# Presidential Election Poll Bias Spatial Analysis

## ST533 - Fall 2020

Rebekah Colonnese, Andrew Freedman, and Megan Tabor

# About our Project

**Motivation**: study the bias in state-level Presidential election polls from the elections of 2012, 2016 and 2020.

**Objective**: Use spatial modelling techniques to:

1. Devise a method to combine individual polls to forecast election results in each state and year
2. Test whether there is systematic polling bias assuming bias is constant over state and election
3. Test whether bias varies by state and/or election

# Data Description

- There are four main variables used throughout this project
  - $Y_{it}$ is the percentage of actual votes for each year and state for the GOP candidate
  - $X_{it}$ is the polling average calculated in objective 1
  - $Z_{it}$ is the difference between $Y_{it}$ and $X_{it}$ and in Objective 3 follows a CAR model with Leroux covariance
  - $B_{it}$ is the average of the Z's

$$B_{it} = \mathrm{E}(Y_{it} - X_{it})$$

# Objective 1

Devise a method to combine the individual polls to forecast the election results in each state and each year

Polling Average, $X_{it}$ is calculated as shown below:

$$X_{it} = \sum_{j=1}^{N_t} w_{itj} P_{jt},$$

Where $N_t$ is the total number of polls in election year t, $P_{jt}$ is poll j's estimated percent GOP support, and the weights $w_{ijt}$ sum to 1.

# Objective 1

To examine the sensitivity to the definition of the polling average, we devised 3 weighting methods to apply to our spatial models

**Method 1 - "`all`"**

1. The weights for "`all`" were calculated by finding the number of days between poll date and the election (`num`)
2. For each state, the number of days was summed up to determine the total number of days (`tot`)
3. Created a new variable, `prop`, which divides `num` by `tot`
4. Lastly, `prop` was summed up for each state (`summ`) and then each `prop` was divided by the `summ` to make a new variable `weight1`.
   a. This added step was to make sure the weights summed to 1

