2024 general election MRP model post-mortem

Chris Hanretty

2024-07-05

Introduction

In the 2024 general election, the Labour party won a large majority of seats on a relatively small share of the vote. The Conservative party had its worst result in modern history, but won over 100 seats. Our MRP model had predicted a much larger Labour majority on a larger Labour share of the vote. We also predicted that the Conservative party would win fewer than 100 seats. Not only were our MRP predictions wrong, they were over-confident. The model thought that the probability of Labour winning less than 418 seats (the number they won in 1997) was close to zero, and we adopted this probability in our own writing.

The purpose of this note is to examine the accuracy of the MRP predictions, and to disentangle, as far as is possible, whether the failure of the model was due to a failure to estimate national vote shares or a failure to map national patterns onto local patterns. Our preliminary conclusion is that we failed to estimate national vote shares correctly.

Our headline figures

Seat tallies

The primary use of MRP is estimate how many seats each party will win. Table 1 therefore shows our forecast seat tallies against the actual number of seats won in the data we have at present. The results make uncomfortable reading. We over-estimated the Labour seat count by 65 seats, and the actual number of Labour seats was lower than the lower end of our 95% forecast interval. The reverse is true for the Conservatives. The Liberal Democrats slightly exceeded the upper end of our forecast range. Only for the Greens, the SNP and Plaid Cymru were the actual results in our forecast range. We over-estimated the number of seats that Reform would win.

Table 1: Error on seat tallies

Party	Actual	Forecast	Lower end	Upper end	Error
Lab	411	470	439	500	59
Con	121	68	40	100	-53
LDem	71	59	48	70	-12
SNP	10	14	7	23	4
Other	6	0	0	1	-6
Green	4	4	1	7	0
Plaid	4	3	1	5	-1
Reform	4	15	5	27	11

Vote shares

One reason why forecast seat tallies can be wrong is because the national vote shares are wrong. These two things aren't separate in an MRP model: we make predictions at the constituency level and add them up to give national vote shares. However, the national vote shares depend in part on the model and in part on the data used to estimate the model. The national vote shares therefore might be wrong because of the incoming data rather than any deficiencies specific to the model.[1]

Table 2: Error on vote shares

Party	Forecast share	Lower end	Upper end	Actual share	Error
Lab	40.7	38	43	34.7	6.0
Con	22.5	20	24	24.4	-1.9
Reform	14.5	12	17	14.7	-0.2
LDem	11.0	9	13	12.5	-1.5
Green	5.5	4	7	6.9	-1.4
Other	2.7	2	4	3.5	-0.8
SNP	2.6	2	3	2.5	0.1
Plaid	0.6	0	1	0.7	-0.1

Table 2 therefore shows the error on our national vote shares. We overestimated the Labour share of the vote dramatically, being almost six points off. We also underestimated the Conservative party share of the vote, meaning that our error on the Lab - Con gap was off by

[1] Again, these two things aren't separate – MRP is supposed to account for sample non-representativeness by adjusting for factors which affect non-response and vote choice – but we have found it helpful to separate these two factors out.

eight percentage points. We got the Reform share of the vote almost exactly right, and so our over-estimation of their seat tally could be a result of our under-estimation of the Conservative vote or of problems in modelling the geographical distribution of the party's vote share. We slightly underestimated the Liberal Democrat share of the vote, which might also explain why that party's seat share was under-estimated.

Accuracy

We use two measures of the accuracy of our seat-level predictions: root mean squared error (RMSE) and a multi-class Brier score. We also assess our errors visually by plotting our forecast against the result.

Visual assessment

Figure 1 shows our forecasts against the observed vote shares, for all seats for which we have data. A black solid line shows the best-fitting straight line through these points; the grey dotted line shows what would happen if there were a perfect one-to-one correspondence between our predictions and the observed vote shares.

The best-fitting line for Labour is always below and to the right of the dotted line indicating a 1:1 correspondence. This means that our predictions for Labour were on average too high, for all levels of Labour support. For the Conservatives, we underestimate the party's vote shares for almost all seats; in a small number of seats where the party's vote share is low we overestimated their share. The trend lines for the Green party and the SNP show that we underestimated these parties where they were wrong. The same pattern is present in the Reform vote, although not to the same extent. There were no particular problems with the Liberal Democrats or Plaid Cymru.

Over-estimating a party's strength where it is weak and under-estimating it where it is strong is a hallmark of *attenuation*. Attenuation is a particular problem in MRP models, and has led to a number of attempts to correct for attenuation ("unwinding"). We did not use any unwinding algorithm, and so we cannot comment on whether unwinding in general is a good idea. We can, however, show the pattern clearly in Figure 2, which shows the error as a function of the predicted vote share.

