Bike Sharing Station Placement Leveraging Heterogeneous Urban Open Data

Longbiao Chen[1,2], Daqing Zhang[2,4], Gang Pan[1], Xiaojuan Ma[3], Dingqi Yang[2,6], Kostadin Kushlev[5], Wangsheng Zhang[1], Shijian Li[1]

[1]Zhejiang University, China; [2]Institut Mines-Télécom, Télécom SudParis, CNRS SAMOVAR, France
[3]Hong Kong University of Science and Technology, Hong Kong; [4]Peking University, China
[5]University of British Columbia, Canada; [6]University of Fribourg, Switzerland
[1]{longbiaochen,gpan}@zju.edu.cn; [2]daqing.zhang@telecom-sudparis.eu

ABSTRACT

Bike sharing systems have been deployed in many cities to promote green transportation and a healthy lifestyle. One of the key factors for maximizing the utility of such systems is placing bike stations at locations that can best meet users' trip demand. Traditionally, urban planners rely on dedicated surveys to understand the local bike trip demand, which is costly in time and labor, especially when they need to compare many possible places. In this paper, we formulate the bike station placement issue as a bike trip demand prediction problem. We propose a semi-supervised feature selection method to extract customized features from the highly variant, heterogeneous urban open data to predict bike trip demand. Evaluation using real-world open data from Washington, D.C. and Hangzhou shows that our method can be applied to different cities to effectively recommend places with higher potential bike trip demand for placing future bike stations.

Author Keywords

Open data; urban computing; bike sharing system

INTRODUCTION

In recent years, an increasing number of cities have introduced bike sharing programs to promote environmental sustainability and encourage a healthy lifestyle [1, 2]. Such bike sharing programs allow people to pick up and drop off public bikes at self-service stations to make short trips within a city. Given the large investment in infrastructure necessary to support a bike sharing program, such as arranging parking facilities and making the roads more bike friendly, it is essential for urban planners to maximize the utility of public bikes. One of the key factors for promoting citizen participation in a bike sharing program is placing bike stations at locations that can best meet the trip demand of potential users [3].

Traditionally, urban planners use surveys to collect information on local *bike trip demand (BTD)* to guide bike station placement [4, 5]. Although existing literature has identified a large set of factors that may affect bike trip demand in general [1, 6, 7], each city has its own environmental, social, and cultural characteristics, resulting in different adoption patterns of bike sharing programs. Therefore, it is necessary for urban planners to understand the needs and tastes locally. Each time urban planners want to extend the bike sharing program to a new area in the city, they send out investigators to conduct a user survey on site. However, this approach consumes a great amount of time, labor, and money, especially when planners need to compare a large number of possible places.

The increasing availability of heterogeneous, fine-grained *urban open data* provides the opportunity to inexpensively assess bike trip demand across a city [8]. For instance, areas with popular restaurants as determined by Foursquare check-in number might generate high bike trip demand. In this paper, we propose a data-driven approach to predict bike trip demand to assist bike station placement. However, due to the considerable volume and variety of urban open data, it is not straightforward to directly select representative features related to bike trip demand. Therefore, we identify the most relevant datasets from a large pool of urban open data sources based on prior knowledge, and extract customized features to characterize bike station utilization in individual cities. Then, we feed these features into predictive models to rank the potential of locations for placing future bike stations. The main contributions of this paper include:

1. A novel use case of the heterogeneous urban open data, namely bike sharing station placement.
2. A semi-supervised feature selection method to extract customized (city-specific) features from the highly variant, heterogeneous urban open data for bike trip demand prediction. First, by exploiting prior knowledge from bike sharing program surveys, we identify three key factors related to bike trip demand from the corresponding open data sources: area functions from online map services, human activity from location-based social networks, and demographics from open government data. We then construct a set of customized features for each city by selecting the most informative dimensions of each factor via correlation analysis, rather than adopting a static feature selection method regardless of city context.
3. A performance evaluation using real-world bike sharing system data in two cities (Washington, D.C., USA and Hangzhou, China). The results show that our method

UbiComp '15, September 7–11, 2015, Osaka, Japan.

http://dx.doi.org/10.1145/2750858.2804291

can be applied to different cities to effectively recommend places with higher potential bike trip demand for placing future bike stations.

