ST 533 Final Exam:

Creating a SpatioTemporal Model for the U.S. Election

Emine Fidan
Hunter Jiang
Vaidehi Dixit

NC STATE
UNIVERSITY

Data Processing

To prepare the poll and election data for spatiotemporal modeling, several processing steps were taken:

1. State, GOP support, election year, starting poll date, and ending poll date were the delineated variables within our election dataset.
2. The polls captured voter preference within a state over a range of time. Thus within this analysis, the median date that a poll was conducted was used as the temporal variable.
3. Poll and election data for Alaska and Hawaii were removed since they do not physically neighbor any state in the contiguous U.S.
4. The spatiotemporal CAR model allows NA observations in the response, thus NAs were kept within the dataset.

NC STATE UNIVERSITY

Poll Weights: Method 1

A geometric sequence can be used to upweight polls closer to election

$$a_n = a_1 \times r^{n-1}$$

- a_n: nth term value
- a_1: 1st term
- r: common ratio
- $n-1$: term number

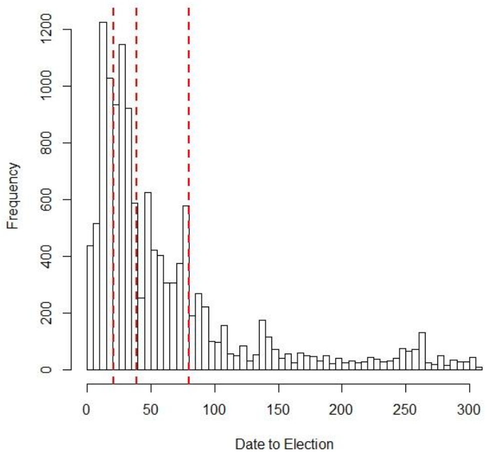
Here, a = (1-r)/(1 - r^n) and r = 0.85

For example, processed Arkansas 2012 election polls yielded:

Poller	GOP Support (%)	Year	Median Poll Date	Weight
The Arkansas Poll	58.0	2012	10 / 11 / 2012	0.5405405
Talk Business Poll	56.0	2012	09 / 17 / 2012	0.4594595

The weights successfully sum to 1
0.54+0.46=1.00

Poll Weights: Method 2



Another set of weights were calculated taking into account days until the election

Based on the temporal distribution for each state and year, the raw weights were assigned as:

$$w_{ij} = 1 * I(1 < t < 21) + 0.9 * I(21 < t < 39) + 0.8 * I(39 < t < 80) + 0.7 * I(80 < t < 309) + 0.6 * I(t > 309)$$

where $I(.) = 1$ if the expression $(.)$ is True

Then, each w_{ij} was normalized over each state and election year.

NC STATE UNIVERSITY

Poll Weights: Method 2

Example: For the state of Arkansas, 2012 election polls

Raw weights: Two election polls took place

$w_{i1} = 0.9$, $w_{i2} = 0.8$ (Based on time to election)

Normalized weights:

$w_{i1} = 0.9 / (0.9+0.8) = 0.5294$

$w_{i2} = 0.8 / (0.9+0.8) = 0.4706$

The weights sum to 1

For example, processed Arkansas 2012 election polls yielded:

Poller	GOP Support (%)	Year	Median Poll Date	Time to election	Weight
The Arkansas Poll	58.0	2012	10 / 11 / 2012	25	0.5294
Talk Business Poll	56.0	2012	09 / 17 / 2012	50	0.4706

Model set-up

Several approaches to building a spatiotemporal model using the 2012, 2016, and 2020 election data:

1. The {CARBayesST} package:
   a. Use the `CARlinear()` and `CARanova()` functions to build a model that represents the spatio-temporal pattern in the data
2. The {spBayes} package:
   a. Transform the areal to point-referenced data by using state centroids.
   b. Use the `spDynLM()` function to build a spatiotemporal model where space is continuous but time is discrete data
3. The {spTimer} package:
   a. Transform the areal to point-referenced data by using state centroids.
   b. Use the `spT.Gibbs()` function to build a spatiotemporal model and draw MCMC samples using the Gibbs sampler.

