Relieving despair at the prospect of the general elections using Bayesian modelling

Gianluca Baio

University College London
Department of Statistical Science

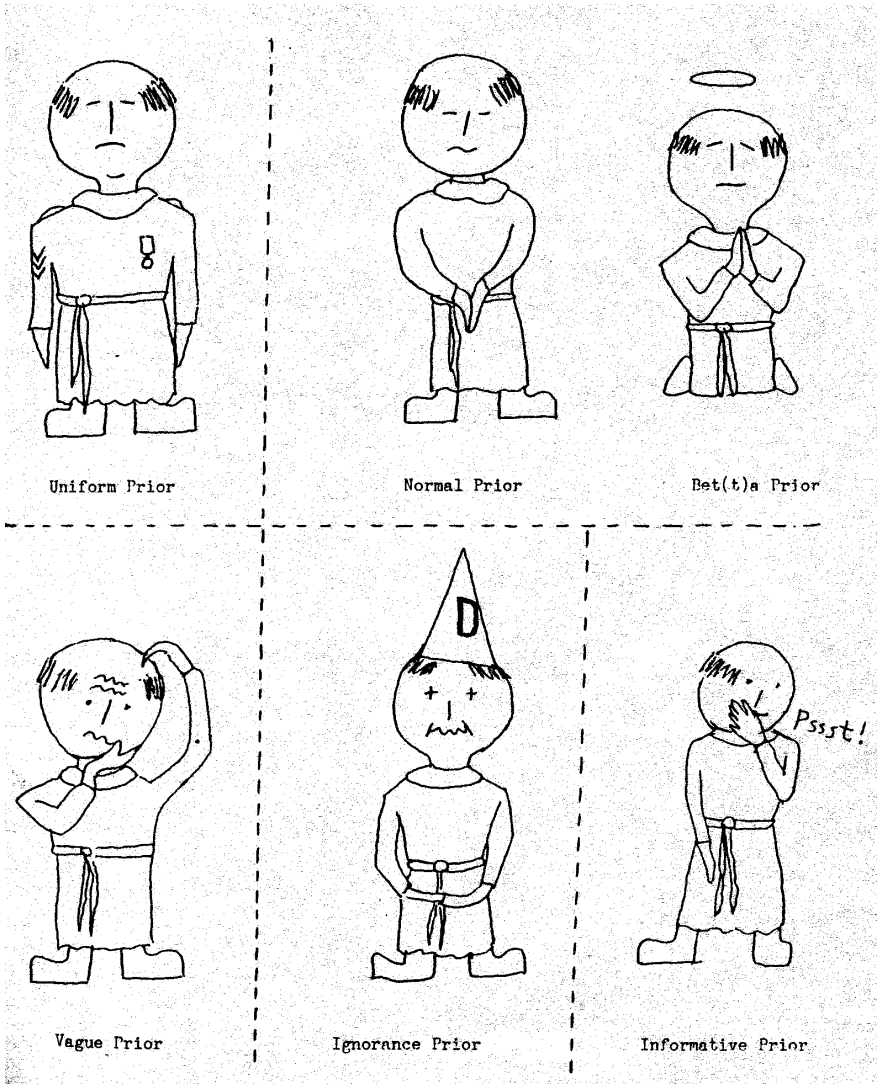
g.baio@ucl.ac.uk
http://www.ucl.ac.uk/statistics/research/statistics-health-economics/
http://www.statistica.it/gianluca
https://github.com/giabaio

RSS 2018 International Conference
Cardiff

Thursday 6 September 2018

- I'm not a professional Psephologist, so I don't know much about pebbles and ballots...

- I'm not a professional Psephologist, so I don't know much about pebbles and ballots...
- **But**: I am a professional (Bayesian!) Statistician and I have a strong interest in Politics and Elections



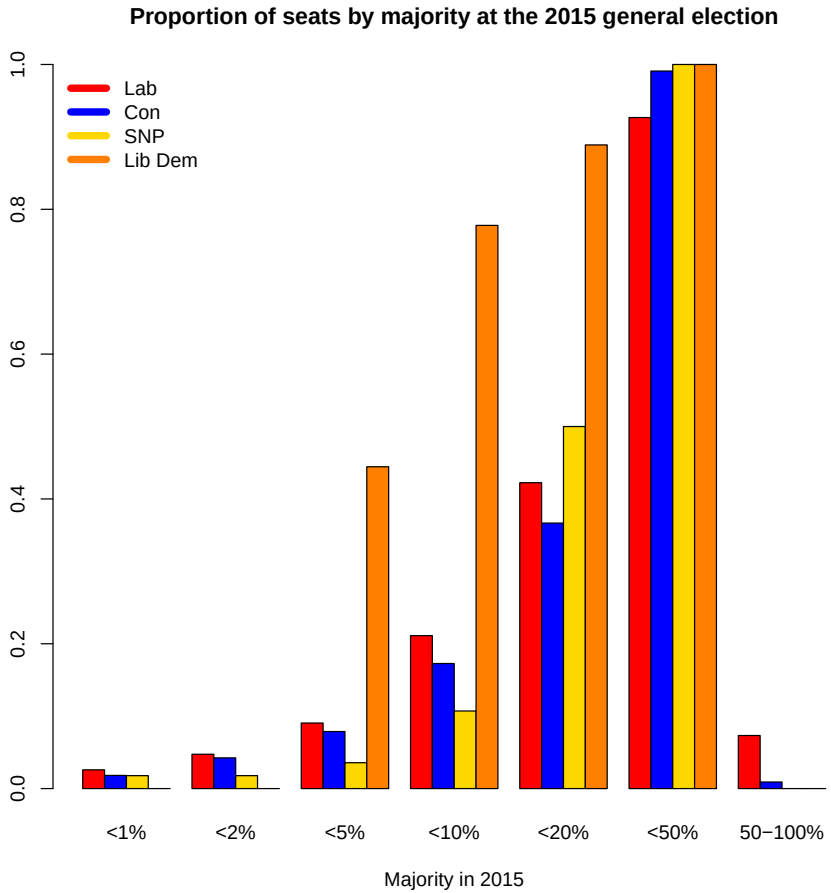
Outline

1. **Background**
   - Snap election
   - Small majorities
   - All about Europe
2. **Statistical modelling**
   - Data
   - Bayesian specifications
   - Simulating the elections
3. **Results**
   - Time & space
   - Postmortem
4. **Conclusions**

