Analysis of US elections

Marco Galliani

May 2024

1 Introduction

The story of US elections has always been the story of the two candidates running for the election, of their personal history, of their beliefs, of their way of thinking. And in the end this is completely reasonable: americans are deciding between one of the two to run the country! The fact that is often undertold is that in a large portion of the country elections are already decided before the vote.

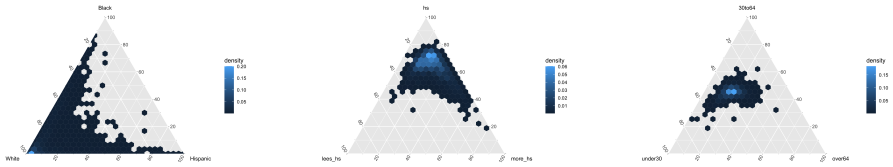
In this work I will try to point out how it is possible to explain a large portion of the uncertainty connected to the electoral results just by looking at demographics data. By doing so I'll focus on answering two main questions:

- Which population segments are more polarized? What are the demographic groups that have the strongest effect on the election outcome?
- Which states consistently vote Democratic and which vote Republican? Is it possible to spot potential swing-states?

The remaining of this report is structured as follows: (1) description of the data and of the basic preprocessing steps, (2) description of the analytical procedure used in the analysis and (3) results and conclusions

2 Data

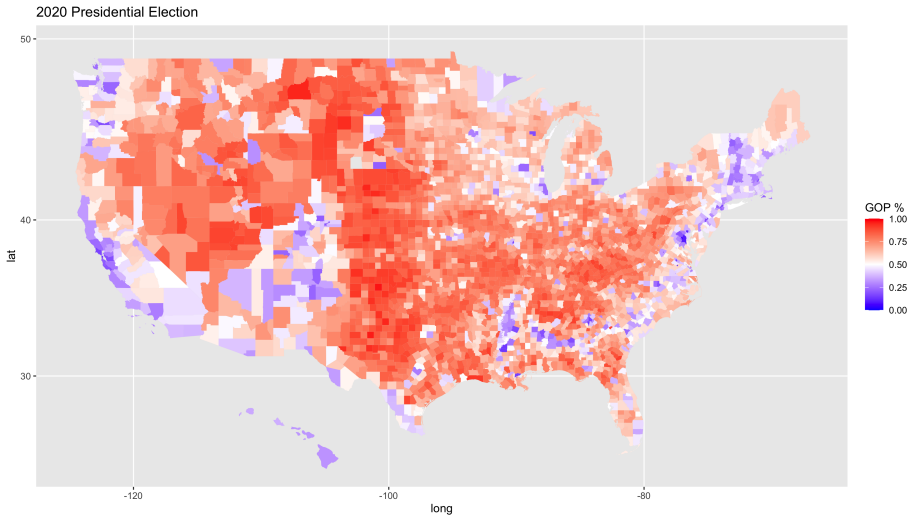
The data used in this analysis come from the "MIT Election Data + Science Lab"[1] and constists in a collection of public available demographics data at county-level. Indeed, the first thing to consider about this data is that the statistical units throughtout the course of the whole analysis will be counties. Secondly, almost all the variables in the dataset will be compositions related to the counties. For instance, let's consider the ethnical composition: in the dataset there are multiple variables containing the percentage of each ethnic group in each county, clearly these variables will be related and a proper approach should be taken while analyzing them.



(a) Ethnicity (b) Education (c) Age

Figure 1: Compositions are well-represented by tern-plots. In this graph is reported the density of the compositions

The dataset already contains the electoral results of the presidential elections of 2012 and 2016. In addition I joined the results from 2020.



[1] data are available here

Moreover, to model a possible spatial autocorrelation of the data I included the adjacency matrix of US states, available in the package `spdep`.

Summing up, the data used in the analysis contains:

- Compositional variables: ethnicity, age, gender, education
- Summary variables: median income, total number of electors, population
- Electoral results: presidential elections 2012, 2016, 2020
- Spatial adjacency matrix of states

3 Analysis

As written, the goal is to understand the relationship between electoral results and the population segments. This task can be framed as a regression problem targeting the percentage obtained by one of the two parties. Before getting to regression a preliminary analysis based on depths and permutaional tests is gonna be performed to inspect the distributions and the significance of the variables for the problem at hand.

