Analysis of big data set of urban traffic data

M. Koroliuk, university of Warwick
C. Connaughton, Warwick centre for Complexity , Warwick Data Science institute

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 fleet of telemetry-enabled vehicles operating in a large UK urban area. The primary objective is to devise informative summary statistics 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 (trip legs, complete trips, working days etc) of increasing temporal duration. We apply hierarchical clustering methods to some simple 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.

Formulation

We have 700000 records that show car position, speed, current time and state of ignition with small time span.

Since the data set is so big, it is hard to study it all together. Thus we separate it to smaller units- **trajectories.** Those are the *longest possible sequences, such that the car stops at the beginning and at the end and never stops at the middle for a time bigger than some threshold dT.* From one hand, it allows us to compress data set to a much smaller one, from the other hand it also allows to consider additional traits of the trip.

The main question is which traits contains the most information about the trajectory and if it is possible to separate quantitively different trajectories.

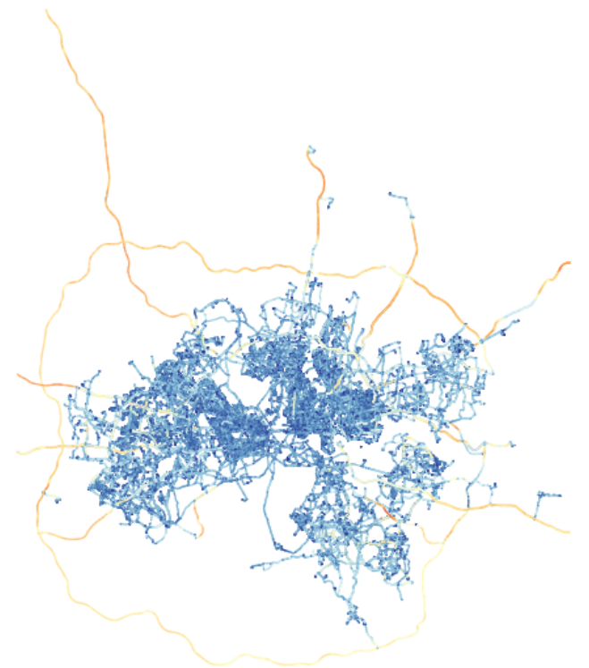


FIG1: Available data

Acknowledgement

Erasmus Mundus
L&A Consultants
centre for complexity science

Variables

Geometrical approach

- lengths
- distance between the end and the begin
- diameter (max possible distance between two points)
- coverage area

