1. INTRODUCTION

In January 2005, the Vehicle and Operator Services Agency (VOSA, now the DVSA) introduced a computerised system for reporting the MOT (roadworthiness) test results, which was fully implemented by the beginning of April 2006. In November 2010, an anonymised version of the dataset was made publicly available, and there have been subsequent data releases since that time.

Between April and June 2011, members of this project team undertook a scoping study (research grant EP/J004758/1) to investigate potential uses of the data in relation to transport, energy and climate change, and to develop techniques for manipulating and analysing the data (see Cairns et al, 2014). This led to the award of the current EPSRC project (EP/K000438/1), which began in October 2012, and will run to the end of 2016. This project is formally supported by the UK Department for Transport (DfT) and the UK Department of Energy and Climate Change (DECC). It is led by the University of Leeds, and involves the Transport Research Laboratory, University of the West of England, University of Bristol, UCL and, initially, the University of Aberdeen (www.MOTproject.net).

The scoping study identified the value of obtaining additional information from vehicle licensing statistics collected by the Driver & Vehicle Licensing Agency (DVLA), which could be used in conjunction with the MOT data to generate an enhanced dataset. After a successful data application to DVLA, an enhanced
version of the dataset was received by the project team in July 2015, with a number of further updates since then.

The primary aim of the work is to explore the potential for this dataset to inform our understanding of car ownership and use, and related issues such as energy consumption, air pollution and climate change emissions from vehicle use.

2. DETAILS OF THE CORE DATASET

The MOT records, available from the DVSA, consist of data collected during the annual vehicle inspection tests for all vehicles (up to 3.5 tonnes) more than three years old in Great Britain (i.e. excluding Northern Ireland). About 35 million tests are recorded every year, meaning that over 300 million records are being analysed for this project. In the raw dataset, data are organised by test, and include details of the nature of the vehicle (date of first use, make, model, colour and engine size), an odometer/mileage reading, and details of the test (date, the postcode area where the testing station was located and test results (pass, fail, test refused or abandoned). There are related tables providing more details of test failures, although those have not yet been utilised in this project, (DfT, 2014).

The vehicle licensing data from the DVLA is, instead, organised by vehicle, and includes details of about 55 million vehicles from 2003 onwards. Unlike the MOT records (where test dates are vehicle dependent), licensing data are reported at regular, quarterly intervals. For this project, key details provided (which are not available from the MOT records) include the location of the registered vehicle keeper (by small area census unit ~700 households); an indication of whether the vehicle is in personal or commercial ownership; and the vehicle’s CO₂ rating (according to the manufacturer).

Key activities which have been undertaken to generate usable information from these data sources include:

- Hosting the data appropriately, specifically on a stand-alone machine held in a secure environment, with access via a virtual private network, and downloads only permitted after independent verification that they do not breach the data protection requirements.
- Fusing and re-organising the data into a master table, which is organised by vehicle, but includes quarterly information for fields that change over time.
- Reconciling inconsistencies between data gathered at different MOT tests that should theoretically be unchanging (such as vehicle engine size).
- Generating an estimate of the mileage travelled by each vehicle for each quarter.
3. GENERATION OF MILEAGE ESTIMATES

The generation of mileage estimates for individual vehicles has been a major analytical task of the project, and is reported in more detail elsewhere (see, for example, early work on this topic by Wilson et al. 2013a, 2013b, Cairns et al 2014a). Particular challenges include the fact that although tests occur on a roughly annual basis, the intervals between tests vary; there are erroneous readings recorded (due to both accidental and deliberate error); there will be variation in vehicle use between test intervals (not least due to changes in the weather and associated activities - e.g. the traditional British summer holiday); there are seasonal patterns of car purchasing (which generate an alternative set of seasonal impacts); and, finally, most vehicles do not have a test - and therefore provide a mileage reading - until they are at least three years old. Processes for dealing with these range from complex heuristics to identify erroneous values, through to simple assumptions (e.g. that mileage increases linearly from zero to the three year value). It has also been necessary to define thresholds to screen out unexpectedly high mileages (currently set at 55,000 miles p.a. based on sensitivity testing during the scoping study) since these would skew all future analysis. Given seasonal issues and the processes used, the mileage data is only considered to be robust for annual estimates. However, using the quarterly data, it is possible to generate data for different annual periods, depending, for example, on whether calendar or financial year figures are of most value. As the project progresses, mileage estimation processes will be refined further, enabling sensitivity testing of initial conclusions.