RELATED WORK

Researchers have conducted extensive surveys on the bike sharing programs deployed in various cities to examine the history [7], facilities [1], user characteristics [4], etc. In order to guide the design and implementation of bike sharing programs, urban planners usually conduct dedicated biker-oriented surveys of potential users to understand the actual demand in their cities [5, 4]. Based on these mobility surveys, Garcia et al. proposed a method to estimate the spatial distribution of bike trip demand [3]. In general, however, such surveys are time-consuming and expensive to conduct.

Recently, many researchers have resorted to urban open data to address challenges in urban development. Such data are usually easy-accessible and free of cost [9]. Examples include sensing city dynamics from bike sharing system data [10], inferring air quality [11] and urban noise [12] based on public air monitoring data and 311 data, and evaluating container port performance of a harbor city leveraging ship GPS traces and maritime open data [13]. In this paper, we present a first attempt to address the bike sharing station placement problem with urban open data. The work closest to ours is presented in [14], where the optimal retail store location is recommended only using Foursquare check-in data for New York City. Our work is different from [14] in terms of data source, data fusion approach, and application domain.

FACTOR ANALYSIS

To understand what factors may impact the bike trip demand in cities with bike sharing programs, researchers and urban planners have conducted a series of surveys [2, 3, 4, 5]. These surveys have shown that the following factors are key in determining the bike trip demand of an area:

1. *Area function* [3]. High bike trip demand often occurs in or near residential areas, transition hubs, and tourist attractions, but relatively less in industrial areas.
2. *Human activity* [2, 4]. People tend rent a bike for certain activities, such as commuting, shopping, entertainment, and personal errands.
3. *Demographics* [4, 5]. Bike sharing system user community tends to be considerably younger, highly educated, and relatively less affluent.

Taking the above results as prior knowledge, we first select a set of relevant urban open datasets to these factors, and then analyze the most informative dimensions of each factor via correlation analysis.

Data Selection

The usage data of many bike sharing programs are publicly accessible [2], such as the Capital Bikeshare System [15] of Washington, D.C. and the Public Bicycle System of Hangzhou [16]. We define the *daily utilization* of a bike station as the average number of bike rentals and returns per day. In the following analysis, we use the daily utilization of a bike station as the proxy[1] for BTD in its *service area*, which is a circular area around the station [3].

Table 1: Top-10 POI categories most relevant to bike station utilization with the corresponding correlation coefficient.

	Washington, D.C.	Hangzhou
1	Cafe and Bakery (0.53)	Residential area (0.65)
2	Bar and Restaurant (0.52)	Vegetable market (0.57)
3	Hotel and Hostel (0.49)	Hospital (0.55)
4	Work Place (0.45)	KTV (Karaoke) (0.51)
5	Residential area (0.38)	Hotel and Hostel (0.49)
6	Retail store (0.35)	Retail store (0.45)
7	Bank and ATM (0.34)	Work place (0.41)
8	Law firm (0.32)	Bar and Restaurant (0.38)
9	Gym and Yoga (0.31)	Hair salon and Spa (0.31)
10	Museum and Gallery (0.25)	Movie theater (0.29)

We identify the following open datasets to characterize the above-mentioned factors.

1. *Point of Interest (POI) dataset*. POI distributions have been used to describe area functions [17, 18]. For instance, an area where a large number of retail stores are located has a high probability to be a business area. We retrieve POI data using the Google Places API [19].
2. *Check-in dataset*. User check-ins in Location Based Social Networks (LBSNs) can serve as an indicator of human activities [20]. For instance, check-ins at a restaurant are likely to associate with dining activities. We retrieve check-in data using the Foursquare API [21], and calculate the *daily check-in number* for an area over a period of time.
3. *Demographics dataset*. Demographics data of an area, such as the median household income, median age, and education level, come from the open data catalogs of government portal [22].

Area Functions and Bike Trip Demand

We characterize an area's functions by the categorical distribution of POIs, and analyze the correlation between area functions and BTD. Specifically, for each bike station, we first retrieve all POIs within its service area, and group them into a set of POI categories (e.g. restaurants). For each category, we then compute the Spearman's correlation coefficient [23] between the POI number and the daily utilization across all the bike stations to measure their monotonic relationship.

