Taylor & Francis
Taylor & Francis Group

Applied Artificial Intelligence
An International Journal

ISSN: 0883-9514 (Print) 1087-6545 (Online) Journal homepage: www.tandfonline.com/journals/uaai20

Predicting Trip Duration and Distance in Bike-Sharing Systems Using Dynamic Time Warping

Ahmed Ali, Ahmad Salah, Mahmoud Bekhit & Ahmed Fathalla

To cite this article: Ahmed Ali, Ahmad Salah, Mahmoud Bekhit & Ahmed Fathalla (2025) Predicting Trip Duration and Distance in Bike-Sharing Systems Using Dynamic Time Warping, Applied Artificial Intelligence, 39:1, 2474786, DOI: 10.1080/08839514.2025.2474786

To link to this article: https://doi.org/10.1080/08839514.2025.2474786

© 2025 The Author(s). Published with license by Taylor & Francis Group, LLC.

Published online: 04 Mar 2025.

Submit your article to this journal

Article views: 986

View related articles

View Crossmark data

Full Terms & Conditions of access and use can be found at
https://www.tandfonline.com/action/journalInformation?journalCode=uaai20

APPLIED ARTIFICIAL INTELLIGENCE
2025, VOL. 39, NO. 1, e2474786 (29 pages)
https://doi.org/10.1080/08839514.2025.2474786

Taylor & Francis
Taylor & Francis Group

RESEARCH ARTICLE

OPEN ACCESS Check for updates

Predicting Trip Duration and Distance in Bike-Sharing Systems Using Dynamic Time Warping

Ahmed Ali [a,b*], Ahmad Salah [c,d*], Mahmoud Bekhit [e,f*], and Ahmed Fathalla [g*]

[a]Department of Computer Science, College of Computer Engineering and Sciences, Prince Sattam Bin Abdulaziz University, Al-Kharj, Saudi Arabia; [b]Deparment of Computer Science, Higher Future Institute for Specialized Technological Studies, Cairo, Egypt; [c]College of Computing and Information Sciences, University of Technology and Applied Sciences, Ibri, Ad-Dhahirah, Sultanate of Oman; [d]Department of Computer Science, Faculty of Computers and Informatics, Zagazig University, Zagazig, Sharkeya, Egypt; [e]Peter Faber Business School, Australian Catholic University (ACU), Sydney, Australia; [f]Faculty of Engineering and Information Technology, University of Technology Sydney (UTS), Sydney, Australia; [g]Department of Mathematics, Faculty of Science, Suez Canal University, Ismailia, Egypt

ABSTRACT

Bike-sharing systems (BSSs) have recently become important in urban transportation due to several factors, such as their cost-effectiveness and environmental considerations. The BSS provides an enormous amount of data that is recorded regarding trips. This huge volume of bike sharing data raises various challenges and opportunities. Many research studies have used bike sharing datasets to understand the geographical, social, financial, and behavioral aspects of bike user behaviors. While existing literature primarily focuses on predicting the number of rentals and returns per station, this study addresses the complementary aspect of predicting the trip duration and distance of the trip. Accurate prediction of ride duration allows a better estimate of bike availability at stations, while distance predictions assist in maintenance planning based on bike usage patterns. The contribution of this work is twofold. First, the proposed work clusters the BSS dataset into k sub-datasets based on similarity of dataset instances. Then, the predictive model is trained to predict the data of each sub-dataset separately. Thus, there will be k models for the k sub-dataset. Next, the performance of the proposed method, the average score of the k models, will be compared to the performance of a model trained on the complete dataset on predicting BSSs ride duration and distance of the trip. The rationale for splitting the dataset into k sub-datasets is to separate similar patterns in one sub-dataset. Second, the utilization of the dynamic time warping (DTW) algorithm on the BSSs data was proposed for the clustering purpose, as the DTW usage is very limited in the current literature of BSSs. The dataset clustering is based on the similarity of the curves representing the number of trips

ARTICLE HISTORY
Received 26 January 2024
Revised 17 February 2025
Accepted 18 February 2025

CONTACT Ahmed Ali a.abdalrahman@psau.edu.sa Department of Computer Science, College of Computer Engineering and Sciences, Prince Sattam Bin Abdulaziz University, Al-Kharj,11942Saudi Arabia

*These authors contributed equally to this work.

© 2025 The Author(s). Published with license by Taylor & Francis Group, LLC.
This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited. The terms on which this article has been published allow the posting of the Accepted Manuscript in a repository by the author(s) or with their consent.

between each pair of bike stations throughout the day hours. Then, the DTW algorithm is used to measure the curve similarity between these bike station pairs' curves. These two contributions of the proposed approach complement existing prediction models for rentals and returns, providing a comprehensive solution for BSS optimization. The proposed method was thoroughly evaluated on two real datasets of different sizes. For the two datasets, the obtained results show that the best improvements of the predictive model's accuracy are 30% and 42% on average for predicting trip duration and distance of the trip, respectively.

Introduction

Recent exponential growth in the transformation industry has enabled researchers to develop effective programs to maintain pace with this development. Specifically, bike riding is regarded as an essential and inexpensive means of transportation on which people used to rely. Meanwhile, the bike-sharing concept has been applied globally to reduce traffic congestion, promote physical activity, and combat climate change. Starting in 2016, at least one thousand bike-sharing systems are in use in sixty various countries, with economic returns reaching $7.7 billion by 2022. The bike-sharing market is anticipated to have 930 million consumers in 2026 (Albuquerque, Sales Dias, and Bacao 2021; Otero, Nieuwenhuijsen, and Rojas-Rueda 2018). The surging demand for carbon-neutral bike-sharing has exacerbated mismatches in urban transportation. As a consequence of the growing need for carbon-neutral environments, there has been an increase in the demand for bike-sharing systems (Zhou et al. 2022).

In a variety of practical domains, machine learning methods are considered powerful tools (Adel et al. 2022; Ali et al. 2021; Indah Lestari et al. ; Salah et al. 2023). In the context of BSSs, one of the well-studied problems using machine learning is forecasting travel demand and create dynamic forecasting models. Determining the optimal number of bikes to be rented at different locations and predicting citywide bike utilization for the next period is another well-studied problem that can be achieved using machine learning methods (Li and Zheng 2019; Wang and Kim 2018; Zhou et al. 2022) While most existing research in bike-sharing systems focuses on predicting the number of rentals and returns at stations (Rudloff and Lackner 2014; Wang 2016; Yang et al. 2018), there are other crucial aspects where machine learning can provide valuable insights. In this context, predicting trip duration and distance of the trip, which is the focus of the current work, provides complementary information that is essential for system optimization. The proposed approach complements existing rental prediction models by providing insights into individual trip characteristics (e.g., trip distance and trip duration), which are crucial for

maintenance scheduling and capacity planning. These metrics improve the station-centric models by exposing the system strain and user behavior patterns invisible to counts alone.

The Toronto bike share dataset has been observed in numerous studies (Butt et al. 2023; Caggiani et al. 2021; Cheng et al. 2020). The utilization of real-time data commonly serves multiple goals, including investigating the factors that influence bike rides (i.e., temperature), the assessment of bike stations' effectiveness, and establishment of operational approaches for public BSS. A study was proposed to estimate the fluctuating bike counts at each station within the San Francisco Bay Area, utilizing the Ford GoBike dataset (Ashqar et al. 2022). The primary objective of this study is to enhance the predicting precision for the number of accessible bikes at a designated bike-sharing station. The same dataset was trained using the Long Short-Term Memory (LSTM) model to predict the bike density at road intersections in San Francisco (Dubey et al. 2019). The Divvy Trips dataset is a real-world dataset collected in Chicago and evaluated using machine learning in Kumar Das, Manoj Joshi, and Dhal (2020) and Zhang et al. (2019). Using machine learning techniques, one of the proposed studies based on the Divvy dataset seeks to assist riders in selecting a suitable bicycle based on their travel needs. A subsequent analysis was conducted on the same dataset to determine the bike-sharing system's catchment area. Using the New York bike-sharing dataset (Chen et al. 2020), the recurrent neural network (RNN) algorithm is utilized for forecasting rental and return demand to provide online balancing plans. The forecasting of the short-term usage of bikes using the LSTM was also investigated using the New York bike-sharing system in Li et al. (2021).

Current bike-sharing systems are frequently based on machine learning and deep learning techniques. To predict or classify bike trips, datasets such as Toronto Bike Share, Divvy Trip, Ford Gobike, and New York City Bike Share have been utilized. The Toronto bike share dataset has been utilized in numerous studies, including Caggiani et al. (2021); Dong et al. (2016); and El-Assi, Salah Mahmoud, and Nurul Habib (2017). This real-time data is used for a variety of purposes, such as investigating the factors affecting bicycle trips, such as temperature, evaluating the efficiency of bicycle stations, and determining an operational strategy for public bike-sharing systems. The Divvy Trips dataset is a real-world dataset collected in Chicago and evaluated using machine learning in Kumar Das, Manoj Joshi, and Dhal (2020) and Zhang et al. (2019). Using machine learning techniques, one of the proposed studies based on the Divvy dataset aims to assist riders in selecting a suitable bicycle based on their travel needs. Another study analyzed the same data to determine the bike-sharing system's catchment area. The LSTM model was employed on the Ford GoBike dataset to predict the bike density at road intersections in San Francisco (Dubey et al. 2019). Using

the New York bike sharing dataset (Chen et al. 2020), the recurrent neural network (RNN) algorithm is utilized for forecasting rental and return demand to provide online balancing plans. The forecasting of the short-term usage of bikes using the LSTM was also investigated using the New York bike-sharing system in Li et al. (2021).

