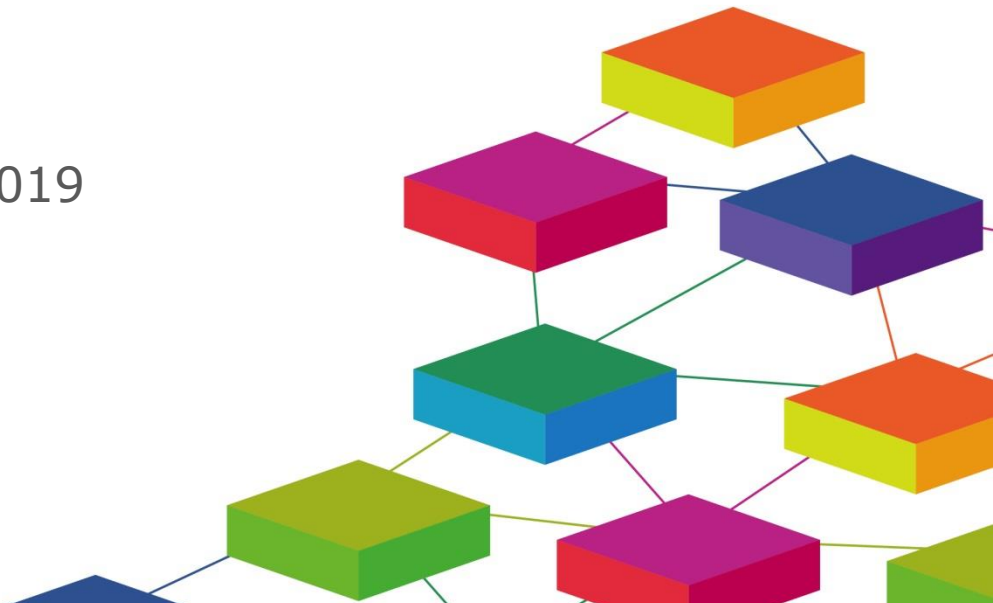


# Using Novel Data to Provide Local Insights

Nik Lomax  
University of Leeds

LIDA Seminar | Leeds | 14 Nov 2019

[@niklomax](https://twitter.com/niklomax)





# Rationale

- Much of the current discussion in data analytics is about 'Big Data' and Big Data methods
- There is a lot of information out there which is very useful for research, but isn't necessarily big data
- I argue that we should use a looser term: 'Novel Data' to provide more flexibility
- The bonus is that much of these data have spatial attributes

# Motivation

- Vision for research does not always equal reality
- A 'Medium Data Toolkit' instead of 'Big Data'

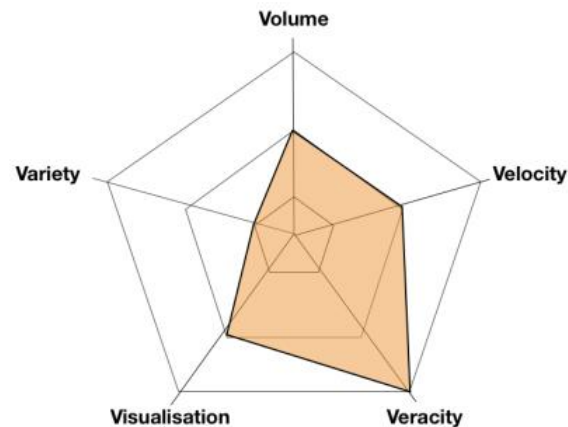
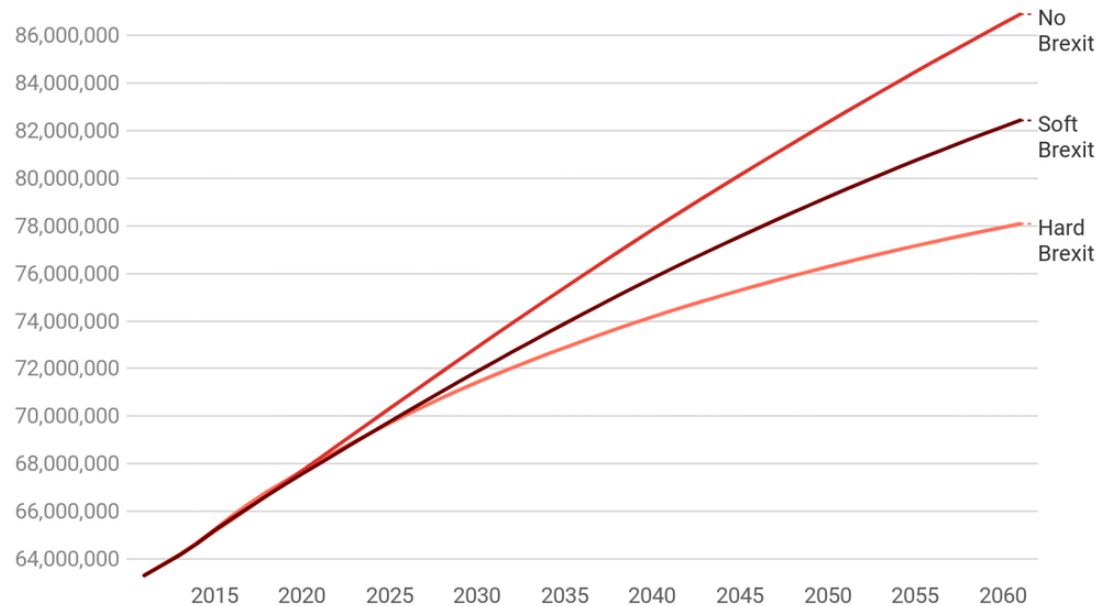


Figure 1: The footfall data from Smart Street Sensor is truly “big” only on the veracity dimension. Otherwise it is mainly a medium sized data.

# Motivation

- As a Geographer, always looking for the spatial dimension to explain phenomena

## Size of UK population under different Brexit scenarios

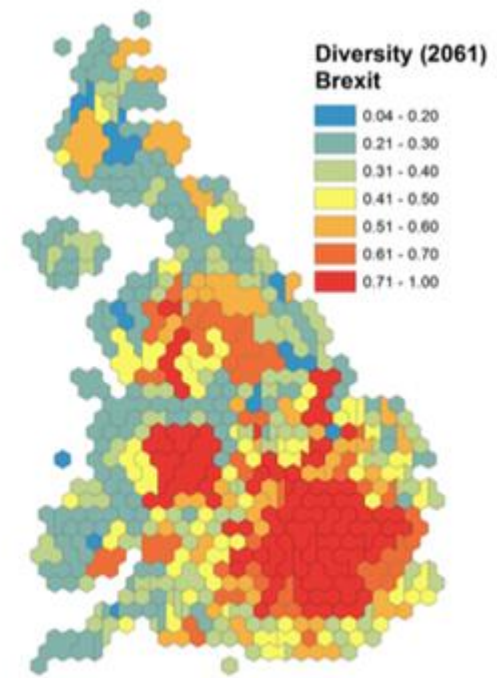
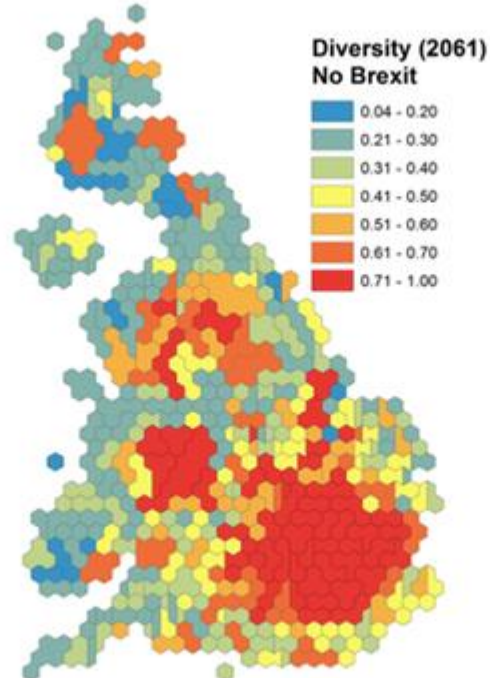
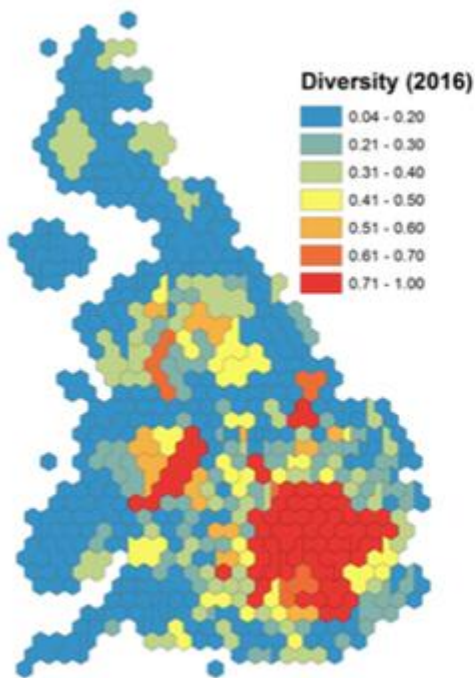


Source: Lomax et al. • [Get the data](#)

Source: Lomax (2019) What the UK population will look like by 2061 under hard, soft or no Brexit scenarios, *The Conversation*, <https://bit.ly/2YUzwCT>

# Motivation

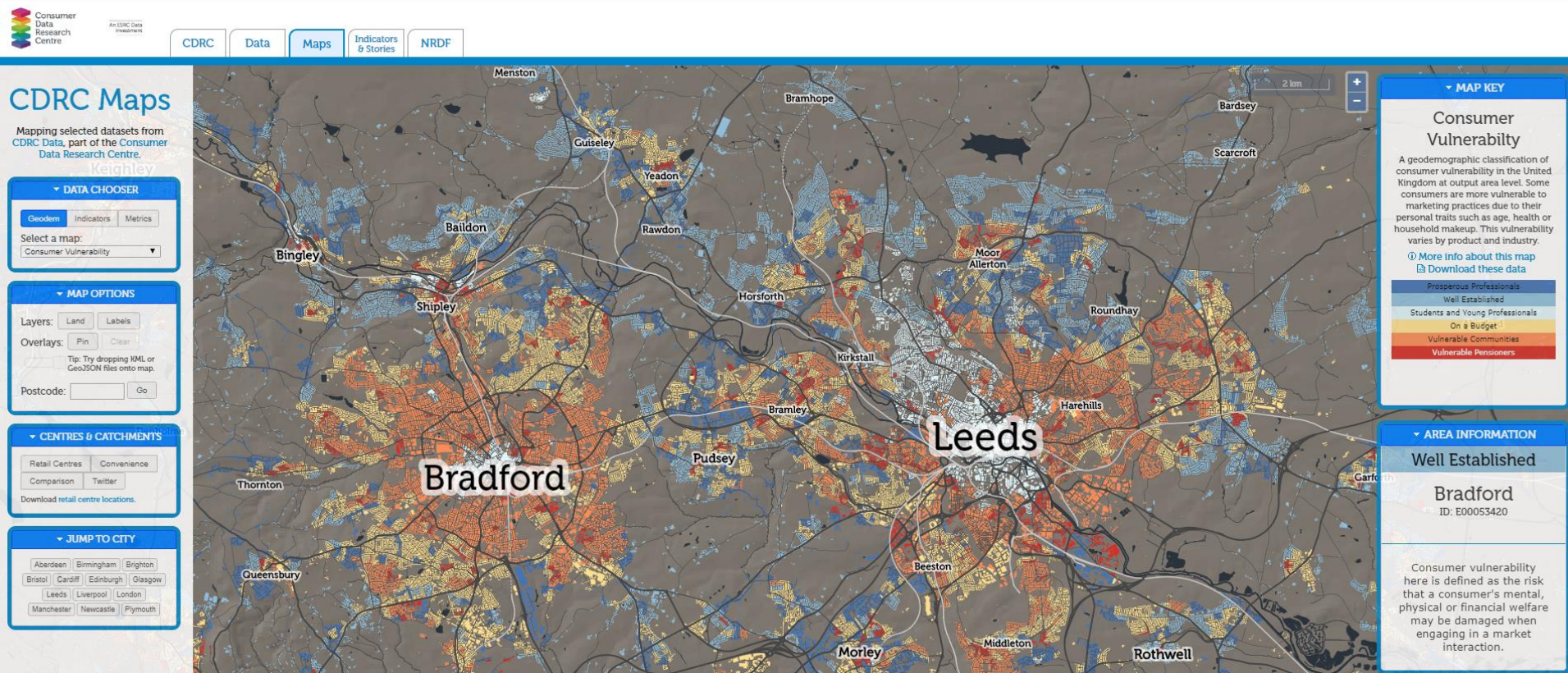
- As a Geographer, always looking for the spatial dimension to explain phenomena



Source: Lomax (2019) What the UK population will look like by 2061 under hard, soft or no Brexit scenarios, *The Conversation*, <https://bit.ly/2YUzwCT>

# Motivation

- People engage with spatial information
- And there is plenty of it



Source: Adcock and Lomax (2018)

<https://maps.cdrc.ac.uk/#/geodemographics/vulnerability/>

# Examples

1. A dataset from a commercial provider and reports the characteristics of properties in the sales and rentals market. Used to **assess local variation in rental prices** ~~and in calculating **rent/price ratios**.~~
2. A dataset from the UK Government's e-petitions website. Used to **estimate the Brexit referendum vote share** for Westminster Parliamentary Constituencies ~~and to **create a classification of Constituencies**.~~

# Example 1: Sales and rental data

A mass market appraisal of the rental market

Calculating rent/price ratio for English housing sub-markets using matched sales and rental data

