Presidential Election Polling Analysis

By Cole Adams, Sibo Peng and Mariya Harris

ST 433

Final Project

Nov. 20, 2020

Introduction and Project Goals

- The main goal of this analysis is to determine whether or not there exists systematic polling bias in presidential elections
    - Polling bias is defined as the difference between the actual percentage of votes that went to a GOP presidential candidate and the estimated percentage by polling methods
- Analysis will be broken up into three interweaving sections:
    - Calculation of polling bias using weighting methods
    - Method for testing the existence of polling bias
    - Analysis of method results

Polling Bias

Polling Bias Formula:

- $B_{it} = E(Y_{it} - X_{it})$ where:
  - B_{it} = Polling bias in state *i* and election year *t*
  - Y_{it} = Actual percentage of votes that went to a GOP presidential candidate in state *i* and election year *t*
  - X_{it} = Estimated percentage of votes that went to a GOP presidential candidate using polling averages in state *i* and election year *t*

Polling Average Formula:

- $X_{it} = \sum_{j=1}^{Nij} w_{itj} P_{jt}$ where:
  - w_{ij} = weight for poll *j* in state *i* and election year *t*
  - P_{jt} = Polling average for poll *j* in election year *t*

Y_{it} Collection:

- Actual vote percentages were collected off the final results in years 2012 and 2016 and the current percentages for 2020 as they were on November 7th

Weighting Method

- Goal: Create a weighting procedure for all polls in a specific state and election, to eventually be used to calculate polling bias
- $w_{ij} = \frac{\frac{1}{m_{itj}/z_{it}}}{\sum_{j=1}^{N}(\frac{1}{m_{itj}/z_{it}})}$ where:
  - m_{ijt} is the months away from the election month that poll *j* was conducted + 1 in state *i* and election year *t*
    - Ex: October Poll: 1 month away from November + 1 = $m_j = 2$
  - z_{it} is the total number of months away for all polls in state *i* and election year *t*

- Example: Suppose North Carolina in Election Year 2016 had 3 polls:

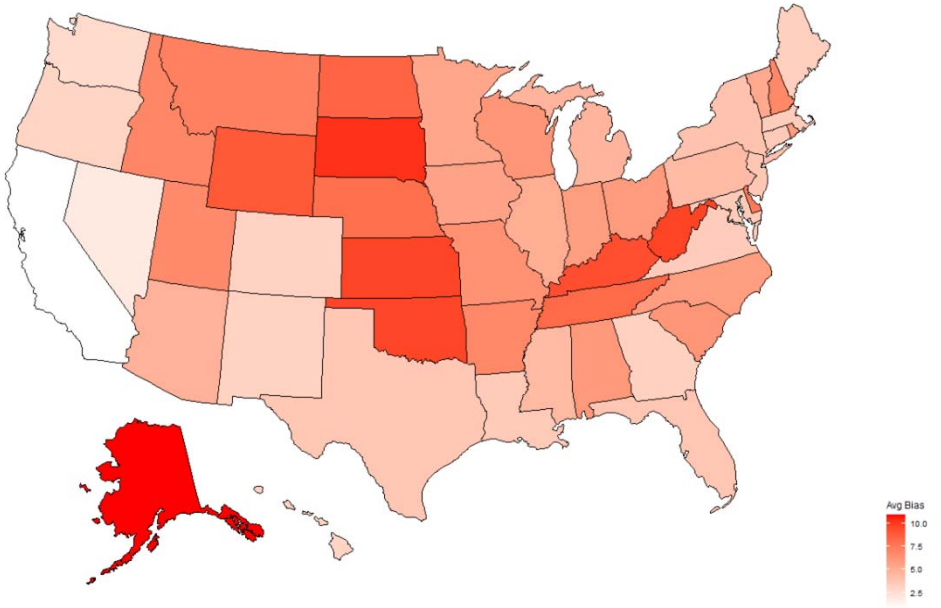
		Top Fraction: $\frac{1}{m_{itj}/z_{it}}$	Bottom Fraction: $\sum_{j=1}^{N}(\frac{1}{m_{itj}/z_{it}})$	Weight:	Weighted Poll:
Poll A: 51% GOP Conducted in October	m = 2 Running Z_{it} = 2	1 / (2 / 10) 1 / .2 = 5		5 / 10.33 w = .484	.51 * .484 wp = .247
Poll B: 48% GOP Conducted in September	m = 3 Running Z_{it} = 5	1 / (3 / 10) 1 / .3 = 3.33	5 + 3.33 + 2 = 10.33	3.33 / 10.33 w = .322	.48 * .322 wp = .155
Poll C: 45% GOP Conducted in July	m = 5 Total Z_{it} = 10	1 / (5 / 10) 1 / .5 = 2		2 / 10.33 w = .194	.45 * .194 wp = .087
				Sum: 1	49.9% GOP

