Draft – Policy Report + Snippets

Liam Bolton

Table of Contents

Executive Summary

- Machine learning and big data present new opportunities for the transport industry. This report demonstrates how big data and machine learning can be used to model car dependency, which also has significant implications for sustainable transport policy and planning
- Big data and machine learning offer the promise of more efficient transport models based on rich, finely-grained data across a wider range of indicators
- Barriers to the uptake of innovative new technologies in transport include: access to skills, investment in R&D and organisational culture
- Machine learning and big data should not supplant domain expertise or contextual knowledge in transport policy and planning. They are best used as aides to human decision-making rather than as replacements
- Car dependency is the cause of a variety of social, environmental and economic issues. Creating more sustainable regions by reducing car dependency requires a multi-faceted approach that involves both 'hard' and 'soft' policy interventions

Transport modelling is rooted to methods that are highly inefficient. The ability to model travel behaviour and mode choice, for example, is reliant on data sources that are inaccurate and costly to produce. As such, models that rely purely on traditional sources such as census data are resource-intensive, often taking many years to complete. This policy report presents an alternative methodology, one that draws on innovative advances in machine learning and big data to model car dependency in West Yorkshire, UK. Big data and machine learning offer transport modellers and decision-makers an unparalleled opportunity to analyse the driving forces behind higher car ownership using rich, finely-grained information across a range of variables. The aims of this accessible and brief policy report are two-fold. First, it will

provide a practical analysis of the promise and pitfalls of big data and machine learning for the transport sector and policy-making. Second, this report will explore the policy implications of car dependency including how we, as practitioners and researchers, might address the barriers to equitable mobility and create a more sustainable transport system.

First, it is important to provide a very brief overview of how machine learning and big data have been used in this study. Three types of input data – origin-destination data, socio-demographic data and spatial data – were used to explore the driving forces behind car dependency in West Yorkshire. In this context, car dependency is defined as the proportion of people in a zone who use a car or van as their main mode of transport relative to distance. The origin-destination data, collected from the 2011 census, was used to determine commuter flows between Middle-Super Output Areas (MSOAs). The geographically aggregated socio-demographic data, which contains a wide variety of variables, was also sourced from the 2011 census. The spatial data was derived from the OpenStreetMap platform and the National Public Transport Access Nodes (NaPTAN) dataset. This contains the location of motorways, train stations, coach stations and bus stops. Applying machine learning models to the input data allowed for this study to analyse the variables behind car dependency. Given that this is predominantly a standard regression problem to determine transport behaviour, the coefficient of determination was chosen as the metric of interest. The MLP and XGBoost models were found to be the best performing and most useful models. Assessing causal relationships was carried out using Directed Acrylic Graphs (DAG) (see Figure 1). The models were principally used to determine variable importance, which lends itself toward policy-making aimed at reducing car dependency.

1. Big Data and Machine Learning: A New Paradigm for Transport

1.1 Big Data

Big data represents a new paradigm for transport and decision-making. We are witnessing the rise of new and innovative technologies that are disrupting the transport industry in the United Kingdom. Big data and machine learning are the vanguard of these rapid advances.

Transport modelling has traditionally relied solely on survey data, which involves a very small sample of the population being asked a set of questions to determine mobility. This method is resource-intensive. Surveys are costly to produce, often taking many years to come to fruition. Transport models would benefit from individual-level data about day-to-day transport behaviour. Big data offers the chance to build more accurate and efficient models. This study demonstrates how big data and machine learning can be used to create more effective models of car dependency in West Yorkshire. On a broader note, big data and machine learning can be utilised to generate valuable insights about carbon dioxide emissions and mobility flows, for example.

