Optimizing Hubway Bicycle Docks Allocation in Boston City

Congyang Wang, Fangyu Lin, Meng Li
Department of Computer Science, Department of Data Science – Worcester Polytechnic Institute, MA, 01609
Email: {cwang8, flin, mli6}@wpi.edu

***Abstract*—As the number of people living in Boston grows, the traffic environment has become more and more crowded. In recent years, local government has already built up the bicycle system to ameliorate traffic conditions and create a green environment. The idea is to design and implement a bicycle system that is accessible and convenient to the public by planning station locations satisfying the public needs. Residents can rent a bike from one Hubway station and return it at another Hubway station which is closed to their destinations. However, it is important and challenging for the government to arrange the number of bicycle docks between stations appropriately so that the system can benefit almost everyone and balance traffic flows in Boston as much as possible. Our approach is to arrange bicycle docks for each station such that maximizes the coverage of the bicycle system and increase penetration of bicycle usage, based on the Hubway shared bike data in Boston Area between 2011 and 2013. Fortunately, the total number of capacity for each Hubway station in Boston area is known, we are going to estimate available bikes from historical Hubway bike trips data, which contains information on stations and every trip. Our goal is to optimize the number of docks for each Hubway station to satisfy the demands and supplies, and at the same time to mitigate traffic congestion as well as to create a green environment in Boston Area.**

***Index Terms*—Bike-sharing program; Bike-station location; Location Optimize model;**

I. Introduction

Sustainable living and energy conservation have been populous across the globe in recent decades since now people are aware of severe damage on environment inflicted by industrialization. To promote renewable energy, a significant number of developed countries launched a shared bicycle program in the hope that it will facilitate people to switch to green transportation. The Hubway is such program launched by Boston government in 2011 with 61 stations and 600 bicycles initially. By the end of 2013, the system has been expanded to 130 stations and 1200 bicycles. However, the current Hubway system is designed mainly for entertainment purpose. Obviously, the bicycle system has far more benefits than just for fun. It even has high potentials to solve urban problems of oversized and overpopulated cities, for instance, traffic congestion, air pollution, etc.

To explore more potentials usage of Hubway Shared Bike, a new bicycle facility plan has to be designed based on trip demands, regardless of trip purposes. As a consequence, the new system will be more efficient and attractive compared to the current one as well as produce more beneficial outcomes. In the first place, reasonable bicycle facility planning will increase the usability of the system and correspondingly reduce usage of private automobiles and emission of carbon dioxide. Plus, the decline in consumption of fossil fuels will help to conserve the non-renewable resources. Secondly, traffic congestion will be mitigated if people switch to bicycles from automobiles. For low-income residents, the Hubway system can provide them an alternative and healthy commute travel solution. For college students, cycling is also a cost-effective and flexible option which better fits their school schedules compared to the subway. Additionally, an efficient bicycle system is likely to ease mass transit pressures for a metropolis.

However, optimizing the entire design, including location planning of stations and re-allocation of docks and bicycles, requires multiple data sources, such as trip records and GIS information, and is way too large for a one-month project. To tailor it to meet workload that we can handle in half semester, the problem is simplified to optimize station capacities based on trip demands and current design. In other words, our goal is to modify current design with the least efforts and costs, focusing on re-allocation of docks for stations so that the Hubway system can satisfy public demands as much as possible. To minimize costs of optimizing the re-allocation of docks, the total number of docks is kept untouched. To tackle the problem, we will use a dataset of the Boston Hubway 2011 Trips And Stations. Figure[1] shows the Hubway Station in the Boston Area.