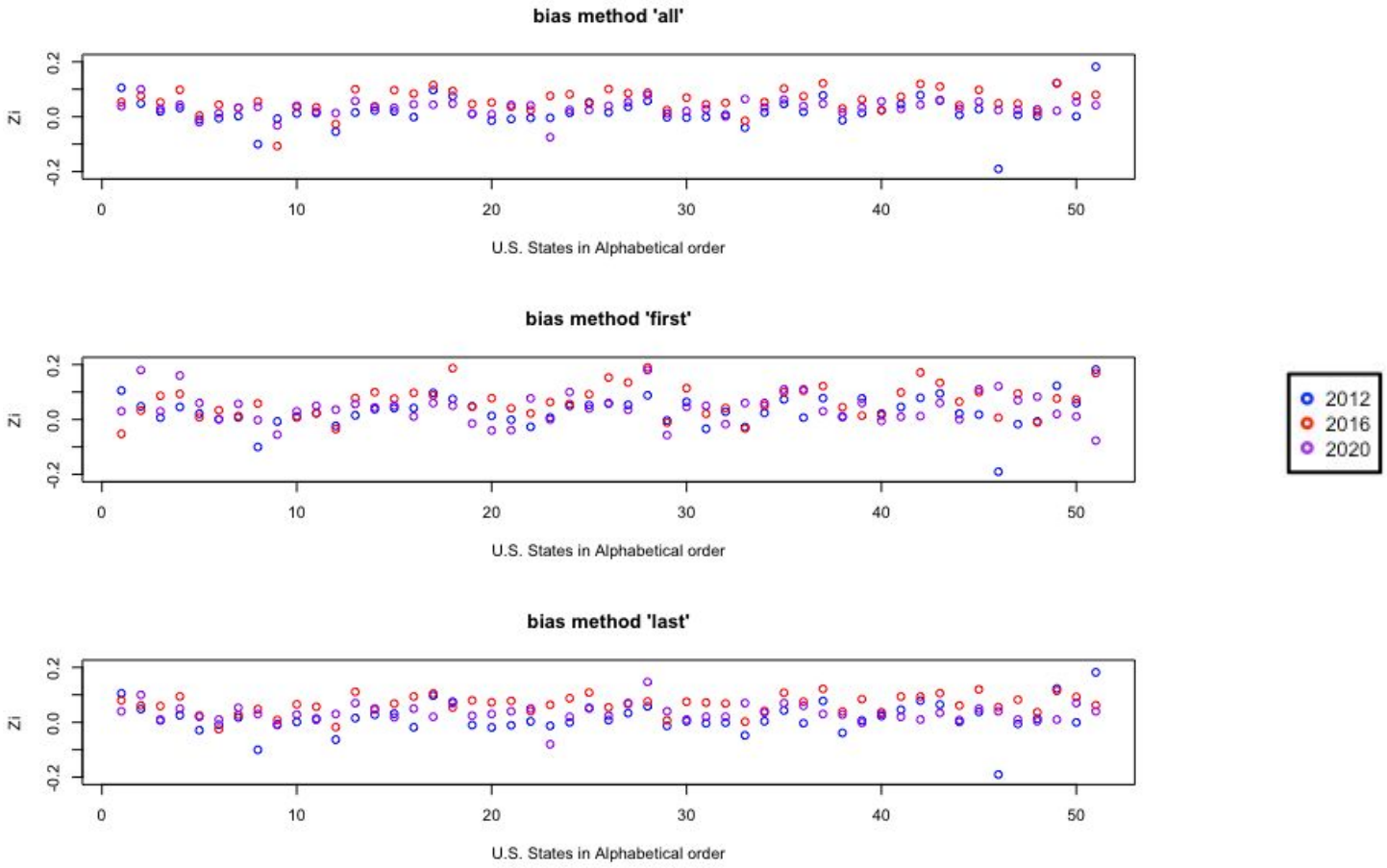
# Objective 1

**Method 2- "`first`"**

1. The first poll in each state was the only poll used to determine the polling average

**Method 3- "`last`"**

1. The last poll in each state was the only poll used to determine the polling average


bias method 'all'
Zi
0.2
0.0
-0.2
0
10
20
30
40
50
U.S. States in Alphabetical order
bias method 'first'
Zi
0.2
0.0
-0.2
0
10
20
30
40
50
U.S. States in Alphabetical order
bias method 'last'
Zi
0.2
0.0
-0.2
0
10
20
30
40
50
U.S. States in Alphabetical order
2012
2016
2020

# Objective 2

Test whether there is systematic polling bias under the assumption that the bias is constant over state and election

**Approach**:

- Model Fit: Spatial Generalized Linear Model (gaussian link)
- Formula: Y~1
  - Where the Y's are the Z for each state averaged over year

**Motivation**:

- Test whether the mean is zero ie. the intercept is zero
- If it is not zero, this will tell us if there is systematic polling bias

# Objective 2

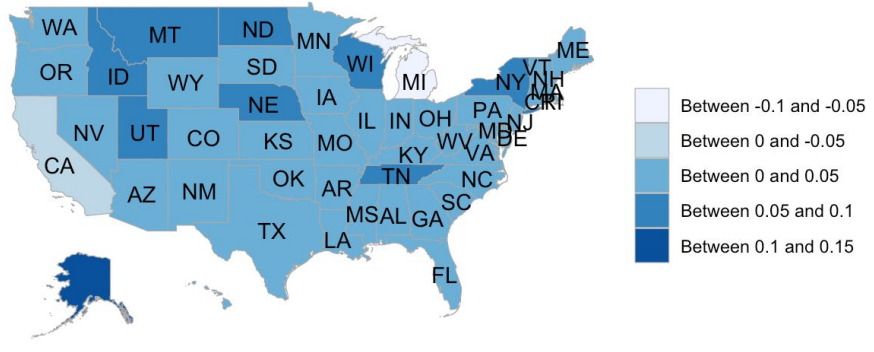
## Weight Method - "all"

| YEAR | GOP under (U) or over (O) performed |
|---|---|
| 2012 | GOP U in west coast, CO, and IA<br>GOP O elsewhere, especially in WY, AL, and WV |
| 2016 | GOP U in NY<br>GOP O elsewhere, especially in central US |
| 2020 | GOP U in CA and MI<br>GOP O everywhere else |

