Forecasting multiparty by-elections using Dirichlet regression

Abstract

By-elections, or special elections, play an important role in many democracies – but whilst there are multiple forecasting models for national elections, there are no such models for multiparty by-elections. Multiparty by-elections present particular analytic problems related to the compositional character of the data and structural zeros where parties fail to stand. I model party vote shares using Dirichlet regression, a technique suited for compositional data analysis. After identifying predictor variables from a broader set of candidate variables, I estimate a Dirichlet regression model using data from all post-war by-elections in the UK (n=468). The cross-validated error of the model is comparable to the error of costly and infrequent by-election polls (MAE: 4.0 compared to 3.6 for polls). The steps taken in the analysis are in principle applicable to any system which uses by-elections to fill legislative vacancies (7,258 words)

Keywords: Dirichlet regression, by-elections, special elections, election forecasting, compositional data, polling

1. Introduction

Where members of a legislature die, resign, or cannot for other reasons continue in their role, special procedures are used to replace them. The most popular procedure is to call a special election, or a by-election, to replace the departing legislator (Feigert and Norris 1990). By-elections are frequent in systems that use them (in the United Kingdom there are on average four per year) and can have broader consequences for the political system than simply replacing former legislators. They can herald the emergence of new parties, crystallize changes in voting behaviour, and wound incumbent governments (Cook and Ramsden 1973). By-elections can have these consequences because they are not just the translation to a particular local context of changes seen in national opinion polling (Price and Sanders 1998, 141), but rather combine substantial idiosyncratic elements and elements of the broader category of second-order elections (Reif, Schmitt, and Norris 1997).

Three examples from the past twelve years of British political history illustrate well the complexities involved in predicting by-election outcomes. In 2008, an incumbent Conservative politician resigned his seat in protest at the "erosion of civil liberties in Britain,"[1] only to contest the resulting by-election. The governing Labour party declined to field a candidate, regarding the by-election as a stunt; the opposition Liberal Democrats supported Davis' stance and similarly declined to field a candidate. With no major party alternatives, Davis was re-elected with 72% of the vote (an increase of 24 percentage points), with 25 other candidates splitting the remainder of the vote. Six years later, another Conservative MP, Douglas Carswell, resigned his seat only to contest the resulting by-election as a UKIP candidate. As a result, the UKIP share of the vote increased from a notional 0% in 2010, when they failed to field a candidate, to 60%, the greatest absolute change in vote share in British political history. The by-election demonstrated the strength of the UKIP challenge to the Conservatives, and presaged some of the elements of the Brexit campaign. The issue of Europe has also affected candidacy decisions: in 2019, the Liberal Democrats won a by-election in Brecon and Radnorshire thanks in large part to the decision of the Welsh nationalist party Plaid Cymru and the Green Party not to field candidates. In these contests, candidacy decisions, secondary dimensions of political competition and dramatic changes of vote share were all present. Although these are exceptional cases, these different elements are present to lesser degrees in many other by-elections both in Britain and in other countries which use such special elections.

Because by-elections can have dramatic effects on political systems, there is value in being able to forecast by-election outcomes. Unfortunately the literature on election forecasting has not considered by-elections, and the literature on by-elections is not suitable for forecasting multiparty outcomes. In this article, I

[1] Jenny Percival, Deborah Summers and agencies, "Tories in turmoil as David Davis resigns over 42-day vote", *The Guardian*, 12th June 2008, available online at https://www.theguardian.com/politics/2008/jun/12/daviddavis.conservatives

set out a Dirichlet regression model which is suitable for forecasting multiparty outcomes where all vote shares must sum to 100%. I estimate this model on data for all British by-elections in the post-1945 period. The accuracy of the model in leave-one-out (LOO) forecasts is within 10% of the accuracy of (rare, expensive) by-election polls.

I begin the article by describing the existing academic literature on by-elections (§2.1) and election forecasting (§2.2), and the literature on the analysis of compositional data (§2.3). I then describe the data I use (§3), providing summary statistics on the outcome variable (§3.2) and the candidate predictor variables (§3.3). I describe the procedure I use to select predictor variables from a longlist of candidate predictor variables (§4.1), and the Dirichlet regression itself (§4.2). I give details on model accuracy in absolute terms (§§5.1, 5.2) and in comparison with other methods of forecasting (§§5.4, 5.5). The penultimate section (§6) describes the performance of Dirichlet regression models when compared to models of log ratios. A short conclusion notes possible extensions.

2. Literature review

2.1. The literature on by-elections

Past scholarship on by-elections has determined how common they are as a means of filling legislative vacancies and how frequent they are in systems which use by-elections. Feigert and Norris (1990) found that 63 out of the 128 countries in their analysis always used by-elections to fill parliamentary vacancies, whilst a further 30 countries used by-elections for at least some parliamentary vacancies. The remaining countries used designated substitutes. Because it is easier to to fill legislative vacancies by using substitutes in systems which use party-list proportional representation, by-elections are disproportionately found in systems which use single member districts. The frequency of by-elections varies according to the size of the legislature and the frequency of general elections. In the UK, which has a large legislature (650 seats) and long maximum inter-election periods (5 years), there are on average four by-elections/year. Rates are lower for Canada (2.7/year: Loewen and Bastien (2010)) and Australia (1.2/year: author's calculations) but higher for the United States (4.7/year for the House for the period 2000-2020).

The literature on by-elections has focused on the four countries just mentioned (Australia, Canada, the United Kingdom, and the United States), except that studies of the US talk of "special elections". There have been both single country studies (Australia: Economou (1999); Canada: Loewen and Bastien (2010); United Kingdom: Mughan (1988) and Norris (1990); United States: Knotts and Ragusa (2016)) and comparative studies (Studlar and Sigelman 1987; Feigert and Norris 1990) which have identified common features of by-elections. By-elections have also been studied under the broader heading of second-order elections (Reif, Schmitt, and Norris 1997) or "barometer elections" (Anderson and Ward 1996), since turnout in these elections is generally lower than in general elections, and since often governing parties lose votes (cf. Feigert and Norris 1990).

Table 1: Findings from the literature on by-elections

Level	Factor	Effect	Discussed in
National	Opposition party vote share in national polls	Negative effect on governing party	Mughan (1988); Price and Sanders (1998)
	Government party majority in the legislature	Negative effect on governing party	Mughan (1988); Price and Sanders (1998)
	Governing party status	Negative effect on governing party	Feigert and Norris (1990); Studlar and Sigelman (1987); Norris and Feigert (1989)
	Inflation	Negative effect on governing party	Mughan (1988)
	Autumn/winter by-election	Negative effect on governing party mediated by turnout	Mughan (1988)
	Third party status	Positive effect on third parties	Norris and Feigert (1989)
Local	Candidacy decisions	Positive effects on parties who stand candidates; negative effects for all other parties	Studlar and Sigelman (1987)
	Nation	Conditions effect of other variables	Mughan (1988)
	Party vote share in preceding election	Positive effect on focal party	Norris and Feigert (1989)

Table 1 gives a stylised overview of the conclusions of this literature. Generally, governing parties do poorly, and some factors which operate at the national level (inflation, large legislative majorities which might represent the peak of the pendulum's swing) also operate at the local level. However, much of the research underpinning these claims is somewhat dated, and some of the factors authors have included in their models of by-election outcomes are not suitable for forecasting since they are only known at or after the result of the election is known (for example: by-election turnout, and inflation in the quarter of the election itself).

2.2. The election forecasting literature

There is a considerable literature on election forecasting, but many of the different approaches taken in the literature are, for different reasons, difficult to use in forecasting by-elections. There is literature on long-range forecasting of national vote shares on the basis of polls (Fisher 2015; Jennings, Lewis-Beck, and Wlezien 2020), but by-elections are rarely polled because of the practical difficulties and considerable cost involved in polling small areas.[2] There is literature on the long-range forecasting of national vote shares on the basis of "fundamentals" such as economic growth (Arnesen 2012; Nadeau and Lewis-Beck 2020) or the incumbent party's time in office (Abramowitz 2012), but often the required measures are not available at the local level, or the same fundamentals have unclear implications at local level. Finally, there is literature on the short-range forecasting of local vote shares given national general election polls (Lauderdale et al. 2020; Munzert 2017; and, on a quite different basis, Murr 2011), but there are reasons to think that voting behaviour in by-elections is distinct from voting behaviour in general elections (Price and Sanders 1998).

2.3. The methodological literature on the analysis of compositional data

All analysis of vote shares is the analysis of compositional data. In some limited contexts, the compositional character of vote shares can safely be ignored. In pure two-party systems, modelling a reference party's vote share automatically gives a prediction for the omitted party, since the two vote shares must sum to one. However, most by-elections are not pure two-candidate contests. Consequently, the analysis of multiparty by-elections must take one of several approaches to dealing with compositional data. These are:

- to ignore the compositional character of the data. Upton (1991), for example, argues that running separate party-specific regressions is justifiable since methods for the analysis of compositional data result "in a considerable increase in programming complexity that does not seem justified in [a] context where the aim is explore a phenomenon rather than provide fully efficient estimates" (113). Ignoring the compositional character of

[2]There is local polling data for some (71/468) of the by-elections studied here. I discuss these polls below.

the data is undesirable because separate regressions can produce logically impossible forecasts where the sum of forecast vote shares exceeds 100%.

