Using Machine Learning and Big Data to Model Car Dependency: an Exploration Using Origin-Destination Data

Robin Lovelace, Ilan Fridman-Rojas and Liam Bolton

2017-06-01

Contents

1 Introduction **1**
1.1 Car dependency: motivations and definitions 2
1.2 Input data 3

2 Visualisation and traditional statistical analysis **6**

3 Machine Learning **8**

4 Discussion **9**
4.1 Policy Implications 10

5 Recommendations **10**

References **11**

1 Introduction

The primary purpose of this project is to show the power and potential benefits of machine learning for transport data analysis and policy development. 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 important to understand what we mean by the terms Big Data and Machine Learning, so some definitions are in order. We take a broad view of both. By Big Data we mean "unconventional datasets that are difficult to analyze using established methods" (Lovelace et al. 2016). Thus we are not necessarily talking about size. It's also about data that are unconventional in terms of "form, format, and complexity", making them difficult to analyse using a spreadsheet, for example. By machine learning, we mean simply that the functional form of the model is not specified by the user, but by an algorithm which uses trial and error

to find the most appropriate model. A model is simply a mathematical or computer-encoded representation of a phenomenon (in this case car dependency), so the techniques outlined in this report could apply to a wide range of transport (and non transport) issues: car dependency is a useful, and highly policy relevant, case study to illustrate what is now possible and potential future directions of travel.

This report summarises the work undertaken over the course of a 3 month project funded by the UK's Department for Transport (DfT) under the Transport Technology Research Innovation Grant (T-TRIG) stream. This funding stream is part of the Chief Scientist's office, which, among other things "helps develop the department's links with the wider science, engineering, technology and innovation community".

The report is self-standing but links to 3 Appendices, which can be consulted for futher details on the input data, methods and policy implications of the research, respectively. They are:

- Progress Report 1, which is focussed on the input data and data generation process and how official datasets were augmented with geographical data from OpenStreetMap.
- Progress Report 2, which describes the Machine Learning algorithms used and provides code snippets and results to show how they performed.
- A report, The Policy Implications of the research, which provides an up-to-date account of the state-of-the-art in terms of machine learning in public policy.
- A tutorial on Big Data and Machine Learning for transport data analysis which explains how to use some of the methods outlined in this report, targetted at transport planners, consultancies and policy-makers relatively new to data science.

1.1 Car dependency: motivations and definitions

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.

Car dependency should be easy to define: a measure of the extent to which people *depend* on their cars. But what does this actually mean? Unless your car doubles as a life-support system, it is unlikley that you literally depend on it for survival. More likely you depend on the car for trips that are *seen as essential*, for modern life. Whether it's for doing the 'weekly shop', dropping children off at school or taking a sick relative to hospital at an affordable, car use can easily become integral modern life.

What is actually meant by car dependency in the context of machine learning? There are many possible ways of defining and measuring car use. Anable (2005), for example, defines

car dependency in attitudinal terms The paper is also relevant methodologically, illustrating the potential benefits of applying cluster analysis, then a new method, to the question of 'car addicts' as the author puts it. 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).

The importance of the spatial element in car dependency was highlighted in a recent article by Clark and Rey (2017), who found that an area's likelihood of becoming or remaining car dependent was greatly increased if surrounding areas had high levels of car ownership. This work was made possible by recent methodological developments, illustrating how new approaches to existing methods, and data, can yield new policy insights. Before progressing to explore the methods, it is vital to understand the input data.

1.2 Input data

During the first stage of the project (2017-02-06 to 2017-03-20) the priority was to access and organise the data. Data preparation may not sound like exciting work, but it can consume a large proportion of analysts' time, so is worth taking seriously (Gillespie and Lovelace 2016; Grolemund and Wickham 2016).

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.

1.2.1 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`.

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 (Lovelace and Ellison 2017). 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.