Background

- Back in April last year, the PM called a Snap Election
  - Much to my despair, at the time, polls were giving the Tories on 43% (up from 37% at the 2015 General Election), with Labour on 25% (down from 30%)

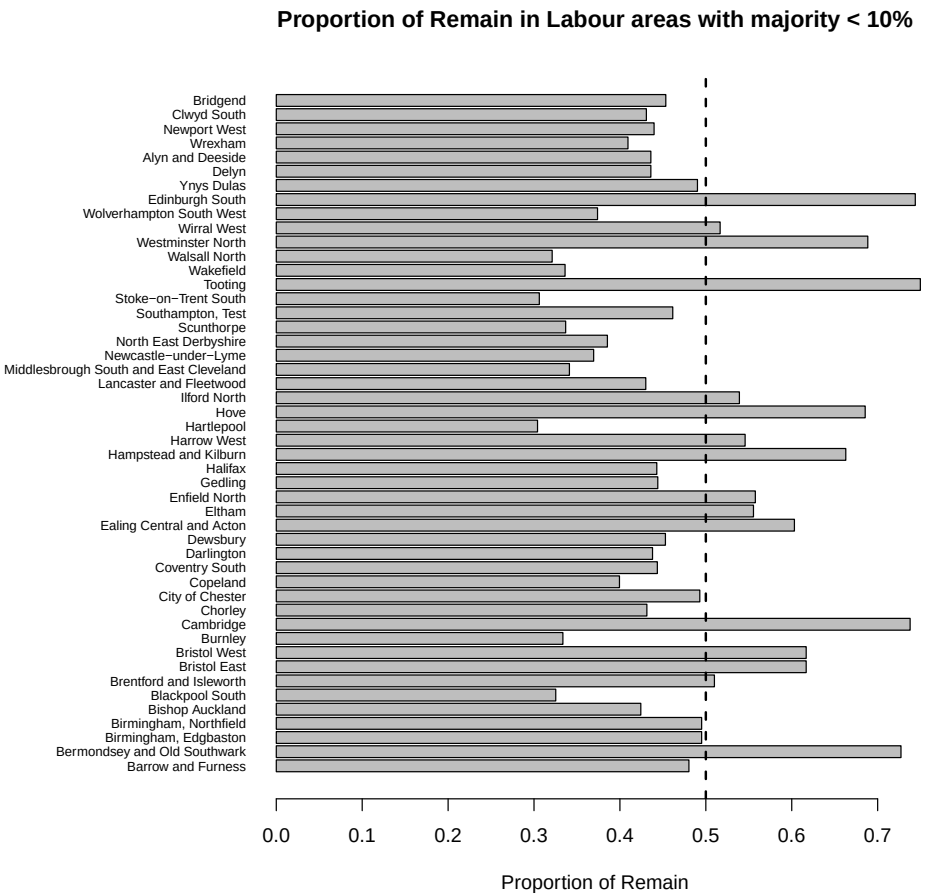
- **At the time**, Brexit was also very confused and confusing and it wasn't really clear what was going on...
  - But, by all commentators' accounts, the election would be mainly driven by Brexit

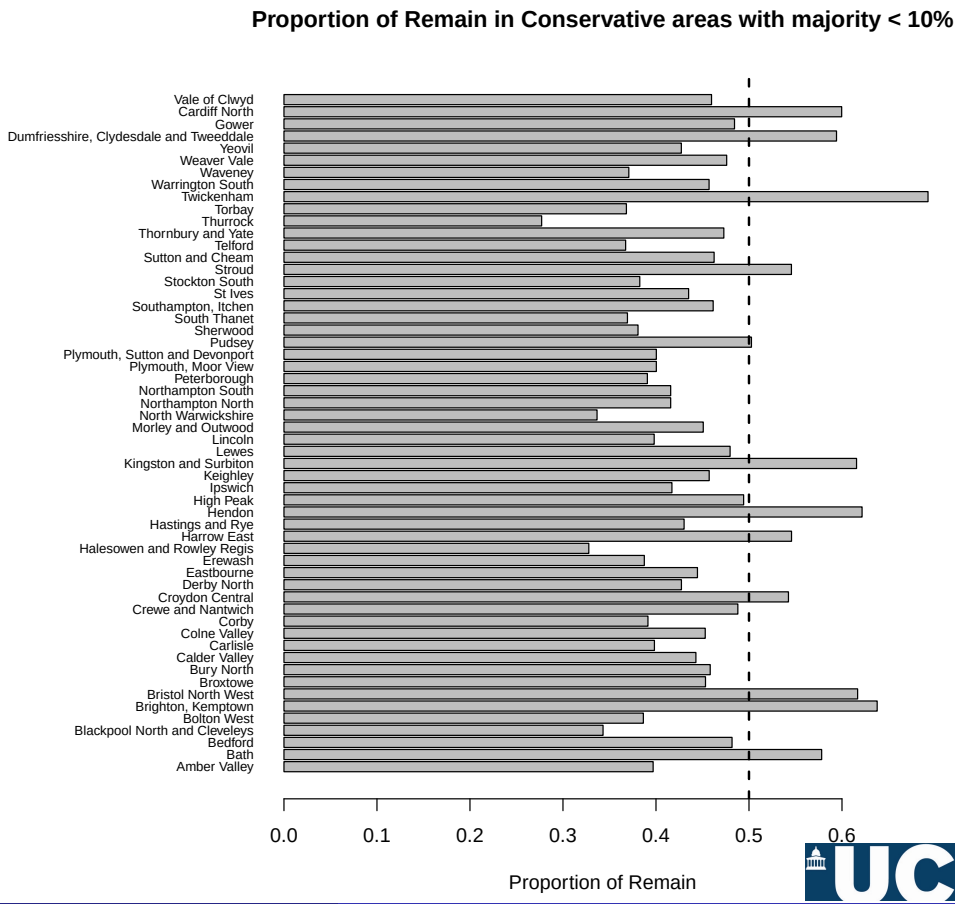


Background

- Back in April last year, the PM called a Snap Election
  - Much to my despair, at the time, polls were giving the Tories on 43% (up from 37% at the 2015 General Election), with Labour on 25% (down from 30%)

