Analysis of big data set of urban traffic data

M. Koroliuk [1]*, C. Connaughton [2]

Abstract
Modern vehicles are increasingly capable of reporting location and status information in real time using GPS-enabled on-board telemetry boxes which connect directly into a vehicle's control and diagnostic systems. We perform an exploratory analysis of such data obtained from a vehicle operating in a large UK urban area. The primary objective is to devise informative summary statistics that allow different "types" of vehicle activity to be identified and abnormal behaviour to be quantified. We use position, speed, time and engine status (ignition on/off) to organise the data into working units of increasing temporal duration. We apply hierarchical clustering methods to some functions of these working units to identify different types of vehicle paths in addition to quantifying how the periodic daily variation in traffic conditions in a modern city affects fleet movements and behaviour.

Keywords
driving mission recognition— unsupervised methods — map constructing

[1] *University of Warwick, Centre for Complexity Science*
[2] *Warwick Centre for Complexity Science and Warwick Data Sciences Institute*
***Corresponding author**: maria.korolyuk@gmail.com

Contents

Introduction and motivation

In the modern world we have more and more data about moving vehicles and an interesting question that arises here is how informative the data are. For example, current technology of GPS installed on phones allow to track a vehicle's position on a regular basis, therefore monitor (or even spy) where the persons spend their time. In fact, there is much information hidden in such data than location history. This information could be valuable for companies to build optimal paths or other real time monitoring applications. In addition, it is useful to know how much can be inferred from the raw data with respect to security: If due to a leak or intended spying a third party gets gets unauthorised access to the information, how much could be inferred from it?

The aim of this research is to obtain information that is not directly interpretable from the raw vehicle data, using only a basic data set provided. In particular we are interested in inferring information that can afterwards be validated by other sources. Our research questions can be divided into two sets; one set concerning location and another set concerning driving patterns. Firstly, we aim to answer the following groups of questions concerning location:

- ⋆ Can we build a personalised map using the data set
  - What are the places/regions visited with high frequency (home bases,gas stations, shops, parking lots) and how can we distinguish them?
  - What are the places and times where speed is limited? (regular traffic jams, vulnerable roads)

Even though some of this information (such as traffic lights) is freely available on internet , the process of extraction holds scientific interest, as basically it allows to build maps from the set of records. This might be useful for constant automatic updating and compressing the information about the structure into a smaller data set. Also it might have application in journey planning, designing architecture of the roads, predicting possible traffic jams.

Secondly, we tried to identify the driver purpose. For example, how different is the behaviour of a driver who is delivering packages from the behaviour when he is simply driving to advertise a company.

This is a challenging aim which we divide into smaller steps:

1. Separating the data available into smaller quantifiable units- which are easier to work with
2. Identifying units that share common purpose (such as delivery, driving,etc.)
3. Identifying units that are possible outliers/ missed or damaged data.

In particular I want to empathise the last aim, as the task of identifying suspicious records in a set of this size is non-trivial, as being an outlier is not a property of a record itself, it is a property of connection on a record with previous and future ones. Therefore the procedure of identifying the outlier should take it into account.

The structure of the paper is as follows: Firstly, the data will be described in detail. Thereafter, we give an overview of some literature and explain why we need to develop new methods. Afterwards, we will try to perform some basic statistics methods, and evolve it towards map construction. Then, we will divide the data into working units and will apply clustering methods on some features of this units.

Data

The data for this project are provided by the company L&A Consultants which is based in East London and specialises in the acquisition and analysis of spatial data for vehicle fleet management. They manage small devices that can be installed into cars and can acquire and transmit information about recorded values of speed, position and time into a data file within a small time interval. Their system, iR3 enables advanced tracking data and reporting metrics. The data set provided to us consists of 7×10^5 records in the form of unique ID of vehicle, time and date, Longitude, Latitude, Speed and Ignition. The time interval between records is not fixed. If the values have not changed, the record is not sent to the receiver, whilst system attempt to send records regularly if any position had changed. The data cover a period of two years, with expected value between records of 20 seconds.

Background

Literature overview

The problem of driving pattern recognition is common in the literature, however the discussion focuses around local intention such as turning or changing the lane, little is known about defining and identifying the global purpose.

Most of articles introduce feedback (classification) variable, while we do not have an opportunity to use it and are strictly limited by the data provided by the customer.

For example in, Hongwen, Sun [1] a driving pattern identifier based on a learning vector quantization were introduced, analysysis of six selected representative standard driving cycles. is held. Micro-trip extraction and principal component analysis methods are applied to ensure the magnitude and diversity of the training samples. We will use a similar approach of calculating traits and making inferences from them, but we will apply this technique as an input for unsupervised learning that tries to identify the different intentions of the driver.

In Liao, Patterson [2] we can see a way of extracting important location from of the record set, using a system that can learn personal maps customised for each user and infer their daily activities and movements from the raw GPS data. However, as it is based on training data and adapted for individuals, its describe a system where car visits a small amount of places regularly. In our research we work with the system that disagree with this assumption: every location is equally probable at the future.

Therefore we will incorporate some insights from the literature, such as looking at some features of the records rather that the whole record, developing our own approach.

Methods

In the project we used the python program language with the following packages: pandas (high-performance, easy-to-use data structures and data analysis tools) , scikit (machine learning package) and seaborn (visualisation library).

Some of the questions might be answered using simple statistical methods, such as frequency analysis, for estimating the probability density function of a random variable from the sample. The derivation of patterns in the behaviour is partially done by using Agglomerative Hierarchical clustering with complete-linkage method [4].

The biggest advantages of this methods is that it is an unsupervised learning algorithm: no feedback is needed and no a priori information about the number of clusters or structures is required. The method is based on merging two classes that are similar, until only one class is left. The disadvantages are that it is time consuming (complexity of at least $O(n^2 logn)$), does not produce rules for future prediction and it is not clear how to estimate the performance level, as no objective function is directly minimised. Moreover it it requires to decide on many groups to be defined, which opens an possibility for wrong interpretation.

