shared bike usage services for the public, have been launched in many big cities. In bicycle-sharing systems, people can borrow and return bikes at any stations in the service region very conveniently. Therefore, bicycle-sharing systems are normally used as a short-distance trip supplement for private vehicles as well as regular public transportation. Meanwhile, for stations located at different places in the service region, the bike usages can be quite skewed and imbalanced. Some stations have too many incoming bikes and get jammed without enough docks for upcoming bikes, while some other stations get empty quickly and lack enough bikes for people to check out. Therefore, inferring the potential destinations and arriving time of each individual trip beforehand can effectively help the service providers schedule manual bike re-dispatch in advance. In this paper, we will study the individual trip prediction problem for bicycle-sharing systems. To address the problem, we study a real-world bicycle-sharing system and analyze individuals' bike usage behaviors first. Based on the analysis results, a new trip destination prediction and trip duration inference model will be introduced. Experiments conducted on a real-world bicycle-sharing system demonstrate the effectiveness of the proposed model.**

***Index Terms*—Trip Prediction, Bicycle-Sharing System, Mobile Data Mining**

I. Introduction

Bicycle-sharing system refers to a public transportation service system in urban areas offering bicycles for shared use to individuals in a relatively short period of time (about $30-45$ minutes) for free or with very low charges [8]. In bicycle-sharing systems, people can borrow bikes from stations near them and return the bike to any stations in the city, which can be used as a short-distance trip supplement for private vehicles as well as regular public transportation (e.g., buses and metro trains). Bicycle-sharing system is green and of low carbon, and each bike can be used by several people per day. What's more, due to the widely spread branches and stations available in the city, people can usually borrow and return the bikes very conveniently without wasting time on waiting (needed for the public transportation) or concerns about parking issues in cities (of private vehicles). As a result, bicycle-sharing systems are becoming more and more popular nowadays, which have been adopted in many large cities, e.g., Chicago (Divvy Bike), New York (Citi Bike), San Francisco (Bay Area Bike Share), Washington, D.C. (Capital Bikeshare).

Bicycle-sharing system allows people to borrow bikes with either "one-day pass" or "annual subscribed membership". "One-day pass" is usually preferred by people for temporary usages, e.g., tourist for short-time sightseeing, but the charges per day are slightly higher. Meanwhile, "subscribed membership" is a great option for people with frequent travel needs, e.g., office worker and students. Generally, trips completed by one-day pass/membership holders within 30 minutes are included in the pass/membership, but trips longer than 30-minutes may incur overtime fees. More information about the detailed pricing rules is available at Divvy's official website[1].

Unlike traditional fixed-route public transportation at pre-scheduled time, services provided by bicycle-sharing systems are more flexible and can meet the daily travel needs of different categories of users. Bicycle-sharing system provides a more microscopic perspective to understand individuals' travel behaviors, which include various aspects about the trips, e.g., trip origin station and start time, as well as trip destination stations and end time. Generally, the travel behaviors of different categories of people with various travel purposes can be quite different. For instance, tourists with one-day pass tend to use the bike to travel among attraction spots, while registered subscribers (like workers and students) mainly travel between companies/schools and homes with the bike.

Meanwhile, for stations located at different places in the city, the bike usage can be quite skewed and imbalanced [5]. Some stations that individuals like to borrow bikes from will lack enough bikes for people to check out, while some other stations that people normally return the bikes to will get jammed easily without enough docks for upcoming bikes. To support such a claim, we also analyze the real-world bicycle-sharing system data (to be introduced in Section II), and count the numbers of bikes borrowed from/returned to each stations respectively. According to the analysis results, among all the 474 stations, 470 of them have historical usage records: 235 stations have more bikes being checked out (i.e., # bikes checked out>\# bikes returned), 234 of them have more returned bikes (i.e., # bikes checked out<\# bikes returned), and only one station (station ID: 449) has balanced usages (i.e., # bikes checked out=# bikes returned). Therefore, one of the most challenging task for the effective operations of bicycle-sharing systems is to manually shift and rebalance the bikes from the jammed stations to the empty ones. Monitoring the bike usage and inferring the potential destinations of