3.1 Preliminary Analysis

3.1.1 The compositional analysis approach

Before getting to the actual analysis I'll briefly explain how compositional data are gonna be handled, following the classical techniques of compositional data analysis. Let's start by defining what a composition is, namely a set of variables which sums up to a constant value. Since the value to which the variables sum up is constant we may always normalize the variables and set the sum to 1. The space of closed compositions is called Aitchinson simplex:

$$\mathbf{S}^D = \{\mathbf{x} = (x)_{i=1}^D : x_i \geq 0, \sum_i x_i = 1\} \tag{1}$$

As proved by Aitchison, this space is actually provided with sum and multiplication to a scalar operations and also with a scalar product, making it an euclidean space of dimension $D-1$. For this reason it is also possible to map the Aitchison simplex to the one and only $(D-1)$-dimensional euclidean space, $\mathbb{R}^{D-1}$, by means of the Isometric Log Ratio transformation (ilr). Being an isometry angles and distances are preserved and we're getting a more suitable space-setting for classical statistical analysis.

The go-to procedure whenever dealing with compositional data will be to map the data from the Aitchison simplex to an euclidean space, perform the analysis by using proper methods for multivariate numerical data and then get back to the Aitchison simplex to interpret the results of the analysis.

3.1.2 Depth measures

First of all I computed Tukey depths to explore the distributions of the main groups of variables. The fact that the exploration is perfomed by groups is a limitation due to the feasibility of computations (depth measures are computationally inefficient in higher dimensions), but the grouping of the variables is actually pretty straightforward for the data I'm working with, as compositional data are essentially multivariate and each composition is associated to multivariate data. I used the computed depths to get three different visualizations of the variables distributions:

1. the usual bagplot, plotting the D-dimensional compositions in the $(D-1)$-dimensional euclidean space to look for possible outliers

2. a *depth heatmap ternplot*, filling the points in the ternary scatterplot with the depth computed in the lower dimensional euclidean space

3. a *depth map*, plotting the computed depths in the original dimension of our data, that is to say the spatial domain

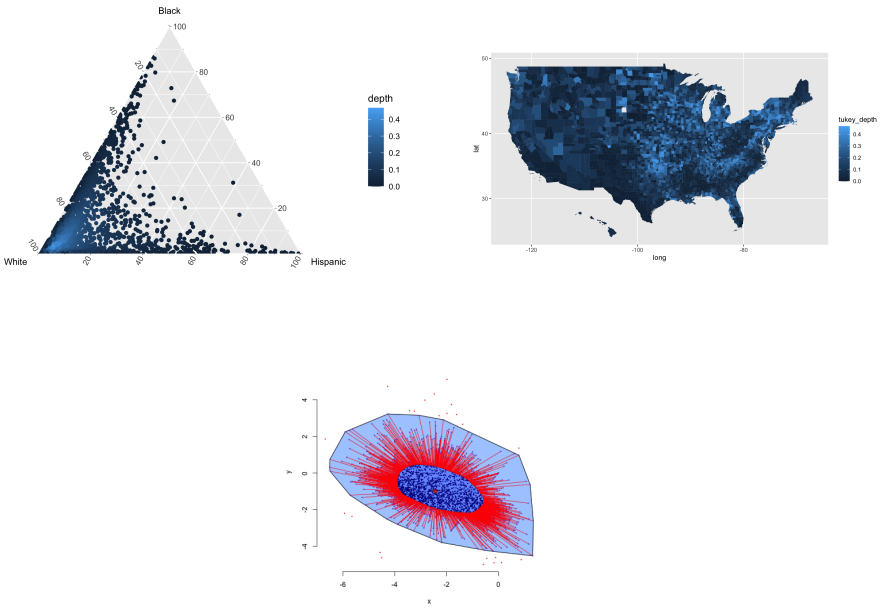


Figure 2: **Depth measures for the ethnical composition**. We observe few outliers w.r.t. to the ethnical composition. Moreover from the depth map we observe some spatial autocorrelation, observing that the counties of the mid-west are actually deeper than the others. It is important to recall that we're not performing any weigthing by the population to compute depth measures: extending the reasoning to get interpretations at national level could be misleading. For instance, it is true that the most representative counties are the prevalent white ones, but the overall USA population is not as prevalently white

Figure 3: **Depth measures for the educational composition**. From the tern plot we see that the deepest observations w.r.t. the age composition are the ones with a prevalent component of people with only a high school diploma.

