Techniques for the inference of mileage rates from MOT data

R. E. Wilson[a]*, S. Cairns[b,c], S. Notley[b], J. Anable[d], T. Chatterton[e] and F. McLeod[a]

[a] *Transportation Research Group, University of Southampton, Southampton, UK*; [b] *TRL, Wokingham, UK*; [c] *Centre for Transport Studies, University College London, London, UK*; [d] *The Centre for Transport Research, University of Aberdeen, Aberdeen, UK*; [e] *Air Quality Management Resource Centre, University of the West of England, Bristol, UK*

Mathematical and computational techniques are developed for the processing and analysis of annual MOT (roadworthiness) test data that the UK Department for Transport has placed in the public domain. Firstly, techniques are given that clean erroneous records and a linking procedure is provided that permits the inference of an individual vehicle's mileage between consecutive tests. Methods are then developed that analyse aggregate mileage totals, as a function of vehicle age, class and geography. The inference of aggregate mileage rates as a function of time is then considered.

Keywords: car ownership and use; mileage rates; MOT tests; open data; mathematical methods

1. Introduction

Understanding car ownership and use is a key issue in transport policy, and appropriate data is essential for analysis. In particular, data on car ownership and use is a core component of geographically disaggregated transport modelling. For example, in the UK, the Department for Transport has commissioned a number of official traffic forecasting models such as the Regional Highway Traffic Model (RHTM), the National Road Traffic Forecasts (NRTF) and the National Transport Model (NTM) all of which contain car ownership (de Jong et al., 2004; Whelan 2007) and car use models that need to be parametrised by data. Data also enables the exploration of historic trends (e.g., Millard Ball & Schipper 2010) and it is essential for understanding the links between travel behaviour and other factors (e.g., Dargay, 2001). Data is also key for understanding the impacts of policy, including those aimed at reducing carbon emissions in the transport sector, ranging from large-scale national projects, such as the promotion of electric vehicles (Office for Low Emission Vehicles, 2011; Element Energy 2009), through to small-scale projects, such as schemes that make it more attractive to walk or cycle in a given local area (Sloman *et al.* 2010).

To date, robust data concerning vehicle ownership and use has come from two main sources. These are (i) surveys of individuals or households and (ii) on-street traffic counts. Surveys are potentially subject to bias due to misreporting (either deliberate or accidental) and who chooses to reply; Mokhtarian and Cao (2008) provide a good summary of the issues. For instance, problems

*Corresponding author. Email: R.E.Wilson@soton.ac.uk

are likely to be caused in such datasets because trips and vehicle data can be only be drawn from the observed sub-sample of participants which made trips during the travel-diary period. Several analytical biases also result from issues such as self selection bias whereby residents choose locations consistent with their travel pre-dispositions; simultaneity bias whereby residential location and travel behaviour decisions influence each other; and omitted variable bias whereby unobserved variables such as attitudes produce incorrect associations. In the general absence of true panel data, pseudo panel datasets have also been constructed by using cohort averages of repeated cross-sections (Huang 2005; Dargay 2002; Dargay & Vythoulkas 1999). Whilst allowing for some dynamic analysis, this method also incurs a number of methodological problems related to the use of cohort means (de Jong et al., 2004; Zegras 2010). Other studies have attempted to understand car ownership using Census data or large scale travel surveys and applying geographically weighted regression to estimate local income elasticities (Adjemian et al. 2010; Clark 2004 & 2007).

In the UK, on-street traffic counts are used to derive national road traffic estimates (Department for Transport 2010a) of the total vehicle miles driven in a given year. Counts of vehicles are multiplied up by the length of the links to which they apply in order to estimate the total miles driven on that link. In practice, the strategic road network (motorways, major trunk roads) has a very good coverage of automatic counters based on inductance loop technology, so that the total mileage driven on it is known quite robustly. Unfortunately, automatic counters cannot robustly disaggregate traffic over vehicle classes, or (for example) identify the fuel type or engine capacity of the counted vehicles, and hence the associated emissions cannot be estimated. Away from the strategic network, roads are poorly instrumented and so flow on them must be estimated by manual observations for a very small sample of links and days. This procedure is thus potentially subject to huge bias and/or uncertainty due to the way in which the sites for manual counting are chosen.

Note that these existing data collection techniques are expensive. As discussed by Eddington (2006), there is a belief that small-scale local transport measures are cost effective in achieving travel behaviour change. However, such smaller schemes are unlikely to have the budgets required to assess their worth rigorously using existing techniques. Thus there is a demand for new estimation techniques which are either cheap, or perhaps even free (because they are based on recycled data that was originally collected for other purposes).

In 2005, the Vehicle and Operator Services Agency (VOSA) introduced a computerised system for reporting annual MOT (roadworthiness) test results and storing them in a Department for Transport (DfT) database. In November 2010, the DfT published this data (Department for Transport 2010b) — consisting of the results of approximately 150 million MOT tests from 2005 to the spring of 2010. Some fields, such as vehicle registration plates and unique VTS (vehicle test station) identities have been withheld from the published data. However, what remains still contains a wealth of information that is not available elsewhere. In addition to the results of the MOT test itself (including detailed reasons for failure), the data include: the vehicle odometer (mileage) reading; the vehicle manufacturer, model and engine capacity; the vehicle's year of first use; and the top-level postal area (first letters only from the postcode) of the VTS.

