Techniques for Inferring Mileage from the Department for Transport's MOT Data Set

R. Eddie Wilson

Jillian Anable (Aberdeen), Sally Cairns (TRL/UCL), Tim Chatterton (UWE), Oliver Turnbull (Bristol) and others

EPSRC grants EP/J004758/1 EP/K000438/1

Faculty of Engineering
University of Bristol

March 25, 2015

UK MOT (Ministry of Transport) test

VT20 MOT Test Certificate VOSA

This certificate has been issued according to the conditions and notes on the back of this certificate.

Vehicle & Operator Services Agency

Note: If you have doubts as to whether this certificate is valid, please use the service described in note 3 overleaf to check.

MOT test number: 761710136293
Make: VAUXHALL
Odometer reading: 105420 Miles
Registration mark: T203UNP
Model: ASTRA
Test class: IV
Vehicle identification or chassis number: WOLOTGF35X8091395
Colour: WHITE
Approximate year of first use: 1999
Expiry date: AUGUST 25th 2007 (ZERO SEVEN)
Issue date/time: AUGUST 18th 2006 (ZERO SIX) 13:30
Fuel type: Petrol
Authentication number
Design gross weight (goods vehicles): kg
Advisory Notice Issued: NO
Test station number: 80572

084907914489518556410227

For all vehicles with more than 8 passenger seats
Seat belt installation checked this test: N/A
Number of seat belts fitted at time of installation check: N/A
Previous installation check date: N/A

Issuer's name in CAPITALS: D. S. BRYANT
Signature of Issuer

Warning: A test certificate is not evidence that the vehicle is in a satisfactory condition.

Check carefully that the above details are correct.
Do not accept a certificate which has been altered.

Reg Mark: T203UNP
Make: VAUXHALL
VTS Number: 80572
MOT Expiry: AUGUST 25th 2007 (ZERO SEVEN)

Peel here

Inspection Authority
HANHAM MOTOR COMPANY
126 BRYANTS HILL
ST GEORGE
BRISTOL
BS5 8RJ

- MOT: the UK's annual safety inspection for all road vehicles older than 3 years
- Since 2005: the results have been captured and stored digitially
- Since November 2010 — the DfT has published this data online - spanning back to 2005.
- Key interest: the *odometer reading* recorded at each test.

A sample of the published data

```
626966|2010-01-18|4|N|P|38198|DE|BMW|523I SE TOURING AUTO|GREEN|P|2494|1998
626977|2010-03-03|4|N|P|25864|ST|LAND ROVER|FREELANDER HSE TD4|BLACK|D|2179|2007
626984|2010-03-04|4|N|P|32884|YO|LAND ROVER|RANGE ROVER SP HSE TDV8 A|BLACK|D|3628|2007
626991|2010-03-26|4|N|F|91196|PL|MERCEDES|ML 320 AUTO|SILVER|P|3199|2000
627020|2010-02-02|4|N|PRS|29180|DH|MERCEDES|ML 320 CDI SE AUTO|SILVER|D|2987|2006
627023|2010-02-24|4|F|P|62713|MK|BMW|325I SE AUTO|SILVER|P|2494|2001
627024|2010-02-24|4|N|F|62713|MK|BMW|325I SE AUTO|SILVER|P|2494|2001
627025|2010-02-22|4|N|F|62647|LU|BMW|325I SE AUTO|SILVER|P|2494|2001
627041|2010-03-04|4|PL|P|230304|IP|MERCEDES|300TE AUTO|GREY|P|2962|1990
627042|2010-03-04|4|N|F|230304|IP|MERCEDES|300TE AUTO|GREY|P|2962|1990
627050|2010-01-25|4|N|PRS|62624|IP|UNCLASSIFIED|UNCLASSIFIED|GREY|P|5300|2006
627058|2010-02-08|4|N|P|88480|SS|JAGUAR|S-TYPE V6 SE AUTO|BLUE|P|2967|1999
627109|2010-01-29|1|N|P|1244|CO|UNCLASSIFIED|UNCLASSIFIED|WHITE|P|125|1959
627145|2010-03-25|7|N|P|35194|LE|AUSTIN|UNCLASSIFIED|BLUE|D|0|1963
627185|2010-02-18|4|PL|P|170507|EX|VOLVO|850|MAROON|P|2435|1997
627186|2010-02-15|4|N|F|170449|EX|VOLVO|850|MAROON|P|2435|1997
627227|2010-02-24|4|N|P|73195|NW|MERCEDES|E430 AVANTGARDE AUTO|BLACK|P|4266|2002
627242|2010-02-01|4|N|P|38225|IP|TOYOTA|HILUX INVINCIBLE D-4D A|BLACK|D|2982|2007
627280|2010-03-08|4|PR|P|44132|B|AUDI|TT QUATTRO (180 BHP)|BLACK|P|1781|2000
627281|2010-03-08|4|N|F|44132|B|AUDI|TT QUATTRO (180 BHP)|BLACK|P|1781|2000
```

- But the tests are grouped by year and do not “link” the vehicles (a problem fixed in more recent releases — at my prompting!)

Here's a trick . . .

- Concatenate all files and sort by the "mystery" identifier.
  You get lots of blocks like this:

Here's a trick . . .

- Concatenate all files and sort by the "mystery" identifier. You get lots of blocks like this:

```
118173532|2009-08-05|4|N|P|132299|BS|VAUXHALL|ASTRA LS 8V|WHITE|P|1598|1999
118173533|2008-08-11|4|PR|P|123259|BS|VAUXHALL|ASTRA LS 8V|WHITE|P|1598|1999
118173534|2008-08-11|4|N|F|123259|BS|VAUXHALL|ASTRA LS 8V|WHITE|P|1598|1999
118173535|2007-08-13|4|N|P|113709|BS|VAUXHALL|ASTRA LS 8V|WHITE|P|1598|1999
118173536|2006-08-18|4|N|P|105420|BS|VAUXHALL|ASTRA LS 8V|WHITE|P|1598|1999
118173537|2005-08-26|4|N|P|99777|BS|VAUXHALL|ASTRA LS 8V|WHITE|P|1598|1999
```

- We can follow individuals around and infer their mileage (rate) between consecutive test dates!!!!

Here's a trick . . .

- Concatenate all files and sort by the "mystery" identifier.
  You get lots of blocks like this:

```
118173532|2009-08-05|4|N|P|132299|BS|VAUXHALL|ASTRA LS 8V|WHITE|P|1598|1999
118173533|2008-08-11|4|PR|P|123259|BS|VAUXHALL|ASTRA LS 8V|WHITE|P|1598|1999
118173534|2008-08-11|4|N|F|123259|BS|VAUXHALL|ASTRA LS 8V|WHITE|P|1598|1999
118173535|2007-08-13|4|N|P|113709|BS|VAUXHALL|ASTRA LS 8V|WHITE|P|1598|1999
118173536|2006-08-18|4|N|P|105420|BS|VAUXHALL|ASTRA LS 8V|WHITE|P|1598|1999
118173537|2005-08-26|4|N|P|99777|BS|VAUXHALL|ASTRA LS 8V|WHITE|P|1598|1999
```

- We can follow individuals around and infer their mileage (rate) between consecutive test dates!!!!

