Available online at www.sciencedirect.com

SciVerse ScienceDirect

Procedia - Social and Behavioral Sciences 80 (2013) 139 – 156

Procedia
Social and Behavioral Sciences

On the Estimation of Temporal Mileage Rates

R. E. Wilson[a,*], J. Anable[b], S. Cairns[c,d], T. Chatterton[e], S. Notley[c], J. D. Lees-Miller[a]

[a]*Transport and Mobility Modelling Group, University of Bristol, UK*
[b]*Centre for Transport Research, University of Aberdeen, UK*
[c]*TRL, UK*
[d]*Centre for Transport Studies, University College London, UK*
[e]*Air Quality Management Centre, University of the West of England, UK*

Abstract

Mathematical and computational techniques are developed for the analysis of annual MOT (roadworthiness) test data that the UK Department for Transport has placed in the public domain. This paper focusses on the development of a new theory which has the potential to estimate fine-scale temporal variations (e.g., monthly) in vehicle mileage at a population level, that we call the *spot rate* — derived from coarse-scale (e.g., annual) mileage data at an individual vehicle level. Due to the availability of data, the focus is on the UK situation, but the theory has applications to any data set internationally, where odometer readings of individual vehicles are monitored on an occasional basis. Numerical time-stepping schemes are derived from the theory and are tested on synthetic data to permit comparison with a known ground-truth mileage rate. It is found that for practical applicability, the methods need to pre-process data with smoothing filters (a full investigation of which is beyond the scope of this paper). Finally, we consider first steps in applying the methods directly to the MOT data set and the remaining problems that must be solved for them to become a practical reality.

Selection and peer-review under responsibility of Delft University of Technology

Keywords:

1. Introduction

Driven by a need to reduce emissions of greenhouse gases in order to mitigate climate change, there is an increasing policy interest in initiatives that reduce car use and encourage people to own cleaner, more efficient vehicles. These initiatives range from large-scale national projects, such as the promotion of electric vehicles through to small-scale projects, such as schemes that make it more attractive to walk or cycle in a given local area. But how can we tell whether these initiatives are effective, and if they are, how can we determine the scale of their effects? And how do the effects vary over time and with geography? (For example, between regions; between urban and rural areas; between different towns?)

*Corresponding author. Full address: Department of Engineering Mathematics, Faculty of Engineering, University of Bristol, BRISTOL BS8 1TR, United Kingdom.

Email address: RE.Wilson@bristol.ac.uk (R. E. Wilson)

Selection and peer-review under responsibility of Delft University of Technology
doi:10.1016/j.sbspro.2013.05.010

To answer these questions, we require robust data concerning vehicle ownership and usage, and to date there have been two main sources. These are (i) surveys of individuals or households — which are potentially subject to bias due to misreporting (either deliberate or accidental) and who chooses to respond; and (ii) on-street traffic counts, which are subject to bias in the manner in which counting sites are chosen — since an estimation procedure is required to convert spot counts into some measure of total vehicle kilometres travelled [2]. Furthermore, automatic traffic counts cannot robustly disaggregate traffic over vehicle classes, or (for example) identify the fuel type or engine capacity of the counted vehicles.

Thus there is a demand for new estimation techniques which are either cheap, or perhaps even free — because they are based on secondary analysis of data that was originally collected for other purposes. In the United Kingdom, one such source of data originates from the annual roadworthiness test (known as the MOT test, since it used to stand for 'Ministry of Transport') which since 1967 has been compulsory for vehicles over three years old. In 2005, a computerised system was introduced for reporting the MOT test results and storing them in a Department for Transport (DfT) database. In November 2010, the DfT published this data [3] — consisting of the results of approximately 150 million MOT tests from 2005 to the Spring of 2010. Although some fields, such as vehicle registration plates and unique Vehicle Test Station (VTS) identities, have been withheld from the published data, what remains still contains a wealth of information that is not available elsewhere. In addition to the results of the test itself, the data include: the date of the test; 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[1]. Consequently one may infer the mileage of a vehicle between a pair of tests on known dates. Our purpose is to analyse this data to provide an important missing link in the analysis of vehicle usage. Note that although the focus here is on the UK situation because of data availability, the methods that we develop are generic and could be applied internationally to any data set where the odometer readings of individual vehicles are monitored from time to time. (A situation where vehicles are monitored at regular intervals is directly addressable by our technique; irregular monitoring intervals would require further extensions to the theory.)

The simplest sorts of questions that we may address with this new data source are *static* (i.e., time-independent) ones, see [1, 4]. For example, we might compare vehicle usage in different towns, or between different classes of vehicle (e.g., SUVs versus minis) — in such examples, we might expect the comparison to be relatively slowly varying in time, thus justifying the use of a static measure assigned to a single calendar year. To answer these questions, a statistic that we develop in Section 2, that we call the *straddling rate*, is sufficient.

However, more interesting questions are dynamic (i.e., time-dependent) and concern how usage is changing with time. If these changes occur over rapid time-scales, the *straddling rate* is unable to capture them and the goal of this paper is to develop and demonstrate more sophisticated measures for calculating the temporal dependence of vehicle usage from the MOT data. As an over-arching challenge, we would like methods that might identify sharp changes in vehicle usage due to (for example) fluctuations in the price of fuel, or perhaps more severely, economic crises such as that resulting from the collapse of the banking system in Autumn 2008. Since data from automatic counters will give some measure of these effects, the potential added value in the MOT data is examining how these changes distribute themselves over vehicles of different ages and classes and the locations in which they reside.

At first sight, the challenge seems impossible, and the fact that it might potentially be viable is the major surprise and fresh contribution of this paper — to the best of our knowledge, this idea is in entirely virgin territory. The seemingly intractable problem is that we cannot know from MOT data how an individual vehicle's mileage is distributed between two consecutive tests (typically one year apart): most likely that distribution is highly non-uniform. For example, a typical vehicle may have a rather constant level of underpinning mileage due to (e.g.) regular commuting trips, that is interspersed with a heavy tail of occasional

[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).

long distance trips, whose occurrence is essentially unmodellable. So how can we use this kind of annual mileage data to infer anything about mileages over time-scales of months?

The answer is that whilst a fine-scale temporal analysis is impossible at the level of individual vehicles, it *might be* possible at the population level. To this end, we shall introduce the concept of the population *spot rate*, which is a mileage rate common to all vehicles in some segment that we wish to analyse, and which captures in average terms how their usage is modulated in time. Clearly, individual vehicles have different levels of total usage, which we may model by assigning to them different (time-independent) proportional factors of the population spot rate. Further, individual vehicles have apparently random temporal fluctuations in their usage as we have described above. However, our assertion is that in some kind of population-averaged way, the spot rate modulates the usage of all vehicles in the given segment, and describes how that usage depends on time. The *spot rate* is formulated in mathematical terms in Section 3, which also develops a calculus that computes the *straddling rate* in terms of the *spot rate*, in the form of an integral equation.

However, since the straddling rate can be computed directly from the MOT data, the desired result is rather a formula that gives the spot rate in terms of the straddling rate, and this theory is the subject of Section 4. The output is a possible time-stepping scheme for evolving the spot rate forward in time. However, a drawback is that two years of initial data for the spot rate are required to start the scheme up. Hence Section 5 considers generalisations to the definition of the straddling rate, which are also directly computable from MOT data, but which have a crisper relationship with the spot rate, and which consequently might be more amenable to time-stepping schemes: a detailed presentation of the resulting schemes is then given in Section 6.

Section 7 then tests the proposed time-stepping schemes on synthetic data. The value in synthetic data is that it may be generated from a known (i.e., ground-truth) spot rate that we may then attempt to reconstruct. Because the synthetic data is 'idealised' (this means that certain approximations to the real-world data that are used in the development of the theory hold exactly), we may focus attention on a particular difficulty of the schemes which involves a trade-off between truncation error and statistical sampling error. In particular, although our method appears feasible for large enough data sets (which tend to reduce the problems due to sampling error), it requires the use of smoothing filters — a full investigation of which are beyond the scope of this paper.