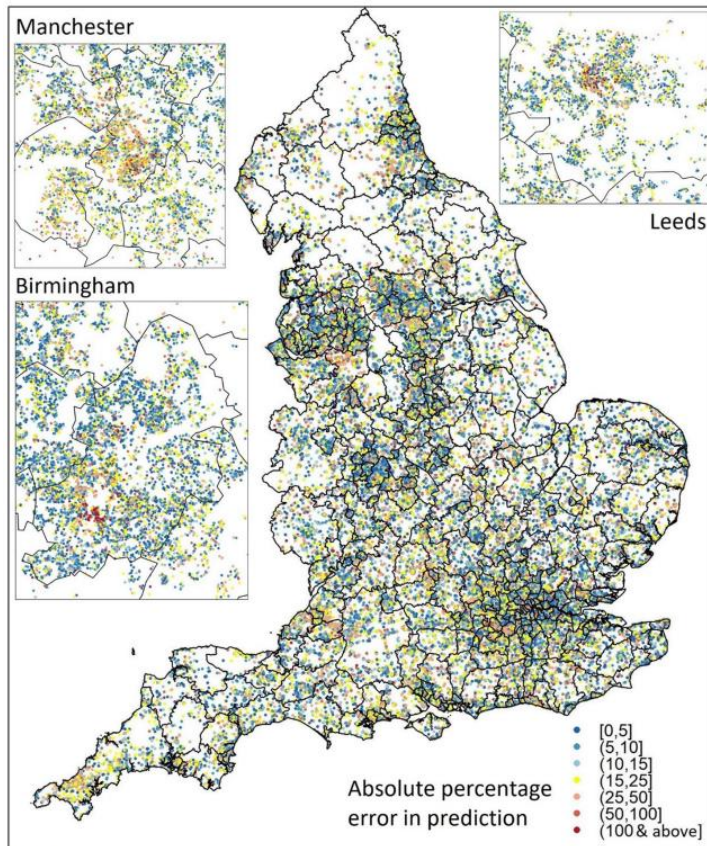


Fig. 2 Absolute percentage prediction error from cubist model

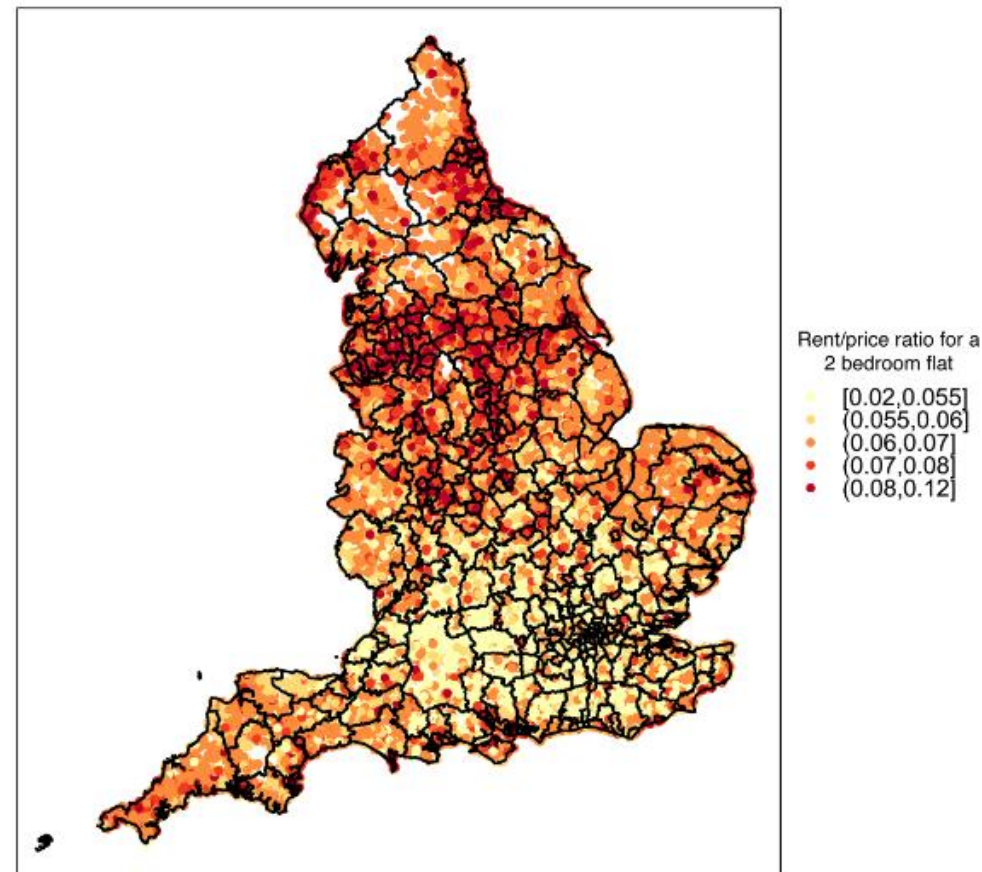


FIGURE 2 Estimated rent/price ratio for two-bedroomed flats for a sample of English postcodes.



# Data (are inherently spatial)

Zoopa

For sale To rent House prices New homes Commercial Overseas

Is2 9jt (within ¼ mile)

Property type

Zoopa > To rent > West Yorkshire > Leeds > Woodhouse Lane property to rent

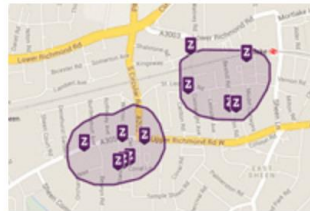
## Property to rent in Woodhouse Lane, Leeds LS



LEEDS  
ACCOMMODATION  
BUREAU

**£525 pcm**

Dene House Court, Off Leicester ...



Pinpoint exactly where to live by using our SmartMaps search

for

No results found

25

Most recent



13 images

## Features

3 bedrooms

1 bathroom

Available from 14th Sep 2018

Well presented

Excellent location

Integrated appliances

Gas central heating

1 reception room

Unfurnished

Garden with lawn and patio

Modern bathroom with P shaped bath

EPC E

Good commuter links

# Introduction

- Mass appraisal of house **sales** market well established
  - Needed for levying of local property taxes
  - Well established field in the literature
- Broad approaches to appraisals:
  - (hedonic) valuation models
  - cost models (based on the materials, design and labour used)
  - use of comparable sales data
  - land value estimations

# Introduction

- Far less emphasis on mass market appraisal in **rental** market
  - But necessary to place a rental value on a property that reflects current market conditions
  - Has received little academic study
  - Primarily due to lack of available data on such transactions



# Introduction

- **Banzhaf and Farooque (2013)** rental values correlate with access to public goods and income levels in Los Angeles
- **Löchli (2010)** accessibility and travel time most important for explaining rents in Zurich
- **Fuss and Koller (2016)** neighbouring property price is most important using hedonic models for Zurich
- **Baron and Kaplan (2010)** impact of 'studentification' on rent is negative in Haifa
- **Prunty (2016)** difference in hedonic features in comparative study of New York and California
- **McCord et al (2014)** use GWR, find a high level of segmentation across localised pockets of the Belfast rental market

# Rationale and contribution

- A lack of insight hampers commercial organisations and local and national governments in understanding rental market.
- We offer a **practical guide** for property professionals and academics wishing to undertake such appraisals and looking for **guidance on the best methods** to use.
- We provide insight in to the property characteristics which most influence rental listing price.



- Rental data from online property search engine Zoopla, cleaned and supplied by WhenFresh
  - 652,454 listings in 2014 and 552,459 in 2015 After cleaning n= **1,063,419**
  - Range of attributes including listing price, number of beds, type of property
- Important to note that listing price  $\neq$  final rental price

- Additional environmental variables
  - Distance from railway station (DFT)
  - Access to Healthy Assets and Hazards (CDRC)
  - School performance (DfE)
  - ACORN – commercial geodemographic profile (CACI)

1. Quasi Poisson generalised linear model (GLM)
2. Machine learning algorithms
  - Tree based: gradient boost (GB) and Cubist
  - Specialist non-linear models: support vector machines (SVM) and multiple adaptive splines (MARS)
3. Practitioner based approach (PBA)
  - rental price is a summary of recently rented similar properties in neighbourhood



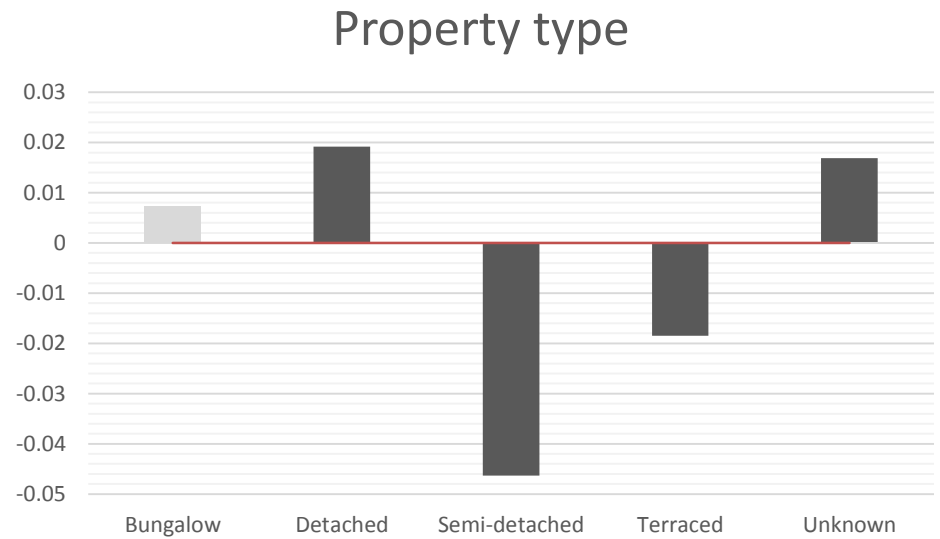


# GLM Results

- quassi Poisson generalised linear model (GLM) used because:
  - skewed distribution of the rental price
  - possible over-dispersion
- Essential step prior to Machine Learning – Does the data capture **dynamics of the housing market in a sensible manner?**
- 63 variables
- Squared correlation between observed and in-sample predicted  $r^2 = 0.738$  on log of rental price
- $r^2$  drops to 0.54 on original scale

# GLM Results

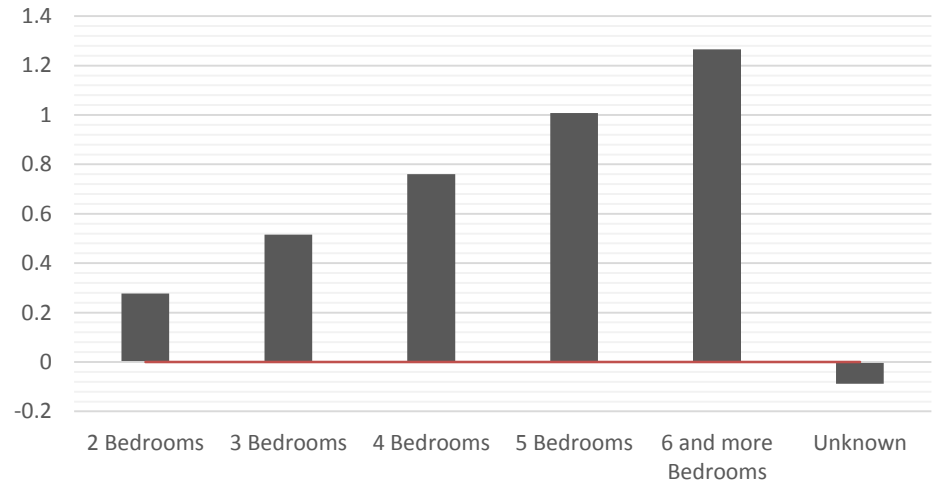
Attribute	N/median	estimate	std error	t	
Intercept	487253	6.451	0.0067	957.7	***
Flat	212275				
Bungalow	11617	0.0073	0.0059	1.2	
Detached	31996	0.0192	0.0037	5.2	***
Semi-detached	54410	-0.0463	0.0032	-14.5	***
Terraced	111087	-0.0185	0.0025	-7.4	***
Unknown	65868	0.0169	0.0026	6.4	***



# GLM Results

Attribute	N/median	estimate	std error	t	
Intercept	487253	6.451	0.0067	957.7	***
1 Bedroom	94379				
2 Bedrooms	192236	0.2772	0.0024	116.8	***
3 Bedrooms	123546	0.5157	0.0028	186.7	***
4 Bedrooms	41505	0.7607	0.0033	228.6	***
5 Bedrooms	12558	1.008	0.0043	235.7	***
6 and more Bedrooms	7097	1.265	0.0051	248.3	***
Unknown	15932	-0.0881	0.005	-17.7	***

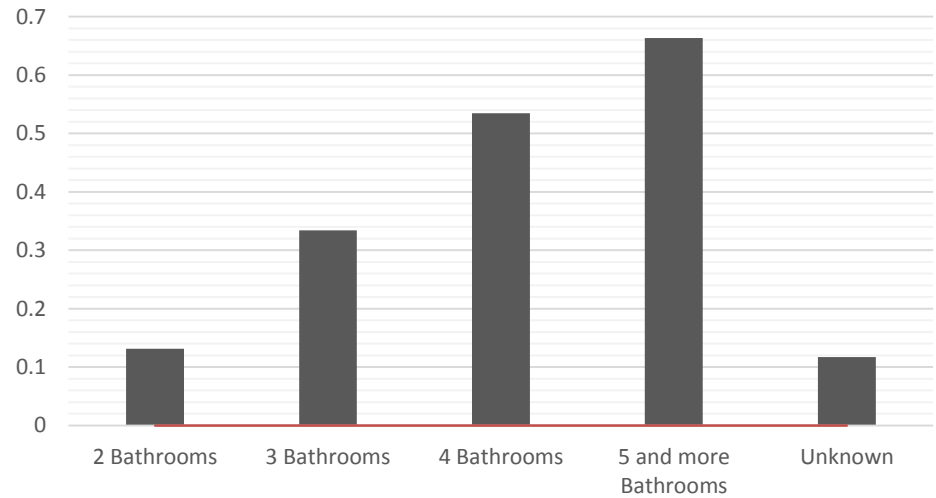
## Number of bedrooms



# GLM Results

Attribute	N/median	estimate	std error	t	
Intercept	487253	6.451	0.0067	957.7	***
1 Bathroom	194157				
2 Bathrooms	45440	0.1314	0.0026	50.8	***
3 Bathrooms	6767	0.3343	0.0047	71.2	***
4 Bathrooms	1150	0.5347	0.0085	63.3	***
5 and more Bathrooms	622	0.6633	0.0107	62	***
Unknown	239117	0.1169	0.0024	48.2	***

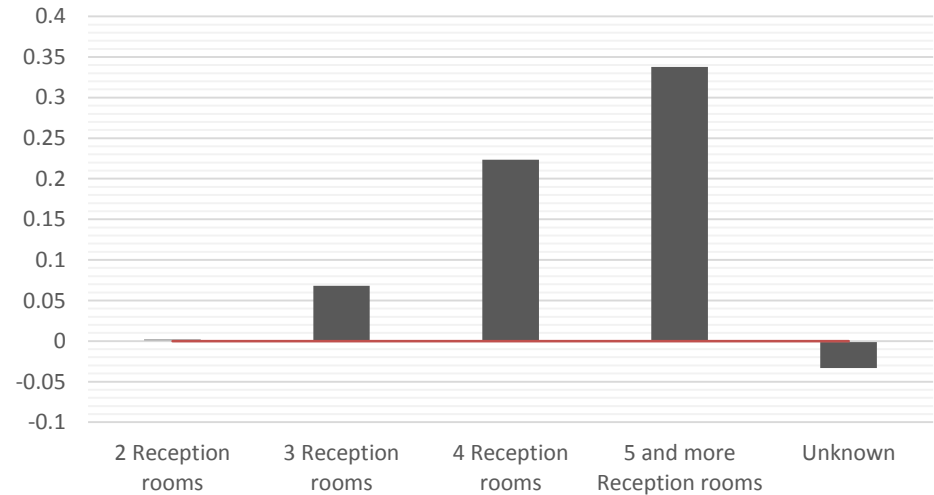
## Number of bathrooms



# GLM Results

Attribute	N/median	estimate	std error	t	
Intercept	487253	6.451	0.0067	957.7	***
1 Reception room	159999				
2 Reception rooms	41912	0.002	0.003	0.7	
3 Reception rooms	4921	0.0681	0.006	11.4	***
4 Reception rooms	723	0.2235	0.0113	19.8	***
5 and more Reception rooms	191	0.3379	0.0189	17.9	***
Unknown	279507	-0.0333	0.0024	-13.9	***

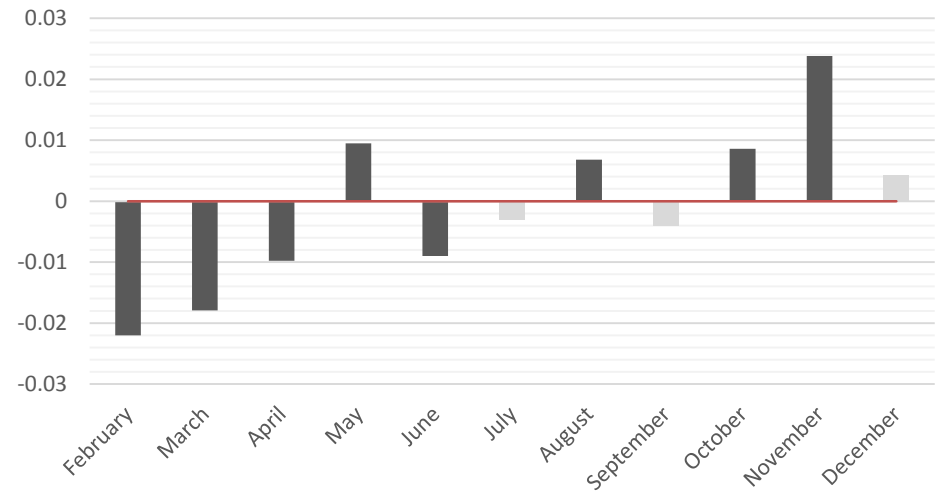
## Number of reception rooms



# GLM Results

Attribute	N/median	estimate	std error	t	
Intercept	487253	6.451	0.0067	957.7	***
January	50988				
February	37309	-0.022	0.0036	-6.2	***
March	39601	-0.0179	0.0035	-5.1	***
April	38037	-0.0098	0.0035	-2.8	**
May	40414	0.0095	0.0034	2.8	**
June	42095	-0.009	0.0034	-2.7	**
July	44808	-0.0031	0.0033	-0.9	
August	39791	0.0068	0.0035	2	*
September	37994	-0.0041	0.0035	-1.2	
October	43005	0.0086	0.0034	2.5	*
November	42037	0.0238	0.0034	7	***
December	31174	0.0042	0.0038	1.1	

