Incorporating spatial network information to improve demand prediction for bike share system expansion

Zhaoyu Kou
Purdue University
West Lafayette, USA
kouz@purdue.edu

Hua Cai
Purdue University
West Lafayette, USA
huacai@purdue.edu

ABSTRACT

In order to better take advantage of the benefits from bike share systems (BSSs), such as traffic congestion reduction and emission reduction, many cities keep expanding their station-based BSSs. Demand prediction for BSS expansion plays an important role in sizing the new stations and preparing the operations when planning the BSS expansion. There are limited studies focusing on the BSS-expansion demand prediction. Such studies mainly rely on external socio-demographic and point-of-interest data, which lacks the transferability across different cities. This study incorporates spatial network information into the prediction models, which shows that the spatial structure of BSS station networks contains important information for the BSS-expansion demand prediction. A Spatial-Eccentricity-Quantile-based Ensemble Model (SEQEM) is also proposed, which requires no external data but yields better prediction performance than external-data-based models.

CCS CONCEPTS

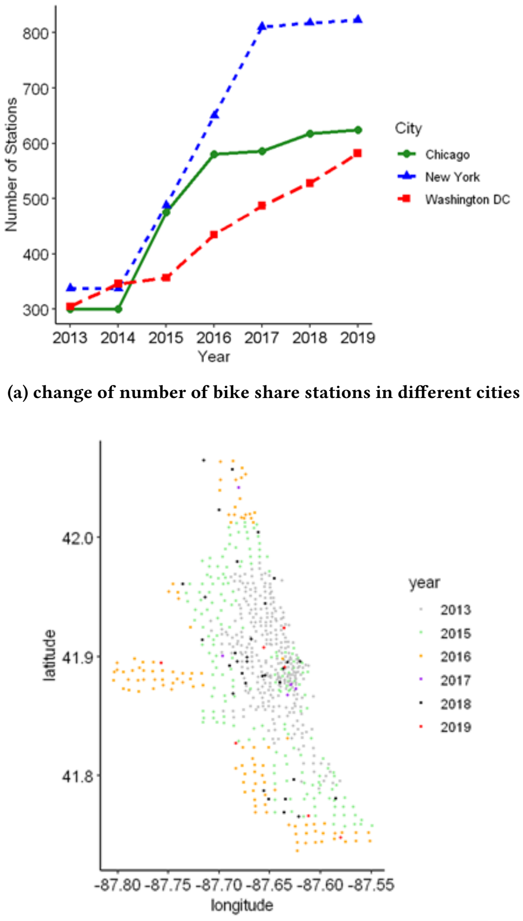
• **Applied computing** → *Forecasting*; *Transportation.*

KEYWORDS

bike share, system expansion, demand prediction, spatial network

1 INTRODUCTION

Bike share can potentially benefit the cities and the society from multiple aspects: reducing public transit travel time [1], alleviating traffic congestion [2], enhancing multimodal transport connections [3], saving travel cost [4], and reducing greenhouse gas (GHG) emissions [3]. To better take advantage of these potential benefits from bike share systems (BSSs), system operators are implementing different strategies to improve the BSSs. One of the most important strategies is to expand the system. As shown in Figure 1(a), many cities kept adding new stations to their station-based systems to expand the overall spatial coverage of the system or increase the density of stations at high demand areas. Figure 1(b) shows the locations of the BSS stations in Chicago as well as their year of launch, illustrating the progress of expanding the station network to the outskirt areas and also increasing the station density in the central areas.



(a) change of number of bike share stations in different cities

(b) locations of the stations in Chicago and their launch years

Figure 1: Examples of expansion process of BSSs

When planning the expansion of the station network, it is crucial to estimate the demand of the newly added stations, which can guide the decision of choosing the station capacity and allocating bikes [5]. However, very few studies have modeled the demand prediction for BSS expansion. The major challenge for such prediction problems is that no historical trip records exist for the new stations, while the demand of existing stations can be predicted based on historical trip records. Therefore, existing studies highly rely on external information such as socio-demographic and Point-of-Interest (POI) attributes to make the predictions [6, 7]. Such external data requires time and effort to collect and process; in addition, because different models used different data and the data availability may differ across cities, the model transferability is often limited when applying the models to different cities.

