NC STATE UNIVERSITY

Spatiotemporal Analysis of Polling Bias in the 2012, 2016, and 2020 US Presidential Elections

Yang Bai, Enrique Pena, and Shitao Fan

ST533 – Applied Spatial Statistics

Dr. Brian Reich

Introduction

Bias is defined as a disproportionate weight in favor of or against an idea or thing

Systematic polling bias has been evident in past US elections

It is of interest to study the polling bias in favor of GOP support in the past three elections

Objectives

1. Devise a method to calculate polling averages and forecast the election results in each state and each year

2. Test whether systematic polling bias exists

3. Test whether the polling bias varies by state and/or by election

Data Sources

- 2012 Polling Data obtained from:

https://en.wikipedia.org/wiki/Statewide_opinion_polling_for_the_2012_United_States_presidential_election

- 2016 Polling Data obtained from:

https://www.kaggle.com/fivethirtyeight/2016-election-polls?select=presidential_polls.csv

- 2020 Polling Data obtained from:

https://projects.fivethirtyeight.com/polls/president-general/

- Demographics obtained from:

https://www.census.gov/data/datasets/time-series/demo/popest/2010s-state-detail.html

Big Scope Methods and Data Tidying

1. Remove all data before September 1st in each polling dataset
2. Average polls with same Poll ID
3. Create "time weights"
4. Average the polls weights within each state
5. Calculate the polling bias
6. Run a spatio-temporal model in R

Objective #1: Devise a method to calculate polling averages and forecast the election results in each state and each year

Methods for Objective #1

The response is going to be the polling bias, which can be calculated as:

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

Where,

Y_{it} is the GOP percentage of actual votes for state i in year t

And,

$$X_{it} = \sum W_{itj} P_{jt}$$

Where,

W_{itj} is the temporal weight

P_{jt} is the GOP percent support seen in poll j and year t

Objective #1: Devise a method to calculate polling averages and forecast the election results in each state and each year

Methods for Objective #1 (cont'd)

- Define the temporal score, S_{tij}
  - Choose four weight models and test them
    1. $S_{tij} \in [1,2,3,4,5]$
    2. $S_{tij} \in [1,2,3,4,10]$
    3. $S_{tij} = e^{-0.1 time_{tij}}$
    4. $S_{tij} = 0.95^{time_{tij}}$

Where,

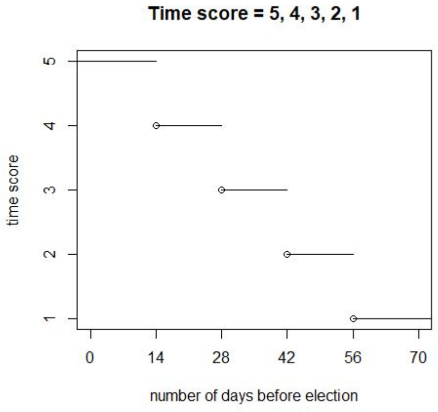
$$\text{time} = election\ date\ - poll\ date$$