- to re-define the dependent variable so that there are only two components, either by modelling the "two-party vote share" (Knotts and Ragusa 2016) or by modelling composite vote shares such as the vote share of incumbent parties (see the discussion in Walther 2015, 2–3) or the vote share of centre-left parties (Nadeau and Lewis-Beck 2020). Re-defining the dependent variable is undesirable because it generates *vague* forecasts: we might (for example) know that the governing party (parties) is fated to do poorly, but we want to know *which* opposition party will reap the benefit.
- to transform the compositional data into an unbounded space, perhaps by modelling the log ratio of each party to a reference party and estimating $n - 1$ party-specific regressions. This is the most common approach in political science (Katz and King 1999; Philips, Rutherford, and Whitten 2016), but is undesirable for forecasting because the regressions minimize error on the log ratio scale rather than the original untransformed scale.[3]
- To use multivariate regression models for compositional data, and specifically Dirichlet regression. Dirichlet regression was deprecated by the pioneers of compositional data analysis because (conditional on covariates) the components are independent of each other (Aitchison 1985, 136). More recent (2004 -) work has seen a resurgence in the use of Dirichlet regression, both generally and for election forecasting (Stoetzer et al. 2019).

3. Data

3.1. Sources of data

I draw on two principal sources of data. These cover by-election results and national and local opinion polling respectively. The principal source of by-election data is a data-set compiled by Professor Pippa Norris.[4] This data-set provides information on by-election results (and the preceding general election results) for by-elections from 1945 to 2012 inclusive. Using information from the House of Commons Library,[5] I have extended this data-set to cover by-elections held before the 31st December 2019. The source of opinion polling data is

[3] Consider an election where the reference party ("Blue") wins 30% of the vote, and where Red and Green parties win 31% and 1% of the vote. The log ratio of the Red vote compared to the Blue vote is log(.31/.30) = 0.03279. The log ratio of the Green vote compared to the Blue vote is log(.01/.30) = -3.4012. Now suppose that we correctly estimate the Blue share of the vote, but over-estimate the Red and Green vote by one percentage point. The error on the ALR-transformed scale for the Red party is log(.32/.30) - log(.31/.30) = 0.0317. The error on the ALR-transformed scale for the Green party is log(.02/.30) - log(.01/.30) = 0.693, over twenty times greater. If our loss function in forecasting is mean squared loss on the original scale, this feature of the log ratio transform is not helpful.

[4] "BRITISH BY-ELECTION RESULTS 1945-2012", available at https://www.pippanorris.com/data, accessed October 2020.

[5] These include "By-elections 2010-15" SN-05833; "By-election results since the 2015 General Election" CBP-7417, and "By-elections in the 2017 Parliament" CBP-8280

PollBase, a data-set compiled by Dr Mark Pack.[6] This data-set includes national opinion polls from 1945 onward, together with information on a limited number of by-election polls.

3.2. Outcome variable

My outcome variable is a five-component composition, where the components are the vote shares won in by-elections by the Conservative party, the Labour Party, the Liberal party (or its Liberal Democrat successor party), the two nationalist parties (the Scottish National Party and Plaid Cymru) taken together, and "all other" parties. I have collapsed the vote shares for the two nationalist parties because these parties never compete against one another. The dependent variable is plotted over time in Figure 1. The figure shows the decline over time of the combined two-party share of the vote, an increase and subsequent decline in the Liberal vote, and an increase towards the latter period in the combined "Other" category, an increase which is almost entirely due to the rise of UKIP and the Brexit Party rather than the Green Party (the other seat-winning party during this period).

The data includes certain structural zeros where parties did not stand a candidate. However, Dirichlet regression requires modelled shares to be strictly positive. In this respect, Dirichlet regression is similar to modelling log ratio transformed data. There are different methods for handling zeros in compositional data. I treat these zeros as if they were rounded zeros, or quantities too small to detect by the measurement process. In the by-election context, this means replacing these zeros with quantities smaller than one vote in the average voting population. Since the average voting population in by-elections is around 40,000, I replace zeros with values of 1/40,000. Nothing hinges on this value, and because it minimizes loss on the value scale rather than the log-transformed ratio scale, Dirichlet regression is not affected by this zero-replacement strategy in the same way that the analysis of log ratio transformed data is.

3.3. Predictor variables

I compile information on the several different predictor variables. These variables are generally the application to the UK case of the factors previously summarized in Table 1. For each predictor variable I discuss the rationale for inclusion and report means and standard deviations in parenthesis.

Considering first the factors which operate at the *national* level: I begin by including the current polling for each party, less its national vote share in the preceding general election. Whilst by-elections may not *just* be the translation to local contexts of changes seen in national polling, it would be foolish to ignore this rich source of data concerning parties' fortunes. I include information for Labour (mean 0; SD 0.07), Liberal (mean 0.01; SD 0.08), and "all other" parties (mean 0.01; SD 0.03). I omit polling changes for the Conservative party to avoid collinearity. I omit polling changes for the nationalist parties because these

[6] Available at https://www.markpack.org.uk/opinion-polls/. I use the 2020 quarter 2 release.

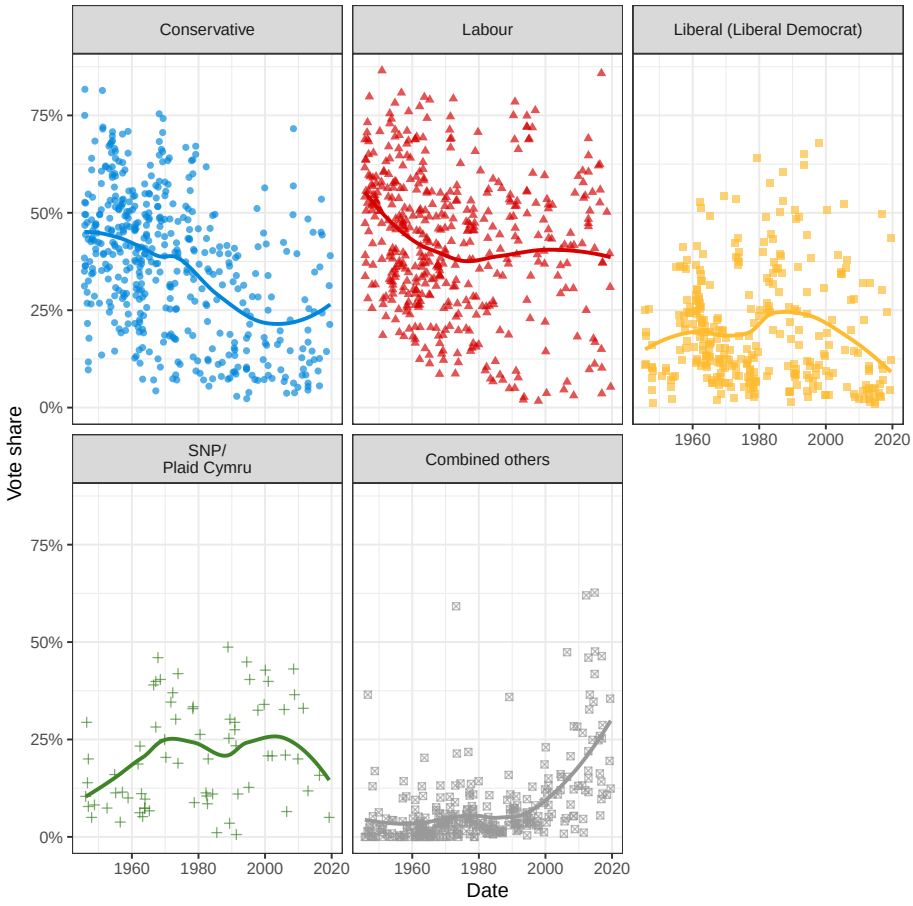

Figure 1: Vote shares in by-elections in the post-war period

parties' polling is often not reported. "Current polling" is a right-aligned seven day average of polls. Missing values are replaced with the last non-missing figure. I go on to include dummies for participation in government by the Labour party (mean 0.28) and the Liberal Democrats (mean 0.04), since previous literature has suggested that governing parties do poorly in by-elections. Once again, there is no variable for the Conservative party, because whenever the Labour party is in government the Conservative party is in opposition. I also include some miscellaneous national factors, such as the rate of retail price inflation over the twelve months preceding the election (ONS series ID: CZBH) (mean 5.3; SD 3.82), and a dummy variable which has a value of one if the by-election was held in winter (mean 0.4).