We used the Ramer-Douglas-Peucker (RDP) algorithm reduce the number of points in a curve that is approximated by a series of points to simplify the path [5] ,[6].

Hypothesis, limitations and challenges:

For the purposes of this project, we assume that the transition between recorded states is smooth, continuous and uniform, and the records are ordered in time. We assume the existence and uniqueness of the aim of the trip in every fixed moment. Finally, we assume that the car makes long stops between different missions, moreover the mission does not change while driving.

The main challenge in the described problem is the absence of response variable, which makes impossible the use of well-known techniques from supervised machine learning. Moreover, it makes it difficult to do a hypothesis test to determine the nature of the difference between trajectories.

The second limitation is that due to security measures from the company providing the data, the values for Longitude and Latitude are encrypted in a way that saves distance between them which makes it impossible to use true coordinates. The encrypted value of the position makes it hard to check the significance of important locations found and also does not allow to deduct information about the relevant locations. However, this contribute towards the goal of inferring information using limited access for sources of inferring.

1. Exploratory analysis

1.1 Summary statistics and data validation

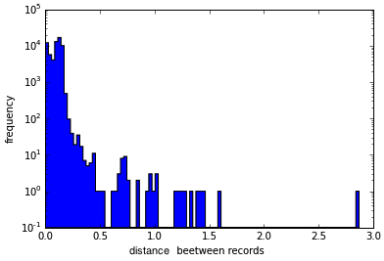


Figure 1. Example of histogram on original data set: distance between consecutive records (miles)

The data we have, as might be inferred from 1 are not smoothly recorded: if we look at the tail, we see that sometimes the distance between consecutive records gets up to 3 miles. Considering the time between records is no more than 20 seconds, 3 miles is not realistic if transmitting works as expected. Therefore we conclude that data are subject to natural errors and biases created at recording. This is an additional source of problems while working with the data set. This is also a reason for some results are sometimes not so smooth. However, in general the records make sense and are appropriate to use for the aims defined.

There is a source of over-defining in the data we have: namely, speed might be inferred only thought time and location as the first derivative of position changing over time. Therefore we might try to compare the recorded value for the speed with the value that car is approximately moving with. Nevertheless, it turned out to be, that such a way of approximating speed does not work well. If we look at the histogram of the difference of reported and estimated speed figure 2 we found, that this difference is significant and could be modelled by normally distributed value with $\mathcal{N}(-3,20)$ (therefore standard deviation of 4.5 miles/hour). The fact that the difference in figure 2 could be modelled so good by normal distribution is not expected and interesting to observe. This is important, because otherwise it might be considered to save memory and no longer record the value for speed.

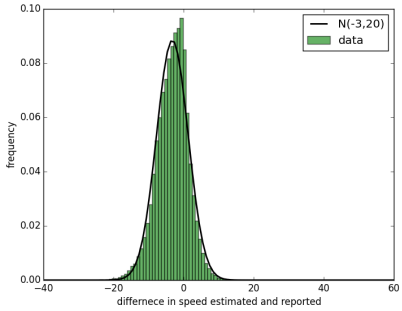


Figure 2. Histogram of the difference between reported and estimated speed

There are several possible explanations for this issue:

1. The most important reason is that in small timespans like 20 seconds even very small difference in position recorded makes such a huge difference. In fact, if we calculate the same difference for the estimated and real change in position (in miles) it appears to be modelled $\mathcal{N}(-0.02, 0.001)$. This small error is enough to cause the big differences in speed in 2.
2. The negative expected value could be explained by the fact that modern speedometers are tending to underestimate real speed to provide lower level of speeding.
3. A factor that comes from acceleration.
4. A factor that comes from the 3rd dimensional in location (altitude) .

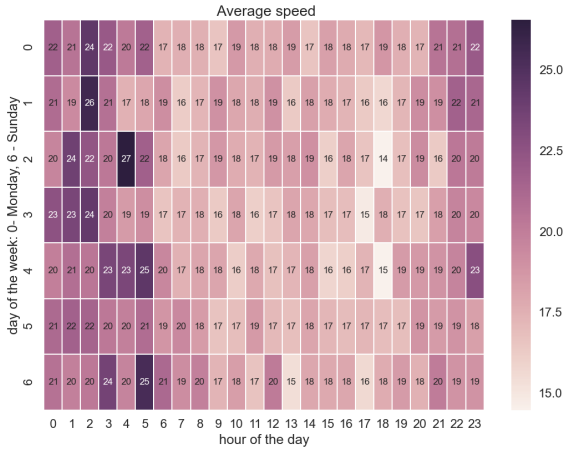


Figure 3. Heat map of average recorded speed recorded on different parts of the day. We can see that during the night speed is higher, while during the rush hour is it hard to move quickly.