Section 8 is concerned with the potential application of our method to the real-world MOT data [3]. Because of the difficulties with sample sizes and smoothing techniques, this work is at a very early stage. In addition, due to the breakdown of simplifying assumptions used in the theory, various extensions are desirable that are discussed in Section 9, that also presents the conclusions.

2. Intervals and the straddling rate

Full details of the UK data, the linking procedure, and cleaning methods are provided in [4]. In short — the basic output of those procedures is a set of (approximately 76 million) *intervals*, between consecutive pairs of tests for the same vehicle. Each interval i has

- the date $t_1^{(i)}$ and odometer reading $x_1^{(i)}$ (miles) at the first test;
- date $t_2^{(i)}$ and odometer reading $x_2^{(i)}$ (miles) at the second test; note that $t_2^{(i)} > t_1^{(i)}$ and $x_2^{(i)} \geq x_1^{(i)}$;
- (coarse) locations of tests, giving an indication of where the vehicle typically resided; and vehicle specific attributes, such as make, model, engine size, fuel type etc. Our only use for these attributes is to consider sub-selections of the full data-set. For example, we might be interested in comparing the usage of vehicles of different ages that are tested in a particular rural area. In this situation, we shall let $\mathcal{I}$ denote the subset of intervals whose vehicles match our specification. In such applications, $\mathcal{I}$ may be very much smaller than the full set of intervals.

Note that due to regulatory requirements, $t_2^{(i)}$ is typically about one year after $t_1^{(i)}$ (but not always so). Throughout we use units of years, so we may write $t_2^{(i)} \simeq t_1^{(i)} + 1$. For a given interval i, the average

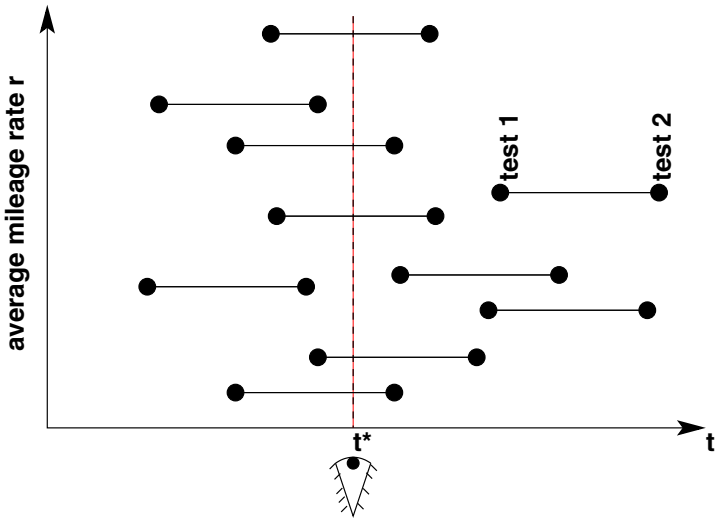

Fig. 1. Analysis of mileage rates at an *observation time* t^*. Those intervals which *straddle* t^* are selected. Each then contributes an average mileage rate r_i to subsequent computations.

(annual) mileage rate $r^{(i)}$ of the vehicle in question in this interval is thus given by

$$r^{(i)} := \frac{x_2^{(i)} - x_1^{(i)}}{t_2^{(i)} - t_1^{(i)}}. \tag{1}$$

How should such inter-test rates be processed? The simplest question concerns how one should infer characteristics of mileage at a given fixed *observation date* t^*. The natural procedure is to select all intervals i which contribute mileage on that date — that is those which *straddle* the observation date, in that $t_1^{(i)} \le t^* < t_2^{(i)}$, where the half-open interval assures the correct counting. See Figure 1. As remarked above, we may narrow the selection further according to vehicle attributes by also requiring that i is a member of some special set $\mathcal{I}$.

In practice, this procedure will yield a (typically large) number M of intervals and corresponding average mileage rates $r_1, r_2, \ldots, r_M$. One may then characterise the corresponding vehicle usage by the average average mileage rate

$$\bar{r} := \frac{1}{M} \sum_{i=1}^{M} r_i, \tag{2}$$

where we mean average over time for each vehicle, averaged over the relevant vehicle population. Because of the way in which $\bar{r}$ is generated, we shall refer to it as the *straddling rate* (computed at time t^*). As we demonstrated in [4], it is a satisfactory measure for making (static) comparisons of usage between different segments of the vehicle population.

However, the focus of this paper is on the temporal dependence of mileage over short time scales. Can the straddling rate be used for this purpose? The obvious approach is to consider a sequence of observation times t_j^*, $j = 1, 2, \ldots$, and for each follow through the straddling procedure developed above, so as to compute the corresponding average average mileage rates $\bar{r}_j$. Thus via the pairs $(t_j^*, \bar{r}_j)$, we reconstruct $\bar{r}(t)$. Since there appears to be no constraint on the choice of observation times t_j^*, this method has apparently arbitrary temporal resolution, even down to the level of a single day.

However, the approach is flawed, in that the temporal resolution is not what it claims to be. In practice, each $\bar{r}_j$ incorporates miles driven over a two-year span $t_j^* - 1 \le t < t_j^* + 1$, incorporating intervals where the first test is about one year before the given observation date ranging through to intervals where the second test is one year after the given observation date. So rather than give a 'spot' (or instantaneous) value for vehicle usage, $\bar{r}(t)$ gives instead a kind of locally time-averaged picture.

3. From the spot rate to the straddling rate

Suppose that there exists a spot population average mileage rate $\phi(t)$ that we aim to discover. How are $\phi(t)$ and the straddling rate $\overline{r}(t)$ related? Recall our discussion of the spot rate in Section 1. Then let us envision that each individual vehicle k has an instantaneous mileage rate $\phi_k(t)$ which is modulated by $\phi(t)$ in the same way, so that

$$\phi_k(t) = c_k\phi(t) + \text{noise}, \tag{3}$$

where c_k 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. We require $\langle c_k\rangle = 1$ and $\langle\text{noise}\rangle = 0$, so that $\phi = \langle\phi_k\rangle$ holds in the natural way.

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

$$\psi_k(\tau) = \int_{\tau-1/2}^{\tau+1/2} (c_k\phi(s) + \text{noise})\ \mathrm{d}s, \qquad = c_k \int_{\tau-1/2}^{\tau+1/2} \phi(s)\,\mathrm{d}s, \tag{4}$$

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 $\psi_k(\tau)$, for τ running from $t - 1/2$ to $t + 1/2$. By their construction, the constants c_k average out and we may write

$$\overline{r}(t) = \int_{t-1/2}^{t+1/2} \int_{\tau-1/2}^{\tau+1/2} \phi(s)\,\mathrm{d}s\,\mathrm{d}\tau. \tag{5}$$

Note that behind the scenes in this calculation are some implicit assumptions which are idealistic approximations of the real-world MOT data, namely

A1 We assume that tests are exactly one year apart: i.e., for each interval i, $t_2^{(i)} = t_1^{(i)} + 1$ *exactly*. In practice, there is some variation in the lengths of intervals, due to owners presenting their vehicles for testing either ahead of the regulatory deadline, or by narrowly missing it.

A2 We assume that tests occur at the same frequency on average throughout the year. (In practice there is a lull over the Christmas holiday period and a double peak in the Spring and Autumn. In the UK, new car sales have traditionally peaked at these times to correspond with dates on which new registration plate marks are released. Vehicles are presented for their first test at three years old, and subsequently every one year, approximately — hence the frequency of MOT test dates is related to the frequency of first registration dates.)

A3 We assume that a vehicle's mileage rate is independent of the time of year at which it takes its MOT tests.

Possible extensions to the theory in which these assumptions are relaxed are considered in Section 9.