4. ADDITIONAL FIELDS FOR THE CORE DATASET

Given the project team’s interest in issues relating to energy use and emissions, a number of further values have been generated for each vehicle (see Chatterton et al 2015b). These are:

- Emissions of carbon dioxide (CO₂), nitrogen oxides (NOx) and particulate matter (PM), using generic emission factors derived from fuel type, engine size and vehicle age according to a methodology developed by UWE.
- Fuel economy, using assumptions based on both detailed data from the Irish vehicle fleet, as well as less detailed data from the UK.
- Energy use, using Defra guidelines.
- Annual fuel costs, derived from the fuel economy figures, together with DECC pricing information.
- Annual VED, based on MOT vehicle characteristics.

In some cases, LPG, LNG, CNG, hybrid and/or electric vehicles have required additional input data for calculations.

5. AREA LEVEL STATISTICS

From the master table, which provides details of all vehicles, it has then proved useful to generate an area level table, which provides average figures and/or totals for various key parameters for each areal unit (such as km
driven, age, engine size, fuel types, CO₂ emissions, etc.). To date, we have been working on a version of this table for the 2011 calendar year (in order to match Census data). This uses the 30th June (Q2) time point for information about vehicle stock, and as the central time point for annual vehicle mileages.

We have excluded vehicles which do not have a valid location (which are also often those which are between keepers), and have simplified the vehicle class information into three categories - 2 and 3 wheelers: (Classes 1-3); cars (Class 4 and 4A); and minibuses and small vans (Classes 5-7), as shown in Table 1. For insights on personal car use, analysis is also focused on vehicles in private ownership since these figures are relatively closely correlated to the Census numbers of ‘cars or vans owned or available for use by members of the household’. Although some vehicles recorded as being in commercial ownership will undoubtedly be available for personal use, there is no obvious way to include them in local area statistics without also including a large number of vehicles which are not of this nature.

Table 1: Vehicle classes recorded in the MOT test data

<table>
<thead>
<tr>
<th>Class</th>
<th>Vehicle Description</th>
<th>1st Test</th>
<th>MOT Cost</th>
</tr>
</thead>
<tbody>
<tr>
<td>Class 1</td>
<td>Motorbike (engine size up to 200 cm³)</td>
<td>3 Years</td>
<td>£ 29.65</td>
</tr>
<tr>
<td></td>
<td>Motorbike with sidecar (engine size up to 200 cm³)</td>
<td>3 Years</td>
<td>£ 37.80</td>
</tr>
<tr>
<td>Class 2</td>
<td>Motorbike (engine size over 200 cm³)</td>
<td>3 Years</td>
<td>£ 29.65</td>
</tr>
<tr>
<td></td>
<td>Motorbike with sidecar (engine size over 200 cm³)</td>
<td>3 Years</td>
<td>£ 37.80</td>
</tr>
<tr>
<td>Class 3</td>
<td>3-wheeled vehicles (up to 450kg unladen weight)</td>
<td>3 Years</td>
<td>£ 37.80</td>
</tr>
<tr>
<td></td>
<td>3-wheeled vehicles (over 450kg unladen weight)</td>
<td>4 Years</td>
<td>£ 54.85</td>
</tr>
<tr>
<td>Class 4</td>
<td>Cars (up to 8 passenger seats)</td>
<td>3 Years</td>
<td>£ 54.85</td>
</tr>
<tr>
<td></td>
<td>Motor caravans</td>
<td>3 Years</td>
<td>£ 54.85</td>
</tr>
<tr>
<td></td>
<td>Quads (max unladen weight 400kg - for goods vehicles 550kg and max net power of 15kw)</td>
<td>3 Years</td>
<td>£ 54.85</td>
</tr>
<tr>
<td></td>
<td>Dual purpose vehicles</td>
<td>3 Years</td>
<td>£ 54.85</td>
</tr>
<tr>
<td></td>
<td>Private hire and public service vehicles (up to 8 seats)</td>
<td>3 Years</td>
<td>£ 54.85</td>
</tr>
<tr>
<td></td>
<td>Ambulances and taxis</td>
<td>1 Year</td>
<td>£ 54.85</td>
</tr>
<tr>
<td></td>
<td>Private passenger vehicles and ambulances (9 to 12 passenger seats)</td>
<td>1 Year</td>
<td>£ 57.30</td>
</tr>
<tr>
<td>Class 4A</td>
<td>Goods vehicles (up to 3,000kg design gross weight)</td>
<td>3 Years</td>
<td>£ 54.85</td>
</tr>
<tr>
<td></td>
<td>Class 4 vehicles (9 to 12 passenger seats) with a seat belt installation check</td>
<td>n/a</td>
<td>£ 64.00</td>
</tr>
<tr>
<td>Class 5</td>
<td>Private passenger vehicles and ambulances (13 to 16 passenger seats)</td>
<td>1 Year</td>
<td>£ 59.55</td>
</tr>
<tr>
<td></td>
<td>Private passenger vehicles and ambulances (more than 16 passenger seats)</td>
<td>1 Year</td>
<td>£ 80.65</td>
</tr>
<tr>
<td>Class 5A</td>
<td>Class 5 vehicles (13 to 16 passenger seats) with a seatbelt installation check</td>
<td>n/a</td>
<td>£ 80.65</td>
</tr>
<tr>
<td></td>
<td>Class 5 vehicles (more than 16 passenger seats) with a seatbelt installation check</td>
<td>n/a</td>
<td>£ 124.50</td>
</tr>
<tr>
<td>Class 7</td>
<td>Goods vehicles (over 3,000kg up to 3,500kg design gross weight)</td>
<td>3 Years</td>
<td>£ 58.60</td>
</tr>
</tbody>
</table>