[1] https://www.divvybikes.com/pricing

2375-0324/16 $31.00 © 2016 IEEE
DOI 10.1109/MDM.2016.35

TABLE I
PROPERTIES OF THE DIVVY DATASET

datasets	trip	station
2013 Q3-Q4	759,788	300
2014 Q1-Q2	905,699	300
2014 Q3-Q4	1,548,935	300
2015 Q1-Q2	1,096,239	474

individuals' trips in advance (e.g., at the moment when individuals borrow a bike and start their trips) can help the service providers schedule the manual bike re-dispatch beforehand.

Problem Studied: In this paper, we propose to predict the potential destination station and arriving time when people start their trips and check out bikes from the origin station at the very beginning. The problem is formally defined as the "trip prediction" problem.

The trip prediction problem is an interesting yet important research problem, which is also very challenging to address as individuals' bike trips can be quite complicated and depend on various factors:

- *Users Composition*: The user composition of bicycle-sharing systems can be quite diverse, which include both (1) long-term registered subscribers and short-term temporary users, (2) male users and female users, as well as (3) young, mid-aged and senior users. The trip prediction problem can be strongly correlated to user categories, and a clear categorization of the bike users can be the prerequisite for addressing the problem.
- *Temporal Travel Patterns*: Start time of a trip is another important factor that may influence individuals' travel behavior as well as the trip prediction problem. Consider, for example, when a registered member (e.g., a student) borrows a bike in the morning on workdays, it is highly likely that he/she will go to a school for classes. Analyzing the individuals' historical temporal travel patterns will help predict the trips more accurately.
- *Spatial Travel Patterns*: Besides the time factor, the origin location is another important factor affecting the trip and individuals' travel behaviors. For instance, if a temporary one-day pass holder borrows a bike from a station at the entrance of a sight-seeing trail, he may want to go to the end of the trail. Studying and utilization the historical spatial travel patterns of individuals can help improve the trip prediction performance a lot.

To address the trip prediction problem, in this paper, we will analyze the user composition, individuals' temporal and spatial travel behavior patterns of a real-world bicycle-sharing system. Based on the analysis results, we will formulate the trip prediction problem and introduce new models to infer both the trip destination station and trip duration.

The remaining part of the paper is organized as follow. In Section II, we first introduce the Divvy bicycle-sharing system dataset and give some basic statistical information about the dataset. The dataset analysis results about the user decomposition, individuals' temporal and spatial travel patterns are available in the complete version of this paper [14]. Based on the analysis results, we formulate the trip prediction problem in Section III and introduce the trip prediction model in Section IV, which is evaluated in Section V. Finally, we discuss the related works in Section VI and conclude the paper in Section VII.

II. DIVVY DATASET DESCRIPTION

Before analyzing the individuals' travel behaviors, we will introduce the dataset about a real-world bicycle-sharing system first in this section. The dataset used in this paper is about the Divvy bicycle-sharing system initially launched in the Chicago city on June 28, 2013. At the very beginning, Divvy had about 750 bikes at 75 stations (operating in an area spanning from the Loop north to Berwyn Avenue, south to $59th$ Street, west to Kedzie Avenue, and east to the Lake Michigan coast). A quick expansion has been made at early 2015, and Divvy now operates $4,760$ bicycles at 474 stations (in an area bounded by $75th$ Street on the south, Touhy Avenue on the north, Lake Michigan on the east, and Pulaski Road on the west).

The Divvy bicycle-sharing system datasets are public and new datasets are released every two quarters, which can be downloaded at its official website[2]. We downloaded the Divvy bicycle-sharing system data on November 2, 2015, which contains 4 separate datasets time ranging from the middle of 2013 to the middle of 2015 respectively. The downloaded datasets include the complete historical trip records as well as the station information, whose statistical information and detailed descriptions are available in Table I and as follows.

