Exploring the Bias in Polls

Md Mehedi Hasnat & Richard Watson

The Setup

What is Bias?

$$B_{ij} = E\left[Y_{ij} - \sum_{k=0}^{k(ij)} W_{ijk} P_{ijk}\right]$$

- Bias is the expected difference between the weighted polls and the actual vote percentages
- **Why the expectation?** Bias is systematic error. Usually we assume the error to have a mean of 0. If not, we say the model is biased and the mean of the error is the Bias

- Y is the percentage of GOP votes
- W are the weights for the polls
- P are the polled percentage of those in favor of GOP
- Indices: i = state, j = year, k = polls, k(ij) = number of polls for ith state and jth year

How to weight the polls? $W_{ijk}(p) = pR_{ijk} + (1-p)S_{ijk}$

- We chose to weight the polls by recency and sample size
- Recency is defined by how many days before the election the poll ended
- p controls how much we care about recency over sample size and is between 0 and 1

$$R_{ijk} = \frac{1}{D_{ij+}^{-1}} \sum_{k=0}^{k(ij)} D_{ijk}^{-1}$$

$$S_{ijk} = \frac{1}{N_{ij+}} \sum_{k=0}^{k(ij)} N_{ijk}$$

Note: D^{-1} are the inverted days until election for each poll. N is the sample size.

Review of temporal AR and areal CAR model

- Both models attempt to describe the correlation of the random effects across time/space
- φ is the correlation for time
- ρ is the correlation for space
- **M** is a diagonal matrix where the diagonal is the number of neighbors for region i
- **W** is the weight matrix

Temporal AR

$$
\begin{aligned}
Z_1 &\sim \text{Normal}(0, \sigma^2) \\
Z_t | Z_{t-1} &\sim \text{Normal}\{\varphi Z_{t-1}, (1-\varphi^2)\sigma^2\} \\
&\text{Var}(Z_t) = \sigma^2 \text{ for all } t
\end{aligned}
$$

Areal CAR

Let Z_{-i} be the collection of the $n-1$ other spatial terms

Further, define $\bar{Z}_i$ as the mean of Z_j over the m_i regions that neighbor region i

$$Z_i | Z_{-i} \sim \text{Normal}(\rho \bar{Z}_i, \sigma^2 / m_i)$$

Leroux parametrization for covariance of joint

$$\Sigma = \sigma^2 [(1-\rho) I_n + \rho(\mathbf{M} - \mathbf{W})]^{-1}$$

Spatiotemporal CAR AR model

- The first two lines show the autoregressive time element described in the AR model
- **Q** is the spatial covariance structure described in the CAR model
- Implementable using CARBayesST package

$$Z_{s,t}|Z_{s,t-1} \sim N(\phi Z_{s,t-1}, \sigma^2 \boldsymbol{Q}(\boldsymbol{W}, \rho)^{-1})$$
$$Z_{s,1} \sim N(0, \sigma^2 \boldsymbol{Q}(\boldsymbol{W}, \rho)^{-1})$$
$$\sigma^2 \sim \text{Inverse} - \text{Gamma}(a, b)$$
$$\rho, \phi \sim \text{Uniform}(0, 1)$$
$$\boldsymbol{Q}(\boldsymbol{W}, \rho) = \sigma^2 \left[(1-\rho) I_n + \rho(\boldsymbol{M} - \boldsymbol{W})\right]$$

Questions and Answers

Is there systematic bias?

Parameters of the model using only the constant term -->

The mean intercept suggests that there is bias between the weighted polls and the actual votes. The bias on average is positive (%vote > weighted polls)

Spatial dependence (rho.S) is close to 1. There are spatial dependence in the bias.

Also, there are moderate temporal dependence in the bias (rho.T > 0.5)

	Mean	Median	2.50%	97.50%
(Intercept)	0.036937	0.0371	0.0324	0.0417
tau2	0.001447	0.0014	0.0008	0.0021
nu2	0.00079	0.0007	0.0005	0.001
rho.S	0.81412	0.8699	0.5535	0.9782
rho.T	0.507163	0.5425	0.1272	0.8769

Weights = 0.5*time + 0.5*sample

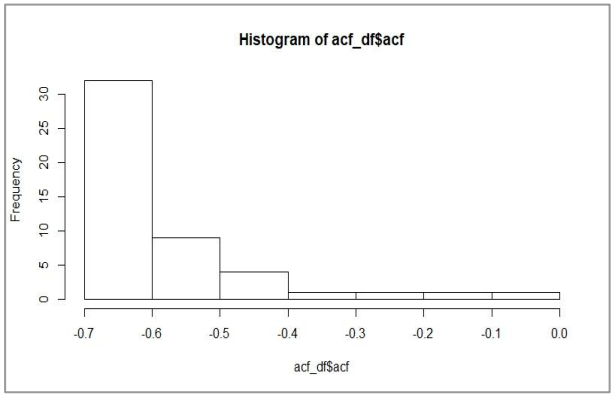
Is the bias spatially autocorrelated?

