PollBasePro: Daily Estimates of Aggregate Voting Intention in Britain from 1955 to the Present Day[*]

Jack Bailey[†] Mark Pack[‡] & Luke Mansillo[§]

This version compiled 21 June 2021 using PollBasePro v0.1.0

Abstract

Political scientists often use public opinion polls to test their theories. Yet these data present some difficulties. First, they are noisy. Second, they occur at irregular intervals. Third, they measure both public preferences and pollsters' survey design choices. We introduce a new dataset, PollBasePro, that accounts for these issues. It includes voting intention estimates for each of Britain's three largest parties on each day between the 1955 general election and the present. As the dataset covers 24,133 days in total, it is perhaps more comprehensive than any other existing time series of British political opinion. We then use our estimates to test a question of pressing importance: how daily deaths attributable to COVID-19 have influenced support for the governing Conservative Party.

Keywords: Opinion polls; public opinion; British politics; Coronavirus; COVID-19; Bayesian methods.

Introduction

Beleaguered politicians often remark that the "only poll that matters is the election itself." Like much that politicians say, this is not true. Though elections decide who governs, public opinion polls are important too: they tell us how popular parties are in the time between elections. Polling is especially important in Britain. Unlike in some countries, British governments can control the election schedule. As a result, the governing party's standing is often one of the only factors that decides when an election will occur (Smith 2003).

Since the advent of universal suffrage in 1928, Britain has held only 24 general elections. This is not a large sample. Political scientists have, thus, turned to public opinion polls to test their theories. Still, these data present some difficulties. Three issues are most serious. First, that polling data are noisy and may be prone to bias. Second, that they occur at irregular time intervals. And, third, that they measure both real changes in the electorate's preferences *and* the design choices of the companies that run them.

We introduce a new dataset—*PollBasePro*—that overcomes these issues. It includes daily estimates of aggregate voting intention for each of Britain's three largest parties on each day between 26 May 1955 and 20 June 2021. Covering 24,133 days in total, PollBasePro permits a degree of specificity and flexibility beyond that of any existing time series. In the sections that follow, we elaborate on the dataset. First, we describe the methods and the data that we use to derive our estimates. Next, we consider what PollBasePro reveals about British politics since 1955 and show how to use it to test answer a question of pressing concern: how do deaths related to COVID-19 affect public support for the governing Conservative Party? We show that COVID-related deaths have led to a "rally 'round the flag" effect that has boosted Conservative support, though that this relationship has weakened, then reversed, as the pandemic has progressed. Finally, we provide some initial conclusions and then remark on how we might develop PollBasePro over the coming years.

Data, Estimation, and Validation

Polls have long informed political science. Goodhart and Bhansali (1970), for instance, used Gallup and NOP polls to show that economic change affects incumbent party support, launching the economic vote. Yet though plentiful, these data have three important problems.

First, polling data are noisy. This is an unavoidable consequence of sampling methodologies. As we cannot poll every person in a country, polls cannot provide exact estimates of public opinion. Rather, they represent draws from a distribution of *possible* estimates of public opinion. Once introduced into formal statistical models, this sampling error becomes a form of measurement error. And failing to account for it can have serious consequences. Where we use polls as an outcome, it reduces our statistical power. Where we use them as a predictor, it pulls real and existing effects towards zero.

Second, polling data occur at irregular intervals, covering any period of time from a single day to several weeks. Political scientists often assume that these figures measure public opinion on the poll's last day in the field. Of course, that is almost never true. It can also be a problem when major events occur partway through the data collection process. This is a major practical problem: polls can be difficult to match to other time series.

Third, polls do not only measure changes in aggregate political preferences. They also measure systematic biases that arise due to the design choices of the companies that run them. In the past, these biases have been so large that they have cast doubt on the efficacy of the 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 (Sturgis et al. 2018; Prosser and Mellon 2018; Mellon and Prosser 2017). These biases were even more serious at the 2019 Australian Federal Election. As in the British case, Labor had a healthy lead in the polls. Yet, again, the conservative Coalition beat them by around 8 percentage points (Mansillo and Jackman 2020).