Big data is characterised by high volume, variety and velocity. This report draws on big data in the form of spatial data from OpenStreetMap, NaPTAN and the 2011 census. In terms of variety, big data allows for a richer analysis across wide range of social, demographic and spatial variables. The velocity of big data lends itself to faster decision-making based on more relevant data (Batty, 2013; Bettencourt, 2013). In addition, big data is high in volume. However, as Lovelace et al. (2015) argue, spatial modelling brings another factor into play: veracity. Indeed, unlike traditional data sources, big data feeds into models that are more representative of a population than traditional data sources such as surveys. On the other hand, as Boyd and Crawford (2012) argue, big data is not always better data.

Until relatively recently, the value of big data in the transport industry was limited mainly due to its size and the availability of inexpensive technology. With the advent of new tools – including R, an open source statistical programming language used extensively in this study and throughout organisations like the Department for Transport and the RAC Foundation – practitioners and researchers can now store and analyse vast quantities of transport data. As Lovelace et al. (2015) and Anda et al. (2016) demonstrate, this could be extended to include other big data sources such as mobile phones, smart cards and social media.

The rise of open data and shared data has also enabled the development of more transparent and accessible transport tools and models. This research draws exclusively on datasets that are free, publicly-available and 'big'. OpenStreetMap, for instance, is helping transport researchers and practitioners develop tools that are powered by open data (Semanjski et al., 2016). One of the primary outputs of this study is an open source, accessible R package that can be utilised by transport planners and policy-makers to perform basic modelling functions. However, the value of open, big data resides in the combination of multiple datasets from different sources. We show that census data can complement data generated by OpenStreetMap and NaPTAN. Linked data is advantageous in the sense that it transcends the weaknesses of any single dataset. Combining multiple datasets leads to more accurate transport models and better decision-making.

Big, open and linked data can be utilised to develop a better understanding of mobility. The abundance of rich, finely-grained data presents new opportunities for transport modellers and policy-makers (Cottrill and Derrible, 2015; De Gennaro et al., 2016; Semanjski et al., 2016). Big, open and linked datasets can lead to better models and support more effective planning and decision-making (Semanjski et al., 2016). This research shows how an innovative approach to modelling car dependency - driven by big data and machine learning - could be used to develop a more sustainable, equitable transport system. Demonstrating the value of big data and machine learning while highlighting some of the key barriers and risks for the transport sector could go some way towards addressing these challenges.

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 (2017) 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. In the following section, this report identifies several other barriers to the effective use of emerging technologies like machine learning in the transport industry.

1.2 Machine Learning

Machine learning is powering some of the most exciting developments in the transport sector (Budka and Salvador, 2017). From autonomous vehicles to smart cities and Mobility as a Service (MaaS), machine learning is at the forefront of transport innovation. This study demonstrates how machine learning can be used to model the driving forces behind car dependency in West Yorkshire. In other use cases, Clarke et al. (1998) use decision trees to explore the extent to which behavioural factors contribute to road accidents and Omrani (2015) employs machine learning to determine the mode choice of individuals. In addition, Mattioli (2016) uses a sequence pattern mining study of UK time use data on car dependence. Machine learning is being used to track congestion using intelligent video surveillance systems. Autonomous vehicles – which may have a role to play in reducing congestion by increasing the space available for new cycling and walking infrastructure (WYCA, 2017) – are powered by machine learning algorithms.

Machine learning is most valuable to policy-makers as a means to find correlations between variables (Athey, 2017). Indeed, this study is innovative in the sense that it draws on machine learning to determine the correlations between car dependency and a wide variety of socio-demographic and spatial variables. We use machine learning to determine the importance of variables. This has significant implications for transport policy-making and planning. Machine learning helps us to answer questions like, what social factors contribute to car dependency in West Yorkshire and what effect does infrastructure, such as motorways and bus stops, have on car-dependent areas? In the next section, the results of our models will be used to explore such questions as, how do we encourage higher-represented groups to take up more sustainable modes and what other interventions can be made to develop a more equitable transport system? It is important to point out, however, that machine learning is most valuable as an aide to human decision-making (Walport and Sedwill, 2016). Machine learning cannot perform causal inference and should not supplant domain knowledge or contextual expertise in public policy-making. Grimmer (2015) and Athey (2017) argue that machine learning is most useful to policy-makers in conjunction with causal inference.