Additionally, existing studies only focused on the new stations themselves without considering how new stations may also influence the demands of existing stations [6, 8]. Some empirical studies have shown that interactions exist between new stations and existing stations [8, 9]. New stations may compete with existing stations when they are very close; while new stations could also complement existing stations and increase their demands since the new stations increase the choices of locations to pick up or return bikes, thus encouraging more usages at some existing stations. Considering such station interactions, the demand prediction for BSS expansion should not only focus on the new stations, but also account for the resulted demand change of existing stations. Models ignoring the station interactions may lead to incorrect decisions when allocating docks and bikes.

Therefore, demand prediction models that have low dependence on external data and consider the station interactions in the BSS expansion process are greatly needed. To address such model needs, this study proposes new strategies to improve BSS expansion prediction by incorporating spatial network information. First, to consider the stations' interactions with other stations in different distance ranges, features of spatial station density in fine-grained concentric bands are integrated into BSS expansion demand prediction models. Then, a Spatial Eccentricity Quantile based Ensemble Model (SEQEM) is proposed that requires no external data but yields better prediction performance than the model using external data.

The rest of this paper is organized as follows: Section 2 introduces the data and methodology. Then, the results are presented in Section 3. Lastly, Section 4 concludes the findings and discusses the limitations.

2 DATA AND METHOD

2.1 Data processing

This study aims to analyze how new stations impact the demand of existing stations. Therefore, the demand information needs to be first extracted from historical bike share trip data. This work selects Chicago as the case study city because the station-based BSS in Chicago has kept expanding over recent years and has a long enough history to study its expansion (Figure 1(a)).

The demand variable of interest is the average daily bike withdrawals w_{iym} at station i in month m of year y. The focus of bike withdrawal is consistent with most of the previous studies modeling bike share demand [9–13], but the same models can also be applied to estimate bike returns. We chose to predict the average daily demand because, for BSS expansion planning, the steady-state demand level matters more to the decision makers than the short-term (e.g., hourly) demand fluctuations [7, 9]. The average daily bike withdrawals w_{iym} at station i in month m of year y is computed as:

$$w_{iym} = c_{iym}/a_{iym} \tag{1}$$

where c_{iym} is the total bike withdrawals at station i in month m of year y, and a_{iym} is the number of days that station i is active in month m of year y. The first available date of each station is identified as the date that the station first appeared in the trip record.

2.2 Demand prediction for system expansion

Based on the observations from previous studies [8, 9, 14], the distances between stations is a key factor for station interactions. Inspired by this observation, this section integrates the distance information into the demand prediction model to improve prediction performances.

2.2.1 Model task. The task of the BSS expansion demand prediction is that, given the dependent variable $\{w_{iym}|y = Y, m = M\}$ – the average daily bike withdrawals for each station i in month M of year Y, train a demand prediction model using some independent variables $\{v_{iym}|y = Y, m = M\}$ associated with the stations. Then the model is applied to predict the average daily bike withdrawals $\{w_{iym}|y = Y + 1, m = M\}$ based on the independent variables $\{v_{iym}|y = Y + 1, m = M\}$ of all stations in the same month M in the next year $Y + 1$. Year $Y + 1$ contains more stations than year Y because new stations have been added.