Our intention is simple: to produce voting intention estimates that account for these issues. To do so, we adapt the method 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 as a random walk between these two points. On any given day, the party's support depends on its support the day before, pollster-specific "house" effects, and random shocks. Jackman and Mansillo (2020; 2016) have used the model to track voting intention in Australia and Louwerse (2016) have used it to estimate aggregate voting intention in Ireland. The full model specification is as follows, though note that we also provide a more complete explanation of how the model works and the steps that we have taken to produce our estimates in the accompanying appendix:

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

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

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

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

$$\delta_j \sim \text{Normal}(0, 0.05) \text{ for } j \text{ in } 1, ..., J \qquad \text{Prior on house effects, } \delta$$

$$\omega_t \sim \text{Normal}(0, 0.025) \text{ for } t \text{ in } 1, ..., T-1 \qquad \text{Prior on random shocks, } \omega$$

$$\tau \sim \text{Normal}(0, 0.05)^+ \qquad \text{Positive prior on scale of innovations, } \tau$$

$$\sigma \sim \text{Exponential}(20) \qquad \text{Prior on residual error, } \sigma$$

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

The PollBase data come from a wide range of sources. This includes books published after each general election, polling almanacs, data shared amongst academics and pollsters, contemporary media reports, and figures from polling company websites. The data since the 1983 general election are complete, baring possible but rare individual errors. Prior to 1983,
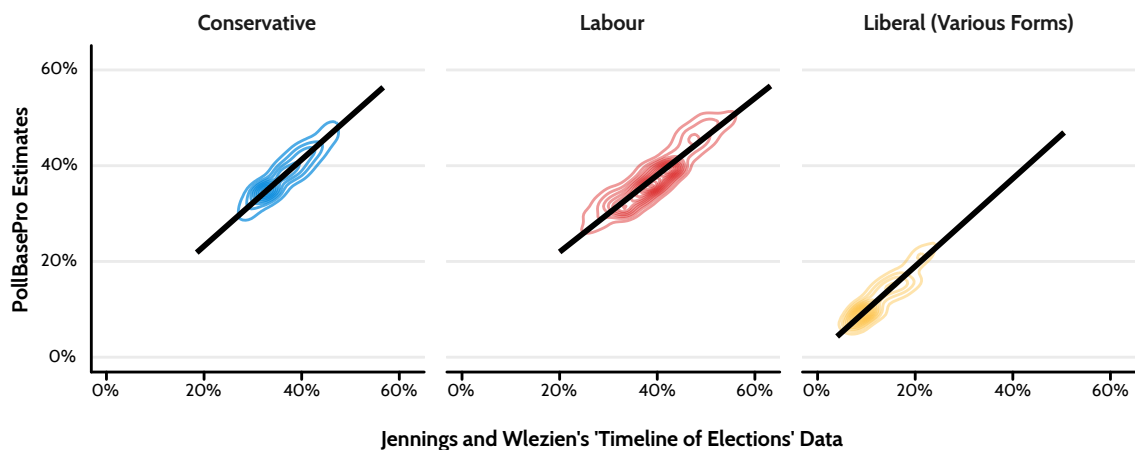
Figure 1: In all cases, estimates from PollBasePro appear well-validated when compared to raw polling data from Jennings and Wlezien's 'Timeline of Elections' dataset (2016).

the data is likewise complete for each general election campaign period. For periods between general elections prior to 1983, coverage is more comprehensive for some pollsters than others[1]. Gallup and NOP data, in particular, are well-covered as the two companies have collated and published their results. The PollBase data start in June 1945. However, we take as our starting point the 1955 Parliament given the gaps in the earlier data.

We use the Wikipedia data to cover the period from 2010 onwards. This is a pragmatic choice. Volunteers update the 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. For the PollBase data, we impute likely sample sizes to allow us to estimate these errors[2]. But as the Wikipedia data include sample sizes, imputation is not necessary. 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. Wikipedia's coverage of public opinion polling since 2010 is also of a very high quality, and almost all figures on the website including links to external data sources that corroborate them.

We validate them against Jennings and Wlezien's "Timeline of Elections" dataset (2016). These data contain 4,302 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. Even so, the

---

[1] We encourage any readers who might know of polls missing from the PollBase dataset to contact its author via his website (https://www.markpack.org.uk)

[2] Again, see the appendix for more information on this process.

Timeline data provide a good test as Jennings and Wlezien compiled them independently. As figure 1 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 94.0% (95% CI: 93.6% to 94.3%), an MAE of 2.16 percentage points, and an RMSE of 0.03; Labour, a correlation of 92.9% (95% CI: 92.5% to 93.3%), an MAE of 2.75 points, and an RMSE of 0.03; and the Liberals, a correlation of 95.0% (95% CI: 94.6% to 95.3%), an MAE of 1.41 points, and an RMSE of 0.02.

