DOI: 10.5281/zenodo.8056334

The CARS Project.
A tale spanning the entire history of data science

R. Eddie Wilson and James Thomas

RE.Wilson@bristol.ac.uk

16th May 2023

University of BRISTOL
UNIVERSITY OF LEEDS
ADRUK
UKRI Economic and Social Research Council
UKRI Engineering and Physical Sciences Research Council

UK MOT (Ministry of Transport) test

VT20 MOT Test Certificate

VOSA Vehicle & Operator Services Agency

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

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: 4V
Vehicle identification or chassis number: WL0LTGF35X8091395
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: 084907914489518356410227
Design gross weight (goods vehicles): kg
Advisory Notice issued: NO
Test station number: 80572
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)
Print 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 (originally) 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
```

- Note the tests are grouped by year, sorted by test identifier, 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:

```
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 now follow individuals around and infer their mileage (rate) between consecutive test dates
- This feels like a serious privacy issue — but turned out not to be — discuss

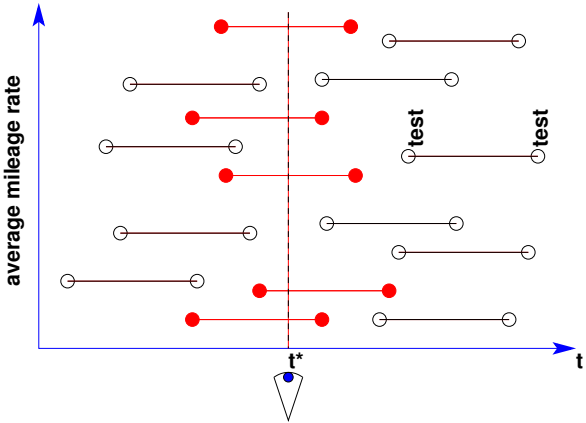
Basic analysis object: an interval and its attributes

- 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

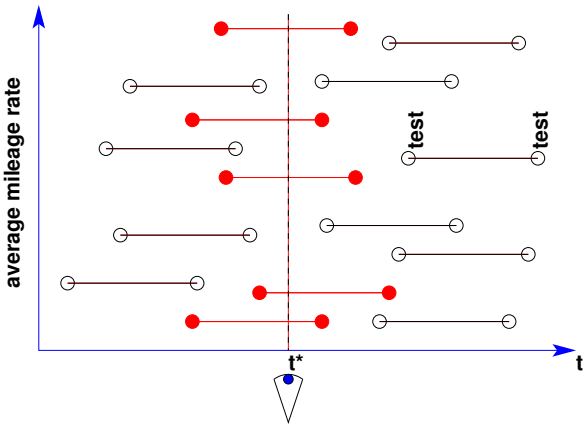
- To which can be linked vehicle-specific attributes:
VAUXHALL, ASTRA LS 8V, WHITE, P (fuel), 1598 (cc), 1999 (year)
- For example, in the **interval** from 2008-08-11 to 2009-08-05
(359 days), I drove 132,299-123,259 = 9,040* miles,
at an **average rate** of 25.18 **miles per day**.
- These mileage 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* $\overline{r}(t)$



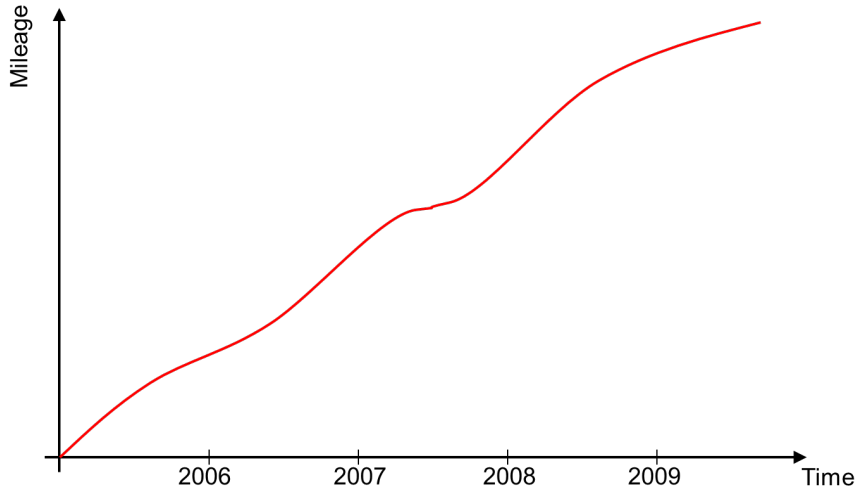
- Select all N intervals that *straddle* a given *observation date* t^*
- Each interval yields an average (per vehicle) rate r_i