2.2.2 Features. The features (independent variables) considered in the demand prediction model can be classified into the following categories in Table 1: In this study, prediction models are trained using different categories of features or a combination of different categories of features to analyze which features are more predictive for the demand prediction. The models that have been analyzed include

- "GEO", "GEO+DENS", and "EXT": "GEO+DENS" model (i.e., the model trained using both "GEO" and "DENS" features) is compared with "GEO" model to evaluate whether adding the information of concentric station density helps improve the predictions. Because the "GEO+DENS" model performed better for stations in central areas and the "GEO" model performed better in outskirt areas of the city (discussed in Section 3.1), a Spatial-Eccentricity-Quantile-based Ensemble Model (SEQEM) is proposed to ensemble the prediction results of the "GEO+DENS" model and "GEO" model in different spatial ranges (the ranges are determined by a threshold of "ECCQ", more details in Section 2.2.4). The major objective of the "GEO", "GEO+DENS", and SEQEM is to predict the demands for BSS expansion only based on the information of station locations and the spatial structure of station networks, which avoids using the external data as in "EXT".
- "GEO+DENS+ECC" and "GEO+DENS+ECCQ": Instead of using ECCQ as a threshold in the SEQEM, this study also directly adds the ECC and ECCQ features into the prediction models to train the "GEO+DENS+ECC" and "GEO+DENS+ECCQ" models, respectively. The performances of these two models are compared with SEQEM to evaluate the necessity of using the ensemble method of SEQEM.
- "DENS" and "DENS2": The prediction performances of models using only DENS and only DENS2 as features will be compared to evaluate whether the more fine-grained bands improve the predictions (Section 3.3).

2.2.3 Model training, hyper-parameters, and performance evaluation. This study applies the XGBoost [20] algorithm to train all the models with different features. We chose XGBoost because it can handle high-dimension data with feature-selection ability [21], especially for the "DENS" features that have 45 variables for all the distance bands. For different models, the hyper-parameters are tuned using grid search [22] with 10-fold cross validation. In this study, prediction models are trained using the data in month M

Table 1: Categories of features used in this study

Feature category (acronym)	Feature details
DENS	Station density (we propose to use such features to represent station network structure). For each station, first construct the concentric distance bands around the station at 0.1-mile interval, i.e., 0 − 0.1, 0.1 − 0.2, . . . , 4.4 − 4.5 miles. Then, for each distance band, compute the number of other stations (including both new and old stations) that are located in the band. We only computed the station density within 4.5 miles from the target station. This threshold is selected based on the trip distance of historical trips. In all trips from 2013 to 2019, the travel distances of 99% of the trips are within 4.5 miles.
DENS2	Station density (literature method). Construct concentric distance bands and compute station density in a similar way as "DENS", but with only two coarse distance bands: 0 − 0.5 and 0.5 − 3.1 miles, which is the arbitrarily selected bands from a previous study [9].
GEO	Geographic information. It includes the latitude and longitude of the stations.
EXT	External information. It denotes the external information around the locations of the stations, which covers the following variables that are found to be important for bike share demand prediction in the previous studies [9, 15, 16]: (1) demographic information - collected from American Community Survey [17] at the census-tract level, the values of one census tract is assigned to a station if the station is located within the census tract. Such features include population density and per capita income; (2) point of interest - collected using Google Maps Places API. Such variables are the counts of points of interest (e.g., bus stations) within a 1,000 feet (305 meters) buffer around a station. The 1,000 feet buffer is selected as an appropriate distance that people can walk between a bike share station and surrounding points of interest [8, 18]. Such features include number of bus stations, subway stations, restaurants, parks, parking lots, museums, and schools (each point-of-interest category corresponds to one feature).
ECC	Spatial eccentricity. The spatial eccentricity of a certain station i is defined as the average distance from station i to all other stations [19]. Note that the spatial eccentricity of a certain station could be different in different months since new stations are added in the expansion process.
ECCQ	Spatial eccentricity quantile. For year y and month m, the ECCQ of station i is the quantile value of the ECC_{iym} value of station i corresponding to the ECC distribution of all stations in year y and month m.

in year Y and evaluated using the data of month M in year $Y + 1$. Instead of only evaluating the prediction performance on new stations [6], the predictions on all stations (including both new and existing stations) were evaluated to analyze the effects of station interactions. Root-mean-square error (RMSE) and mean-absolute-percentage error (MAPE) are adopted to evaluate the performance of predictions.