- **At the time**, Brexit was also very confused and confusing and it wasn't really clear what was going on...
  - But, by all commentators' accounts, the election would be mainly driven by Brexit





Modelling

Data

- Rolling opinion polls (voting intention), by several providers (YouGov, ICM, Opinium, ...)
- For each poll i, the data are $\mathcal{D}_i = (\boldsymbol{y}_i, n_i, t_i)$ with:
  - $\boldsymbol{y}_i = (y_{i1}, \ldots, y_{iP})$ = number of people reporting intention to vote for one of the main parties $p = 1, \ldots, P(= 8)$ in poll i
  - n_i = total sample size for the $i-$th poll
  - t_i = time (day the $i-$th poll is released)
- **NB**: Voting intention available **by Brexit vote**: $\boldsymbol{y}_i = (\boldsymbol{y}_i^L, \boldsymbol{y}_i^R)$
  - Effectively, consider data $\mathcal{D}_i^j = (\boldsymbol{y}_i^j, n_i^j, t_i)$, for $j = 0(= L), 1(= R)$

Model

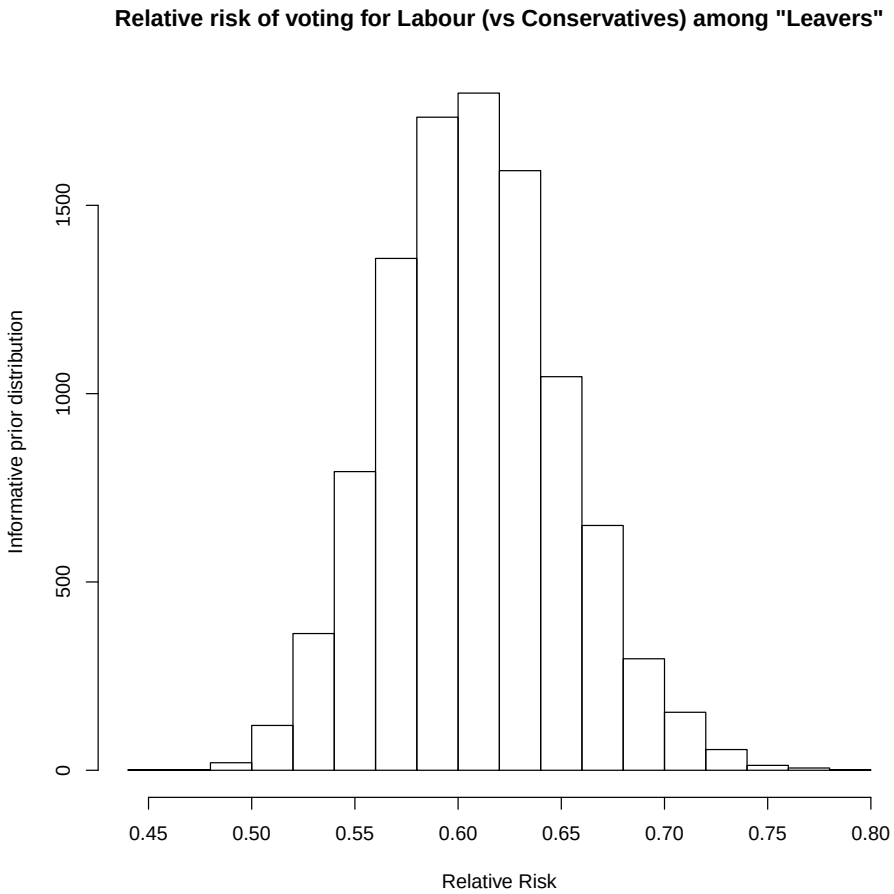
- $\boldsymbol{y}_i^j \sim \text{Multinomial}(\boldsymbol{\pi}^j, n_i^j)$
  - $\boldsymbol{\pi}^j = (\pi_1^j, \ldots \pi_P^j)$ = vector of overall proportion of votes among "$j-$ers", pooled across polls
- $\pi_p^j = \dfrac{\phi_p^j}{\sum_{p=1}^P \phi_p^j}$ and $\log(\phi_p^j) = \alpha_p + \beta_p j$
  - α_p = party-specific "Leave" effect (baseline)
  - $\alpha_p + \beta_p$ = party-specific "Remain" effect (incremental/decremental wrt "Leave")

