Polling bias and undecided voter allocations in recent US Presidential elections

Joshua J. Bon with

T. Ballard & **B. Baffour**

JSM July 30th 2019

Mathematical Sciences, Queensland University of Technology

Motivation: Undecided voters to election day (2004–2016)

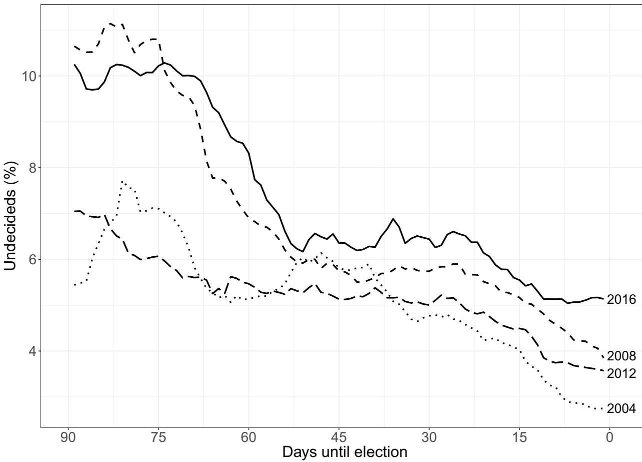

Figure 1: Mean level of undecided voters from US presidential elections. Weighted average from national polls that occur within a two-week window centred at x.

Motivation: Polling error versus undecided voters

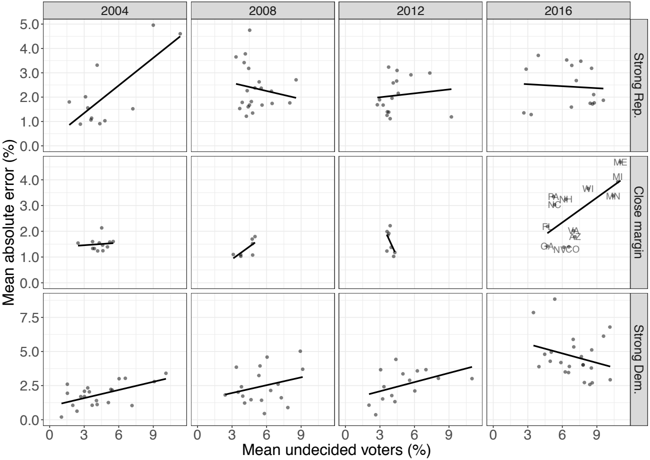

Figure 2: State-level mean absolute error versus mean undecided voters. Polls within 35 days of elections. "Close margin" categorises state-level elections with absolute margin $\leq 6\%$.

How do we assess state-level polling error?

A multilevel model for polling error[1]

$$
\begin{aligned}
y_i &\sim \mathcal{N}(p_i, \sigma_i^2) \\
\text{logit}(p_i) &= \text{logit}(v_{r[i]}) + \alpha_{1r[i]} + t_i \beta_{1r[i]} \\
\sigma_i^2 &= \frac{p_i(1-p_i)}{n_i} + \tau_{1r[i]}^2
\end{aligned}
\tag{1}
$$

- $r[i]$ indexes poll i to state-year r
- v_r is the actual poll result in state-year r
- $\alpha_{1r} + t_i \beta_{1r}$ time varying bias away from truth (election result)
- t_i is time until election day
- τ_{1r}^2 accounts for the excess variance above a SRS

[1] H. Shirani-Mehr *et al.*, *Journal of the American Statistical Association* **113**, 607–614 (2018).

How do we incorporate undecided voters?

Standardising polls and undecided voters

Standard to assume proportional allocation of undecided voters by

$$p_i = \frac{R_i}{R_i + D_i} \tag{2}$$

however you may include undecideds by letting

$$p_i' = \frac{R_i + \lambda U_i}{R_i + D_i + U_i} \tag{3}$$

where $0 \leq \lambda \leq 1$ allocates the undecided voters. The values p_i and p_i' coincide under the assumption of static proportionate allocation:

$$\lambda = \frac{R_i}{R_i + D_i} \tag{4}$$

Incorporating undecided voters into the model

We would like to include uncertainty in undecided allocations.

Assuming there is some bias away from proportionate splitting

$$\lambda = \frac{R_i}{R_i + D_i} + \theta_i \qquad (5)$$

leads to the identity

$$p_i' = p_i + u_i\theta_i \qquad (6)$$

which can be incorporated into the mean of the original model...

