BIKE DISTRIBUTION MODEL FOR URBAN DATA APPLICATIONS

MARICICA NISTOR AND ANDRÉ DIAS
CEiiA/Centre of Engineering and Product Development, Portugal

ABSTRACT

Bike sharing systems are fundamental sources of data for creating applications of monitoring the city and guiding the user's choice for bike usage. Although many related works analyse the generated data by these urban bike systems with the scope of finding bike usage patterns, the variety of the cities and the lack of impact factors require further deeper investigations. We propose a simple, but efficient mathematical approach based on a Markovian model to predict the bike distribution for an urban sharing bike system considering the weather and event impacts. The model is applied for data collected from the New York city bike system. The main findings are relevant for the urban applications and are summarized as follows: (a) the model results substantially address the city's characteristics, i.e., for the New York city, in terms of weather, only the temperature influences the bike usage, while regarding the events, the impact is insignificant, (b) the hourly bike distribution is predicted 1 day-ahead that is of particular interest to the city manager and (c) to the user who is able to know 1 day in advance the probability of finding an available bike or a free parking space at a specific station. Further city comparison analysis in terms of traffic, vehicle utilization and population density is provided for future purposes. Finding the precise station's capacity is a forthcoming feature of the proposed model.
Keywords: data applications, mathematical modelling, urban bike distribution

1 INTRODUCTION

The mobility in the cities constantly changes once with the digitalization and the appearance of sustainable actions promoted by the United Nations goals [1]. Among the main modalities of transportation, bike utilization is gaining very much attraction mainly because of the simplicity and easiness of travelling in crowded cities as well as the economic impact for the users. Many city councils have already approved the installation of small to large bike sharing systems and they incentive people to use the bike as a mean of transportation instead of cars primarily because of the traffic jam and the CO_2 emission reduction. Data generated by the bike systems is of particular interest to both the city managers and the users for a large range of urban applications. Part of these applications is related to traffic and CO_2 reduction, as well as optimizing public services and several business models. Understanding the user behaviour within the city is equally vital for the city. Progressively cities make available their bike data and, thus, more research works have been publishing their results and the first insights. In particular, we select some of the works related to the bike sharing system data analysis. Basically, the literature review contains various models with respect to the bike prediction or the demand analysis, as shown in Table 1. The previous works aim to find the demand prediction (including the check-in and check-out) and consider the bike data from several cities. Moreover, the main impact factors are related to time and weather and very few consider the event impact.

ISSN: 2058-8305 (paper format), ISSN: 2058-8313 (online), http://www.witpress.com/journals
DOI: 10.2495/TDI-V3-N1-67-78

Table 1: Literature review.

Work	Objectives	Model	Factors	City
[2]	Demand analysis	Quantification (O-D pairs)	Traffic, trip, slopes	Coimbra
[3]	Predict bike availability in stations	Queuing theoretical time-inhomogeneous Markovian model	Time	Paris
[4]	Demand (check-in/out) prediction in stations and neighbourhoods	Log-log regression model	Taxi usage, weather, spatial variable	New York
[5]	Demand (check-in/out) prediction in stations	Graph convolutional neural network	Time, day, location	New York
[6]	Demand (check-in/out) prediction in stations	Decision tree random forest Adaboost	Weather, time location, day, week	Seattle
[7]	Demand (check-in/out) prediction in stations	Random forest, pruning	Time, bike availability, weather, location, events	Hanghzou
[8]	Demand (check-in/out) prediction in stations in clusters	Weighted correlation network, Monte Carlo, label propagation	Time, weather, social events, traffic, user	New York, Washington
[9]	Trip prediction	Regression models	User, location, time, trip	Chicago

Since most of the works are limited to the demand prediction and the impacts, there is room for improvements and new metrics analysis. Thus, our main contributions are enumerated as follows:

- *Bike distribution*: reveals the city hourly bike distribution, considering also the weather and events impact, i.e., sport games, concerts. The bike distribution is predicted for 1 day-ahead using a simple Markovian approach. This metric is mainly attractive to the city managers.
- *Probability to find at least one bike available in a specific station*: creates generally advantages for the users that aim to find a bike available at the bike parking station. Also, probability to find free space in a station is as well needed for the users in order to not spend too much time looking around for a park space.
- *City comparison*: provides insights on city features to consider for the prediction bike model.