Decision-makers need to evaluate a range of ethical and operational issues – including privacy, data quality, transparency and evaluating social impact - before applying machine learning and big data to policy problems (Kleinberg et al., 2016). For one, Boyd and Crawford (2012) contend that big data is not always better data: big data can be highly unstructured and noisy (Cabinet Office, 2016). O'Neil (2016) provides a powerful account of the potential for machine learning to reinforce disparities based on socio-economic background. It is therefore imperative to adhere to strict ethical guidelines including effective privacy and data standards (Royal

Society, 2017). Improving the transparency of machine learning algorithms is another way of addressing ethical issues. This report is accompanied by open source code that reveals how machine learning – including a deep learning model – have been used to model car dependency. In addition, as the Royal Society (2017, p.8) states, “continued efforts are needed in a wave of ‘open data for machine learning’ by Government to enhance the availability and usability of public sector data”. Although transparency is not always possible, making algorithms open can be useful for designing public policies (Walport and Sedwill, 2016). It is beyond the scope of this study to explore these challenges in more depth with respect to the transport industry.

Maximising the value of machine learning in the UK transport industry requires addressing a number of key barriers. According to a recent report by the Transport Systems Catapult 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. The report focuses on access to skills, investment in the research and development landscape and multidisciplinary collaboration. There is a sizeable data skills gap in the UK economy. In a recent report, the Royal Society (2017) argues that the disruptive potential of new technologies like machine learning is hindered by a shortage of data skills. The Transport Systems Catapult (2016) 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). Fostering a culture of multidisciplinary collaboration could go some way towards addressing barriers involving data quality, skills and access. Finally, as Lovelace et al. (2015) and the Transport Systems Catapult (2016) state, there needs to be significant investment in the research and development landscape in to assist the UK in becoming a leader in transport innovation.

2. Car Dependency: Implications for Sustainable Transport Policy and Planning

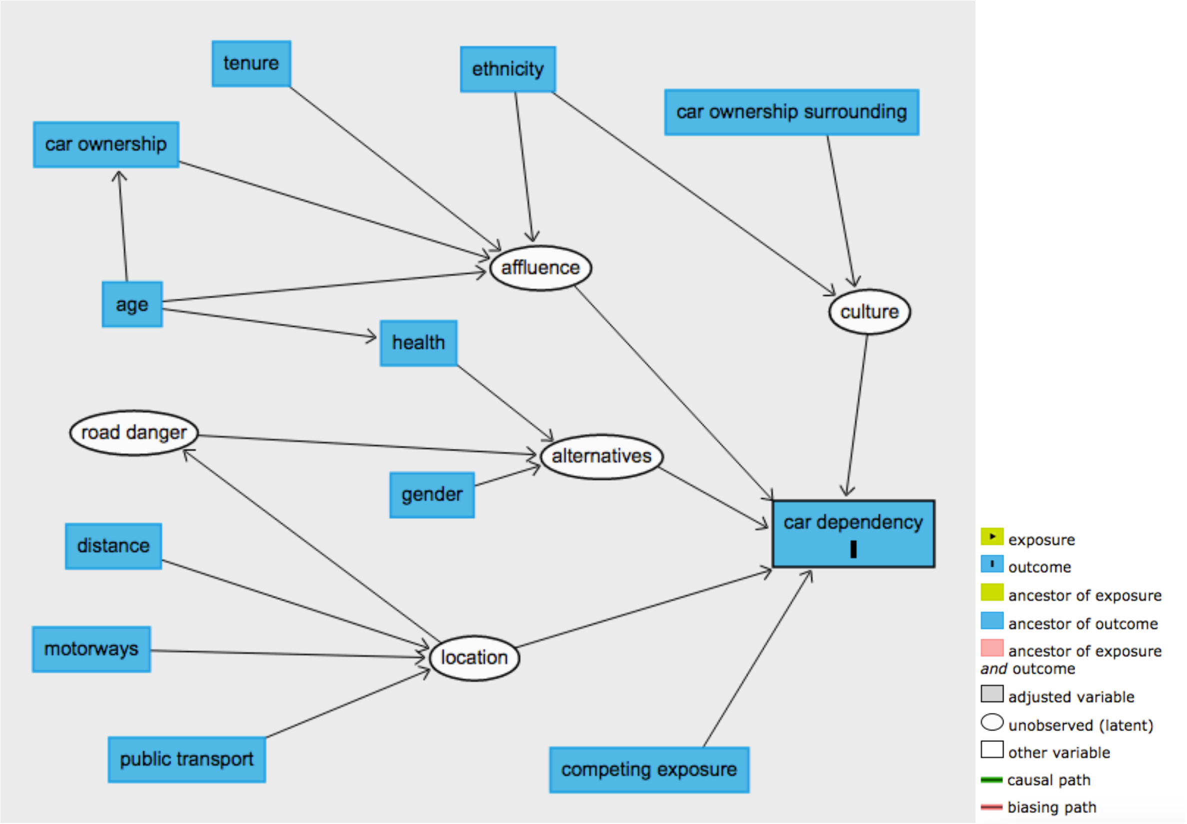
2.1 What is Car Dependency?