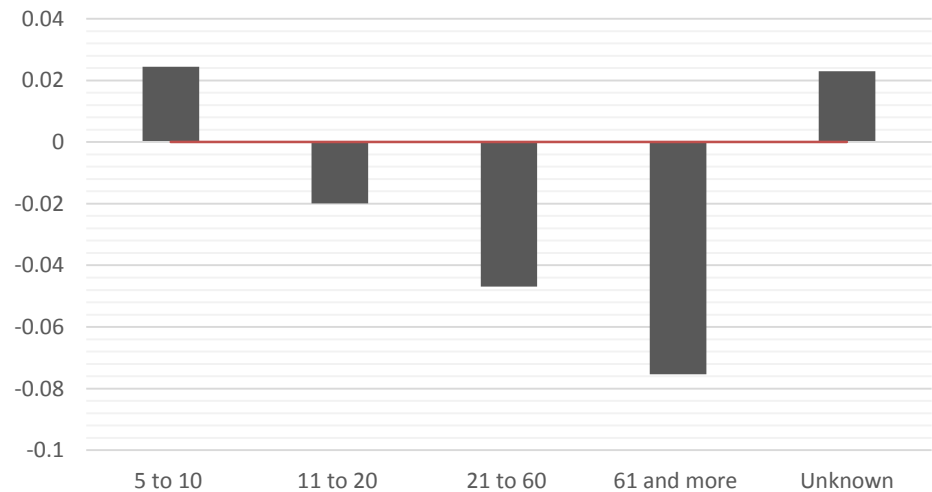
Month of listing



# GLM Results

Attribute	N/median	estimate	std error	t	
Intercept	487253	6.451	0.0067	957.7	***
Up to 4	24094				
5 to 10	14610	0.0244	0.0055	4.4	***
11 to 20	23114	-0.0199	0.005	-3.9	***
21 to 60	39969	-0.0469	0.0046	-10.3	***
61 and more	29423	-0.0754	0.005	-15.2	***
Unknown	356043	0.023	0.0037	6.2	***

## Webpage visits per day

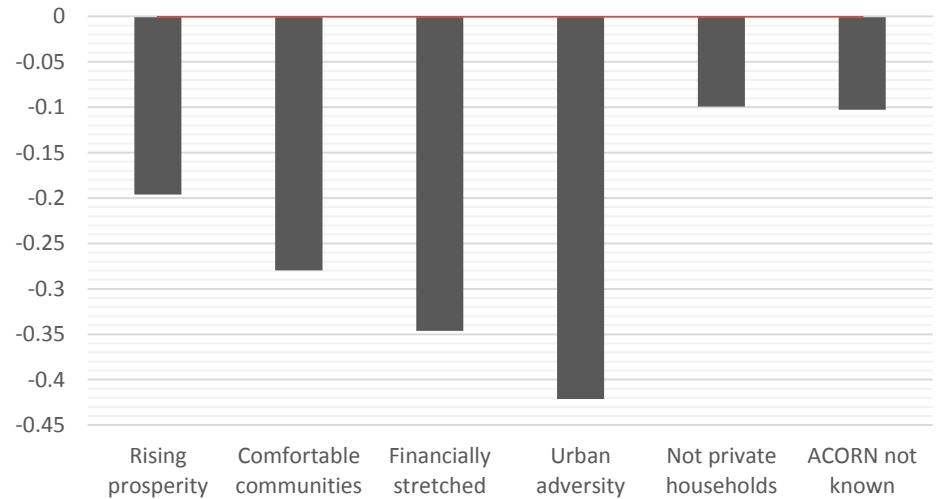




# GLM Results

Attribute	N/median	estimate	std error	t	
Intercept	487253	6.451	0.0067	957.7	***
Affluent achievers	60017				
Rising prosperity	136624	-0.1961	0.0026	-74.5	***
Comfortable communities	98779	-0.2798	0.0028	-99.7	***
Financially stretched	92146	-0.3463	0.0031	-112.9	***
Urban adversity	96472	-0.4212	0.0031	-134.3	***
Not private households	3008	-0.0994	0.009	-11.1	***
ACORN not known	207	-0.1028	0.0274	-3.8	***

## Acorn classification



# GLM Results

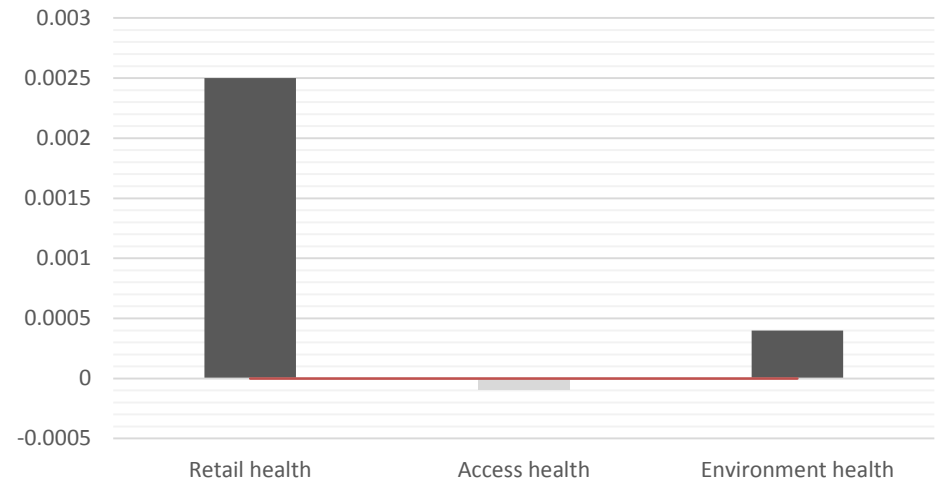
Attribute	N/median	estimate	std error	t	
Intercept	487253	6.451	0.0067	957.7	***
Log Distance from the City of London	113.95km	-0.2862	0.00079	-363.2	***
Log Distance from railway station	1.11km	-0.0204	0.001	-20	***

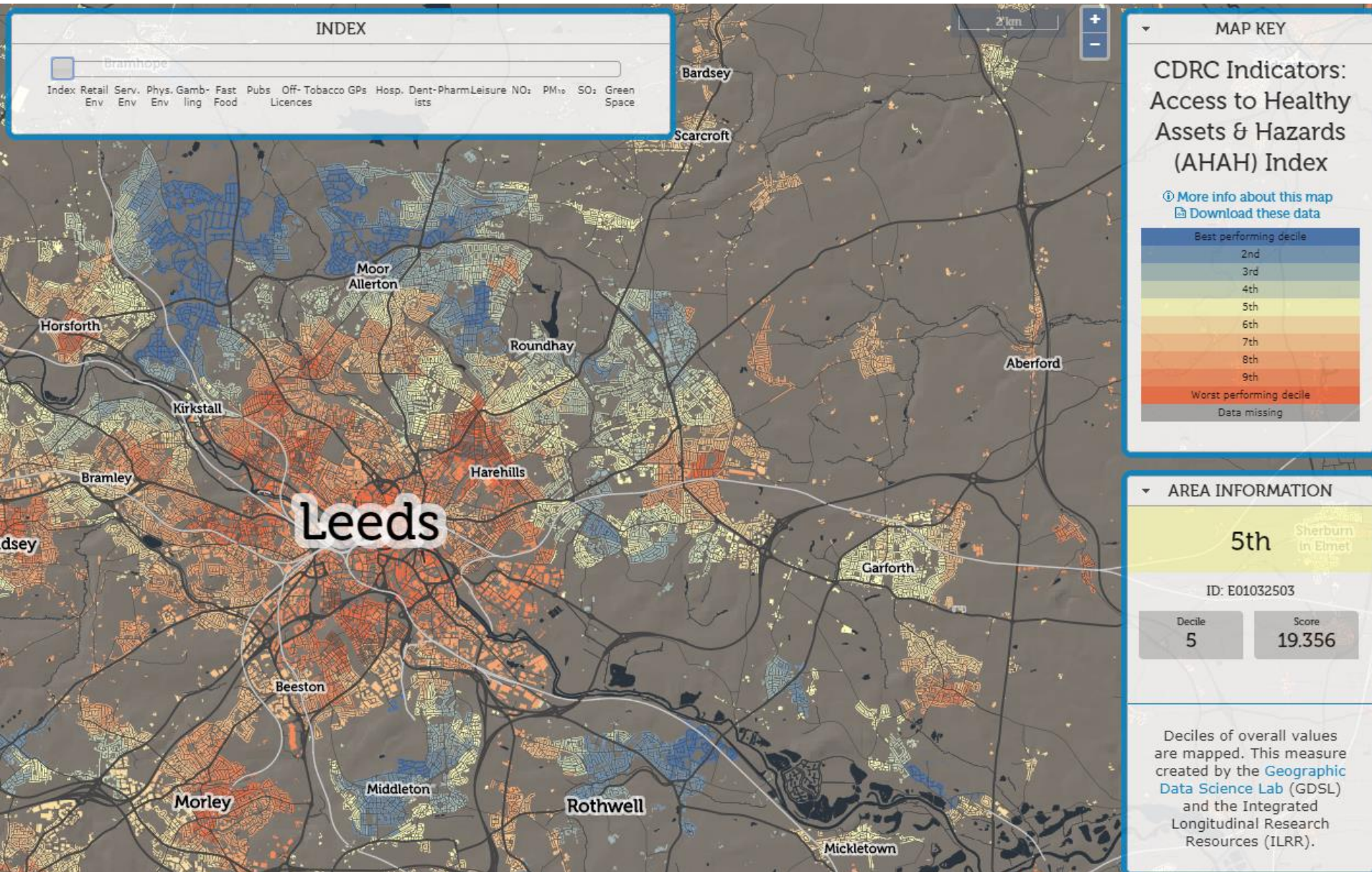


# GLM Results

Attribute	N/median	estimate	std error	t	
Intercept	487253	6.451	0.0067	957.7	***
Retail health	30.53	0.0025	0.00005	52.2	***
Access health	7.21	-0.0001	0.00008	-1.9	
Environment health	25.32	0.0004	0.00004	10.5	***

## Environment and amenity





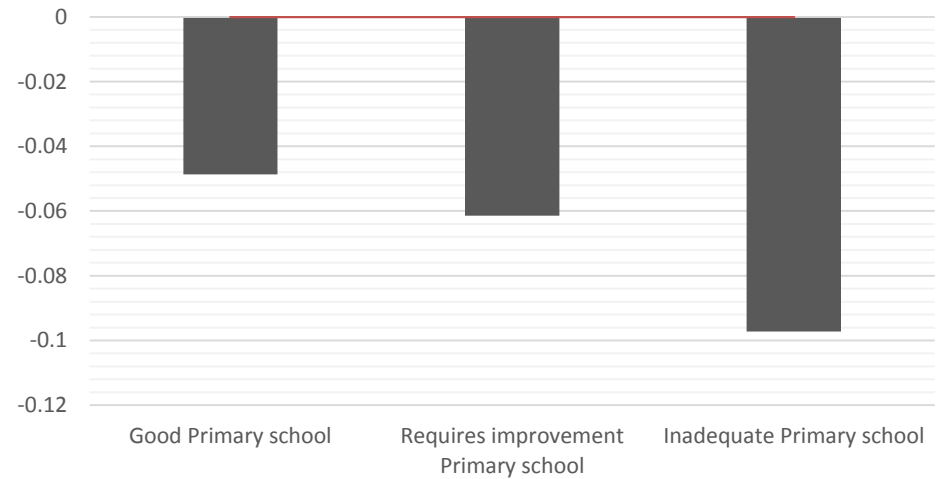
## Access to Healthy Assets and Hazards (AHAH)

Daras, Konstantinos; Green, Mark; Davies, Alec; Singleton, Alex; Barr, Benjamin. (2017).

# GLM Results

Attribute	N/median	estimate	std error	t	
Intercept	487253	6.451	0.0067	957.7	***
Outstanding Primary	91869				
Good Primary	308287	-0.0487	0.0019	-26.2	***
Requires improvement Primary	79841	-0.0614	0.0026	-24	***
Inadequate Primary	7256	-0.0972	0.0071	-13.7	***

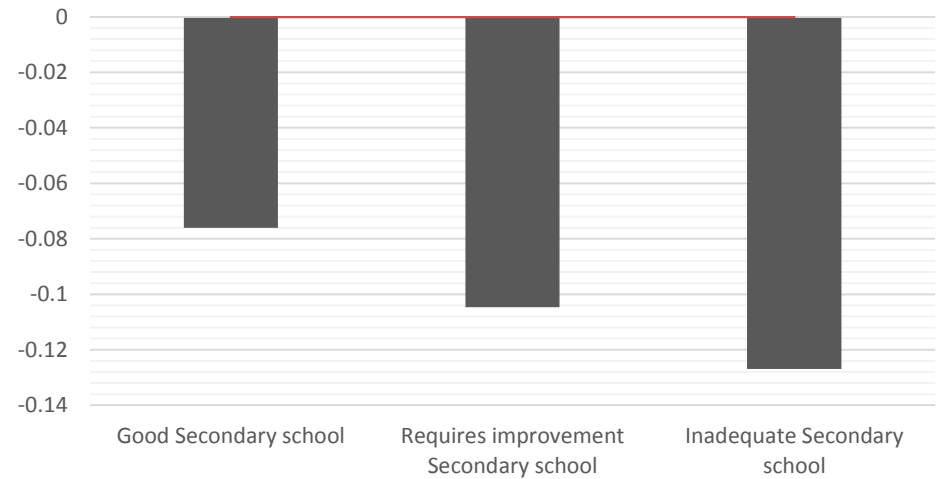
## Primary school Ofsted score



# GLM Results

Attribute	N/median	estimate	std error	t	
Intercept	487253	6.451	0.0067	957.7	***
Outstanding Secondary	1119014				
Good Secondary	245070	-0.076	0.0018	-43.2	***
Requires improvement Secondary	96715	-0.1047	0.0024	-44.6	***
Inadequate Secondary	26454	-0.1269	0.0044	-28.9	***

## Secondary school Ofsted score



# Machine Learning

- Algorithms fitted within the machine learning paradigm of the caret package in R
- Primarily tree based algorithms:
  1. Gradient boost (GB)
  2. Cubist
- Specialist non-linear models:
  3. Support vector machines (SVM)
  4. Multiple adaptive splines (MARS)



# Practitioner approach

- Combines price of recently rented similar properties in neighbourhood
- Comparable properties must be of the same property type, have the same number of bedrooms, bathrooms and reception rooms and be in the same ACORN group.
- Inverse distance weight used (closer properties contribute more)



# Results – comparing r2

Testing	PBA	GLM	GB	SVM	Cubist	MARS	Ensemble
Jan	0.55	0.56	0.62	0.56	<b>0.65</b>	0.47	0.67
Feb	0.53	0.55	0.61	0.57	<b>0.64</b>	0.50	0.65
Mar	0.48	0.49	0.52	0.48	<b>0.56</b>	0.43	0.57
Apr	0.52	0.55	0.58	0.55	<b>0.65</b>	0.47	0.65
May	0.41	0.44	0.48	0.44	<b>0.50</b>	0.39	0.51
Jun	0.53	0.59	0.63	0.60	<b>0.67</b>	0.52	0.68
Jul	0.55	0.58	0.66	0.61	<b>0.66</b>	0.53	0.69
Aug	0.51	0.53	0.58	0.56	<b>0.62</b>	0.48	0.63
Sep	0.52	0.57	0.64	0.57	<b>0.68</b>	0.51	0.69
Oct	0.49	0.56	0.59	0.57	<b>0.63</b>	0.49	0.64
Nov	0.52	0.57	0.63	0.54	<b>0.64</b>	0.48	0.66
Dec	0.51	0.56	0.61	0.57	<b>0.66</b>	0.51	0.67
ALL	0.51	0.54	0.59	0.55	<b>0.63</b>	0.48	0.64