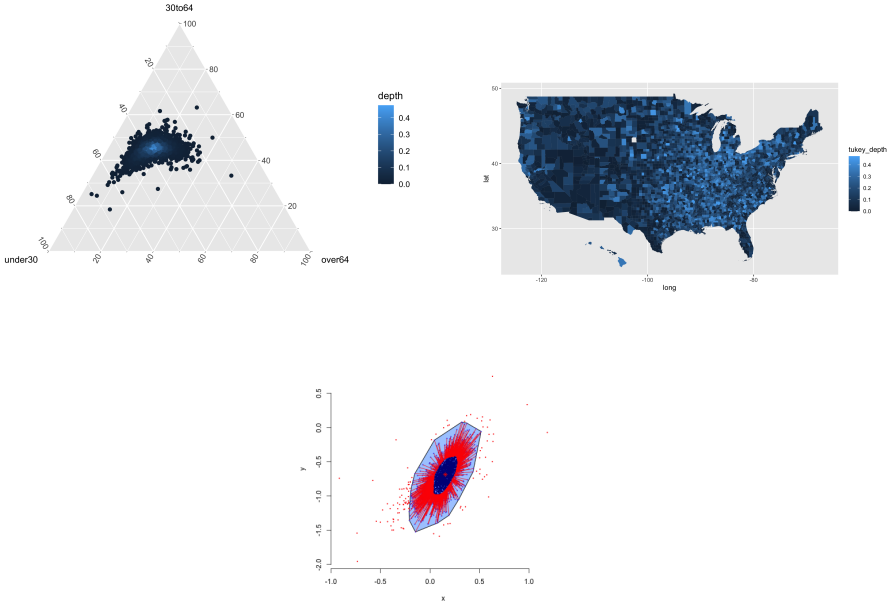


Figure 4: **Depth measures for the age composition**. The median county w.r.t. age composition has equal proportion of under 30 and of people between 30 and 64 and a lower proportion of over 64. Some regions, like the midwest, seems to be more deeper than others

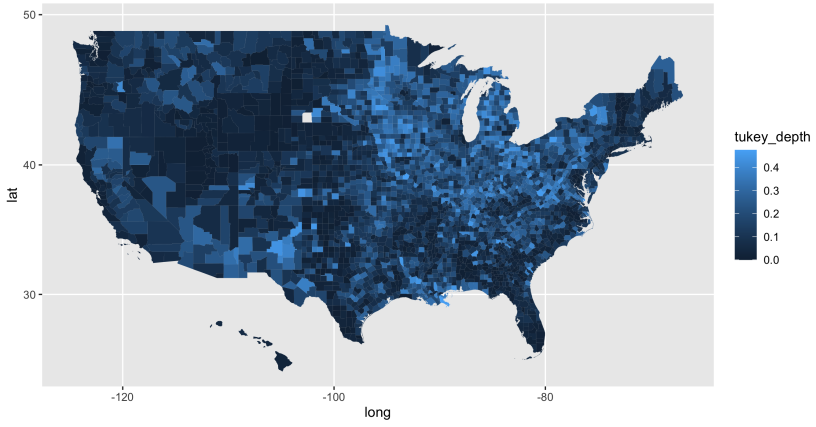


Figure 5: **Depth map for the 2020 electoral results**. Even electoral results can be interpreted as a composition of democratic, republican and third-party voters. In this case too we observe some spatial autocorrelation

3.1.3 Permutational tests

Moving to hypothesis testing I employed t-tests to compare counties having a Democratic majority to counties having a Republican majority. To do so I followed this pipeline: firstly I checked if the two populations were generated by the same distribution by means of DD-plots, then I computed permutational t-tests, using the distance between depth medians in $\mathcal{L}_\infty$ as a test statistic:

$$T(\mathbf{m}_d, \mathbf{m}_r) = \|\mathbf{m}_d - \mathbf{m}_r\|_\infty \tag{2}$$

Moreover, I computed bootstrap 5% confidence intervals for the depth medians of the considered compositions in the lower dimensional space. The resulting confidence intervals are a join of Bonferroni's simultaneous confidence intervals along the dimensions of the mapped space and for this reason are rectangular in the mapped space. However, by mapping them back to the Aitchison simplex we lose the straightforward interpretation of rectangular intervals, as we cannot interpret them as confidence intervals for each dimension of the simplex.

