Insights on exploring a small-scale electric bicycle sharing data set

Shyam Sundar Rampalli, Andreas Keler, Georgios Grigoropoulos

Abstract— **Cities can benefit to a great extent from the deployment of bicycle-sharing services (BSS). Conceived as being an alternative for taxi services and public transport, BSSs are used by daily commuters, as well as tourists. Open data sharing policies in the US allow data scientists to make use of relatively detailed and anonymized data extracts of often station-based, bicycle-sharing services. Origin destination matrices (OD matrices) allow the representation of flows while leaving out detailed information on the specific trajectory and the traffic control that influences individual movements. This paper aims at reviewing the existing methods for analyzing sharing services that operate predominantly using GPS data. We aim to use unsupervised clustering learning techniques to analyze GPS data and develop insights. We found that combining more than one clustering technique is more effective as compared to individual techniques. We discuss the merits and demerits of individual and combined techniques and their relevance in analyzing bike-sharing GPS data.**

I. Introduction

Bicycle usage patterns can help us understand the city's behaviour and geography [1]. City planners could understand the movement of people within and in-out of the city to a certain extent. Bicycle usage patterns of a dockless bicycle-sharing service are different from a docked bicycle-sharing service [2]. A new domain of bicycle sharing is electric bicycle sharing which is predominantly docked [3]. This research is on analysis of GPS data from a dockless electric bicycle sharing service implemented in a university campus area. Much of this work is on the application of unsupervised machine learning techniques where each technique is applied followed by a discussion with an emphasis on transport. Clusters, which correspond to the locations of interest or hotspots, formed with these techniques are related to different modes of transport. Key performance indicators have been defined for such service with both an operational and performant perspective.

II. State of the Art

In 2021, more than 1,900 different bike-sharing services were operating worldwide with around 9.5 million used bikes [4]. One of the advantages, these services provide is options for last-mile trips between a public transport stop and the place of destination [5]. Buck et al. [6]point out the mostly work-related purpose for BSS bikes in the group of long-term members. Besides daily commuters, Krauss et al. [7] identify the bike-sharing user group focusing on leisure activities during the day for short trips or longer trips up to six hours.

In general, we can distinguish between two different types of Bike Sharing Systems (BSS): (1) the traditional Station-Based BSSs (SBBSSs), where users can only start and end their trips at static stations and (2) Free-Floating BSSs (FFBSSs) where no restriction exists in start and endpoints of trips.

Besides this, it is known that rebalancing methods for BSSs are the focus of many research studies, but mainly on SBBSSs. Only a few studies focus on the rebalancing of FFBSSs [8]. The redistribution of bicycles by the provider usually follows the optimization for reducing trips and resulting costs, which is more accessible in the case of the usage of fixed dock-based stations for bikes [9].

The first generation of BSS appeared around 1965 with the option of freely available bikes with no fixed stations throughout the city [10]. On the other hand, fixed dock-based stations usually are implemented near points of interest such as restaurants, locations of businesses, leisure activities or stops of public transportation [11]. Rixey [12] and Faghih-Imani et al. [11] discover a relationship between the number of rented and returned bikes at locations near points of interest and those areas with a higher population or job densities. Some studies show that most customers of bike-sharing services do not use bike sharing frequently, which is supported by studies from for example Chicago [13], Cologne [14] or Paris [15]. Additionally, detecting different usage patterns (a) on weekends and weekdays or (b) during rush hours and the rest of the day are also part of these studies, as well as (c) changing demand depending on the weather and course of the year and (d) also the duration of bike trips [14]. Etienne et al.[16] identify groups of stations with similar usage profiles without using the station occupancy details.