Proceeding with the calculation, the order of the double integral (5) may now be reversed by noting (see Figure 2) that

$$\int_{t-1/2}^{t+1/2} \int_{\tau-1/2}^{\tau+1/2} \mathrm{d}s\,\mathrm{d}\tau = \int_{t-1}^{t} \int_{t-1/2}^{s+1/2} \mathrm{d}\tau\,\mathrm{d}s + \int_{t}^{t+1} \int_{s-1/2}^{t+1/2} \mathrm{d}\tau\,\mathrm{d}s. \tag{6}$$

This reversal simplifies matters, because $\phi(s)$ may then be pulled outside of the inner integrals, which can be solved directly (since their integrands are unity). We may thus obtain

$$\overline{r}(t) = \int_{t-1}^{t+1} K(s;t)\phi(s)\,\mathrm{d}s, \tag{7}$$

where $K(s;t)$ is a kernel function with triangular shape, defined by

$$K(s;t) = \begin{cases} s - (t-1) & \text{for} \quad t-1 < s < t, \\ (t+1) - s & \text{for} \quad t < s < t+1, \\ 0 & \text{otherwise.} \end{cases} \tag{8}$$

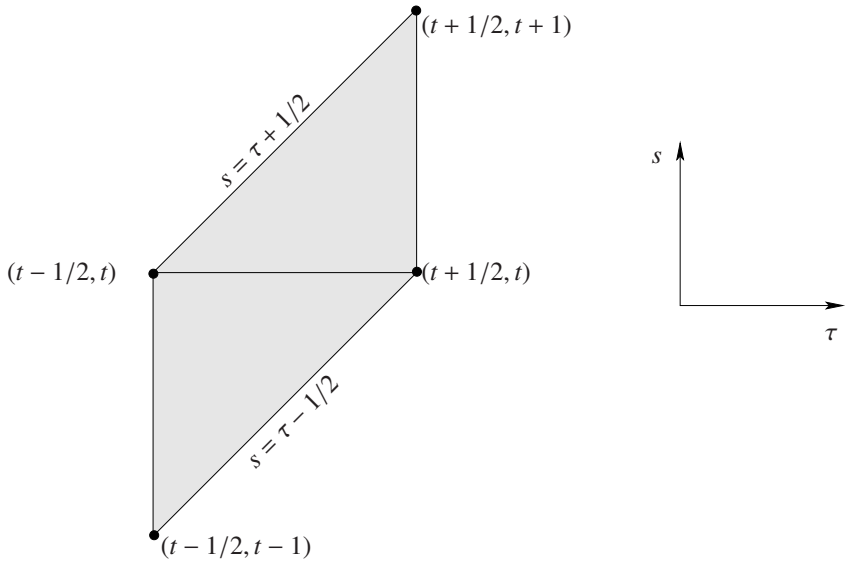

Fig. 2. The domain of integration for the double integral in equation (5). To reverse the order of integration so that s is the outer integral, the domain needs dividing into two pieces to constrain discontinuities to the end-points of the consequent one-dimensional integrals. See equation (6).

We have therefore derived a formula for the straddling rate $\overline{r}(t)$ in terms of the spot rate $\phi(t)$. However, we require the reverse direction, namely a formula for the spot rate in terms of the straddling rate (since the latter is derived easily from data), and we will turn to this question in Section 4.

Note that a slight simplification of (7,8) is obtained by exploiting the translational invariance, so that with $u = s - t$, we obtain

$$\overline{r}(t) = \int_{-1}^{+1} k(u)\phi(u + t)\,\mathrm{d}u, \tag{9}$$

where the new kernel function is given by

$$k(u) = \begin{cases} u + 1 & \text{for} \quad -1 < u < 0, \\ 1 - u & \text{for} \quad 0 < u < +1, \\ 0 & \text{otherwise.} \end{cases} \tag{10}$$

Finally, note that the 'triangular-shaped' kernel obtained by these formula should not come as a surprise. It simply expresses how the straddling rate $\overline{r}(t^*)$ incorporates mileage over a two-year period, with a strong weighting to the mileage incorporated at the observation time t^* (since all the selected intervals contain this value) and a weighting that diminishes to zero at the extremes $t^* - 1$ and $t^* + 1$ of the two-year window. This is because only a small proportion of the one-year windows that straddle t^* will also straddle values just above $t^* - 1$ (or alternatively values just below $t^* + 1$).

4. From the straddling rate to the spot rate

Because t appears only inside the integrand of (9), we may differentiate with respect to t and carry the differentiation through the integral to obtain

$$\overline{r}'(t) = \int_{-1}^{+1} k(u)\phi'(u + t)\,\mathrm{d}u, \tag{11}$$

where we suppose that the spot rate $\phi(t)$ is continuously differentiable. This step alone does not isolate $\phi(t)$, and we must continue with further applications of differentiation combined with integration by parts. For this procedure we shall suppose that $\phi(t)$ is twice continuously differentiable, but care is required because $k'(u)$ has a discontinuity at the midpoint of the domain. The correct approach is to partition the range of integration in (11) so that $\overline{r}'(t) = I_- + I_+$, where I_- and I_+ are respectively the integrals from -1 to 0 and from 0 to $+1$, so that the integrands in each integral are smooth in the interior of their respective domains.

We then have

$$I_+(t) = \int_0^{+1} k(u)\phi'(u+t)\,\mathrm{d}u, \qquad = \Big[k(u)\phi(u+t)\Big]_{u=0}^{u=+1} - \int_0^{+1} k'(u)\phi(u+t)\,\mathrm{d}u, \tag{12}$$

on applying integration by parts. Since $k(0) = 1$, $k(+1) = 0$ and $k'(u) = -1$ on the given range, we obtain

$$I_+(t) = -\phi(t) + \int_0^{+1} \phi(u+t)\,\mathrm{d}t, \tag{13}$$

so that

$$I'_+(t) = -\phi'(t) + \int_0^{+1} \phi'(u+t)\,\mathrm{d}u, \tag{14}$$

$$= -\phi'(t) + \phi(t+1) - \phi(t). \tag{15}$$

A similar calculation yields $I'_-(t) = +\phi'(t) + \phi(t-1) - \phi(t)$, so that

$$\overline{r}''(t) = \phi(t+1) - 2\phi(t) + \phi(t-1). \tag{16}$$

In effect, we have achieved our end goal in that we may re-arrange (16) and replace t with $t-1$ to obtain

$$\phi(t) = \overline{r}''(t) + 2\phi(t-1) - \phi(t-2), \tag{17}$$

so that we have expressed the spot rate in terms of the straddling rate (or rather, its second derivative) and values of the spot rate at earlier times. Since in principle $\overline{r}''(t)$ can be derived from $\overline{r}(t)$, which in turn can be derived from the source data, we have a procedure for time-stepping the spot rate $\phi(t)$ as new source data accumulates, *provided* two years of initial data for it are also provided. This is an insuperable problem: one cannot determine the spot rate in absolute terms unless one has at least *some* initial data for it.

In practice, initial data for the spot rate must be derived by triangulating from other sources, such as traffic counts or surveys, and will thus be prone to error. This error will propagate forward to the absolute value of $\phi(t)$ at later times. However, we are more interested in measuring *changes* in the spot rate, and it will turn out that schemes based on (17) *might* be able to do this robustly, even if the absolute value is wrong.

Finally, let us make some observations concerning the structure of (16). Firstly, note that the right-hand side is effectively a discrete second derivative for $\phi(t)$, and a coarse approximation to $\phi''(t)$ when it is sufficiently slowly varying. In concise terms, Taylor's theorem gives

$$\phi(t+1) - 2\phi(t) + \phi(t-1) = \phi''(t) + \frac{1}{3}\phi'''(\xi), \tag{18}$$

for some ξ with $t-1 < \xi < t+1$. So in particular, if $\phi(t)$ were quadratic in t, we would have $\overline{r}''(t) = \phi''(t)$, and hence $\phi(t) = \overline{r}(t) + At + B$. Alternatively, in situations where ϕ were linear in t (for at least a two-year duration), then we would obtain $\overline{r}''(t) = \phi''(t) = 0$, and in this (rather trivial) case $\phi(t) \equiv \overline{r}(t)$.

However, a more realistic situation is that the spot rate combines a slow trend (over many years) with a strong periodic component due to seasonal effects. We might model this profile in the form

$$\phi(t) = At + B + f(t), \tag{19}$$

where f has period one year (so that $f(t+1) = f(t)$ for all t) and mean zero. For such a profile, (16) also yields $\overline{r}''(t) = 0$. The consequences of this result are two-fold. Firstly, this implies that if one's interest is in simple static (time-independent) statistics, then one may compute $\overline{r}(t^*)$ for a single value of t^*, and not worry about time-of-year effects in the choice of t^*, since they are effectively averaged out to zero by the triangular kernel. The second consequence is a negative one: there is unfortunately no way of using $\overline{r}(t)$ by itself to infer the seasonal structure of the spot rate. The problem (assumption **A1** page 5) is that the interval between tests (one year) is exactly the same as the period of the effect that we are trying to discover. In real-world data, assumption **A1** does not hold exactly, and potentially one might exploit this feature to infer seasonal structure.