Table 1 shows the top-10 POI categories most relevant to the daily utilization of all bike stations in Washington, D.C. and Hangzhou. Such results align well with surveys regarding bike trip origins and destinations in these two cities [4, 6]. Note that some POI categories in the top-10 list, such as *vegetable market* and *KTV (Karaoke)*, are unique to Hangzhou, and others to Washington, D.C. Such variability indicates the need to construct customized features for different cities.

Human Activity and Bike Trip Demand

We take LBSN check-ins as the semantic proxy of human activities. We first retrieve all check-ins within the service area of each bike station and map them to a pre-defined list of human activities (in this paper we adopt the activity list in the

[1]We note that due to the limit of station capacity and bike availability, the actual demand might be larger than the observation.

Table 2: Top-10 human activities most relevant to bike station utilization with the corresponding correlation coefficient.

	Washington, D.C.	Hangzhou
1	Transit (Bus, Metro) (0.58)	Dinner and Meal (0.49)
2	Entertainment (Bar) (0.57)	Shopping (Food) (0.45)
3	Dinner and Meal (0.54)	Sight-seeing (0.41)
4	Personal errands (0.53)	Meeting (Tea house) (0.38)
5	Shopping (Clothes) (0.51)	Entertainment (Karaoke) (0.36)
6	Exercise (Gym, Yoga) (0.44)	Personal errands (0.36)
7	Visiting (Museum) (0.42)	At school (0.31)
8	At work (0.39)	Transit (Bus) (0.29)
9	At home (0.26)	At work (0.24)
10	At school (0.24)	At home (0.21)

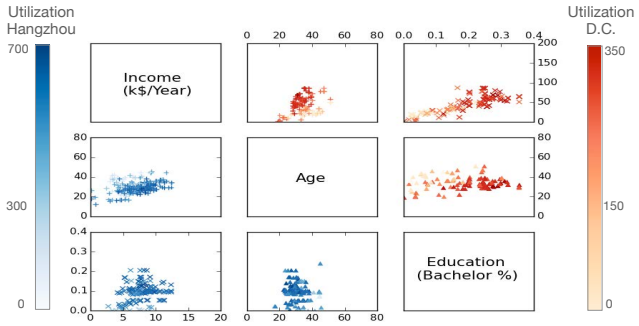


Figure 1: Correlation matrix between bike station utilization and demographics features in Washington, D.C. (top-right) and Hangzhou (bottom-left), respectively.

2013 Capital Bikeshare survey questionnaire [4]). We then use the daily check-in number to approximate the frequency of each activity. Finally, we determine the most pertinent activity types to BTD by comparing the Spearman's correlation coefficient between activity frequency and utilization.

Table 2 shows the top-10 human activities most relevant to the daily utilization of all bike stations in Washington, D.C. and Hangzhou; these activities are consistent with the bike trip purposes reported in surveys [4, 6]. As with Table 1, some activities are unique to a specific city, such as *meeting at tea house* and *entertainment at Karaoke* in Hangzhou.

Demographics and Bike Trip Demand

We validate the correlation between demographic factors and BTD. Specifically, for each bike station, we map its service area to a census tract [24] and retrieve the corresponding *median household income*, *median age*, and *education level (bachelor's degree percentage)* information.

Figure 1 illustrates the correlation matrix [11] in Washington, D.C. and Hangzhou, where each row/column denotes one demographic factor and each data point represents the daily utilization of a bike station. We can see that stations with higher utilization (i.e. darker points) are usually located in neighborhoods with younger population, moderate income level, and/or higher education level. These findings are consistent with the bike sharing system user surveys [4, 6], and provide insights on how neighborhood demographics affect bike sharing system adoption.

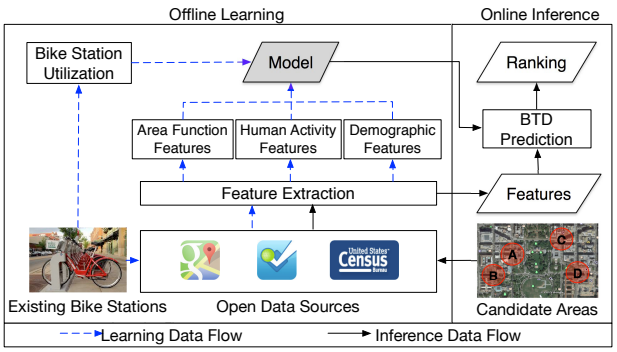


Figure 2: Overview of the framework.