The rest of the work is organized as follows. Section 2 provides the input data analysis and Section 3 presents the model description, while Section 4 the numerical results. The city comparison analysis is provided in Section 5 and the work concludes with Section 6.

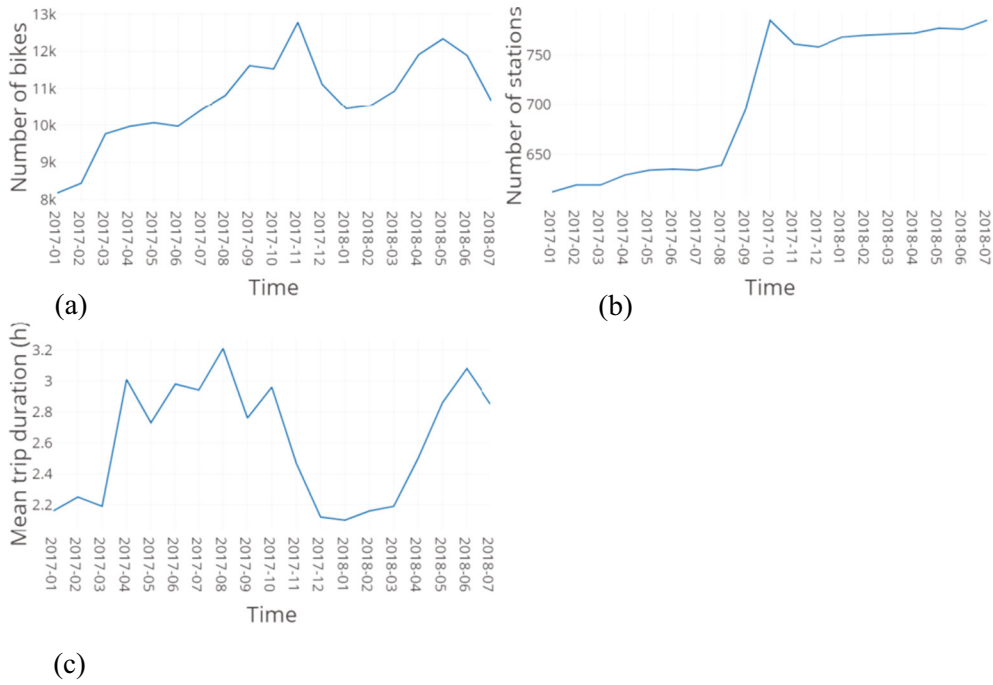


Figure 1: Bike data collected from January 2017 until July 2018. (a) Number of bikes used; (b) Number of stations used; (c) Average trip duration (h).

2 URBAN DATA SETS

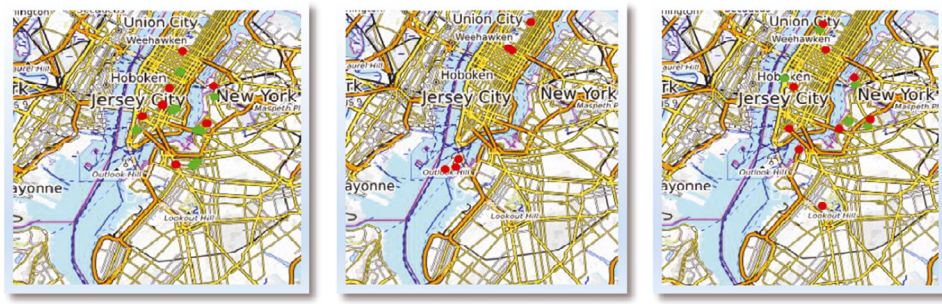
The data collected for the analysis can be divided in (a) bike data, (b) weather data and (c) event data. All data sets are collected from 1st of January, 2017 until 31st of July, 2018. The details for each data set is provided in the following.

2.1 Bike Data

The bike data set has been downloaded from Citi Bike [10]. The bike system in New York city covers the areas of Manhattan, Brooklyn and Jersey City. The consistency of data is not always accurate as it seems that some fails in the system may cause interruption of data record.