Turning now to the factors which operate at the *local* level, I begin by including the vote share of each party in the constituency in the preceding general election. This gives information on the relative "starting points" of each party, starting points which may be modified by parties' (national) polling performance. I include information for Labour (mean 0.46; SD 0.17), Liberal (mean 0.09; SD 0.1), and nationalist parties (mean 0.01; SD 0.05), together with the combined vote share of "all other" non-Conservative parties (mean 0.02; SD 0.08). As before, I omit figures for the Conservative party to avoid collinearity. Because zeros in these variables can indicate extremely low local support or structural zeros, I also include dummy variables which have a value of one if each party stood a candidate in the by-election. Values are close to one for the Conservative (mean 0.99) and the Labour parties (mean 0.99), and lower for the Liberal (mean 0.57), nationalist (mean 0.12), and "all other" parties (mean 0.44).

Although the party with the highest vote share will (necessarily) win the seat, parties can win seats with higher or lower absolute vote shares, and so I also include two dummy variables which record whether the Labour party (mean 0.54) or the Liberal party (mean 0.01) held the seat prior to the by-election. Ordinarily, the incumbent does not stand for re-election (perhaps the by-election has been caused by their death or resignation), but in cases of party defection incumbents often stand under a different label. I therefore include dummy variables which have a value of one if the Liberal or "Other" candidate was the incumbent MP. These dummy variables have very low means, but reflect situations where candidates called by-elections on their own initiative, and sometimes as a response to party defections. Further local predictors are a variable which records the turnout in the seat in the preceding general election (mean 0.73; SD 0.09), and dummy variables which record whether the by-election was held in a Scottish (mean 0.15) or Welsh (mean 0.07) constituency.

The last set of local predictors is not shown in Table 1, perhaps because it was considered too obvious to study: if a party does not stand a candidate in a by-election, its vote share must be zero. Studies which have modelled two-party vote share, or the vote share of parties taken one at a time, have not typically had to deal with this complication. I therefore include as additional predictors whether or not each party stood a candidate in the by-election. As with the "general election candidacy" predictors, values are close to one for the

Conservative (mean 0.99) and the Labour parties (mean 0.99), and lower for the Liberal (mean 0.68), nationalist (mean 0.15), and "all other" parties (mean 0.65).

All of these variables, including dummy variables, were standardized to have a mean of zero and a standard deviation of one. This standardization is necessary for the variable selection that follows, but also helps in the specification of prior distributions for coefficients.

4. Analysis

4.1. Variable selection

I carried out variable selection using as candidate variables all of the variables listed in Table 2 and all of the pairwise interactions between those variables. All candidate variables were standardized before variable selection was carried out. Variable selection was carried out using the `glmnet` package (Friedman, Hastie, and Tibshirani 2010) for the R statistical environment (R Core Team 2020). `glmnet` implements cross-validated multiresponse lasso regression. I use *lasso* regression, which shrinks the value of certain coefficients to zero, and which can be used as the first stage in a two-stage modelling process. I prefer lasso regression to other techniques (ridge regression, horseshoe priors) which perform continuous shrinkage. Models where certain variables are selected (have non-zero values) are preferable for forecasting applications where the model has to be deployed in a light-weight computing environment. I use *multiresponse* lasso regression (Simon, Friedman, and Hastie 2013) to select the same predictors for all components of the compositional outcome. This is not necessary – different predictors can be used for different components of a Dirichlet-distributed outcome – but the use of the same set of predictor variables across all components further reduces model complexity. Finally, I perform cross-validated variable selection to guard against over-fitting, with 10-fold cross-validation and mean square error as the loss function.

Table 2 shows the variables which are included in the model specification which minimizes cross-validated mean squared error. The same variables feature in the rows and the columns; the entries in the cells give the order in which the variables (other than the intercept) were selected at progressively smaller levels of λ. The best model includes 31 different predictors.

Table 2: Order of inclusion

Variable	Main term	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)	(17)	(18)	(19)	(20)
(1) Labour GE share	1				13			21													
(2) Liberal candidate	2																				
(3) Nat. candidate	3																			22	
(4) Liberal GE share	4															20			17		
(5) Nat. GE share	4																			8	

Variable	Main term	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)	(17)	(18)	(19)	(20)
(6) Labour poll change	4																				
(7) Labour party inc.	7			22																	
(8) Other poll change	8																				
(9) Other candidate(s)	8																				26
(10) Other GE share	14														26					26	
(11) Turnout at GE	14								22												
(12) Labour govt	17			26																	
(13) Lib poll change	22		14																		
(14) Lib Dems in govt			8														8				
(15) Inflation														26							
(16) Conservative candidate																					17
(17) Labour candidate																				26	
(18) Lib. candidate in GE																					
(19) Personal incumbent running as other																					
(20) Labour candidate in GE																					

4.2. Dirichlet regression

In the preceding section, I used a multiresponse linear model to carry out variable selection. That model would be inappropriate for forecasting, since it ignores the compositional character of the data. I therefore use the same predictors identified in the previous step, but in a Dirichlet regression model (Campbell and Mosimann 1987). The Dirichlet distribution is a continuous multivariate distribution. The sum of a Dirichlet-distributed random variate is always one. Ordinarily, the Dirichlet distribution has, as parameters, a number of categories k and a length-k vector of concentration parameters α. Alternately, the Dirichlet distribution can be re-parameterised to allow the concentration parameters to depend on component-specific mean parameters μ_i and a common precision parameter ϕ, such that $\alpha_i = \mu_i \phi$. This in turn allows the parameters μ to be linked to covariates using the same softmax link used for multinomial regression:

$$\mu_j = \frac{exp(X^j \beta^j)}{1 + \sum exp(X\beta)}$$

where β_1 is set to zero for the purposes of identification.[7] Although Dirichlet regression models can be estimated using a variety of methods, including maximum likelihood, the estimation of Dirichlet regression models using Bayesian methods is a natural choice for forecasting applications. Because Bayesian methods generate a posterior distribution of parameter values, the probabilities of particular outcomes can be calculated easily by iterating over draws from the posterior distribution. Although this process can be mimicked using frequentist methods (King, Tomz, and Wittenberg 2000), the naturalness of the link between Bayesian estimation and probability statements means that Bayesian methods are an obvious choice in this context. Accordingly, I estimate the model using the `brms` package (Bürkner 2017), which is in turn built on Stan (Stan Development Team 2020), a probabilistic programming language for Bayesian inference.[8]

When modelling Beta- or Dirichlet-distributed percentages, specifying priors becomes particularly important because common default priors can lead to prior predictive distributions which place large probability masses on percentages close to zero or one. Because these vote shares are a priori implausible, researchers need to carefully specify prior distributions to avoid this scenario. I use party-specific priors for the intercept terms which are based on *a priori* assumptions about the parties' strength over the post-war period which would have been reasonable at the start of the period covered.[9] I then use a prior on the precision term ϕ which ensures that there are non-zero chances both of the largest parties winning their deposit and of "all other parties" winning a contest.[10] Finally, I set N(0, 0.25) priors on the coefficients, with the exception of coefficients relating to candidacy variables, where I set a N(0, 1) prior on candidacy variables for the relevant party, and N(0, 0.25) priors for candidacy variables' effect on other parties. A detailed rationale for each of these prior choices can be found in the Appendix. Since these priors are tailored to the British context, they would need to be reviewed before being applied to other countries.

Coefficients from the model and 95% credible intervals are plotted in Figure 2. Coefficients for the Conservative party are zero by construction. The intercept for the precision parameter ϕ are the same for all components ($\hat{\phi} = 20$; 95% CI: 18.4 to 21.5). Interpretation of the coefficients is made difficult by the fact that many terms feature not just in their own right but as part of interactions, and even for coefficients which feature only in their own right it would be wrong to interpret these coefficients as having any causal significance. Predictions from the model can be generated by repeatedly drawing from the posterior distribution of parameters relating to μ ($= X\beta$) and ϕ, and using these parameters to generate

[7] The precision parameter ϕ (which must be strictly positive) can also be modelled using covariates, but I do not do this here.

[8] Each model was run over five chains for 1,300 iterations, with the first 1,000 iterations discarded as warmup iterations. There were no divergent transitions, no iterations exceeding the maximum tree-depth, and no problems with convergence as monitored by the $\hat{R}$ statistic.

[9] Labour: N(0, 0.429); Liberal: N(-0.95, 0.584); Nat.: N(-2.457, 0.714); Other: N(-0.95, 0.584).

[10] Specifically, $\phi \sim N(11.5, 3.25)$.