# Results – comparing median percentage prediction error

Testing	PBA	GLM	GB	SVM	Cubist	MARS	Ensemble
Jan	<b>7.95</b>	16.62	16.07	13.80	13.59	20.73	13.44
Feb	<b>8.17</b>	16.55	15.22	13.30	13.46	20.66	13.04
Mar	<b>8.35</b>	16.28	15.24	13.32	13.22	20.66	13.14
Apr	<b>8.47</b>	15.83	15.00	13.13	13.31	20.49	12.95
May	<b>8.62</b>	15.94	14.85	12.99	13.04	20.01	13.32
Jun	<b>8.82</b>	16.02	15.07	13.39	13.36	19.83	13.04
Jul	<b>9.23</b>	15.68	14.82	12.97	12.91	19.69	12.87
Aug	<b>9.26</b>	15.70	14.74	13.02	12.90	19.92	12.91
Sep	<b>9.26</b>	15.12	14.40	12.55	12.38	19.25	12.40
Oct	<b>9.80</b>	16.14	15.17	13.40	13.39	19.67	13.39
Nov	<b>9.95</b>	16.70	15.76	13.83	13.89	19.64	14.46
Dec	<b>9.73</b>	15.77	14.76	13.20	12.35	19.36	13.00
ALL	<b>9.07</b>	16.04	15.11	13.25	13.18	20.01	13.06

# Results – distribution of percentage error

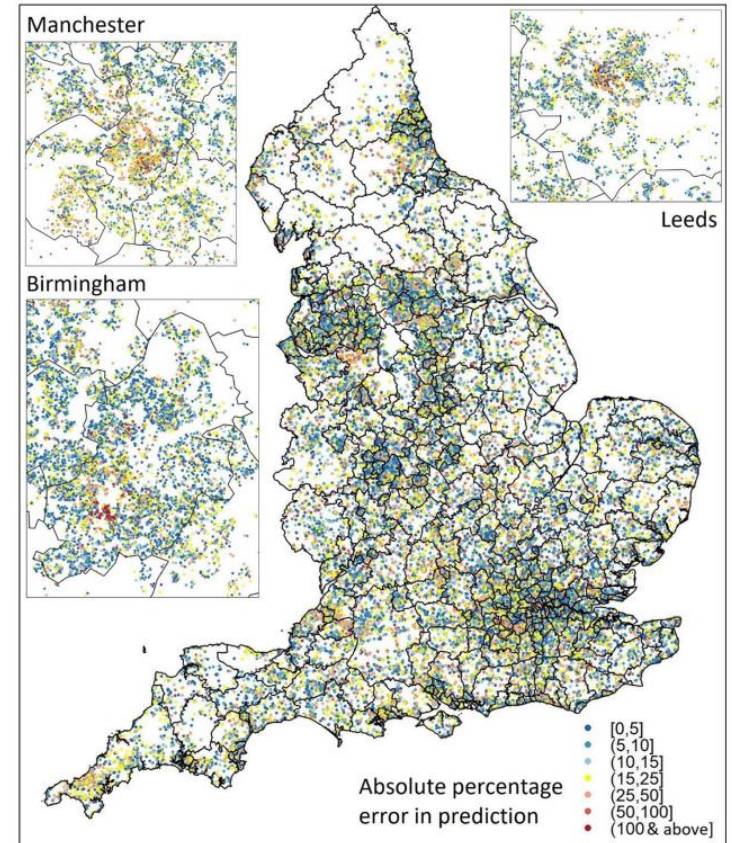
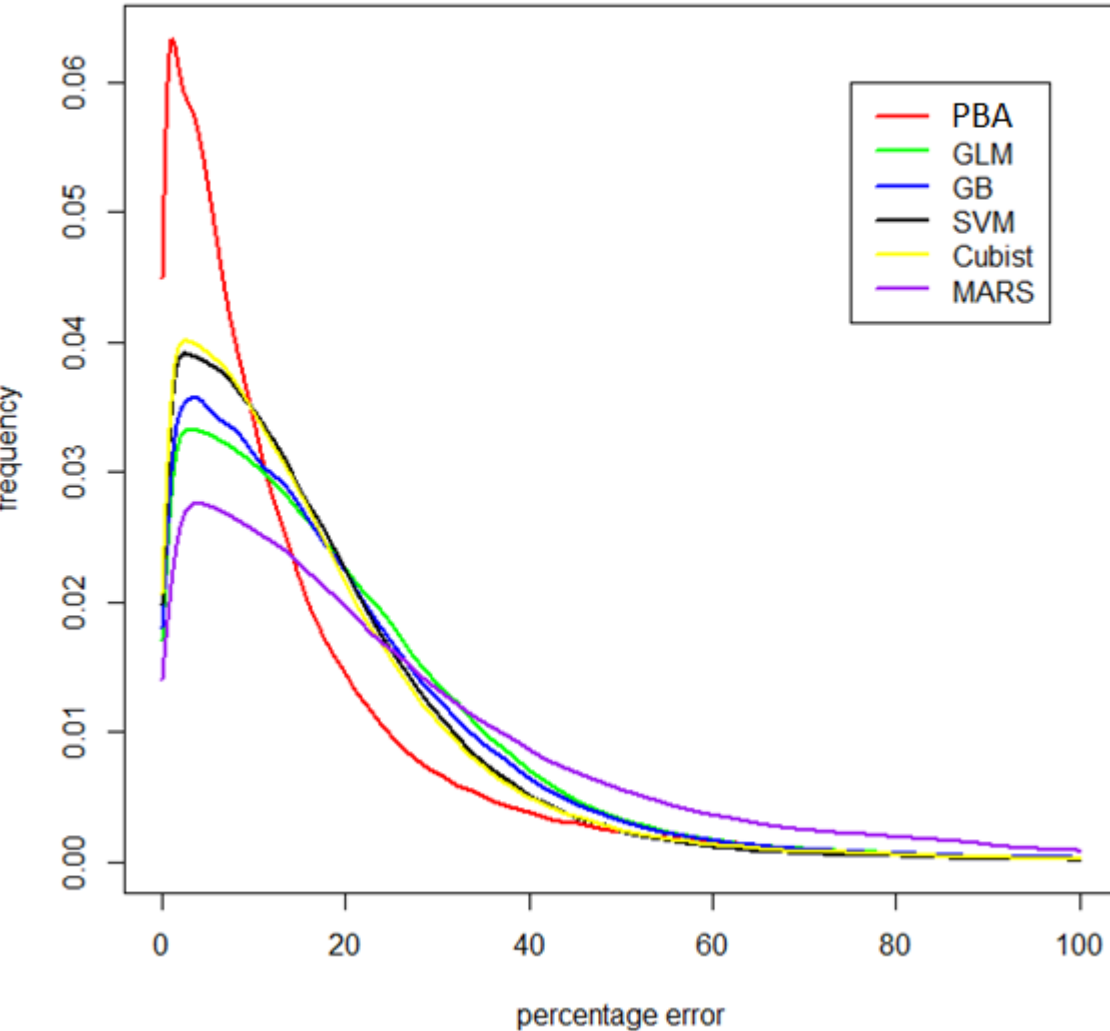


Fig. 2 Absolute percentage prediction error from cubist model

# Conclusions

- What increases rental price (from GLM):
  - Number of rooms in the property
  - proximity to central London
  - Proximity to railway stations
  - being located in more affluent neighbourhoods
  - being close to local amenities
  - Being close to better performing schools



# Conclusions

- Practitioner approach produced appraisals that have much smaller percentage error whilst the other approaches have better  $r^2$
- Our preferred Machine Learning Algorithm is Cubist



# And conclusions from the other study...

- An investor with £10million to invest and looking to maximise their gross rental yield would, rather than investing in a couple of properties in West London, be better off investing in hundreds of properties in the less affluent areas of the Midlands and North.

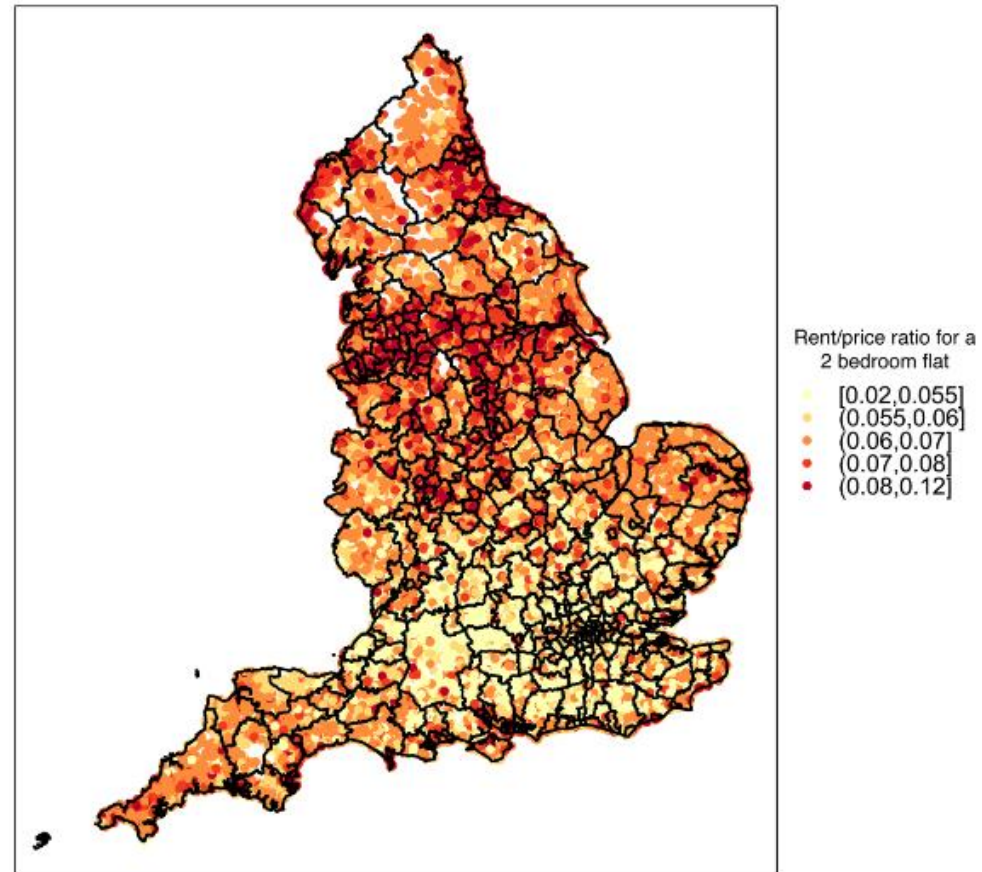
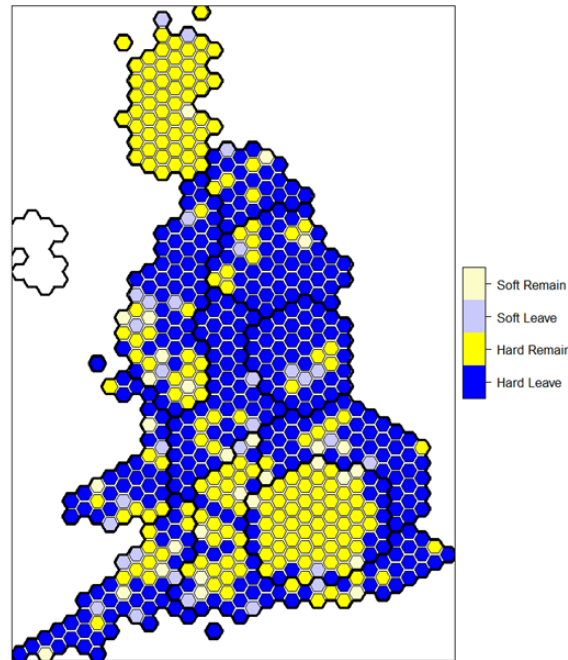
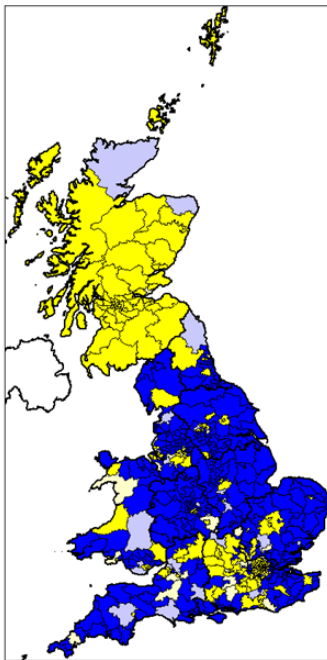


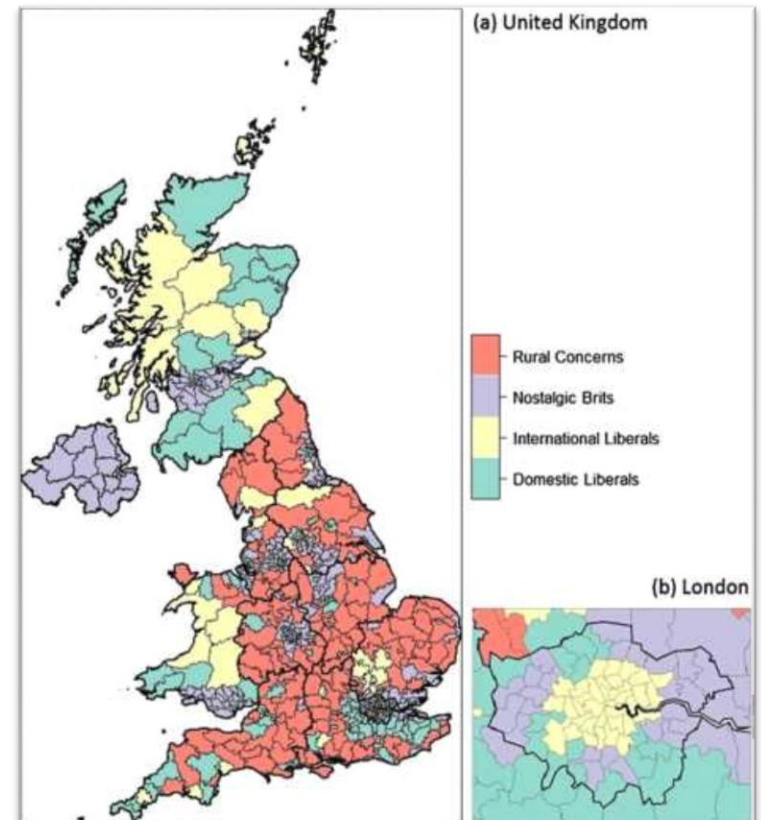
FIGURE 2 Estimated rent/price ratio for two-bedroomed flats for a sample of English postcodes.

# Example 2: E-Petition Data

Estimating the outcome of UKs referendum on EU membership using e-petition data and machine learning algorithms



Classification of Westminster Parliamentary constituencies using e-petition data



- On 23 June 2016, 52% voted in favour of leaving the EU (turnout 72% of registered voters)
- Results published for ‘Counting Areas’
- But not for Westminster Parliamentary Constituencies (WPCs)
- WPCs are geography that elected members of Parliament are held to account by their constituents.

<b>Referendum on the United Kingdom's membership of the European Union</b>	
<b>Vote only once</b> by putting a cross <input checked="" type="checkbox"/> in the box next to your choice	
Should the United Kingdom remain a member of the European Union or leave the European Union?	
<b>Remain a member of the European Union</b>	<input type="checkbox"/>
<b>Leave the European Union</b>	<input type="checkbox"/>



*“for the purpose of examining dyadic representation ... results at the level of Westminster parliamentary constituencies would be far more useful than results from local authority areas.” (Hanretty 2017, p. 466)*

Our study uses e-petition data and machine learning algorithms to estimate the Leave vote percentage for Westminster Parliamentary Constituencies.