$$W_{tij} = \frac{S_{tij}}{\sum_{j=1}^{N_{ti}} S_{tij}}$$

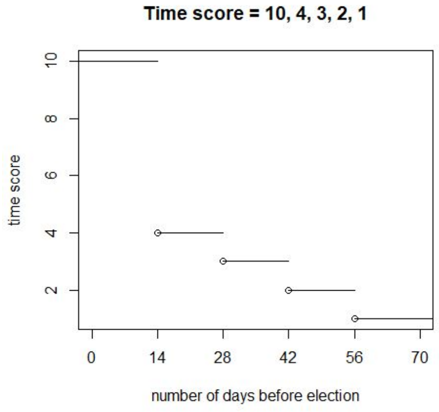
Where,

N_{ti} is the number of polls taken in state i in election year t

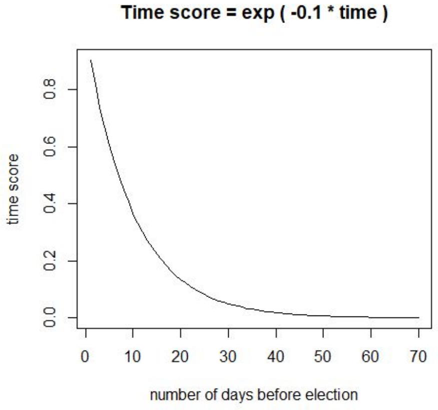
Case 1



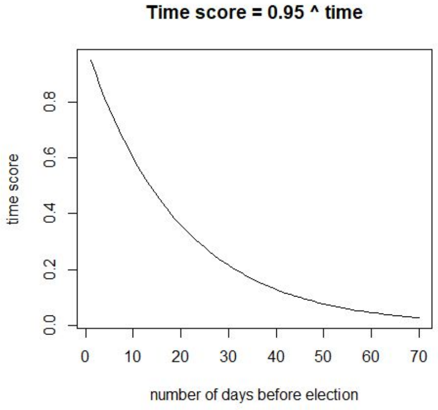
Case 2



Case 3



Case 4

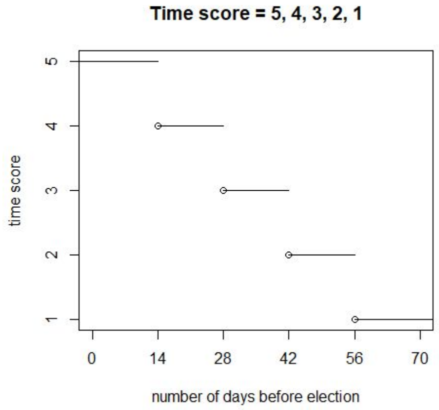


CAR Model

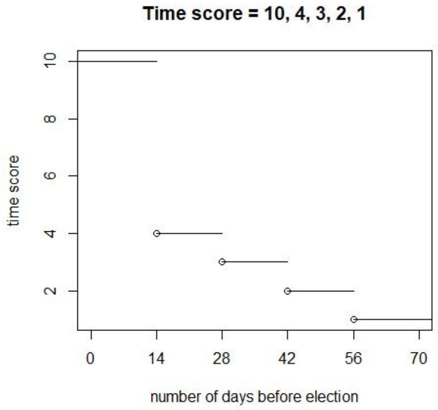
- S.CARleroux() is a conditionally autoregressive model
- Use the S.CARleroux() function from the CARBayes package to determine what weight is best in terms of DIC and nu2
- Small DIC and nu2 are preferred

Weight test results

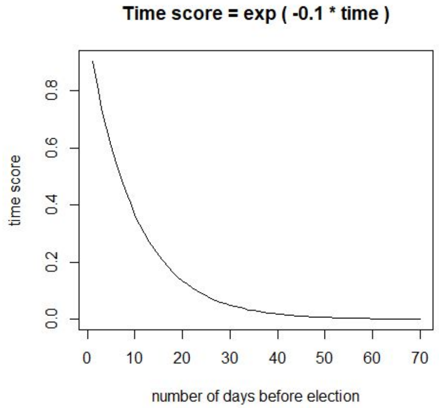
Case 1



Case 2



Case 3



Case 4

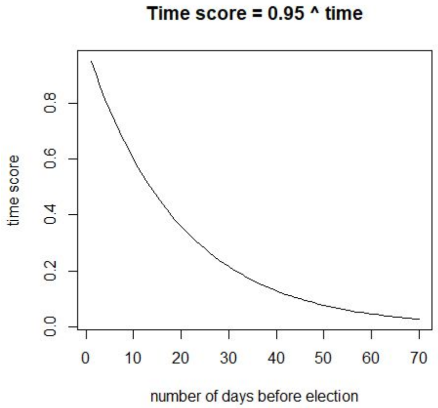


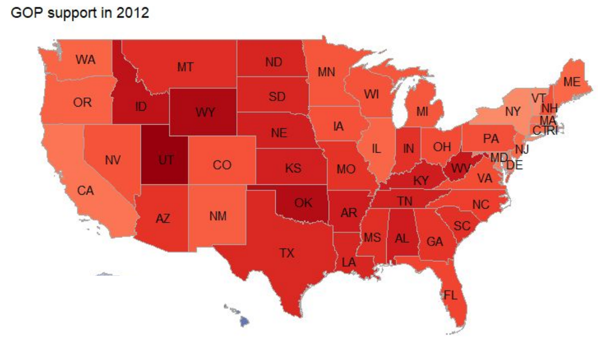
Table 1: Criteria to choose weight using CAR model

		case 1	case 2	case 3	case 4
2012	**nu^2**	5.86	6.01	6.73	6.25
	DIC	193.04	193.50	198.20	195.27
2016	**nu^2**	15.59	13.21	12.30	13.79
	DIC	278.47	269.41	266.17	271.69
2020	**nu^2**	5.10	4.99	4.63	4.85
	DIC	223.42	222.34	218.24	220.91

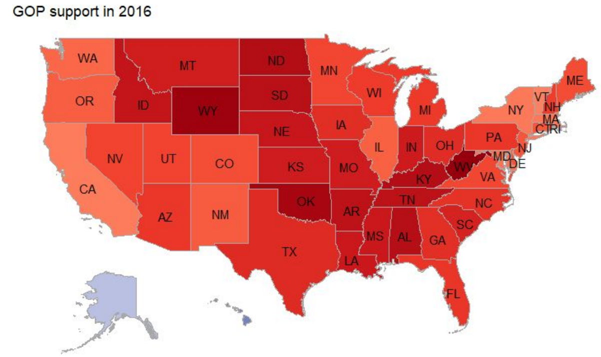
Real GOP Support Results vs Polling Average Results