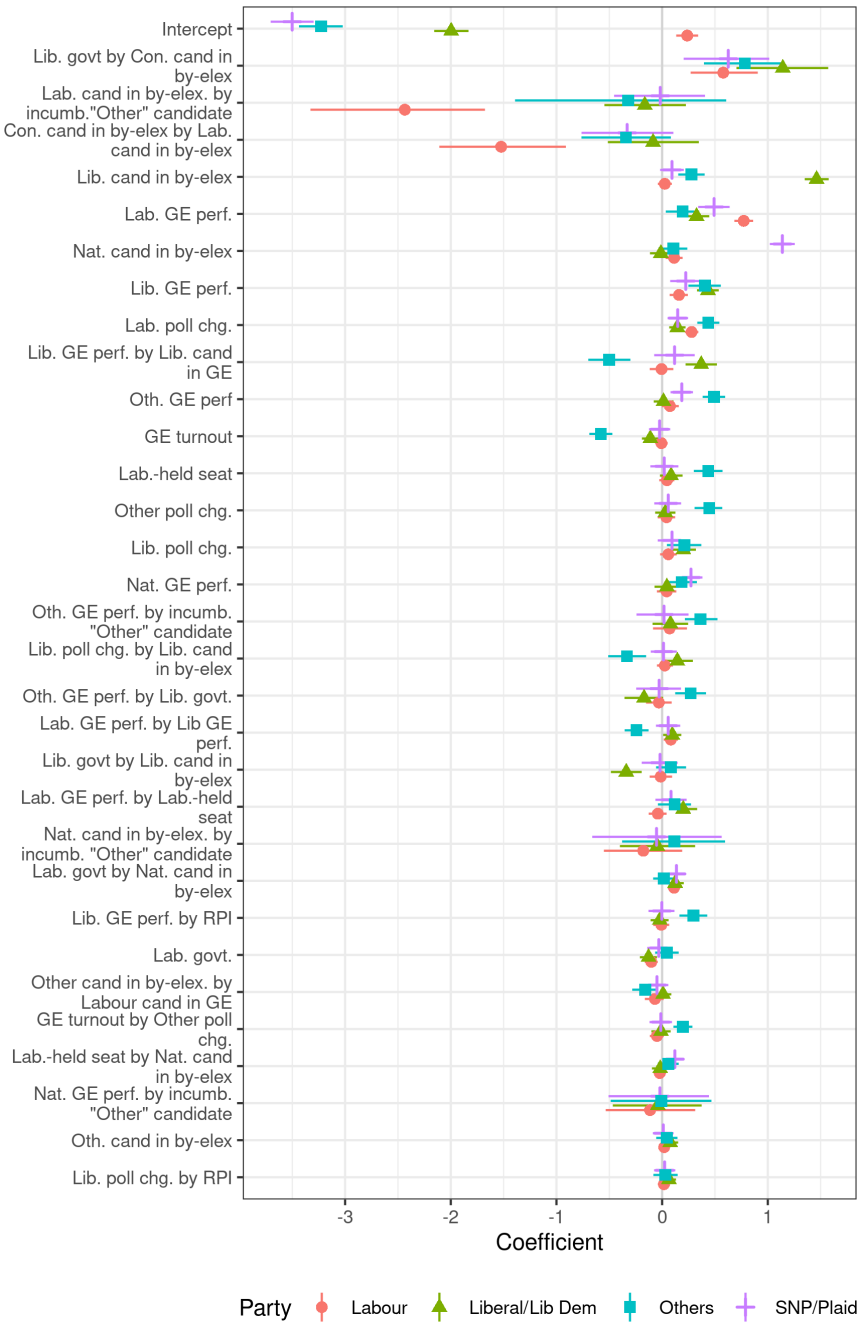

Figure 2: Coefficient plot from Dirichlet regression. Points are plotted in separate shapes and colours for each component. Coefficients are ordered from highest average absolute magnitude to lowest.

Table 3: Performance measures based on the model and other forecasting techniques

Statistic	In sample	LOO	National polls	Local polls
Seats correctly predicted	85.04	83.76	82.48	88.29
Mean absolute error	3.64	4.00	5.76	3.60
Median absolute error	1.79	1.88	2.72	2.70
Multiclass Brier score	25.04	24.96	NA	NA
Predictions inside 95% interval	95.94	94.96	51.88	59.05

Note:
Figures for local polls refer to a subset of 111 by-elections for which there is local polling.

a draw from a Dirichlet distribution.

5. Performance

Table 3 gives selected performance metrics for the model and two rival ways of producing forecasts of by-elections based on local polling and national polls respectively. I begin by discussing the in-sample performance of the model, since patterns found in the in-sample performance also apply more generally. I then move on to evaluate the leave-one-out performance of the model and rival procedures.

5.1. In-sample performance

The first column of table 4 gives in-sample performance. The model predicts 398 out of 468 contests correctly. Although the model does not systematically over- or under-predict vote shares for any party, the model under-predicts Liberal/Liberal Democrat winners (17 wins predicted, 35 actual). This pattern is consistent with Liberal Democrat support being highly local rather than national (Rallings and Thrasher 1999) and (for the period before 2010) able to channel a variety of protest votes (Curtice 2007). The model mean absolute error averages across all parties, and absolute error for nationalist parties and all other parties (1.85 and 2.95 respectively) is much smaller than the average across all parties. This is because both predictions and values are often exactly zero (in cases where no such party stands) or close to zero (for all other parties). As is to be expected, almost exactly 95% of in-sample predictions are within the 95% prediction interval, and as such we would describe the model as well-calibrated on the basis of in-sample model performance.

5.2. Leave-one-out performance

As is well known, performance metrics calculated on the basis of in-sample data are *optimistic*, and the best test of a model is performance on out-of-sample data. I therefore calculate the performance of the model by leaving one observation out, estimating the model, and generating predictions for the

withheld observation. Leave-one-out (LOO) performance is inferior to evaluating the model on new data to which the researcher has had absolutely no access, but since new data only arrives at the rate of three to four observations per year, it is not possible adequately to test the model on forthcoming by-elections. The LOO performance of the model reveals that the in-sample performance measures were moderately optimistic: the figures for mean and median absolute error are 10% and 5% worse than the corresponding figures for in-sample performance. However, the figures for correct prediction, the multiclass Brier score, and calibration are almost unaffected.

5.3. Forecasting ten elections ahead

One possible concern about this model is that is estimated on a relatively large number of observations. Models of Australian or Canadian by-elections might only have one half of one-quarter as many observations as are modelled here. The performance of the model might therefore result from the greater ability of the model to incorporate a large number of predictor variables and to estimate the effects of these predictor variables precisely.

In order to address this concern, I estimate the model at different points over the post-war period. I start by estimating the model on the first ten post-war by-elections, and generate out-of-sample predictions for the following ten by-elections. I record performance metrics for these out of sample predictions, and then repeat the exercise moving forward ten by-elections. This procedure is designed to mimic the predictions that might have been made in real time, *except that* the model specification is still the result of a process of variable selection using data from the whole period. The results of this procedure are shown in Figure 3, for three different metrics (mean absolute error, percentage correctly predicted, and Brier score).

The figures show that sample size is far from the primary driver of out-of-sample model performance, and that instead the primary driver is a secular trend towards more unpredictable by-elections. Model performance is highest in the earliest part of the post-war period, and gets steadily worse. This finding that more recent by-elections are less predictable matches current accounts of increasing (individual-level) volatility in general elections (Fieldhouse et al. 2019, 9–72). This is therefore a mixed finding for election modellers: high performing models can be trained on limit data, but forecasting in the present period is difficult.

5.4. Comparison to national polling

Are these figures for LOO performance good or bad? The answer must depend on the performance of alternative methods of forecasting by-elections. Here I consider two alternative methods based on national and local polling respectively. Column 3 in Table 4 reports the performance of a simple method of by-election forecasting which involves taking results in the preceding general election, and adding on the uniform national swing implied by national polling in the period preceding the by-election. For example: if the Conservatives won

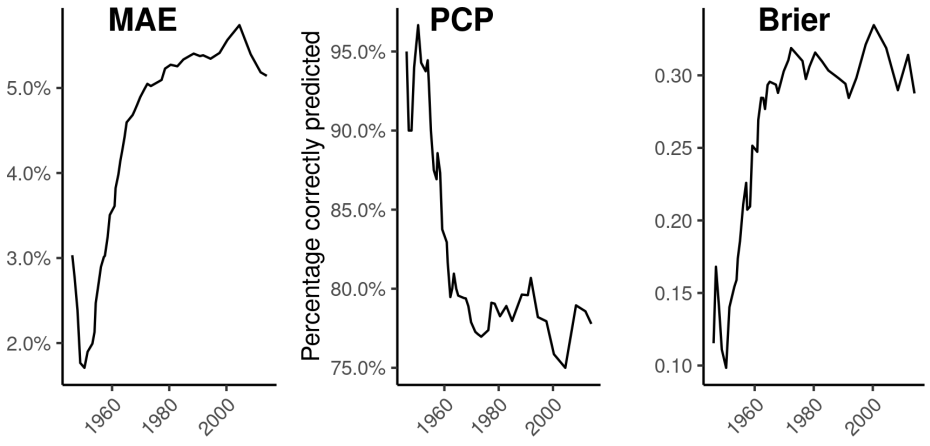

Figure 3: Model performance over next ten by-elections, for models estimated on progressively larger windows of data from 1945 onwards

40% of the vote in the seat, and if national polling shows their national vote share has increased seven percentage points compared to their vote share in the last general election, then the prediction for the Conservatives is 47%. This method of forecasting respects the compositional nature of the outcome (past election results are compositions and changes in polling sum to zero), but can result in predictions of negative vote share (where a party performs poorly in an area and then suffers a further national loss in popularity) and positive vote share predictions for parties which do not stand. To construct a prediction interval, I add (subtract) three percentage points, since this is the margin of error associated with a standard national polling sample of 1,000 respondents. This method of forecasting elections often predicts the correct winner, but it does so with much greater error: the figures for mean and median absolute error are very much larger than the leave-one-out performance of the model, and relatively few predictions are within the 95% forecast interval.