III. Overview and Data Description

A. Overview of the bicycle sharing experimentation

The data refers to the NTU Free2move which is an e-bicycle sharing experiment operated within the campus of Nanyang Technological University (NTU) in Singapore. The users are students and staff of the university who unlock the bicycle by scanning the QR code present on the bicycle via an android application. The experiment is free of charge and works only on android devices. It is based on a credits system where each registered user is given a total of 100 credits and each minute of a ride costs 1 credit. Charging stations powered by solar energy for charging the batteries of the bicycles and incentivizing the users by giving extra credits to bring the

Shyam Sundar Rampalli is with TUM Asia, Singapore (e-mail: rampallishyam@outlook.com)

Andreas Keler is with the Chair of Traffic Engineering and Control, Technical University of Munich, Munich, 80333 Germany (e-mail: andreas.keler@ tum.de)

Georgios Grigoropoulos is with the Chair of Traffic Engineering and Control, Technical University of Munich, Munich, 80333 Germany (e-mail: george.grigoropoulos@ tum.de)

bicycles to the charging station are used to ease up the operation.

The data is collected via a GPS device installed on the bicycle which uses GSM to share the data to the cloud. The data contains the GPS coordinates of the start and end location of the ride. Also, intermediate GPS points in the ride are collected which can suggest the approximate route of each ride. The intermediate GPS points during the ride may not be accurate since the bicycle is in motion and the GPS was not accurate enough. With that assumption, the ride starting and ending points which also use the GPS location of the user's mobile phone were found to be relatively accurate. This experiment is implemented for the first time in Singapore.

B. *Description of the data set*

The data used for this research is from bicycle-sharing experimentation implemented in the Nanyang Technological University campus located in Singapore. It has an uneven topography with a total land area of 2 square kilometres. The experimentation was rolled out in three phases. The first phase was from 5th Feb 2018 to 20th March 2018. It is of 42 days which has 30 weekdays and 12 weekend days (trip count: 301) The second phase is from 10th April 2018 to 10th May 2018 i.e., 22 weekdays and 8 weekend days (trip count: 405). And finally, the third phase, which is the longest, from 1st Aug 2018 to 31st Dec 2018 consisted of 108 weekdays and 44 weekend days (trip count: 710). In the first and the second phases, the experimentation was open even on weekend days which is not the case with the final phase. During the third phase of the experimentation, the experimentation was only on weekdays operating from 09.00 hrs. to 18.00 hrs.

As this research solely relies on the accuracy of the GPS data, it is important to understand the accuracy of the data. The location of the bicycle when no one is riding is tracked by the on-device GPS unit. But the GPS coordinates of the start and endpoints of the rides are collected from the GPS coordinates indicated by the mobile phones of the users. Hence the accuracy of these coordinates highly depends on the signal strength and telecom provider. We tried to test the accuracy of the data through observations and follow-up questions from the riders. We concluded that the actual coordinates lie in a range of 50-100m from the original data obtained. This assisted us to filter and processing the GPS data during the further spatial analysis of the data.

C. *Trip analysis*

a. *Land-use*

We categorized the NTU campus by land use to understand the popular start and end locations. Open street map data was used to classify the buildings and parts of the campus based on land use. Certain places on campus were not available with the data obtained from the open street maps. Google maps and the map given to campus users are used as a basis to understand other parts of the campus. Certain corrections and updates are made wherever necessary.

b. *Trip description*

i. *Average distance*

The average distance of all the trips is about 0.9 km. Which represents 25% of the campus loop i.e., the length of the road encircling the campus. Since this experiment is used to fulfil first and last-mile trips, the data represented a similar pattern. All the trips are used to move from one location to another and we did not observe any intermediate stops between the start and end location of a trip. Figure 1 shows the frequency plot of the distance with an interval of 2 kilometres indicating the average distance in red.

