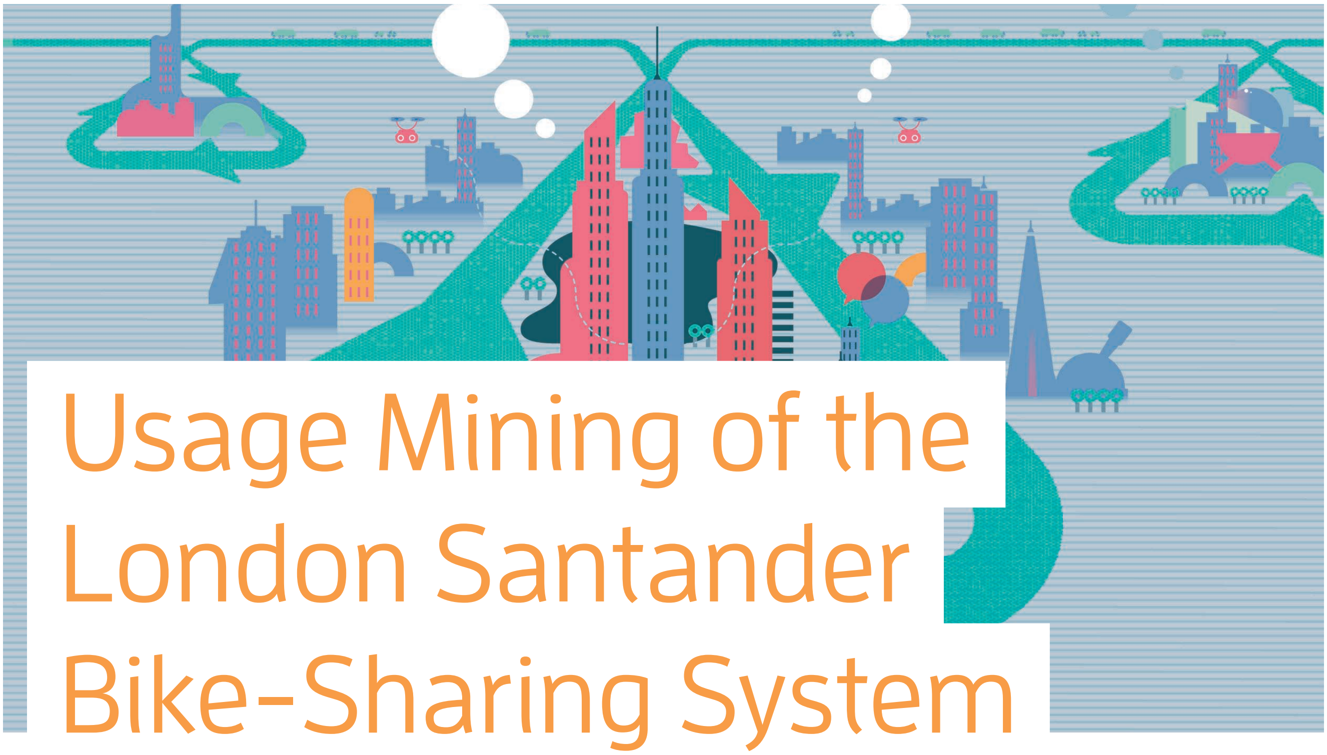


Usage Mining of the London Santander Bike-Sharing System

Suparna De, University of Surrey
Wei Wang, Xi'an Jiaotong Liverpool University
Usamah Jassat, University of Surrey
Klaus Moessner, Chemnitz University of Technology

With cycling moving from being a pastime to a mainstream form of mobility and transport, bike-sharing systems (BSSs) are increasingly being deployed in many cities. Analysis of BSS usage data can provide insights into factors that shape the patterns of trips, uncovering latent city dynamics.

Increasing environmental pressures and limited urban resources, such as roads and public transport, call for the development of more sustainable urban mobility strategies.[1] To lessen the soaring impacts of urban mobility demands, public bike-sharing systems (BSSs) have been implemented in more than 450 cities worldwide.[2] BSSs are characterized by short-term bike rentals available through a network of unattended bike docking stations. With a dense deployment of BSS stations that offer seamless connectivity with existing public transport infrastructure, such as bus stops as well as tube and train stations, BSSs offer a softer public transport alternative that is more affordable, is healthier and less polluting, and provides good opportunities for meeting the last-mile commuting challenge.[2,3]

With the 2021 Intelligent Transport Systems Congress[4] identifying sustainability and a modal shift in transport as priorities to meet the challenge of reducing CO_2 emissions, bike sharing forms a part of the solution for smart and zero-emission mobility in cities. For decision makers in city municipalities to implement BSSs in their urban policy plans, it has been recognized that, in addition to the BSS infrastructure provision, a step change is needed. This means the provision of software, platforms, and applications for managing

Digital Object Identifier 10.1109/MC.2022.3193308
Date of current version: 15 November 2022

and analyzing the bike fleets, such as cyclists and BSS traffic flows in various city regions.[4]

As cities continue to grow with sustained migration, connections between different neighborhoods become more complex with the vast array of transport options available to the public in large cities, such as London.[1] Finding functional areas in a city through mobility data mining can help in understanding these connections and give urban planners insights into the urban infrastructure.

Mobility mining for discovering urban areas and their functions has been explored in the literature through taxi trips data,[5] social network posts,[6] and GPS call data records (CDRs).[7] However, the development of new techniques more suitable for the BSS transport data, which are mainly about short-distance trips, is needed. The spatiotemporal nature of the correlations of the BSS data with the neighborhoods also needs to be considered.[8] It has been recognized that the clustering of BSS stations is tied to their different functions, which are, in turn, linked to the city's activities, for example, residential, leisure, and employment.[2]

For this, we present a method for BSS station clustering using the expectation maximization (EM) generative mixture model and Poisson distribution in its construction to better reflect the event-based nature of bike check-ins at stations. The method can uncover spatiotemporal trends in terms of bike arrivals and departures, with distinct temporal usage in each cluster, owing to their spatial distribution and demographic characteristics. Additionally, we derive station-pair clusters to find the strongest pairwise flow movement patterns between stations, which are, in turn, related to different social activities, such as commuting or going out for lunch. We validate the model on data collected from the London public BSS, named *Santander Cycles*.[21]

In contrast to existing work on BSS clustering that uses station occupancy data,[9,10] our method uses departure/arrival count series, which are more detailed and able to distinguish periods of high and low (or no) activity. The proposed mixture model also handles the differences in weekday/weekend behavior directly rather than through data preprocessing or feature construction.