Shading divides the classes into the three categories being used for analysis.

In England and Wales, the primary areal unit being used is the Lower-layer Super Output Area (LSOA), a geographical unit designed for the UK Census in England and Wales. It relates to a relatively socially homogeneous area of, on average, around 700 households, and 1,600 persons. There are 34,753 of
these. In Scotland, data have been generated for Data Zones (which are similar, though have slightly smaller populations than LSOAs) of which there are 6,976.

Lookup tables have been created, to enable the analysis of data at other spatial scales, including Middle-layer Super Output Areas (of which there are 7,201 in England and Wales, together with 1,279 Intermediate Geography Zones in Scotland), and local authority districts (of which there are 380).

As part of this process, it has become necessary to deal with missing values (i.e. when vehicle information is incomplete). This has been done in various ways. For example, the work on calculating emissions and energy consumption has been done by averaging values from the other vehicles in the area (LSOA) for any field that is missing.

It has also been necessary to identify areal units with clearly anomalous values, which can then be excluded from analysis. Checks are currently taking place to define a definitive list of these areas, based on all potential variables of interest.

6. PARALLEL DATA SETS

In order to make sense of the project dataset, it has been necessary to fuse it with a wide variety of other datasets, available at a range of spatial scales. Of these, the most important is the UK Census, making it possible to move from average figures for each area about vehicle characteristics, to figures which provide these data in relation to the number of people or households (e.g. to go from total number of cars in the area, to number of cars per person in the area) as well as looking at this information in association with the types of people who live in those areas.

One of the key benefits of the MOT data is its availability on a rolling basis, rather than, say, every 10 years like the Census. However, this raises the issue that much of the other data that is useful to use for analysing with the MOT data is not available annually. Going forward, the mid-year population estimates will be important in this regard (which provide information on people, and age bands, but not household numbers).

When combining with socio-demographic data, particular issues also arise in relation to people in communal establishments (who are included in population totals, but not in Census-related measures about vehicle ownership and certain other household characteristics). Issues also arise where geographical boundaries do not match (due to other data being provided in grid squares, or aligned to 2001 rather than 2011 boundaries).

Other datasets being used in the project are as follows:

- ONS urban-rural area classification
- ONS output area classification
- Experian median household income data (at LSOA level)
- ONS mean household income data (at MSOA level)
- DfT ‘Journey Time Statistics’ (previously called ‘Accessibility Statistics’), available separately for England, Scotland and Wales
• DECC subnational statistics on domestic gas and electricity consumption
• Air pollution data from Defra and the National Atmospheric Emissions Inventory
• Poverty metrics developed from the Breadline Britain Index
• Topography measures derived from OS Terrain data.
• Density of public transport provision measures derived from the Traveline National DataSet (TNDS), including number of bus departures per week (MSOA level) and distance to nearest rail station (LSOA level)
• Data from the National Public Transport Data Repository

7. RESEARCH TOPICS AND FINDINGS TO DATE
A variety of research topics are being explored through the project using the data. This section attempts to give some insights into the range of topics, without being able to provide full insights on any. All references are available through the project website (www.MOTproject.net).