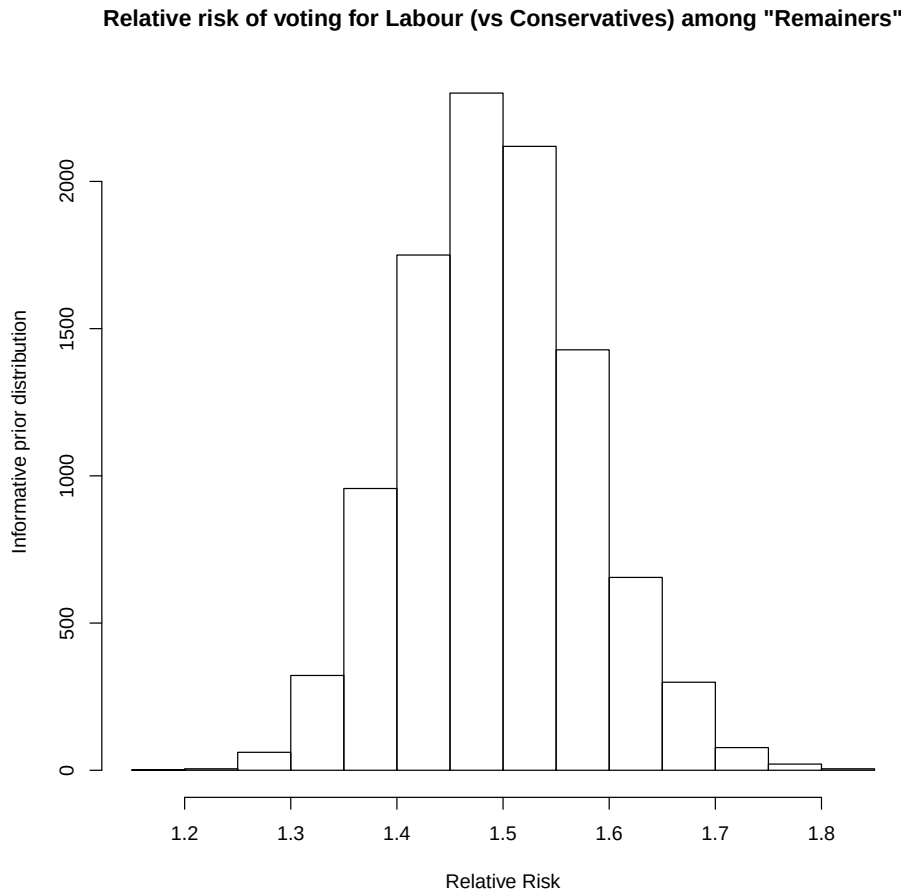
- Identifiability constraints: $\alpha_1 = \beta_1 = 0$
  - Essentially fixes the effect for a reference party to 0
  - I chose the Tories — psychologically, it felt comforting...

- $\alpha_2, \ldots, \alpha_P \sim \text{Normal}(0, \sigma_\alpha)$ and $\beta_2, \ldots, \beta_P \sim \text{Normal}(0, \sigma_\beta)$
  - Model the effect for the other parties wrt the baseline
  - Negative values for α_p indicate that party $p \neq 1$ is less likely to grab votes among Leavers than the Tories
  - Similarly, positive values for β_p mean that party $p \neq 1$ is more popular than the Tories among Remainers

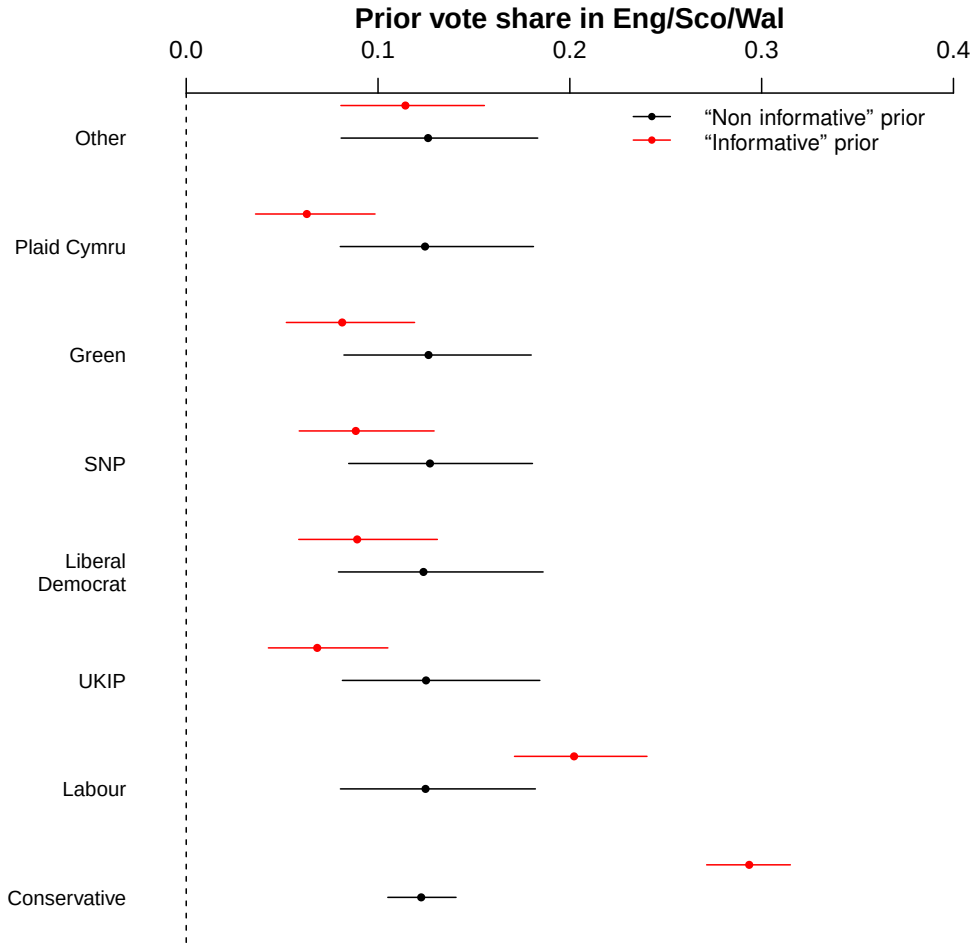
- Consider relatively informative priors by defining $\sigma_\alpha = \sigma_\beta = \log(1.5)$
  - This implies that it is unlikely to observe massive deviations (NB: α_p, β_p are on the log scale!)