# e-petitions (X data)

- Hosted by UK Parliament
- Create or sign a petition that asks for a change to the law or to government policy.
- Use e-petitions between May 2015 to April 2016 (25 petitions)
- JSON files of raw counts in WPCs
- Size of WPC electorate varies from 22k to 110k
- Normalise by dividing by the size of the 2015 electorate



## Petitions

UK Government and Parliament

[Other lists of petitions](#)

### Petitions debated in Parliament



56 petitions

[Authorise open book examinations for GCSE English Literature 2017](#)

110,876 signatures

Debated 24 April 2017

[End the badger cull instead of expanding to new areas](#)

108,320 signatures

Debated 27 March 2017

[Put a max of £1200 on car insurance for 18-25 year olds](#)

185,175 signatures

Debated 20 March 2017

[April's Law](#)

127,296 signatures

Debated 13 March 2017

[Make it illegal for a company to require women to wear high heels at work](#)

152,420 signatures

Debated 6 March 2017

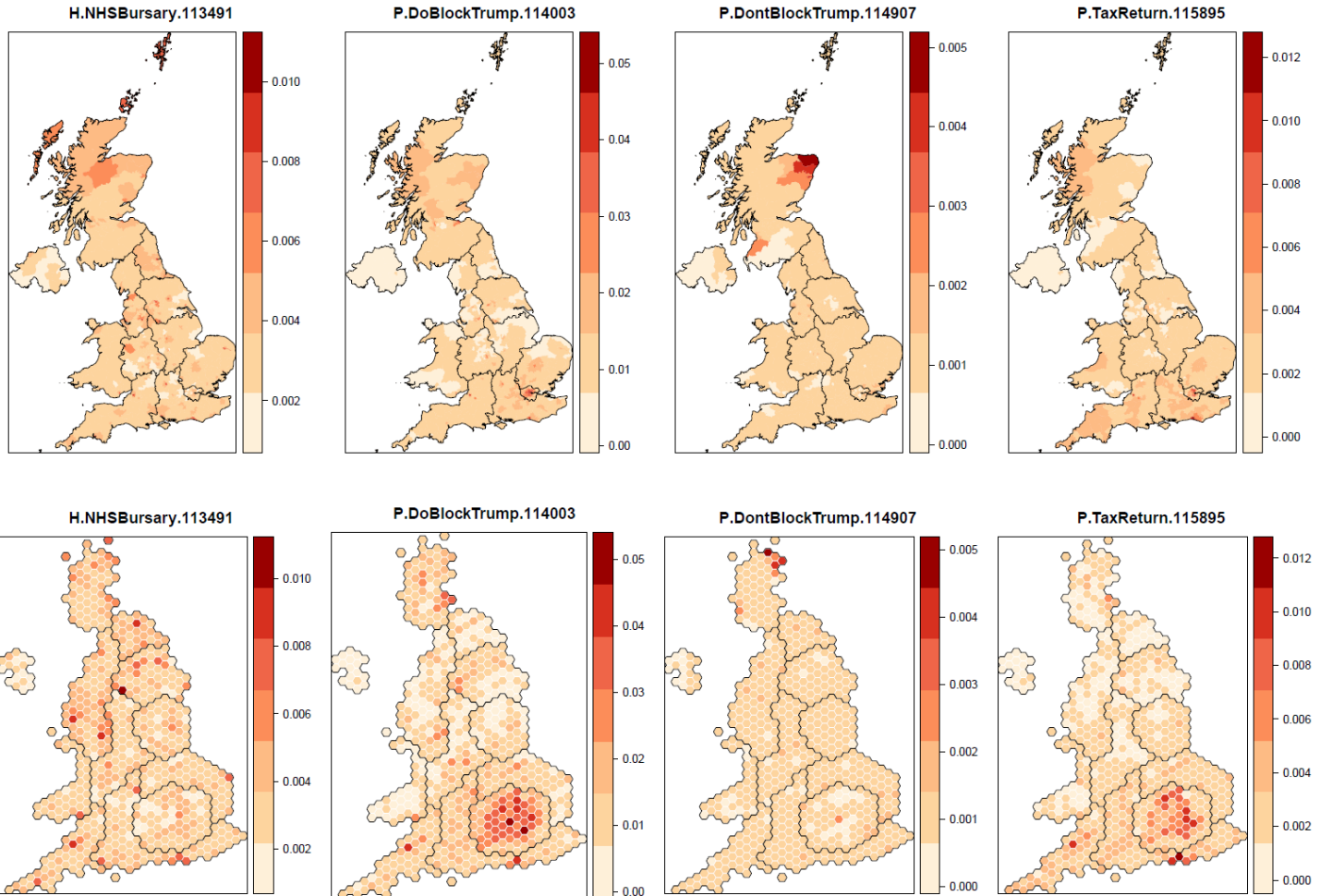
# e-petitions used

**Table 1.** List of e-petitions used in this study.

Petition id	Topic	Signatories	Open	Closed
104334	To debate a vote of no confidence in Health Secretary the Right Hon Jeremy Hunt.	231,136	20/07/2015	20/01/2016
104349	Make the production, sale and use of cannabis legal.	236,995	21/07/2015	21/01/2016
104796	Don't kill our bees! Immediately halt the use of Neonicotinoids on crops.	99,909	24/07/2015	24/01/2016
105560	Fund more research into brain tumours, the biggest cancer killer of under-40s.	120,129	03/08/2015	04/02/2016
105991	Accept more asylum seekers and increase support for refugee migrants in the UK.	450,287	13/08/2015	14/02/2016
106133	Make an allowance for up to 2 weeks term time leave from school for holiday.	127,199	14/08/2015	15/02/2016
106477	Stop allowing immigrants into the UK.	216,949	25/08/2015	26/02/2016
106651	Introduce a tax on sugary drinks in the UK to improve our children's health.	155,516	26/08/2015	27/02/2016
108072	Give the Meningitis B vaccine to ALL children, not just new born babies.	823,348	14/09/2015	15/03/2016
108570	Free Sergeant Alexander Blackman.	34,440	16/09/2015	17/03/2016
108782	The DDRB's proposals to change Junior Doctor's contracts CANNOT go ahead.	110,065	22/09/2015	23/03/2016
108944	Save British Steel making. Scunthorpe, Teesside, Port Talbot etc.	18,429	24/09/2015	25/03/2016
109383	Stop the scathing cuts to the Police budget.	9,947	05/10/2015	06/04/2016
109649	Prevent the scrapping of the maintenance grant.	133,069	02/10/2015	03/04/2016
109702	Restrict the use of fireworks to reduce stress and fear in animals and pets.	104,038	02/10/2015	03/04/2016
110776	Make fair transitional state pension arrangements for 1950's women.	193,186	20/10/2015	21/04/2016
111731	Include expressive arts subjects in the Ebacc.	102,499	09/11/2015	10/05/2016
112342	Stop the destructive 'building our future' office closure programme in HMRC.	2,585	16/11/2015	17/05/2016
113064	Vote no on military action in Syria against IS in response to the Paris attacks.	227,745	20/11/2015	21/05/2016
113231	No UK airstrikes on Syria.	190,223	22/11/2015	23/05/2016
113491	Keep the NHS Bursary.	162,568	24/11/2015	25/05/2016
114003	Block Donald J Trump from UK entry.	586,930	08/12/2015	09/06/2016
114907	Don't ban Trump from the United Kingdom.	46,622	09/12/2015	10/06/2016
115895	Scrap plans forcing self-employed & small business to do 4 tax returns yearly.	114,504	16/12/2015	17/06/2016
116762	STOP CAMERON spending British taxpayers' money on Pro-EU Referendum leaflets.	221,866	22/12/2015	23/06/2016



# e-petitions: geography



## Counting areas (Y data)

- EU votes counted for Counting Areas (CAs) (380)
  - Same as Local Authority Districts (LADs)
  - ex Orkney/Shetland
- Most political interest at Westminster Parliamentary Constituencies (WPCs) (650)
- Some CAs are co-terminus with WPCs
- Some LADs released counts for WPCs/Wards
  - Issue of allocation of postal votes to WPCs

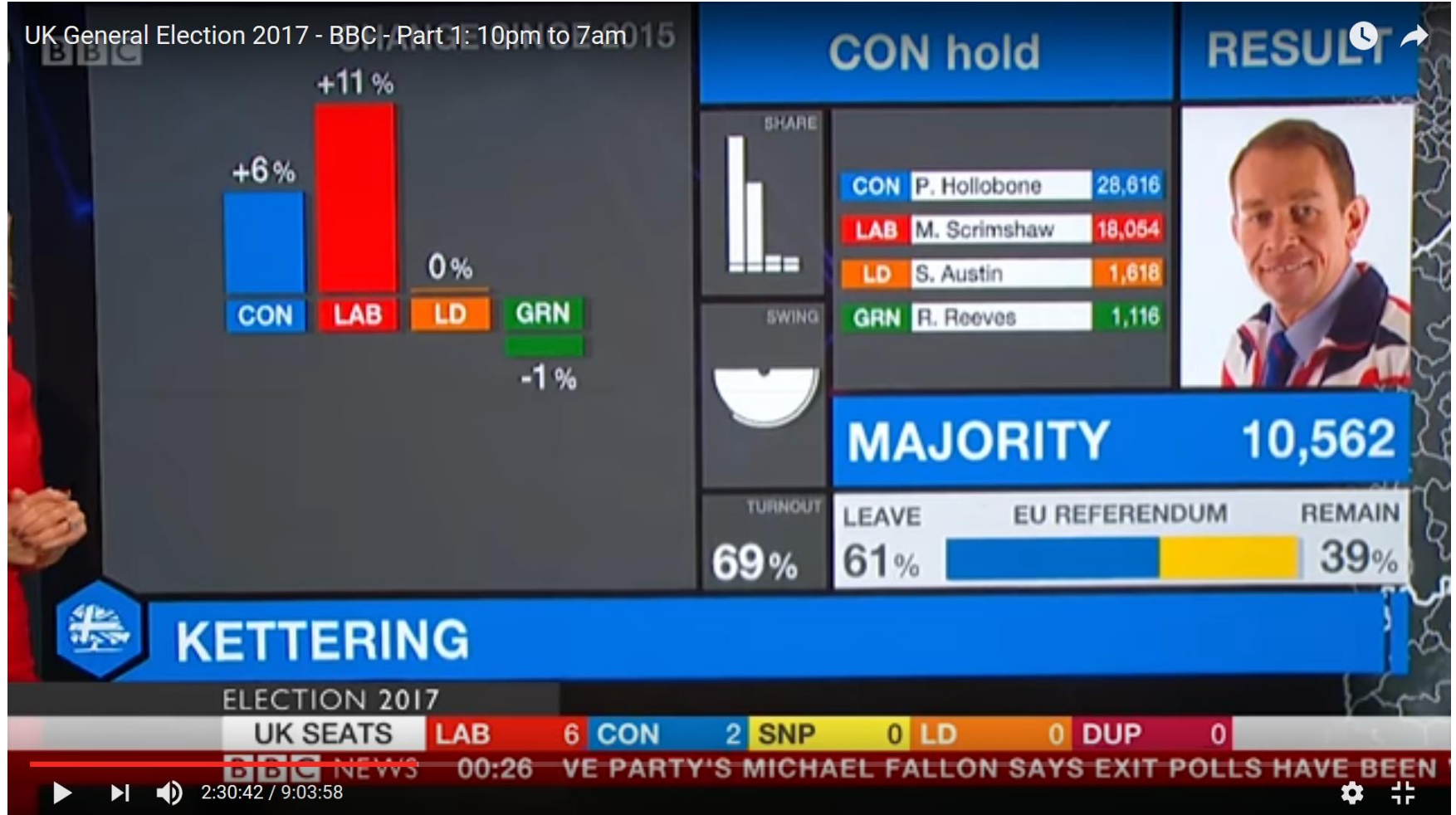
# Incompatible geographies

- Referendums results from 382 CAs
- E-petition counts from 632 WPCs (exclude NI)
- A new geography needed where aggregations of CAs are the same as aggregations of WPCs
- 173 Data Zones

Description		Number of DZ	Number of CA	Number of WPC
An aggregation of CAs same as a WPC	$\sum CA \equiv WPC$	1	2	1
CA same as a WPC	$CA \equiv WPC$	35	35	35
CA same as an aggregation of WPCs	$CA \equiv \sum WPC$	55	55	158
An aggregation of CAs same as an aggregation of WPCs	$\sum CA \equiv \sum WPC$	82	288	438
<b>Total</b>		<b>173</b>	<b>380</b>	<b>632</b>

