London Santander cycles: A Visual Analysis Approach

Pablo Quinoa

Abstract— This paper aims to understand mobility patterns of cyclers in London. For this purpose, Santander cycling shared system data for a week in 2017 was used. Visual analytical techniques were used to perform the analysis, thus combining some computing with our own interpretation of the generated visualizations, to finally model London cycling mobility as accurately as possible.

1 Problem Statement

This study aims to analyse cyclist's mobility patterns of a big city like London using geo visualization tools. We will try to find common patterns in the way Londoner riders move around the city's central area, and for this we will be using data from Santander bike hires during an entire week in April 2017.

For the purpose of the analysis we will try to respond to a few specific questions regarding cyclist's mobility around the city, so we will divide the analysis in some sections or questions to help its understandability. Thus, we will try to respond to the following:

- Can we see any important differences in terms of cycle hires during week working days vs weekends? How are those time periods different?
- What areas of London have higher cycles hire volume? Is there any logic behind the findings?
- Where in London (which journey start station) people hire bicycles for longer durations? Can we see any pattern on the findings?
- Can we find some patterns across different journeys? E.g. do cyclists in east London travel to central for work? Are people starting a cycle journey in the same area all travelling in a similar direction?

The dataset used for the analysis covers a normal week of the year, with 220k journeys recorded between 785 different cycle hire points. One week's data was found to be enough, as we think that the behavior of the cyclers will repeat week after week during a year without substantial differences.

2 State of the Art

To help with the analysis proposed in this paper, we first read some papers covering geo special data using visual analytics. We found two papers specially interesting, as they provide some useful insights into the domain and this type of data.

First, *Gennady Andrienko et al. [1]* wrote an article where they summarized the outcomes of the discussions that took place during a workshop where they discussed the state of the art in visually-enabled spatial decision support, and identified major problems or challenges. From this article we found specially interesting how they discussed the complex nature of geographic space for its analysis. They highlighted how the metric properties of physical spaces are very different from abstract mathematical spaces, so that distances in geographical physical spaces are not the same as Euclidean distances on an abstract space like a plane. This will be important to consider in our analysis, as our data is inside a geographic area (London) where the distance between two docking stations is not just its Euclidean distance or a straight line, but realistically a longer path as cyclists need to follow different roads and avoid objects on their way.

We also found interesting how the paper by *Gennady Andrienko et al. [1]* covers the scalability topic of geospatial data. They mention it is important that geo visual analytics methods and tools are scalable with the amount of data, dimensionality, resolution… To cope with this problem, they proposed the use of a strong links between the geo visualization and data mining. For our analysis, due to the high volume of data being analysed we will generate geo visualizations as well as some light data mining and combine both to find some patterns.

The other paper we found very interesting was that one by *Oliver O'Brien et al. [2]*. This paper studies mobility of bike sharing system riders across 38 cities, and although it is more about comparing the bike systems in between them rather than focusing on one, it uses some interesting approaches to analyse the data. They performed nearest neighbours analysis to find which cities have more concentration of stations. They used Euclidean distances which like mentioned previously is not representative of a real network distance in a geographic area. But because they wanted to compare average distance between stations across different cities the distance error was found to be ok, as they only want to capture proportion. For our analysis proportion is not a valuable metric, we would be calculating

average distance between stations for just London, thus the error in distance would provide invalid information that we don't want. But from this paper, what we found very useful for our analysis is how they divide the time scale into week and weekends and try to find the peak times for both. For London, they found a pattern with two weekday commuter peaks and a broad afternoon peak at weekends.

3 Properties of the Data

For this paper's analysis we used some open source data from Transport for London, which was found under https://cycling.data.tfl.gov.uk/. Although there is data for each week from 2015 all the way to 2018, we decided to use data from just a week outside of holiday season to model better a normal week. We used all Santander cycle journeys registered from the 29th March to the 4th of April 2017. This was in the form of a single csv file which contains 219k journeys. Each row or journey provides an index (rental id), journey duration in seconds, start and end of journey timestamp, and start and end station id and name for the journey.

In order to place the journey's dataset docking stations in a map, we used an additional spreadsheet (csv) with latitude and longitude information for each station which was downloaded from tfl.gof.uk as well. The spreadsheet lists 785 docking stations across London during 2017. Each row contains the name and id of a docking station, its latitude and longitude, and a count for the total bike hires for that station during that year. The station id was the join index used to map both spreadsheets; thus, journeys would now have geographic coordinates.