- The problem with this model structure is that, effectively, it implies that the prior vote share is approximately $1/P$ for each party ("vague", but hardly reasonable!)
- Use informative prior on the log RR of voting for p vs baseline
  - Combine (weighted 3:2:1) past elections + subjective opinion





- The problem with this model structure is that, effectively, it implies that the prior vote share is approximately $1/P$ for each party ("vague", but hardly reasonable!)
- Use informative prior on the log RR of voting for p vs baseline
  - Combine (weighted 3:2:1) past elections + subjective opinion



- A nice feature of Bayesian modelling (eg based on MCMC simulations) is that we can retrieve a full (posterior) probability distribution **for any function of the basic parameters**
- Can use the estimated party- and Brexit-specific probabilities to compute a "relative risk" wrt the observed overall vote at the 2015 election

$$\rho_p^j = \frac{\pi_p^j}{\pi_p^{15}}$$

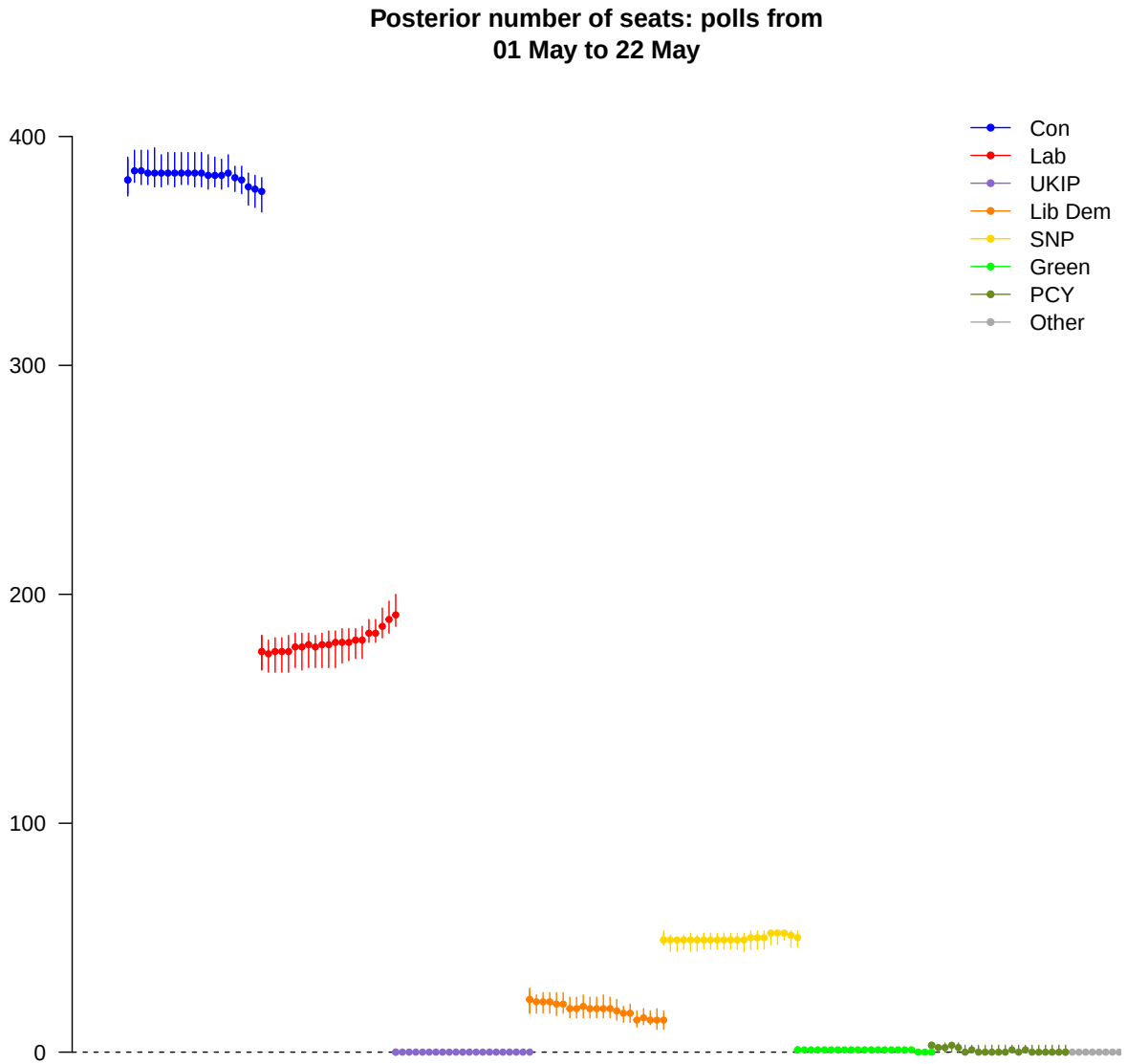
  - Estimates how better/worse parties are doing wrt the last election among "$j-$ers"

- Then can distribute the information from the current polls and the EU referendum to each constituency $c = 1, \ldots, C$ by estimating the predicted share of votes at the next election as the mixture