An unadvertised feature of the published data is that an internal database index may be used to track many individual vehicles from year to year throughout their test history. In consequence one may infer the mileage of a vehicle between a pair of tests. Therefore our proposition is that MOT odometer data may provide an important missing link in the analysis of vehicle usage.

This paper outlines computational and mathematical techniques that use large sets of MOT records to infer mileage rates across the vehicle population. The paper is organised as follows.

Section 2 describes the DfT public data release and the procedure which may be used to follow individual vehicles from test to test. Section 3 introduces the key data structure, namely the *interval* between two tests of a given vehicle, on which the remainder of this study is based. The development of analytical techniques is then divided according to (i) static questions (section 4) which analyse mileage rates as a function of attributes such as geography, vehicle age, vehicle class etc., but at a fixed point in time; and (ii) how to compute time-dependent mileage rates (section 5). Finally, section 6 presents conclusions and discusses possibilities for future work.

2. Description of the public data release

Here we give further details of the data that the DfT released into the public domain in 2010 (Department for Transport 2010b). In addition to documentation and look-up tables, there are two main groups of plain text files: type `mdr_test_result` which contain one data record for each MOT carried out (pass or fail) for the study period 2005–2010; and type `mdr_test_item` which can be cross-linked with `mdr_test_result`, and with a separate index file to examine in detail the advisories / reasons for failure in any particular MOT test.

Our interest is in the `mdr_test_result` data. For convenience this is segmented into files which are organised by the year in which the given MOT tests took place, (for example, `mdr_test_result_2005.txt`, `mdr_test_result_2006.txt`, etc.). Each line of each of these files describes a single MOT test, and takes the form (for example):

```
1738409|2007-07-07|4|PR|P|85436|MK|VAUXHALL|ASTRA CLUB 8V|WHITE|P|1598|1999
```

The 13 fields here have the following meanings.

- (i) **1738409**: the identifier of the individual test (a unique integer across all the files provided).
- (ii) **2007-07-07**: the date of the test.
- (iii) **4**; (iv) **PR**; (v) **P**: these represent the class of the vehicle, the test type and the result (pass/fail).
- (vi) **85436**: the odometer reading reported by the Vehicle Testing Station (VTS).
- (vii) **MK**: the two-digit postcode area of the VTS where the test was performed — note this is not necessarily the same as where the vehicle is registered.
- (viii) **VAUXHALL**; (ix) **ASTRA CLUB 8V**; (x) **WHITE**: the make, model and colour of the vehicle according to the DVLA database.
- Similarly (xi) **P**; (xii) **1598**: fuel type (here petrol), and engine capacity; (xiii) **1999**: year of first use.

Each of the provided files contains many, many such lines of data, but they are not of much use in this form — because the records corresponding to different tests of the same vehicle are split amongst separate files. Furthermore, because the VRM (vehicle registration mark) was suppressed in the initial (2010) public data release, it was not clear how to link the test records belonging to an individual vehicle and thus compute its mileage between tests[1].

[1]Since the initial submission of this paper, there has been a second release of the MOT data (in March 2012) that provides additional fields, including a unique vehicle identifier that links tests to individual vehicles. A further release is planned in Autumn 2012, which is expected to provide further enhancements to the data (DfT, personal correspondence).

Table 1: A sample of concatenated test data sorted by the test identifier index. This block of data gathers together four consecutive years of MOT tests (pass and fail) for an individual vehicle, in reverse-time order. We may read off the reported odometer history at each of the test dates: 2006-06-27, 71,803 miles; 2007-07-07, 85,436 miles; 2008-07-07, 96,592 miles; and 2009-08-28, 107,094 miles.

```
1738412|2006-06-27|4|N|F|71803|MK|VAUXHALL|ASTRA CLUB 8V|WHITE|P|1598|1999
1738411|2006-06-27|4|PR|P|71803|MK|VAUXHALL|ASTRA CLUB 8V|WHITE|P|1598|1999
1738410|2007-07-07|4|N|F|85436|MK|VAUXHALL|ASTRA CLUB 8V|WHITE|P|1598|1999
1738409|2007-07-07|4|PR|P|85436|MK|VAUXHALL|ASTRA CLUB 8V|WHITE|P|1598|1999
1738408|2008-07-07|4|N|F|96592|MK|VAUXHALL|ASTRA CLUB 8V|WHITE|P|1598|1999
1738407|2008-07-07|4|PR|P|96592|MK|VAUXHALL|ASTRA CLUB 8V|WHITE|P|1598|1999
1738406|2009-08-28|4|N|F|107094|UB|VAUXHALL|ASTRA CLUB 8V|WHITE|P|1598|1999
1738405|2009-08-28|4|PR|P|107094|UB|VAUXHALL|ASTRA CLUB 8V|WHITE|P|1598|1999
```

However, for this analysis, we established the following procedure. If one concatenates the six test result files (one for each year of MOT tests, 2005–2010 inclusive) and re-sorts the data by the test identifier field, one obtains many blocks in the form shown in table 1. Thus it was possible to track the mileage of individual vehicles, together with the type of the vehicle and its (rough) location. For example, we may infer that this particular vehicle (a Vauxhall Astra, first registration 1999) drove 85,436-71,803 = 13,633 miles in the 375 days from 2006-06-27 to 2007-07-07, at an average rate of 36.35 miles per day. During this time it most probably resided in the Milton Keynes area (postcode MK). Note that we cannot say how those 13,633 miles were spread amongst the 375 days, and in practice that distribution is likely to be highly non-uniform; nor can we say with confidence where those miles were driven.