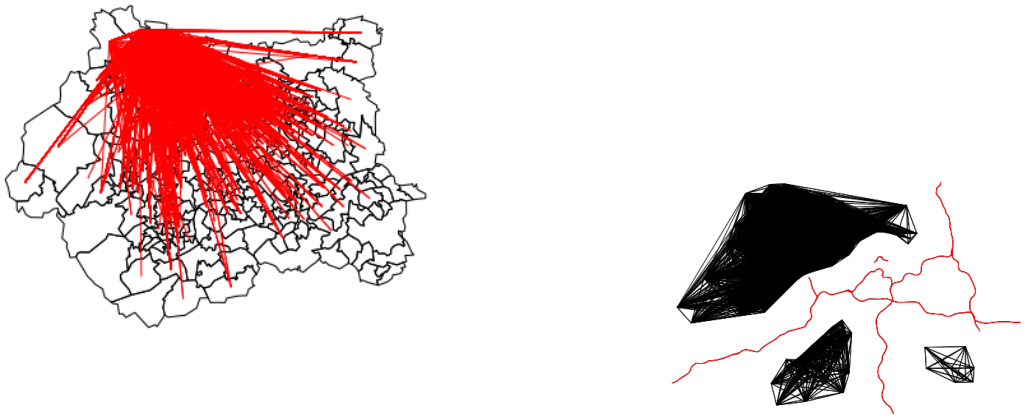

Figure 1: A sample of work commute flows in the West Yorkshire region (left), and a sample of the result of the code which calculates the distance between motorways (red) and flows (black), in this case discarding all flows within a certain distance of the motorway, as an example.

1.2.2 Socio-demographic data

The geodemographic data relevant to each origin (home) MSOA consists of the following variables: number of people in particular age brackets, number of 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 (see Appendix 1).

1.2.3 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.

The input data are described in more detail in Appendix 1.

2 Visualisation and traditional statistical analysis

Visualisation and descriptive statistics are tried and tested techniques for understanding data. The importance of gaining an overview of the data is heightened for large and complex datasets that will be analysed using machine learning, because errors can have large un-intended consequences, that may be hidden by the 'black box' nature of the algorithms.

For analysis aimed at informing locally targetted interventions, maps are an excellent starting point. Even though the data operate at a high geographical resolution (at the OD level), it is worth starting with visualisations at a low level of detail, to ensure that the data makes sense.