THE ACTIVITY BEHAVIOR OF BIKE STATIONS IS MODELED THROUGH A STATISTICAL APPROACH THAT DESCRIBES BIKE STATION USAGE IN TERMS OF ARRIVAL AND DEPARTURE COUNT STATISTICS.

RELATED WORK

The monitoring of long-term trends in personal mobility patterns has traditionally been achieved through annual household surveys, such as the National Travel Survey (NTS) in England.[22] The growth of the urban computing paradigm, which uses statistical and machine learning techniques for deriving patterns in large-scale urban data sets, has led to the mining of mobility patterns from taxi trip data,[5] location-based social network data,[6] and mobile CDRs.[7] Though the NTS informs policy on personal transport, such targeted surveys are reliant on user participation. The short-distance (also short-duration) nature of bike trips requires the development of techniques that take into account the specific event-based nature of bike rentals/returns from stations as well as their spatiotemporal correlation with the bike station neighborhood.[2]

Research using BSSs has employed either clustering methods to find bike station partitions that have similar usage or prediction techniques to forecast the occupancy of stations and station traffic toward bike rebalancing and scheduling optimization. Initial studies on temporal pattern mining from bike usage data consider statistical features, such as historical averages/trends, with Bayesian networks[9] or time series analysis,[10] with the most salient feature derived being the repeating three-pronged spike corresponding to the morning, lunch, and evening commutes across all weekdays. The BSS data elements in these studies are the station location, the number of available cycles, and the number of vacant parking slots.

In contrast to these studies that use station occupancy data, our method uses departure/arrival count series, which are more detailed and able to distinguish periods of high and low (or no) activity. Additionally, we consider trip data, rather than station occupancy, ensuring that the derived trends relate to actual bike journeys rather than the BSS load balancing measures via trucks.[9]

Subsequent works involving bike trip data (similar to our approach) include the research by Etienne and Latifa,[3] who proposed a count series model to predict hidden station clusters. The resulting clusters include those that are related to commuting (that is, stations located close to public transport and mostly active during the morning and evening on weekdays). Another BSS mobility study[11] investigated the spatial analysis of bike trips by visualizing the activity in each station separately and then identifying the main characteristics of the flow between stations.

Our model is motivated by these works that consider bike trips as count series. We further extend the clustering of stations according to their temporal usage profiles as conducted in these studies to pairwise traffic flows between stations that correlate the cycle journeys to work/social travel patterns. Another approach whose objective is close to the one proposed here looks at station function discovery[2] by modeling a station as a document in a latent Dirichlet allocation algorithm, with station functions derived as the topics of a document. Other studies utilize spatiotemporal features, such as the impact of points of interest (POIs),[12,13] POI categories,[8] or weather conditions[14] on station-level traffic prediction.

In contrast to these studies, our proposed method encodes the differences in weekday/weekend behavior directly into the mixture model parameters rather than through preprocessing or feature engineering methods. Moreover, our model encodes the trip data for each available day over a long period rather than a summary of the statistics, which takes into account factors of seasonality.

Different clustering approaches applied to BSS data include studies for traffic prediction that involve a clustering step at the city, cluster, or station level (for example, bipartite clustering,[15] Xgboost,[16] Gaussian mixture model,[17] or hierarchical clustering[18]) to divide bike stations into groups and counteract traffic fluctuations at individual stations. While different clustering methods offer different performance benefits, our proposed model considers the specific count-based nature of the data, whereas previous solutions do not use this particularity.

COUNT SERIES CLUSTERING MODEL

Mixture model

Observed data can be utilized to infer the underlying unseen probability density distribution. The activity behavior of bike stations is modeled through a statistical approach that describes bike station usage in terms of arrival and departure count statistics.

The underlying mixture model f is a mixture of K component distributions $P_1, P_2, \ldots P_k$, where each component is a Poisson distribution to match the count series data based upon mixing weights π_k.

The mixture model has the general form

$$f(x) = \sum_{k=1}^{K} \pi_k P_k(x) \tag{1}$$

with K representing the number of station clusters that need to be obtained, being latent (that is, not directly observed) in the observed bike trip data.

Regarding notation, in this section, we employ lowercase letters for variables and their corresponding uppercase equivalents for the overall summation value for the variable. For instance, k represents a station cluster, and K represents the total number of station clusters, that is, $k \in \{1,\ldots.,K\}$.

The observed data for a station s can be represented by a count series of the number of departures X_s^{out}, and arrivals X_s^{in} at a given hour $h \in \{1, 2, \ldots 24\}$ on a given day $d \in \{1, \ldots D\}$. The quantization of 1 h is deemed as a good tradeoff between data resolution and fluctuations in departure/arrival counts, in line with existing literature on bike usage modeling.[3,11] These arrival and departure count series are concatenated to X_{sd}, denoting the arrival and departure activity of a station s on day d.

All of the bike check-in data can be represented as a 3-dimensional tensor X of size $S \times D \times T$, where S is the total number of stations, D is the total number of days available in the data set (corresponding to the data collection period January 2015 to 2017 May), and T is 48 because the arrival and departure counts in a day are computed in 1-h nonoverlapping windows (that is, 24 × 2). The parameters for the model are arranged as arrays of varying dimensions and represent the probability that a given station belongs to a particular cluster. An intermediary parameter m of size $S \times K$ [with K

as specified in (1)] is used to calculate these parameters.

Although the most popular distribution to use in the construction of mixture models is the Gaussian distribution, the Poisson distribution is used in this work. This is because the Poisson distribution fits the count nature of the observations. The discrete Poisson distribution expresses the probability of a number of events occurring in a given time period based on a mean. In this work, the bike check-ins in a given hour on a day are the events, and we model their probability distribution to cluster them.