PollBasePro and British Politics Since 1955

Figure 2 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 jumps out: that Labour and the Conservatives were almost perfectly matched. Each averaged support from around 40% of voters, this support varied by around 7 percentage points, and each took the lead around 50% of the time. The Liberals—Britain's third most popular party for much of the period—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.

As well as summarising facts about the parties, our estimates also allow us to make general claims about the relationships that exist *between* the parties. As figure 3 shows, these are all negative. Consider the effect of rising Liberal support on the Conservatives and Labour. The correlation between the Liberals and the Conservatives is strong at -45.0% (95% CI: -46.0% to -43.9%) and between the Liberals and Labour -21.0% (95% CI: -22.2% to -19.8%). Contrary to popular arguments that the Liberals serve to split the centre-left vote (e.g. Jenkins 2019), our figures suggest that rising Liberal support has tended to hurt the Conservatives more than

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

Party	Median	Mean	Error	Lowest	Highest	Leader
Conservative	39.6%	39.4%	6.8%	20.0%	57.4%	50.9% of days
Labour	39.6%	39.8%	6.7%	21.6%	57.4%	48.5% of days
Liberal	13.3%	14.0%	6.0%	2.8%	45.5%	0.6% of days

Labour. The correlation between Labour and the Conservatives is also sizeable, at -37.7% (95% CI: -38.9% to -36.6%). Given that they are Britain's two main parties, this is perhaps unsurprising: Labour and the Conservatives are the only parties likely to form a government after any election and punishing one often requires voting for the other.

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 8 April 1968 when our estimates show that 57.4% (95% CI: 54.8% to 60.0%) of the electorate would vote for them at the next election. Similarly, we can assert that the Conservative's peak in the polls came on 8 April 1968 when 57.4% (95% CI: 54.8% to 60.0%) of the electorate intended to back them.