- For example, in the **interval** from 2008-08-11 to 2009-08-05 (359 days), I drove 132,299-123,259 = 9,040* miles, at an **average rate** of 25.18 **miles per day**.

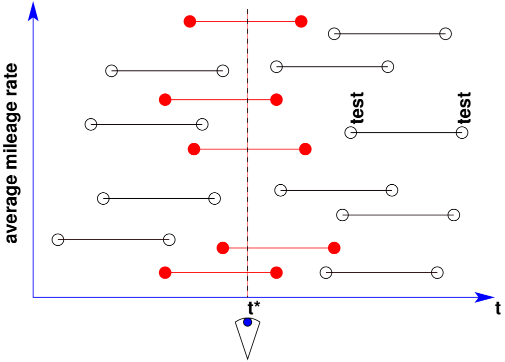
Basic analysis object: *intervals* and their attributes

- Re-arrange blocks of same-vehicle data into consecutive pairs of tests:

Interval	First test			Second test		
	date t_1	miles x_1	place$_1$	date t_2	miles x_2	place$_2$
1	2005-08-26	99777	BS	2006-08-18	105420	BS
2	2006-08-18	105420	BS	2007-08-13	113709	BS
3	2007-08-13	113709	BS	2008-08-11	123259	BS
4	2008-08-11	123259	BS	2008-08-11	123259	BS
5	2008-08-11	123259	BS	2009-08-05	132299	BS

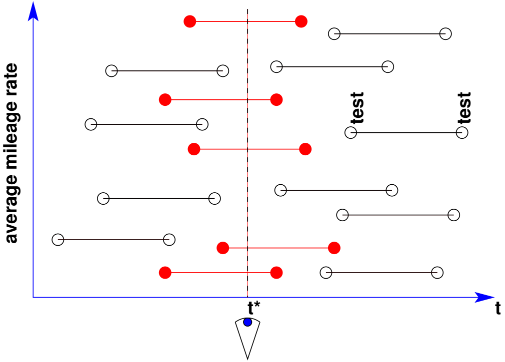
- To which can be linked vehicle-specific attributes:
  VAUXHALL, ASTRA LS 8V, WHITE, P (fuel), 1598 (cc), 1999 (year)
- (Eg) during *interval* 3 — I drove at an average rate of $(123259 - 113709)/364 = 26.24$ miles per day, but we don't know how my mileage was *distributed* during that period.
- These mileage rates are (more or less) complete across the vehicle population — even after cleaning.

Population level statistics: *straddling rate* $\bar{r}(t)$



- Select all N intervals that *straddle* a given *observation date* t^*
- Each interval yields an average (per vehicle) rate r_i.

Population level statistics: *straddling rate* $\overline{r}(t)$

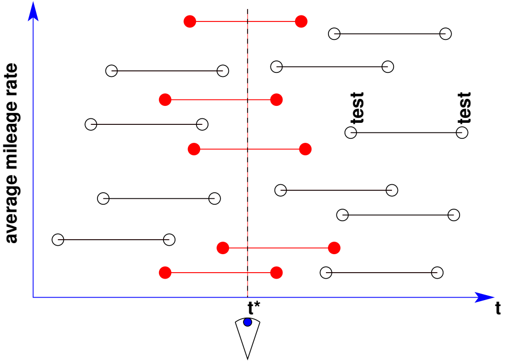


- *Straddling rate* $\overline{r}(t^*)$ is then defined by the **average average**

$$\overline{r}(t^*) = \frac{1}{N}\sum_{i=1}^{N} r_i .$$

- Select all N intervals that *straddle* a given *observation date* t^*
- Each interval yields an average (per vehicle) rate r_i.

Population level statistics: *straddling rate* $\bar{r}(t)$



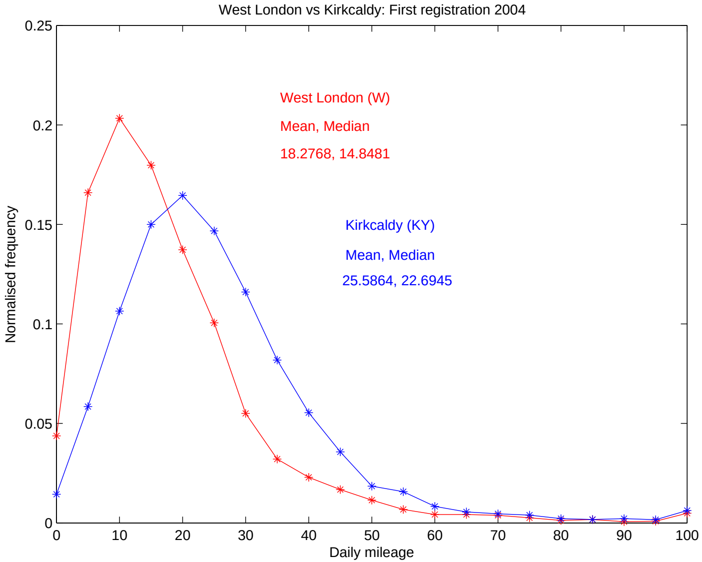
- Select all N intervals that *straddle* a given *observation date* t^*
- Each interval yields an average (per vehicle) rate r_i.

- *Straddling rate* $\bar{r}(t^*)$ is then defined by the **average average**

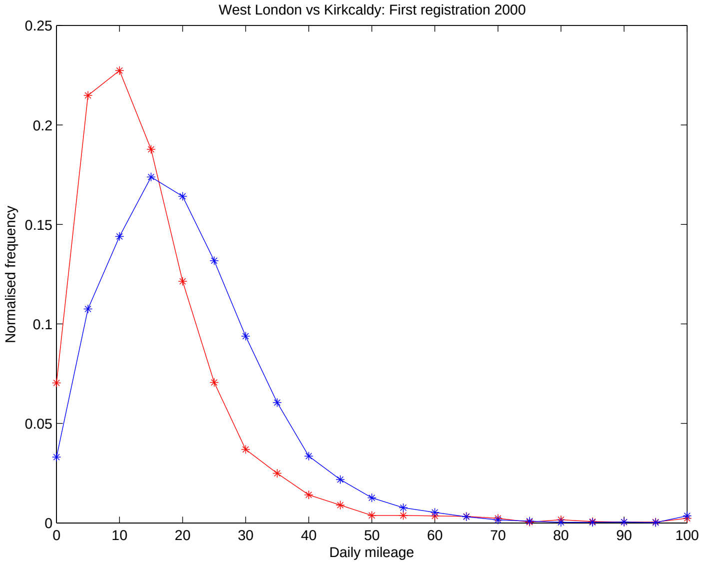
$$\bar{r}(t^*) = \frac{1}{N}\sum_{i=1}^{N} r_i .$$

- It is fine for annual statistics: choose $t^* = 1/7/2007$, $1/7/2008$, $1/7/2009$ etc.
- But $\bar{r}(t^*)$ actually incorporates miles driven over the two year span $t^* - 1 \leq t < t^* + 1$.

