Detecting Spatial Patterns Through Data Mining Techniques: a Cluster Analysis of the London Cycle Hire System

Andrea Sibilia[*1] and James Haworth[†2]

[1]Smart Consulting, WSP, 70 Chancery Ln, London WC2A 1AF
[2] SpaceTimeLab, Department of Civil Environmental & Geomatic Engineering, University College London

January 17th 2019

Summary

The study of bicycle hire systems is an active field of research due to the large amount of data they are able to generate. However, most of the studies develop their analysis by looking at the flows recorded at the docking stations without exploring the actual movements carried out by single bikes. This research aims to fill this gap through a cluster analysis of the bikes which belong to the London Cycle Hire System, in order to explore their activity pattern.

KEYWORDS: bike sharing, data mining, clustering, Origin-Destination data, flow map

1. Introduction

Bike-sharing systems are a very popular and sustainable alternative to intensive car use. Since Paris launched the world's largest cycling system in 2007, hundreds of cities all over the world have followed suit, with London being a notable example (DeMaio, 2009). The increasing number of cycle hire systems was followed by new technologies used to collect real time cycling data (Côme *et al.*, 2014; Corcoran *et al.*, 2014). This has led to an increase in research aiming to explain the different aspects of cycling systems.

A wide range of studies have investigated bike-share systems across the world using a range of data mining approaches such as clustering (Côme *et al.*, 2014) to detect spatio-temporal patterns or to create forecasting models. The majority of these studies use aggregate data, which means that the information is collected and/or aggregated at the docking station level. However, there is still a lack of knowledge about how individual bikes move within the cycling network. Understanding these movements can be useful in order to understand the self-organisation of the network, and how to optimise redistribution.

1.1. Research questions

The hypothesis tested is that individual bikes generate specific spatial patterns at the aggregate level through their journeys within the London Cycle Hire System (LCHS). Therefore, this study seeks to address the following questions:

1. Are there meaningful spatial patterns or does each single bike move randomly within the cycling network?
2. How can patterns of movements of individual bikes be clustered to reveal spatial patterns at the network level?
3. Does the pattern recognition lead to significant findings that could be helpful in terms of planning and network management?

[*] Andrea.Sibilia@wsp.com
[†] J.haworth@ucl.ac.uk

2. Data

The datasets used in this study are the public TfL cycle hire journeys and the BikePoint dataset which contains the location of the LCHS docking stations. The TfL cycle hire journeys are origin-destination (OD) pairs and associated journey times, and are stored in a PostgreSQL database originally created by SpaceTimeLab at University College of London (UCL). The database is organised in a single table which contains the single hires which occurred within the LCHS from April 2012 to February 2018. This is equivalent to 49,901,598 records. The raw data are available at the TfL website (https://cycling.data.tfl.gov.uk/).
The BikePoint dataset is downloadable from the TfL unified API as a JSON script. This data provides the docking stations' coordinates in WGS84 geographic coordinate system expressed in decimal degrees, as well as the numerical code that identifies each station.

3. Methodology

The approach employed in this research follows three main steps:

1- Data exploration: an essential statistical descriptive analysis of the data;
2- Cluster analysis: the core of the analysis which produces the main results that are interpreted in the following step. This method allows research question number 1 to be answered;
3- Interpretation of the results: the final stage of the analysis, which returns the findings in an easy-to-read fashion. This step allows the usefulness of the outcomes to be assessed, providing the answer to research question number 2.

