Consumer Data Research Centre

An ESRC Data Investment

Using Novel Data to Provide Local Insights

Nik Lomax
University of Leeds

LIDA Seminar | Leeds | 14 Nov 2019

@niklomax

Consumer Data Research Centre

An ESRC Data Investment

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'

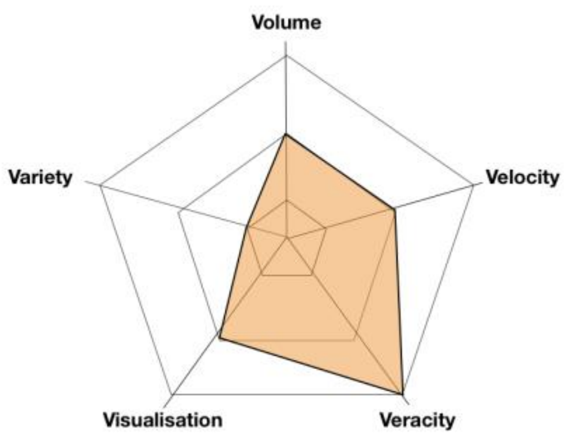
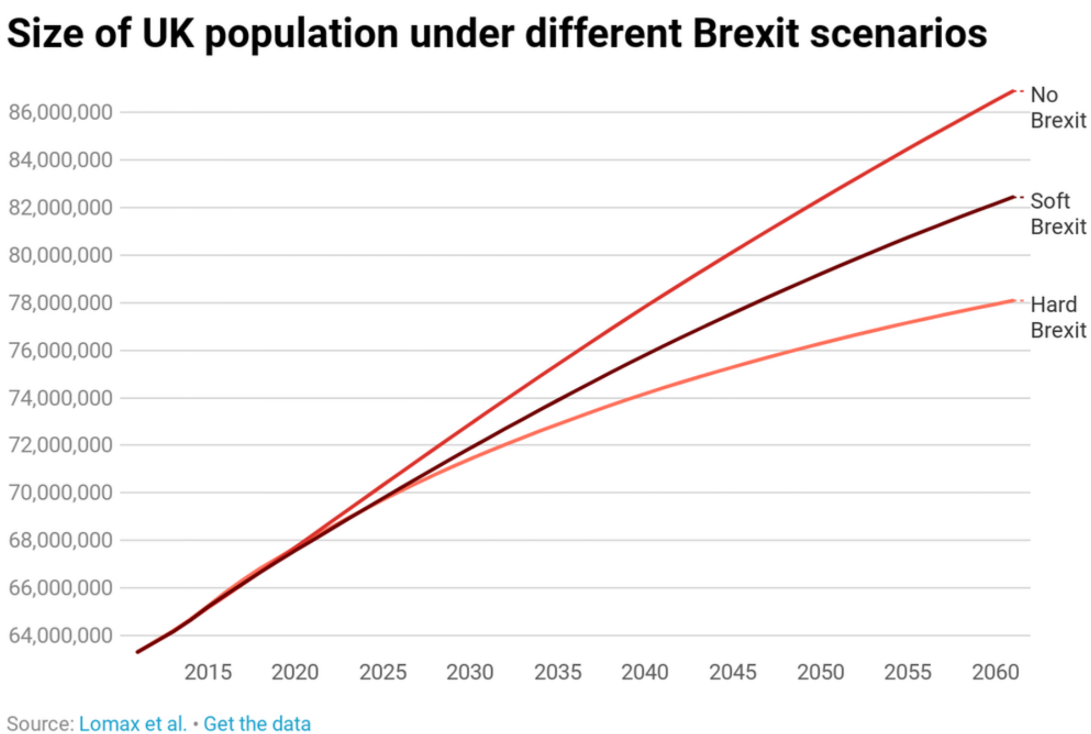


Figure 1: The footfall data from Smart Street Sensor is truly "big" only on the veracity dimension. Otherwise it is mainly a medium sized data.

Source: Soundararaj, B., Cheshire, J. and Longley, P. (2019) Medium Data Toolkit - A Case study on Smart Street Sensor Project. Presentation at GISRUK, Newcastle, 24-26 April.

Motivation

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



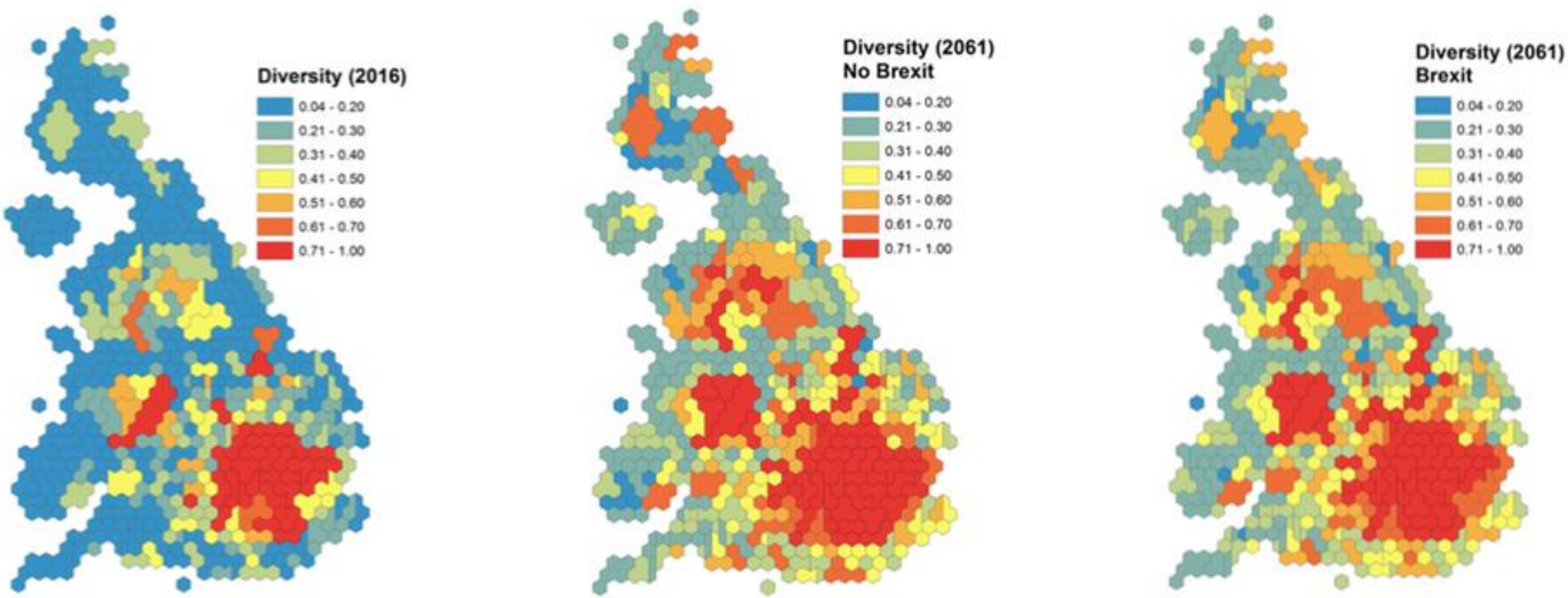
Source: Lomax (2019) What the UK population will look like by 2061 under hard, soft or no Brexit scenarios, *The Conversation*, https://bit.ly/2YUzwCT

Consumer Data Research Centre

An ESRC Data Investment

Motivation

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



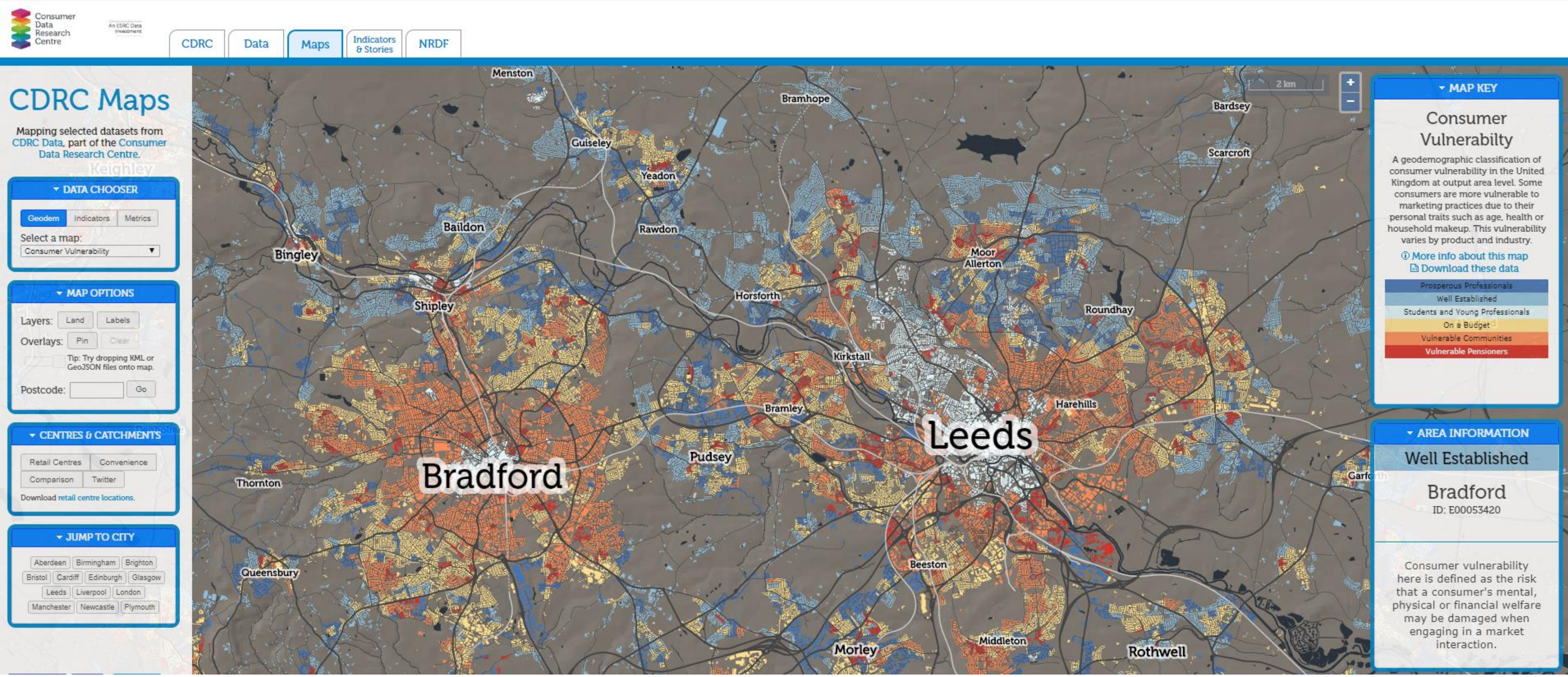
Source: Lomax (2019) What the UK population will look like by 2061 under hard, soft or no Brexit scenarios, *The Conversation*, https://bit.ly/2YUzwCT

Consumer Data Research Centre

An ESRC Data Investment

Motivation

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



Source: Adcock and Lomax (2018)
https://maps.cdrc.ac.uk/#/geodemographics/vulnerability/

Consumer Data Research Centre

An ESRC Data Investment

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**~~.