In addition to using the Poisson mixture to build the generative model, two indicator variables are defined. The first, W_{dl}, is used to take into account the difference in the bike station usages on weekdays and weekends, as these present very different usage profiles, with $W_{d0} = 1$ indicating that the day d is a weekend and $W_{d1} = 1$ indicating a weekday; that is, l denotes the day (weekday/weekend) cluster membership of the station. $D_l = \Sigma_d W_{dl}$ denotes the number of days in day cluster l. The second indicator variable π_k encodes the cluster membership of a station, with K denoting the number of station clusters and applied as a component of the model as specified in (1).

A scaling factor α_s, specific to each station s, is used to represent the global activity (total volume of arrival/departure counts) at a station. It is used to distinguish between stations that may have a common usage profile but show wide differences in activity (arrival/departure) volume, and it is calculated as

$$\alpha_s = \frac{1}{DT}\sum_{d,t} X_{sdt} \tag{2}$$

where X_{sdt} represents the arrival and departure activity of a station s on day d and timeframe t, and D and T are as explained previously in this section. As seen in (2), α_s of station s is calculated as the average of its activity vectors along all of the timeframes and days.

Consideration is also made for the variation in activity at different times during the day. The difference in activity in different clusters is modeled through the mean used in the Poisson distribution λ, with λ_{klt} representing the temporal variations of arrivals/departures for each station cluster k, day type l (that is, weekend/weekday), and timeframe t. The following constraint is placed on the λ in order for the model parameters to be calculated:

$$\sum_{l,t} D_l \lambda_{klt} = DT,\ \forall k \in \{1, \ldots, K\}. \tag{3}$$

Taking (3) into account, the conditional density of the activity vector X_{sd} can be derived as

$$P_k(X_s) = \prod_{d,t,l} p(X_{sdt}; \alpha_s \lambda_{klt})^{W_{dl}} \tag{4}$$

where $p(., \lambda)$ is the density of the Poisson distribution with mean λ. The generative model makes the assumption that the departure/arrival counts for each hour are independent and follow a Poisson distribution of parameter $\alpha_s \lambda_{klt}$. Estimation of the model parameters and station clustering can be performed by the maximum likelihood estimates (MLEs) of these parameters. For this, the log-likelihood is first derived by substituting P_k from (4) into the mixture model equation (1), summing over all k, and taking the logarithm of the function. Instead of estimating the parameters' MLEs directly through numerical optimization, the EM algorithm is used to maximize the log-likelihood.

EM algorithm

The parameters of the mixture model can be estimated by using the EM algorithm,[19] which is used for obtaining the MLEs of parameters when there is latent—that is, unobserved—data. It is an iterative algorithm with two steps: in the E step, soft assignment is done for each of the data points to one of the clusters based on the current model parameters. This is done by estimating the a posteriori probabilities of each cluster m_{sk}, given by

$$m_{sk} = \frac{\pi_k P_k(X_s)}{\sum_k \pi_k P_k(X_s)}. \tag{5}$$

Thus, the E step computes the expectation of the log-likelihood of the conditional density given in (4). This provides the lower bound of the log-likelihood. The M step updates the parameters in such a way so as to maximize the log-likelihood of the model based on the results from the E step. The parameters are updated according to the following rules:

$$\pi_k = \frac{1}{N}\sum_{s=1}^{S} m_{sk} \tag{6}$$

$$\lambda_{klt} = \frac{1}{\sum_s m_{sk}\alpha_{sk}\sum_d W_{dl}} \sum_{s,d} m_{sk} W_{dl} X_{sdt}. \tag{7}$$

Equation 6 depicts how π_k, which encodes the cluster membership of a station, is updated using the a posteriori probabilities of each cluster. Equation 7 shows the calculation of λ_{klt} as a weighted mean of the activity of cluster k stations in day cluster l

and timeframe t. The E and M steps are iterated until the parameters converge to the local maximum of the log-likelihood function.

DATA SET

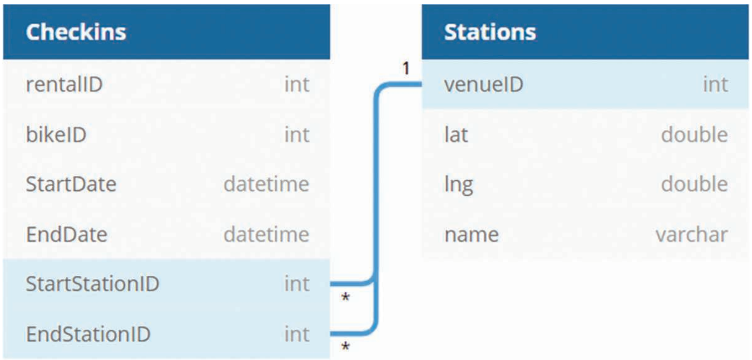
The data set is sourced from the "usage-stats" section of the Transport for London (TfL) cycling open data website,[23] which has data on all Santander Cycles journeys. The TfL data are available as downloadable comma-separated values files, each containing bike journeys for a 15-day period. Each bike journey is described as shown in the "Checkins" table of Figure 1, with the start/end date and time and start/end station IDs.

We collected bike trip data for a 3-year period, from 4 January 2015 to 16 May 2017, which, after parsing and cleaning, contains more than 15 million trips. Cleaning the data set included removing erroneous or invalid trip data, that is, removing any journeys with a duration of 0 s. Journeys that took longer than a day, constituting 0.06% of all journeys, were also removed, as these possibly point to misuse of bikes. This seems like an appropriate threshold to use to retain only appropriate and normal bike usage, as the majority of the journeys in the data set (98%) were less than an hour. The station IDs are mapped to the station name and location (latitude/longitude) using the TfL unified application programming interface and querying for "BikePoint."[24] The resulting data schema is shown in Figure 1.