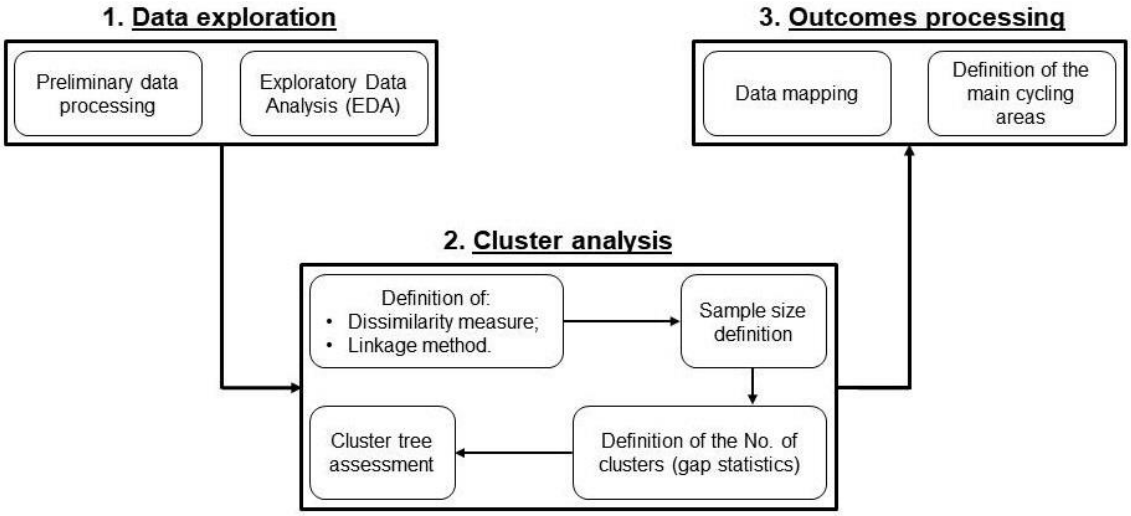

Figure 1 Diagram of the general workflow used to analyse the data.

The main research question aims to detect meaningful spatial patterns generated through the OD paths travelled by the bikes and the data exploration highlights the patterns that exist. It proves that some areas support a higher level of traffic, suggesting that bikes potentially move through a defined scheme.

Even though the general pattern appears homogeneous across the whole LCHS, it is not possible to assume that all the bikes move together within the system; this would oversimplify the problem. It is logical to expect that bikes tend to form groups with a similar behaviour, meaning that bicycles parked in the same area are likely to travel between similar destinations. Therefore, bikes which have similar behaviours will generate similar spatial patterns. This essential consideration leads to employing a clustering method to analyse the data, in order to find similar group of bikes and then proceed with the pattern recognition.

3.1. Cluster analysis

Clustering is the process of grouping data. It is one of the most important data mining techniques for discovering knowledge in complex datasets (Kaufman and Rousseeuw, 1990; Kassambara, 2017).
The method used in this research is agglomerative clustering, which is the most common type of hierarchical clustering. It is also known as AGNES (Agglomerative Nesting) and it is fully described by Kaufman and Rousseeuw (1990). This technique is chosen due to its flexibility and because in this case data needs to be classified with a process which is similar to the one used in biological application, such as the classification of animals and plants. Therefore, the algorithm results in the best choice in terms of outcomes. The main disadvantage resides in the computation time required to run the algorithm which is quadratic i.e. $O(n^2)$. Hence, the method proves to be less efficient for large datasets, therefore the analysis is performed on a subset of bikes.

4. Results and discussion

The clustering process returned six clusters of bikes which moved through specific schemes within the LCHS during the month of July 2017. Note that the bikes which belong to a cluster did not move together within the cycling network, but their routes share common docking stations. Therefore, within each cluster it is possible to detect the most frequently visited nodes. In other words, the way the clusters are built allows detection of the main cycling area for each group of bikes.

Examining figures 2 to 7, it can be seen that Clusters 1 and 7 are centred on Hyde Park, indicating that the bikes in those clusters remained in the Hyde Park area for most or all of the month. This demonstrates the popularity of the area for cycle hire trips and has implications for redistribution; if bikes continue to be picked up and dropped off at the same locations then intervention is required to balance the system. This rebalancing may be reflected in other clusters such as cluster 6, which has a more spatially dispersed pattern. Examining the clusters in general terms, it can be hypothesised that Hyde Park and the Olympic Park at Stratford act as sinks for bikes, which must be actively rebalanced.