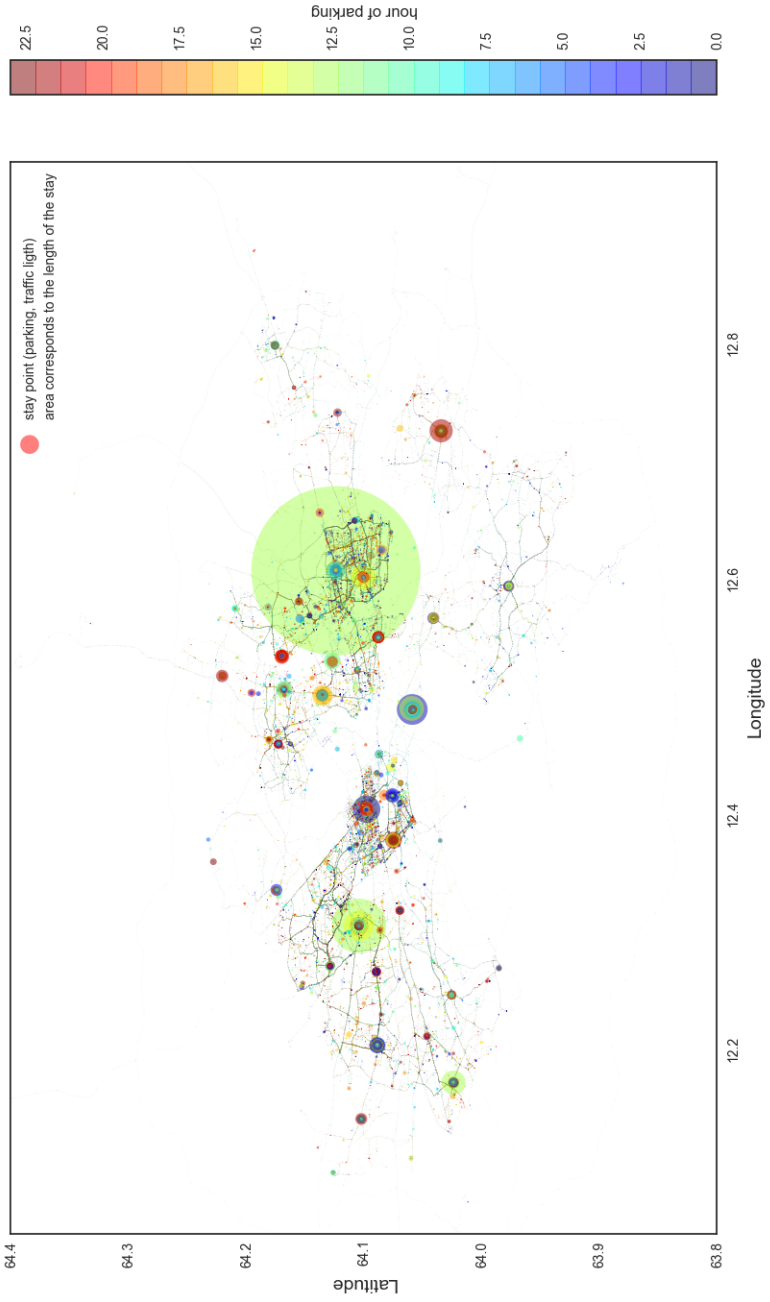


Figure 4. Personalised map of point where car is likely to stop . It is helpful to identify some important positions, such as parking, traffic light, parts of the town with hard and easy traffic.

One more description interesting to look at is the Figure 3 where we might see the average recorded speed (taking only records that have positive value) in different times/ days. Original reason for it to be done is that it is natural to expect that high speed during night bears different meaning that high speed during the day, therefore it is not the speed itself that is informative but its difference from average recorded level. It turned out not to be true (or there is not enough data for enough accuracy) , however the Figure 3 is a good way to find out peak hours and the best hour for traffic, which is important for journey planning.

1.2 Maps

In this section we present some of the maps made out of records. Having access to the coordinates allows us to plot them with the colour, that corresponds to a particular value (such as time, average recorded speed, etc.).

Figure 5 displays the road structure along with the popular locations, while the hue shows average recorded speed at this position (where blue means slow and pink means fast). We observe centre-based arrangement, with predictably slower speed at the centre, and higher outside the internal ring. Also in this figure we mark the most popular locations , were identified from the frequency of the records made.

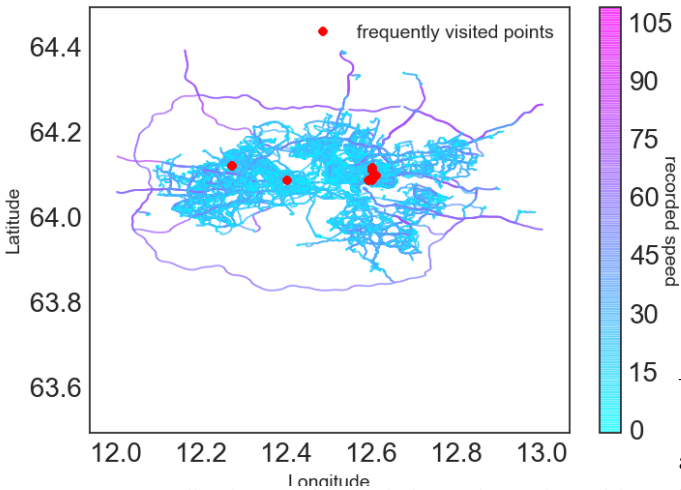


Figure 5. Personalised map of recorded speed together with marked bases.

In figure 4 we can observe several interesting aspects of car behaviour. Every point on the figure represents a position where the car speed had dropped to the zero value. The area of the point represents the time spent at the location ,and the colour the hour of the day when the parking was done. Therefore we can independently identify locations where parking was done for a long time , which are likely to be parking spots (big circles) and locations that have frequent short term stops, likely to be traffic lights (lots of small circles). Moreover, we can observe places with slow motions- likely to be the regular traffic jams. Another interesting observation is that long term parkings are likely to happen at lunchtime or earlier at night.

As mentioned before, the data are encrypted in a way that makes the true location unaccessible, however we believe that the map helped to identify the place records were made at, therefore such a way of encryption is not sufficient.

Please refer for appendix 3 for other examples of maps constructed.

2. Identification of different behaviour

2.1 Trajectories

One of the first steps in identifying different types of car behaviour is to separate all the data into working units, named ”trajectories”. A natural way is to say that everything that happens within two consecutive records when speed drops to 0 is one working unit. However this is not the best approach as a driver might stop without the intention to finish the current trip. Therefore we have to allow the driver to make stops up to a particular threshold, and still count it as the same unit.