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

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

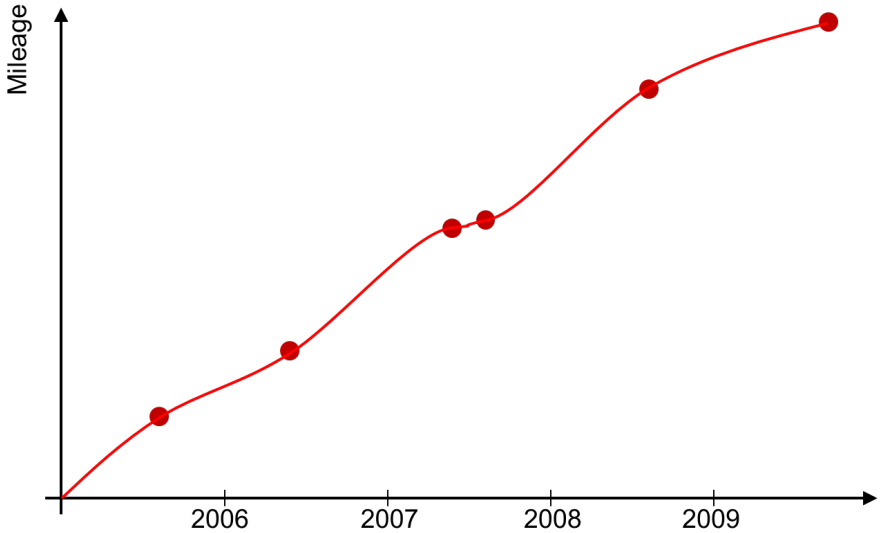
- It is *ok* for annual statistics: choose $t^* = 1/7/2007, 1/7/2008, 1/7/2009$ etc.
- But $\overline{r}(t^*)$ actually incorporates miles driven over the two year span $t^* - 1 \leq t < t^* + 1$
- I spent a lot of time and math fun worrying about this problem

From the Straddling Rate to the Census Date Rate



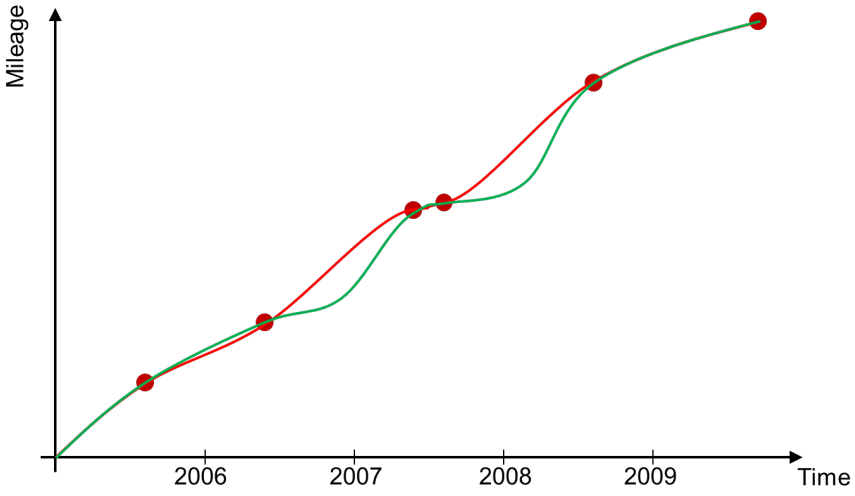
- Progression of a vehicle's odometer with time

From the Straddling Rate to the Census Date Rate



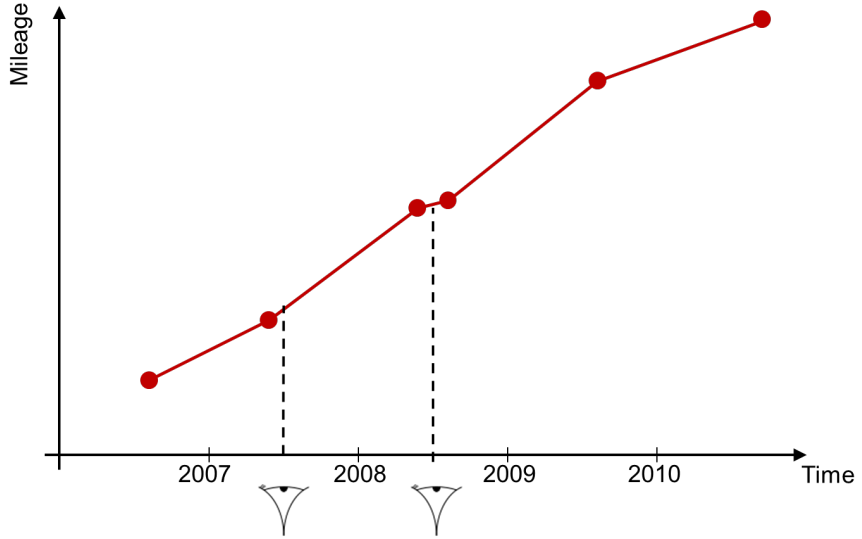
- Progression of a vehicle's odometer with time — with tests