First coined by Newman and Kenworthy (1989), 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 into account the distance of trips”. In terms of distance, 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 use a car for trips of 20 miles. The meaning of the concept is highly contested, however.

2.2 What are the Driving Forces Behind Car Dependency?

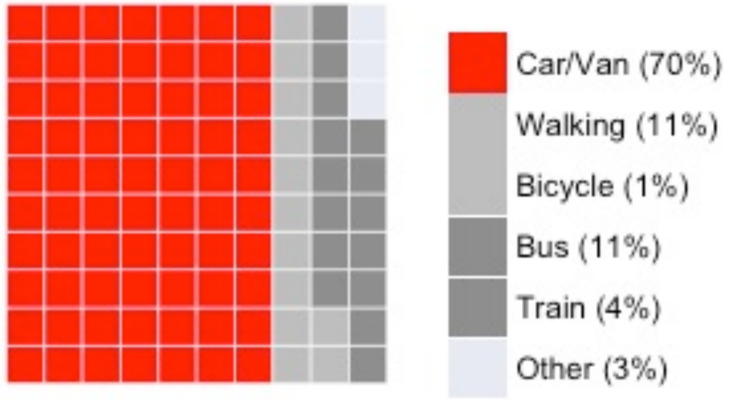
Car dependency is connected to a range of socio-economic, demographic, transport and land use variables (Newman and Kenworthy, 1989; Potoglou and Kanaroglou, 2008; McIntosh et al., 2014; Newman et al., 2016; Wiersma et al., 2016). While our models yielded different results in terms of variable importance, the Random Forest model finds that car dependency in West Yorkshire is strongly associated with the age band 35-49, followed by the age band 50-64 and males. This research finds that the motorway and coach stations play an important role in car dependency, which has implications for future land use planning. Directed Acrylic Graphs (DAG), a tool for visualising causal relationships between variables, was used to show the dynamic interplay between car dependency and the driving factors behind it (see Figure 1).

Figure 1. Directed Acrylic Graph (DAG)



The West Yorkshire Combined Authority (WYCA) Transport Strategy 2016-2036 found that car use was comparatively higher across a number of key socio-economic variables. The results from the WYCA Transport Strategy are broadly consistent with the results from this research. The WYCA compared car use with bus use and train use, finding that there was a significant difference between the modes. Using machine learning and big data to model cycling, walking and public transport use in West Yorkshire is a potential avenue for future research. To further illustrate, according to the 2011 census approximately 70% of commutes to work are made by car (see Figure 2). This is followed by 11% bus and 11% walking.