Algorithm 1 Trajectories detection

```
1:  procedure DETECT TRAJECTORIES
2:      begin,end ← 0
3:      df ← DataFrame
4:      for k in DataFrame.index do
5:          if DataFrame[k].Speed=0 then
6:              end ← k
7:              while df[k].Speed=0 do
8:                  k ← k+1
9:                  end new ← k
10:             if df[end new].time-df[end].time > threshold
    then
11:                 trips indexes ← range(begin,end)
12:                 begin ← end new+1
13:                 while df[begin].Speed=0 do
14:                     begin ← begin+1
15:         end ← k
```

Currently, trajectories are extracted as described at the algorithms 1. In summary, it is defining the longest possible consecutive non-empty subset of raw data, such that for all points of $p_i, p_2, p_3, ..., p_j$ it satisfyed: $p_i.ignition = 0$,$p_j.ignition = 0$, and for $k = i..j$

$$p_{k+1}.time - p_k.time < \Delta t$$

for some threshold Δt.

2.2 Features

Driving style varies when people have different goals. For example, going shopping is significantly different from trying to catch a plane on time. We know that the car we work with made distinct kinds of jobs, that differ in objective and therefore vary in average speed, curvature, lengths, etc. However, this distinction is not very clear as it might happen that

Name	Description	Mean value
Distance (TD)	Distance between the start and the finish of the journey (miles)	1.49
Total lengths (TL)	Sum of the distances between any two recorded points (miles)	5.75
Total time (TT)	Time from the beginning to the end, not including stops (hours)	1.38
Diameter (D)	Maximal distance between two possible points	2.22
Maximal speed (MS)	please note that minimal speed is always zero	41.67
Curvature (C)	Amount of time trajectory crosses itself	24
Sinuosity (S)	A inverse ratio of actual path length to shortest path length	0.52
Number of stops (NS)	Speed reach zero	8
Path complexity (TP)	Length of simplified path/ total length	0.85
Average speed (AS)	The average value for the speed (mile/hours)	9.83
Effective speed	Total distance/ Total time	1.90

Table 1: Main representative features

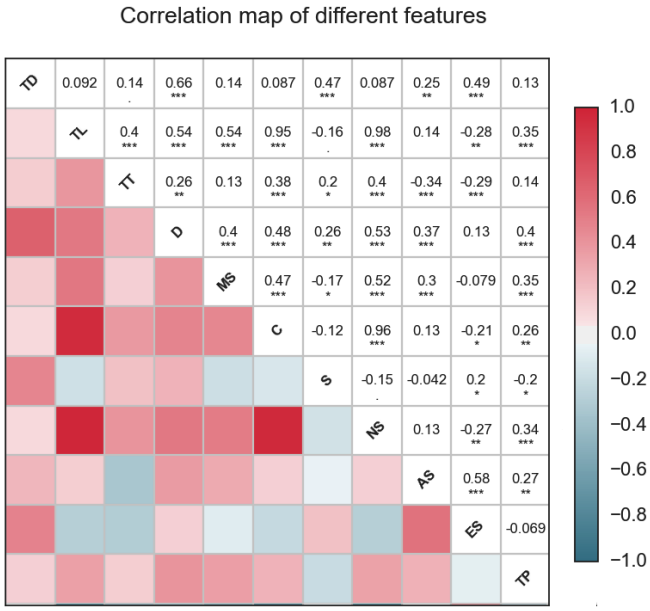


Figure 6. Correlation between different features (please refer table 1 for explanation of variables). For example, number of stops correlate a lot with lengths, and curvature and average speed appeared to be independent.

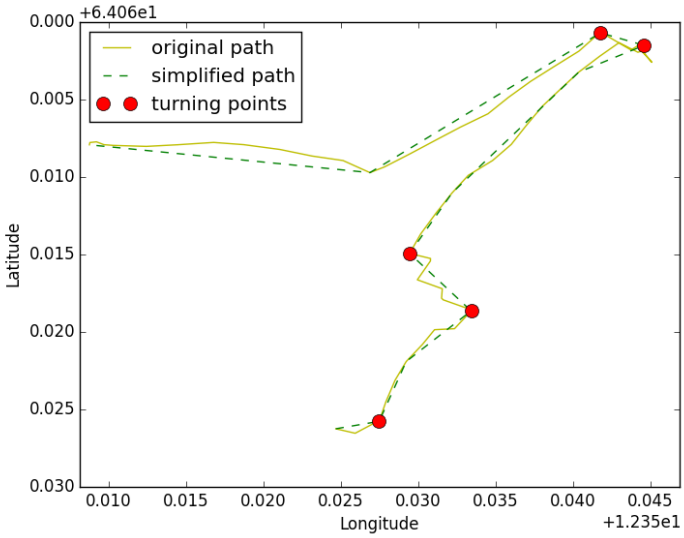


Figure 7. Example of a trajectory with marked explanation for path complexity.

speed is lower not because of the type of job car is doing, but because traffic was busy.

We believe that taking lots of characteristics,gives us the information about type of the trajectory. Therefore we created a list of variables (see Table 1), that can be calculated at different sequences of points. Most of them are self-explanatory, however, some of them need explanation. For example, path complexity (figure 7) is constructed in a few steps: we simplify the path according to Ramer-Douglas-Peucker algorithm (It relys on determining turning points – rapid change in the direction) [5] and [6] , then we calculate its length , and look how much this value differs fro the total path length. For calculating the length of the path we use the Haversine formula ,[8] ,giving great-circle distances between two points on a sphere from their longitudes and latitudes.

