PollBasePro: Daily British Post-War Voting Intentions[*]

Abstract

Students of British politics often use polling data to test their theories. Yet these data present some difficulties. Polls are noisy, occur at irregular intervals, and measure both voting intention and survey design choices. By pooling the polls together, we can mitigate the worst of these problems. To this end, we introduce a new dataset: PollBasePro. Drawing on decades' worth of data, spanning 24,277 daily voting intention estimates for Britain's three largest parties, from 26 May 1955 to 11 November 2021. Out estimates track British political history well and offer new insights on post-war British politics. As such, we expect them to offer scholars of British politics a degree of specificity and flexibility beyond that of any existing data.

Introduction

When it comes to polling data, students of British politics are more fortunate than most. Outside of the US, Britain is perhaps the most-polled country in the world. Yet, like all data, polls present some difficulties. First, polling figures include some noise due to random sampling variation. Second, polls occur at irregular intervals and can be hard to match to more regular data. And, third, polls measure changes in voting intention but are subject to systematic bias from the design choices that the companies that run them make. For a single poll, these issues are all but unavoidable. But by pooling the polls together, we can separate the signal from the noise (Jackman 2005).

To this end, we introduce a new dataset—PollBasePro—that includes daily voting intention estimates for each of Britain's three largest parties. Our estimates span 24,277 days, from 26 May 1955 to 11 November 2021. To inform them, we rely on a rich dataset of almost all post-war British voting intention polls. As a result, the PollBasePro data track British political history well, suggesting that they are both reliable and provide us with the ability to identify discrete events. Further, given its size, we expect our data set to permit a degree of specificity and flexibility beyond that of any existing British voting intentions measure.

In the sections that follow, we elaborate on PollBasePro in more detail. First, we describe our underlying polling data. Next, we discuss in detail the methods that we use to derive our estimates. We then summarise our dataset and consider what PollBasePro tells us about British politics since 1955. Finally, we offer some initial conclusions on the data and the study of British politics then remark on how we will develop PollBasePro in the coming years.

Source Data

Two data sources inform our estimates. First, the PollBase dataset of historic British voting intention polls (Pack 2021). Second, data compiled by volunteers on Wikipedia (see, for example, Wikipedia 2021). Both are comprehensive, high-quality, and track British voting intention over the past several decades.

The PollBase data have many sources (Pack 2021). These include books published after each general election, polling almanacs, contemporary media reports, and figures from polling company websites[1]. The data since the 1983 general election are almost complete, barring rare individual errors. Before 1983, the data is likewise complete for each general election campaign period. For periods between general elections before 1983, coverage is more comprehensive for some pollsters than others. Gallup and National Opinion Poll, in particular, are well-covered as the two companies publish their results. The data start in June 1945, however, our estimates begin at the 1955 general election as data before this point are scarce.

We use publicly-collected data from Wikipedia to cover the period from 2010 onwards. Volunteers update these data in real time and include sample sizes that are missing from PollBase. This is important as our model assumes that we know the sampling error present in each estimate and this data is missing from PollBase. While some might doubt Wikipedia's reliability, we do not think that it is an issue. Polling figures are verifiable and likely of interest only to a very small group of people. Further, almost all figures on the website including links to source material that corroborate them. What's more, past research shows that Wikipedia articles are reliable too, at least compared to sources like the *Encyclopaedia Britannica* (Giles 2005).

Estimation

Polls are an essential tool in political science. They allow us to test theories that we otherwise could not. Yet polling data nevertheless face three problems.

First, polls are noisy. We cannot interview an entire population. As such, we use sampling methodologies to interview a smaller number that we can then use to make estimates about the target population. Importantly, these estimates are probabilistic and include error. This error affects models that test political theories in one of two ways. Where we use polls as outcomes, it reduces our statistical power. Where we use them as predictors, it biases real effects towards zero.

Second, polling data occur at irregular intervals. In fact, they may cover any period of time from a single day to several weeks. Too often we assume that polls measure public opinion on

---

[1] For a list of polling companies included in the data, plus relevant statistics, see table A1 in the appendix.

the final fieldwork date. Of course, this is almost never true. As far as assumptions go, this is not an unreasonable one. But it does ignore that events occur partway through the data collection process. Likewise, these irregular time intervals can also make polls difficult to match to other, more regular, time series data.

Third, polls are the product of more than opinion alone. Polls are subject to systematic biases due to the design choices of the companies that run them. Like any bias, it is a problem as it obscures the measurement of the phenomenon of interest. In the past, these biases have been so large that they have cast doubt on the efficacy of the entire polling industry. At the 2015 UK general election, most polls suggested that Labour had a good chance of becoming the largest party. But, on the night, the Conservatives won a small majority instead, leading to an industry-wide enquiry (Sturgis et al. 2018; Prosser and Mellon 2018; Mellon and Prosser 2017).

