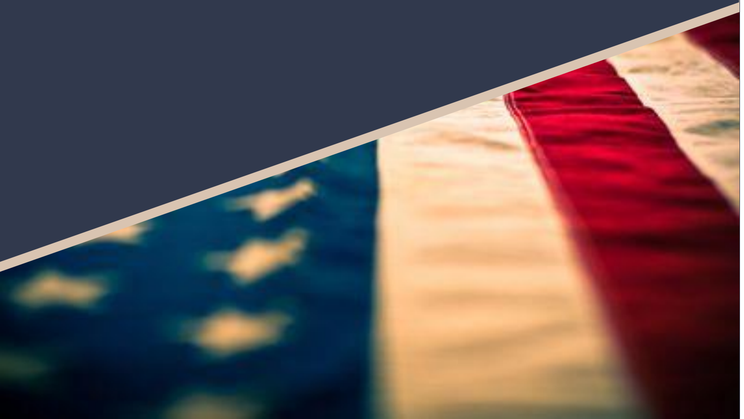
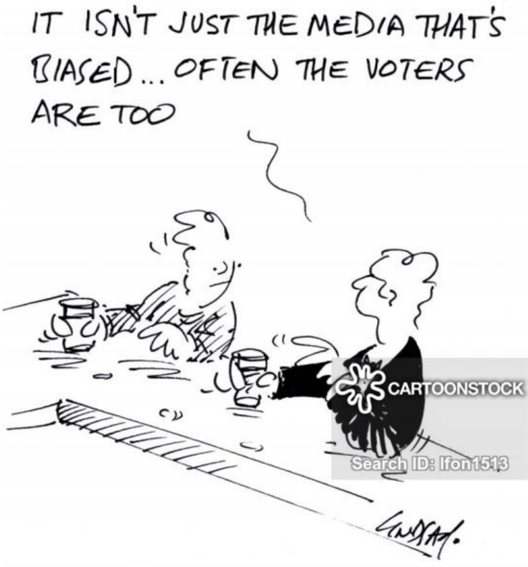
Polling Bias in US Presidential Elections?

Mariella Carbajal Carrasco, Laura Goodman,
& Andrew Hutchens

ST533 - Final Exam

Background




IT ISN'T JUST THE MEDIA THAT'S BIASED... OFTEN THE VOTERS ARE TOO
CARTOONSTOCK
Search ID: lfon1513

Why important:

- Many decisions within the USA are based on who is elected to lead in DC
- Polling is important since it predicts the election outcome based off the opinions of a sample of the population

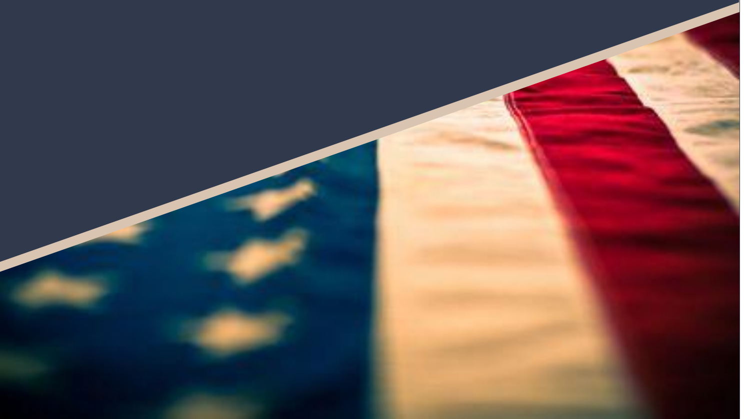
Goal:

- Determine whether polling was biased for US presidential elections in 2012, 2016, and 2020.

How:

- Compare public polls from each state in those years to election outcomes in each state.

How Polling Averages and Data Cleaning are done



- Since we are using publicly available polling information and election outcomes, need to:
  - Format data into the same way
  - Remove logs without enough information
  - Figure out how to average all polls into one value per state
  - Format election outcome to be comparable to polls

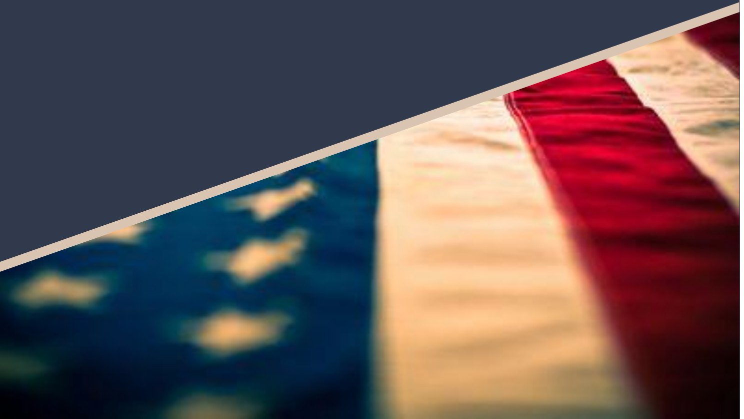
Cleaning Election Data:

- Corrected misspelled “New Hampshire” and “Pennsylvania” entries.
- Combined 259 write-in votes for Trump in 2016 in MD with regular Republican vote count.

