COS424 Final Project

Joe Bartusek
jfb4@princeton.edu

Shun Yamaya
syamaya@princeton.edu

May 2019

Abstract

This report exploits Ballot Image Data – a complete record of how the South Carolina Electorate voted (n = 1,569,831) – and voter file data to interpret latent dimensions in the 2016 general election. We first use unsupervised methods to observe hidden structures in vote patterns, and subsequently use demographic information to predict component loadings in a regression analysis. We find that in both national-state and local races, liberal-conservative ideology associated with partisanship explain around 75% the variance observed. Other important structures were national vs. state contests, and competitiveness in individual local races. Furthermore we find that ethnicity predicted partisan ideology well in the federal and state races, but performed significantly worse in more local races, supporting the view that different mechanisms govern voter considerations in sub-state politics. These results indicate that while party labels are strong forces across different levels of government, future political science research can benefit from more substantive local politics research.

1 Introduction

What latent structures exist in voting records? In general elections, citizens vote for a slew of offices across different levels of government, ranging from President to State Senator to County Council to Probate Judge. In this paper we try to understand and interpret patterns in voting records across multiple offices with two sets of analyses on a data set created from electronic ballot images generated from voting machines in South Carolina[1]. First, we apply three unsupervised methods – Principal Component Analysis (PCA), Latent Dirichlet Allocation (LDA), and Non-Negative Matrix Factorization (NMF)

[1]This data set is being developed by graduate student Shiro Kuriwaki and has been used by Shun in his independent work. See Section 3 for more details about the data.

– to find latent structures in the population. Next, we try to substantively interpret the structures by finding demographics associated with the latent structures. We regress precinct-level demographics, taken from L2 voter file data, on the PCA component loadings. We find that apart from the variance caused by partisan ideology, split-ticketing on the basis of national/state/local office is an important source of variance.

The paper is organized as follows: we first review the related literature and how our work fits with it. Then, we describe the data set in more depth, outline our analytic strategy, and show our results. We end with a brief discussion and conclusion to contextualize our findings.

2 Related Literature

Decades of political science research has documented the polarization of national politicians (Poole and Rosenthal, 1984, 2000). Using roll-call votes, academics have mapped estimated ideologies of federal politicians to show that, over time, Democrats have become more liberal and Republicans more conservative. Recent work suggests that state politics has also become increasingly partisan, following this national trend (Hopkins, 2018). This line of work finds that in the backdrop of polarized partisan politics, citizens are more keen to pick up on national party cues and apply them in the course of determining their votes on state-level contests.

However, traditional political science research has often overlooked the multi-faceted nature of ballots, and has often focused on only national or state level contests[2]. In reality, voters in the general election are required to make decisions for multiple offices – often more than 10 – across different levels of government, including county, municipal, and district-level offices. Voters make these series of choices sequentially, so understanding how the partisan nature of national and state level affects voters' choices for more local races is an increasingly important research agenda. This is especially true when many critical services of the local community (e.g. schools, civil court judges, waste services, large-scale infrastructures) are frequently managed by locally-elected officials.

There are a few reasons to believe that local level politics may differ substantively from national politics. Oliver et al. (2012) has suggested local politics may operate in a completely different style of democracy than state and national politics. He argues that vote choice for local contests is based on the personal ties the local community has to its

[2]Berry and Howell (2007) write that between 1980-2000, out of articles related to elections and published in the top 5 political science journals, only 6% have focused on state elections and 1% on sub-state elections.

TABLE 1. *Contests included in this analysis*

President	U.S. Senate	U.S. House	State Senate	State House	County Council	Sheriff	Clerk of Court	Coroner
National	National	National	State	State	County	County	County	County
46	46	46	46	46	35	35	35	35

leaders and their leaders' custodial performance. Furthermore, unlike national politics, researchers have found that local politics can be more ideology-based. Boudreau et al. (2015) demonstrates a correlation between voters' and candidates' estimated ideal points; further, Sances (2018) shows that voters learn the ideological positions of candidates over a campaign which in turn causally affects voter's vote choices.

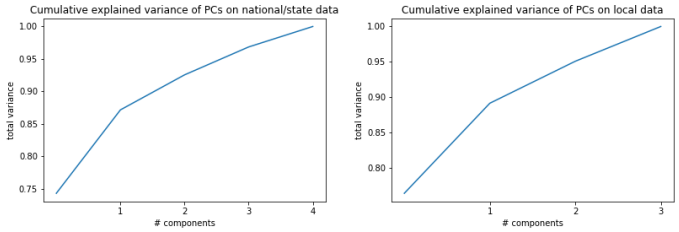
Our work in this paper situates itself to understand what dimensions may exist to voting in federal-state politics and local politics. Most recently, Kuriwaki (2019) has developed and analyzed ballot image data from South Carolina to show that a) there are sizable shares of voters who deviate from their national partisan preference and their local partisan preference and b) incumbents seem to have an advantage in local politics. We use this same data and extend his framework in three ways; first, we subset the data into national-state and sub-state contests; second, we apply two additional unsupervised models to compare with PCA; third, we use the computed principle components as dependent variables in a regression on precinct-level demographic data.

