Big Data Analysis of Population Flow between TfL Oyster and Bicycle Hire Networks in London

N. Sari Aslam[a], J. Cheshire [b], T. Cheng [c]

[a b c] University College London(UCL), Department of Civil, Environmental and Geomatic Engineering, Gower St, London, UK

10 March 2015

Summary

This study seeks to undertake an initial analysis of the likely flow of people between the Tube to bicycle hire network in London. Data for the two networks were extracted for a month (April and June 2012) in order to establish the strength of the relationship between them. The results quantify the extent to which Tube commuters impact the capacity utilization of the bicycle network. We expect this research to have implications in the expansion and maintenance of bicycle hire in London and similar schemes around the world.

KEYWORDS: Big Data, Oyster Data, Bicycle share system, Time Series Analysis, Regression Analysis

1. Introduction

In London, 24 million journeys are completed on Transport of London's (TfL's) network each day (Transport for London, 2012). The vast majority of these are by bus and Tube but a growing number of travellers make use of London's bicycle hire (O'Brien et al. 2014).The nature of a bicycle hire network is completely unique compared to other transport networks, because the service providers have little or no control over the key resource i.e. bicycle. The challenge is to optimise this network resource utilization based on usage behaviour of the bicycle users.

The rapid pace of technological advances and the availability of huge amounts of data from transport networks have made it possible to analyse population flows in great detail. Billions of rows of continuous and non-invasive data with spatial and temporal dimensions is now available in the public domain (Beecham & Wood, 2013; Blythe & Bryan, 2007; Kusakabe, Iryo, & Asakura, 2010; Lathia, Ahmed, & Capra, 2012; Páez, Trépanier, & Morency, 2011). Given the huge volumes of data now available, it has become challenging to undertake analysis using conventional statistical software (Blackwell & Sen, 2012). It is therefore often necessary to either subset the data or perform some kind of aggregation to reduce data size.

[a]n.aslam.11@ucl.ac.uk,
[b]james.cheshire@ucl.ac.uk
[c]tao.cheng@ucl.ac.uk

2. Data Description

The users of London transport network whose usage behaviour is repetitive and can be modelled are the main focus of this study. Integrating and analysing the data from train and bicycle hire networks provide an opportunity to understand the usage behaviour and allow efficient allocation of the resources.

Tube to bicycle: Because of the relative size of the networks, a large influx of Tube users can impact significantly on the capacity utilization of a bicycle hire network. To show the strength of this relationship, the analysis considered exit (Tube stations) to exit (docking stations) data.

Bicycle to train: To gauge the strength of the relationship in the reverse direction, i.e. from bicycle on to train to establish if users coming into the station on bicycle are continuing with their commute via trains. The analysis will consider entry (docking stations) to entry (train stations) data.

The cycle hire data is for the individual journeys from one docking station (origin) to another (destination). In order to focus on the specific time windows it was decided that journeys should be aggregated into 15 minute time intervals. It resulted in two records per docking station per 15 minute period. One record for aggregate 'entry' terminating at the station and the second for aggregate 'exit' from a docking station. For the purpose of this analysis the multiple docking station data have been aggregated based on proximity to the station.

Oyster data available for this analysis was aggregated at 15mins intervals and were provided by TfL. In order to match the bicycle data, all the journeys terminating at a given station within a period were aggregated into one 'entry' record. All the journeys starting from a station within a period were aggregated in one 'exit' record.

The processed data could be classified as below, and were used for the following analysis:

- Aggregate Tube exists
- Aggregate Tube entries
- Aggregate bicycle docking station exits
- Aggregate bicycle docking station entries
- Capacity of the bicycle docking station (the number of bicycles available for use/exit and return/entry)

3. Methodology

Central London provides the focus for this study. It sees the largest peaks in commuter flows as well as being the primary destination for tourism and leisure.

The study started with the trend analysis of the time series of the two networks, to understand obvious patterns in the data. To look further into this relationship, the Pearson correlation coefficient was calculated for time series data for the two networks, this quantified the strength of the relationship between the two networks. This was followed by linear regression to show the trend lines using docking station data as the dependent variable. The study started with a daily, followed by weekly and monthly analysis of the three selected Tube stations and their corresponding docking stations.