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 in the polls 48.5% 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; partisan non-response could have led Conservative voters to drop out of polls in the period between elections; or, most simply, the Conservatives might just be better at campaigning.
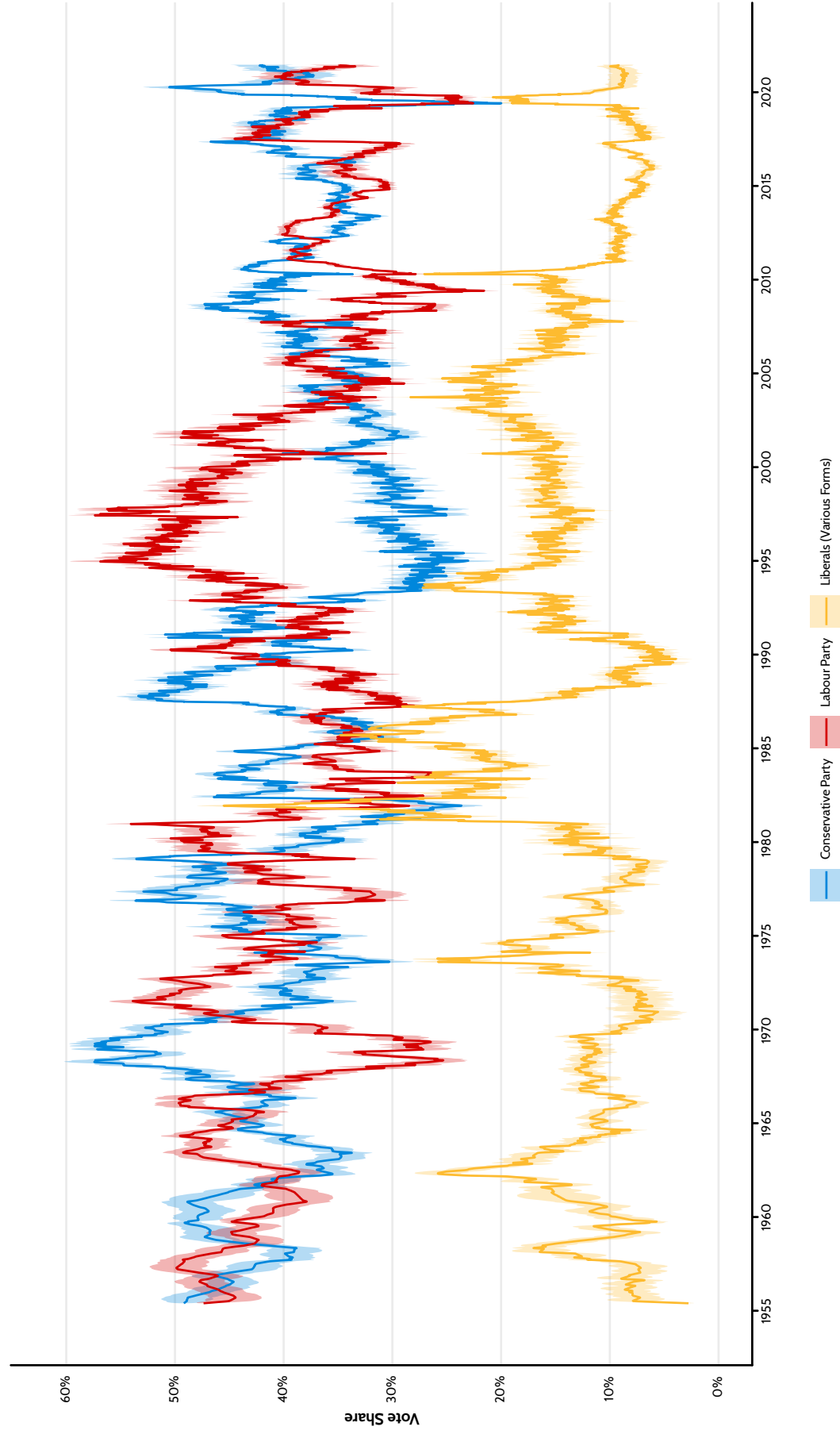
Figure 2: PollBasePro includes 24,133 daily estimates of aggregate voting intention for each of Britain's largest parties: the Conservatives, Labour, and the Liberals in their various forms.
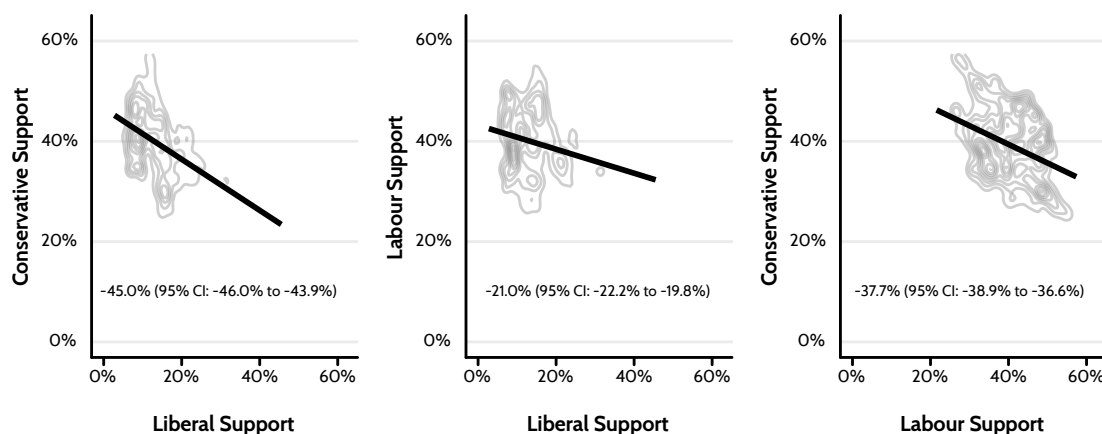
Figure 3: Historic correlations between each of Britain's three largest parties, 1955 to 2021

Using PollBasePro to Answer Political Questions

It is one thing to show that our estimates are well-validated; it is quite another to show that they can advance our understanding of politics. As we mention above, raw polling data cover irregular time intervals and can be difficult to match to other time series. Our data face no such problem. Instead, as PollBasePro includes daily voting intention estimates, it is simple to merge the data into any other time series, whether they be daily, weekly, monthly, quarterly, or yearly.

To demonstrate their usefulness, we use our data to answer a question of timely importance: how are deaths attributable to COVID-19 related to public support for the governing Conservative Party? To answer this question, we merge our data into the UK Government's tracker of daily COVID-related deaths (2021a). These data run from 2 March 2020 to 18 June 2021, though we begin our analysis at the 2019 General Election and mark any deaths before the start of the pandemic as zero. We then compute daily changes in Conservative support and daily changes in COVID-related deaths across a four week window, before modelling the former as a function of the latter using the following error-in-variables model:

$$\Delta C_t \sim \text{Normal}(\mu_t, \sqrt{\sigma^2 + S_t^2}) \qquad \text{Likelihood function}$$