2.2.4 Spatial-Eccentricity-Quantile-based Ensemble Model (SEQEM). After observing the predictions of "GEO" and "GEO+DENS" model, it is found that the "GEO+DENS" model performed better in the central area of the station network (regarding RMSE), while the "GEO" model performed better at those outskirt low-demand stations (regarding MAPE). Therefore, to take advantage of the good performance of both "GEO+DENS" and "GEO" models in different areas, the Spatial-Eccentricity-Quantile-based Ensemble Model (SEQEM) is proposed, which switches the applied prediction models based on the spatial eccentricity quantile threshold Q. Besides potentially improving the overall prediction performance, another objective of the SEQEM is to identify a spatial boundary within which the station interactions should be considered. The detailed procedure of SEQEM is presented in Algorithm 1. Note that a quantile threshold Q is used as the threshold instead of an absolute value of spatial eccentricity, because the spatial eccentricity of each station can change when new stations are added. In this study, SEQEM is applied on all the data with a list of Q in $QL = \{0.02, 0.04, ..., 0.98\}$. By analyzing the RMSE and MAPE changes with varying Q, the Q value that yields the best overall performance can be identified.

3 RESULTS AND DISCUSSIONS

In this section, we first compared the performances of the "GEO" and "GEO+DENS" models in Section 3.1 to evaluate whether integrating the spatial structure information can improve the predictions. Then, Section 3.2 compares the performances of the proposed SEQEM with the baseline "EXT" model. Lastly, the performances of models using only "DENS" and only "DENS2" as features are compared to explore whether the fine-grained distance bands improve the predictions in Section 3.3.

3.1 Performance of "GEO", "GEO+DENS", and "EXT" models

Figure 2 presents the overall prediction performances of the "GEO", "GEO+DENS", and "EXT" models in different years. For all three models, the RMSE in 2015 is much higher than other years, because the expansion in 2015 was in a much larger scale than other years – there were 174 new stations in July 2015, while in July 2014, the system only had 300 stations. In contrast, there were 82, 8, 9, and 32 new stations in July of 2016 to 2019, respectively. When the number of stations is quickly increased by 58% in July 2015, the information learned from the previous year's data is insufficient to provide a desirable prediction.

Algorithm 1: Spatial-Eccentricity-Quantile-based Ensemble Model (SEQEM)

```
Input  : (1) the average daily demand $w_{iym}$ for each station
         $i$ in different months and years as well as the
         corresponding "GEO", "DENS", and "ECCQ"
         features; (2) $QL$: a list of candidate spatial
         eccentricity quantile threshold $Q$, in this study,
         $QL = \{0.02, 0.04, ..., 0.98\}$; (3) $Metric$: the model
         evaluation metric(s), which in this study are RMSE
         and MAPE.
Output : $\{EV_Q\}$: The evaluated model performances
         corresponding to the $Q$ values in $QL$
for any month $M$ and year $Y$ in the training data do
    Train models $MD_{YM}^{GEO}$ and $MD_{YM}^{GEO+DENS}$ using data
      $\{w_{iym} | y = Y, m = M\}$ and the corresponding "GEO"
      and "GEO+DENS" features, respectively
end for
for $Q$ in $QL$ do
    for any month $M$ and year $Y$ in the training data do
        for each station $i$ do
            if $ECCQ_{iYM} < Q$ then
                Predict the demand $f_{iYM}$ using
                  $MD_{(Y-1)M}^{GEO+DENS}$
            else
                Predict the demand $f_{iYM}$ using $MD_{(Y-1)M}^{GEO}$
            end if
        end for
    end for
    Compute an overall performance $EV_Q$ based on all the
      demand rates $\{w_{iym}\}$ and predicted values $\{w_{iym}\}$
      using $Metric$, which is then recorded in $\{EV_Q\}$
end for
```