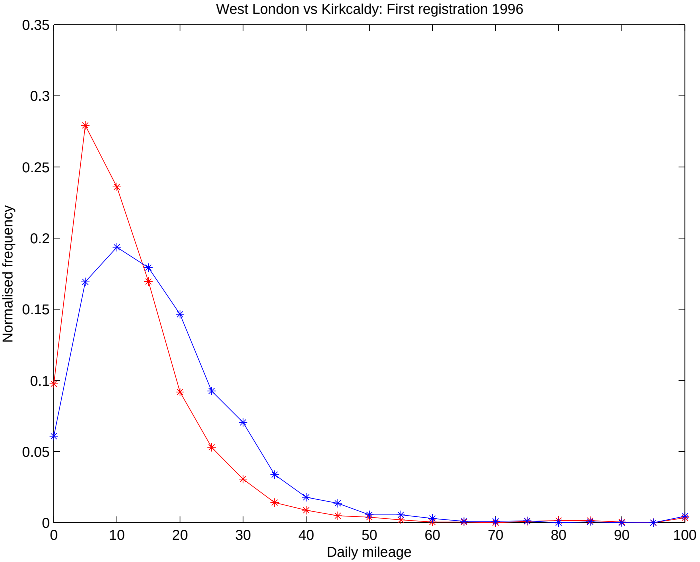
Mileage distributions: new(ish) vehicles



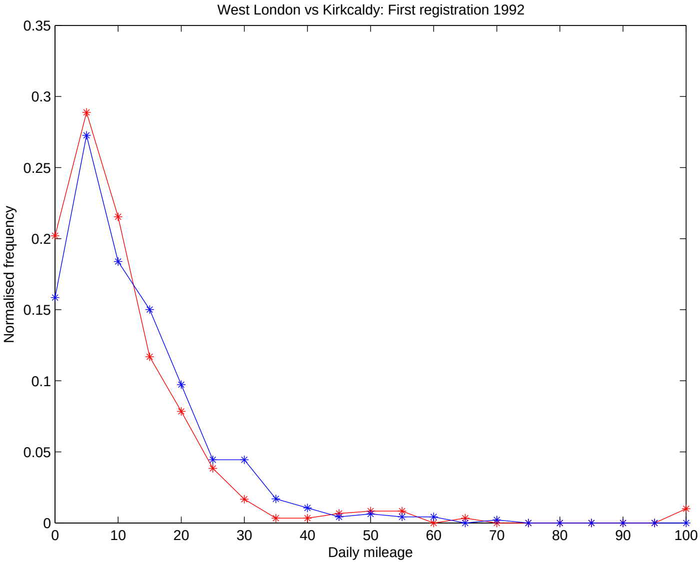
Mileage distributions: older vehicles



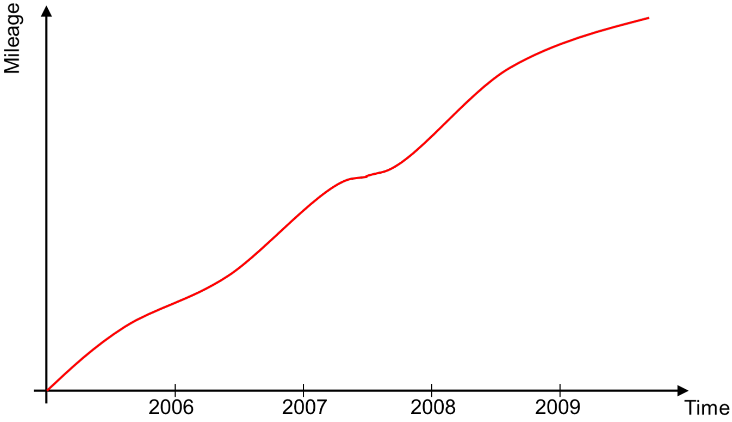
Mileage distributions: even older vehicles



Mileage distributions: old vehicles

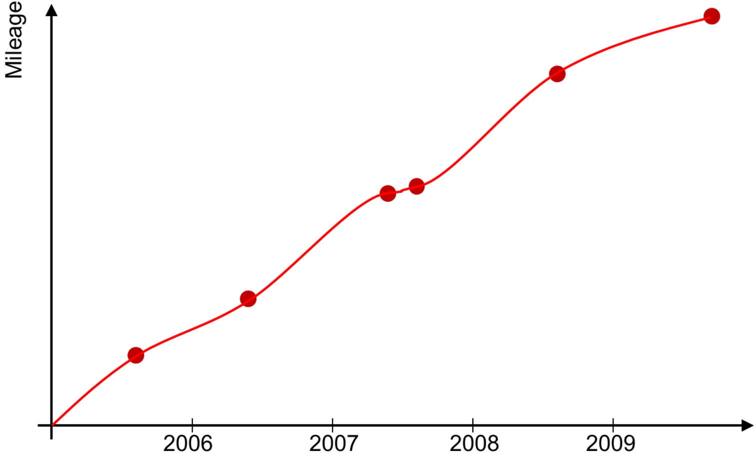


From the Straddling Rate to the Census Date Rate



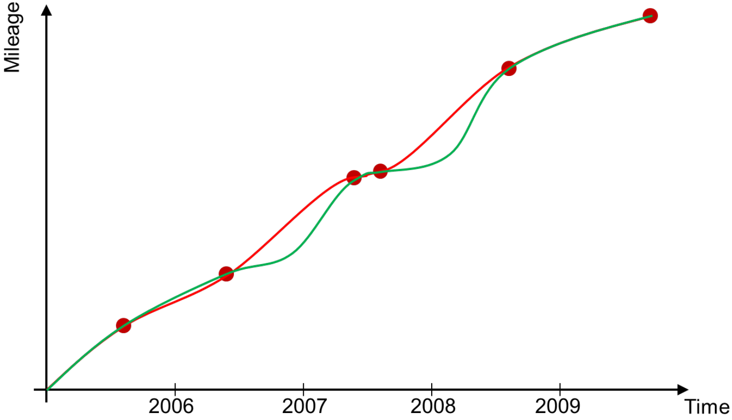
- Progression of a vehicle's odometer with time

From the Straddling Rate to the Census Date Rate



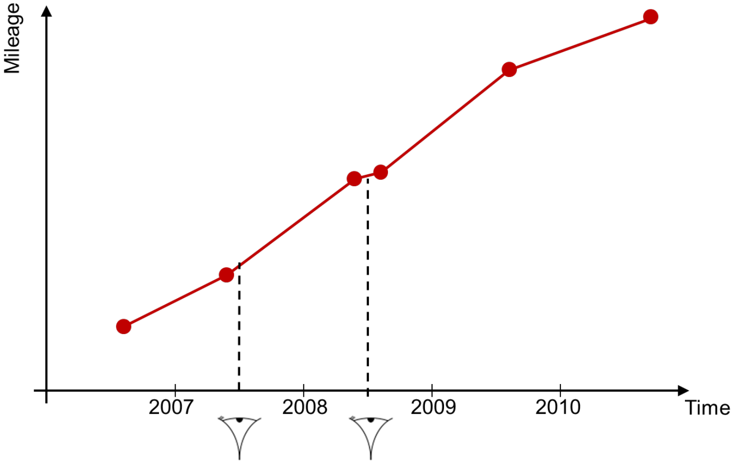
- Progression of a vehicle's odometer with time — with tests