Z's for 2012 'all' Weight Method


Z's for 2016 'all' Weight Method


Z's for 2020 'all' Weight Method



# Objective 2

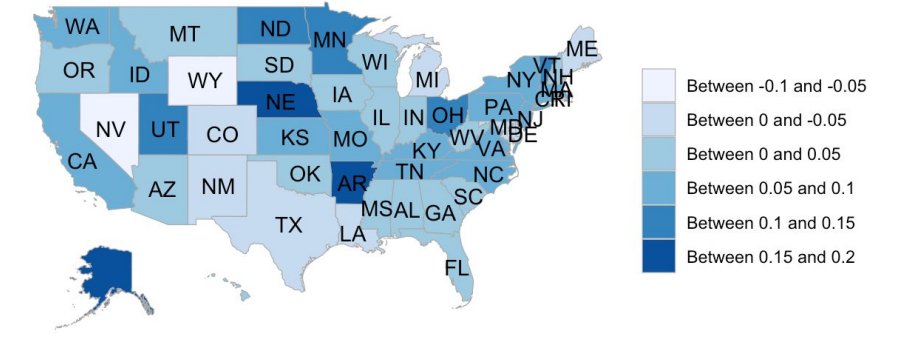
## Weight Method - "first"

| YEAR | GOP under (U) or over (O) performed |
|---|---|
| 2012 | GOP U in NV and WA<br>GOP O everywhere else, especially in WY |
| 2016 | GOP U in NV, WA, and AL<br>GOP O everywhere else, especially in midwest |
| 2020 | GOP U in various places, but especially in WY, NV<br>GOP O everywhere else, especially in NE, AR, OH, MN, ND |

Z's for 2012 'first' Weight Method


Z's for 2016 'first' Weight Method


Z's for 2020 'first' Weight Method

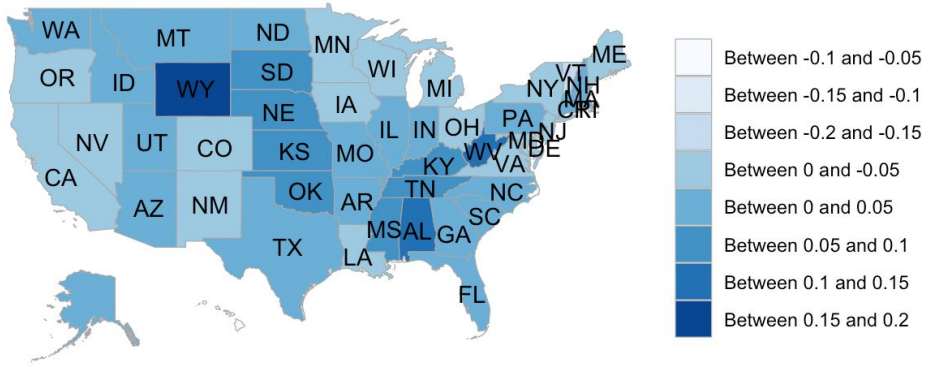


# Objective 2

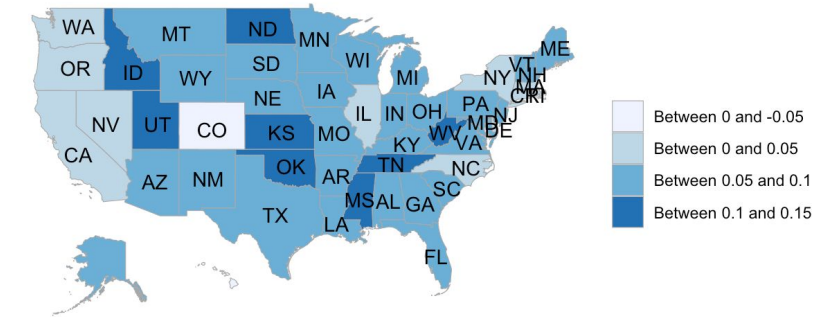
## Weight Method - "last"

| YEAR | GOP under (U) or over (O) performed |
|---|---|
| 2012 | GOP U on west coast and midwest<br>GOP O especially in central US and south |
| 2016 | GOP U on west coast, NC, NY, IL, and especially CO<br>GOP O in south, midwest, and central US |
| 2020 | GOP U in PA and MI<br>GOP O **everywhere else** |