FRAMEWORK

We propose a two-phase framework to determine optimal placement of bike sharing stations leveraging urban open data, as shown in Figure 2. In the *offline learning phase*, we first identify representative data sources that correspond to the factors critical to BTD as determined by the bike sharing program surveys in a target city. We then extract city-specific features for each factor to learn a model for BTD prediction. In the *online inference phase*, we extract the same set of features for each candidate area and feed them into the learned model to predict its potential BTD. The candidate areas with higher potential BTD are considered to be better locations for placing future bike stations.

Feature Extraction

We extract a set of city-specific features for each factor based on the analysis in the previous section.

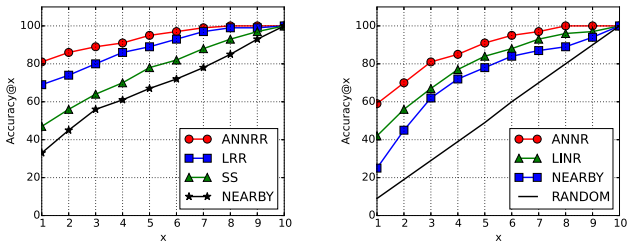
1. *Area Function Feature* F_f. We select the top-j POI categories most relevant to BTD for a specific city, i.e., $F_f = (f_1, f_2, ..., f_j)$. Each element of the feature denotes the significance of a function, which is the total number of POIs of the corresponding category.
2. *Human Activity Feature* F_h. We select the top-k human activities most relevant to BTD for the city, i.e., $F_h = (h_1, h_2, ..., h_k)$. Each element of the feature corresponds to the intensity of a human activity, which is the total number of daily check-ins of the corresponding type.
3. *Demographics Feature* F_d. We construct the demographics feature vector $F_d = (d_i, d_a, d_e)$ for the city, where d_i corresponds to the median household income, d_a corresponds to the median age, and d_e corresponds to the education level (bachelor's degree percentage).

Model Selection

We need to select models that are capable of incorporating heterogeneous features to effectively predict the potential BTD of candidate areas and rank them. We adopt the *regression-and-ranking* methodologies. Specifically, we first train a supervised regression model with existing bike station utilization, and then rank the candidates according to the predicted BTD using that model. We compare two supervised learning algorithms in our evaluation, i.e., Linear Regression-and-Ranking (LRR) and Artificial Neural Network Regression-and-Ranking (ANNRR) [25].

Table 3: Summary of datasets from two cities.

	Washington, D.C.	Hangzhou
Data collection period	2010–2013	2011–2012
Bike stations	203	2,115
POIs	16,520	145,119
Check-ins	17,356,179	1,553,354
Census Tracts	181	882



(a) Washington, D.C. **(b) Hangzhou**

Figure 3: $Accuracy@x$ of the baselines and proposed methods.

EVALUATION

Experiment Settings

We collect datasets about bike station utilization and the relevant factors from Washington, D.C., USA and Hangzhou, China, as summarized in Table 3.

Baseline Methods

We use the following two baseline methods in comparison with the proposed LRR and ANNRR algorithms. (1) *Nearby Station Average (NEARBY)*, which uses the average utilization of four nearby bike stations to estimate the BTD of the candidate areas. (2) *Single Data Source Static Model (SS)*. We adapt the methodology of [14] to extract a set of static features from only Foursquare check-in data for predicting bike trip demands in different cities. The major features include neighborhood density and diversity modeled by check venue types, and area popularity modeled by check-in numbers.

Parameter Settings

We use a service area radius $r = 200m$ for Washington, D.C., and $r = 250m$ for Hangzhou, respectively, based on the corresponding surveys [4, 6]. To obtain optimal features of top-k POI categories and top-j check-in types, we experimentally select $k = j = 10$ for Washington, D.C., and $k = j = 15$ for Hangzhou, as Hangzhou is more diverse in terms of city functions and human activities. We repeat our experiments 1,000 times in both cities. For each experiment, we randomly select 10 stations as candidate areas and use the rest for training in the offline learning phase.