Our intention is simple: to improve estimates by accounting for these issues. To do so, we adapt the method outlined in Jackman (2005). Jackman's model has estimates for a given party start and end at known results from any given pair of elections. It then treats the party's level of support in the intervening period as a random walk between these two known points. On any given day, the party's support depends on its support the day before, pollster-specific "house" effects, and random shocks. Others have already used this method to good effect. Jackman and Mansillo (2020; 2018), for instance, use it to model voting intention in Australia. Likewise, Louwerse (2016) use it to estimate aggregate voting intention in Ireland.s

Imputing Missing Sample Sizes

Jackman's (2005) approach requires that we know each poll's sample size. But our data do not include sample sizes for most polls before the 2010 general election. To solve this problem, we use data from Jennings and Wlezien's (2016) "Timeline of Elections" dataset. Though less comprehensive than PollBase, these data do include sample sizes and, furthermore, from countries other than Britain.

We model the sample size of poll i in the Timeline data, n_i, as Poisson-distributed according to the rate parameter, λ_i. We then fit a simple linear function to the logarithm of this parameter

that includes an intercept, $\alpha_{Country}$, and a slope on the effect of time, $\beta_{Country}$, both of which we allow to vary over countries. We then relate these two parameters to one another by modelling them as though they come from a multivariate normal distribution. This allows the parameters to be correlated and, thus, to share information. The model is as follows:

$$n_i \sim \text{Poisson}(\lambda_i) \qquad \text{Likelihood function}$$

$$log(\lambda_i) = \alpha_{Country[i]} + \beta_{Country[i]}T_i \qquad \text{Linear model on } \lambda$$

$$\begin{bmatrix} \alpha_{Country} \\ \beta_{Country} \end{bmatrix} \sim \text{MVNormal}(\begin{bmatrix} \alpha \\ \beta \end{bmatrix}, \mathbf{S}) \qquad \text{Multivariate prior on varying effects}$$

$$\mathbf{S} = \begin{pmatrix} \sigma_\alpha & 0 \\ 0 & \sigma_\beta \end{pmatrix} \mathbf{R} \begin{pmatrix} \sigma_\alpha & 0 \\ 0 & \sigma_\beta \end{pmatrix} \qquad \text{Covariance matrix on varying effects}$$

Our choice to use all data from all countries in the Timeline data is a prudent one. The dataset does not contain sample sizes for British polls conducted before the early 1960s, but it does contain values for other countries as early as the mid-1940s. Pooling all available information for all countries across the entire time series, thus, allows us to impute reliable estimates of likely sample sizes in Britain across the full range of dates by drawing on persistent differences between polls in Britain and polls in all other countries in the data.

Figure 1 shows our model's estimate of the likely sample size of the average British voting intention poll between 1955, the first year in our data, and 2013, the point at which known sample sizes become available. The model estimates that the average British voting intention poll included around 1,198 respondents in 1955. By 2013, the model suggests that this value had increased by 946 to 2,144 respondents per poll, on average. Or, more simply, that sample sizes for British public opinion polls have increased over time.

We use the model to produce a time series of estimated sample sizes between 1955 and 2013. This includes all dates for which we intend to produce a voting intention estimate. Where our
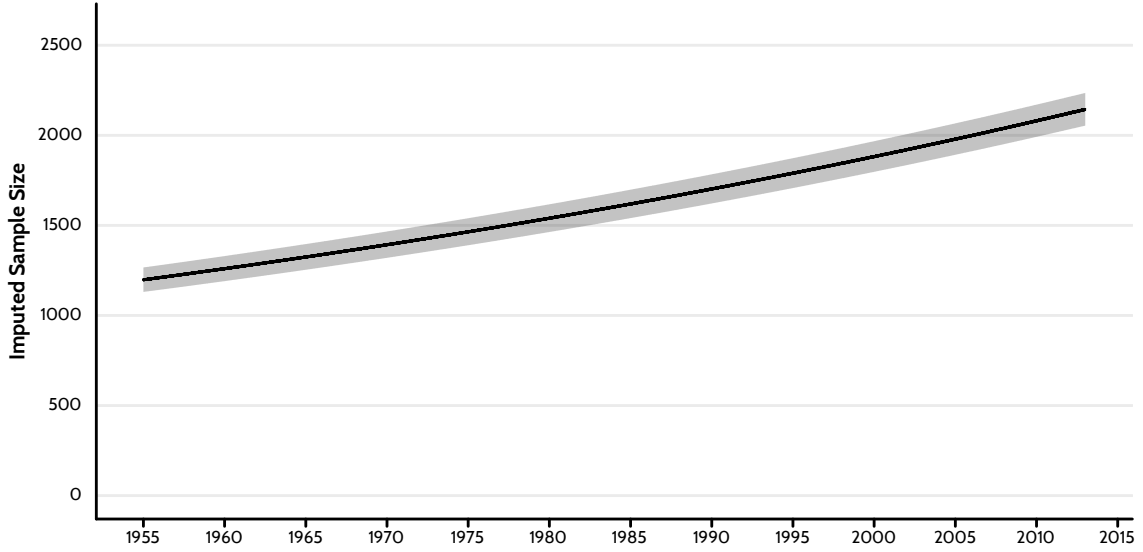
Figure 1: Imputed sample sizes in Britain between 1955 and 2013, estimated using sample size data from Jennings and Wlezien's "Timeline of Elections" dataset (2016).