Calculating the Bias

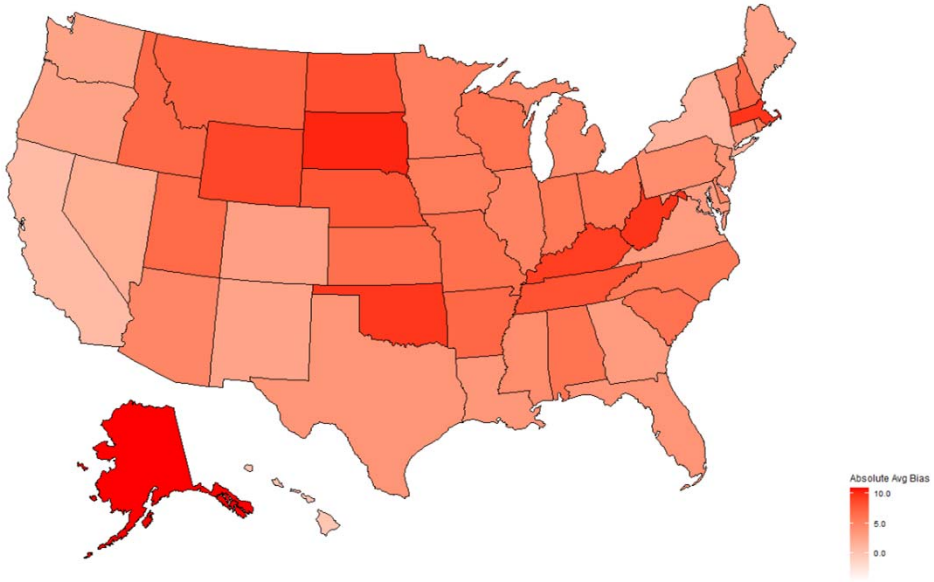
- To test for systematic polling bias we looked at both absolute average bias per state and the average bias per state
  - Absolute Average: Add up the absolute value of biases for each election and calculate the average.
  - Average: Add up the values of biases for each election and calculate the average
- For example, suppose a state had biases of 3, -1, and 2
  - The absolute average would be (3 + 1 +2) / 3 = 2
  - The average would be (3 - 1 + 2) / 3 = 1.33
- This was done so that we could analyze whether there was bias in general and then which way it swung

Bias Visualizations

Average Bias Across 3 Elections



Absolute Average Bias Across 3 Elections



- Appears to be some mild spatial correlation across certain areas:
  - Midwest and mideast tend to have higher average/ absolute biases
  - Coastal states tend to have lower average bias/ absolute biases

Conceptual Model

- Goal: To test whether there is a systematic deviation from the average and absolute average bias under the assumption that the bias is constant over state and election. To do so we constructed a gaussian CAR model with the intercept and the error as shown below.
- $B_{it} = B + \varepsilon_{it}$
- B_{it} is the predicted bias for each state and election
- B is the intercept
- ε_{it} is the error with a spatiotemporal correlation

Model in Code

Gaussian CAR Model

- The function used to create the model in R was the S.CARleroux() function located inside the CARBayes package
- Adjacencies used in the model were neighbor adjacencies, meaning that in the adjacency matrix, if two states neighbored one another they would receive a value of 1; if they did not they would receive a value of 0
- For both the absolute average value and the average values of Bias, we fit the models and computed the results.
- n.sample and burnin were adjusted to to generate a better model

```
#absolute value of B
cm <- S.CARleroux(B~1, family="gaussian", W=W1, burnin=1000000, n.sample=10000000, thin=10, verbose=FALSE)
```