However, our chief interest is to detect strong deviations from the usual seasonal profile, over relatively short time-scales, and this, rather than analytical solutions to pathological special cases, is the main focus of the remainder of the paper.

5. Weighted straddling rates

As we have discussed, formula (17) provides the basis for methods that time-step solutions to the spot rate $\phi(t)$ as new data accumulates. However, such methods require two years of initial data for $\phi(t)$, which seems over-strong from the practical point of view. The requirement is of course a consequence of the use of the straddling rate which as presently defined incorporates spot mileage over a two year interval. The present section is devoted to a more general definition of the straddling rate which leads to time-stepping schemes that may operate with a shorter start-up interval.

As before, we wish to consider inter-test intervals (t_1, t_2) that *straddle* a given observation time t^*, so that $t_1 < t^* < t_2$. However, whereas the previous definition of $\overline{r}(t)$ averaged the mileage rates of all such matching intervals in a uniform way, we now wish to allow that they be weighted according to the values of t_1 and t_2 relative to t^*. In our present idealised setting, we assume that $t_2 = t_1 + 1$, and hence we introduce the weight function

$$w(v) \quad \text{for} \quad -1/2 < v < 1/2 \quad \text{where} \quad v := (t_1 + t_2)/2 - t^*, \tag{20}$$

which is used to weight the contributions of respective intervals in a straddling rate that we now denote $\overline{r}_w(t)$. For our various integral arguments to work through gracefully, we require

$$\int_{-1/2}^{+1/2} w(v)\,\mathrm{d}v = 1. \tag{21}$$

In particular, the previous case where intervals were unweighted corresponds to the trivial weight function

$$w(v) \equiv 1 \quad \text{for} \quad -1/2 < v < +1/2. \tag{22}$$

The idea is now to re-work the previous analysis to link the spot rate with the weighted straddling rate. We then hope to find that weight functions which focus attention on intervals that only *just* straddle t^*, so that t_2 is only slightly beyond t^*, will turn out to be useful, and to reduce the requirement for initial data for $\phi(t)$.

In this vein, the basic steps of Section 3 may be re-worked to obtain

$$\overline{r}_w(t) = \int_{-1}^{+1} k_w(u)\phi(u+t)\,\mathrm{d}u, \tag{23}$$

analogous to (9), where now

$$k_w(u) = \begin{cases} \int_{-1/2}^{u+1/2} w(v)\,\mathrm{d}v & \text{for} \quad -1 < u < 0, \\ \int_{u-1/2}^{+1/2} w(v)\,\mathrm{d}v & \text{for} \quad 0 < u < +1, \\ 0 & \text{otherwise.} \end{cases} \tag{24}$$

In particular, note that $k_w(-1) = k_w(+1) = 0$ is guaranteed automatically, and that $k_w(0) = 1$ is a result of condition (21). Furthermore, note that k_w is piece-wise continuously differentiable, with

$$k'_w(u) = \begin{cases} w(u+1/2) & \text{for} \quad -1 < u < 0, \quad \text{and} \\ -w(u-1/2) & \text{for} \quad 0 < u < +1. \end{cases} \tag{25}$$

Note that the previous results of Section 4 are special cases of these formula using the weight function (22).

Analogous to (11), we have

$$\overline{r}'_w(t) = \int_{-1}^{+1} k_w(u)\phi(u+t)\,\mathrm{d}u, \tag{26}$$

and we proceed as before by partitioning the range of integration so that discontinuities in the integrand are limited to the end points. Analogous to (12), we have

$$I_+(t) = \int_0^{+1} k_w(u)\phi'(u+t)\,\mathrm{d}u, \qquad = \Big[k_w(u)\phi(u+t)\Big]_{u=0}^{u=+1} - \int_0^{+1} k'_w(u)\phi(u+t)\,\mathrm{d}u, \tag{27}$$

$$= -\phi(t) + \int_0^1 w(u-1/2)\phi(u+t)\,\mathrm{d}u, \tag{28}$$

on applying integration by parts, and noting the derivative formula (25). As before, we may proceed by a further differentiation with respect to time, yielding

$$I'_+(t) = -\phi'(t) + \int_0^1 w(u-1/2)\phi'(u+t)\,\mathrm{d}u. \tag{29}$$

The difficulty that we now have is that we would like to continue with a further integration by parts, by integrating the $\phi'(u+t)$ term, and by differentiating the $w(u-1/2)$ term. However, the differentiability of the weight function w is not given. Moreover, the tractability of the earlier calculation was due to the fact that the weight function (22) yielded $w' \equiv 0$.

The way forward to counter these difficulties is to choose a piecewise-constant weight function, whose derivative (where it exists) is zero. The only remaining problem is then to partition up the integral according to the places where the weight function 'jumps'. A particularly good choice is

$$w_\alpha(u) = \begin{cases} 1/\alpha & \text{for} \quad -1/2 < u < -1/2 + \alpha \quad \text{and} \\ 0 & \text{otherwise,} \end{cases} \tag{30}$$

with $0 < \alpha \leq 1$. For α small, this weight function does indeed focus attention only on the intervals which *just* straddle past the given observation time.

With choice (30), equation (29) becomes

$$I'_+(t) = -\phi'(t) + \frac{1}{\alpha}\int_0^\alpha \phi'(u+t)\,\mathrm{d}u, \tag{31}$$

$$= -\phi'(t) + \frac{1}{\alpha}[\phi(t+\alpha) - \phi(t)]. \tag{32}$$

If we work through the process to compute I_-, that is, the contribution to $\overline{r}'(t)$ from the range -1 to 0 in (26), we obtain

$$I'_-(t) = \phi'(t) - \frac{1}{\alpha}[\phi(t-1+\alpha) - \phi(t-1)]. \tag{33}$$

Hence

$$\overline{r}''_w(t) = \frac{1}{\alpha}[\phi(t+\alpha) - \phi(t)] - \frac{1}{\alpha}[\phi(t-1+\alpha) - \phi(t-1)], \tag{34}$$

analogous to (16), which may be obtained from here by the special choice $\alpha = 1$. The remainder of the paper will apply this choice of weight function, and so henceforth we will use the notation $\overline{r}_\alpha(t)$ to denote the straddling rate with the weight function (30).

To repeat our earlier remarks — note that irrespective of the absolute value of $\phi(t)$, schemes based on (34) will detect relative changes in the spot rate, *compared to what those changes were in the past*. For example, in the case where $\alpha = 3$ months, formula (34) enables a robust comparison of the quarter-on-quarter change in the spot rate compared to the quarter-on-quarter change one year earlier. Via a simple re-grouping of the terms, this is equivalent to a comparison of the year-on-year change for two successive quarters. Finally, note that for α small, a simple Taylor expansion yields

$$\overline{r}''_\alpha(t) \simeq \phi'(t) - \phi'(t-1), \tag{35}$$

emphasising how $\overline{r}''_\alpha$ may be used to make year-on-year comparisons of the rate of change of the spot rate.

6. Implementation as a time-stepping scheme

We now consider how to implement time-stepping schemes based on the weighted straddling rate formula (34). The trick is to re-arrange so that the most recent value of ϕ appears as the 'subject' of the equation, that is

$$\phi(t+\alpha) = \phi(t) + \left[\phi(t-1+\alpha) - \phi(t-1)\right] + \alpha\overline{r}''_\alpha(t). \tag{36}$$

Similar to (17), this equation provides a procedure for advancing the spot rate $\phi(t)$ forward in time as new data accumulates and is incorporated through $\overline{r}''_\alpha(t)$. Whereas two years of initial data were required previously, now only one year seems to be required — up to a technical constraint involving the computation of $\overline{r}''_\alpha(t)$ that we discuss shortly.