3 Data & Methods

The dataset we use is created and developed by Shiro Kuriwaki. South Carolina uses electronic voting machines which output anonymous sets of vote choices, which are later publicized by the Election Commission. For every election cycle, we have access to the state population's record of how every voter voted across multiple races. This means, for example, in the 2016 data set, we have 1,569,831 observations across 10-12 races, depending on the county.

We decided to focus on the 2016 general election for two main reasons. First, the 2016 election includes a wide range of races specified in Table 1 and thus gives us the opportunity to focus on multiple levels of contests across federal, state, and local level. Other years such as midterm or special election years lack the same kind of diversity. Second, the controversial election of Donald Trump makes 2016 a substantively interesting year to study voter behavior.

Figure 1. *Component tuning for PCA*



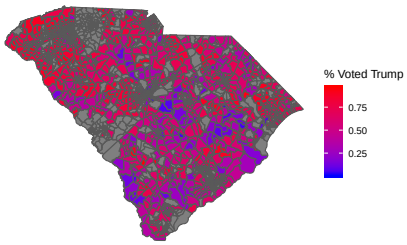
3.1 Pre-Processing of Data

Following recent work in political science which argues state politics to have become increasingly nationalized (Hopkins, 2018), we separate the ballot image data into two smaller data sets that consist of five national and state level races, and four sub-state races. In the first data set, we use the full 1,569,831 vote records for US President, US Senate, US House, SC Senate, and SC House. For the sub-state local races we focus on partisan races of which at least 5% of the races where contested. This is because one feature of local politics is that many races can non-partisan (ex. school board elections) and uncontested. As such, we try to circumvent analyzing races in which voters do not have a choice to begin with. We are left with 1,358,982 vote records for the 35 counties that participated in elections for County Council representative, Sheriff, Clerk of Court, and Coroner[3].

We transform each data set by replacing the name of the candidate with the party label. Since PCA, LDA, and NMF do not allow for missing data, we code Democrat, Republican, and other types of votes as -1, 1, or 0 respectively. "Other types of votes" includes abstentions, third-party candidates, and write ins. Finally, for the purposes of fitting a LDA and MNF models, we one-hot-encode the data into whether a voter voted Democrat or not, Republican or not for every relevant office.

We also use voter file data provided by L2. We generate precinct aggregates of percent non-white, senior, female, and college educated vote-registered citizens, as well as their median income. We later use these as independent variables to predict the PCA component loadings computed earlier.

Figure 2. *Which precincts voted for Trump?*



3.2 Methods

Our analytic approach is twofold. First, we apply three unsupervised methods (PCA, LDA, and NMF) to each of the federal/state level contests and local-level contests. Our baseline method is Principal Component Analysis, following (Kuriwaki, 2019). We tuned the number of components for PCA based on cumulative explained variance, visualized in Figure 1. Both graphs show that for each data set, with three components, 95 to nearly 100 percent of the variance is explained. Similarly, we explored using log-liklihoods and cross validation to hypertune component/topic numbers for LDA and NMF. We concluded that it would be appropriate to fit the same number of components for each classifier.

In our second set of analyses, we calculate the mean value for each principal component loadings by precinct. We then merge that data with voter file data to do regression analysis. We take each precinct-average principal component loading as a dependent variable and use precinct-level demographic proportions of senior, female, non-white, college educated citizens and their median income as independent variables to explore what kind of demographics are associated with each latent dimension.