Incorporating undecided voters

...but, there several issues:

1. Undecided voter levels are time-varying
2. Undecided voters are not reported in $\approx 10\%$ of polls
3. Undecided voter levels are themselves poll estimates $\Longrightarrow$ measurement error
4. θ_i is a parameter for every poll

Incorporating undecided voters

Model the undecided voters with

$$u_i \sim \mathcal{N}\left(\alpha_{2r[i]} + t_i\beta_{2r[i]}, \tau^2_{2r[i]}\right) \tag{7}$$

- α_{2r} is the the election day mean for each state-year
- Polls that don't include u_i are accounted for since we estimate (and use) the state-year parameters

Addresses time varying, missing data, and measurement error concerns by using state-year estimates of undecided voters on election day.

A model with undecided voters (and house effects)

$$
\begin{aligned}
y_i &\sim \mathcal{N}(p_i, \sigma_i^2) \\
\text{logit}(p_i) &= \text{logit}(v_{r[i]}) + \alpha_{1r[i]} + t_i\beta_{1r[i]} - \alpha_{2r[i]}\gamma_{g[i]} + \kappa_{h[i]} \\
\sigma_i^2 &= \frac{p_i(1-p_i)}{n_i} + \tau_{1r[i]}^2
\end{aligned}
\tag{8}
$$

- α_{2r} is the election day estimate of undecided voters for each state-year (estimated by (7) concurrently)
- γ_g controls the amount of biasing effect from undecided voters in each election-year$\times$result-margin g
- κ_h is the house-effect from polling firm (or conglomerate) h

Data sources

- State level polling data
  - 2012, 2016 from Pollster API[2]
  - 2004, 2008 from US Election Atlas[3]
- Polls up to 35 days prior to their respective election included
- 2,044 state-level polls total ($\approx$ 90% had undecideds reported)
- No 2000 or earlier polls with sufficient data on undecided voters were found.

[2] Huffington Post, *Pollster API V2*, http://elections.huffingtonpost.com/pollster/api/v2, Accessed: 2016-12-20, Huffington Post, 2016.

[3] D. Leip, *Atlas of US Presidential Elections*, http://uselectionatlas.org/, Accessed: 2016-12-20, 2008.

So what did we find?

Model estimates - sources of error

Table 1: Average election-level absolute bias and average election-level standard deviation across state-elections in given year(s) from model (8) with assumption of proportional allocation of undecided voters.

	2004	2008	2012	2016	*Overall* 2004–2016
Average absolute bias	0.8% (0.11)	1.0% (0.10)	1.3% (0.10)	2.6% (0.10)	1.7% (0.06)
Average absolute election day bias	0.8% (0.12)	0.9% (0.11)	1.3% (0.14)	2.4% (0.12)	1.6% (0.07)
Average absolute undecided voter bias	0.3% (0.17)	0.4% (0.17)	1.0% (0.29)	2.1% (0.25)	1.1% (0.11)
Average absolute house effects	0.6% (0.15)	0.4% (0.12)	0.2% (0.08)	0.2% (0.09)	0.3% (0.09)
Average standard deviation	2.2% (0.04)	2.2% (0.04)	2.1% (0.04)	2.4% (0.05)	2.2% (0.03)
Average election day undecided	3.3% (0.24)	3.8% (0.21)	3.0% (0.21)	5.5% (0.28)	4.2% (0.14)

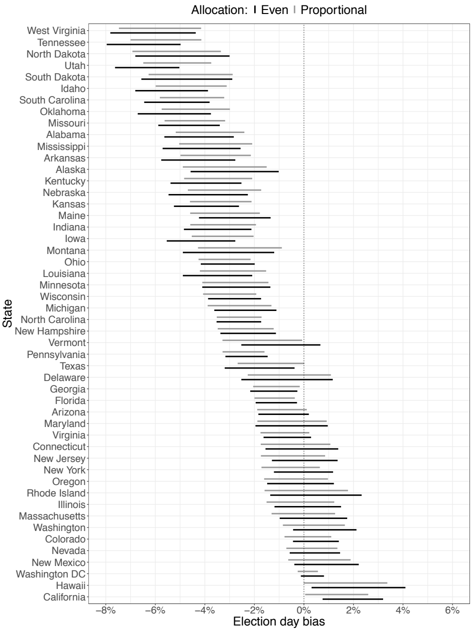

Figure 3: 95% Credible intervals for state election day bias (2016).