Finally, note that this technique supposes that the entry of odometer data at the VTS is itself robust. In practice, it may only be approximate or indeed may be wholly fabricated. Techniques for identifying anomalous records are outlined shortly.

3. Basic data structure: intervals and their attributes

Once individual vehicles can be tracked through a sequence of tests, a a huge data resource becomes available for longitudinal analysis. For example, could a change in a vehicle's usage be explained either by a change in ownership, or be a change in the number of vehicles at its registered address? Unfortunately, to answer these questions would require individual vehicle ownership and registration data which are not in the public domain. Furthermore, individual vehicle histories tend to be quite short, owing to the relatively short duration of the public release data set (2005–2010).

Hence to date our analysis has been concerned with distributions of mileage across the vehicle population, and how these disaggregate over various attributes of the vehicles involved. This analysis includes results which are longitudinal (time-dependent) at the population level, but does not include analysis which is longitudinal at the level of the individual vehicles.

To this end — the key data structure that we use is that of an *interval* between a pair of consecutive tests for an individual vehicle (and in particular, we do not use linking at the level of triples, quadruples etc. of tests). For example, the eight tests in table 1 yield seven inter-test intervals,

Table 2: Inter-test *intervals* generated from table 1. Note that there is redundancy in this data format because most tests appear twice, both as a first test (i.e., 'left-hand end') and as a second test (i.e., 'right-hand end'). However, the advantage of this redundancy is that for a given search date, the set of all intervals which 'span' that date can be found without table look-ups.

Interval	First test			Second test		
	date t_1	miles x_1	place$_1$	date t_2	miles x_2	place$_2$
1	2006-06-27	71803	MK	2006-06-27	71803	MK
2	2006-06-27	71803	MK	2007-07-07	85436	MK
3	2007-07-07	85436	MK	2007-07-07	85436	MK
4	2007-07-07	85436	MK	2008-07-07	96592	MK
5	2008-07-07	96592	MK	2008-07-07	96592	MK
6	2008-07-07	96592	MK	2009-08-28	107094	UB
7	2009-08-28	107094	UB	2009-08-28	107094	UB

as shown in table 2. Here we have removed the test outcomes (pass, fail etc.), as they are not of primary interest, and to this data may be linked vehicle-specific fields (namely `|4|VAUXHALL|ASTRA CLUB 8V|WHITE|P|1598|1999`) which do not vary from test to test. In practice, the vehicle's colour is not of interest, nor is the precise make and model. We thus retain (i) the fuel type; (ii) the engine capacity and (iii) the year of first registration. In addition, each vehicle is classified into one of 15 categories, namely super-mini, small family, large family, executive, van, SUV, MPV, sports-cabrio, unknown, motorcycle, city car, luxury, taxi, or bus — listed here in descending order of their frequency of occurrence. This classification is achieved by applying a hand-built look-up table that maps each of the (approximately) 33,000 combinations of make and model that occur in the MOT data.

Note that in practice many intervals (e.g., intervals 1, 3, 5 and 7 in table 2) have a span of zero days, because they correspond to a failed test and re-test on the same day. The mileage rate is thus either infinite or undefined, and such intervals are removed from the data before analysis continues. The data is cleaned further by removing intervals for which:

- the span is less than 30 days (The mileage rates for these intervals are anomalously low — presumably because they mostly correspond to vehicles which are off-road between a failed test and subsequent re-test.);
- either test involves an anomalous odometer reading such as 0, 99999 or 999999 (These occur with a relatively high frequency that suggests poor practice at the VTS.);
- if $x_2 < x_1$, i.e., an apparent negative mileage rate (This may be due to fraud, but is more likely erroneous data entry at the VTS — e.g., due to a missing digit in x_2.);
- the mileage rate exceeds an upper threshold — which in practice we set at 500 miles per day. In practice, very few vehicles drive this far consistently over extended periods. Most of these records are probably due to erroneous data entry (e.g., a missing digit in x_1).

The chief computational task that we have performed is thus the construction of a very large

number of such *intervals*. The methods that we have developed for identifying contiguous same-vehicle blocks of data yield approximately 113 million intervals, of which about 76 million survive the cleaning criteria laid out above.

