# It's not fair: gerrymandering by whom?

Bruce Mallory

12/2/2020

## ABSTRACT

This project set out to analyze whether or not the political party in control of redistricting for federal congressional districts had an effect on the fairness of the outcome. For this analysis, a metric for fairness (proportional-political-representation or PPR) was calculated. Election results for the years 2002 through 2018, along with data about who did the redistricting prior to those elections, was collected and analyzed. Though it was clearly demonstrated that there was a difference in the proportional-political-representation when one or the other political party drew congressional districts, when the number of congressional districts being drawn was added to the analysis, the effect of politically controlled redistricting was no longer apparent.

## INTRODUCTION

Every ten years, after the census is completed, the 435 House Representatives are reapportioned among the states based on the current populations. Then each state redraws its electoral districts so the districts within a state have approximately equal populations. Aside from mandating that the districts have equal populations, the Constitution is silent about the how the drawing of districts should be done by each state. Redistricting laws in the different states vary as does the responsibility for redistricting. In some states redistricting is done by the courts, in other states there are independent commissions. And in the majority of states, redistricting is done by the state legislature and the governor. Of these states, the last time congressional districts where redrawn (2011), twenty-three had a state legislature and a governor who belonged to the same political party.

In assessing the fairness of the redrawn electoral districts, there are four overlapping and often conflicting, perspectives.

First is the hope that districts will be politically proportional, yet that doesn't always happen. In North Carolina, in 2018, the Democratic candidates (as a group) received 48.3% of the votes cast. That year, only 3 of the 13 representatives (23%) were Democratic. And in Massachusetts in 2018, Republican candidates received 36% of the votes cast, yet none of the 9 congressional representatives were Republican.

A second concern in redistricting is the competitiveness of the districts. Clearly there is a relationship between the number of competitive districts and proportional-political-representation. In the appendix I discuss my observations about the number of uncontested and competitive districts that I observed in the years that I've analyzed (2002 to 2018).

The third perspective about fairly drawn congressional districts is stipulated by the Voting Rights Act of 1965 and focuses on a minority groups' ability to elect representatives of their choice. The law addresses instances where this ability is diminished during redistricting and focuses on assuring that there are "majority-minority districts."

And, the fourth perspective on drawing "fair" electoral districts is most clearly articulated in California's 2010 voter initiative that established a redistricting commission tasked with creating "communities of interest." The goal was to create districts that allow like-minded communities to have representation in congress.

The analysis that I am doing in this paper will only focus on proportional-political-representation and will not address community representation or minority representation. All the while understanding that the Voting Rights Act impacts the drawing of districts in a manner that can skew proportional-political-representation (ironically, assuring majority-minority districts is a form of packing and can dilute Democratic representation). Also the goal of keeping communities of interest together can also impact proportional-political-representation.

In my current analysis, I will focus on the proportion of seats to votes. Specifically, I define the proportional-political-representation coefficient (PPR) to be the (% of seats) divided by the (% of votes). From a proportional-political-representation perspective, if an election is "fair," then the proportion of seat to votes should be 1. And if you look at the seats and votes from one party's perspective, if the proportion is greater than 1 then that party was advantaged. As an example, in Massachusetts the PPR for Democrats in 2018 was (100% seats) / (66% votes) = 1.52. In the appendix, I introduce another measure of political fairness called the "efficiency gap." As I continue to work on this paper, I will look at how the "efficiency gap" compares with the PPR outcome that I'm currently using.

In my analysis, I am looking for differences in states where redistricting was controlled by Republicans, by Democrats, and by Others.

Who draws the districts can and often does have an impact on the political fairness of the districts. There is no need for statistical analysis to convince you of this statement, nor do I need to argue the fact that gerrymandering is an aspect of politics. But there are other factors aside from who drew the lines, that can impact the "fairness" of the resulting districts.

Gerrymandering writ-large, is accomplished by either "packing" (drawing district lines to pack one group into a small number of districts, which gives the other group an advantage in the remaining districts), or by "cracking" (drawing district lines that distribute a group of voters among numerous districts, thus diluting that group's vote).

Though we know that those who are drawing political boundaries can pack a district, voters also pack districts when they choose where they will live. And in a political environment where Democratic correlates with urban and Republican correlates with rural, this natural clumping can have an impact on proportional-political-representation. In the appendix I discuss some measures of clumping that I analyzed, and others that I will look at further. For this version of my analysis though, I have not included clumping as a predictor in my model fitting, and fully realize that any model that doesn't include clumping is incomplete. I will include clumping as I continue to refine my model.