LONDON BSS STATION CLUSTERING

The mixture model for clustering the bike stations is applied to the count-based trip data, with the model parameters estimated using the EM algorithm. The mixture model and EM algorithm are implemented in Python 3 on a laptop with an Advanced Micro Devices Ryzen5 5600X CPU and 16 GB of random-access memory. Loading and preprocessing (parsing and cleaning) of the data are performed using the Pandas library,[20] as it provides functionality to load large amounts of table data and to easily filter and sort them. Numerical calculations for tensor manipulation and operations were performed using the Numpy library.[25]

The maximum probability of each station was used to determine the cluster that a station belonged to. Check-in data across all stations for a period of one week were used for parameter estimation in the model. A range of cluster numbers were experimented with, with the most appropriate value selected by plotting the mixture model's log-likelihood against the cluster numbers. This was then analyzed by the elbow method heuristic,[3] which shows an elbow in the curve at $K = 5$. Hence, the value of five is chosen for the number of clusters. The MLE of the mixture model parameters is taken as the best of the set of local maxima obtained from the various runs of the EM algorithm. The time to reach convergence was 18 min, with each EM run taking between 90 and 134 s, with run durations increasing with the later runs.



Checkins	
rentalID	int
bikeID	int
StartDate	datetime
EndDate	datetime
StartStationID	int
EndStationID	int

Stations	
venueID	int
lat	double
lng	double
name	varchar

FIGURE 1. The schema of the bike journey data.

STATION CLUSTERING RESULTS

Figure 2 shows a map of London with the location of the bike station colored by the cluster that it belongs to. Figures 3 and 4 show the temporal activity profiles of the stations in the five computed clusters, given by the parameter λ of the model. The temporal plots are organized according to the nature of the count (arrivals/departures) and the day type (weekday/weekend), with the 24-h scale on the x-axis and the y-axis depicting the normalized count of bike arrivals/departures (corresponding to the normalized X_s^{in}/X_s^{out}, respectively, of all stations in that cluster).

Cluster A: Leisure and tourism

Figure 3(a) shows the activity pattern for cluster A. On weekdays, the arrival and departure patterns were very similar, with peaks in the morning (around 8 a.m.) and evening (7 p.m.). These two peaks most likely corresponded to commuting times, which is one of the main uses of the bikes in London. In between the peaks, however, the activity still stays relatively high in comparison to some of the other clusters. Additionally, the peaks during the weekend activity are very similar in magnitude to the weekday peaks. This shows that these stations are used as much on weekends as they are used during commuter times on weekdays. This suggests that these stations are also used heavily for tourism and leisure activities in addition to commuting. Looking at the locations of these stations shows that they are close to either public transport or tourist attractions, such as Madame Tussauds and Hyde Park as well as commercial areas.

Cluster B: Transport

Figure 3(b) shows the activity patterns for stations in cluster B, which were similar to the patterns seen in cluster A; however, distinctions can be seen in the difference in peak activity during weekdays and the activity in between these peaks. The difference is a lot larger, with activity between the peaks being significantly smaller, suggesting that these stations are predominantly used during commuting hours. However, unlike stations in cluster A, they are used in both directions. Looking at the locations of these stations shows that they are located close to public transport, including several of London's largest train stations, such as London Euston, Victoria, London Marylebone, and Paddington.

Clusters C: Work and leisure and Cluster D: Work

Figure 4 shows the activity patterns for stations in clusters C and D, which are similar to each other, with both having high peaks in arrivals during the morning and in departures during the evening. This is opposite to what is seen in cluster E, suggesting that these stations are predominantly being used as the destination stations for work commuting in the morning and as the origins for commuting back home in the evenings. The locations of these stations are

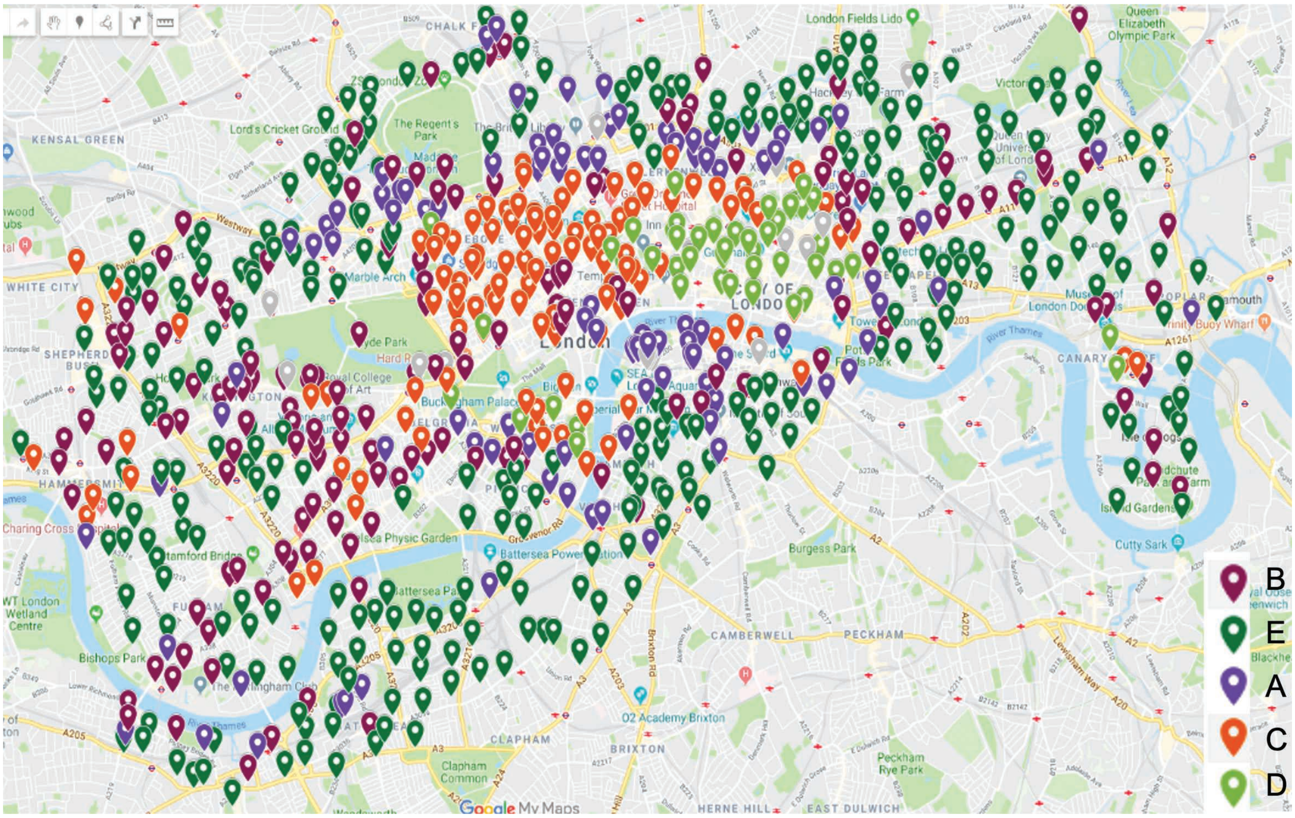


FIGURE 2. A map of the bike station clusters.