Accuracy on RMSE

Root mean squared error is similar to the mean absolute error ("how off were you on average across all parties and all seats?"), but penalizes bigger mistakes more. RMSE is the measure that you would adopt if you care equally about all vote shares, no matter whether they are the

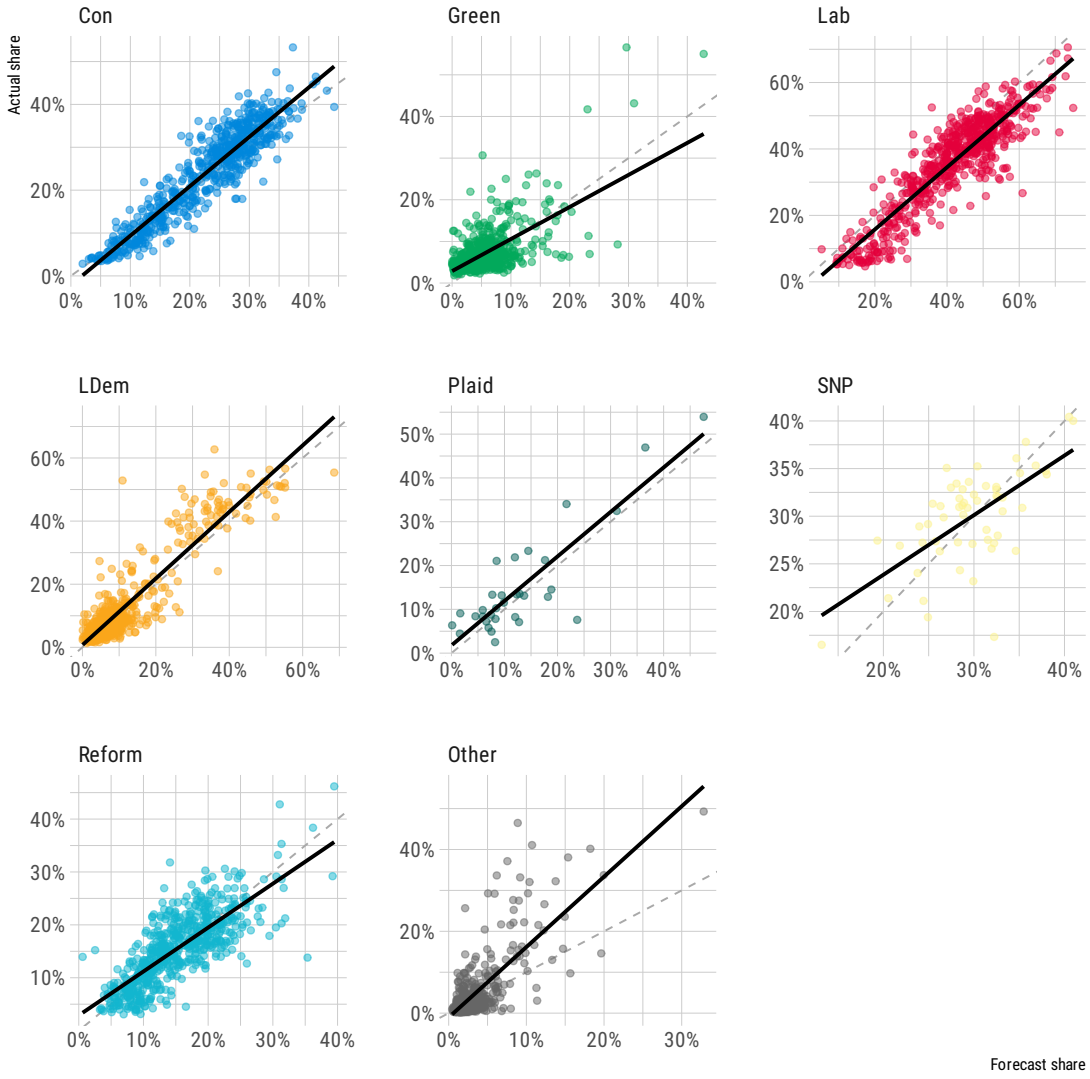

Figure 1: Scatter plot of forecast (horizontal axis) against observed vote shares (vertical axis), faceted by party