This study addresses the understudied aspect of trip duration and distance of the trip prediction in BSSs, complementing the existing body of work that focuses on rental and return predictions. Particularly, a new clustering method to divide the BSS data into k subsets using DTW to measure similarity between temporal demand curves of station pairs. Training specific models on each subset isolates latent behavioral patterns, improving prediction accuracy over suitable approaches. The initial application of DTW to BSS data addresses the issue of temporal misalignment in station-pair usage patterns where there is a critical gap in prior distance-based clustering. Experiments demonstrate that the proposed ensemble of k models outperforms single-model benchmarks, validating the value of temporal alignment for BSS optimization. The proposed work has two main objectives. First, this work proposed clustering bike station pairs with similar patterns over the day hours. To achieve this goal, the DTW (Berndt and Clifford 1994) was utilized to find the distance matrix between all pairs of bike stations. The rationale for using DTW is its ability to compare two temporal sequences and drive similarity among the raw data being compared. Due to this rationale, the DTW algorithm has been applied in several applications, including the performance analysis of karate skills (Fathalla et al. 2023). Another use of DTW is to translate real-time hand gestures captured by Kinect sensors (Kowdiki and Khaparde 2021). It has also been used as a preliminary procedure to recognize speech in machine translation (Jiang and Chen 2023). BSS dataset clustering can divide the dataset into smaller datasets where each sub-dataset includes similar data points, i.e., trip data. Predicting data with similar patterns is easier than predicting data with different patterns. Thus, the second purpose of this work is to improve the trip information prediction, i.e., trip distance and duration of the trip, by training the predictive model on a sub-dataset of similar data patterns which can improve the overall prediction accuracy. Predicting the duration of the trip should help the operators to estimate when a bike will arrive at different stations that allows them to plan the number of bikes at stations and guarantee an equitable number of bikes across the network. Prediction of the distance traveled by each bike helps, by monitoring usage patterns. It is valuable information for taking care of maintenance and scheduling: replacement of worn-out parts or the decision to discard a bike, which will raise the overall effectiveness and help the bike-sharing system to last longer.

The list of contributions of this work is as follows:

(1) To the best of the authors' knowledge, this is the first work to utilize the DTW algorithm to improve the prediction of BSS data (the datasets and source code are available online)[1]
(2) The current method proved that training the predictive model on each cluster individually produces an overall prediction accuracy better than training the same predictive model on the full BSS dataset.
(3) The proposed method was thoroughly evaluated and tested on two real datasets.

The rest of the paper is organized as follows. In Section 2, some of the existing methods are discussed. Section 3 exposes the proposed methods. The results are presented and discussed in Section 4. Finally, the paper is concluded in Section 5.

Literature Review

There is a very limited utilization of the DTW algorithm with the BSS data. The literature has few conducted research works in this direction with a limited use of clustering BSS datasets only to improve the predictive models' accuracy. This discussion will delineate the constrained studies to identify the research gap that inspired the present work.

In many cities across the world, bike-sharing systems have recently emerged as one of the most well-liked modes of transportation. These systems provide an affordable and convenient way for people to travel within the city. Consequently, a vast quantity of data is produced that describes the users' trips. Therefore, it is imperative to examine and comprehend this data to enhance the overall effectiveness and operation of bike-sharing systems.

The authors in Chabchoub and Fricker (2014) used the DTW distance metric along with the K-means clustering method to assess the similarity among stations. This study aimed to identify patterns and group stations based on their similarities. This approach was used to anticipate the availability of free docks and bikes at Velib stations in Paris. The similarity among stations is comprehended through clusters, and the centers of each cluster are elected using the DTW Barycenter Averaging (DBA) technique. In this study, the clusters are classified into three classes: balanced, overloaded, and unloaded. The bike loads during the working days and weekends are taken into account when analyzing the dataset; however, the authors only take into account the bike loads during working days. The study examined 121,709 journeys, encompassing 1,225 Velib stations during a single day. Nevertheless, this work lacks the analysis of bike loads during the weekends or the management of bike distribution to control the BSS systems.

The authors in Gao et al. (2020) classified the public bike stations of Yangcheng City using the DTW algorithm to study the spatial dynamics of

the bike-sharing system. To evaluate data with non-time series properties, the authors expanded the proposed method by identifying the data points of interest and developing a set of data format conversion rules. These aid in uncovering the correlation between the spatial and temporal attributes of cyclists and the types of land use. The dataset, which was obtained from the Yancheng local government's transport department, includes 424,581 valid trip records from 420 public bicycle stations, or one month's worth of travel. The dataset is partitioned into weekday and weekend data to examine both the temporal and spatial patterns of bike activities. However, this study does not examine the spatial distribution of stations, which is subject to user dispersion and meteorological conditions. Similarly, Lee and Leung (2023) examined the spatial and temporal patterns and the land usage of BSS in urban regions, mainly NYC. In addition, they employed machine learning models to forecast stations by analyzing the local characteristics. The data is obtained from the Citi Bike dataset, which displays distinct temporal patterns and is categorized into eight clusters. Extended travel times between stations and destinations, which impact the use of bike sharing, were not taken into account in this study, but only the shortest distance between a station and points of interest.

The authors in Li, Zhao, and Li (2019) conducted a study that analyzed a dataset of bike sharing over 3 months. The dataset included information from 572 bike stations obtained from the BSS in Chicago. Nevertheless, they focused on reducing the dimensionality of the raw data and identifying patterns from the time-series representations. The Discrete Wavelet Transform was utilized to eliminate errors generated from the raw time series. The DTW algorithm was subsequently employed to measure similarity among the clustered time series, thereby unveiling the fundamental features of the raw time series. The experimental results demonstrated that the DWT algorithm effectively reduced the overall size of the raw dataset by a factor of 4. This implies that there were only six data points representing daily usage as opposed to the 24 data points in the raw dataset. The clustering and analysis of the significant features of the multi-seasonal time series should be considered.

The emergence of free-floating (i.e., dockless) has increased the growth of shared bike fleets in China. These systems offer users a higher degree of freedom by eliminating the need to rent and return bikes at designated stations. In this study, Xu et al. (2019) employed time-series analysis and DTW approaches to investigate the issue of parking in the free-floating bike-sharing system (FFBS). The primary goal is to achieve a substantial spatial dispersion of Beijing's subway stations with varying levels of density. The time series was utilized to analyze parking densities, identify the FFBS parking pattern, and categorize these stations. The dataset used in this study comprises the rental data of 297 subway stations obtained from the Chinese Municipal Commission of Transport.

The authors in Tong et al. (2023) proposed using a DTW distance-based approach to identify the spatiotemporal patterns required to organize the use

of dockless station usage in a BSS. Additionally, the authors employed explainable-boosting machine learning to determine common factors among the various patterns. They examined urban regions in Wuhan that exhibit a significant volume of cycling activity, specifically roughly 88% of the city's overall utilization of shared bicycles. The dataset was derived from seven-day data collected between November 1 and November 7, 2019. The data was gathered during a period of favorable weather conditions and the absence of any notable events. The dataset was divided into six patterns for weekdays and four patterns for weekends that aid in the identification of issues, such as the inability to satisfy users' requests, to prioritize efforts, and to determine well-rounded bike usage strategies. Both studies require exploring correlations between specific events or activities happening in proximity to popular biking routes or stations during favorable weather periods outlined in the dataset collection period.

Similarly, the bike usage patterns are also studied for the Guangdong province, especially small and medium cities as reported in Wang, Wu, and Li (2019). The authors utilized a hierarchical clustering algorithm using DTW to extract the spatial data from station-based data, along with a random forest algorithm to evaluate the significant factors.

The authors in Zhao et al. (2019) introduced a methodology called density-based spatial clustering of applications with noise (DBSCAN) to determine patterns of bike usage using DTW. To analyze the spatiotemporal datasets, a data mining approach was utilized to find patterns of bike users' behavior in urban areas. The main goal was to change the dataset from space-time sequences to time-series sequences using time-series data mining techniques. This would then help sort the users' travel needs into groups. The Lambert equal-area grid was employed to partition the BSS into groups. Subsequently, demands in distinct areas were represented as a system where bike users can get incentives known as "bouns bikes," which would effectively mitigate bike accumulation in certain locations. However, further research is needed to enhance the analysis of data dimensionality and extract multiple features from the generated clusters. The authors in Hulot, Aloise, and Dominik Jena (2018) proposed building a specific model per station, not a cluster of station. The authors utilized a predictive model, i.e., ML-model, at the station level and then utilized a statistical model to produce the final prediction. This work of Hulot, Aloise, and Dominik Jena (2018) predicts the number of trips while ours predicts the trip's distance and time.

Clustering Techniques for BSS Datasets

Clustering datasets is an essential activity in BSS to get insights into utilization trends, enhance operational efficiency, and elevate service quality. Current approaches for clustering BSS datasets mostly rely on quantitative variables,

such as distance between stations, to redistribute the bikes. The efforts can be classified into seven main clustering criteria which are discussed in the following text.

First, temporal criteria include the process of grouping data based on various time intervals, such as different times of the day, weekdays vs weekends, and seasonal fluctuations. This approach aims to capture and analyze temporal use trends. The work in Li, Zhao, and Li (2019) conducted research that created a sophisticated pattern recognition model employing time-use data. This model was able to detect groups with similar daily activity patterns. This approach successfully captures the temporal dynamics of bike-sharing use, allowing for the identification of peak usage hours and seasonal fluctuations.

Second, geographical factors prioritize the physical positions of bike stations, and distance-based metrics are used to categorize stations that are near one another. The authors in Brown, Scott, and Páez (2022) used GPS trajectories to create a model for bike share traffic volumes, with a focus on spatial autocorrelation and the influence of physically segregated cycling infrastructure on traffic volumes. This spatial clustering methodology facilitates the optimization of station locations and infrastructure planning.

Third, the clustering method used in Hafezi, Liu, and Millward (2019), which is based on use patterns, encompasses factors such as the length of trips, the frequency of trips, and the intensity of usage. Dynamic Time Warping (DTW) is used to quantify similarities in time series data of Blind Source Separation (BSS). They employed k-means clustering and DTW barycenter averaging (DBA) for clustering. This approach efficiently decreases the number of dimensions and eliminates noise, revealing significant use patterns that are essential for optimizing the system.