The first thing we did was to investigate the quality of the data, so we checked for null values. The stations spreadsheet was found to be complete; no missing values were found. But for the journeys data we were missing some values, more specifically 2354 end stations and 2332 end date/duration were missing (see Figure 1 below).

```
firstWeekApril2017.isnull().sum()
```

```
Rental Id            0
Duration          2332
Bike Id              0
End Date          2332
EndStation Id     2354
EndStation Name   2354
Start Date           0
StartStation Id      0
StartStation Name    0
dtype: int64
```

```
stations.isnull().sum()
```

```
Site Name      0
Latitude       0
Longitude      0
Total Hires    0
Total Docks    0
StationID      0
dtype: int64
```

Figure 1: Null values for each column. Journeys on the left, stations on the right

These results were not unexpected, it sounds logical that some riders fail to check out the bike correctly when reaching their destination, thus some journeys did not register the end station or duration. So, because it was only 2354 out of 219k journeys, we decided to remove all those journeys or rows containing a missing end station or duration.

From the station's dataset, given that each docking station has a number of total hires for that year, we calculated the mean across all stations. We found that each docking station has 35 cycle hires per day on average (we normalized the year count into hires per day dividing by 365 days).

We also checked for any time slot during the week analysed which could be missing, and we found no time slot during that week for which there was no data. We had continuous data for the 24 hours of the day during the 7 days of the week for that week in March - April 2017.

4 Analysis

4.1 Approach

Our analysis will try to respond to the questions presented in section 1 (Problem Statement), and for that we will be using a combination of computing capabilities and our own interpretation on some generated visualizations. The following diagram summarizes very accurately how computing and humans interact in this analytical process:

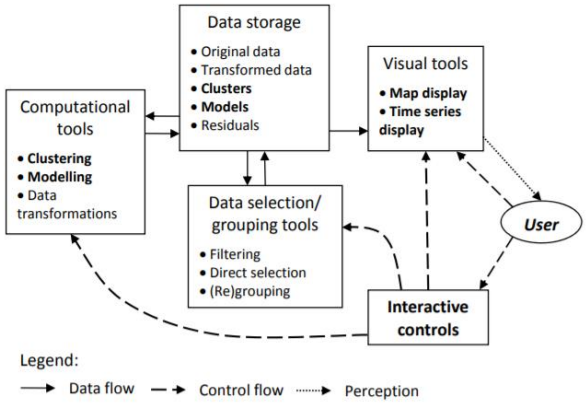

Figure 2: Diagram for geo spatial and temporal Visual Analysis workflow by Gennedy Adrienko et al. [3]

To solve each of our proposed analytical questions, we iterated through the above workflow one or more times, until we as humans were satisfied with the results gathered through one or more of these cycles. So, any analytical suggested question would start with some data transformations, for example each journey data sample start or end date was transformed from a

string into a valid timestamp to be able to divide the time space accordingly.

In our case, the three boxes on the left side from Figure 2 was a combined process in python where we would store the data in the local virtual memory, apply transformations and filter or group data as needed. The purpose of these data transformations, grouping and filtering was to feed the visualization tools with the data in a required format. We would then use mainly two python visualization libraries (Folium and Pyplot) to generate some 2-D visualizations that would represent the feed data either in a map or in a time series.

The visual tools would generate a visualization, and then us users (humans) would use our perception to interpret the plotted data. If we users would be satisfied with what we found from that initial visualization, we would close the cycle (Figure 2), otherwise we would go back to transform the data again or filter it in a different way and generate a new visualization until some findings were met.

4.2 Process

We divided the analysis in four sections or sub analysis to try to respond to the four main questions presented in section 1 (Problem Statement):

Weekdays vs Weekends:

We started our analysis comparing how are weekdays different from weekends in terms of cycle hire volumes. For this purpose, we needed to visualize the data in a time series space, so we started the workflow cycle presented in the previous section transforming the data given times.

Each row or journey in the journeys data frame has a start date and end date (up to minute precision), which is read as a string. We first transformed all those values under start or end date into valid date objects. To associate a time to each journey we decided to take only the start date, so we would be comparing all journeys by the time those started.

We then grouped all journeys departures into the two groups that we wanted to compare, weekdays and weekends. It is obvious that to be able to compare two time spaces of different sizes (5 days for weekdays vs 2 days for weekends) we needed to visualize the data normalized in just a day period. We would now have the average departure times per day for both weekdays and weekends, in other words, both visualizations with equally scaled x axis (equal time space). On the y axis we would plot average number of cycle departures per time slice. We divided the 24 hours of a day into 72 time slices of 20 min each, this would allow us to observe more in detail patterns in the peak hours.

We then generated two bar chart visualizations from the pyplot python library, one for the average departure time volumes during the week and one for the weekends (see figure 3 below).