3.1. Network Analysis

The population flow between the Tube and Bicycle hire network depends upon the proximity of the bicycle docking station to the Tube station. 'Closest facility' network analysis was conducted to find the docking stations in close adjacency to the Tube stations for the available data (747 Docking stations and 163 Tube Stations) and to identify the shortest path between them. A maximum of five docking stations were selected within 300m (walking distance) of the selected Tube stations.

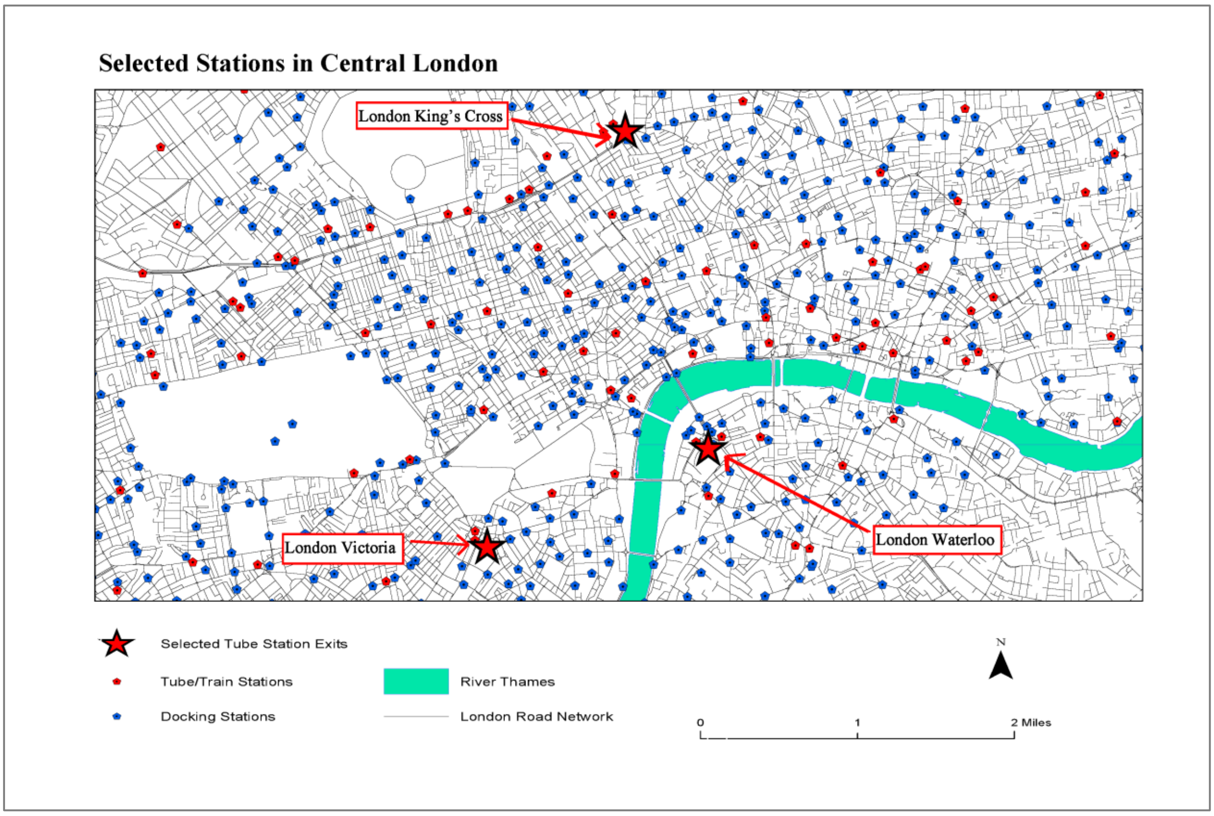

Figure 1: Study Area of the selected stations of London Waterloo, Victoria and King's cross.

In order to undertake more detailed analysis, three Tube stations - London Waterloo, London King's Cross and London Victoria - were selected (shown as stars in Figure 1) and their details described below.