Real results

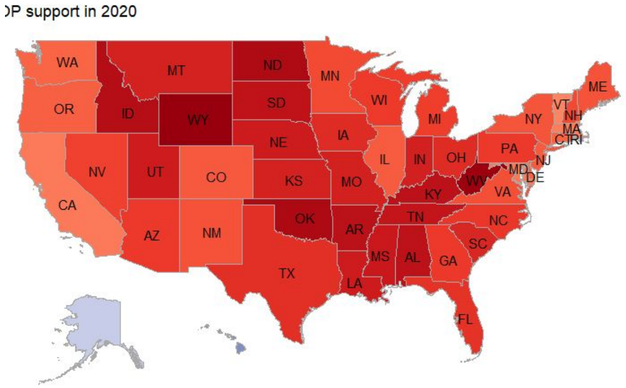
2012



2016

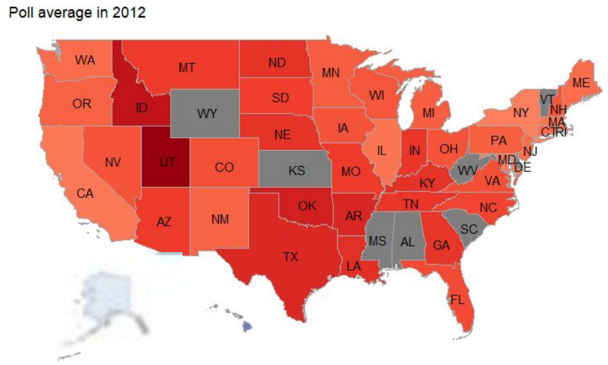


2020

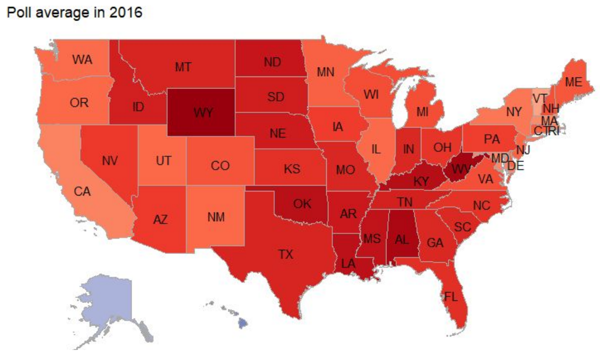


Polling average results

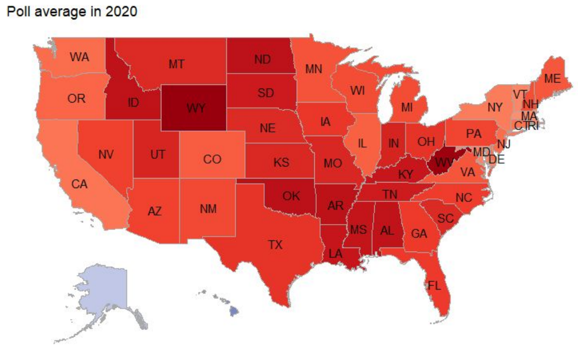
2012



2016



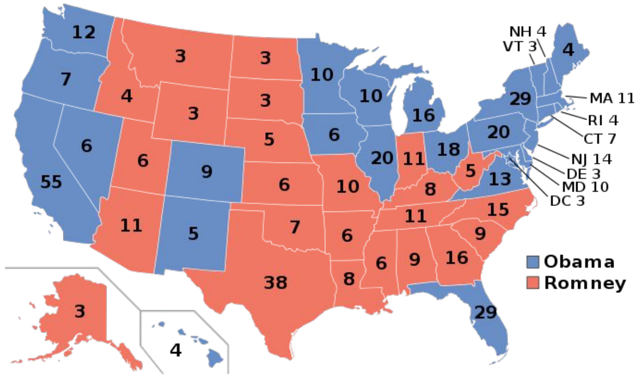
2020



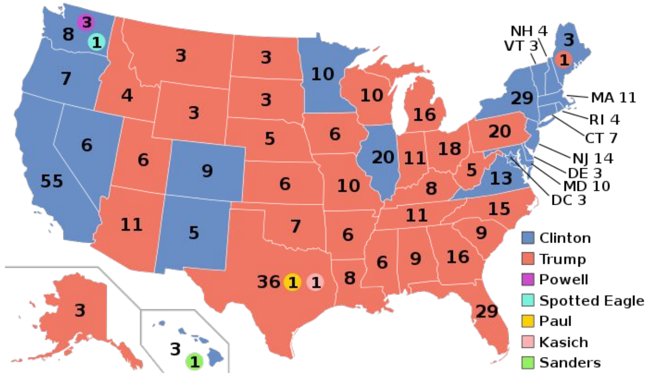
Real Election Results vs Predicted Polling Average Results