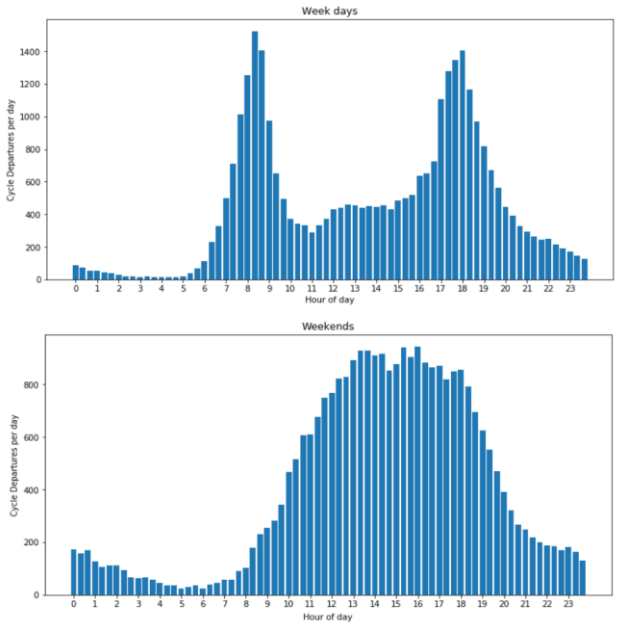

Figure 3: Cyclist volumes Weekdays on the top vs Weekends on the bottom

The results coincide with those of *Oliver O'Brien et al. [2]*, where they found a pattern with two weekday commuter peaks and a broad afternoon peak at weekends. But if we look more in detail into the range of the y axis which represent the number of departures, we can see that the weekdays peaks are twice the size of the weekends one (1400 vs 800 departures aprox in 20 min time slices). This suggests that the bike share system analysed is used for commuting to work more than anything, and not so much for recreational purposes.

The morning peak hours for the weekdays was found to be in between 7:30 and 9 am, and 17:30pm to 18:30 pm for the evening peak. This suggests a very strong demographic group of commuters that use this service to go to work or go back home after work.

Cycle hires across London Areas:

In this section of our analysis we wanted to find what areas of London have higher bike hire volumes than other areas; and try to find a pattern for that distribution if there is any.

We decided to plot the data in a geo spatial plane, so positioning each docking station in a London map. For this

purpose, not much data transformations or grouping was required; the stations data set already contains total number of hires (per year) per station as well as a latitude and longitude for each station. In order to be able to categorize hire volumes in a two-dimensional map, we decided to use three different colours depending on each station's number of cycle hires. We then set some thresholds to decide the colour of the station point to plot, green for the stations with less hires, orange for medium hire volume, and red for very demanded stations. This would allow the interpretation of the visualization, so that we can find patters from just an overall picture of London.

We then generated a visualization from the python library Folium which clusters stations by location so that initially from a zoom out perspective you can see which areas have more concentration of stations, but then once you zoom in those clusters, they split into original data points or stations where we can see the colour as defined previously, representing the hire volume.

Figure 4: Stations clustered zoom in and out

In the above visualization we can see in the top zoomed out London map that the centre area above the Thames river contains the greatest number of docking stations (two red points, each one clustering 127-128 docking stations). If we then zoom into an area, we can start to see how the clusters split and we can start seeing each station (circles without numbers) with its appropriate colour for the number of hires that the station receives.

This visualization helped us understand where in London there are more docking stations, and although you can zoom in to specific areas to see hire volume per station, is not good enough to compare station's hire volumes across the total London scope. For that, we went one step back in the workflow diagram proposed to generate a new visualization. We now generated the same folium map with the same thresholds to decide on different colours, but this time without applying clustering.

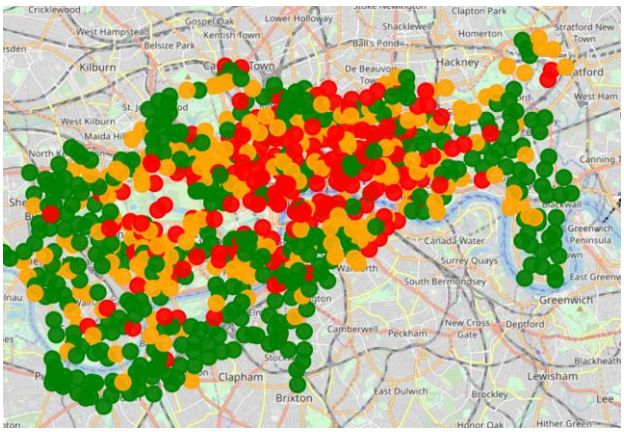

Figure 5: Stations coloured by total number of hires

This time we can see how central London, from Hyde Park to London city above the Thames, as well as Waterloo area under the Thames have the docking stations with the bigger number of bike hires. The surroundings of these more central areas receive a smaller number of hires, with some outliers like Stratford in the north east of the map.