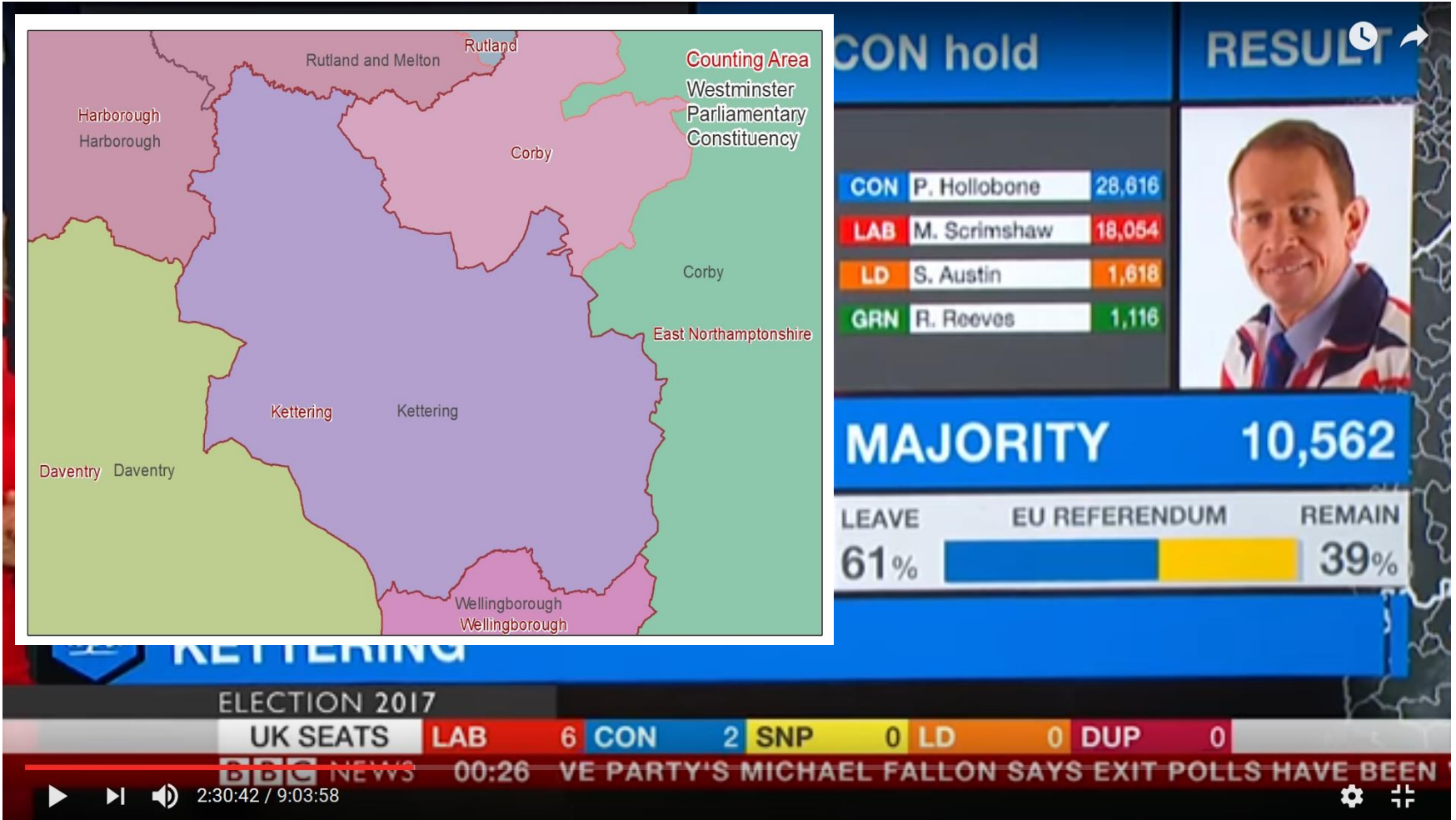
# Here one CA = one WPC

UK General Election 2017 - BBC - Part 1: 10pm to 7am





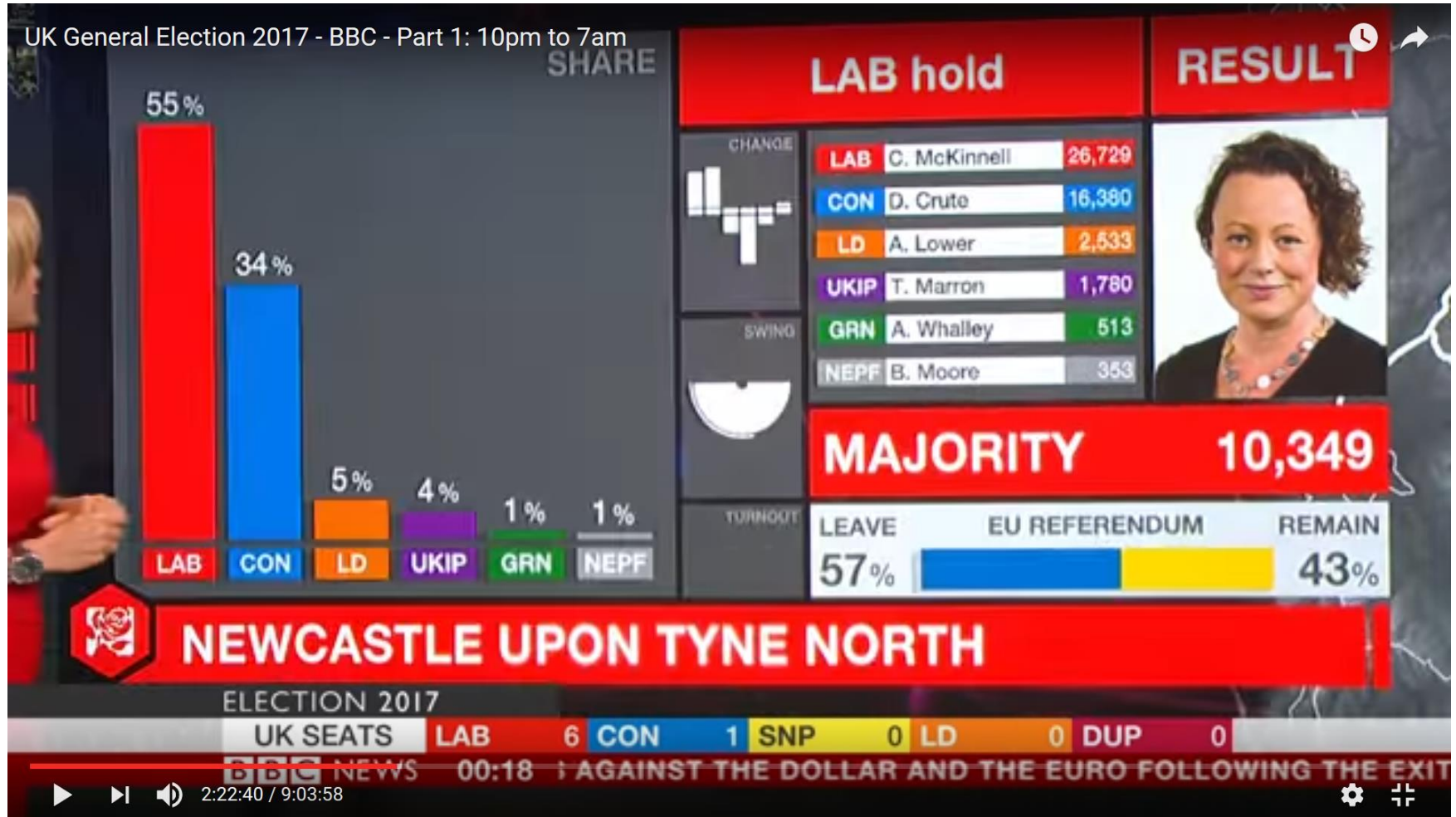
# Here one CA = one WPC



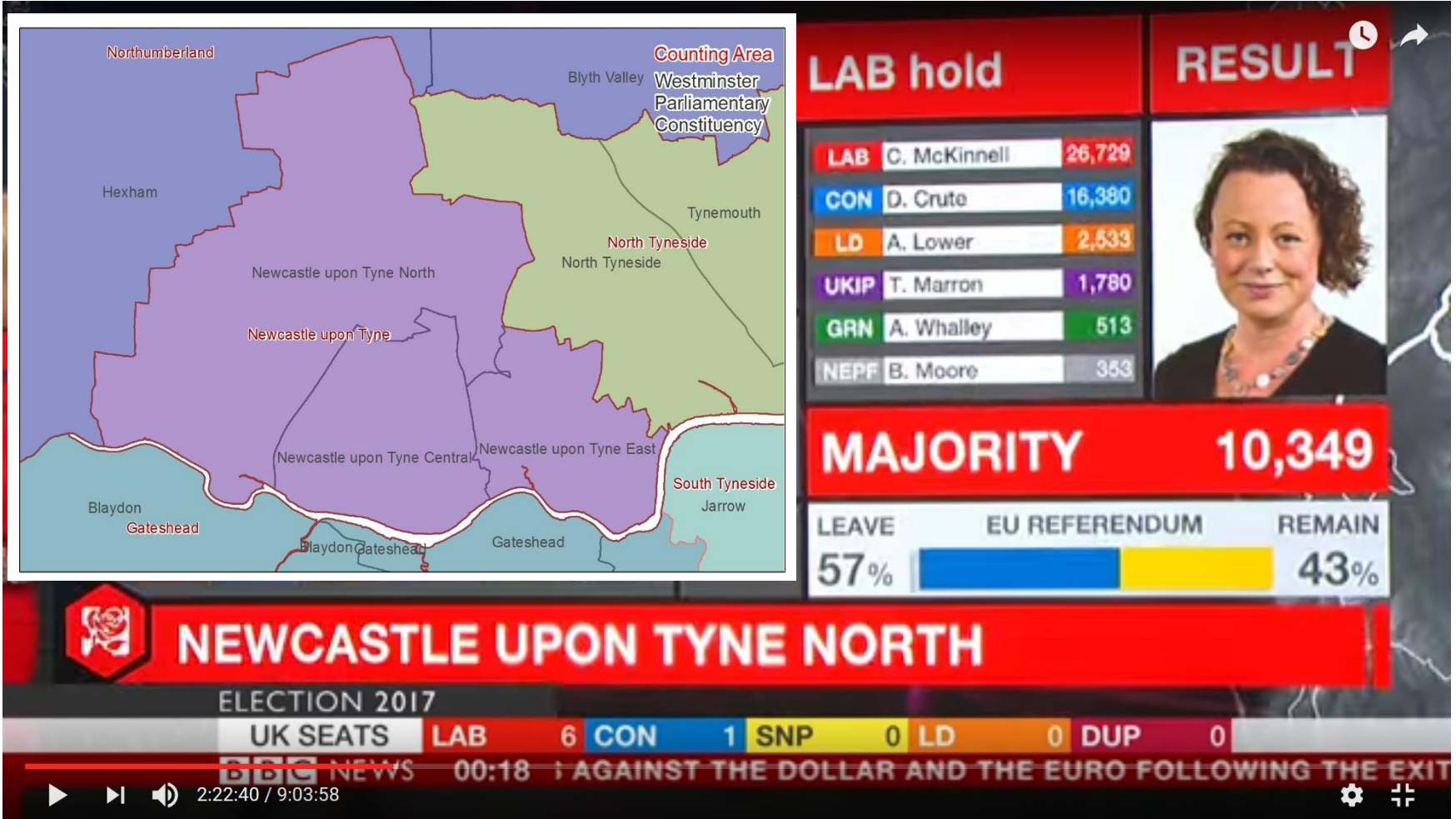


# Here one CA = three WPCs

UK General Election 2017 - BBC - Part 1: 10pm to 7am

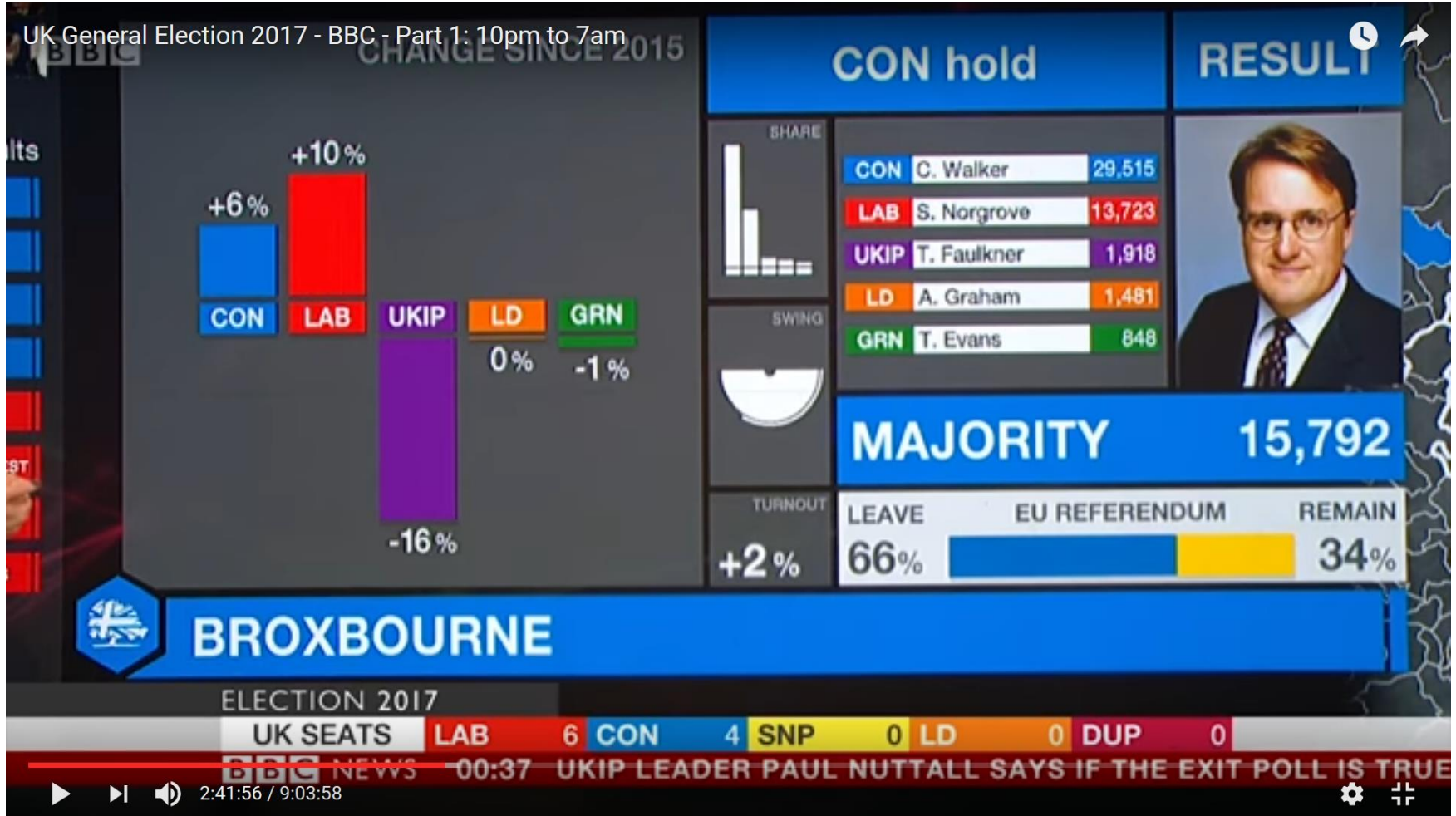


# Here one CA = three WPCs

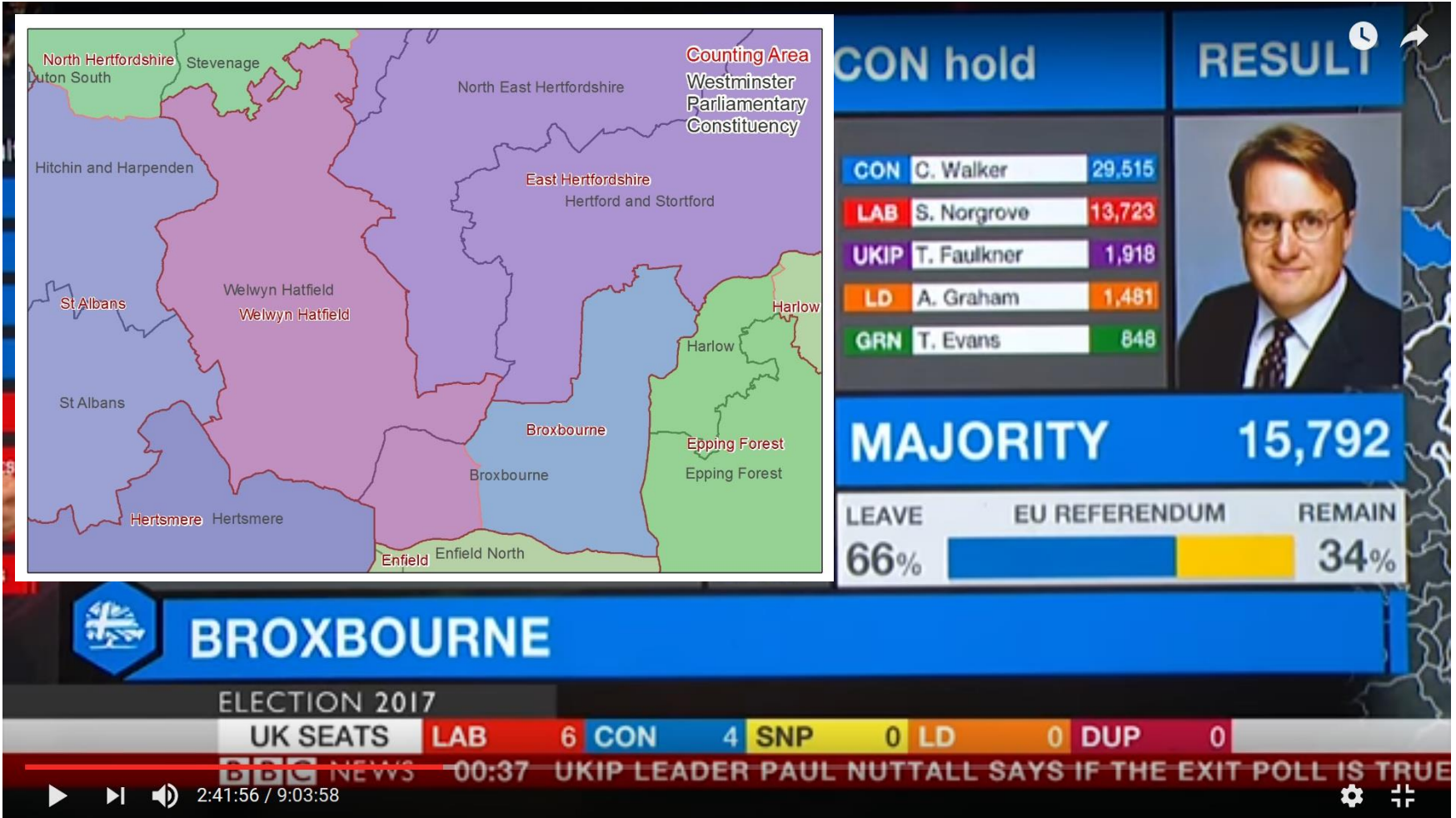


# Here two CA = two WPCs

UK General Election 2017 - BBC - Part 1: 10pm to 7am



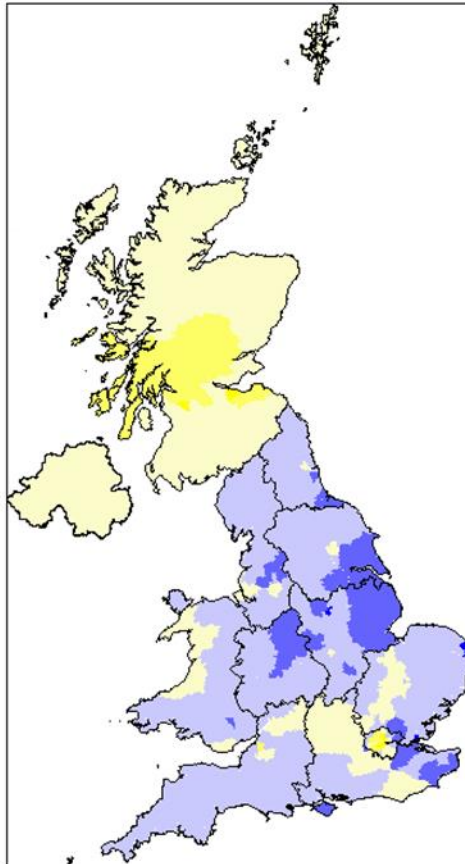
# Here two CA = two WPCs



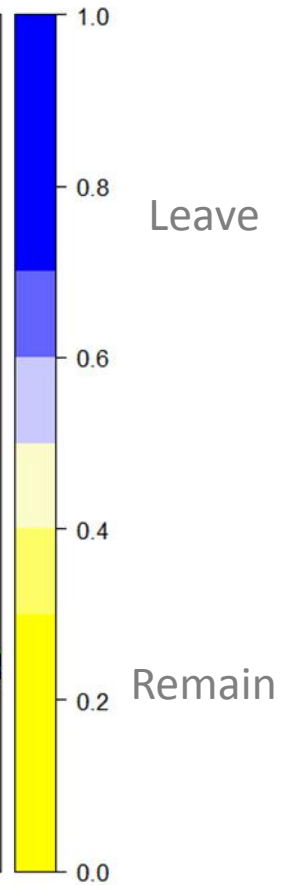
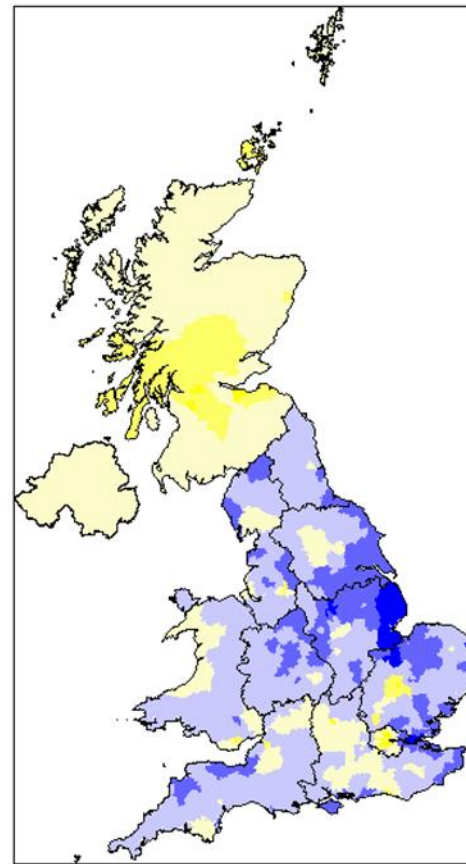


# Remapped outcomes

Outcome (DZ)



Outcome (CA)



# Machine learning algorithms

- **Lazy Learners**
  - K nearest neighbours
  - Self-organising maps
- Characterised by capturing learning through a set of similarity relationships in multidimensional ‘space’

# Machine learning algorithms

- **Divide and Conquer**
  - Random forests
  - Gradient Boost Machines
- Largely tree-based algorithms, consisting of nodes which act as routing paths leading to a leaf (with if-then conditions)

# Machine learning algorithms

- **Regression**
  - Support Vector Machines
  - Artificial Neural Networks
  - MARS (BagEarth)
- Designed to capture non-linear relationships





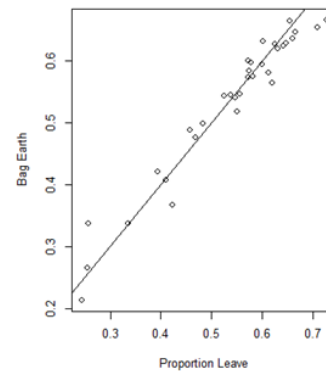
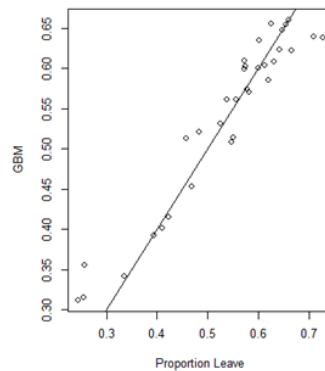
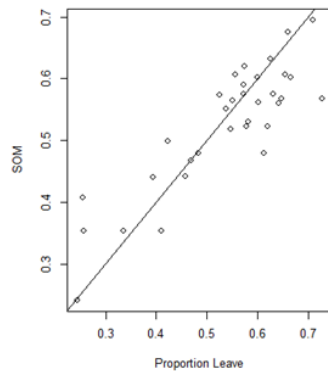
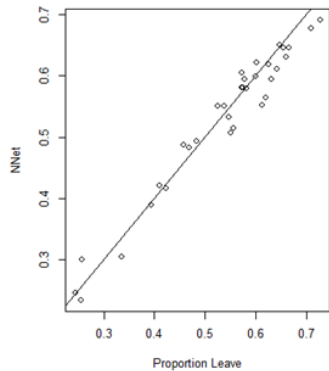
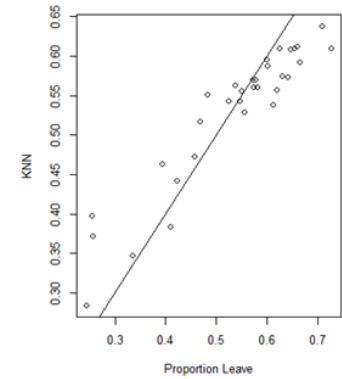
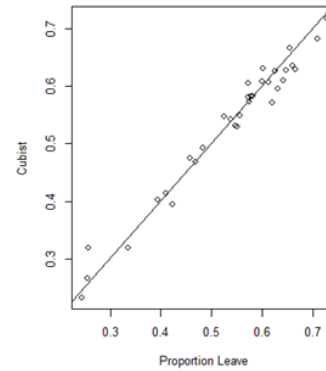
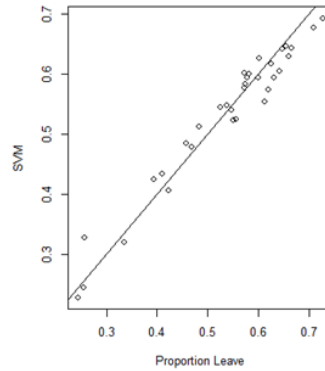
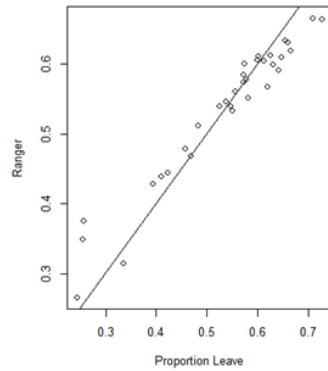
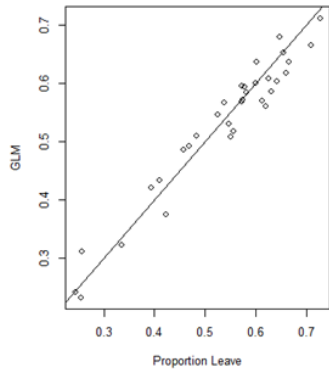
# Machine learning algorithms

- **Hybrid**
  - Cubist
    - Combination of a traditional decision tree and regression equations