Fourth, in Brandtzag, Heim, and Karahasanovi´c (2011), the authors concluded that user's characteristics, such as user classification, age, and gender, substantially impact the clustering of BSS information. A cluster analysis is conducted using survey data from Eurostat to identify five distinct user categories in Europe. This method emphasizes how individuals use the Internet, drawing a parallel to the various ways different demographic groups utilize bike-sharing systems. This comparison helps identify specific areas where service enhancements may be made.

Fifth, characteristics of the Station: Effective system management necessitates clustering based on station characteristics such as capacity, accessibility, and attractiveness. The authors in Lin, He, and Peeta (2018) conducted a study to examine the impact of weather conditions and calendar events on the utilization of bike-sharing services at various stations in Daejeon. The study showed that station-specific factors need to be taken into account to predict demand better and manage the system. This was done by putting stations with similar consumption patterns together and looking at how their different external influences affect them.

Sixth, characteristics of the trip, understanding user behavior relies heavily on trip features, such as origin-destination pairings and trip reasons. The work in Nair et al. (2019) examined urban cycling behavior by using GPS traces and available datasets to deduce trip aims and scrutinize route selections. This approach offers a comprehensive comprehension of riders' preferences, hence improving route design and optimizing system efficiency.

Finally, factors outside of a certain situation or context, weather conditions, and special events have a substantial impact on the utilization of bike-sharing. The authors in Kim (2018) proposed a method using a graph convolutional neural network to forecast the hourly demand at the station level, by considering the spatial and temporal relationships. The research emphasized the need to include weather and other external variables in demand forecast models.

The theoretical weakness in the current methods is two-fold. First, there is a lack of evidence in the literature investigating the impact of clustering BSS datasets on the prediction accuracy of predictive models. Second, there is no documented research on clustering BSS information by analyzing graphical patterns representing the frequency of travels between station pairs. This research gap motivated the current study, highlighting the need for a focused analysis using the DTW algorithm for clustering purposes. Consequently, addressing this gap is the primary aim of this work.

Methodology

The main motivation of this work is to improve the prediction accuracy of a given model. The justification is that the complete BSS dataset includes many patterns; thus, simplifying this complete dataset by separating different patterns would help improve the prediction accuracy. In this context, separating data with similar patterns into different clusters was proposed. Then, the predictive model will train and predict the cluster data separately. This idea should ease the predictive model's task, as the patterns to be used in training are similar. In theoretical machine learning, this approach is known as Mixture of Experts which was published in the highly cited paper (Jordan and Jacobs 1994). Thus, the idea is to divide the BSS dataset into subsets, each with similar patterns, and then one expert, i.e., model, will train on each subset.

To achieve this goal, presenting the number of trips over day hours between two stations with a curve was proposed. Then, a set of curves is obtained as each pair of stations will result in one curve. Next, the DTW was used to calculate the similarity between each two curves, i.e., two station pairs. This step will produce the distance matrix to be used to cluster the curves. In the following subsections, this proposed methodology will be exposed in detail.

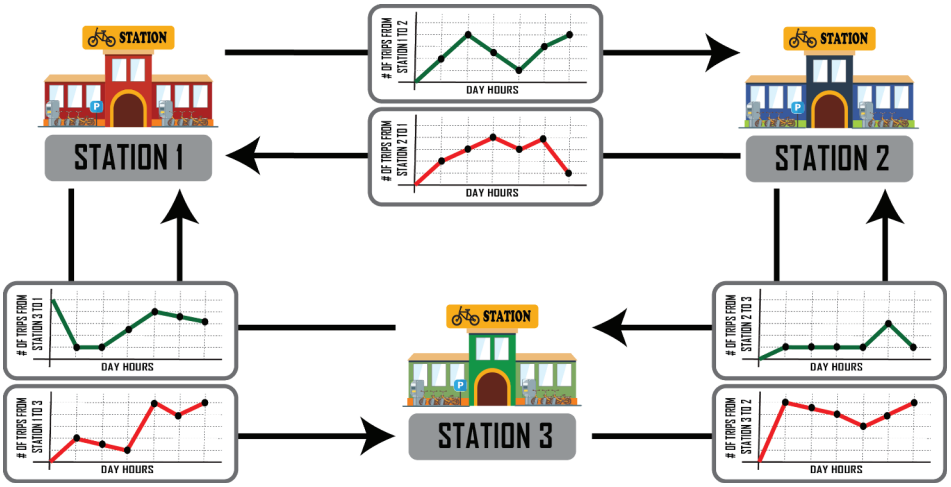
Overview

This method primarily examines the trip patterns between pairs of bike stations and measures their similarity by employing DTW. The DTW method is employed for this comparison, resulting in a quantification of similarity, i.e., distance matrix, between the curves of the station pair. Finally, the distance matrix is used to cluster the curves, as depicted in Figure 1. The approach efficiently groups the pairs of stations according to their usage patterns, which can subsequently be employed for predictive modeling. Through the examination of these trends, a more comprehensive comprehension of the dynamics of the utilization of bike-sharing systems in urban areas was gained.

In the proposed method, the BSS data are collected from a genuine BSS repository to ensure their quality and usability. The utilized dataset contains information about the user, including age and gender, as well as trip details. Subsequently, the proposed method divides the dataset-based calculated DTW distance matrix. This stage of data clustering permits the identification of distinct tendencies and patterns within each sub-dataset. In Table 1, a statistical summary is provided for the outcomes features, i.e., trip duration and trip distance.

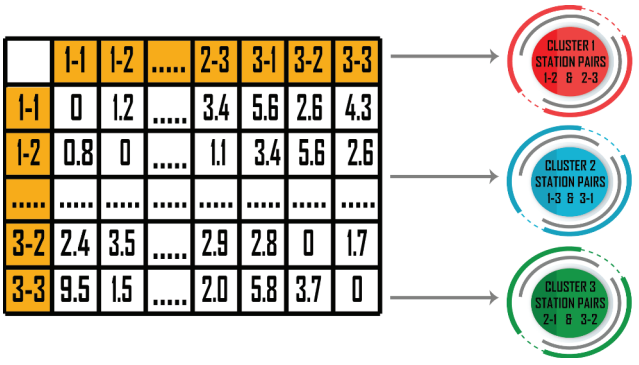
The proposed methodology can be summarized into the following phases:

- **Curve Formulation**: - In phase 1 of Figure 1a, the flow of bicycles between each pair of bike stations is shown by a time series curve, illustrating the movement throughout the day, i.e., 24 h. The curves are derived from bike trip data obtained from the BSS dataset, which records the number of bicycles traveling between any two stations over time. In the depicted curves, the x-axis represents the number of day hours, i.e., 0 to 23, and the y-axis represents the number of trips from station *a* to station *b*, where *a* and *b* can be any station number, e.g., 1, 2, etc.
- **DTW for curve similarity**: - In phase 2 of Figure 1b, the DTW algorithm is used to assess the similarity between these time series curves. DTW, unlike other algorithms, permits an elastic adjustment of the time axes, enabling the alignment of two sequences that may differ in speed or encounter time shifts. The DTW technique computes the distance between two curves by allowing one curve to be temporally stretched or compressed to align with the other curve. This is seen in the magnified inset, where the lines connecting the curves illustrate the process of "warping" that aligns corresponding points on the curves, even if they are temporally moved. Thus, the results of measuring the DTW distance between all possible pairs of stations are known as the distance matrix.
- **Clustering**: - In phase 3 of Figure 1c, the resultant similarity measures, i.e., DTW distance matrix, which assesses the degree of similarity in number of trip patterns between each pair of stations serves as the



(a) The first phase involves collecting trip patterns between station pairs and presenting them as curves over the day.

(b) The second phase used curves of phase to calculate the distance matrix of all curves using DTW.



(c) The third phase cluster similar curves.

Figure 1. The proposed methodology for enhancing the predictive performance of machine learning models in bike-sharing systems.

distance matrix for clustering the BSS dataset. This should facilitate the identification of clusters of station pairs exhibiting comparable use patterns, a critical factor for operational planning, such as the redistribution of bicycles to ensure supply and demand equilibrium. The technique uses the adaptability of DTW to examine intricate time-dependent patterns in BSS data, offering insights that are not immediately evident with more

Table 1. Statistical summary of the trip distance and trip duration of the utilized dataset.

	Trip distance	Trip duration
Mean	1.11e+03	9.34e+02
Standard Deviation	2.023e+04	4.06e+04
Min	0.00e+00	6.10e+01
Max	8.67e+06	2.03e+07

inflexible, linear approaches. The objective is to enhance the efficiency and user satisfaction of bike-sharing systems by improving their administration and planning.

Once the data has been clustered, three machine learning models, namely Random Forest, CatBoost, and Bagging, and a deep learning model (i.e., GRU model) are trained on each sub-dataset to predict the trip's distance and duration. This procedure entails the development of predictive models that are tailored to the particular characteristics of each sub-dataset. Following the model training phase, the performance of the proposed method is evaluated on each sub-dataset by testing the trained models to identify discrepancies and inconsistencies. Then, the most accurate model for each sub-dataset is reported, ensuring the most reliable and accurate outcomes possible. Then, the same machine learning models are trained on the complete dataset, and the trained models are evaluated as well. It is worth mentioning that, the reason behind choosing the aforementioned machine and deep learning models is that they achieved state-of-the-art performance in numerous applications (Fathalla et al. 2021; Fathalla, Salah, and Ali 2023; Indah Lestari et al. 2021; Salah et al. 2023, 2023).