The other predictor variable that I did include in my model fitting, aside from who did the redistricting, is the number of districts in the state. Clearly if there are only 2 congressional seats, proportional-political-representation is unlikely. And at the other end of the spectrum, such as the 61 seats the Democratically controlled legislature in California had to work with when they drew district boundaries in 2001, large numbers of seats give numerous opportunities for cracking and packing and could lead to more politically skewed districts.

## METHODS

In building the data frames for my analysis I've worked from: 1) election results from the Harvard Dataverse website, 2) An analysis of who was responsible for redistricting in each state, done by the Brennan Center for Justice and 3) population and land area data for 2000 and 2010, downloaded from Census.gov.

Aside from some data cleaning and data frame joining, the major data wrangling issue that I dealt with was imputing vote numbers for candidates who ran unopposed. The vote tallies for unopposed candidates are not reflective of the number of Democratic and Republican voters in a district. In it's extreme, there are Florida's vote records where the winning candidate is noted as having just 1 vote. Since my outcome measure was comparing % of seats (an accurate number) I needed a method get a more accurate total of votes within a state.

For each unopposed election, I calculated the maximum number of votes received by candidates of that party in the other districts in that state and then assigned that maximum vote total to the unopposed candidate. And for the party that didn't field a candidate, I created a candidate named "Imputed" and assigned them the minimum vote total from the other candidates of that party. (When I do my analysis using the "efficiency gap" and wasted votes, I will revisit this method of imputing uncontested elections.)

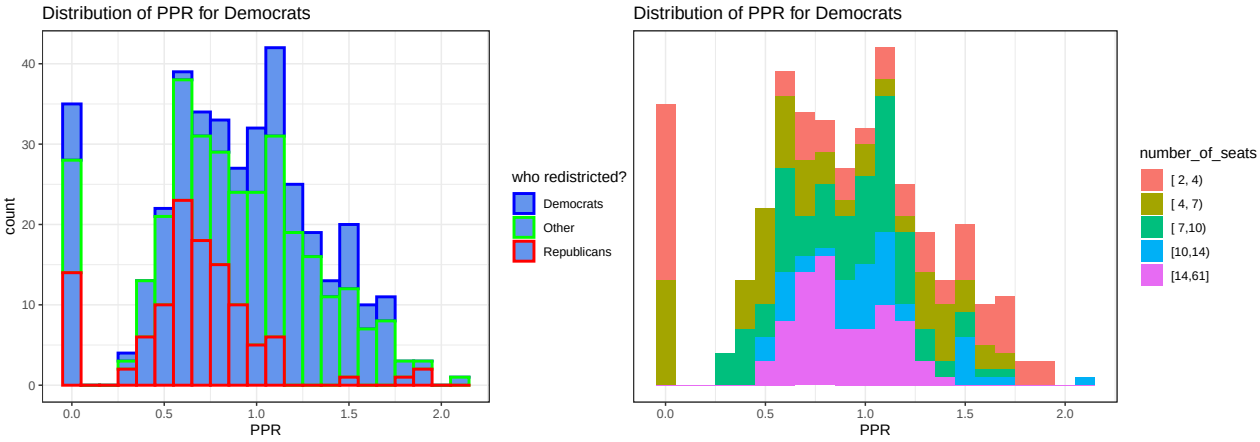
Finally, for each county within each state I calculated the population density for each year. Those values were used in my attempt to build a "clumping" metric (see the appendix).

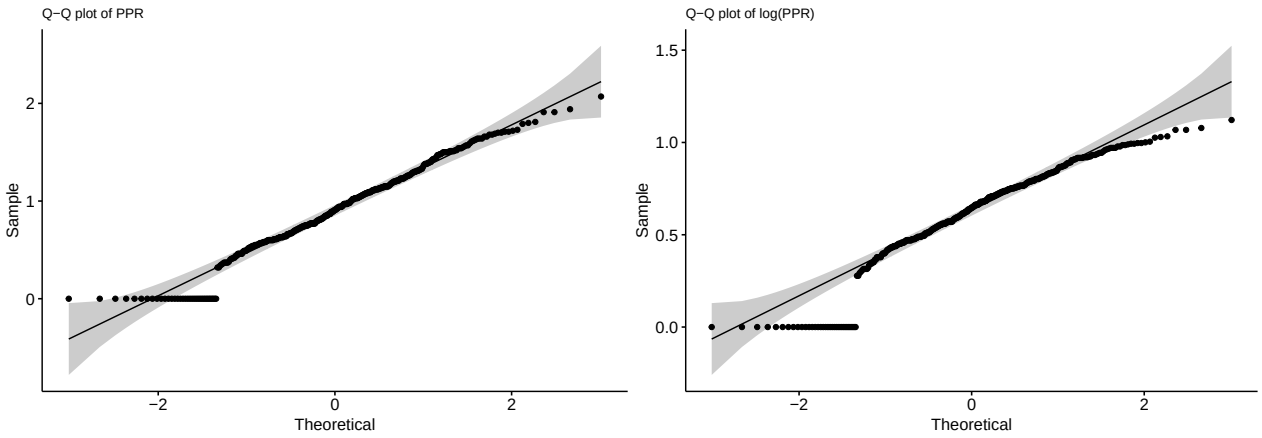
When done, I had a data.frame for both Democratic results and Republican results. And I did my data analysis from the Democrat's perspective, using the Democratic Results data frame. A subset of that data frame is shown below.