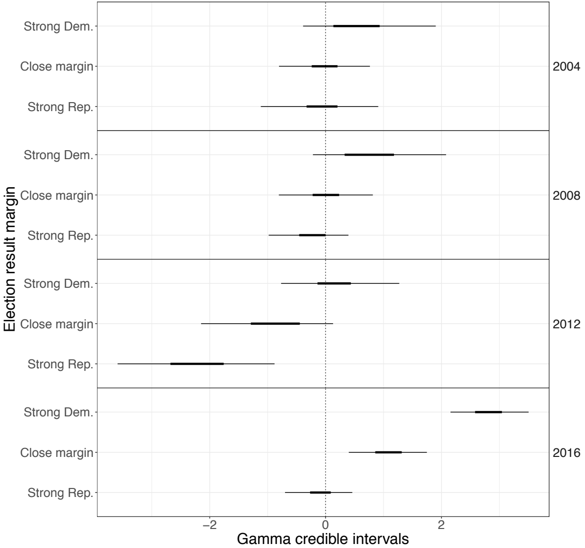

Figure 4: 95% and 50% credible intervals for γ_g on logit scale. A positive value indicates a bias away from proportional allocation of undecided voters in favour of the Republican candidate.

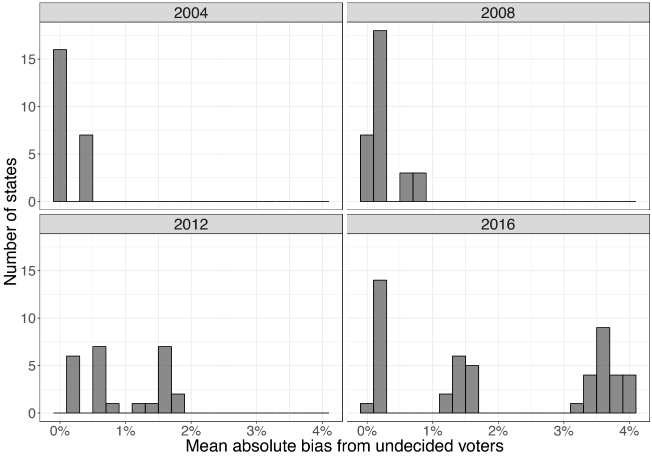

Figure 5: Histograms of the average absolute bias from undecided voters for each state-level election, separated by year. The bias from undecided voters is the quantity $\alpha_{2r}\gamma_g$ in the model. A positive value indicates a bias away from proportional allocation of undecided voters in favour of either candidate.

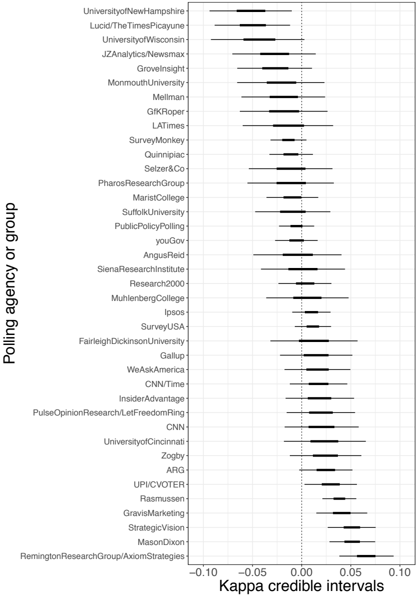

Figure 6: 95% (outer line) and 50% credible intervals for house effects bias from polling organisations in the model (κ_h), on the logit scale

Concluding remarks

- In 2016, 5.5% of voters were undecided on election day, up from 3.0–3.8% in previous years
- Undecided voters biased polls in the 2016 US presidential election by 2.1 percentage points on average
- A static, proportionate split in undecided voters between leading candidates was a bad assumption in 2016, less so in previous years
- Pollsters and modellers should move towards stochastic allocation methods to allow uncertainty from undecided voters to propagate through models
- Every poll should report undecided level

Key references

H. Shirani-Mehr *et al.*, *Journal of the American Statistical Association* **113**, 607–614 (2018)

J. J. Bon *et al.*, *Journal of the Royal Statistical Society: Series A (Statistics in Society)* **182**, 467–493 (2019)

Code and data available: `https://github.com/bonStats/undecided-voters-us-pres-elections`

Appendix

Table 2: Priors used in models for analysis of state polls.