However, a significant shortcoming of the data is that private vehicles which are less than three years old are typically not required to undergo MOT tests — and yet these vehicles represent a significant proportion of the total miles driven across the vehicle population. In principle, the average mileage rate of such vehicles could be estimated at the point when they are first tested, on reaching three years of age. However, at present we only have the date of first registration to the nearest year, which leads to inaccuracy in the measurement of the time span to the first test, and consequently to inaccuracy in the computed average mileage rate. Moreover, the temporal resolution problem is severe as potentially significant changes in usage patterns might have occurred during a three year span. Finally, such an analysis can only be performed up to 3 years after some of the associated miles were driven. (A similar problem occurs elsewhere in our analysis, where there is a roll-in and roll-out period of one year, as this is the typical span between consecutive MOT tests.) In consequence, the analysis in this paper is restricted to the usage of vehicles that are more than three years old.

4. Analysis of intervals: time-independent questions

For convenience, we divide our analytical development according to the following types of question:

1. How do mileage rates depend on attributes such as location, vehicle type, year of first registration etc.?

2. How do mileage rates depend on time?

Of course, we may also ask questions of a compound nature. But to simplify matters, questions of type 1. will be asked at a single point (or possibly a small number of points) in time, and questions of type 2. will be asked without a fine disaggregation over attributes. This section deals with type 1. questions; section 5 deals with type 2.

For questions of type 1., the first step is to choose an *observation time* t^*. One then selects those intervals which *straddle* the observation time — that is for which $t_1 \leq t^* < t_2$, where the half-open interval assures the correct counting — and which match the required attributes. Each relevant interval contributes an *average* mileage rate r, where, in the notation of section 3,

$$r := \frac{x_2 - x_1}{t_2 - t_1}.$$

Here *average* means average over time for the corresponding individual vehicle. The process is illustrated graphically in figure 1. In practice, a large number N of intervals will match a given observation date and attribute criteria, and a large set of corresponding average mileage rates $r_1, r_2, \ldots, r_N$ will be computed. One may then characterise the corresponding vehicle use either

- (crudely) by the average average mileage rate

$$\overline{r} := \frac{1}{N} \sum_{i=1}^{N} r_i,$$

  where we mean average over time for each vehicle, averaged over the relevant vehicle population;

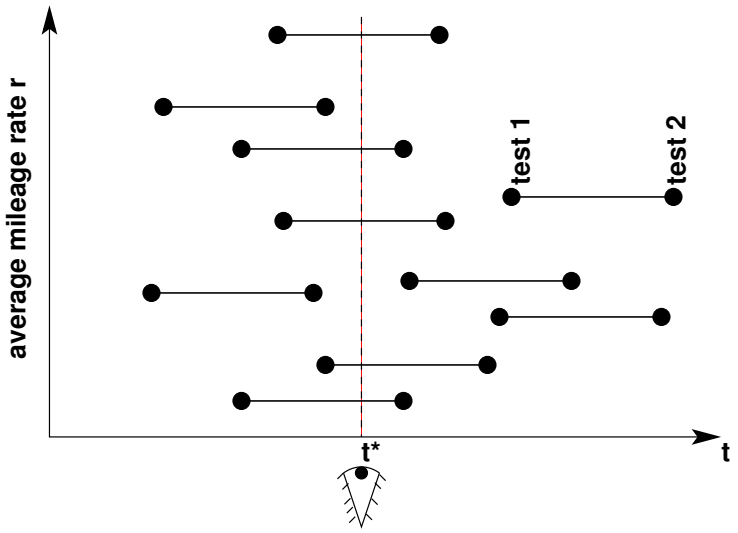

Figure 1: Analysis of mileage rates at an *observation time* t^*. Those intervals which *straddle* t^* are selected. Each then contributes an average mileage rate r_i to subsequent computations. Here r_i is the average rate at which the given vehicle accrues miles between the relevant consecutive tests.

Table 3: Sample sizes N and average average mileage rates $\overline{r}$ (miles per day) for the postcode areas KY (Kirkcaldy) and W (West London), disaggregated over years of first registration.

N	1992	1996	2000	2004
KY	818	4820	10555	12191
W	1020	3137	5777	4748

$\overline{r}$	1992	1996	2000	2004
KY	13.6	18.0	22.0	26.5
W	12.5	13.8	15.8	19.2

- or (with more sophistication) by computing higher moments of r_i, $i = 1, 2, \ldots, N$, or by estimating a (parametric) distribution for them.

Example. We consider a comparison of vehicle usage in rural versus urban areas and how it depends upon vehicle age. We choose an observation date of 2008-1-1, and we compare the postcode areas KY (Kirkcaldy, Scotland) and W (West London) as the most extreme opposites of rural and urban geographies. (It is a significant drawback that the spatial resolution is so coarse, since KY includes the towns of Dunfermline and Kirkcaldy itself, each of which has a population of approximately 50,000. KY is therefore not entirely rural, but rather it is the most rural postcode area available.) We then select intervals that (i) match the respective postcode areas at both the 'left-hand' and 'right-hand' tests; (ii) span t^*; and (iii) we disaggregate them over vehicle age, focussing on 1992, 1996, 2000, and 2004 as years of first registration. Table 3 presents sample sizes N and average average mileage rates $\overline{r}$, whereas figure 2 displays the results in the form of the distributions of the (individual vehicle) average mileage rate across the relevant populations.