From the Straddling Rate to the Census Date Rate



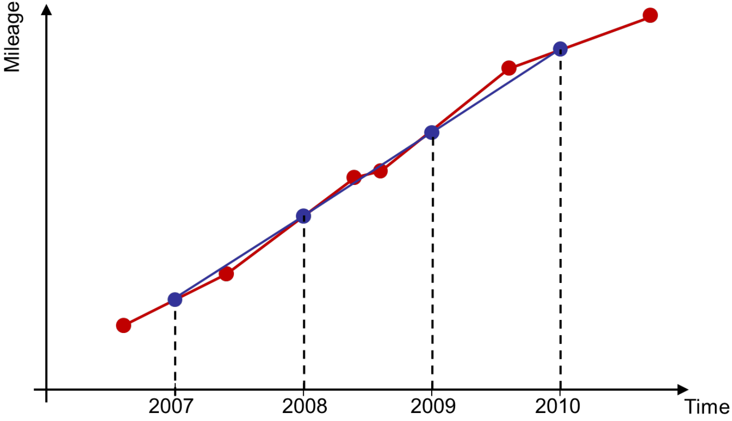
- The tests do not allow you to distinguish the 2 trajectories.

From the Straddling Rate to the Census Date Rate



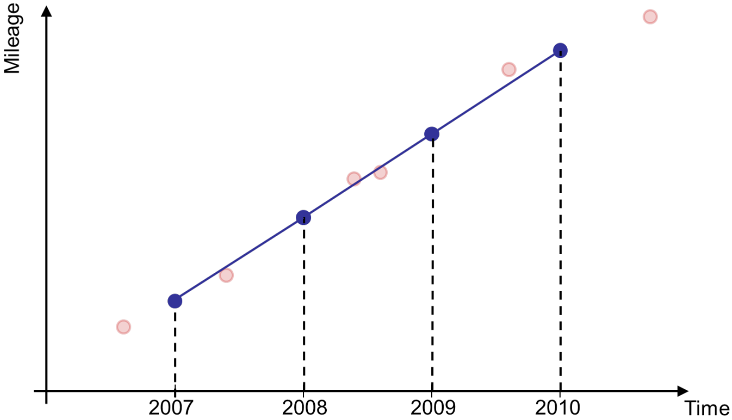
- *Distributions* derived from straddling rate suffer anomalous variance because some intervals are very short

From the Straddling Rate to the Census Date Rate



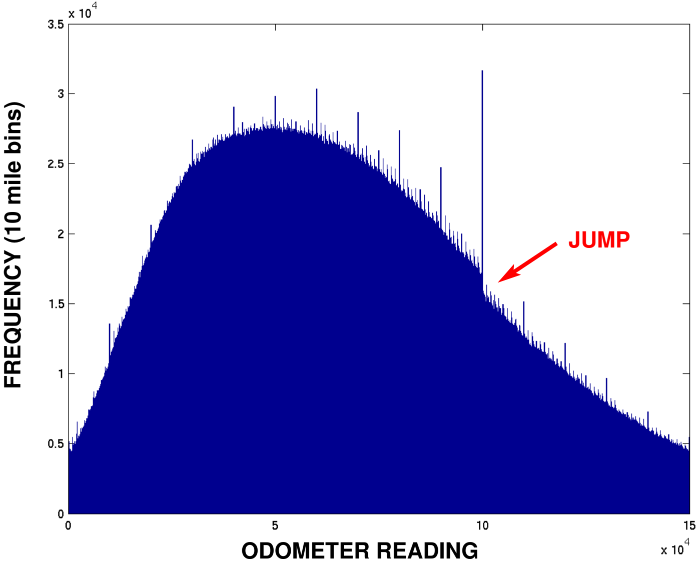
- Solution is to interpolate onto some given *census dates* . . .

From the Straddling Rate to the Census Date Rate



- ... and use the rates between the census dates. (Also neatly synchronises the data into calendar year comparisons.)

Five digit odometer problem



Cleaning: How to Deal with Bad Odometers

Solution 1: don't worry about it too much

Cleaning: How to Deal with Bad Odometers

Solution 1: don't worry about it too much

- Compute rates as if all odometers are perfectly correct

Cleaning: How to Deal with Bad Odometers

Solution 1: don't worry about it too much

- Compute rates as if all odometers are perfectly correct
- Reject intervals (*) if rates which are outside a reasonable range:
  - Below 0
  - Above 150 miles per day (?)

Cleaning: How to Deal with Bad Odometers

Solution 1: don't worry about it too much

- Compute rates as if all odometers are perfectly correct
- Reject intervals (*) if rates which are outside a reasonable range:
  - Below 0
  - Above 150 miles per day (?)
- Scale population statistics up for the intervals of vehicles thus discarded

(*) Nomenclature: will talk of intervals as **B**ad or **G**ood.

Cleaning: How to Deal with Bad Odometers

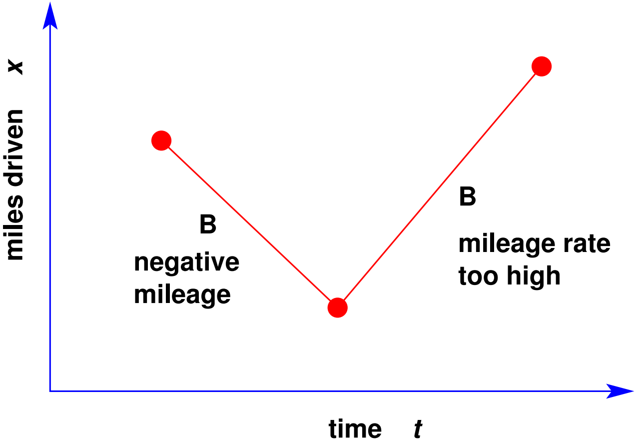
Solution 1: don't worry about it too much

- Compute rates as if all odometers are perfectly correct
- Reject intervals (*) if rates which are outside a reasonable range:
  - Below 0
  - Above 150 miles per day (?)
- Scale population statistics up for the intervals of vehicles thus discarded

(*) Nomenclature: will talk of intervals as **B**ad or **G**ood.

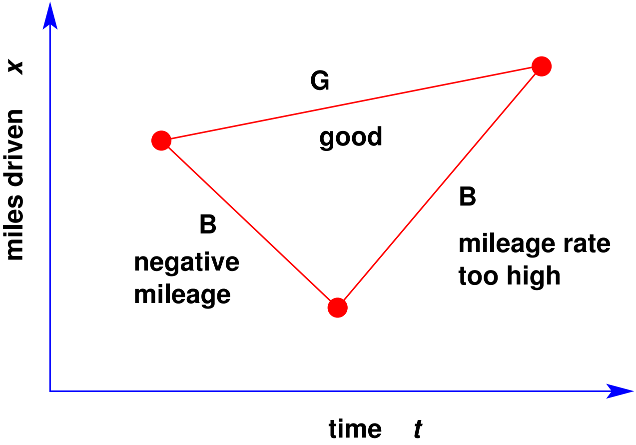
Solution 2: try to identify which individual odometer entries are bad and remove them instead

When two (or more) **B**ads make a **G**ood



- The middle odometer entry is (probably) erroneous — due to a missing digit in the data entry?