| year | state | party | seats | who | seat_percentage | vote_percentage | PPR |
|---|---|---|---|---|---|---|---|
| 2002 | AL | democrat | 7 | D | 28.57 | 45.50 | 0.63 |
| 2002 | AR | democrat | 4 | O | 75.00 | 57.06 | 1.31 |
| 2002 | AZ | democrat | 8 | O | 25.00 | 40.91 | 0.61 |
| 2002 | CA | democrat | 53 | D | 62.26 | 53.55 | 1.16 |
| 2002 | CO | democrat | 7 | O | 42.86 | 50.67 | 0.85 |

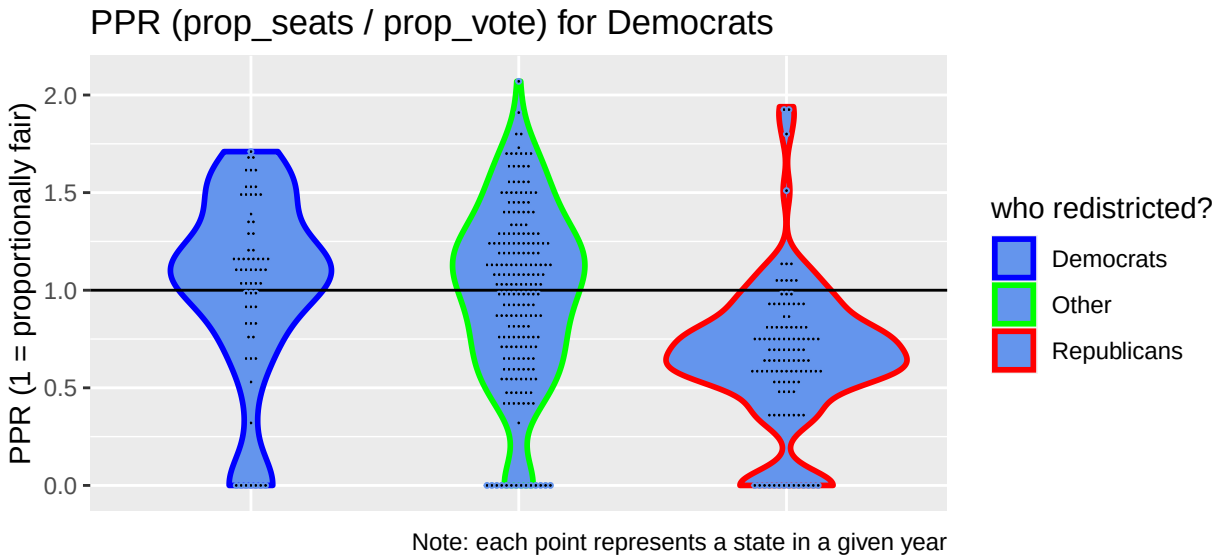
## EDA

**What's the distribution of PPR look like?:** Visually the distribution of PPR looks normal, with the exception of the 35 elections where the Democrats won zero of the congressional seats in that state (regardless of their vote percentage). And when subseting the PPR's by quantile, it's clear that with more congressional seats, the results are less varied and move toward a mean PPR of 1 (a politically proportional result). Finally a Q-Qplot of PPR (and log(PPR)) and a Shapiro-Wilk normality test (p=.00001) show that PPR is normally distributed (noting the exception of the 35 PPR=0 results). Given this, I am assuming normality of my outcome variable for my regression fitting, and am using PPR instead of log(PPR).