Cleaning Polling Data:

- Remove polls without dates
- Give same poll end date format
- Make state naming consistent for polls
- Remove AK, and HI since 0 neighbors

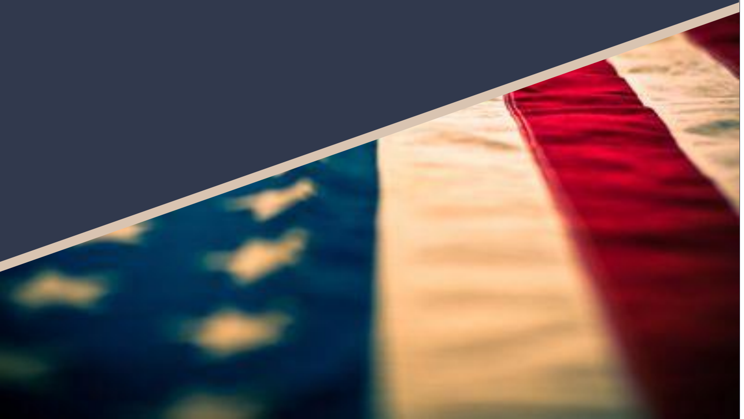
Cleaning the polling data



Examples of Cleaning the Polling Data:

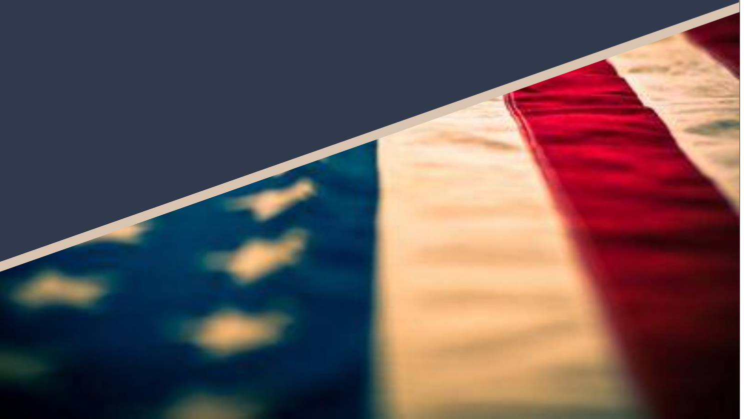
- 2020
  - Removed entries with no state identified.
  - Renamed “Maine CD-1”, “Maine CD-2” observations to “Maine”; renamed “Nebraska CD-1”, “Nebraska CD-2” observations to “Nebraska”.
- 2016
  - Dropped national-level polls (“U.S. polls”), renamed “Maine CD-1”, “Maine CD-2” observations to “Maine”, & renamed “Nebraska CD-1”, “Nebraska CD-2”, “Nebraska CD-3” observations to “Nebraska”.
- 2012
  - Format dates and choose an end date
  - Remove logs without full data
- All years
  - Dropped AK and HI since they have 0 neighbors.

Poll weighting methods



- Since there are often many polls per state, need to find a weighted average to turn this into one value.
  - Weighting based only on date of poll
  - (did not consider sample size, polling institution, manner of polling, polling demographics, etc since info not always available).
- Within each state and year, each poll observation's unnormalized weight u_it is calculated via one of the following 3 methods, where d is the number of days between the polling end date and the election date:
  - **Inverse exponential: u_it = e^(-d)**
  - **Inverse square root: u_it = 1/(sqrt(d))**
  - **Inverse logarithmic: u_it = 1/(log(d))**
- All u_it values are summed by state to obtain totals T_it. Then each observation (within each state) is normalized so that each state's set of weights w_it sums to 1 in each year: w_it = (u_it)/(T_it)
- We present results using the **inverse square root method** because its estimated models have the lowest DIC.

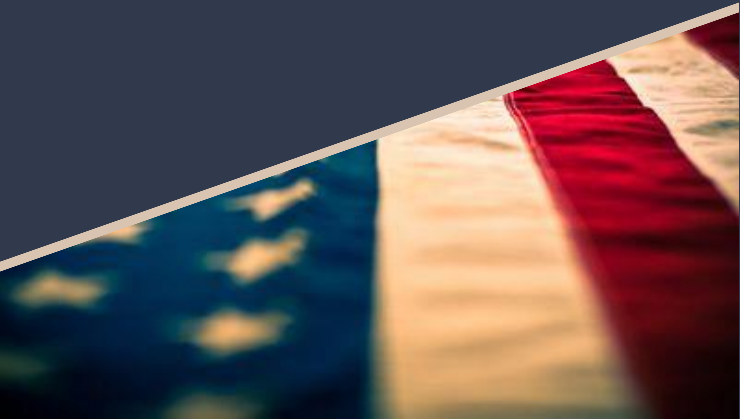
Data



- Final panel dataset contains 147 observations spanning 49 states (contiguous + DC) and 3 years of data (2012, 2016, and 2020).
  - 2020 election results were obtained from https://www.nbcnews.com/politics/2020-elections/president-results on 11/11/2020 (thanks Tyler!).
- State-level covariates used in (3) are: number of residents aged 18-30, number of male residents, number of Hispanic residents, number of White residents, number of Black residents, number of American Indians or Alaska natives, number of Asian residents, and number of native Hawaiians or Pacific Islanders.
  - Covariates data was obtained from the U.S. Census Bureau, Population Division.
- State neighbors data used for creating the state adjacency matrix was obtained from
  - https://writeonly.wordpress.com/2009/03/20/adjacency-list-of-states-of-the-united-states-us/