Consumer Data Research Centre

An ESRC Data Investment

Example 1: Sales and rental data

A mass market appraisal of the rental market

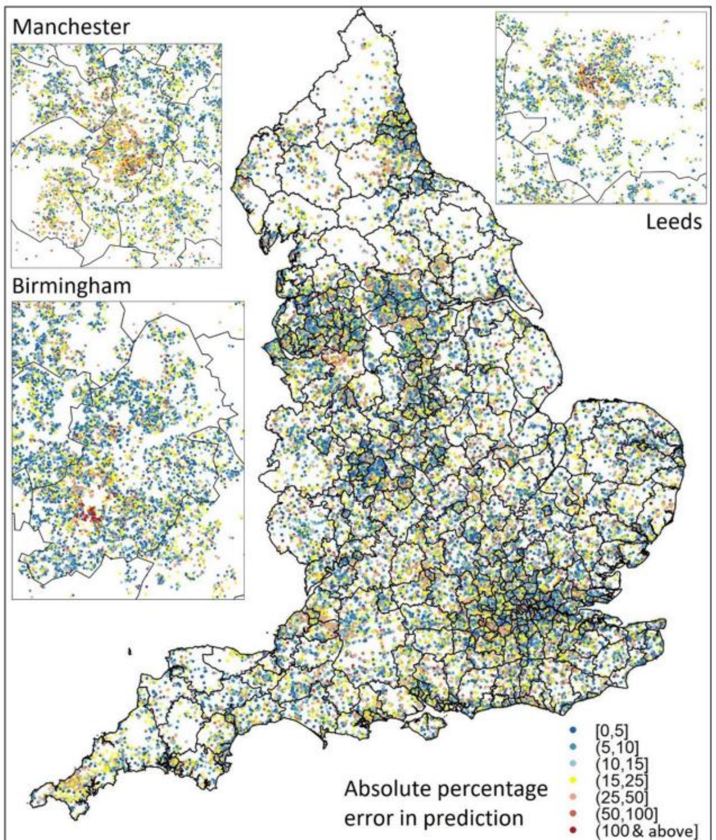


Fig. 2 Absolute percentage prediction error from cubist model

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

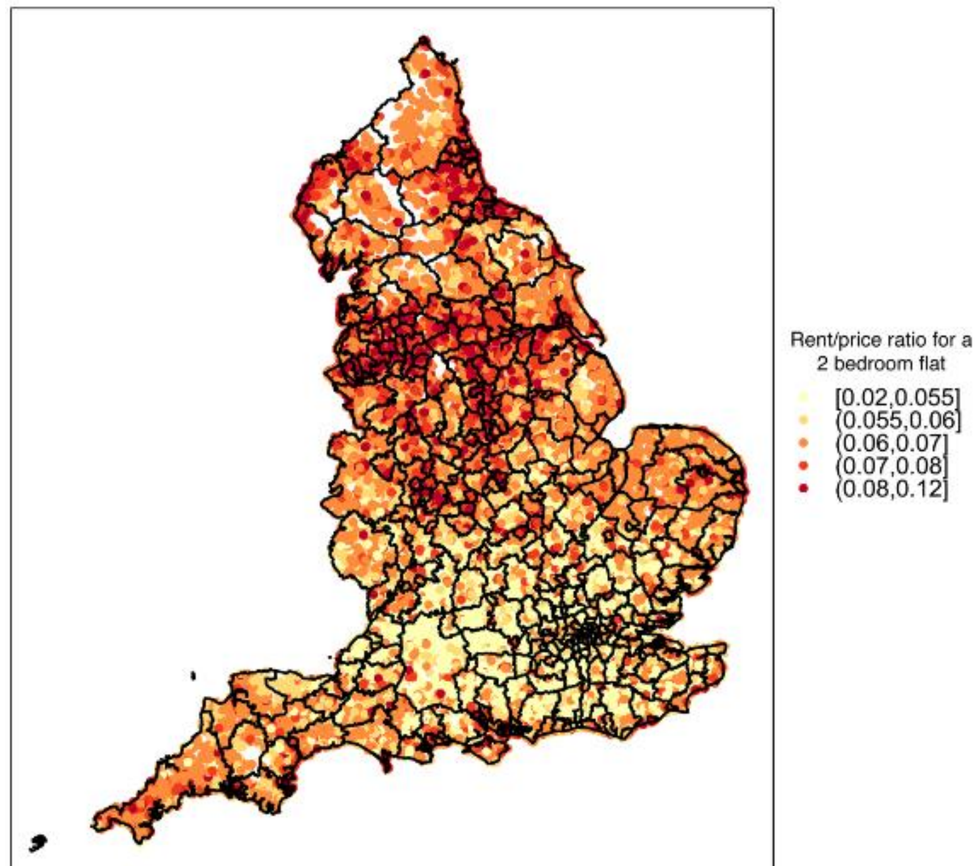
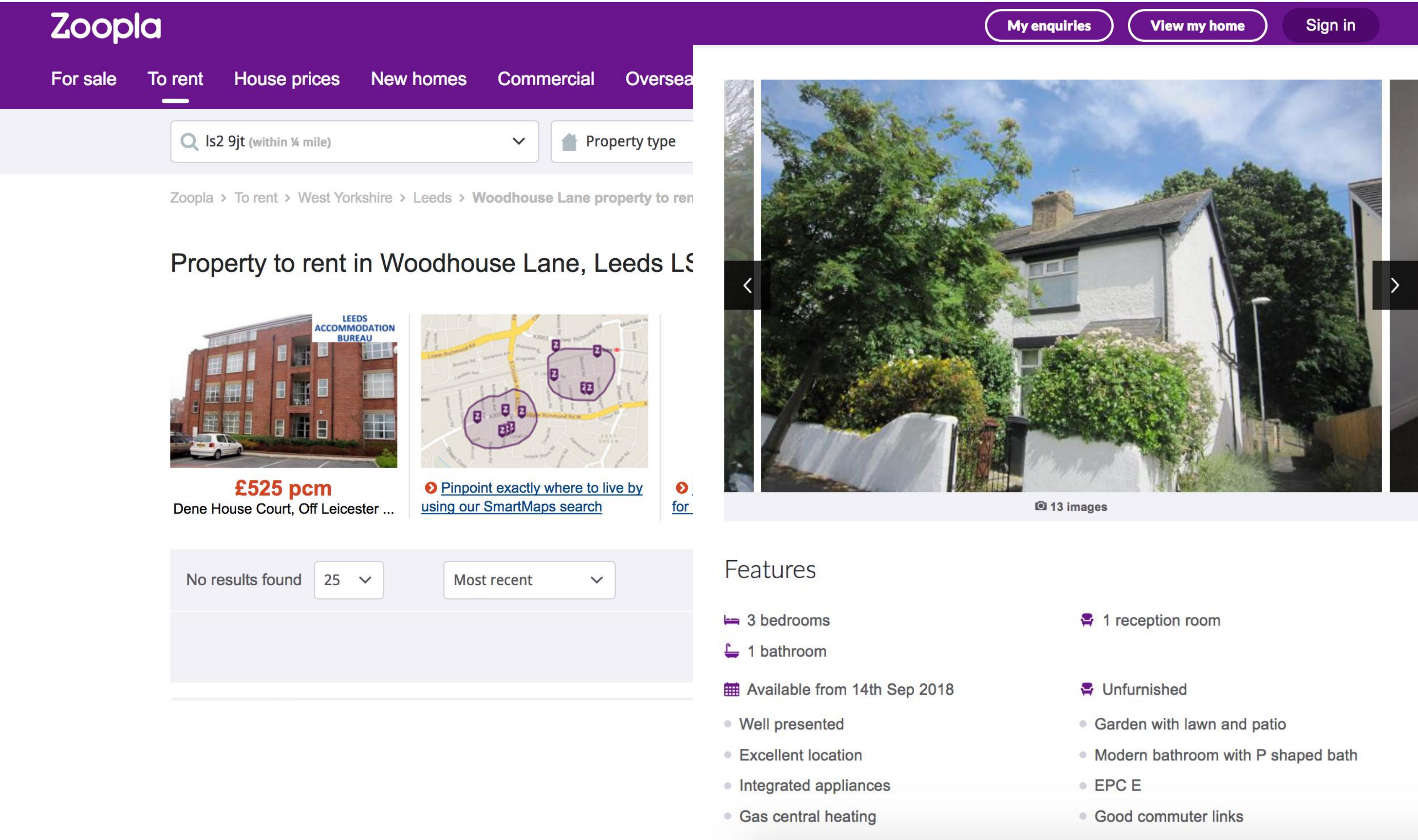


FIGURE 2 Estimated rent/price ratio for two-bedroomed flats for a sample of English postcodes.

Consumer Data Research Centre

An ESRC Data Investment

Data (are inherently spatial)



Consumer Data Research Centre

An ESRC Data Investment

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

Consumer Data Research Centre

An ESRC Data Investment

Introduction

- **Banzhaf and Farooque (2013)** rental values correlate with access to public goods and income levels in Los Angeles
- **Löchl (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

Consumer Data Research Centre

An ESRC Data Investment

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.

Consumer Data Research Centre

An ESRC Data Investment

Data

- 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 ≠ final rental price

Consumer Data Research Centre

An ESRC Data Investment

Data

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

Consumer Data Research Centre

An ESRC Data Investment

Methods

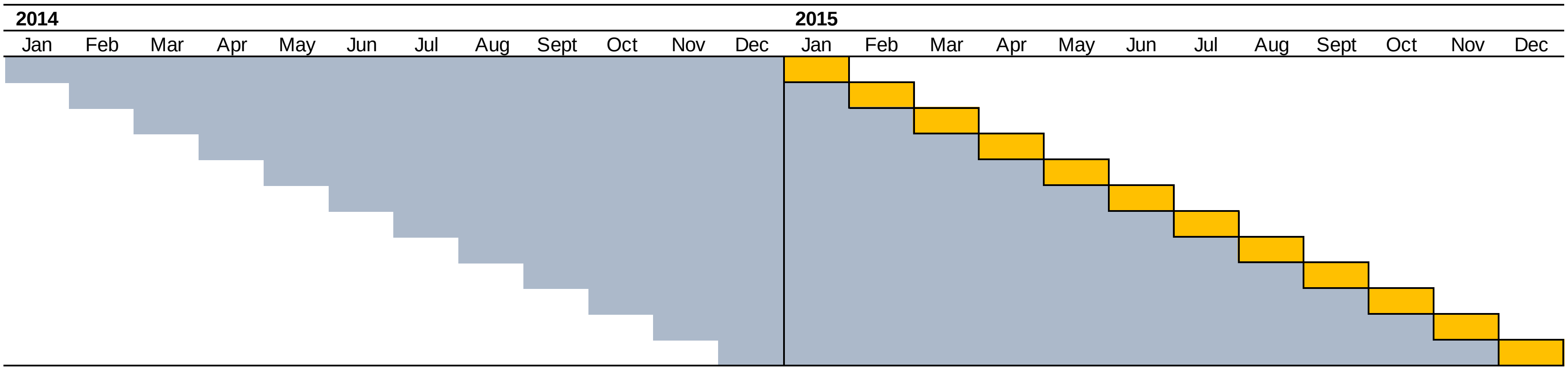
1. Quassi Poisson generalised linear model (GLM)

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

Experimental procedure

- All methods are applied in a consistent manner akin to a moving window
- Information from the previous 12 months used predict the out-of-sample rental prices



Consumer Data Research Centre

An ESRC Data Investment

GLM Results

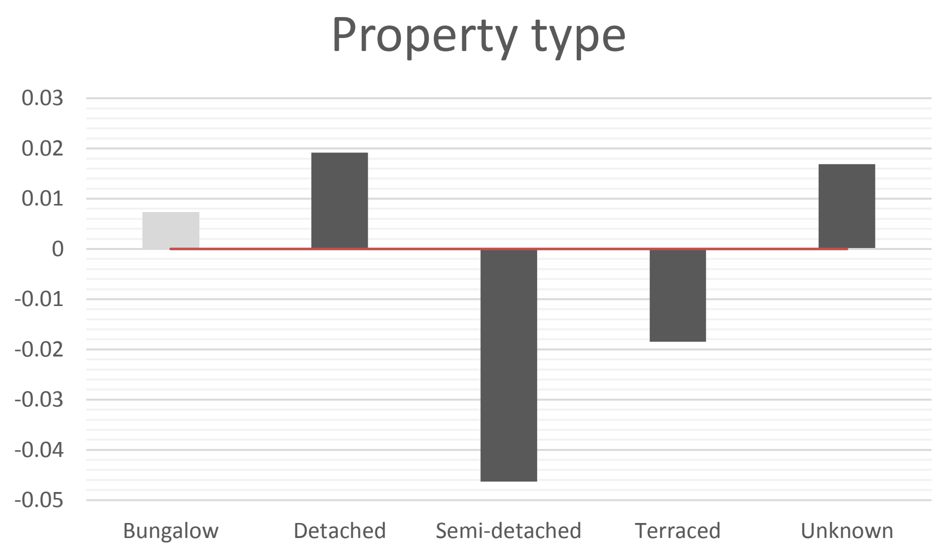
- quassi Poisson generalised linear model (GLM) used because:
    - skewed distribution of the rental price
    - possible over-dispersion
- Essential step prior to Machine Learning – Does the data capture **dynamics of the housing market in a sensible manner?**
- 63 variables
- Squared correlation between observed and in-sample predicted r2 = 0.738 on log of rental price
- r2 drops to 0.54 on original scale