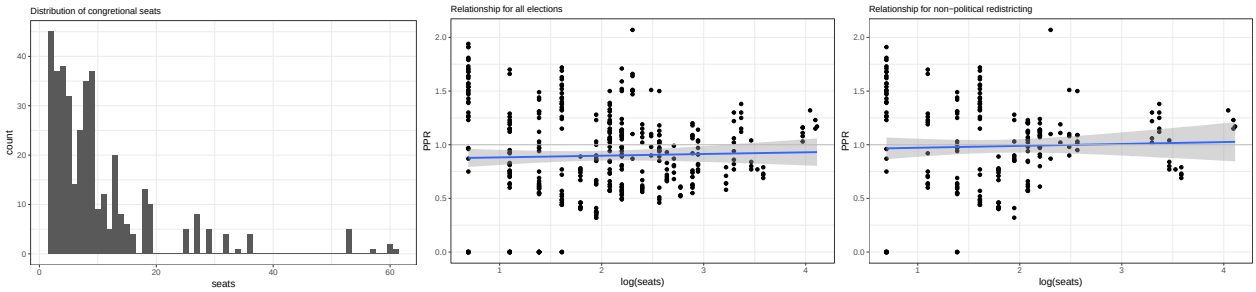





**Who vs. PPR (seat/vote)?:** Just looking at the comparison between the three groups who redistricted, it's visually clear that the Democrats were advantaged when Democratic controlled states redistricted, and Democrats where disadvantaged when Republican controlled states redistricted. This is validated by the specific and historical examples that we've all seen of political gerrymandering. And a simple t-test of the PPR means for Democratically vs. Republican controlled redistricting shows a significant difference (p=.00000003).




**Size of district vs. political proportionality (seat/vote):** In looking at the relationship between the size of the district and PPR, there is no visual pattern in the aggregate, and the correlation coefficient is only r=.07. In my analysis, I've used the log of the number of congressional seats to spread out the data, given that there is a strong right skew to the number of seats (as seen in the first graph below). Given that there is a demonstrable difference in the PPR based on who did the redistricting (see above), I also looked at the relationship between the size of the district and the PPR just for the states where neither political party was responsible for the redistricting. Here there was also no clear visual pattern, nor a significant correlation (r=.05).

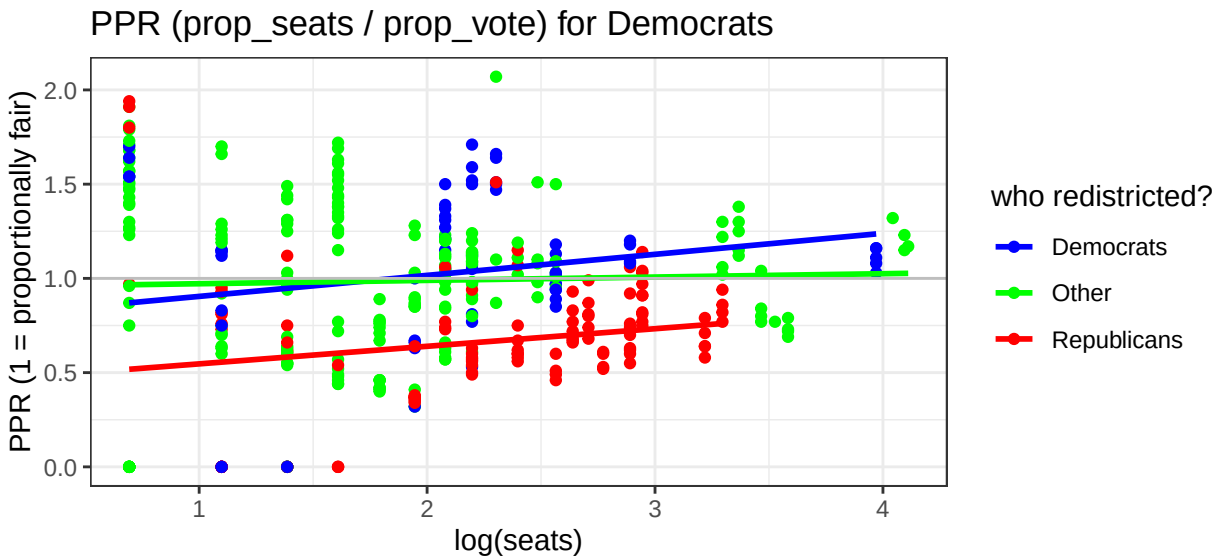



## MODEL FITTING / RESULTS

Wanting to look at how the relationship between the number of congressional seats and who did the re-districting effects the PPR, I built a model using the interaction between the continuous predictor variable (log(seats) and the categorical predictor variable (who).