In the years from 2016 to 2019, when the expansion is more gradual, the RMSE is lower and in an identical level in different years. In these four years, overall, the "GEO+DENS" outperforms "GEO" regarding the RMSE metric (Figure 2(a)). However, the "GEO+DENS" model has larger overall MAPE error than the "GEO" model (Figure 2(b)). In Figure 3(d), the spatial distribution of the absolute percentage errors is plotted, which shows that stations with very large percentage errors are located in the Southern outskirt area of the BSS (bottom in the figure).

In the following sections (Section 3.2 and 3.3), this study will only focus on "gradual expansion" and only compare the prediction performances of different models evaluated using 2016-2019 data.

After further exploration, it is found that the "GEO+DENS" model outperforms "GEO" model in the central area of the city by considering the changes of station density in the concentric buffers, but it tends to overestimate the demands in the outskirt areas, which leads to larger MAPE for these low-demand stations. Considering the entire system, the benefits from better predictions in the central high-demand areas outweigh the poor predictions in the outskirt low-demand areas. Therefore, the "GEO+DENS" model pays more attention to the spatial structures of the stations in the central areas, but such knowledge learned from the central areas does not apply to the outskirt areas. In contrast, the simpler model, "GEO", which only considers the latitude and longitude of stations, yields better predictions for those low-demand stations.

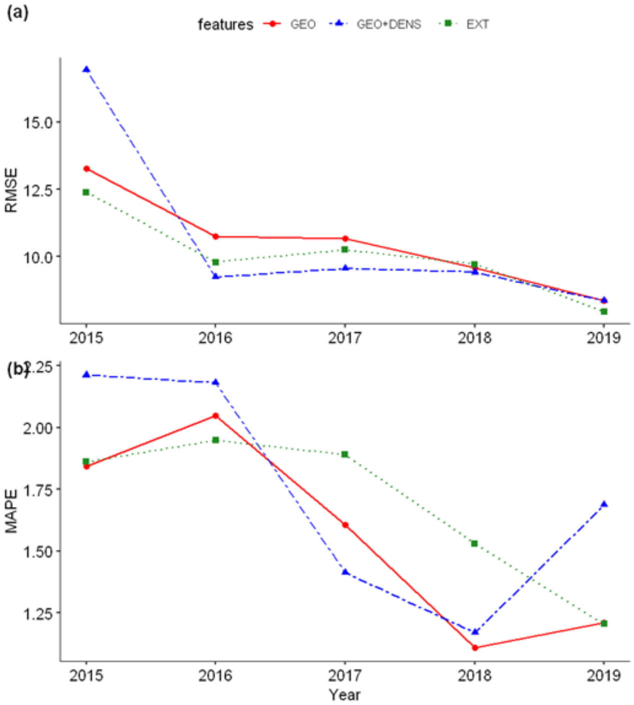


Figure 2: Overall model performance on all stations in different years: (a) RMSE, (b) MAPE

3.2 Performance improvement by SEQEM

Based on the observations from the performances of "GEO" and "GEO+DENS" models, the proposed SEQEM is applied to further improve the prediction performances. Figure 4 shows the overall performance of SEQEM with different spatial eccentricity quantile threshold (evaluated using all the 2016-2019 data). Considering both RMSE and MAPE, a quantile of 0.78 is identified as a quantile threshold that yields the best overall performance. This threshold can also be viewed as the spatial boundary only within which the station interactions should be considered. Using the 0.78 quantile threshold, the SEQEM achieved a better performance than the "EXT" model regarding both RMSE and MAPE. In addition, compared with the "EXT" model, the SEQEM model has better transferability to be applied to other cities, because it does not require external data.