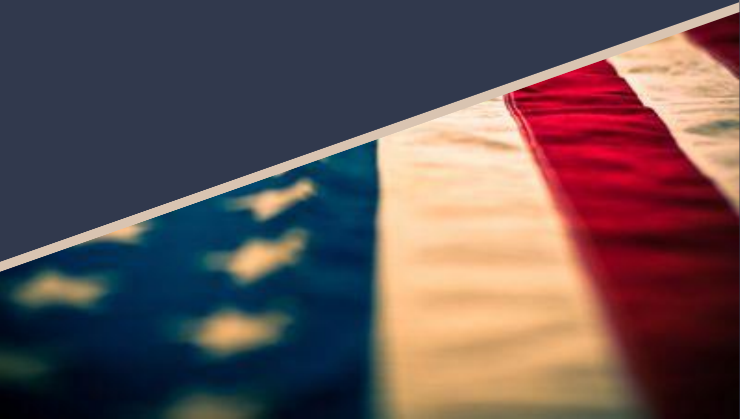
Testing for systematic polling bias

assuming estimated bias is constant across states and elections



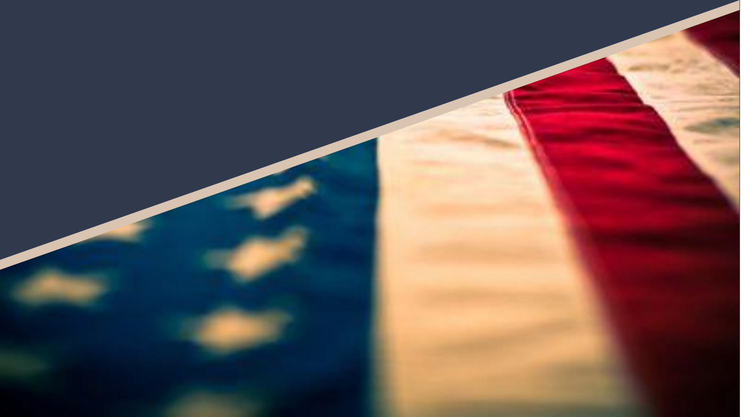
- A spatiotemporal conditionally autoregressive (STCAR) model with **no covariates** was fit to the data using the CARBayesST package in R.
  - The ST.CARar() function was used. The model estimated by this function is $y_{it} = \beta_0 + \xi_{it}$
    - A Gaussian prior on the dependent variable was chosen and the model assumes a Gaussian prior for the common intercept (mean) term.
    - 1,000,000 samples, a burn-in of 200,000, and a thinning parameter of 10 were specified.
  - The state adjacency matrix was created using border adjacency.
  - High spatial dependence and low temporal dependence was found.
  - Convergence was good for beta_0, the spatial and temporal dependence parameters, and CAR variance.
- Once the 720,000 MCMC samples of the mean term beta_0 were estimated, we ran a one-sample t-test on the samples to determine if the average bias was significantly different from 0.
  - The t-test produced a t-statistic of 18,934 and a near-zero p-value. Thus we **reject the null hypothesis** of beta_0's mean equaling 0 and conclude that **there is systematic polling bias** *(assuming that the estimated bias is constant across states and elections)*.

Testing polling bias' variability across space & time: Setup



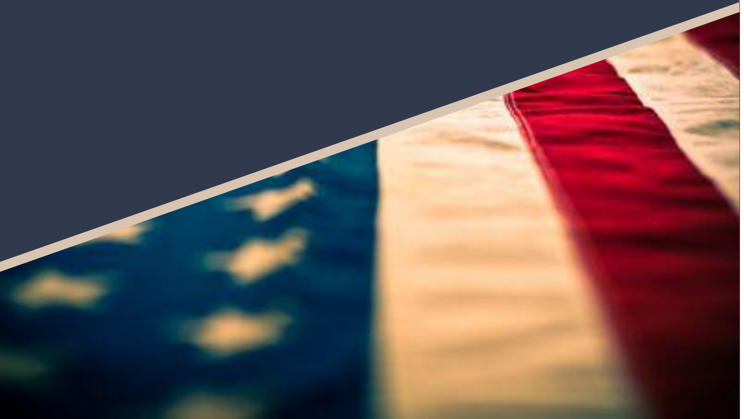
- A spatiotemporal conditionally autoregressive (STCAR) model with **8 covariates** was fit to the data using the CARBayesST package in R.
    - The ST.CARar() function was used. The model estimated by this function is
        - A G $y_{it} = \mathbf{X}_{it}^T\beta + \xi_{it}$ e dependent variable was chosen and the function assumes a Gaussian prior for the covariates.
        - 1,000,000 samples, a burn-in of 200,000, and a thinning parameter of 10 were specified.
    - The state adjacency matrix was created using border adjacency.
    - High spatial dependence and low spatial dependence was found.
    - Convergence was not good for covariates but good for spatial and temporal dependence parameters and variance parameters.
- After estimating all 720,000 MCMC beta samples, we constructed the estimated bias "Xb" by calculating the mean of each beta from its respective set of samples and multiplying the vector of mean beta values by the covariate matrix.
    - This generated an estimated bias for each state in each year (i.e. 3 total observations per state, one for each election year). All of the estimated biases were placed into a 147x1 vector ordered by year group (i.e. the first 49 entries were 2012 estimates, the next 49 were 2016 estimates, and the last 49 were 2020 estimates).