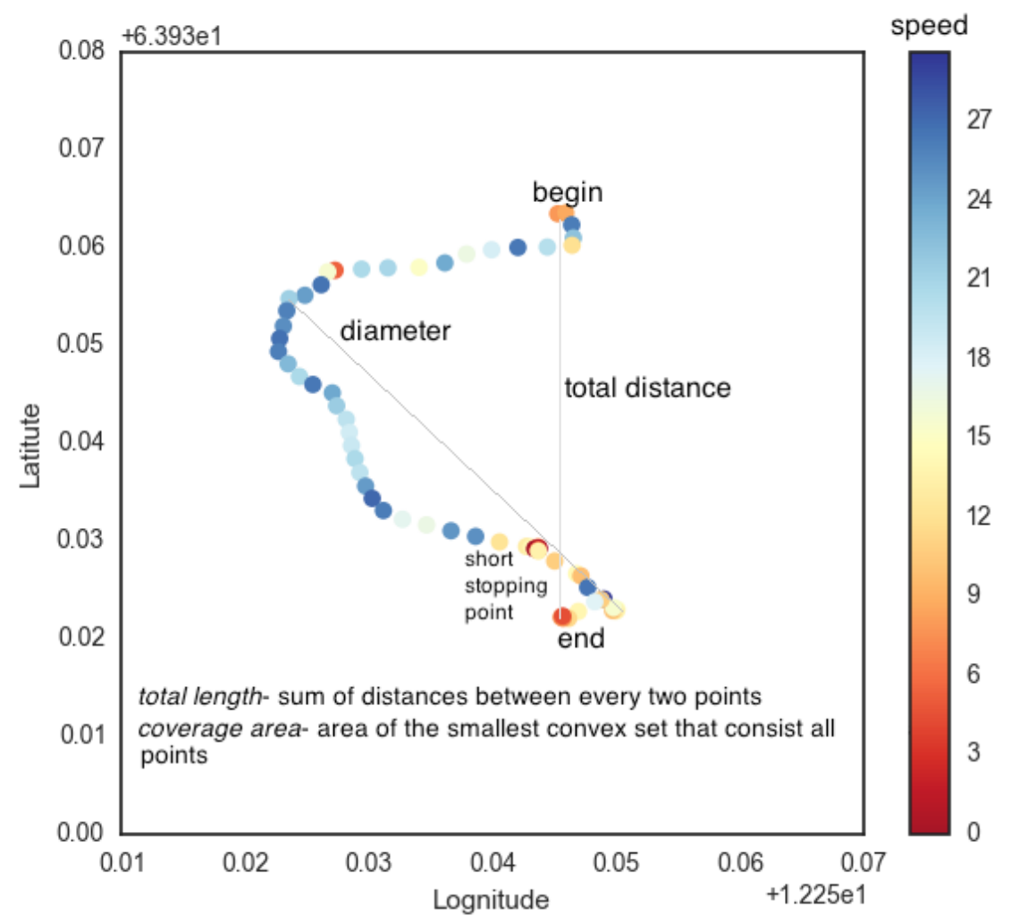


FIG 2: Geometrical