London areas where longer journeys start from:

In this section of our analysis we wanted to find what areas of London people start journeys with longer duration. In this case we decided to only use data from the morning peak time during weekdays, this would allow us to draw some conclusions around commuters demographic group.

For this purpose, we used all peak morning weekdays journeys data samples from the journeys dataset, where each journey has a feature or attribute duration in seconds. We started by converting these values from seconds into minutes, as we are better able to understand time results in minutes. In addition to converting time measures, we also executed some statistics on the data and found that on average people ride their bike for 15 minutes and a half in the morning, presumably on their way to work.

This time we started visualizing the data with a Folium heatmap, where the heat areas would depend on the latitude and longitude of the station where the journey was started, as well as depend on journey duration.

Figure 6: Heatmap based on stations duration and location

From the above figure, we can observe how the warmer areas (in red) match with those places where there is a higher concentration of docking stations (blue dots). As the heat depends on location coordinates as well as the journey duration, it was obvious that we cannot draw any conclusions on journey durations, for its lack of independence with the location variables.

We decided to go one step back in the workflow diagram, and generate a new visualization where duration is visibly independent from other variables. This time we would use the same technique proposed in Figure 5, a folium map with coloured data points or stations where the colours are decided based on some thresholds for the duration of each journey attribute.

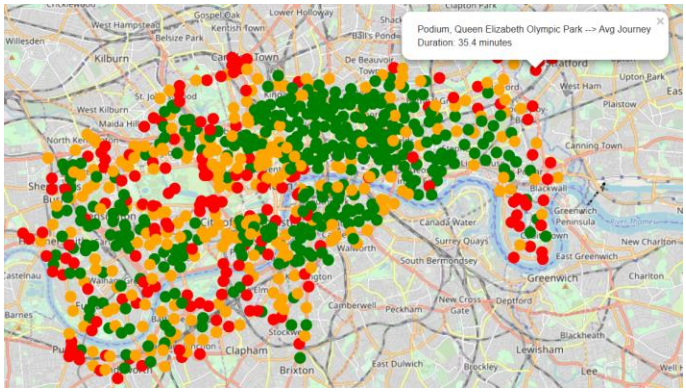

Figure 7: Start journey stations coloured by duration

We could observe how those people who start their journey from the most edge locations of this map ride for a longer time, and those starting in more centric locations travel for a shorter time.

Journey locations and directions:

In this section of our analysis we wanted to find some patterns across different journeys. We decided to plot a network with nodes (start and end stations) and edges (straight lines representing the journeys). But having over 200k journeys meant over 200k lines drawn on a limited space, so a visualization where we would not be able to see anything. Therefore, we decided to analyse just those journeys happening on weekdays during the morning peak. To differentiate those journeys (duplicated journeys that start and finish on the same station) that are more frequent than others, we grouped unique journeys and associate a journey counter or weight to it. We then plotted all journeys, where the edges or lines between start and end station would be thicker based on the journey popularity.

Figure 8: Nodes and edges representing journeys

We could see how in the mornings on weekdays many riders go from Kings Cross all the way central (south), or from Victoria or Waterloo Station all the way towards central (north).

In order to draw some conclusions in terms of the direction of these journeys we added arrows to the edges. No clear pattern was observed, see Figure 9 below.

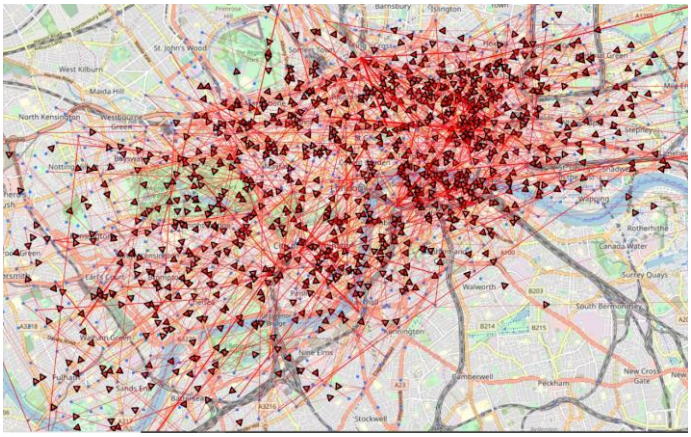

Figure 9: Nodes and edges with arrows representing journeys directions

4.3 Results

We concluded that London Santander riders match quite closely to the demographic group of commuters, with a morning and evening peak on weekdays that clearly match the office hours.