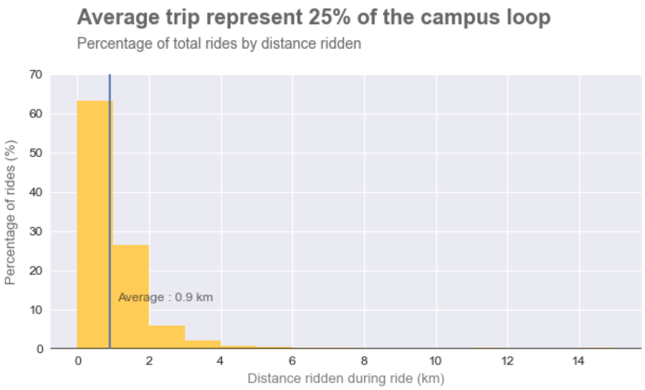


Figure 1. Trip distance frequency plot.

ii. *Average duration*

The Predominant trip duration is below 10 mins. The average duration of trips turned out to be 9.6 mins as shown in Figure 2. Trip duration frequency plot. The average speed of all the rides is about 6 kilometres per hour which about lower than double the walking speed of humans on average.

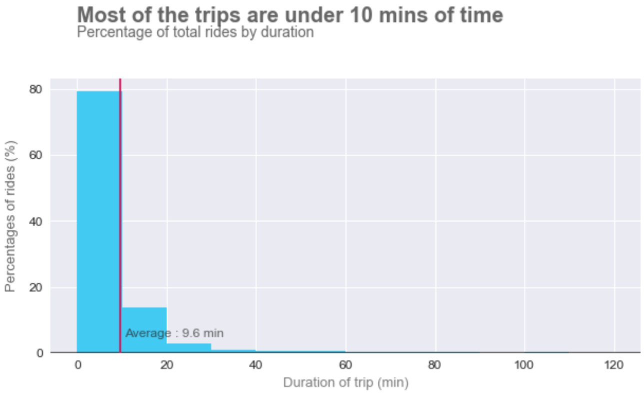


Figure 2. Trip duration frequency plot.

iii. *Temporal Variation*

We observed the peak usage in the afternoon between 12.00 hrs. and 14.00 hrs., followed by evening time between 17.00 hrs. and 19.00 hrs. Figure 3 shows the trip count by the hour of the day. We observed that this experimentation was not used for morning commutes as compared to the lunch trips and evening commutes.

The next sub-section presents the quality of the GPS data as it is common for GPS data to have accuracy issues.

We intend to apply unsupervised machine learning techniques, as mentioned before, to the GPS data hence obtained and are interested to learn the advantages and disadvantages of applying such techniques on a smaller data set. Accordingly, we describe the techniques used for this research, and analysis followed by possible limitations in the next section.

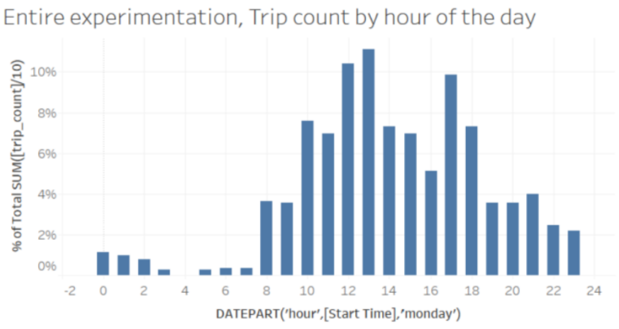
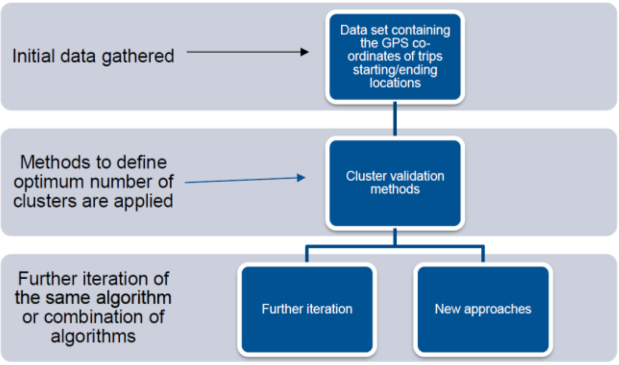


Figure 3. Trip count by the hour of the day.