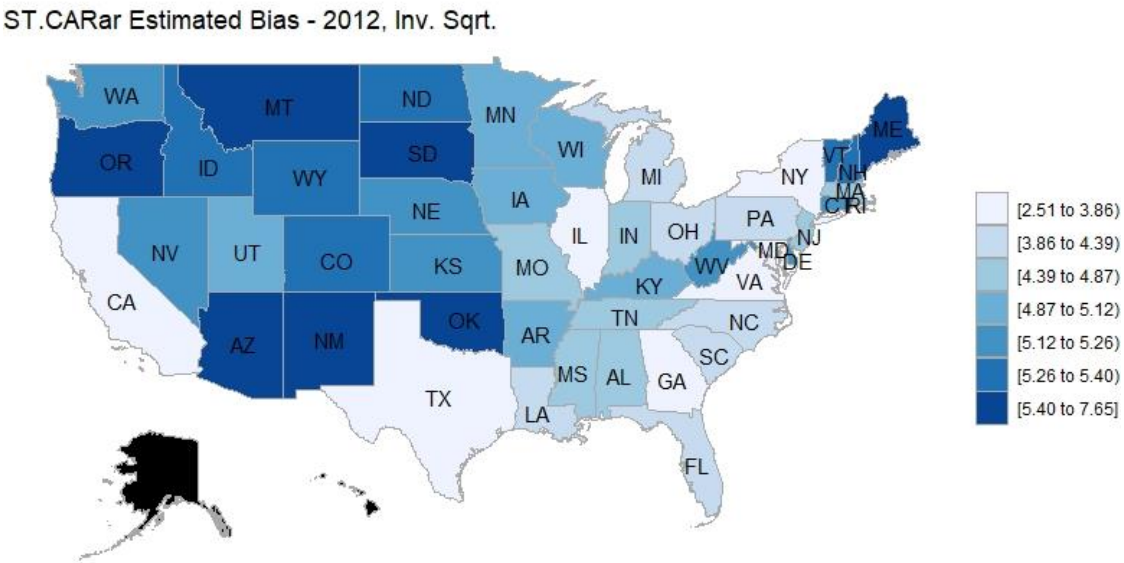
Testing polling bias' variability across space & time: Results



- To test whether the estimated bias varies by year, we added the corresponding year to each entry in the 147x1 vector of estimated biases and ran an ANOVA by year to compare the average bias in each year.
  - We obtain an F-stat of 0.013 and a p-value of 0.987, meaning that we cannot conclude that there are significant differences between the average estimated bias across years.
  - That is, **the average estimated bias does not appear to vary across years**.
- To test whether the estimated bias varies by state, we added the corresponding state to each entry in the 147x1 vector of estimated biases and ran an ANOVA by state to compare the average bias in each state.
  - We obtain an F-stat of 531.3 and a near-zero p-value, meaning that we can conclude that there are significant differences in the average bias between states across all 3 years of estimated biases.
  - That is, **the average estimated bias varies across states**.

