Combining national and constituency polling for forecasting

Chris Hanretty, Ben Lauderdale, Nick Vivyan

Abstract

We describe a method for forecasting British general elections by combining national and constituency polling. We reconcile national and constituency estimates through a new swing model.

1. Introduction

This note sets out a method for forecasting the 2015 British general election based on national and constituency polling data. It comprises three steps: a model for forecasting national public opinion, a model for current constituency public opinon, and a method of reconciling these two sets of estimates through a new swing model. With only minor changes, this method has been used to make daily forecasts of the election outcome from September 2014 onwards. These forecasts have been published online at `www.electionforecast.co.uk`.[1] This note provides our forecast from the morning of the election.

2. National model

We begin by estimating current vote intention for the seven main parties (Conservative, Labour, Liberal Democrats, SNP, Plaid Cymru, the Greens, and UKIP) and all other parties combined. To do so, we use all publicly available national[2] polls published from May 2014. Where possible, we use information on the weighted number of respondents intending to vote for each party, rather than the percentages reported. Like other contributors to this volume, we combine these polls using a state space model estimated using Markov Chain Monte Carlo methods (Jackman 2005), except that in order to account for the compositional nature of this data, we perform an additive log ratio transform (Aitchison 1986) on the vote shares of all parties save the reference party.[3] With this model we recover estimates of party support that sum to 100%, together with the associated posterior distributions.

Specifically, where y_i is a vector of length 8 which stores the (weighted) number of respondents to poll i intending to vote for each party, and where n_i is the (weighted) number of respondents in each poll ($n_i = \sum y_i$) we model the outcome of each poll as

$$y_i \sim Multinom(\mu_i, n_i)$$

where the probability of voting for each party μ_{ij} is modelled as follows:

$$log(\mu_{ij}) = \delta_{jh_i} + \alpha_{jt_i}$$

where t indexes time from a year before the election ($t = 1...365$), h indexes polling companies,[4] δ_{jh} reflects the house effect for party j, and where α_{jt_i} represents the latent support for party j. That level of latent support can in turn be modelled as a random walk for all parties save the reference party:

$$\alpha_{jt} \sim N(\alpha_{jt-1}, \omega^2), j = 2...8, t = 2, ...365$$

For the reference party, $\alpha_{1t} = 0$ for all values of t. Initial relative values of party support for all other parties are drawn from a diffuse uniform prior bounded between -10 and +10 on this log ratio scale. In order to transform party support back to

[1] Here we describe our predictions for the 632 mainland constituencies. Our website includes predictions for Northern Irish seats, but these are derived from a very different model.

[2] By "national" polls, we mean polls that cover Great Britain, but not Northern Ireland.

[3] We use the Conservative party as our reference party. The choice of reference party does not affect the results.

[4] More accurately, h indexes combinations of polling companies and methodologies, such that a polling company which changes its methodology is akin to a "new" polling company.

national support for a party as a proportion (V_{jt}), we calculate

$$V_{jt} = \frac{e^{\alpha_{jt}}}{\sum_{j=1}^{8} e^{\alpha_{jt}}}$$

In this way, we ensure that our estimates of current party support sum to one. House effects in this model are identified by ensuring that the house effects of "active" polling company-methodology combinations have mean zero for each party.[ˆactive]

Using the same state-space model, we have also estimated party support for three main parties (Conservative, Labour and Liberal Democrat, or predecessor parties) for the eight elections from 1979 onwards, from a year before the election.[5] We use these estimates to calculate how much levels of support for a party will revert back to their performance in the previous election. Specifically, for each of the 365 days preceding the election ($t = 1...365$), and combining data across parties and elections (such that $N = 3 \times 8 = 24$ for each day), we estimate the following regression equation:

$$\text{Observed swing} = \gamma_t(\text{Poll implied swing}) + \epsilon_t$$

$$\epsilon_t \sim N(0, \sigma_t)$$

recovering 365 values of γ, which we then smooth and store as $\widetilde{\gamma}$.[6] We also store each of the 365 values of σ, and also smooth these.

