Hierarchical spatial aggregation of arrival-destination pairs from vehicle telemetry data

Supervisor: Colm Connaughton (Complexity and Mathematics)

Co-supervisor(s): Roozbeh Pazuki

External partner: L& A Consultants Ltd.

Scientific background

L&A Consultants Ltd, is a London-based data analytics company specialising in providing telemetry-based fleet management software to large organisations, particularly the emergency services. Their flagship product, IR3, is a software platform which combines mapping, event processing and data visualisation tools with streaming telemetry data to allow fleet operators to monitor the state of their fleet in real time. Date is acquired by custom telemetry boxes installed in vehicles and streamed via the mobile data network to a central database. The boxes provide updates of position, velocity and other diagnostics several times per minute. For any reasonably sized fleet of vehicles, the data set generated by this systen rapidly becomes very large, typically of the order of terabytes. L & A seek ways to extract actionable information - ultimately in real time - about the behaviour of the fleet from this data which can be used by its clients' fleet managers to inform decision making.

Datasets on the terabyte scale cannot be analysed using conventional computing. They are too large to fit on a typical hard disk and the time to stream through a single processor's memory for the purposes of analysis is prohibitively long. Analysis on this scale requires both a distributed storage system and a distributed computing framework which allows analysis to be done in parallel. Apache Spark [1] is a large scale data analysis platform which provides the necessary functionality. During an MSc project [2] and subsequent internship with L & A one of last year's MSc students, Mariia Koroliuk, wrote a Spark application which processes the raw data into a collection of individual trips by vehicles and provides functionality to extract summary statistics and characteristics of these trips such as total length, origin-destination, total travel time etc. These trip datasets are of a much more manageable size, typically less than a gigabyte, depending on how many features are used to characterise the trip. This data will be the starting point for this project.

Research challenge

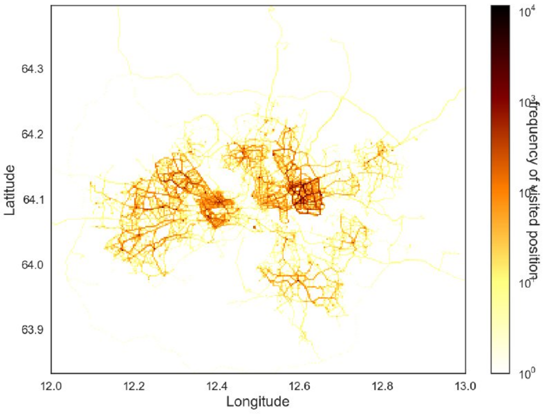

Figure 1: Typical heatmap showing frequency of visits of fleet vehicles in London.

A gigabyte of data about individual vehicle trips is still far too much information to provide a picture of the behaviour of the fleet as whole which is comprehensible to a human. The purpose of this project is to design an algorithm to aggregate this data in space and time in a way which helps to visualise the behaviour at the system level. More specifically we wish to estimate and visualise the flows between different parts of the city in time. The algorithm should be hierarchical so that we can zoom in and out in order to see structure at different scales. Fig. 1 shows a first step in this direction: a heat map which shows the frequency with which a single vehicle visited different locations in the city over a period of a year. In this project, we will do something similar for trips:

- Since GPS coordinates are continuous, the first step is to perform a low-level spatial coarsegraining at an appropriate resolution and map all the arrival-destination pairs from the trips database onto this grid. This can then be used to define heat-maps of arrival and destination locations over different timescales. Such a heat-map will presumably identify hot-spots like in Fig. 1.
- The next step is to define a sensible rule to aggregate squares to form "basins" of each hot-spot based on the number of trips passing through nearby squares en route to each hotspot. Some thinking will be required here to come up with a sensible and informative algorithm for doing this. The idea is that repeated

application of the aggregation rule should allow us to uncover the structure contained in the data at larger and larger spatial scales.

- Once a sensible aggregation algorithm has been implemented we will define "transition matrix" at each level of resolution which will characterise the flows between different spatial regions. Our hope is that such flows will uncover functional information at the large scale about the systemic properties of the fleet in a spatial context.

Pre-requisites

The primary skills required for this project are the ability to manipulate, analyse and visualise data. Ideally the project will be done in Python since this is the language of choice for data analytics in the real world and will use Warwick's new data analytics platform, Chiron. The data management aspects should be reasonably straightforward initially since the analysis will be done on the trips dataset rather than the full raw dataset. As the project progresses, particularly during any follow-on PhD project, the student will be expected to learn Spark programming and acquire the hardware and software expertise needed for doing distributed analytics on large data sets. This is an ideal project for a student who plans to get more seriously into data science at the next stage in their career.

Additional considerations

If the project goes well there will be a follow-on PhD project available and the possibility of an internship with L&A as part of an on-going collaboration between L&A, the Centre for Complexity Science and the Warwick Impact Fund. Since forecasting is a current business priority for the company there is every possibility for the project to have immediate real-world impact, even at the MSc level.

References

[1] Apache Software Foundation. Apache spark - lightning-fast cluster computing. `https://spark.apache.org/`, 2016.

[2] M. Koroliuk. Analysis of big data set of urban traffic data. *Erasmus Mundus Masters in Complex Systems MSc. dissertation*, 2015.