As we might expect, the average mileage rates are higher in the rural postcode, and moreover they decline with the age of vehicles irrespective of the postcode area — that is, older vehicles tend

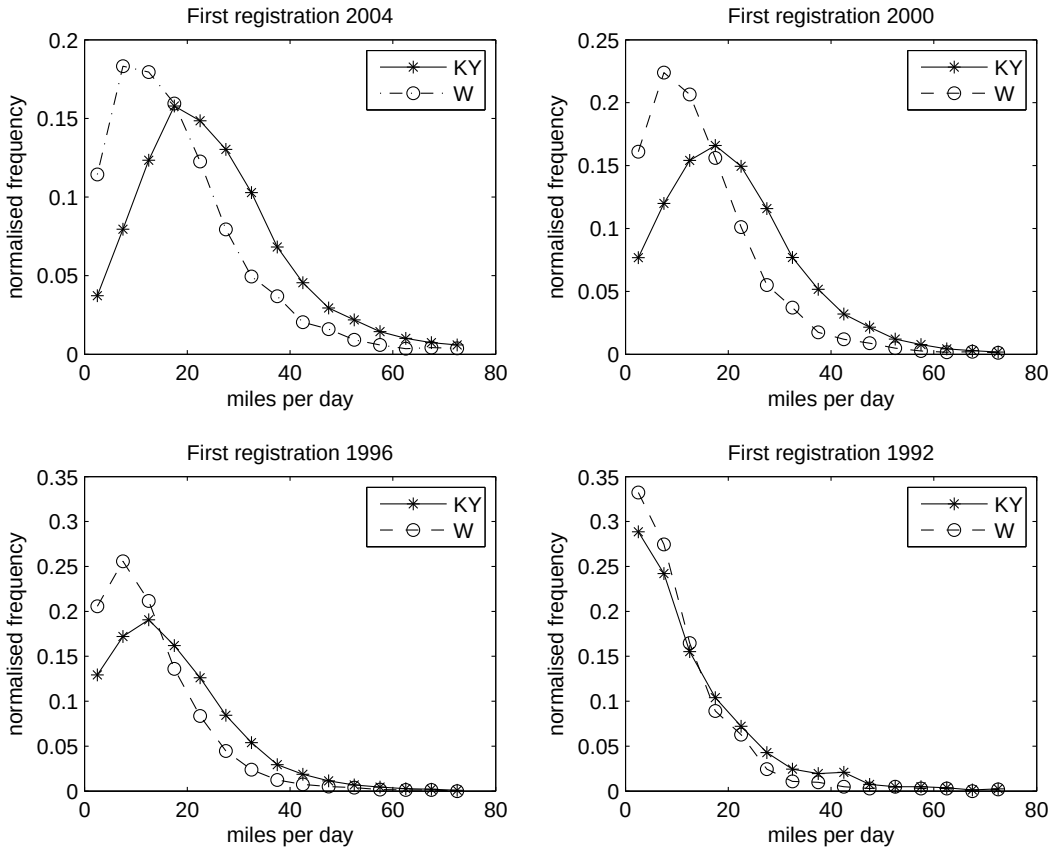

Figure 2: Comparison of the mileage rate distribution in the rural postcode area KY (Kirkcaldy) versus the urban postcode area W (West London), as a function of vehicle age. Frequencies are computed in bins of width 5 miles per day. Owing to the relatively small sample sizes in this example, grouping with narrower bins tends to be dominated by noise.

to be driven less than newer ones[2]. But the precise parametrisation of these relationships has not previously been available in such fine detail.

However, the chief interest is in the dependence between the age and geography relationships, as illustrated via the progressive change in the distributions in figure 2. For newer vehicles, the higher mileage in KY displays itself via a 'right-shift' of the distribution. However, the difference between the usage patterns of rural and urban vehicles tends to disappear as the vehicles age. We speculate that this trend might be due to patterns in household vehicle ownership — for example, are the older vehicles mainly second vehicles that are used for non-essential journeys? Could it be the case that non-essential mileage is similar in rural and urban postcodes? These sorts of question demand vehicle registration data to which we presently lack access.

Finally — we may ask whether there is a simple parametric distribution that fits the profiles shown in figure 2, whose shapes are reminiscent of classical probability distributions such as the Log-normal, Gamma and Weibull types. However, our data can only be explained by modelling

[2]Further analysis is required to establish whether the mileage rates of individual vehicles decline as they age, or whether this effect is due to the incremental retirement of high mileage vehicles due to wear.

complex patterns in human behaviour, so there is no analytical reason to suppose that a particular parametric distribution will explain the data. Rather, this is an exercise in fitting distributions to achieve data compression. To clarify, if a distribution with (say) three parameters can be fitted to the KY data for 2004, then those three parameter values can be used in place of the 12,191 data values from which they are derived. Moreoever, when data is sparse, the uncertainty in conclusions may be expressed formally in terms of confidence bounds for the parameters. Furthermore, those parameters may themselves be regressed over changes in the attributes.