NC STATE UNIVERSITY

Model set-up

Let the polling bias $B_{it} = Y_{it} - X_{it}$

$B_{it} = \beta_0 + \phi_k + \delta_t + \epsilon \quad ; k = 1, 2, ..., K \quad t = 1, ..., N$

$K = 49, N = 3, \quad \beta_0 = E(B_{it})$

ϕ_k : spatial random effect

δ_t : temporal random effect

$$\phi_k \mid \phi_{-k}, \mathbf{W} \sim N\left(\frac{\rho_S \sum_{j=1}^{K} w_{kj}\phi_j}{\rho_S \sum_{j=1}^{K} w_{kj} + 1 - \rho_S}, \frac{\tau_S^2}{\rho_S \sum_{j=1}^{K} w_{kj} + 1 - \rho_S}\right),$$

$$\delta_t \mid \delta_{-t}, \mathbf{D} \sim N\left(\frac{\rho_T \sum_{j=1}^{N} \delta_{tj}\delta_j}{\rho_T \sum_{j=1}^{N} \delta_{tj} + 1 - \rho_T}, \frac{\tau_T^2}{\rho_T \sum_{j=1}^{N} \delta_{tj} + 1 - \rho_T}\right),$$

$\beta_0 \sim N(0, 100000)$

$\epsilon \sim N(0, \tau_I^2),$

$\tau_S^2, \tau_T^2, \tau_I^2 \; Inverse\ Gamma(a = 1, b = 0.01)$

$\rho_S, \rho_T \sim Uniform(0, 1)$

NC STATE UNIVERSITY

Model explanation

- The `ST.CARanova()` allows a random spatiotemporal interaction term, but due to lack of identifiability between the interaction and the Gaussian term we only include $\in$, random error.
- The conditional priors for the spatial and temporal random effects are as proposed by Leroux et al. (2000).
- Parameters (ϱ_S, τ_S^2) and (ϱ_T, τ_T^2) account for the strength of spatial correlation and the temporal correlation respectively.
- ϱ and (1 - ϱ) terms are basically weights assigned to the neighbors versus the non-neighbors.

NC STATE UNIVERSITY

Spatial Adjacencies

Generate an n x n matrix representing each state in our analysis

- If two states border, assign n_i x n_i a value of 1. Otherwise assign 0.
- Hawaii and Alaska were excluded (n=49).
- Diagonals were assigned a value of 0, rather than 1.

For example, the adjacency matrix for the first five states:

	Alabama	Arizona	Arkansas	California	Colorado
Alabama	0	0	0	0	0
Arizona	0	0	0	1	1
Arkansas	0	0	0	0	0
California	0	1	0	0	0
Colorado	0	1	0	0	0

California and Colorado border Arizona!

Temporal Adjacencies

Generate an m x m matrix representing each election year in our analysis

- If two elections occurred within a lag, assign m_i x m_j a value of 1. Otherwise assign 0.
- Diagonals were assigned a value of 0, rather than 1.
- This process was conducted automatically within {CARBayesST}

The temporal adjacency matrix for the 2012, 2016, 2020 election data:

	2012	2016	2020
2012	0	1	0
2016	1	0	1
2020	0	1	0

For the 2012 election year:
2012 and 2016 are within 1 lag, but 2020 is within 2 lags. Therefore, 2016 is assigned 1 and 2020 is assigned 0.

NC STATE UNIVERSITY

Objective 1

1. Combine polls into an average using two weighting schemes
2. Build a spatiotemporal model to forecast election results

To address point 1:

Use two methods to upweight polls closer to election

Example Code:

```
library(bsts)
GeometricSequence(length = n, initial.value = a,
                  discount.factor = r)
```

where n, a, and r represent the variables n, a, and r from our geometric sequence equation.