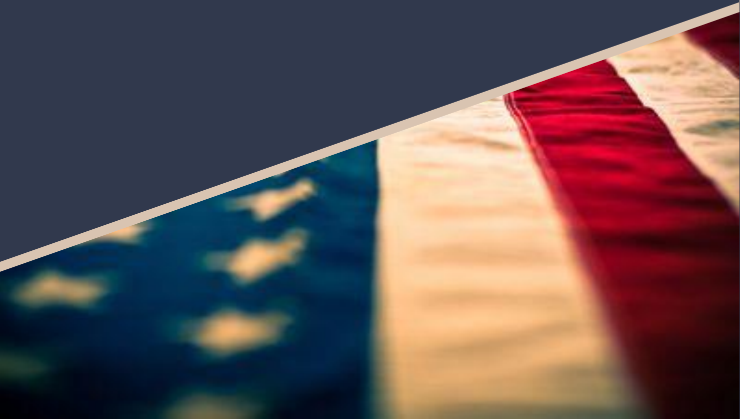
Plots of estimated biases



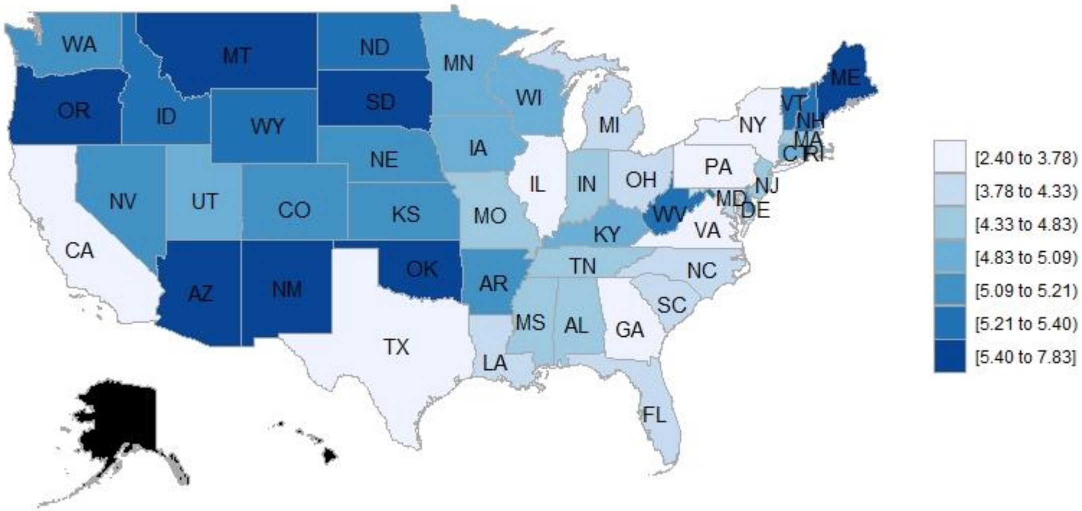


ST.CARar Estimated Bias - 2012, Inv. Sqrt.