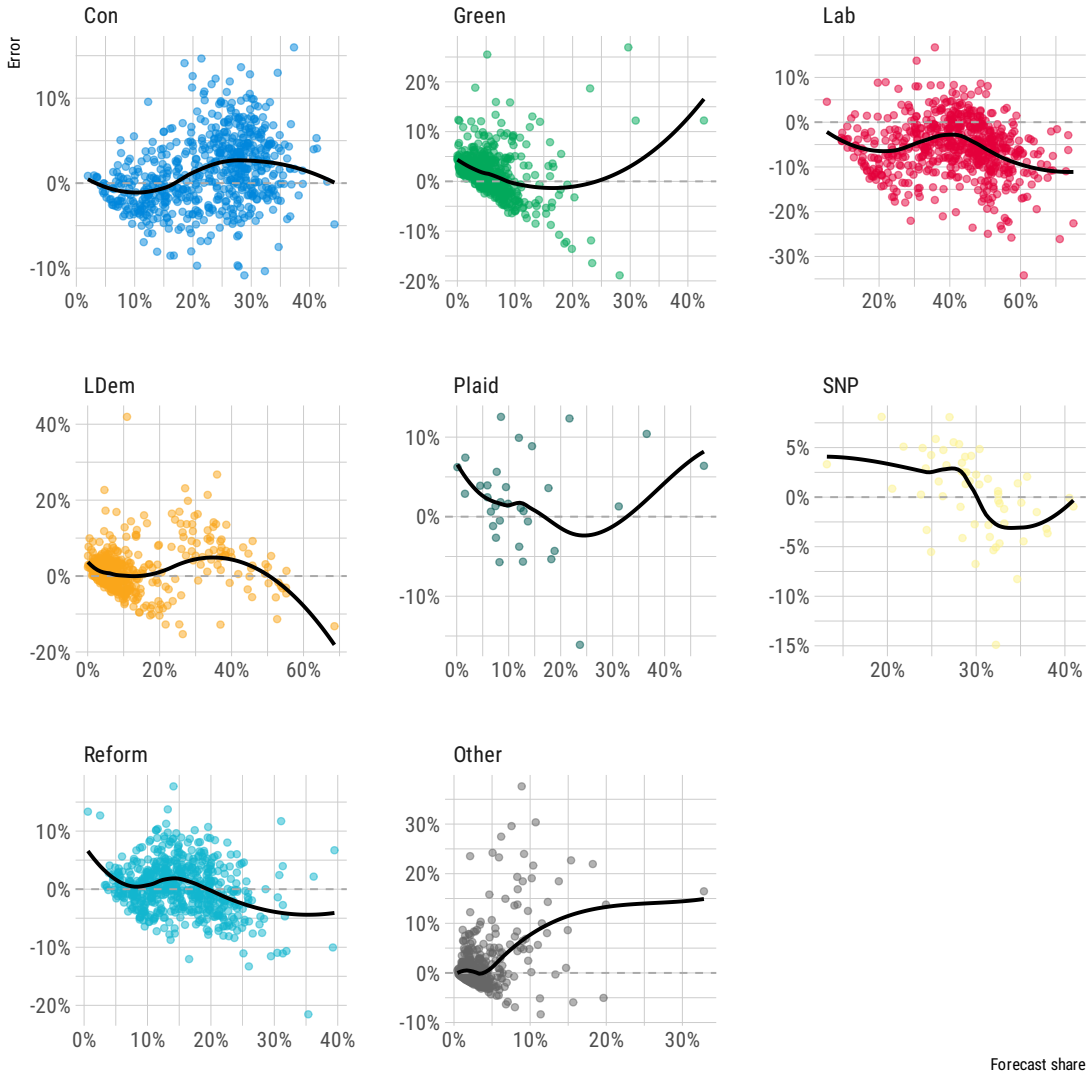

Figure 2: Scatter plot of forecast error (vertical axis) as a function of forecast vote shares, faceted by party

vote shares of a party which is well-placed to win a seat or a party which is in fourth or fifth place. It is calculated as follows:

$$RMSE = \sqrt{\frac{1}{632}\frac{1}{8}\sum_{i=1}^{I=632}\sum_{p=1}^{P=8}(f_{i,p} - o_{i,p})^2}$$

where i keeps track of the 632 constituencies in Great Britain, where p keeps track of the eight modelled parties (Conservatives, Labour, Liberal Democrats, SNP, Plaid Cymru, Reform, Greens, and "all others"), and where $f_{i,p}$ stands for our forecast vote share for party p in seat i, and $o_{i,p}$ stands for the observed vote share of that party in that same seat.

Table 3 gives an example of this calculation for Aldershot. Note that the forecasts and observed shares are all on the same 0-1 scale rather than the 0-100 scale, but that the final figure for RMSE has been multiplied by 100 for ease of reading.

Table 3: Root mean square error calculations for Aldershot

Party	forecast	observed	Error	Squared error	Sum sq. errs	RMSE
Con	0.3090271	0.2900667	0.0189604	0.0003595	0.0033229	5.764475
Green	0.0311611	0.0443927	-0.0132316	0.0001751		
LDem	0.0971084	0.0834707	0.0136378	0.0001860		
Lab	0.3600941	0.4071358	-0.0470417	0.0022129		
Other	0.0178474	0.0058092	0.0120383	0.0001449		
Reform	0.1847618	0.1691249	0.0156369	0.0002445		

Table 4 gives this calculation overall and by party, together with a comparison with the equivalent figures for 2019.

Table 4: Root mean square error overall and by party, with a comparison to 2019. 2019 figure for Reform is the figure for the Brexit Party.

Party	RMSE	RMSE (2019)
Overall	5.55	
Con	4.39	4.69
Green	4.54	1.77
Lab	8.36	4.78
LDem	5.31	3.58
Plaid	6.22	4.06
SNP	4.26	7.27
Reform	4.34	2.89

Table 4: Root mean square error overall and by party, with a comparison to 2019. 2019 figure for Reform is the figure for the Brexit Party.

Party	RMSE	RMSE (2019)
Other	5.24	2.46

The table shows that our estimates of Conservative vote share were actually more accurate than our estimates in 2019. The same is true of our estimates for the SNP. However, we made bad errors on the Labour shares and on the Liberal Democrats. We also made bad errors on "all others", although this figure is particularly affected by a small number of notable candidates.