Our numerical experiments have shown that the standard classical distributions underestimate the proportion of vehicles with low mileage rates, since typically their probability density functions (pdfs) satisfy $f(0) = 0$. However, it appears that the data are fitted well by a shifted Gamma distribution whose pdf takes the form

$$f(r) = c(r + a)^k \exp(-\lambda r).$$

Here r is the mileage rate, whereas a, k, $\lambda \geq 0$ are the parameters to be fitted, and $c = c(a, k, \lambda)$ is a normalisation constant required to ensure that $\int_0^\infty f(x)\,\mathrm{d}x = 1$ (which in practice may be computed by numerical quadrature). The optimal fits of a, k, and λ may be obtained by numerical maximisation of the corresponding log likelihood function using library routines for non-convex optimisation. Likewise, confidence intervals may be derived most directly (but with significant computational cost) by numerical bootstrapping of the MLE procedure.

5. Temporal analysis

We now consider the question of how to compute mileage rates as a function of time. Our chief problem is that consecutive MOT tests are typically one year apart. Hence nothing can be said about how the mileage of an individual vehicle distributes itself over time scales shorter than one year. But can something better be achieved at the population level?

The obvious (but flawed) approach is to consider a sequence of observation times t_i^*, $i = 1, 2, \ldots$, and for each follow through the straddling procedure developed in section 4, so as to compute corresponding average average mileage rates $\overline{r}_i$. Thus via the pairs $(t_i^*, \overline{r}_i)$, we reconstruct $\overline{r}(t)$, which is the object of interest. Since there appears to be no constraint on the choice of observation times t_i^*, this method has apparently arbitrary temporal resolution, even down to the level of a single day. Figure 3 shows the sorts of results that can be obtained by following this procedure.

However the approach outlined above has a fundamental problem, as follows. Consider figure 1 and let us assume a simple model where consecutive MOT tests are always exactly one year apart[3]. Thus when we compute data for a given observation time t^*, we gather data from intervals in which the first test occurred up to one year before t^*, and where the second test occurred up to one year after t^*. Thus $\overline{r}(t^*)$ as derived by the straddling procedure incorporates mileage incurred on the interval $t^* - 1 \leq t < t^* + 1$, where time is expressed in years. In practice, patterns of vehicle usage might change substantially during a two year period — for example, due to fluctuations in the economy or the price of fuel. In consequence, the straddling method does not recover the true *spot* mileage rate, which represents variations in aggregate usage over short time scales.

[3]Note that this assumption is only an approximation and in practice there is a distribution of inter-test intervals which peaks at one year but which has some spread. Furthermore, there is an implicit assumption in these calculations that MOT tests occur at a constant rate throughout the calendar — whereas in fact they peak in the Spring and Autumn, to coincide with the anniversary of the peaks in the distribution of the dates of first registration. The modelling of these details requires further work.

