Using Machine Learning and Big Data to Model Car Dependency: Progress Report 1 - Input Data

Robin Lovelace and Ilan Fridman-Rojas

2017-03-20

Introduction

The primary purpose of this project is to show the power and potential benefits of machine learning algorithms in analysing transport data. By demonstrating previously impossible or inaccessible methods we aim to show how new techniques, combined with new and newly open datasets, can generate a strong evidence base for transport planning and policy. The aspiration is that the work will filter into policies, to make them data-driven, transparent, reproducibible and encouraging of citizen science and innovation.

It is well-known that car dependency is close to the root of many problems that are exacerbated by the transport system, economic, environmental and social. However, car dependency is a complex phenomena linked to multiple interrelated factors, the root causes of which are not well understood. This makes it an ideal topic for investigation by machine learning. Therefore, beyond the methodological insights gained by this project, we hope that the results lead to policies that are more evidence-based and effective that decisions based on basic indicators, intuition, and experience alone.

During the first month-and-a-half of the project (2017-02-06 to 2017-03-20) we have spent the majority of the time accessing and organising the data, exploring cutting edge methods in machine learning and researching the literature on car dependency. This **first Progress Report** focuses on the input data, the vital raw material for machine learning algorithms to work. First, however, it is worth considering what is actually meant by car dependency in the context of machine learning, which depends on a categorical or (more frequently) quantitative dependent variable which we are seeking to understand.

Definition of car dependency

There are many possible ways of defining and measuring car use. In this work we interpret car dependency as more about current behaviour than hypothetical future behaviours. Specifically, we define car dependency as:

> The proportion of people in a zone or along a particular desire line who drive, or are passengers in a car or van, as their main mode of transport, taking account of the distances of those trips.

By 'taking account of the distances of those trips' we mean that a zone in which 50% of the population use a car for trips of 5 miles is more car dependent than a zone in which 50% of the population uses a car for trips of 20 miles. We define car dependency *relative to distance* because the policy motivation of this project is to help identify zones and routes where interventions could reduce car dependency. It is easier to intervene to reduce car use for 5 mile trips (e.g. by providing good walking and cycling facilities) than for 20 mile trips (e.g. by creating new bus routes or public transport stations).

Input data

For simplicity and maximum accessibility this project uses only datasets which are open (publicly available). There are three main input dataset types:

- Origin-destination (OD) commute data from the 2011 Census data.
- Geographically aggregated socio-demographic data from the 2011 Census.
- Geographic variables associated with each OD pair.

The Census was used as the primary input dataset because it provides so many variables relevant to car dependency. To our knowledge, the breadth of input datasets have never been analysed together in a single project.

The case study region used for this project was **West Yorkshire**. This case study was selected due to the wide range of social and geographical environments linked to car dependency found here: it includes rural, urban, deprived and privaledged areas.

Origin-destination data

The fundamental input dataset was a table reporting the number of people travelling, by main mode of transport, commuting to work between Middle-Super Output Areas (MSOAs, average population: ~7,500). This is an open dataset (available online from the official WICID data portal) (dataset WU03EW). The file was downloaded as `wu03ew_v2.zip`, an 11.8 MB zipfile which when unzipped creates the file 109 MB plain text file `wu03ew_v2.csv`, read-in with the following command:

```
odall = readr::read_csv("data/wu03ew_v2.csv")
```

The result is a data frame the first two columns of which contain a the code for the MSOA of origin and destination, respectively. The subsequent columns report number of people whose main mode of travel to work was:

- work at home (no transport used)
- some form of metro
- train
- bus or coach
- taxi
- motorcycle or scooter
- drive a car or van
- passenger in car or van
- bicycle
- walk
- other

We grouped together people who drive or are passengers to estimate car dependency. Further, we filtered out flows with an origin or destination outside West-Yorkshire, resulting in **53,807** OD pairs in the study region. This amount of data is sufficient for machine learning algorithms to work and extract complex insights, but small enough for experimentation and fast iteration of methods.

To convert the non-geographical OD dataset into geographic data we used the function `od2line()` from the **stplanr** R package. The result is straight lines that can be plotted on the map and about which geographical variables, such as proximity to motorways, can be extracted.

We will hereon refer to each home-work (origin-destination) pair and its corresponding number of commuters, in geographic form, as a *flow* (see Figure 1).

To this base table there are two ways to further increase the data included for modelling: the Census provides numerous other demographic and economic measures and indicators which can be linked to the commuter's home MSOA (available here), and it also provides workplace data which a group of researchers at the University of Southampton have conveniently created a classification of (by work type, available online as well) which can be linked to the workplace MSOA.