Finally, the performance of the model trained on the full data set is compared to that of the same machine learning model trained separately on each sub-dataset. Comparing the two training approaches, if the model trained on the sub-dataset obtained higher prediction rates, then the proposed method enhanced the prediction task. The objective of this methodology is to enhance the predictive performance of machine learning models in bike-sharing systems. The proposed method contributes to more accurate and reliable predictions of journey distance and trip duration by taking into account the unique characteristics of different observation groups.

Problem Formulation

This sub-section provides a detailed and formal description of the problem concerning the computation of the distance between two temporal sequences. The problem is primarily focused on analyzing the utilization of a bike-sharing system throughout the day.

Problem Setup

Consider two vectors **a** and **b**, each of length n, which represent the number of bike journeys taken throughout each hour of the day. Therefore, the value of n is 24, and both vectors have equal dimensions. The main objective is to reduce the Euclidean distance between these two curves that are aligned in time using the DTW algorithm. The vectors **a** and **b** can be described as follows:

$$\mathbf{a} = (a_1, a_2, \ldots, a_n), \tag{1}$$

$$\mathbf{b} = (b_1, b_2, \ldots, b_n), \tag{2}$$

where a_i and b_i denote the number of trips during the *ith* hour of the day for vectors **a** and **b**, respectively.

Objective Function

The objective is to minimize the distance between these curves as aligned via DTW, represented formally as follows:

$$\text{Minimize } : D(\mathbf{a}, \mathbf{b}), \tag{3}$$

where $D(\mathbf{a}, \mathbf{b})$ is the total path cost to align the trajectories of **a** and **b**.

DTW Distance

The DTW algorithm computes the optimal match between two provided sequences under specific constraints. The sequences are non-linearly "warped" in the time dimension to find a measure of similarity unaffected by specific non-linear fluctuations in the time dimension. This is especially beneficial when comparing time series with dynamic speeds.

For two time series **a** and **b**, the DTW distance is defined by the recursive function:

$$D(i,j) = d(a_i, b_j) + \min(D(i+1, j), D(i+1, j+1), D(i, j+1)) \tag{4}$$

where $D(i,j)$ is the distance up to the i^{th} element of **a** and the j^{th} element of **b**, $d(a_i, b_j)$ is the Euclidean distance between the i^{th} and j^{th} elements of **a** and **b**, respectively. The recursion begins from $D(n,n)$ and proceeds backward to $D(1,1)$, and $D(0,0) = 0$ serves as the boundary condition. The calculated DTW distance between daily usage curves can help identify similar trends across many stations or days, which facilitates more accurate demand prediction and optimal resource distribution.

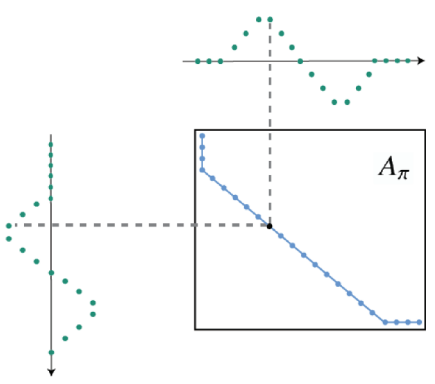
A Motivational Example for Curve Similarity in BSS

The DTW algorithm is a useful tool as it can collect and compare the changes in bike-sharing usage over time between different pairs of stations. This

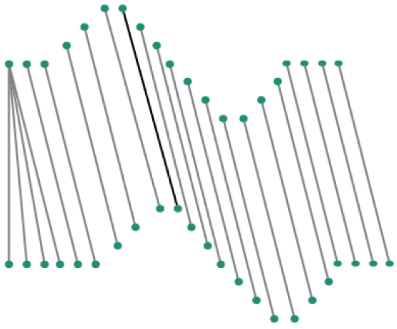
provides valuable information for improving the efficiency and user experience of bike-sharing systems. In Figure 2 the DTW algorithm's main two steps are visually depicted into two sub-figures.

The graphic in Figure 2a displays the concept of DTW applied to two-time series; the graph includes two-time series shown on a two-dimensional plane. First, the vertical axis, which represents one-time series, and second, the horizontal axis, which represents the other. According to the determination by DTW, the lines linking the points on these axes represent the ideal path of alignment that minimizes the total distance between these two series. Moving to Figure 2b, which illustrates a typical warping path associated with DTW. The figure shows the alignment of individual data points from two time series, accommodating differences in time scale and sequence. This includes aligning numerous points from one series to a single point in the second series.

To demonstrate the idea of curve similarity between bike stations, Figure 3 and 4 depict two curves with high and low similarity on the DTW distance, respectively. In Figure 3, the red line represents the number of trips over the day hours, i.e., 0 to 23, between two stations, namely, "City Hall" and "MLK Light Rail" while the green line represents the same data for two different stations (i.e., "Baldwin at Montgomery" and "Lafayette Park"). The proposed method includes measuring the distance, i.e., curve similarity, between the red and green curves of Figure 3, where each curve representing the number of trips between two stations over the day hours. Thus, the proposed method applies the DTW method to calculate the distance between the two time-series curves, i.e., the red and green curves



(a) Application of DTW to two time series, showing the optimal alignment path in a two-dimensional plane.



(b) Trajectory of warping for DTW, illustrating the alignment of multiple points from one time series with a single point from another.

Figure 2. The two main steps of the DTW algorithm.

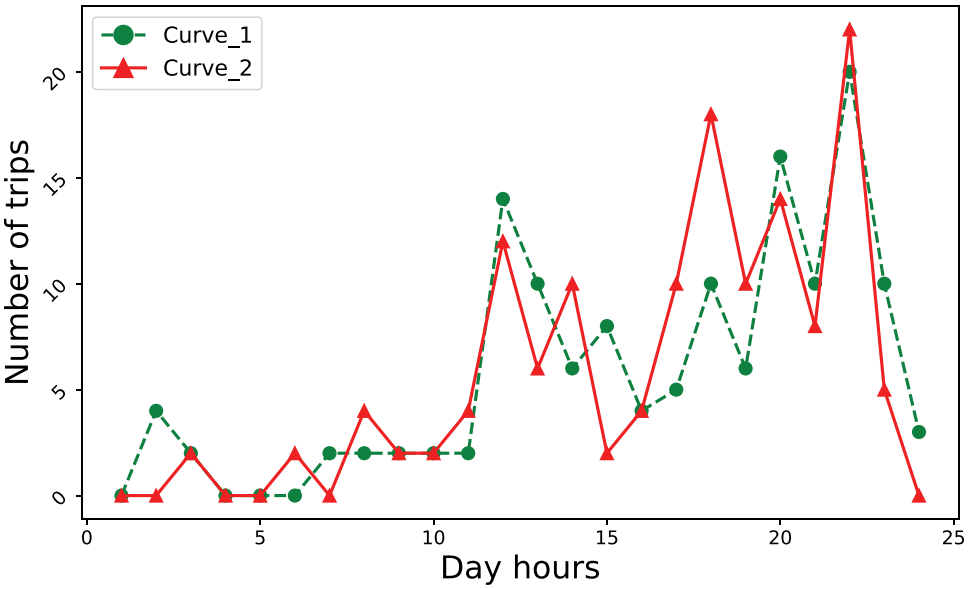


Figure 3. Two curves with low DTW distance and high similarity.

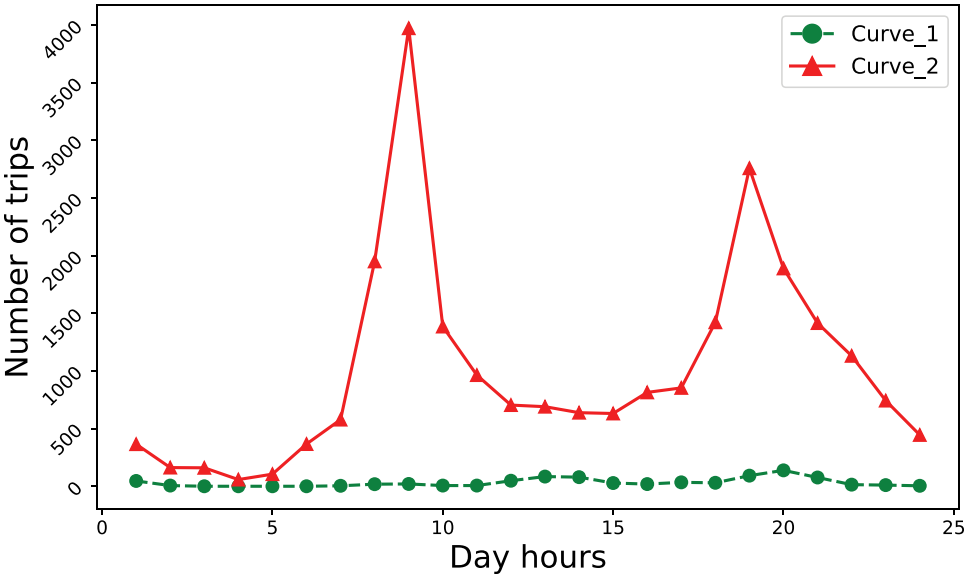


Figure 4. Two curves with high DTW distance and low similarity.

of Figure 3. The DTW similarity method measures the non-linear alignment between the curves by minimizing the cumulative distance of their points. The score obtained from the DTW algorithm represents how similar or different the two time-series curves are, with a lower score indicating higher similarity and a higher score indicating greater dissimilarity. The distance scored for the two curves of Figure 3 is 48 which indicating a very

high level of similarity. In contrary, applying the same explanation to Figure 4, the obtained distance between the two curves, i.e., red and green curves, is 22,435 which show a huge level of dissimilarity between the two curves of Figure 4.

Thus, Figure 3 shows that the two curves of Figure 3 will likely fall under the same sub-dataset (i.e., cluster) while the two curves of Figure 4 will likely to fall in different sub-datasets/clusters because of the high distance values of Figure 4's two curves. Of note, the DTW similarity score is calculated using the "similaritymeasures" library (https://pypi.org/project/similaritymeasures/).

