Department: Head
Editor: Name, xxxx@email

Usage Mining of London Santander Bike Sharing System

Suparna De
Dept. of Computer Science, University of Surrey, UK

Wei Wang
Xi'an Jiaotong Liverpool University

Usamah Jassat
University of Surrey, UK

Klaus Moessner
Chemnitz University of Technology

***Abstract*—With cycling moving from being a pastime and sport to a mainstream form of mobility and transport, bike sharing systems (BSS) are increasingly being deployed in many cities. Analysis of the BSS usage data can provide insights into factors that shape the patterns of trips, uncovering latent city dynamics. A Poisson mixture model is proposed to cluster the stations according to their usage profiles and reveal latent links between the social and economic activities of BSS station neighbourhood type and the generated mobility patterns. It reveals the varying functions of different urban areas that induce specific bike trip patterns. Pairwise clustering of bike station with appreciable trip activity between them further advance the understanding of urban neighbourhoods with the strongest mobility patterns. The results are showcased through an analysis of 15 million bike journeys of the London Santander Cycles BSS over a 3-year period.**

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 (BSS) have been implemented in more than 450 cities worldwide [2]. BSS are characterised 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, tube and train stations, BSS offer a softer public transport alternative which is more

 Published by the IEEE Computer Society © 2022 IEEE

affordable, healthy and less polluting, and good opportunities for meeting the last mile commuting challenge [2], [3].

With the 2021 ITS 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 BSS in their urban policy plans, it has been recognised that in addition to the BSS infrastructure provision, a step change is needed. This means the provision of software, platforms and applications for managing and analysing the bike fleets, e.g. cyclists and BSS traffic flows in various city regions [4].

As cities continue to grow with sustained migration, connections between different neighbourhoods become more complex by 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, as well as giving 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, development of new techniques more suitable for the BSS transport data that is mainly about short-distance trips is needed. The spatio-temporal nature of the correlations of the BSS data with the neighbourhoods also needs to be considered [8]. It has been recognised that the clustering of BSS stations is tied to their different functions that are in turn linked to the city's activities, e.g. residential, leisure, employment [2].

For this, we present a method for BSS station clustering using the expectation maximisation (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 spatio-temporal 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, that 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[1].

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) activities. The proposed mixture model also directly handles the differences in weekday/weekend behaviour, rather than through data pre-processing or feature construction.

The remainder of this paper is organised as follows: we begin with a survey of related work. Then we present our proposed model based on count series clustering. This is followed by a description of the London BSS dataset. Implementation details of the clustering model and BSS station clustering results are presented next, followed by the station pair clustering approach and corresponding results. The paper concludes with a discussion of the achieved results, limitations of the modelling approach and the potential applications of this research.

RELATED WORK

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[2]. The growth of the urban computing paradigm, that uses statistical and machine learning techniques for deriving patterns in large-scale urban datasets, has led to mining of mobility patterns from taxi trips 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 spatio-temporal correlation with the bike station neighbourhood [2].

[1] https://tfl.gov.uk/modes/cycling/santander-cycles
[2] https://www.gov.uk/government/collections/national-travel-survey-statistics

Research using BSSs has employed either clustering methods to find bike station partitions that have similar usage, or prediction techniques for forecasting the occupancy of stations and station traffic towards bike rebalancing and scheduling optimisation.

Initial studies on temporal pattern mining from bike usage data consider statistical features such as historical average/trend 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) activities. 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 *et al.* [3] that proposed a count series model to predict hidden station clusters. The resulting clusters include those that are related to commuting (i.e. stations located close to public transport and mostly active during the morning and evening on weekdays). Another BSS mobility study [11] investigates the spatial analysis of bike trips by visualising 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 modelling a station as a document in a Latent Dirichlet Allocation (LDA) algorithm, with station functions derived as the topics of a document. Other studies utilise spatio-temporal features, such as impact of points-of-interest (POI) [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 behaviour directly into the mixture model parameters, rather than through pre-processing 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 city, cluster or station-level (e.g. bipartite clustering [15], Xgboost [16], Gaussian mixture model (GMM) [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 utilised to infer underlying unseen probability density distribution. The activity behaviour of bike stations is modelled 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, $P_1, P_2, ...P_k$, 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 (i.e. not directly observed) in the observed bike trip data.

Notation: in this section, we employ lowercase letters for variables and its corresponding uppercase equivalent for the overall summation value for the variable. For instance, k represents a station cluster and K represents the total number of station clusters, i.e. $k \in \{1, ..., K\}$.