polling data come from before the 2010 general election, or are otherwise missing, will fill in the gaps with these imputed values. To do so we match our polling data to the imputed values from the model based on their respective dates.

Estimating Daily Voting Intention Figures

As Jackman's (2005) model is complex, we build it up step-by-step. We assume that each voting intention estimate in our data, y_i, is normally-distributed according to two parameters. The first is some mean, μ_i, of the source polling estimates. The second is some error that leads each estimate to deviate from the expected value. In many models, this error parameter would measure only random residual error. But as each voting intention estimate is a proportion, we can use the equation for the standard error of a proportion to compute the error in the estimate due to random sampling variation, where $S_i = \sqrt{\frac{y_i(1-y_i)}{\nu_i}}$. Note that ν_i is the sample size of y_i, n_i, divided by the number of days the poll spent in the field, k_i. Thus, we assume that an equal number of people took each poll on each day that it was in the field[2]. We can then include both in our model to account for any known error, S_i, and any random residual error, σ:

---

[2]We think that this is a fair assumption given an absence of any other information. Mansillo and Jackman (2020) also make the same assumption, and note that it also produces smoother estimates than using the median field date.

$$y_i \sim \text{Normal}(\mu_i, \sqrt{\sigma^2 + S_i^2}) \hspace{3cm} \text{Likelihood function}$$

The next step is to fit a linear function to μ_i. This allows us to decompose the variance and produce an estimate of the electorate's *latent* voting intention on each day. We assume that μ_i is a linear function of two variables: $\alpha_{Day[i]}$, the electorate's latent voting intention for y_i on the day that it was fielded, and $\delta_{Pollster[i]}$, the persistent "house effects" that arise due to the survey methodological and design choices that inform how the company that ran the poll collected its data. Updating our model specification to include these assumptions, gives the following:

$$y_i \sim \text{Normal}(\mu_i, \sqrt{\sigma^2 + S_i^2}) \hspace{2cm} \text{Likelihood function}$$

$$\mu_i = \alpha_{Day[i]} + \delta_{Pollster[i]} \hspace{2cm} \text{Measurement model on } \mu$$

At present, all values of α_{Day} are independent, throwing away valuable information. Instead, we assume that voting intention today is most similar to voting intention yesterday and tomorrow. This links our estimates and informs any days that lack data of their own. We constrain α_1 to equal the vote share that a given party received at a given election. We also constrain α_T to equal the vote share that the same party received at the following election. Next, we fit a dynamic model to α_t for all other days in our time series. This acts as a sort of "chain" that links together all values of α[3]. Thus, when the value of one estimate changes during the model fitting process, so too do all others. The model assumes that α_t is equal to α_{t-1}, plus any random shocks that take place between the two days, ω_{t-1}. These random shock parameters are themselves scaled according to τ. This gives:

---

[3]More technically, this "chain" is an example of a Weiner process.

$$y_i \sim \text{Normal}(\mu_i, \sqrt{\sigma^2 + S_i^2}) \qquad \text{Likelihood function}$$

$$\mu_i = \alpha_{Day[i]} + \delta_{Pollster[i]} \qquad \text{Measurement model on } \mu$$

$$\alpha_t = \alpha_{t-1} + \tau\omega_{t-1} \text{ for } t \text{ in } 2, ..., T-1 \qquad \text{Dynamic model on } \alpha_t$$

$$\alpha_T \sim \text{Normal}(\alpha_{T-1}, \tau) \qquad \text{Adaptive prior on } \alpha_T$$

After specifying our model, we loop over our data and fit the model to each pair of elections for each party. We do this for every election pair between 1955 and the present for each party. Note that the most recent election by definition has no subsequent election. In this case, we leave α_T unconstrained and estimate its value from the data. Users should, thus, treat any estimates between the most recent election and the present as provisional.