IV. Methodology

We focused on implementing unsupervised machine learning techniques for pick-up and drop-off locations of the experimentation. The approach is presented in Figure 4.


Initial data gathered

Data set containing the GPS co-ordinates of trips starting/ending locations

Methods to define optimum number of clusters are applied

Cluster validation methods

Further iteration of the same algorithm or combination of algorithms

Further iteration

New approaches

Figure 4. Flow chart describing the method.

A. GPS data: Initial insights

The first step as mentioned in Figure 4 is post-processing of GPS data of trip origins and destinations. Initially, we observed that data appeared as a point cloud with concentration at a few locations as described in Figure 5.

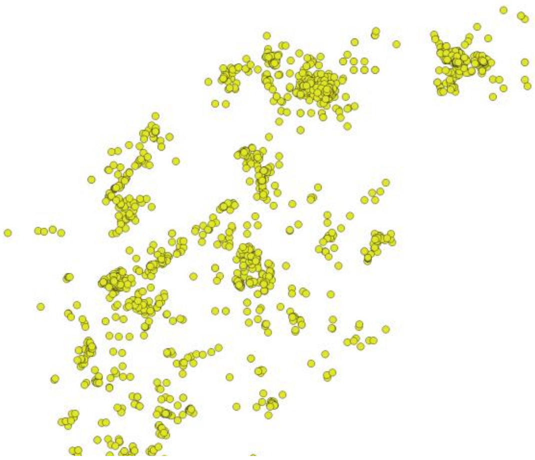


Figure 5. Unprocessed GPS data.

Although this did give an insight into the location with the greatest number of trips, it is not clear about the popular start and end locations. Hence, we investigated further to find out the popular locations i.e., administration building, bus stop, etc.

B. Data classification

Various fields depending on whether image processing, data mining or spatial data analysis require specific data classification methods. In this research, we are curious to implement popular unsupervised clustering techniques to identify popular locations. It is reasonable to observe the spatial data based on land use and identify the hotspots. While such a process is considered reasonable, we intend to use unsupervised machine learning techniques to observe if it is possible to automate the analysis of similar types of datasets. If such a method, it is then possible to extend this analysis for larger-scale bike-sharing systems implemented on a city level for example.

Determining the optimal number of clusters in a data set is a fundamental issue in partitioning clustering, such as k-means clustering, which requires the user to specify the number of clusters [17]. The methods to find the optimum number of clusters for a data set are categorized into two types. They are direct methods and statistic testing methods. The Elbow method, a direct method, is used for this research to determine the optimum number of clusters. The results found do not define a hotspot or a place of importance. Hence, they are used as the starting point upon which further method has been developed.

In this research, we used the elbow method considering the simplicity of implementation and higher computational efficiency [18].

Upon using this method, we identified the number of clusters into which the entire data was classified. Figure 6 shows the clusters as suggested by the Elbow method.

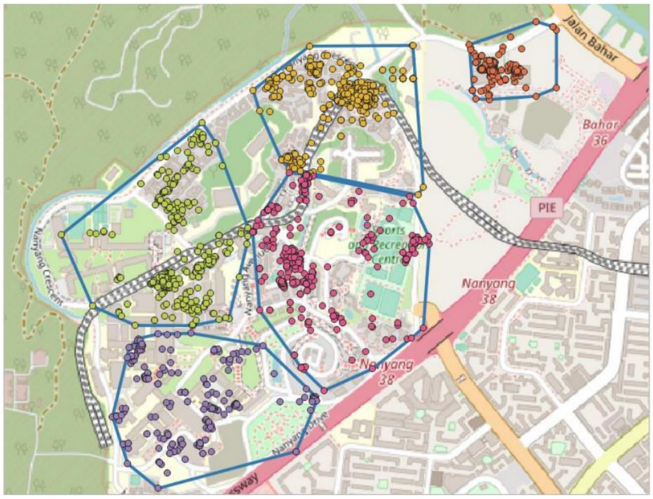


Figure 6. Five clusters as suggested by the Elbow method.