characteristics

Frequency of locations

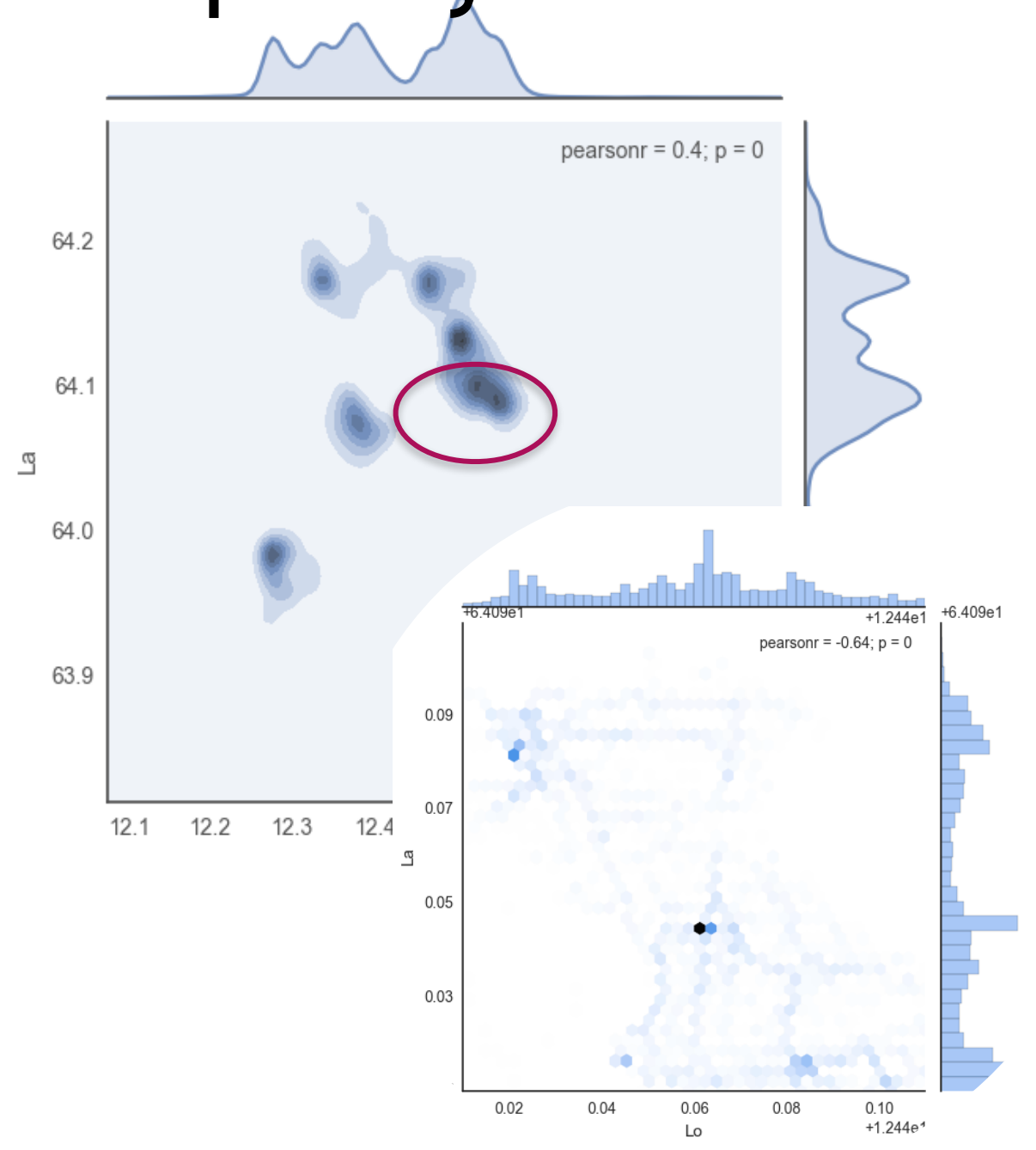