Real Results

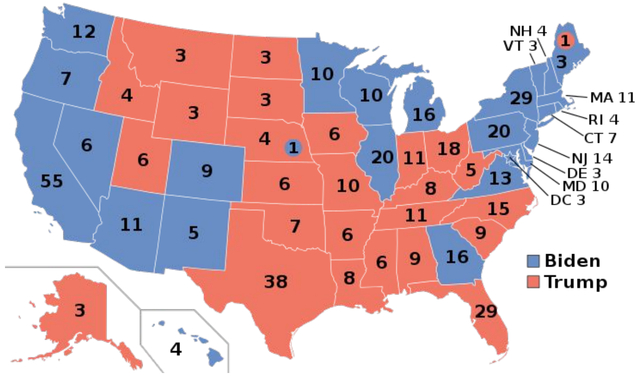
2012



2016

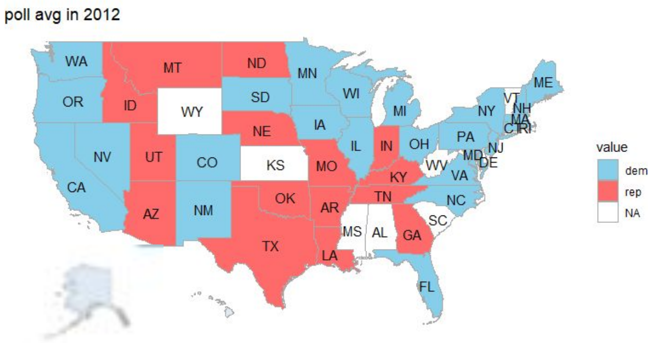


2020

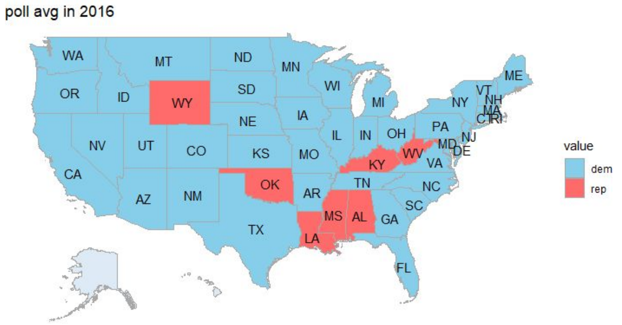


Polling Average Results

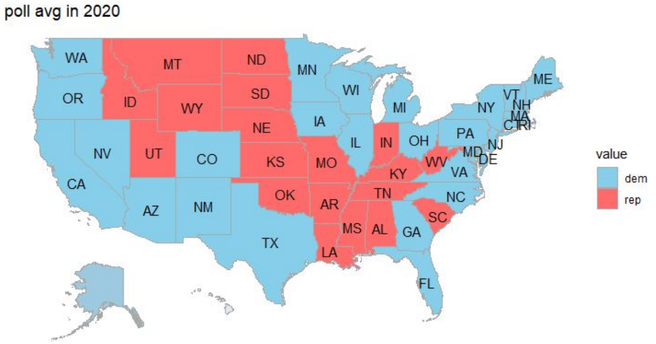
2012



2016



2020



Objective 2: Test whether systematic polling bias exists

Methods for objective 2

- Bias was assumed to be constant across states and years

- A linear regression was used to test whether systematic polling bias existed

$$B_{it} = \mu + \varepsilon$$

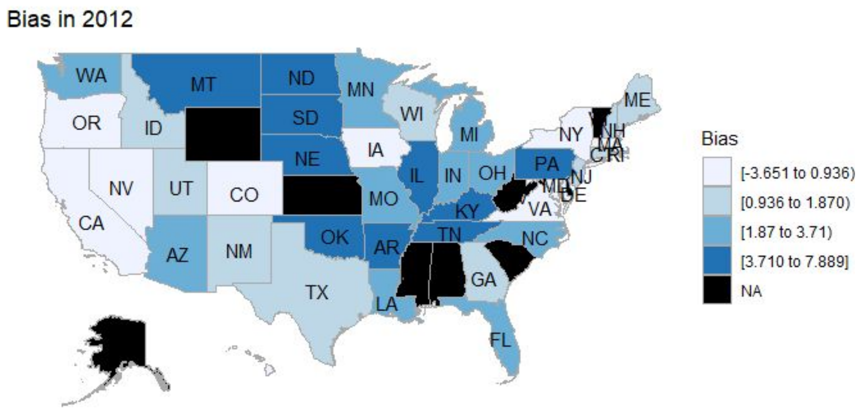
- Conduct a hypothesis test where:
  - H_o: $\mu = 0$
  - H_a: $\mu \neq 0$

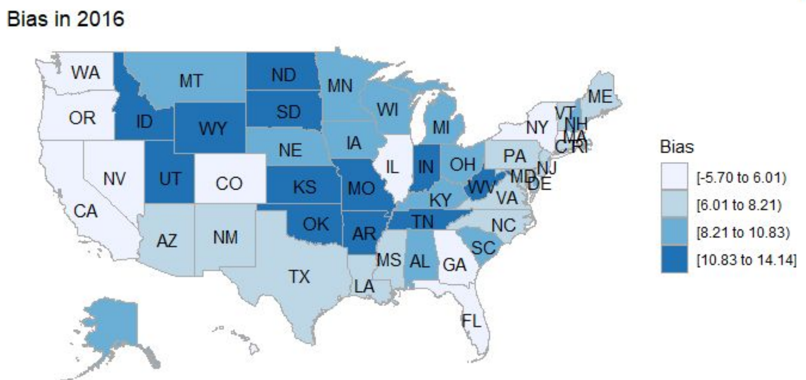
- Test stastistic was calculated as: $t = \frac{\mu}{\frac{s}{\sqrt{n}}}$

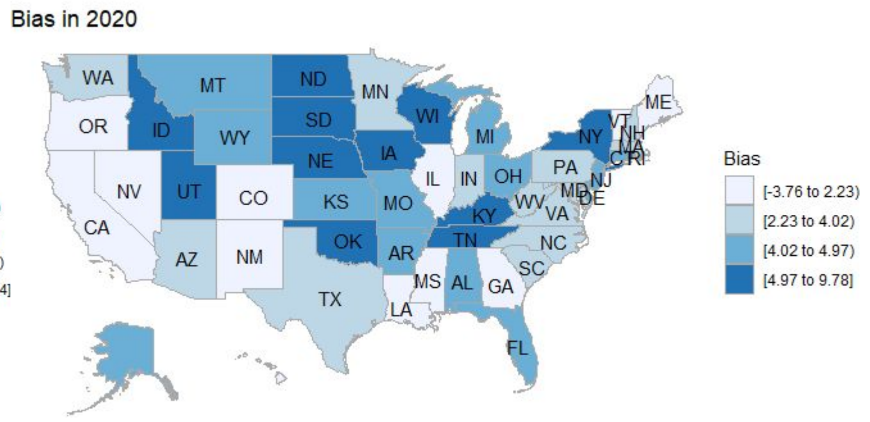
Test Results for Existence of Systematic Bias