Validation

We validate our data against Jennings and Wlezien's "Timeline of Elections" dataset (2016). These data contain 4,302 pooled polls from Britain from 15 June 1943 to 6 June 2017. Given that our data are so comprehensive, it is likely that most polls appear in both datasets[4]. Even so, since Jennings and Wlezien compiled their data independently to us, they provide a good test case against which to validate our data. As figure 2 shows, our estimates are well validated. Correlations between the two series are strong and positive. Their mean absolute error (MAE) and root-mean-square error (RMSE) are also low in all cases. The Conservatives showed a correlation of 0.94 (95% CI: 0.94 to 0.94), an MAE of 2.16 percentage points, and an RMSE of 0.03; Labour, a correlation of 0.93 (95% CI: 0.92 to 0.93), an MAE of 2.75 points, and an RMSE of 0.03; and the Liberals, a correlation of 0.95 (95% CI: 0.95 to 0.95), an MAE of 1.41 points, and an RMSE of 0.02.

---

[4]Note, however, that the Timeline data are not strictly polls but instead small-scale poll aggregations. As such, it is not possible to compute the degree of overlap.
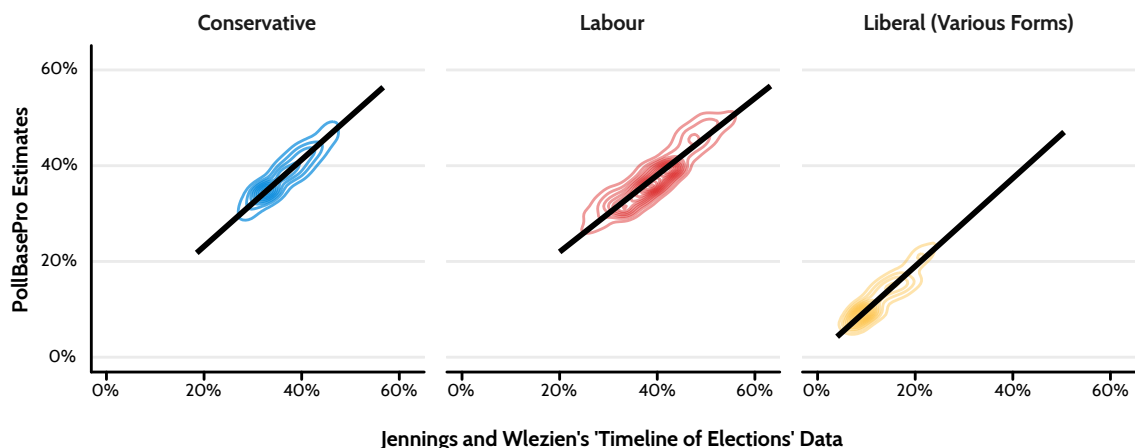
Figure 2: In all cases, estimates from PollBasePro appear well-validated when compared to raw polling data from Jennings and Wlezien's 'Timeline of Elections' dataset (2016).

PollBasePro and British Politics Since 1955

Figure 3 shows that our estimates track British political history well. From 1955 to 1980, we see the heyday of the two-party system. Here, around 43% of the electorate supported Labour and another 44% the Conservatives. In the 1980s, we see the rise and fall of the SDP-Liberal alliance. Labour's slow rise to power in 1997 soon follows, as does their loss of support over the next decade and a half. More recently, the data show a "blip" in Liberal support that coincides with "Cleggmania" in 2010, Labour's surge in 2017, and the Conservative landslide in 2019.

Table 1 summarises our estimates. One fact is most apparent: that Labour and the Conservatives are almost perfectly matched across the time period as a whole. Each averaged support from around 40% of voters, this support varied by around 7%, and each took the lead around 50% of the time. The Liberals—Britain's third most popular party—were not so fortunate. Their support averaged around 14%, though this figure varied between a low of 3% and a high of 46%, with the party taking the lead only around 1% of the time.