The data available for the analysis is given by the transitions between two stations. The data format is recorded as the trip duration, the start and end time, the start and end station, i.e., identification and location, the bike identification, and the user gender and age. Hence, we can process the number of bikes, and stations per day. Fig. 1 shows the number of bikes and stations used and the average trip duration during the mentioned period. The bike system increases significantly from 2017 in terms of number of stations and bikes. Curiously, the average trip duration is around 2 h during the cold day and more than 2.5 h during the warm days, most probably because people keep the bike close while having stops.

We observe some patterns in the bike data usage. For instance, the top ten common routes, i.e., most used stations for check-in and check-out, are calculated during the periods of morning, afternoon and night. The periods of morning, afternoon, nights mean the



(a) (b) (c)

Figure 2: Top ten common routes during morning, afternoon, night (green points = check-in, red points = check-out). (a) Morning; (b) Afternoon; (c) Night.

following time intervals: morning between 04:00 am and 12:00 pm, afternoon between 12:00 pm and 08:00 pm, and night between 08:00 pm and 04:00 am. The location of the most common stations is shown in Fig. 2, where during morning the most frequented stations are used between 245 and 332 times, in the afternoon between 482 and 826 times with the same check-in and check-out stations, while during the night between 85 and 169 times. One station located in the south of Central Park, named Central Park S & 6 Ave, is common for all periods. This analysis reveals the most common transitions between Manhattan and Brooklyn during morning and night and between Manhattan stations and Governors Island during afternoon.

2.2 Weather

The weather data is based on the online service provider named Weather Underground [11]. We consider two metrics for the bike usage pattern, i.e., the daily cumulative trip duration and the daily number of bikes used per day. We analyse the dependence between these two variables in a daily basis manner. The correlation and R-squared results are shown in Table 2 and the match between the weather conditions versus the trip duration and bike usage in Fig. 3. As it can be seen, the temperature is a key factor for the bike usage pattern choice, while the rain and the wind have a weak relationship with the bike usage and trip duration for the New York city.

Table 2: Correlation coefficient/R-squared between the duration of trips/number of bikes and the weather implications.

Data	Temperature	Precipitation	Wind
Trip duration	0.79/0.63	–0.13/0.016	–0.16/0.025
Number of bikes	0.77/0.6	–0.19/0.038	–0.16/0.025

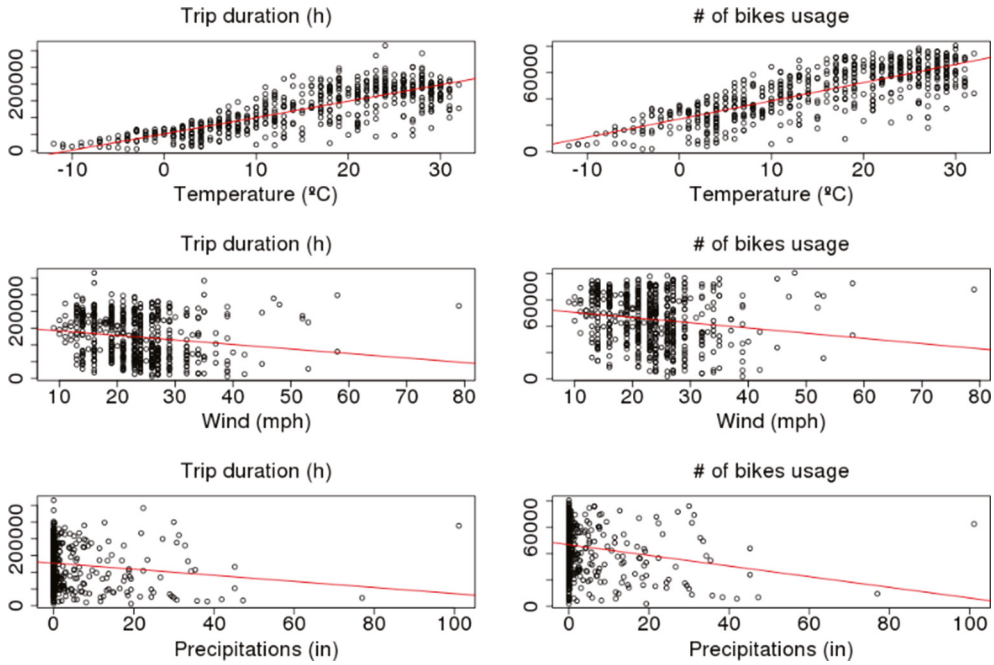
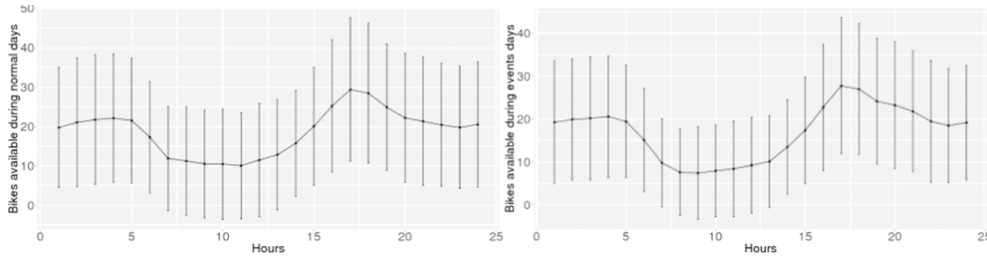


