Analysis of geolocated vehicle flows

Ivan Vallejo Vall

September 26, 2017

1 Overview

This document contains the analysis of a real-world dataset. The dataset aggregates data transmitted from a number of cars while they are driving around the United Kingdom.

The dataset contains a row for each trip taken by a device user. Each row contains, *inter alia*, the fields enumerated in Table 1.

device key	duration (s)	end point city	end point country	end point state
end point zipcode	end point lat	end point lon	end point time	gps distance
start point city	start point country	start point state	start point zip	start point lat
start point lon	start point time	top speed	car nickname	

Table 1: Data fields available for the log record of each trip.

The underlying data cannot be shared because of privacy issues. However, the aggregate results of an analysis of the vehicle flows of these fleet are presented in this report.

The data analysis is implemented using R. The file 'code_vehicles.R' contains the code used to perform the analysis and generate the charts.

2 Data cleaning

Using the library 'sparklyr' I create a local Spark context, load the parquet files and output them into an R dataframe. The rest of the analysis is carried out in R because there is no need to use Spark given the size of this dataset.

The following checks are carried out (sequentially in the order stated) to ensure the integrity of the dataset:

1. There are 55 unique devices transmitting the data. There are no missing values in the device field and the format of the device key seems consistent.

2. There a large number of rows with the same device key, same latitude and longitude for the start and end points and same time stamps for the start and end of the trip. In the absence of an explanation on the possible meaning of these records and their integrity, I filter them out.

3. There are 57 observations with zero duration time and with no information on the end point. I delete them as they do not seem to correspond to actual car trips.

4. There are 95 observations with either zero start latitude or zero end latitude, which cannot be correct given that the observations are from the UK. Considering these as erroneous measurements, I delete them. Longitude measurements are within the boundaries of the UK so no cleaning is needed.

5. There are 118 observations with top speed equal to zero. It may be that the speed meter did not work and the car actually moved, although a majority of them start and end in the same area (i.e. same zip code). To be on the safe side, I remove them as there is a risk of measurement error.

6. There are 360 observations with the field "gps_distance" not available. Half of them start and end in the same latitude and longitude coordinates. The durations of these trips span from some seconds to several minutes. If the gps did not work for the distance, it might also have failed for the latitude/longitude measurement. Therefore, I remove them from the dataset.

7. I convert the character time field into a proper time stamp. The observations span from 10 August 2016 to 13 December 2016. I add a flag for weekend observations and for the day number (useful later on). All dates seem to be coherent so there are no corrections needed.

8. There are 316 observations with no information on the name of the start city and 334 with no information on the end city. I use the Google Maps API for R (Kahle and Wickham, 2013) to retrieve the actual city name based on the latitude and longitude coordinates and fill in the missing information on the name of the city.

9. There are 16 observations with no information on the starting zip code and 18 observations with a blank ending zip code. Google Maps does not provide zip codes for the majority of these locations, so I use instead a lookup table from Free Map Tools. For each missing zip code, I look for the closest point in the lookup table based on the longitudes/latitudes using the package 'rgeos' (Bivand et al., 2017). I attribute the each point the postcode of the closest entry in the lookup table. Given that the table includes 1.74 million registered postcodes in the UK, the approximate location is rather precise (I manually check some entries on a map to have a sanity check).

3 Analysis

Given that the dataset contains geolocation data for the vehicles, I perform an analysis of the data flows of these vehicles. The aim is to identify some patterns that may point to solutions leading to a more efficient and environmental friendly mobility of the people using these cars.

In particular, my objective is to see whether the mobility patterns suggest that some sort of transport sharing (be it based on public transportation or car sharing) could be proposed to these users.

3.1 Implementation of the predictive analytics

I start by separating the dataset into weekdays and weekend observations, given that it is to be expected that the traffic flows will be different for them.

Next I calculate the origin-destination matrix of the vehicles during the whole period. In order to present the flows graphically, I produce two chord diagrams using the package 'circlize' (Gu et al., 2014).

Since the full origin-destination matrix contains too many locations for it to be displayed meaningfully as a whole, I create two reduced versions of the origin-destination matrix: one for weekdays including 29 locations and one for weekends with 20 locations. The latter includes

fewer locations because with them it is possible to represent a significant percentage of the total flows, whereas on weekdays more locations are needed to achieve a similar coverage.