Model	Component	*Prior*	*Hyper-prior* Mean	Variance
Polling	Mean	$\alpha_{1r} \sim \mathcal{N}(\mu_{1\alpha}, \sigma^2_{1\alpha})$	$\mu_{1\alpha} \sim \mathcal{N}(0, 0.2)$	$\sigma_{1\alpha} \sim \mathcal{N}_+(0, 0.2)$
		$\beta_{1r} \sim \mathcal{N}(\mu_{1\beta}, \sigma^2_{1\beta})$	$\mu_{1\beta} \sim \mathcal{N}(0, 0.2)$	$\sigma_{1\beta} \sim \mathcal{N}_+(0, 0.2)$
		$\gamma_g \sim \mathcal{L}(0, 0.05)$		
		$\kappa_h \sim \mathcal{N}(\mu_\kappa, \sigma^2_\kappa)$	$\mu_\kappa \sim \mathcal{N}(0, 0.05)$	$\sigma_\kappa \sim \exp(1/0.05)$
	Variance	$\tau^2_{1r} \sim \mathcal{N}_+(0, \sigma^2_{1\tau})$		$\sigma_{1\tau} \sim \mathcal{N}_+(0, 0.05)$
Undecided voters	Mean	$\alpha_{2r} \sim \mathcal{N}(\phi_{y[r]}, \sigma^2_{2\alpha})$	$\phi_y \sim \mathcal{N}(0.04, 0.01)$	$\sigma_{2\alpha} \sim \mathcal{N}_+(0, 0.02)$
		$\beta_{2r} \sim \mathcal{N}(\mu_{2\beta}, \sigma^2_{2\beta})$	$\mu_{2\beta} \sim \mathcal{N}(0, 0.02)$	$\sigma_{2\beta} \sim \mathcal{N}_+(0, 0.02)$
	Variance	$\tau^2_{2r} \sim \mathcal{N}_+(0, \sigma^2_{2\tau})$		$\sigma_{2\tau} \sim \mathcal{N}_+(0, 0.01)$

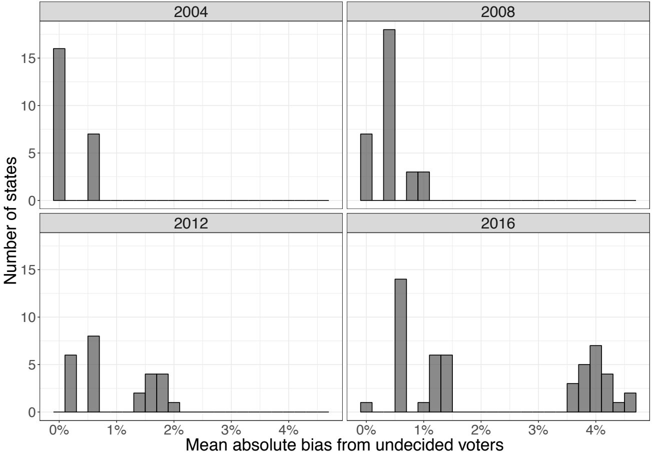

Figure 7: Histograms of the average absolute bias from undecided voters for each state-level election, separated by year. The bias from undecided voters is the quantity $\alpha_{2r}\gamma_g$ in the model. A positive value indicates a bias away from 50/50 allocation of undecided voters in favour of either candidate.

Table 3: Average house effects across elections. Only those polling agencies with absolute mean posterior greater than 0.5% are shown.

	Posterior	
Polling agency or group	mean	s.d.
ARG	0.61	0.35
CNN	0.50	0.48
Gravis Marketing	1.01	0.33
Grove Insight	-0.67	0.49
JZ Analytics / Newsmax	-0.69	0.54
Lucid / The Times Picayune	-1.23	0.49
Mason Dixon	1.27	0.29
Monmouth University	-0.51	0.55
Rasmussen	0.95	0.22
Remington Research Group / AxiomStrategies	1.64	0.35
Strategic Vision	1.27	0.31
University of Cincinnati	0.58	0.53
University of New Hampshire	-1.28	0.53
University of Wisconsin	-1.06	0.59
UPI/CVOTER	0.69	0.32
Zogby	0.60	0.46