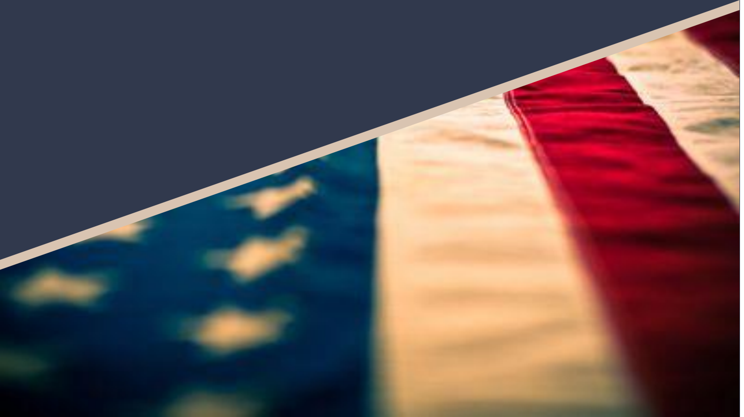
Plots of estimated biases

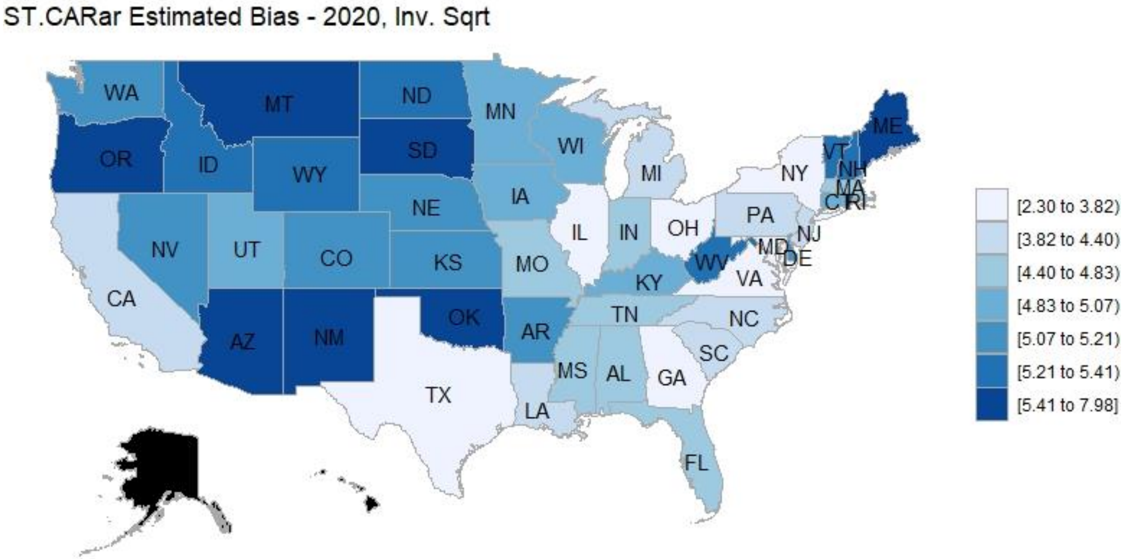


ST.CARar Estimated Bias - 2016, Inv. Sqrt



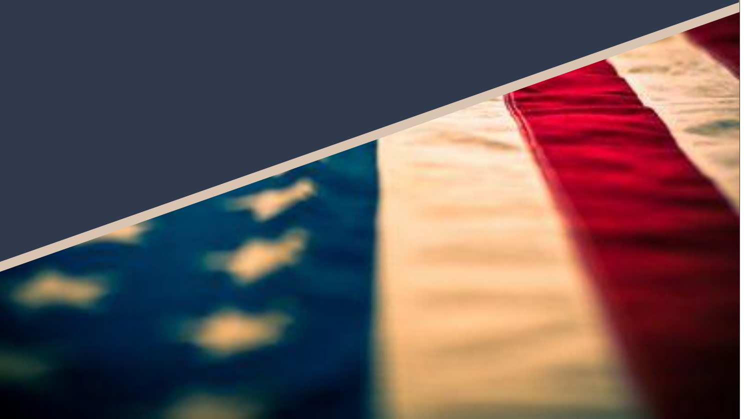
Plots of estimated biases





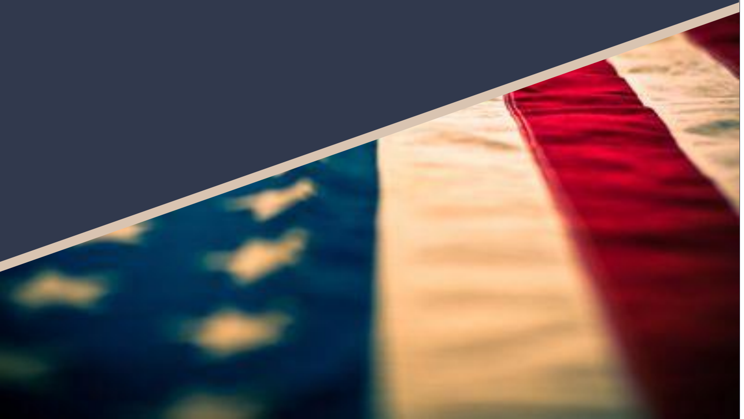
ST.CARar Estimated Bias - 2020, Inv. Sqrt

Sensitivity to polling average method



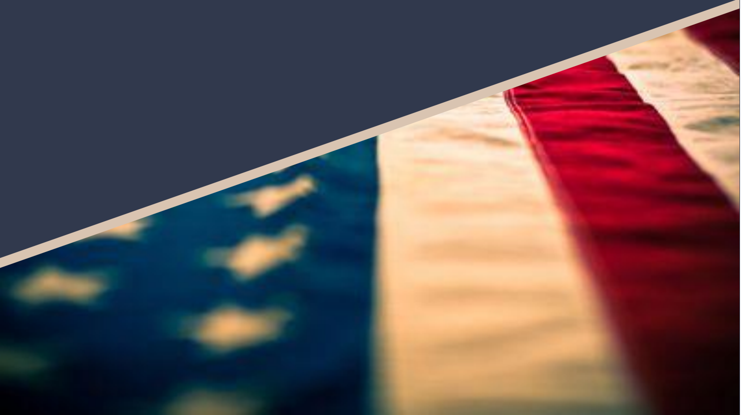
- Using the other two poll weighting methods, we
  - Still conclude that there is systematic polling bias in the case where we assume that the bias is constant across states and elections.
  - Still cannot conclude that there are significant differences in the average estimated bias across years.
  - Still can conclude that there are significant differences in the average estimated bias across states.

