A continuation on the data analysis for the Dublin Bike rental scheme
Joe Timoney*1, Carlos Siqueira Do Amaral †1, Thanh Thoa Pham Thi ‡2, and
Adam Winstanley §1

1 Department of Computer Science, Maynooth University.
2 School of Management, Dublin Institute of Technology

January 12, 2018

Summary
Analysis of Dublin bike rental usage patterns, presented at the last conference in this series, has been
extended. The purpose is to help in management of the scheme including better rebalancing of stock
and other logistical improvements. A systematic data analysis framework is described. Results of
analysis from a full year of data are presented with a focus on identifying usage characteristics of
each station or clusters of stations. This provides insights into the flow of bikes between stations. A
discussion on the possibilities for building prediction algorithms is developed with a proposition for
the most suitable approach.

KEYWORDS: public bike data analysis, spatial clustering and data visualisation, bike scheme
prediction.

1. Introduction
An analysis of data collected over a 3-month autumnal period from the JCDecaux API (JCDecaux,
2018) for Dublin bikes (Dublin bikes, 2018) was presented previously (Pham Thi et al, 2017). The
work included clustering analysis applied to the busiest and quietest bike stations that illustrated
significant time-dependent differences in usage. However, more data was required to extend the
analysis and make firmer conclusions across the whole network. Since then a year’s worth of data has
been gathered over the period from 14th October 2016 until 14th October 2017. The data was filtered
to handle missing and corrupted values. Due to its volume, a systematic approach for storing data and
analysis was needed. This would allow repeated analysis as new data became available and
generalisation of the method for similar cases. The framework developed incorporates the analysis
process and new software tools. Results are viewed via ‘Heatmaps’ that show the activity around
different stations and clusters are derived that summarise the patterns associated with four basic
station groupings. The geographical relationship between the clusters can be shown cartographically.
We then describe the framework and its output as well as discussing how it helps for inferring
patterns of flow across the network. This is followed by a consideration of the choices for developing
a predictor algorithm. Finally, conclusions and ideas for future work are given. The overall
motivation for this work is that it can result in a deeper understanding of the scheme. This is intended
to assist in making improvements for both users and the operator by ensuring that bike availability at
stations will always be in line with demand, and that operator can plan load balancing strategies well
in advance.

* Joseph.timoney @mu.ie
† carlos.siqueiradoamaral.2015@mumail.ie
‡ Thoa.pham@dit.ie
§ Adam.winstanley@mu.ie
2. Analysis Framework
Dublin Bike data was collected from JCDecaux (JCDecaux, 2018) for a period of one year. Each real time data download is a JSON file which contains 100 records corresponding to 100 stations. Each record consists of information such as station name, station location, total stands at that station, the number of stands available, and the number of available bikes at the most recent timestamp. The data is stored in a Mongo database (MongoDB, 2017) for systematic calculations, the results of which are output to the analysis programs written in R. Figure 1 depicts the analysis framework.

![Image](https://example.com/figure1.png)

**Figure 1 Analysis Framework**

From the last update timestamp the information of date, time and day of week is derived. The number of check-ins (drop bike) and check outs (take bike) are calculated to identify the degree of activity at each station. Table 1 shows a sample of data prepared for analysis.

<table>
<thead>
<tr>
<th>(1)</th>
<th>(2)</th>
<th>(3)</th>
<th>(4)</th>
<th>(5)</th>
<th>(6)</th>
<th>(7)</th>
<th>(8)</th>
<th>(9)</th>
<th>(10)</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>CHARLEMONT PLACE</td>
<td>12/09/2016</td>
<td>15:30:00</td>
<td>Mon</td>
<td>40</td>
<td>2</td>
<td>2</td>
<td>0</td>
<td>34</td>
</tr>
<tr>
<td>5</td>
<td>CHARLEMONT PLACE</td>
<td>12/09/2016</td>
<td>15:40:00</td>
<td>Mon</td>
<td>40</td>
<td>-9</td>
<td>0</td>
<td>9</td>
<td>25</td>
</tr>
<tr>
<td>5</td>
<td>CHARLEMONT PLACE</td>
<td>12/09/2016</td>
<td>15:50:00</td>
<td>Mon</td>
<td>40</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>25</td>
</tr>
<tr>
<td>5</td>
<td>CHARLEMONT PLACE</td>
<td>12/09/2016</td>
<td>16:00:00</td>
<td>Mon</td>
<td>40</td>
<td>11</td>
<td>11</td>
<td>0</td>
<td>36</td>
</tr>
</tbody>
</table>

Columns names: (1) – Station ID, (2) – Station Name, (3) – Date, (4) – Time, (5) – Day of week, (6) – Bike stands, (7) – Previous period difference, (8) – Check-in number, (9) – Check-out number, (10) – Available stands.

Some R functions are available as an open-source special library that can be used for other applications: [https://github.com/amaralcs/dublin_bikes](https://github.com/amaralcs/dublin_bikes).