mainly based around central London, which is the busiest part of the city, especially in terms of industry and jobs, reinforcing the idea that these stations are used as destinations for work commutes.

The main difference between the activities at these clusters is their activity outside of commuting hours and on weekends. Stations in cluster C show more activity during these times in comparison to the commuting peaks, suggesting that these stations are also used for other purposes, such as tourism and entertainment. The spread of these stations across London shows that, although they are both mostly located around central London, cluster D heavily occupies the east side, while cluster C heavily occupies the west side. Looking at the map, although both sides have a lot of industries, the west side has more entertainment facilities, such as shopping districts and theaters.

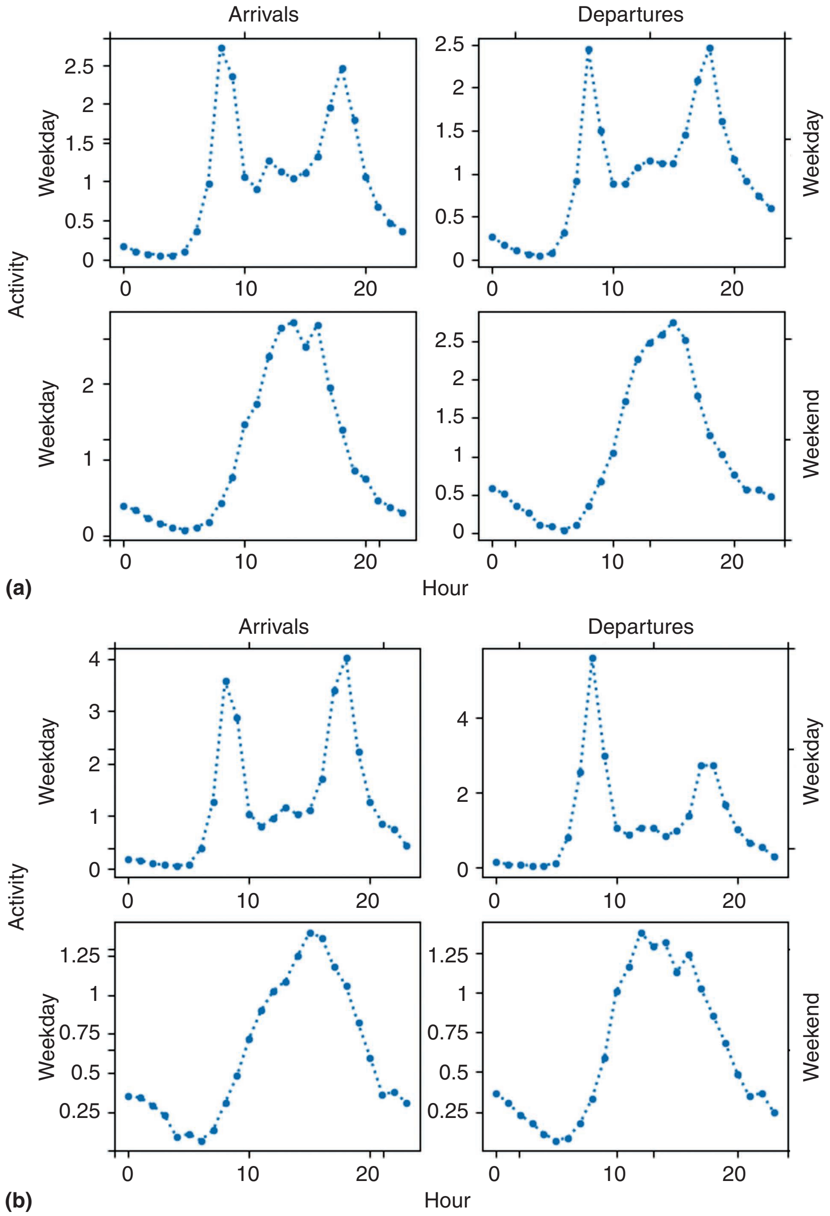


FIGURE 3. The mean activity pattern for stations in (a) cluster A: leisure and tourism (close to tourist attractions, commercial areas, and public transport) and (b) cluster B: transport (close to public transport and train stations).

Cluster E: Residential

Figure 4(c) shows the activity patterns for stations in cluster E. It contains a significant difference in arrival and departure activity during the weekdays, with peak activity occurring during the evening for arrivals and in the morning for departures. These peaks appear at commuting times and suggest that these stations are in residential areas and are used by individuals leaving for work in the morning and returning home in the evening. Looking at the locations of these stations on the map shows that they are predominantly located on the outskirts of London, which are less industrially dense than the center and contain more residential areas.

These stations are also usually close in proximity to a station from another cluster, usually one that is close to public transport. This indicates that these are most likely the destination stations for the morning commute and origin stations for the evening commute. The stations are also relatively active during the weekends, with the peak activity only dropping by 50%

from weekday to weekend, suggesting the stations are still in heavy use during the weekend by individuals living in these residential areas.

LONDON BSS STATION-PAIR PATTERN

Pair pattern approach

To quantify the relationship of the temporal characteristics of the stations (derived from the EM model) to the social and economic activities of the station neighborhood type, we spatially cluster the pairwise flows between source–destination station pairs. To this end, to find the pattern of bike journeys that relate to different socioeconomic activities, a second clustering approach is applied to group station pairs that serve as the source and destination stations of the bike journeys contained in the data set to determine the strongest movement patterns between pairs of stations.