7.1 International experience of undertaking similar analysis
At a relatively early stage of the work, a review of international experience of undertaking this sort of analysis was undertaken (Cairns et al., 2014b). This was complemented by a discussion at the Association of American Geographers conference in April (Chatterton & Anable, 2016). We are currently running a survey to systematically obtain information about international experience:
https://www.smartsurvey.co.uk/s/Vehicleinspections/

7.2 Modelling the determinants of car ownership and use
Various strands of work have taken place to try to model the underlying determinants of car ownership and use at an areal level.

This has included models regressing only income and density against car ownership and car use at MSOA and LSOA levels, and more complex models with a greater number of potential explanatory variables (such as average age and occupation profiles, household structure, public transport availability and average distances to work). We have also compared the outcomes using global linear regression models with more localised geographically-weighted regression models, to explore the importance of both overarching, and locally-specific factors (Yeboah et al., 2015; Emmerson et al., 2016). Comparisons have also been made with conventional transport models.

In all models, our analyses confirm previous studies that reveal the dominant influence of population density as a key variable for car ownership and use - underlining the importance of land-use planning in any transport policy. For example, a simple model (at MSOA level) using the square-root of population density as the only input variable is able to account for nearly 70% of the
variance in car miles per person. (Further work is needed to explore the relationship between population density and public transport provision.)

We have also seen differences at different geographical levels of analysis. Interestingly, for car ownership, density becomes less important when examining differences between smaller geographical areas, suggesting that local factors become relatively more important at that scale.

Use of geographically weighted regression (GWR), where models are set up to take account of locally varying relationships, also has an effect on predictive power. For instance, a global model using income and density values is able to account for 66% of the variance in car ownership per person in England. Figure 1 shows that this model fit is not spread evenly, with blue areas depicting relatively poor fit and red areas relatively good fit. When GWR is applied, 84% of variance is explained, with a much more equal model fit across the country. Further GWR analysis with a wider set of variables also reveals that the strength of certain relationships is different in different parts of the country. For instance, occupation is a much stronger predictor of car ownership in the north of England than in the south-east. Our analysis will continue to identify the most important locally relevant variables.

Figure 1: Relative ‘fit’ of a global car ownership model based on income and population density.
7.3 Understanding intra-areal variation

As well as undertaking analysis using area averages, the available data also makes it possible to explore how vehicle use varies within an individual area. This, in turn, makes it possible to look at issues such as whether the same areal average mileage can reflect different underlying profiles of use; how much any average is being skewed by a group of high mileage vehicles within the area; and whether there are statistics other than the average that could be more sensitive for distinguishing between areas, (Ball et al., 2016a and b). Figure 2 indicates the sorts of profiles that can be generated.

![Figure 2: Cumulative frequency of population, depending on proportion of population in households with no car and annual mileage per person for those in households with at least one car (Regions in Great Britain, 2011, MOT dataset and Census KS404 & LC4109 combined)](image)

7.4 Understanding the significance of vehicle characteristics and use to total emissions

Early exploration of the data highlighted the overriding importance of the distance that people drive in determining overall vehicle emissions and energy use, particularly when compared to the efficiency or cleanliness of vehicles owned (Chatterton et al 2015). When analysis is undertaken at the LSOA level (i.e. looking at differences in aggregate values for the areas), the difference between the areas (excluding outliers) with the highest average values and the lowest average values is in the order of 15-20% for average emissions (from about 152 to 172gCO₂ per km), but 10x for car use (from about 1,000 to 10,000km p.a. per person). Figure 3 shows the relationships between total energy use, average car use, and average vehicle emissions to illustrate the nature of the relationships.
7.5 Understanding the contribution of vehicle use to total household energy use

Data from DECC on direct electricity and gas use has been linked with data from this project to look at total direct household energy use. As shown in Figure 4, this work highlights that high vehicle energy use is typically correlated with high electricity and gas energy use. It also helps to highlight the dramatic increase in electricity supply needed if all petrol and diesel vehicles are to become electrified, (Chatterton et al, 2016b).