    - At the leaf there is an estimated regression equation rather than a constant.

# Machine learning (approach)

- Use `caret` package in R to optimise parameters
- 10 fold cross-validation repeated 10 times
- Learn on Data Zone geography - aggregate up both CAs and WPCs to DZs
  - Keep 20% (33) back for out-of-sample performance
- Use best algorithm to predict on WPC geography

# Machine learning (performance)



Algorithm	RMSE	R <sup>2</sup>
Cubist	<b>0.0224</b>	<b>0.971</b>
Nnet	0.0270	0.959
SVM	0.0279	0.955
Bag Earth	0.0296	0.949
Ranger	0.0378	0.945
GLM	0.0307	0.944
GBM	0.0382	0.926
kNN	0.0547	0.885
SOM	0.0642	0.759

# Comparison against other studies

- **Hanretty (2017)** uses areal interpolation
  - Scaled Poisson regression incorporates demographic information from lower level geographies.
  - Estimated 400 WPCs voted Leave whilst 232 voted Remain
  - Demonstrates geographic distribution of signatures to a petition for a second referendum strongly associated with how constituencies voted in the actual referendum.

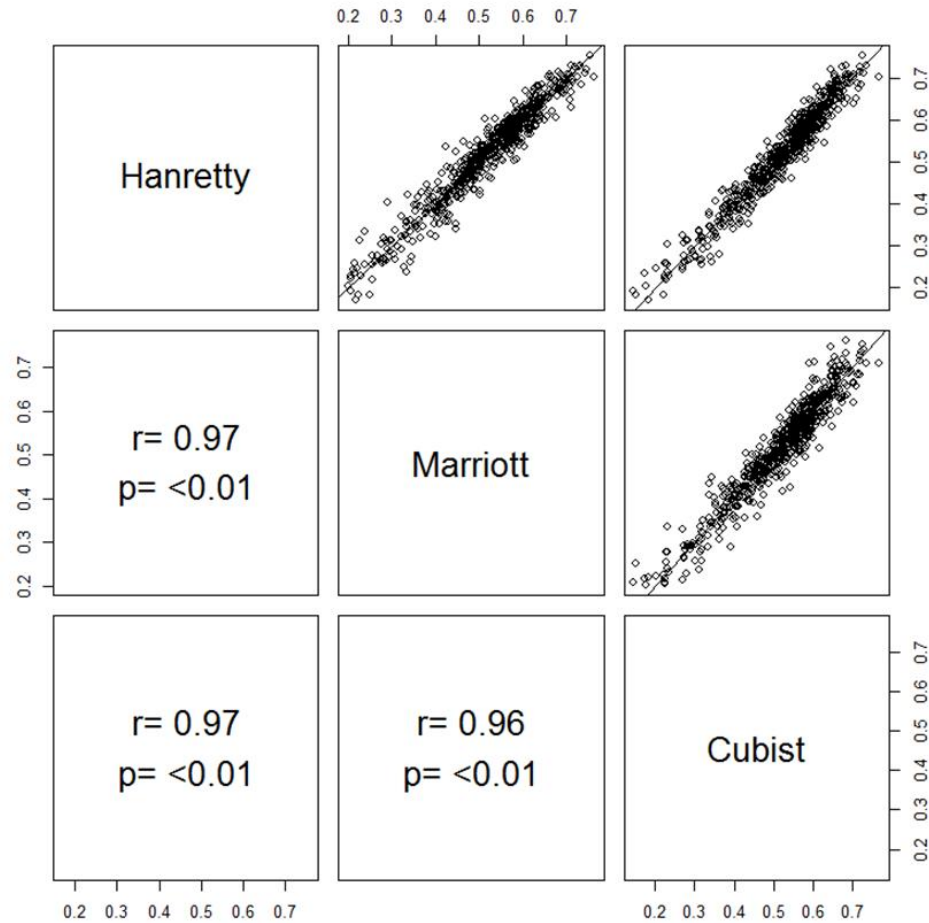
Hanretty, C. 2017. "Areal interpolation and the UK's referendum on EU membership." *Journal of Elections, Public Opinion and Parties*:1-18. doi: 10.1080/17457289.2017.1287081.

# Comparison against other studies

- **Marriott (2017)** uses a look-up table of WPCs to CAs and then a method to re-allocate votes to a WPC based on a 'classification' of each WPC.
- Estimated a Leave vote for 403 WPCs (later updated to 400)

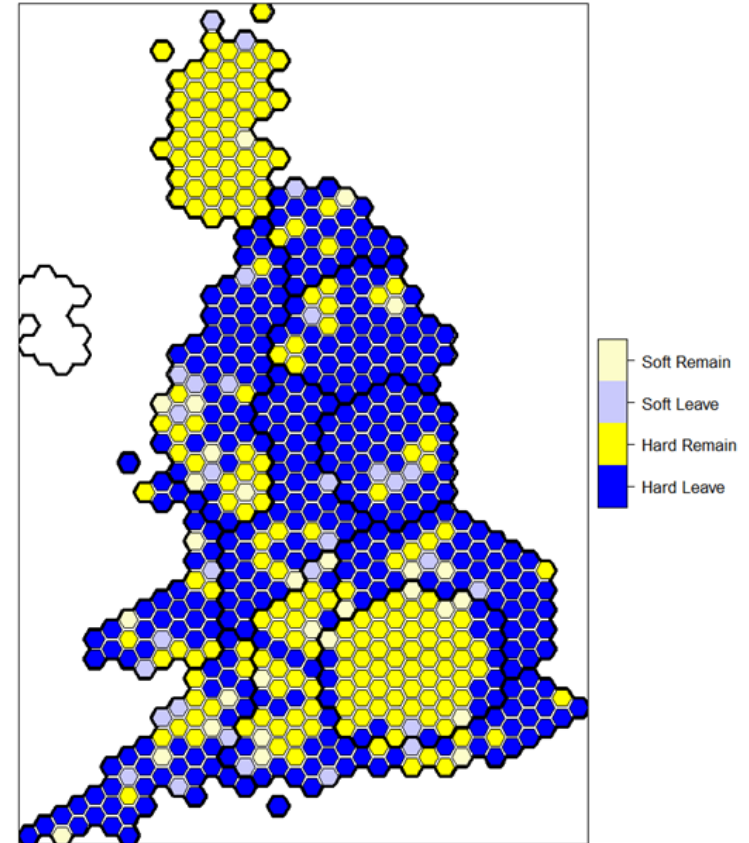
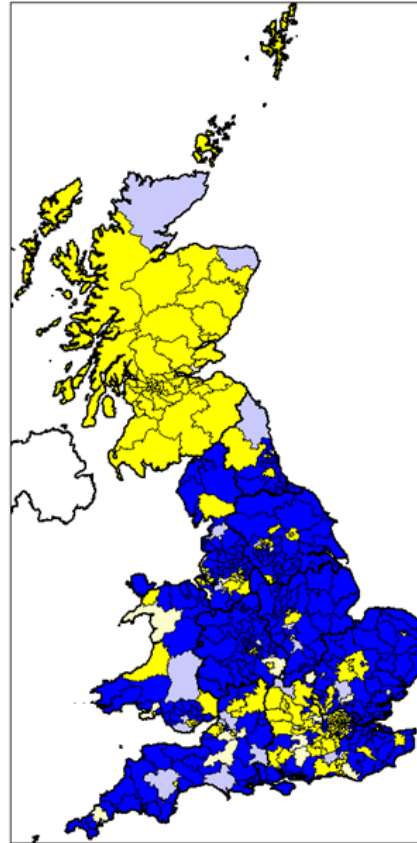
Marriott, J. 2017 "EU Referendum 2016 #1 – How and why did Leave win and what does it mean for UK politics? (a 4-part special)." <https://marriott-stats.com/nigels-blog/brexit-why-leave-won/>.