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 $h \in \{1, 2, ...24\}$ on a given day $d \in \{1, ...D\}$. The quantisation of 1-hour is deemed as a good trade-off between data resolution and fluctuations in departure/arrival counts, in line with existing literature on bike usage modelling [11], [3]. 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 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 dataset (corresponding to the data collection period: January 2015 – May 2017), and T is 48, since the arrival and departure counts in a day are computed in 1-hour non-overlapping windows (i.e., 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 $S \times K$, (K as specified in Equation 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 in order to cluster them.

In addition to using 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 stations usage on weekdays and weekends, as these present very different usage profiles, with $W_{d0} = 1$ indicating that the day d is a weekend and $W_{d1} = 1$ indicates a weekday, i.e. l denotes the day (weekday/weekend) cluster membership of the station. $D_l = \sum_d W_{dl}$ 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 Equation 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 is calculated as below:

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

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

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

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

Taking into account the above, 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 (MLE) of these parameters. For this, the log-likelihood is first derived by substituting P_k from Equation 4 into the mixture model equation (Equation 1), summing over all k and taking the logarithm of the function. Instead of

estimating the parameters' MLE directly through numerical optimisation, the EM algorithm is used to maximise the log-likelihood.

EM Algorithm

The parameters of the mixture model can be estimated by using the expectation maximization (EM) algorithm [19], which is used for obtaining MLE of parameters when there is latent, i.e. 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 $aposteriori$ 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 Equation 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 $aposteriori$ 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 time frame t. The E and M steps are iterated until the parameters converge to the local maximum of the log-likelihood function.

DATASET

The dataset is sourced from the Transport for London (TfL) cycling open data website[3], usage-stats section, that has data on all Santander Cycles journeys. The TfL data is available as downloadable comma-separated-values (CSV) 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 trips' data for a 3-year period, from 4th January, 2015 to 16th May, 2017, which after parsing and cleaning contains over 15 million trips. Cleaning the dataset included removing erroneous or invalid trip data, i.e. removing any journeys with a duration of 0 seconds. 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 majority of the journeys in the dataset (98%) were less than an hour. The station IDs are mapped to the station name and location (latitude/longitude) using the TfL Unified API and querying for 'BikePoint'[4]. The resulting data schema is shown in Fig. 1.

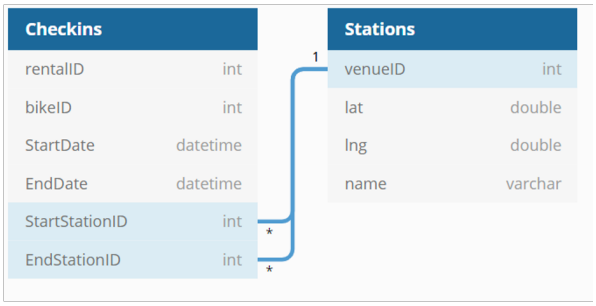

Figure 1. Schema of bike journey data

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 AMD Ryzen5 5600X CPU and 16GB RAM. Loading and pre-processing (parsing and cleaning) of the data is performed using the Pandas library [20] as it provides functionality to load large amounts of table data and to easily filter and sort it. Numerical calculations for tensor manipulation and operations were performed using the Numpy library[5]. The maximum probability of each station was used to determine the cluster