The benefit of this model – which can be described as a "change on change" model, and which has no intercept – is that it can be applied to any party, even parties for which we lack historical polling information (for example, UKIP), and that, because it treats parties equivalently, it ensures that vote shares continue to sum to 100%. Additionally, the parameter γ can be interpreted quite naturally as the weight to place on poll-implied vote shifts. Values of γ increase from about 0.45 to 0.80 over the year before the election.

Because γ is always less than one, our model suggests that, before the election, parties which are polling badly in the run up to the election (compared to the previous election) tend to recover some of the support they have lost. Conversely, parties which are polling well in the run up to the election lose some of what they have gained. Some of this "swingback" and "fallback" occurs before the election, but because the maximum value of γ is less than one, our model assumes that some of it occurs between the final polls and the election result.

On any given day t, a tentative forecast for each party's national vote share, $\hat{V_j}$, can thus be approximated by the following equation:

$$\hat{V_j} = \text{Previous vote share} + \widetilde{\gamma_t}(\text{Poll implied swing}_t)$$

Our actual forecast is more complicated, because we must incorporate not just uncertainty surrounding the poll-implied swing, but also the uncertainty present in the relationship between polling and outcomes captured in σ. At the same time, we must restrict our estimates to fall between 0 and 100%.

We therefore forecast vote shares for each party by drawing them from a beta distribution with parameters a and b, which are defined based upon the unconstrained forecast given above ($\hat{V}$) and the stored and smoothed values of σ:[7]

$$V_j \sim Beta(a, b)$$

$$a = \frac{\hat{V}(\hat{V} - \hat{V}^2 - \sigma\frac{\hat{V}(1-\hat{V})}{0.21})}{\sigma(\frac{\hat{V}(1-\hat{V})}{0.21})}$$

$$b = \frac{(1-\hat{V})(\hat{V} - \hat{V}^2 - \sigma\frac{\hat{V}(1-\hat{V})}{0.21})}{\sigma(\frac{\hat{V}(1-\hat{V})}{0.21})}$$

3. Constituency model

The national model provided us with a forecast of the national share of the vote won by each party. In this section, we describe a model for estimating current constituency opinion. In the section that follows, we describe how to reconcile these estimates of current constituency opinion with our forecast national vote share.

We begin by describing our data. We use data from 187 published constituency polls. The vast majority of these (169) were commissioned by Lord Ashcroft.

[5] We start in 1979 because previous research has suggested that this election represents a break (Fisher 2014), and because the polling record for the previous October 1974 election is truncated by the February 1974 election.

[6] Specifically, we fit a local linear regression with a window around date of poll t that runs from $1.5t - 10$ to $0.5t + 10$.

[7] The intuition here is to approximate well vote shares which are normally distributed with a mean of 30%.

We also use data from YouGov national samples. We have information on the constituency and 2010 vote of each respondent, as well as limited demographic information. We reweight these constituency samples to match constituency characteristics as reported by the 2011 Census. Specifically, we reweight on the basis of gender, age group, highest educational qualification, social grade and 2010 vote. We then use the implied sample shares to create a pseudo-sample of the size implied by Kish's effective sample size formula (Kish 1965).

On average, the information from these constituency polls and subsamples is 68 days old.

[Table 1 about here.]

Because many of the constituency-specific polls ask respondents about their vote intention under two different prompts – one generic ("If there was a general election tomorrow, which party would you vote for?"), one constituency-specific ("Thinking specifically about your own *parliamentary* constituency at the next General Election and the candidates who are likely to stand *for election to Westminster* there, which party's candidate do you think you will vote for in your own constituency?") – we can investigate the relationship between party support under these two conditions. Party-specific regressions relating these different levels of support are shown in Table 1 for the three main parties only.