- *Trip*: Each trip record in the datasets has a unique ID. From the trip record data, we can know the trip start and end time as well as the corresponding origin and destination bike stations. The trip record also indicates whether the user is an annual membership holder or just an one-day pass holder, who are called the "subscriber" and "customer" respectively. For the annual membership subscribers, the trip record data also includes their gender and birth year information, which is helpful for categorizing the users (into male vs female, as well as youth vs senior) and allows us to study the bike-usage behaviors of different categories of people.
- *Station*: For each station, we can know its ID, name as well as its specific location, which is represented as a (latitude, longitude) coordinate pair in the dataset. At stations, bikes are locked at the docks and the numbers of docks available at the stations are called the station capacities, which are also available in the datasets.

As shown in Table I, the numbers of trips in these 4 separate datasets are 2013 Q3-Q4: $759,788$; 2014 Q1-Q2: $905,699$; 2014 Q3-Q4: $1,548,935$; 2015 Q1-Q2: $1,096,239$ respectively. Generally, the Chicago people like to use the Divvy bike a lot and, on average, $179,610$ trips were taken in each month during the past two years. Meanwhile, the number of stations doesn't change in the first 3 datasets (which are all

[2]https://www.divvybikes.com/data

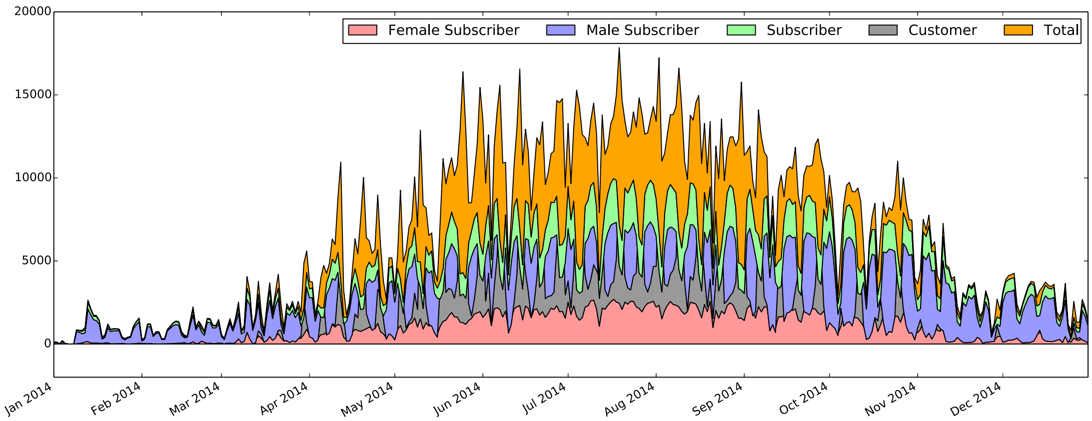


Fig. 1. Trip statistics on each day of the 2014 year (X axis: each day of 2014; Y axis: number of trips in one day).

300), and increases to 474 in the last dataset because of the scheduled expansions at the beginning of 2015. The trips taken by different categories of users during the 2014 year is shown in Figure 1.

Due to the limited space, the dataset analysis sections about the bike user composition, individuals' temporal travel patterns, and spatial travel patterns are omitted. For more information about the analysis results, please refer to the complete version of this paper [14].

III. Trip Prediction Problem Formulation

Based on the analysis results [14], we will introduce the trip prediction problem in this section. The trip prediction problem studied in this paper aims at inferring the destination station and trip end time, given that a user borrows a bike from a Divvy station at certain time. We propose to formulate the problem as an origin and destination station pair prediction problem in this paper.