Consumer Data Research Centre

An ESRC Data Investment

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	***

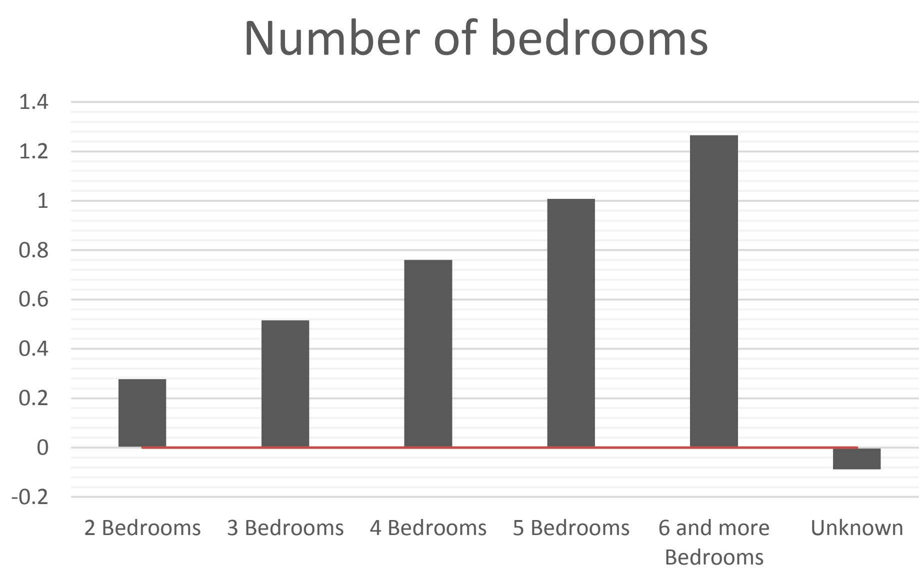


Consumer Data Research Centre

An ESRC Data Investment

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	***

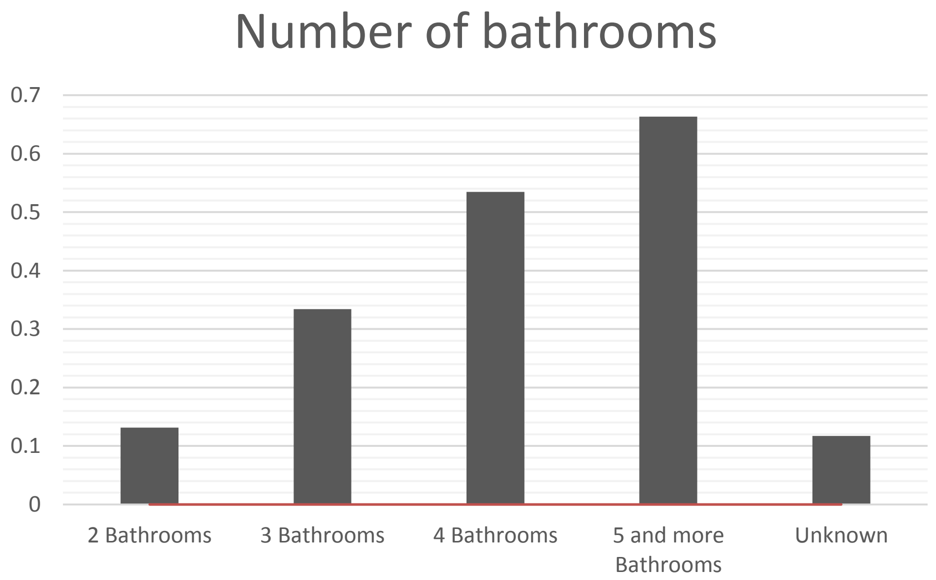


Consumer Data Research Centre

An ESRC Data Investment

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	***

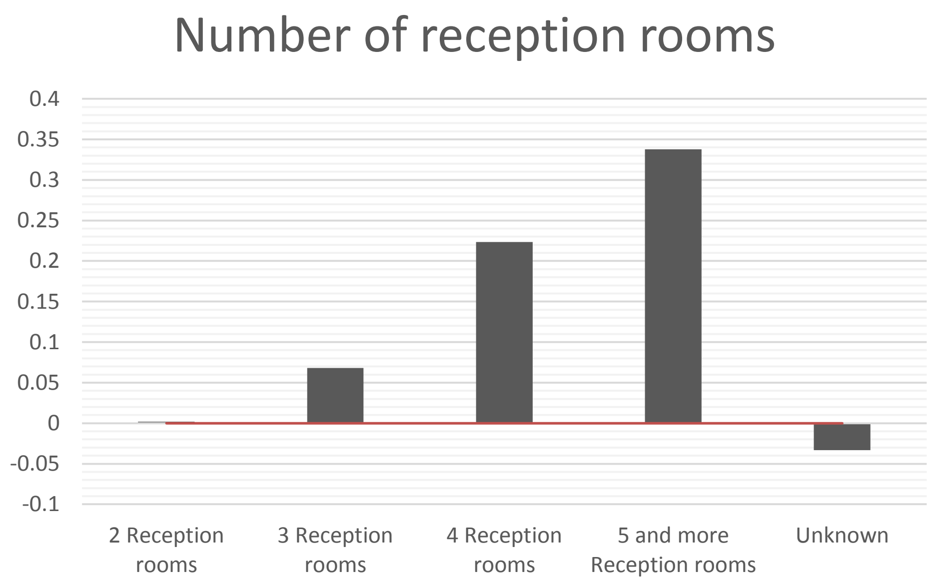


Consumer Data Research Centre

An ESRC Data Investment

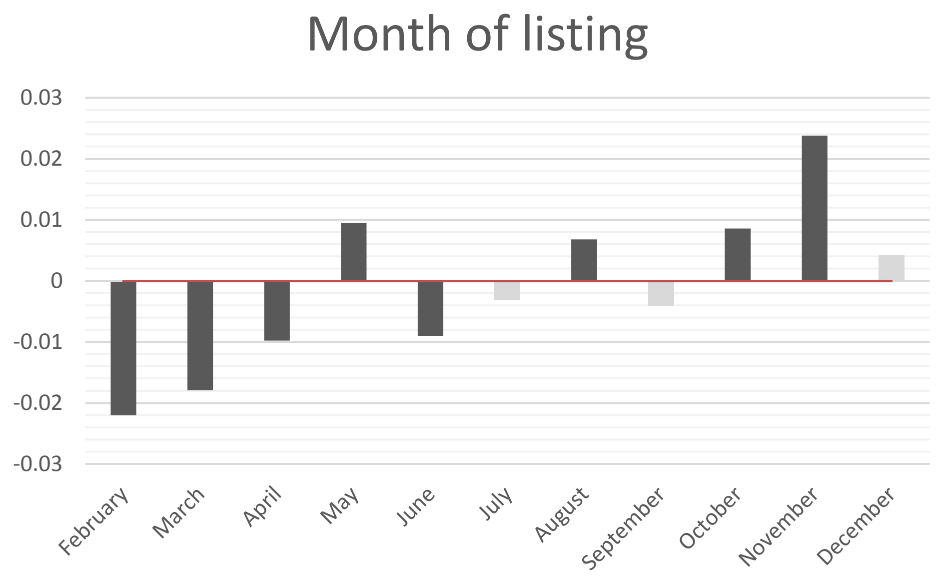
GLM Results

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	***



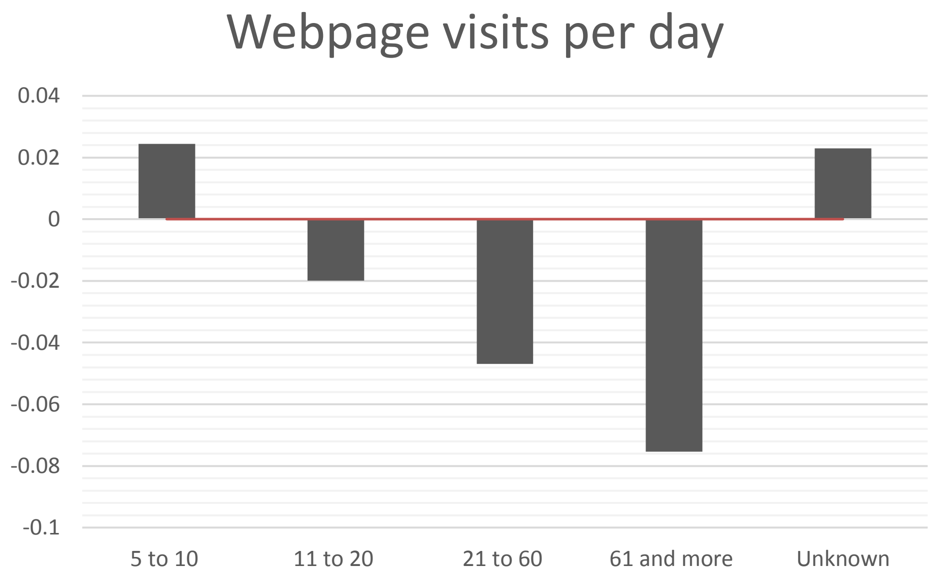
GLM Results

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	



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	***

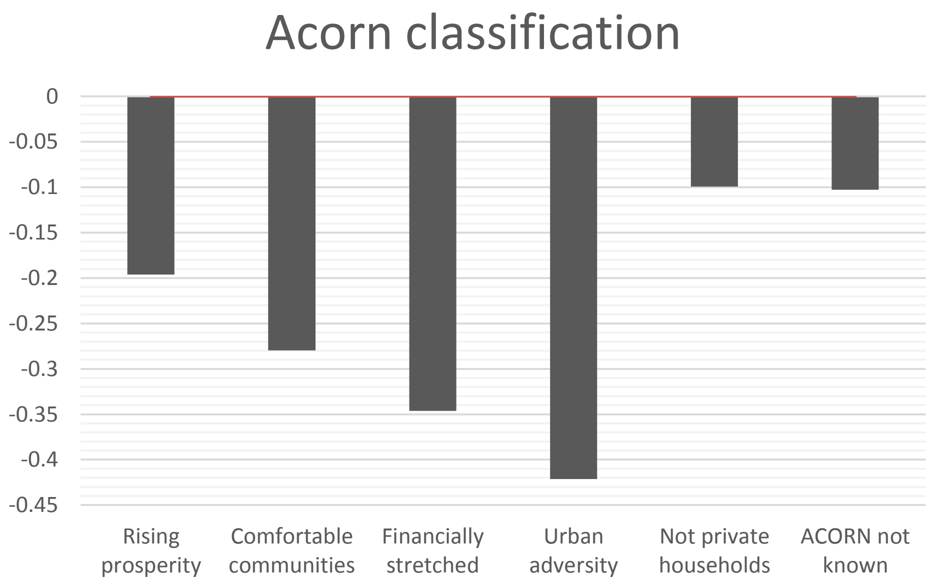


Consumer Data Research Centre

An ESRC Data Investment

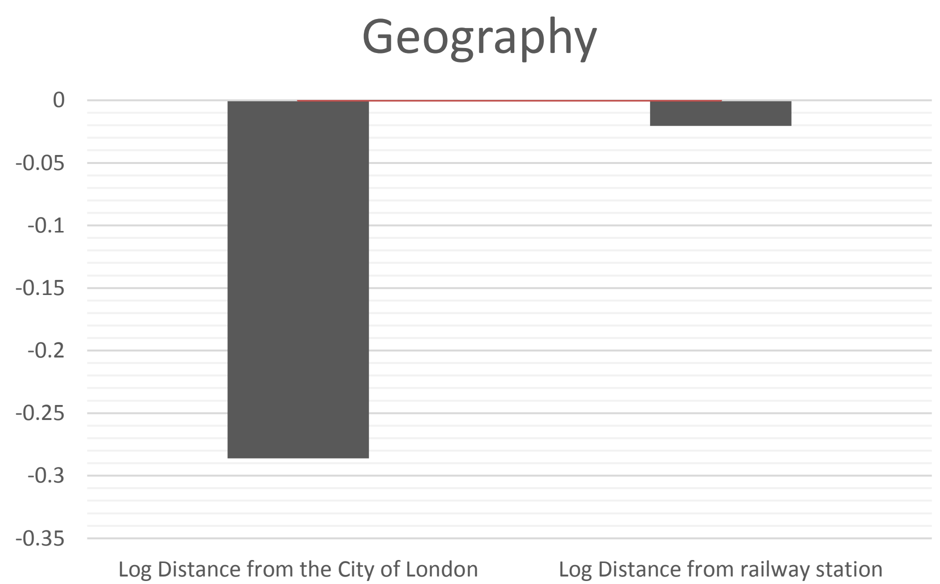
GLM Results

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	***



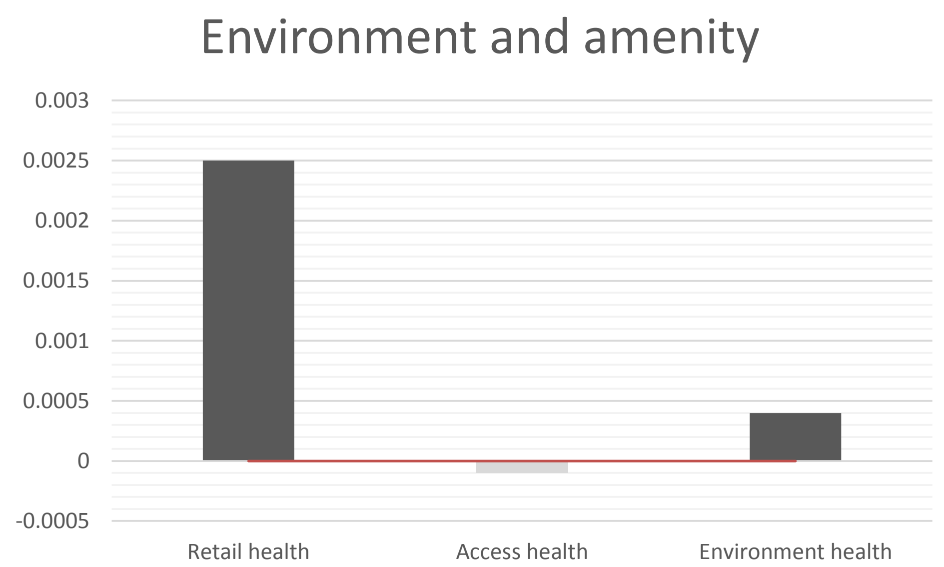
GLM Results

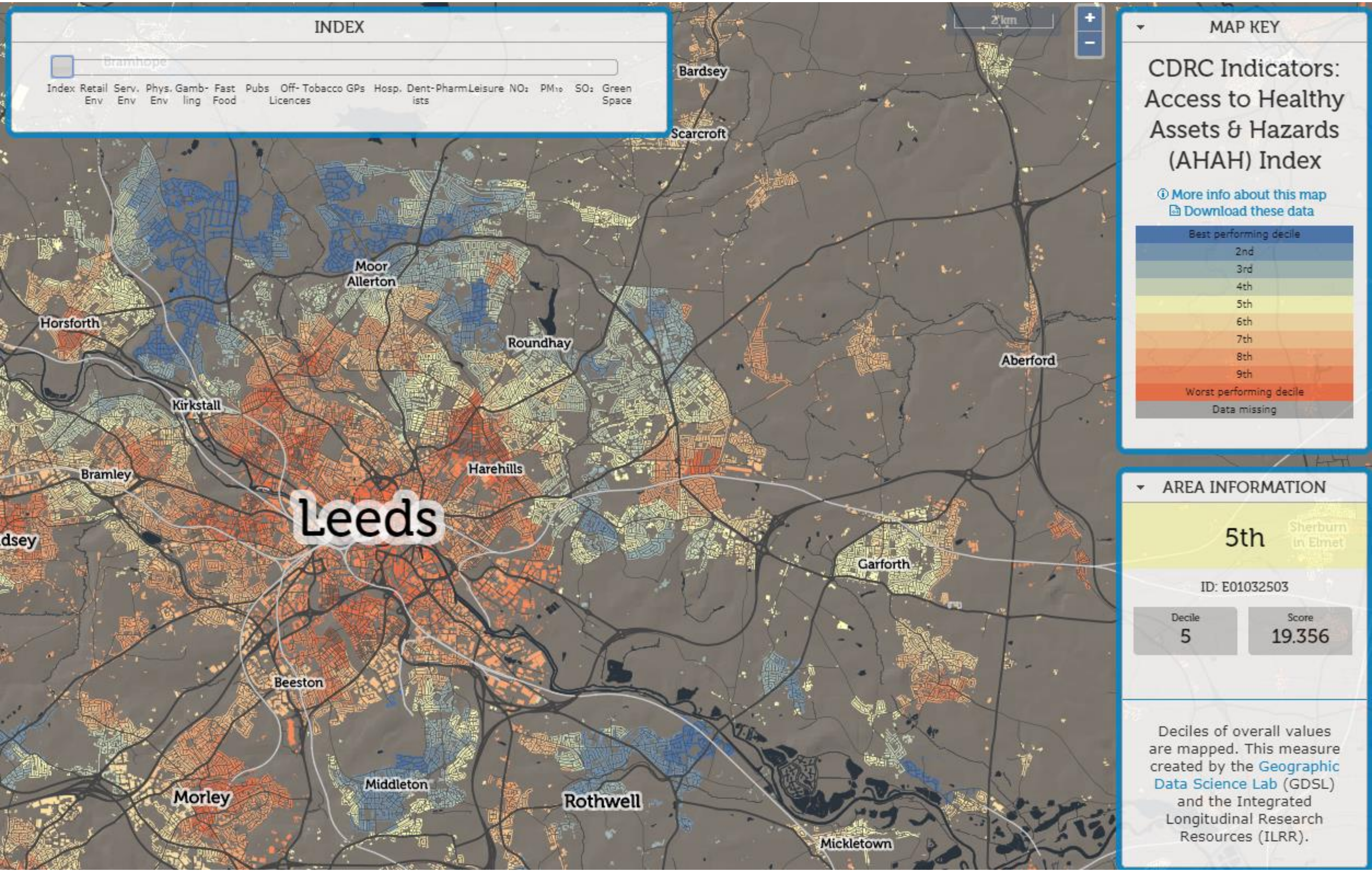
Attribute	N/median	estimate	std error	t	
Intercept	487253	6.451	0.0067	957.7	***
Log Distance from the City of London	113.95km	-0.2862	0.00079	-363.2	***
Log Distance from railway station	1.11km	-0.0204	0.001	-20	***



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	***





Access to Healthy Assets and Hazards (AHAH)

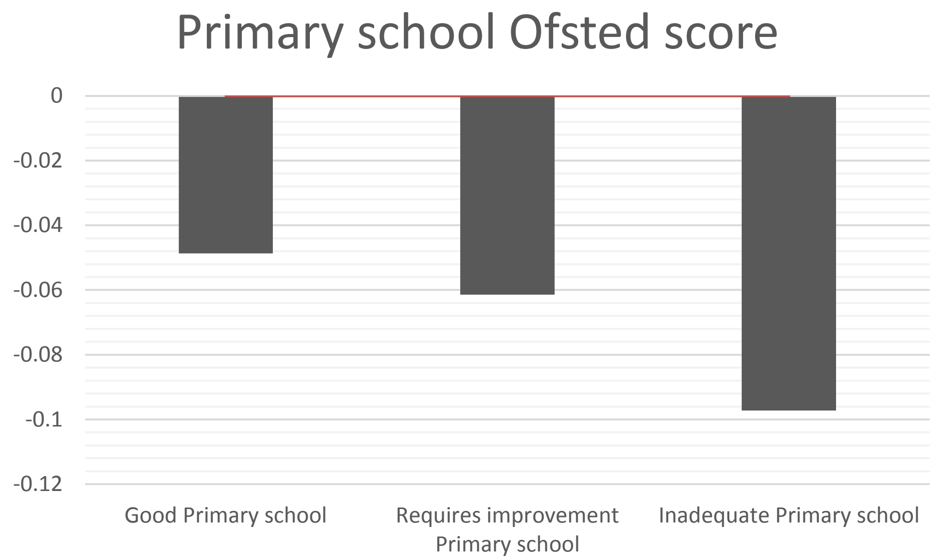
Daras, Konstantinos; Green, Mark; Davies, Alec; Singleton, Alex; Barr, Benjamin. (2017).