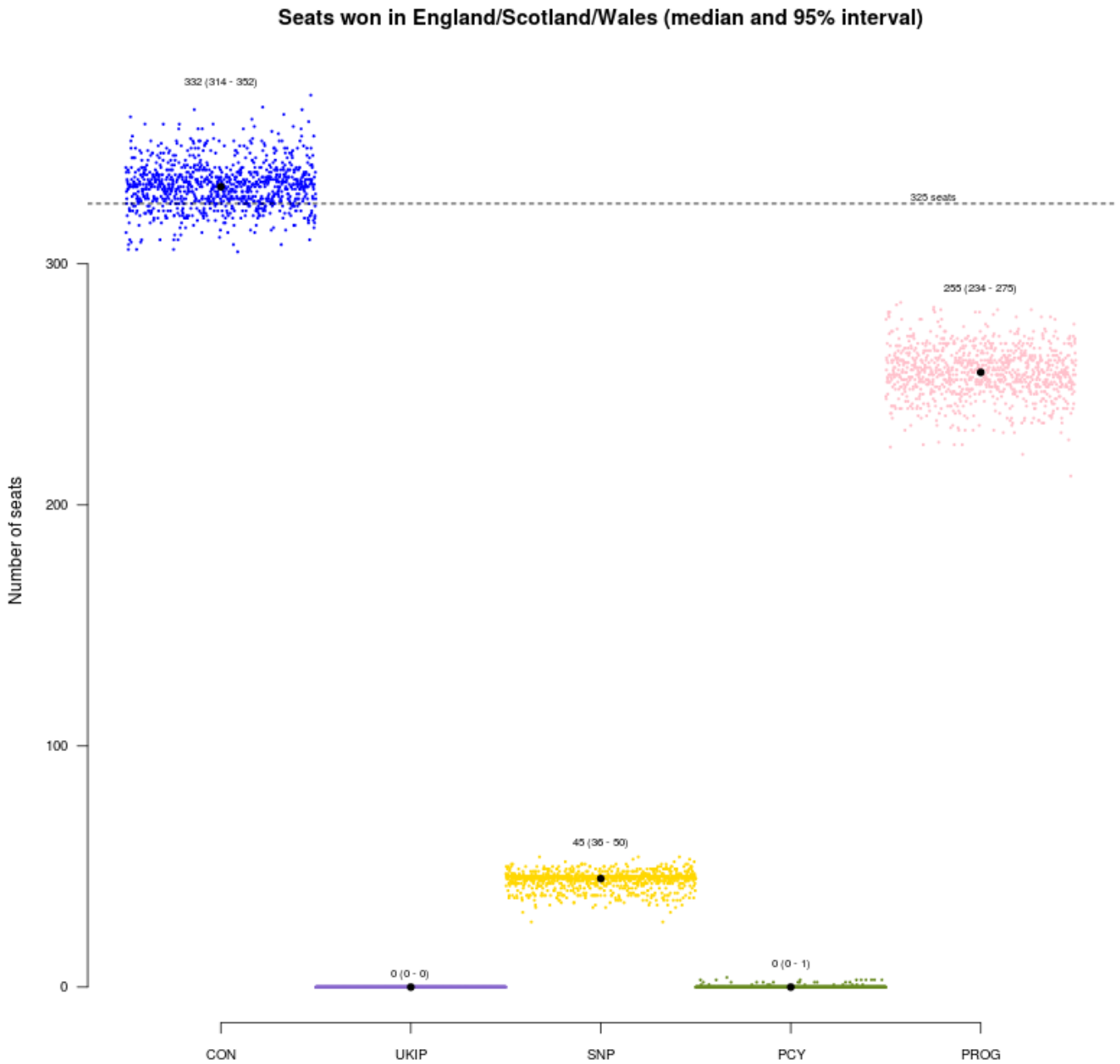
$$\pi_{cp}^{17} = (1 - \gamma_c)\pi_p^{15}\rho_p^L + \gamma_c \pi_p^{15}\rho_p^R$$

  - Can use the full posterior distribution of π_{cp}^{17} to simulate the General Election, given the current level of information provided by the polls (up to a given day)
  - Quick & dirty discounting of data $\tilde{y}_{ip}^j = \frac{y_{ip}^j}{(1+\delta)^\tau}$ (**NB**: Can do **much** better than this!)