From the Straddling Rate to the Census Date Rate



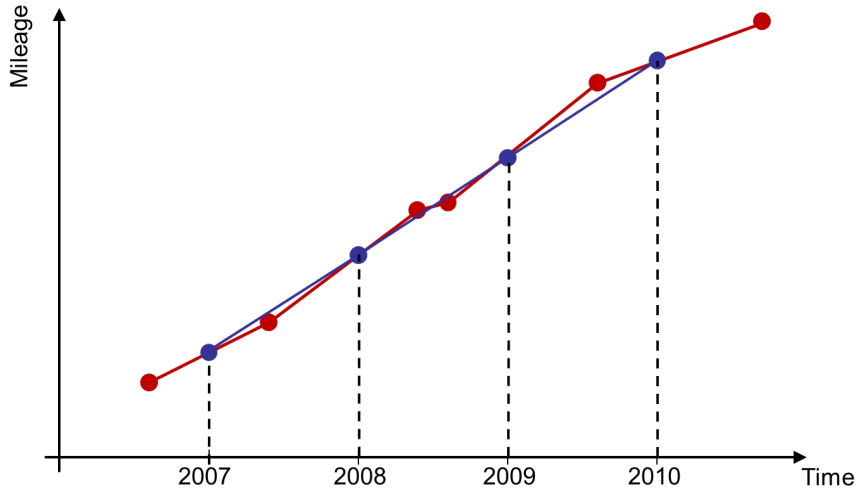
- The tests do not allow you to distinguish the 2 trajectories

From the Straddling Rate to the Census Date Rate



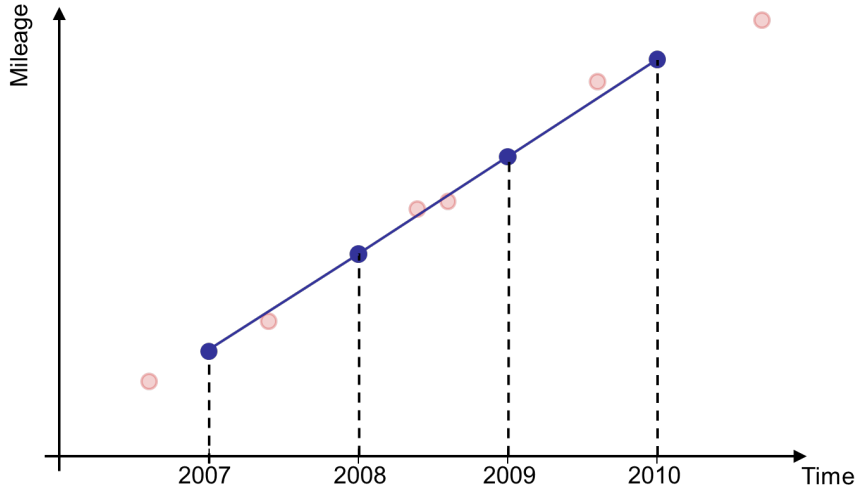
- *Distributions* derived from straddling rate suffer anomalous variance because some

From the Straddling Rate to the Census Date Rate



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

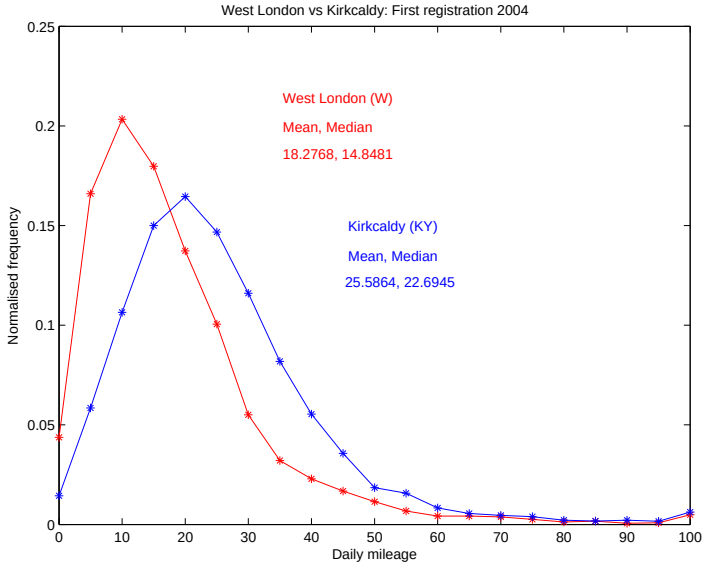
From the Straddling Rate to the Census Date Rate