5.5. Comparison to local polling

What about local rather than national polls? I motivated this article by noting that by-elections are rarely the subject of polls. However, this does not mean that there are *no* by-election polls. Mark Pack has compiled results from 111 by-election polls in 71 contests. Performance measures for these local polls average over different observations from the measures in columns (1) - (3), in two senses. First, these local polls covered contests which were more newsworthy. Second, these local polls often did not report figures for other parties. When calculating performance measures, I average over all 111 polls and all parties for which a polling figure was reported. To construct a prediction interval, I add (subtract) 3.3 percentage points, since this is the margin of error associated

with a local polling sample of 855 respondents, which is the close to the average sample size reported in Hanretty, Lauderdale, and Vivyan (2018), 586.

The performance of local polls is better than the out-of-sample performance from the model on two metrics (seats correctly predicted and mean absolute error) but has worse median absolute error and is over-confident (lower calibration score). Considering just these performance metrics, then, local polls are generally to be preferred, except for those who are willing to write off occasional errors as outliers, who may prefer instead to use the model. Of course, the choice between model-based forecasting and local polling depends not just on loss functions like mean and median absolute error, but on *cost* functions: local polls cost money to commission, whereas the forecasting model set out here is freely available. The model is therefore a useful option where polling is not feasible or not affordable.

6. What about log ratio regressions?

In section 2.3 I noted that the most common method for analysing compositional data in political science is to log transform the ratio of each component to a reference component and perform (potentially correlated) regressions on these log ratios. I argued against using log ratio regressions on the grounds that these regressions minimized error on the log-ratio scale rather than on the scale we actually care about. Nevertheless, it is reasonable to ask whether Dirichlet regression performs better for analysis of vote shares than log ratio regressions.

To test whether Dirichlet regression is superior to an analysis of log ratios, I transform my multivariate dependent variable by taking the log ratio of parties' vote shares relative to the reference (Conservative) party. I carry out variable selection in the same way I did for Dirichlet regression, and estimate a multivariate regression with correlated errors across components. In order to generate predictions, I draw from the posterior distribution of regression coefficients, and reverse the log ratio transformation.

The in-sample performance on an ALR model is very much poorer than the performance of a Dirichlet regression model, and is a worse guide to outcomes than a model based on national polling. The mean absolute error is almost twice as large as the in-sample error from the Dirichlet regression model (6.2 compared to 3.5), and the percentage of outcomes correctly predicted is extremely low (75%). Given that the in-sample performance of the ALR model is so much poorer, comparing the out-of-sample performance is redundant. For forecasting of vote shares, Dirichlet regression is superior to regressions on log ratios.

7. Conclusion

In this paper I have set out a method for forecasting by-elections in the United Kingdom. The method produces forecasts which are in some respects comparable with forecasts derived from local polling without the cost of commissioning polling. The method uses data from the entire post-war period (the period for which national polling is available) and selects a subset of variables from a candidate list of variables suggested by previous literature.

The method can in principle be extended to other jurisdictions which use by-elections to replace members of the legislature. The same set of candidate variables would apply to other jurisdictions, although sparser data in other jurisdictions (the UK has a large legislature and long parliamentary terms) may make the variable selection stage more difficult. The model includes features also found in other jurisdictions, such as the presence of separatist parties, which are found in Canada just as they in Great Britain.

In applying this method to other jurisdictions researchers will have to answer a number of questions about the set of candidate predictor variables and the way in which results are recorded. In this paper, I model elections over a long period in which the UK introduced and abolished different forms of second-order elections (elections to national parliaments, elections to local assemblies; elections to the European Parliament) and reorganized local government areas. Where instead sub-national elections are held regularly on a systematic basis (as is certainly the case in Canada and Australia), researchers may be able to incorporate results from *other* second order elections. Researchers in jurisdictions with or without other regularly held second order elections will also have to consider the coding of the "other" category. Here, I have modelled parties which won more than two seats over the period, and included all other parties, including electorally relevant parties like UKIP and the Green party, in the residual "other" category. Clearly the forecast for all other parties sets an upper bound on the forecast for a specific other party, but researchers who are interested in making forecasts for a range of parties may have to subset their data to a period where a named party is consistently recorded in polling data.

Future research may address one of three outstanding issues: performing variable selection within a Dirichlet regression framework rather than selecting those variables selected through a multiresponse linear regression; dealing with zero-inflation in compositional data due to varying patterns of candidacy (Tang and Chen 2019); and modelling not just the mean of the Dirichlet distribution but also the precision.

Acknowledgements

REMOVED FOR REVIEW.

References

Abramowitz, Alan. 2012. "Forecasting in a Polarized Era: The Time for Change Model and the 2012 Presidential Election." *PS, Political Science & Politics* 45 (4): 618.

Aitchison, John. 1985. "A General Class of Distributions on the Simplex." *Journal of the Royal Statistical Society: Series B (Methodological)* 47 (1): 136–46.

Anderson, Christopher J, and Daniel S Ward. 1996. "Barometer Elections in Comparative Perspective." *Electoral Studies* 15 (4): 447–60.

Arnesen, Sveinung. 2012. "Forecasting Norwegian Elections: Out of Work and Out of Office." *International Journal of Forecasting* 28 (4): 789–96.

Bürkner, Paul-Christian. 2017. "brms: An R Package for Bayesian Multilevel Models Using Stan." *Journal of Statistical Software* 80 (1): 1–28. https://doi.org/10.18637/jss.v080.i01.

Campbell, G, and J Mosimann. 1987. "Multivariate Methods for Proportional Shape." In *ASA Proceedings of the Section on Statistical Graphics*, 1:10–17. Washington.

Cook, Chris, and John Ramsden. 1973. *By-Elections in British Politics*. Springer.

Curtice, John. 2007. "New Labour, New Protest? How the Liberal Democrats Profited from Blair's Mistakes." *The Political Quarterly* 78 (1): 117–27.

Economou, Nicholas M. 1999. "The Trouble-Maker's Ballot Box? A Note on the Evolving Role of the Australian Federal by-Election." *Australian Journal of Political Science* 34 (2): 239–47.

Feigert, Frank B, and Pippa Norris. 1990. "Do by-Elections Constitute Referenda? A Four-Country Comparison." *Legislative Studies Quarterly*, 183–200.

Fieldhouse, Edward, Jane Green, Geoffrey Evans, Cees van der Eijk, Hermann Schmitt, Jonathan Mellon, and Christopher Prosser. 2019. *Electoral Shocks: The Volatile Voter in a Turbulent World*. Oxford University Press, USA.

Fisher, Stephen D. 2015. "Predictable and Unpredictable Changes in Party Support: A Method for Long-Range Daily Election Forecasting from Opinion Polls." *Journal of Elections, Public Opinion & Parties* 25 (2): 137–58.

Friedman, Jerome, Trevor Hastie, and Robert Tibshirani. 2010. "Regularization Paths for Generalized Linear Models via Coordinate Descent." *Journal of Statistical Software* 33 (1): 1–22. http://www.jstatsoft.org/v33/i01/.

Hanretty, Chris, Benjamin E Lauderdale, and Nick Vivyan. 2018. "Comparing Strategies for Estimating Constituency Opinion from National Survey Samples." *Political Science Research and Methods* 6 (3): 571–91.

Jennings, Will, Michael Lewis-Beck, and Christopher Wlezien. 2020. "Election Forecasting: Too Far Out?" *International Journal of Forecasting*.

Katz, Jonathan N, and Gary King. 1999. "A Statistical Model for Multiparty Electoral Data." *American Political Science Review*, 15–32.

King, Gary, Michael Tomz, and Jason Wittenberg. 2000. "Making the Most of Statistical Analyses: Improving Interpretation and Presentation." *American Journal of Political Science*, 347–61.

Knotts, H Gibbs, and Jordan M Ragusa. 2016. "The Nationalization of Special Elections for the US House of Representatives." *Journal of Elections, Public Opinion and Parties* 26 (1): 22–39.

Lauderdale, Benjamin E, Delia Bailey, Jack Blumenau, and Douglas Rivers. 2020. "Model-Based Pre-Election Polling for National and Sub-National Outcomes in the US and UK." *International Journal of Forecasting* 36 (2): 399–413.