Table 1: The list of the Tube and docking stations

Name	Station Exits	Bicycle Hire Docking Stations
Waterloo	Shell Gates Main Gates Auxiliary Gates W&C Validators Jubilee Gates	Waterloo Station 1, Waterloo Waterloo Station 2, Waterloo Waterloo Station 3, Waterloo
Victoria	District Gates Main Gates	Ashley Place, Victoria Cardinal Place, Victoria
Kings Cross	Met Main Entry Gates Tube Gates Thameslink Gates Northern Ticket Hall	St. Chad's Street, King's Cross Belgrove Street , King's Cross Northdown Street, King's Cross

3.2. Daily Data (7 June, 2012)

To understand the daily pattern of people flow between the two networks, data points are plotted for the three selected stations in Central London. From Tube station to docking station AM exit data (6am-10am) has strong relationship for all three stations. From Docking stations to Tube station PM entry data (5pm-9pm) has strong relationship, but not as significant as in AM peak. Capacity utilization graphs also support the same results.

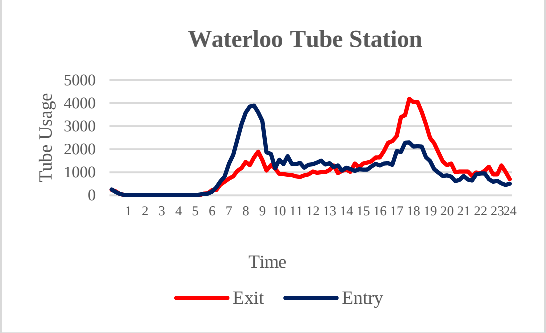

Figure 2 Waterloo Stations one day data (7 June 2012)

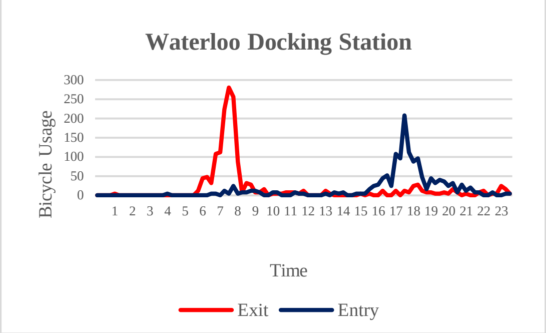

Figure 3 Waterloo Docking Stations, 24 hours data (7 June 2012)

Capacity utilization of the bicycle hire network, as shown in Figures 4 and 5, follows the trend of the Tube network rush hours, but also highlights how the lack of capacity impact the relationship (adversely).

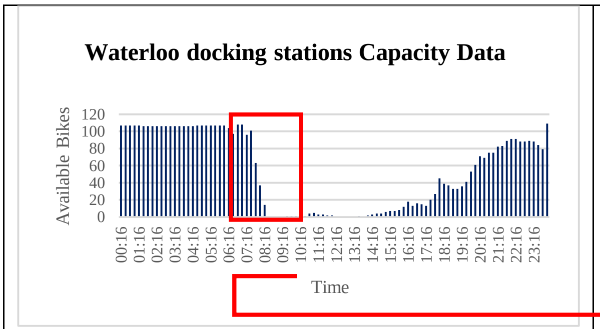

Figure 4 Capacity Usage of Waterloo docking stations with 24 hours data (7 June 2012).

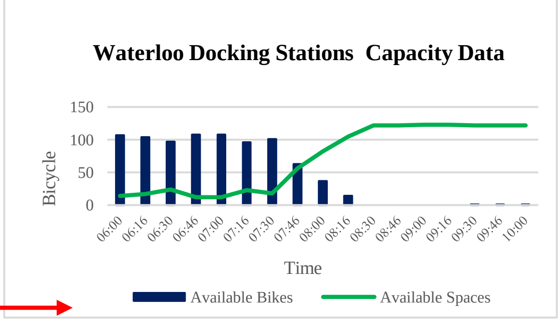

Figure 5 Capacity Usage of Waterloo docking stations with four hours data (7 June 2012, 6:00am– 10:00am data).