With this data, and an idea of the relationship between generic and specific support, we can construct our dependent variable, y_i, a vector of length eight which stores the (weighted) number of poll respondents intending to vote for each party in each constituency. Subscript i indexes unique combinations of constituency and polling company (i=1... 818). As before, let n_i stand for the number of respondents in each row of y.

$$y_i \sim Multinom(p_i, n_i)$$

The probabilities of respondents in each constituency (sub)sample voting for each party j can be modelled as follows:

$$p_{ij} = g_i\alpha_j + g_i\beta_j\frac{\pi_{kj}}{\sum\pi_k} + (1-g_i)\frac{\pi_{kj}}{\sum\pi_k}$$

where g_i is an indicator which has the value 1 if the poll used a "generic" prompt rather than a constituency-specific prompt, and where α and β are the intercept and slope of a regression of vote intention given a generic prompt against vote intention under a constituency-specific prompt (as plotted in Table 1). π_{ij} is in turn modelled as a function of μ_{jc_i}, or today's latent level of support for party j in constituency c, plus house effects specific to house h δ_{jh_i}, minus a "shift", λ_{jt_i}. That shift is equal to the change in the log-ratio of national support for party j, relative to the reference party, between the day of the poll i and the current day.[8]

$$log(\pi_{ij}) = \mu_{jc_i} + \delta_{jh_i} - \lambda_{jt_i}$$

Constituency vote shares are modelled as draws from a normal distribution with mean equal to a linear function of logit-transformed past vote shares (v_{jc}^{2010}) of all parties and explanatory variables X_{jc} which are all measured at the level of constituency c.

$$\mu_{jc} \sim N(\alpha_j v_{jc}^{2010} + X_{jc}\beta_j, \sigma_j^2)$$
$$\sigma_j \sim Unif(0,1)$$

The explanatory variables used include:

- political variables (logit-transformed vote share of party j in the European Parliament elections of 2014; logit-transfomed vote share of party j in the most recent local authority elections, and dummy variables recording whether party j currently holds the seat, whether party j's MP is standing down, whether party j's incumbent MP is a first-term MP).[9]
- geographic variables (the government operating region)
- demographic variables taken from the 2011 Census (average highest level of education on a seven-point scale; average NRSM social grade scored one to four; average age in years; the percentage of residents who are Christian, of no religion, of another non-Christian religion; the percentage of residents who are female, married, own their own home, and who are in the private sector) and from the 2013 Annual Survey of hours and Earnings (log of median earnings in pounds).

[8] In practice, this means that we assume a uniform national swing in the log-ratio transform of party vote shares between the day of a constituency poll and the day we generate our forecasts. We return to this issue in the following section.

[9] Vote shares obtained in different geographic areas have been mapped onto Westminster constituency boundaries in proportion to area.

- public opinion variables: the estimated proportion of respondents in each constituency who support British exit from the European Union

Most multinomial logistic regression models (of which this is a variant) are identifed by constraining the coefficients for the reference outcome category to zero. Here, we identify the model through tight priors on α.

At each new release of Ashcroft polls, we have compared our estimates from this model to the new polling data, finding that our estimates are only modestly overconfident once we take into account poll and model uncertainty. In order to extract estimates of constituency support (v_{jc}) from this model, we calculate:

$$v_{jc} = \frac{e^{\mu_{jc}}}{\sum_{j=1}^{8} e^{\mu}_{jc}}$$

4. Reconciliation

In order to produce a forecast of constituency vote shares, we must combine our forecast of election-day national vote shares with our estimates of current constituency vote shares.

One way of combining these two sets of estimates is to calculate, for each party, the difference between the party's national vote share at the time the constituency votes shares were estimated, and the forecast national vote share, and to then add on this difference to the estimated vote share in each constituency. This re-creates the logic of uniform national swing (UNS), except that instead of adding on (subtracting) a uniform national swing from past constituency *results*, we add on (subtract) a uniform national swing from constituency *estimates*. This also recreates the problems of UNS, in that it leads to negative vote shares, particularly when making predictions for all other parties and parties with low estimated vote share, which in turn creates the potential for inconsistency between national and constituency estimates.