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   4.8043     0.3228   14.88   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Bias Maps







Objective 3: Test whether the polling bias varies by state and/or by election

Methods for objective 3

- An alternative to the model proposed by Knorr-Held (2000) was used
- Random effects are decomposed into three components:
  - Spatial component
  - Temporal component

ST.CARanova() from the **CARBayesST** package

- Fit a spatio-temporal model with and without covariates

```
model <- ST.CARanova(bias~., family="gaussian", W=ADJ,
burnin=10000, n.sample=50000,thin=10,data = newdata)
```

```
model <- ST.CARanova(bias~1, family="gaussian", W=ADJ,
burnin=10000, n.sample=50000,thin=10,data = newdata)
```

Covariates	
Longitude	
Latitude	
Longitude_Latitude	
Age Percent 65 and Older	
Education: Bachelor's Degree or Higher	
Nonfarm Employment	
Black Population Percentage	
Median Household Income	
Population Percent Change	

Spatio-temporal generalized linear mixed model

$$
\begin{aligned}
Y_{kt} &\sim \mathrm{N}(\mu_{kt}, \nu^2) \text{ and } \mu_{kt} = \mathbf{x}_{kt}^{\top}\boldsymbol{\beta} + O_{kt} + \psi_{kt}. \\
\boldsymbol{\beta} &\sim \mathrm{N}(\boldsymbol{\mu}_{\beta}, \boldsymbol{\Sigma}_{\beta}) \\
\psi_{kt} &= \phi_k + \delta_t + \gamma_{kt}, \\
\phi_k|\boldsymbol{\phi}_{-k}, \mathbf{W} &\sim \mathrm{N}\left(\frac{\rho_S \sum_{j=1}^{K} w_{kj}\phi_j}{\rho_S \sum_{j=1}^{K} w_{kj} + 1 - \rho_S}, \frac{\tau_S^2}{\rho_S \sum_{j=1}^{K} w_{kj} + 1 - \rho_S}\right), \\
\delta_t|\boldsymbol{\delta}_{-t}, \mathbf{D} &\sim \mathrm{N}\left(\frac{\rho_T \sum_{j=1}^{N} d_{tj}\delta_j}{\rho_T \sum_{j=1}^{N} d_{tj} + 1 - \rho_T}, \frac{\tau_T^2}{\rho_T \sum_{j=1}^{N} d_{tj} + 1 - \rho_T}\right), \\
\gamma_{kt} &\sim \mathrm{N}(0, \tau_I^2), \\
\tau_S^2, \tau_T^2, \tau_I^2 &\sim \text{Inverse-Gamma}(a, b), \\
\rho_S, \rho_T &\sim \text{Uniform}(0, 1).
\end{aligned}
$$

$$
\mathbf{D} = (d_{tj}), \text{ where } d_{tj} = 1 \text{ if } |j - t| = 1 \text{ and } d_{tj} = 0 \text{ otherwise.}
$$

- Spatio-temporal generalized linear mixed model to areal unit data, where the response variable can be binomial, Gaussian or Poisson (Lee et al. 2018)