We can also use our estimates to make specific claims about British politics. For example, we can assert with a reasonable degree of certainty that the Conservative Party received the largest degree of support of any party in Britain between 1955 and 2021 on 10 April 1968 when our estimates show that 57.5% (95% CI: 54.8% to 60.1%) of the electorate would have said that they would have voted for them at the next election. Similarly, we can also assert that Labour Party's

Table 1: Overall summary of daily voting intention estimates, 1955 to 2021

Party	Median	Mean	Error	Lowest	Highest	Leader
Conservative	39.5%	39.4%	6.8%	20.2%	57.5%	51.9% of days
Labour	39.5%	39.8%	6.7%	21.6%	57.3%	47.5% of days
Liberal	13.2%	14.0%	6.0%	2.8%	45.6%	0.6% of days

peak in the polls came under former Prime Minister Tony Blair on 7 June 1997 when 57.3% (95% CI: 55.3% to 59.3%) of the electorate intended to back them at the next election.

These data and their summaries raise an interesting question: if Labour and the Conservatives have tended to be so well-matched in the polls, why have the Conservatives done so much better at election time? Despite leading 47.5% of the time, Labour has gained the highest share of the vote in only 6 of the period's 18 general elections (33%). This phenomenon is not entirely unprecedented. Jackman (1994) shows that the Australian Labor Party has suffered at elections due to pervasive electoral bias. Though this might also be the case in Britain, our figures reflect *vote* shares, not *seat* shares. As such, electoral bias should be a concern only insofar as it affects the parties' overall popularity. Other factors likely explain Labour's poor performance: Britain's press leans right and tends to support the Conservatives; successive Conservative governments could have scheduled elections to maximise their chances of winning; partisan non-response could have led Conservative voters to drop out of polls in the period between elections; or, most simply, the Conservatives might be the better campaigners.
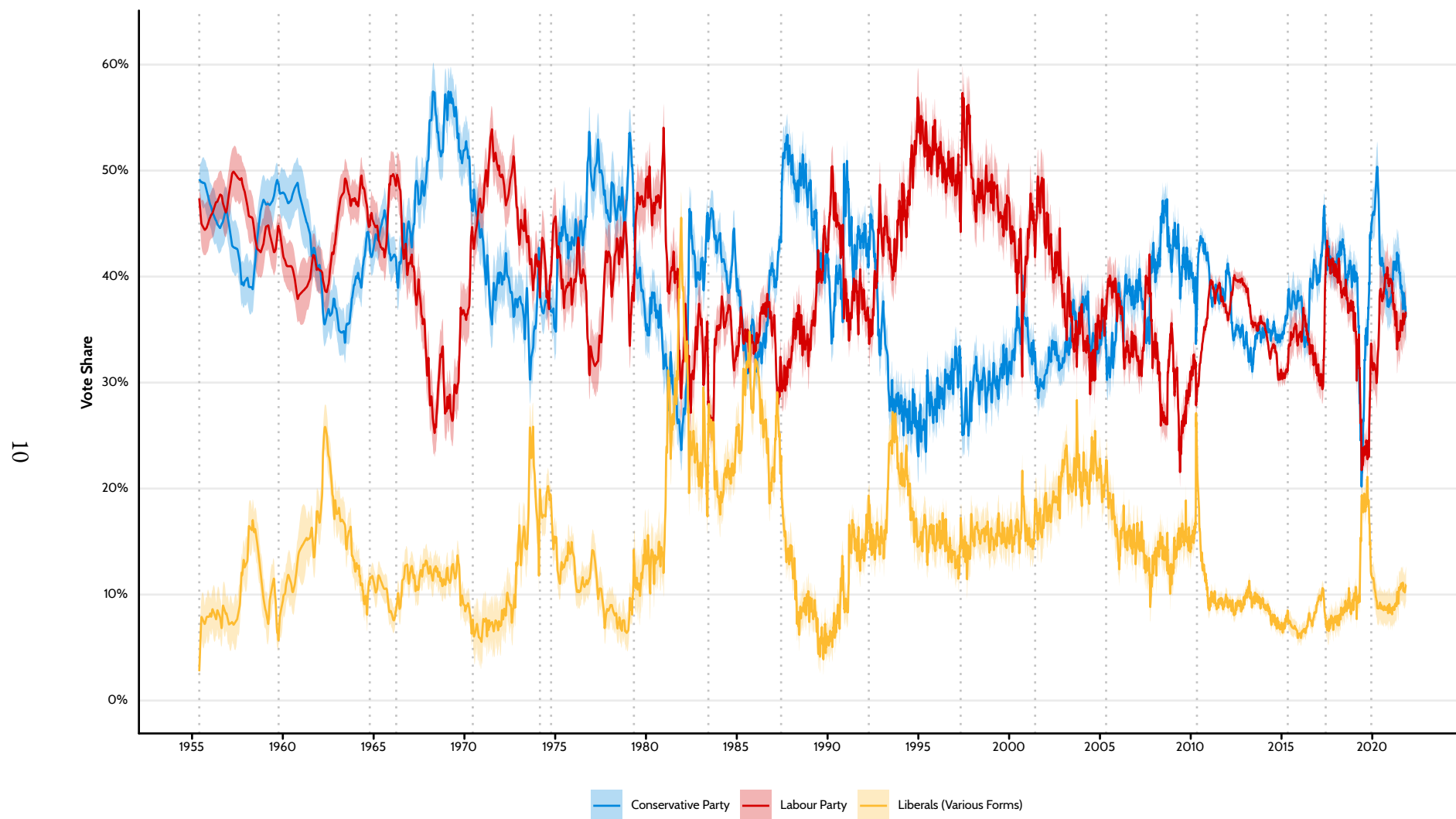
Figure 3: PollBasePro includes 24,277 daily estimates of aggregate voting intention for each of Britain's largest parties: the Conservatives, Labour, and the Liberals in their various forms. Dotted lines show election dates.