3.3. Weekly Data

Results plotted for the same three stations from the 18 June to the 24 June show the separation between the weekend and weekday trends at 15 minute intervals. Figure 6 shows two peaks for Tube in both AM and PM and PM peak is higher than the AM. Figure 7 highlight the AM peak data for bicycle users. The PM peak is less obvious for the bicycle population flow as it may be due to national rail commuters using Tube for the first leg of their journey. The relationship is AM peak between two networks.

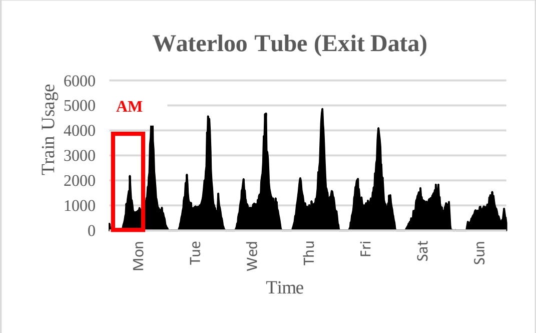

Figure 6 Weekday and Weekend data for Waterloo station. Highlighted peaks indicate possible relationship.

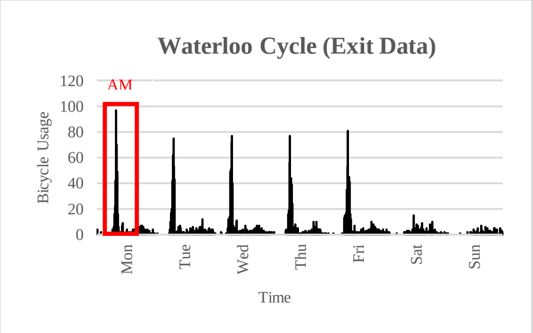

Figure 7 Week day and Weekend data for the docking stations adjacent to Waterloo station. Highlighted peaks indicate possible relationship

3.4. Monthly Data (April, 2012)

The analysis for the month was carried out to highlight the relationship between docking station and train station over a longer duration. The results for the four weeks period further emphasised the weekday and weekend trends visible in weekly data.

After investigating the daily, weekly and month trends, further insight can be gained through calculating the Pearson correlation coefficient (defined in Equation 1).

$$\text{Correlation}(r) = \frac{Cov(x,y)}{\text{std.dev (x) ´ std. dev (y)}} \qquad \text{(Equation 1)}$$

where x is the number of population using train stations at an hour interval (train usage), and y is the number of population using bicycle docking stations at an hour interval (bicycle usage).

Table 2: Correlation coefficients using month AM peak data all days (6:00am - 10:00 am) excluding weekends in Waterloo Tube/Train Stations and Docking Stations

Morning Peak/Morning Activities				
Station to Docking S.	**0600-0700**	**0700-0800**	**0800-0900**	**0900-1000**
Waterloo (Exit)	0.88534	0.90249	0.90056	0.39092
Evening Peak/Evening Activities				
Station to Docking S.	**1700-1800**	**1800-1900**	**1900-2000**	**2000-2100**
Waterloo (Exit)	0.23922	0.42573	0.27059	0.41861

Description	Range
Very Weak	0.01 – 0.19
Weak	0.20 – 0.39
Modest	0.40 – 0.69
Strong	0.70 – 0.89
Very Strong	0.90 – 0.99

Linear regression was conducted by assuming docking stations time series data as the dependent variable and train station data as the independent variable. The chart (Figure 8) shows two-hour intervals over a period of months excluding weekends. It shows a good fit with a coefficient of determination of 72% explaining all the variability of the data around its mean.