In other words, for a given user, who has borrowed a bike from a known Divvy station A at time t, the trip prediction problem aims at returning a set of potential destination Divvy station candidates in the decreasing order of their likelihood that u will ride the bike to as well as the trip duration τ. The trip end time can be represented as $t+\tau$. In the trip prediction problem, we can represent trip origin and destination stations as pairs (s_o, s_d), where s_o denotes the origin station and s_d represents the trip destination station. Based on the existing historical data, a set of features that depict either the user or the characteristics of stations s_o, s_d can be extracted, which can be represented as vector $\mathbf{x}(s_o, s_d) \in \mathbb{R}^k$ of length k (the features will be introduced in Section IV). Pair (s_o, s_d) can be labeled with relevance scores between the origin station s_o and the potential destination station s_d, which can be represented as $y(s_o, s_d)$ ($y(s_o, s_d) = +1$ if the trip ends at station s_d and 0 otherwise). Meanwhile, the time duration of trip from s_o and s_d can be denoted as $t(s_o, s_d) \in \mathbb{R}$.

Formally, let $\mathcal{T}$ be the training set containing labeled station pairs. We can represent the features and labels extracted for pairs in $\mathcal{T}$ as $\mathcal{D} = \{\mathbf{x}(s_o, s_d), y(s_o, s_d), t(s_o, s_d)\}$. The trip prediction problem can be formalized as building two functions $f : \mathbb{R}^k \to \{1, 0\}$ and $h : \mathbb{R}^k \to \mathbb{R}$, where function f maps the station pair feature vector to their inferred relevance score (i.e., the likelihood for the trip to finish at the potential destination stations), while function h maps the feature vector to the inferred trip duration time. These two regression functions will be applied to the potential stations in the test set and can return the predictive confidence scores $\{y(s_o, s_d)\}_{(s_o,s_d)}$ and duration length $\{t(s_o, s_d)\}_{(s_o,s_d)}$ for station pairs in the test set.

IV. Trip Prediction Model

To address the trip prediction problem, in this part, we will introduce the prediction model in detail. First, we will introduce the features extracted for station pairs based on information about users, start time and stations. Next, we will briefly talk about the specific models used in this paper.

A. *Features about the user*

Throughout the previous data analysis sections, the type of users (i.e., "customer" vs "male subscriber" vs "female subscriber") have significant influences on the bike trips in both destination stations and trip duration. As a result, based on the user personal information, we propose to extract 3 features about the users, which include

- *User Type*: "Customers" normally behave very differently from the "subscribers" in Divvy bike usage. To differentiate them from each other, based on the user type information, we propose to extract feature x_1. If user u is a subscribed user, then $x_1 = +1$; otherwise, $x_1 = -1$.
- *User Gender*: "Male" uses Divvy bike more often and their activity region concentrates around the Chicago loop area, which is different from the "female" users. To denote the gender about the subscribed users, we define feature x_2, where $x_2 = +1$ for "male subscribers" and $x_2 = -1$ for "female subscribers". For "customers", we have no idea about their gender and we will assign $x_2 = 0$.
- *User Age*: In addition, the birth year information is available for subscribers. Young people and mid-aged people tend to use Divvy bike more often. We propose to extract

feature x_3 to represent the user age. For "customers", we set $x_3 = 0$, as we don't know their ages.

B. Features about the departure time

Besides the users information, the Divvy bike usage is also correlated with the trip start time a lot. For instance, people use Divvy bike more often in the summer; "customers" tend to use Divvy bike at weekends; "subscribers" mainly use the Divvy bike during the rush hours. Therefore, 3 different features are extracted based on the trip start time:

- *Month of the trip time*: Winter and early spring in Chicago are not suitable for bike riding. To denote the month of the trip start time, we define feature x_4 in the paper, where $x_4 = 1$ if the trip is at January, $x_4 = 2$ if it is at February, and so forth.
- *Weekday of the trip time*: For subscribed users and customers, they have totally different bike usage patterns on different weekdays. To utilize this information, a new feature x_5 is introduced. We set $x_5 = 0$ if the trip starts on Sunday, and set $x_5 = 1$ for Monday, and so forth.
- *Hour of the trip time*: Another time-related feature extracted is the specific hour of the start time, as the start hour can show the purpose of the trip a lot. For simplicity, we divide each day into 24 hours and define another feature x_6 to represent the specific trip start hour, where $x_6 = 0$ if it starts within [12AM, 1AM); $x_6 = 1$ if at [1AM, 2AM); and so forth.