Accuracy on Brier score

The multi-class Brier score is a measure of predictive accuracy for categorical outcomes. In our case, we use it to measure how good our probabilistic forecasts of victory in each seat were. This is the measure you would use if you only cared about your ability to predict winners rather than your ability to predict the seat shares that lead them to victory. The Brier score is different to the percentage of seats correctly predicted because it rewards confident successful predictions and penalizes confident mistakes. It is calculated as follows:

$$Brier = \frac{1}{632} \sum_{i=1}^{I=632} \sum_{p=1}^{8} (f_{i,p} - o_{i,p})^2$$

where f stands for the forecast probability of victory for party p in some seat i, and where o stands for whether or not that party won that seat, with o having a value of one if the party won and a value of zero if the party list. Lower values of the Brier score are better scores; a score of zero is a perfect score.

The value of the Brier score for our forecast was 0.243; the percentage of seats correctly predicted was 83.8. Neither of these numbers is good.

Accuracy in terms of RMSE across constituencies

Figure 3 gives a map of root mean squared error by constituency. What are noticeable are odds spots of error, which are themselves listed in Table 5. Most of these constituencies involve a strong independent candidate. One exception is North Shropshire, which we had a

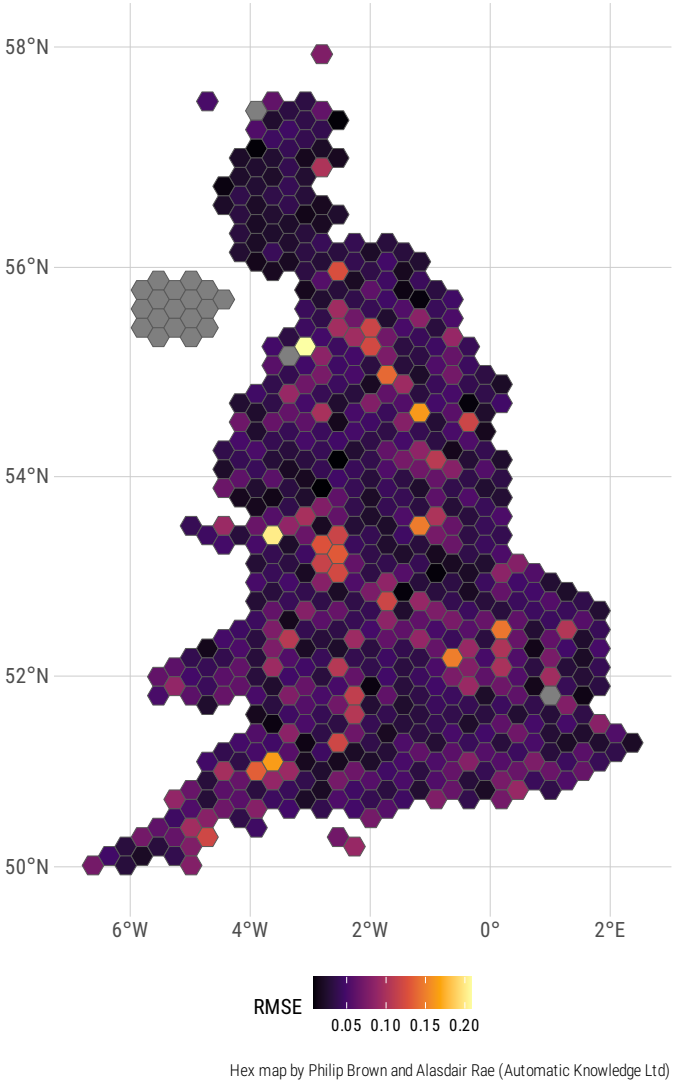

Figure 3: Hex map of RMSE

hard time modelling as a result of the marked difference between the by-election result and the preceding general election result, which formed the basis for the post-stratification frame.

Table 5: Top ten seats by RMSE

PCON24CD	Seat	RMSE
E14001102	Blackburn	0.2083440
E14001398	North Shropshire	0.1998838
E14001132	Bristol East	0.1635204
E14001196	Dewsbury and Batley	0.1620704
E14001086	Bethnal Green and Stepney	0.1487533
E14001327	Leicester South	0.1476362

Figure 4 shows the major predictors of RMSE at constituency level, according to their partial correlation with the error. The plot shows that we were more accurate in areas with more people with entry-level qualifications, with Labour councillors, in high income areas, and where lots of people are economically active. We are less accurate in areas with a greater ethnic minority population, where the "Other" share of the vote in 2019 was greater, where there are more people with level 4 qualifications or above, and where there are more people who do not own their accommodation.

Improvements over UNS

We used an MRP model in order to estimate vote shares at the constituency level. One alternative to MRP is to use a uniform national swing. We would hope that the predictions from our MRP model would be more accurate than using the national swings implied by our poll. Although the MRP model might have performed poorly in an absolute sense, it might have performed well relative to what we might have predicted had we gone with UNS.

In order to calculate seat level outcomes under UNS, we use the model-derived vote shares in Scotland, England and Wales. It might seem strange using outputs from the MRP model as part of a comparison with uniform national swing. However, our results would not change if we were to take subsamples from the most recent national Survation polling.