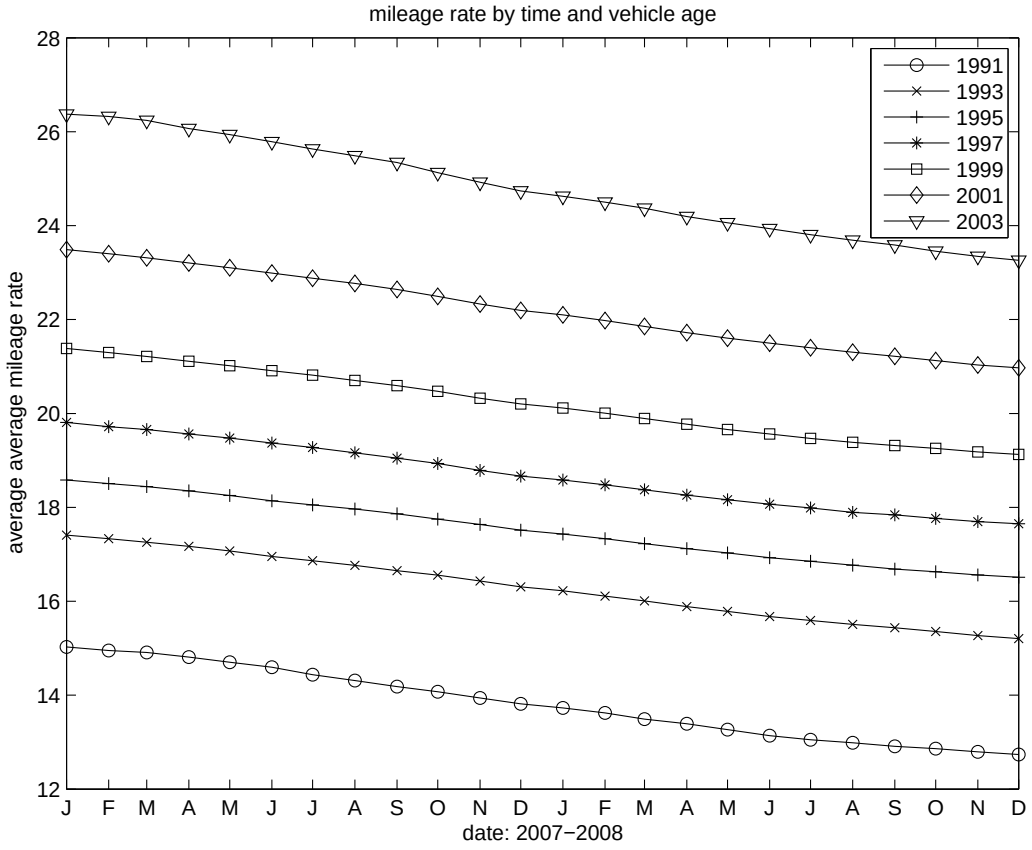

Figure 3: The average average mileage rate $\overline{r}$ as a function of time t, through 2008–2009, derived by straddling intervals, and disaggregated over the year of first registration. By rolling each successive plot two years to the right, one can visualise the gradual decline in mileage rates as vehicles age over a fourteen year span. However, as we have discussed in the main text, these results do not represent true *spot* mileage rates.

We shall proceed as follows. Let us suppose that there exists a spot average mileage rate $\phi(t)$ that we aim to discover, and let $\overline{r}(t)$ denote the mileage rate which is easily computed by the straddling procedure as described above. How are the two quantities related?

To answer this question, we need a little more precision on what the average spot mileage rate should mean. Specifically, we envision that each individual vehicle shall have an instantaneous mileage rate $\phi_i(t)$ which is modulated by $\phi(t)$ in the same way. We may write

$$\phi_i(t) = c_i \phi(t) + \text{noise},$$

where c_i is a constant (for that vehicle) which denotes its level of usage relative to the rest of the vehicle population. Here we incorporate a noise term to model the random (short time-scale) fluctuations of the individual that are not incorporated in this simple idealisation. We require $\langle c_i \rangle = 1$ and $\langle \text{noise} \rangle = 0$, so that $\phi = \langle \phi_i \rangle$ holds in the natural way.

Let $\psi_i(\tau)$ denote the miles driven by vehicle i between tests at $\tau - 1/2$ and $\tau + 1/2$, so that

$$\begin{aligned}\psi_i(\tau) &= \int_{\tau-1/2}^{\tau+1/2} (c_i\phi(s) + \text{noise})\ \mathrm{d}s, \\ &= c_i \int_{\tau-1/2}^{\tau+1/2} \phi(s)\,\mathrm{d}s,\end{aligned}$$

if the noise has the appropriate zero average property. The point is that $\overline{r}(t)$ may now be written in terms of sums (over the vehicle population) of terms of type $\psi_i(\tau)$, for τ running from $t - 1/2$ to $t + 1/2$. If we make the simplifying assumptions that tests occur at a constant rate throughout the year (which needs some refinement) and that the time of year of the test does not effect the distribution of c_i, then those constants average out and we may write

$$\overline{r}(t) = \int_{t-1/2}^{t+1/2} \int_{\tau-1/2}^{\tau+1/2} \phi(s)\,\mathrm{d}s\,\mathrm{d}\tau.$$

This integral may be simplified by reversing the order of integration, to yield

$$\overline{r}(t) = \int_{t-1}^{t+1} w(s;t)\phi(s)\,\mathrm{d}s,$$

where $w(s)$ is a kernel function with triangular shape, defined by

$$w(s;t) = \begin{cases} s - (t-1) & \text{for} \quad t-1 < s < t, \\ (t+1) - s & \text{for} \quad t < s < t+1, \\ 0 & \text{otherwise.} \end{cases}$$

In consequence, we may easily pass from the spot mileage rate $\phi(t)$ to the 'straddling' rate $\overline{r}(t)$, but the reverse direction (i.e., the required one) requires the inversion of an integral equation. This is a non-standard problem because both the integrand and the limits involve the independent variable t and the (non-smooth) kernel is zero at the end-points of the domain. However, one may show that

$$\overline{r}''(t) = \phi(t+1) - 2\phi(t) + \phi(t-1),$$

where $''$ denotes the double derivative, and in principle this equation can be used to time-step a solution for ϕ. However, this method requires the provision of two years of initial data for $\phi(t)$. Moreover, the (numerical) differentiation of $\overline{r}(t)$ is problematic, since the computation of $\overline{r}(t)$ is itself subject to statistical sampling error. The solution to these problems lies in the development of a new $\overline{r}(t^*)$ that weights the intervals that straddle t^* differently according to their end-points. The details are in development and beyond the scope of this paper.

6. Conclusions and further work

We live in a society in which digital data has become ubiquitous. It may often be the case that data collected either accidentally, or for a single narrow purpose may have far-reaching benefits that could not have been anticipated originally. It is in this spirit that the UK government has developed an open data initiative (see http://data.gov.uk) on the premise that government-owned data should be released in to the public domain unless there is a pressing reason to the contrary (e.g., individual

privacy, or protection of over-riding commercial interests) — since only with exposure to many eyes will the full potential of any given data set be realised. The release of MOT results data, and in particular the inclusion of odometer readings in this data, has enormous potential in this regard.