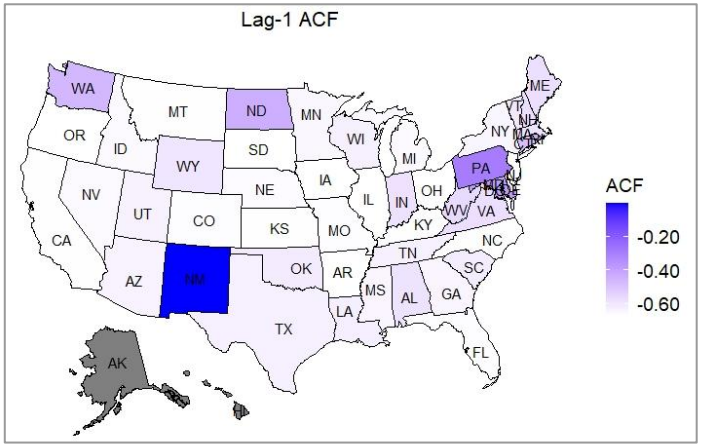
- Moran's I is centered around 0 (no autocorrelation) and goes from -1 (negative autocorrelation) to 1 (positive autocorrelation)
- Geary's C is centered around 1 (no autocorrelation). Values greater than 1 are negatively correlated and less than 1 are positively correlated
- None of our tests were significant, but we found that the values point toward negative correlation.

	Moran's I	Geary's C
Bias 2012	-0.179	1.147
Bias 2016	-0.195	1.32
Bias 2020	-0.147	0.827

Weights = 0.5*time + 0.5*sample

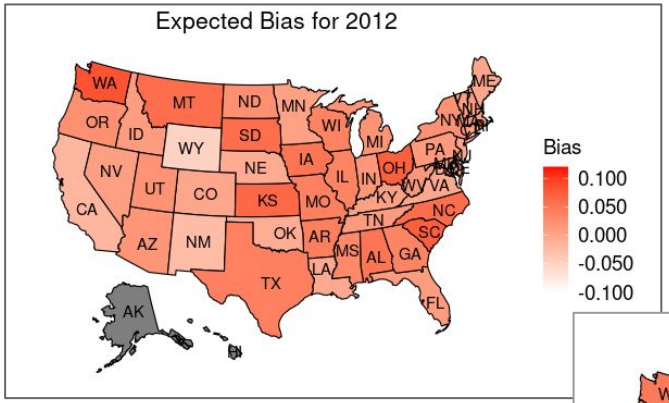
Is the bias temporally autocorrelated?

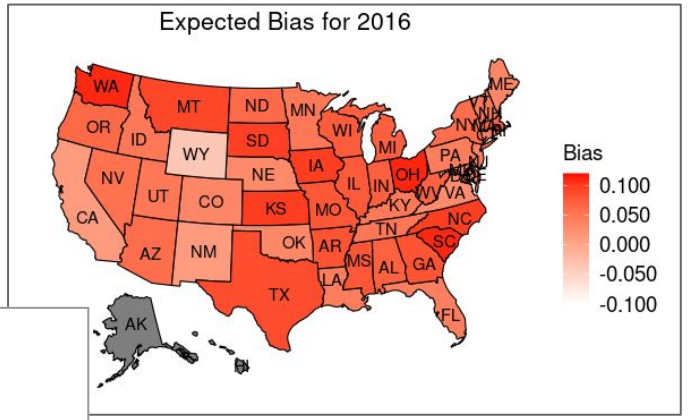


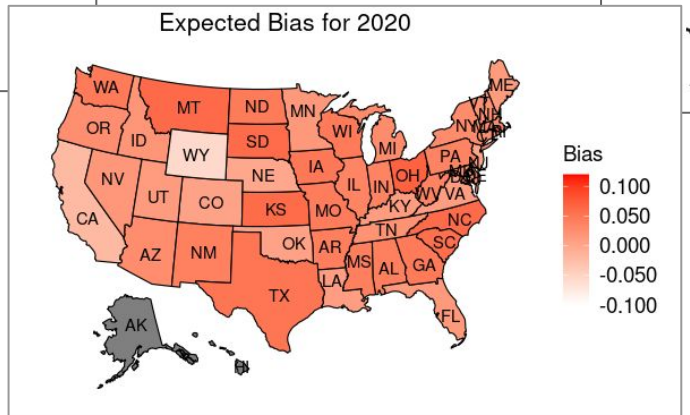


Negative autocorrelation for all the states in different times

Is there a spatial pattern common among election years?







Does the bias significantly vary over time/space?