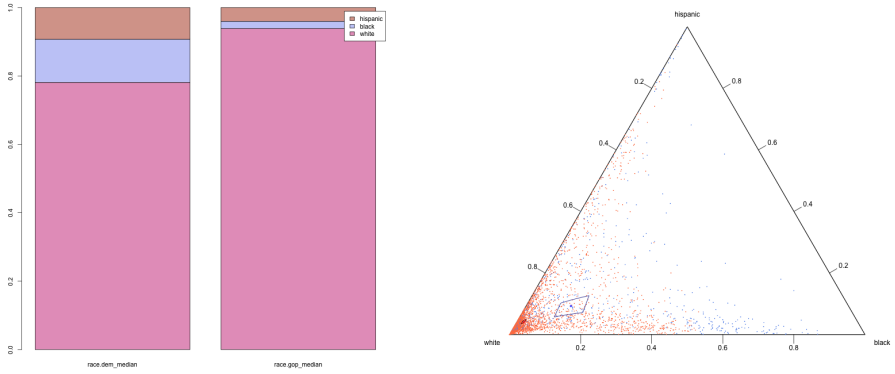


Figure 6: **Ethnical composition**. The computed p-value for the test on the difference of the depth medians is 0, thus the difference is significant. By plotting the medians for the two populations we observe that republican counties are characterized by a stronger prevalance of white people and a lower presence of both black and hispanic people. By looking at the ternplot we observe that democratic counties have a higher variance and a larger confidence interval for the median

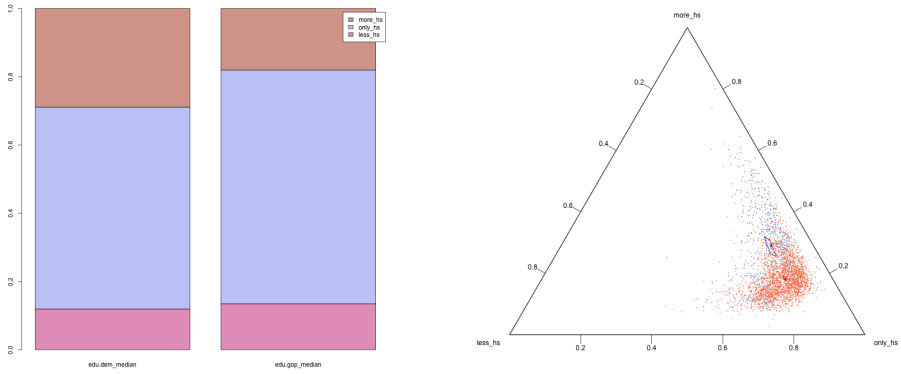


Figure 7: **Educational composition**. Even in this case the p-value of the t-test is 0. The main difference that we note from the plots is that the presence of persons with a college education is higher in democratic counties while republican counties have more people with only a high school diploma

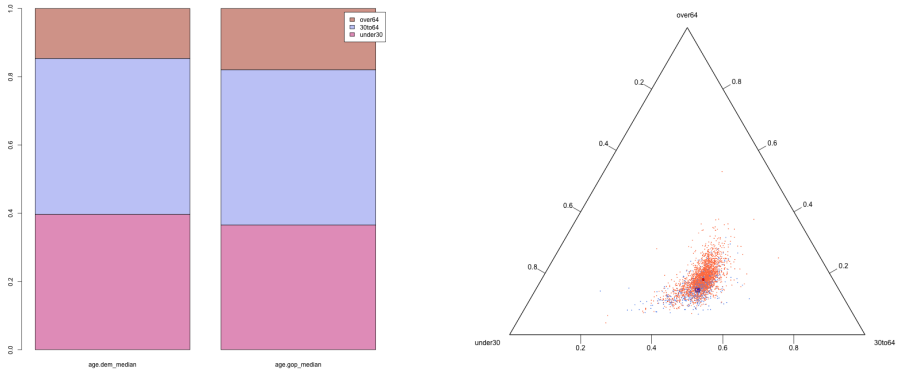


Figure 8: **Age composition**. Once again we got a significant p-value equal to 0. Republican counties appear to have more people over 64 and less under 30.

3.2 Regression

3.2.1 Building the target

The first issue going into regression is which should be the target variable of the analysis. Indeed, in the dataset I have electoral results relative to the 2012, 2016 and 2020 elections. If we use just one year as a target, the risk is to learn the relationship between the candidates running for elections in that year and the population segments. Aiming for a more general result, that goes beyond the result of a single election the first idea was to average the three electoral results I have in my dataset. Later on, I settled on a slightly more refined approach (that actually gives almost the same results). What I did was to compute a PCA for the three variables representing the electoral results: as the variables are highly correlated it is possible to reduce them to just one component, that accounts for 97% of the original variability of the three. The final target of the regression will be the computed component, that is nothing more than a weighted average of the electoral results of the past three elections. I interpreted this average as an indicator of how consistently a county vote republican or democratic.