An important aspect which stands out from the results is represented by the random component related to the paths travelled by the bikes. Clusters are based on the similarities between the routes that the bicycles have covered; these similarities are mainly focused on the most frequented docking stations. It is extremely complex to detect a kind of pattern affinity for the stations which are visited once or twice per month, since those visits are principally caused by random factors (i.e. the bikes are used by different customers who have different behaviours, and bikes can be moved in other areas of the city by Serco's staff who need to maintain the network in a balanced manner).

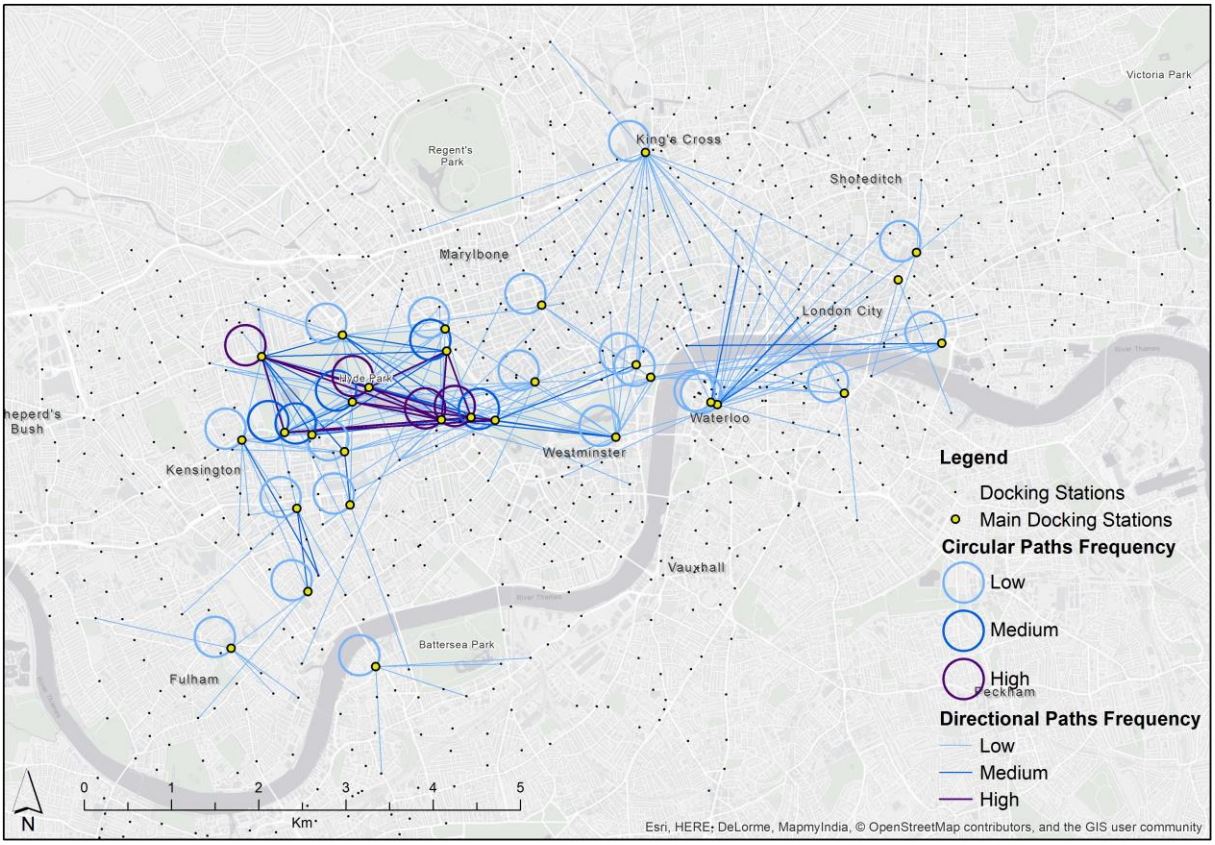

Figure 2 Cycling areas of Cluster 1. The yellow dots highlight the main docking stations, while lines and circles (journeys with same origin and destination point) represent paths covered by the bikes.