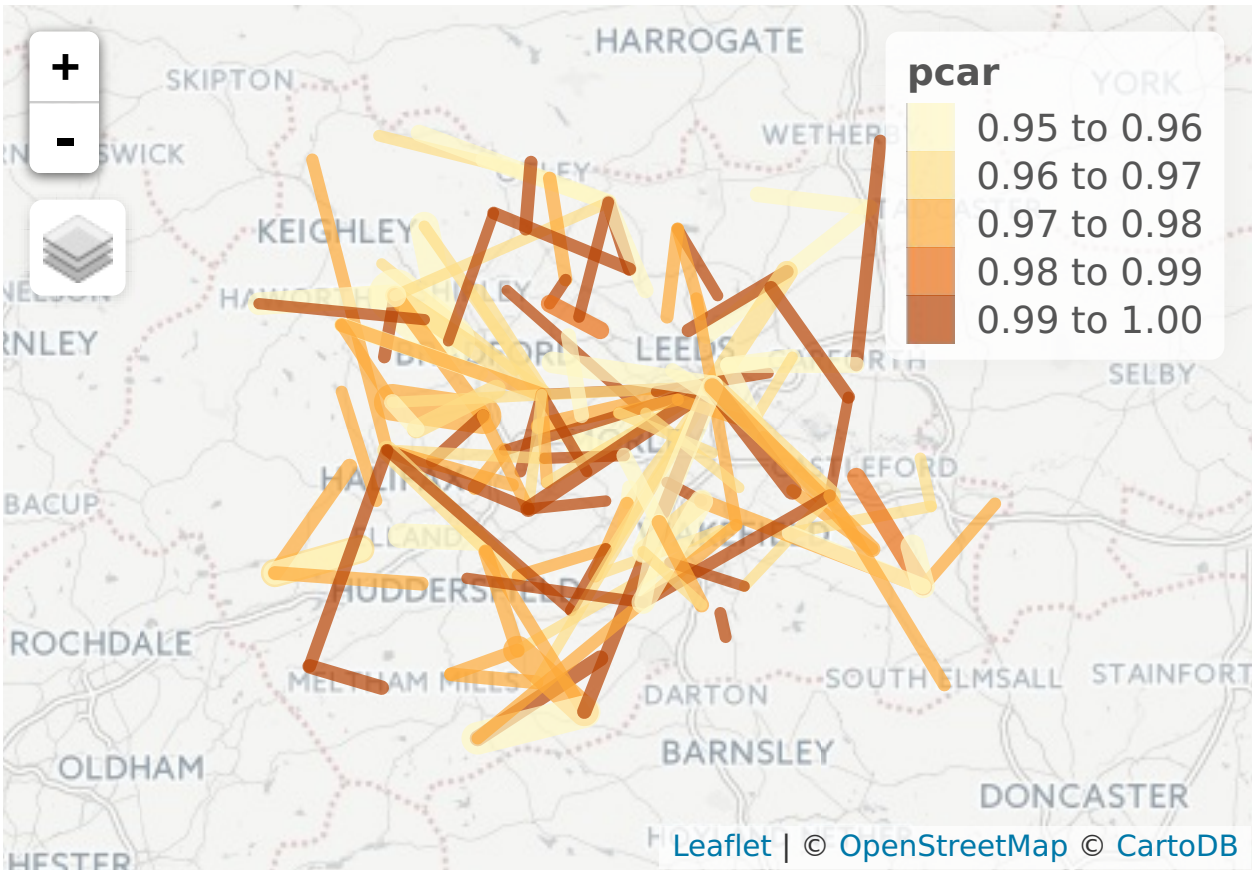

Figure 2: Flows in West Yorkshire in which a high proportion of trips (95%+) of trips are made by commuting, for flows along which which 30+ people overall commute.

Figure 2 shows the benefits of converting the data into a spatial format. It shows that car dependency is a problem across West Yorkshire, with particular issues in the suburbs between Leeds and Bradford and in rural areas such as Ripon in which there are comparatively few job opportunities.

Another way to visualise the same data is in terms of absolute numbers of people using the car (Figure 3).