These 3 extracted features show the information about trip start time in 3 different cyclic patterns.

C. Features about the stations

In the individuals' spatial travel pattern analysis section (available in [14]), we have shown the some top frequently commuted station pairs by different categories of users. Therefore, the trip origin station can provide important information to help us infer the destination station as well. Three different features about the stations are extracted in the experiments:

- *Station Pairs*: In the spatial travel pattern analysis section (available in [14]), we show that some station pairs can be frequently traveled by the users. The first station features extracted for (s_o, s_d) is the station ID pairs, i.e., $x_7 = ID((s_o), ID(s_d))$.
- *Station geographic information*: Besides the ID information, we also have the geographic information about the stations, which can be represented as the (latitude, longitude) pairs. The coordinate pairs are also used as a feature, which can be represented as $x_8 = \text{latitude}(s_o), \text{longitude}(s_o), \text{latitude}(s_d), \text{longitude}(s_d)$.
- *Geographic Distance*: The majority of Divvy bike trips are of length $0.5 - 5$KM and trips that are too short (around 0KM) or too long (longer than 10KM) are very rare. We propose to extract feature x_9 to denote geographic distance between stations (s_o, s_d), where *Manhattan Distance* is used as the distance measure.

Based on the above extracted features, we can represent the feature vector for certain station pairs (s_o, s_d) as $\mathbf{x}(s_o, s_d) = [x_1, x_2, \cdots, x_9]$ of length 13 in total (as x_7 and x_8 are of lengths 2 and 4 respectively), which together with the label $y(s_o, s_d)$ and time $t(s_o, s_d)$ can be used to build the confidence score and trip duration prediction models.

D. Trip Destination Station Inference Model

For the trip destination station prediction problem, we propose to map it to a binary classification in the experiments. For each trip origin and destination station pairs (e.g., (s_o, s_d)), we assign it with different labels $\{1, 0\}$ to denote whether a certain trip starting at s_o will end at s_d or not. To address the problem, we propose to apply a state-of-the-art pairwise based regression algorithm, namely MART (Multiple Additive Regression Trees) [13], to develop a regression function. MART is based on the stochastic gradient boosting approach described in [3], [4] which performs gradient descent optimization in the functional space. In our experiments, we used the log-likelihood as the loss function, steepest-descent (gradient descent) as the optimization technique, and binary decision trees as the fitting function. For more information about the MART model, please refer to [3], [4], [11].

E. Trip Duration Inference Model

To predict the time length of the trip, the same set of features are applied to build the trip duration inference model. Different regression models can be used as the base prediction model, and, without a loss of generality, we will apply the Lasso regression model as the base regression model in this paper, which fits a linear equation $\hat{t}(s_o, s_d) = \sum_{i=1}^{k} b_i x_i + b_0$, where $\hat{t}(s_o, s_d)$ is the inferred trip length between stations s_o and s_d, term b_i denotes the coefficient of feature x_i and b_0 represents the bias term.