When two (or more) **B**ads make a **G**ood



- The middle odometer entry is (probably) erroneous — due to a missing digit?
- The spanning interval without the middle test is (probably) ok.

Syntactic games

- Represent each vehicle's intervals as a sequence of **B** and **G**. For example **BGGGBBGGBGG**.

Syntactic games

- Represent each vehicle's intervals as a sequence of **B** and **G**. For example **BGGGBBGGBGG**.
- Try to remove tests to end up with a sequence that is all **G**.

Syntactic games

- Represent each vehicle's intervals as a sequence of **B** and **G**. For example **BGGGBBGGBGG**.
- Try to remove tests to end up with a sequence that is all **G**.
- Multiple consecutive **B**s should be replaced with the spanning interval which is either **G** (problem solved) or perhaps **B**.

Syntactic games

- Represent each vehicle's intervals as a sequence of **B** and **G**. For example **BGGGBBGGBGG**.
- Try to remove tests to end up with a sequence that is all **G**.
- Multiple consecutive **B**s should be replaced with the spanning interval which is either **G** (problem solved) or perhaps **B**.
- Only remaining problem is singleton **B** — which end of the bad interval should be removed?

Syntactic games

- Represent each vehicle's intervals as a sequence of **B** and **G**. For example **BGGGBBGGBGG**.
- Try to remove tests to end up with a sequence that is all **G**.
- Multiple consecutive **B**s should be replaced with the spanning interval which is either **G** (problem solved) or perhaps **B**.
- Only remaining problem is singleton **B** —
  which end of the bad interval should be removed?
  - Endpoint **B**: delete the end test (yes, you then need infill)

Syntactic games

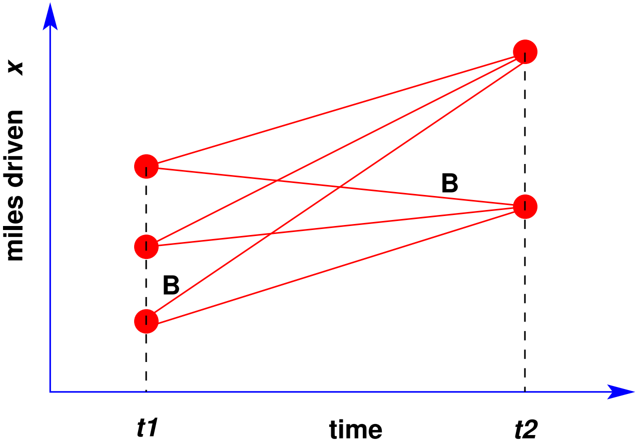
- Represent each vehicle's intervals as a sequence of **B** and **G**. For example **BGGGBBGGBGG**.
- Try to remove tests to end up with a sequence that is all **G**.
- Multiple consecutive **B**s should be replaced with the spanning interval which is either **G** (problem solved) or perhaps **B**.
- Only remaining problem is singleton **B** —
  which end of the bad interval should be removed?
  - Endpoint **B**: delete the end test (yes, you then need infill)
  - Interior **B**: a messy mixture of clocking events; clock rollover; (mild) *centrally bad* cases etc.

Syntactic games

- Represent each vehicle's intervals as a sequence of **B** and **G**. For example **BGGGBBGGBGG**.
- Try to remove tests to end up with a sequence that is all **G**.
- Multiple consecutive **B**s should be replaced with the spanning interval which is either **G** (problem solved) or perhaps **B**.
- Only remaining problem is singleton **B** — which end of the bad interval should be removed?
  - Endpoint **B**: delete the end test (yes, you then need infill)
  - Interior **B**: a messy mixture of clocking events; clock rollover; (mild) *centrally bad* cases etc.
  - Look at removing either or both ends so as to generate **G**. Repeat

How to deal with multiple tests on the same day (I)

(need to pare down to a single odometer reading per test day)



- We want to complete previous syntactic procedure *before* deciding which test to select for each date.

How to deal with multiple tests on the same day (II)

- Compute 4 rates, from the odometer pairs

$$(x_1^{\min}, x_2^{\min}) \quad (x_1^{\max}, x_2^{\max}) \quad (x_1^{\min}, x_2^{\max}) \quad (x_1^{\max}, x_2^{\min})$$

How to deal with multiple tests on the same day (II)

- Compute 4 rates, from the odometer pairs

$$(x_1^{\min}, x_2^{\min}) \quad (x_1^{\max}, x_2^{\max}) \quad (x_1^{\min}, x_2^{\max}) \quad (x_1^{\max}, x_2^{\min})$$

- We call the interval
  - Certainly **B**ad, if all 4 rates are **B**ad
  - Certainly **G**ood, if all 4 rates are **G**ood
  - **D**on't know — if there is a mix

How to deal with multiple tests on the same day (II)

- Compute 4 rates, from the odometer pairs

$$(x_1^{\min}, x_2^{\min}) \quad (x_1^{\max}, x_2^{\max}) \quad (x_1^{\min}, x_2^{\max}) \quad (x_1^{\max}, x_2^{\min})$$

- We call the interval
  - Certainly **B**ad, if all 4 rates are **B**ad
  - Certainly **G**ood, if all 4 rates are **G**ood
  - **D**on't know — if there is a mix
- The **D** are rare — no great loss in calling them **B**

How to deal with multiple tests on the same day (II)

- Compute 4 rates, from the odometer pairs

$$(x_1^{\min}, x_2^{\min}) \quad (x_1^{\max}, x_2^{\max}) \quad (x_1^{\min}, x_2^{\max}) \quad (x_1^{\max}, x_2^{\min})$$

- We call the interval
  - Certainly **B**ad, if all 4 rates are **B**ad
  - Certainly **G**ood, if all 4 rates are **G**ood
  - **D**on't know — if there is a mix
- The **D** are rare — no great loss in calling them **B**
- Note: for certainly **B**ad: there might be a good interval if there are 3 or more distinct tests at both t_1 and t_2: also rare

How to deal with multiple tests on the same day (II)

- Compute 4 rates, from the odometer pairs

$$(x_1^{\min}, x_2^{\min}) \quad (x_1^{\max}, x_2^{\max}) \quad (x_1^{\min}, x_2^{\max}) \quad (x_1^{\max}, x_2^{\min})$$

- We call the interval
  - Certainly **B**ad, if all 4 rates are **B**ad
  - Certainly **G**ood, if all 4 rates are **G**ood
  - **D**on't know — if there is a mix
- The **D** are rare — no great loss in calling them **B**
- Note: for certainly **B**ad: there might be a good interval if there are 3 or more distinct tests at both t_1 and t_2: also rare
- Proceed with previous procedure using certainly **B**ad and **G**ood.