[3]https://cycling.data.tfl.gov.uk

[4]https://api.tfl.gov.uk/bikepoint

[5]https://numpy.org

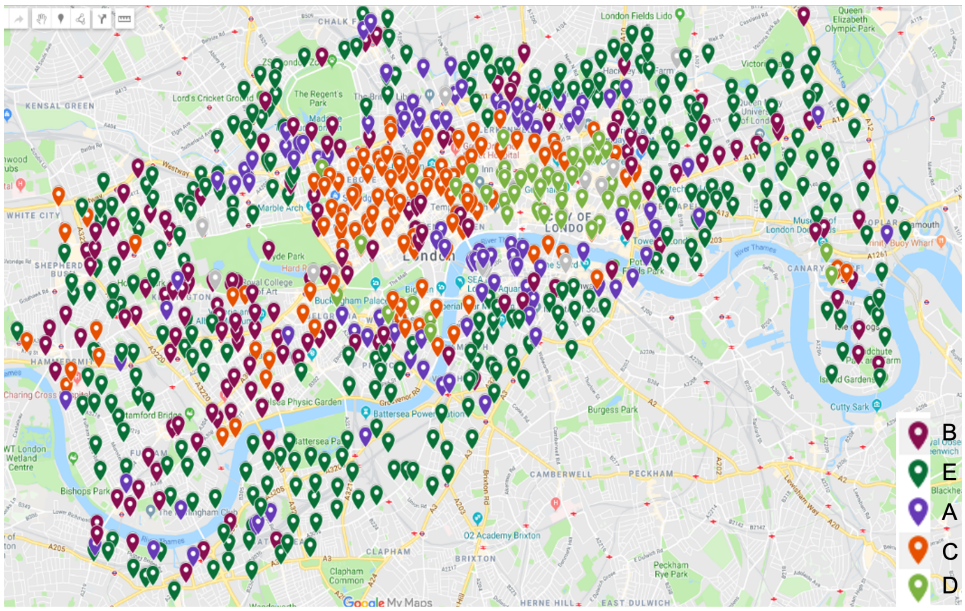

Figure 2. Map of bike station clusters

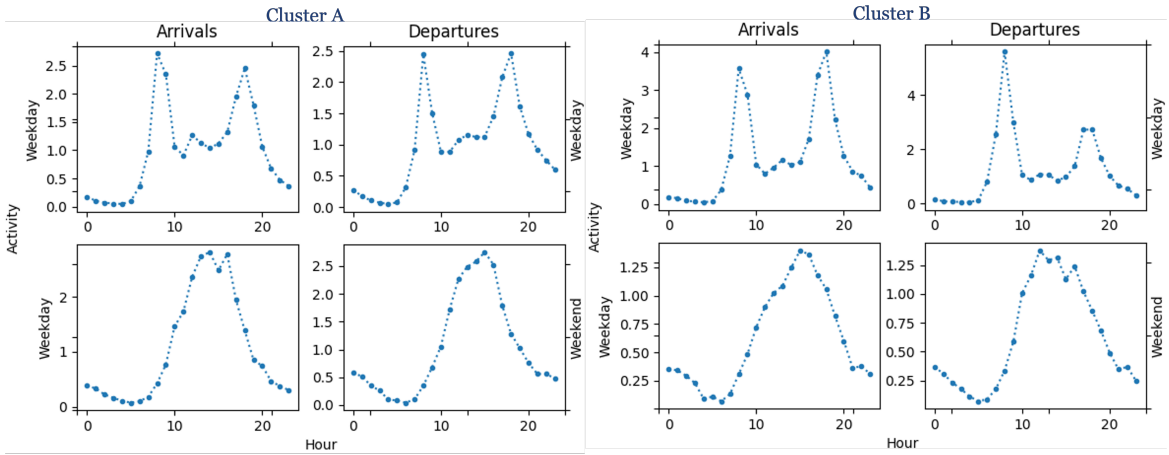

Figure 3. Mean activity pattern for stations in Cluster A: Leisure & Tourism (close to tourist attractions, commercial areas and public transport) and Cluster B: Transport (close to public transport and train stations).

that a station belonged to. Check-in data across all stations for a period of one week was 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 analysed by the elbow-method heuristic [3] which shows an elbow in the curve at $K = 5$. Hence, the value of 5 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. Time to reach convergence was 18 minutes, with each EM run taking between 90 - 134 seconds with run durations increasing with the later runs.

STATION CLUSTERING RESULTS

Figure 2 shows a map of London with the location of the bike station coloured by the cluster that they belong 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-hour scale on the x-axis and the y-axis depicting the normalised count of bike arrivals/departures (corresponding to the normalised X_s^{in}/X_s^{out}, respectively, of all stations in that cluster).

Cluster A - Leisure and Tourism

The left half of Figure 3 shows the activity pattern for cluster A. On weekdays, the arrivals and departures patterns were very similar with peaks in the morning (around 8am) and evening (7pm). 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 were used during commuter times on weekdays. This suggests that these stations are also used heavily for tourism and leisure activities besides 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

The right half of Figure 3 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, it 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 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 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 origins for commuting back home in the evenings. The locations of these stations are 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 this cluster is their activity outside of commuting hours and on the 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 whilst 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 theatres.