Figure 1 and Figure 2 show the results of the two chord diagrams. The analysis of the flows shows that there are no dominating routes, i.e. none has a much higher volume of vehicles than the rest.

Moreover, the largest traffic flow in most routes is the one beginning and ending in the same city (arrows bending back in Figure 1 and Figure 2). This is not only the case of big cities such as such as London and Edinburgh, but also of smaller towns such as Bracknell and Scunthorpe. This holds true for both weekdays and weekends.

This suggests that the traffic flows included in the dataset correspond in most cases to intra-local routes of single vehicles (i.e. there is no big concentration of in-sample cars in a given town). In this context, the possibilities of switching to a common transport (i.e. a bus covering a specific route or car sharing) are limited.

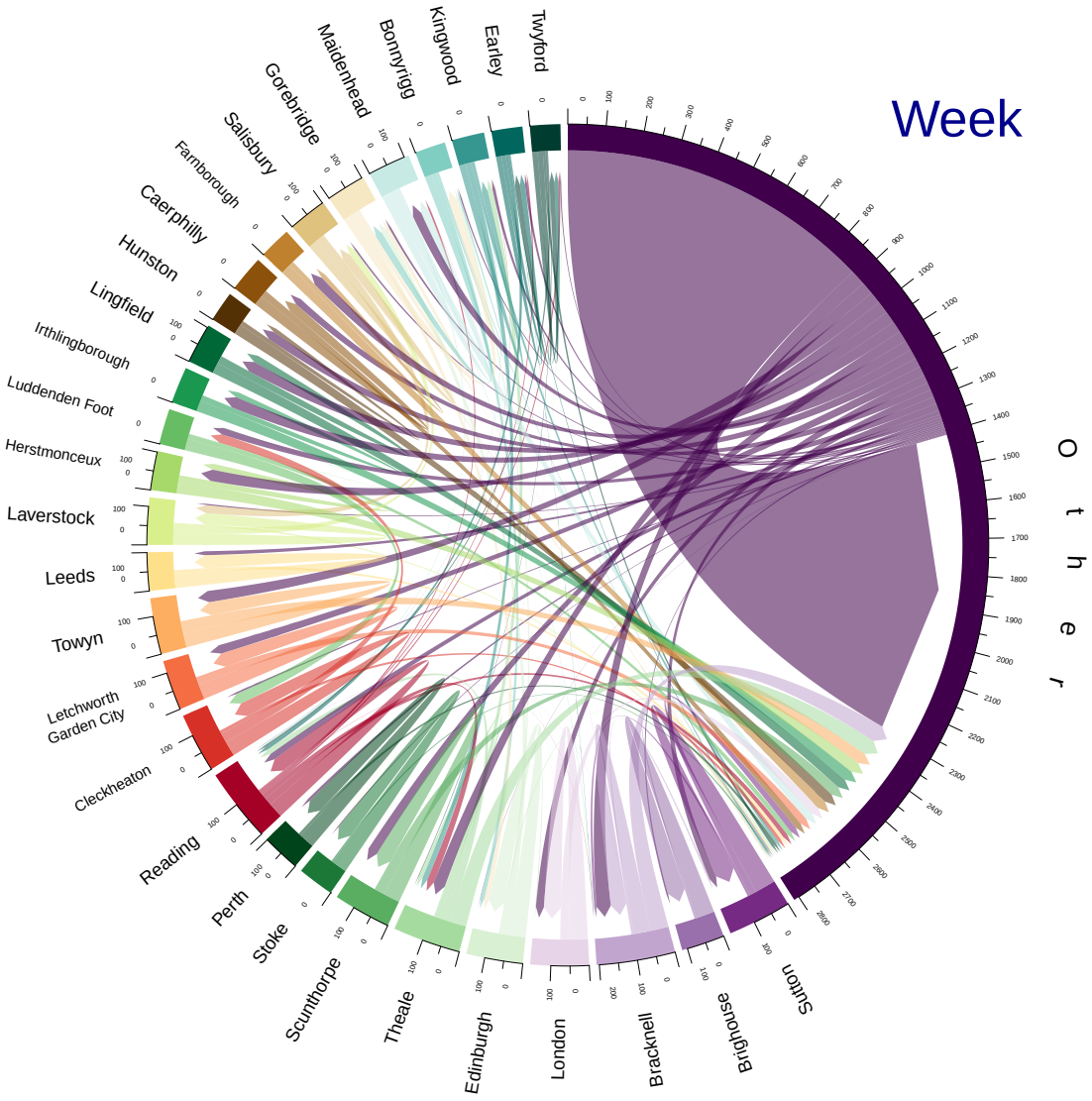

Figure 1: Chord diagram of traffic flows during weekdays in the period from 10 August to 13 December 2016.

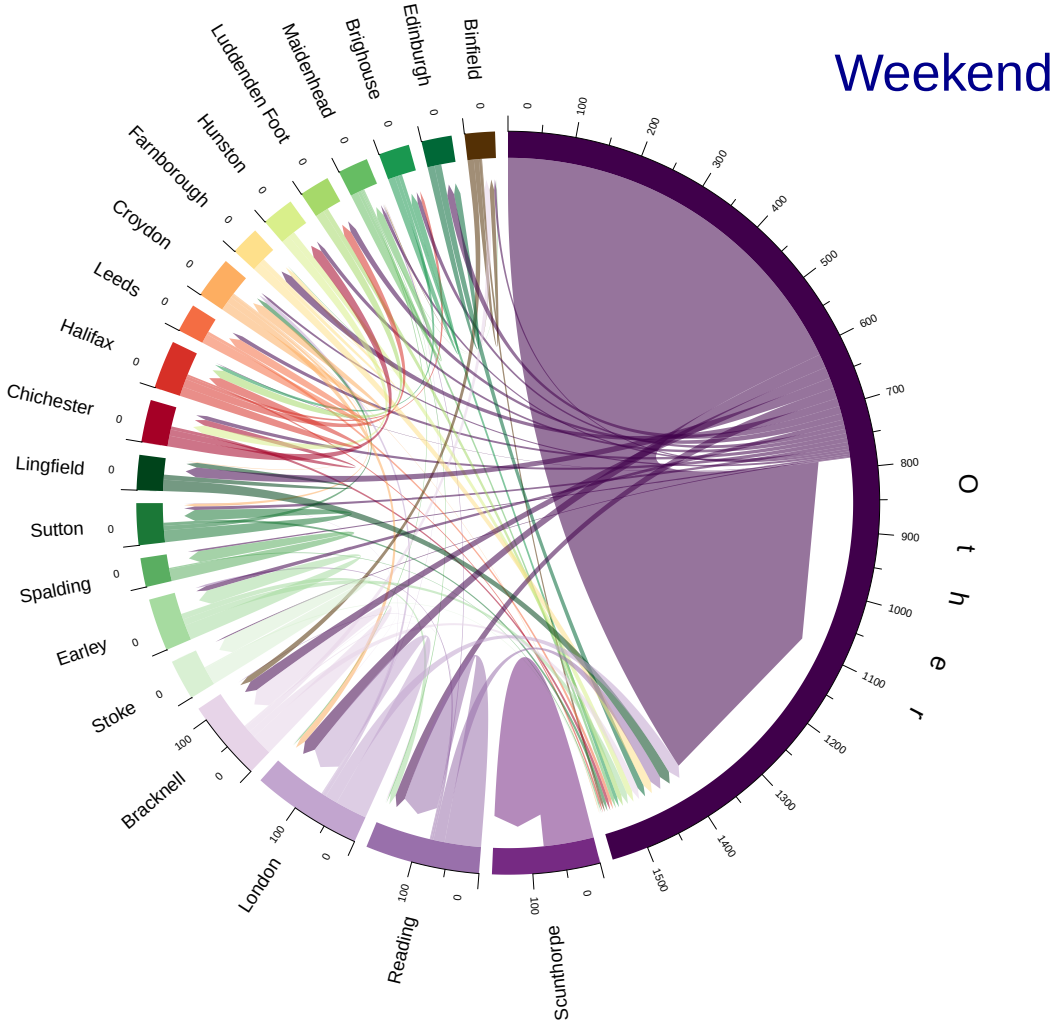

Figure 2: Chord diagram of traffic flows during weekends in the period from 10 August to 13 December 2016.