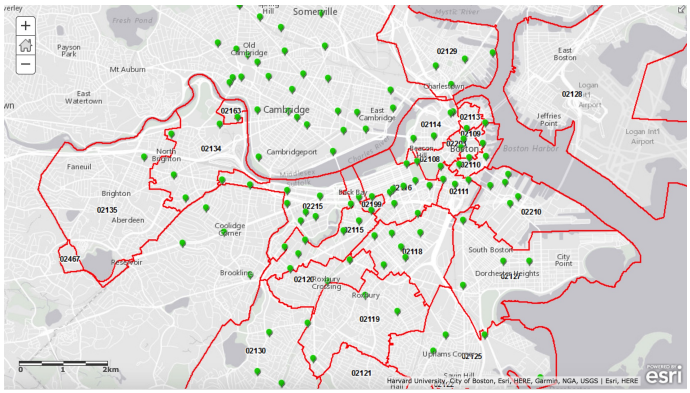

Fig. 1: Hubway Station Location In Boston Area

According to the ”Wisconsin Bicycle Planning Guidance Handbook (WBPG)” [2], it is necessary to take both supplies and demands of shared bicycles into consideration in the shared bicycle facility planning. In the paper, we proposed our methodology based on the demands and supplies from the Hubway Bike Trips Data in Boston Area between 2011 and 2013. The rest of paper is organized as follows. Section 2 is the related work in recent years, which talks about what other research projects are doing by using the Hubway Bike Trips Data. Section 3 is going to introduce the Methodology we are proposed, and how can we predict and optimize the docks at each Hubway Stations grouped by Regions. Section 4 Evaluation is going to use the real Hubway Bike Trips data to test our methodology and compare the results with real data. And the last section 5 is the Conclusion for this paper, what we have observed by the Evaluation Experiment results.

II. Related Work

The research paper we have reviewed currently is the one [3] that analyzes the bike rental system using the P-Median (minimize-impedance solution), which is to minimize ”weighted costs between

demand points and solution facilities" to maximize coverage. Moreover, in the paper, it uses the "GIS-based multi-criteria analysis tools." As we know, this is a popular tool for data visualization on a map creating a heat map rendering with other features. However, the tool is inadequate and flawed in certain aspects considering our problem. So we are going to apply a more advanced method. To estimate the capacity of each bike station, first, aggregate the number of bicycles rented from one station on a daily basis. Then select the maximum over daily total from each station as the estimated capacity. Regarding accessibility and regional population, some stations will have more bicycles that are rented than others. Based on the estimated needs, we can make a right decision to allocate bikes to stations to maximize overall usage of bicycles. Besides, we are reviewing other methodologies, such as multi-criteria decision analysis and exploratory spatial data analysis. We will make comparisons across them to figure out the best one for our problem and evaluate performance by comparing it with baseline algorithms.

III. Methodology

A. Problem Setting

Technically, we intend to estimate the number of docks installed at each station that can satisfy demands of the station. To achieve the goal, we propose a two-step framework:

- Step 1: Predict the number of available bicycles at each station, i.e., the number of bicycles allocated to each station in the very beginning, which meets trip demands of the station.
- Step 2: Estimate capacity of each station to accommodate predicted the number of available bicycles, keeping the total number of docks of the entire system unchanged.

Intuitively, the sum of available bicycles at one station and bicycles that have traveled to it should meet the bicycle demands of the station. On the other hand, each station also has to provide sufficient empty docks to accommodate bicycles that have traveled to the station from other stations. Thus, the number of available bicycles and the capacity of each station should be predicted subject to these constraints.

The two-step framework is essentially developed to create a dock re-allocation plan taking trip demands into consideration. Since the number of docks at each station is estimated by the number of available bicycles which is predicted by trip demands, the station capacity estimates take trip demands into account spontaneously. In the first step, bicycle demands are predicted with inbound and outbound trips. In the second step, the Great Boston Area is split into 35 regions by zipcodes. Docks are assigned to each region first corresponding to its bicycle demands. For example, suppose the city has two regions A and B. Region A accounts for 20% bicycle demands of the total demands and region B accounts for 80%. Then 20% docks of the total will be allocated to the region A and the remaining 80% docks will be allocated to the region B. Within each region, docks are allocated to stations following the same procedure.

To make notations in our methods simple and clear, following terms are defined.

$Outbound_s$: the number of bicycles outbound from the station s.

$Inbound_s$: the number of inbound bicycles to the station s.

Avl_s: the number of available bicycles at the station s in the beginning.

$Capacity_s$: the capacity of station s, i.e., the total number of docks at the station s.