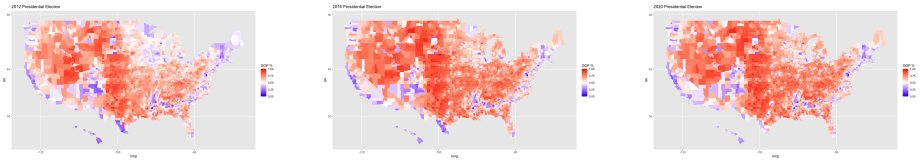


Figure 9: The electoral results show high correlation between the years

A few notes:

- For simplicity I did not consider the votes obtained by candidates that are not from the two main parties. This choice hasn't a big impact on the analysis since the percentages obtained by other candidates are negligible. However, some fixes had to be done: adjusting the percentages obtained by the democrats and by the republicans and excluding third-parties voters from the count of the electors
- The resulting target is itself a proportion. Indeed, by normalizing the first component of the PCA by the loadings we're actually computing a weighted average of a proportion that is still a proportion.
- Being the target a proportion a proper regression model should be used: I ended up with a beta regression model, more on that will be discussed later.

3.2.2 Handling spatial autocorrelation

As seen in the previous analysis we observe spatial autocorrelation for some variables in the dataset and in the target. Thus, the issue should be handled in

some way. For this reason I decided to add a Markov Random Field Smoother, available in the package `mgcv`. This smoother can be represented by a classical mixed effect model: let $\mathbf{Y} \in \mathbb{R}^n$ a vector of responses, $\mathbb{X} \in \mathbb{R}^{n \times p}$ a matrix storing the n covariates, $\epsilon \sim \mathcal{N}(\mathbf{0}, \sigma^2 I_p)$ the random errors, $\beta \in \mathbb{R}^p$ the vector of coefficients. The model is:

$$\mathbf{Y} = \mathbb{X}\beta + \mathbb{Z}\mathbf{u} + \epsilon \tag{3}$$

Where $\mathbb{Z} \in \mathbb{R}^{n \times m}$ with $m \leq n$ design matrix for the random effects that are modeled as $\mathbf{u} \sim \mathcal{N}(\mathbf{0}, \sigma_u^2 \Sigma)$.

The spatial modeling is done by parametrising the covariance matrix of the random effects Σ:

$$\Sigma^{-1} = \text{diag}(w_i^r) - W \tag{4}$$

where W is a matrix of spatial weights and w_i^r are the row sums of W. The spatial weigths used in this analysis consider just the adjacency of states, but other choices could have been made while specifying W, taking into account distances for instance. Moreover, the modeling of spatial autocorrelation could have been expanded to counties, not limiting ourselves to states, but this would have required to fit a random effect for each county, obtaining a really high dimensional and computationally expensive model.

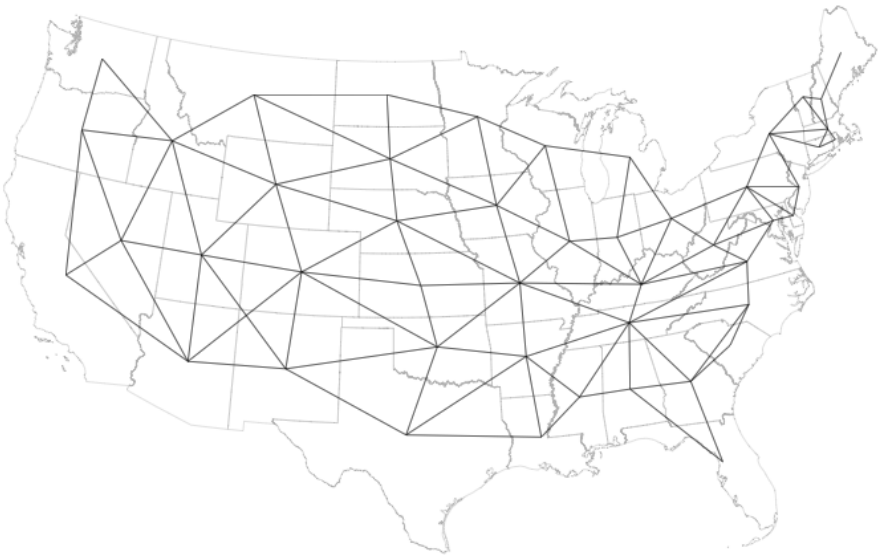


Figure 10: **Adjacency of the states**. The Markov Random Field smoother in `mgcv` comes with the limitation of having an underlying connected graph. For this reason counties belonging to non-connected states had to be excluded

3.2.3 Generalized Additive Model

To model possible non linear associations between the input and the target I used the GAM framework. In particular, the final model uses:

- A Markov Random Field Smoother, using the grouping induced by states to model a portion of the spatial autocorrelation of the data. Another justification for a state-level random effect is that the electoral system of US presidential elections is strongly based on states: each state elects a delegation of big electors and these big electors will all vote Democratic or Republican depending on the majority in the state. Thus, we can hypothize the belonging to a state to have an effect on the vote of the county.
- Thin-plate splines smoothers for compositions. To mantain the interpretability given by additivity, each composition was smoothed separately. Recall that before the smoothing, the usual Isometric Log Transformation has been applied to map the composition to a suitable euclidean space. The statistical significance of using bidimensional-smoothers w.r.t. simple linear terms was assed by using proper Chi-square tests. The compositions included in the model are the ethnical, educational and age ones
- One-dimensional cubic-splines smoothers to model the effect of the median income of the county and of the unemployment rate in the county

Moreover, since the target of the regression is a proportion, a beta regression model was used, specifying as target the weighted percentages obtained by the two parties in the counties. Another possibility would have been to use a binomial regression model, passing the number of voters for each party to estimate the probability parameter of the binomial distribution. The reason why I opted for a beta regression model is that the latter allows me to specify different weigths, while the binomial regression weighted each observation by the number of voters. Weighting by the number of voters is not ideal due to electoral law based on the big electors system, that we have previously discussed speaking about the state-level random effect. The final weights I used were given by:

$$w_i = \frac{n_i}{N_{S_i}} B^{el}_{S_i} \tag{5}$$

where n_i is the population of the i-th county, N_{S_i} the population of the state S_i to which the county i belongs and $B^{el}_{S_i}$ the number of big electors elected by that state.

The final model is:

$$\begin{aligned} \text{logit}(\mathbb{E}(\%\text{rep})) = & \sum_{\text{comp} \in \{\text{ethn, edu, age}\}} f_{tp}(\text{ilr}(\text{comp})) && (6) \\ & + f_{cr}(\text{unempl rate}) + f_{cr}(\text{med income}) && (7) \\ & + f_{\text{mrf}}(\text{state}) && (8) \end{aligned}$$

The model achieves very good performance at explaining the target, with a percentage of deviance explained equal to 89.8%. Let's now look to the estimated effects.

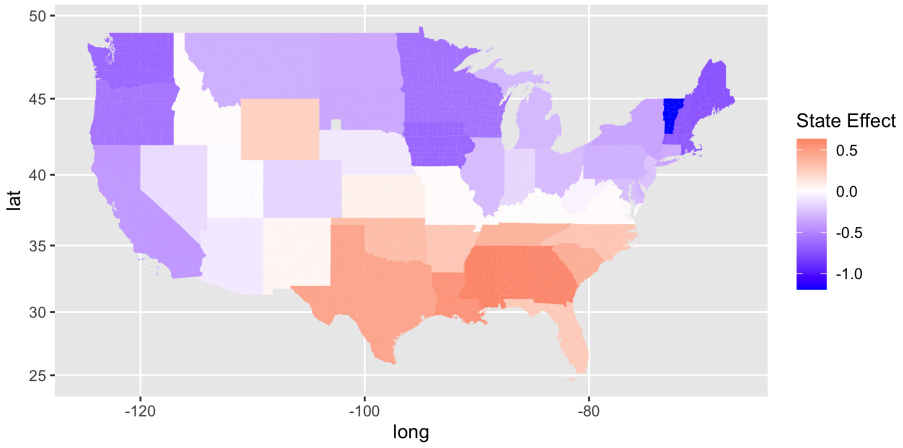


Figure 11: **Estimated state effect**. We observe a strong effect toward the republican party in the states of the south, while on the coasts and especially on the nothern eastern coast we observe a positive effect towards the democratic party

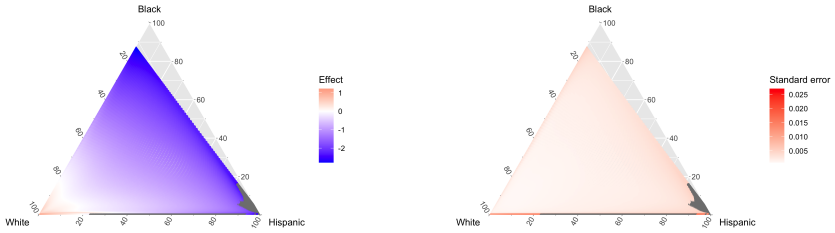


Figure 12: **Estimated ethnical composition effect**. A stronger prevalence of white people in the county is associated to a positive effect on the republican vote, while a more etherogeneous composition relates to a positive effect towards the democratic party