In the next step, I look at the time patterns of one of the main routes: London intracity car flows. Figure 3 shows the aggregated results for the whole month.

As could be expected, there is a clear difference between weekdays and weekends: on the weekend most trips happen between 10h and 14h, whereas on weekdays most traffic concentrates in the interval from 15h to 20h.

Although the overall intracity traffic flow is weak, considering the size of the city and the fact that the period covers 125 days, there is some overlap in the transit times of the different vehicles. This suggest that there may still be some opportunities for vehicle sharing.

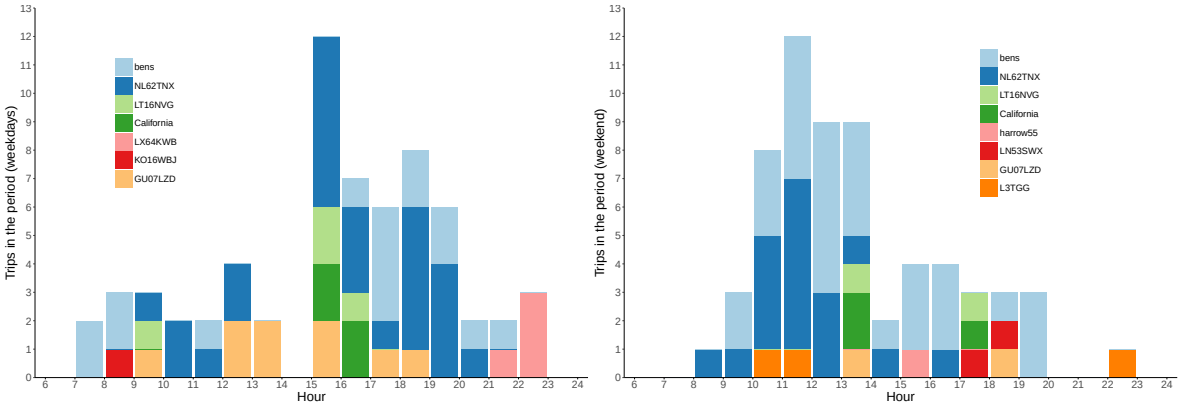

Figure 3: London intracity routes, weekdays (left) and weekends (right). Trips per hour and unique car in the period from 10 August to 13 December 2016

In the next and final step of the analysis, I try to predict those users which could benefit from car sharing by applying the following filtering criteria:

1. Vehicles that traveled to the same city on the same day at similar departure times. I allow for a time gap of one hour, assuming that if they had been aware of the car-sharing possibility they might have been able to concert a common departure time within that hour range.

   AND

2. The departure location of each car is within a given distance. I obtain the number of car-sharing matches for a range of values of the distance parameter (from 2 to 20 km). This parameter is a proxy for the willingness to travel a given distance to meet/pick up the other person in exchange of the transport efficiency derived from the car sharing (i.e. reduced travel costs and reduced environmental footprint).

When implementing the criterion (2) above, it is important to consider which projection is

used to transform the difference in the longitude and latitude coordinates into a planar distance in meters. Given that the data observations correspond to the UK, I use as a CRS for the projection the OSGB 1936 / British National Grid (EPSG:27700).

Figure 4 shows the results of the matchings between different vehicles. Initially, as the start distance allowed between the two cars is increased, the number of matches rises. However, starting at 16 km, (moderately) increasing the distance does not change the number of matches given that all possible car sharers in the area are already covered, the rest of the cars being in other urban agglomerations.

It is worth noting that the number of possible car-sharing opportunities on the weekend is lower but still remarkable in relative terms. Indeed, we are comparing absolute numbers and in the period covered there were obviously more weekdays than Saturdays and Sundays.