Table 6: Root mean square error overall and by party under UNS

Party	RMSE
Overall	7.61
Con	8.15
Green	4.49

Table 6: Root mean square error overall and by party under UNS

Party	RMSE
LDem	6.49
Lab	11.34
Other	7.64
Plaid	3.84
Reform	5.98
SNP	3.60

Table 6 shows that our root mean squared error is substantially higher under UNS. We can therefore be confident in saying that our forecasts of constituency vote share would have been more inaccurate had we used uniform national swing.

Table 7: Seat tallies under UNS

party	Forecast	Actual	Error
Con	195	121	-74
Green	1	4	3
LDem	36	71	35
Lab	379	411	32
Other	3	6	3
Plaid	3	4	1
SNP	15	10	-5

The seat tallies from uniform national swing would also have been less accurate than the seat tallies from MRP, although here the difference is less marked. The sum of absolute errors from Table 7 is slightly higher than the corresponding sum from Table 1.

Turnout filters

In our MRP model we model "not turning out to vote" as a choice on a par with choosing a party to vote for. Those who won't turn out to vote are, for us, those who, when asked to report how likely they are to turn out to vote on a scale from zero to one, give a figure of less than eight. This threshold therefore affects the model by changing the data fed to it.

Our full model takes hours to run, and so it is not feasible in the short term to re-run the model with different threshold values. We do, however, have a quicker approximation to the model,