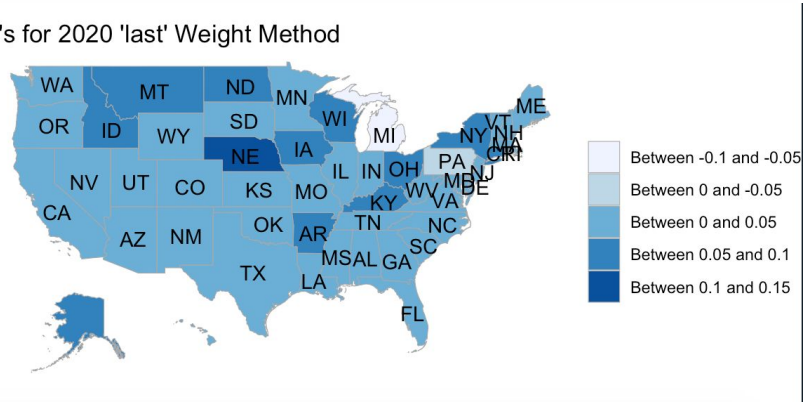
Z's for 2012 'last' Weight Method




Z's for 2016 'last' Weight Method




Z's for 2020 'last' Weight Method



# Objective 2

## Results

| Weighting Method | Lower Bound (2.5%) - Intercept | Upper Bound (97.5%) - Intercept | Statistical Significance |
|---|---|---|---|
| 1 - `all` | 0.0259 | 0.0475 | Yes |
| 2 - `first` | 0.0301 | 0.0575 | Yes |
| 3 - `last` | 0.0263 | 0.0475 | Yes |

All three weighting methods suggest that Z's are statistically different from zero, which suggests there is systematic polling bias under the assumption that the bias is constant over state and election.

# Objective 3

Test whether the bias varies by state and/or election and display the estimated bias

**Approach**:

- Model Fit: CAR using basic state adjacency matrix (0 if not adjacent, 1 if adjacent)
- Formula: Y ~ Latitude + Longitude + Unemployment_Rate + I(Election Year)
  - Y is the $Z_{it}$ for each state and election and follows a spatial CAR model with Leroux covariance
  - (Latitude,Longitude) coordinates are the centroids of each state
  - Unemployment_Rate (UE) is the annual rate for each state and election year
  - Election Year 2012 - baseline (intercept)
  - Election Year 2016 (‘16) effect and Election Year 2020 (‘20) effect

**Motivation**:

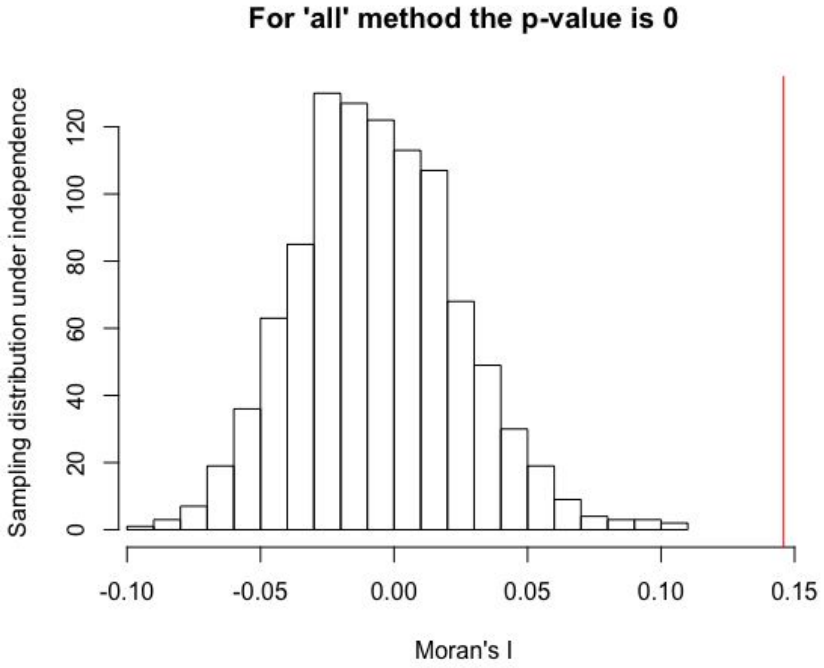
- We used a CAR model because we have areal data. The explanatory variables we selected so that we could see if there is a time and/or space dependence in Z’s

