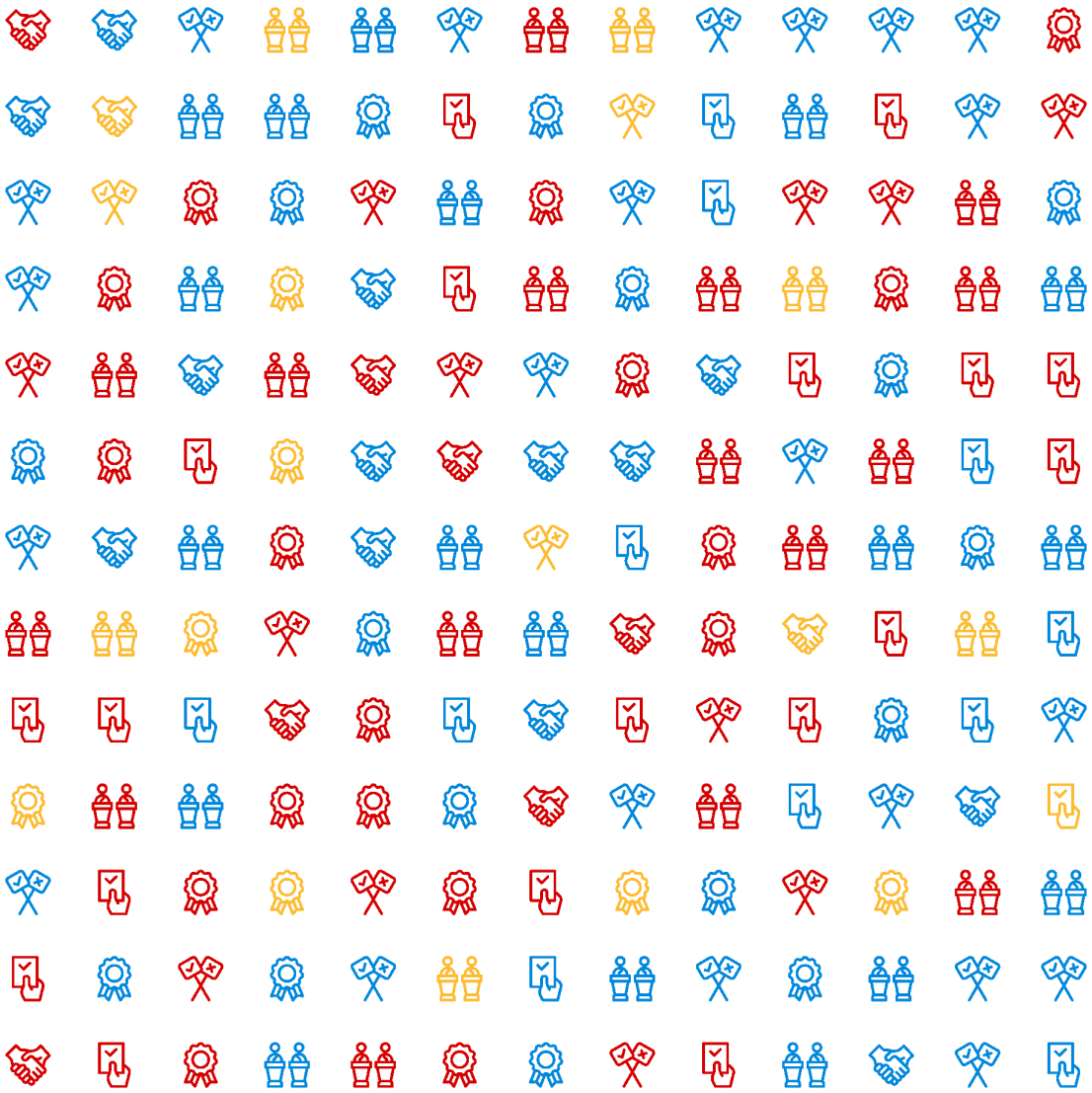
PollBasePro v0.1.0 (Beta)

User Guide and Codebook

This version: 12 March 2021



Contents

About PollBasePro 2

Citing PollBasePro 2

Using PollBasePro 3

Technical Details: Estimating Daily Voting Intention 5

Imputating Missing Sample Sizes 5

Estimating Daily Voting Intention Figures 7

Validating Our Estimates 9

Open-Source Data Pipeline 10

Variable List: PollBase (Historical British Election Polls from 1945 to 2021) 11

Variable List: PollBasePro (Daily British Voting Intention from 1955 to 2021) 12

Change Log 13

References 14

About PollBasePro

PollBasePro is a data package for the `R` programming language. It includes two datasets: `pollbase` and `pollbasepro`. The first provides a long-format and ready-to-analyse version of Mark Pack's (2021) dataset of historic British public opinion polls combined with more recent polling data from Wikipedia. The second provides 24,033 daily estimates of voting intention figures for each of Britain's three largest parties between 26 May 1955 and 12 March 2021.