We therefore create a new swing model which satisfies the constraint that constituency estimates must, when multiplied by constituencies' share of the voting population T_c (which is a result both of the eligible population and the rate of turnout), sum up to national estimates. To do so, we assume that the relative rates of turnout across constituencies stay as they were in 2010.

Let us begin by paraphrasing the naive approach in a more formal way which begins to take account of differential turnout. Our problem is to find the value of x (i.e., the right "uniform swing") which minimizes the following function:

$$f(x) = \frac{T_c}{\sum T_c}(v_{jc} + x) - V_j$$

where $f(x) = 0$ means that our two estimates are perfectly reconciled. Because, in this statement of the problem, x is always and everywhere a uniform shift, the problem of non-negative vote shares arises.

In order to avoid this issue, we can re-state the formula above by *transforming* the vote shares using a further function $G(.)$.

$$f(x) = \frac{T_c}{\sum T_c} G^{-1}(G(v_{jc}) + x) - V_j$$

[Figure 1 about here.]

$G()$ can be any invertible sigmoidal function which transforms real numbers into numbers in the range (0,1). Figure 1 shows three such functions, and how they deal with swings of 5 and 20% respectively. The top panel shows the effect of these swings under uniform national swing (i.e., the identity function). The middle panel shows what happens if we use the logistic function, in which case instead of adding on a value of x measured in percentage points, we add on a value of x measured in logits. After experimenting with a number of functions, we have opted to use the cumulative distribution function of the generalized normal distribution, which has additional parameters α (scale) and β (shape). We set $\alpha = 1$ and $\beta = 10$. This is shown in the bottom panel of Figure 1.

With this function, we then optimize to find, for each party, the value of x which minimizes the above function, ensuring the closest possible match between our constituency estimates and our national forecasts.

5. Forecasts

Table 2 gives our forecasts of vote shares and seat counts, along with the respective 90% credible intervals.

[Table 2 about here.]

Note that the vote shares reported are predictions of the vote shares won by parties considering votes cast in Great Britain only, and therefore excluding Northern Ireland.

Because we use Bayesian methods to generate the national forecast, the constituency current estimates, and apply the reconciliation on an iteration-by-iteration basis, we can also calculate probabilities of arbitrary events. Thus, the probability that the Conservatives will be the largest party in terms of seats is 63.2%, but in 2000 simulations Conservatives had a majority on 0 occasions.

Indeed, in no simulation run did either party win a majority of 326 seats or more, and it is likely (41%) that no two parties *combined* (short of a grand coalition) will be able to command 326 seats. Thus, although it is extremely difficult to forecast which party will be the largest party – something which might be thought to be an important desideratum of any forecasting model – we can be relatively confident that the eventual outcome is likely to be "messy".

6. References

Aitchison, John. 1986. *The Statistical Analysis of Compositional Data*. London: Chapman; Hall.

Fisher, Stephen D. 2014. "Predictable and Unpredictable Changes in Party Support: A Method for Long-Range Daily Election Forecasting from Opinion Polls." *Journal of Elections, Public Opinion & Parties* (ahead-of-print): 1–22.

Jackman, Simon. 2005. "Pooling the Polls over an Election Campaign." *Australian Journal of Political Science* 40(4): 499–517.

Kish, Leslie. 1965. *Survey Sampling*. John Wiley; Sons.