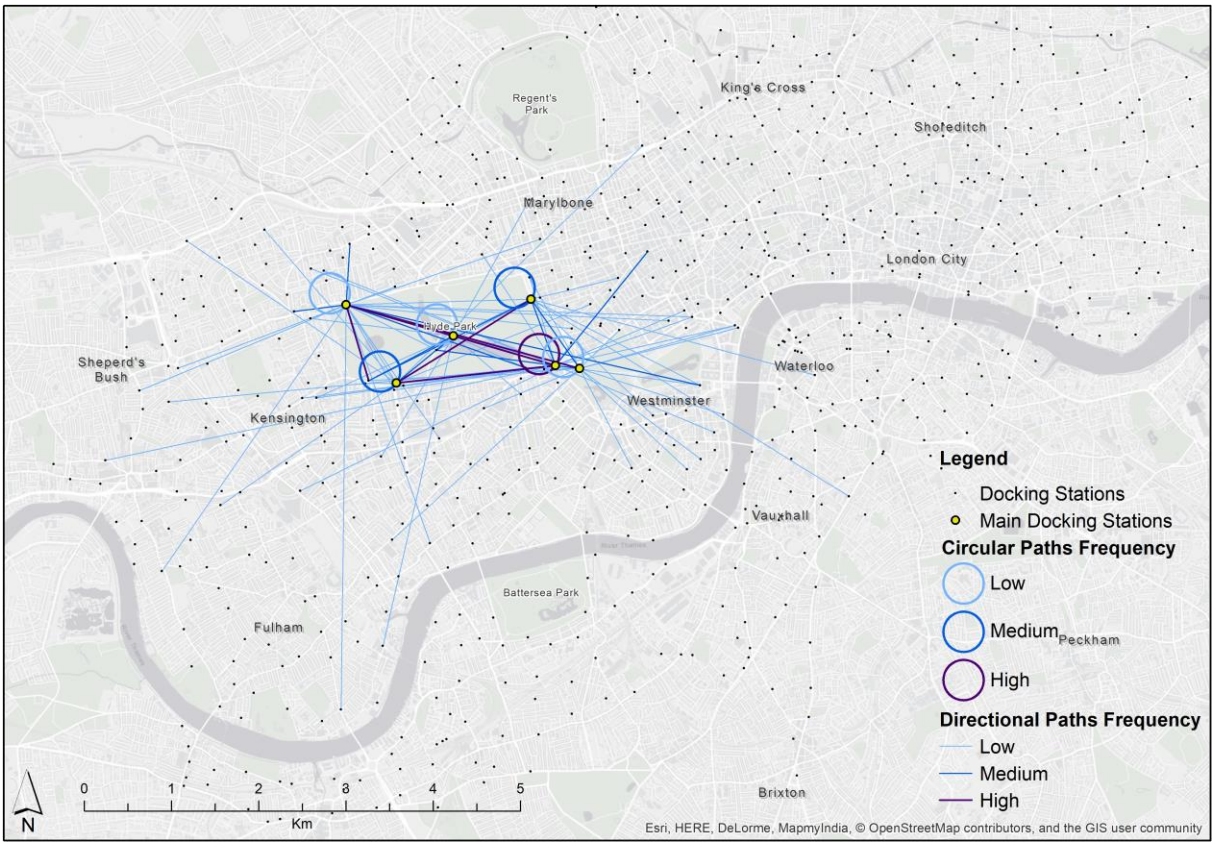

Figure 3 Cycling areas of Cluster 2.