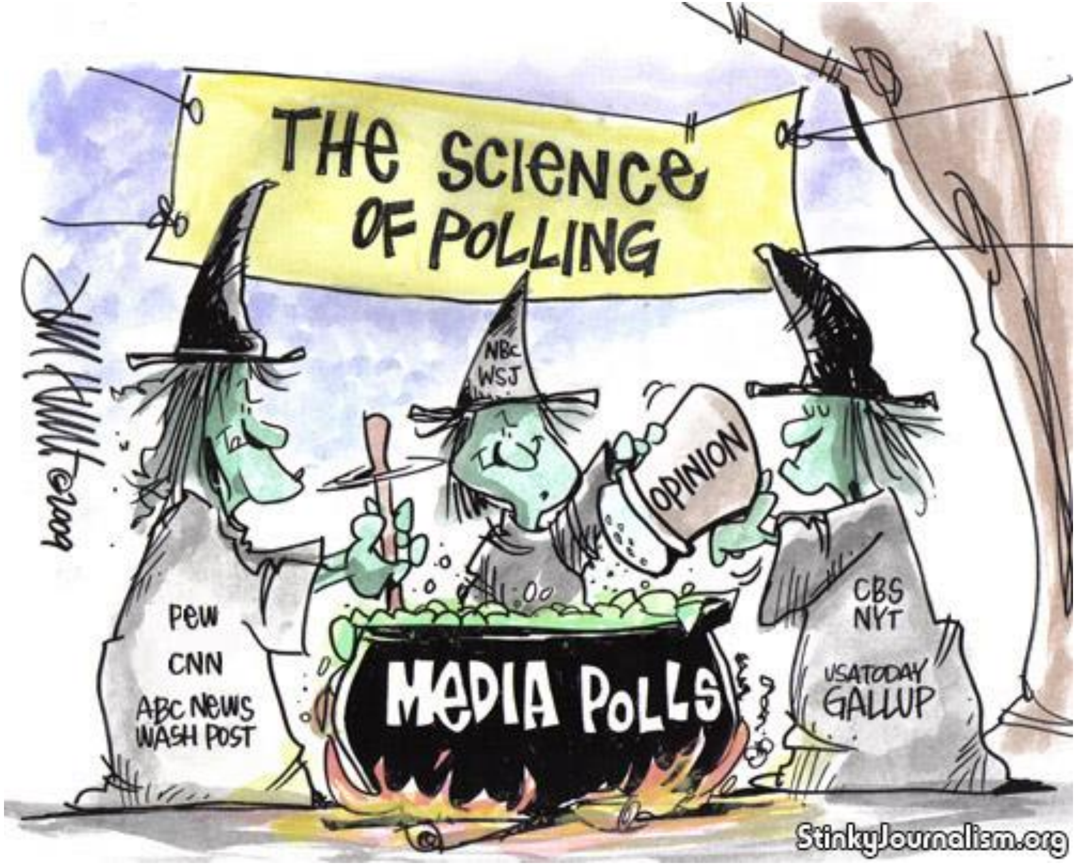
Conclusions



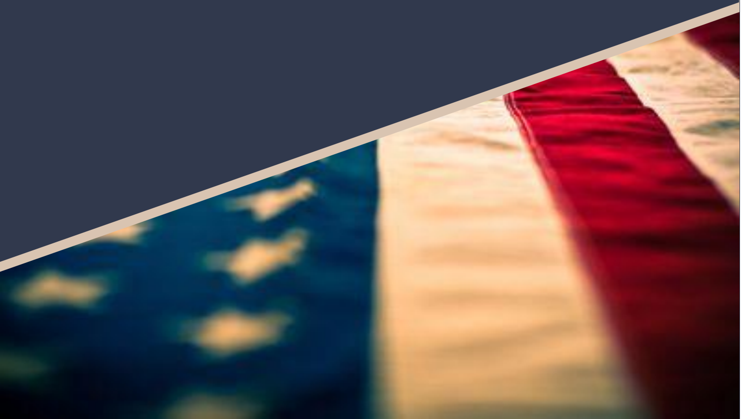
- **There is systematic polling bias in the US presidential elections** *(assuming that the estimated bias is constant across states and elections)*.
    - Spatiotemporal conditionally autoregressive (STCAR) model with no covariates (only knows border adjacency about states)
- When considering covariates, that average estimated bias **does not appear to vary across years**
    - Raises question as to whether polling method has changed in the last 8 years, and if it has, whether that was effective.
- However, the average estimated bias **does appear to vary between states.**
    - Polling in different states carry different amounts of bias
    - *covariates:*
        - *number of residents aged 18-30*
        - *number of male residents*
        - *number of Hispanic residents*
        - *number of White residents*
        - *number of Black residents*
        - *number of American Indians or Alaska natives*
        - *number of Asian residents*
        - *number of native Hawaiians or Pacific Islanders*

Any Questions?



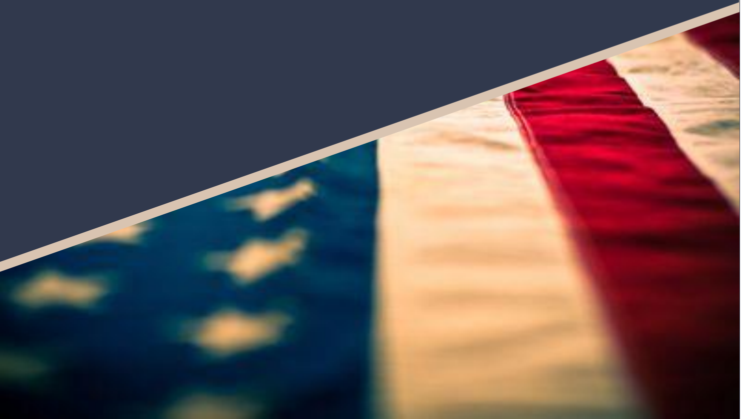


Resources

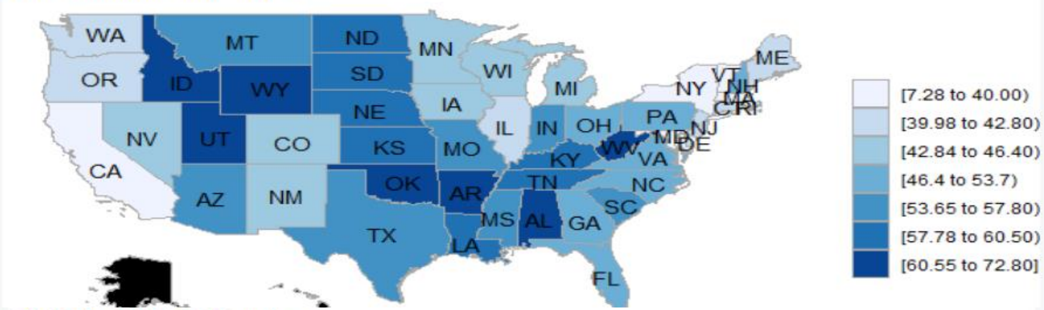


- Ideas for Weighted Averages based on date:
    - https://stackoverflow.com/questions/7041867/how-to-decide-on-weights
- How to do a weighted average:
    - https://www.indeed.com/career-advice/career-development/how-to-calculate-weighted-average
- CARBayesST documentation: https://cran.r-project.org/web/packages/CARBayesST/vignettes/CARBayesST.pdf