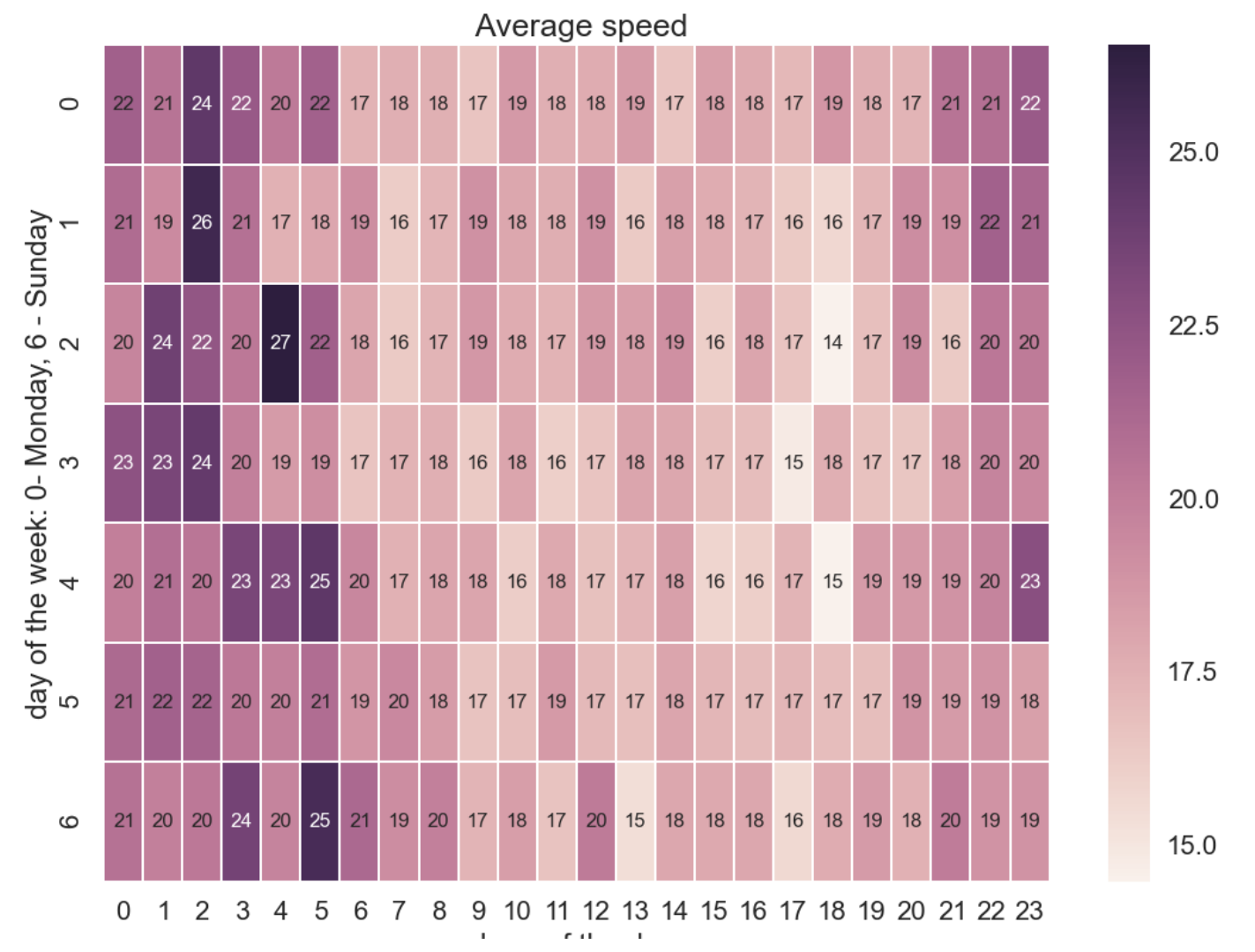