The Proposed Algorithm

The proposed algorithm is listed in Algorithm 1. The *main* function in Algorithm 1 consists of four nested loops to iterate between all possible station pairs. Besides, the algorithm has two auxiliary functions, namely, *StationPairCurve* and *DTWDistance*. The *StationPairCurve* function calculates the curve representing the trips flow between stations x and y while The *DTWDistance* function calculates the DTW distance between two curves;

Algorithm 1: Analyzing Bike-Sharing System Data using the DTW.

```
Data: Bike-sharing dataset
Result: Similarity matrix for station pairs
Function Main():
    Load bike-sharing dataset
    Extract list of Stations
    Initialize an empty similarity matrix
    for s ∈ Stations do
        for t ∈ Stations do
            curve_st ← StationPairCurve(s, t)
            for a ∈ Stations do
                for b ∈ Stations do
                    curve_ab ← StationPairCurve(a, b)
                    distance((s, t), (a, b)) ← DTWDistance(curve_st, curve_ab)
                    Update similarity matrix with distance((s, t), (a, b))
                end
            end
        end
    end
    return similarity matrix
Function StationPairCurve(x, y):
    Extract travel data between stations x and y
    Formulate the curve representing daytime flow
    return curve
Function DTWDistance(curve1, curve2):
    Calculate the DTW distance between curve1 and curve2
    distance((s, t), (a, b)) ←
      DTW(stationPairCurve(s, t), stationPairCurve(a, b))
    return distance((s, t), (a, b))
```

each curve represents the trips flow between two stations. The main steps of Algorithm 1 can be summarized as follows:

- Step 1: In the first loop, the algorithm travels through each station. Let us designate s as the present station.
- Step 2: In the second loop, the algorithm iterates through each second station t and links it to station s to form the station pair (s, t). In the second loop, a curve representing the trip flow pattern for the station pair (s, t) is calculated.
- Step 3: In loops 3 and 4, the trip flow pattern for the station pair (a, b) is calculated as well. Then, for each station pair (s, t) and (a, b), the DTW distance should be measured, in the fourth loop.
- Step 4: The distance matrix is updated with the distance value of (s, t) and (a, b) which indicates the similarity between these pairs of stations.

Implementation Details

In this study, three machine learning models were used, namely Random Forest, CatBoost, Bagging, and one deep learning model (i.e., GRU), to predict trip distance and trip duration in BSSs. The selection of these models is due to their superior performance reported in recent works such as Guidon, Reck, and Axhausen (2020); Aydin, Erdem, and Cicek (2023); Tekouabou et al. (2021); Zhou et al. (2022) for Random Forest, CatBoost, Bagging, and GRU, respectively. The Random Forest model is known for being reliable and able to handle many different kinds of data. It is made up of many decision trees that work together to make accurate predictions. CatBoost is an innovative gradient-boosting model that excels at categorical features and generalization performance. The bagging model employs an ensemble of base learners trained on random subsets of the dataset to improve stability and reduce overfitting. By employing these cutting-edge machine learning models, the proposed approach ensures accurate predictions tailored to the unique characteristics of each user group, thereby enhancing the overall performance and reliability of predictive models in the context of bike-sharing systems.

The default Random Forest model hyperparameters are the number of trees in the forest, the maximum tree depth, and the minimum sample size needed to split an internal node. CatBoost model's default hyperparameters are the learning rate, the depth of the trees, and the number of iterations of the gradient boosting process. The Bagging model's default hyperparameters include the number of base estimators, the maximum number of samples for each base estimator, and the way the predictions of the base estimators are added together. Although set to their default values, these hyperparameters provided a solid foundation for the models and aided in achieving the desired level of predictive accuracy in the study.

The Mean Absolute Error (MAE) was employed as the accuracy metric for evaluating the performance of the machine learning models. In regression problems, MSE is a commonly employed metric for measuring the difference between the predicted and actual values. It computes the square root of the mean squared differences between the predicted and actual values, providing an interpretable and scale-sensitive performance metric for the model. Using RMSE, the study was able to compare the outcomes of the proposed approach to those of the standard training approach, ultimately demonstrating the superior performance of the proposed method in predicting trip distance and trip duration for bike-sharing systems.

The MAE is denoted for a set of N observations as follows:

$$MAE = \sum_{i=1}^{N} |x_i - y_i| \tag{5}$$

where x_i represents the predicted value and the y_i represents the true value.

Results and Discussion

Experimental Setup

The experiments were conducted on a computer running a 64-bit Windows 10 OS with two 2.6 GHz Intel 6-core processors. All of the utilized predictive models were implemented in the Python programming language version 3.9.16. Moreover, two Python packages are utilized to implement machine learning models, namely, Scikit-learn (Pajankar and Joshi 2022) and CatBoost (Dorogush, Ershov, and Gulin 2018) packages. Furthermore, the following libraries are utilized in the proposed work: *pandas* (Bantilan 2020), *Numpy* (Unpingco 2021), *Matplotlib* (Cao et al. 2021), and *similaritymeasures 1.1.0*. The utilized clustering algorithm is *KMedoids* (Schubert and Rousseeuw 2019).

Dataset

In this paper, two datasets were utilized. The first dataset is the New York CityBike Share Dataset. It is a publicly available dataset that affords bike trips throughout the boroughs of New York City (NYC). This dataset (New York City Bike Share Dataset, 2023) contains information on 735,502 anonymized trips collected between January 2015 and June 2017. In the following text, this dataset is called "dataset-1."

Dataset-1 consists of 17 features. Four of these features (i.e., Start time, End time, Trip duration, and Trip duration in minutes) were used to label the first

predictive task, i.e., trip duration. For the trip distance label generation, another four features were utilized, namely, start station latitude, start station longitude, end station latitude, and end station longitude. Then, the dataset includes the station ID and station name features; a station's name was picked to represent these two features in data splitting. The user type, gender, and birth year features were selected to split the data as well. The only unused feature in data splitting was Bike ID, as it can produce a huge number of groups.

The second dataset is the Citi Bike Sharing of New York City over the period of January 2020 to April 2021. This dataset includes 15 features, and the number of records is 393,312. In the following text, this dataset is called "dataset-2." By utilizing two datasets, the aim of this study is to try two datasets of different numbers of features and different sizes to evaluate the proposed method. Dataset-2 is utilized to analyze the BSS usage patterns during the abnormal period of the COVID-19 pandemic. Thus, the two datasets include covering the normal and abnormal patterns of bike usage.

In all of the experiments, the following features were utilized as an input to the predictive models: trip start station, time stamp of trip's starting time, user type, user's age, and user's gender. The "User type" attribute's value can be customer or subscriber. The intuition of including the users' gender and age as input features to the model is that they can help the model in the prediction task. These features can significantly influence riding patterns, trip duration, and preferred routes (i.e., distance). For instance, age may correlate with varying levels of physical activity and stamina, impacting the length and frequency of bike trips. Similarly, gender differences can shed light on distinct preferences and behaviors in cycling, such as the choice of biking routes or trip timing. Both datasets users have age and gender data.

The utilized data split rate for all experiments is 80%–20% for the training and test sets, respectively. All of the reported results are for the test sets only. The baseline results used to report the achieved improvements are the prediction evaluation metrics of the predictive models trained on the complete dataset. For the proposed method results, the reported improvements were calculated by taking the average evaluation scores of the k models trained on the k clusters/sub-datasets and then comparing them to the evaluation score of the model trained on the complete dataset.

Results

The obtained results outline the changes in the predictive accuracy of trip duration and distance. Hence, various machine learning models were utilized (RF, Bagging Regressor, CatBoost, and GRU) to examine the performance of the proposed method due to the superior performance in relevant machine

learning applications (Akilandesvari Ramesh et al. 2021; Ve and Cho 2024; Zhou et al. 2022).

The changes include the MAE metric values and the percentage of the change, where the positive values mean an improvement and the negative values mean a decline of the predictive model. The results listed in Table 2 show the improvement of baseline MAE values for three ML models, namely, RF, Bagging Regressor, and CatBoost, and a deep learning model (GRU). Of note, the best MAE value was achieved with five clusters for the ML models and four clusters for the GRU model. The changes in the MAE values of predicting the bike trip duration are listed as percentages in Table 3 and the same percentages are depicted in Figure 5.

Table 4 lists four MAE values of four baseline models for predicting the bike trip distance, in the second column. In the third to 11th columns, the MAE values for each model are listed, but using different numbers of clusters varying from 2 clusters to 10 clusters. Then, the percentage of changes in the MAE values using different numbers of clusters relative to the baseline MAE values are listed in Table 5 and depicted in Figure 6.

In Figure 6, the number of clusters, k, is a parameter in the proposed method that significantly impacts the overall performance of this proposed method. The k is a hyperparameter that should be tuned based on the nature of the dataset. Changing the number of clusters alters the dataset's structure, which subsequently affects the final results. Just as a dataset can have multiple local optima and one global optimum of improvement, the number of peaks in a clustering analysis can vary. Figure 6 with two peaks aims to visually identify and select the global peak for optimal clustering.

Table 2. A comparison of the trip duration prediction on the MAE metric for different ML models on different numbers of clusters for dataset-1.

		Number of Clusters								
Model	Baseline	2	3	4	5	6	7	8	9	10
RF	11.02	7.93	6.23	**5.96**	6.67	6.56	7.58	6.54	6.07	9.44
BaggingReg	11.43	8.09	6.46	**6.15**	6.43	6.29	7.52	6.71	6.52	8.60
CatBoost	18.49	15.15	12.53	**10.77**	12.50	16.33	18.20	17.00	14.57	20.29
GRU	12.94	12.85	10.50	**9.20**	10.81	12.29	11.04	13.16	12.16	15.32

Table 3. Percentage improvement of trip duration prediction on the MAE metric using the proposed method for dataset-1.