There are other options such as simply adding the spatial eccentricity or spatial eccentricity quantile as a feature into the model (i.e., the "GEO+DENS+ECC" model and "GEO+DENS+ECCQ" model, respectively). These two models are also evaluated. The results show that, for RMSE (Figure 5(a)) and MAPE (Figure 5(b)), both "GEO+DENS+ECC" (RMSE: 10.37, MAPE: 1.52) and "GEO+DENS+ECCQ" (RMSE: 9.22, MAPE: 1.54) yield poorer performance than SEQEM

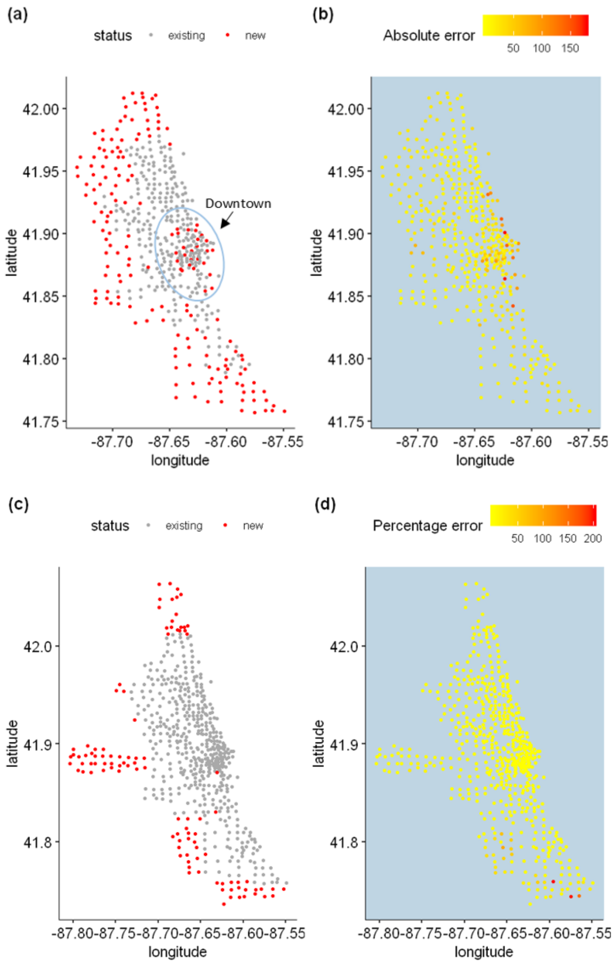


Figure 3: Spatial visualization of the prediction errors of "GEO-DENS" model in July 2015 ((a) locations of new stations, (b) absolute error) and August 2016 ((c) locations of new stations, (d) absolute percentage error)

(RMSE: 9.12, MAPE: 1.42). Therefore, the spatial eccentricity quantile should only serve as a threshold to identify a spatial boundary, but its value does not add more useful information for the prediction.

3.3 The impact of concentric band construction

By comparing the performance of models that only uses "DENS" or "DENS2" (literature method) features, it is found that the "DENS" model has a much lower RMSE of 10.31 than that of the "DENS2" model (22.97); "DENS" model also has a lower MAPE (2.47) than that of "DENS2" (2.99). Overall, the more fine-grained distance bands can better reflect the structure of the station network and lead to better prediction performances.

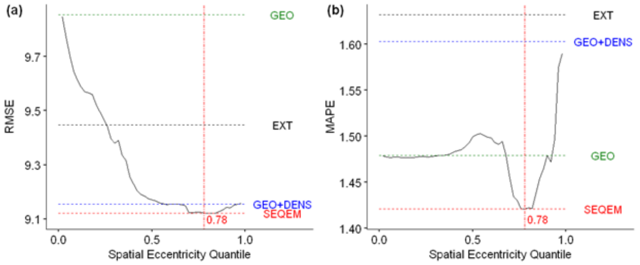


Figure 4: Performance of SEQEM by varying the spatial eccentricity quantile based on (a) RMSE and (b) and MAPE. The horizontal lines indicate the performance of the corresponding models in the right-hand side). The quantile 0.78 provides the overall best performance considering both RMSE and MAPE.