Loewen, Peter John, and Frédérick Bastien. 2010. "(In) Significant Elections? Federal by-Elections in Canada, 1963-2008." *Canadian Journal of Political Science/Revue Canadienne de Science Politique*, 87–105.

Mughan, Anthony. 1988. "On the by-Election Vote of Governments in Britain." *Legislative Studies Quarterly*, 29–48.

Munzert, Simon. 2017. "Forecasting Elections at the Constituency Level: A Correction–Combination Procedure." *International Journal of Forecasting* 33 (2): 467–81.

Murr, Andreas Erwin. 2011. "'Wisdom of Crowds'? A Decentralised Election Forecasting Model That Uses Citizens' Local Expectations." *Electoral Studies* 30 (4): 771–83.

Nadeau, Richard, and Michael S Lewis-Beck. 2020. "Election Forecasts: Cracking the Danish Case." *International Journal of Forecasting*.

Norris, Pippa. 1990. *British by-Elections: The Volatile Electorate*. Oxford University Press.

Philips, Andrew Q, Amanda Rutherford, and Guy D Whitten. 2016. "Dynamic Pie: A Strategy for Modeling Trade-Offs in Compositional Variables over Time." *American Journal of Political Science* 60 (1): 268–83.

Price, Simon, and David Sanders. 1998. "By-Elections, Changing Fortunes, Uncertainty and the Mid-Term Blues." *Public Choice* 95 (1-2): 131–48.

Rallings, Colin, and Michael Thrasher. 1999. "Local Votes, National Forecasts–Using Local Government by-Elections in Britain to Estimate Party Support." *International Journal of Forecasting* 15 (2): 153–62.

R Core Team. 2020. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Reif, Karlheinz, Hermann Schmitt, and Pippa Norris. 1997. "Second-Order Elections." *European Journal of Political Research* 31 (1-2): 109–24.

Simon, Noah, Jerome Friedman, and Trevor Hastie. 2013. "A Blockwise Descent Algorithm for Group-Penalized Multiresponse and Multinomial Regression." *arXiv Preprint arXiv:1311.6529*.

Stan Development Team. 2020. "RStan: The R Interface to Stan." http://mc-stan.org/.

Stoetzer, Lukas F, Marcel Neunhoeffer, Thomas Gschwend, Simon Munzert, and Sebastian Sternberg. 2019. "Forecasting Elections in Multiparty Systems: A Bayesian Approach Combining Polls and Fundamentals." *Political Analysis* 27 (2): 255–62.

Studlar, Donley T, and Lee Sigelman. 1987. "Special Elections: A Comparative Perspective." *British Journal of Political Science* 17 (2): 247–56.

Tang, Zheng-Zheng, and Guanhua Chen. 2019. "Zero-Inflated Generalized Dirichlet Multinomial Regression Model for Microbiome Compositional Data Analysis." *Biostatistics* 20 (4): 698–713.

Upton, Graham JG. 1991. "The Impact of by-Elections on General Elections: England, 1950-87." *British Journal of Political Science*, 108–19.

Walther, Daniel. 2015. "Picking the Winner (S): Forecasting Elections in Multiparty Systems." *Electoral Studies* 40: 1–13.

Forecasting multiparty by-elections: Appendix

Abstract

By-elections, or special elections, play a special role in many electoral democracies – but whilst there are multiple forecasting models for national elections, there are no such models for multiparty by-elections. Multiparty by-elections, like those in the UK, present particular analytic problems. I model party vote shares using Dirichlet regression, a technique suited for compositional data analysis. After identifying a set of predictors from a broader set of candidate variables, I estimate a Dirichlet regression model using data from all post-war by-elections in the UK (n = 468). The cross-validated error of the forecasting model is comparable to the error of costly and infrequent by-election polls. The steps taken in the analysis are in principle applicable to any system which uses by-elections to fill legislative vacancies.

1. Performance of ALR models

In the main text I discussed two strategies for modelling compositional data: Dirichlet regression, and running (potentially correlated) multivariate regressions of log ratios of components to a reference component ("additive log ratios"), which I refer to as ALR regression. I argued that Dirichlet regression was better for forecasting purposes, because it minimizes errors on the scale of the original variables rather than a transformed version of those variables.

This claim is born out by carrying out the same analytic steps in the main article, but for ALR regression instead of Dirichlet regression. I estimate a multivariate lasso regression on four ALR-transformed continuous variables (log ratios of Labour, Liberal, Nationalist and "all other" shares relative to the Conservative share). I take the model specification which minimizes the cross-validated error on the ALR scale. I use that specification in an ALR model of by-election outcomes. I back transform the fitted values of the model, and calculate the same fit statistics as are shown in the table in the main article.

In all respects, the within sample performance of the ALR model is inferior to the within-sample perform of the Dirichlet regression model. The MAE is higher (6.9 compared to 3.7), as is the proportion of seats correctly predicted (73.5% compared to 83.8%). The Brier score for the ALR model is considerably worse (0.437 compared to 0.256). Given that the in-sample performance of the ALR model is so much poorer, comparing the out-of-sample performance is redundant.

2. Modelling across-election variance

I note in the main text that it is possible to model the mean and precision of Dirichlet-distributed outcomes. Because there is no obvious way simultaneously to perform variable selection for the mean and precision of the response, I do not model the precision in the models reported in the main text. In this supplementary analysis, I discuss the changes in in-sample performance that result from including researcher-selected variables as predictors of excess variance across elections. This technique only allows researchers to model unpredictable *elections*, rather than unpredictable *parties*.

The variables I use to model variance are:

- the number of polls included in the rolling seven-day average of opinion polls;
- whether the by-election took place in Scotland or in Wales;
- the "other" share of the vote in the preceding election.

The leave-one-out performance of this model is slightly better than the performance of the model presented in the main text for two important measures, the Brier score (0.251 versus 0.266), and mean absolute error (4.0 versus 4.1). Performance is identical for median absolute error and seats correctly predicted, but worse for calibration, although the calibration is still nominal (96.3 versus 97.1).

Modelling variance across elections therefore results in performance that is marginally but not clearly better across a range of measures, at least for this selection of variables.

3. Prior choice

In this section I describe the reasoning behind my choice of priors. I begin by describing the reasoning surrounding the selection of priors for the intercept terms in the model. In the model, these intercepts are expressed as log ratios of vote shares relative to a reference party. I have, however, found it more helpful to start by expressing my intuitions regarding vote shares directly, and only subsequently converting these to priors concerning log ratios.

I start with the two largest parties, Conservative and Labour. At the beginning of this period it was reasonable to assume that the Conservative and Labour parties would be the two largest parties. I think it also reasonable to assume that there was no expectation that one party would *generally* be larger than the other party. Work on electoral systems has established that the seat share of the largest party is approximately equal to the seat product to the power of negative one-eighth; with around 630 seats in mainland Great Britain and a median district magnitude of one, the seat share of the largest party in expectation is 44.5% (Shugart and Taagepera 2017). The vote share of the largest party is smaller than the seat share ($v_1 = (s_1^{-2} + 1)^{-1/2}$) at 40.7%. There is no clear expectation for the seat or vote share of the second party in a system:

Taagepera and Allik (2006) suggest that where the number of parties is given by p, then $v_2 = \frac{1-v_1}{\sqrt{p-1}}$, which gives a figure of approximately 0.3. If we are ambivalent about whether the Conservatives or Labour are the larger party, but if there is equal probability of either outcome, then our best guess as to their vote share is the average of the vote shares of the first and second-placed parties, at roughly 35%.

What might be the standard deviation of the vote shares of the two largest parties? One way of mining our intuitions is to consider scenarios which would be "unlikely but possible", and which should therefore be in the bulk of the probability mass. One scenario which is unlikely but possible is one party winning a majority of votes cast. If our priors follow a normal distribution, and this outcome is regarded as a "two-sigma" event, then our standard deviation given a mean of 35% is 7.5%.

I now turn to the three minor "parties": the Liberals, the Nationalist parties, and all others. We know, from our priors on the two main parties, that the sum of these parties' vote shares should not, in expectation, exceed 30%. We also know, given the population of Scotland and Wales relative to England, that the vote share of the Nationalist parties logically cannot exceed 12%. Somewhat arbitrarily, I set the estimate for the average vote share of nationalist parties at one-quarter of this maximum, or 3%. This leaves 27% to be divided between the Liberals and all others. Without strong intuitions as to how to divide the remaining share of the vote between these two parties, and in recognition of the fact that the "all others" vote share will be split between multiple parties, I split the remaining vote equally, 13.5% and 13.5%.