- **No,** most of the states and years are insignificant other than Wyoming and 2016
- This could be due to the amount of data or it could be that the bias is mainly systematic

	Median	2.50%	97.50%		Median	2.50%	97.50%
(Intercept)	0.046	-0.0117	0.1042	xsNM	-0.0227	-0.1099	0.0582
xsAR	-0.0269	-0.0984	0.0418	xsNV	0.0111	-0.0694	0.087
xsAZ	-0.0054	-0.0849	0.0729	xsNY	-0.0179	-0.1024	0.0602
xsCA	-0.0663	-0.1537	0.016	xsOH	0.0233	-0.0539	0.1005
xsCO	-0.0467	-0.1214	0.0263	xsOK	-0.0512	-0.126	0.0217
xsCT	-0.0378	-0.1267	0.0477	xsOR	-0.0222	-0.1051	0.0592
xsDC	-0.0256	-0.124	0.0689	xsPA	-0.029	-0.1072	0.0465
xsDE	-0.0324	-0.1164	0.05	xsRI	0.0091	-0.0797	0.1016
xsFL	-0.0361	-0.1098	0.0342	xsSC	0.022	-0.057	0.1016
xsGA	-0.0074	-0.0748	0.0568	xsSD	0.0118	-0.0654	0.0867
xsIA	-0.035	-0.1112	0.0392	xsTN	-0.0335	-0.098	0.0298
xsID	-0.0106	-0.0902	0.0664	xsTX	-0.0029	-0.0829	0.073
xsIL	-0.0236	-0.1044	0.052	xsUT	-0.0289	-0.1055	0.0456
xsIN	0.0041	-0.0831	0.0891	xsVA	-0.0348	-0.1093	0.035
xsKS	0.0157	-0.0629	0.0918	xsVT	-0.0469	-0.1357	0.0371
xsKY	-0.0389	-0.113	0.0323	xsWA	0.0247	-0.0646	0.1134
xsLA	-0.0461	-0.1244	0.0288	xsWI	-0.0255	-0.1088	0.0541
xsMA	-0.045	-0.1285	0.0348	xsWV	-0.0086	-0.0906	0.0731
xsMD	-0.0442	-0.1237	0.031	**xsWY**	**-0.1007**	**-0.1771**	**-0.0267**
xsME	-0.0307	-0.1338	0.0666	**xt2016**	**0.0373**	**0.0133**	**0.0613**
xsMI	-0.0239	-0.1074	0.0579	xt2020	0.0071	-0.024	0.0362
xsMN	-0.0388	-0.1275	0.0455				
xsMO	-0.0123	-0.085	0.0568				
xsMS	-0.0054	-0.0765	0.0631				
xsMT	0.0113	-0.0685	0.0896				
xsNC	-0.0522	-0.1252	0.021				
xsND	-0.0366	-0.1196	0.0462				
xsNE	-0.0349	-0.1108	0.0412				
xsNH	-0.0449	-0.1357	0.0422				
xsNJ	-0.0562	-0.1461	0.0231				

How do our answers change for different p's?

	w= 1*time	w= 0.5*time + 0.5*sample	w= 1*sample
(Intercept)	0.03653	0.03717	0.03773
tau2	0.00139	0.00138	0.00138
nu2	0.00077	0.00075	0.00075
rho.S	0.85909	0.84161	0.82035
rho.T	0.50069	0.52952	0.54524

No significant changes with different weights.

Putting equal weights to recent polls and polls with larger sample sizes produces better model fit.

Conclusion

- There is bias in polls. On average the bias is positive (%GOP vote > weighted polls).
- There is some negative spatial correlation in the bias.
- Different weights did not have much impacts on the results.
- Different covariates did not show any significant impacts on the bias estimation.

Appendix: Implementation, Convergence, & Covariates

Implementation of CARBayesST package

- This package was based off of the CARBayes package, so the implementation and output is very similar, if not the same
- **Note:** CARBayesST expects a vector where the spatial points for each time point is stacked (i.e. Vec[1:n] <- 2012_data; Vec[n+1:2n] <- 2016_data; etc)

```
model1 <- ST.CARar(B~1,"gaussian",W=W, burnin=20000,
                   n.sample=100000,thin=10,verbose=TRUE)
```

```
> model1$summary.results
            Median   2.5%  97.5% n.sample % accept n.effective Geweke.diag
(Intercept) 0.0369 0.0322 0.0417     8000     100.0       8000.0         1.1
tau2        0.0014 0.0009 0.0022     8000     100.0       6340.8        -0.4
nu2         0.0008 0.0005 0.0011     8000     100.0       6410.8        -0.8
rho.S       0.8420 0.5123 0.9740     8000      44.1       5002.5         0.7
rho.T       0.5118 0.1098 0.8533     8000     100.0       5911.9        -0.4
```