For a practical implementation of (36), the simplest approach is to suppose that α divides exactly into one year, so that

$$\alpha = \frac{1}{N} \qquad \text{for some natural number } N = 1, 2, \ldots \text{etc., with units years}^{-1}. \tag{37}$$

We then let the scheme's time-steps be denoted by

$$t_n := t_0 + \frac{n}{N}, \tag{38}$$

for $n = 1, 2, \ldots$, and we let ϕ_n denote the approximations to ϕ that we will derive by this scheme, that is

$$\phi_n :\simeq \phi(t_n). \tag{39}$$

In this notation, formula (36) becomes

$$\phi_{n+1} = \phi_n + \left[\phi_{n-N+1} - \phi_{n-N}\right] + \frac{1}{N}\overline{r}''_{\alpha,n}, \tag{40}$$

for $n = 0, 1, 2, \ldots$, where $\overline{r}''_{\alpha,n}$ denotes the approximation to $\overline{r}''_\alpha(t_n)$ that we will derive from the interval data. The requisite one year of initial data for the spot rate must be supplied in the form

$$\phi_{-N},\ \phi_{-N+1},\ \phi_{-N+2},\ \ldots\ ,\phi_{-1}, \phi_0. \tag{41}$$

How should we estimate $\overline{r}''_\alpha(t_n)$ in (40)? As we remarked earlier, we may in principle use the weighted straddling procedure to compute $\overline{r}_\alpha(t)$ at any time t. Then we may apply

$$\overline{r}''_\alpha(t) \simeq \frac{1}{\epsilon^2}\left[\overline{r}_\alpha(t+\epsilon) - 2\overline{r}_\alpha(t) + \overline{r}_\alpha(t-\epsilon)\right], \tag{42}$$

for $\epsilon > 0$ sufficiently small. In fact, the approximation due to (42) is the sole source of truncation error for scheme (40), which is otherwise exact, since no approximations have been used in the earlier integration-based arguments. In principle, it seems that the truncation error can almost be eliminated by choosing as small a value of ϵ as possible. However, the values of $\overline{r}_\alpha(t)$, $\overline{r}_\alpha(t-\epsilon)$ and $\overline{r}_\alpha(t+\epsilon)$ themselves are derived

by sampling interval data, where the resolution of $t_1^{(i)}$ and $t_2^{(i)}$ is one day. Hence in practice, $\epsilon = 1$ day is an absolute minimum value, and a sensible choice is much larger than this, as we now explain.

In addition to the truncation error, the estimation of $\overline{r}''_\alpha$ is subject to statistical sampling error due to the finite set of interval data on which it is based. In contrast to the truncation error, the sampling error becomes *worse* as ϵ is reduced. To see this, we group terms in (42) to obtain

$$\overline{r}''_\alpha(t_n) \simeq \frac{1}{\epsilon^2}[\overline{r}_\alpha(t+\epsilon) - \overline{r}_\alpha(t)] - \frac{1}{\epsilon^2}[\overline{r}_\alpha(t) - \overline{r}_\alpha(t-\epsilon)]. \tag{43}$$

Each of the bracketed terms has a clear interpretation. Let us focus on the first bracket. Here $\overline{r}_\alpha(t+\epsilon)$ is derived from intervals i that satisfy $t+\epsilon < t_2^{(i)} < t+\epsilon+\alpha$, and $\overline{r}_\alpha(t)$ is derived from intervals i that satisfy $t < t_2^{(i)} < t+\alpha$. Since we are interested in the small ϵ limit, we suppose that $\epsilon < \alpha$. If we consider the overlap of these two ranges for $t_2^{(i)}$, it follows that the first bracket is the average mileage rate for intervals i with $t+\alpha < t_2^{(i)} < t+\alpha+\epsilon$, minus the average mileage rate for intervals i with $t < t_2^{(i)} < t+\epsilon$. As $\epsilon \to 0$, the number of intervals that contribute to each of these terms will become smaller and smaller, and thus their respective averages will become dominated by statistical sampling error. (The same conclusion follows for the second bracket in a similar way.)

Hence in general there is a play-off in the choice of ϵ. Choosing a small value for ϵ will tend to reduce the truncation error, whereas a larger value of ϵ will help reduce fluctuations due to the statistical sampling error. However, the optimal choice for ϵ is problem dependent. For example, a possible objective is to reconstruct variations in the spot rate over minimum time-scales of order one month. In practice, we anticipate that these variations will be rather slight; hence that the nonlinearity of $\overline{r}_\alpha(t)$ will not be very pronounced; and hence that the truncation error may well be acceptable with rather large values of ϵ, perhaps as large as one month. Whether this is a statistically robust choice depends on the underpinning data: for example, if we aim to analyse trends in a specific locality and/or for a specific vehicle class, the total number of contributory vehicles and hence contributory intervals may be very limited, thus exacerbating the sampling error problems for small ϵ. In contrast, if our interest is in trends for the full UK vehicle fleet, the sample size is very large and $\epsilon = 1$ month might turn out to be a robust choice. However, we shall see that the sampling error is a severe problem that we have yet to resolve fully.

Taking the above discussion into account, a particularly neat choice that we will consider is $\epsilon = 1/N$ years^{-1}, to match the time-step of the scheme. Then if we define

$$r_n := \overline{r}_\alpha(t_n), \tag{44}$$

scheme (40) becomes

$$\phi_{n+1} = \phi_n + [\phi_{n-N+1} - \phi_{n-N}] + N\,[r_{n+1} - 2r_n + r_{n-1}]\,. \tag{45}$$

Let us consider the operation of this scheme in an 'on-line' application, where new interval data is constantly accumulating. In this situation, r_{n+1} can only be computed after one has waited one further time step $1/N$ beyond t_{n+1}. The consequence is that in addition to supplying one year of initial data for ϕ, there is a further lag of $1/N$ in reporting the spot rate. So in an on-line application, where e.g. $N = 12$, the method will compute the spot rate one month in arrears.

7. Tests on synthetic data

Potentially there are two quite distinct sources of approximation or error in the time-stepping schemes (40) and (45). Firstly, there may be problems with the underpinning assumptions **A1**, **A2** and **A3** introduced in Section 3. Or more seriously, one may question the very existence of a spot rate that modulates the whole population's mileage. These are difficult issues and very relevant when one applies the method to real-world data, as we shall see in Section 8.

Secondly, however, there are potential problems with the implementation of the schemes, even if all of the underpinning modelling assumptions hold exactly. These problems are (i) the estimation of $\overline{r}''(t)$ and the consequent balancing of truncation error and sampling error that we introduced in Section 6; and (ii)

the propagation of errors from imperfect initial data for the spot rate. To investigate these problems, we test the time-stepping scheme on synthetic data where assumptions **A1**, **A2** and **A3** hold exactly. In addition to simplifying the modelling problems, this approach has the advantage that the synthetic data may be built from a 'ground-truth' spot rate that we can attempt to reconstruct and thus test the scheme normatively.

Our synthetic data is generated by a Matlab code which generates an ensemble of vehicles whose day-by-day mileages are simulated over a given extended time interval. The key inputs to this simulator are:

1. The spot rate $\phi(t)$.
2. A probability distribution from which vehicles' total usage factors c_k are drawn — see (3). Here, we have used a Gamma distribution with shape and scale parameters 2 and $1/2$ respectively, because this is a simple distribution with $\langle c_k \rangle = 1$ which incorporates some of the tail-like phenomena that we have observed in real-world mileage distributions.
3. A probability distribution which incorporates multiplicative noise on a day-to-day basis in an attempt to model the strong temporal fluctuations in individual vehicles' mileages. The MOT data is unable to inform the choice of this distribution. In this study we have used a simple uniformly distributed random variable on $[0, 2]$, but in future, we should probably experiment with heavy-tailed distributions, such as the log-normal distribution, which might more faithfully model the occurrence of infrequent long trips.
4. A function which generates test dates for each vehicle. Here, each vehicle is assigned its first test at a random point in the first year of the simulation, and each subsequent test follows at exact multiples of one year later.

The day-by-day mileages are then post-processed in combination with the simulated test dates to generate interval data which resembles the form of the real-world MOT data, to which we can attempt to apply the time-stepping methods.