Cluster E - Residential

The rightmost part of Figure 4 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 these stations are in residential areas and 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, it shows that they are predominantly located on the outskirts of London which is less industrially dense than the centre and contains 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 station for the morning commute and origin station for the

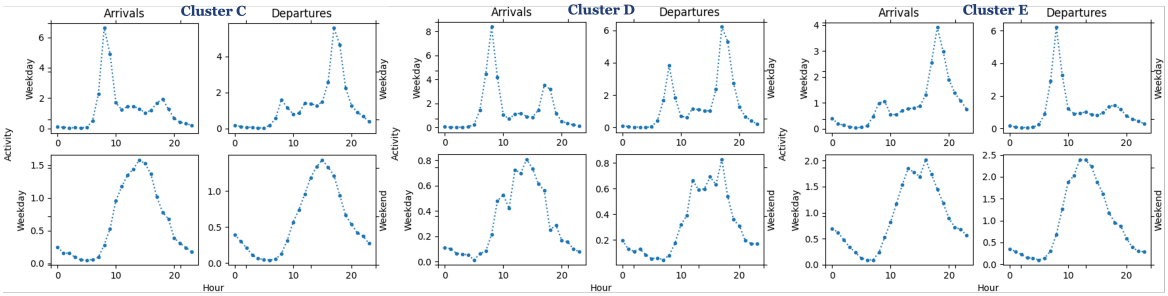

Figure 4. Mean activity pattern for stations in Clusters C: Work & Leisure (high morning arrivals and high evening departures, more weekend activity, predominantly in west central London); D: Work (high morning arrivals and high evening departures, less weekend activity, predominantly in east central London); and E: Residential (peak arrivals in the evening and high departures in the morning).

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

In order 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 neighbourhood 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 socio-economic 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 dataset, 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 grows rapidly with the number of stations. In order to do this, the first step is to find the principal components of the activity patterns, which enable to summarise the data without loss of information by transforming correlated attributes into non-correlated 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 7 days of the week and 24 hours per day. The principal components are found by using Principal Component Analysis (PCA), and selected such that 90% of the variance within the data is 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 activities between them. Stations might have similar temporal activity patterns; however, this pattern may vary in scale from station to station. So, before clustering, the stations' data is 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 (i.e. $K = 3$) 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 at which the graph forms an 'elbow', i.e. 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. The left half of the plot shows the activity pattern between stations in cluster 0 and 1. The figure shows 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 am whilst stations in cluster 0 peak fractionally later at 8am.

From the 3 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, whilst those in cluster 0 are used more at the beginning or for shorter commutes.

The right half of Figure 5 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" whilst only 5% contain a station that was in cluster D, which almost exclusively showed commuter patterns. This would suggest station pairs in this cluster are mainly used for tourist and leisure purposes.

CONCLUSION

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 behaviour 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 difference 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 its start and end.

With the inclusion of the scaling factor in the developed model, both the temporal characteristics as well as magnitude of activity at stations is captured. This enables a good insight into urban mobility in the city, such as comparison of commuter behaviour before and after certain events, such as the coronavirus lockdown. In addition to facilitating recognition of evolving functions of particular stations (i.e. for commuting or leisure), new emerging transportation hubs can also be identified. The latter can also help the business profile of the BSS: to optimise the bike redistribution policy and plan for citing new stations (location as well as 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 neighbourhood.

ACKNOWLEDGMENT

This work has been supported through the UKRI Strategic Priority Fund as part of the wider Protecting Citizens Online programme (Grant number: EP/W032473/1) and is associated with the National Research Centre on Privacy, Harm Reduction and Adversarial Influence Online (REPHRAIN).

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.
2. Y. Guo, X. Shen, Q. Ge, and L. Wang, "Station function discovery: Exploring trip records in urban public

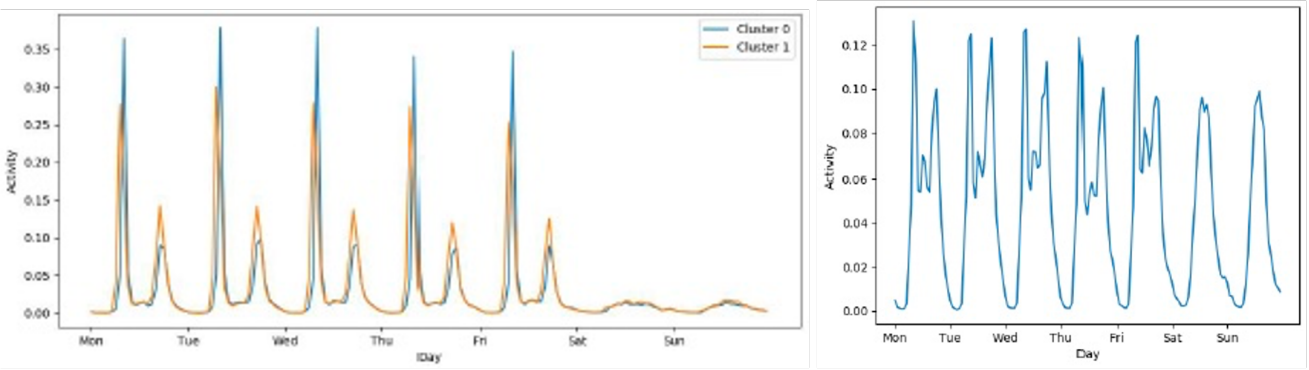

Figure 5. Weekly activity pattern for station pair cluster 0 and 1 (left) and cluster 2 (right)