Model comparison and diagnostic

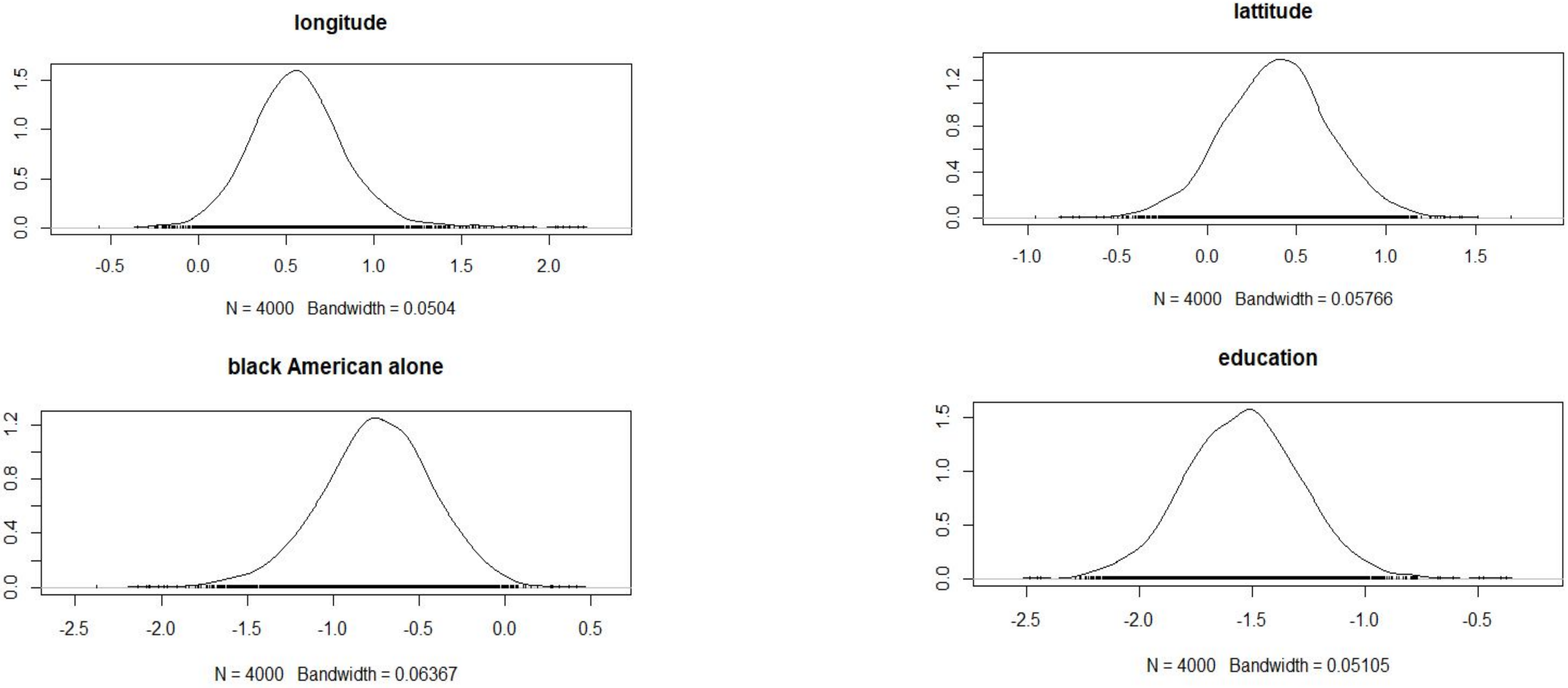
model\criteria	DIC	WAIC
without covariates	632.78	638.42
with covariates	638.48	642.20

The spatio-temporal model without covariates had lower DIC results. We pick this model to find the spatial and temporal effect. However, we are still interested in the model with covariates because we want to explore the covariate effects on bias.

Model diagnostic for spatio-temporal model without covariates

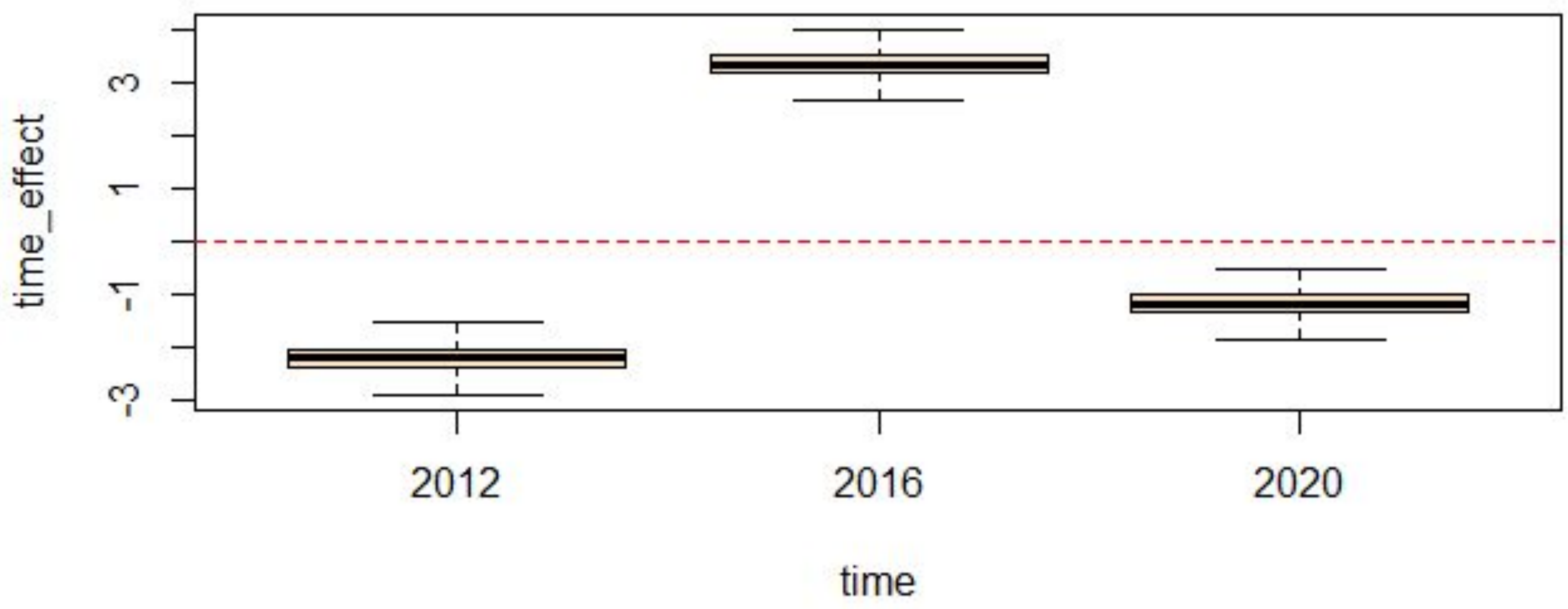
Parameters	2.5%quantile	median	97.5%quantile	effective sample size(>1000)	Geweke.diag (abs<2)
(Intercept)	4.43	4.78	5.13	4000	0.2
tau2.S	2.98	6.28	12.30	3718.7	-0.4
tau2.T	1.69	5.90	32.55	4000	0.1
nu2	3.21	4.24	5.81	4544.2	0.5
rho.S	0.07	0.44	0.88	3119.1	-0.8
rho.T	0.006	0.19	0.83	3693.7	0.2