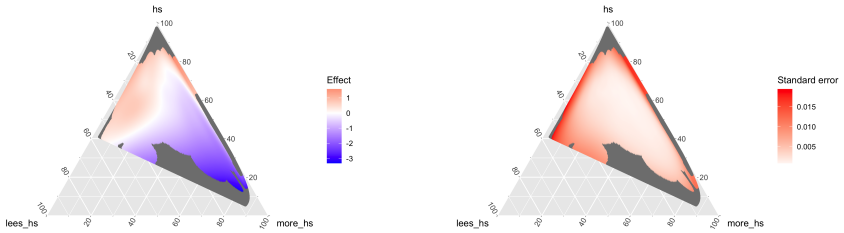


Figure 13: **Estimated educational composition effect**. A stronger presence of people with a college education is associated to a positive effect towards the democratic party, coherently to the t-test in figure 6

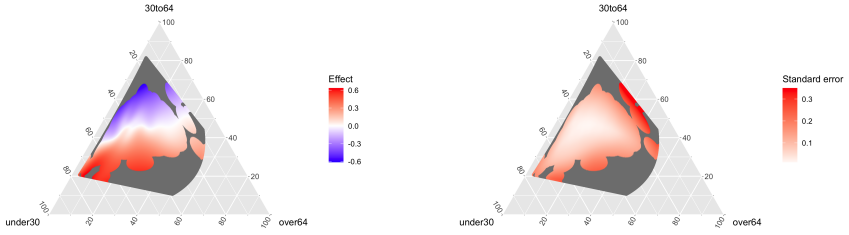


Figure 14: **Estimated age composition effect**. Here the interpretation seems to be in contradiction to what was observed in figure 8, with a positive effect for a stronger presence of under 30 in the county. However the standard error plot has to be taken into account as the variance of the linear predictor grows as we approach to the borders of the domain and the estimation becomes less trustworthy. The main effective direction seems to be the vertical one, as a larger presence of people between 30 and 64 is associated to a positive effect on the Democrats result

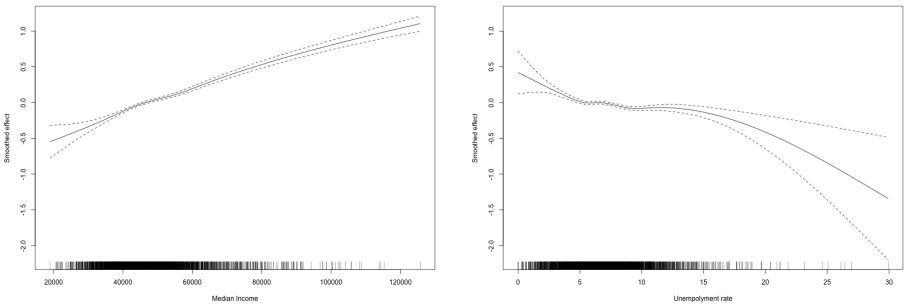


Figure 15: **One dimensional smoothers**. We observe that a higher median income is associated to a positive effect towards the republican party. Even if the effect seems visually linear, the smoothed model turns out to be significant in comparison to its linear alternative in a chi-squared test. The effect of the unemployement rate is reported in the left image

3.2.4 State-level predictions

Finally I used the predictions extracted from the previously described model to get state-level predictions. To do so, I extracted the county-level predictions for the proportions of voters and computed for each state an average weighted by the number of voters in each county, obtaining the state-level prediction. I did the same for the upper and lower bounds of the confidence intervals based on the standard error to get state-level confidence intervals.

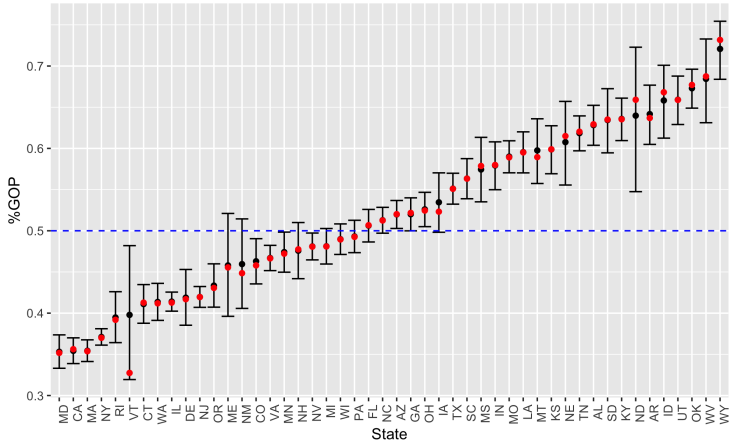


Figure 16: The computed confidence intervals for the weigthed percentage obtained by the republican party at state level. In red the true weigted percentages.