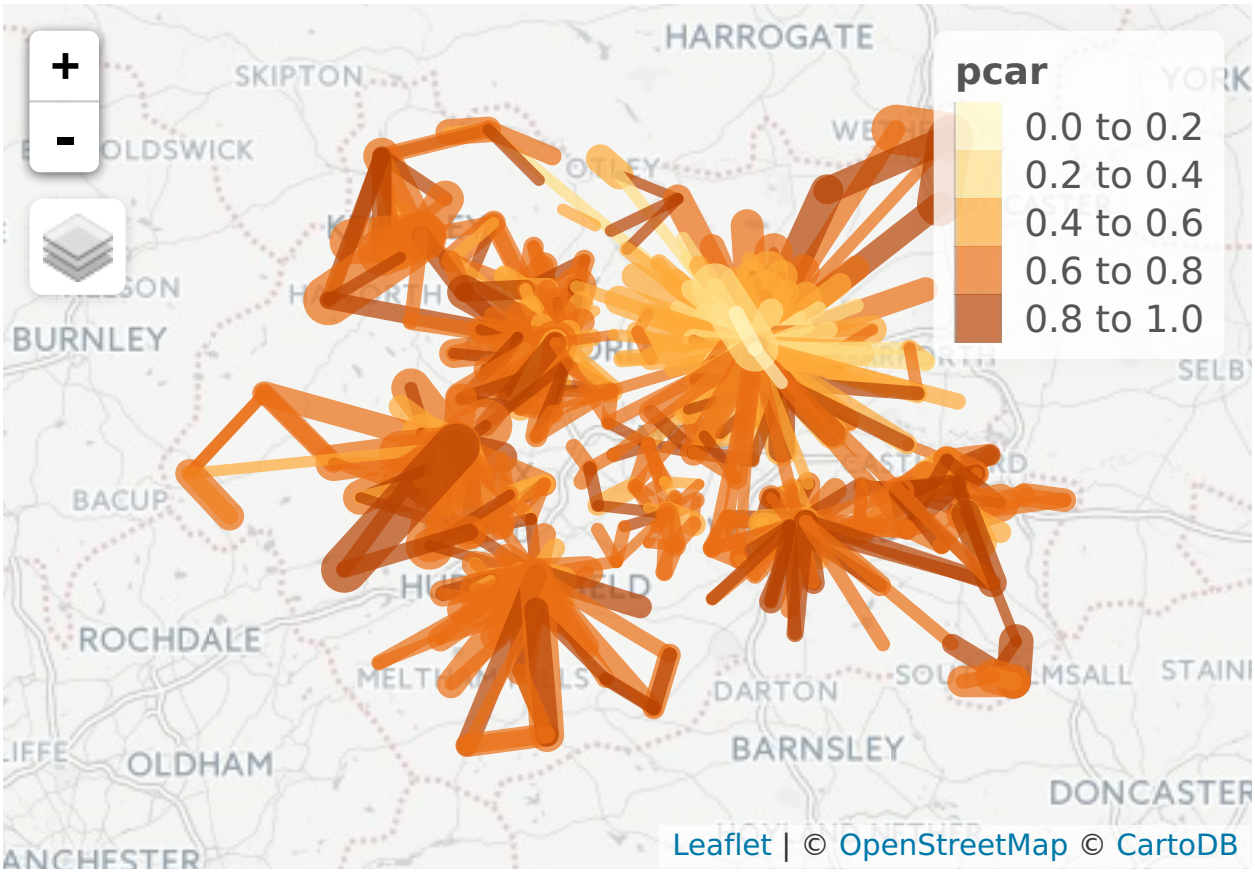

Figure 3: Visulisation of the most flows in West Yorkshire in which over 100 people travel to work by car. Red lines have a higher percentage travelling by car, wide lines have a higher overall number of people travelling by car.

Figures 2 and 3 demonstrate the importance of decisions around visualisation. Both plots are useful in that they provide a high level overview of the issue. However, despite our definition of car dependency in proportional terms, Figure 3 is probably more useful. It shows how small changes to visualisation options (in this case visualising only flows with a high absolute number of car drivers, > 100, and setting width proportional to the same variable) can have a great impact on how the data is interpretted.

Figure 3 shows that each settlement has car dependency issues. However, they are different in nature, with Leeds suffering from high car use primarily due to high numbers of commuters but with smaller settlements having both high numbers and high proportions of people driving to work. Another difference compared with Figure 2 is that Figure 3 clearly shows that in terms of absolute numbers, short car trips dominate: the average distance of flows in which 100+ people drive is less than 5km straight line distance.

Many linear regression models were run to explain the variability in the number of people driving to work. This code is presented in the file `vignettes/traditional-stats.Rmd` but not discussed further in this document as the focus of the project is on Machine Learning.[1] Suffice to say that the 'features' of greatest importance according to Machine Learning (particularly age) cooincided with the most statistically significant explanatory variables. This suggests that traditional statistical techniques should be used along-side machine learning techniques to ensure understanding of the data and cross-checking of results.

3 Machine Learning

Appendix 2 demonstrates and explains the Machine Learning methods that have been applied to the input data. The Python code underlying this work, and the results generated, are available in IPython notebooks (`.ipynb` files) called `Regression_toolkit.ipynb` and `Deep_MLP.ipynb`. These are stored in the Python folder of the mlCars GitHub repository. To ease accessibility, GitHub renders these files so they are human readable on its interface - see https://github.com/Robinlovelace/mlCars/blob/master/Python/Deep_MLP.ipynb an example of this.

The overall finding was that there is a very wide number of software pacakges, in both R and Python. Python was used for the majority of the Machine Learning work because the analyst working on the project had most experience in this field, but either can be used. For very large scale data, specialist software may be necessary, such as Theano (The Theano Development Team et al. 2016), Spark (Meng et al. 2016) or TensorFlow (Abadi et al. 2016). Fortunately, high level interfaces to such 'low level' Machine Learning programs have been developed in both R and Python, such as the **kerasR**, **tensorflow** and **sparklyr** packages.

A demonstration of the Theano approach is demonstrated in Appendix 2. While this proof-of-concept is useful, it did not provide results that were more useful than the simpler algorithms used in ealier sections of that report. In the context of the comparatively small dataset size

[1] Thanks to Rob Long for undertaking the statistical analysis using traditional linear regression techniques.

and simplicity of the problem (compared with machine translation and image recognition), using systems such as Theano on this particular tasks is probably overkill. Metaphorically speaking it's like using a machine gun to swat a fly.