It is clear by now that the number of clusters does not represent significant independent locations as single points but rather, but t a group of popular or significant locations. Therefore, this method, although computationally efficient, is not effective for this data set. Hence, we pursue further clustering algorithms such as K-means and density-based clustering.

C. K-means clustering

K-means algorithm clusters data by trying to separate samples in n-groups of equal variances, minimizing a criterion known as the inertia of within-cluster sum-of-squares. This algorithm requires the number of clusters to be specified. It scales well to many samples and has been used across a large range of application areas in many different fields [19]. There is numerous research effort in the field of K-means clustering algorithm ranging from image processing, banking, and spatial data analytics. In this research, we implement K-means to bike-sharing data and evaluate the applicability of this clustering method to bike-sharing data in general.

D. DBSCAN: Density-based clustering

The DBSCAN algorithm views clusters as areas of high density separated by areas of low density. Due to this rather generic view, clusters found by DBSCAN can be any shape, as opposed to k-means which assumes that clusters are convex shaped. The central component of the DBSCAN is the concept of core samples, which are samples that are in areas of high density. A cluster is therefore a set of core samples, each close to each other (measured by some distance measure) and a set of non-core samples that are close to a core sample (but are not themselves core samples). There are two parameters to the algorithm, min_samples and eps, which define formally what we mean when we say dense. Higher min_samples or lower_eps indicate higher density necessary to form a cluster [19].

E. Combining K-means and density-based clustering

In addition to testing the dataset with K-means and density-based clustering, we observed the benefit of combining both the clustering algorithms. Bike-sharing data, as we have observed, form concentrations at specific locations. While that stands true, there are cases where users start and end trips at locations that do not have any importance. As K-means clustering includes all the points of the data set, it is difficult to form realistic clusters. When we observed the properties of density-based clustering, we observed that it is highly parameter sensitive, and it is unable to identify different point groups with different densities. For example, it is unable to differentiate a bus stop from an administrative building on the university campus.

Upon realizing this, we categorized the data set into specific clusters using K-means first and then applied density-based clustering to each cluster. This method was found beneficial to using one of the algorithms alone as it can differentiate between different locations which usually have different densities or number points which translates to the number of trips.

V. Results

In this section, we present the effectiveness of the k-means clustering method, density-based clustering and the combination of both K-means and density-based clustering in the same order.

A. K-means clustering

We tested the dataset for different K-values from 1 to 25 and we observed that no K-value defines all the significant locations. For example, we observed that for a k-value of 5 as shown in Figure 6, we observed 5 clusters and none of these clusters represent a location. When tested with a k-value of 15 as shown in Figure 7, in cluster 1 there are activities in two different places but still, K-means sees it as a single cluster. Cluster 2 has significant activity, but the noise points are included. Cluster 3 is having few points to be considered a hotspot. A lot of noise points are also considered which was understood as a limitation of this algorithm.

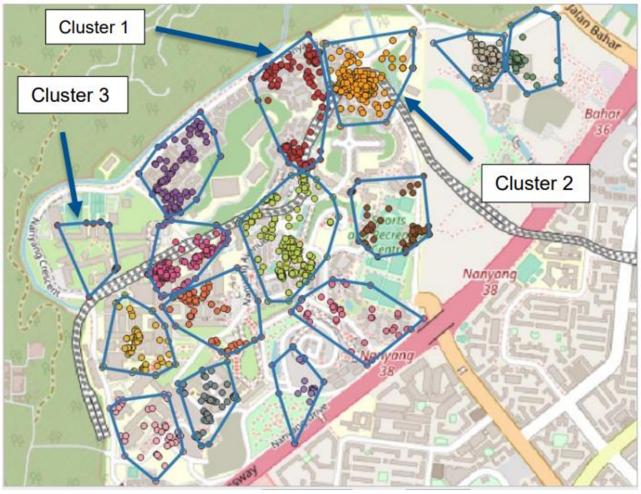


Figure 7. K-means result with k=15.