What should the standard deviation of these vote shares be? It would be surprising if the variability of the Liberal vote share were greater than the variability of the much larger vote shares of the Conservative or Labour parties. A standard deviation of five percent ensures that two standard deviations either side do not create negative vote shares or vote shares which approach the expected vote share of the second-placed party (30%) less one standard deviation. I apply this standard deviation to "all other" parties as well. For the combined nationalist vote, I use a standard deviation of 1.5%, which is the largest standard deviation which does not create negative vote shares at two standard deviations below the mean.[1]

Given these means and standard deviations, what are our expectations about log-ratios? Katz (1978) has suggested that for two normally distributed variables x and y, the log ratio T is distributed:

$$T \sim \log_e\left(\frac{\mu_x}{\mu_y}\right) + \mathbb{N}\left(0, \frac{\sigma_x^2}{\mu_x^2} + \frac{\sigma_y^2}{\mu_y^2}\right)$$

[1] This does create the possibility of negative vote shares. Whilst other distributions bounded on [0, 1] would avoid this risk, it would also make the process of generating intuitions about vote shares much more difficult.

Our expectation for the mean in this term is zero (the ratio is 1:1, and the log of one is zero). Our expectation for the standard deviation is $2 \times \frac{0.075}{0.35} = 0.429$. We can repeat this for the remaining parties. The resulting prior distributions are:

- Labour: N(0, 0.429)
- Liberal: N(-0.95, 0.584)
- Nat.: N(-2.457, 0.714)
- Other: N(-0.95, 0.584)

This completes the reasoning for the priors surrounding the intercepts in the model. I now turn to the priors concerning the precision parameter. Recall our expectations about vote shares:

$$[0.35, 0.35, 0.135, 0.025, 0.135]$$

If we start from these vote shares and multiply them by a precision parameter, we get the input to a Dirichlet distribution which we can simulate. Figure 3 shows the distribution of the *first* (i.e., Conservative) component at different values of the precision parameter ϕ. The sub-title below each panel shows the probability that the Conservative share will be below 5% (the threshold below which candidates lose their deposit). This is equivalent to the predictions we would make before seeing any of the data.

We can use this figure to set a soft lower bound on the value of the precision parameter. Note how when ϕ is set to 5, the probability of a Conservative vote share less than 5% is close to 2.5%. A Conservative candidate losing their deposit is the kind of "unlikely but possible" outcome we would wish to assign low but non-zero probability too.

At the same time, we do not want to set the precision parameter so high that some events become impossible. Figure 4 shows the probability that "other" vote share is the largest vote share, as a function of the precision parameter. As the value of ϕ increases above 18, the probability of others winning dips below 2.5%.

Since we do not wish to exclude the possibility either that a Conservative candidate loses their deposit, or that an "other" candidate wins, values of the precision parameter somewhere between 5 and 18 seem reasonable. Rather than place a uniform prior on this range, I prefer to place a normal prior with a mean of 11.5 and a standard deviation of one quarter of the range (3.25).

What does this give when applied to the whole data? A prior predictive check using a model with an intercept and precision parameter alone gives the distributions shown in Figure 3. The distribution of vote shares is slightly tighter than the distribution shown previously in the pure Dirichlet case.

This prior predictive check is for a simple model with just an intercept, rather than the more complex model with multiple parameters. In order to illustrate some of the difficulties that come with including more parameters, I carry out a prior predictive check, placing a N(0, 1) prior on all the coefficients in the model.

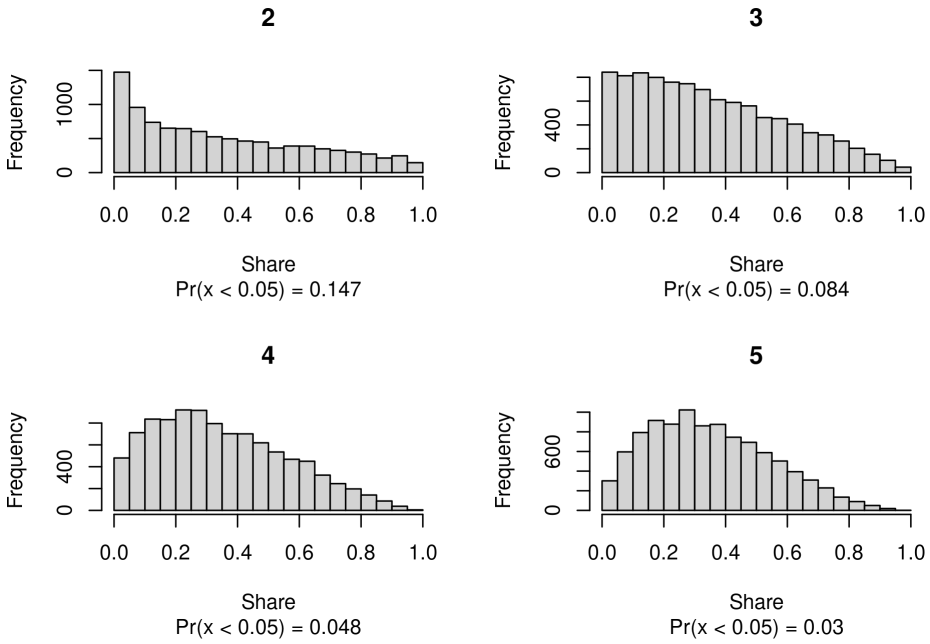

Figure 1: Histogram of the distribution of the first component of a Dirichlet distribution with varying values of the precision parameter.

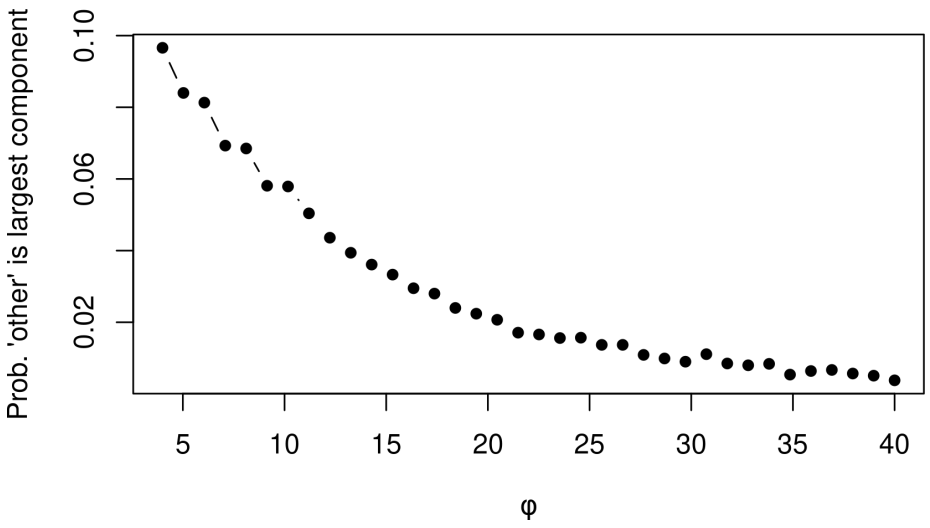

Figure 2: Probabilities of the other component being the largest at different values of the precision parameter.

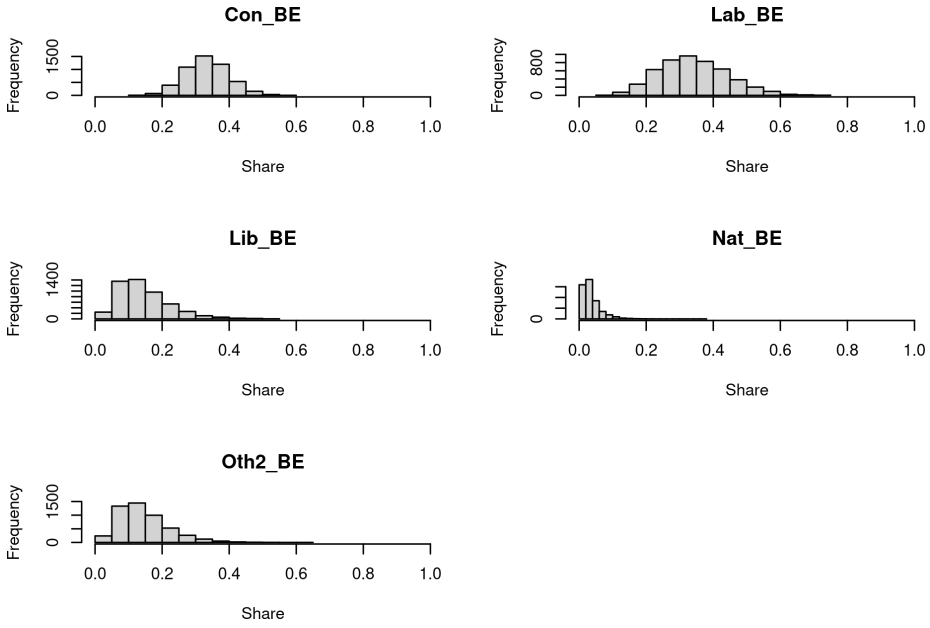

Figure 3: (#fig:ppc_no_data)Prior predictive check, intercepts and precision parameters only

This prior would in many contexts be regarded as informative enough to avoid extreme predictions.[2]

This prior predictive check is shown in Figure 4. The addition of further parameters on a relatively unconstraining scale has pushed the predictions out towards the extremes of the scale, such that the modal prediction for each party is always close to zero, with mean predictions for other parties generally greater than the mean prediction for the Conservatives.