NC STATE
UNIVERSITY

Objective 1

1. Combine polls into an average using two weighting schemes
2. Build a spatiotemporal model to forecast election results

To address point 2:

Use the R package {CARBayesST} to build a spatiotemporal model

Example Code:

```
library(CARBayesST)
ST.CARanova(formula = B ~ X, family = "gaussian", data =
            polls, W = W, burnin = 20000, n.sample =
            1500000,thin = 100)
```

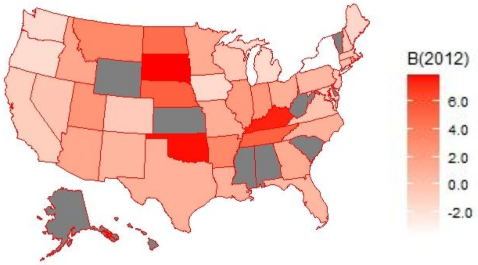
where `B` is the response variable, `X` contains the covariates, `polls` is the variable containing our dataset, and `W` is the spatial adjacency matrix.

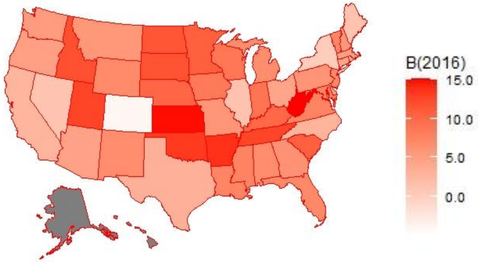
NC STATE UNIVERSITY

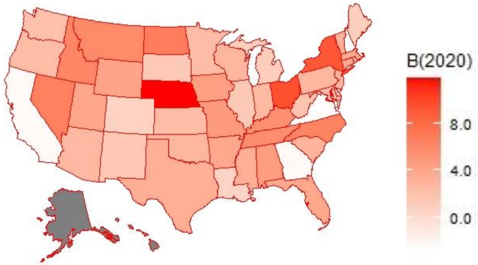
Results for B~1 Model

	Weighting scheme 1 (n.sample = 1.5M)		Weighting scheme 2 (n.sample = 1.5M)	
	Median	95% CI	Median	95% CI
(Intercept)	3.74	(3.17, 4.32)	5.12	(2.67, 5.56)
τ_S^2	0.01	(0.00, 3.91)	8.31	(3.61, 16.53)
τ_T^2	4.34	(1.11, 25.68)	9.76	(2.79, 54.41)
τ_I^2	11.57	(8.48, 14.92)	7.06	(5.36, 9.66)
ϱ_S	0.38	(0.02, 0.92)	0.51	(0.10, 0.92)
ϱ_T	0.21	(0.01, 0.83)	0.20	(0.01, 0.82)

B_{it} values for the election years 2012, 2016, 2020, respectively



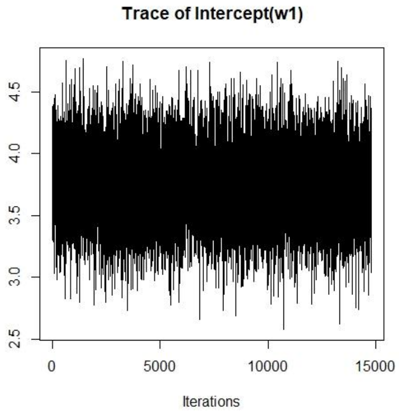


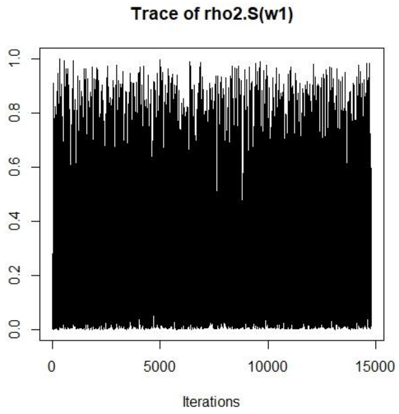


NC STATE UNIVERSITY

Model Convergence

	Weighting scheme 1 (n.sample = 1.5M)		Weighting scheme 2 (n.sample = 1.5M)	
	n.effective	Geweke	n.effective	Geweke
(Intercept)	14800	1.5	14530	0.4
τ_s^2	3137	-0.5	9827	1.1
τ_T^2	14800	0.9	14800	-1.4
τ_I^2	8222	1.0	4565	-0.2
ϱ_S	14800	1.6	14800	0.2
ϱ_T	14800	0.8	14800	-0.4





NC STATE UNIVERSITY

Objective 2

- To test whether systematic bias exists assuming it is constant over state and election,

  We fit the CARBayes model using ONLY the intercept as,

$$E(B_{it}) = \beta_0$$

  Which is constant over state and election.