Both `pollbase` and `pollbasepro` are *living datasets*. Thus, users should endeavour to use only the most recent version of the data when conducting their analyses. This will become more important in the future, as we intend to add new polls and estimates over time. Likewise, we cannot rule out that minor mistakes might have crept into our data processing pipeline. To guard against this, we have made all of our materials available for others to inspect. If you find an error in the code or wish to make a recommendation for a future update, we invite you to raise an issue on the PollBasePro GitHub repository.

Citing PollBasePro

If you use PollBasePro, you should also cite it. This is good practice and also allows the PollBasePro team to monitor how the data have been used. The project comprises three items: the data, this user guide, and a companion paper. The citations for each item are as follows:

- **Data:** Bailey, J., M. Pack, and L. Mansillo (2021) PollBasePro: Daily Estimates of Aggregate Voting Intention in Great Britain from 1955 to 2021 v.0.1.0 [computer file], March 2021. Retrieved from doi.
- **Documentation:** Bailey, J., M. Pack, and L. Mansillo (2021) PollBasePro v0.1.0: User Guide and Codebook. Retrieved from doi.
- **Paper:** Bailey, J., M. Pack, and L. Mansillo (2021) PollBasePro: Daily Estimates of Aggregate Voting Intention in Great Britain from 1955 to 2021. Retrieved from doi.

Using PollBasePro

Getting started with PollBasePro in R is simple. It requires only three short steps: first, to install the package; second, to load the package; and, third, to load the data. Once you have taken these three steps, you can then begin analysing the data. Note that PollBasePro is not yet available on CRAN, the service that hosts most R packages. As such, it is not yet possible to install PollBasePro with R's standard `install.packages()` function. Instead, we must install it directly from its GitHub repository. Thankfully this is straightforward and requires only that you run the following code in your R console:

```
# 1. Install the PollBasePro package from GitHub
devtools::install_github("jackobailey/PollBasePro")

# 2. Load the PollBasePro package in R
library(PollBasePro)

# 3. Load the pollbase and pollbasepro datasets
data("pollbase")
data("pollbasepro")
```

Though PollBasePro is first and foremost an R data package, we have sought to accommodate those who use Stata and SPSS too. To make this as easy as possible, we have made both the `pollbase` and `pollbasepro` datasets available as .dta and .sav files. These include all necessary value and variable labels and should work seamlessly with both software packages. You can download the latest version of the data by clicking the links below:

- `pollbase_0.1.0.dta`
- `pollbasepro_0.1.0.dta`
- `pollbase_0.1.0.sav`
- `pollbasepro_0.1.0.sav`

Figure 1: The posterior uncertainty in the estimates that we include in the pollbasepro dataset are correlated with time. This is because polls have become more frequent and have included larger sample sizes as time has passed. Thus, our estimates also become more precise.

Note that the estimates that we include in the `pollbasepro` dataset are *probabilistic*. As such, we include in the dataset both an estimate of the posterior mean of aggregate voting intention for each party on each day *and* the posterior uncertainty in these estimates. As figure 1 makes clear, our uncertainty estimates are not random. Instead, they are correlated with time. This occurs because polls have become more numerous and have tended to include larger sample sizes as time has passed. Thus, our estimates for more recent years are more certain than our estimates for years long far into the past.

We advise all those who use the `pollbasepro` dataset to include this uncertainty in their analyses wherever possible. This is important both because propagating our uncertainty forward is good practice and because such uncertainty serves both to reduce statistical power and to attenuate real and existing relationships in the data. This is possible using "errors-in-variables models." These models work much like regular generalised linear models, though account for measurement error in either the dependent variable, the independent variables, or both. McElreath (2020) provides a good introduction to the intuition behind error-in-variable models. Similarly, Bürkner (2017) provides an easy-to-use interface for fitting such models in `R` using the `brms` package (see also chapter 15.1 in Kurz 2020 for an applied example).