We also found out that those areas which overall looks like London fare zone 1, receive the greatest number of bike hires. Outside of this more central area some outliers with high volume of hires like Shepherds Bush or Stratford were found.

It was also found that people would start longer journeys outside of this central area, presumably to commute a longer distance all the way to central London. But another interesting finding in terms of journey durations, was that people hire bikes for longer periods of time during night-time (see Figure 10 below). This is probably because during the night after 12 there is limited public transport available.

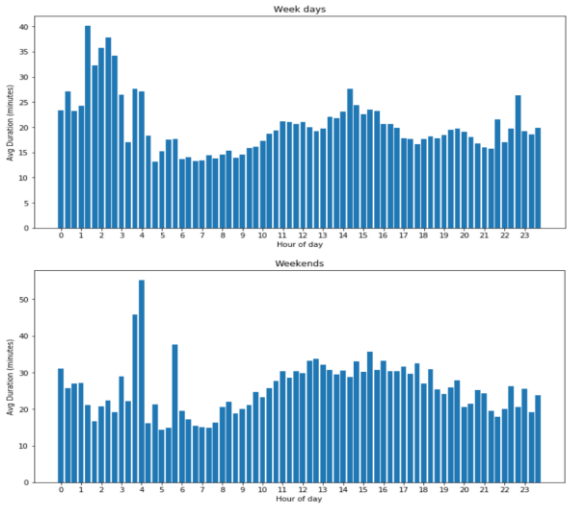

Figure 10: Journey durations Weekdays on the top vs Weekends on the bottom

Finally, from the network of nodes and edges plotted, we found that riders in the peak morning time on weekdays use the bike a lot to travel from main train stations like Victoria, Kings Cross or Waterloo to the more central area where most offices are.

5 Critical Reflection

This study has been efficient thanks to the cooperation of computing capabilities together with our own interpretation of the visualizations as humans, who also know the context of this data well. Knowing London well helped very much to draw conclusions. This can be generalized to the Visual Analytics field overall, where actors with domain knowledge is preferred for a richer analysis.

In this analysis process, we humans were able to understand the space better than machines. Geographic areas like London are not simple mathematical abstract spaces where mathematics and computing would do a very accurate job. But in fact, we understand what parks or rivers are, or how the centre of London is where most people work. Thus, performing a Visual Analytics approach where we use computers to execute transformations, filter the data, run statistics or generate visualizations, then letting humans interpret these visualizations, is the most efficient way of such analysis.

To improve the range of our analysis, we could use in the future some machine learning algorithm to support the analysis, thus potentiate the use of the computing tools during the workflow presented before.

But more specifically, when we analysed the journeys on a network with nodes and edges, we draw the edges like straight lines between origin and destination, thus accepting some error in the way these journeys really are in a geo spatial space. We only had data that would locate riders at start and end of their journey, making it impossible to represent the real travelled space. To compensate the error of these plotted assumptions, humans interpreting the visual data need to keep this in mind. For example, if a rider must cross the Thames, he might have to ride for a longer distance in order to use one of the available bridges; therefore when comparing edges on the geographic space this information needs to be kept in mind before we can draw some conclusions.

Overall, we can draw the inference that analytical problems involving geographic spaces are good candidates for this type of Visual Analytics framework where humans are the ones interpreting the space with their knowledge on it. A solution to better capture the real space, is having more data samples which means much more data. If we had the position of a rider collected every 30 seconds, we could much better represent the

travelled space. But this comes with the price of what it involves collecting so much more data.

Table of word counts

Problem statement	247
State of the art	484
Properties of the data	443
Analysis: Approach	303
Analysis: Process	1500
Analysis: Results	197
Critical reflection	408

REFERENCES

[1] Gennady Andrienko, Natalia Andrienko, Piotr Jankowski, Daniel Keim, Menno-Jan Kraak, Alan MacEachren, Stefan Wrobel. *Geovisual Analytics for Spatial Decision Support: Setting the Research Agenda*, 2017.

[2] Oliver O'Brien, James Cheshire, Michael Batty. *Mining bicycle sharing data for generating insights into sustainable transport systems,* 2014.

[3] Andrienko, N. and Andrienko, G. (2013). *A visual analytics framework for spatiotemporal analysis and modelling. Data Mining and Knowledge Discovery,* 27(1), pp. 55-83. doi: 10.1007/s10618-012-0285-7.