During this study, it has been possible to undertake a detailed study of odometer readings and mileage rates provided in the dataset. A large cleaned database (consisting of 76 million records) of inter-test intervals has been created, which can be polled to answer questions concerning mileage rates. The development of a straddling method to extract distributions of mileage rates for given vehicle populations at a given observation point in time provides numerous opportunities. Here we have presented a simple example which illustrates differences between vehicle usage in rural and urban areas. In particular, this example shows that older vehicles are driven far less than was previously thought, with potentially significant implications for policies which promote scrappage of those vehicles in order to reduce carbon emissions. In addition, methods for modelling the statistical distribution of mileage rates across a vehicle population have been developed. However, the challenge of inferring changes in mileage over short time scales (e.g., of the order of a month or so) remains open.

Many of the remaining challenges concern ways in which this study can be enhanced by the addition of extra data, which is presently not in the public domain. For example, this might include the addition of vehicle emissions data to the dataset, finer scale locational information about where vehicles are registered, and additional information relating to the vehicle keeper and their households. Integration and interrogation of this information, in combination with other data sets, not least the 2011 Census data, would also provide further insights. Finding ways of understanding the travel of vehicles of less than 3 years old, is also key in order to provide a complete picture of national vehicle ownership and use.

In brief, then, the technical challenges involved in using MOT data as an alternative to more conventional sources of information on car ownership and use are considerable. However, this paper demonstrates that useful information can be derived, and future activities are planned to explore the potential to develop this work, in collaboration with relevant Government agencies. The end goal — a readily available set of tools for understanding changes in car ownership, use and emissions — from information that is already available — could be of value in all countries where this sort of information is collected as a matter of course.

Acknowledgments

This work was supported by EPSRC grant ref. EP/J004758/1 *Using MOT test data to analyse travel behaviour change: part 1 - scoping study.* REW also acknowledges the support of an EPSRC Advanced Research Fellowship (grant ref. EP/E055567/1). Grateful thanks are due to members of DfT, VOSA, DVLA and DECC, who have provided advice and support for this and for future work in this area. Finally, many thanks to Ecolane and nextgreencar.com for their advice on vehicle classification systems.

References

Adjemian, M.K., Cynthia Lin, C.-Y. and Williams, J. 2010. *Estimating spatial interdependence in automobile type choice with survey data.* Transportation Research Part A 44, pp661-675.

Cairns, S., Wilson, R.E., Chatterton, T., Anable, J., Notley, S. and McLeod, F., 2011. *Using*

MOT data to analyse travel behaviour change — scoping report. TRL PPR578. Bracknell: IHS, ISSN 0968-4093.

Clark, S.D. 2004. *Estimating car ownership using geographically weighted regression.* Traffic Engineering and Control 45(12), pp416–420.

Clark, S.D. 2007. *Estimating local car ownership models.* Journal of Transport Geography 15, pp84–197.

Dargay, J. and Vythoulkas, P. 1999. *Estimation of a dynamic car ownership model, a pseudo panel approach.* Journal of Transport Economics and Policy 33(3), pp287–302.

Dargay, J., 2001. *The effect of income on car ownership: evidence of asymmetry.* Transportation Research Part A 35(9), pp807-821.

Dargay, J., 2002. *Determinants of car ownership in rural and urban areas: a pseudo-panel analysis.* Transportation Research Part E: Logistics and Transportation Review 38(5), pp351-366.

de Jong, G., Fox, J., Pieters, M., Vonk, L. and Daly, A., 2004. *Comparison of car ownership models.* Transport Reviews 24(4), pp379–408.

Department for Transport, 2010a. *Annual road traffic estimates: methodology.*
Downloaded from `http://www.dft.gov.uk/statistics/series/traffic/`.
Page checked 25 July 2012.

Department for Transport, 2010b. *Anonymised MOT tests and results.*
Downloaded from `http://www.dft.gov.uk/data/release/10007`.
Page checked 3 January 2012.

Eddington, R., 2006. *The Eddington Transport Study.* HMSO, London.

Element Energy, 2009. *Strategies for the uptake of electric vehicles and associated infrastructure implications for the Committee on Climate Change final report.* Cambridge: Element Energy.

Huang, B. 2005. *Car demand forecasting using dynamic pseudo panel model.* Proceedings of the European Transport Conference, 2005.

Millard Ball, A. and Schipper, L. 2010. *Are we reaching peak travel? Trends in passenger transport in eight industrialized countries.* Transport Reviews 31(3), pp357–378.

Mokhtarian, P. and Cao, X. 2008. *Examining the impacts of residential self-selection on travel behaviour: a focus on methodologies.* Transportation Research Part B 42, pp204–228.

Office for Low Emission Vehicles, 2011. *Making the Connection. The Plug-in Vehicle Infrastructure Strategy.* London: OLEV.

Sloman, L.; Cairns, S., Newson, C., Anable, J.; Pridmore, A. and Goodwin, P., 2010. *The effects of smarter choice programmes in the Sustainable Travel Towns.* Report to the Department for Transport, London.

Whelan, G. 2007. *Modelling car ownership in Great Britain.* Transportation Research Part A 47, pp205–219.

Zegras, C. 2010. *The built environment and motor vehicle ownership and use: evidence from Santiago de Chile.* Urban Studies 47(8), pp1793–1817.