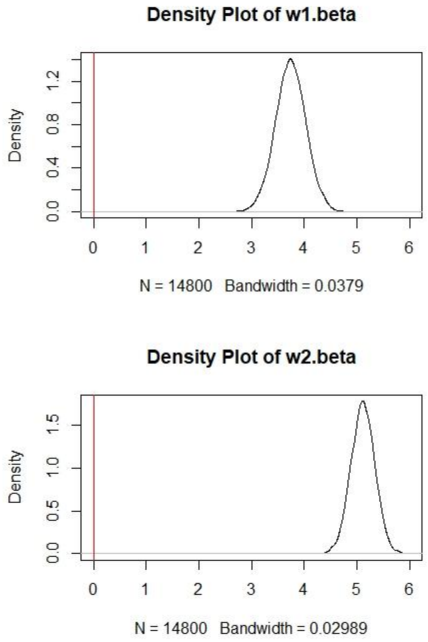
$$\text{To test } H_0: \beta_0 = 0 \text{ vs. } H_1: \beta_0 \neq 0$$

NC STATE UNIVERSITY

Objective 2

Using both the weighting schemes, we found that the 95% credible interval of the intercept did not include 0, and is distributed over a range of 3-6, indicating positive bias.

Conclusion : There is evidence of systematic polling bias, assuming it is constant over state and election.



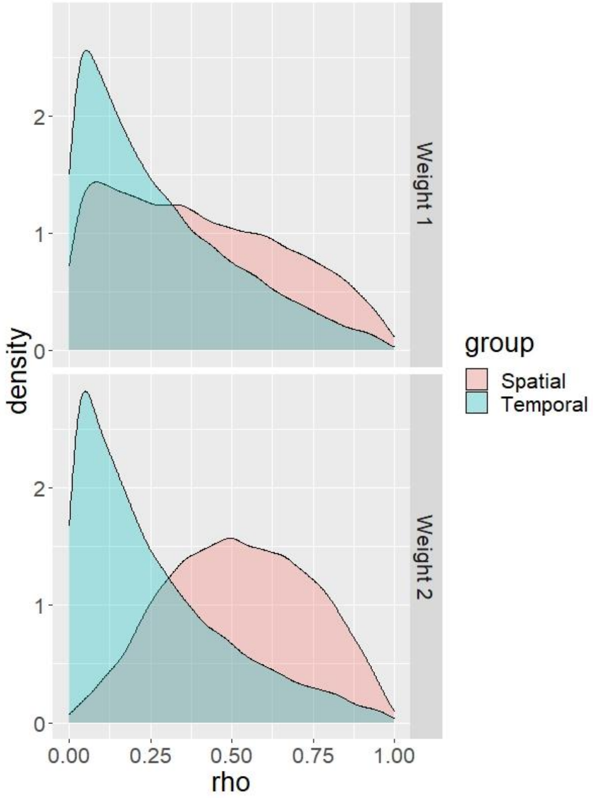
Density Plots for betas

NC STATE UNIVERSITY

Objective 3

To address objective 3:

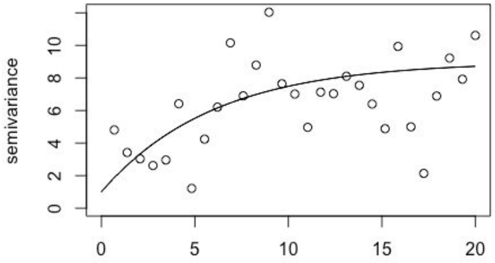
- From the distribution of strength parameters, we can find there is sign of spatial and temporal autocorrelations.
- We will first check the variograms and the Moran's I statistics for each year.
- To allow the bias to depend on space and time we add appropriate covariates as Xβ