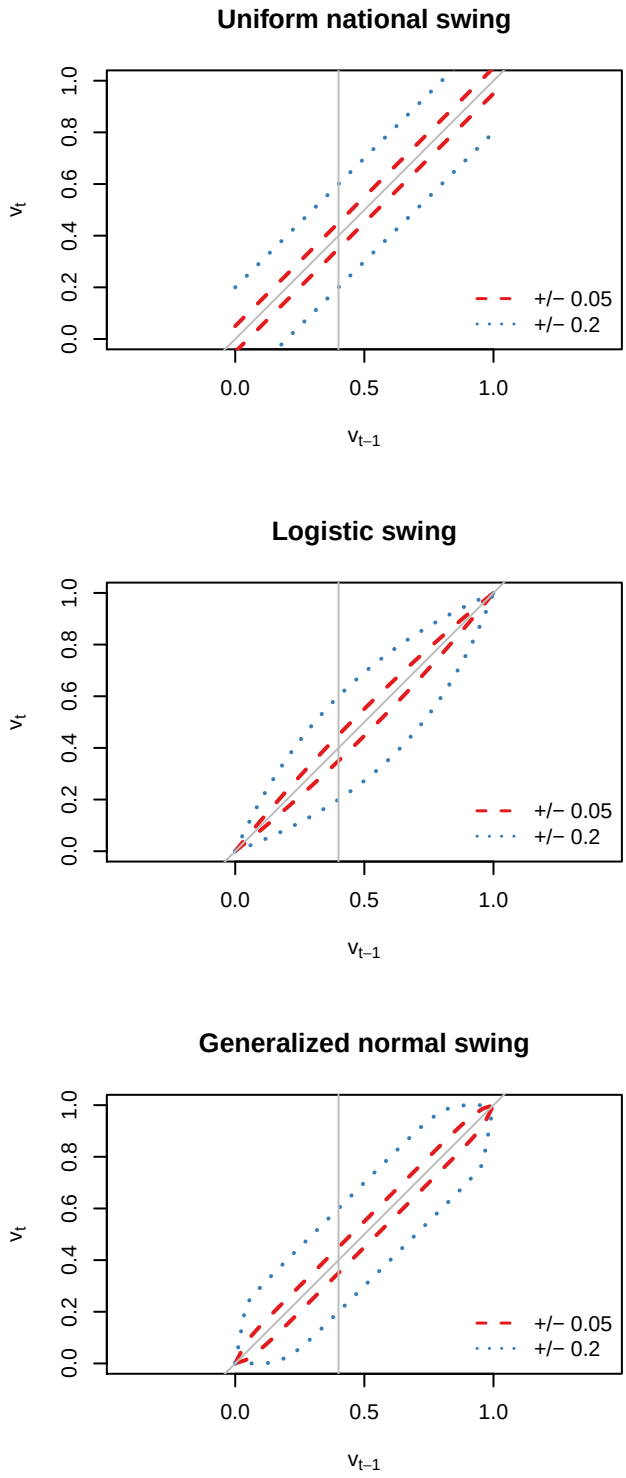

Figure 1: Different swing functions

Table 1: Specific v generic support

	Dependent variable:		
	Con (spec.)	Lab (spec.)	LD (spec.)
	(1)	(2)	(3)
Con (generic)	1.048***		
	(0.019)		
Lab (generic)		1.117***	
		(0.015)	
LDem (generic)			1.644***
			(0.024)
Constant	−0.027***	−0.041***	−0.016***
	(0.006)	(0.005)	(0.003)
Observations	225	225	225
R^2	0.931	0.960	0.955
Adjusted R^2	0.931	0.960	0.955
Residual Std. Error (df = 223)	0.028	0.025	0.027
F Statistic (df = 1; 223)	3,017.000***	5,353.000***	4,702.000***
Note:		*p<0.1; **p<0.05; ***p<0.01	

Party	Mean	Lo	Hi	Mean	Lo	Hi
Conservatives	34.4	31.8	37.1	278	252	305
Labour	32.8	30.0	35.6	267	240	293
Liberal Democrats	11.7	9.8	13.9	27	21	33
SNP	4.0	3.5	4.5	53	47	57
Plaid Cymru	0.6	0.5	0.7	4	2	6
Greens	4.1	2.9	5.5	1	0	1
UKIP	10.6	8.7	12.6	1	0	2
Other	1.7	0.9	2.7	1	1	1

Table 2: Forecast GB vote and seat shares