PROJECT 3: RACIALLY POLARIZED VOTING

MGGG

We'll be walking through some methods for performing racially polarized voting (RPV) analysis. Two of the leading statistical methods are called *ecological regression* (ER) and *ecological inference* (EI). ER is just a simple linear regression which many of you learned about in Stats 101 (and if not, you'll learn about here). EI is a bit fancier, but not hugely different in the kinds of answers that come out. In this project we walk through some methods for analyzing racial polarization in real elections.

1. Warmup

Our Lab has performed racially polarized voting analysis and written reports in several cities and counties, including Santa Clara, CA, Yakima County, WA, and Chicago, IL. Here are some links to reports that include RPV analysis.

- MGGG Santa Clara report (public: regarding remedial phase for lawsuit)
- MGGG Yakima report (commissioned by civil rights litigators for challenge letter)
- MGGG Chicago report (for community organizers to boost reform conversation)

Some friends of MGGG also serve as experts in VRA litigation around the country. These are what some full-fledged expert reports look like. (If you are curious, ask– we have literally thousands more examples!)

- Fred McBride Sumter County, GA report (expert report filed in court case)
- Matt Barreto Orange County, FL report (expert report filed in court case)

Your Task: Skim (or read!) one of the MGGG reports, pick a few elections, and replicate the findings.

We'd like you to use the **Shiny EI app** that some of our students made (`https://vrdi.shinyapps.io/ei-app/`). You can find data in the Project 3 github repo.

Shiny is a package that lets you build interactive apps based on the R programming language (which is one of the most popular languages for statistics). Some of you have installations of R on your laptops, in which case you can run Shiny locally, or skip Shiny entirely and run EI packages directly in R. For everyone else, you can run it in your browser. We'd like *everyone* to try running EI and ER in Shiny. (We'll start with the User Demonstration walk-through.)

Note: The Shiny app may slow down and/or crash with too many people using it simultaneously. If this happens, one person per cohort can use it and screen share.

2. Explore prepared data

The github contains prepared data for Chicago, Lowell, Everett, Yakima, and Santa Clara. Pick one or two places to focus on.

Not all of the localities we are studying here have data that is formatted in the same way, so you'll have to pay attention to what columns are available. You may even need to compute some new columns before you're able to do everything you want in the Shiny app.

Your Task:

- Find the total population, and figure out what share of population belongs to various minority groups. Do the same for CVAP.
- What is the distribution of population across the precincts? (Are some much more populous than others or are they all fairly even?)
- Choose a minority group (or a coalition) of significant size and explain how you chose it.
- What is the distribution of the minority group's CVAP across the precincts?
- For a few key elections—primary and general—how does turnout vary across the precincts and how does that correlate to racial composition of precincts?

- Run EI/ER in several elections and see what you find in terms of polarization.
- Can you find examples where EI and ER don't agree? Any thoughts on why, or which you would trust more?

Challenge Question: Create made-up datasets that illustrate how things can go wrong. For each of the following try to construct a simple sample input file (made-up data for 10-20 precincts) for the Shiny app that demonstrates a situation where your EI/ER inferences could be misleading.

- What kinds of problems can occur in your conclusions when the CVAP varies too narrowly? (For instance, if the minority population is between 20 and 30% in every precinct.)
- How can differential turnout impact your findings? Think about this open-endedly first. We've prepared a sample spreadsheet to explore turnout effects here. Dive into this as a cohort and discuss what you find!
- How can unequal population across the units impact your findings? What might be reasons for or against trimming out very low-population precincts?

3. From raw data to RPV

You have the following raw materials for Santa Monica, CA and Denham Springs, LA.

- Precinct shapefile
- Census block shapefile with demographics
- Census block group shapefile with CVAP
- Tabular election results

Your Task: Assemble this data into a unified CSV by joining election results to precincts, disaggregating CVAP data from block groups to blocks, aggregating from blocks to precincts, and calculating totals and proportions as needed. The MAUP activity may help review the aggregation steps; Project 0 may be helpful to review joins and other data merging tips.

Spend some time exploring the Santa Monica or Denham Springs data. Use EI/ER, but also use various other kinds of visualizations like choropleths. What's the story of racial polarization there? Is there potential for a VRA case?

4. Extras: A Texas example

Here's a big data example: the entire state of Texas. We'll look at the 2018 Democratic runoff election for Governor between Lupe Valdez and Andrew White.

Materials: Tabular data from this election can be found in the Project 3 GitHub and a shapefile for Texas precincts can be found here.

Your Task:

- One person per cohort should run a full state EI /ER on precincts for the GOV18 Democratic runoff to assess statewide candidate support for each candidate among Black voters (using BCVAP). Tip: if this is taking a very long time, try removing the precincts with a small number of votes (say fewer than 10) and rerunning.
- Interpret the EI results, and use the ER plots to visualize/analyze.
- Everyone in cohort should run EI on precincts for each of the state's two largest counties: Harris County (which contains Houston) and Dallas County. For each county, assess candidate support for each candidate among Black voters (using BCVAP) in that county.
- Make choropleths of support for these candidates, and compare support in Harris and Dallas county.
- Explore and discuss!

5. Extras: Supplemental data sources

5.1. **Voter files.** Self-reported race is on the voter registration file for several states, including Florida. We've provided a real voter file from Broward County, FL (broward_voter_file). Take a look at the README and the columns in the voter file to familiarize yourselves with the contents.

The column called "Voted in Pres 16" tells you whether the voter participated in the 2016 Presidential general election. Note that the file lists each voter's party affiliation, but not who they voted for in the election. The file also includes a Race column; this is the voters' self-reported race. The race codes are in the README file.

We can use the voter file to estimate *turnout by race*, or the share of a particular race of voters that cast ballots in the election. This information is critical for doing an accurate RPV analysis.

5.2. **Voter files plus surname data.** However, many voter files do not include self-identified race. While ER and EI are often used to estimate turnout by race, there are techniques to predict a voter's race by other information about them in the voter file.

Bayesian Improved Surname Coding (BISG) is a way to predict a voter's race using their surname and address. We ran BISG on the Broward voter file and the results are also included in the Broward data, in the file called geocoded_and_bisged. More details can be found in the README.

Try using both self-reported race and BISG-predicted race to estimate turnout-by-race for PRES16 in Broward county. How do the estimates compare? Any theories about why they may be different?

5.3. **Court cases.** We've provided some legal materials in the github from two small pieces of Dallas County that had interesting court cases: Farmer's Branch and Irving. The cases are interesting because the plaintiff's experts had pushback on their Gingles demonstrations. Explore that data and see what you find.

6. Extras: Supplemental analysis

In our Chicago report, we showed some statistics for 2×2 EI. That is, we chose one racial group at a time and one candidate at a time—for instance, if focusing on Black support for Chuy Garcia, the groups would be Black vs non-Black and Garcia vs. non-Garcia. However, ecological inference can also be performed in $R \times C$ form, with any number of rows and columns. We built a shiny app for $R \times C$ EI (`https://vrdi.shinyapps.io/ei-app-RxC/`) but it may be buggy, so this is probably best run in R. (You can find the appropriate packages linked at the app.)

Compare point estimates in Chicago (Rahm Emanuel vs. Chuy Garcia) if the three major racial groups are handled with three 2×2 EI runs or one 3×2 run.

Note: We have not actually tried this and are very curious to see what you find!