On the other hand, simpler techniques such as random forest and boosted regression trees (as implemented in the package **xgboost**) are easy to set-up and run. Perhaps too easy. It is vital that users understand at least some of the theory underlying the code to ensure appropriate input parameter and interpretation. One advantage of **xgboost** is that it prevents 'overfitting' and deals well with co-linearity, highly correlated predictor variables.

One interesting and somewhat unexpected finding from the Machine Learning experience was that the results in terms of feature importance were sensitive to both the algorithm used and the set-up of the model run.

4 Discussion

We have demonstrated that a new generation of statistical tools are available, for free and packaged in mature open source software, for transport researchers.

The uptake of new technologies like big data and machine learning in the transport industry has been comparatively slow Department for Transport, 2016. The Transport Systems Catapult makes several recommendations for improving the use of data in the transport sector including: the establishment of a Policy Advisory Group; development of contract and licensing templates; a framework for data sharing; the continued publishing of open data; and more training for public sector professionals Transport Systems Catapult, 2017.

Maximising the value of machine learning in the UK transport industry requires addressing a number of key barriers. According to a report by the Transport Systems Catapult (2017) and the Open Data Institute (ODI), tackling these barriers and making better use of transport data could unlock £14 billion per annum for the UK by 2025. There is a sizeable data skills gap in the UK economy. The Royal Society argues that the disruptive potential of new technologies like machine learning is held back by a shortage of data skills. The Transport Systems Catapult also argues that the development of Intelligent Mobility in the UK is hindered by the lack of an effective skills strategy. The report suggests an increase in specialist apprenticeships and postgraduate degrees would narrow the skills gap. In addition, designing policies based on data-driven insights requires decision-makers have a foundational knowledge of statistics as well as the ethical challenges underpinning new data-driven technology. Failing to narrow the skills gap in Intelligent Mobility could lead to an estimated £50 billion loss in GDP per annum Transport Systems Catapult (2016). There there needs to be significant investment in the research and development landscape in order to assist the UK in becoming a leader in transport innovation Lovelace et al. (2015); Transport Systems Catapult (2016).

4.1 Policy Implications

As an inactive, highly polluting mode of transport, driving is associated with a range of health problems such as obesity, diabetes, coronary heart disease, asthma and lung cancer. In order to reduce car dependency, policy-makers and planners need to develop better infrastructure for walking and cycling. This could be achieved using open source planning systems like the Propensity to Cycle Tool (Lovelace et al. 2017). To illustrate, the West Yorkshire Combined Authority (WYCA) is improving walking and cycling networks as well as investing in new infrastructure and redesigning the layout of streets to reduce conflict between drivers, cyclists and pedestrians. Rather than designing cities around the car, land use and transport planners should aim to reduce car dependency. This research has found that car dependency in West Yorkshire is higher near motorways. Building new services and housing close to dense, existing centres rather than on the periphery would encourage shorter journeys Newman et al. (2016). In addition, increasing density and improving public transport would encourage multi-modal mobility over car use (Newman and Kenworthy, 1989; Newman et al., 2016; Bertolini and Le Clercq, 2002). This would have the added benefit of reducing urban sprawl in peripheral zones. The introduction of road pricing schemes, pioneered by London and the Congestion Charge, could reduce car dependency in designated parts of West Yorkshire. Improvements to transport planning should be supplemented by 'soft' interventions such as promotional campaigns and 'nudging' towards walking, cycling, public transport and car sharing. For example, 'soft' interventions could be aimed at more car dependent socio-economic groups. Implementing 'soft' and 'hard' measures to reduce car dependency would go a long way towards developing multi-modal mobility in the region. In many respects, the WYCA is at the forefront of sustainable transport innovation. By championing 'One System Public Transport' and prioritising the environment, health and wellbeing as part of its transport strategy, the WYCA is leading the way in the development of sustainable, multi-modal mobility in the UK. The WYCA is also distinguished by its commitment to implementing 'Smart 'Futures' as part of their strategy for future sustainable transport. However, implementing the vision of a more integrated, sustainable transport system means overcoming barriers such as institutional conditions, political factors and market demand Forward et al., 2014. This report ultimately argues that building a more sustainable transport system by reducing car dependency should be a priority for policy-makers.