How to deal with multiple tests on the same day (II)

- Compute 4 rates, from the odometer pairs

$$(x_1^{\min}, x_2^{\min}) \quad (x_1^{\max}, x_2^{\max}) \quad (x_1^{\min}, x_2^{\max}) \quad (x_1^{\max}, x_2^{\min})$$

- We call the interval
  - Certainly **B**ad, if all 4 rates are **B**ad
  - Certainly **G**ood, if all 4 rates are **G**ood
  - **D**on't know — if there is a mix
- The **D** are rare — no great loss in calling them **B**
- Note: for certainly **B**ad: there might be a good interval if there are 3 or more distinct tests at both t_1 and t_2: also rare
- Proceed with previous procedure using certainly **B**ad and **G**ood.
- Finally — decide which odometer at each t to use at the end. (For example: the median value.)

Central Question for Remainder of Talk

Recall that I cannot possibly say anything about an individual's mileage on finer time scales than one year.

But can I derive something about <u>population level</u> mileage over shorter time scales — eg a month?

Central Question for Remainder of Talk

Recall that I cannot possibly say anything about an individual's mileage on finer time scales than one year.

But can I derive something about <u>population level</u> mileage over shorter time scales — eg a month?

Possible application: detect the sharp drop in driving in Autumn 2008 following Lehman brothers collapse.

How to compute temporal evolution of mileage rates?

How to compute temporal evolution of mileage rates?

- Erm, isn't it obvious?

How to compute temporal evolution of mileage rates?

- Erm, isn't it obvious?
- Take a given sequence t_i, $i = 1, 2, \ldots$

How to compute temporal evolution of mileage rates?

- Erm, isn't it obvious?
- Take a given sequence t_i, $i = 1, 2, \ldots$
- Compute corresponding $\overline{r}(t_i)$ using straddling procedure

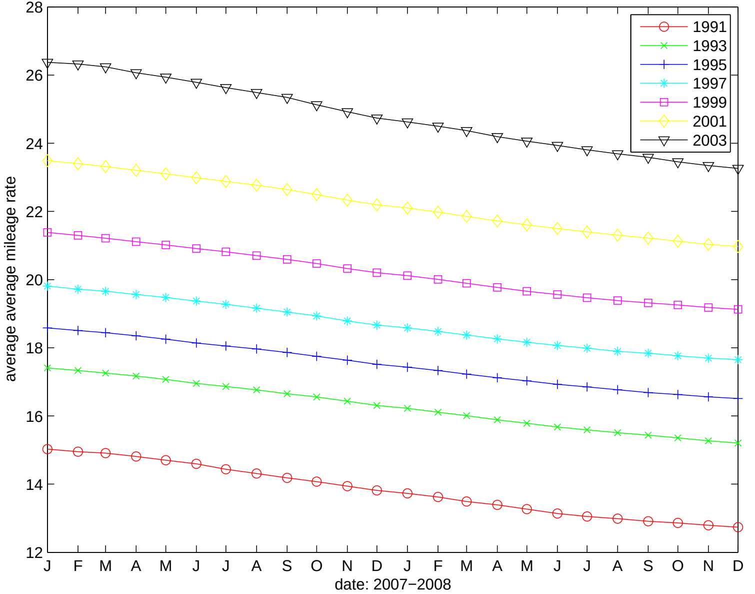
How to compute temporal evolution of mileage rates?

- Erm, isn't it obvious?
- Take a given sequence t_i, $i = 1, 2, \ldots$
- Compute corresponding $\overline{r}(t_i)$ using straddling procedure
- Pairs $(t_i, \overline{r}(t_i))$ *reconstruct* $\overline{r}(t)$

How to compute temporal evolution of mileage rates?

- Erm, isn't it obvious?
- Take a given sequence t_i, $i = 1, 2, \ldots$
- Compute corresponding $\overline{r}(t_i)$ using straddling procedure
- Pairs $(t_i, \overline{r}(t_i))$ *reconstruct* $\overline{r}(t)$
- Actually … this process is flawed…
  But just look what we can do with it!!!

Example of temporal evolution via straddling (WRONG)



Basic postulate: the population *spot rate* $\phi(t)$

- Suppose there is a population-level *spot rate* $\phi(t)$ that modulates *all* vehicles' mileage (alt. restrict to a population segment).

Basic postulate: the population *spot rate* $\phi(t)$

- Suppose there is a population-level *spot rate* $\phi(t)$ that modulates *all* vehicles' mileage (alt. restrict to a population segment).

- Then each vehicle i has an individual spot rate $\phi_i(t)$ with

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

Here c_i=const.; $\langle c_i \rangle = 1$; and $\langle \text{noise} \rangle = 0$, so that $\phi = \langle \phi_i \rangle$.

Basic postulate: the population *spot rate* $\phi(t)$

- Suppose there is a population-level *spot rate* $\phi(t)$ that modulates *all* vehicles' mileage (alt. restrict to a population segment).

- Then each vehicle i has an individual spot rate $\phi_i(t)$ with

$$\phi_i(t) = c_i \phi(t) + \mathrm{noise}.$$

Here c_i=const.; $\langle c_i \rangle = 1$; and $\langle \mathrm{noise} \rangle = 0$, so that $\phi = \langle \phi_i \rangle$.

- Let $\psi_i(\tau)$ denote miles driven by i between tests at times $\tau - 1/2$ and $\tau + 1/2$. Then

$$\psi_i(\tau) = \int_{\tau-1/2}^{\tau+1/2} (c_i \phi(s) + \mathrm{noise})\, \mathrm{d}s, \qquad = c_i \int_{\tau-1/2}^{\tau+1/2} \phi(s)\, \mathrm{d}s.$$

From the *spot rate* to the *straddling rate*

- Thus by averaging over tests that straddle t:

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

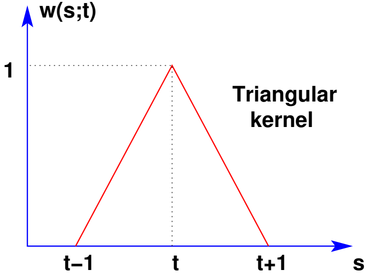
From the *spot rate* to the *straddling rate*

- Thus by averaging over tests that straddle t:

$$\bar{r}(t) = \int_{t-1/2}^{t+1/2} \langle \psi_i(\tau) \rangle_i \, \mathrm{d}\tau = \int_{t-1/2}^{t+1/2} \langle c_i \rangle \int_{\tau-1/2}^{\tau+1/2} \phi(s) \, \mathrm{d}s \, \mathrm{d}\tau.$$

- Simplify integral by $\langle c_i \rangle = 1$ and reverse the order of integration

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



- Thus $\phi(t)$ leads to $\bar{r}(t)$.
  But we want to derive $\phi(t)$ from $\bar{r}(t)$ (which is derivable from data).