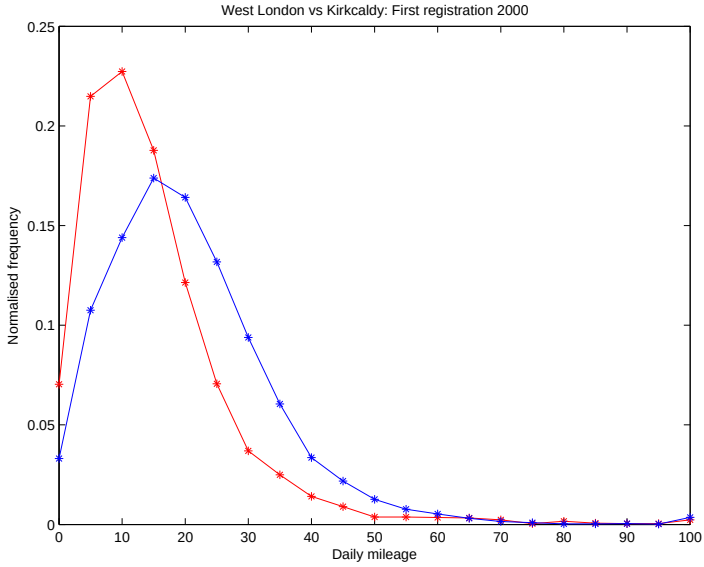
- ... and use the rates between the census dates.

(Also neatly synchronises the data into calendar year comparisons.)

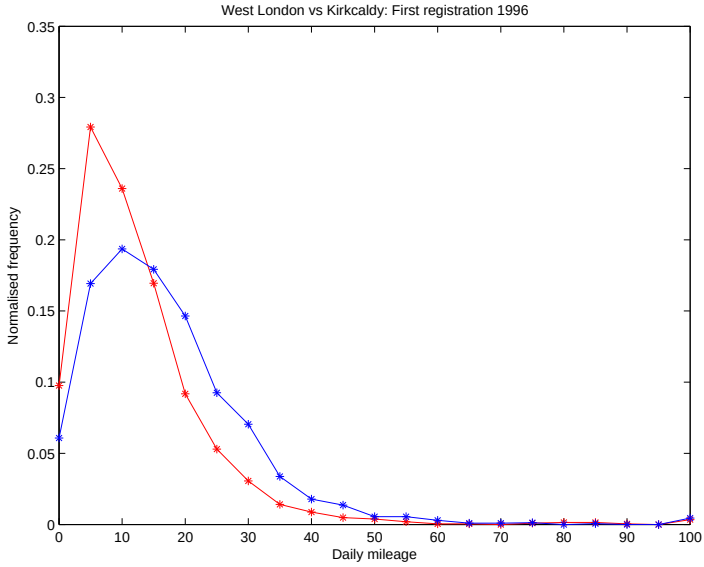
Early impacts — Mileage distributions: new(ish) vehicles



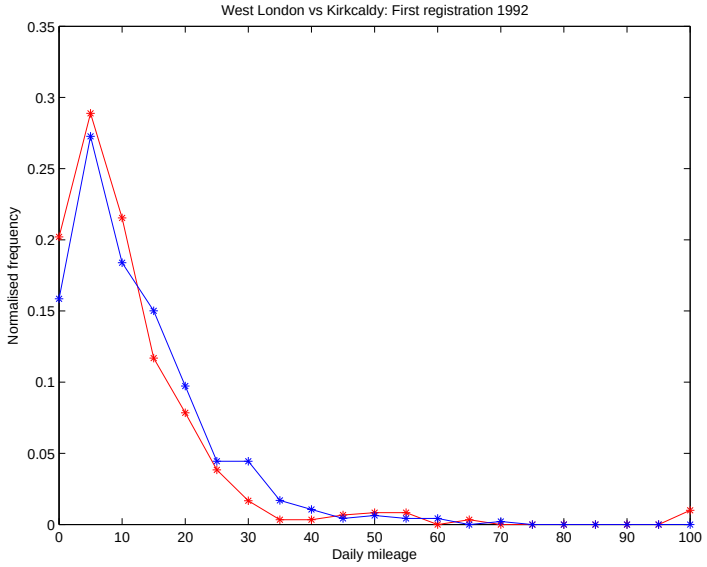
Early impacts — Mileage distributions: older vehicles