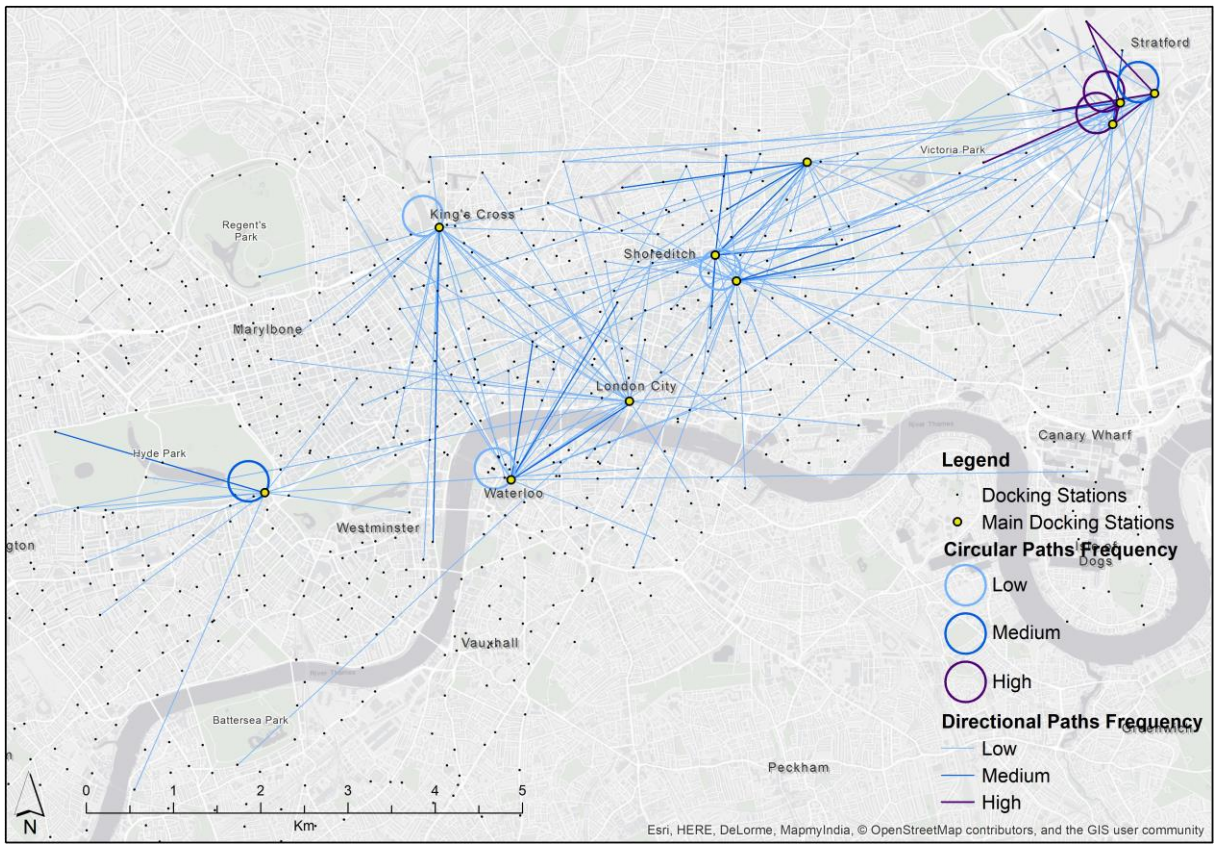

Figure 4 Cycling areas of Cluster 3.