Figure 2. Work Commutes in West Yorkshire by Mode

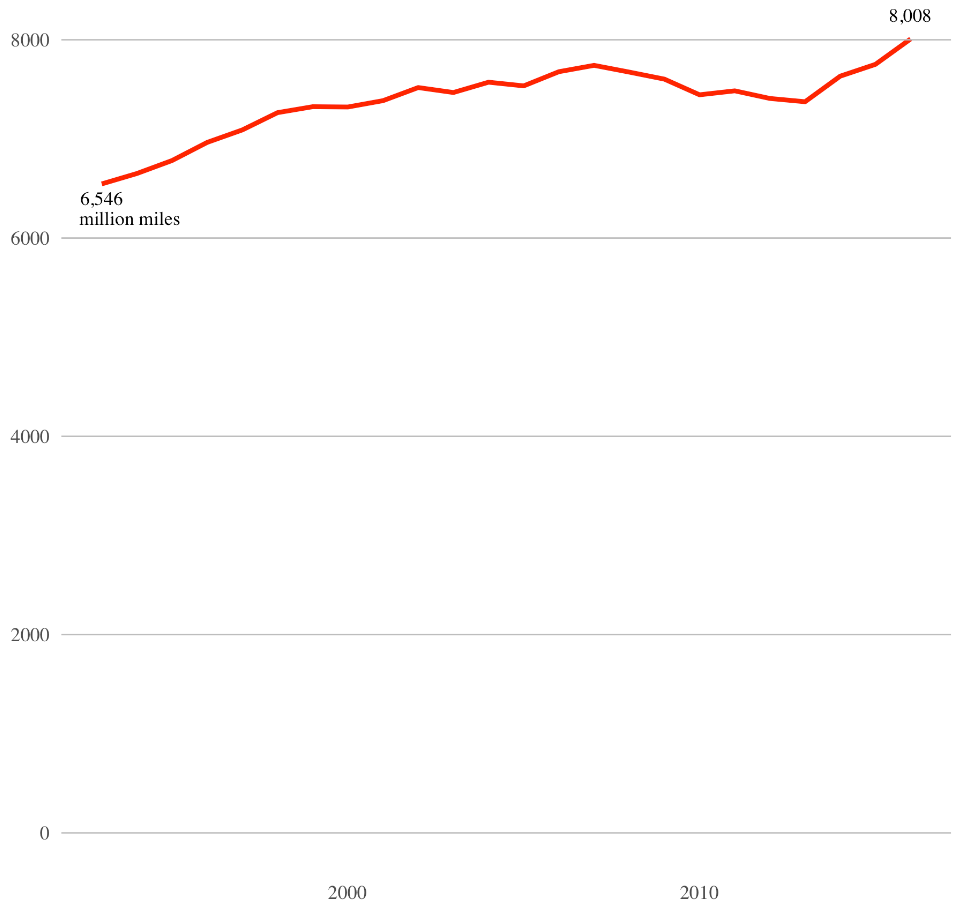


Potoglou and Kanaroglou (2008) model car dependency across a number of land use and socio-economic variables, ultimately finding that urban form and household socio-economic factors are strong determinants of car dependency in Hamilton, Canada. For example, urban sprawl was strongly associated with automobile ownership in Hamilton (Potoglou and Kanaroglou, 2008). Meanwhile, Wiersma et al. (2016) focus on the spatial factors underpinning car dependency in the Netherlands. Wiersma et al. (2016) argue that the use of cycling and public transport over cars is heavily dependent on spatial factors. Building on Newman and Kenworthy's (1989) analysis of car dependency in global cities, McIntosh et al. (2014) stress the importance of density and public transport. For Newman and Kenworthy (1989) and McIntosh et al. (2014), policy aimed at reducing car dependency should be geared toward land use and public transit interventions – including building up mixed use density and improving urban transit – rather than socio-economic, demographic and cultural factors. This report elects to take a broader perspective, one that draws on innovative methods of data analysis and collection. The next section will offer a very brief, broad and selective exploration of the policy implications of car dependency including how we, as transport practitioners and researchers, might develop a more sustainable transport system for the future.