$$\mu_t = \alpha + \beta_1 T_t + \beta_2 \Delta D_t + \beta_3 T_t \times \Delta D_t \qquad \text{Linear model on } \mu$$

$$\alpha \sim \text{Normal}(0, 0.1) \qquad \text{Prior on } \alpha$$

$$\beta_j \sim \text{Normal}(0, 0.1) \text{ for } j \text{ in } 1, ..., J \qquad \text{Prior on } \beta$$

$$\sigma \sim \text{Exponential}(10) \qquad \text{Prior on } \sigma$$

This is a simple linear model, where change in Conservative support at time t, ΔC_t, is a function of the passage of time, T_t, and change in the number of COVID-related deaths, ΔD_t. We also interact the two variables so that the effect of these deaths can change over time. As the intervals in figure 2 make clear, our data are *probabilistic*. Thus, it is important that we propagate this uncertainty forward so as to maximise our statistical power. This is what differentiates our error-in-variables model from a standard linear model: rather than include only the residual error, σ, we also include any known uncertainty in our estimates at time t, S_t.

It is possible to propagate this error forward because PollBasePro includes *two* estimates for each party for each day that the data covers: estimates of public support for each party *and* an estimate of the standard error of this estimate. Table 2 shows the first 5 rows of the PollBasePro dataset. Variables ending "_est" show estimated voting intention and those ending "_err" their associated error. Notice that the error on the first date is zero. This is because this is the date of the 1955 general election, so we have absolute certainty what the result really was. As the days pass, this error gradually increases, reflecting our own epistemic uncertainty. These errors provide useful information in their own right, so it is prudent to also include them in our analysis.

According to the literature, there are good reasons why the relationship between the two variables might be positive and good reasons why it might be negative. The literature on "rally 'round the flag" effects would favour a *positive* relationship. Here, voters rush to support the governing party in times of crisis, such as wars (Kuijpers 2019; Mueller 1970). The literature on

Table 2: The first 5 rows of the PollBasePro dataset

date	con_est	con_err	lab_est	lab_err	lib_est	lib_err
1955-05-26	0.492	0.000	0.474	0.000	0.028	0.000
1955-05-27	0.492	0.002	0.473	0.002	0.029	0.003
1955-05-28	0.492	0.003	0.473	0.003	0.030	0.004
1955-05-29	0.492	0.004	0.472	0.003	0.031	0.005
1955-05-30	0.492	0.004	0.471	0.004	0.033	0.005

retrospective voting (Healy and Malhotra 2013; Fiorina 1981), instead, would favour a *negative* relationship. It holds that voters reward governments for positive policy outcomes and punish them for negative ones (i.e. for avoidable deaths).

Figure 4 shows that the effect is more "rally 'round the flag" than "reward and punishment." That said, our model suggests at least some preliminary evidence of a transition from one mechanism to the other. In March 2020, when the pandemic started to gain pace in Britain, there was a strong positive relationship between change in the number of deaths attributable to COVID-19 and change in support for the governing Conservative Party. For every 100 extra COVID-related deaths over the past 4 weeks, Conservative support over the same period increased on average by 0.6 points (95% CI: 0.5 to 0.7). Six months later, in September 2020, this relationship had weakened markedly. Though still positive, it had fallen to only 0.2 points (95% CI: 0.2 to 0.3). A year after the pandemic began, the relationship had reversed. In March 2021, every 100 additional COVID-related deaths over the past 4 weeks led Conservative support to fall on average by -0.1 points (95% CI: -0.2 to -0.1).