Consumer Data Research Centre

An ESRC Data Investment

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	***

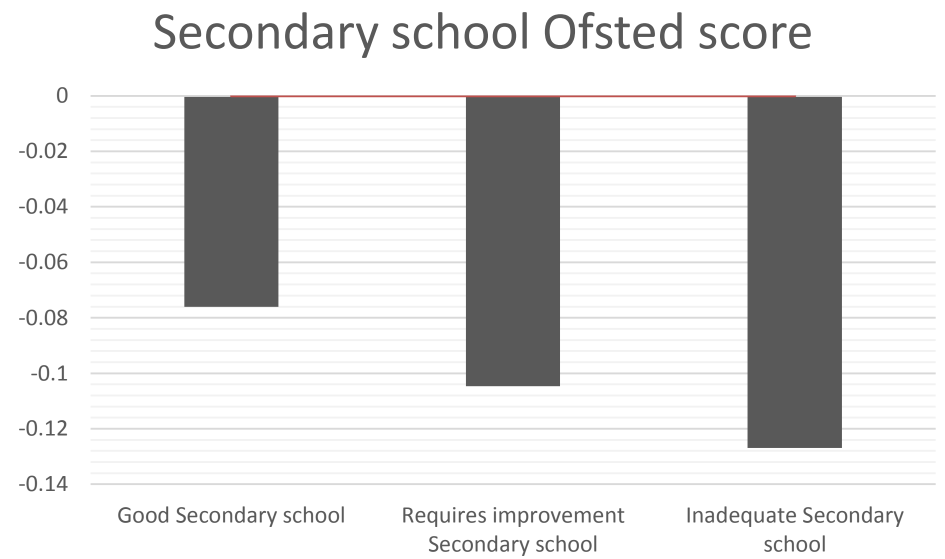


Consumer Data Research Centre

An ESRC Data Investment

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	***



Consumer Data Research Centre

An ESRC Data Investment

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

Consumer Data Research Centre

An ESRC Data Investment

- 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)

Consumer Data Research Centre

An ESRC Data Investment

Results – comparing r2

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

Consumer Data Research Centre

An ESRC Data Investment

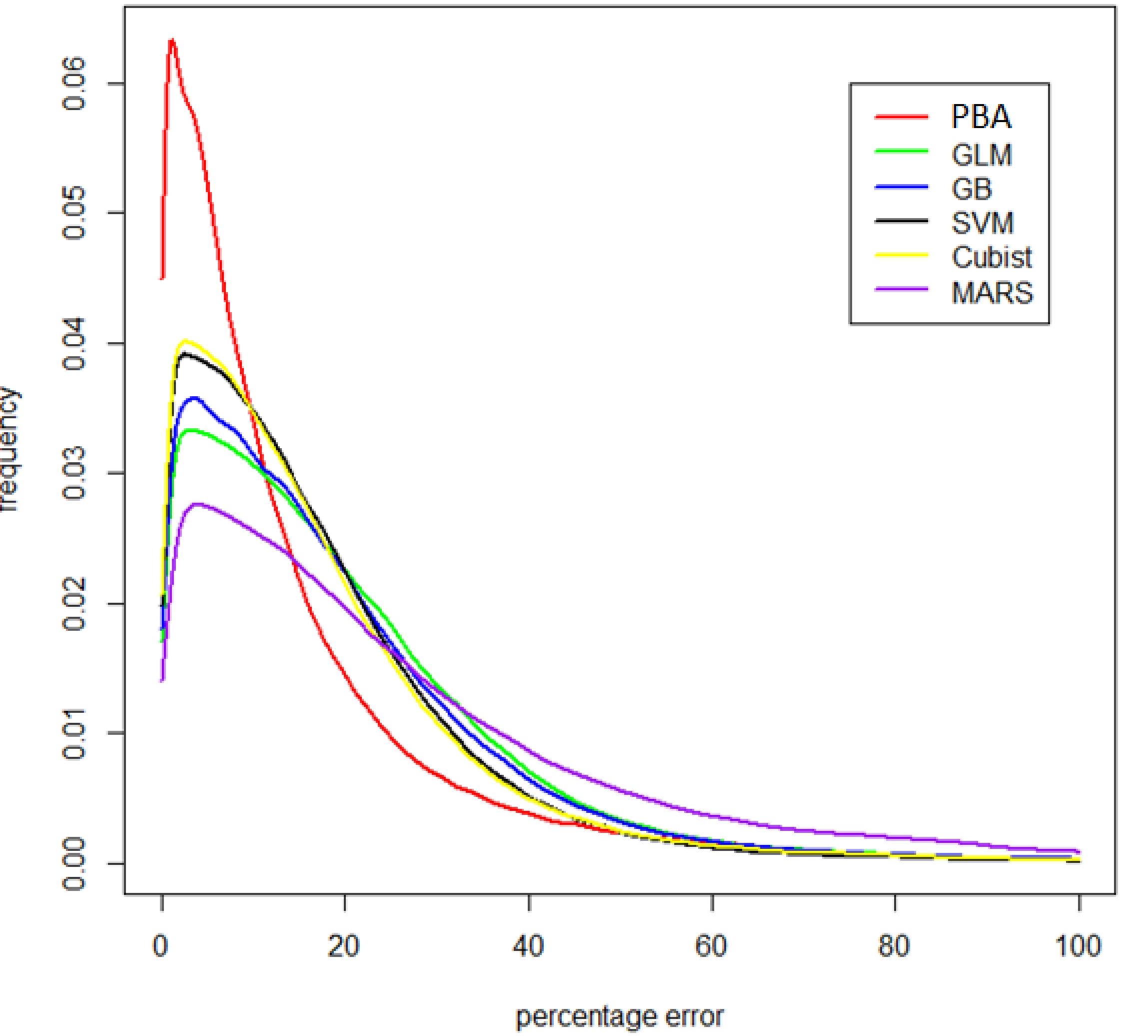
Results – comparing median percentage prediction error

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

Consumer Data Research Centre

An ESRC Data Investment

Results – distribution of percentage error



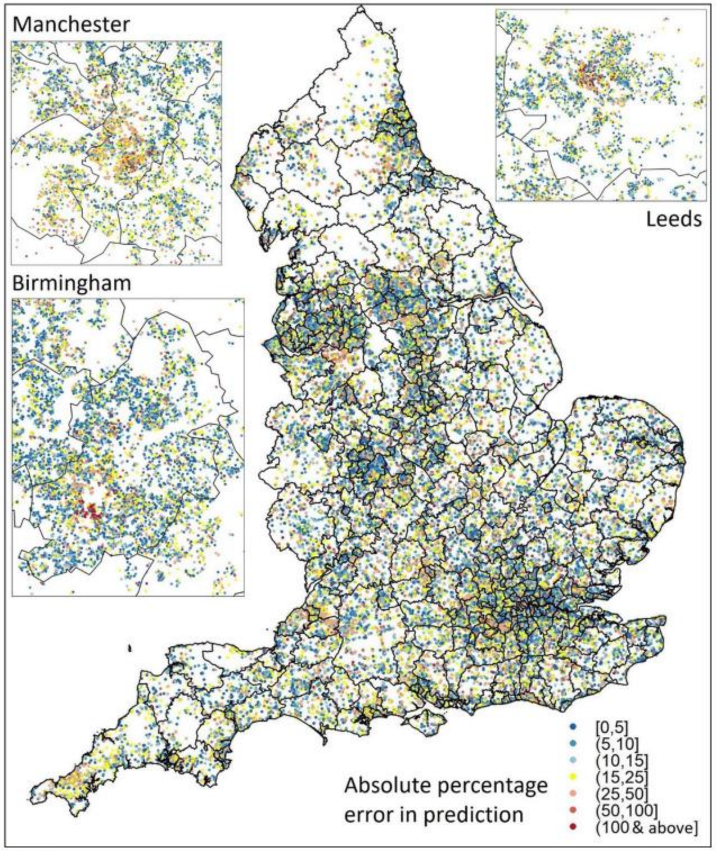


Fig. 2 Absolute percentage prediction error from cubist model

Consumer Data Research Centre

An ESRC Data Investment

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 r2

- 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.

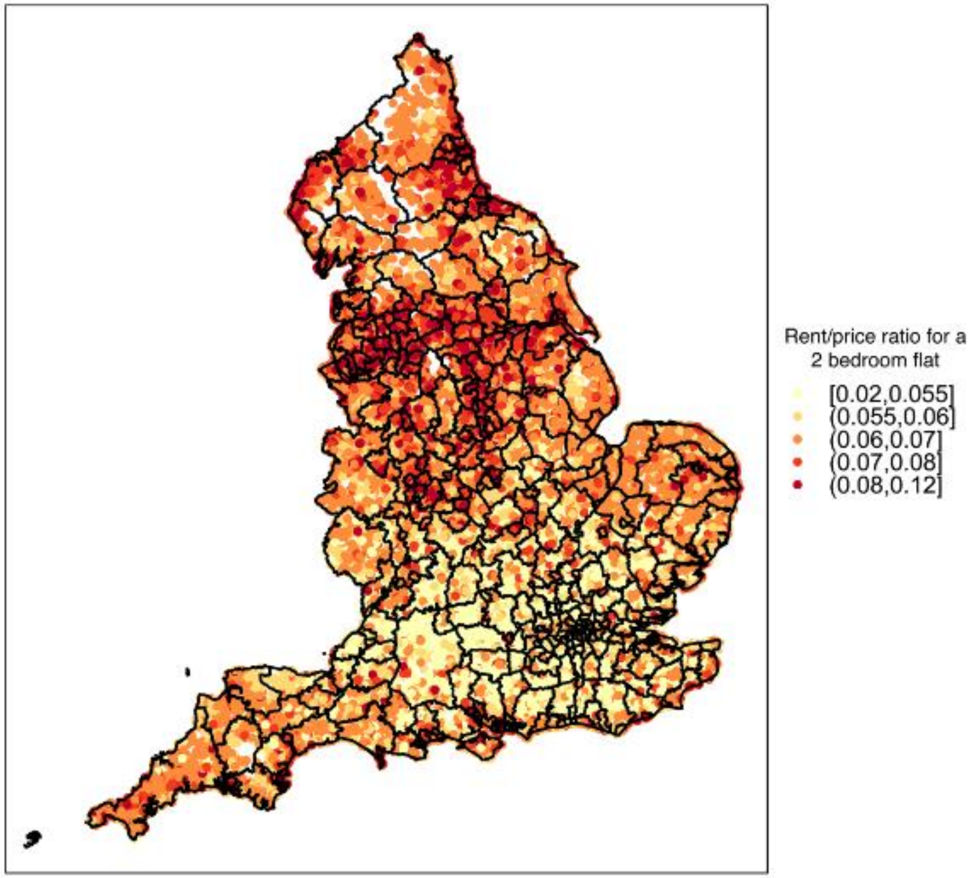
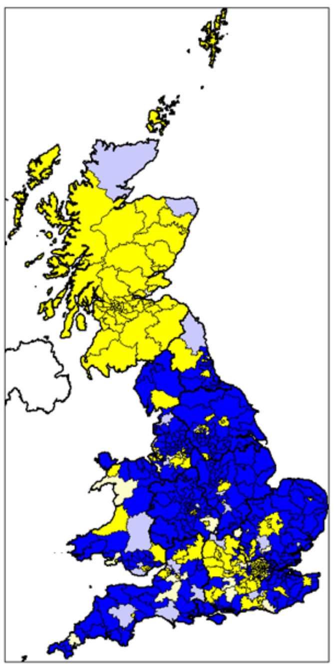
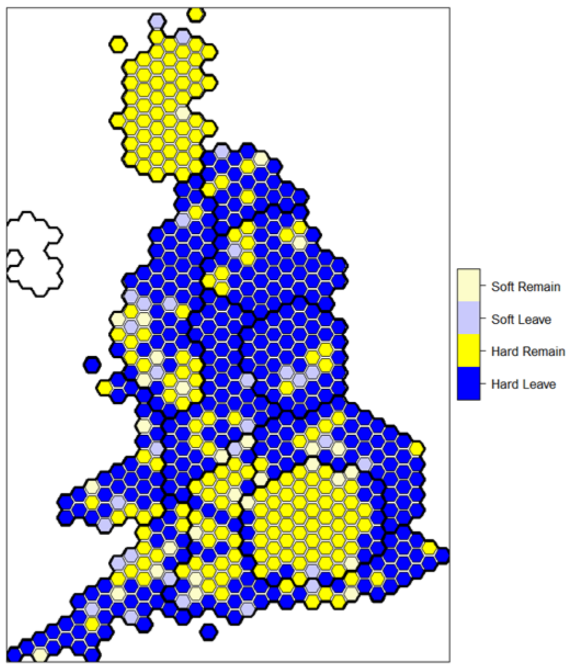


FIGURE 2 Estimated rent/price ratio for two-bedroomed flats for a sample of English postcodes.