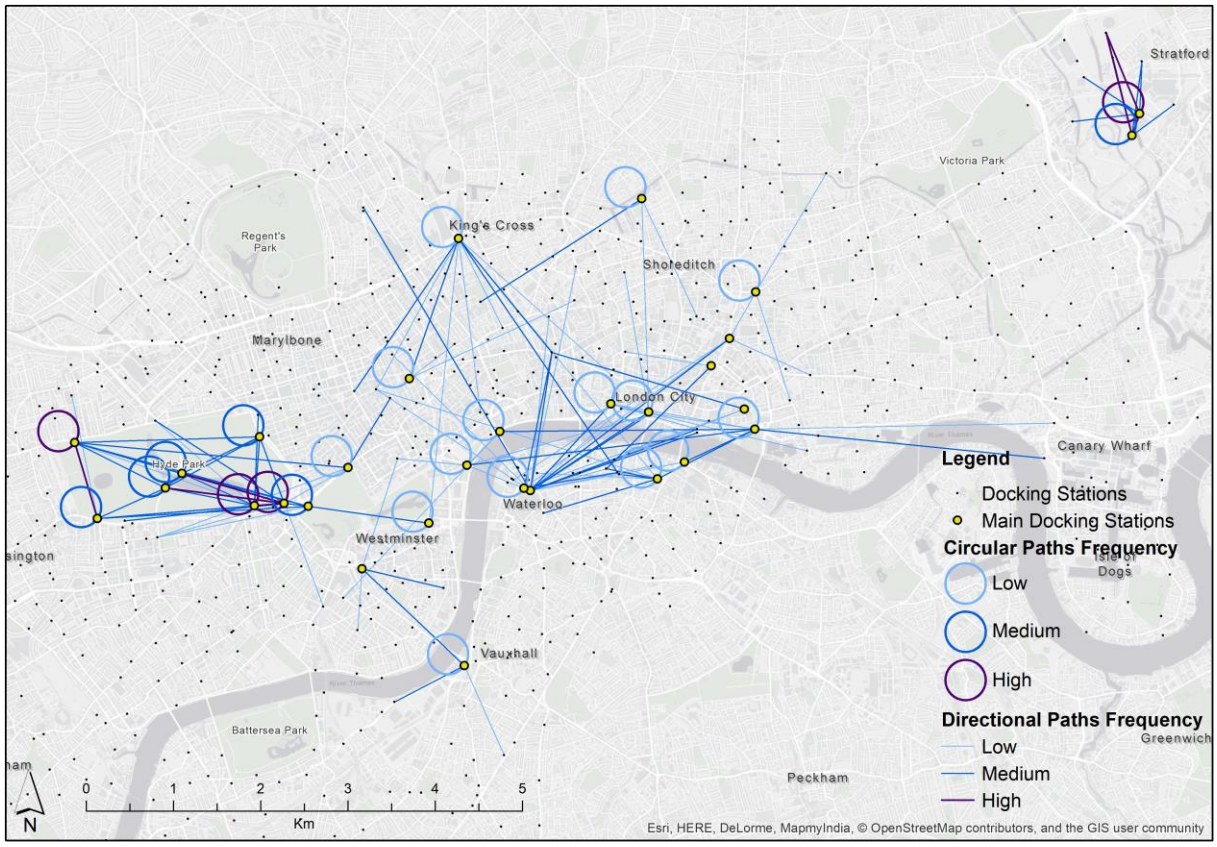

Figure 5 Cycling areas of Cluster 4.