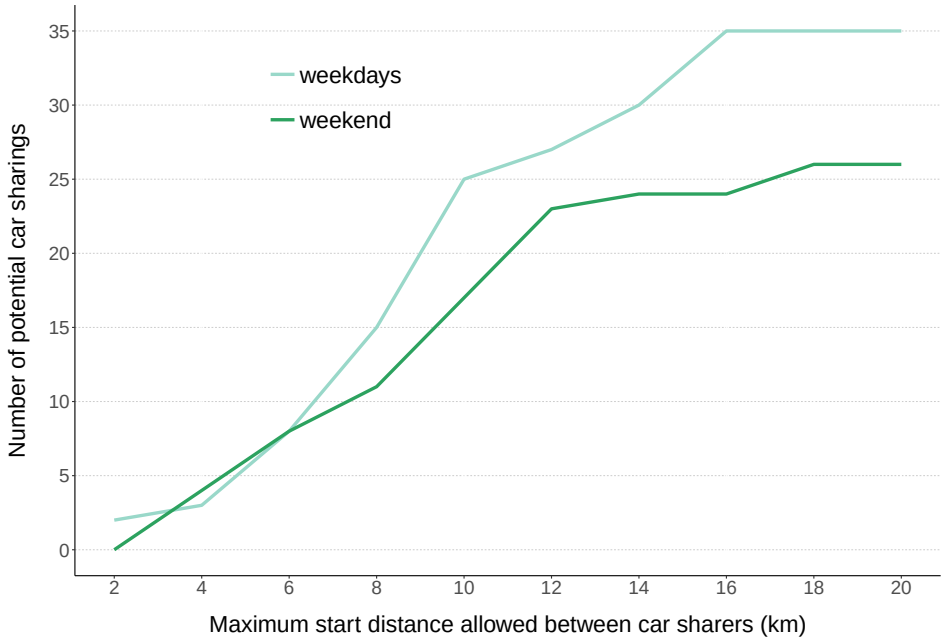

Figure 4: Number of car sharing matches in the period from 10 August to 13 December 2016 depending on the threshold distance considered between the starting points.

Based on these results, there are a total of about 60 possible car-sharing opportunities that could be proposed to these users. Given that the dataset covers a period of 125 days, it is a non-negligible number.

4 Future work

Based on the information contained in the dataset, the following additional analyses could be envisaged:

1. The trajectories and times of those cars moving from one city to another could be interpolated in order to have a more complete estimation of car sharing opportunities. That is, in the path A-B-C, a car may start at A and end at C, yet pass by B at some point in time in which another car may be starting its journey from B to C. Different methods to interpolate and keep track of paths between signal points have been used, for instance, in the context of the tracking of the Ebola outbreak in West Africa (ITU, 2017a,b).

2. Explore the possibility of linking this dataset with external sources based on the geographical location. For instance, the dataset contains no information on the socio-economic characteristics of the users of these vehicles. However, based on their paths, it could be predicted which area could be considered their home location and from there extract some social and economic parameters. This kind of socio-spatial linkage has been explored, for instance, in (Riddlesden and Singleton, 2014) precisely for the UK.

3. Beyond the mobility analysis, there are other possible predictive analytics that could be envisaged based on the available records. An obvious option would be to exploit the speeding, hard-brake and hard-accelerator counts to determine the driver's level of risk. This kind of approach has been implemented by several car insurance companies, as well as by technology firms active in the sector.[1]

[1]For and example, see https://www.cnbc.com/2016/10/23/more-auto-insurers-want-to-track-your-driving-behavior-in-exchange-for-lower-rates.html.

References

Bivand, R., Keitt, T., and Rowlingson, B. (2017). *rgdal: Bindings for the Geospatial Data Abstraction Library*. R package version 1.2-7.

Gu, Z., Gu, L., Eils, R., Schlesner, M., and Brors, B. (2014). circlize implements and enhances circular visualization in r. *Bioinformatics*, 30:2811–2812.

ITU (2017a). Call detail record analysis: Republic of liberia. http://www.itu.int/en/ITU-D/Emergency-Telecommunications/Documents/2017/Reports/LB/D012A0000C93301PDFE.pdf.

ITU (2017b). Call detail record analysis: Sierra leone. http://www.itu.int/en/ITU-D/Emergency-Telecommunications/Documents/2017/Reports/SL/D012A0000CA3301PDFE.pdf.

Kahle, D. and Wickham, H. (2013). ggmap: Spatial visualization with ggplot2. *The R Journal*, 5(1):144–161.

Riddlesden, D. and Singleton, A. D. (2014). Broadband speed equity: A new digital divide? *Applied Geography*, 52:25–33.