It is informative to look at the connection between different features. We constructed a correlation map (figure 6) that shows the dependence between variables. For example, we might infer that total length and curvature are correlated, which tells us that for long paths a car is more likely to intersects its own trajectory. However, there is no causation from long trajectories towards self-crossing points, it is more a identifier that a type of job that refer to long units have factors that causes lots of crosses.

During the process of analysis , we have defined more

features that are not listed in Table 1, such as:

- Coverage area (area of the smallest convex figure that covers all the points).
- How often the trajectory crosses the line from the start to the finish.
- Speed comparing to average recorded speed speed at this time/day.
- Speed comparing to average recorded speed at this particular location.
- How often speed get higher that particular threshold (such as speed limitation).
- How often speed get lower that particular threshold (such as speed limitation).
- How often speed changes its trend in (such as from increasing to decreasing).

All the above turned out not to be informative about final result.

2.3 Clustering

We have used hierarchical clustering to classify the trajectories into different groups, representing different car behaviour. This car behaviour can later be mapped to different trip purposes. The technique of hierarchical clustering is an appropriate technique for classification, because it is a natural way to separate data and does not need any feedback variable. Hierarchical clustering starts by treating every unit as different , and merging similar units together, forming a tree structure.

We used all the characteristics from Table 1 as an input for the clustering algorithms. The number of clusters is chosen from the tree structure. Figure 8 shows the resulting tree from the hierarchical clustering. This figure shows that there are several groups emerging within which there is similarity between the units. The similarity between units of different groups, however, is smaller. This is what causes the group structure. Please refer to Appendix 1 for details on the different groups of trajectories detected. Please refer Appendix 2 to see the distribution of different features into resulting groups detected. From the resulting tree structure, we can also detect some abnormal paths. These are coloured blue and green in Figure 8. These paths are mainly data error/ biases, and will not be considered in the clusters afterwards. For details and examples of such abnormal paths, please see Appendix 1.

2.4 Connection with handwriting recognition

A part of this research involves the application of methods designed for handwriting recognition [3] in the path structure. A method, which uses iterated integral signature,i s designed for identifying the intended characters from the movement of a pen or stylus, which had lots of similarities with trajectories of a car.

The way humans write the letters by connecting the set of consecutive points have similarity with the path that cars make . On figure 11 we can see visual similarity between both examples. If is is possible for artificial intelligence algorithms

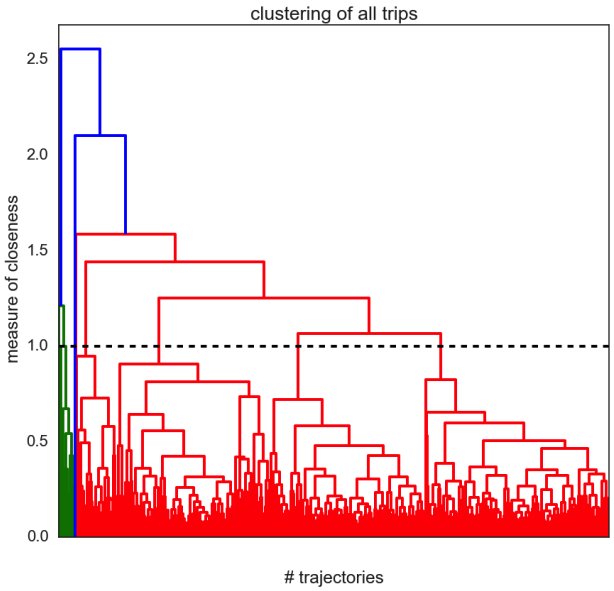


Figure 8. Main tree: emergence of the groups with similar structure (appendix 1 for examples) inside of it. Outliers are marked with blue and green color, while main groups are with red.

to classify handwriting values into alphabet, why wouldn't it be possible for it to classify trajectories into types?

The project from which we used methods, undertaken by Jeremy Reizenstein and Dr Ben Graham [3] , aimed to identify a methodology which improves the accuracy and efficiency of machine recognition of handwritten characters using the iterated integral signature. This method has a few advantages , in particular that it gives the same result to congruent figures, works with any amount of input points and allows to look at the input at different "levels". However,as originally the method is developed specifically for handwriting recognition, this might cause some problems with the type of path identification.

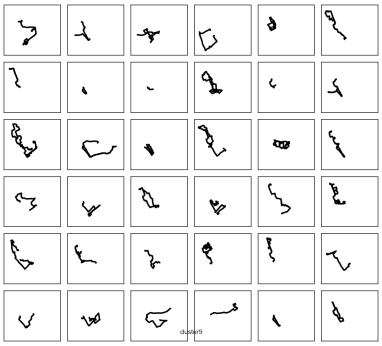
In practice using handwriting recognition towards path structure is successful, but the problem arising here is that the meaning we put into similar handwriting symbols is not perfectly analogous to the trajectories.

3. Conclusions

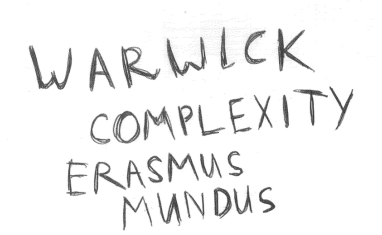
Overall, this research project demonstrates that by using raw data of simple car variables we can infer a lot of unexpected information. In particular, we observe that:

1. It is possible to build different types of maps with significant locations specified on it. Moreover, some of the maps we constructed display patterns that are not very obvious from prior knowledge, such as high traffic roads and busy junctions.

2. It is possible to detect and identify different types of car



A. Trajectories of cars



B. Handwriting

Figure 9. The similarity between handwrittng letter and car trajectories gives motivation to use methods from handwriting recognition in the clustering.