As well as estimating latent vote intention, our approach also estimates "house effects" over time. Figure 4 pools house effects across polling companies for each party over each parliament[5]. In general, polling companies have tended to underestimate support for the Conservatives and overestimate it for Labour. House effects for the liberals have, more-or-less, always been close to zero. In terms of polling misses, the three largest average underestimates were for the Conservatives in April 1992 (-3.91 percentage points), Labour in June 2017 (-3.66 percentage points), and the Conservatives in May 2015 (-3.66 percentage points). Likewise, the three largest average overestimates were for Labour in June 1970 (4.07 percentage points), Labour in June 2001 (3.92 percentage points), and Labour in April 1992 (3.52 percentage points).

Conclusions

We introduce PollBasePro: the most comprehensive time series of British voting intention data assembled to date. It contains 24,277 daily voting intention estimates for each of Britain's three largest parties, beginning at the 1955 General Election on 26 May 1955 and ending on 11 November 2021. What's more, these data are well validated and do a good job of tracking the ebb and flow of British political history.

PollBasePro is a living dataset and likely to change over time. Thus, we would stress that users should always endeavour to use the most recent version of the data. As is sensible, we intend to take some steps to ensure that our data remain available for the foreseeable future. First, we will host our code online for others to use, explore, and expand upon. In particular, we will do so at the following GitHub repository: [censored due to anonymity]. GitHub is well-equipped to handle software development and provides a suite of useful support features so that our users can flag any problems. Second, we will host a stable version of our data on the Harvard Dataverse at [censored due to anonymity]. Relying on the Dataverse over, say, temperamental institutional websites should ensure that our data remain available in the long-term.

Though we believe that our data can help to answer all manner of questions, PollBasePro still has room to grow. Three new features are most obvious. First, to provide voting intention