To get the coefficient values in training the model, Lasso uses the L_1 prior as the regularizer, and the optimal coefficients can be learned by solving the following equation

$$\arg\min_{\mathbf{b}} \sum_{(s_o, s_d)} (\hat{t}(s_o, s_d) - t(s_o, s_d))^2 + \alpha \left|\mathbf{b}\right|_1,$$

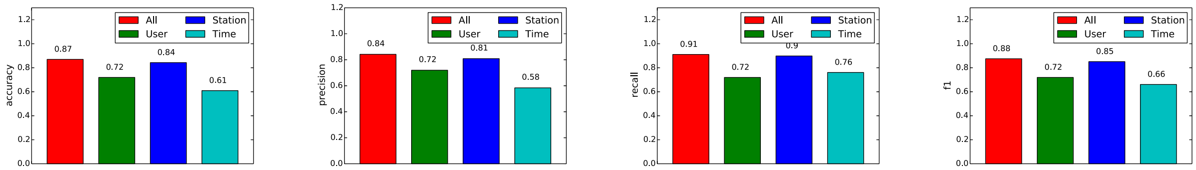
where $t(s_o, s_d)$ is the real duration of the trip between s_o and s_d and α is the weight of the regularizer term.

V. Experiments

To test the effectiveness of these two models in addressing the trip prediction problem, we conduct experiments on the real-world bicycle-sharing system Divvy (introduced in Section II). In this section, we will first introduce the experiment settings, which include the experiment setups, comparison methods and evaluation metrics. Next, we will show the experiment results and give detailed analysis.

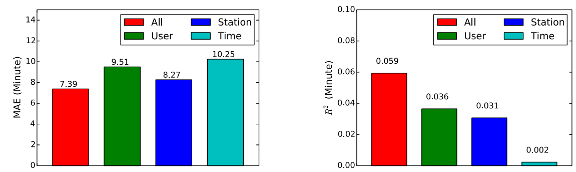
A. Experiment Settings

1) Experiment Setups: From the dataset, we extract the trip tuples (user, origin station, destination station, departure time) as the existing trip set, where each tuple contains the complete information about the trip. In the trip destination station prediction problem, the existing trip set is used as the positive



(a) Accuracy (b) Precision (c) Recall (d) F1

Fig. 2. Trip destination station prediction results evaluated by different metrics.



(a) MAE (b) R^2

Fig. 3. Trip duration inference results.

set (i.e., assigned with labels 1), and a equal-sized negative trip tuple set is random generated, instances in which are assigned with labels 0. In the negative tuple set, (1) users can be "customers" and "subscribers" of equal chance, and the gender of "subscribers" is randomly assigned with either "male" and "female", whose ages are random selected from $\{1, 2, \cdots, 100\}$; (2) the origin and destination are randomly selected from the whole station set; and (3) the negative trip departure time is random selected from July $1st$, 2013 to June $30th$, 2015. Both positive and negative trip sets are divided into two parts according to ratio $4:1$ based on the time order, where 4 folds are used as the training set and 1 fold is used as the test set. A set of features are extracted for each instance in the training and test sets. We train the trip end prediction model MART with the training set, which will be applied to the test set to infer the labels of the test pairs.

Meanwhile, in the trip duration inference problem, similarly, we divide the existing trip set into two parts according to ratio $4:1$ based on the time order, where 4 folds are used as the training set and 1 fold is used as the test set. However, the setting of trip duration inference is slightly different: (1) no negative trip set is needed; and (2) the instances in the training and test set are assigned with their real-trip duration as their labels. The same set of features are extracted to build the trip duration inference model (i.e., Lasso) based on the training set, which will be applied to infer the trip time duration of instances in the test set.

2) Comparison Methods: The comparison methods used in trip destination and duration inference can be divided into two categories depending on the information used:

Models using all information

- ALL: Method ALL builds the trip end prediction and trip duration inference models with all the three categories of features extracted, which include *user*, *station* and *time* based features.

Models using partial information

- USER: Method USER builds the trip end prediction and trip duration inference models with the features about *users* only.
- STATION: Method STATION builds the trip end prediction and trip duration inference models with the features about *stations* only.
- TIME: Method TIME only uses the features about the *time* only to build the trip end prediction and trip duration inference models.

3) Evaluation Metrics: To evaluate the performance of these different methods in addressing the trip prediction problem, we apply different evaluation metrics to measure their prediction results.