To simplify the problems and optimize allocation of bicycles and docks with the lowest costs and the least efforts, we made several assumptions.

- The total numbers of docks of the bike-sharing system is fixed and equal to the total of the current Hubway system throughout the study.
- Given trip data with start time and end time, regional population is assumed to be redundant for trip demand forecasts and will not be considered in the study.

B. Bike Demand Prediction For Stations

Since demands of bicycles are different from station to station, a strategy that distributes bicycles uniformly across all stations as adopted by most of bike-sharing systems is unreasonable. It may create a surplus of bicycles for stations with low demands and a shortage of bicycles for stations with high demands. To balance bicycle supplies and demands across stations, the number of available bicycles in theory should satisfy the equation below,

$$Avl_s + Inbound_s \geq Outbound_s$$

C. Estimation Of Dock Allocation

Since capacity and quantity of available bicycles at each station of the current Hubway System are known, and the totals of docks and bikes are assumed unchanging, the total of docks and bicycles across the city can be obtained.

With predicted bike demands of individual stations, a demand ratio of bikes at a given region r can be calculated as,

$$Ratio_r^{Bike} = \frac{Predicted\ Avl_r}{\sum_{i=1}^{n} Predicted\ Avl_i}$$

where n is the number of regions in the Great Boston Area.

For region r, dock allocation should follow

$$Estimated\ Capacity_r = Ratio_r^{Bike} * \sum_{i=1}^{n} Capacity_i$$

where $Capacity_i$ are data of the current system.

For a specific region r, dock allocation for a station s located in the region is estimated as,

$$Ratio_s^{Bike} = \frac{Predicted\ Avl_s}{\sum_{i=1}^{m} Predicted\ Avl_i}$$

where m is the number of stations in the region r.

$$Estimated\ Capacity_s = Ratio_s^{Bike} * Capacity_r$$

where $Capacity_r$ is the number of docks in region r estimated previously.

D. Data

The analysis is conducted with three datasets: hubway_stations.csv, hubway_trips.csv, and the most important and the largest data file, stationstatus.csv. Key attributes are summarized below,

Attribute	Data Type	Description Of Usage
seq_id (sid)	Character (Primary key)	trip index, counting trips
hubway_id	Character (Reference key)	Classification for station count sum
start_station	Integer	Demand-based counter
end_station	Integer	Supply-based counter
start_date	Timestamp	Classification for day, week, month, year
end_date	Timestamp	Classification for day, week, month, year
zip_code	Character	Classification for regions

IV. Experiments

A. Data Preparation

First of all, for the data we are using, which is the Hubway bikes shared trip data in Boston. We assume that the user information is not useful in this paper. It is necessary to simplify the data file before doing the data analysis. We are going to remove the column with attribute 'subs_ctype', 'zip_code', 'birth_data', and 'gender'. Load the data file into a database for further calculation and data analysis.