From the *straddling rate* to the *spot rate*

- See TR-E 2013 for a whole bunch of Mathematics!!! - upshot:

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

From the *straddling rate* to the *spot rate*

- See TR-E 2013 for a whole bunch of Mathematics!!! - upshot:

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

- Isolate $\phi(t+1)$ to derive a time-stepping scheme to evolve $\phi(t)$, with a time-step Δt ($= 1$ month, say)

From the *straddling rate* to the *spot rate*

- See TR-E 2013 for a whole bunch of Mathematics!!! - upshot:

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

- Isolate $\phi(t+1)$ to derive a time-stepping scheme to evolve $\phi(t)$, with a time-step Δt ($= 1$ month, say)
- Compute $\bar{r}(t)$ from data at a mesh of points t_i, and estimate $\bar{r}''(t)$ by the divided difference — a natural step size is Δt.
  - in practice: $\bar{r}(t)$ is noisy, so the difference is applied to a smoothing least squares fit spline.

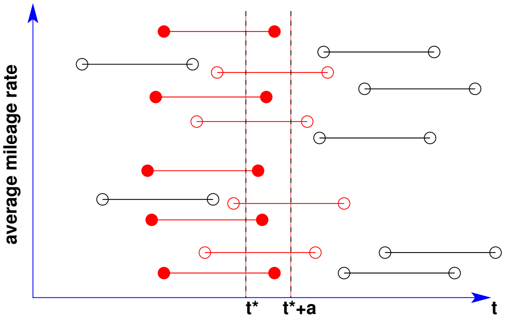
From the *straddling rate* to the *spot rate*

- See TR-E 2013 for a whole bunch of Mathematics!!! - upshot:

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

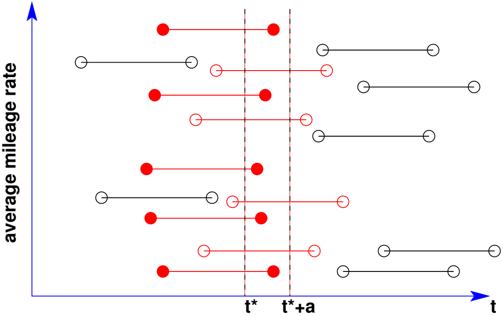
- Isolate $\phi(t+1)$ to derive a time-stepping scheme to evolve $\phi(t)$, with a time-step Δt ($= 1$ month, say)
- Compute $\overline{r}(t)$ from data at a mesh of points t_i, and estimate $\overline{r}''(t)$ by the divided difference — a natural step size is Δt.
  - in practice: $\overline{r}(t)$ is noisy, so the difference is applied to a smoothing least squares fit spline.
- Unfortunately: 2 years of initial data for $\phi(t)$ are required — at the fine scale resolution Δt.

Refinement of the *straddling rate* idea



- Select only the intervals that *straddle* t^* and with right hand ends before $t^* + \alpha$, with $\alpha \leq 1$ year.
- Call resulting average average *straddle rate* $\bar{r}_\alpha(t)$

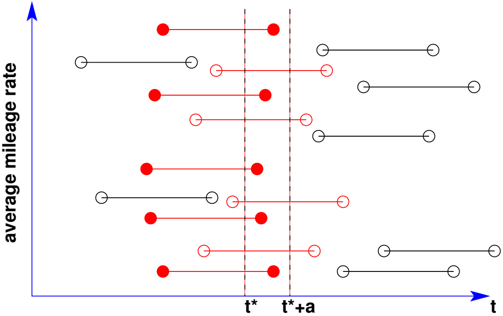
Refinement of the *straddling rate* idea



- Select only the intervals that *straddle* t^* and with right hand ends before $t^* + \alpha$, with $\alpha \leq 1$ year.
- Call resulting average average *straddle rate* $\bar{r}_\alpha(t)$
- Crank the handle to give:

$$\bar{r}''_\alpha(t) = \frac{1}{\alpha}\left[\phi(t+\alpha) - \phi(t)\right] - \frac{1}{\alpha}\left[\phi(t-1+\alpha) - \phi(t-1)\right]$$

Refinement of the *straddling rate* idea



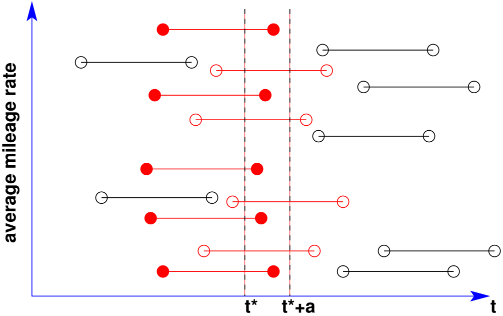
- Select only the intervals that *straddle* t^* and with right hand ends before $t^* + \alpha$, with $\alpha \leq 1$ year.
- Call resulting average average *straddle rate* $\bar{r}_\alpha(t)$

- Crank the handle to give:

$$\bar{r}''_\alpha(t) = \frac{1}{\alpha}\left[\phi(t+\alpha) - \phi(t)\right]$$
$$-\frac{1}{\alpha}\left[\phi(t-1+\alpha) - \phi(t-1)\right]$$

- Gives time-stepping scheme: but only $1+\alpha$ years of initial data required.

Refinement of the *straddling rate* idea



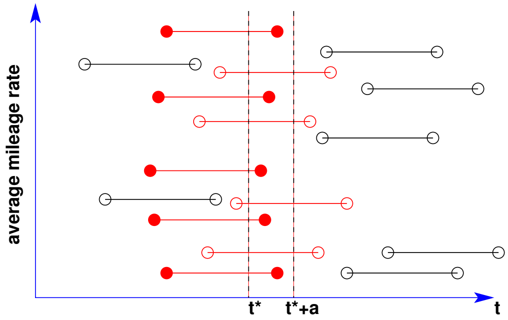
- Select only the intervals that *straddle* t^* and with right hand ends before $t^* + \alpha$, with $\alpha \leq 1$ year.
- Call resulting average average *straddle rate* $\bar{r}_\alpha(t)$

- Crank the handle to give:

$$\bar{r}''_\alpha(t) = \frac{1}{\alpha}\left[\phi(t+\alpha) - \phi(t)\right] - \frac{1}{\alpha}\left[\phi(t-1+\alpha) - \phi(t-1)\right]$$

- Gives time-stepping scheme: but only $1+\alpha$ years of initial data required.
- So interest is in $\alpha \to 0$, which gives $\bar{r}'_\alpha(t) \simeq \phi'(t) - \phi'(t-1)$ (natural meaning)

Refinement of the *straddling rate* idea



- Select only the intervals that *straddle* t^* and with right hand ends before $t^* + \alpha$, with $\alpha \leq 1$ year.
- Call resulting average average *straddle rate* $\overline{r}_\alpha(t)$

- Crank the handle to give:

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

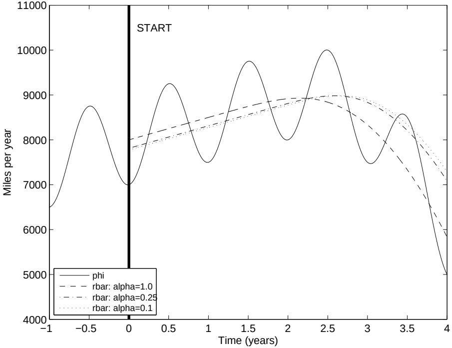
- Gives time-stepping scheme: but only $1 + \alpha$ years of initial data required.
- So interest is in $\alpha \to 0$, which gives $\overline{r}'_\alpha(t) \simeq \phi'(t) - \phi'(t-1)$ (natural meaning)
- $\alpha \to 0$ means fewer and fewer intervals, means noisy $\overline{r}_\alpha(t)$