Figure 3: Weather impact for data collected from January 2017 until July 2018.

2.3 Events

To better understand the impact of the events in bike usage system, we focus on the events that occur in Madison Square Garden since the arena capacity is significant, i.e., around of 20,000 seats. Along the analysed period, from January 2017 until July 2018, we collect 88 concerts, 64 basketball games and 72 ice hockey games which represents more than 200 days out of 577 total number of days for the selected period [12–14]. The bike stations evaluated are the ones closest to the Madison Square Garden, as Fig. 4 shows, namely 8 Ave & W 33

Figure 4: Stations closest to the Madison Square Garden.



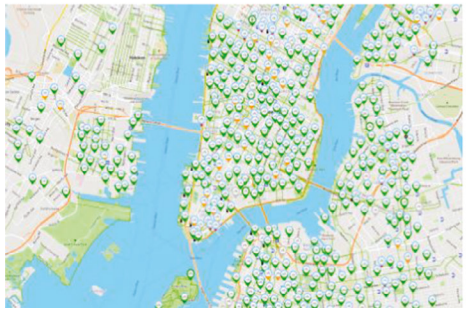
(a) (b)

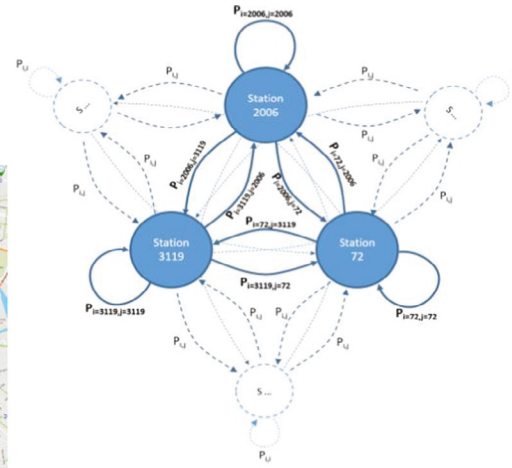
Figure 5: Normal day vs. Event day of the bikes available in Madison Square Garden. (a) Normal day; (b) Event day.

St, 8 Ave & W 31 St, W 31 St & 7 Ave and W 33 St & 7 Ave. Around 70% of the events started between 07:00 pm and 08:00 pm and the number of attendances was around 19,000 for the basketball games and more than 13,000 for the concerts. For the ice hockey games the attendance is not available, but the arena has a capacity of 18,006. Fig. 5 shows the average bike available in these stations during a normal day and an event day. As it can be seen, since New York is an uninterrupted city with various events every day, the event impact in this area is not significant.

3 MODEL DESCRIPTION

At the heart of our model, a Markovian approach is applied where the representation is provided in Fig. 6. In the left image, we show the map with part of the bike sharing system in New York. The bike stations have fixed geographic location and the bikes are moving between the stations. This representation was the criterion for the Markov model exemplified in right figure, where the states are represented by the stations and the transitions between





(a) (b)

Figure 6: Markov model description. (a) Part of the Citi bike system; (b) Part of Markov Model.

the states are described by the bike transitions between the stations. Here, for instance, we pick three bike stations, i.e., ID 2006, ID 3119, ID 72, and represent part of the bike transitions between them and the rest of the stations from the system. In the following we describe in detail the Markov model and the insights on the complexity of the proposed model.