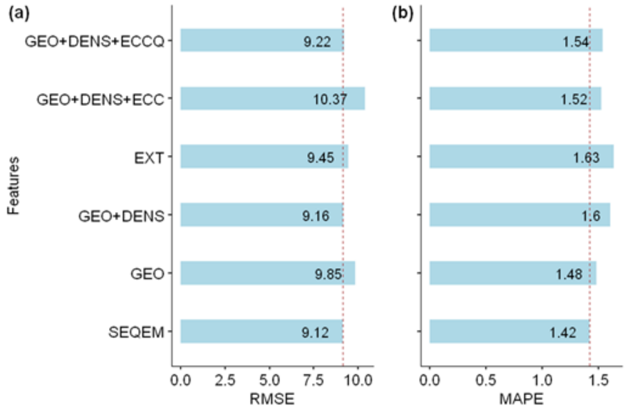


Figure 5: The performances when adding spatial eccentricity ("ECC") and spatial eccentricity quantile ("ECCQ") features: (a) RMSE, (b) MAPE (the vertical dashed lines indicate the performance of SEQEM)

4 CONCLUSIONS AND LIMITATIONS

This study focuses on improving the demand prediction for bike share system expansion. Features of spatial station density in fine-grained concentric bands around a station are constructed to represent the number of stations in different distance ranges that the station can interact with. A Spatial Eccentricity Quantile based Ensemble Model (SEQEM) is proposed to further improve the prediction performances and also identify the spatial range that the station interactions take effects. The results of the demand prediction models show that: integrating the station density in concentric distance bands improves the prediction performance by considering the spatial structure of the station network. The "GEO+DENS" model performed well for central areas but has poor performance for outskirt low-demand stations. The proposed SEQEM addresses this limitation and improves the prediction performance for the entire system. With the 0.78 spatial eccentricity quantile threshold, the SEQEM yields better performance than the "EXT" model regarding both RMSE and MAPE. This indicates that the spatial structure of the station network contains very important information for the small-scale expansion demand prediction, which can also save the effort to collect and process the external data.

This study provides insights for the station interactions in the BSS expansion process and practical suggestions to integrate spatial station network information to improve the BSS expansion demand prediction. However, there are some limitations that need to be improved in future research. First, the prediction results indicate that all models have poor performances for the large-scale expansion in 2015, which should be improved in future research. Second, this study only focuses on the BSS in Chicago. Future studies can apply the SEQEM model to other BSSs to evaluate the model transferability in different cities.

REFERENCES

[1] Sakari Jäppinen, Tuuli Toivonen, and Maria Salonen. Modelling the potential effect of shared bicycles on public transport travel times in greater helsinki: An open data approach. *Applied Geography*, 43:13–24, 2013.

[2] Ahmadreza Faghih-Imani, Sabreena Anowar, Eric J Miller, and Naveen Eluru. Hail a cab or ride a bike? a travel time comparison of taxi and bicycle-sharing systems in new york city. *Transportation Research Part A: Policy and Practice*, 101: 11–21, 2017.

[3] Susan Shaheen and S Guzman. et zhang, h.(2010). bikesharing in europe, the americas, and asia: past, present, and future. *Transportation Research Record: Journal of the Transportation Research Board,(2143)*, pages 159–167.

[4] Ralph Buehler and Andrea Hamre. Business and Bikeshare User Perceptions of the Economic Benefits of Capital Bikeshare. *Transportation Research Record: Journal of the Transportation Research Board*, 2520:100–111, 2015. ISSN 0361-1981. doi: 10.3141/2520-12. URL http://trrjournalonline.trb.org/doi/10.3141/2520-12.