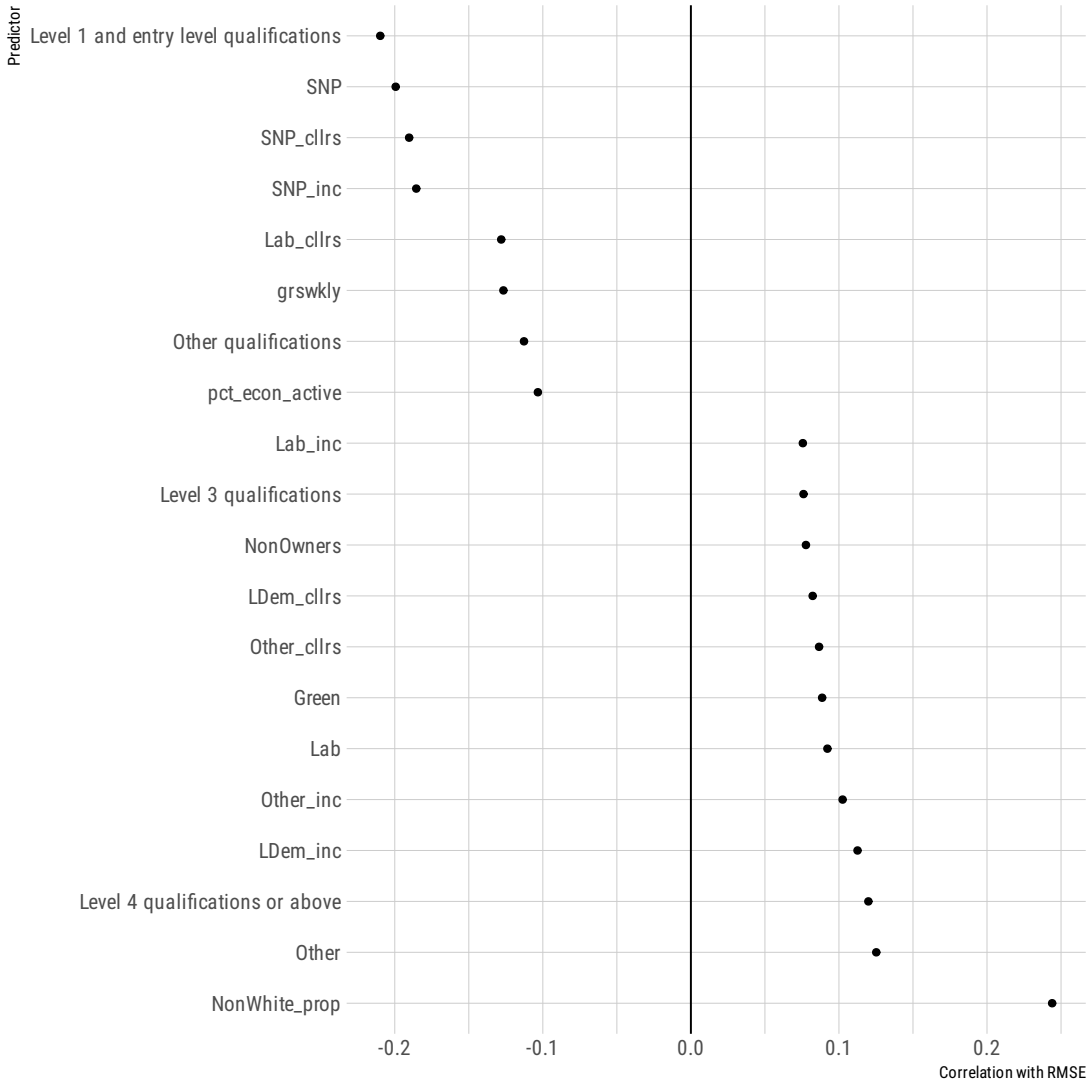

Figure 4: Strongest associations with RMSE at constituency level

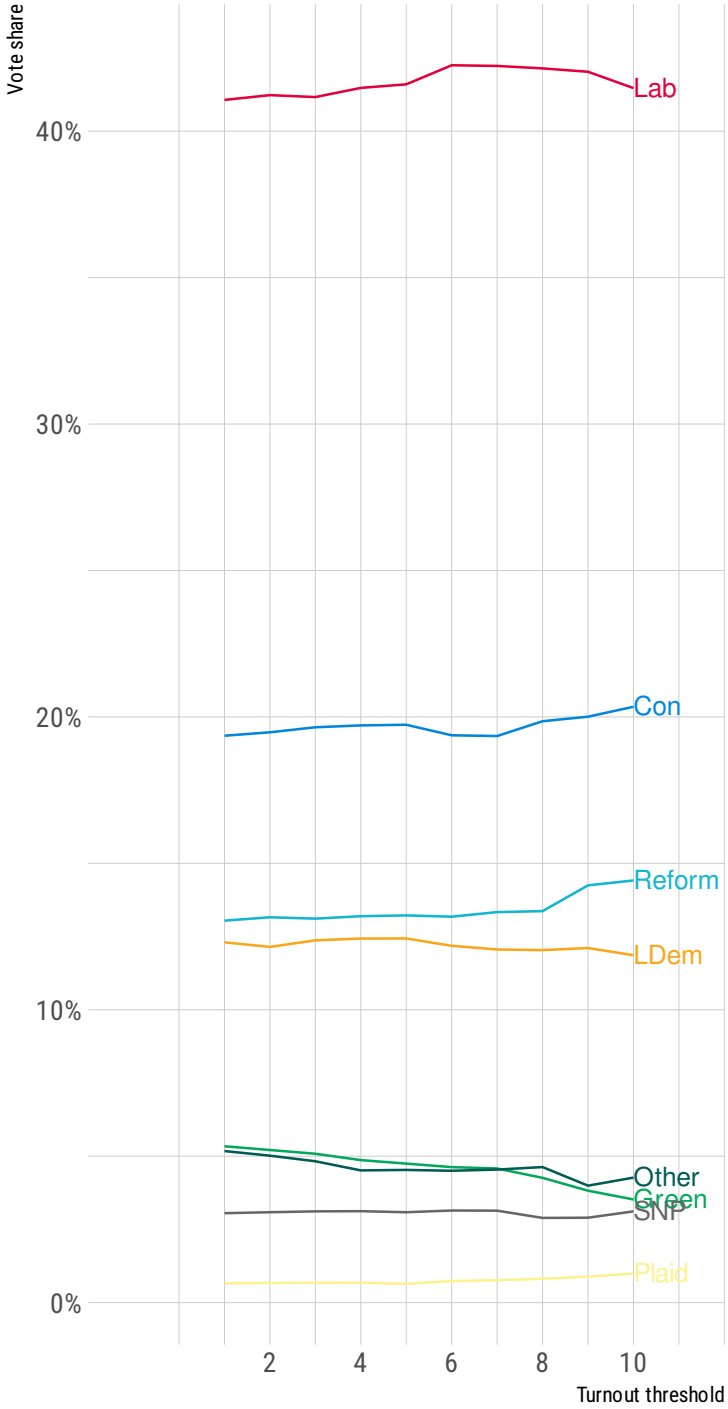

Figure 5: Modelled vote share at different turnout thresholds

which runs in minutes, not hours. We therefore re-ran the model, calculating national vote shares at different turnout thresholds. The results are shown in Figure 5.

As the figure shows, the effects of different turnout thresholds are negligible. The estimated Conservative share of the vote increases slightly when we raise the turnout threshold, but the effects are tiny compared to the overall underestimation of the Conservative.

Data recency

In our model we include data from several fieldwork periods (batches of five or six days). We include data from several fieldwork periods because it is not possible to collect large (> 20,000) representative samples over the course of a small number of days, and because MRP requires large samples. When we model, we allow for effects of time. We include a random walk, which allows for the "effect" of each fieldwork period to depend on the effect of the last fieldwork period, plus a small increment. This should, in theory, allow us to capture level shifts in support for different parties. It would not, however, allow us to capture interactions between fieldwork period and particular characteristics. If, for example, the relationship between past vote and vote intention changed over time, such that past Labour voters became more likely to vote Green *over time*, then our model would capture some average of the effect of having voted Labour on voting Green before and after this putative switch.

We therefore re-estimated our model approximation, cutting our data down from all seven fieldwork periods in the data to between six and one fieldwork periods. The results are shown in Figure 6.

The figure shows that whilst the Labour and vote share is largely constant over time, the Conservative vote share increases the more (older) data we include. Data recency does not, therefore, explain the over-estimation of the Labour vote, and using more recent data alone would under-estimate the Conservative vote share by even more than we already do.