Covariates from spatio-temporal model with covariates



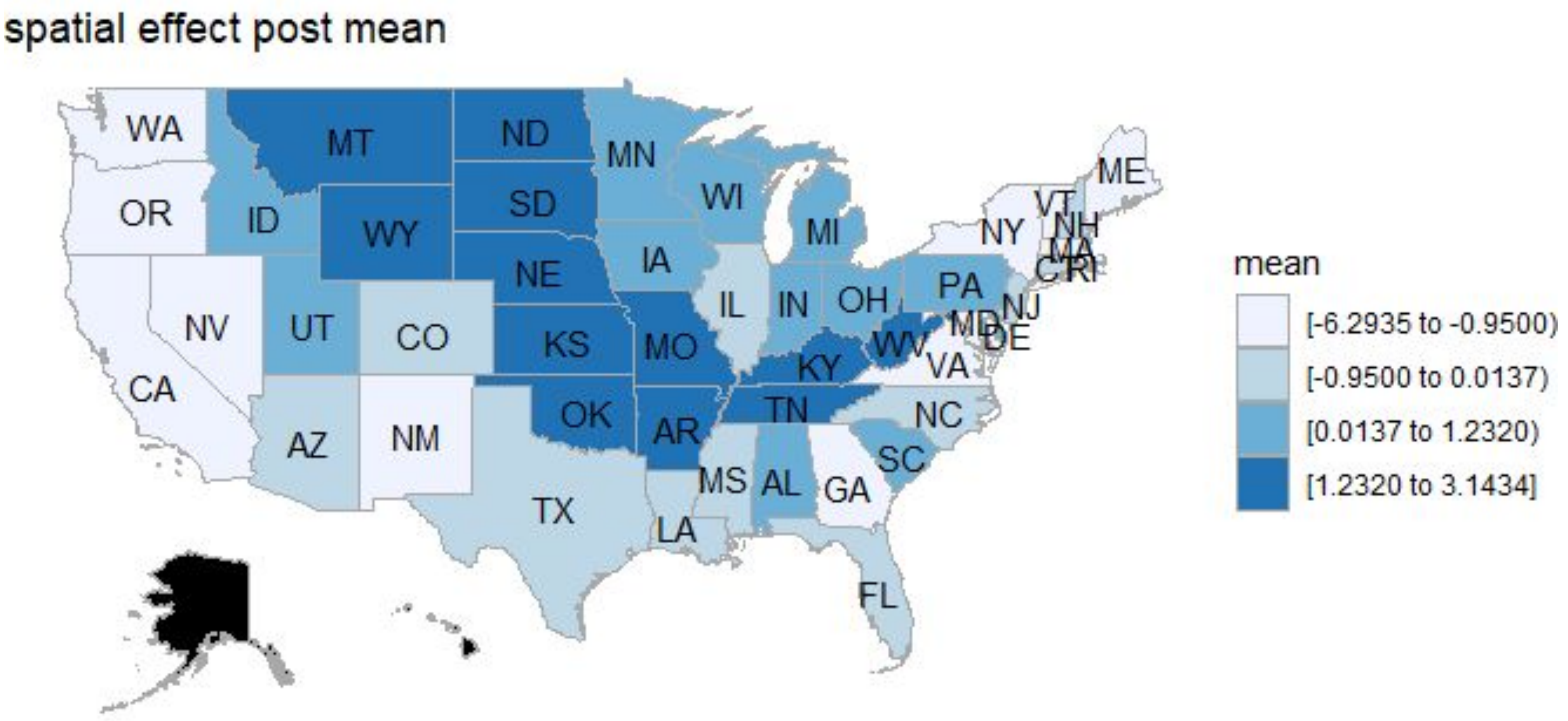
1. Education level and black American population are significant in modeling the GOP support bias.
2. Results suggest that states that have higher education levels will have less systematic polling bias
3. Results also suggest that states that have higher black american population levels will have less systematic polling bias