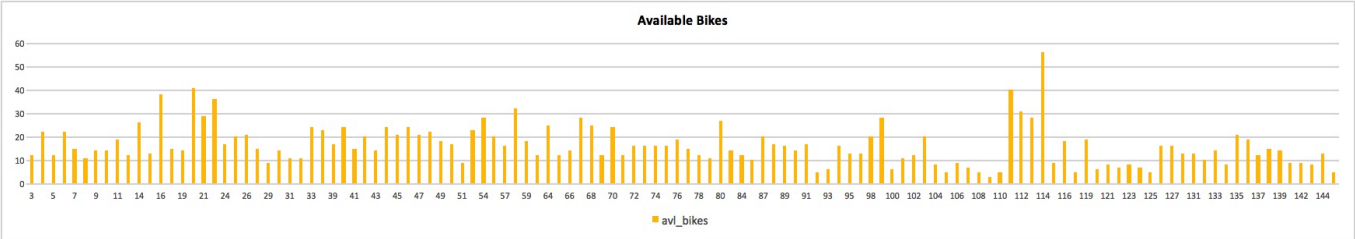

Fig. 2: Docks Re-allocation Result Plot

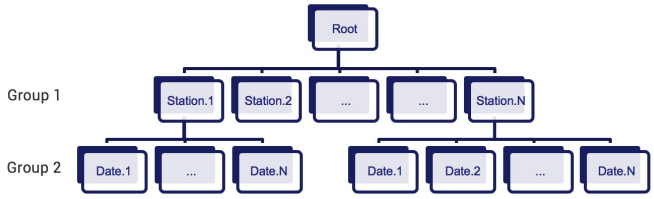

Fig. 3: Tree Level Group Structure

Second, the other datafile that we are going to use has the Hubway Station information, which contains the removed bike stations. In this case, we assume that removed Hubway station is no longer available, so we use a filter to remove those stations data before processing. Furthermore, we add the zipcode as another attribute into this Hubway Station location file because we need it to do the region classification.

Third, the last datafile that we used has information of docks number, with the station_ID and zipcode. We extract a couple of useful columns from this file to do the optimization, which is to group by regions (zipcode) and re-allocate docks for each station in the same region. The detail analysis will show in the subsection (c).

B. Prediction of Available Bikes

In this section, we use the methodology to calculate the available number of bikes at each station. First, group by the station number as well as group by the start_date for one table. Second, group by the station number as well as the start end_date for another table. The Two level Tree group structure is shown in the Figure[3]. Third, use the methodology function which we have mentioned to join the two tables with same station number and subtract the end_date to the start_date. We saw this result as a two level classification. First level is classified by station number. Second level, based on each station, which is classified by date. And the result we got have positive and negative numbers. It is very clear that the positive number represent the leftover bikes on that date at that station. On the other hand, the negative number means the minimum number of bikes available at the docks at the beginning of each data at each station. So, we assume the positive number has enough supplies, we set those positive data to zero. However, for the positive data, we take the absolute value which represent the available bikes at the docks each day at different stations.

$$Avl_s + Inbound_s \geq Outbound_s$$

The result table has three attributes, which are station_ID, Date, and the number of bikes needs at the station on each date. Next, we find the max number of the date available bikes for each station, which means that gets the biggest number through all dates in one station. On the other word, group by station and looking for the maximum number,

zipcode	id	terminal	station	municipal	lat	lng	status	avl_bikes	docks	assign_docks	differ
02127	107	C32017	South Boston Library - 6	Boston	42.336	-71.04	Existing	7	15	7	-8
02119	123	E32004	Egleston Square at Colun	Boston	42.316	-71.1	Existing	8	15	8	-7
02130	121	E32002	JP Centre - Centre Street	Boston	42.313	-71.11	Existing	8	15	8	-7
02138	89	M32020	Harvard Law School at M	Cambridge	42.379	-71.12	Existing	16	19	17	-2
02116	42	D32007	Boylston St. at Arlington	Boston	42.352	-71.07	Existing	20	23	21	-2
02138	130	M32021	Harvard University Gund	Cambridge	42.376	-71.11	Existing	13	15	14	-1
02111	81	D32019	Boylston St / Washingtor	Boston	42.352	-71.06	Existing	14	15	15	0
						...	...				
02115	46	B32005	Christian Science Plaza	Boston	42.344	-71.09	Existing	24	19	26	7
02210	64	C32010	Congress / Sleeper	Boston	42.351	-71.05	Existing	25	19	27	8

Fig. 4: Hubway Docks Re-allocation Results

which is the absolute value from previous data processing. The result histogram plot table is shown in Figure[2].

C. Hubway Docks Re-allocation

In this section, we are using the predicted available number of bikes in the previous section to re-allocate the bike dots for each station. We import another data file "stationstatus.csv" which contains the docks numbers for 145 stations. Second, we join the previous table results by the same primary key and foreign key, which are station_ID number. Next, we use the methodology function from the previous section to do a ratio of each station in a same group of the region, which is group by zipcode. The result table is shown in the figure[4].

$$Estimated\ Capacity_r = Ratio_r^{Bike} * \sum_{i=1}^{n} Capacity_i$$

$$Ratio_s^{Bike} = \frac{Predicted\ Avl_s}{\sum_{i=1}^{m} Predicted\ Avl_i}$$

Second, use the methodology to re-allocate each station, but the sum of total stations in each region is constant. Third, compare the new docks number in each station. The final re-allocation results are ployed as histogram and scatters in the figure[6]. The histogram is sorted by the re-allocate docks values, which is from negative to positive. From this graph, we can find out some stations have good dock numbers, but some have to re-allocate to satisfied the requirement of a minimum of demands.