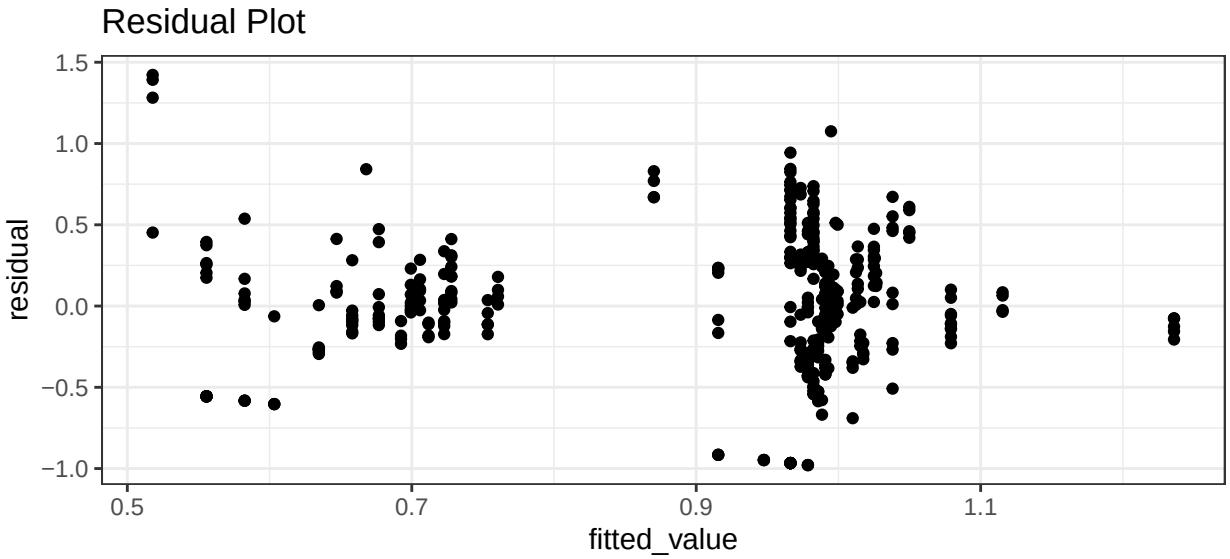
PPR ~ who + who:log(seats) -1

The model is visually represented by the ggplot below.




The model's residual plot shows no pattern, and it has a high Adjusted R-squared value (.82). But there is a great deal of variability in the residuals. In fact the standard deviation of the residuals (.42) is only slightly lower than the standard deviation of the PPR variable (.45) which we are trying to predict.

```
##
## Call:
## lm(formula = PPR ~ who + who:log(seats) - 1, data = DemYearlyResults)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.97854 -0.23161  0.01125  0.25601  1.42217
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## whoD             0.79288    0.14875   5.330 1.68e-07 ***
## whoO             0.95397    0.07092  13.452  < 2e-16 ***
## whoR             0.45317    0.12848   3.527 0.000472 ***
## whoD:log(seats)  0.11160    0.06588   1.694 0.091073 .
## whoO:log(seats)  0.01772    0.03600   0.492 0.622902
## whoR:log(seats)  0.09328    0.05520   1.690 0.091840 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4246 on 380 degrees of freedom
## Multiple R-squared:  0.8241, Adjusted R-squared:  0.8214
## F-statistic: 296.8 on 6 and 380 DF,  p-value: < 2.2e-16
```




In looking at the coefficients for the model, we see that all of the slope coefficients are positive and that they all have confidence intervals that cross zero. So for each of the three groups controlling redistricting (D, O, R), we can't say that there is a relationship between PPR and who redistricts when we include the number of seats being redistricted.