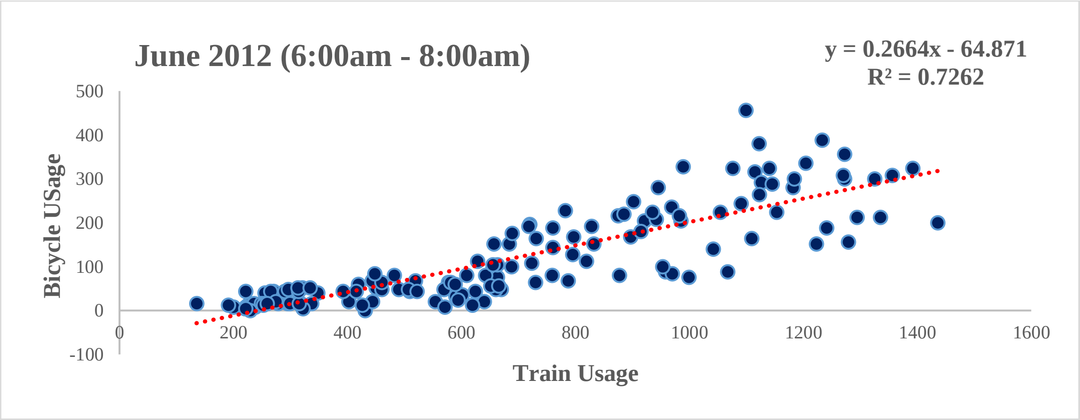

Figure 8 Linear regression applied to Waterloo AM Exit (0600-0800)

4. Conclusions

This paper was set out to study the relationship between the bicycle hire network and the TfL Oyster network. The results contain few surprises with the strength of the relationship closely linked to the rush hour commuting patterns.

There is a dip in the correlation after the rush hours, initial perception was that it is due to the drop in the number of commuters from the Tube, but it has been observed that the lack of available bicycles also adversely impact the relationship as shown in figure 4. Although the current report only focuses on three major Tube stations, the methods are easily expanded to cover the entire London Underground network.

Cycling as a government policy always has a positive impact on the environment, health and economy. Journeys made from bicycles instead of other modes of transport makes cities less congested, cut transport and delivery costs, reduce illness-related expenditure and make people fitter and the environment cleaner ("Position Paper of the European Cyclists ' Federation," n.d.). The analysis in this paper provides an insight into the user behaviour of the bicycle hire network and it will allow future infrastructure investment decision to be made in an informed manner.

5. Acknowledgements

I am grateful to the Economic and Social Research Council for funding my studentship at UCL. Furthermore, I owe a special thanks to Prof Paul Longley for his support during the paper preparation.

6. Biography

Nilufer Sari Aslam completed MSc Geographic Information Science at UCL in 2013. Currently she is in MRes at Urban Sustainability and Resilience at UCL to contribute how big data can be analysed using statistical approaches. Nilufer's research interests are big data analysis, spatial temporal analysis, network analysis and demand modelling.

7. References

Beecham, R., & Wood, J. (2013). Exploring gendered cycling behaviours within a large-scale behavioural data-set. *Transportation Planning and Technology*, *37*(1), 83–97. doi:10.1080/03081060.2013.844903

Blackwell, M., & Sen, M. (2012). Large Datasets and You: A Field Guide. *Manuskript*, *14627*, 1–8. Retrieved from http://www.mattblackwell.org/files/papers/bigdata.pdf\npapers3://publication/uuid/9B8CCBB4-E848-4D7E-9F6E-14AE5A96FC0C

Blythe, P., & Bryan, H. (2007). Understanding behaviour through smartcard data analysis. *Proceedings of the ICE - Transport*, *160*(4), 173–177. doi:10.1680/tran.2007.160.4.173

Kusakabe, T., Iryo, T., & Asakura, Y. (2010). Estimation method for railway passengers' train choice behavior with smart card transaction data. *Transportation*, *37*(5), 731–749. doi:10.1007/s11116-010-9290-0

Lathia, N., Ahmed, S., & Capra, L. (2012). Measuring the impact of opening the London shared bicycle scheme to casual users. *Transportation Research Part C: Emerging Technologies*, *22*, 88–102. doi:10.1016/j.trc.2011.12.004

O'Brien, O., Cheshire, J., & Batty, M. (2014). Mining bicycle sharing data for generating insights into sustainable transport systems. *Journal of Transport Geography*, *34*, 262–273. doi:10.1016/j.jtrangeo.2013.06.007

Páez, A., Trépanier, M., & Morency, C. (2011). Geodemographic analysis and the identification of potential business partnerships enabled by transit smart cards. *Transportation Research Part A: Policy and Practice*, *45*(7), 640–652. doi:10.1016/j.tra.2011.04.002

Position Paper of the European Cyclists ' Federation. (n.d.).

Transport for London Transport for London. (2012), 1–3.