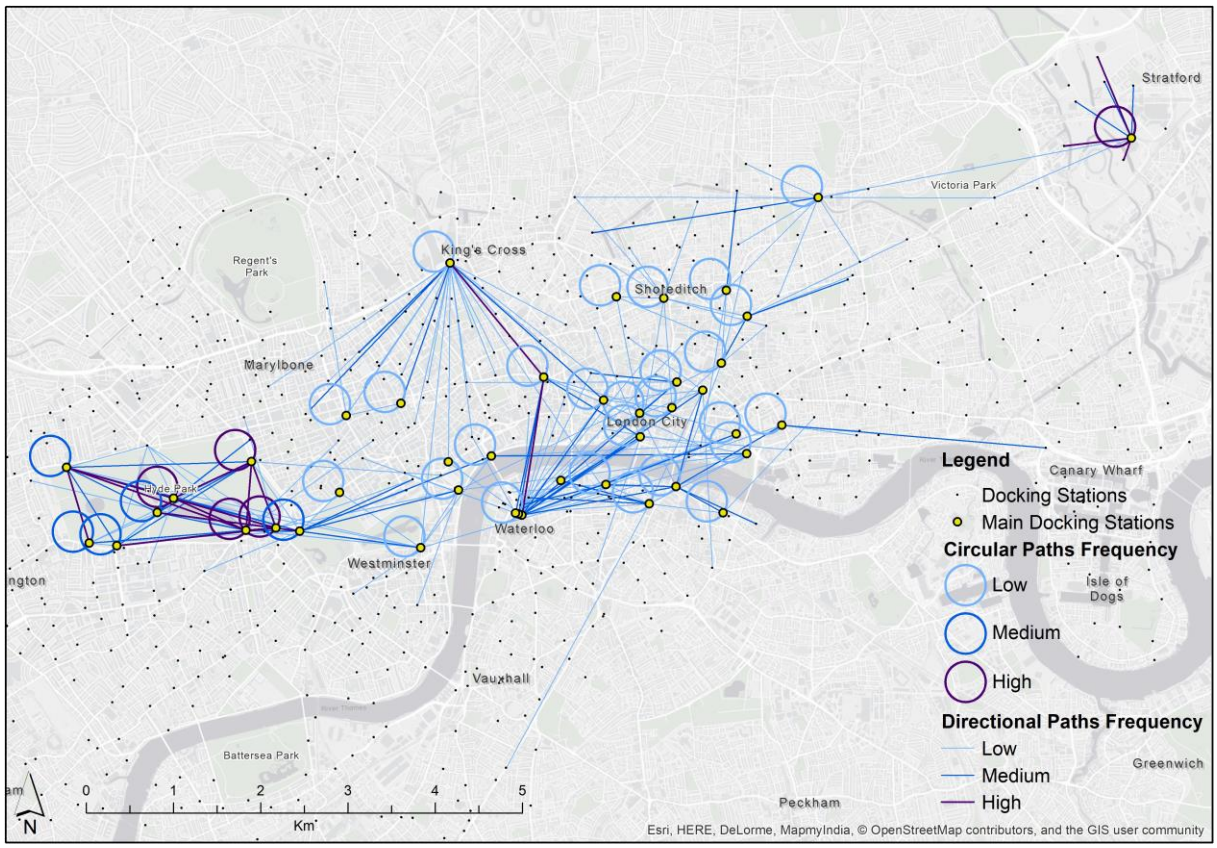

Figure 6 Cycling areas of Cluster 5.