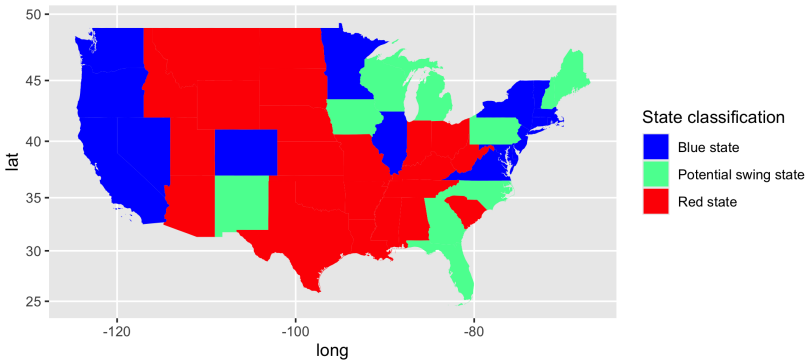
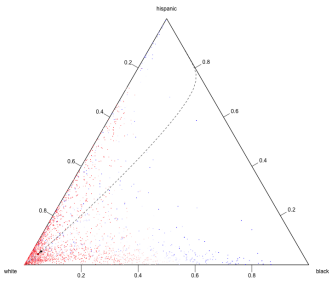


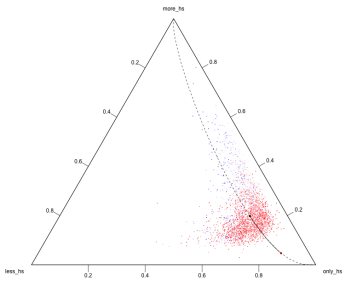
Figure 17: **Map of swing states**. An intersting application of the resulting model is the identification of possible swing states: here I identify a state as a swing state if 0.5 falls inside the confidence interval shown in the previous figure

3.2.5 Comparison to a robust model

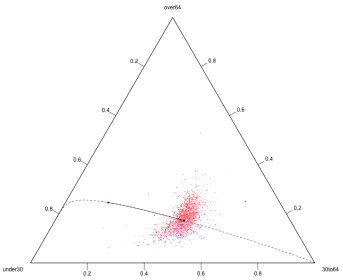
To conclude I fitted a weaker robust model to compare the estimated effects to the ones of the GAM. The robust model is weaker in the sense that the estimating infrastructure I used did not allow to fit the smoothers I used in the GAM. For this reason all the effects are assumed to be linear in the link space and even the markov random field smoother is not present, so no state effect is taken into account. Moreover, beta regression was not supported so I had to switch to binomial regression to estimate the proporition of reoublican voters. The GLM robust algorithm I used is implemented in the `glmrob` function from the `robustbase` package.



(a) Ethnic composition



(b) Educational composition



(c) Age composition

Figure 18: The interpretation of these plots is the following: the segments that ends up with a red marker describe the direction in which we have to perturb a composition to get the highest positive effect towards the percentage obtained by the republican party. We observe that the interpretations of the effects are coherent to what we saw from figures 12, 13

4 Results and Conclusions

Let's now go back to where we started and answer the questions that guided this work:

- *Which population segments are more polarized? What are the demographic groups that have the strongest effect on the election outcome?*
  The most clear observed effects are the ones given by the ethnic and by the educational composition. Having a stronger presence of white people relates to a positive effect towards the result of the republican party, while a more etherogenous ethnical composition relates to a positive effect towards the percentage obtained by the democratic party. Coming to the educational composition, having a larger percentage of college-educated people relates to a positive effect towards a vote for the democrats.

- *Which states consistently vote Democratic and which vote Republican? Is it possible to spot potential swing-states?* Using the proposed model as shown in section 3.2.4 I identified possible swing states with the ones whose confidence interval for the predicted percentage contains 0.5. The identified states are: Maine, New Mexico, New Hampshire, Michigan, Wisconsin, Pennsylvania, Florida, North Carolina, Georgia and Iowa. A succesful campaign should target these states, remembering to keep into account their contribution in terms of big electors.

In conclusion, the performances of the model are pretty satisfying at predicting an indicator of how consistently a county votes for one of the parties. However, there are some limitations and possible improvements with respect to the approach I used:

- The modeling of the temporal relationship: I assumed demographic data to be constant over the time window I considered (from 2012 to 2020) and fitted a model to predict a weighted average of the electoral results. This is not really an ideal approach if we would like to extend the model to time series data where to each year are associated demographic data for that year.

- Investigation of the state effect: we saw that state have an impact on the target, but still we cannot understand what causes this effect, further investigations should be conducted