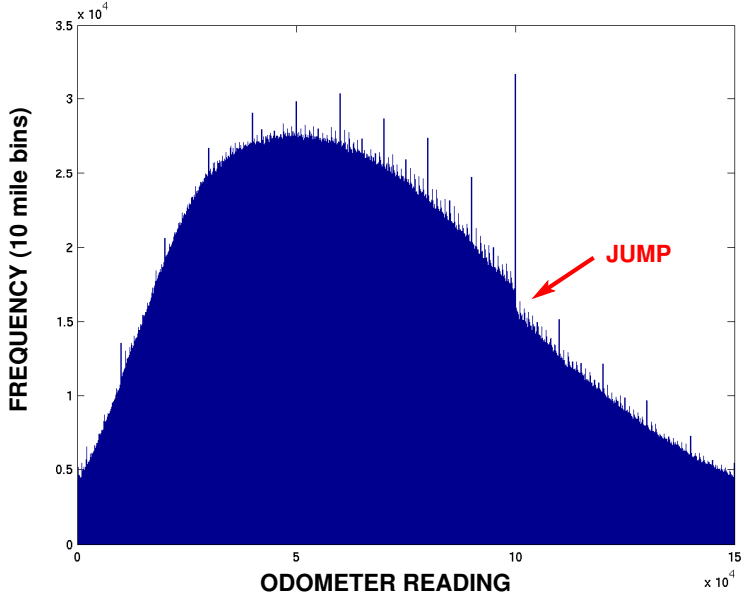
Early impacts — Mileage distributions: even older vehicles



Early impacts — Mileage distributions: old vehicles



It's never as neat as you want (1): Five digit odometer problem



It's never as neat as you want (2): Data entry problems

Solution 1: don't worry about it too much

- Compute rates as if all odometers are perfectly correct
- Reject intervals (*) if rates 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.

It's never as neat as you want (2): Data entry problems

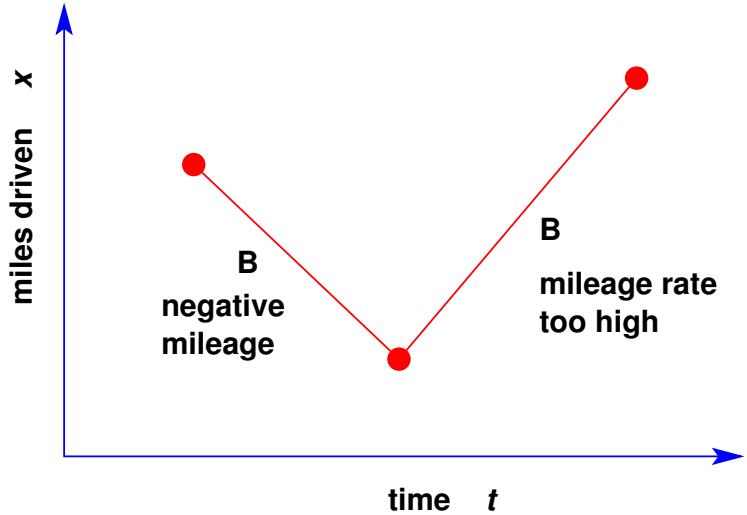
Solution 1: don't worry about it too much

- Compute rates as if all odometers are perfectly correct
- Reject intervals (*) if rates 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.

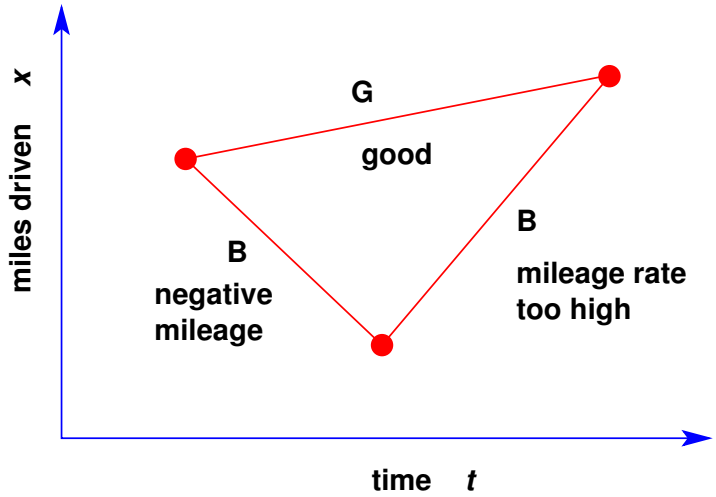
Solution 2: try to identify which individual odometer entries are bad and remove them instead