To think through priors for the regression coefficients in the model, I consider two model terms which enter the model in their own right, rather than as part of an interaction term. These are the change in Labour polling, and the presence of a Labour candidate in the by-election.

Consider first changes in Labour polling. Recall that a standard deviation change in Labour polling (one unit change on the standardized variable) is an increase or decrease of seven percentage points. Suppose that Labour polling changes feed through perfectly to by-election results, such that if Labour is up seven points nationally it ought also be up by seven points locally. Suppose further that in the average by-election, the Conservative and Labour shares are as we have assumed above ([0.35, 0.35]). If Labour are up by seven points, and if we know nothing about patterns of party competition except that party vote

[2] https://github.com/stan-dev/stan/wiki/Prior-Choice-Recommendations describes the N(0, 1) prior as a "Generic weakly informative prior".

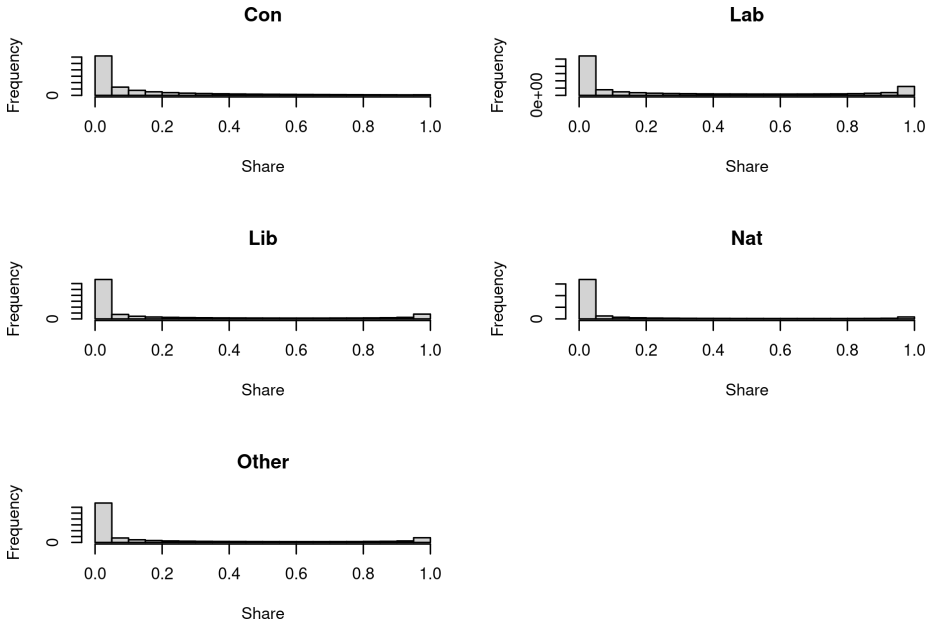

Figure 4: Prior predictive check, full data

shares must sum to one, then it is reasonable to expect that the correlation between changes in the Labour vote share and changes in the Conservative vote share follow the correlation patterns of components of a multinomial distrbution, and that therefore the correlation between the two vote shares p_i and p_j is:

$$-\sqrt{\frac{p_i p_j}{(1-p_i)(1-p_j)}}$$

When both p_i and p_j are 0.35, this evaluates to -0.290. If Labour is up seven points, we might therefore expect the Conservatives to be down by two percentage points. The expected change in the log ratio is therefore:

$$log(\frac{.35+0.07}{0.35-0.02}) = 0.241.$$

This expected change is therefore our best guess as to the likely value of the coefficient on the Labour polling change and the Labour/Conservative log-ratio. In order to produce a prior standard deviation, I note that we would regard it is as very unlikely for an increase in Labour's national polling position to yield a decrease in local vote share, and so the standard deviation should not be greater than 0.12, since otherwise negative effects would be more than surprise events.

If Labour polling changes have implications for the Conservative share of the vote, then they also have implications for the ratio of vote shares of the Liberal party or national parties or other parties to the Conservative share of the vote. A decrease of 2% in the Conservative share of the vote would lead

us to expect an improvement of 0.06 in the log ratios for the remaining parties. Since we would regard it as very unlikely that the effect of Labour polling changes on *other parties* would be greater than the average effect of Labour polling changes *on Labour*, I set the standard deviation on these effects to be $(.241 - 0.06)/2 = 0.0905$.

Now consider the coefficient on the presence of a Labour candidate in the by-election. Suppose that a Labour candidate in an "average seat" (i.e., one with a Labour share of the vote equal to Labour's national share of the vote, assumed to be ~35%) drops out. In this case, the Labour vote decreases to the smallest detectable amount (1/40,000), and the log ratio between Labour and Conservative vote share changes from 0 to -9.5. Because Labour candidacy is almost ubiquitous (Labour only failed to field candidates in three by-elections), the range of the standardized "Labour by-election candidacy" variable is large, as 12.5 units. We might therefore expect the coefficient on this variable to be roughly 0.75.

Considering these two variables guides our prior choice concerning the remaining variables. Candidacy variables, or variables interacted with candidacy, are likely to have large magnitudes for the focal party – in the Labour case, close to one, but possibly larger for other parties which have more dramatic shifts and more balanced patterns of candidacy (nationalist parties, for example, which score zero in England but substantial vote shares in their respective countries). For these candidacy variables a generic weakly informative prior like N(0, 1) seems reasonable when modelling the log ratio of the party in question.

For the remaining variables not involved with candidacy, or for candidacy variables used in modelling the vote shares of other parties (for example: the prior on the effect of Labour candidacy on the Liberal vote) the prior size from the Labour polling change variable shows that substantively large effects may have coefficient values of around 0.25. It would seem extraordinary if the effect of a standard deviation increase in the Retail Price Index had the same effect on the Labour vote share as did a standard deviation increase in Labour's polling, and yet the N(0, 1) prior discussed above places considerable probability on effects much larger than that. For all remaining coefficients, I therefore use a N(0, 0.25) prior.

The prior predictive check for this set of priors is shown in Figure 5. The mode of the distribution for the Conservative party is non-zero (which is desirable), and vote shares of close to one are very infrequent. The distribution for Labour is less well-shaped, as the mode is close to zero and there is a spike in the probability mass at vote shares close to one. For other parties, modal values close to one may seem more reasonable, especially given that for these parties non-candidacy (and a vote share of close to zero) is an ever present risk. Generally the prior predictive distributions are much more plausible than the distributions which result from assigning a N(0, 1) prior to all coefficients.

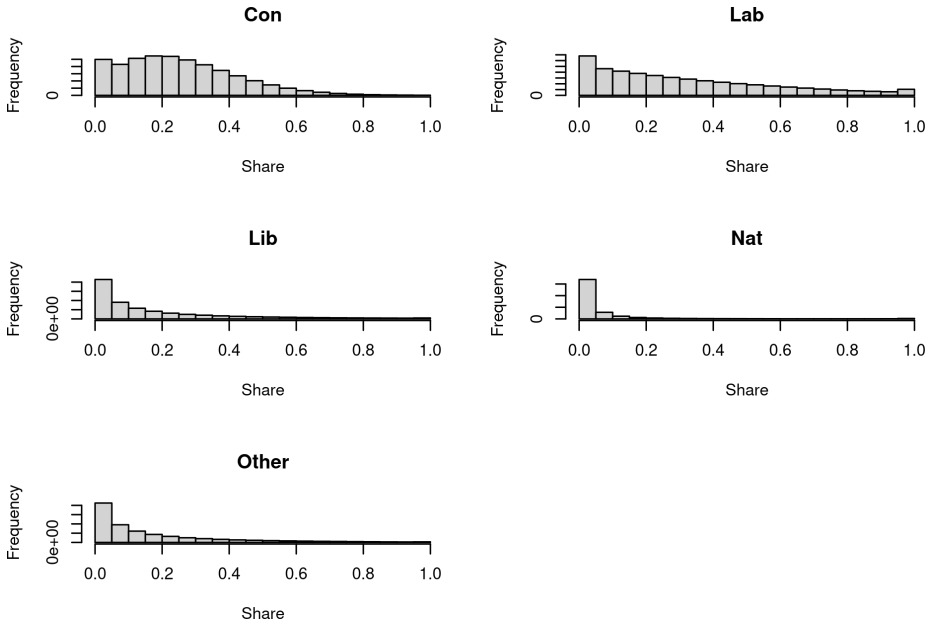

Figure 5: Final set of priors

References

Shugart, Matthew S, and Rein Taagepera. 2017. *Votes from Seats: Logical Models of Electoral Systems*. Cambridge University Press.

Taagepera, Rein, and Mirjam Allik. 2006. "Seat Share Distribution of Parties: Models and Empirical Patterns." *Electoral Studies* 25 (4): 696–713.