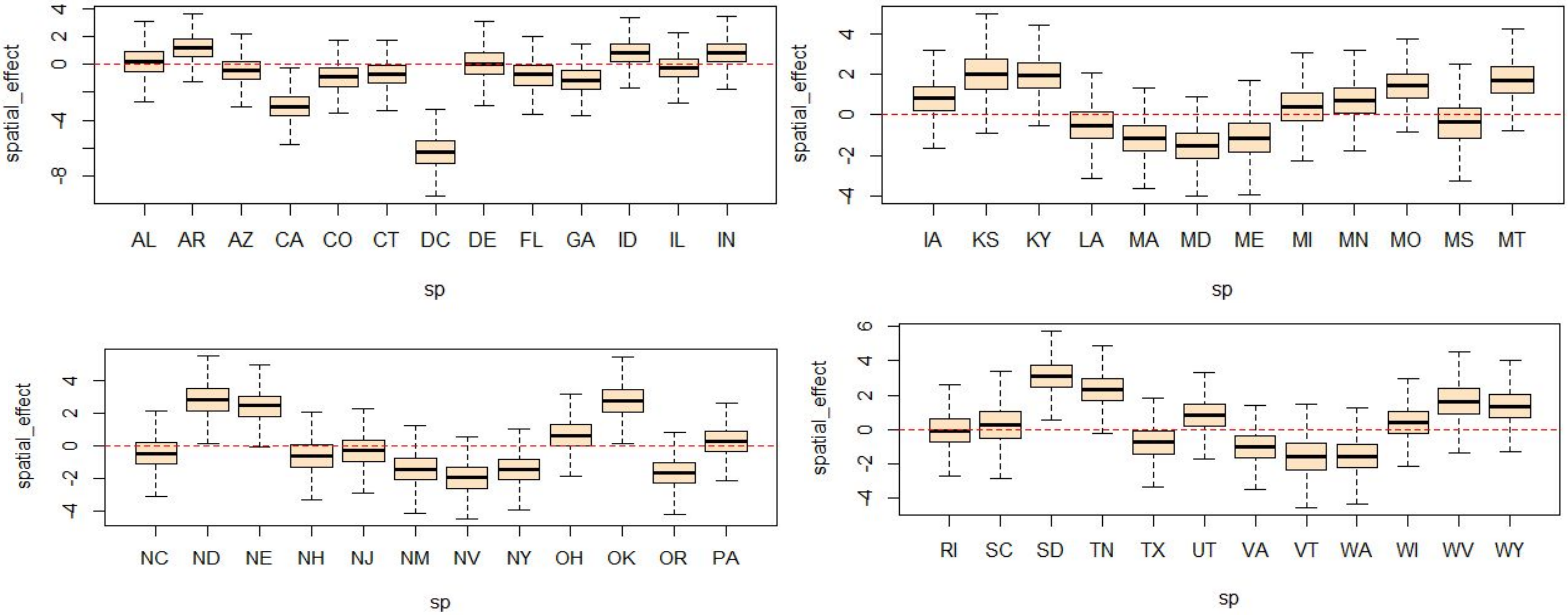
Temporal Effect from spatio-temporal model without covariates



1. Time has a strong effect for the GOP support bias.
2. 2016 is very different from the other years.

Spatial Effect from spatio-temporal model without covariates





1. Spatial effect exits in the sup GOP bias.
2. Blue states: CA, DC
3. Red states: ND, NE, MT, OK, SD, TN

Discussion and Conclusion

1. Easy and intuitive method was used to define the temporal scores
2. Bias was found to be dependent to the poll location and polling year
3. Largest polling bias was found in Midwestern states
4. Spatiotemporal model without predictors had a lower DIC than the model with predictors
5. Education level and black American population were significant predictors of bias
6. Type of voter (lv, rv, a) could be added to the model to further refine poll predictions

References

- Lee D, Rushworth A, Napier G (2018). “Spatio-Temporal Areal Unit Modeling in R with Conditional Autoregressive Priors Using the CARBayesST Package.” _Journal of Statistical Software_, *84*(9), 1-39. doi: 10.18637/jss.v084.i09 (URL: https://doi.org/10.18637/jss.v084.i09).
- Knorr-Held L (2000). “Bayesian Modelling of Inseparable Space-Time Variation in Disease Risk.” Statistics in Medicine, 19(17–18), 2555–2567. doi:10.1002/1097-0258(20000915/ 30)19:17/183.0.co;2-\%23.

Appendix 1: Table for choosing temporal score

		case 1	case 2	case 3	case 4
2012	mean	2.304	2.450	2.944	2.620
	res sd	2.458	2.474	2.622	2.528
2016	mean	8.745	8.212	8.062	8.477
	res sd	4.026	3.676	3.555	3.758
2020	mean	3.660	3.589	3.425	3.539
	res sd	2.298	2.272	2.180	2.241

We use linear regression with intercept only. We prefer a small mean and small residual standard deviation. Case 2 also gives a good results.

Appendix 2: Model diagnostic for spatio-temporal model without covariates

	Median	2.5%	97.5%	n.sample	% accept	n.effective	Geweke.diag
(Intercept)	4.7918	4.4185	5.1707	4000	100.0	4000.0	0.8
centroid.lon	0.5696	0.0596	1.3121	4000	100.0	715.8	-0.3
centroid.lat	0.3890	-0.2271	0.9945	4000	100.0	4000.0	-1.6
Education.Bachelor.s.Degree.or.Higher	-1.5377	-2.0300	-1.0083	4000	100.0	1185.2	0.1
Ethnicities.Black.Alone	-0.7551	-1.4901	-0.1397	4000	100.0	628.0	0.0
tau2.S	0.0163	0.0023	3.7385	4000	100.0	67.8	0.0
tau2.T	5.5745	1.6197	30.8576	4000	100.0	3756.5	0.1
nu2	5.2259	3.5688	6.9236	4000	100.0	123.8	0.0
rho.S	0.4442	0.0225	0.9498	4000	46.6	471.0	0.4
rho.T	0.1944	0.0061	0.8185	4000	58.7	4000.0	-1.0

The convergence of tau2.S, nu2 and rho.S are not good. The covariates cancel the spatial effect.