To find patterns between different source–destination station pairs, it is important to reduce the size of the data, as the combinations of stations grow rapidly with the number of stations. To do this, the first step is to find the principal components of the activity patterns, which enables the data to be summarized without loss of information by transforming correlated attributes into noncorrelated components that are a linear combination of the original features. Activity patterns between stations are aggregated per day of the week and hour (treating this as a timestamp) and represented in a vector of size 168 (7 × 24), corresponding to the seven days of the week and 24 h per day. The principal components are found by using principal component analysis and selected such

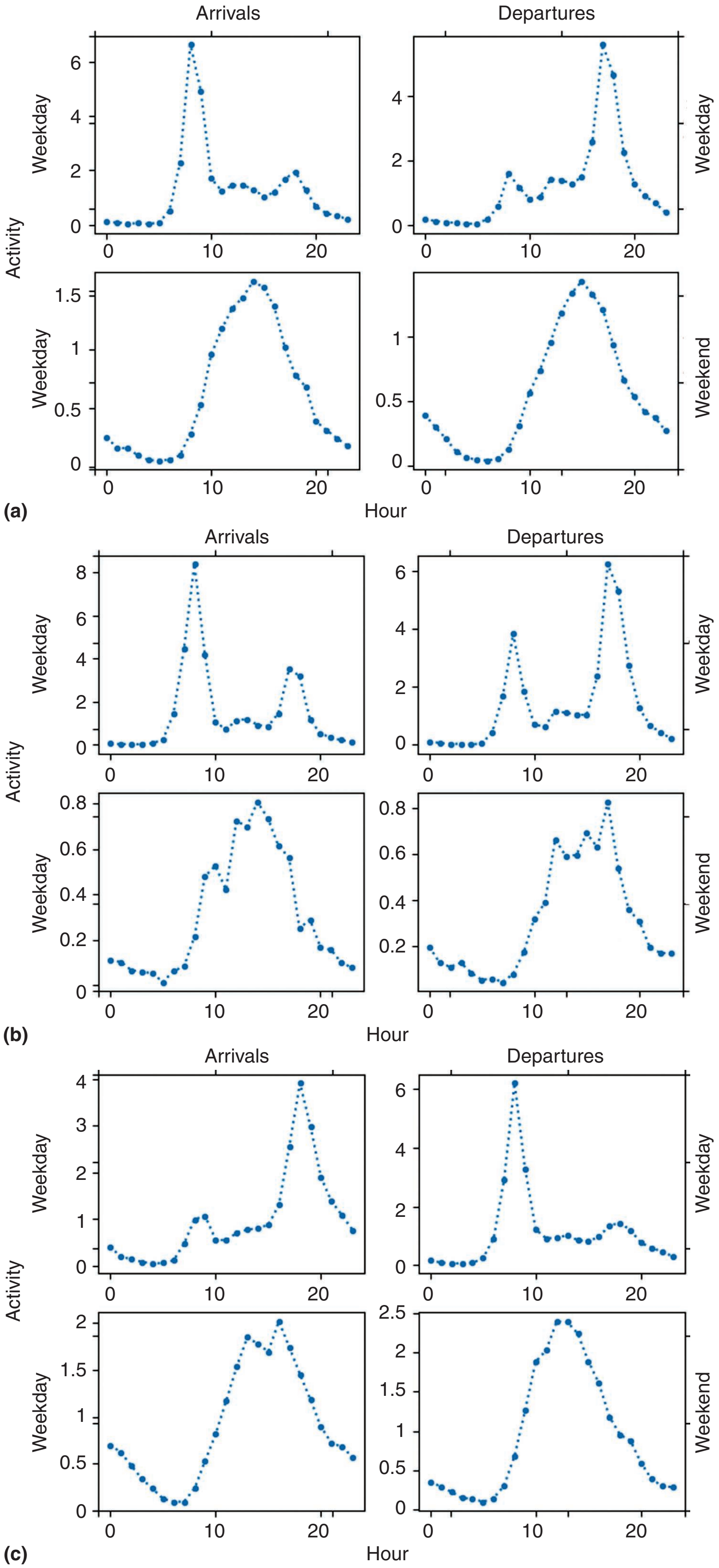


FIGURE 4. The mean activity pattern for stations in (a) cluster C: work and leisure (high morning arrivals and high evening departures as well as more weekend activity, predominantly in west central London) (b) cluster D: work (high morning arrivals and high evening departures as well as less weekend activity, predominantly in east central London); and (c) cluster E: residential (peak arrivals in the evening and high departures in the morning).

that 90% of the variance within the data are kept.

To further reduce the data, station pairs are filtered based on their maximum travels during any timestamp. This removes station pairs with little to no activity between them. Stations might have similar temporal activity patterns; however, this pattern may vary in scale from station to station. Therefore, before clustering, the stations' data are normalized. The final step applies the *k*-means algorithm to the normalized data to find clusters between stations. The most appropriate value for the number of clusters (that is, $K = 3$) is determined by the elbow method heuristic. This is done by calculating the sum of squared error (SSE) between the cluster centroid and each cluster member for each value of K and plotting the SSE values against K. The K value is chosen where the graph forms an "elbow," that is, the SSE decreases abruptly.

Station-pair pattern results

Figure 5 shows the weekly activity pattern for the station pair clusters, with the days plotted on the *x*-axis and mean activity levels on the *y*-axis. Figure 5(a) shows the activity pattern between stations in clusters 0 and 1. The figure shows that the activity relates very closely to commuter patterns, with a lot of activity on weekdays peaking in the morning and in the afternoon. The two clusters show a small difference in the timing of the morning peak, with stations in cluster 1 peaking at 7 a.m., while stations in cluster 0 peak marginally later at 8 a.m.

From the three clusters found, cluster 0 contains the most station pairs, with 36%. Comparing the types of stations that form the pairs in this cluster reveals that 40% of the pairs contain a station that was previously classified as "residential" and another 40% contain a station that was previously classified as "transport." On top of this, the majority of the pairs, 54%, had a station that was in one of the two clusters previously identified to be commuter destinations (clusters C and D). This shows that one of the most predominant uses of the London shared-bike scheme is for commuting purposes.