[5] Jiawei Zhang, Xiao Pan, Moyin Li, and Philip S Yu. Bicycle-sharing systems expansion: station re-deployment through crowd planning. In *Proceedings of the 24th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, pages 1–10, 2016.

[6] Junming Liu, Leilei Sun, Qiao Li, Jingci Ming, Yanchi Liu, and Hui Xiong. Functional zone based hierarchical demand prediction for bike system expansion. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 957–966, 2017.

[7] Bryan C Watson and Cassandra Telenko. Predicting demand of distributed product service systems by binomial parameter mapping: A case study of bike sharing station expansion. *Journal of Mechanical Design*, 141(10), 2019.

[8] Ying Zhang, Tom Thomas, MJG Brussel, and MFAM Van Maarseveen. Expanding bicycle-sharing systems: lessons learnt from an analysis of usage. *PLoS one*, 11 (12):e0168604, 2016.

[9] Michael Hyland, Zihan Hong, Helen Karla Ramalho de Farias Pinto, and Ying Chen. Hybrid cluster-regression approach to model bikeshare station usage. *Transportation Research Part A: Policy and Practice*, 115:71–89, 2018.

[10] Yexin Li, Yu Zheng, Huichu Zhang, and Lei Chen. Traffic prediction in a bike-sharing system. In *Proceedings of the 23rd SIGSPATIAL International Conference on Advances in Geographic Information Systems*, pages 1–10, 2015.

[11] Lei Lin, Zhengbing He, and Srinivas Peeta. Predicting station-level hourly demand in a large-scale bike-sharing network: A graph convolutional neural network approach. *Transportation Research Part C: Emerging Technologies*, 97:258–276, 2018.

[12] Divya Singhvi, Somya Singhvi, Peter I Frazier, Shane G Henderson, Eoin O'Mahony, David B Shmoys, and Dawn B Woodard. Predicting bike usage for new york city's bike sharing system. In *Workshops at the twenty-ninth AAAI conference on artificial intelligence*, 2015.

[13] Bo Wang and Inhi Kim. Short-term prediction for bike-sharing service using machine learning. *Transportation research procedia*, 34:171–178, 2018.

[14] Jueyu Wang and Greg Lindsey. Do new bike share stations increase member use: A quasi-experimental study. *Transportation research part A: policy and practice*, 121:1–11, 2019.

[15] Zhaoyang Liu, Yanyan Shen, and Yanmin Zhu. Inferring dockless shared bike distribution in new cities. In *Proceedings of the eleventh ACM international conference on web search and data mining*, pages 378–386, 2018.

[16] Cyrille Médard de Chardon, Geoffrey Caruso, and Isabelle Thomas. Bicycle sharing system 'success' determinants. *Transportation research part A: policy and practice*, 100:202–214, 2017.

[17] American Community Survey. ACS Demographic and Housing Estimates - 5-Year Estimates Data Profiles, 2017. URL https://data.census.gov/cedsci/table?g=0400000US17{_}0500000US17031{&}d=ACS5-YearEstimatesDataProfiles{&}tid=ACSDP5Y2017.DP05.

[18] Edoardo Croci and Davide Rossi. Optimizing the position of bike sharing stations. the milan case. 2014.

[19] Rémy Cazabet, Pierre Borgnat, and Pablo Jensen. Using degree constrained gravity null-models to understand the structure of journeys' networks in bicycle sharing systems. In *ESANN 2017-European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning*, 2017.

[20] Tianqi Chen et al. Guestrin, c.: Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'16)*, pages 785–794, 2016.

[21] Jie Wang, Jing Xu, Chengan Zhao, Yan Peng, and Hongpeng Wang. An ensemble feature selection method for high-dimensional data based on sort aggregation. *Systems Science & Control Engineering*, 7(2):32–39, 2019.

[22] Iwan Syarif, Adam Prugel-Bennett, and Gary Wills. Svm parameter optimization using grid search and genetic algorithm to improve classification performance. *Telkomnika*, 14(4):1502, 2016.