I also did an ANOVA to compare the regression model that only used the log(seats) as a predictor and the model that used both log(seats) and who redrew the districts. This ANOVA showed a significant difference in these two models ($p<.001$), and I believe that I am correctly interpreting this result when I conclude that adding the "who" grouping variable to the model results in a model that is significantly better at predicting the PPR. But, that improved model is not able to discern a difference in PPR based on who controls redistricting.

```
##                       2.5 %     97.5 %
## whoD             0.50041072 1.08534887
## whoO             0.81453499 1.09341222
## whoR             0.20054455 0.70580040
## whoD:log(seats) -0.01792906 0.24112420
## whoO:log(seats) -0.05306916 0.08850498
## whoR:log(seats) -0.01524443 0.20181386
```

```
## Analysis of Variance Table
##
## Model 1: PPR ~ who + who:log(seats) - 1
## Model 2: PPR ~ log(seats)
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1    380 68.497
## 2    384 78.629 -4   -10.133 14.053 1.053e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

## DISCUSSION

I've simplified the complicated question of whether or not congressional districts are fairly drawn by focusing on the ratio of the proportion of seats won to the proportion of votes cast (PPR). This simplified metric is informative for specific instance of unfairness. A PPR=.47 for the Democrats in N.C. or a PPR=1.5 for Democrats in MA, correctly describe the perceived unfairness of the political representation in those states. And the metric corroborates what we would expect when comparing politically drawn congressional districts versus non-politically drawn districts. Mainly that politically drawn districts tend to be further from "proportionally fair" (PPR=1), a result that is seen in the significant difference between the PPR scores for Democratically versus Republican drawn districts. Further the metric behaves as we would expect when the number of electoral districts being drawn increases - the larger the number of seats, the less variability in the PPR.

But any effect the metric is able to highlight disappears when the analysis is expanded to include the number of seats that a given state is redistricting. It's not clear if this is due to usefulness of the PPR metric, or the diluting and overlapping influences of the number of seats being redrawn and who's drawing the district.

As I work on this question further, my hope is to get a clearer idea of how the choice of who does the redistricting influences the political fairness of the results. To this end I plan to add in a measure of how the populations in state "clump" and do my analysis using the "efficiency gap" as a measure of political fairness.

## BIBLIOGRAPHY and APPENDIX

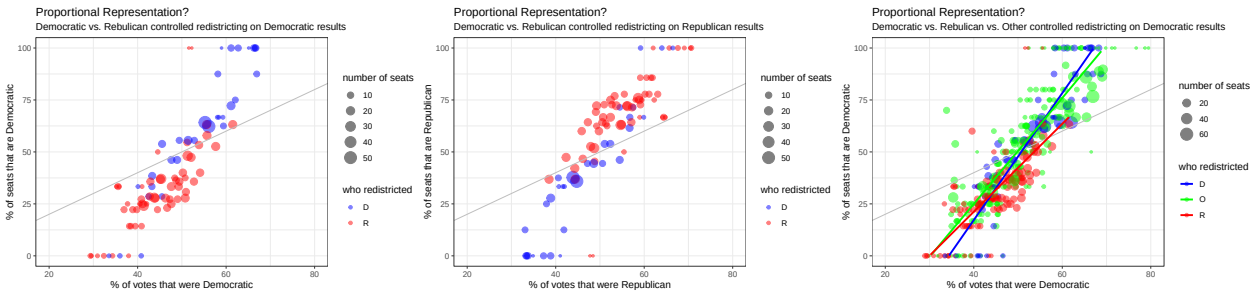
### Metrics

### (I) Who to include?

In looking at vote results I had to decide which parties to include. I combined Minnesota's "democratic-farmer-labor" party with the "Democrat" party. I looked at the entirety of listed parties, to see if there were political parties aside from the Democrats and the Republicans that were significant either in the number of candidates or the number of votes. The only party that was worthy of consideration aside from the two major parties was the Libertarian party. The Libertarians ran candidates in 2,018 elections between 1990 and 2018. But never got higher than 31% of the vote and their mean vote percentage was 3.4% (with a distribution of results that was heavily skewed right with "skewness" = 3.4). As such I have not included the "libertarian" party as a major party, and have confined my analysis to Democrats and Republicans.

### (II) Is it politically fair?

**Vote_percentage vs seat_percentage:** Originally, I looked at representing "politically fair" where the outcome variable was the proportion of house seats won and the predictor was the proportion of votes won. If an election was fair then these two proportions would be equal. State-years where the vote_proportion and the seat_proportion fell on the y=x line (grey line in the graphs below), would be proportionally equal/fair. In the graphs I've used color to note who did the redistricting and how many seats were decided (with the notion that for a state with a small number of seats, it would not be surprising to have a big difference in vote and seat proportions, e.g. NH-2016 democrats had 46% of the vote but 0% of the two congressional seats). Ultimately I decided that this representation was harder to visualize. And the clumping variable was not a predictor variable for the proportion of seats won, but rather a possible predictor variable for political proportionality. As such I decided to center my analysis around a measure of political fairness (seat_proportion / vote_proportion).

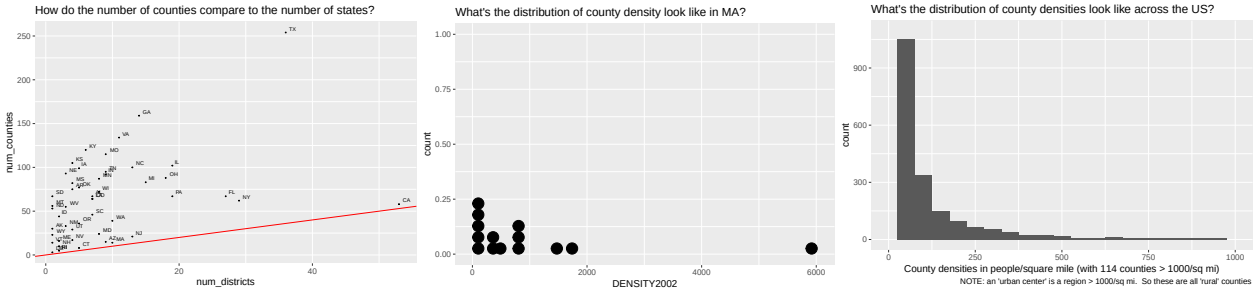



**The efficiency gap & wasted votes** When "cracking" or "packing," voters are moved into or out of districts to dilute the power of their vote. The efficiency gap is a mathematical calculation that "counts the number of votes each party wastes in an election to determine whether either party enoyed a systematic advantice in turning votes into seats." ("How the Efficiency Gap Works," Petry, 2015, http://bettergov.nc.lwvnet.org/files/how_the_efficiency_gap_standard_works.pdf). I have not yet taken the time to write the code to calculate the efficiency gap and then explore how illustrative it is as a measure of political fairness.

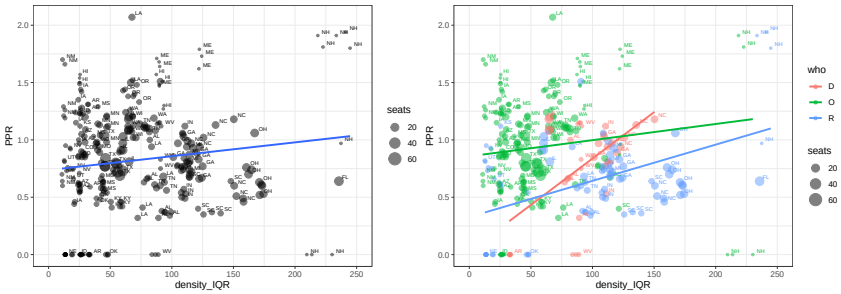
### (III) How to measure clumping.

I'm still trying to wrap my head around how to appropriately measure "clumping" so that the measurement will have meaning in the context of packing and cracking. I investigated using the IQR and standard deviation of county densities within a state as a measure of clumping. Neither of these felt informative, so I did not include a clumping variable in my current analysis. As I continue to expand on the analysis of political fairness, I will need to incorporate a measure of clumping.

### Exploration of the distribution of county densities, to help structure a "clumping" metric .

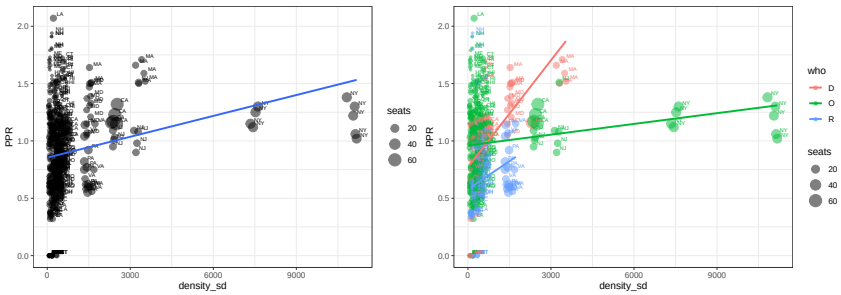



### Exploring IQR of county level population density as a predictor variable .