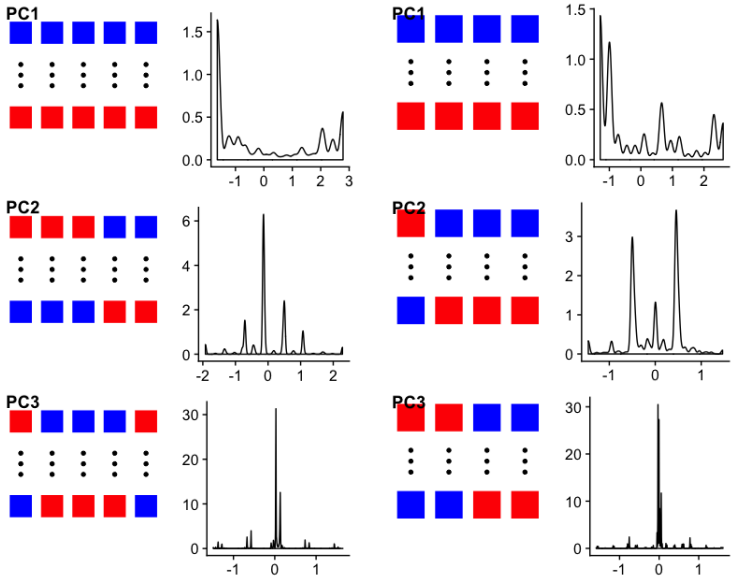
4 Results

4.1 South Carolina's Politics

We briefly contextualize our results within South Carolina's political landscape. In the seminal text, Key Jr (1949) describes southern politics to be centered around race. As a southern state, South Carolina is not exempt from this characterization. In South Carolina, 67% of the population is White/Caucasian and 27% is Black/African American. Whites are known to be more Republican and Blacks more Democratic. Overall, the state is a

[3]Some counties do not have certain races because they had already elected a Sheriff/clerk of court/coroner in the previous election cycle

Figure 3. *Extreme ballots and densities of PCs*



National/state races are represented by the ballots on the left. Local races are represented by the ballots on the right.

Republican stronghold that has voted for a Republican president in the past 10 election cycles (do you remember Jimmy Carter?) and for a Republican Congressional Senator since 1956. The only Democrat to serve a federal office in the state is Jim Clyburn, a U.S. House representative, who represents the predominantly black, southern, coastal "Lowcountry" regions.

In Figure 2 we show the precinct-level returns for the proportion of voters who voted for Trump in the 2016 election. We can see that other than the few hints of blue and purple in the southern areas, the state is predominantly red.

4.2 Principal Component Analysis

We first visualize the components produced by PCA. In Figure 3, we show the density of each component and visualize what the extremes of each component would look like in terms of actual votes. Interpretation of the dimensions produced by PCA is the most difficult part of the process, but the extreme ballot races (in the order specified in Table 1) allow us to make educated guesses. The first component for both federal-state and local level races seems to be liberal-conservative ideology associated with partisanship. This explains

TABLE 2. *Correlation of national and state contests on PCs*

	President	U.S. Senate	U.S. House	State Senate	State House
PC1	-0.49	-0.49	-0.50	-0.35	-0.38
PC2	0.33	0.29	0.28	-0.61	-0.60
PC3	0.06	-0.07	-0.03	-0.71	0.70

TABLE 3. *Correlation of local contests on PCs*

	County Council	Sheriff	Clerk of Court	Coroner
PC1	-0.29	-0.57	-0.54	-0.55
PC2	0.95	-0.18	-0.20	-0.12
PC3	0.02	0.77	-0.20	-0.61

the largest amount of variation across ballots, and dwarfs other components. The second largest source of variation among ballots at the federal-state level comes from voters that split their ticket between national-level and state-level offices. The third component is a little bit more difficult to interpret, but looking at Table 2, we can see that the correlation coefficients are largest and opposite at State Senate and State House races, suggesting variation produced by people who split their ticket between the two legislative houses at the state level.

At the sub-state level, structural variation seems to be coming from competitiveness in individual races. The second component seems to explain voters who deviate at County Council races, and the third those who deviate at County Council and Sheriff. Out of the 4 races, both County Council and Sheriff races were the most contested, with 22% and 46% of races, respectively, having more than one candidate running. On the other hand, only 20% of Clerk races and 6.9% of Coroner races were contested, resulting in less variability in the outcome of votes for these races. Accordingly, we think that the second and third components of the local contests are very much influenced by how contested each race is.

We find that NMF and LDA largely confirm the results we find with PCA. NMF and LDA produce very similar analyses of the latent structure of our data; a strong first component/topic, clearly denoting partisanship, and weaker second and third components/topics, denoting split-ticketing between national- and state-level races.

4.3 Regression on Precinct-level Demographics

In Appendix A we show results from our multivariate OLS regression. Here, we regress loadings on the six PCA components produced in the earlier section (three components ×

two data sets) upon precinct-level demographic variables. The aim of this analysis is to supplement our understanding of components by exploring what kind of demographics are associated with the latent dimensions. It is also worth mentioning upfront that our unit of analysis is no longer individuals, but precincts ($N = 2022$).

On the federal level, we find that the proportion of non-white voters strongly predicts the first component loading. The effect size is extremely large: moving from a completely white precinct to a completely non-white precinct[4] increases the loading on principal component 1 by 3.568 (a more Liberal/Democratic vote on this scale). Putting into perspective that the range of values principal component 1 can take is 4.42 (between -1.64 to 2.78), we can see that this effect size easily dwarfs other predictors. As mentioned earlier, politics in South Carolina is very much race-based, in which white precincts vote Republican and black precincts vote Democratic. We also find that for the second component on the federal level older, male, non-white, and less educated precincts are associated with voting Republican at the federal level races.