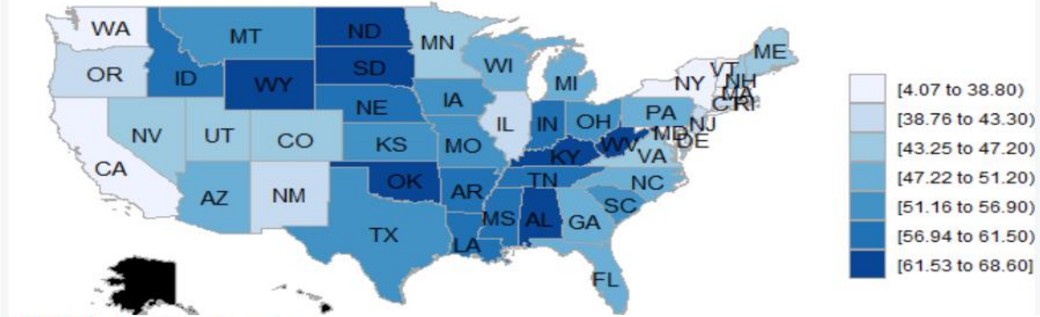
Actual Election Results (GOP %)



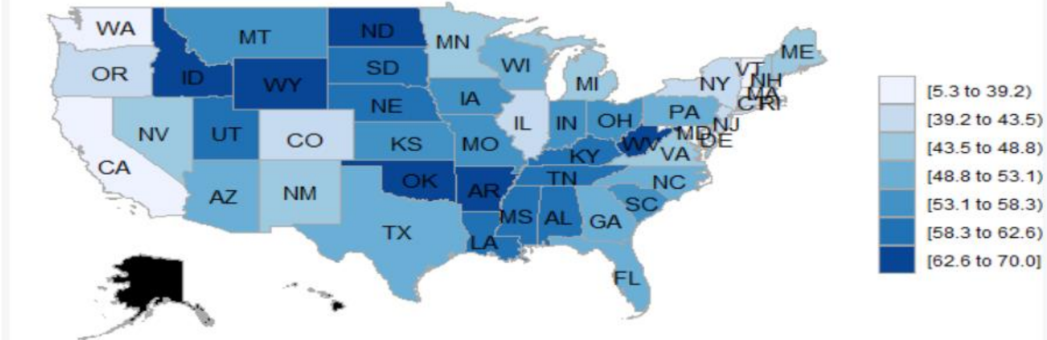
2012 Actual GOP votes



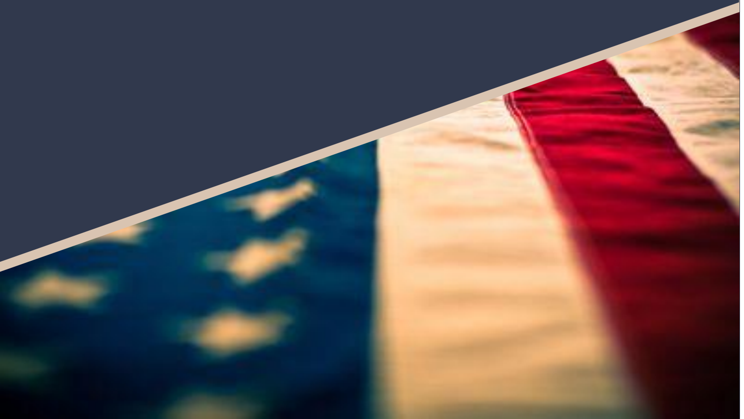
2016 Actual GOP votes



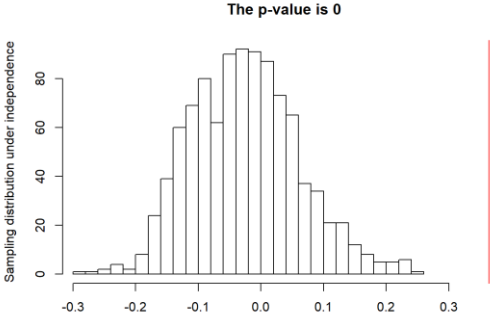
2020 Actual GOP votes



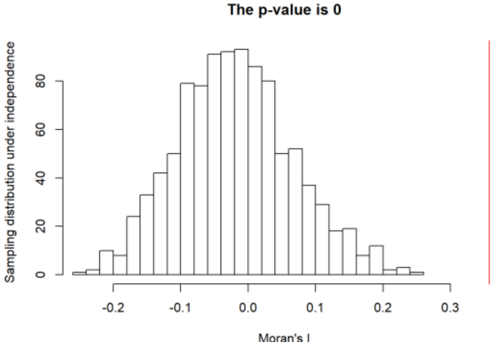
Testing whether bias in US election (assuming constant) Moran’s I



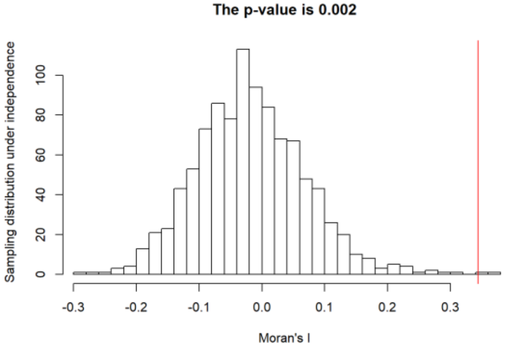
2012:



2016:



2020:


The p-value is 0.002
Sampling distribution under independence
Moran's I