Synthetic data set-up

- Choose *spot rate*

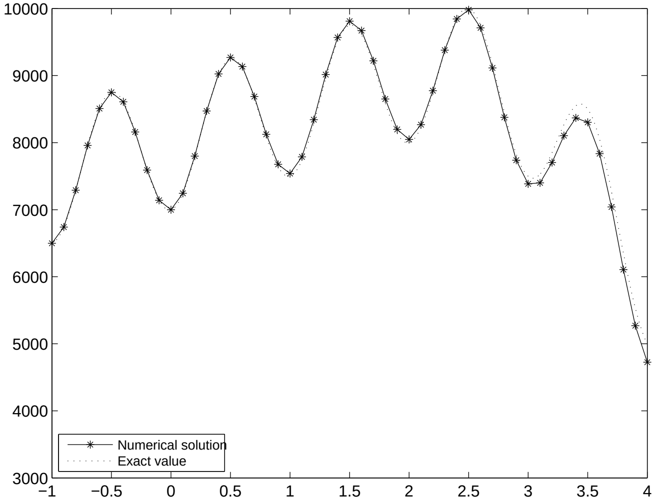
$$\phi(t) = 8000 + 500t - 1000\cos 2\pi t - 1000\left[t-2\right]_+(t-2)^2,$$

- 10^6 vehicles with tests 1 year apart, test dates uniformly distributed through calendar year
- Vehicle i daily mileage drawn from a distribution modulated by $\phi(t)$ and (random) c_i.
- Odometer readings on test dates are synthesised by adding individual vehicle daily totals



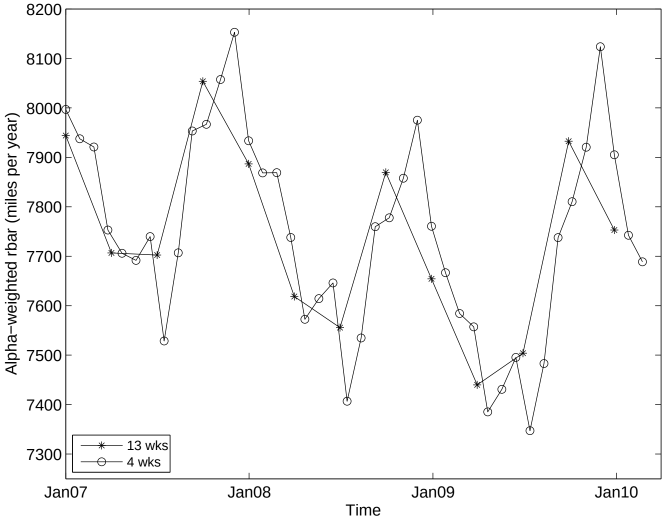
- Periodic component in spot rate $\phi(t)$ is suppressed in straddling rates $\bar{r}_\alpha(t)$

Results with synthetic data: $\alpha = \Delta t = 0.1$ years



- Reconstructed $\phi(t)$ almost indistinguishable from ground truth.

Straddling rates $\bar{r}_\alpha(t)$ for real-world data



- Seasonal component shouldn't be there: underlying assumptions of the theory are broken

Implicit assumptions in the theory...

A1 We assume that tests (odometer readings) are exactly one year apart.

Implicit assumptions in the theory. . .

A1 We assume that tests (odometer readings) are exactly one year apart.

- OKish — theory can be generalised.

Implicit assumptions in the theory...

A1 We assume that tests (odometer readings) are exactly one year apart.

- OKish — theory can be generalised.
- In fact — marginal failure of this assumption can be used to quantify seasonal variation.

Implicit assumptions in the theory...

A1 We assume that tests (odometer readings) are exactly one year apart.

- OKish — theory can be generalised.
- In fact — marginal failure of this assumption can be used to quantify seasonal variation.

A2 We assume that tests occur at same frequency on average throughout year.

Implicit assumptions in the theory...

A1 We assume that tests (odometer readings) are exactly one year apart.

- OKish — theory can be generalised.
- In fact — marginal failure of this assumption can be used to quantify seasonal variation.

A2 We assume that tests occur at same frequency on average throughout year.

- Not true — but easy to fix theory.

Implicit assumptions in the theory...

A1 We assume that tests (odometer readings) are exactly one year apart.

- OKish — theory can be generalised.
- In fact — marginal failure of this assumption can be used to quantify seasonal variation.

A2 We assume that tests occur at same frequency on average throughout year.

- Not true — but easy to fix theory.

A3 We assume that a vehicle's mileage rate is independent of the time of year of at which it is tested (and its odometer is read).

Implicit assumptions in the theory...

A1 We assume that tests (odometer readings) are exactly one year apart.

- OKish — theory can be generalised.
- In fact — marginal failure of this assumption can be used to quantify seasonal variation.

A2 We assume that tests occur at same frequency on average throughout year.

- Not true — but easy to fix theory.

A3 We assume that a vehicle's mileage rate is independent of the time of year of at which it is tested (and its odometer is read).

- Completely wrong. And very hard to fix.

On **A3**: fails because a pattern in new vehicle registrations throughout the year (in the UK).

Conclusions and Further Work (I)

- Incidental data is beautiful! (and useful and cheap)

Conclusions and Further Work (I)

- Incidental data is beautiful! (and useful and cheap)
- (Inadvertently) the MOT set provides vehicle usage data — not intentioned by its release — which is not available elsewhere

Conclusions and Further Work (I)

- Incidental data is beautiful! (and useful and cheap)
- (Inadvertently) the MOT set provides vehicle usage data — not intentioned by its release — which is not available elsewhere (at least in this quantity and detail)

Conclusions and Further Work (I)

- Incidental data is beautiful! (and useful and cheap)
- (Inadvertently) the MOT set provides vehicle usage data — not intentioned by its release — which is not available elsewhere (at least in this quantity and detail)
- Other data sources might enable huge extensions:
  1. Per vehicle emissions data
  2. Fine scale data (month?) for point of first use
  3. Fine scale location data (LLSOA of registered keepers?)
  4. Link vehicles with same registered keeper / address

Conclusions and Further Work (II)

- Methods developed which extract population-level *spot rate* mileage from widely spaced individual vehicle odometer readings. Success with synthetic data.

Conclusions and Further Work (II)

- Methods developed which extract population-level *spot rate* mileage from widely spaced individual vehicle odometer readings. Success with synthetic data.
- UK MOT data set: some fixes/patches to theory are needed.

Conclusions and Further Work (II)

- Methods developed which extract population-level *spot rate* mileage from widely spaced individual vehicle odometer readings. Success with synthetic data.
- UK MOT data set: some fixes/patches to theory are needed.
- Please contact me if you know of other datasets (international) in which odometer readings are systematically collected.
- These methods have the potential to complement / replace existing survey-based / link-flow techniques for estimating population-level mileage.