5 Recommendations

Based on our work we offer the following tentative recommendations to researchers in the field:

- Pay attention to the vital stage of data pre-processing. Mistakes made here can have knock-on consequences throughout the project. As with any modelling work, rubbish in will ensure that rubbish comes out so it is important to spend time visualising and statistically summarising the input data before proceeding to the modelling stage.
- Look for opportunities to augment official datasets with novel data sources. An example

from our work was the addition of proximity to motorways using OpenStreetMap data as an explanatory variables.

- Making transport data geographical by pre-processing (as illustrated in our conversion of OD data to geographic lines) can add value to the analysis and allow locally specific knowledge be generated.
- Before moving to Machine Learning algorithms such as **xgboost** ensure that traditional statistical models have been run and interpretted. This mitigates against the problem than machine learning models are 'black box' and provides a cross-check to ensure that the results make sense.
- Visualise the data in new ways and interactively. By allowing policy makers and the public to play with the data, new insights can be gleaned. A transparent, interactive visualisation can be more valuable for generating local insight than the results of a complex machine learning method.
- Take care not to allow machine learning methods to hide the raw data or lead to confusing English that risks people being 'blinded by science'.
- Now that the best machine learning algorithms are open source and can be scripted in code, there is no excuse for results not to be reproducible. **All publicly funded transport models using machine learning should be open access, with code published alongside results, and reproducible.**
- Machine learning and big data present new opportunities for the transport planners and policy-makers. However, they should not supplant domain expertise or contextual knowledge in transport policy and planning. Machine learning and big data are best used as aides to human decision-making rather than as replacements.
- Barriers to the uptake of innovative new technologies in the transport industry include: access to skills, organisational culture and investment in R&D.

References

Abadi, Martín, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, et al. 2016. "TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems." *ArXiv:1603.04467 [Cs]*, March. http://arxiv.org/abs/1603.04467.

Anable, Jillian. 2005. "'Complacent Car Addicts' or 'Aspiring Environmentalists'? Identifying Travel Behaviour Segments Using Attitude Theory." *Transport Policy* 12 (1): 65–78. doi:10.1016/j.tranpol.2004.11.004.

Clark, Stephen D., and Sergio Rey. 2017. "Temporal Dynamics in Local Vehicle Ownership for Great Britain." *Journal of Transport Geography* 62 (June): 30–37. doi:10.1016/j.jtrangeo.2017.05.007.

Gillespie, Colin, and Robin Lovelace. 2016. *Efficient R Programming: A Practical Guide to*

Smarter Programming. O'Reilly Media. https://csgillespie.github.io/efficientR/.

Grolemund, Garrett, and Hadley Wickham. 2016. *R for Data Science*. 1 edition. O'Reilly Media.

Lovelace, Robin, and Richard Ellison. 2017. "Stplanr: A Package for Transport Planning." *Under Review*. https://github.com/ropensci/stplanr.

Lovelace, Robin, Mark Birkin, Philip Cross, and Martin Clarke. 2016. "From Big Noise to Big Data: Toward the Verification of Large Data Sets for Understanding Regional Retail Flows." *Geographical Analysis* 48 (1): 59–81. doi:10.1111/gean.12081.

Lovelace, Robin, Anna Goodman, Rachel Aldred, Nikolai Berkoff, Ali Abbas, and James Woodcock. 2017. "The Propensity to Cycle Tool: An Open Source Online System for Sustainable Transport Planning." *Journal of Transport and Land Use* 10 (1). doi:10.5198/jtlu.2016.862.

Meng, Xiangrui, Joseph Bradley, Burak Yavuz, Evan Sparks, Shivaram Venkataraman, Davies Liu, Jeremy Freeman, et al. 2016. "Mllib: Machine Learning in Apache Spark." *Journal of Machine Learning Research* 17 (34): 1–7. http://www.jmlr.org/papers/volume17/15-237/15-237.pdf.

The Theano Development Team, R. Al-Rfou, G. Alain, A. Almahairi, C. Angermueller, D. Bahdanau, N. Ballas, et al. 2016. "Theano: A Python Framework for Fast Computation of Mathematical Expressions." *ArXiv E-Prints*, May.