```
##
## Call:
## lm(formula = PPR ~ density_IQR, data = DemYearlyResults)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.89770 -0.27926  0.01272  0.30302  1.22094
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 8.288e-01  2.567e-02  32.289  < 2e-16 ***
## density_IQR 2.993e-04  5.608e-05   5.338 1.61e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4368 on 384 degrees of freedom
## Multiple R-squared:  0.06908,    Adjusted R-squared:  0.06665
## F-statistic: 28.49 on 1 and 384 DF,  p-value: 1.612e-07
```

### Exploring stDev of county level population density as a predictor variable .



**Comparing state level population density, with county level population density, with Census defined "Urban centers"** This will be my next clumping variable attempt.

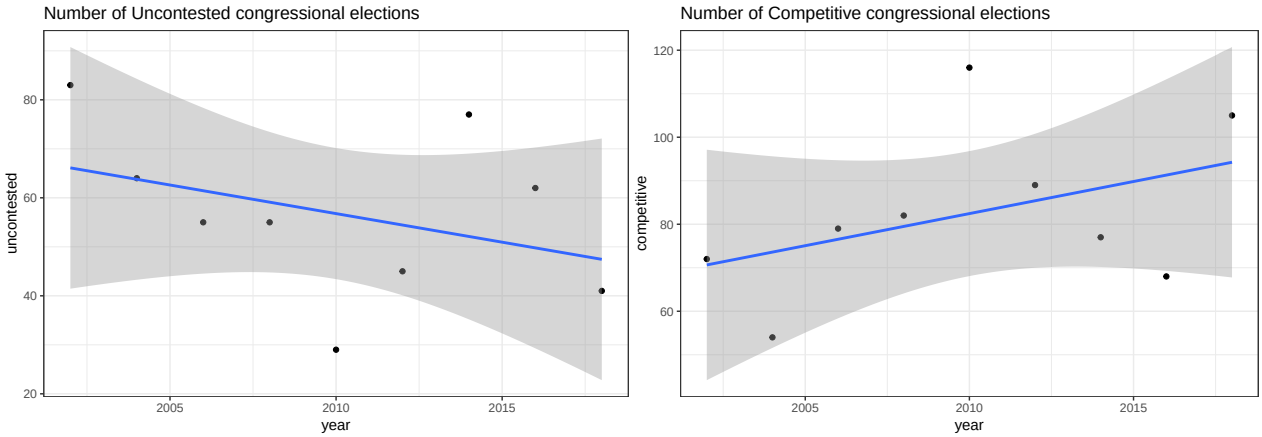
### (IV) Who is making the redistricting decisions

These are the sources I used to create the group variable "who" for each election decade (2002-2010 & 2012-2018). The "who" variable attributes the responsibility for the redistricting to either D, R, or O (Democratically controlled state legislatures, Republican controlled state legislatures, or Other). I've defined "Other" as not Democratically controlled nor Republican controlled. "Other" includes court ordered redistricting, split-party redistricting, and non-political commissions.

https://www.ncsl.org/documents/statevote/2010_Legis_and_State_post.pdf
https://www.brennancenter.org/our-work/research-reports/redistricting-and-congressional-control-first-look
https://www.brennancenter.org/sites/default/files/2019-08/Report_CGR-2010-edition.pdf

### (V) Looking for a pattern in the number of uncontested and competitive congretional races by year

Though there is a great deal that's been written about how political gerrymandering is increasing the number of uncontested elections, and decreasing the number of competitive elections, my analysis doesn't bear this out. My quick look is looking at all elections and is not looking at the grouping variable of who is doing the redistricting.