For the other component loadings, we were not able to find statistically significant relations between observed demographics. This is particularly surprising for the first component in the local races. Contrary to what we find in national races, ethnicity does not seem to predict partisanship for local races. This suggests that there are different forces at work, apart from demographics, driving voting at the local level. Some potential avenues for exploration would be candidate quality, amount of money fund raised, and length of incumbency.

5 Discussion & Conclusion

Our inspiration going into this project was to explore and test the efficacy of latent structures in ballot image data; we have confirmed the importance of split-ticket voting along national/state/local lines, and have demonstrated the influence of key demographics on certain components.

Our results are limited in their capacity to be generalized because our analysis is restricted to South Carolina, a state which itself is a solidly Republican region in the American political landscape. Given access to ballot image data and voter file records from other states, particularly states on the other end of the political spectrum (Democratic) and battleground states, and applying the same analysis, we could get a much clearer picture of voter behavior in American politics as a whole.

[4] In the context of South Carolina, non-white generally means Black/African American

Additionally, we could have extended our analysis by including more factors relating to candidates – for example, conducting PCA on the gender, age, and incumbency of candidates along with their party and the contest they are participating in. This would lead to more insightful observations about the ways voters are influenced by candidates in their choices on the ballot.

References

Berry, C. R. and Howell, W. G. (2007). Accountability and local elections: Rethinking retrospective voting. *The Journal of Politics*, 69(3):844–858.

Boudreau, C., Elmendorf, C. S., and MacKenzie, S. A. (2015). Lost in space? information shortcuts, spatial voting, and local government representation. *Political Research Quarterly*, 68(4):843–855.

Hopkins, D. J. (2018). *The Increasingly United States: How and Why American Political Behavior Nationalized*. University of Chicago Press.

Key Jr, V. O. (1949). *Southern politics*.

Kuriwaki, S. (2019). Partisan allegiance on the long ballot. *Working Paper*.

Oliver, E. J., Ha, S. E., and Callen, Z. (2012). *Local Elections and the Politics of Small-Scale Democracy*. Princeton University Press.

Poole, K. T. and Rosenthal, H. (1984). The polarization of american politics. *The Journal of Politics*, 46(4):1061–1079.

Poole, K. T. and Rosenthal, H. (2000). *Congress: A political-economic history of roll call voting*. Oxford University Press on Demand.

Sances, M. W. (2018). Ideology and vote choice in us mayoral elections: Evidence from facebook surveys. *Political Behavior*, 40(3):737–762.

Appendix

A Regression Tables for OLS

TABLE 4. *Regression output*

(a) *PCA loadings on federal and state races*

	Dependent variable:		
	PC1	PC2	PC3
	(1)	(2)	(3)
Percent Senior	0.256*	0.445***	−0.007
	(0.148)	(0.131)	(0.090)
Percent Female	1.175**	−1.576***	0.201
	(0.549)	(0.487)	(0.334)
Percent Non-White	3.568***	0.845***	−0.014
	(0.073)	(0.065)	(0.045)
Percent College Educated	0.096	−0.820***	0.031
	(0.205)	(0.182)	(0.124)
Median Income	−0.123*	−0.136**	−0.072*
	(0.071)	(0.063)	(0.043)
Constant	−0.627	2.321***	0.670
	(0.823)	(0.730)	(0.500)
Observations	2,022	2,022	2,022
R^2	0.690	0.228	0.004
Adjusted R^2	0.689	0.226	0.002
Residual Std. Error (df = 2016)	0.560	0.496	0.340
F Statistic (df = 5; 2016)	897.855***	118.987***	1.615
Note:		*p<0.1; **p<0.05; ***p<0.01	

(b) *PCA loadings on local races*

	Dependent variable:		
	PC1	PC2	PC3
	(1)	(2)	(3)
Percent Senior	0.147	−0.124	0.028
	(0.320)	(0.110)	(0.049)
Percent Female	−0.196	−0.194	0.051
	(1.212)	(0.418)	(0.187)
Percent Non-White	0.269	0.022	0.048*
	(0.164)	(0.057)	(0.025)
Percent College Educated	−0.079	−0.154	−0.079
	(0.453)	(0.156)	(0.070)
Median Income	0.228	0.012	0.082***
	(0.155)	(0.053)	(0.024)
Constant	−2.475	0.045	−0.923***
	(1.805)	(0.623)	(0.278)
Observations	1,729	1,729	1,729
R^2	0.003	0.003	0.011
Adjusted R^2	0.0002	0.0003	0.008
Residual Std. Error (df = 1723)	1.157	0.399	0.178
F Statistic (df = 5; 1723)	1.077	1.100	3.662***
Note:		*p<0.1; **p<0.05; ***p<0.01	