---

[5]For more information on the full distribution of pollster-by-pollster historical house effects, see the `house_effects` data set that accompanies our data.
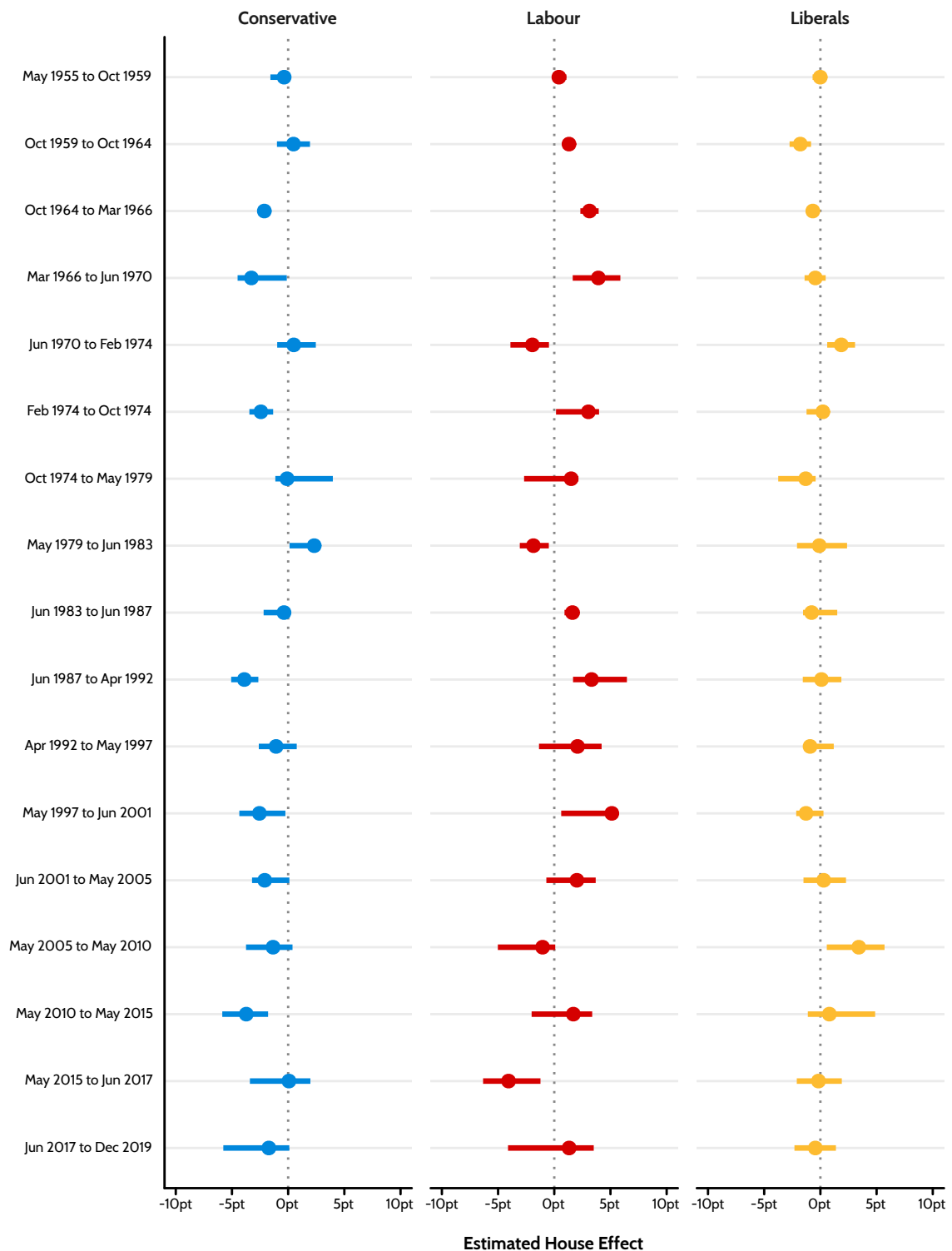
Figure 4: Dot plot distributions of our estimated house effects for each of the three main parties between 1955 and 2019. Here each dot represents one one-hundredth of the total distribution.

figures for the UK's referendum on leaving the European Union and Scotland's referendum on leaving the UK. Both time series would be long and benefit from known outcomes to which we could anchor our estimates. Second, to expand our data beyond the three main parties. This would be particularly useful for those who study, for example, the rise of UKIP. Third, to collect and incorporate any missing sample sizes. As we discuss above, we use the Timeline of Elections dataset (Jennings and Wlezien 2016) to impute sample sizes for all polls that occurred before the 2010 general election. This is a pragmatic and reasonable decision given present data limitations. Still, were we to have the necessary resources, we would like to collect the sample sizes associated with these polls to ensure that our estimates are as accurate as possible.

No matter how PollBasePro develops in the future, we are clear that it provides many opportunities in the present. One is to improve the quality of political journalism too. The stories that we use to make sense of the polls often come to shape our politics (Barnfield 2020). And many of these stories emerge from the simple rolling averages that journalists and commentators often use to make sense of the polls. As we mention above, our estimates account for shortcomings that these simple methods cannot. They might, thus, act as a kind of "sanity check" when it comes to reporting on the latest polling figures. Whether this comes to pass or not remains to be seen. Still, we remain optimistic about our data's potential. Indeed, we expect PollBasePro to become a valuable resource for students of British politics in the years to come.

References

Barnfield, Matthew. 2020. "Think Twice Before Jumping on the Bandwagon: Clarifying Concepts in Research on the Bandwagon Effect." *Political Studies Review* 18 (4): 553–74. https://doi.org/10.1177/1478929919870691.

Giles, Jim. 2005. "Internet Encyclopaedias Go Head to Head." *Nature* 438 (7070): 900–901. https://doi.org/10.1038/438900a.

Jackman, Simon. 1994. "Measuring Electoral Bias: Australia, 1949." *British Journal of Political Science* 24 (3): 319–57. https://doi.org/10.1017/S0007123400006888.

———. 2005. "Pooling the Polls over an Election Campaign." *Australian Journal of Political Science* 40 (4): 499–517. https://doi.org/10.1080/10361140500302472.

Jackman, Simon, and Luke Mansillo. 2018. "The Campaign That Wasn't: Tracking Public Opinion over the 44th Parliament and the 2016 Election Campaign." In *Double Disillusion: The 2016 Australian Federal Election*, edited by Anika Gauja, Peter Chen, Jennifer Curtin, and Juliet Pietsch, 720. Canbarra, ACT, Australia: ANU Press.