3.1 States

The states are defined by the identification of the bike stations, as represented in Fig. 6, e.g., *Station i* and *Station j*, where the number of states varies as defined by $i,j = \{0,..,S\}$, with S meaning the total number of stations. Thus, the number of states is given by the number of bike stations.

3.2 Probability

Two probabilities are available for each state. $P_{i,j}$ means the probability to transit from *Station i* to *Station j*, while the $P_{i,i}$ represents the probability to transit in the same state, where $i,j = \{0,..,S\}$.

3.3 Transitions

The probability to transit from *Station i* to *Station j* is given by

$$P_{i,j} = \frac{N_{i,j}}{N_i}, \tag{1}$$

where $N_{i,j}$ represents the number of bikes that transits from *Station i* to *Station j* and $N_i = \sum_{j=0}^{S} N_{i,j}$ represents the total number of bikes that transits from *Station i* to any other stations, with $i = \{0,..,S\}$. In particular, if $i = j$ then $P_{i,j} = 1$ means that no bike transits from the bike *Station i* to the other stations and $P_{i,i} = 0$ means no bike exists at the bike *Station i*. Thus, one row from the transition matrix translates to the probability of one station to transit to the rest of the bike stations from the system. The size of the transition matrix is dictated by the number of bike stations.

3.4 Prediction Model

The mathematical model previously described performs the transition matrix which is hourly calculated given the bike data set, i.e., for each hour a transition matrix is valid. All transitions between two stations are counted. Since 1 day-ahead is predicted, 24 transition matrices are performed.

The following steps are considered for the prediction model:

- Calculate the initial bike distribution for the desired starting hour, as summarized in Algorithm 1. The initial distribution is performed starting from the initial hour and looking backward at the transitions from the past data sets. The bikes are identified for each transition and, depending on the trip duration between each two stations, counted to the corresponded bike stations. The algorithm stops when all the available bikes are allocated

to the stations. Using sampling distribution, the initial bike distribution is calculated for the initial hour.
- Calculate iteratively the bike distribution among the stations using the sampling distribution for the rest of the hours, as summarized in Algorithm 2.
- The weather impact is reflected in the transition of the bikes. Only temperature data is considered for the model since it is the parameter that mostly influences the bike usage. In particular, we assume that the temperature affects equally all the New York regions, e.g., the bike stations. Linear regression is calculated between the trip duration and the future temperature for the predicted day. We consider the differences in probability variation between the future temperature and the average one. This variation is then multiplied to the bikes distribution. Future temperatures higher than the maximum temperature or lower than the minimum temperature are limited by the maximum and minimum temperatures, respectively.
- Even though the events have insignificant impact for the New York bike model, the events can be introduced by considering the ratio of the bike usage during an event day to the bike usage during a normal day.

Algorithm 1 Initial bike distribution: bike distribution
Data: bike data set, number of bikes, initial hour, initial hourly transition matrix, station IDs, station capacity
Result: initial bike distribution
for each transition starting from initial hour and going backward in data set **do**
  if check if all bikes allocated **then**
    leave the loop
  end if
  if check if start transition time < initial hour **then**
    bikes at stations: count bikes on start station
  else
    bikes at stations: count bikes on end station
  end if
end for
initial bike distribution: sample distribution (stations IDs and capacity, bikes at stations, transition matrix for initial hour)

Algorithm 2 Prediction model: bike distribution
Data: hourly transition matrices, station IDs, station capacity, initial bike distribution
Result: bike distribution
initial bike distribution
for each hour (starting from the initial distribution) **do**
  bike distribution: sample distribution (stations IDs and capacity, bike distribution in previous hour, hourly transition matrix)
end for

3.5 Computational Complexity

The complexity for the proposed model depends on the number of states of the model, which is given by the number of bike stations, and the number of transition matrices. Although the New York bike system is extensive, the number of bike stations reaches in 2018 almost 800

stations, which in turn means a reasonable size for the transition matrix. We compute hourly the transition matrices, which translates to 24 transition matrices. Alternatively, for simplicity and instead of hourly transition matrices, the transition matrices can be computed for defined periods of time, as follows.