Our basic set-up involves a continuously differentiable artificial spot-rate

$$\phi(t) = 8000 + 500t - 1000 \cos 2\pi t - 1000[t-2]_+(t-2)^2, \tag{46}$$

in units miles per year, that incorporates a rather simple linear trend plus a periodic term designed to model seasonal effects. In addition, at $t = 2$ years, a strong decreasing term cuts in to represent (perhaps) some kind of economic catastrophe in our imaginary world. We consider the time interval $-1 < t < 4$ years, see Figure 3, which also displays analytical computations for $\overline{r}_\alpha(t)$ derived from this choice of $\phi(t)$. Note that $\overline{r}_\alpha(t)$ may only be computed after $t = 0$, because one year of simulation is required before any intervals have elapsed.

However, the real difficulties begin when we consider the estimation of $\overline{r}_\alpha(t)$ from synthetic data, that is required in the time-stepping scheme. Let us focus on the case $\alpha = 0.1$. Figure 4 displays a comparison of $\overline{r}_\alpha(t)$ derived by sampling with its analytical computation, for sample sizes ranging from $M = 1,000$, to $M = 1000,000$ vehicles. The plots for the smaller sample sizes are extremely 'crinkly' — yet our scheme requires that we estimate their second derivative, via the finite difference formula (42). As we remarked earlier, smaller sample sizes imply more noise, implying that the 'hop' ϵ in the finite difference formula must be increased to secure robust values for the second derivative. Alternatively — some kind of smoothing algorithm (such as a spline) must be applied prior to differentiation — with the caveat that this may well introduce its own side-effects. However, note that these issues might become less severe if one chooses a larger value of α, because then more intervals contribute to each value of $\overline{r}_\alpha$.

Let us demonstrate how our time-stepping scheme works on an example in which the sampling error problem should be relatively mild. We choose $M = 1,000,000$ vehicles, $\alpha = 0.1$ and $\epsilon = \alpha$, so that in effect we adopt scheme (45). Let us suppose that we also know the initial data perfectly. Figure 5(a) shows the raw results, where no smoothing tricks have been applied in the estimation of the second derivative. Reassuringly, the time-stepping method captures the basic form of the spot rate including its rapid decline and seasonal variations — in particular, the unprocessed straddling rate by itself cannot capture the latter. However, as time progresses, the noise that is amplified in the computation of the second derivative causes the numerical solution to degrade cumulatively.

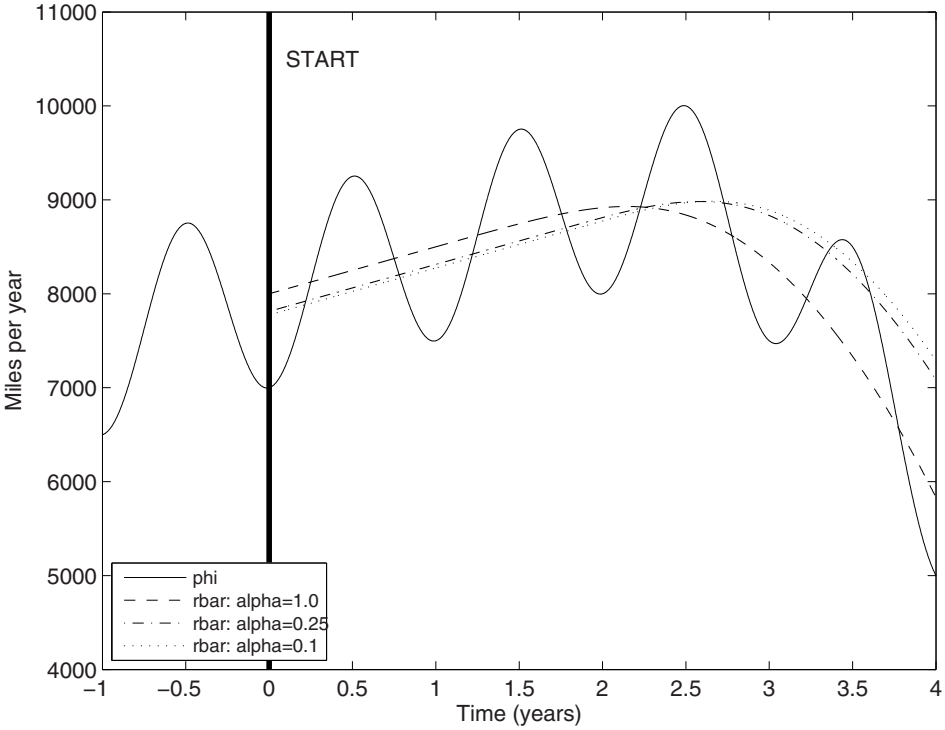

Fig. 3. Plots of the spot rate $\phi(t)$ used to generate our synthetic data and of analytical computations of the corresponding straddling rates $\bar{r}_\alpha(t)$ for $\alpha = 1.0$, $\alpha = 0.25$ and $\alpha = 0.1$ years respectively. Note that as we remarked earlier, the seasonal structure of the spot rate is not recoverable in $\bar{r}_\alpha(t)$ for any α.

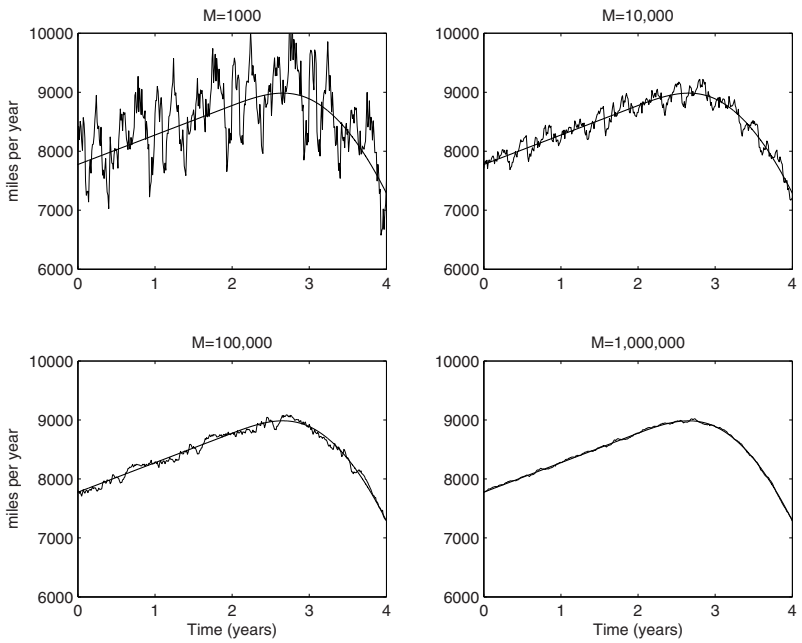

Fig. 4. A comparison of the analytical computation for $\bar{r}_\alpha(t)$ with $\alpha = 0.1$ and its estimates as extracted from synthetic data for sample sizes ranging from $M = 1,000$ to $M = 1,000,000$ vehicles.

Observation date	$\overline{r}$ miles/year (3sf)
January 2007	7,840
January 2008	7,720
January 2009	7,620

Table 1. Average annual mileage estimated via the straddling rate $\overline{r}(t)$.

Thus in Figure 5(b), we show improved results where a smoothing filter has been applied to reduce the effects of sampling error. Specifically, the samples $\overline{r}_\alpha(t)$ for $t = 0.0, 0.1, 0.2, \ldots$ etc. are smoothed by a least-squares-fit cubic B-spline, with knots chosen at $t = 1.0, 2.0, 3.0, \ldots$, and so on. Each raw value $\overline{r}_\alpha(t)$ is then replaced with an evaluation of the spline at the corresponding time t, *before* its input to the finite difference estimation of $\overline{r}''_\alpha(t)$. The numerical estimation of $\phi(t)$ now follows the ground truth with a high degree of fidelity.

Potentially, the use of a smoothing method such as this destroys (true) high frequency characteristics in $\phi(t)$, or implies a latency in on-line applications which may need to gather data some time after the estimation point in order that the filter can be applied. However, a detailed investigation of the filter design remains for future work.