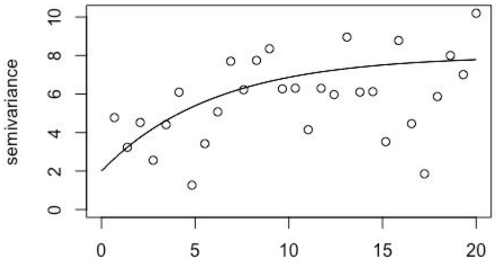
NC STATE UNIVERSITY

2012:
Moran's I = 0.249
p value = **0.008**

2012 Variogram using weighting scheme 1



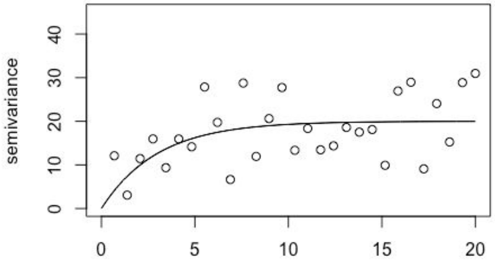
2012 Variogram using weighting scheme 2



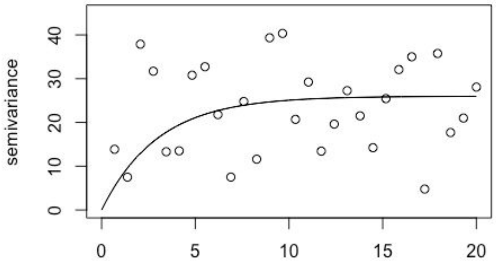
2012:
Moran's I = 0.170
p value = **0.044**

2016:
Moran's I = 0.087
p value = 0.158

2016 Variogram using weighting scheme 1



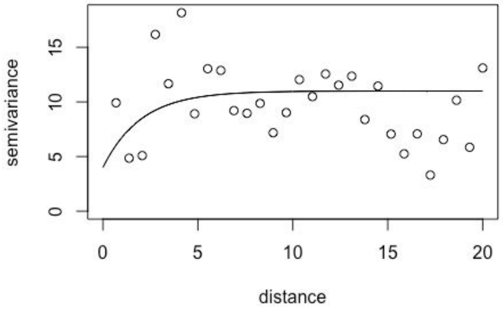
2016 Variogram using weighting scheme 2



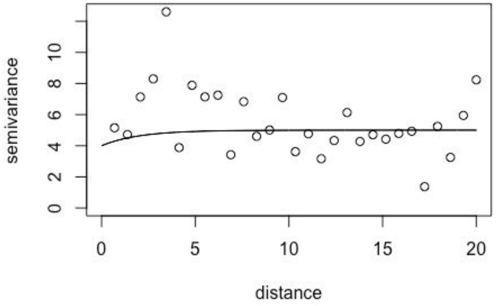
2016:
Moran's I = 0.325
p value = **0.003**