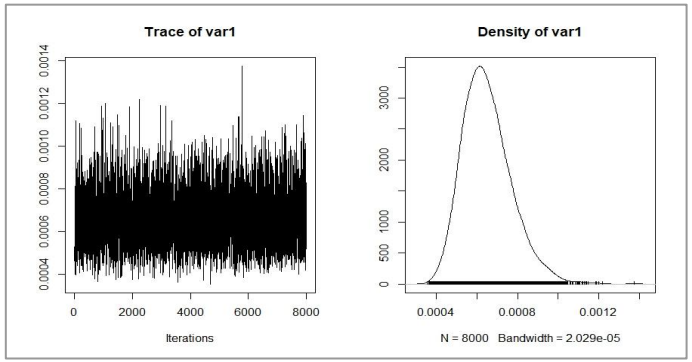
MCMC Convergence

Convergence:

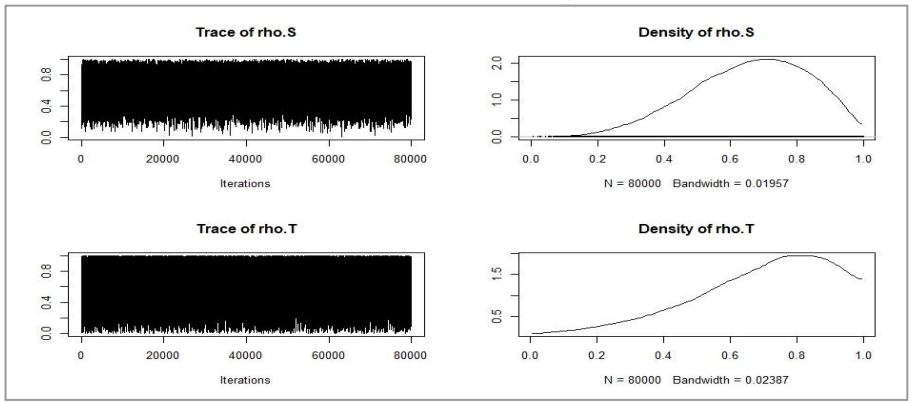
a. Trace plot.
b. n.effective- close to 1000 for each covariate,
c. Geweke.diag – between -1.96 to +1.96;

For our case, trace plot looks OK, also the n.effective and Geweke.diag are good with 100k sample (20k burn).

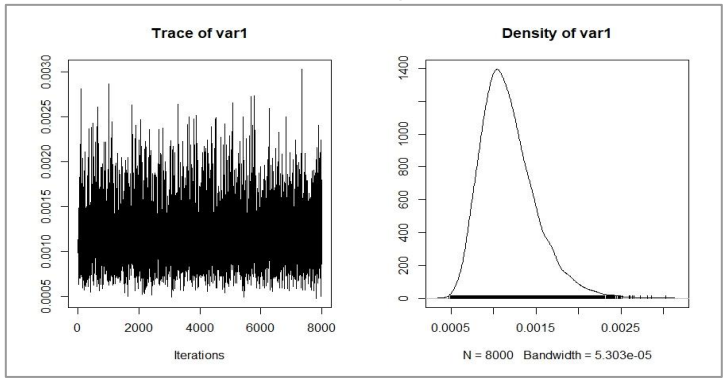
Nugget variance



Spatial/temporal dependence parameter



Variance parameter



Some covariates we tried

Covariates tested

- Unemployment rate for state/year
- Demographics (percentage of each sex/race in state/year)
- Average GOP support in state up until year
- Total GOP events in state/year
- Total GOP spending in state/year
- Total GOP spending / Total DEM spending

None of the covariates tested were significant, however, the only one that improved the fit of the model (though slightly) was **unemployment**

Covariates	DIC	WAIC
Base	-574.53	-599.13
Base+Unen	**-577.71**	**-602.75**
Base+Demo	-556.19	-584.16
Base+Hist_GOP	-572.07	-596.45
Base+Event_GOP	-571.85	-596.37
Base+Spend_GOP	-572.12	-595.83
Base+Ratio_GOP	-572.26	-596.38

Data sources:
2012- candidate fundraising by state: https://graphics.latimes.com/usmap-presidential-candidate-contributions/
2016 & 2020 candidate fundraising by state: https://www.fec.gov/data/candidates/president/presidential-map/
2016 and 2020 campaign events: https://www.nationalpopularvote.com/map-general-election-campaign-events-and-tv-ad-spending-2020-presidential-candidates
https://docs.google.com/spreadsheets/d/1oR_x3wGpFi1wO2V0BNMV529s_V-AgGH7tKd66DD7rrM/edit#gid=2025398596
2012 events: https://archive3.fairvote.org/assets/CNN-city-visits.xlsx