Evaluation Metrics

We compare the ranking results with the ideal ranking list $\bar{R}$, where candidates are sorted by their actual daily bike utilization. We evaluate the performance with the following two metrics frequently used in information retrieval:

1. To assess the *accuracy of the top recommendation*, we use $Accuracy@x (1 \leq x \leq 10)$, which measures the frequency the top recommendation is appearing among the top-x of $\bar{R}$.

Table 4: $nDCG@3$ of the baseline and proposed methods.

	Washington, D.C.	Hangzhou
NEARBY	0.47	0.38
SS	0.63	0.51
LRR	0.77	0.75
ANNRR	0.86	0.85

2. To further assess the *quality of the overall ranking* in the top-k recommendations, we adopt the top-k Normalized Discounted Cumulative Gain ($nDCG@k$) metric [26]:

$$nDCG@k = \frac{DCG@k}{IDCG@k}, DCG@k = \sum_{i=1}^{k} \frac{2^{rel_i} - 1}{log_2(i+1)}$$

where $IDCG@k$ is the $DCG@k$ value of $\bar{R}$, and rel_i is the relevance score of item i. To calculate rel_i, we use the position of i in $\bar{R}$, i.e., $rel_i = |\bar{R}| - p_i + 1$, where p_i is the position of item i ($1 \leq p_i \leq |\bar{R}|$).

Evaluation Results

Results on $Accuracy@x$ and $nDCG@3$ are shown in Figure 3 and Table 4, respectively. Our methods (ANNRR and LRR) outperform the two baseline methods in terms of top recommendation accuracy and overall ranking quality in both cities. Specifically, ANNRR is better than LRR in both measures, achieving more than 80% accuracy of top recommendation and over 0.85 $nDCG@3$ for both cities. In contrast, the baseline methods (SS and NEARBY) do not work as well in Hangzhou as in Washington, D.C., implying that the two cities have different characteristics in urban development and human flow. In summary, our method effectively adapt to the specific contexts of individual cities by applying semi-supervised customization of dynamic features.

We compare the top recommendation accuracy between AM hours (6:00 - 10:00) and PM hours (15:00 - 21:00) and find out that the accuracy is higher in PM hours (84% vs. 71% in Washington, D.C.). One possible reason is that most check-ins are performed in PM hours [27], and thus characterizing the human activities more accurately in PM hours. We also validate that our method runs efficiently on different scales of cities. The running time (including a training and 1,000 times of prediction) of ANNR on a server equipped with an Intel Xeon CPU is 4.3 seconds for the middle-scale bike sharing system in Washington, D.C., and 12.5 seconds for the world's largest bike sharing system in Hangzhou.

CONCLUSION

In this paper, we leverage open data to predict bike trip demand and recommend optimal placement of bike sharing stations. We propose a two-phase feature selection method to extract customized features from heterogeneous urban open data for bike trip demand prediction. The evaluation results show that our semi-supervised method outperforms the state-of-the-art baseline approaches on recommending locations for optimal bike station placement. Specifically, customized factor identification and feature selection based on city characteristics can achieve consistent performance when applying to heterogeneous urban open data in different cities.

ACKNOWLEDGMENT

We would like to thank the shepherd and the reviewers for their constructive suggestions. Jérémie Jakubowicz, Hua Lu, Chenwei Zhang and Leye Wang contribute useful comments and inputs to this paper. This research was supported by the Program for New Century Excellent Talents in University (NCET-13-0521) and Zhejiang Provincial Natural Science Foundation of China (LR15F020001). This work was done when Longbiao Chen was visiting Institut Mines-Télécom; CNRS, France.