Does the model do well when given the true vote shares?

Given that we now know the results of the election, and thus the true vote shares, we can reweight the data to match the known vote shares. We assign each observation a weight which is equal to the actual vote share for the respondent's preferred party, divided by the forecast vote share for the respondent's preferred party. Where we underestimated a party, we up-weight observations; where we overestimated a party, we down-weight observations. Respondents who did not intend to vote or who did not pass the turnout filter were given a weight of one. We then re-run the model and rely on the model picking up the changes in party support through the model intercept for each party.

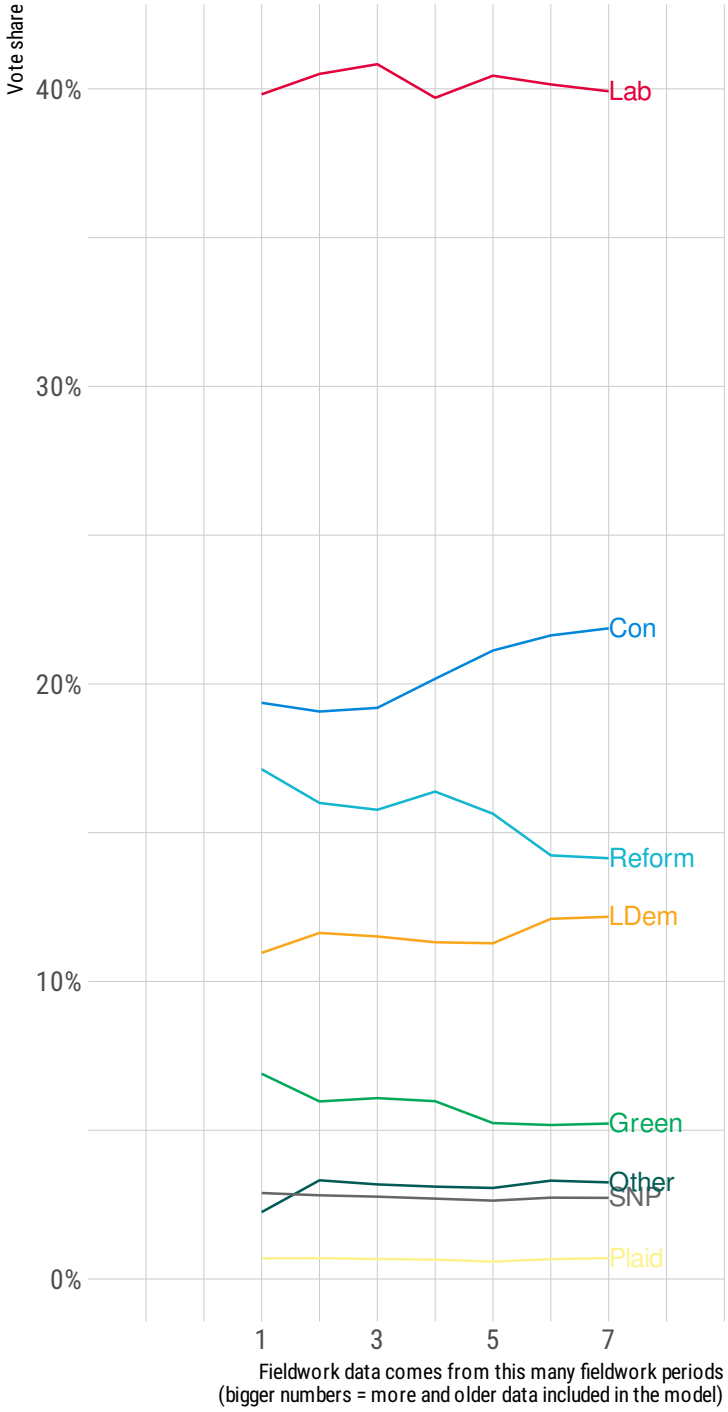

Figure 6: Modelled vote share using different lengths of fieldwork

When we do this the model predicts results that are very close to the final result, and close also to the results of the exit poll, which forecast a greater number of Reform seats than they eventually won. Generally the model forecasts a higher number of seats for smaller parties (Greens, Reform) than were in fact won, but all results save that for Reform are within the 95% forecast intervals.

Table 8: Seat predictions after reweighting to match known shares

Party	Avg seats	Median seats	Lower end	Upper end
Lab	405.5	406	366	448
Con	112.8	112	72	154
LDem	65.8	66	55	79
Reform	18.0	17	6	33
SNP	15.5	15	7	27
Green	8.2	8	3	14
Plaid	3.9	4	2	6
Other	2.4	2	1	5

Figure 7 shows the forecast seat shares from this model compared to the actual results. For the three largest parties by seats won, there is no evidence of attenuation, or underprediction of vote share at high levels. There is evidence of attenuation for the Green party, and for Reform. We believe this is a consequence of our lower ability to identify where these parties are competitive, particularly with the Reform party, where the absence of any past track record unaffected by stand-down arrangements makes it difficult to link past results for related party and current success.