3. Results
Heatmaps were generated by taking all the data and grouping it by the weekday. The average is plotted using a scheme where the colour closer to yellow means more bikes are available and the deep purple colour means the opposite. Figure 2 shows examples from contrasting stations. Charlemont Street station has very low bike availability from 7:00am to 5:00pm Monday to Friday, while for Herbert Street is opposite and has a high availability in that time period. Herbert Street is close to a very busy business area, while Charlemond is situated in a residential area, leading to different activity patterns.
There are 100 bike stations in total and visual inspection shows that many similar patterns appear among their heatmaps. K-means analysis identified four distinct clusters. Figure 3 shows the mean plot of each cluster.

As shown in Figure 3, Clusters 2 and 3 have almost opposite behaviour while Clusters 1 and 4 also show contrasting behaviour. The cluster data is being analysed further to identify pattern differences between weekdays and weekends.

Figure 4 shows where the stations in each of the four clusters lie. Cluster 4 is mostly in the east of the city in the business district, having the quietest stations. Clusters 2 and 3 are in the main city centre shopping and business area while Cluster 1 is concentrated on the south side of the river.
4. Discussion
Based on these results, it can be seen that there are flows of bikes moving from certain stations to other stations at the same time. For instance during weekdays (Mon-Thurs), the patterns of cluster 1 and cluster 3 are completely opposite (Figure 4 - left), whereas cluster 4 is almost unchanged over the day. Cluster 1 consists of stations mostly from the South along the river. Cluster 3 consists of stations mostly from the North and city centre. This shows that people pick up bikes from Northern stations and drop them at the Southern stations during business hours during week days, and they do the opposite out of business hours. Meanwhile, cluster 2 gets almost the same number of available stands all day, which means they are relatively quiet stations. There is a need for rebalancing these quiet stations with those that are busier close by.

The analysis can be used to develop a predictive model that can help in planning and management, such as notifying users in advance where to pick up or drop off bikes, for load balancing across the network, or managing overall service commitments. Many prediction models use statistical regression techniques (for example Singhvi D et al., 2015) and also account for external variables such as the weather (Mahmoud M et al, 2015). Other approaches (Gast N et al, 2015; Schuijbroek J et al, 2013) employ probabilistic methods derived from the M/M/1/k Queueing Model. These can be implemented using closed-form expressions (Tarabia A, 2002; Al Hanbali A and Boxma O, 2009), but more recent paper adopts a Runge-Katta technique (Schuijbroek J et al, 2017). Another approach gaining popularity is connected with classification trees in the form of Random Forests, which has been shown to bring improved results (Wang W, 2016; Yang Z et al, 2016). Random Forest is a meta-algorithm that combines a large number of decision-tree models, each individually built on bootstrapped samples of the data. This process of sampling the data and combining the individual decision-trees is called bagging, and reduces prediction variance without increasing the bias. The final prediction is formed as the mean of the individual predictions.

A drawback regarding the use of linear regression-based techniques is that highly non-linear interactions between the variables cannot be modelled (Feng Y at al, 2017). Additionally, the point estimators produced are not always practically useful: a user is not interested in the actual number of bikes available at some point in the future, rather, they only want to know the probability they could pick one up (Gast N et al, 2015). However, the queueing theory approach also has difficulties. It was observed that it can be challenging to compute the probability of certain stations (Li Q and Fan R, 2016) and deviations from the theoretical model do occur, i.e. non-Poisson arrival processes and non-exponential riding-bike times. One solution (Feng C et al, 2016) is to make a more elaborate model by representing the system as a Population Continuous Time Markov Chain (PCTMC) with time-dependent rates, but more work is required to properly capture neighbouring station interactions. The advantages of the Random Forest technique is that there are very few assumptions attached to it, it is considered to be robust against overfitting, it can handle highly nonlinear variables and categorical interactions, and that it ranks each variable’s individual contributions in the model (Yang Z et al, 2016). These make it very attractive and it seems the best choice for Dublin bikes. Regression should not be discounted entirely as a recent work (Ashqar H et al, 2017) used Random Forest for univariate modelling to predict the number of available bikes at each station, it also looked at the prediction of the number of available bikes in the larger network using a Partial Least-Squares Regression (PLSR) to account for spatial correlation between stations which produced reasonably good results.

5. Conclusion and future work
This is an on-going project from which we have derived intermediate results. It is an extension of work presented previously (Pham Thi et al, 2017) which has been expanded as follows:

- Busy and quiet stations can be identified through the whole-network analysis not just by considering individual stations
- Using heat maps provides a better visualisation of the behaviour of each station
- Clustering allows us to identify stations with similar temporal characteristics, allowing
identification of phenomena such as the flow of bikes over the network.

Immediate future work Future work will use these models to predict bike and stand availability by time and location. Weather-dependent and season-dependent behaviour are also being investigated and modelled.

Biography
Joe Timoney is a lecturer in the Computer Science Department, Maynooth University. Adam Winstanley is Professor of the same Department. Thanh Thoa Pham Thi is a lecturer at the school of Management in Dublin Institute of Technology. Carlos Siqueira Do Amaral is a final year student in Computer Science at Maynooth University. All are interested in developing spatial analysis tools for public bike schemes.

References