We formulate the trip end prediction problem as a binary classification problem, and all these 4 comparison methods can output the predicted labels of trip pairs in the test set. By comparing them with the ground-truth labels, we can evaluate their performance with 4 frequently used metrics: Accuracy, Precision, Recall and F1-score.

We formulate the trip duration inference problem as a regression problem, and the comparison methods will output the inferred the time duration of trips in the test set. Meanwhile, we also have the real-world trip duration from the dataset, i.e., the ground-truth. Different metrics used for regression problems can be applied here, and we use the MAE (Mean Absolute Error) and R^2 (i.e., Coefficient of Determination) as the evaluation metrics.

B. Experiment Results

The experiment results are available in Figure 2 and Figure 3. Figure 2 show the results of trip end prediction and Figure 3 gives the results of trip duration inference.

By comparing ALL with the other methods in Figure 2, we can observe that ALL can outperform other methods with significant advantages consistently evaluated by different metrics. For instance, in Figure 2(a), the Accuracy achieved by ALL is 0.87 which is about 21% higher than the Accuracy gained by USER (i.e., 0.72); 4% higher than the Accuracy score achieved by STATION (i.e., 0.84); and 43% larger than the Accuracy score obtained by TIME (i.e., 0.61). Similar results can be observed in Figures 2(b)-2(d), where Precision, Recall and F1 are used as the evaluation metrics.

Generally, among the comparison methods, ALL utilizing all these 3 categories of features perform the best. Among the 3 methods using one category of feature only, STATION can outperform USER, while USER performs better than TIME. It is also easy to understand, as the task is to infer the trip destination station, historical trip station pair information can only provide more direct information for addressing the task.

Meanwhile, the features about users and trip start time can provide the indirect hints, as they are about the bike user and time, not directly about the stations.

In Figure 3(a)-3(b), we show the results about trip duration inference problem, which are evaluated by both MAE and R^2 metrics. Compared with other comparison methods, ALL achieves better performance with the much smaller MAE and larger R^2 score. For example, the MAE introduced by ALL is 7.39 (minute), which is 22.23% lower that the MAE introduced by USER, 10.64% smaller than the MAE introduced by STATION and 28% lower than the MAE introduced by TIME. For the R^2 metric, the R^2 score achieved by ALL is 0.059, which is nearly the double of the R^2 scores gained by USER and STATION. The advantages of ALL against TIME is more obvious: the MAE of ALL accounts for only 72% of that achieved by TIME; and the R^2 score obtained by ALL is as large as the 30 times of the R^2 achieved by TIME.

Therefore, by utilizing the complete information available about the trips, ALL can outperform other comparison methods with significant advantages in both predicting the trip destination stations and inferring the trip duration time.

VI. Related Work

Bicycle-sharing has received increasing attention in recent years with initiatives to increase cycle usage improve the first mile/last mile connection to other modes of transit, and lessen the environmental impacts of our transport activities. DeMaio gives a complete introduction about the history, impacts, models of provision, and future of bicycle-sharing systems in [1]. Midgley provides a complete overview work about the bicycle-sharing schemes, management, policies, and challenges as well as opportunities in [8]. A large number of other review and case-study works on bicycle-sharing systems have appeared so far [7], [15], [12], [2], which study the bicycle-sharing systems from different aspects and directions.

Recently, urban computing has become a hot research area and lots of works have been done by Zheng et al. already [16], [9], [5]. The bicycle-sharing systems are an important part in urban computing. Many research works have been done on bicycle-sharing systems and other transportation systems to study the system design problem [6], load balance problem [10], and bicycle traffic prediction problem [5]. Lin et al. [6] introduce a strategic design problem for bicycle sharing systems incorporating bicycle stock considerations, which is formulated as a hub location inventory model. The problem studied in [6] covers the design work about various aspects of the bicycle-sharing system, e.g., the number and locations of bicycle stations, the creation of bicycle lanes, the selection of paths, etc. Pavone et al. develop methods for maximizing the throughput of a mobility-on-demand urban transportation system and introduce a rebalancing policy that minimizes the number of vehicles performing rebalancing trips [10]. The optimal rebalancing policy can be found as the solution to a linear program effectively in the proposed model. Li et al. propose a hierarchical prediction model to predict the number of bikes that will be rent from/returned in a future period for bicycle-sharing systems [5], which focus more on the macroscopic bike traffic flow in the bicycle-sharing system and is different from the microscopic trip destination and duration prediction problem of a specific trip studied in this paper.