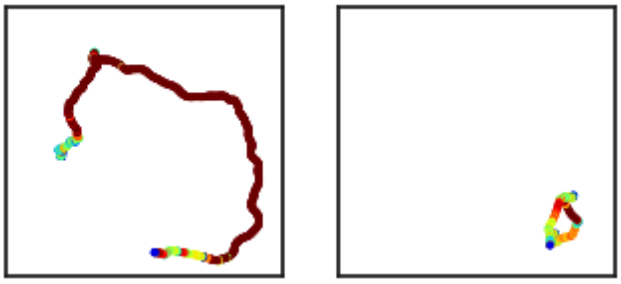


Figure 10. Possible problem that might occur with applying handwriting recognition.

behaviour which we can map to different trip purposes. Unfortunately we were not able to do this mapping, and check for its validity because of the lack of data on actual trip purposes. However, this is impossible for the company providing us the data.

3. The techniques developed here allow to detect outliers. Detecting outliers is not an obvious task, because being an outlier is not a property of a record but a property of connections between previous and consecutive records.

For future research, it would be interesting to;

1. From other sources, validate the separation made in this research.
2. Work with multiple cars:
   (a) They might have correlated behaviour (thus the type of jobs different cars are doing might be the same).
   (b) It will allow to build the prediction of important location that does't depend on individual factors.
3. Predictive power: Generate posterior distributions of likelihood on the car doing a particular type of job based on input of current location, speed and time.
4. Compare with random situation, where all the values are generated from a similar distribution to the real ones and show that the structure in clustering we observe might not emerge.

Acknowledgments

This research was funded by Erasmus Mundus program and carried in collaboration with the L&A consultant, based at London. My gratitude goes to my supervisor C. Connaughton (Warwick Centre for Complexity Science and Warwick Data Sciences Institute) for the guidance and support provided. Thanks also goes to Roozbeh Pazuki from the company for his productive discussions on the project. I thank our colleagues from university of Warwick who provided insight and expertise that greatly assisted the research, although they may not agree with all of the interpretations/conclusions of this paper. I thank Jeremy Reizenstein, university of Warwick for assistance with integral signatures.

References

[1] Hongwen He, Chao Sun, (2012)." A Method for Identification of Driving Patterns in Hybrid Electric Vehicles Based on a LVQ Neural Network". National Engineering Laboratory for Electric Vehicles, Beijing Institute of Technology, Beijing, Energies 2012, 5, 3363-3380; doi:10.3390en5093363 .

[2] Lin Liao and Donald J. Patterson, (2005)."Building Personal Maps from GPS Data". Report from department of Computer Science and Engineering,University of Washington.

[3] Jeremy Reizenstein , Ben Graham (2014). "Signatures in online handwriting recognition". Report of Department of Statistics and Centre for Complexity Science University of Warwick.

[4] Ward, Joe H. (1963). "Hierarchical Grouping to Optimize an Objective Function". Journal of the American Statistical Association 58 (301): 236-244. doi:10.2307/2282967. JSTOR 2282967. MR 0148188.

[5] Urs Ramer(1972), "An iterative procedure for the polygonal approximation of plane curves", Computer Graphics and Image Processing, 1(3), 244-256 doi:10.1016/S0146-664X(72)80017-0

[6] David Douglas ,Thomas Peucker (1973), "Algorithms for the reduction of the number of points required to represent a digitized line or its caricature", The Canadian Cartographer 10(2), 112-122 doi:10.3138/FM57-6770-U75U-7727

[7] Rosenblatt, M. (1956). ”Remarks on Some Nonparametric Estimates of a Density Function”. The Annals of Mathematical Statistics 27 (3): 832.

[8] W. Gellert, S. Gottwald, M. Hellwich, H. Kastner, and H. Kustner,” The VNR Concise ” Encyclopedia of Mathematics, 2nd ed., ch. 12 (Van Nostrand Reinhold: New York, 1989).

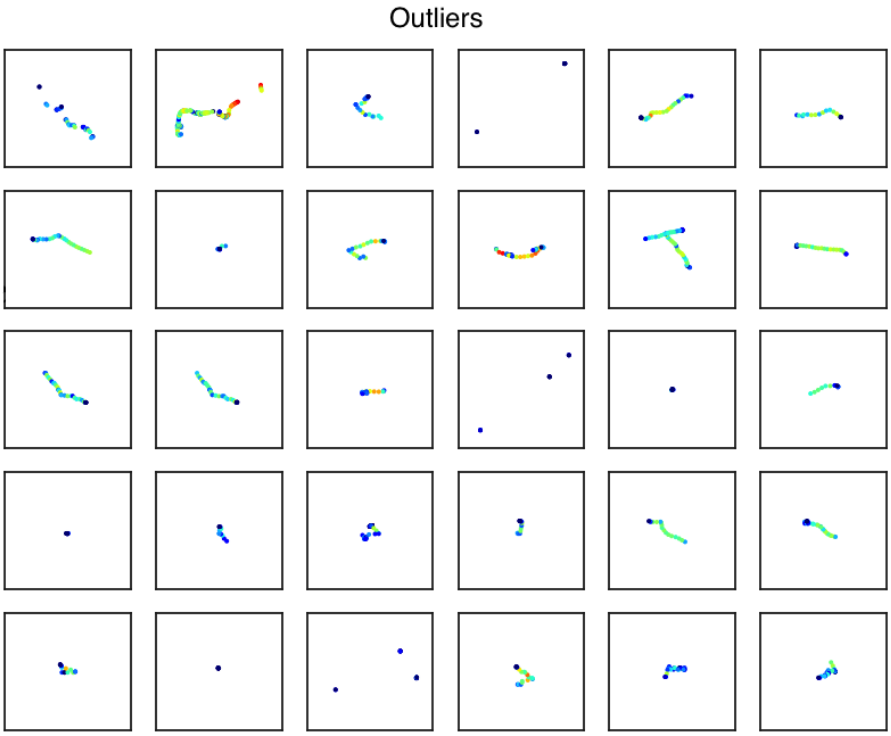
[9] M. Ahmed, B.T. Fasy, K.S. Hickmann, C. Wenk. Path-Based Distance for Street Map Comparison. In ArXiv: 1309.6131, 2013.