Absolute Average Value Model Comparisons

n.sample = 1000000,
burnin = 200000, thin= 10

```
#################
#### Model fitted
#################
Likelihood model - Gaussian (identity link function)
Random effects model - Leroux CAR
Regression equation - B ~ 1
Number of missing observations - 0

############
#### Results
############
Posterior quantities and DIC

            Median   2.5%  97.5% n.effective Geweke.diag
(Intercept) 5.6264 4.9737 6.2809     80000.0        -1.2
nu2         5.5050 3.8180 8.3408     76493.1         0.7
tau2        0.0083 0.0021 0.0947      8521.2         0.9
rho         0.7353 0.0499 0.9977     21639.3         0.6

DIC =  236.2557       p.d =  2.20325       LMPL =  -118.07
```

n.sample = 10000000,
burnin = 1000000, thin= 10

```
#################
#### Model fitted
#################
Likelihood model - Gaussian (identity link function)
Random effects model - Leroux CAR
Regression equation - B ~ 1
Number of missing observations - 0

############
#### Results
############
Posterior quantities and DIC

            Median   2.5%  97.5% n.effective Geweke.diag
(Intercept) 5.6261 4.9735 6.2819    904981.3        -0.5
nu2         5.5021 3.7519 8.3343     21342.8         0.3
tau2        0.0084 0.0021 0.1164      1537.7        -0.4
rho         0.7319 0.0456 0.9976    150893.4         0.2

DIC =  233.5703       p.d =  1.135926       LMPL =  -117.85
```

- Both models achieved similar results, with both clearing 1000 n.effective for each parameter
- All four parameters are significant as they do not include 0 in their 95% confidence intervals

Average Value Model Comparisons

n.sample = 1000000, burnin = 200000, thin= 10

```
#################
#### Model fitted
#################
Likelihood model - Gaussian (identity link function)
Random effects model - Leroux CAR
Regression equation - B_na ~ 1
Number of missing observations - 0

############
#### Results
############
Posterior quantities and DIC

            Median   2.5%   97.5% n.effective Geweke.diag
(Intercept) 5.2983 4.4921  6.1094     77468.2        -1.9
nu2         8.3087 5.2561 12.5741       415.3        -0.2
tau2        0.0085 0.0021  0.3522       100.5         0.1
rho         0.7249 0.0448  0.9977      7243.1         0.0

DIC =  247.0381       p.d =  -2.059476       LMPL =  -128.42
```

n.sample = 10000000, burnin = 1000000, thin= 10

```
#################
#### Model fitted
#################
Likelihood model - Gaussian (identity link function)
Random effects model - Leroux CAR
Regression equation - B_na ~ 1
Number of missing observations - 0

############
#### Results
############
Posterior quantities and DIC

            Median   2.5%   97.5% n.effective Geweke.diag
(Intercept) 5.2979 4.4914  6.1028    881630.1        -0.7
nu2         8.3411 5.7171 12.6337     34031.9         1.5
tau2        0.0084 0.0021  0.1127      1490.5        -1.7
rho         0.6157 0.0359  0.9865    244083.4         1.7

DIC =  255.2879       p.d =  1.253284       LMPL =  -129.09
```

- The model on the right is preferred as it had more n.effective and cleared 1000 for all parameters whereas the left model did not
- All four parameters are significant as they do not include 0 in their 95% confidence intervals

Parameter Interpretations

Model Results

Component	Average	Absolute Average
Intercept	5.2983	5.6264
nu^2	8.3411	5.5050
tau^2	.0084	.0083
rho	.6157	.7353

- The intercept represents the expected bias for each state as there are no covariates
- nu^2 is the nugget variance, the variance occurring from non-spatial error. This composes a large majority of the total error
- tau^2 is the spatial variance, the variance occurring from spatial error. This makes up a very small small percentage of the total error
- rho is the spatial dependence parameter. A positive rho indicates positive spatial dependence, meaning an increase one neighbor indicates a corresponding increase in the other. A lower rho indicates that there is not a strong spatial dependance

Final Conclusion

Intercept Values

Component	Median	2.5%	97.5%
Average	5.2983	4.4921	6.1094
Absolute Average	5.6264	4.9737	6.2809

- The model using absolute averages had a significant intercept. Thus, it can be interpreted that there is significant bias in the polling methods
- The model using averages also had a significant intercept. Thus, it can be interpreted that there is significant bias in the polling methods in the direction that the GOP candidate receives a higher percentage of votes than forecasted by the polling methods