2.3 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. This is particularly important given that car vehicle traffic has increased in West Yorkshire from 6,546 million vehicle miles in 1993 to 8,008 million miles in 2016 (see Figure 3). 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; Bertolini and Le Clercq, 2002; Newman et al., 2016). This would have the added benefit of reducing urban sprawl in peripheral zones. 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 the prioritising of 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. 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). Policy-makers would do well to remove the barriers to Intelligent Mobility and smart cities, for example, as they may have an important role to play in the development of sustainable transport in the UK (Transport Systems Catapult, 2016). This report ultimately argues that building a more sustainable transport system by reducing car dependency should be a priority for policy-makers.

Figure 3. Car Vehicle Traffic in West Yorkshire, 1993-2016



3. Conclusion

Machine learning and big data present new opportunities for policy-makers and planners to model car dependency and develop sustainable transport. This report has provided a brief exploration of the policy implications of car dependency. Car dependency is the cause of a variety of social, environmental and economic issues. A critical analysis of the promise and pitfalls of machine learning and big data has also been outlined. Machine learning and big data have an important role to play in the future of the transport industry in the UK. However, a variety of barriers – including access to skills, investment in R&D and organisational culture – prevent the full value of new technology from being realised.

Bibliography

Anda, C., Fourie, P. and Erath, A. 2017. Transport Modelling in the Age of Big Data. *International Journal of Urban Sciences*. **21**(1), pp.19-42.

Athey, S. 2017. Beyond Prediction: Using Big Data for Policy Problems. *Science*. **355**(6324), pp. 483-485.

Batty, M. 2013. Big Data, Smart Cities and City Planning. *Dialogues in Human Geography*. **3**(3), pp. 274-279.

Bettencourt, L.M.A. 2013. *The Uses of Big Data in Cities*. Online. Sante Fe: Sante Fe Institute. Available from: http://samoa.santafe.edu/media/workingpapers/13-09-029.pdf

Boyd, D. and Crawford, K. 2012. Critical Questions for Big Data: Provocations for a Cultural, Technological and Scholary Phenomenon. *Information, Communication and Society.* **15**(5), pp. 662-679.

Budka, M. and Salvador, M.M. 2017. Harvesting Big Data Could Bring About the Next Transport Revolution. *The Conversation*. Online. 11 May. Available from: http://theconversation.com/harvesting-big-data-could-bring-about-the-next-transport-revolution-right-now-77261?utm_campaign=Echobox&utm_medium=Social&utm_source=Twitter#link_time=1494512851

Cottrill, C.D. and Derrible, S. 2015. Leveraging Big Data for the Development of Transport Sustainability Indicators. *Journal of Urban Technology*. **22**(1), pp.45-64.

De Gennaro, M., Paffums, E. and Martini, G. 2016. Big Data for Supporting Low Carbon Road Transport Policies in Europe: Applications, Challenges and Opportunities. *Big Data Research*. **6**, pp.11-25.

Department for Transport, 2016. *Transport Infrastructure Skills Strategy: Building Sustainable Skills*. Online. London: Department for Transport. Available from: https://www.gov.uk/government/publications/transport-infrastructure-skills-strategy-building-sustainable-skills

Forward, S. 2014. *Challenges and Barriers for a Sustainable Transport System: State of the Art Report*. Online. Koln: Transforum. Available from: http://www.rupprecht-consult.eu/uploads/tx_rupprecht/TRANSFORuM_D4-2_APPROVED_Challenges-and-barriers.pdf

Grimmer, J. 2015. We Are All Social Scientists Now: How Big Data, Machine Learning and Causal Inference Work Together. *Political Science and Politics*. **48**(1), pp. 80-83.