[10] M. Ahmed and C. Wenk. Constructing street networks from GPS trajectories. In Proc. European Symp. on Algorithms, pages 60-71, 2012.

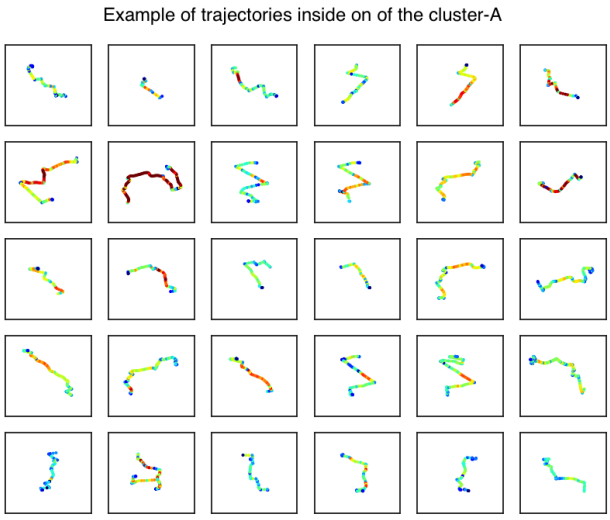
[11] M. Ahmed, S. Karagiorgou, D. Pfoser, and C. Wenk. A Comparison and Evaluation of Map Construction Algorithms. GeoInformatica, 19(3):601-632, 2015.

[12] S. Karagiorgou and D. Pfoser. On vehicle tracking data-based road network generation. In Proc. 20th ACM SIGSPATIAL, pages 89-98, 2012.

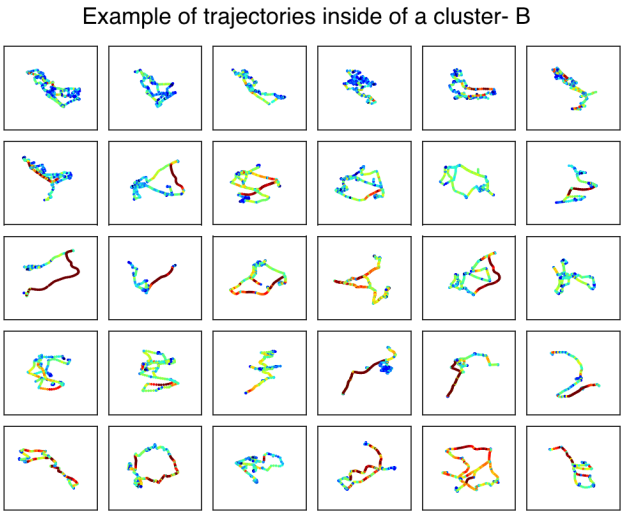
4. Appendix 1: Examples of trajectories in defined clusters



(a) Example of outliers detected: almost all of them correspond to the missed or damaged data, therefore it helps to detect a problem with the transmitting device