With further increase in k, for example, k=25, we observed that it defines a more clearly location as shown in Figure 8. However certain clusters such as cluster 4 show two locations which are not realistic.

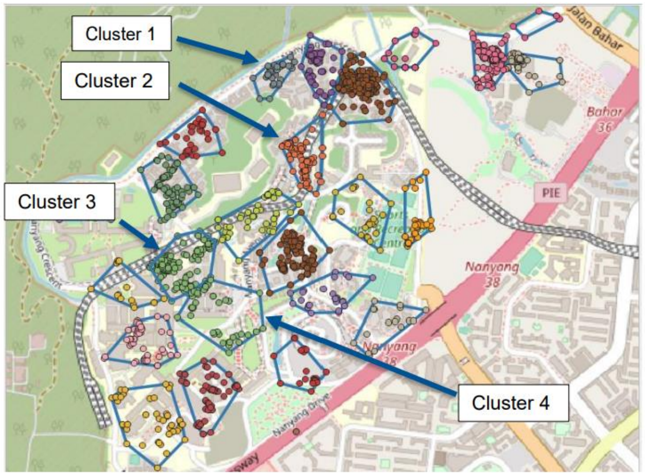


Figure 8. K-means result with k=25.

B. Density-based clustering

In a similar fashion, as mentioned in section 0, we investigated the effectiveness of density-based clustering. We observed that density-based clustering effectively filtered the redundant trip points, or in better terms, noise. However, while the algorithm filtered noise, it filtered points that refer to a specific location. For example, Figure 9 shows high filtering when applied to the data set as shown in Figure 5. We observed it selects a cluster that meets a specific density between the points. Such a method is only useful if the density throughout the region is the same. It does not work for locations with varying land use.

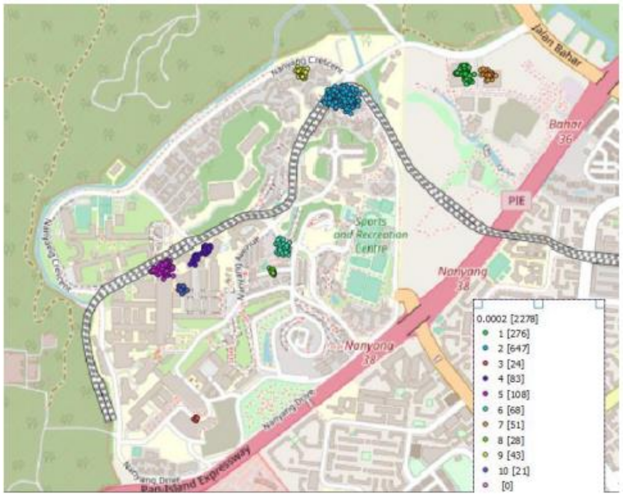


Figure 9 Density-based clustering with high filtering

C. Combining K-means and density-based clustering

K-means gave the poorest clustering result referring to the three clusters formed. The DBSCAN algorithm or density-based clustering gave better results by removing the noise as compared to K-means. The approach which combined K-means and DSBCAN gave better results compared to the other two approaches as shown in Figure 10 which indicates the clusters by number and location each cluster represents.