Table 9: Root mean square error overall and by party given known vote shares, for an MRP model and for uniform national swing

Party	RMSE	RMSE (UNS)
Overall	5.08	7.07
Con	4.31	7.92
Green	4.89	4.18
LDem	5.37	6.38
Lab	6.30	9.59
Other	4.90	7.50
Plaid	6.21	3.44
Reform	4.41	5.94
SNP	3.93	3.59

The root mean squared error for a model with known vote shares is also substantially reduced,

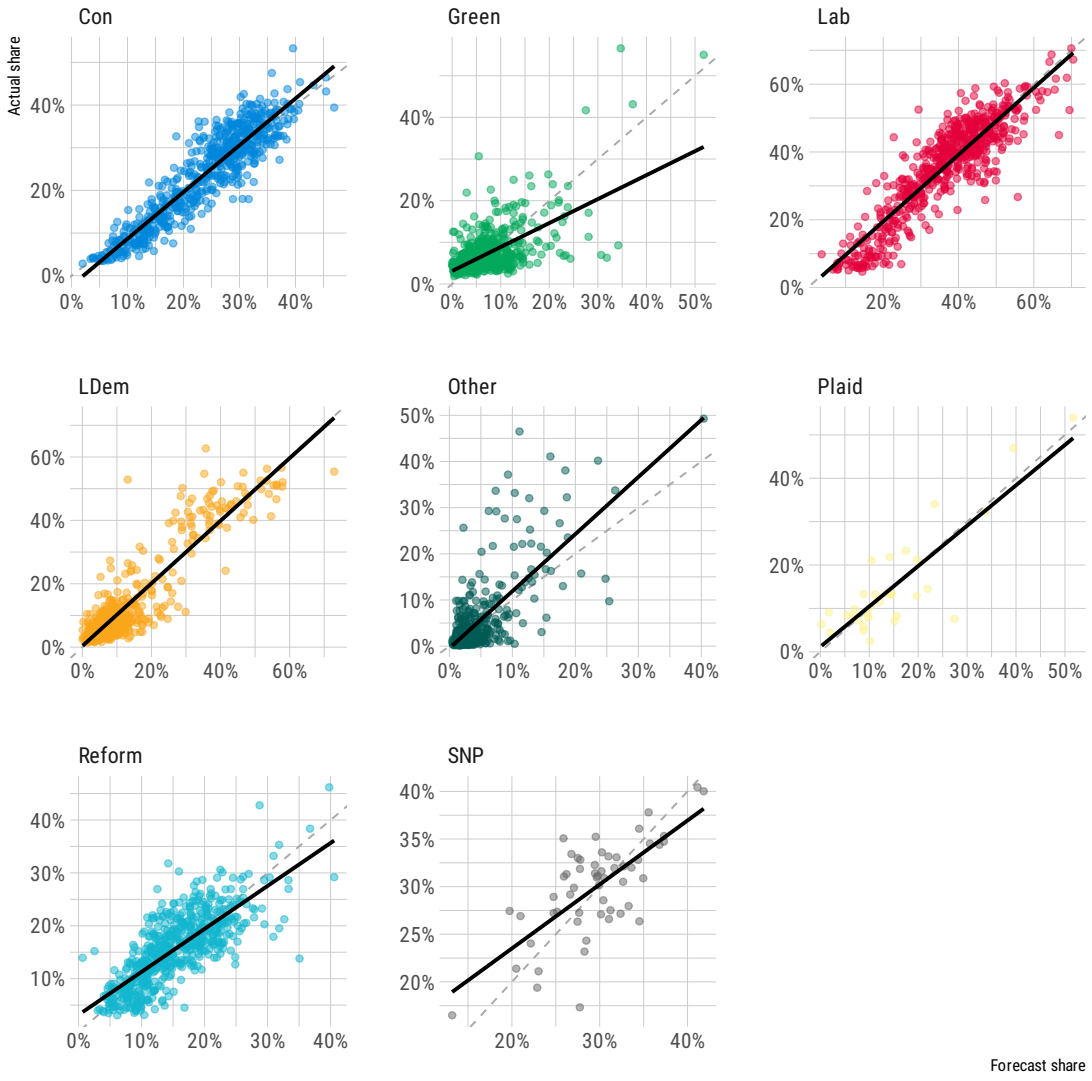

Figure 7: Scatter plot of forecast (horizontal axis) against observed vote shares (vertical axis), faceted by party, under a model where observations are reweighted to match known vote shares

although here the errors are larger than they were in 2019, when we were broadly accurate on the headline figures.

What is important to note that the root mean square errors under the MRP model are much, much smaller than the root mean square errors assuming a uniform national swing within England, Wales and Scotland. This is shown in Table 9. The difference between the success of the MRP model given known shares and UNS is so stark that we cannot see how "unwinding" predictions to resemble the predictions of uniform national swing more closely would improve the accuracy of the seat predictions. Although "unwinding" may have helped some companies achieve a more accurate headline seat figure, we believe that unwinding masks problems in the original composition of the sample. Phrased differently, errors in sample composition and the translation of votes to seats partially offset each other.