VII. Conclusion

In this paper, we have studied the trip prediction problem for bicycle-sharing systems to infer the potential trip destination station and trip duration. Extensive analysis about the user composition of a real-world bicycle-sharing system, individuals' temporal bike usage behavior patterns and spatial bike usage behavior patterns have been done. Based on the analysis results, two new regression based inference models have been introduced in this paper to predict the potential trip destination station and trip duration respectively. Experiments conducted on the real-world bicycle-sharing system dataset demonstrate the effectiveness of the proposed model.

VIII. Acknowledgement

This work is supported in part by NSF through grants III-1526499, and CNS-1115234, and Google Research Award.

What's more, this research was partially supported by the grant from the Natural Science Foundation of China (No. 61303017), the Natural Science Foundation of Hebei Province (No. F2014210068).

References

[1] P. DeMaio. Bike-sharing: History, impacts, models of provision, and future. *Journal of Public Transportation*, 2009.

[2] C. Etienne and O. Latifa. Model-based count series clustering for bike sharing system usage mining: A case study with the vÉlib system of paris. *ACM Transaction Intelligent Systems and Technology*, 2014.

[3] J. Friedman. Greedy function approximation: A gradient boosting machine. *Annals of Statistics*, 2000.

[4] J. Friedman. Stochastic gradient boosting. *Computational Statistics & Data Analysis*, 2002.

[5] Y. Li, Y. Zheng, H. Zhang, and L. Chen. Traffic prediction in a bike-sharing system. In *SIGSPATIAL*, 2015.

[6] J. Lin, T. Yang, and Y. Chang. A hub location inventory model for bicycle sharing system design: Formulation and solution. *Computers and Industrial Engineering*, 2013.

[7] P. Midgley. The role of smart bike-sharing systems in urban mobility. *Land Transport Authority JOURNEY*, 2009.

[8] P. Midgley. Bicycle-sharing schemes: Enhancing sustainable mobility in urban areas. *United Nations Department of Economic and Social Affairs*, 2011.

[9] B. Pan, Y. Zheng, D. Wilkie, and C. Shahabi. Crowd sensing of traffic anomalies based on human mobility and social media. In *SIGSPATIAL*, 2013.

[10] M. Pavone, S. Smith, E. Frazzoli, and D. Rus. Load balancing for mobility-on-demand systems. *International Journal of Robotics Research*, 2012.

[11] S. Tyree, K. Weinberger, K. Agrawal, and J. Paykin. Parallel boosted regression trees for web search ranking. In *WWW*, 2011.

[12] P. Vogel and D. Mattfeld. Strategic and operational planning of bike-sharing systems by data mining a case study. *Computational Logistics*, 2015.

[13] Q. Wu, C. Burges, K. Svore, and J. Gao. Adapting boosting for information retrieval measures. *Information Retrieval*, 2010.

[14] J. Zhang, X. Pan, M. Li, and P. Yu. Bicycle-sharing system analysis and trip prediction. *CoRR*, 2016.

[15] L. Zhang, J. Zhang, Z. Duan, and D. Bryde. Sustainable bike-sharing systems: Characteristics and commonalities across cases in urban china. *Journal of Cleaner Production*, 2015.

[16] Y. Zheng, X. Yi, M. Li, R. Li, Z. Shan, E. Chang, and T. Li. Forecasting fine-grained air quality based on big data. In *KDD*, 2015.