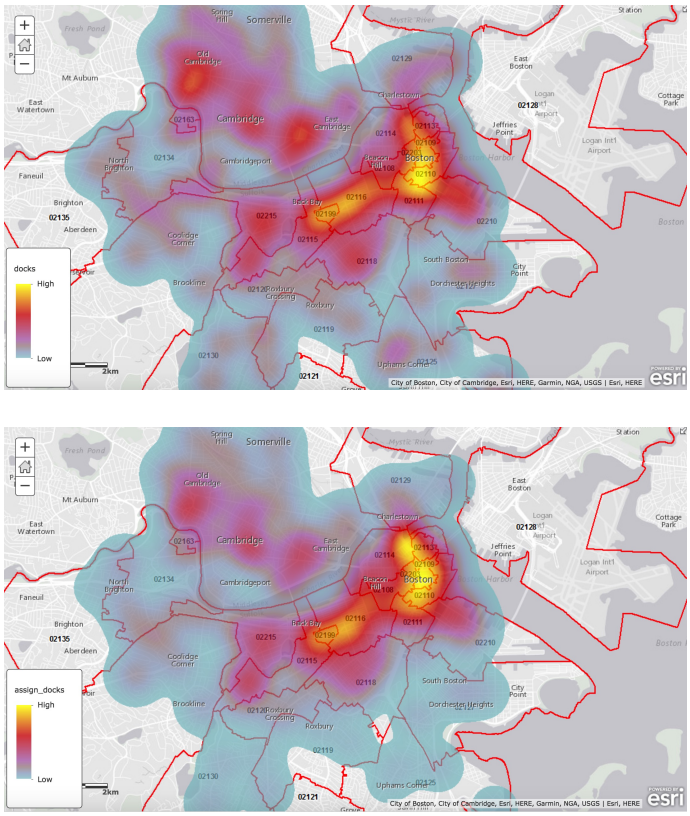

Fig. 5: Heat Map of Docks Re-allocation Result

D. Result Analysis

According to the results in each plot, our design matches the real situation in Boston Area. For example, the minimum Docks Required to be added at "TD Garden - Causeway at Portal Park #1" is at least 33 docks. As it is known that TD Garden is a multi-purpose arena which can be used as a Basketball Field as well as a ice rink and for other entertainment. Also, it is a commuter rail point, connecting inter-city trains and the subway. As a consequence, trip demands at TD Garden are ought to be high, which is consistent with our optimization results: the minimum 33 docks have to be installed. On the other hand, there are some area has redundant docks which is wasting the resources. For example, the Hubway Station at "Jackson Square T at Center St" has about 14 empty docks. As we