FIG 3: Distribution of recorded locations shows that there are high correlation between longitude and latitude together with area of high frequency of some particular points. Also we might notice road maps structure on zooming in.

Position variables is informative in terms of how often they was used in the past. For example ,visiting new locations usually means different activities that location with long history

Speed characteristics



- Average speed
- Relative speed(length/time)
- difference between average speed of trip and speed at particular time and day
- percentage of the time when speed is getting higher or lower particular threshold

FIG 4: Distribution of average speed of different hours and days. Clearly visible difference between nighttime and day time, together with peak times and non-weekend

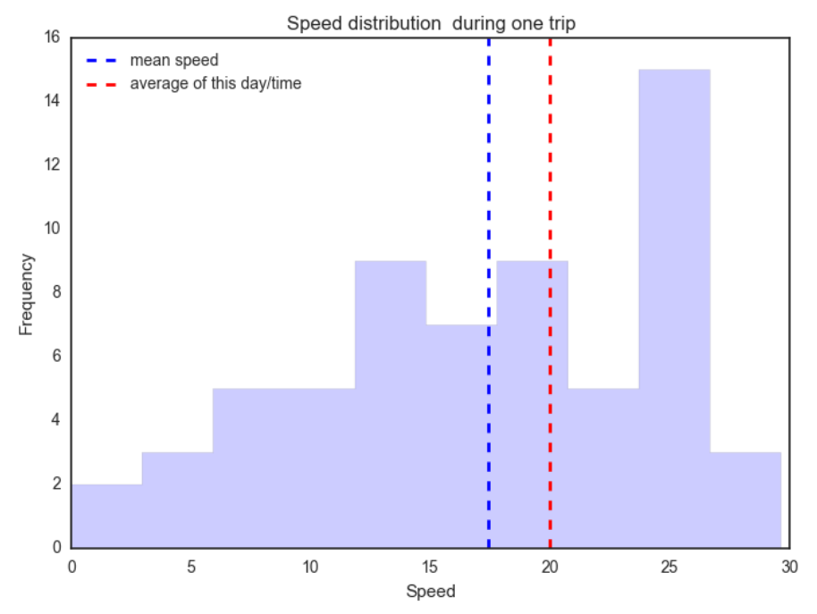


FIG 5: Speed distribution inside one trip

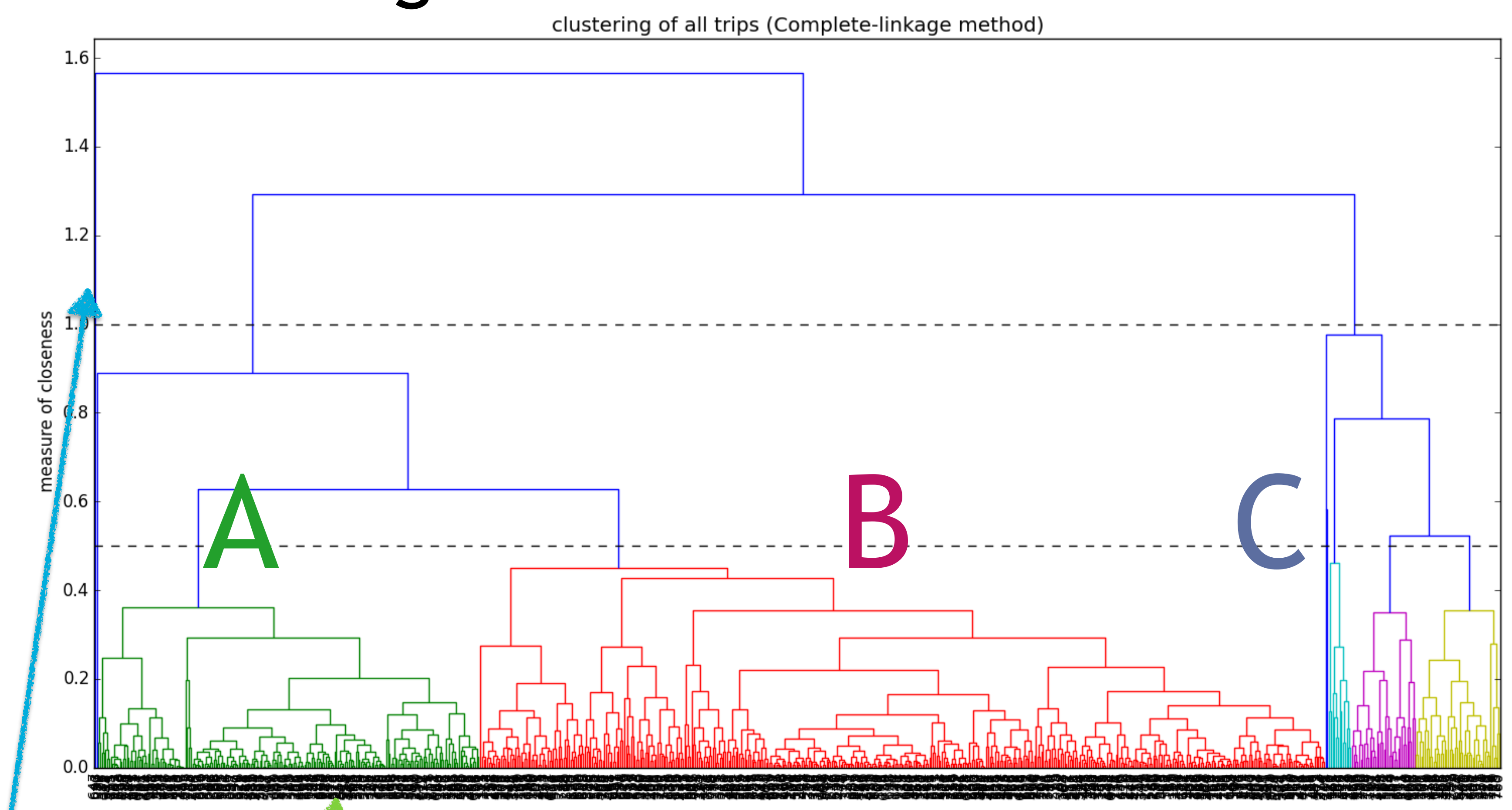
Final choice of variables

An important questions is which of the variables are the best to use. Practice shows, that best obtained result need variables to be:

- not complicated for definition and computations
- Robust to small changes
- covering all aspects (geometrical, speed and frequencies)

For example, a reasonable choice might use difference from average speed, ratio of total length to total distance and ratio of total distance to total time.