2020:
Moran's I = -0.068
p value = 0.640

2020 Variogram using weighting scheme 1



2020 Variogram using weighting scheme 2



2020:
Moran's I = 0.075
p value = 0.162

NC STATE
UNIVERSITY

Objective 3

```
> model1_result$summary.results
                    Median    2.5%   97.5% n.sample % accept n.effective Geweke.diag
(Intercept)         1.2328  0.1703  2.2876    14800    100.0     14800.0         0.7
as.factor(Year)2016 5.5576  4.0881  7.0243    14800    100.0     14800.0         0.4
as.factor(Year)2020 1.9155  0.4611  3.3796    14800    100.0     14432.4         0.2
tau2.S              0.0096  0.0022  3.8257    14800    100.0      2660.8         0.5
tau2.T              0.0085  0.0021  0.0887    14800    100.0     14145.2         0.0
nu2                11.5175  8.4976 14.8358    14800    100.0      7523.4        -0.3
rho.S               0.3728  0.0166  0.9202    14800     45.2     14800.0         0.2
rho.T               0.3814  0.0179  0.9128    14800     82.4     15198.3        -0.1
> model2_result$summary.results
                     Median     2.5%    97.5% n.sample % accept n.effective Geweke.diag
(Intercept)        -51.7196 -75.3461 -28.1945    14800    100.0     14800.0        -0.7
as.factor(Year)2016  5.5330   4.1524   6.9384    14800    100.0     14334.3         0.4
as.factor(Year)2020  1.8997   0.5117   3.2844    14800    100.0     14800.0         0.5
lon                 -1.1241  -1.6393  -0.6182    14800    100.0     14800.0        -0.8
lon2                -0.0058  -0.0086  -0.0031    14800    100.0     14800.0        -0.8
tau2.S               0.0084   0.0022   0.1137    14800    100.0      7402.5        -0.8
tau2.T               0.0083   0.0021   0.0890    14800    100.0     11593.6         0.8
nu2                 10.1562   8.0782  13.0380    14800    100.0     14800.0        -0.6
rho.S                0.3681   0.0179   0.9163    14800     45.0     14800.0         0.2
rho.T                0.3743   0.0165   0.9180    14800     82.5     14800.0        -1.3
```

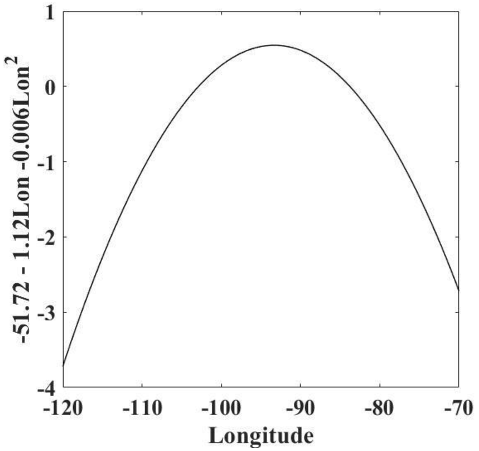
All significant (no change of signs); Enough samples; and good geweke statistics.

Objective 3

- Model 1 (weighting scheme 1):

  B_{it} = **1.23** + 5.56 I(year = 2016) + 1.92 I(year = 2020)

  $\tau_S^2 = 0.01$, $\tau_T^2 = 0.01$, $\tau_I^2 = 11.51$, $\varrho_S = 0.37$, $\varrho_T = 0.37$.



- Model 2 (weighting scheme 1):

  B_{it} = **-51.72** + 5.53 I(year = 2016) + 1.90 I(year = 2020) **-1.12lon - 0.006lon^2**

  $\tau_S^2 = 0.01$, $\tau_T^2 = 0.01$, $\tau_I^2 = 10.16$, $\varrho_S = 0.37$, $\varrho_T = 0.37$.

- Residuals from the random effect model can be explained by a mixed model with a factorial variable indicates election and a quadratic relationship to longitude.

NC STATE UNIVERSITY

Summary

- For objective 1, we built two different spatiotemporal models using **two types of weighting schemes**. One type weighted the recent polls more heavily than the other.
- For objective 2, both the weighting schemes indicated **positive systematic bias** but the results from the second weighting scheme were more prominent.
- {CARBayesST} gives fast efficient results and all parameters converge reasonably, barring the temporal component which could be better if more time points are involved.
- For objective 3, starting from the strength parameters in the CARanova models, we inspect variograms for each elections and run several models with more covariates. **The coefficients turned out to be significant indicating that the mean bias from the polls varies among election and states**.

NC STATE UNIVERSITY

References

[1] 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.

[2] Leroux BG, Lei X, Breslow N (2000). Statistical Models in Epidemiology, the Environment, and Clinical Trials, chapter Estimation of Disease Rates in Small Areas: A new Mixed Model for Spatial Dependence, pp. 179–191. Springer-Verlag, New York. URL http: //dx.doi.org/10.1007/978-1-4612-1284-3_4.

[3] R Core Team (2020). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/

[4] Bivand, R. S., Pebesma, E. J., Gómez-Rubio, V., & Pebesma, E. J. (2008). Applied spatial data analysis with R (Vol. 747248717, pp. 237-268). New York: Springer.

NC STATE UNIVERSITY