	Number of Clusters								
Model	2	3	4	5	6	7	8	9	10
RF	28%	43%	46%	**39%**	40%	31%	41%	45%	14%
BaggingReg	29%	44%	46%	**44%**	45%	34%	41%	43%	25%
CatBoost	18%	32%	42%	**32%**	12%	2%	8%	21%	−10%
GRU	1%	19%	**29%**	16%	0%	0%	−2%	6%	−18%

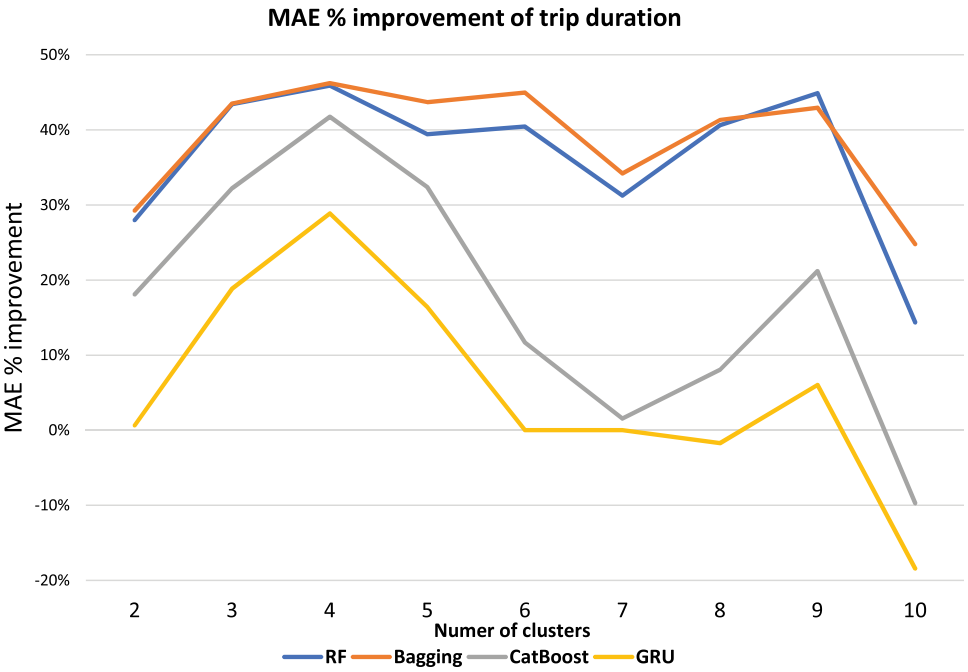


Figure 5. Percentage improvement of trip duration prediction on the MAE metric using the proposed method for dataset-1.

Table 4. A comparison of the trip distance prediction on the MAE metric for different ML models on different number of clusters for dataset-1.

		Number of Clusters								
Model	Baseline	2	3	4	5	6	7	8	9	10
RF	97.75	78.33	67.72	58.94	**53.93**	67.78	66.65	65.89	62.35	62.37
BaggingReg	106.06	84.82	72.82	62.41	**57.44**	71.67	70.03	69.16	65.39	65.47
CatBoost	223.23	184.05	158.31	132.59	**122.38**	150.06	148.22	146.61	143.67	145.91
GRU	310.68	292.63	256.60	**233.94**	238.43	272.43	284.10	292.08	310.93	315.08

Table 5. Percentage improvement of trip distance prediction on the MAE metric using the proposed method for dataset-1.

	Number of Clusters								
Model	2	3	4	5	6	7	8	9	10
RF	20%	31%	40%	**45%**	31%	32%	33%	36%	36%
BaggingReg	20%	31%	41%	**46%**	32%	34%	35%	38%	38%
CatBoost	18%	29%	41%	**45%**	33%	34%	34%	36%	35%
GRU	6%	17%	**25%**	23%	12%	9%	6%	0%	−1%

Table 4 presents a comparison of the MAE for several ML models as the number of clusters varies for dataset-1. The Random Forest (RF) model initially has a baseline MAE of 97.75. As the number of clusters rises, the model consistently improves, reaching its lowest MAE of 62.37 when using 10 clusters. The Bagging Regressor (BaggingReg) model has a larger initial MAE

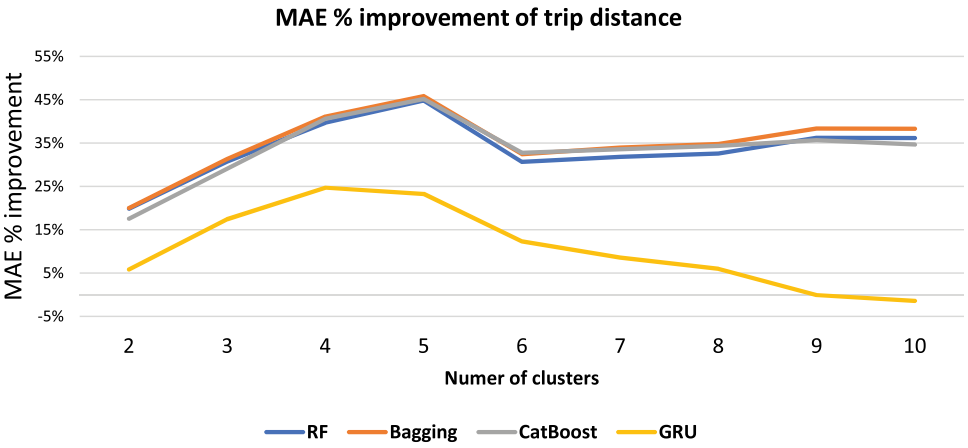


Figure 6. Percentage improvement of trip distance prediction on the MAE metric using the proposed method for dataset-1.

compared to the RF model. However, it shows a similar pattern of improvement and achieves its lowest MAE of 65.47 when using 10 clusters. CatBoost model's initial MAE is 223.23, which serves as the baseline. As the number of clusters rises, there is a notable decrease in MAE. The optimal performance is achieved when there are five clusters, resulting in an MAE of 122.38. However, adding additional clusters leads to a minor increase in error.

The GRU model has the greatest initial MAE value of 310.68. However, the MAE lowers with an increase in the number of clusters, albeit not as much as the other models. The enhancement reaches a peak at around five clusters, after which further increases in the number of clusters only lead to negligible improvements in MAE.

Table 5 displays the percentage improvement achieved in predicting trip distance for dataset-1. Table 5 displays the percentage enhancement in the MAE metric for the same models, in comparison to the baseline for dataset-1. The RF: model demonstrates a consistent enhancement, with the percentage rising as more clusters are used, reaching 45% at five clusters, and retaining a comparable improvement as the number of clusters increases to 10. The The Bagging Regressor, like the Random Forest, exhibits a 46% improvement with 5 clusters and maintains a steady 38% improvement with 10 clusters. The CatBoost model has the greatest percentage enhancement at 4 and 5 clusters, with improvements of 41% and 45%, respectively. However, the degree of increase diminishes gradually with the addition of additional clusters. Finally, the GRU model's degree of enhancement is less significant compared to other models, beginning at 6% with two clusters and reaching its highest point of 25% with four clusters. Furthermore, the enhancement diminishes and becomes negative when there are 10 clusters, suggesting a decline in performance compared to the initial state. Thus, the best achieved improvements in

the prediction accuracy are 44% and 46% for both trip duration and distance, respectively.

The reported results for dataset-2 are for the RF, CatBoost, and Bagging models only; the GRU model was excluded due to the smaller dataset size relative to dataset-1. For dataset-2, the percentages of improvements in the MAE metric for trip duration and distance prediction were depicted in Figure 7 and 8, respectively. In Figure 7, the improvement of the MEA value started from six clusters. The performance of the three models was close except for 10 clusters, the CatBoost clearly outperformed the other two methods. In Figure 8, the pattern of trip distance prediction improvement is increasing as the number of cluster increases. The Bagging and RF models have slightly performed better than the CatBoost model.

For dataset-2, the conclusion reached was that the proposed method achieved the lowest improvement for predicting the trip duration, out of the two datasets especially for a small number of clusters. This may be linked to varying levels of impact on bike riders due to health issues related to the

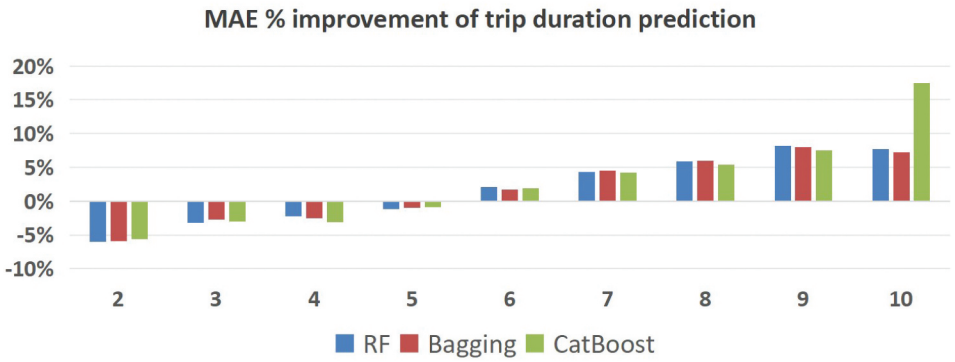


Figure 7. Percentage improvement of trip duration prediction on the MAE metric using the proposed method for dataset-2.