# Objective 3

## Weight Method 1 - "`all`"

| Parameter | Median | Lower Bound (2.5%) | Upper Bound (97.5%) | Statistically Significant |
|---|---|---|---|---|
| Intercept | 0.0577 | -0.0584 | 0.1766 | No |
| Latitude | -0.0006 | -0.0029 | 0.0016 | No |
| Longitude | -0.0003 | -0.0012 | 0.0007 | No |
| Unemployment | -0.0058 | -0.0101 | -0.0007 | Yes |
| Effect of 2016 Election | 0.0259 | 0.0072 | 0.0458 | Yes |
| Effect of 2020 Election | 0.0088 | -0.0080 | 0.0249 | No |
| $nu^2$ | 0.0012 | 0.0007 | 0.0017 | |
| $tau^2$ | 0.0038 | 0.0019 | 0.0077 | |
| rho | 0.6506 | .0.2654 | 0.9534 | |

# Spatial dependence in response var. Z_i (p<0.05)

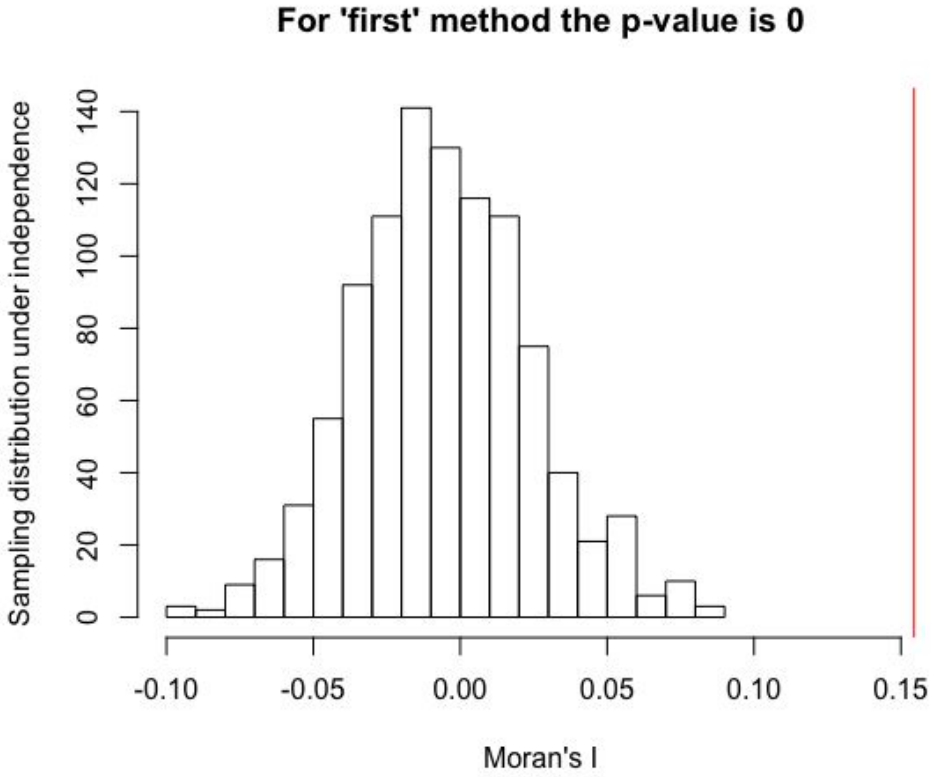


# Objective 3

## Weight Method 2 - "`first`"

| Parameter | Median | Lower Bound (2.5%) | Upper Bound (97.5%) | Statistically Significant |
|---|---|---|---|---|
| Intercept | 0.0601 | -0.1221 | 0.2335 | No |
| Latitude | -0.0002 | -0.0033 | 0.0032 | No |
| Longitude | -0.0005 | -0.0018 | 0.0009 | No |
