# 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. We hope that Parts 1-3 can be done in 3 to 4 hours by someone who has already worked through the Geodata curriculum that precedes this. (The preceding material is needed for part 3, but part 2 stands alone.) This is followed by some supplementary project material. Parts 4-7 are more open-ended and more challenging.

## 1. Setup

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)

Below, we will walk you through using 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. If you don't know how to clone a github repo, here is a Dropbox link that gets you all the same data.

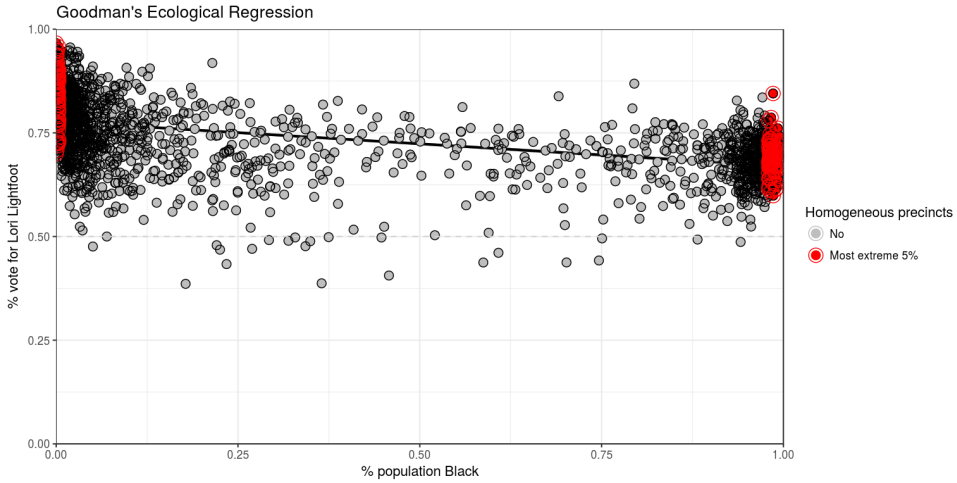
Shiny is a package that lets you build interactive apps based on the R programming language (which is one of the most popular languages for statistics). Some of you have installations of R on your laptops, in which case you can run Shiny locally, or skip Shiny entirely and run EI packages directly in R. For everyone else, you can run it in your browser. We'd like *everyone* to try running EI and ER in Shiny.[1]

The github contains prepared data for Chicago, Lowell, Everett, Yakima, and Santa Clara. Let's start with the Chicago 2019 mayoral runoff and walk through how to use the Shiny app to make some educated guesses about racially polarized voting.

The output includes a scatterplot where every point is one of Chicago's 2000 or so precincts (shown below). The ones colored red are the most homogeneous—the 5% of precincts that have the highest Black share of population, and the 5% that have the lowest. (And hey– this picture also shows that Chicago's Black population is quite segregated! notice how many of the dots are either very close to 0 or 100% Black.)

The summary table at the top contains three kinds of inferences In the scatterplot, a best-fit line (called a *regression line*) is shown for all the points. The fact that it's pretty flat (close to horizontal) shows that Lightfoot's support didn't vary very much depending on the Black share of residents. In particular, looking at how many of the points are above 50%, you can she that Lightfoot got more votes than her opponent (Toni Preckwinkle) in the vast majority of precincts in the city.

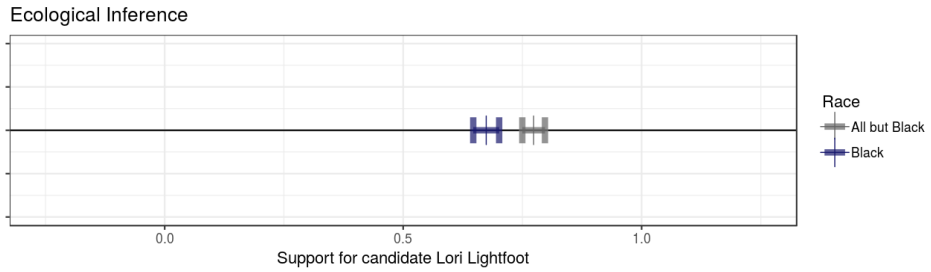
[1] **Note:** The Shiny app may slow down and/or crash with too many people using it simultaneously. If you're part of a group trying to work through these exercises, you might need to have one person at a time working in Shiny.




Looking up to the summary table at the top, you can see that regression predicts about 78% support from non-Black voters and about 67% support from Black voters. The way that corresponds to the picture is by taking that regression line out to the extremes: what does it predict when the precinct population is 0% or 100% Black?