Cluster 1 contains a smaller portion of the station pairs, with 21%, and shows a difference in the station types in the pairs. Only 30% of the pairs contain a station that was previously classified as "residential," and 33% contain a station from "transport." In comparison to the pairs in cluster 0, a significantly larger portion of the pairs contain a station identified as a commuter destination, at 66%. This shows that in this pair cluster, station pairs are used more for the final part of a commute, while those in cluster 0 are used more at the beginning or for shorter commutes.

Figure 5(b) shows the weekly patterns for pairs in cluster 2; this cluster has the smallest portion of the station pairs and shows a very different pattern than any of the other clusters. There are a lot more activities during the weekends and outside of normal commuter hours. Most of the station pairs in this cluster, 55%, contain a station that was previously classified as "tourist," while only 5% contain a station that was in cluster D, which almost exclusively showed commuter patterns. This would suggest that station pairs in this cluster are mainly used for tourist and leisure purposes.

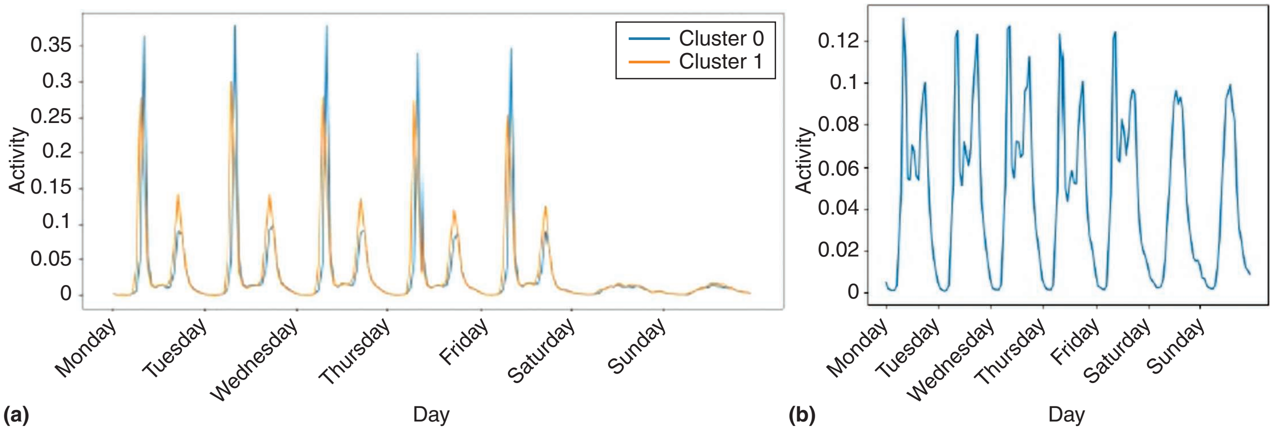


FIGURE 5. The weekly activity pattern for (a) station-pair clusters 0 and 1 and (b) cluster 2.

ABOUT THE AUTHORS

SUPARNA DE is a lecturer in the Department of Computer Science at the University of Surrey, Guildford GU2 7XH, U.K. Her research interests include data and knowledge engineering as well as deep learning applied to text data. De received a Ph.D. in electronic engineering from the University of Surrey. She is a member of IEEE. Contact her at s.de@surrey.ac.uk.

WEI WANG is an associate professor with the Xi'an Jiaotong Liverpool University, Suzhou 215123, China. His research interests include data processing and machine learning applications. Wang received a Ph.D. in computer science from the University of Nottingham, Malaysia. Contact him at wei.wang03@xjtlu.edu.cn.

USAMAH JASSAT is a software developer at Amazon AWS, Cambridge CB1 2GA, U.K. and was previously with the University of Surrey. His research interests include machine learning. Jassat received an M.Sc. in computer science from Loughborough University. Contact him at usamah.jassat@gmail.com.

KLAUS MOESSNER is professor for communications engineering at the Chemnitz University of Technology, Chemnitz 09107, Germany. His research interests include the area of collaborative situation awareness and reliable connectivity for future mobility. Moessner received a Ph.D. from the University of Surrey. He is a Senior Member of IEEE. Contact him at klaus.moessner@etit.tu-chemnitz.de.

The analysis performed on the London BSS data was able to discover multiple unique patterns that can be explained and related to the different usages across the city of London. When looking at the arrivals and departures at the bike stations across London, five unique behavior patterns were discovered. Although all of these patterns displayed two peaks on weekdays (reflecting their use by commuters at the start and end of the working day) and a single peak on weekends (reflecting their use for leisure activities or exercise), the differences in magnitude of these peaks and in between these peaks were strikingly different. The discovered patterns showed that most travel using the bike is dictated by commuters; however, leisure activities in areas also had a strong impact and resulted in notable changes to the usage of the bike stations. The developed model can be easily replicated and applied to other cities where bike trips are captured in terms of the date, time, and location of their start and end.

With the inclusion of the scaling factor in the developed model, both the temporal characteristics and the magnitude of activity at stations are captured. This enables a good insight into urban mobility in the city, such as a comparison of commuter behavior before and after certain events, such as the coronavirus lockdown. In addition to facilitating the recognition of evolving functions of particular stations (that is, for commuting or leisure), new emerging transportation hubs can also be identified. The latter can also help the business profile of the BSS by optimizing the bike redistribution policy and planning for siting new stations (the location as well as the bike fleet size). A limitation of the Poisson construction of the model is the assumption of independence between the arrival and departure events at each hour. However, this may not hold depending upon the mixed nature of the station neighborhood.

ACKNOWLEDGMENTS

This work was supported through the UK Research and Innovation Strategic Priority Fund as part of the wider Protecting Citizens Online program (grant EP/W032473/1) and is associated with the National Research Center on Privacy, Harm Reduction, and Adversarial Influence Online.

REFERENCES

1. S. De, W. Wang, Y. Zhou, C. Perera, K. Moessner, and M. N. Alraja, "Analysing environmental impact of large-scale events in public spaces with cross-domain multimodal data fusion," *Computing*, vol. 103, no. 9, pp. 1959–1981, 2021, doi: 10.1007/s00607-021-00944-8.