| Unemployment | -0.0090 | -0.0157 | -0.0016 | Yes |
| Effect of 2016 Election | 0.0092 | -0.0190 | 0.0394 | No |
| Effect of 2020 Election | 0.0037 | -0.0194 | 0.0262 | No |
| $nu^2$ | 0.0021 | 0.0012 | 0.0033 | |
| $tau^2$ | 0.0091 | 0.0034 | 0.0192 | |
| rho | 0.5860 | 0.2162 | 0.9413 | |

# Spatial dependence in response var. Z_i (p<0.05)

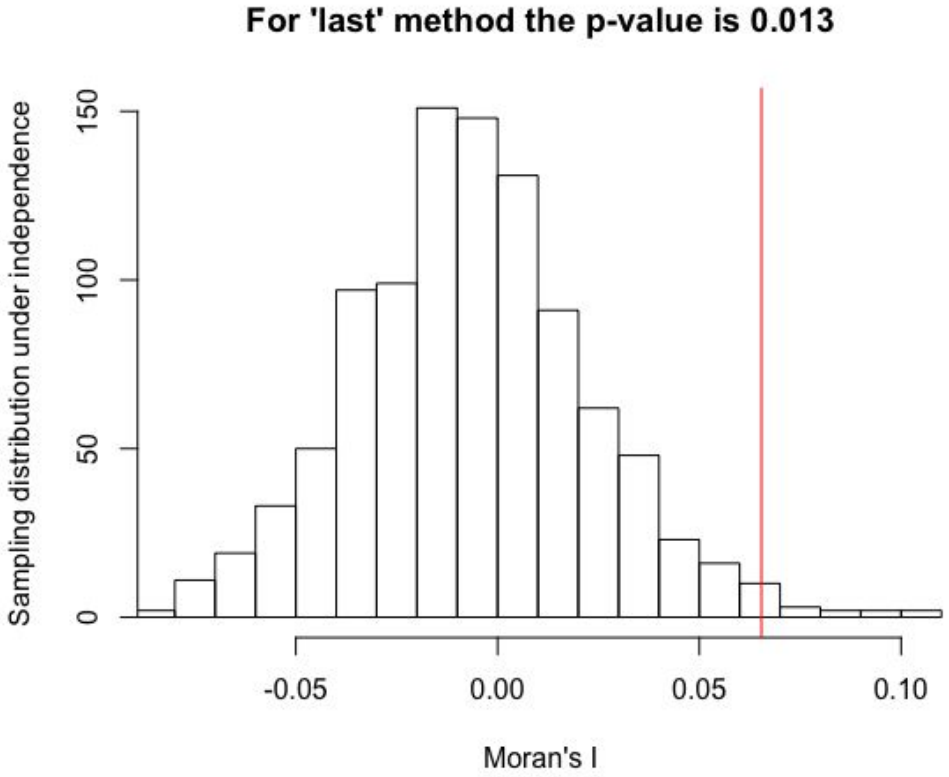


# Objective 3

## Weight Method 3 - "last"

| Parameter | Median | Lower Bound (2.5%) | Upper Bound (97.5%) | Statistically Significant |
|---|---|---|---|---|
| Intercept | 0.0575 | -0.0671 | 0.1788 | No |
| Latitude | -0.0006 | -0.0028 | 0.0015 | No |
| Longitude | -0.0002 | -0.0011 | 0.0007 | No |
| Unemployment | -0.0045 | -0.0095 | 0.0002 | No |
| Effect of 2016 Election | 0.0389 | 0.0157 | 0.0597 | Yes |
| Effect of 2020 Election | 0.0143 | -0.0037 | 0.0325 | No |
| $nu^2$ | 0.0014 | 0.0010 | 0.0019 | |
| $tau^2$ | 0.0029 | 0.0014 | 0.0060 | |
| rho | 0.6994 | 0.2498 | 0.9652 | |