Kleinberg, J., Ludwig, J. and Mullainathan, S. 2016. A Guide to Solving Social Problems with Machine Learning. *Harvard Business Review*. Online. 8 December. Available from: https://hbr.org/2016/12/a-guide-to-solving-social-problems-with-machine-learning?utm_source=twitter&utm_medium=social&utm_campaign=harvardbiz

Lovelace, R., Birkin, M., Cross, P. and Clarke, M. 2015. From Big Noise to Big Data: Toward the Verification of Large Datasets for Understanding Regional Retail Flows. *Geography Analysis*. **48**(1), pp.59-81.

Lovelace, R., Goodman, A. and Alfred, R. 2017. The Propensity to Cycle Tool: An Open Source Online System for Sustainable Transport Planning. *Journal of Transport and Land Use*. **10**(1), pp.502-528.

Mattioli, G., Anable, J. and Vrotsou, K. 2016. Car Dependent Practices: Findings from a Sequence Pattern Mining Study of UK Time Use Data. *Transporation Research Part A: Policy and Practice*. **89**, pp.56-72.

McIntosh, J., Trubka, R., Kenworthy, J. and Newman, P. 2014. The Role of Urban Form and Transit in City Car Dependence: Analysis of 26 Global Cities from 1960-2000. *Transport Research Part D: Transport and Environment*. **33**, pp.95-110.

Newman, P., Kosonen, L. and Kenworthy, J. 2016. Theory of Urban Fabrics: Planning the Walking, Transit/Public Transport and Automobile/Motocar Cities for Reduced Car Dependency. *The Town Planning Review*. **87**(4), pp.429-458.

Newman, P.G. and Kenworthy, J.R. 1989. *Cities and Automobile Dependency: An International Sourcebook*. Brookfield: Gower Publishing.

O'Neil, C. 2016. *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. London: Allen Lane.

Omrani, H. 2015. Predicting Travel Mode of Individuals by Machine Learning. *Transportation Research Procedia*. **10**, pp.840-849.

Potoglou, D. and Kanaroglou, P.S. 2008. Modelling Car Ownership in Urban Areas: A Case Study of Hamilton, Canda. *Journal of Transport Geography*. **16**(1), pp.1-26.

Royal Society. 2017. *Machine Learning: The Power and Promise of Computers That Learn By Example*. Online. London: Royal Society. Available from: https://royalsociety.org/~/media/policy/projects/machine-learning/publications/machine-learning-report.pdf

Semanjski, T., Bellens, R., Gautama, S. and Witlox, F. 2016. Integrating Big Data into a Sustainable Mobility Policy 2.0 Planning Support System. *Sustainability*. **8**(11), pp.1-19.

Transport Systems Catapult. 2017. *The Case For Government Involvement to Incentivise Data Sharing in the UK Intelligent Mobility Sector*. Online. London: Transport Systems Catapult. Available from: https://s3-eu-west-1.amazonaws.com/media.ts.catapult/wp-content/uploads/2017/04/12092544/15460-TSC-Q1-Report-Document-Suite-single-pages.pdf

Walport, M. and Sedwill, M. 2016. *Artificial Intelligence: Opportunities and Implications for the Future of Decision Making*. Online. London: Government Office for Science. Available from: https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/566075/gs-16-19-artificial-intelligence-ai-report.pdf

West Yorkshire Combined Authority. 2017. *Transport Strategy 2016-2036*: Our Policies. Online. Leeds: WYCA. Available from: http://www.westyorks-ca.gov.uk/uploadedFiles/Content/Transport/Transport_Plan/Transport%20Main%20part%202%20reduced.pdf

Wiersma, J., Bertolini, L. and Straatemeier, T. 2016. How does the Spatial Context Shape Conditions for Car Dependency? An Analysis of the Differences Between and Within Regions in the Netherlands. *The Journal of Transport and Land Use*. **9**(3), pp.35-55.