Socio-demographic data

The geodemographic data relevant to each origin (home) MSOA which was deemed relevant and annexed to the flows data consists of the following variables: number of people in particular age brackets, number of

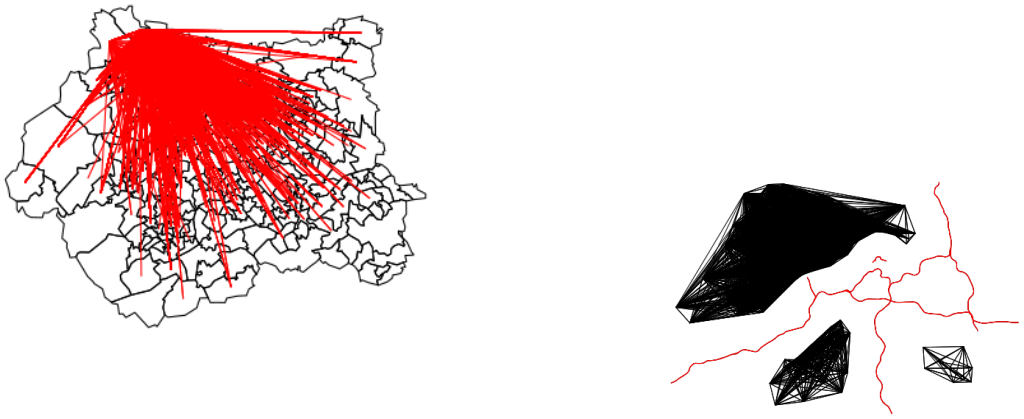

Figure 1: A sample of work commute flows in the West Yorkshire region (left), and a sample of the result of the code which calculates the distance between motorways (red) and flows (black), in this case discarding all flows within a certain distance of the motorway, as an example.

people of each gender, car or van availability (including number of homes with 0,1,2,etc. cars), population density, number of economically active and inactive people, general health (number of people with very good, fair, etc. general health), number of people per ethnicity group, number of people by maximum qualification level, and lastly, average number of rooms, bedrooms and fraction of homes with central heating.

In addition to this we included the workplace zone classification which assigns each destination MSOA (workplace) to one of the following groups:

Table 1: The classification of workplace zones.

Workplace zone classification	Code
Retail	1
Top jobs	2
Metro suburbs	3
Suburban services	4
Manufacturing and distribution	5
Rural	6
Servants of society	7

Spatial data

In addition to this Census data we gathered and computed spatial data as well. We have obtained the location of motorways through the OSM API, and the positions of train stations, coach stations, and bus stops from the NAPTAN dataset. This allowed us to calculate geographical distances between flow lines and motorways, train stations, coach stations, and bus stops. This proximity and accessibility data is arguably vital to best understand and model car usage propensity, and has for the first time become readily available via the OSM platform, and analysis of the type we aim to do next should be cutting-edge data analysis.

Overview of input data

A list of the variables used in the input dataset for the machine learning algorithms is provided below.

Table 2: The variables in the full dataset.

Variable	Description
homeMSOA	MSOA code for origin (place of residence)
workMSOA	MSOA code for destination (place of work)
workhome	Fraction of people who work from home (don't commute)
metro	Fraction of people who use the metro, tram or light train to commute
train	Fraction of people who use the train
bus	Fraction of people who use the bus
taxi	Fraction of people who use a taxi
motorcycle	Fraction of people who use a motorcycle, scooter or moped
car	Fraction of people who use a car or are passengers in a car or van
cycle	Fraction of people who cycle
walk	Fraction of people who walk
othertransp	Fraction of people who use a form of transport not listed above
npeople	Total number of people resident in the homeMSOA
16-24	Fraction of people in the age range 16-24
25-34	Fraction of people in the age range 25-34
35-49	Fraction of people in the age range 35-49
50-54	Fraction of people in the age range 50-54
65-74	Fraction of people in the age range 65-74
75	Fraction of people of age 75+
female	Fraction of people who are female
house0carpct	Fraction of homes with no cars
house1carpct	Fraction of homes with 1 car
house2carpct	Fraction of homes with 2 cars
house3carpct	Fraction of homes with 3 cars
house4carpct	Fraction of homes with 4 or more cars
ppperhect	Population density in people per hectare
econactivpct	Fraction of people who are economically active
econinactivpct	Fraction of people who are economically inactive
vghealth	Fraction of people who are in very good health
ghealth	Fraction of people who are in good health
fhealth	Fraction of people with fair health
bhealth	Fraction of people with bad health
vbhealth	Fraction of people with very bad health
white	Fraction of people of white ethnicity
mixed	Fraction of people of mixed ethnicity
asian	Fraction of people of asian ethnicity
black	Fraction of people of black ethnicity
otherethn	Fraction of people of other ethnicity
noqual	Fraction of people with no qualifications
aptshpqual	Fraction of people with apprenticeship-level qualifications
lev1qual	Fraction of people with level 1 qualifications
lev2qual	Fraction of people with level 2 qualifications
lev3qual	Fraction of people with level 3 qualifications
lev4qual	Fraction of people with level 4 qualifications
otherqual	Fraction of people with other level of qualifications
centheat	Fraction of homes with central heating
nrooms	Average number of rooms in homes