REFERENCES

1. J. Pucher, J. Dill, and S. Handy, "Infrastructure, programs, and policies to increase bicycling: An international review," *Preventive Medicine*, vol. 50, pp. 106–125, 2010.
2. S. Shaheen, S. Guzman, and H. Zhang, "Bikesharing in Europe, the Americas, and Asia," *Transportation Research Record: Journal of the Transportation Research Board*, vol. 2143, no. 1, pp. 159–167, 2010.
3. J. C. Garcia-Palomares, J. Gutierrez, and M. Latorre, "Optimizing the location of stations in bike-sharing programs: A GIS approach," *Applied Geography*, vol. 35, no. 1–2, pp. 235–246, 2012.
4. LDA Consulting, *2013 Capital Bikeshare Member Survey Report*, Washington, D.C., 2013.
5. A. M. Burden, R. Barth, and others, *Bike-Share Opportunities in New York City*. New York: New York Department of City Planning, 2009.
6. S. Shaheen, H. Zhang, E. Martin, and S. Guzman, "China's Hangzhou public bicycle," *Transportation Research Record: Journal of the Transportation Research Board*, vol. 2247, no. 5, pp. 33–41, 2011.
7. P. DeMaio, "Bike-sharing: History, impacts, models of provision, and future," *Journal of Public Transportation*, vol. 12, no. 4, pp. 41–56, 2009.
8. D. Zhang, B. Guo, and Z. Yu, "The emergence of social and community intelligence," *Computer*, vol. 44, no. 7, pp. 21–28, 2011.
9. B. Ubaldi, "Open Government Data," *OECD Working Papers on Public Governance*, vol. 22, no. 1, pp. 1–61, 2013.
10. J. Froehlich, J. Neumann, and N. Oliver, "Sensing and Predicting the Pulse of the City through Shared Bicycling." in *Proc. IJCAI'09*, vol. 9, pp. 1420–1426.
11. Y. Zheng, F. Liu, and H.-P. Hsieh, "U-Air: When urban air quality inference meets big data," in *Proc. KDD'13*, pp. 1436–1444.
12. Y. Zheng, T. Liu, Y. Wang, Y. Zhu, and E. Chang, "Diagnosing New York City's Noises with Ubiquitous Data," in *Proc. UbiComp'14*, pp. 715–725.
13. L. Chen, D. Zhang, G. Pan, L. Wang, X. Ma, C. Chen, and S. Li, "Container throughput estimation leveraging ship GPS traces and open data," in *Proc. UbiComp'14*, pp. 847–851.
14. D. Karamshuk, A. Noulas, S. Scellato, V. Nicosia, and C. Mascolo, "Geo-spotting: Mining online location-based services for optimal retail store placement," in *Proc. KDD'13*, pp. 793–801.
15. Capital Bikeshare, "Washington, D.C. Capital Bikeshare System," 2015. http://www.capitalbikeshare.com/
16. Hangzhou Online, "Hangzhou Public Bicycle System," 2015. http://www.hangzhou.com.cn/hzbike/
17. P. Jensen, "Network-based predictions of retail store commercial categories and optimal locations," *Physical Review E*, vol. 74, no. 3, pp. 1–4, 2006.
18. J. Yuan, Y. Zheng, and X. Xie, "Discovering regions of different functions in a city using human mobility and POIs," in *Proc. KDD'12*, pp. 186–194.
19. Google Inc., "Google Places API," 2015. https://developers.google.com/places/
20. D. Yang, D. Zhang, V. Zheng, and Z. Yu, "Modeling User Activity Preference by Leveraging User Spatial Temporal Characteristics in LBSNs," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 45, no. 1, pp. 129–142, 2015.
21. Foursquare Inc., "Foursquare API," 2015. https://developer.foursquare.com/
22. U.S. Government, "U.S. Government's Open Data Portal," 2015. http://www.data.gov/
23. J. H. Zar, "Significance testing of the Spearman rank correlation coefficient," *Journal of the American Statistical Association*, vol. 67, no. 339, pp. 578–580, 1972.
24. The U.S. Census Bureau, "Census Tracts and Block Numbering Areas," 2015. https://www.census.gov/geo/reference/gtc/gtc_ct.html
25. D. F. Specht, "A general regression neural network," *IEEE Transactions on Neural Networks*, vol. 2, no. 6, pp. 568–576, 1991.
26. K. Jarvelin and J. Kekalainen, "Cumulated gain-based evaluation of IR techniques," *ACM Transaction on Information Systems*, vol. 20, no. 4, pp. 422–446, 2002.
27. E. Malmi, T. M. T. Do, and D. Gatica-Perez, "From Foursquare to My Square: Learning Check-in Behavior from Multiple Sources." in *Proc. ICWSM'13*.