It is clear then that PollBasePro has allowed us to reveal an important—if surprising—answer to the question that we posed above: despite Britain having one of the world's highest levels of COVID-related deaths per capita (Ritchie et al. 2021), these untimely deaths appear to have *benefited* the fortunes of the governing Conservative Party. Yet this relationship has changed over time and, a year into the pandemic, has reversed direction. We might, thus, expect any further deaths attributable to COVID-19 to erode the Conservative's ample base of support. Yet given the pace of Britain's vaccination programme (UK Government 2021b), it is possible that the Conservatives have reaped the electoral rewards of the crisis without risk of punishment.
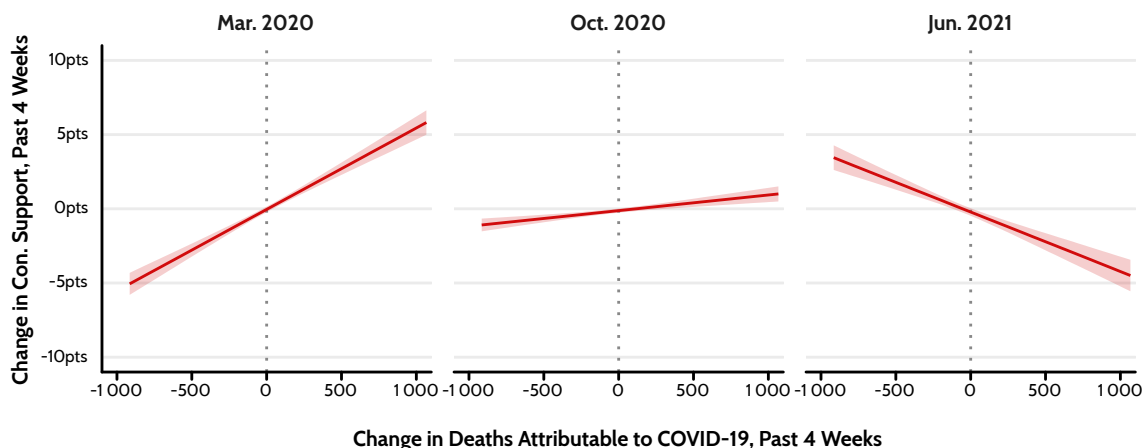
Figure 4: At the start of the COVID-19 pandemic, more deaths were related to greater support for the incumbent Conservative Party. As time has passed, this relationship has reversed.

Conclusions

In this paper, we introduce PollBasePro: the most comprehensive time series of British voting intention data assembled to date. The dataset contains 24,133 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 20 June 2021. Furthermore, our data are well validated and follow the course of British political history.

It is important to stress that PollBasePro is a *living dataset*. As a result, we expect to incorporate new estimates into the data as time goes by. Users should, therefore, endeavour to use the most recent data in any of their analyses. It is also important to stress that—like any project of this nature—our data pipeline might contain minor errors or inefficiencies. To guard against this, we have hosted all of our materials online for others to inspect. If our users find any errors in our code or wish to make recommendations for future updates, we invite them to raise an issue on the project's GitHub repository or to contact the authors directly.

Though we believe that our data can help to answer all manner of questions, PollBasePro still has room to grow. Three new features would be most useful. First, to estimate aggregate support for the UK leaving the European Union and for Scotland leaving the UK. Both time series would be long-ranging and benefit from having known outcomes to which we could anchor our estimates. Second, to expand the data beyond the three main parties. This would

11

be particularly useful for those who study, for example, the rise of the SNP in Scotland or the influence of UKIP. Third, to collect and incorporate known sample sizes into all of our polling data. At the moment, 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 but reasonable decision given that these data do not include sample sizes. Still, were we to have the necessary resources, it would be good to collect the sample sizes associated with these polls to ensure that our estimates are as precise as possible.

No matter how PollBasePro develops in the future, we are clear that it provides many opportunities in the present. We have discussed many of the academic opportunities that the data provide above. But our estimates might stand to improve the quality of polling reporting too. The stories that we tell ourselves to explain the polls often come to shape our politics (Barnfield 2020). And many of these stories emerge from the simple averages that many commentators use to make sense of the polls. As we mention above, our estimates account for shortcomings that these simple rolling averages do not. PollBasePro might, therefore, stand to benefit not only academic researchers, but also the quality of political debate in Britain too. As such, 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.

Fiorina, Morris P. 1981. *Retrospective Voting in American National Elections*. New Haven, CT: Yale University Press.

Goodhart, C. A. E., and R. J. Bhansali. 1970. "Political Economy." *Political Studies* 18 (1): 43–106. https://doi.org/10.1111/j.1467-9248.1970.tb00659.x.

Healy, Andrew, and Neil Malhotra. 2013. "Retrospective Voting Reconsidered." *Annual Review of Political Science* 16 (1): 285–306. https://doi.org/10.1146/annurev-polisci-032211-212920.

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. 2016. "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. Acton, AU: ANU Press.

Jenkins, Simon. 2019. "The Lib Dems Helped the Tories to Victory Again. Now They Should Disband." *The Guardian*. https://www.theguardian.com/commentisfree/2019/dec/16/lib-dems-tories-split-vote-labour.

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.

Kuijpers, Dieuwertje. 2019. "Rally Around All the Flags: The Effect of Military Casualties on Incumbent Popularity in Ten Countries 1990." *Foreign Policy Analysis* 15 (3): 392–412.

https://doi.org/10.1093/fpa/orz014.