Technical Details: 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:

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

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

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

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

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

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

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

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

$$\mathbf{R} \sim \text{LKJ}(2) \quad \text{Prior on correlation matrix, } \mathbf{R}$$

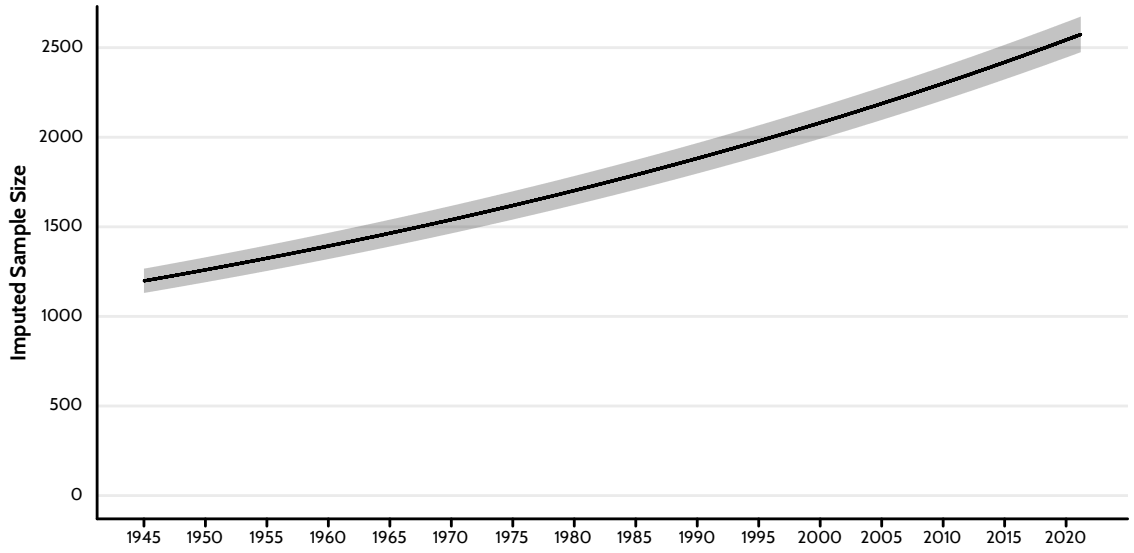


Figure 2: Imputed sample sizes in Britain between 1945 and 2021, estimated using sample size data from Jennings and Wlezien's "Timeline of Elections" dataset (2016).

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 2 shows the model's best estimate of the likely sample size of the average British voting intention poll between 1945 and 2021. These imputed values seem sensible and conform to our expection that sample sizes should increase over time. The model estimates that the average British voting intention poll included around 1,198 respondents in 1945. By 2021, the model suggests that this value had increased by 1,376 to 2,574 respondents per poll on average.

We use the model to produce a time series of estimates sample sizes between 1945 and 2021. 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

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:

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

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

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. This [DOES WHATEVER IT DOES]. Updating our model specification again, we get:

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

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

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

As we rely on Bayesian methods, our final step is to provide the model with a set of prior distributions.

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

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

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

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

$$\delta_j \sim \text{Normal}(0, 0.1) \text{ for } j \text{ in } 1, ..., J \quad \text{Prior on house effects, } \delta$$

$$\omega_t \sim \text{Normal}(0, 0.1) \text{ for } t \text{ in } 1, ..., T-1 \quad \text{Prior on random shocks, } \omega$$

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

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

Model pinned at both ends and includes only one set of polling figures – loop over each pair of elections and each party between 1955 and the present day.

Validating Our Estimates

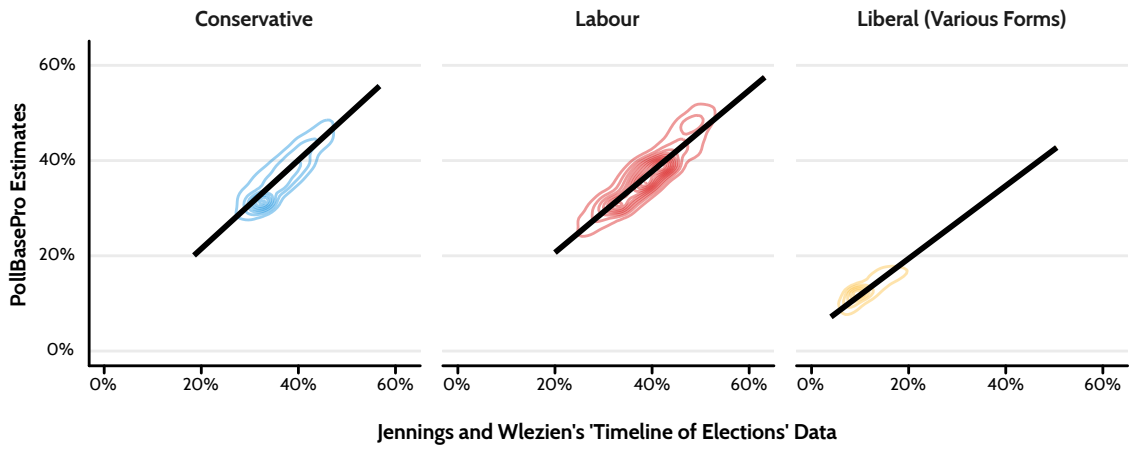


Figure 3: Estimates for each party from the pollbasepro data appear well-validated against raw polling data from Jennings and Wlezien's 'Timeline of Elections' dataset (2016).

