Spatiotemporal Modeling of Voting in North Carolina

Marschall Furman, Andrew Giffin, Matt Miller

North Carolina State University

Project Goal

The 13 voting districts in North Carolina have repeatedly been struck down by the courts for (illegal) racial gerrymandering; and the state has received national attention for its highly (legally) politically gerrymandered districts – which have been accused of guarding the majority party (Republican) seats, despite voting fluctuations.

This project examines the publically available highly local precinct-level voting and demographic data which feed into the district voting data, using a spatial areal-data model.

The Data

- Examining voting data for US House of Representatives races, for years 2002, 2004, ..., 2018.
- Currently 2,704 precincts. Due to boundary changes over time, and missing data, we model a subset of 2,045 of these precincts.
- The publically available election data includes: precinct-level vote counts (the response); as well as precinct-level demographic data on age, race, and sex (the predictors).

- covariate plots?

- Because this project relied of many different publically available datasets, a substantial amount of work was put into cleaning and standardizing the data. As a result of data irregularites, there are a number of caveats that we include with out data:
  - A substantial number of precincts were not reported for each year – either due to the precincts having changed over time, changing of precinct naming schemes, or missing data. All such precincts were excluded from our model.
  - Because absentee and early voting was not often reported at the individual precinct level, we excluded all absentee and early voting voters counts (usually $\sim 3\%$ of the votes) which could introduce bias.
  - Within each precinct, there were often covariates that had missing data. Our estimates simply summed over the available covariates.
  - One particular modeling nuisance was the occurrence of several elections with *unopposed* candidates (resulting in θ_{kt} at exactly 0 or 1.)
- Lastly, we used the full Binomial model, rather than the common normal approximation, because the normal approximation is ineffective for probabilities close to 0 or 1 – which is the case with a substantial number of our observed precincts.

another section?

Spatiotemporal Model

Model form

$$
\begin{aligned}
Y_{kt} &\sim \text{Binomial}(n_{kt}, \theta_{kt}) \\
\log\left(\theta_{kt}/(1-\theta_{kt})\right) &= \mathbf{x}_{kt}^\top \beta + \psi_{kt} \\
\beta &\sim \text{Normal}(\mu_\beta, \mathbf{\Sigma}_\beta) \\
\psi_{kt} &= \beta_1 + \phi_k + (\alpha + \delta_k)\frac{t-\bar{t}}{N} \\
\phi_k|\boldsymbol{\phi}_{-k}, \mathbf{W} &\sim \text{Normal}\left(\frac{\rho_{int}\sum_{j=1}^K w_{kj}\phi_j}{\rho_{int}\sum_{j=1}^K w_{kj} + 1 - \rho_{int}}, \frac{\tau_{int}^2}{\rho_{int}\sum_{j=1}^K w_{kj} + 1 - \rho_{int}}\right) \\
\delta_k|\boldsymbol{\delta}_{-k}, \mathbf{W} &\sim \text{Normal}\left(\frac{\rho_{slo}\sum_{j=1}^K w_{kj}\delta_j}{\rho_{slo}\sum_{j=1}^K w_{kj} + 1 - \rho_{slo}}, \frac{\tau_{int}^2}{\rho_{slo}\sum_{j=1}^K w_{kj} + 1 - \rho_{slo}}\right) \\
\tau_{int}^2, \tau_{slo}^2 &\sim \text{Inverse Gamma}(1, .01) \\
\rho_{int}, \rho_{slo} &\sim \text{Uniform}(0, 1) \\
\alpha &\sim \text{Normal}(0, 1000)
\end{aligned}
$$

where $\bar{t} = N^{-1}\sum_{t=1}^N t$ and the linear trend $(t-\bar{t})/N$ runs over $[-\frac{1}{2}, \frac{1}{2}]$; β_1 is the first element of β; $k = 1, \ldots, K$, where K is the number of precincts; $t = 1, \ldots, T = 9$; $\sum_{j=1}^K \phi_j = \sum_{j=1}^K \delta_j = 0$;

- The random effect ψ_{kt} incorporates both the spatial effect ϕ_k and a linear-trend time component $(\alpha + \delta_k\frac{t-\bar{t}}{N})$ for that individual location k. (In the above, *int* and *slo* correspond to "intercept" and "slope".)
- The terms ϕ_k and δ_k are modeled using their full conditional distributions, as determined by their neighbors (specified in adjacency matrix $\mathbf{W}$.)
- The hyper-parameters were chosen to be uninformative.

Parameter Estimates

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-307.3524	102.7686	-2.99	0.0042
log(GDP)	80.3790	10.1575	7.91	0.0000
locWestern	416.3502	139.9841	2.97	0.0044
locOther	640.5218	128.5153	4.98	0.0000
log(GDP):locWestern	-43.3499	13.6999	-3.16	0.0026
log(GDP):locOther	-73.3333	12.8909	-5.69	0.0000

$R^2 = 0.853$

Results

Year	Model	Data	Reality
2002	6	6	6
2004	7	6	6
2006	6	7	7
2008	6	8	8
2010	6	8	7
2012	3	3	4
2014	3	3	3
2016	4	3	3
2018	3	4	2 or 3

Conclusions