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. Canberra, AU: 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.

Mueller, John E. 1970. "Presidential Popularity from Truman to Johnson." *The American Political Science Review* 64 (1): 18–34. https://doi.org/10.2307/1955610.

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.

Ritchie, Hannah, Esteban Ortiz-Ospina, Diana Beltekian, Edouard Mathieu, Joe Hasell, Bobbie Macdonald, Charlie Giattino, and Max Roser. 2021. "Coronavirus (COVID-19) Deaths." *Our World in Data*. https://ourworldindata.org/covid-deaths.

Smith, Alastair. 2003. "Election Timing in Majoritarian Parliaments." *British Journal of Political Science* 33 (3): 397–418. https://doi.org/10.1017/S0007123403000188.

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.

UK Government. 2021a. "Deaths Within 28 Days of Positive Test by Date of Death." *GOV.UK*. https://coronavirus.data.gov.uk/details/deaths.

———. 2021b. "People Who Have Received Vaccinations, by Report Date (Daily)." *GOV.UK*. https://coronavirus.data.gov.uk/details/vaccinations.

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

Appendix: Estimating Daily Voting Intention

We adapt Jackman's (2005) method to derive our daily estimates. Still, there are issues specific to our case that we must first overcome. We elaborate on our choices below.

Imputating Missing Sample Sizes

Our data do not include sample sizes before the 2010 general election. This is a problem, as our model requires that we know this information. 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 information on sample sizes. What's more, they also include data from countries other than Britain. This lets us pool all available information to improve our estimates.

Sample sizes are count data. As such, we use the following multilevel Poisson regression model to impute likely sample sizes for all of our pre-2010 polling data:

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

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

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

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

$$\alpha \sim \text{Normal}(7, 0.5) \qquad \qquad \text{Prior on average intercept, } \alpha$$

$$\beta \sim \text{Normal}(0, 0.1) \qquad \qquad \text{Prior on average slope, } \beta$$

$$\sigma_\alpha \sim \text{Exponential}(10) \qquad \qquad \text{Prior on uncertainty in the intercepts, } \sigma_\alpha$$

$$\sigma_\beta \sim \text{Exponential}(10) \qquad \qquad \text{Prior on uncertainty in the slopes, } \sigma_\beta$$

$$\mathbf{R} \sim \text{LKJ}(2) \qquad \qquad \text{Prior on correlation matrix, } \mathbf{R}$$
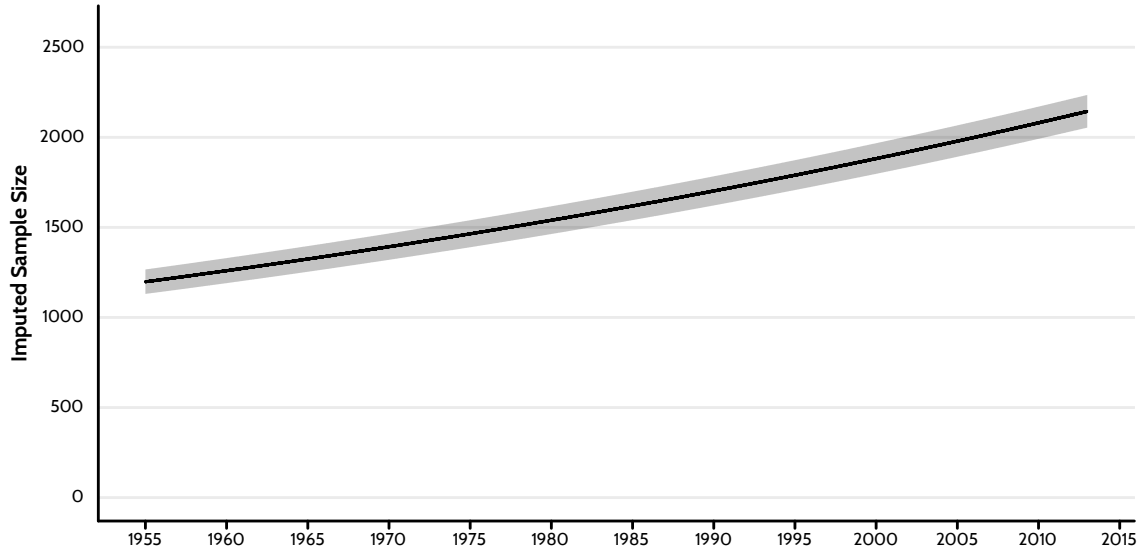
Figure 5: Imputed sample sizes in Britain between 1955 and 2013, estimated using sample size data from Jennings and Wlezien's "Timeline of Elections" dataset (2016).