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



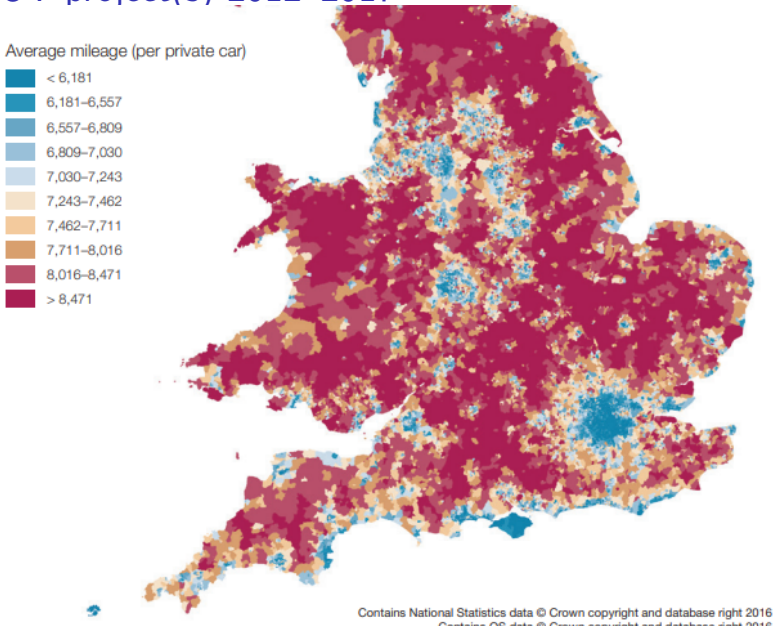
- The middle odometer entry is (probably) erroneous — due to a missing digit in the data entry?

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



- Middle odometer entry is (probably) erroneous — due to a missing digit?
- Spanning interval without the middle test is probably (possibly?) ok

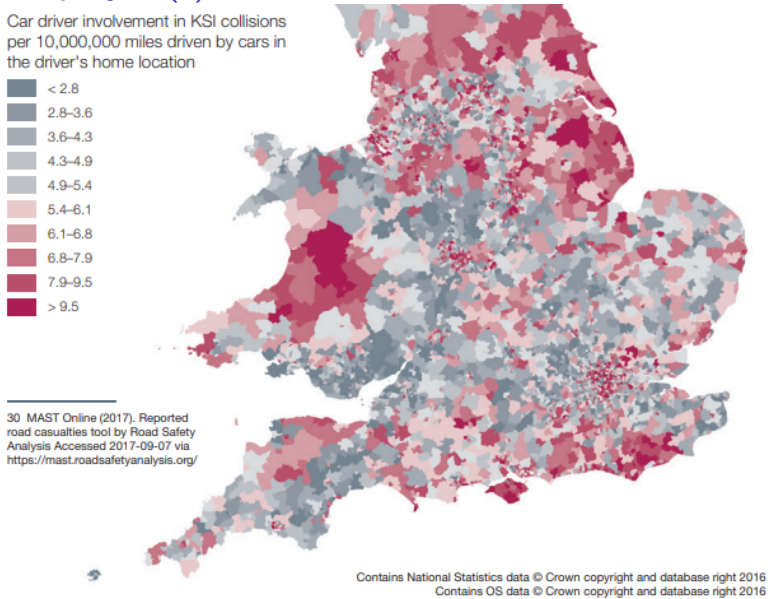
MOT project(s) 2012–2017



- Main trick: (confidential) versions of MOT data linked to vehicle keeper record at LSOA level

- Cairns et al. (2017) *MOToring Along: The lives of cars seen through licensing and test data*

MOT project(s) 2011-2017



- Many surprising correlates: age of vehicles, emissions, poverty indices, safety ...

- Cairns et al. (2017) *MOToring Along: The lives of cars seen through licensing and test data*

And some proper papers (and press coverage)

ELSEVIER

Transportation Research Part D: Transport and Environment

Volume 39, August 2015, Pages 151-164

Use of a novel dataset to explore spatial and social variations in car type, size, usage and emissions

Tim Chatterton [a], Jo Barnes [a], R. Eddie Wilson [b], Jillian Anable [c], Sally Cairns [d]

Present day – what now?

- Opportunity to open up (previously closed) data from DVSA and DVLA
- Provide ongoing access via the ONS TRE

University of BRISTOL
UNIVERSITY OF LEEDS
ADRUK
UKRI Economic and Social Research Council
RAC Foundation

Present day – sources of data (1)