# Results (WPC)



# Results (BREXIT)

- Hard Remain  
= 201
- Hard Leave  
= 372
- Soft Remain  
= 29
- Soft Leave  
= 30



- WPCs are the democratic geography – MPs elected and represent their constituents
- Largely confirms Hanretty's and Marriot's estimates
- Signatories  $\neq$  Electors
- Method can be applied in different contexts
  - For example – plans to reduce the number of WPCs from 650 to 600



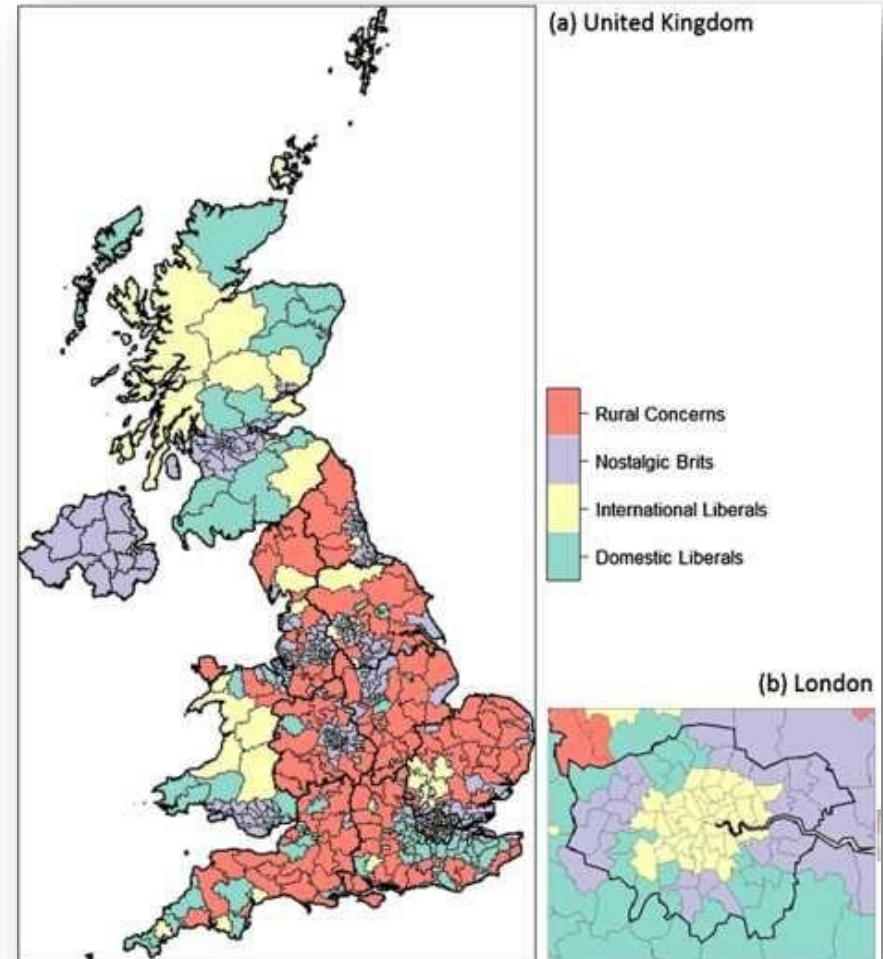
# Conclusion

- e-petition data is an informative and versatile source of information that gauges the political sentiment in a location
- This sentiment can be used to infer other outcomes
- Scope for political scientists to apply machine learning algorithms to gain confirmatory or alternative insight.

# And conclusions from the other study...

There are four distinct classes of Westminster Parliamentary Constituency

Two liberal classes are identified that are concentrated in and around London, one conservative class to be found in the urban centres and a distinct class concerned with rural issues.

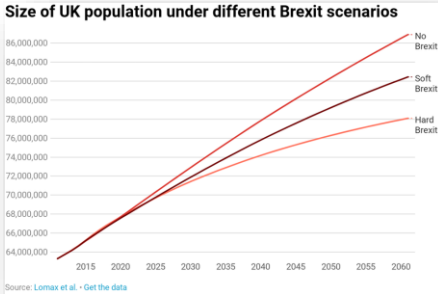




# Final Conclusions

- 'Novel' data is out there
- It is useful and applicable to academic research
- We should be doing interesting things with it
- Don't get hung up on 'big data'!
- Novel data often has a spatial dimension...
- ... which people can relate to

# Links and reading



Clark et al. EPJ Data Science (2017) 6:16  
DOI 10.1186/s13067-017-0113-9  
EPJ.org  
REGULAR ARTICLE  
Open Access

## Classification of Westminster Parliamentary constituencies using e-petition data

Stephen Clark<sup>1</sup>, Nik Lomax<sup>2</sup> and Michelle A Morris<sup>3</sup>

### Abstract

In a representative democracy it is important that politicians have knowledge of the desires, aspirations and concerns of their constituents. Opportunities to gauge these opinions are however limited and, in the era of novel data, thoughts turn to what alternative, secondary, data sources may be available to keep politicians informed about local concerns. One such source of data are signatures to electronic petitions (e-petitions). Such e-petitions have risen greatly in popularity over the past decade and allow members of the public to initiate and sign an e-petition online, with popular e-petitions resulting in media attention, a response from the government or ultimately a debate in parliament. These data are thus novel in their availability and have not yet been widely used for research purposes. In this article we will use the e-petition data to show how semantic classes of Westminster Parliamentary constituencies, fitted as Gaussian finite mixture models via EM algorithm, can be used to typify constituencies. We identify four classes: Domestic Liberals; International Liberals; Nostalgic Bits and Rural Concerns; and illustrate how they map onto electoral results. The findings and the utility of this approach to incorporate new e-petitions and adapt to changes in electoral geography are discussed.

**Keywords:** United Kingdom; Parliamentary Constituencies; Classification; Gaussian finite mixture models; electronic petitions

### 1 Introduction

Knowledge of an area's characteristics is important in gaining an understanding of the needs of those who live in, work in or service the area. Whilst each area is unique, some areas will be very similar to others and some will be distinct. The classification or geodemographic segmentation of areas allows for these areas that are similar in nature to be grouped together as identifiable classes. These classes are usually established by using multi-variate data to characterise an area and then grouping together areas whose characteristics are broadly similar (Everitt et al. [1]). Given the nature of these data, there is the potential for these classes to be dispersed over space, with neighbouring areas belonging to different classes (Berry and Linnér [2]).

Classification can be applied at any level of geographic scale, from small neighbourhoods (Office for National Statistics [3]; Gale et al. [4]) through to municipalities (Office for National Statistics [5]). They can also be designed for general use or bespoke for a particular

JOURNAL OF Big Data  
2016, VOL. 1, NO. 4, 344-357  
http://dx.doi.org/10.1186/s13067-016-0493-6



OPEN ACCESS

## Estimating the outcome of UKs referendum on EU membership using e-petition data and machine learning algorithms

Stephen D Clark<sup>1</sup>, Michelle A Morris<sup>2</sup>, and Nik Lomax<sup>3</sup>

Leeds Institute for Data Analytics, School of Medicine, University of Leeds, LEEDS, UK

**ABSTRACT**  
The United Kingdom's 2016 referendum on membership of the European Union is perhaps one of the most important recent electoral events in the UK. This political sentiment has confounded politicians, media commentators and academic alike, and has challenged elected Members of the Westminster Parliament. Unfortunately, for many areas of the UK this referendum outcome is not known for Westminster Parliamentary Constituencies, rather it is known for the coarse geography of counting areas. This study uses novel data and machine learning algorithms to estimate the Leave vote percentage for these constituencies. The results are seen to correlate well with other estimates.

**ARTICLE HISTORY**  
Received 15 September 2017  
Revised 20 April 2018  
Accepted 15 June 2018

**KEYWORDS**  
EU referendum; e-petitions; estimation; machine learning

### E-petitions and political activism

In a representative democracy it is important that politicians have knowledge of the desires, aspirations and concerns of their constituents. This can be accomplished by them attending meetings; conducting advice surgeries and simply talking with people. The question then arises as to how representative these interactions are? Ideally the conduct of a statistically sound opinion survey would provide a more objective measure of local views. However, in the era of big data thoughts turn to how alternative data can inform politicians about important issues (Bright & Margets, 2016; Karpf, 2016b). One such source of data are signatures to electronic petitions (e-petitions) (Karpf, 2016a). In this study use is made of such e-petition data to provide local politicians with information about possibly the most significant event in recent United Kingdom (UK) politics, the referendum on its European Union (EU) membership.

### The impact of the United Kingdom's European Union membership referendum

On June 23rd 2016 the UK held a referendum on its membership of the EU. The question put to the eligible voters was:

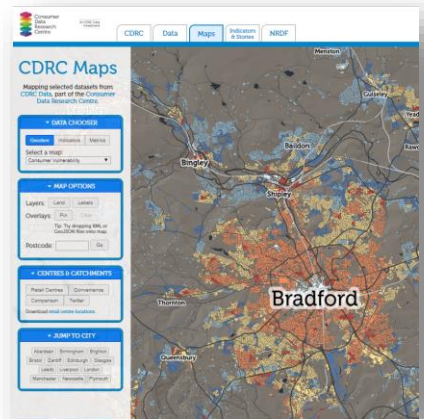
"Should the United Kingdom remain a member of the European Union or leave the European Union?"

with possible answers of Remain or Leave. The vote counts and declarations took place in each of 582 counting areas<sup>1</sup> (CA) and these results were aggregated at the national level to provide a majority vote of 52% to leave the EU.

Both prior to the referendum and afterwards it became apparent how the outcome would influence many important aspects of life (Herburn, 2017), including demography (Coleman, 2016), immigration (Portes & Forte, 2017), financial markets (Yosh, 2016), education (Mayhew, 2017), health (McKenna, 2016) and agriculture (Hein, 2017). Much analysis took place to understand what socio-economic or socio-class factors could explain the result (Becham, Slingsby,

**CONTACT** Stephen D Clark [s.d.clark@leeds.ac.uk](mailto:s.d.clark@leeds.ac.uk)  
Stephen Clark Dr Clark has worked in the public, private and academic sectors, in areas as diverse as software engineering, computer aided design, transport planning and health. His current research interests include the spatial modelling of transportation, public health planning and understanding electoral outcomes. His former role at Leeds is an Academic Fellow in Spatial Data Analytics at the University of Leeds. His research interests include the estimation and prediction of populations and their attributes, which includes election, health, mobility and health.  
**Michelle A Morris** Dr Morris is an Academic Fellow in Health Data Analytics at the University of Leeds. Her research interests include how new forms of data and analytical methods can be used to inform health and policy.  
**Nik Lomax** Nik Lomax is an Academic Fellow in Health Data Analytics at the University of Leeds. His research interests include how new forms of data and analytical methods can be used to inform health and policy.  
© 2016 The Author(s). Published by Taylor & Francis.  
This is an Open Access article distributed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.  
© 2016 The Author(s). Published by Taylor & Francis.  
This is an Open Access article distributed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

Link to *CDRC Maps*  
<https://maps.cdrc.ac.uk>



Clark and Lomax / Big Data (2018) 5:1-4  
DOI: 10.1186/s13067-018-0154-3



CASE STUDY  
Open Access

## A mass-market appraisal of the English housing rental market using a diverse range of modelling techniques

Stephen D Clark<sup>1</sup> and Nik Lomax<sup>2</sup>

**ABSTRACT**  
**Introduction:** Mass appraisals in the rental housing market are far less common than those in the sales market. However, there is evidence for substantial growth in the rental market and this lack of insight hampers commercial organisations and local and national governments in understanding this market.  
**Case description:** This case study uses data that are supplied from a property listings website and are unique in their scale, with over 1.2 million rental property listings available over a 2 year period. The data is analysed in a large data institute using generalised linear regression, machine learning and a pseudo-practitioner based approach.  
**Discussion and evaluation:** The study should be seen as a practical guide for property professionals and academics wishing to undertake such appraisals and looking for guidance on the best methods to use. It also provides insight into the property characteristics which most influence rental listing price.  
**Conclusions:** From the regression analysis, attributes that increase the rental listing price are: the number of rooms in the property, proximity to central London and to railway stations, being located in more affluent neighbourhoods and being close to local amenities and better performing schools. Of the machine learning algorithms used, the two tree based approaches were seen to outperform the regression based approaches. In terms of a simple measure of the median appraisal error, a practitioner based approach is seen to outperform the modelling approaches. A practical finding is that the application of sophisticated machine learning algorithms to big data is still a challenge for modern desktop PCs.  
**Keywords:** Housing; Rental; Regression; Machine learning; Big-data; Commercial

### Introduction

This study is concerned with the operation of a mass market appraisal within the English housing private rental market [1] using a source of novel big data. Mass market appraisal is the ability to make an assessment of the potential rental value that a property can be let at, using an automated approach with little or no intervention by rental professionals, such as estate agents or letting agents [2]. The advantages of such approaches are that they are able to crunch through large volumes of information to provide these appraisals; they are based on an understanding of the current state of the market through the accumulation of information captured by novel data; and they can provide some insight into

Accepted: 3 April 2018  
DOI: 10.1186/s13067-018-0154-3

REGULAR PAPER  
WILEY AREA

## Rent/price ratio for English housing sub-markets using matched sales and rental data

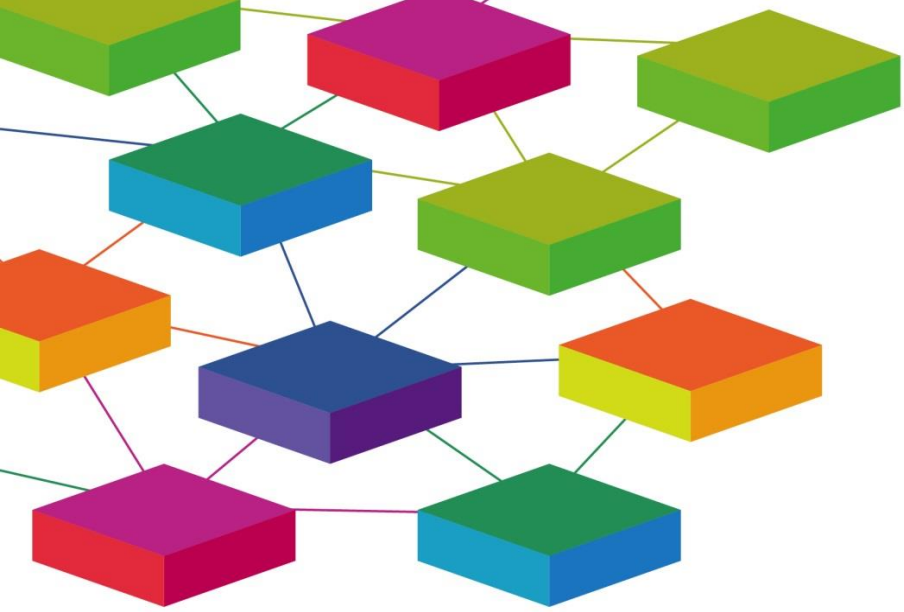
Stephen Clark<sup>1</sup> and Nik Lomax<sup>2</sup>

**ABSTRACT**  
The ratio between the rental and sales values of residential properties are a much studied statistic in the field of real estate economics. When these values do not keep pace with each other, and in particular when the ratio is low, some commentators take this as an indication that there may be a housing bubble. The ratios are also of interest to potential property investors. These ratios are commonly computed on aggregate statistics derived from the housing market and as such rarely provide any indication of sub-market bubbles, that can occur with particular property types or regions of the country. In this study use is made of a data set from a property listings company that provides sales and, potentially, rental prices for the same properties within England. From the matching that takes place it is possible to calculate the rent/price ratio for individual properties. A regression model is then estimated to explain how the characteristics of the properties, the nature of their neighbourhood, and their location influence this ratio. The model consistently validates the hypothesis that the more desirable a property or affluent an area, the lower the rent/price ratio. It also begins to illustrate the range of "normal" rent/price ratios that may exist in housing sub-markets. The regression model is then used to provide a map of the geographical distribution of the ratio for England for one property sub-market.  
**KEYWORDS**  
Rental; sales; ratios; regression; rent/price ratio; sub-markets

### 1 INTRODUCTION

Housing is one of the largest items of household expenditure for families; in the United Kingdom (UK) it is the third largest outgoing, after transport and recreation (Office for National Statistics, 2018). The 2011 Census reports that in England, 31% of households owned their property outright, 33% had a mortgage, 17% were socially rented, 17% were privately rented, and 2% were either shared ownership or rent free. Trends since 2011 however point to a reduction in the proportion of residential properties that are owned, mortgaged, or socially rented, and an increase in those that are privately rented (Lam, 2013). The latest to 2015 show that the proportion of households that are privately rented has risen to 20% (Valuation Office Agency, 2015; Wilcox et al., 2017).

These trends are driven by changes to policy around the support given to renters (Stephens & Whitehead, 2014); the availability of de-socialised housing (Copley, 2014); and the attractiveness of rental properties for landlords (Ronald



# Questions

[@niklomax](#)