Variable	Description
wzclass	Workplace zone classification (these are defined at a smaller geographical lev
distance	The cartesian distance between the centroid of the home MSOA and the workplace
disttrainstn	The minimum cartesian distance between the flow line and the nearest train sta
distcoachstn	The minimum cartesian distance between the flow line and the nearest coach sta
distbusstop	The minimum cartesian distance between the flow line and the nearest bus stop
distmway	The minimum cartesian distance between the flow line and the nearest motorway

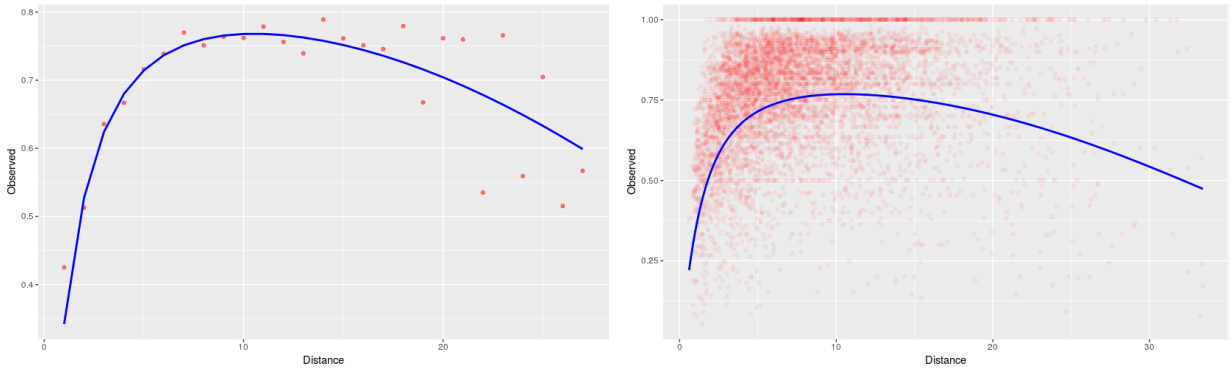

Figure 2: The model fit obtained for car useage fraction as a function of distance. The fit against the mean by distance band is shown on the left, and the fit against all data points is shown on the right.

Preparing the target variable: removing distance-dependence

From the outset one can foresee that out of our dataset one of the most important variables in predicting car dependency will be the distance between the origin (home MSOA) and the destination (workplace MSOA). However given our stated aim of better understanding the driving factors behind car dependency for policy decision-making purposes, distance is not something we can alter by policy and it is therefore of secondary interest, as mentioned in our definitioni of car dependency above. In order to prevent our models from focussing on distance dependence we would therefore like to remove the distance-dependence of the car fraction data as best as possible.

We do this by using a model to fit the distance dependence of the car fraction, ensuring it has a good fit to the data, and then subtracting from the observed car use fractions the car use predicted by the model. The result of this is a variable which has a much diminished distance-dependence, and it is this variable which we will then apply machine learning to model in the next stage.

The model which best fit how the car use fraction varies with distance turned out to be a logistic regression model fit on the mean car fraction by 1km distance band (i.e. a curve where the first point is the average car use fraction for distances of 0-1km, etc.) using the distance and the natural logarithm of the distance as input variables (Figure 2).

After subtractin the car fraction predicted by the model from the observed car fraction we obtain a variable which a much reduced dependence on distance (i.e. it is centred on zero and does not significantly increase or decrease with distance), as shown in Figure 3.

Now that we have approximately removed the dominant influence of the distance, we can proceed to use machine learning to tease out the patterns in the data from all the other more subtle variables.

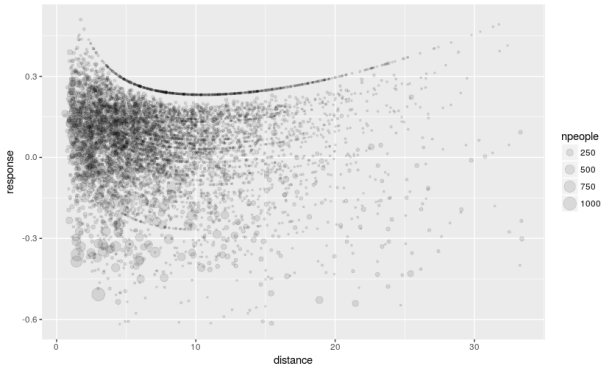

Figure 3: The variable obtained by subtracting predicted car fraction from observed car fraction.

Next steps

Based on the input data documented in the previous sections, we have already made a start on developing and running code to identify the variables associated with car dependency. To do this we will perform model selection of a wide range of machine learning regression models, checking their performance in prediction car dependency (car use fraction as defined above) on a new dataset.

Once the best-performing model has been found we will do some model validation to ensure the model gives sensible predictions on new data, and then proceed to attempt to extract which variables the model deems most important in predicting car dependency, in general, or if possible also for individual predictions.