- Define the three transition matrices which correspond to three periods of the day; morning between 04:00 am and 12:00 pm ($P_{morning}$), afternoon between 12:00 pm and 20:00 pm ($P_{afternoon}$), night between 20:00 pm and 04:00 am (P_{night}). Each of these periods has an equal length.
- Calculate the probability matrix between each two consecutive periods, i.e., between morning and afternoon is P_{ma}, between afternoon and night is P_{an}, and so forth. Given the probability matrix $P_{morning}$ and $P_{afternoon}$, the probability matrix between the periods *morning* and *afternoon* is

$$P_{ma} = P_{morning}\, w_{morning} + P_{afternoon} w_{afternoon}\,, \tag{2}$$

where $w_{morning}$ and $w_{afternoon}$ are weights calculated depending on the time period iteration, i.e., more close to starting morning period, then $w_{morning}$ is close to 1 and $w_{afternoon}$ is close to 0, more close to starting afternoon period, then $w_{morning}$ is close to 0 and $w_{afternoon}$ is close to 1.

$$w_{morning} = 1 - w_{afternoon} \tag{3}$$

$$w_{afternoon} = \sum_{k=1}^{T} \frac{K}{T}\,, \tag{4}$$

where k is an integer between $1 \le k \le T$ and T is the duration of the period defined, i.e., morning, afternoon, night. For instance, for a period of 8 h for morning, afternoon, night, the pair ($w_{morning}$, $w_{afternoon}$) is composed of {(0.875,0.125), (0.75,0.25), (0.625,0.375), (0.5,0.5), (0.375,0.625), (0.25,0.75), (0.125,0.875), (0,1)}. The same reasoning applies for the probabilities between the other periods.

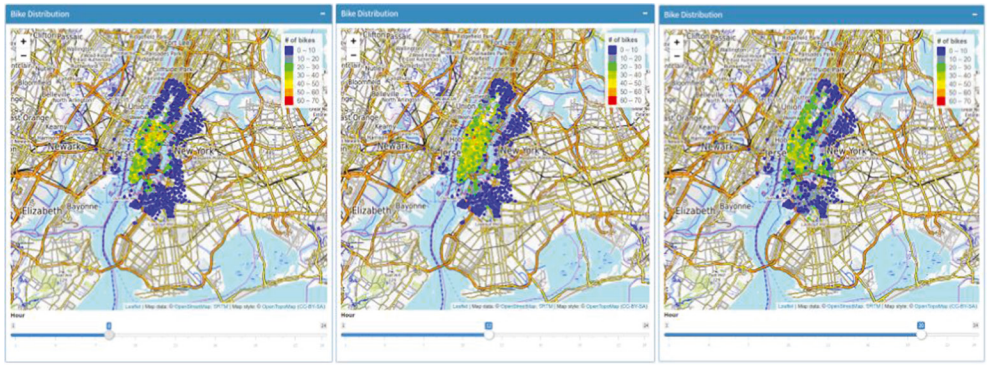
This possibility allows to have only 3 transition matrices instead of 24, but with a higher resolution on the transition between the states of the model.

4 NUMERICAL RESULTS

Given the historical data, the prediction model is performed for 1 day-ahead starting at 01:00 am on 1st of July 2018. The results are divided in two separated categories, the ones for the city manager and the ones for the user. The validation of the results is provided at the end of the section.

4.1 City Manager

The question we want to answer here is *how is the hourly distribution of the bikes in the city tomorrow?* The answer is shown in Fig. 7 with the predicted distribution at three selected hours, i.e., 08:00 am, 12:00 pm and 08:00 pm. At 08:00 am most bikes are distributed in Manhattan downtown, but at 12:00 pm the bikes are dissipating around Central Park and the margin sides between Manhattan and Brooklyn, whereas at 08:00 pm a larger Brooklyn area is covered by bikes. Given the predicted distribution, a variety of offline and online data



(a) (b) (c)

Figure 7: Bike distribution for 1st of July 2018. (a) Bike distribution at 08:00 am; (b) Bike distribution at 12:00 pm; (c) Bike distribution at 08:00 pm.

applications are triggered to further analysis. For instance, the distribution can be further applied to predict the traffic jam and air quality for particular regions. Eventually, based on the bike usage pattern, new areas can be covered by bike stations or the capacity of the installed stations can be accordingly changed during the day. Finding the right capacity for the excessively demand stations is also a possible feature of the model.