# Spatial dependence in response var. Z_i (p<0.05)



# Objective 3

**Results**:

Method 1- “`all`”

- Effect of 2016 - indicates that there is a statistical difference in the polling bias between 2012 and 2016
- Unemployment - indicates that unemployment helps explain the variation in polling bias year-over-year

Method 2 - “`first`”

- Unemployment - indicates that unemployment helps explain the variation in polling bias year-over-year
- No statistically significant difference in polling bias between 2012 & 2016 or 2012 & 2020

Method 3 - “`last`”

- Effect of 2016 - indicates that there is a statistical difference in the polling bias between 2012 and 2016

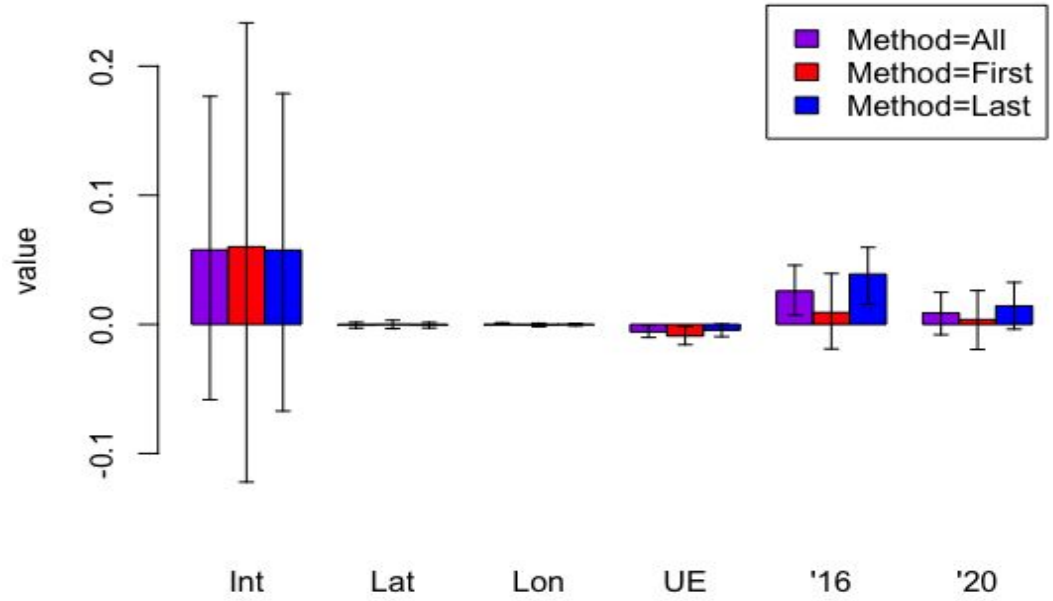
# Objective 3

## Results - Measure of Spatial Dependence

Method 1- "`all`", Method 2 - "`first`", and Method 3 - "`last`"

- Evidence of spatial dependence because:
  - $Tau^2 > nu^2$
    - CAR variance is larger than the nugget variance
  - Rho's are not close to one, but still high (between 0.60 and 0.70)
- Conclude: Bias varies by state

# Assess the Sensitivity to Polling Weights




"Int" is intercept / effect of 2012

# Conclusions

**Objective 2**

All three weighting methods produced the same conclusion: suggesting there is systematic polling bias under the assumption that the bias is constant over state and election.

**Conclusion**: Estimating the polling bias was not sensitive to change weighting methods used

**Objective 3**

Weighting methods 1 and 3 produced same conclusion that the effect of the 2016 election was significant and 2020 was not. Method 2 produced the result that none of the elections has a significant effect. All 3 weighting methods showed spatial dependence. Unemployment's effect was significant using all 3 weighting methods. Lat/long was not significant for any.

**Conclusion**: The results varied on which method was used on the polling average

# Questions?