Through **time** & space

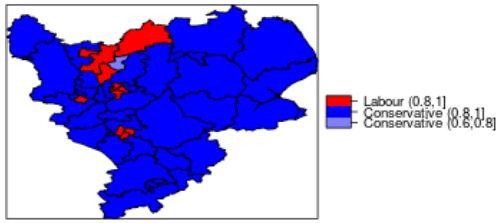


Through **time** & space

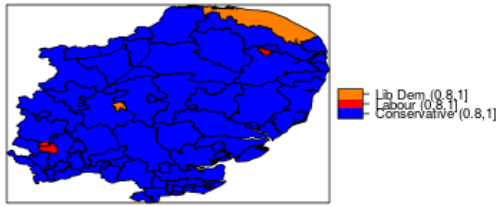


Through time & **space**

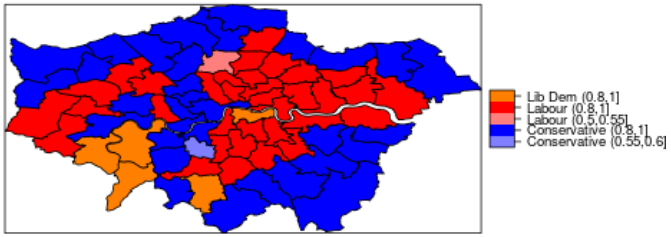
East Midlands



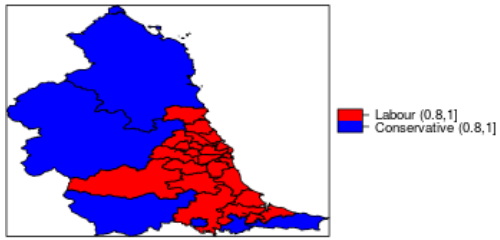
East of England



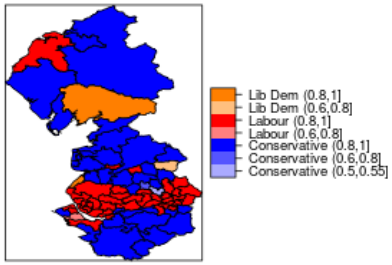
London



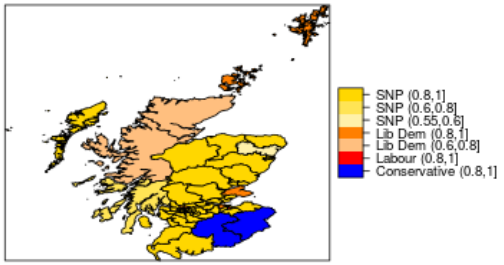
North East



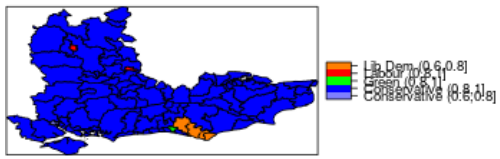
North West



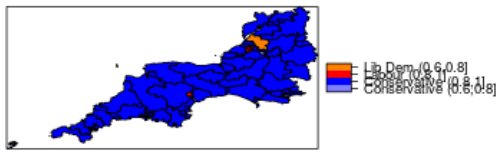
Scotland



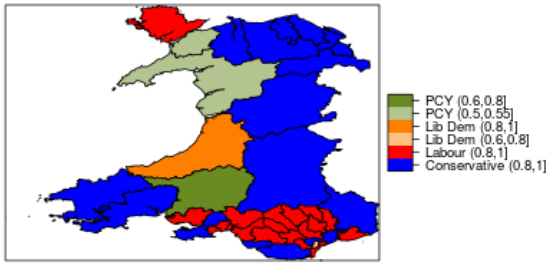
South East



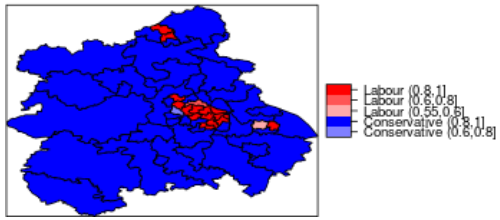
South West



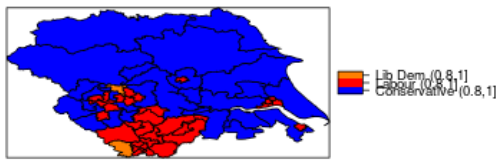
Wales



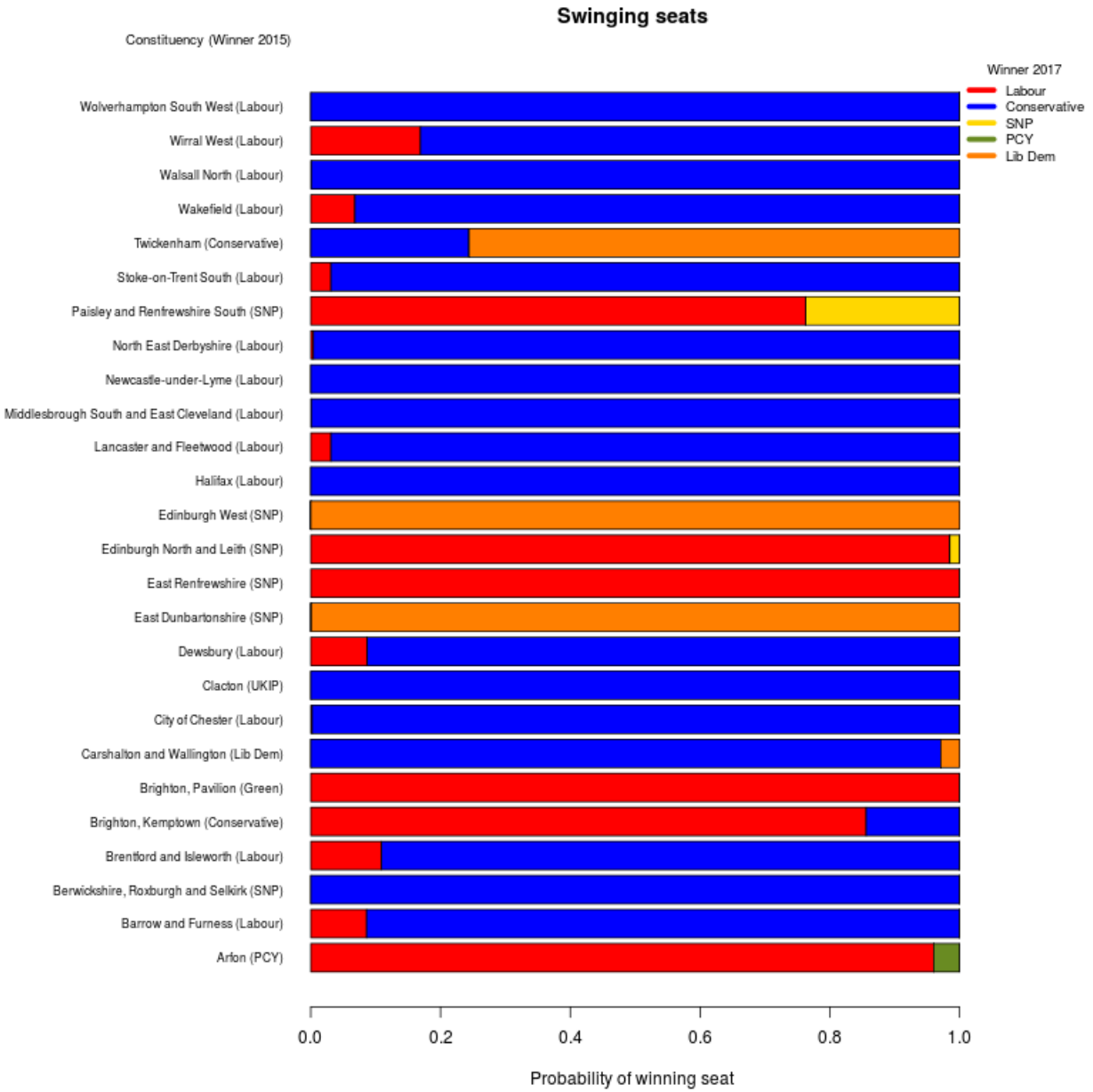
West Midlands



Yorkshire and The Humber



Through time & **space**



♪ There's a new (Bayesian) man in town...♪

- A couple of days before the election, YouGov published the results of a more sophisticated (Bayesian!) model
  - Based on **much** bigger sample size — not from aggregated polls, but more structured survey (over 50,000 responses)
  - Including more information **at the individual level** (age, sex, education, ...)
  - Re-scaled the results to match population-based statistics in the constituencies ("Multilevel Regression and Post-stratification", MRP)
- Surprisingly (?), Yougov's model produced a rather different result from most of the other poll analyses
  - Much worse performance for the Tories, who were estimated to gain only 304 seats (95% credible interval: 265-342 — even worse than the last general election)
  - Labour estimated to get 266 (230-300) seats
  - Suggestion of hung parliament

- My model was somewhere in between the "mainstream" poll analyses (which predicted a Tory clear victory) and Yougov's MRP model
  - But the comparisons weren't super fair/meaningful, as the underlying data were substantially different...