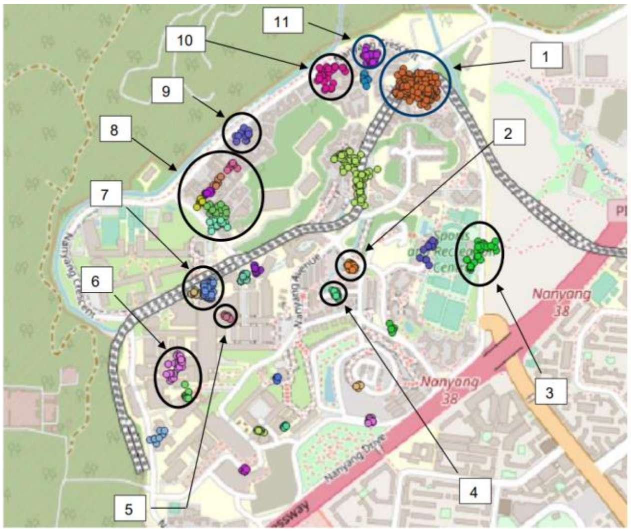


Figure 10. Final clusters on combining K-means and density-based clustering; 1. The charging station, 2. Canteen 2, 3. The wave sports hall, 4. Bicycle parking behind canteen 2, 5. Arc learning hub north bicycle parking, 6. Department of civil and environmental engineering, 7. North Spine Bus stop, 8. Undergraduate student halls, 9. Hall 15 bus stop, 10. Nanyang crescent hall, 11. Graduate Hall 2.

The initial k-value used for the Elbow method and k-means algorithm is applied. Within the individual clusters, data is clustered using a density-based clustering method. The next section describes the research outcomes, and criticisms followed by key insights.

VI. DISCUSSION

For testing, the Python programming language was used to apply different algorithms from [19] and QGIS [20] is used for geospatial analysis. No single algorithm is found to be effective to define a location of interest or a hotspot. The underlying issue is that the density of GPS data differs by the type of location. In a residential location, the points are spaced apart by a longer distance when compared to a train station where the points are closely spaced. This was found to be a hurdle for all the existing algorithms which predominantly use distance as a parameter to classify clusters. A combination of existing algorithms has improved the quality of the clusters. One specific example is a combination of K- means and density-based clustering algorithms that have formed more realistic clusters. However, the choice of the number of clusters is an input for K means algorithm which was not suggested by any existing method. This research does not focus on identifying communities for the clusters as investigated in [21]. We observed that the knowledge of the type of location could significantly improve the results. Also, quantifying such knowledge of the location is a known challenge.

VII. CONCLUSION

With the advancements in computing and the scale of data generated, we must consider using intelligent methods to analyze transport data. It is important to understand that combining multiple unsupervised clustering techniques was achieved earlier by many researchers. We aim to use these methods and evaluate their effectiveness for bike-sharing GPS data. We found that if we understand the pattern of the GPS data, it is possible to apply clustering techniques wisely by adjusting the parameters of the techniques for suiting the data.

REFERENCES

[1] J. Froehlich, J. Neumann, N. Oliver, and others, “Sensing and predicting the pulse of the city through shared bicycling.,” in IJCAI, 2009, vol. 9, no. Jul, pp. 1420–1426.

[2] G. McKenzie, “Docked vs. Dockless bike-sharing: Contrasting spatiotemporal patterns,” 2018.

[3] S. Ji, C. R. Cherry, L. D. Han, and D. A. Jordan, “Electric bike-sharing: simulation of user demand and system availability,” Journal of Cleaner Production, vol. 85, pp. 250–257, 2014.

[4] R. Meddin, “The Meddin Bike-sharing World Map.” 2009.

[5] T. D. Tran, N. Ovtracht, and B. F. d’Arcier, “Modeling Bike Sharing System using Built Environment Factors,” Procedia CIRP, vol. 30, pp. 293–298, 2015.

[6] D. Buck, R. Buehler, P. Happ, B. Rawls, P. Chung, and N. Borecki, “Are Bikeshare Users Different from Regular Cyclists?: A First Look at Short-Term Users, Annual Members, and Area Cyclists in the Washington, D.C., Region,” Transportation Research Record, vol. 2387, no. 1, pp. 112–119, 2013.