To validate our estimates, we compare them against raw polling data from Jennings and Wlezien's "Timeline of Elections" dataset (2016). These data contain 4,302 polls from Britain between 15 June 1943 and 6 June 2017. Given that the data we use to produce our estimates are so comprehensive, it is likely that most polls appear in both datasets. Still, the Timeline data provide a good test as Jennings and Wlezien compiled them independently. As figure 3 shows, our estimates are well validated. Correlations between the two series are strong and positive, as we would expect. Their mean absolute error (MAE) and root-mean-square error (RMSE) are also low in all cases. The Conservatives showed a correlation of 90.1% (95% CI: 89.5% to 90.7%), an MAE of 2.18 percentage points, and an RMSE of 0.03; Labour, a correlation of 92.0% (95% CI: 91.5% to 92.4%), an MAE of 2.89 points, and an RMSE of 0.04; and the Liberals, a correlation of 91.6% (95% CI: 91.1% to 92.1%), an MAE of 2.22 points, and an RMSE of 0.03.

Open-Source Data Pipeline

We recognise that some users will find understanding our modelling decisions more simple if they were able to see our code[1]. This transparency also has other benefits: it allows our users to identify mistakes in our code. After all—and like any project of this nature—our data pipeline likely contains minor errors or inefficiencies that could affect the estimates that we obtain. To guard against this, and provide our users with a more in-depth look at our modelling process, we have hosted our entire data pipeline 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.

[1]Note that we estimate all of our models using `R` version 4.0.4 (2021-02-15) and either version 2.14.4 of the `brms` package (Bürkner 2017) or version 2.26.0 of `cmdstan`, an interface to the Stan probabilistic programming language (Carpenter et al. 2017)

Variable List: PollBase (Historical British Election Polls from 1945 to 2021)

Name	Description
`id`	Unique poll identification number
`election`	Date of last general election
`govt`	Largest party in government after the last general election
`start`	First day of fieldwork
`end`	Last day of fieldwork
`pollster`	Polling company that conducted the poll
`n`	Sample size
`con`	Voting intention: Conservative
`lab`	Voting intention: Labour
`lib`	Voting intention: Liberal
`con_ldr`	Leader of the Conservative Party
`lab_ldr`	Leader of the Labour Party
`lib_ldr`	Leader of the Liberals (various forms)

Variable List: PollBasePro (Daily British Voting Intention from 1955 to 2021)

Name	Description
`date`	Date
`election`	Date of last general election
`govt`	Largest party in government after the last general election
`con_est`	Posterior mean: Conservative voting intention
`con_err`	Posterior error: Conservative voting intention
`lab_est`	Posterior mean: Labour voting intention
`lab_err`	Posterior error: Labour voting intention
`lib_est`	Posterior mean: Liberal voting intention
`lib_err`	Posterior error: Liberal voting intention
`con_ldr`	Leader of the Conservative Party
`lab_ldr`	Leader of the Labour Party
`lib_ldr`	Leader of the Liberals (various forms)
`week`	Weekly subset indicator
`month`	Monthly subset indicator
`quarter`	Quarterly subset indicator
`year`	Yearly subset indicator

Change Log

For the sake of openness and transparency, we provide a change log that lists all updates and changes made to the PollBasePro datasets and documentation over time. If you think that you have found a problem with either, please raise an issue on the project's GitHub repository.

Version 0.1.0 (Beta)

- Beta release of data, user guide, and accompanying paper

References

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.

Carpenter, Bob, Andrew Gelman, Matthew D. Hoffman, Daniel Lee, Ben Goodrich, Michael Betancourt, Marcus Brubaker, Jiqiang Guo, Peter Li, and Allen Riddell. 2017. "Stan: A Probabilistic Programming Language." *Journal of Statistical Software* 76 (1): 1–32. https://doi.org/10.18637/jss.v076.i01.

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

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.

Kurz, A. Solomon. 2020. "Statistical Rethinking with Brms, Ggplot2, and the Tidyverse: Second Edition (version 0.1.1)." https://bookdown.org/content/4857/.

McElreath, Richard. 2020. *Statistical Rethinking: A Bayesian Course with Examples in R and Stan*. Second. CRC Texts in Statistical Science. Boca Raton: CRC Press.

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