Postmortem

Postmortem

Party	Mean	SD	2.5%	Median	97.5%	**Observed**
Conservative	346.827	3.411262	339	347	354	**318**
Labour	224.128	3.414861	218	224	233	**261**
UKIP	0.000	0.000000	0	0	0	**0**
Lib Dem	10.833	2.325622	7	11	15	**12**
SNP	49.085	1.842599	45	49	51	**35**
Green	0.000	0.000000	0	0	0	**1**
PCY	1.127	1.013853	0	2	3	**4**

1. Everybody (including YouGov's MRP) overestimated the seats won by the SNP
   - I used National polls, so not necessarily representative of Scotland — "over-enthusiastic" prior to counter that may have exaggerated the results
2. Aggregate polls data indicated that in leave areas the Tories would have had massive gains, but in fact the former UKIP vote has split nearly evenly between the two major parties
   - In strong leave areas, the Tories have gained marginally more than Labour, but in reality that was not enough to swing and win the marginal Labour seats.
   - Conversely, in remain areas, Labour has done really well (as the polls were suggesting)
3. I missed the Greens' success in Brighton...
   - Mainly down to being a bit lazy and not telling the model that in Caroline Lucas' seat the Lib Dem had not fielded a candidate
   - The model was predicting a big surge in the vote for the Lib Dems (as Brighton Pavilion is a strong remain area), which would eat into the Green's majority and make Labour win
4. ...But correctly predicted the Tories would regain Richmond Park, the Lib Dems Twickenham and Labour Copeland
   - The sort of "skeptical" prior on the Lib Dems managed to discount some of the polls, which seemed to be overly-optimistic

Room for improvement...

1. Could tell the model the whole story...
   - No real account for "tactical" votes, or more or less formal "alliances"
   - Spatial structure in past vote?

2. Could include more predictors in modelling the polls
   - Brexit vote probably a proxy for several composite "profiles" (education, age, ethnic mix...) — so could try to re-balance when estimating the constituency-level probabilities
   - Majority at the last election often a good indicator — massive swings unlikely (but only indirectly accounted for)

3. Eventually, the "young vote" made a massive difference (mostly to Labour)
   - Some information was available on the polls by age group too
   - Probably would have required a much more complex modelling structure...

Conclusions

- Bayesian modelling is particularly suitable for poll data
  - We don't always want to take the polls at face value
  - Including prior information may counter-balance some of the bias intrinsic in surveys (especially about political opinions)

- The nature of Bayesian modelling allows to obtain a full characterisation of uncertainty
  - About the main model parameters
  - **And** about any other related quantity of interest

- The idea of bringing together different sources of information in a formal way is central to the Bayesian paradigm and can be successfully applied to political analysis
  - MRP is a brilliant example — and allows to go beyond aggregated polls data, given suitable information at the individual level

https://www.significancemagazine.com/politics/551-predicting-the-uk-s-snap-general-election
http://www.statistica.it/gianluca/blog

SIGNIFICANCE

Business Culture Politics Science Sports search...

Predicting the UK's snap general election

Written by Gianluca Baio on 30 May 2017.

I have decided to build a model to try to predict the results of the upcoming snap general election in the UK. I'm sure there will be many people attempting this, from various perspectives and using different modelling approaches. But I have set out to develop a fairly simple (though, hopefully, reasonable) model. In the process of describing this to you, I hope to shed some light on how statisticians build predictive models.

We start with the data, which come from national voting intention polls conducted by a number of research agencies, including YouGov, ICM and Opinium.

Arguably, this election will be mostly about Brexit; there surely will be other factors, but because this election comes almost exactly a year after Britain's referendum on membership of the European Union (EU), it is a fair bet to suggest that how people voted back then, and how they still feel now, will massively influence the election result.

Luckily, all the polls I have found report data in terms of voting intention, broken down by Remain/Leave votes during the EU referendum. Using these polls, I am looking to predict the results of the seven main political parties – Conservatives, Labour, UKIP, Liberal Democrats, SNP, Green, and Plaid Cymru – plus all "Others". Also, for simplicity, I'm considering only the results for the 632 constituencies in England, Scotland and Wales, not the 18 Northern Ireland constituencies. This shouldn't be a big problem though, as elections are generally a local affair in Northern Ireland, with the mainstream British parties not playing a significant role.

As well as recent polling data, I also have available data on the results of both the 2015 general election and the 2016 EU referendum. I had to do some work to align these two datasets, as the referendum did not use the same geographical resolution as is typically used during general elections. I therefore mapped the voting areas used in 2016 to the parliamentary constituencies and have recorded the proportion of votes won by my eight parties in 2015, as well as the proportion of Remain votes in 2016.

For each constituency, I therefore have a distribution of election results, which I can use to determine the average outcome, as well as various measures of uncertainty. In a nutshell, my model is all about (a) re-proportioning the 2015 and 2017 votes based on the polls; and (b) propagating uncertainty in the various inputs.

But before I can start making (what I hope will be) reliable predictions, I need to tune the model a little more.

Priors and posteriors

My model is built using a Bayesian approach, one feature of which is that the model starts with a "prior" distribution for each party's vote share, and this prior is then updated based on the results of the polling data to create a "posterior" distribution. This posterior is the model's prediction of a party's likely

Related articles

- 6 Nations Rugby – who's the biggest overachiever?
- Belgium to win Euro 2016? A Q&A on probabilistic predictions
- The 2016 Prediction Games – Part I
- The 2016 Prediction Games – Part II
- A smashing return: Team GB's Olympic host legacy

Latest issue

April 2017
(Vol 14 Issue 2)

From the print edition

- Fairness and transparency in the age of the algorithm
- The story of the Flint water crisis
- In search of Earth analogues
- The "Ferguson effect", or too many guns?

Thank you!