```
##
## Call:
## lm(formula = uncontested ~ year, data = YearlySummary)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -27.778  -6.444  -4.111  12.222  24.889
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 2401.778   2201.245   1.091    0.311
## year          -1.167      1.095  -1.065    0.322
##
## Residual standard error: 16.97 on 7 degrees of freedom
## Multiple R-squared:  0.1395, Adjusted R-squared:  0.01658
## F-statistic: 1.135 on 1 and 7 DF,  p-value: 0.3221
```

```
##
## Call:
## lm(formula = competitive ~ year, data = YearlySummary)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -23.294 -11.344   2.456   3.606  33.556
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -2882.306   2364.110  -1.219    0.262
```

```
## year           1.475      1.176   1.254    0.250
##
## Residual standard error: 18.22 on 7 degrees of freedom
## Multiple R-squared:  0.1835, Adjusted R-squared:  0.0668
## F-statistic: 1.573 on 1 and 7 DF,  p-value: 0.2501
```

## Citations (messy - sorry)

(need to figure out how to use LaTex or ?? to display this citation) @incollection{DVN/IG0UN2/ELBYL3_2017, author = {MIT Election Data and Science Lab}, publisher = {Harvard Dataverse}, title = {1976-2018-house2.tab}, booktitle = {U.S. House 1976–2018}, UNF = {UNF:6:8iuXTceVO5a7EpOwUD5UPw==}, year = {2017}, version = {V7}, doi = {10.7910/DVN/IG0UN2/ELBYL3}, url = {https://doi.org/10.7910/DVN/IG0UN2/ELBYL3} }

https://www.census.gov/library/publications/2011/compendia/usa-counties-2011.html#POP

to get .xls for population and .xls for area:

https://www.google.com/url?sa=t&rct=j&q=&esrc=s&source=web&cd=&ved=2ahUKEwjHkPmp4bftAhV5FlkFHbDHBvv
url=https%3A%2F%2Fwww2.census.gov%2Flibrary%2Fpublications%2F2011%2Fcompendia%2Fusa-counties%2Fexcel%2FMastdata.xls&usg=AOvVaw1ahgb3GWupqOb1UHYbZlMw