Jennings, Will, and Christopher Wlezien. 2016. "The Timeline of Elections: A Comparative Perspective." *American Journal of Political Science* 60 (1): 219–33. https://doi.org/10.1111/ajps.12189.

Louwerse, Tom. 2016. "Improving Opinion Poll Reporting: The Irish Polling Indicator." *Irish Political Studies* 31 (4): 541–66. https://doi.org/10.1080/07907184.2016.1213719.

Mansillo, Luke, and Simon Jackman. 2020. "National Polling and Other Disasters." In *Morrison's Miracle: The 2019 Australian Federal Election*, edited by Anika Gauja, Marian Sawer, and Marian Simms, 125–48. Canbarra, ACT, Australia: ANU Press.

Mellon, Jonathan, and Christopher Prosser. 2017. "Missing Nonvoters and Misweighted Samples." *Public Opinion Quarterly* 81 (3): 661–87. https://doi.org/10.1093/poq/nfx015.

Pack, Mark. 2021. "POLLBASE: OPINION POLLS DATABASE FROM 1943-TODAY." https://www.markpack.org.uk/opinion-polls/.

Prosser, Christopher, and Jonathan Mellon. 2018. "The Twilight of the Polls? A Review of

Trends in Polling Accuracy and the Causes of Polling Misses." *Government and Opposition* 53 (4): 757–90. https://doi.org/10.1017/gov.2018.7.

Sturgis, Patrick, Jouni Kuha, Nick Baker, Mario Callegaro, Stephen Fisher, Jane Green, Will Jennings, Benjamin E. Lauderdale, and Patten Smith. 2018. "An Assessment of the Causes of the Errors in the 2015 UK General Election Opinion Polls." *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 181 (3): 757–81. https://doi.org/10.1111/rssa.12329.

Wikipedia. 2021. "Opinion Polling for the 2019 United Kingdom General Election." *Wikipedia*. https://w.wiki/35RK.

Table A1: List of polling companies in the PollBase data that we use to produce our daily estimates.

Pollster	Count	From	To
Yougov	1816	11 January 2002	4 November 2021
Gallup	664	2 July 1955	6 June 2001
Mori	582	18 March 1976	5 May 2010
ICM	569	13 January 1989	9 December 2019
NOP	511	1 September 1959	3 May 2005
Populus	318	7 February 2003	11 December 2018
Opinium	305	4 September 2009	6 November 2021
Comres	256	20 October 2006	10 November 2019
Harris	169	1 June 1970	9 June 2010
Survation	139	5 January 2013	7 October 2021
Ipsos Mori	121	18 June 2010	4 November 2021
Marplan	96	11 June 1970	6 December 1988
Redfield and Wilton Strategies	74	12 February 2020	8 November 2021
Savanta Comres	71	11 November 2019	7 November 2021
TNS	61	18 February 2010	8 August 2016
ASL	56	16 May 1983	27 January 1994
BMG	55	25 April 2015	26 April 2021
Lord Ashcroft	53	17 May 2013	22 August 2016
Kantar	46	20 April 2017	18 October 2021
Angus Reid Public Opinion	45	15 October 2009	13 April 2013
Deltapoll	38	30 May 2018	5 November 2021
Bpix	33	1 April 2005	1 May 2010
Panelbase	32	24 March 2015	16 September 2021
G9000	17	3 May 1990	28 January 1992
Number Cruncher Politics	15	27 March 2018	18 October 2021
ORC	12	13 June 1970	3 May 1979
Commr	11	23 September 2004	28 April 2005
NMR	10	6 April 1990	24 March 1992
ORB	9	19 April 2017	31 October 2019
Business Decisions	5	7 February 1974	2 October 1974
Qriously	4	28 April 2017	8 December 2019
Rasmussen	4	12 May 2001	3 June 2001
Research Service Ltd	4	3 April 1979	25 April 1979
Find Out Now	3	2 February 2021	8 September 2021
Focaldata	3	27 November 2019	30 April 2021
Hanbury Strategy	3	5 April 2019	4 September 2019
Onepoll	3	31 March 2010	24 April 2010
GFK	2	1 March 2017	14 May 2017
MFS	2	6 September 2000	19 September 2000
Norstat	2	23 April 2017	1 June 2017
RNB	2	4 March 2010	2 May 2010
Forecasting Statistics	1	1 October 1959	4 October 1959
Live St	1	3 January 2003	4 January 2003
Marketing Sciences	1	15 April 2009	16 April 2009
Sky Data	1	19 February 2019	19 February 2019
Surveymonkey	1	30 April 2015	6 May 2015