8. Preliminary inspection of real-world data

As we described earlier, the initial MOT release in Autumn 2010 consisted of approximately 150 million tests from early 2005 to 31st March 2010 inclusive. After tests for the same vehicle have been matched, and the resulting data cleaned (see [4]) 76 million usable intervals remain. In practice, these intervals are not all exactly one year long, contrary to the assumption (**A1** page 5) used to construct our theory. Hence for the purposes of the present analysis, we have further removed all intervals in which the length differs from one year by more than 10 days. About 42 million intervals then remain.

Our approach here is to consider potential use of the scheme (45) where the time-step is chosen equal to the weight function parameter α. Hence the essential input required from the data is computations of the weighted straddling rates $\overline{r}_\alpha$ at time-steps spaced by α. To avoid any possible difficulties with weekly patterns in usage (e.g., less driving at the weekend) which have finer time-scales than we wish to resolve, we shall use α which are an integer number of weeks, specifically $\alpha = 52$ weeks (approximately one year), $\alpha = 13$ weeks (approximately one quarter), and $\alpha = 4$ weeks (a little short of one month).

Whatever the value of α, the computation of $\overline{r}_\alpha(t^*)$ requires intervals that began one year before the observation date t^*. Since the roll-out of the digitisation of MOT data was not completed until January 2006, $t^* =$ 1st January 2007 is the first possible observation date. Moreover $\overline{r}_\alpha(t^*)$ requires intervals that end up to time $t^* + \alpha$. In particular, this means that $\overline{r}_\alpha$ must stop just short of 31st March 2010 for the quarterly and monthly estimates.

More severely, in the case where $\alpha = 52$ weeks, we obtain only three values, corresponding roughly to January 2007, 2008 and 2009 respectively. These numbers are in effect computations of the original (unweighted) straddling rate introduced in Section 2, and are thus robust estimates of vehicle usage over a two-year period centred on these dates, with the triangular kernel weighting discovered in Section 3.

Table 1 gives the results. The data display a clear decline in per-vehicle usage, at the rate of circa 1.5% per year. However, it turns out that the number of contributory intervals (circa 11 million for each year's statistic) is climbing at about 4% per year. Of course, the number of contributory intervals is only a proxy for the level of vehicle ownership. (In total there are approaching 30 million vehicles in circulation in the UK, and in essence our filters that clean the data are selecting the mileage of only one third of them. If there were to be a change in how vehicle owners respond to MOT renewal dates, this proportion would change.) However, these statistics suggest tentatively that increases in the vehicle fleet are outweighing declines in the per-vehicle mileage. (Note that this conclusion excludes vehicles which are less than three years old which do not have to undergo MOT tests.)

However, recall that our principal interest is in temporal information with finer scales. Plots of $\overline{r}_\alpha(t)$ for $\alpha = 4$ weeks and $\alpha = 13$ weeks are displayed in Figure 6. Note that caution is required in the interpretation of these graphs. The peaks in the latter part of the Autumn of each year do not describe a peak in mileage incurred at that time of year, but rather high mileage over the previous year for vehicles that are tested in the late Autumn.

A key interest in Figure 6 is the strong period-one seasonal component, and especially the fine-scale features in the four-week data that appear to be robust on a year-to-year basis. This periodic structure cannot be explained simply by an underlying periodic $\phi(t)$, because as we have explained in theoretical terms and then demonstrated in Section 7, $\overline{r}_\alpha(t)$ does not retain a simple periodic seasonal component in $\phi(t)$, for *any* α. Furthermore, these fluctuations are not due to sampling error (although that will remain a serious issue if we were to progress to implementing a time-stepping scheme) — since we have also computed these pictures for random sub-samples of the data and have found the features to be robust.

If we were to progress to processing the data in Figure 6 with a time-stepping scheme such as (45), clearly some filtering would be required before the second derivative $\overline{r}''_\alpha(t)$ was estimated. However, it seems most likely that we would discover large negative values for $\overline{r}''_\alpha(t)$ in the late Autumn and (possibly) positive values in the Spring / early Summer. If we interpret this result in terms of marginal changes, using (17), it would imply that year-on-year, mileage rates during the summer are getting higher and higher whilst mileage rates during the winter are getting lower and lower. This 'result' does not seem very plausible.

The most likely culprit here is a combination of the breakdown of assumptions **A1**, **A2** and **A3** (page 5). As we have remarked, there is a double peak of MOT tests in the Spring and Autumn owing to a kind of 'hangover' effect of the peak in new car sales at those times due to the release of new registration plates. Because inter-test intervals deviate from one year in practice, these peaks disperse if one considers only aging vehicles. Since newer vehicles tend to be driven more than older ones, it seems that assumption **A3** is compromised — i.e., we *might* expect vehicles that are tested in the Spring and Autumn to have higher usage. However, this hypothesis needs much more careful testing, for example by segmenting the data over the first registration date of the vehicles.

9. Conclusions and possible extensions to the theory

The chief contribution of this paper has been the body of theory presented through Sections 2–5 inclusive. This theory purports to do what at first sight is seemingly impossible: namely, to estimate fine-scale temporal variations in vehicle mileage, albeit at a population level, from coarse-scale (e.g., annual) mileage data at an individual vehicle level. As far as we are aware, this is an entirely fresh problem area: the magic is that we are somehow using low (temporal) resolution data for lots of vehicles to derive high (temporal) resolution data that governs the population average. The theory has been developed with the UK MOT (roadworthiness) test data in mind, but it would have applications to any data set internationally, where odometer readings of individual vehicles are monitored on an occasional basis.

Section 6 has shown how to implement the theory as a time-stepping scheme, which has been tested on synthetic data as illustrated in Section 7. This is where reality hits home: in practice the scheme is subject to sampling error problems that one needs very large numbers of vehicles to overcome. However, it seems equally that the method may be fixed by applying smoothing filters to the straddling rate data and initial results show this approach to be very promising — albeit a full investigation of it is beyond the scope of this paper.

Section 8, which considered preliminary steps in the application to the real-world MOT data set, uncovers other difficulties. In particular, there is evidence to suggest that the simplifying assumptions **A1**, **A2** and **A3** (see page 5), used to derive the theory, are over-strong.

Let us conclude by describing how these assumptions might be relaxed and what the consequences for the subsequent theoretical development might be. Firstly **A1** supposes that the points at which individual vehicle mileages are captured are exactly one year apart. In fact, we can use the MOT data to fit a distribution for the length of the interval between tests. The consequence for Section 3 would be a more involved calculation for $\overline{r}(t)$, since the integrals would need to incorporate the spot mileage rate over a wider span,

according to a weight determined by the distribution for the length of inter-test intervals. However, a likely up-side is that it would be possible to parametrise seasonal effects in the resulting *straddling rates*, without resorting to a time-stepping scheme — which is presently impossible owing to the inter-test interval (one year) corresponding exactly to the seasonal period.

Secondly, **A2** supposes that tests occur at the same rate throughout the year. Actually, since our analysis deals with average per-vehicle mileage, rather than its cumulative sum over the population, this is not an important assumption, except that it is convolved with the failure of **A3**, which supposes that a vehicle's mileage is independent of the time of year at which it is tested. As we have discussed above, this assumption could be very poor, owing to the high proportion of relatively new (i.e., high mileage) vehicles which contribute to the double peak in tests in the Spring and Autumn. Whereas the variation in the testing rate throughout the year is easily parametrised from the MOT data, the dependence of vehicles' mileages on the date of the test is very hard to separate from what may well be genuine dynamic effects of interest. The only solution that we can propose is a careful segmentation of the data over vehicles of different ages — i.e., we should attempt the computation of distinct spot rates for vehicles in different age bands, so that the mileage in each segment is independent of the date of the test. If this technique proves effective, then no further generalisations to the existing theory would be required.

Finally, we should say that this body of work is based upon the assumption that a spot rate which modulates the population mileage actually *exists* — and is sufficiently smooth for the theory to apply. Smoothness of course breaks down at some (fine) temporal level, at least when one focusses down to the level of an individual week, in which there are (strong) day-to-day fluctuations in mileage due to working and weekend days. However, a more significant problem is that we presently have very little data on the daily mileage distribution of individual vehicles — to inform assumptions about the spot-rate and the distribution of the noise term in equation (3). It is possible that new data sets (e.g., from the forthcoming 3000-car study in the USA) may help solve this problem.