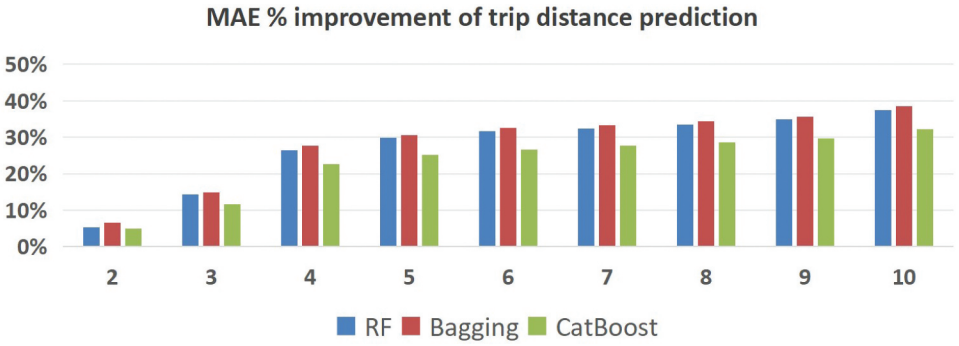


Figure 8. Percentage improvement of trip distance prediction on the MAE metric using the proposed method for dataset-2.

COVID-19 virus. On the other hand, the improvements in the trip distance prediction for the MAE metric are similar to dataset-1, as the distance is fixed between stations for the two datasets. The best achieved improvements for predicting trip duration and distance for dataset-2 are 17% and 38%, respectively. Thus, the average best achieved improvements of the two datasets are 30% and 42% for trip duration and distance, respectively.

Based on these aforementioned results in tables and figures, the following findings can be deduced:

Impact of Clustering

Integrating clustering often enhances the performance of all models. However, the proper number of clusters should be carefully selected, as for certain cases the improper number of cluster can negatively impact the results.

Optimal Number of Clusters

The proposed method shows varying levels of improvement based on the patterns in the dataset. The selection of the appropriate k value, i.e., the number of clusters, significantly impacts the improvement achieved by the proposed method.

Comparing Complex Models to Simple Models

The GRU model, a more complex model, has worst performance compared to simpler models such as RF and Bagging Regressor. This implies that the dataset may not need the complex capabilities provided by GRU to achieve accurate prediction.

Model Selection

When choosing a deployment model, it is essential to take into account not only the percentage of improvement but also the MAE values. An ideal model would have a reduced baseline error and demonstrate steady progress as more clusters are added. These findings are significant for individuals seeking to enhance bike-sharing operations since they emphasize the efficacy of clustering approaches in combination with diverse machine learning models for forecasting trip lengths.

Different Patterns of Datasets

Dataset-2 includes different bike riding patterns due to the COVID-19 pandemic in comparison with to dataset-1. Thus, the pattern of improvements for dataset-2 is slightly different relative to dataset-1. This emphasizes the finding of the research work (Basak et al. 2023; Heydari, Konstantinoudis, and Wahid Behsoodi 2021; Nikiforiadis, Ayfantopoulou, and Stamelou 2020).

Conclusion

To improve the trip distance and duration prediction accuracy of the bike-sharing systems, this work proposed utilizing the DTW method to split the BSS dataset into sub-datasets based on the similarity of curves representing the time-series curve of the number of trips between station pairs. In other words, the primary objective of this study was to investigate whether predictive models for BSS could benefit from training on a set of sub-datasets with more similar data patterns, yielded by the proposed method, rather than training on the complete dataset. To achieve this goal, the proposed method was thoroughly tested on two real datasets of New York City BSS at different periods of time. Several predictive models were trained on the complete BSS dataset, and then those baseline models were compared with the predictive models trained on divided BSS sub-datasets by the proposed method. The metric utilized to assess accuracy was the MAE. The results demonstrate significant improvements in trip duration and distance prediction accuracy rates. In addition, the obtained results demonstrate that the proposed method significantly reduces the average MAE values of the used predictive models for both the trip duration and distance prediction. These findings hold promise for enhancing the predictive capabilities of bike-sharing systems, potentially leading to more efficient and user-friendly services. Future research directions include evaluating the proposed method in free-floating bike-sharing systems and exploring its applicability in other domains.

Note

1. https://github.com/Ahmed-Fathalla/BSS-using-DTW.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

The authors extend their appreciation to the Deanship for Research & Innovation, Ministry of Education in Saudi Arabia for funding this research work through the project number [IF2/PSAU/2022/01/20968].

ORCID

Ahmed Ali http://orcid.org/0000-0003-2775-4104
Ahmad Salah http://orcid.org/0000-0003-3433-7640
Mahmoud Bekhit http://orcid.org/0000-0002-8115-4233
Ahmed Fathalla http://orcid.org/0000-0001-5432-5407

Data availability statement

The data that support the findings of this study are available from the corresponding author, [Ahmed Ali], upon reasonable request.

References

Adel, B., A. Badran, N. E. Elshami, A. Salah, A. Fathalla, and M. Bekhit. 2022. A survey on deep learning architectures in human activities recognition application in sports science, healthcare, and security. In *The International Conference on Innovations in Computing Research* Athens, Greece, 121–134. Springer.

Akilandesvari Ramesh, A., S. Pavani Nagisetti, N. Sridhar, K. Avery, and D. Bein. 2021. Station-level demand prediction for bike-sharing system. In *2021 IEEE 11th Annual Computing and Communication Workshop and Conference (CCWC)* NV, USA, 0916–21. IEEE.

Alaoui, E. A.and Tekouabou, S. C. K. 2021. Intelligent management of bike sharing in smart cities using machine learning and internet of things. *Sustainable Cities and Society* 67:102702. doi: 10.1016/j.scs.2020.102702.

Albuquerque, V., M. Sales Dias, and F. Bacao. 2021. Machine learning approaches to bike-sharing systems: A systematic literature review. *ISPRS International Journal of Geo-Information* 10 (2):62. doi: 10.3390/ijgi10020062.

Ali, A., A. Fathalla, A. Salah, M. Bekhit, E. Eldesouky, and A. M. Khalil. 2021. Marine data prediction: an evaluation of machine learning, deep learning, and statistical predictive models. *Computational Intelligence and Neuroscience* 2021 (1). doi: 10.1155/2021/8551167.

Ashqar, H. I., M. Elhenawy, H. A. Rakha, M. Almannaa, and L. House. 2022. Network and station-level bike-sharing system prediction: A san francisco bay area case study. *Journal of Intelligent Transportation Systems* 26 (5):602–612. doi: 10.1080/15472450.2021.1948412.

Aydin, Z. E., B. I. Erdem, and Z. I. E. Cicek. Prediction bike-sharing demand with gradient boosting methods Pamukkale University Journal of Engineering Sciences 29 8 824–832 2023 https://dergipark.org.tr/tr/download/article-file/3630651 doi:10.5505/pajes.2023.39959.

Bantilan, N. 2020. pandera: Statistical data validation of pandas dataframes. In *Proceedings ofthe Python in Science Conference (SciPy)*, 116–24.

Basak, E., R. Al Balawi, S. Fatemi, and A. Tafti. 2023. When crisis hits: Bikesharing platforms amid the COVID-19 pandemic. *PLOS ONE* 18 (4):e0283603. doi: 10.1371/journal.pone.0283603.

Berndt, D. J., and J. Clifford. 1994. Using dynamic time warping to find patterns in time series. In *Proceedings of the 3rd International Conference on Knowledge Discovery and Data Mining* Seattle, Washington, USA, 359–370.

Brandtzag, P. B., J. Heim, and A. Karahasanović. 2011. Understanding the new digital divide — a typology of internet users in europe. *International Journal of Human-Computer Studies* 69 (3):123–138. doi: 10.1016/j.ijhcs.2010.11.004.

Brown, M. J., D. M. Scott, and A. Páez. 2022. A spatial modeling approach to estimating bike share traffic volume from gps data. *Sustainable Cities and Society* 76:103401. doi: 10.1016/j.scs.2021.103401.

Butt, M. A., S. Danjuma, M. S. B. Ilyas, U. Muneer Butt, M. Shahid, and I. Tariq. 2023. Demand prediction on bike sharing data using regression analysis approach. *Journal of Innovative Computing and Emerging Technologies* 3 (1). doi: 10.56536/jicet.v3i1.52.

Caggiani, L., R. Camporeale, Z. Hamidi, and C. Zhao. 2021. Evaluating the efficiency of bike-sharing stations with data envelopment analysis. *Sustainability* 13 (2):881,765. doi: 10.3390/su13020881.

Cao, S., Y. Zeng, S. Yang, and S. Cao. 2021. Research on python data visualization technology. *Journal of Physics: Conference Series* 1757(1):012122.

Chabchoub, Y., and C. Fricker. 2014. Classification of the vélib stations using kmeans, dynamic time warping and dba averaging method. In *2014 International Workshop on Computational Intelligence for Multimedia Understanding (IWCIM)* Paris, France, 1–5. IEEE.

Chen, P.-C., H.-Y. Hsieh, K.-W. Su, X. Kharis Sigalingging, Y.-R. Chen, and J.-S. Leu. 2020. Predicting station level demand in a bike-sharing system using recurrent neural networks. *IET Intelligent Transport Systems* 14 (6):554–561. doi: 10.1049/iet-its.2019.0007.

Cheng, L., J. Yang, X. Chen, M. Cao, H. Zhou, and Y. Sun. 2020. How could the station-based bike sharing system and the free-floating bike sharing system be coordinated? *Journal of Transport Geography* 89:102896. doi: 10.1016/j.jtrangeo.2020.102896.

Dong, H., G. Singh, A. Attri, and A. El Saddik. 2016. Open data-set of seven Canadian cities. *Institute of Electrical and Electronics Engineers Access* 5:529–543. doi: 10.1109/ACCESS.2016.2645658.

Dorogush, A. V., V. Ershov, and A. Gulin. Catboost: Gradient boosting with categorical features support. arXiv preprint arXiv:1810.11363, 2018.

Dubey, M., A. Peral Ortiz, R. Agrawal, and A. G. Forbes. 2019. Predicting biker den sity at bikeshare station intersections in San Francisco. In *2019 IEEE Global Humanitarian Technology Conference (GHTC)* Seattle, WA, USA, 1–7. IEEE.

El-Assi, W., M. Salah Mahmoud, and K. Nurul Habib. 2017. Effects of built environment and weather on bike sharing demand: a station level analysis of commercial bike sharing in toronto. *Transportation* 44 (3):589–613. doi: 10.1007/s11116-015-9669-z.