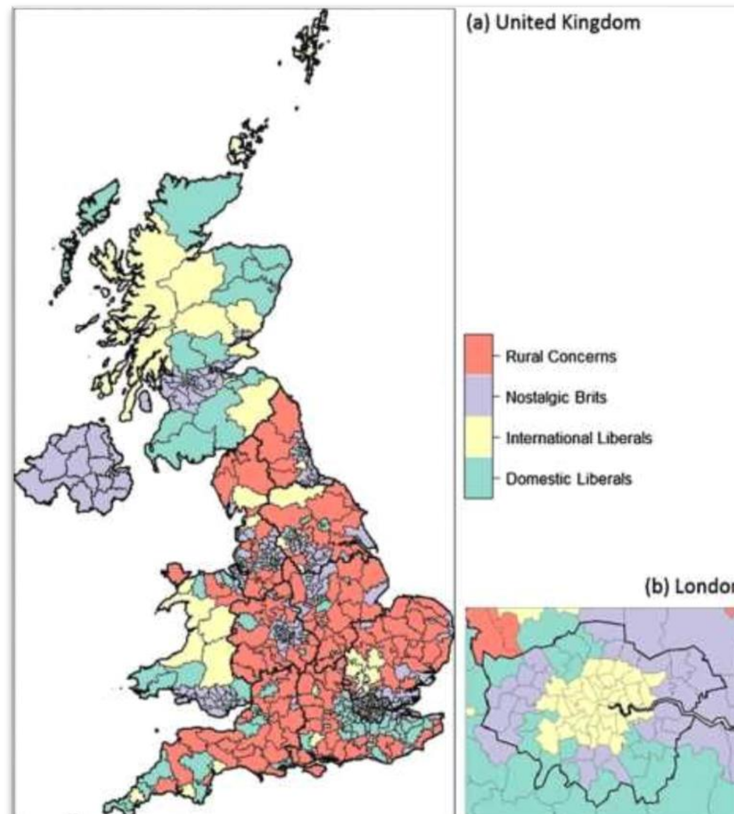
Example 2: E-Petition Data

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

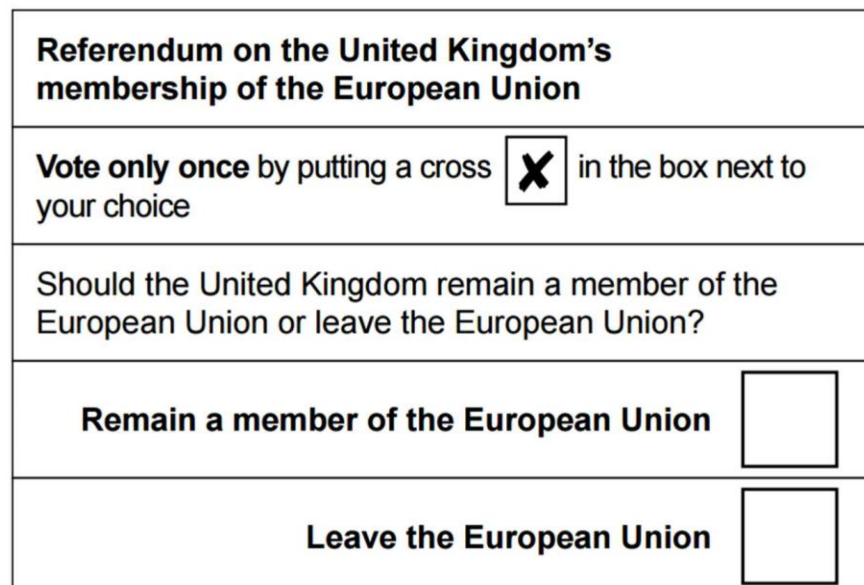


Consumer Data Research Centre

An ESRC Data Investment

Context

- 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.


Referendum on the United Kingdom's membership of the European Union

Vote only once by putting a cross ✗ in the box next to your choice

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

Remain a member of the European Union

Leave the European Union

Consumer Data Research Centre

An ESRC Data Investment

Context

"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.

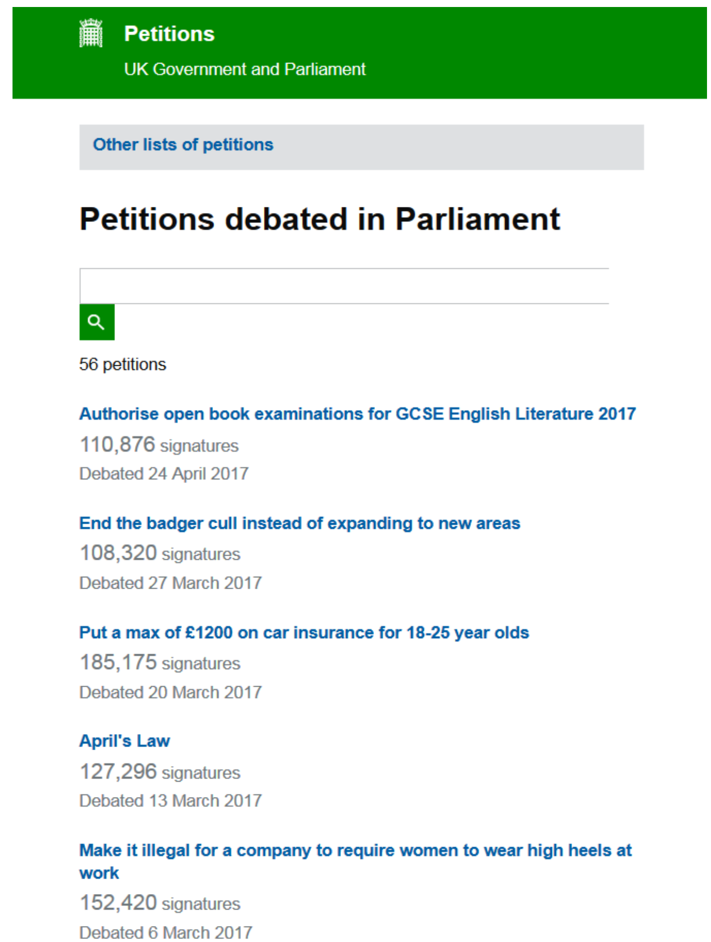
Hanretty, C. 2017. "Areal interpolation and the UK's referendum on EU membership." Journal of Elections, Public Opinion and Parties:1-18. doi: 10.1080/17457289.2017.1287081.

Consumer Data Research Centre

An ESRC Data Investment

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



Consumer Data Research Centre

An ESRC Data Investment

e-petitions used

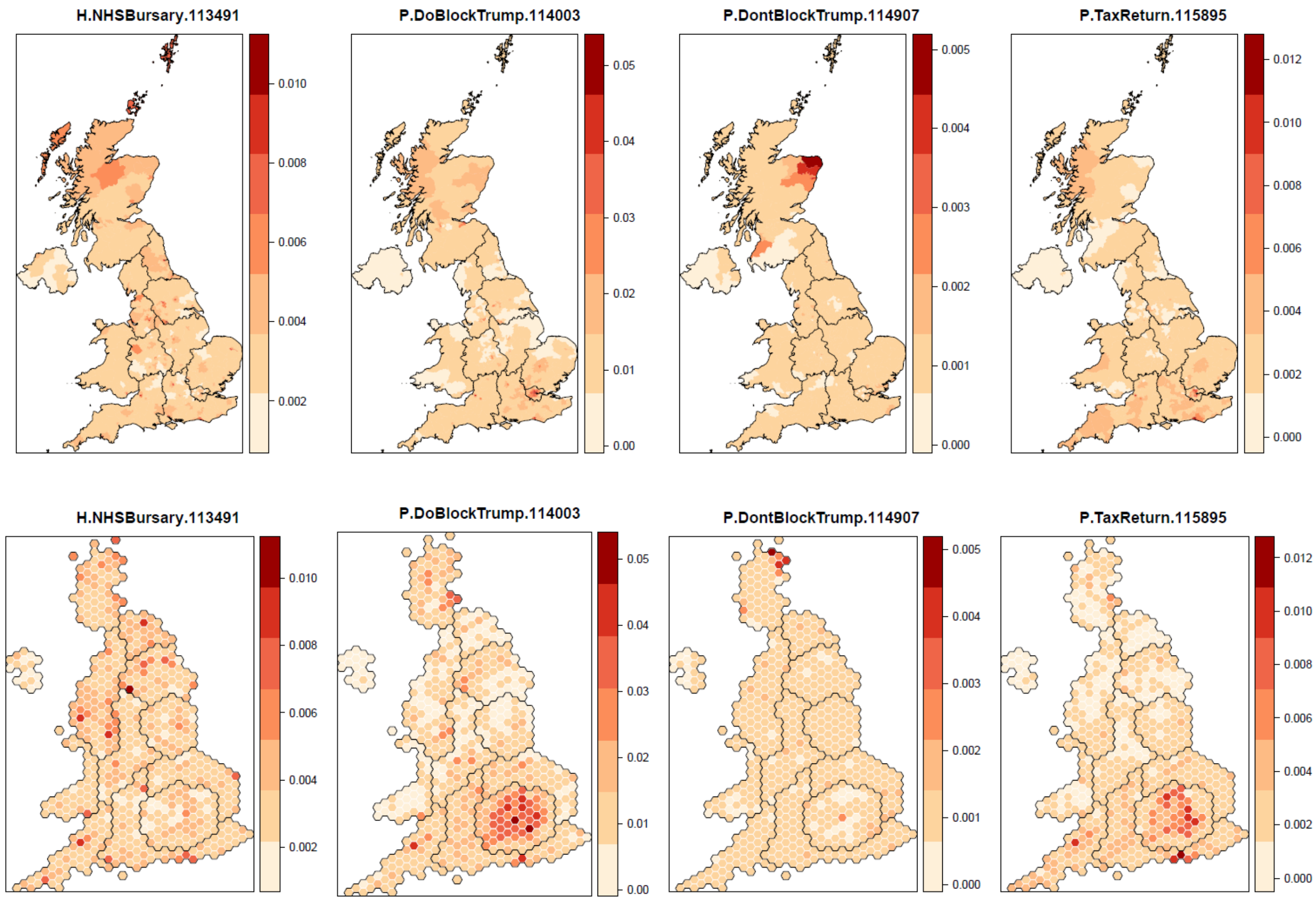
Table 1. List of e-petitions used in this study.

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

Consumer Data Research Centre

An ESRC Data Investment

e-petitions: geography



Consumer Data Research Centre

An ESRC Data Investment

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

Consumer Data Research Centre

An ESRC Data Investment

Incompatible geographies

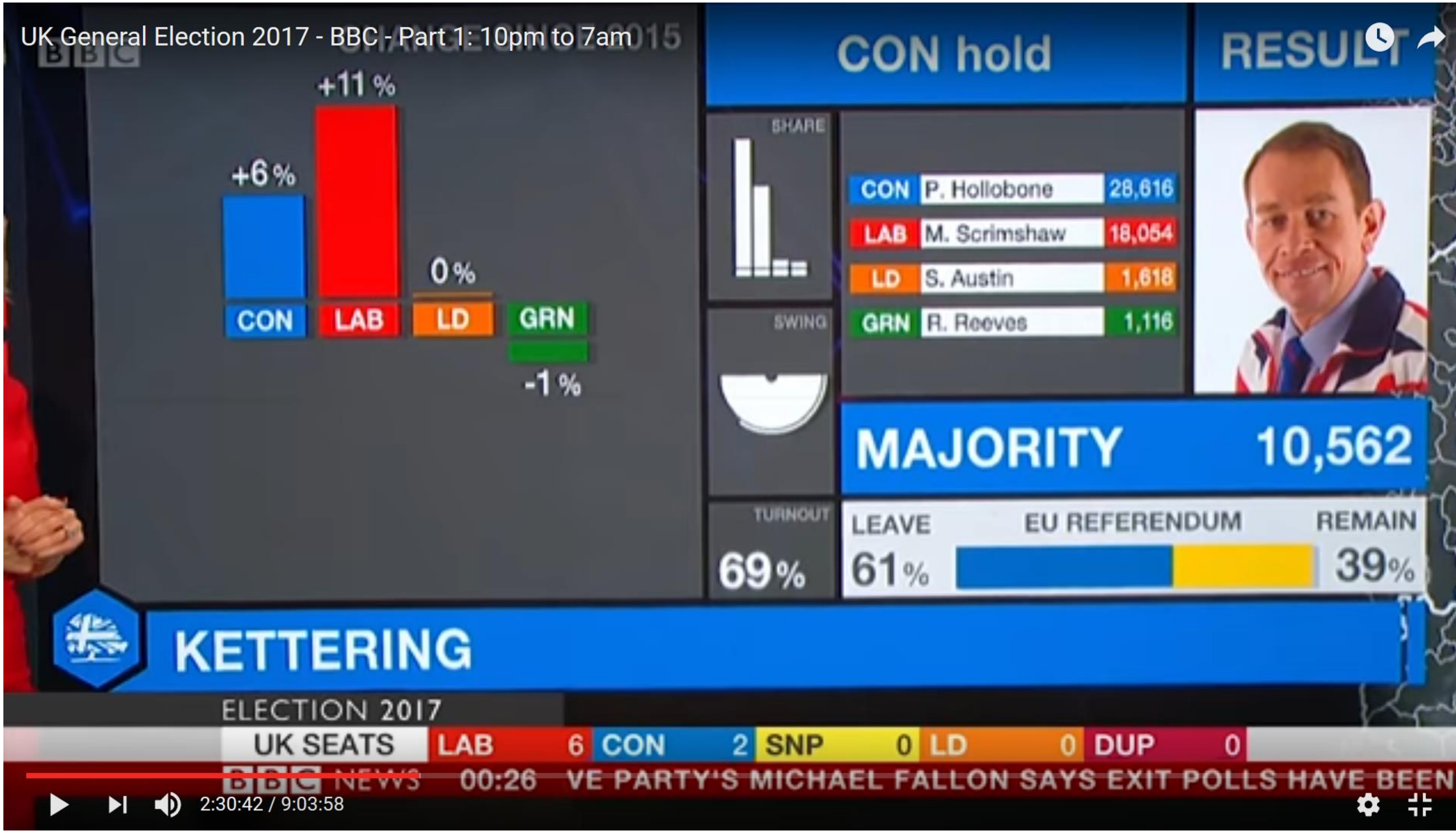
- Referendums results from 382 CAs
- E-petition counts from 632 WPCs (exclude NI)
- A new geography needed where aggregations of CAs are the same as aggregations of WPCs
- 173 Data Zones