Clustering results



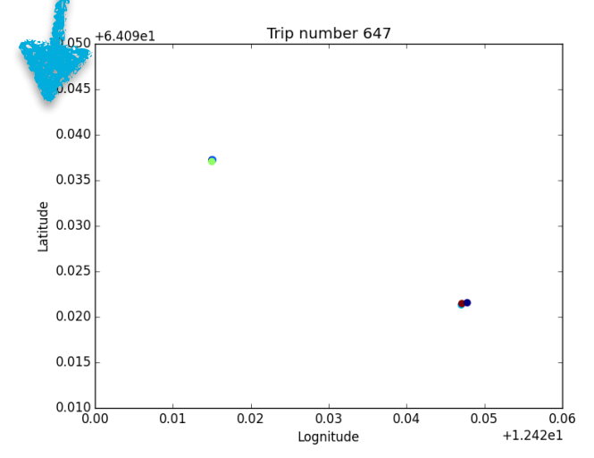
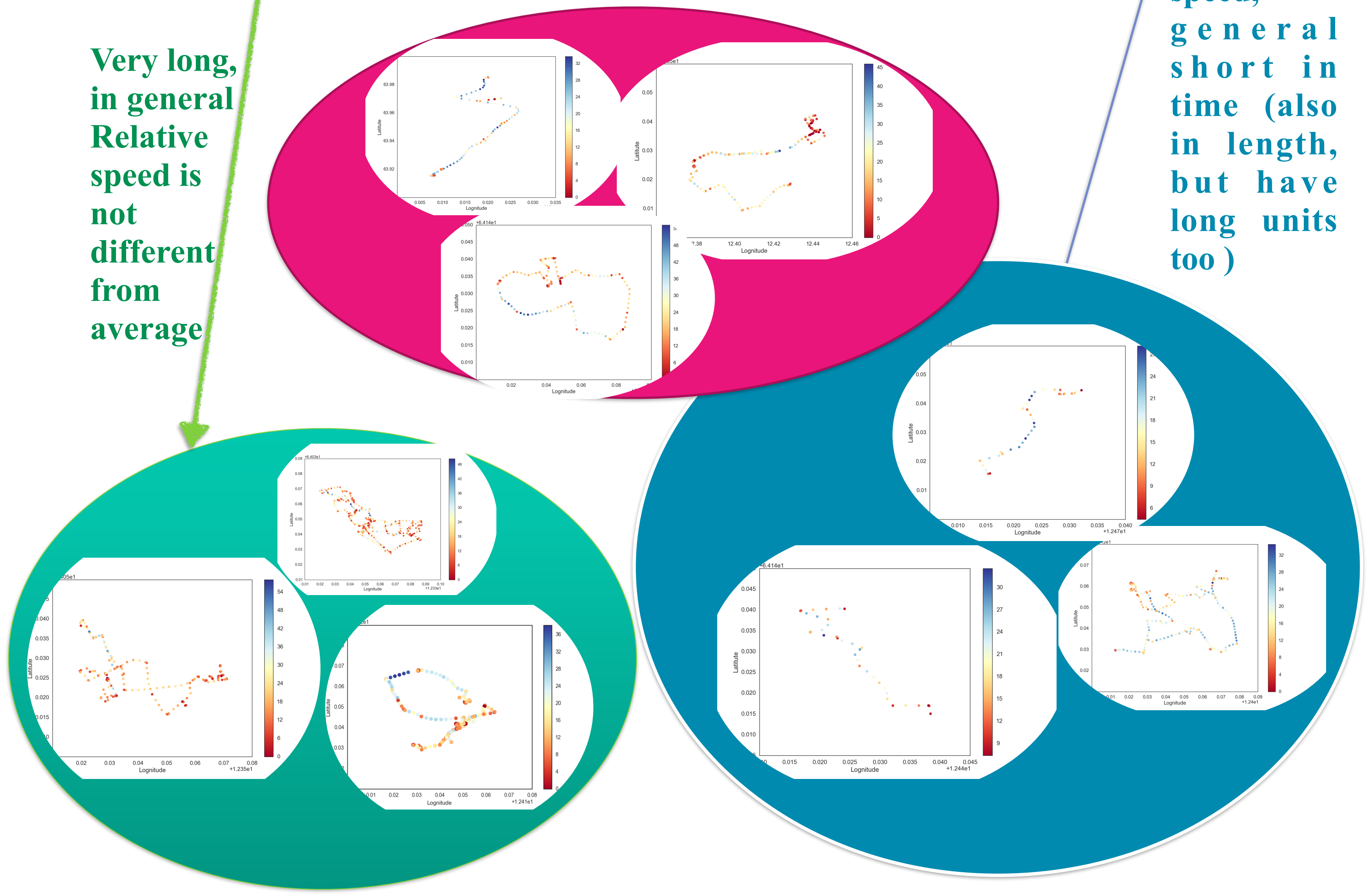


FIG 5: Results of h- clustering - 3 big clusters identified

Low speed, middle lengths

Very high speed, general short in time (also in length, but have long units too)

Very long, in general Relative speed is not different from average



Conclusion

We made an attempt to identify different types of actions of moving vehicles and look at the determining values for such actions. There are few types of variables that are particularly informative (such as geometrical functions, speed functions, frequency variable), however not all of them are good for clustering , for example total length is uninformative towards learning action car is making. There are lots of questions left, for example which other clustering alhoritms we could use (such as Bayesian Hierarchical Clustering) etc.