(b) one of the detected classes



(c) one of the detected classes

Figure 11. Examples from detected classes

5. Appendix 2: Distribution of some of the features inside of different clusters.

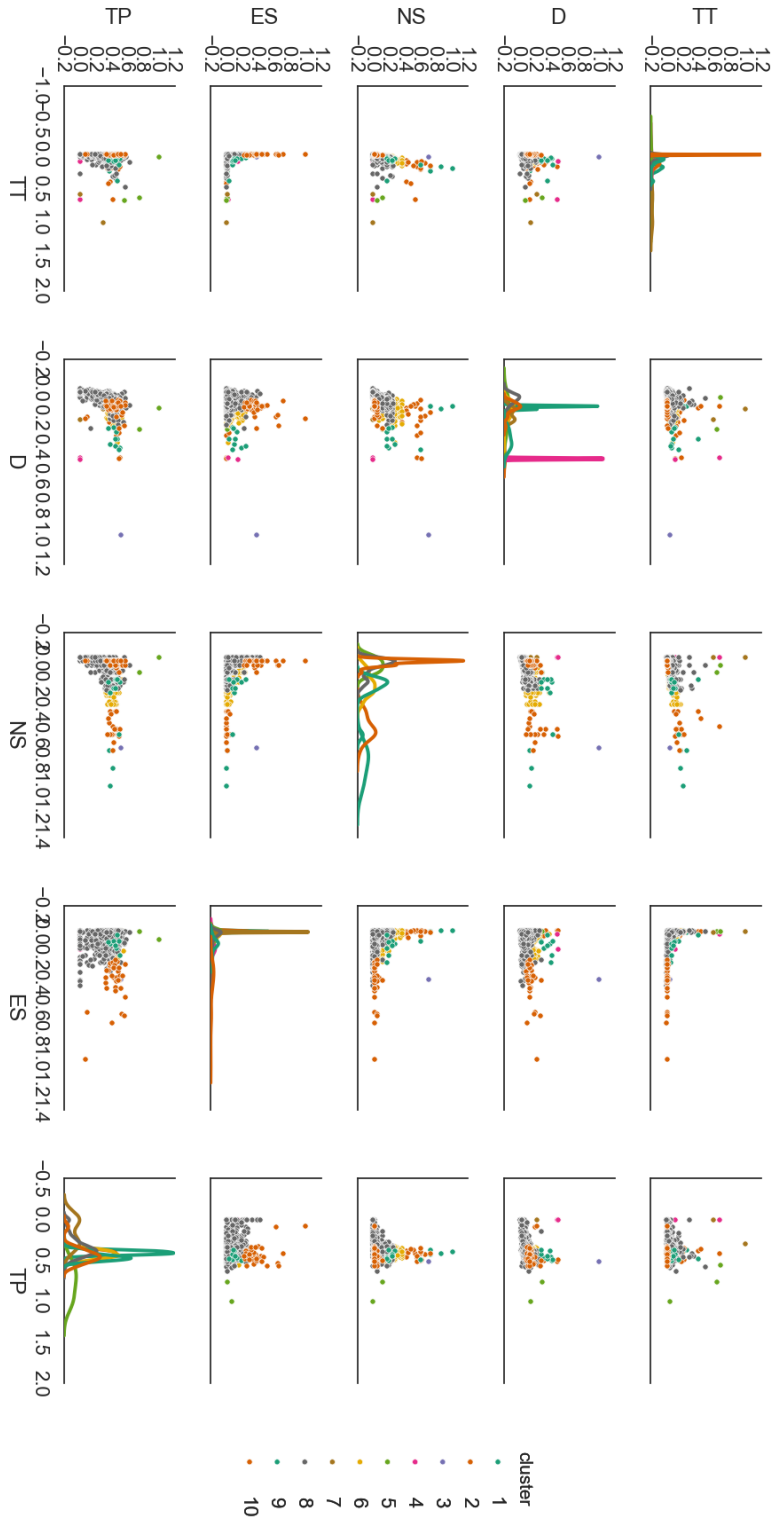
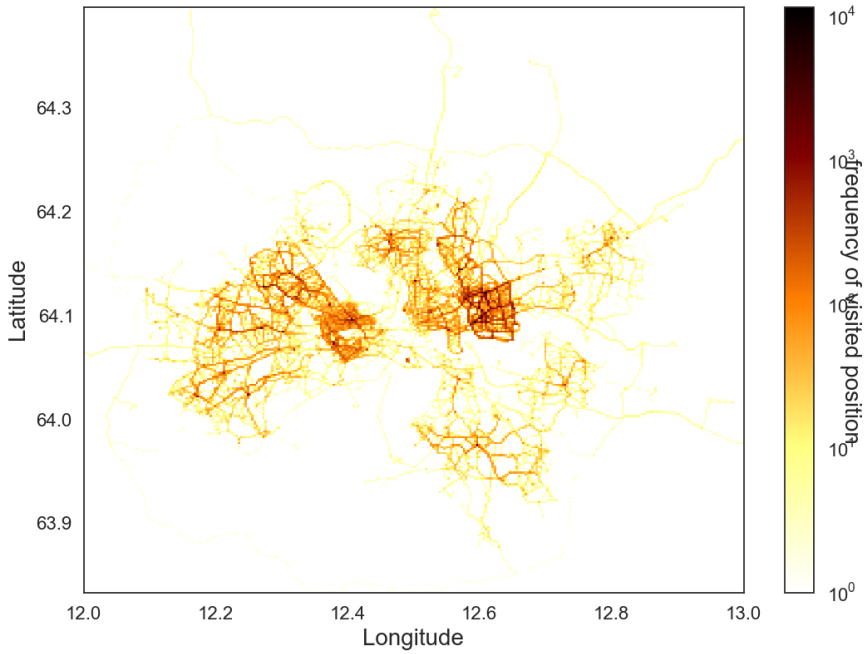
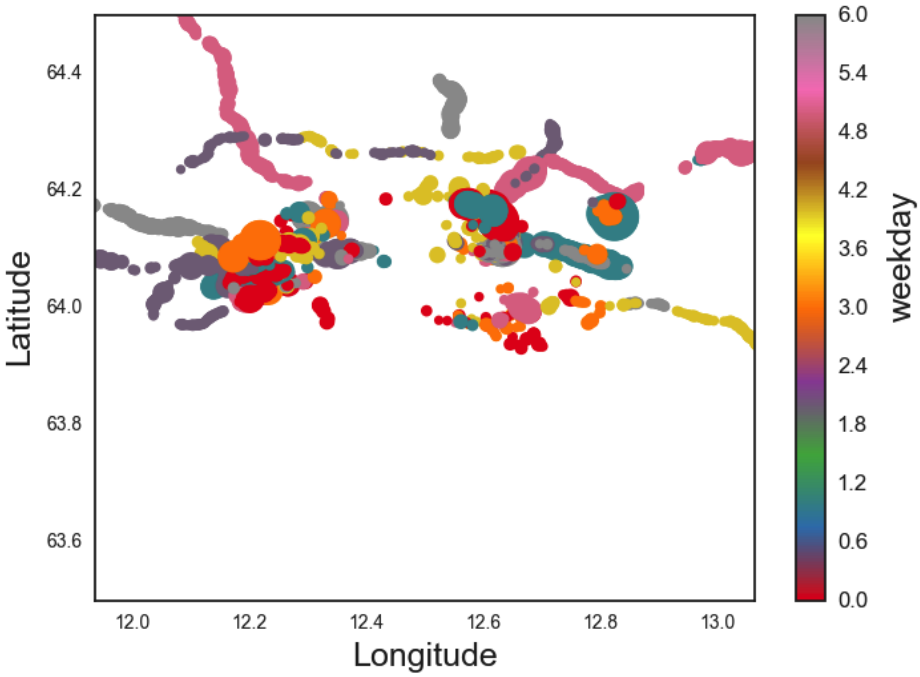


Figure 12. Distribution of some of the features (Total time (TT), Diameter (D), Number of stops (NS),Estimated Speed (ES), Complexity of the path (TP)) inside of found clusters.

6. Appendix 3: Other examples of maps constructed.



(a) Map of the frequency recorded: the hotter the color is, the most frequent the point was visited.



(b) Map of position with high recorded speed (>50 miles per hour) : size of the oint corresponds to the value of the speed, color to the week day when the record was made (0-Monday, 6- Sunday)