bike-sharing system," *IEEE Access*, vol. 6, pp. 71 060–71 068, 2018.

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, 2014.
4. Cycling Industries Europe, "Five takeaways from the ITS World Congress 2021," https://cyclingindustries.com/news/details/five-takeaways-from-the-its-world-congress-2021, 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 Transactions on Big Data*, vol. 3, no. 2, pp. 208–219, jun 2017.
6. Y. Zhang, B. Li, and J. Hong, "Using online geotagged and crowdsourced data to understand human offline behavior in the city," *ACM Transactions on Intelligent Systems and Technology*, vol. 9, no. 3, pp. 1–24, 2018.
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 *11th International Conference on Wireless Communications and Signal Processing (WCSP)*, 2019, pp. 1–6.
8. X. Yang, S. He, and H. Huang, "Station correlation attention learning for data-driven bike sharing system usage prediction," in *IEEE 17th International Conference on Mobile Ad Hoc and Sensor Systems (MASS)*, 2020, pp. 640–648.
9. J. Froehlich, J. Neumann, and N. Oliver, "Sensing and predicting the pulse of the city through shared bicycling," in *Proceedings of the 21st International Joint Conference on Artificial Intelligence*, ser. IJCAI'09, 2009, p. 1420–1426.
10. P. Vogel and D. C. Mattfeld, "Strategic and operational planning of bike-sharing systems by data mining – a case study," in *Computational 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 *ECCS'09*, 2009.
12. Y. Li, Z. Zhu, D. Kong, M. Xu, and Y. Zhao, "Learning heterogeneous spatial-temporal representation for bike-sharing demand prediction," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, pp. 1004–1011, 2019.
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," *Transportation Research Part C: Emerging Technologies*, vol. 97, pp. 258–276, 2018.
14. C. Rudloff and B. Lackner, "Modeling demand for bike-sharing systems: Neighboring stations as source for demand and reason for structural breaks," *Transportation Research Record*, vol. 2430, no. 1, pp. 1–11, 2014.
15. Y. Li, Y. Zheng, H. Zhang, and L. Chen, "Traffic prediction in a bike-sharing system," in *Proc. 23rd SIGSPATIAL International Conference on Advances in Geographic Information Systems*, 2015.
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 of Things Journal*, vol. 8, no. 8, pp. 6416–6424, 2021.
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, 2019.
18. K. Kim, "Spatial contiguity-constrained hierarchical clustering for traffic prediction in bike sharing systems," *IEEE Transactions on Intelligent Transportation Systems*, pp. 1–11, 2021.
19. A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maxi-

mum likelihood from incomplete data via the EM algorithm," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 39, no. 1, pp. 1–22, 1977.

20. T. P. development team, "pandas-dev/pandas: Pandas," Feb. 2020. [Online]. Available: https://doi.org/10.5281/zenodo.3509134

Suparna De is a lecturer in the Department of Computer Science at the University of Surrey, Guildford, GU2 7XH, U.K. Her research interests are in data and knowledge engineering and deep learning applied to text data. She received her PhD and MSc in Electronic Engineering from the University of Surrey, UK. 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, China. His research interests are in data processing and machine learning applications. He received his PhD 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 and was previously with the University of Surrey, U.K. His research interest is in machine learning. He received his MSc in Computer science from Loughborough University and BEng in Electronic Engineering from the University of Surrey, UK. Contact him at usamah.jassat@gmail.com.

Klaus Moessner is professor for communications engineering at the Chemnitz University of Technology, Germany. His research interests are in the area of collaborative situation awareness and reliable connectivity for future mobility. He received his PhD from the University of Surrey, UK, MSc at Brunel University, UK, and Dipl.-Ing.(FH) at the University of Applied Science in Offenburg, Germany. He is a senior member of IEEE. Contact him at klaus.moessner@etit.tu-chemnitz.de.