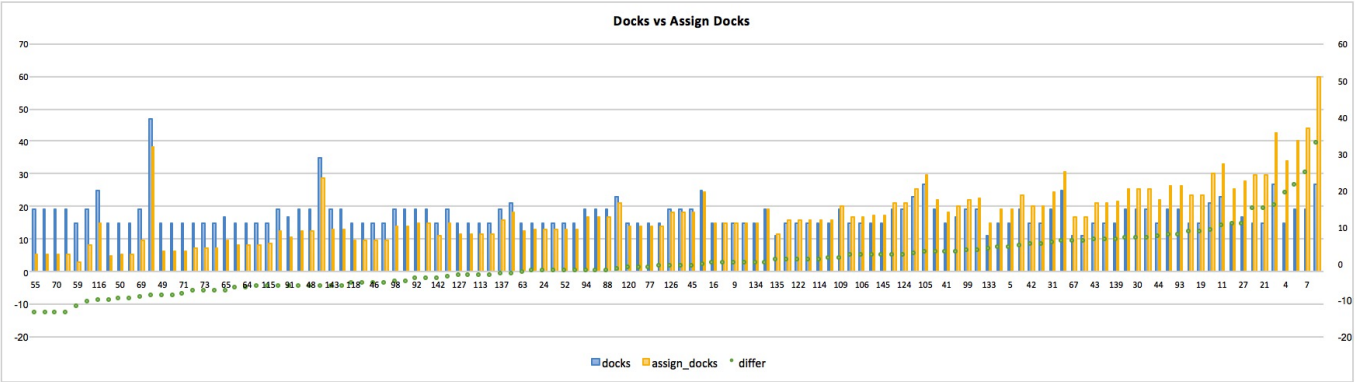

Fig. 6: Docks Re-allocation Result Plot

known from the map that this Hubway Bike Station is associated with Subway Station. However, from the result we know that people may not want use the Hubway bikes here because it is too close for a walking distance to the activities area. In this case, we plan to re-allocate those 14 docks to other Hubway stations. Those are two extreme results which show the correct re-allocation results from our experiment. The figure[5] shows the Heat Map of Docks Re-allocation Result in Boston Area. As we can see directly from the Map that most docks are re-allocated to Boston Downtown Area. Our goal to optimize Hubway Docks for each Station has been achieved.

V. Conclusions

The goal of this paper is to predict the shared bike needs and re-allocate the Hubway Docks in each station to satisfy the minimum requirement of the demands for that stations by day. The data we used in the paper is from the Hubway Shared Bikes in Boston from 2011 to 2013. We develop our Methodology to predict the numbers of bikes in each station needs to satisfy the minimum demand for a day. The evaluation results are shown as expected that many Hubway Docks required re-allocation due to a different ratio between ”Needs” and ”Docks.” For the future work, it may be a good practice to make the Docks motivative. It means that each dock can be installed and uninstalled easily in real practice. Use the real Hubway Shared Bike data in Boston, and we can re-allocate the docks dynamically each Months. And this is also an open challenge to use the real-time bike sharing date to re-allocate the docks for each station dynamically in the future. It is quite essential to predict trip demands, since station capacity is estimated based on the needs. Trip demands, intuitively, are correlated with demographic factors of residents if the information is available. Primarily, regional population has direct impacts on trip demands generation, since large population normally suggests high demands. Secondly, private automobile penetration rate is also related to the bicycle demands, because people who own automobiles are less likely to rent bicycles than people without automobiles. Also, since cycling is popular among the young people, regions with a high proportion of the young may demand more bikes compared to other regions. In addition, weather is another critical factor for trip demands prediction, since people barely cycle during severe weather conditions, such as rain and

snow. To make effective use of the information, one may apply feature extraction techniques to extract latent features from demographic factors and devise prediction models, like random forests and time series models when non-sparse data are supplied, for future endeavors.

REFERENCES

[1] J. C. Garca-Palomares, J. Gutirrez, and M. Latorre, ”Optimizing the location of stations in bike-sharing programs: A GIS approach,” Applied Geography, vol. 35, no. 1, pp. 235-246, 2012.

[2] Tom Huber, ”WISCONSIN BICYCLE PLANNING GUIDANCE” Unpublished, 2003.

[3] N. Baird and D. Kim, ”Bicycle Facility Demand Analysis using GIS,” Spaces and Flows: An International Journal of Urban and ExtraUrban Studies, vol. 1, no. 2, pp. 1-14, 2011.

[4] G. Rybarczyk and C. Wu, ”Bicycle facility planning using GIS and multi-criteria decision analysis,” Applied Geography, vol. 30, no. 2, pp. 282-293, Apr. 2010.

[5] I. Frade and A. Ribeiro, ”Bicycle Sharing Systems Demand,” Procedia - Social and Behavioral Sciences, vol. 111, pp. 518-527, Feb. 2014.

[6] WinNT, Bamboo Bicycles Beijing David Wang, howpublished = https://www.bamboobicyclesboston.org/,Accessed: 2017-02-20