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
Total		173	380	632

Consumer Data Research Centre

An ESRC Data Investment

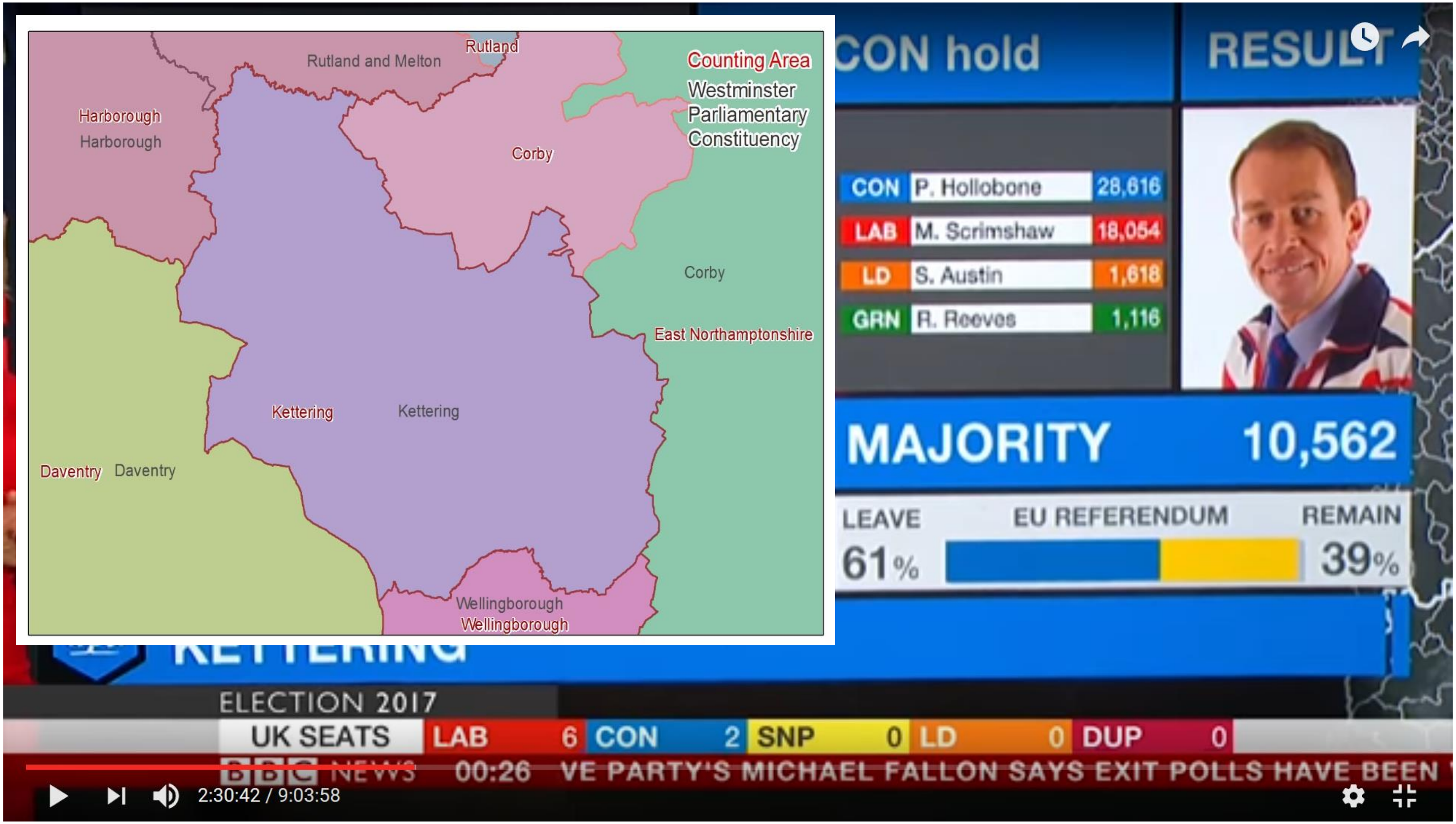
Here one CA = one WPC



Consumer Data Research Centre

An ESRC Data Investment

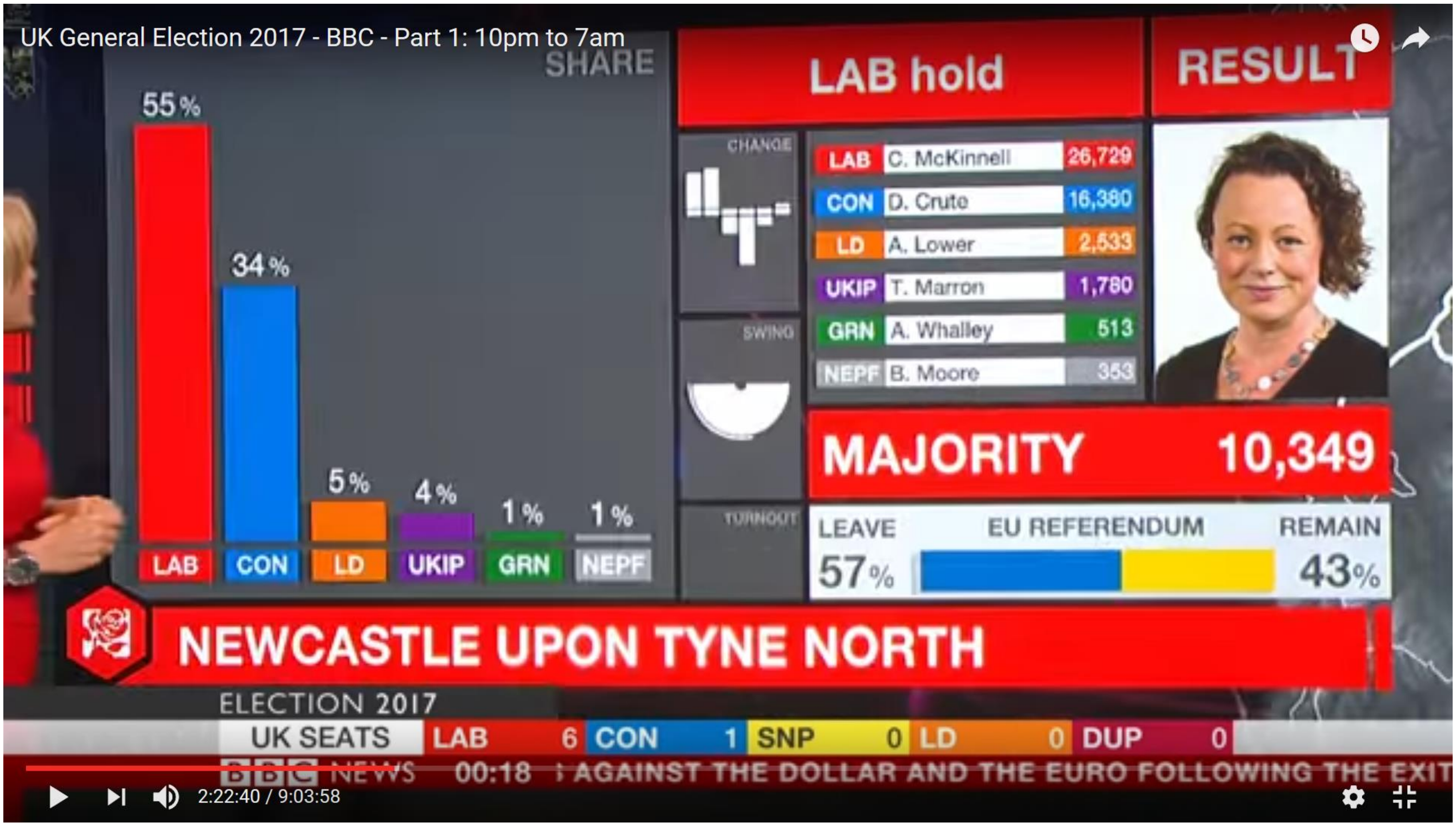
Here one CA = one WPC



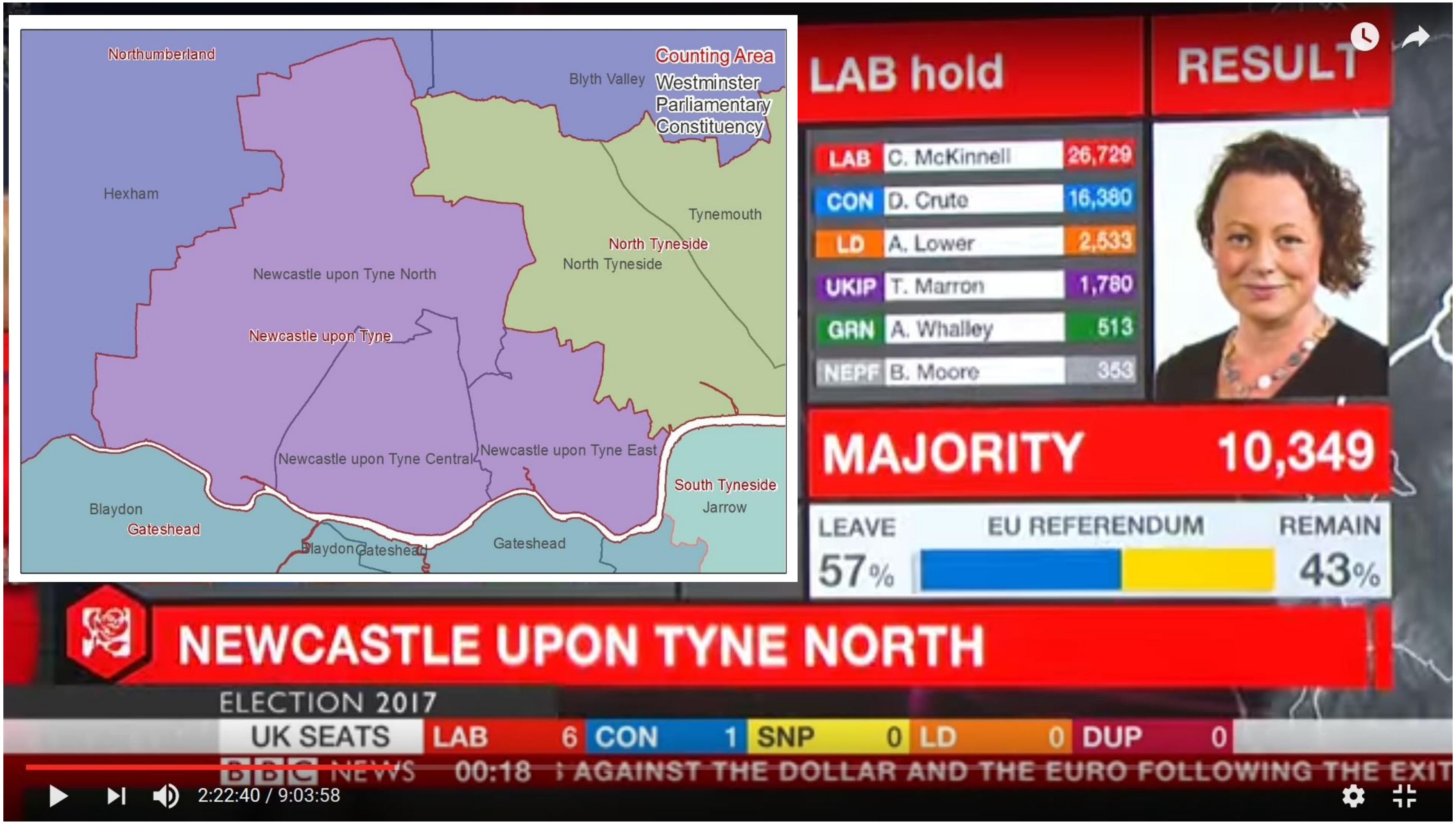
Consumer Data Research Centre

An ESRC Data Investment

Here one CA = three WPCs



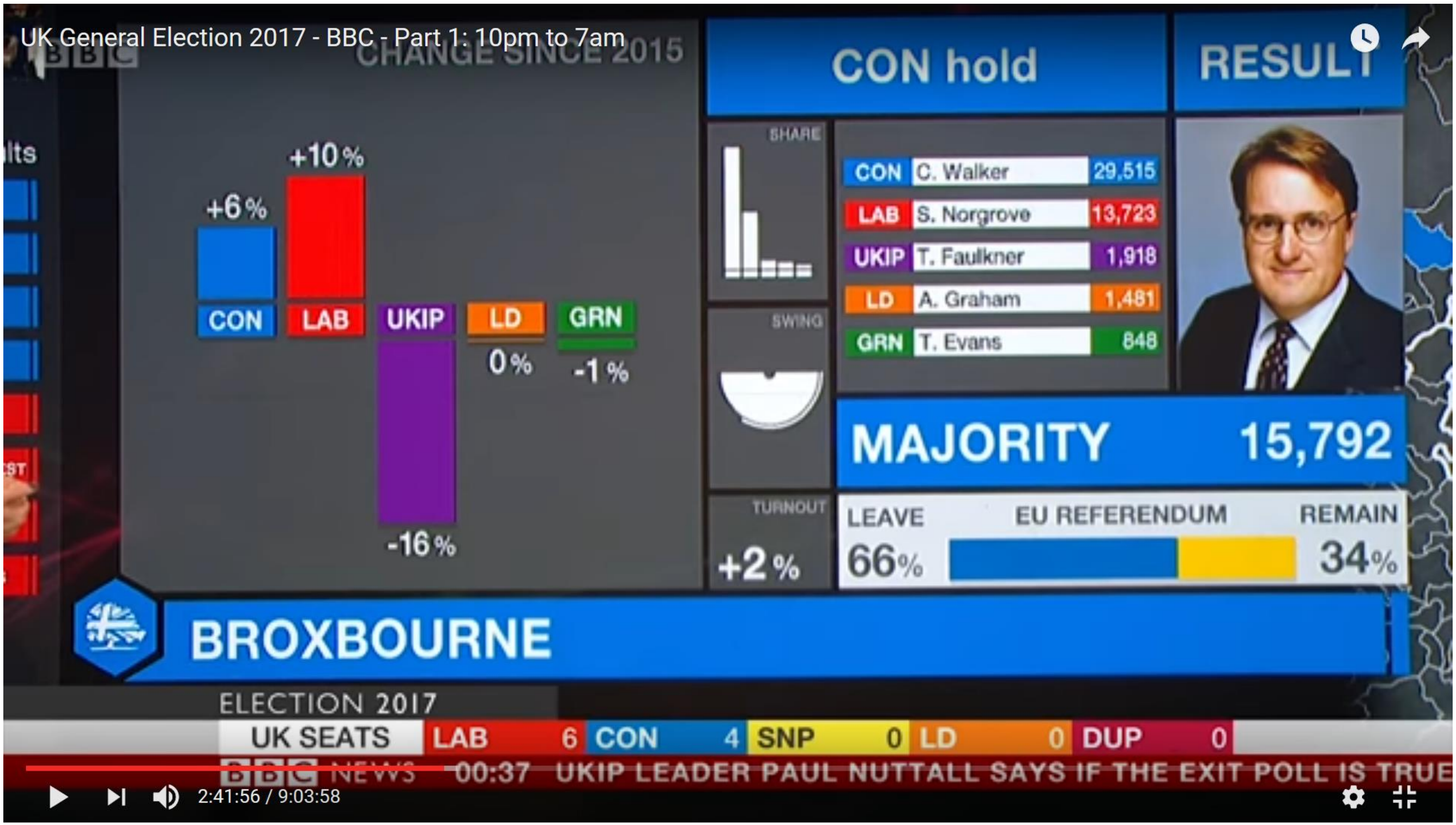
Here one CA = three WPCs



Consumer Data Research Centre

An ESRC Data Investment

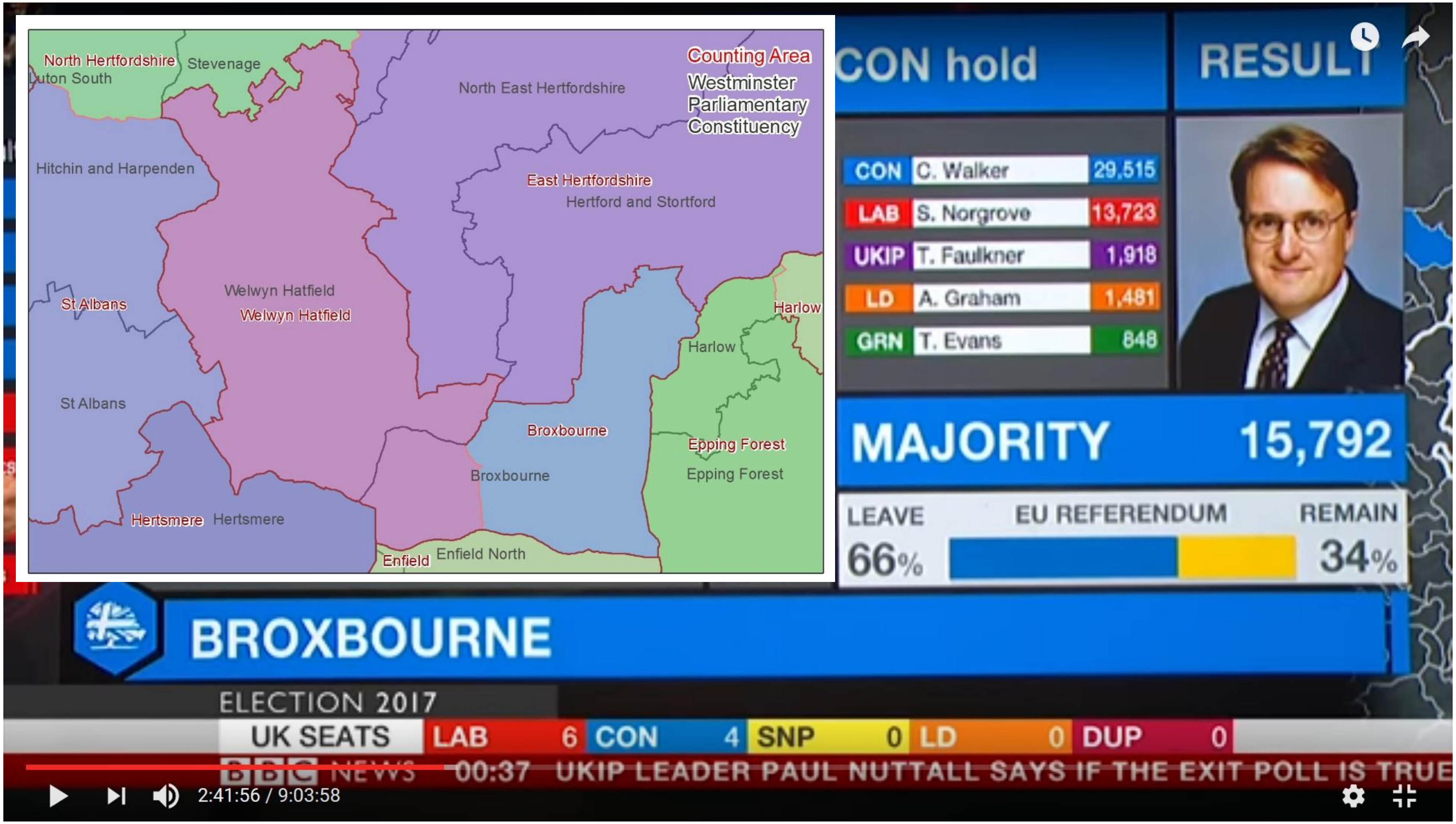
Here two CA = two WPCs



Consumer Data Research Centre

An ESRC Data Investment

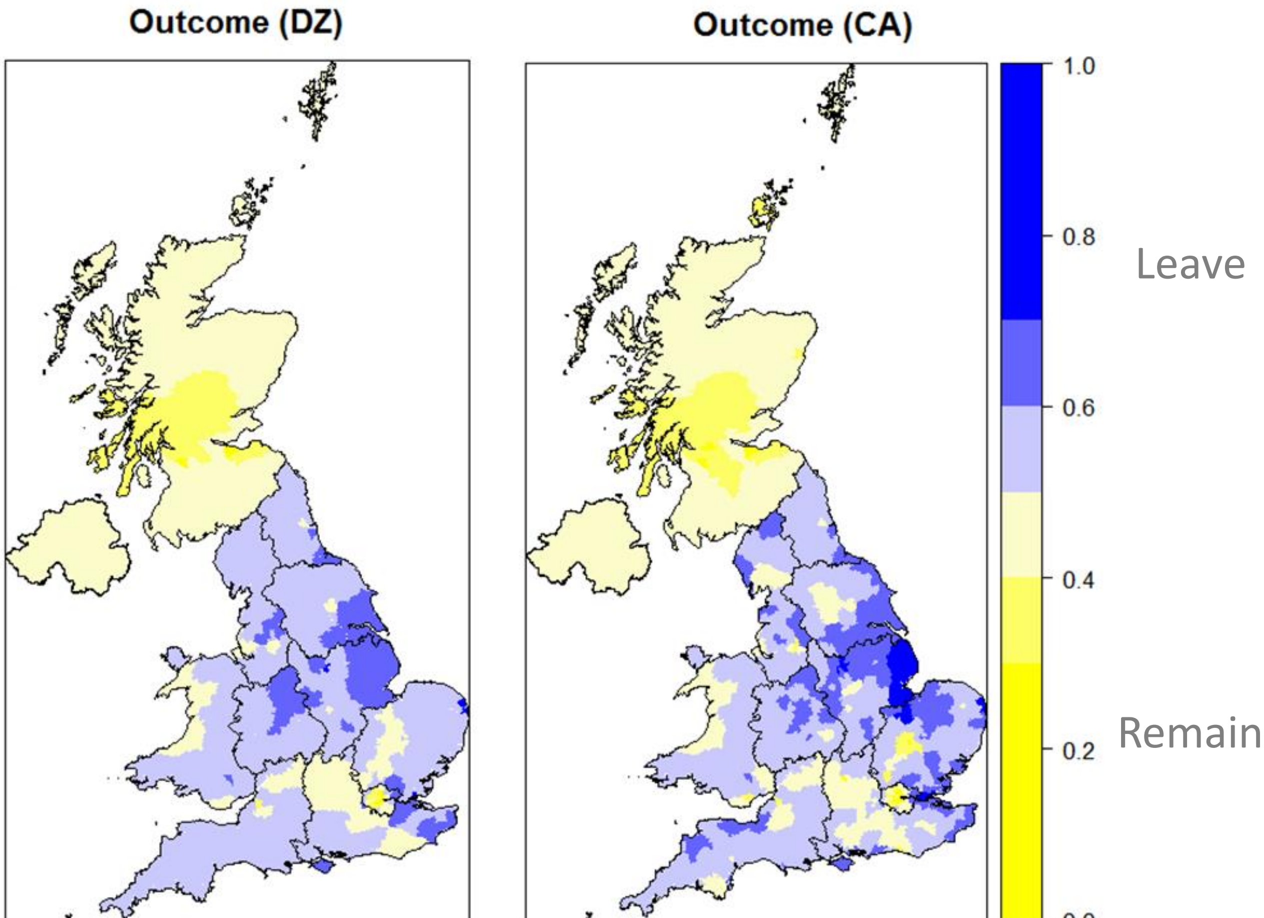
Here two CA = two WPCs



Consumer Data Research Centre

An ESRC Data Investment

Remapped outcomes



Machine learning algorithms