"Homogeneous precincts" tells you want would have happened if you only fit the line to the homogeneous precincts (the red points). This can be a little different from the ER result, because it's choosing to ignore most of the data! And the "EI" results tell you the point estimates that come out of King's EI method, where a bivariate normal distribution is fit to so-called *tomographic lines* from each precinct. Note: the ER and homogeneous precincts results are deterministic, so if you re-run the data you will always get the same answer. But the EI uses a randomized algorithm (MCMC) to estimate a best-fit probability distribution for the data, so your answer may be slightly different when you re-run.

A little lower, you get a visual that shows the EI outputs, with confidence intervals.




What this figure is showing you is not only the point estimates for the two candidates, but also the 95% confidence intervals.[2] The fact that the confidence intervals do not overlap means that the model predicts greater than 95% confidence that non-Black voters went for Lightfoot at a higher rate than Black voters.

## 2. Explore prepared data

Take a look at what is in the files. (You can do this by previewing the CSVs in github, or you can load a file into the Shiny app and use the *data* tab.) Next, use the dropdown to pick column headings and hit run. Once you run it, you'll need to switch back to the *Candidate 1* tab to see outputs.

The Chicago data includes the 2015 mayoral runoff between Rahm Emanuel and Chuy Garcia, and also includes the first round from 2019 as well as the runoff election. Each row of the CSV is a precinct, and the columns tell you either counts or percentages for various race and voting statistics in the precinct.

To replicate the results above, load the 2019 runoff into Shiny, choose 1 for your number of candidates being considered, pick Lori Lightfoot (P_LIGHT gives her percentage), and select Black voters. In the text

[2]Small print: technically, they are credible intervals rather than confidence intervals, but the idea is similar. Warning! this app is running the package eiCompare, which has a known bug in how it computes confidence intervals for EI. So this needs to be fixed before you'd go to court with it, but let's put that aside for now.

fields asking for your data source, just write MGGG github. Then compute! This is a big election so it might take 1-2 minutes to run.

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. Important note: this app wants *percentages* for votes and demographics, and it will fail to run if you choose columns with counts instead. In this case the percentage column was P_LIGHT, and most of the examples here use the prefix P to tell you it's a percentage.

YOUR TASK: Try the Shiny app for a few other CSVs in the prepared data! Here are some suggested column combinations to get you started (Data file/Candidate 1/Demographic variable/Total votes)

- Chicago: Mayoral19_runoff.csv / P_LIGHT / PER_BLACK / TOTVOTES
- Santa Clara: SantaClaraSampleData.csv / pct_for_hardy2 / pct_asian_pop / total2
- Yakima: YA_2016_D2_GEN.csv / P_MANJAR / P_CVAPHISP / TOTVOTE
- Lowell: Lowell_EI_15.csv / PCC_Nuon / CVAPASIAN / TOT_CC
- Everett: Everett.csv/ P_SIMONELLI / P_HISP/ TOT_CCAL

Think about whether the column names are ambiguous and about what naming conventions you want to use when you do it yourself below. Your writeup should include one EI/ER output from the recommended set above, and one or two others that you come up with yourself.

## 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

You may need to compute some new columns before you're able to do everything you want in Shiny.

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: THINGS THAT CAN GO WRONG

Points to discuss!

- What kinds of problems can occur in your conclusions when the share of a minority group (or CVAP percentage) varies too narrowly? (For instance, if the minority population is between 20 and 30% in every precinct.)
- How can unequal population across the units impact your findings? What might be reasons for or against trimming out very low-population precincts?
- How can differential turnout impact your findings? For instance, Asian and Latino populations—as well as other groups with a high share of immigrants—tend to have lower citizenship rates, lower voter registration, and lower turnout than the population at large. Suppose you are working on behalf of a minority group that has low turnout, but you base your statistics on the group's share of population. How can this affect your findings?

CHALLENGE QUESTION: Create made-up datasets that illustrate how things can go wrong. For each of the points above, 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.

For instance, we put a made-up dataset in the github to explore turnout effects: challenge_question_example. Explore this and discuss what you find! (And try something similar for the other bullet points.)

## 5. 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!

## 6. Extras: Supplemental data sources

6.1. **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 (of course—this is private). The file also includes a Race column; this is the voters' self-reported race. The race codes are in the README file.

You 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.

6.2. **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 Geocoding (BISG) is a way to predict a voter's race using their surname (their last name) 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?

6.3. **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.

## 7. Extras: Supplemental analysis

In our Chicago report, we showed some statistics for $2 \times 2$ 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 $2 \times 2$ EI runs or one $3 \times 2$ run.