Fathalla, A., A. Salah, and A. Ali. 2023. A novel price prediction service for e-commerce categorical data. *Mathematics* 11 (8):1938. doi: 10.3390/math11081938.

Fathalla, A., A. Salah, M. Ali Mohamed, N. Indah Lestari, and M. Bekhit. 2021. A novel dual prediction scheme for data communication reduction in iot-based monitoring systems. In *International Conference on Internet of Things as a Service* Sydney, Australia, 208–20. Springer.

Fathalla, A., A. Salah, M. Bekhit, E. Eldesouky, A. Talha, A. Zenhom, A. Ali, and V. Memmolo. 2023. Real-time and automatic system for performance evaluation of karate skills using motion capture sensors and continuous wavelet transform. *International Journal of Intelligent Systems* 2023 (1). doi: 10.1155/2023/1561942.

Gao, Z., S. Wei, L. Wang, and S. Fan. 2020. Exploring the spatial-temporal characteristics of traditional public bicycle use in Yancheng, china: a perspective of time series cluster of stations. *Sustainability* 12 (16):6370. doi: 10.3390/su12166370.

Guidon, S., D. J. Reck, and K. Axhausen. 2020. Expanding a(n) (electric) bicycle-sharing system to a new city: prediction of demand with spatial regression and random forests. *Journal of Transport Geography* 84:102692. doi: 10.1016/j.jtrangeo.2020.102692.

Hafezi, M. H., L. Liu, and H. Millward. 2019. A time-use activity-pattern recognition model for activity-based travel demand modeling. *Transportation* 46 (4):1369–1394. doi: 10.1007/s11116-017-9840-9.

Heydari, S., G. Konstantinoudis, and A. Wahid Behsoodi. 2021. Effect of the covid-19 pandemic on bike-sharing demand and hire time: Evidence from santander cycles in london. *PLOS ONE* 16 (12):e0260969. doi: 10.1371/journal.pone.0260969.

Hulot, P., D. Aloise, and S. Dominik Jena. 2018. Towards station-level demand prediction for effective rebalancing in bike-sharing systems. In *Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining* London, United Kingdom, 378–386.

Indah Lestari, N., M. Bekhit, M. Ali Mohamed, A. Fathalla, and A. Salah. 2021. Machine learning and deep learning for predicting indoor and outdoor iot temperature monitoring

systems. In *International Conference on Internet of Things as a Service* Sydney, Australia, 185–197. Springer.

Jiang, S., and Z. Chen. 2023. Application of dynamic time warping optimization algorithm in speech recognition of machine translation. *Heliyon* 9 (11):e21625. doi: 10.1016/j.heliyon.2023.e21625.

Jordan, M. I., and R. A. Jacobs. 1994. Hierarchical mixtures of experts and the em algorithm. *Neural computation* 6 (2):181–214. doi: 10.1162/neco.1994.6.2.181.

Kim, K. 2018. Investigation on the effects of weather and calendar events on bike-sharing according to the trip patterns of bike rentals of stations. *Journal of Transport Geography* 66:309–320. doi: 10.1016/j.jtrangeo.2018.01.001.

Kowdiki, M., and A. Khaparde. 2021. Automatic hand gesture recognition using hybrid meta heuristic-based feature selection and classification with dynamic time warping. *Computer Science Review* 39:100320. doi: 10.1016/j.cosrev.2020.100320.

Kumar Das, A., A. Manoj Joshi, and S. Dhal. 2020. A machine learning based bike recommendation system catering to user's travel needs. In *2020 IEEE 17th India Council International Conference (INDICON)* New Delhi, India, 1–6. IEEE.

Lee, C. K. H., and E. K. H. Leung. 2023. Spatiotemporal analysis of bike-share de mand using dtw-based clustering and predictive analytics. *Transportation Research Part E: Logistics and Transportation Review* 180:103361. doi: 10.1016/j.tre.2023.103361.

Li, D., Y. Zhao, and Y. Li. 2019. Time-series representation and clustering approaches for sharing bike usage mining. *Institute of Electrical and Electronics Engineers Access* 7:177856–-177863. doi: 10.1109/ACCESS.2019.2958378.

Li, X., Y. Xu, Q. Chen, L. Wang, X. Zhang, and W. Shi. 2021. Short-term forecast of bicycle usage in bike sharing systems: a spatial-temporal memory network. *IEEE Transactions on Intelligent Transportation Systems* 23 (8):10923–10934. doi: 10.1109/TITS.2021.3097240.

Li, Y., and Y. Zheng. 2019. Citywide bike usage prediction in a bike-sharing system. *IEEE Transactions on Knowledge and Data Engineering* 32 (6):1079–1091. doi: 10.1109/TKDE.2019.2898831.

Lin, L., Z. He, and S. Peeta. 2018. 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. doi: 10.1016/j.trc.2018.10.011.

Nair, S., K. Javkar, J. Wu, and V. Frias-Martinez. 2019. Understanding cycling trip purpose and route choice using gps traces and open data. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 3 (1):1–26. doi: 10.1145/3314407.

New York City Bike Share Dataset. 2023. https://www.kaggle.com/akkithetechie/890new-york-city-bike-share-dataset.

Nikiforiadis, A., G. Ayfantopoulou, and A. Stamelou. 2020. Assessing the impact of COVID-19 on bike-sharing usage: the case of Thessaloniki, Greece. *Sustainability* 12 (19):8215. doi: 10.3390/su12198215.

Otero, I., M. J. Nieuwenhuijsen, and D. Rojas-Rueda. 2018. Health impacts of bike sharing systems in Europe. *Environment International* 115:387–394. doi: 10.1016/j.envint.2018.04.014.

Pajankar, A., and A. Joshi. 2022. Introduction to machine learning with scikit-learn. In *Hands-on machine learning with Python: Implement neural network solutions with scikit learn and PyTorch*, 65–77. USA: Apress Berkeley, CA.

Rudloff, C., and B. Lackner. 2014. Modeling demand for bikesharing systems: neighboring stations as source for demand and reason for structural breaks. *Transportation Research Record* 2430 (1):1–11. doi: 10.3141/2430-01.

Salah, A., M. Bekhit, E. Eldesouky, A. Ali, and A. Fathalla. 2023. Price prediction of seasonal items using time series analysis. *Computer Systems Science and Engineering* 46 (1):445–460. doi: 10.32604/csse.2023.035254.

Salah, A., A. Fathalla, E. Eldesouky, W. Li, A. M. Mahmoud Ibrahim, and A. Hošovský. 2023. Forecasting the friction coefficient of rubbing zirconia ceramics by titanium alloy. *International Journal of Intelligent Systems* 2023 (1). doi: 10.1155/2023/6681886.

Schubert, E., and P. J. Rousseeuw. 2019. Faster k-medoids clustering: Improving the pam, clara, and clarans algorithms. In *Similarity Search and Applications: 12th International Conference, SISAP 2019,* October 2–4, 2019. Newark, NJ, USA, vol. 12, 171–187. Springer.

Tong, Z., Y. Zhu, Z. Zhang, R. An, Y. Liu, and M. Zheng. 2023. Unravel the spatio temporal patterns and their nonlinear relationship with correlates of dockless shared bikes near metro stations. *Geo-Spatial Information Science* 26 (3):577–598. doi: 10.1080/10095020.2022.2137857.

Unpingco, J. 2021. Numpy. In *Python programming for data analysis*, 103–126. Switzerland: Springer, Cham.

Ve, S., and Y. Cho. 2024. Season wise bike sharing demand analysis using random forest algorithm. *Computational Intelligence* 40 (1):e12287. doi: 10.1111/coin.12287.

Wang, B., and I. Kim. 2018. Short-term prediction for bike-sharing service using machine learning. *Transportation Research Procedia* 34:171–178. doi: 10.1016/j.trpro.2018.11.029.

Wang, L. Y., J. S. Wu, and W. F. Li. 2019. Usage patterns and driving mechanisms of public bicycle systems in small and medium-sized cities based on space-time data mining. *Journal of Geo-Information Science* 21 (1):25–41. doi: 10.1016/j.ins.2019.05.015.

Wang, W. 2016. Forecasting bike rental demand using new york citi bike data.

Xu, D., Y. Bian, J. Rong, J. Wang, and B. Yin. 2019. Study on clustering of free-floating bike-sharing parking time series in Beijing subway stations. *Sustainability* 11 (19):5439. doi: 10.3390/su11195439.

Yang, H., K. Xie, K. Ozbay, Y. Ma, and Z. Wang. 2018. Use of deep learning to predict daily usage of bike sharing systems. *Transportation Research Record* 2672 (36):92–102. doi: 10.1177/0361198118801354.

Zhang, G., H. Yang, S. Li, Y. Wen, Y. Li, and F. Liu. 2019. What is the best catchment area of bike share station? A study based on divvy system in Chicago, usa. In *2019 5th International Conference on Transportation Information and Safety (ICTIS)*. Liverpool, UK, 1226–1232. IEEE.

Zhao, X., C. Hu, Z. Liu, and Y. Meng. 2019. Weighted dynamic time warping for grid-based travel-demand-pattern clustering: Case study of Beijing bicycle-sharing system. *ISPRS International Journal of Geo-Information* 8 (6):281. doi: 10.3390/ijgi8060281.

Zhou, S., C. Song, T. Wang, X. Pan, W. Chang, and L. Yang. 2022. A short-term hybrid tcn-gru prediction model of bike-sharing demand based on travel characteristics mining. *Entropy* 24 (9):1193. doi: 10.3390/e24091193.

Zhou, Y., Q. Li, X. Yue, J. Nie, and Q. Guo. 2022. A novel predict-then-optimize method for sustainable bike-sharing management: a data-driven study in china. *Annals of Operations Research* 1–33. doi: 10.1007/s10479-022-04965-0.