We assume that the sample size associated with poll i in the Timeline data, n_i, is distributed as Poisson according to some rate parameter, λ_i. We then model the logarithm of this parameter using a simple linear function that includes an intercept, $\alpha_{Country}$, and a slope on the effect of time, $\beta_{Country}$, 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. In effect, this allows the parameters to be correlated and, thus, to share information.

While our data concern Britain alone, we use all available data in the Timeline dataset. This is for good reason. The dataset does not contain reliable sample size values for British polls conducted before the early 1960s. But it does contain reliable 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 British polls and those from other countries.

Figure 5 shows the model's best estimate of the likely sample size of the average British voting intention poll between 1955 and 2013. These imputed values seem sensible and conform to our expectation that sample sizes should increase over time. 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.

We use the model to produce a time series of estimates sample sizes between 1955 and 2013. This includes all dates for which we intend to produce a voting intention estimate. Where our 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 we mention above and in the accompanying paper, we adapt the model in Jackman (2005) to compute our daily British voting intention estimates. The model is complex and has many moving parts, so we will build it up step by step.

We assume that each poll in our underlying data, $Poll_i$, is generated from some normal distribution. This distribution has two parameters. The first is some mean, μ_i. The second is some error that leads the estimates to be higher or lower than the expected value, μ_i. In many models with a normal likelihood function, this error parameter would measure only random residual error and be represented by the Greek letter σ. But, in our case, we have additional information that we can use. We know that each poll is a proportion and represents a draw from some random distribution. Thus, we can use the equation for the standard error of a proportion to calculate the uncertainty in each estimate, where $S_i = \sqrt{\frac{Poll_i(1-Poll_i)}{\nu_i}}$. Note that ν_i is the sample size of $Poll_i$, n_i, divided by the number of days the poll spent in the field, k_i. In effect, this implies that we assume an equal number of people were polled on each day that the model was in the field. We can then include both in our model to account for any known error, S_i, and any random residual error, σ. So far, our model is as follows:

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

The next step is to fit a model to μ_i. This will be a measurement model, as it will allow us to produce an estimate of the electorate's *latent* voting intention on each day. We assume that

18

μ_i is a linear function of two variables: $\alpha_{Day[i]}$, the electorate's latent voting intention for $Poll_i$ on the day that it was fielded, and $\delta_{Pollster[i]}$, the persistent "house effects" that arise due to the methodological and design choices that inform how the company that ran the poll collected its data. If we update our model specification to include these assumptions, we get the following:

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

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

At present, all values of α_{Day} are independent. This is a problem. First, we want estimates closer together to be more similar. Second, some days have no polling data to inform them. To address this problem, we constrain α_1 to be equal to the vote share that a given party received at a given election. We also constrain α_T to be equal to the vote share that the same party received at the following election. Next, we fit a dynamic model to α_t for all days in our time series except for the first and last. This acts as a sort of "chain" that links together all values of α. Because these values are now linked, they can then share information amongst themselves. This means that when the value of one estimate changes during the model estimation process, so too do the values of 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, the scale of innovations parameter. Updating the model again, we get:

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

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

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

As we rely on Bayesian methods, our final step in building the model is to provide a set of conservative and weakly-informative prior distributions[3].

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

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

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

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

$$\delta_j \sim \text{Normal}(0, 0.05) \text{ for } j \text{ in } 1, ..., J \qquad \text{Prior on house effects, } \delta$$

$$\omega_t \sim \text{Normal}(0, 0.025) \text{ for } t \text{ in } 1, ..., T-1 \qquad \text{Prior on random shocks, } \omega$$

$$\tau \sim \text{Normal}(0, 0.05)^+ \qquad \text{Positive prior on scale of innovations, } \tau$$

$$\sigma \sim \text{Exponential}(20) \qquad \text{Prior on residual error, } \sigma$$

After specifying our model, we loop over our data and fit the model to each election pair for each party. As we know with almost absolute certainty what the electorate's voting intention was on election days, we use this to our advantage and constrain α_1 to equal a given party's vote share at the first election and α_T to equal the same party's vote share at the following election. 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.

---

[3]In this case our priors serve only to regularise our parameter estimates. For example, our prior on the pollster house effects allows for a large—indeed, improbable—10 point house effect. They are unlikely, therefore, to constrain our estimates in such a way as to affect our inferences.