2. Y. Guo, X. Shen, Q. Ge, and L. Wang, “Station function discovery: Exploring trip records in urban public bike-sharing system,” *IEEE Access*, vol. 6, pp. 71,060–71,068, Oct. 2018, doi: 10.1109/ACCESS.2018.2878857.
3. 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 Trans. Intell. Syst. Technol.*, vol. 5, no. 3, pp. 1–21, 2014, doi: 10.1145/2560188.
4. “Five takeaways from the ITS World Congress 2021,” Cycling Industries Europe, Brussels, Belgium, 2021. [Online]. Available: https://cyclingindustries.com/news/details/five-takeaways-from-the-its-world-congress-2021
5. S. Jiang, J. Ferreira, and M. C. Gonzalez, “Activity-based human mobility patterns inferred from mobile phone data: A case study of Singapore,” *IEEE Trans. Big Data*, vol. 3, no. 2, pp. 208–219, Jun. 2017, doi: 10.1109/TBDATA.2016.2631141.
6. Y. Zhang, B. Li, and J. Hong, “Using online geotagged and crowdsourced data to understand human offline behavior in the city,” *ACM Trans. Intell. Syst. Technol.*, vol. 9, no. 3, pp. 1–24, 2018, doi: 10.1145/3078851.
7. S. Zheng, S. Xie, and X. Chen, “Discovering urban functional regions with call detail records and points of interest: A case study of Guangzhou city,” in *Proc. 11th Int. Conf. Wireless Commun. Signal Process. (WCSP)*, 2019, pp. 1–6, doi: 10.1109/WCSP.2019.8927961.
8. X. Yang, S. He, and H. Huang, “Station correlation attention learning for data-driven bike sharing system usage prediction,” in *Proc. IEEE 17th Int. Conf. Mobile Ad Hoc Sensor Syst. (MASS)*, 2020, pp. 640–648, doi: 10.1109/MASS50613.2020.00083.
9. J. Froehlich, J. Neumann, and N. Oliver, “Sensing and predicting the pulse of the city through shared bicycling,” in *Proc. 21st Int. Joint Conf. Artif. Intell.*, ser. IJCAI’09, 2009, pp. 1420–1426.
10. P. Vogel and D. C. Mattfeld, “Strategic and operational planning of bike-sharing systems by data mining – A case study,” in *Proc. Int. Conf. Comput. Logistics*, 2011, pp. 127–141.
11. P. Borgnat, E. Fleury, C. Robardet, and A. Scherrer, “Spatial analysis of dynamic movements of Vélo’v, Lyon’s shared bicycle program,” in *Proc. ECCS’09*, 2009, pp. 1–7.
12. Y. Li, Z. Zhu, D. Kong, M. Xu, and Y. Zhao, “Learning heterogeneous spatial-temporal representation for bike-sharing demand prediction,” in *Proc. 33rd AAAI Conf. Artif. Intell.*, 2019, vol. 33, pp. 1004–1011, doi: 10.1609/aaai.v33i01.33011004.
13. L. Lin, Z. He, and S. Peeta, “Predicting station-level hourly demand in a large-scale bike-sharing network: A graph convolutional neural network approach,” *Transp. Res. C, Emerg. Technol.*, vol. 97, pp. 258–276, Dec. 2018, doi: 10.1016/j.trc.2018.10.011.
14. C. Rudloff and B. Lackner, “Modeling demand for bikesharing systems: Neighboring stations as source for demand and reason for structural breaks,” *Transp. Res. Rec. J. Transp. Rec. Board*, vol. 2430, no. 1, pp. 1–11, 2014, doi: 10.3141/2430-01.
15. Y. Li, Y. Zheng, H. Zhang, and L. Chen, “Traffic prediction in a bike-sharing system,” in *Proc. 23rd SIGSPATIAL Int. Conf. Adv. Geographic Inf. Syst.*, 2015, pp. 1–10, doi: 10.1145/2820783.2820837.
16. J. Yang, B. Guo, Z. Wang, and Y. Ma, “Hierarchical prediction based on network-representation-learning-enhanced clustering for bike-sharing system in smart city,” *IEEE Internet Things J.*, vol. 8, no. 8, pp. 6416–6424, 2021, doi: 10.1109/JIOT.2020.3034991.
17. W. Jia, Y. Tan, L. Liu, J. Li, H. Zhang, and K. Zhao, “Hierarchical prediction based on two-level gaussian mixture model clustering for bike-sharing system,” *Knowl. Based Syst.*, vol. 178, pp. 84–97, Aug. 2019, doi: 10.1016/j.knosys.2019.04.020.
18. K. Kim, “Spatial contiguity-constrained hierarchical clustering for traffic prediction in bike sharing systems,” *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 6, pp. 5754–5764, 2021, doi: 10.1109/TITS.2021.3057596.
19. A. P. Dempster, N. M. Laird, and D. B. Rubin, “Maximum likelihood from incomplete data via the EM algorithm,” *J. Roy. Statistical Soc., B (Methodological)*, vol. 39, no. 1, pp. 1–22, 1977, doi: 10.1111/j.2517-6161.1977.tb01600.x.
20. T. P. Development Team. *Pandas-Dev/Pandas: Pandas*. (Feb. 2020). Zenodo. [Online]. Available: https://zenodo.org/record/4394318#.YtvKBKFBxnJ
21. “Santander cycles.” Transport for London. https://tfl.gov.uk/modes/cycling/santander-cycles (Accessed: Jun. 11, 2022).
22. “National travel survey,” GOV. UK, London, U.K., 2013. [Online]. Available: https://www.gov.uk/government/collections/national-travel-survey-statistics
23. “Cycling data.” Transport for London. https://cycling.data.tfl.gov.uk (Accessed: Jun. 11, 2022).
24. “TfL API.” Transport for London. https://api.tfl.gov.uk/bikepoint (Accessed: Jun. 11, 2022).
25. NumPy. [Online]. Available: https://numpy.org (Accessed: Jun. 11, 2022).