4.2 User

The question we want to answer is *what is the probability of having bike available at a specific station?* The answer is given by the ratio of the final distribution of the specific station to the capacity of the station. The confidence interval is calculated using the standard deviation among 100 repetitions. The results are shown in Fig. 8 for two mostly used stations, i.e., station 2006 in the Central park and station 3119 in Brooklyn.

On the contrary, the probability of having free-space for parking at a specific station is given by reversing the results for the probability of bike available. This metric is essential to be available on the user mobile application for consulting at any moment.

4.3 Validation

For the validation of the results, we look at the predicted and real values among all the stations for each hour. The initial distribution of bikes in the system dictates the number of bikes used in the model. However, the initial number of bikes is estimated from the data provided by [10] using the following reasoning: for each transition between start and end station, if the trip duration is less than the average trip duration calculated for the considered period, then the bike is considered to be parked at the start station and otherwise at the end station, as presented by Algorithm 1. We use the Root Mean Square Error (RMSE). which is given by

$$RMSE = \sqrt{\frac{1}{TS}\sum_{i=1}^{T}\sum_{j=1}^{S}\left(\hat{Y}_{ij} - Y_{ij}\right)^2}, \tag{5}$$

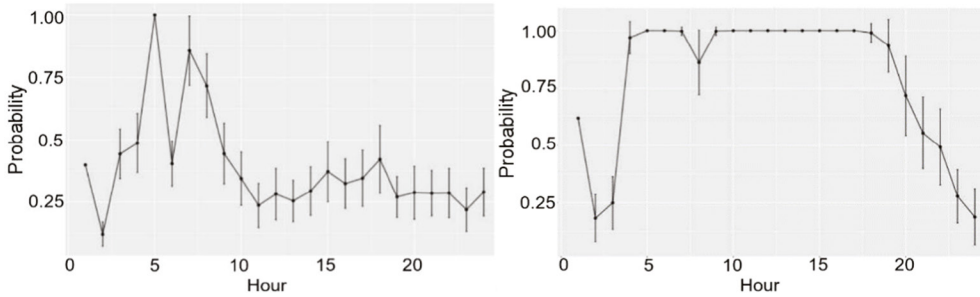


Figure 8: Probability of bike available for 1st of July 2018. (a) Station 3119 in Brooklyn; (b) Station 2006 in Central Park.

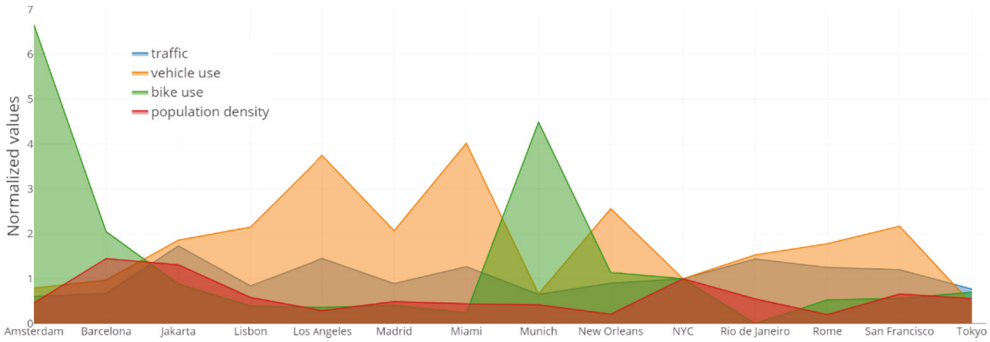


Figure 9: City comparison.

where $\hat{Y}_{ij}$ and Y_{ij} are the predicted and real values, S is the number of stations, and T represents the time period, which is 24 hours.

The numerical value of RMSE calculated following the reasoning above is 1.34 which indicates a reasonable value for the model proposed.