7.6 Air pollution and social justice

New investigation has been undertaken, that builds on previous analyses of air quality and environmental justice (particularly Mitchell and Dorling, 2003). Using data on both emissions and atmospheric concentrations of NOx and NO₂, as well as analysis of the level of poverty in each area (defined according to an estimate of the percentage of households in poverty), it has been possible to show that those in the poorest households typically suffer from the worst air quality in their residential neighbourhoods. Households with
a high proportion of under-5s, and young adults (in their 20s) are also disproportionately exposed. Meanwhile, there is an almost inverse relationship between the areas where the people who generate the greatest emissions live, and those where the highest concentrations of emissions occur (see Figures 5 and 6). There are also significant differences in the type of social groups affected by pollution compared with those that are responsible for emissions (Barnes and Chatterton 2016; Chatterton and Barnes 2016).

Figure 5: For ten population deciles (based on the proportion of households in poverty), graphs show the difference between exposure to NO$_2$ and responsibility for vehicle emissions of NOx.

Figure 6: Cartograms (areas based on population) showing background concentrations of NO$_2$; measured emissions of NOx from road transport; and annual emissions from vehicles registered in each LSOA.
7.7 Financial considerations and social justice

It has been possible to estimate the annual Vehicle Excise Duty (VED) and fuel costs for each individual vehicle in the dataset, in order to explore how the average costs for areas vary in relation to the income. As shown in Figure 7, at the LSOA level, households with cars in low-income areas are likely to spend a higher proportion of their income on these costs than households with cars in higher income areas. In this context, the social justice of future tax changes, which will benefit electric-car owners compared to those buying other types of cars, should be of concern, given that uptake of these vehicles is initially likely to be by higher income groups, (Chatterton et al, 2016a).

![Figure 7: Average household expenditure on VED and road fuel depending on median household income per LSOA.](image)

7.8 Local authority profiles and benchmarks

A variety of cluster analyses have been undertaken as part of this project, together with assessment of how existing ways of categorising areas appears to relate to observed car ownership and use characteristics. One future research priority is to reach consensus about the most useful typology that can be used to group areas, in order to provide input into the following research questions:

a) On what basis can we assess whether areas demonstrate patterns of car ownership and use that are significantly different from others that they might be expected to be similar to, and whether this can be interpreted as evidence of policy impact; and

b) If benchmark or control areas need to be defined to assess the impacts of a particular policy in a specific location, on what basis should this be done?
Availability of data at an LSOA level means that such comparisons could be done at a relatively fine level, although it may be that MSOA level comparisons will also be important, depending on the policy issue, the availability of the complementary data needed for analysis and the potential for misleading results from small sample sizes. This requires further exploration.

8. CONCLUSIONS

This paper has attempted to provide an overview of the activities being undertaken as part of the MOT project, and to provide an indication of some of the findings that are emerging through specific topic work.

To date, a non-trivial amount of effort has been required to convert the data available in the national datasets into a resource that can be used for this type of analysis. Work has focused on a 2011 dataset, involving a large number of data checks, and the development of a range of procedures for dealing with erroneous or missing values. One future priority is to move to multiple years of data, enabling analysis of trends over time.

ACKNOWLEDGEMENTS

This analysis has been undertaken as part of the MOT project (EP/K000438/1), funded by the UK Engineering and Physical Sciences Research Council under the Research Councils UK Energy Programme, leading on from Research Councils UK Energy Programme research grant EP/J004758/1. This project has been led by Prof Jillian Anable at the University of Leeds, and also involves TRL, University of the West of England, University of Bristol, University College London and University of Aberdeen. The work has been formally supported by the UK Department of Transport (DfT) and the former UK Department of Energy and Climate Change (DECC). This project includes use of Census data from National Statistics data © Crown copyright and database right 2012, and OS data © Crown copyright and database right (2016). Project outputs can be found at www.MOTproject.net. Analysis in this report is based on the version 8 project dataset. Grateful thanks to DfT, DVLA, DVSA, DECC, the project advisory board and all those who have worked on the project.
BIBLIOGRAPHY


Chatterton, T., Barnes, J., Wilson, R.E., Anable, J., and Cairns, S. (2015) Use of a novel dataset to explore spatial and social variations in car type, size,

DfT, (2014) MOT testing data user guide v3.1, Department for Transport, August 2014