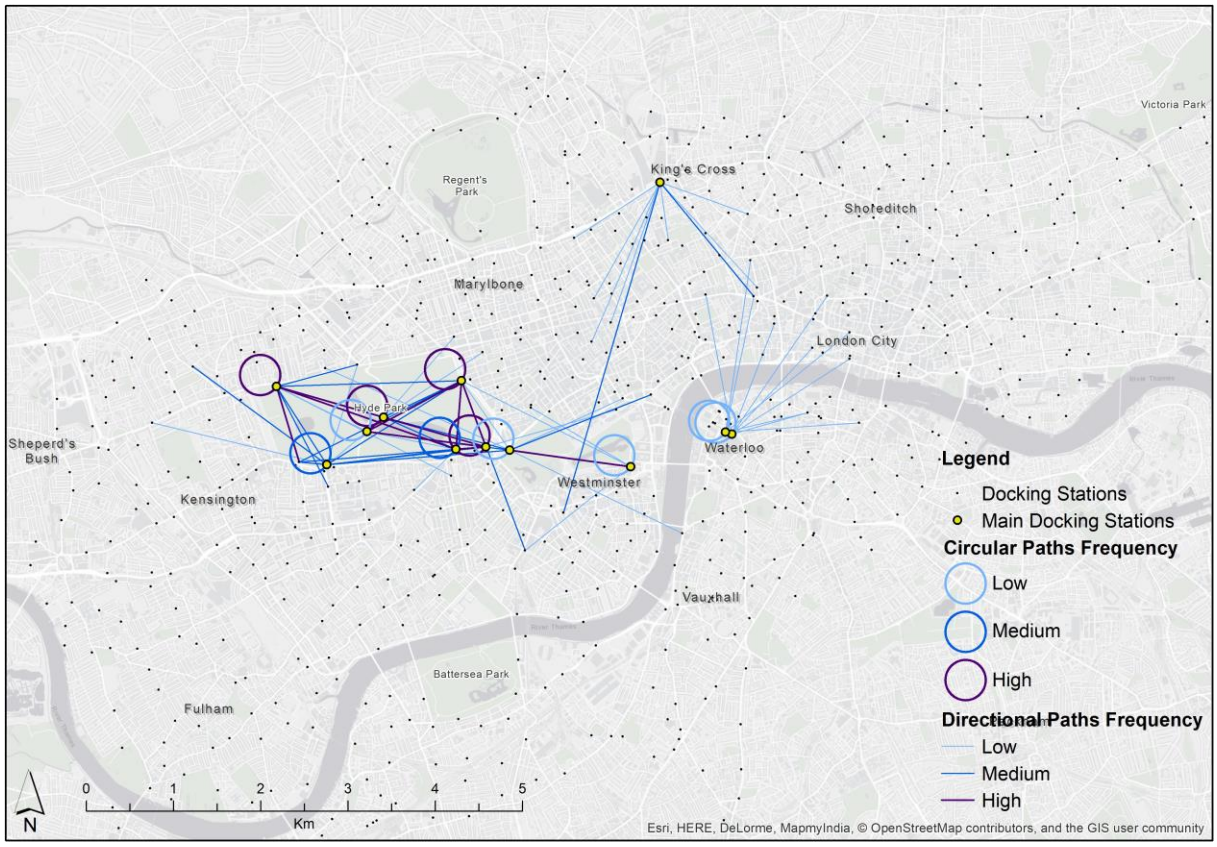

Figure 7 Cycling areas of Cluster 6.

5. Conclusions

The study aimed to analyse OD bike-share journeys freely available on the TfL website, to detect meaningful spatial patterns. Therefore, this project introduced a new perspective in this field of research, by looking at the data from the point of view of the bikes; instead of aggregating data at station level or analysing it from the customers' viewpoint.
The main analysis was performed through a hierarchical clustering which generated six clusters. The results demonstrate that the bikes of the London Cycle Hire System (LCHS) exhibit specific spatial patterns through their journeys, and certain geographic locations appear to act as sinks for bikes.
Overall, the district of central London is the most frequented area by cyclists (King's Cross, Waterloo and Hyde Park) as well as the district of Stratford, which also proved to be a highly frequented area.
Further research is needed to investigate the impact of the patterns of bike-share usage on redistribution strategies.

References

Côme, E. *et al.* (2014) 'Spatio-temporal analysis of Dynamic Origin-Destination data using Latent Dirichlet Allocation . Application to the Vélib ' Bike Sharing System of Paris .', *TRB 93rd Annual meeting*, pp. 0–18.

Corcoran, J. *et al.* (2014) 'Spatio-temporal patterns of a Public Bicycle Sharing Program: The effect of weather and calendar events', *Journal of Transport Geography*. Elsevier Ltd, 41, pp. 292–305. doi: 10.1016/j.jtrangeo.2014.09.003.

DeMaio, P. (2009) 'Bike-sharing: History, Impacts, Models of Provision, and Future', *Journal of Public Transportation*, 12(4), pp. 41–56. doi: 10.5038/2375-0901.12.4.3.

Kassambara, A. (2017) *Practical guide to cluster analysis in R: unsupervised machine learning*. Edition 1. STHDA.

Kaufman, L. and Rousseeuw, P. J. (1990) *Finding groups in data - An introduction to cluster analysis*. Hoboken, New Jersey: Wiley-Interscience. doi: 10.15713/ins.mmj.3.

Biographies

Andrea Sibilia has an academic background in environmental sciences. He covered both the roles of Environmental Consultant and GIS Technician at G.R.A.I.A. srl an Italian engineering company. Currently, he is an Assistant Geospatial Specialist of the Smart-Consulting team of WSP and his main research interests reside in GIS and analytics.

James Haworth is a lecturer in spatio-temporal analytics at SpaceTimeLab in the Department of Civil, Environmental and Geomatic Engineering, UCL. James' main interests lie in the analysis, modelling and forecasting of spatio-temporal data using statistical and machine learning methods.