In summary, we believe that none of the practical difficulties that we have discovered are insuperable. The form of data that we have considered in this paper is, as far as we are aware, bespoke to the UK, although it seems quite likely that other countries collect the mileage data of individual vehicle data as part of their regulatory requirements in vehicle safety, insurance or taxation. Furthermore, in monitoring the effectiveness of local travel schemes, there seems to be a strong case for requesting vehicle odometer readings, in addition to the usual surveys, traffic counts and travel diaries. In this respect, we anticipate that the body of theory presented in this paper will make a significant underpinning contribution to the measurement of vehicle usage in the future.

Acknowledgements

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

[1] S. Cairns, R.E. Wilson, T. Chatterton, J. Anable, S. Notley, and F. McLeod. Using MOT data to analyse travel behaviour change — scoping report. Technical Report TRL PPR578, TRL, TRL, Bracknell, United Kingdom, 2011. IHS, ISSN 0968-4093.

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

[3] Department for Transport. Anonymised mot results. Downloaded from `http://www.dft.gov.uk/statistics/series/traffic/`, 2010. Page checked 3 January 2012.

[4] R.E. Wilson, S. Cairns, S. Notley, J. Anable, T. Chatterton, and F. McLeod. Techniques for the inference of mileage rates from MOT data. *Transportation Planning and Technology*, 36(1):130–143, 2013.

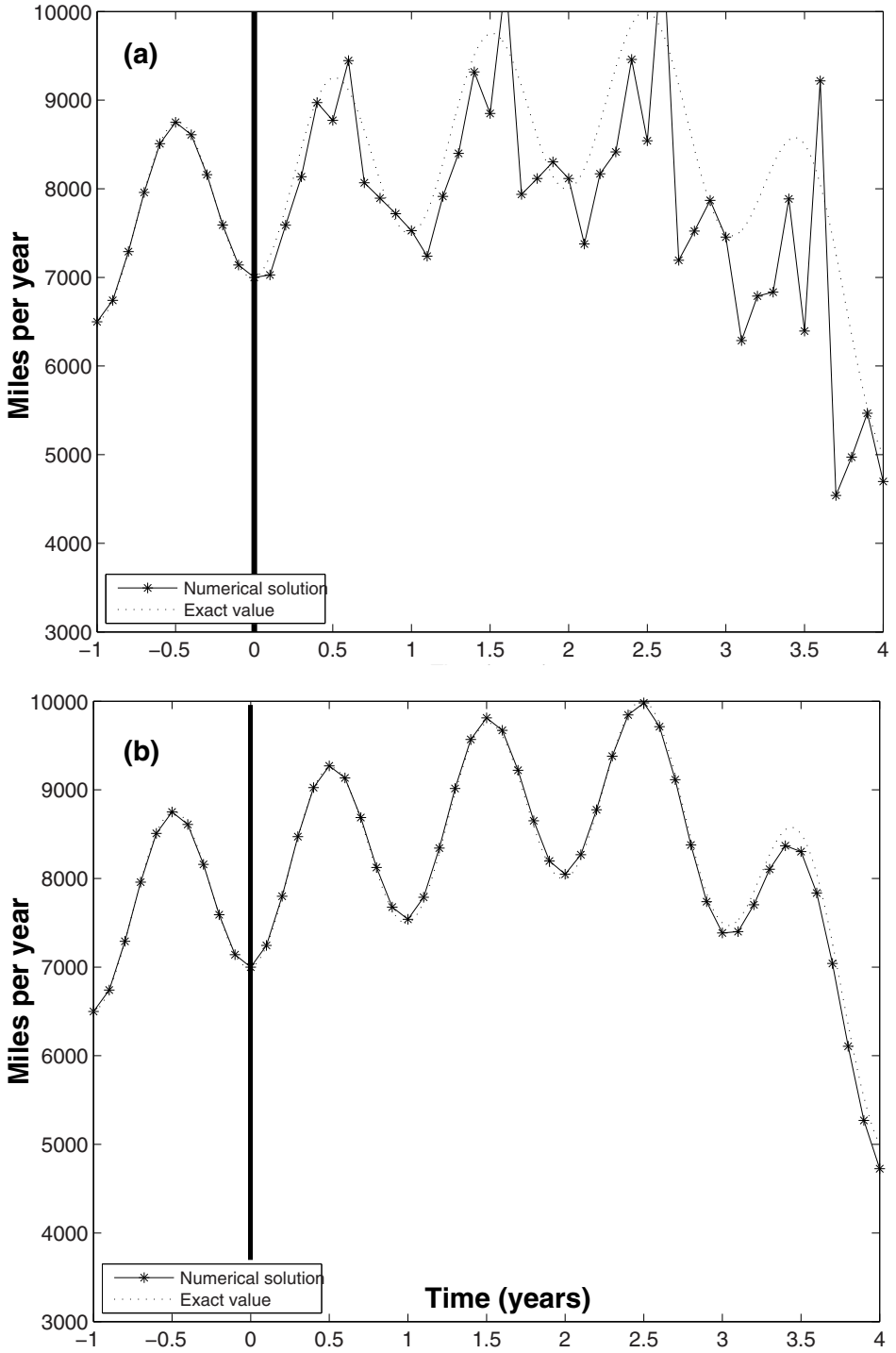

Fig. 5. Comparison of the numerical estimation of $\phi(t)$ with its ground truth, using synthetic data. Plots display (a) results based on raw values of $\overline{r}_\alpha(t)$ and (b) results based on filtered values of $\overline{r}_\alpha(t)$, that are smoothed by a cubic B-spline.

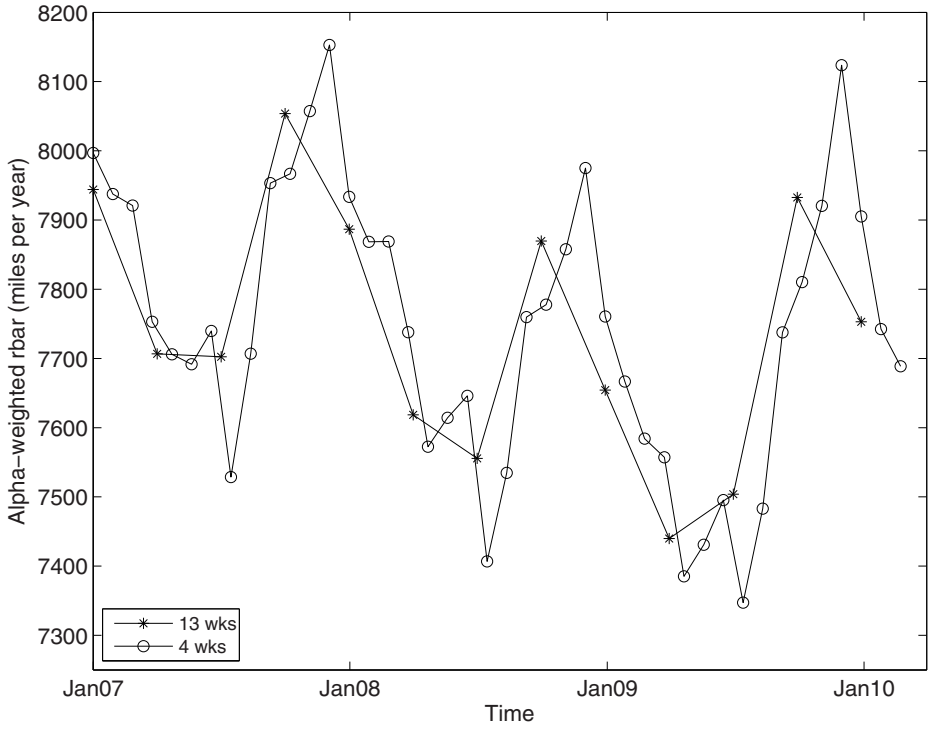

Fig. 6. Plots of $\bar{r}_\alpha(t)$ for the real-world MOT data with time-steps $\alpha = 4$ weeks and $\alpha = 13$ weeks.