[7] K. Krauss, A. Scherrer, U. Burghard, J. Schuler, A. M. Burger, and C. Doll, “Sharing Economy in der Mobilität: Potenzielle Nutzung und Akzeptanz geteilter Mobilitätsdienste in urbanen Räumen in Deutschland,” Fraunhofer-Institut für System- und Innovationsforschung ISI, 2020.

[8] C. S. Shui and W. Y. Szeto, “Dynamic green bike repositioning problem – A hybrid rolling horizon artificial bee colony algorithm approach,” Transportation Research Part D: Transport and Environment, vol. 60, pp. 119–136, 2018.

[9] M. Benedek, C. von Groote-Bidlingmaier, and S. Timpf, “GIS-gestützte Analyse und Optimierung von Bike-Sharing-Systemen.,” Angewandte Geoinformatk 2014. Hrsg.: Strobl, J., Blaschke, T., Griesebner, G. & Zagel, B., 2014.

[10] P. Midgley, “Bicycle-sharing schemes: enhancing sustainable mobility in urban areas,” United Nations, Department of Economic and Social Affairs, vol. 8, pp. 1–12, 2011.

[11] A. Faghih-Imani, R. Hampshire, L. Marla, and N. Eluru, “An empirical analysis of bike-sharing usage and rebalancing: Evidence from Barcelona and Seville,” Transportation Research Part A: Policy and Practice, vol. 97, pp. 177–191, 2017.

[12] R. A. Rixey, “Station-level forecasting of bike-sharing ridership: Station network effects in three US systems,” Transp Res Rec, vol. 2387, no. 1, pp. 46–55, 2013.

[13] X. Zhou, “Understanding spatiotemporal patterns of biking behaviour by analyzing massive bike-sharing data in Chicago,” PLoS One, vol. 10, no. 10, p. e0137922, 2015.

[14] K. Schimohr and J. Scheiner, “Spatial and temporal analysis of bike-sharing use in Cologne taking into account a public transit disruption,” Journal of Transport Geography, vol. 92, p. 103017, 2021.

[15] C. Etienne and O. Latifa, “Model-based count series clustering for bike-sharing system usage mining: a case study with the Vélib’system of Paris,” ACM Transactions on Intelligent Systems and Technology (TIST), vol. 5, no. 3, pp. 1–21, 2014.

[16] C. Etienne and O. Latifa, “Model-Based Count Series Clustering for Bike Sharing System Usage Mining: A Case Study with the Vélib’ System of Paris,” ACM Transactions on Intelligent Systems and Technology, vol. 5, no. 3, pp. 1–21, Oct. 2014, doi: 10.1145/2560188.

[17] I. D. Borlea, R. E. Precup, A. B. Borlea, and D. Iercan, “A Unified Form of Fuzzy C-Means and K-Means algorithms and its Partitional Implementation,” Knowledge-Based Systems, vol. 214, p. 106731, Feb. 2021, doi: 10.1016/J.KNOSYS.2020.106731.

[18] P. Bholowalia and A. Kumar, “EBK-means: A clustering technique based on elbow method and k-means in WSN,” International Journal of Computer Applications, vol. 105, no. 9, 2014.

[19] F. Pedregosa et al., “Scikit-learn: Machine learning in Python,” the Journal of Machine Learning research, vol. 12, pp. 2825–2830, 2011.

[20] Q. D. Team and others, “QGIS geographic information system,” Open source geospatial foundation project, 2016.

[21] P. Borgnat, C. Robardet, P. Abry, P. Flandrin, J.-B. Rouquier, and N. Tremblay, “A Dynamical Network View of Lyon’s Vélo’v Shared Bicycle System,” 2013, pp. 267–284. doi: 10.1007/978-1-4614-6729-8_13.