- **Lazy Learners**
  - K nearest neighbours
  - Self-organising maps

- Characterised by capturing learning through a set of similarity relationships in multidimensional ‘space’

Consumer Data Research Centre

An ESRC Data Investment

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)

Consumer Data Research Centre

An ESRC Data Investment

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 tradition decision tree and regression equations
  - At the leaf there is an estimated regression equation rather than a constant.

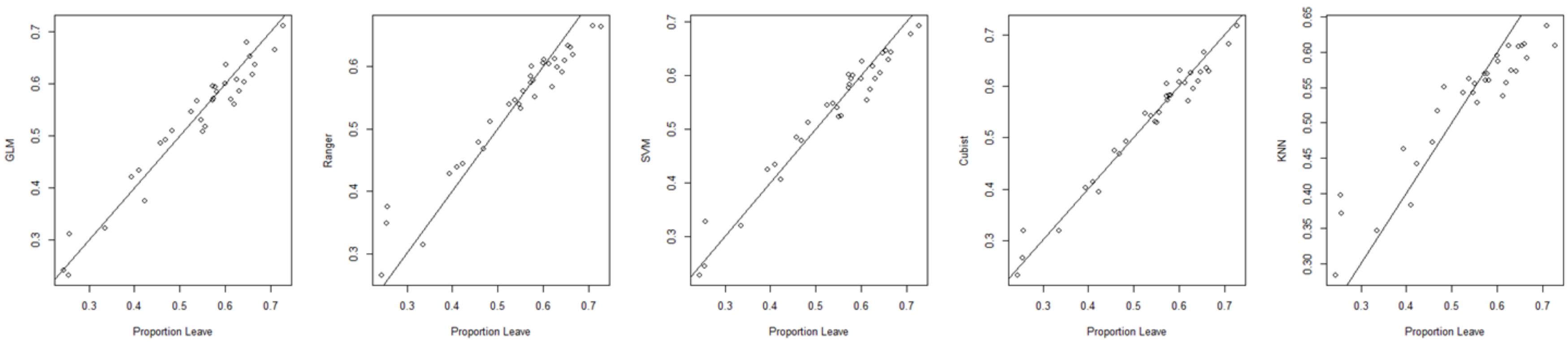
Consumer Data Research Centre

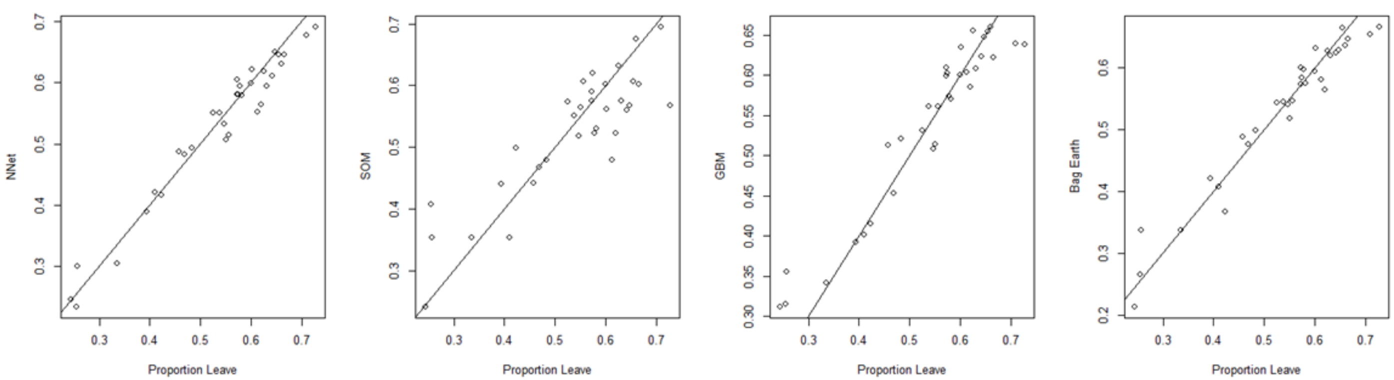
An ESRC Data Investment

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^2
Cubist	**0.0224**	**0.971**
Nnet	0.0270	0.959
SVM	0.0279	0.955
BagEarth	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.