0.2 GB test_result_2005.txt.gz
0.9 GB test_result_2006.txt.gz
0.9 GB test_result_2007.txt.gz
0.9 GB test_result_2008.txt.gz
1.0 GB test_result_2009.txt.gz
1.0 GB test_result_2010.txt.gz
1.0 GB test_result_2011.txt.gz
1.0 GB test_result_2012.txt.gz
1.0 GB test_result_2013.txt.gz
1.0 GB test_result_2014.txt.gz
1.0 GB test_result_2015.txt.gz
1.1 GB test_result_2016.txt.gz
1.1 GB dft_test_result_2018.zip
1.1 GB dft_test_result_2019.zip
1.1 GB dft_test_result_2017.zip
1.1 GB dft_test_result_2020.zip
1.2 GB dft_test_result_2021.zip

```
$ zcat test_result_2007.txt.gz | head -n 4

test_id|vehicle_id|test_date|test_class_id|test_type|test_result|
808298134|151699072|2007-01-01|4|NT|ABR||SK|FORD|MAVERICK|GREEN|P
842444180|1291028996|2007-01-01|4|RT|P|97109|HU|VAUXHALL|ASTRA|WH
348649550|174602976|2007-01-01|4|NT|PRS|28389|M|VAUXHALL|CAVALIER
```

```
$ unzip -l dft_test_result_2020.zip

dft_test_result-[…].csv
dft_test_result-[…].csv
dft_test_result-[…].csv
dft_test_result-[…].csv
```

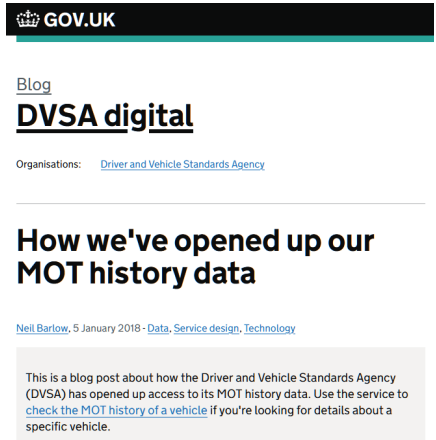
```
$ unzip -p dft_test_result_2020.zip dft_test_result-[…].csv | head -n 4

test_id,vehicle_id,test_date,test_class_id,test_type,test_result,test_mi
677835507,1044704117,2020-04-01,4,RT,P,50331,M,PEUGEOT,EXPERT,RED,DI,156
763132479,1217941099,2020-04-01,7,NT,P,156078,WA,MERCEDES-BENZ,SPRINTER,
635187021,503571165,2020-04-01,7,NT,P,104440,BD,MERCEDES-BENZ,SPRINTER,W
```

- Anonymised data is essentially the same

Present day – sources of data (2)

```
$ curl -H "Accept: application/json+v6" \
       -H "x-api-key: XX" \
       https://beta.check-mot.service.gov.uk/trade/vehicles/mot-tests\?page=1
```


```
"registration": "ZZ99ABC",
"make": "FORD",
"model": "FOCUS",
"firstUsedDate": "2010.11.13",
"fuelType": "Petrol",
"primaryColour": "Yellow",
"vehicleId": "4Tq319nVKLz+25IRaUo79w==",
"registrationDate": "2010.11.13",
"manufactureDate": "2010.11.13",
"engineSize": "1800",
"motTests":[…]
        "completedDate": "2013.11.03 09:33:08",
        "testResult": "PASSED",
        "expiryDate": "2014.11.02",
        "odometerValue": "47125",
        "odometerUnit": "mi",
        "odometerResultType": "READ",
        "motTestNumber": "914655760009",
        "rfrAndComments": []
```

GOV.UK

Blog

DVSA digital

Organisations: Driver and Vehicle Standards Agency

How we've opened up our MOT history data

Neil Barlow, 5 January 2018 - Data, Service design, Technology

This is a blog post about how the Driver and Vehicle Standards Agency (DVSA) has opened up access to its MOT history data. Use the service to check the MOT history of a vehicle if you're looking for details about a specific vehicle.

- DVSA API is accessible via an API key, but is rate limited, and no geography

Present day – newer tools

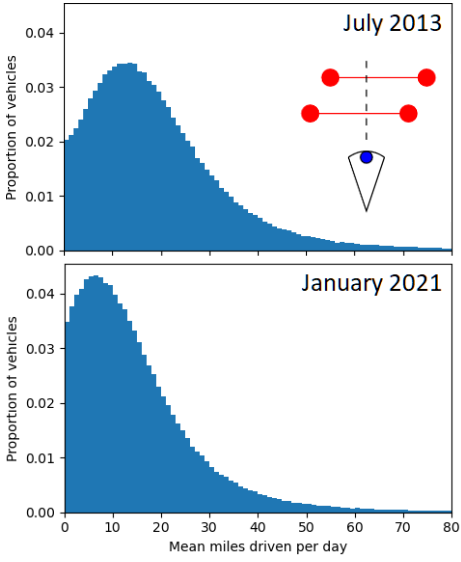
Some tools have changed:

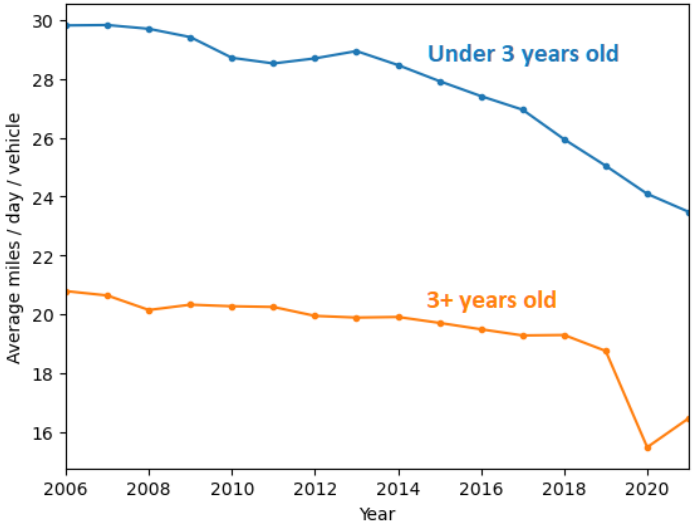
- MATLAB → Python + Pandas/Dask + Matplotlib + Jupyter Notebooks
  → DuckDB
- CSV → Apache Parquet

But other principles remain unchanged:

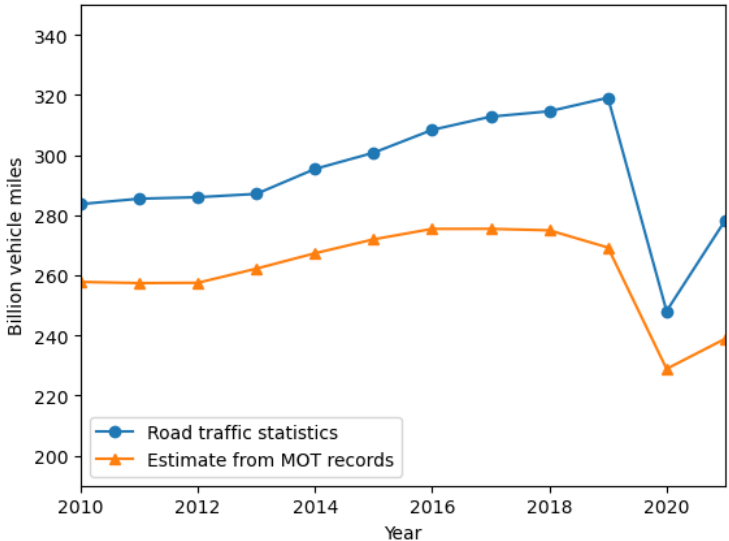
- Work on a sample of the data first
- Use efficient data structures (do you *really* need a 64-bit integer? or a string?)
- Command line tools are still useful: `grep` `head` `tail`
  `ripgrep` `visidata`
  `jq`

The latest data – per vehicle mileage rates have continued to fall. . .

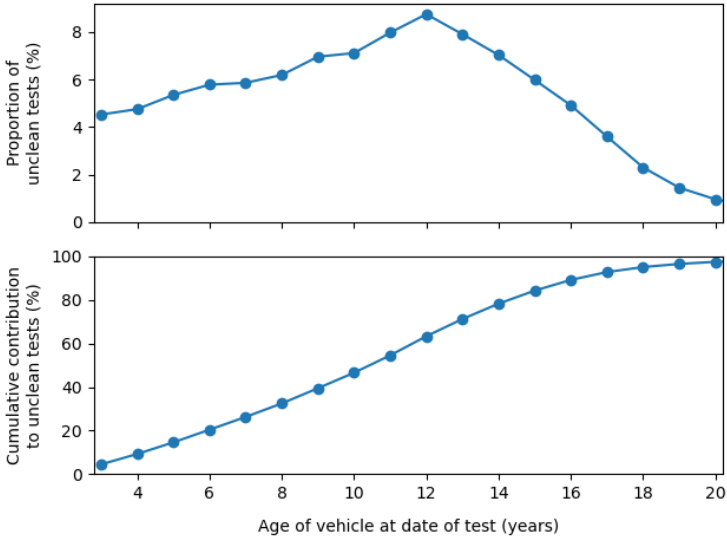




. . . but pre-COVID total miles driven was still increasing



Analysing the latest data – findings (3)



Future aims

Develop a single de-identified, research-ready dataset:

- Linking data on registration and usage patterns of light duty vehicles in GB
  (from DVLA) (from DVSA)
- A resource for addressing transport, environmental and social policy questions
- Inform government action on climate change
  air quality & health
  road safety & taxation

Project timeline: July 2022 – March 2026

University of BRISTOL
UNIVERSITY OF LEEDS
ADR UK
UKRI Economic and Social Research Council
UKRI Engineering and Physical Sciences Research Council