5 CITY COMPARISON

New York city is a different city from many others, mainly because of the high traffic conditions and various social events. Although the bike system covers completely the Manhattan area, the bike usage is significant high everyday disregarding the events that occur, as shown previously. The mobility in the city is also dictated by the other means of transportation as well as the population number. Hence, we collect data from [15,16] for 2018. The normalized values for each city are computed with regard to New York city findings, e.g., the ratio of the values from each category to the New York city value for the same category. By comparison with other cities in terms of traffic index, vehicle and bike use, and population density, we find out in Fig. 9 of few interesting properties about New York city (abbreviated as NYC): it is similar to New Orleans in terms of traffic, to Barcelona for car use, to Jakarta for bike use and population density. The bike usage in Munich is 4.5 times higher than in New York city, while in Amsterdam is almost 6.5 times higher, hence, each city has its own patterns. But the other city's characteristics, e.g., traffic and population density factors, could be as well be considered for more accurate models.

6 CONCLUSION

We proposed a simple, but efficient mathematical model to predict the bike distribution for an urban sharing bike system considering the weather and event impact. The model results are directly dependent on the city's characteristics. We observed for the New York city that only the temperature influences the bike usage pattern and the events have an insignificant impact. But this might not be the case for the other cities as we revealed that other properties characterize the city. The main insights are in particular fundamental for the city manager who needs to monitor the city and to design a more precise bike system for the city, e.g., capacity versus location of the stations and the user who wants to efficiently use the bike sharing system. Creating a complete city's characteristic model is part of our future work.

ACKNOWLEDGEMENTS

This article is a result of the Generation.mobi project (17369), supported by Competitiveness and Internationalization Operational Programme (COMPETE 2020), under the PORTUGAL 2020 Partnership Agreement, through the European Regional Development Fund (ERDF).

REFERENCES

[1] Sustainable Development Goals, availabe at https://www.un.org/sustainabledevelopment/sustainable-development-goals/ (accessed August 2018).
[2] Frade, I. & Ribeiro, A., Bicycle sharing systems demand. *Procedia – Social and Behavioral Sciences,* **111**, pp. 518–527, 2014.
[3] Gast, N., Massonnet, G. & Reijsbergen, D., Probabilistic Forecasts of Bike-Sharing Systems for Journey Planning. *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*, Melbourne, Australia, 2015.
[4] Singhvi, D., Singhvi, S., Frazier, P., Henderson, S., Mahony, E., Shmoys, D. & Woodard, D., Predicting Bike Usage for New York Cities Bike Sharing System. *AAAI Workshop: Computational Sustainability*, Texas, USA, 2015.
[5] Lin, L., Peeta, S., He, Z. & Wen, X., Predicting Station-level Hourly Demands in a Large-scale Bike-sharing Network: A Graph Convolutional Neural Network Approach, available at http://arxiv.org/abs/1712.04997, 2018.
[6] Datta, A., Master thesis, 2014, Predicting bike-share usage patterns with machine learning, University of Oslo.
[7] Yang, Z., Hu, J., Shu, Y., Cheng, P., Chen, J. & Moscibroda, T., Mobility Modeling and Prediction in Bike-Sharing Systems. *MobiSys*, Singapore, 2016.
[8] Chen, L., Zhang, D., Wang, L., Yang, D., Ma, X., Li, S.,Wu, Z., Pan, G., Nguyen, T. & Jakubowicz, J., Dynamic Cluster-Based Over-Demand Prediction in Bike Sharing Systems. *UbiComp*, Heidelberg, Germany, 2016.
[9] Zhang, J., Pan, X., Li, M. &Yu, P., Bicycle-Sharing System Analysis and Trip Prediction. *17th IEEE International Conference on Mobile Data Management*, Porto, Portugal, 2016.
[10] Citi Bike data, available at https://www.citibikenyc.com/ (accessed August 2018).
[11] Weather Underground, available at https://www.wunderground.com (accessed August 2018).
[12] Concert archives, available at https://www.concertarchives.org (accessed August 2018).
[13] Basketball Reference, available at https://www.basketball-reference.com/ (accessed August 2018).
[14] New York Rangers, available at https://www.nhl.com/rangers/schedule (accessed August 2018).
[15] Numbeo, available at https://www.numbeo.com (accessed August 2018).
[16] Wikipedia, available at https://wikipedia.com (accessed August 2018).