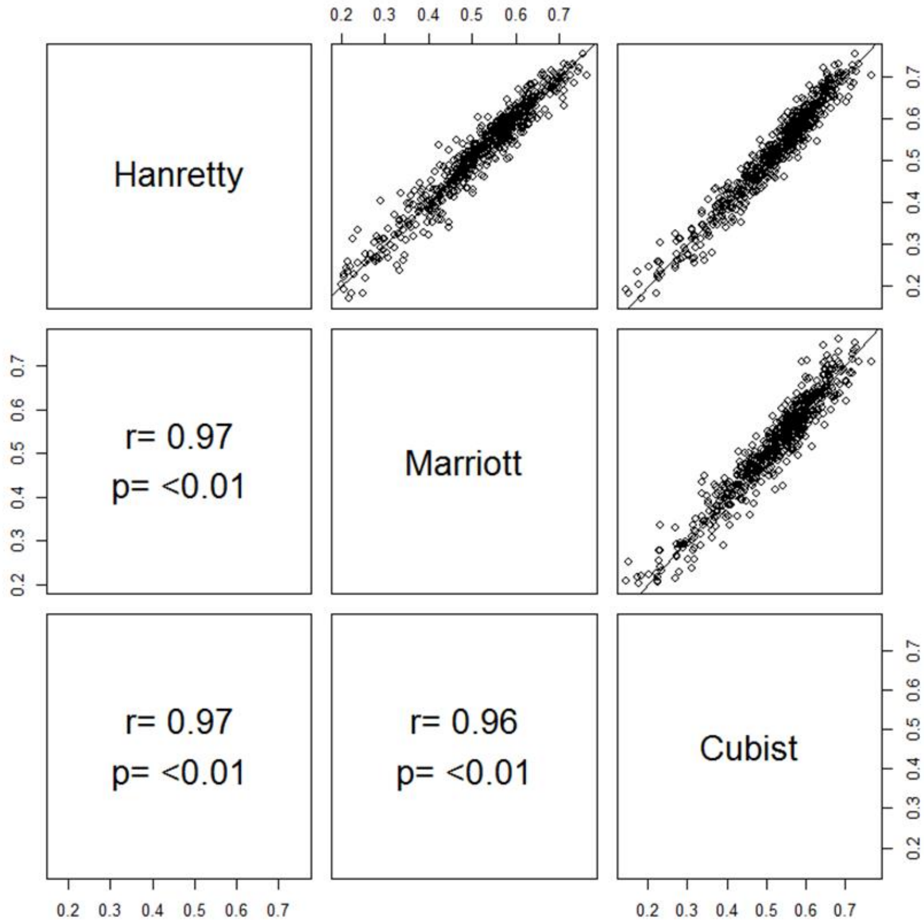
Hanretty, C. 2017. "Areal interpolation and the UK's referendum on EU membership." Journal of Elections, Public Opinion and Parties:1-18. doi: 10.1080/17457289.2017.1287081.

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)

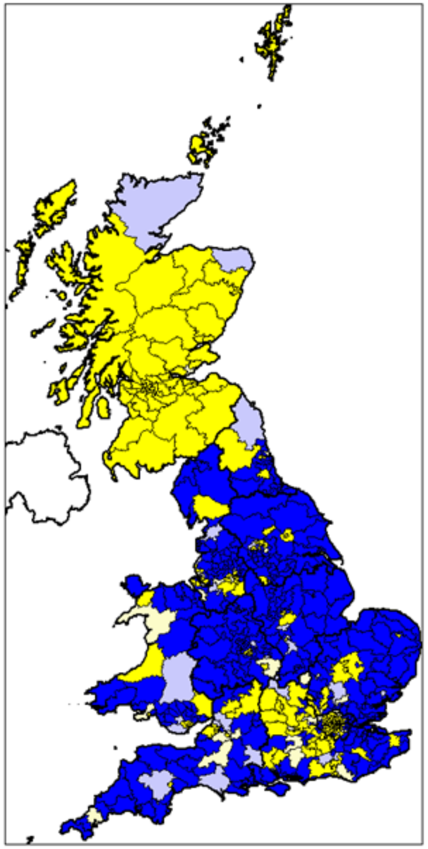

Hanretty
Marriott
Cubist
r= 0.97
p= <0.01
r= 0.97
p= <0.01
r= 0.96
p= <0.01
0.2 0.3 0.4 0.5 0.6 0.7

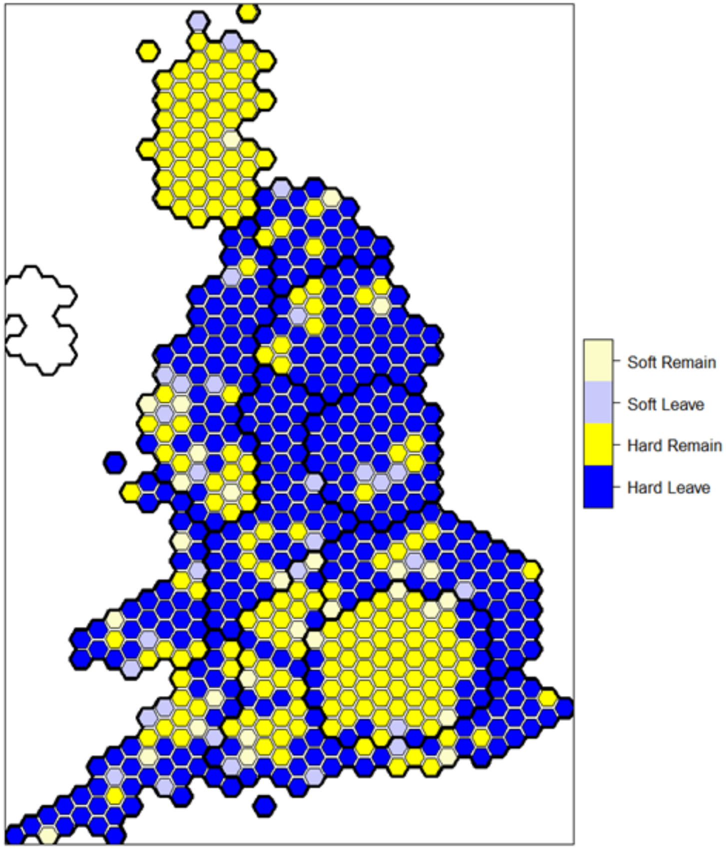
Consumer Data Research Centre

An ESRC Data Investment

Results (BREXIT)

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





Discussion

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

Consumer Data Research Centre

An ESRC Data Investment

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.

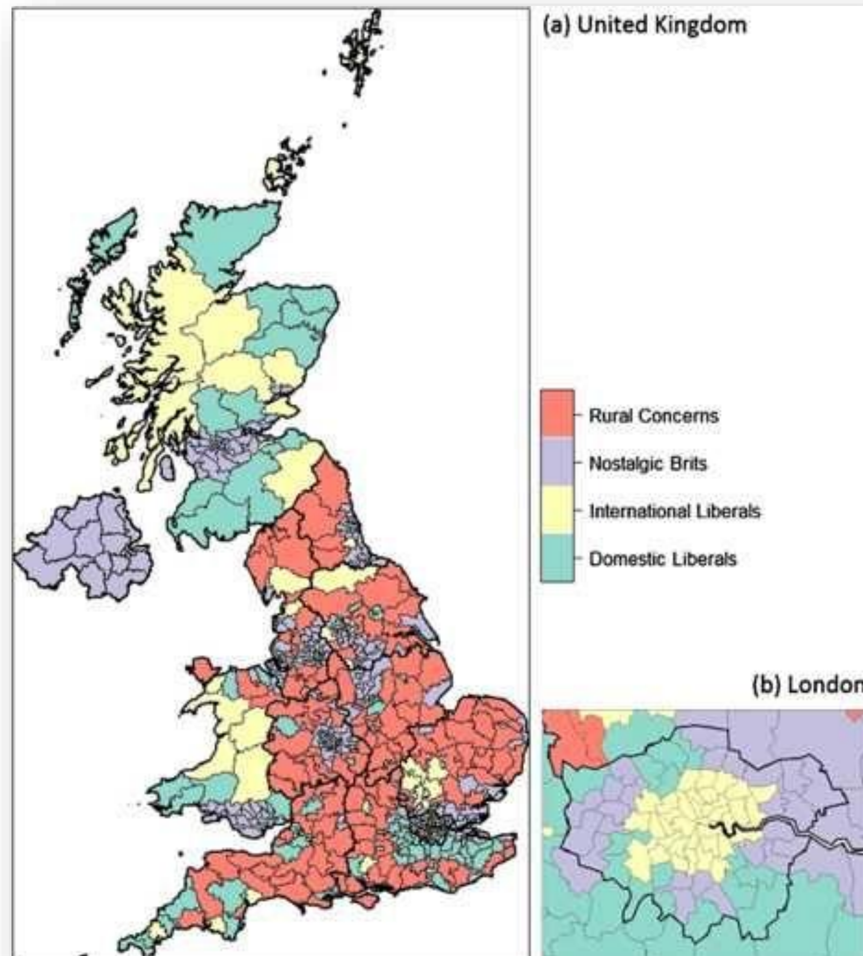
Consumer Data Research Centre

An ESRC Data Investment

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.



Consumer Data Research Centre

An ESRC Data Investment

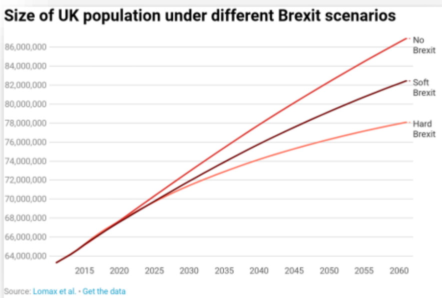
Final Conclusions

- ‘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

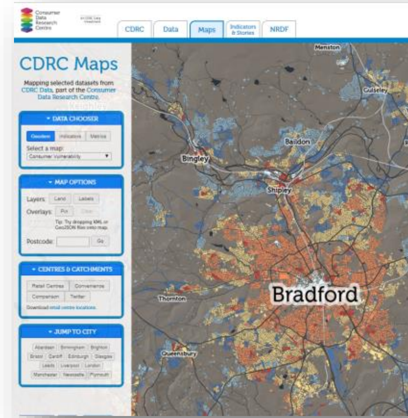
Consumer Data Research Centre

An ESRC Data Investment

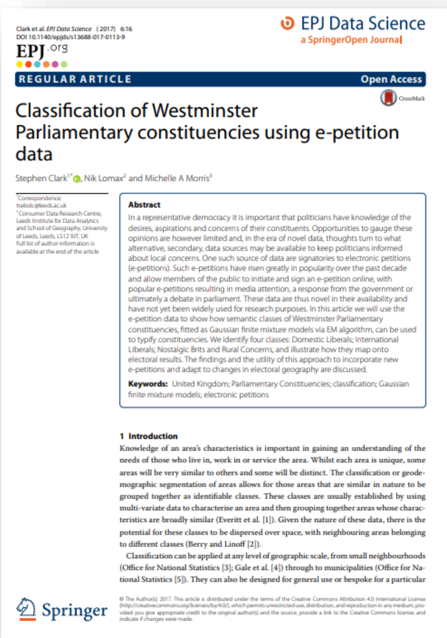
Links and reading



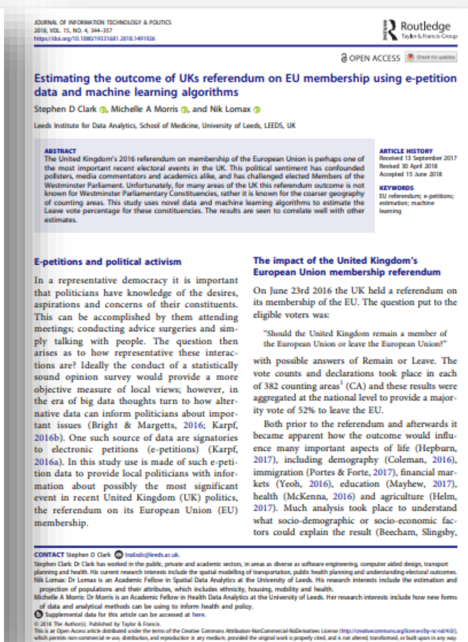
Link to *CDRC Maps*
https://maps.cdrc.ac.uk



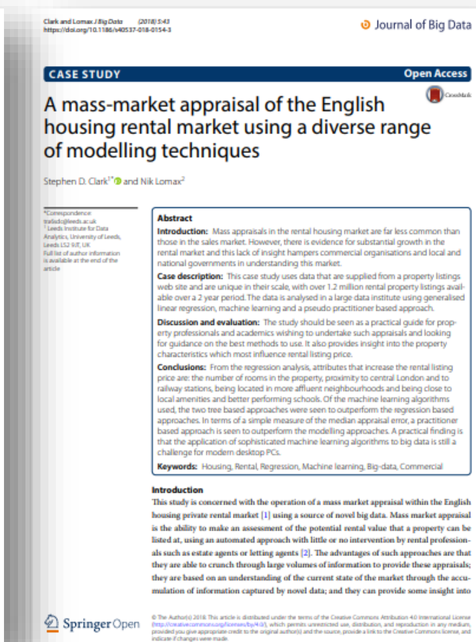
Link to *The Conversation* article
https://bit.ly/2YUzwCT



https://bit.ly/2Z6Meyp

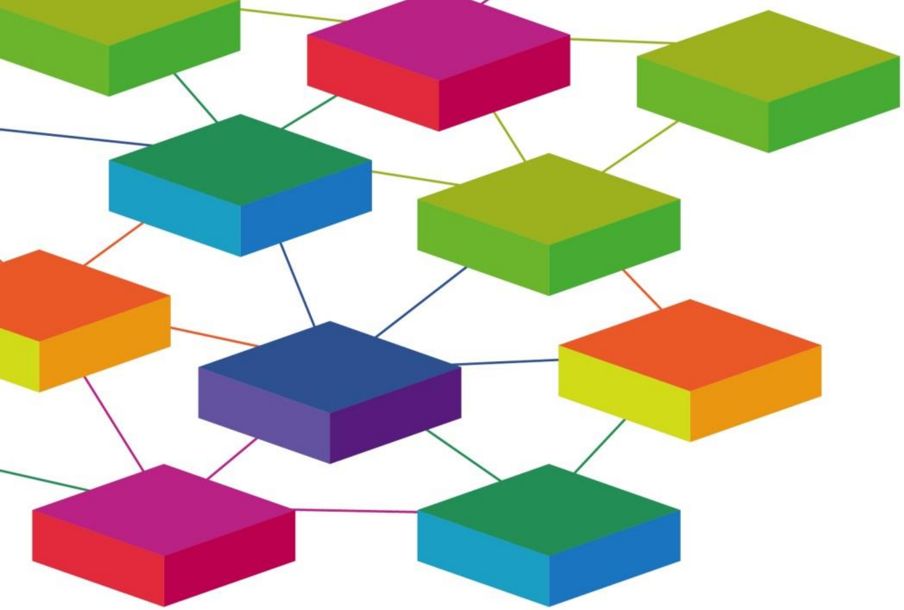


https://bit.ly/2Z96j7d



https://bit.ly/2MvtFCE

https://bit.ly/2JTLt8t



Questions

@niklomax