Statical Learning and Large Data

A.A. 2023/2024

A county study on demographic determinants of 2020 US presidential election results

Students:
Suqi Chen
Giovanni Stivella

May 20, 2024

Abstract

This paper investigates the relationship between demographic variables and voting behavior in the 2020 U.S. presidential election, focusing on county-level data. Our project examines various demographic factors such as ethnic composition, language spoken, family structure, religious affiliation, education, and economic status to uncover patterns in electoral outcomes.

To achieve this, we applied several statistical analysis techniques. For unsupervised learning, we utilized Principal Component Analysis (PCA) and clustering to explore the underlying structures within the data. For supervised learning we employed LASSO regression for feature selection, then refined the result with feature screening, bootstrap and ordinary Least Square.

Contents

1 Introduction

The analysis of US election results, in contrast to those of European elections, has frequently been conducted through the lens of demographics. In a nation as ethnically and religiously diverse as the United States, demographic identities are often deemed at least as significant as ideological affiliations in shaping voting behaviors. Indeed, many even argue that ideology itself is influenced, if not determined, by demographics, a phenomenon less pronounced in other Western democracies where economic divisions have historically dominated the political landscape.

This distinctive political milieu has spurred a lot of studies aimed at identifying relations between demographic characteristics and voting behaviors. Understanding and exploiting these relations have long been the interest of pollsters, politicians and political strategists.

In recent years, these analyses have experienced a surge, driven by two distinct phenomena. Politically, there has been a rise in what is commonly referred to as "identity politics," where belonging to specific groups or categories becomes the primary, if not sole, criterion for political alignment. Methodologically, advancements in statistical techniques have enabled more precise identification of electoral patterns helping the emergence of specialized news outlets such as FiveThirtyEight.

Nevertheless, the perceived ascendance of identity politics does not fully account for the renewed interest in demographic patterns. While media coverage of this phenomenon has surged, its explanatory power may have waned.

Firstly, the influence of demographics on elections is no news in US history and it has already been exploited by political parties on numerous occasions: the main example is given by political strategies on race matters, aimed in different occasions at granting the votes of black people or of white racists. Additionally, predictions from influential works like "The Emerging Democratic Majority" have only been confirmed by subsequent political developments, indicating evolving demographic cleavages. While the Democratic Party has made gains in the popular vote, it has faced challenges among demographic groups once considered solidly democratic, such as white, uneducated, blue-collar workers in the Midwest, and, more recently, Latinx voters.

While demographic identities have certainly maintained their influence in determing how people vote, the association of demographic identity and partisanship is subject to evolutions not always easy to predict.

It is the methodological evolution that may have a greater impact: while polls have been under scrutiny after their failure in predicting Trump's victory in 2016, easy access to data and statistical methodologies has fueled interest in studying the relationship between demographic characteristics and electoral outcomes. Political enthusiasts can now engage in simulations to explore how an elector's partisanship may change based on demographic traits or how electoral results could shift with varying preferences among demographic cohorts.

1.1 Our project

In our project we have taken county data on demographic variables, including ethnic composition, languages spoken, family composition, religious affiliation, education and variables describing economic, occupational and insurance status. We have tried to find relations between these variables and electoral outcomes in 2020 US presidential election. Moreover, we

have investigated other demographic patterns that emerge looking at county data.

Our analysis is focused only on 2020 presidential election because, as we have seen, in considering more elections the evolution of political preferences in time should be taken in account. That would be over the scope of our project but it could be an interesting supplement for future research, as we will discuss more precisely in the last section.

2 Description of dataset

In order to build our dataset, we have downloaded demographic data from IPUMS, the world's largest individual-level population database. We have selected data on race, education, languages spoken, composition of family, age, income, occupation, health insurance. Data from IPUMS are individual-level; obviously, there exists no analogous individual-level dataset of electoral votes. Moreover, IPUMS does not cover religious affiliation: in order to get data of religious affiliation, we have used data from United States Religion Census by the Association of Statisticians of American Religious Bodies.

We then chose the smallest unit of observation that allowed us to aggregate both electoral and demographic data: counties.

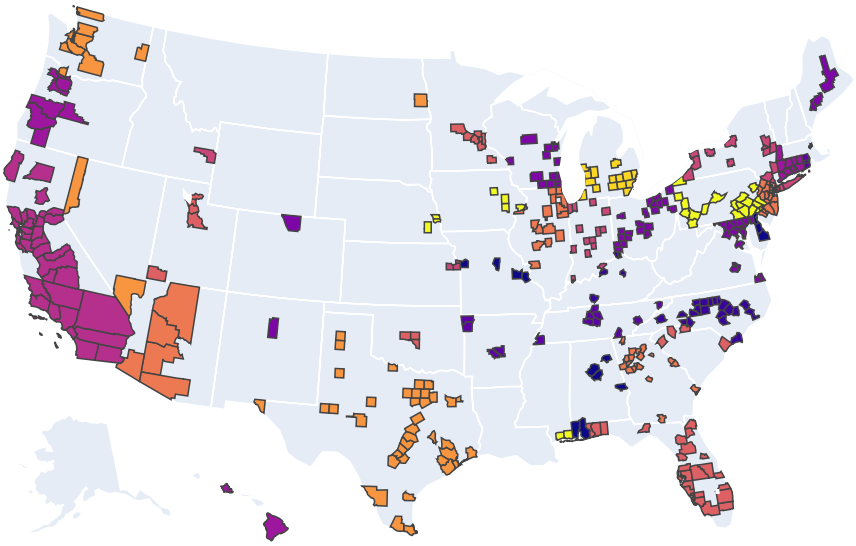
There exist 3244 counties and county-equivalent in the United States. Counties vary greatly by population, ranging from Loving County, Texas and its population of 64 people to Los Angeles County, California and its population of around 10 million people. Clearly, IPUMS cannot cover properly every county; anyway, we assume their survey, using their weights, is representative and not biased.

Considering electoral results by county, religious affiliation data by county, and individual-level data from IPUMS, we aggregated individual data by county, taking the statistics we found more informative: with few exceptions, we have taken percentages where the feature was binary (percentage of people with a mortgage, percentage of black people, e.g.) and averages or medians where values were numerical (average number of people in family, median income,e.g.); all statistics were computed weighting for the weights indicated by IPUMS.

We have assumed that weights indicated by IPUMS permit to have unbiased estimates even at county-level.

In the end, we obtained a dataset of 398 counties with 117 features observed. Among them, there were many collinear features, such as percentages of race or religious affiliation.

We will not present each feature in detail, as it would be unnecessarily tedious: however, we will present features that will appear relevant.



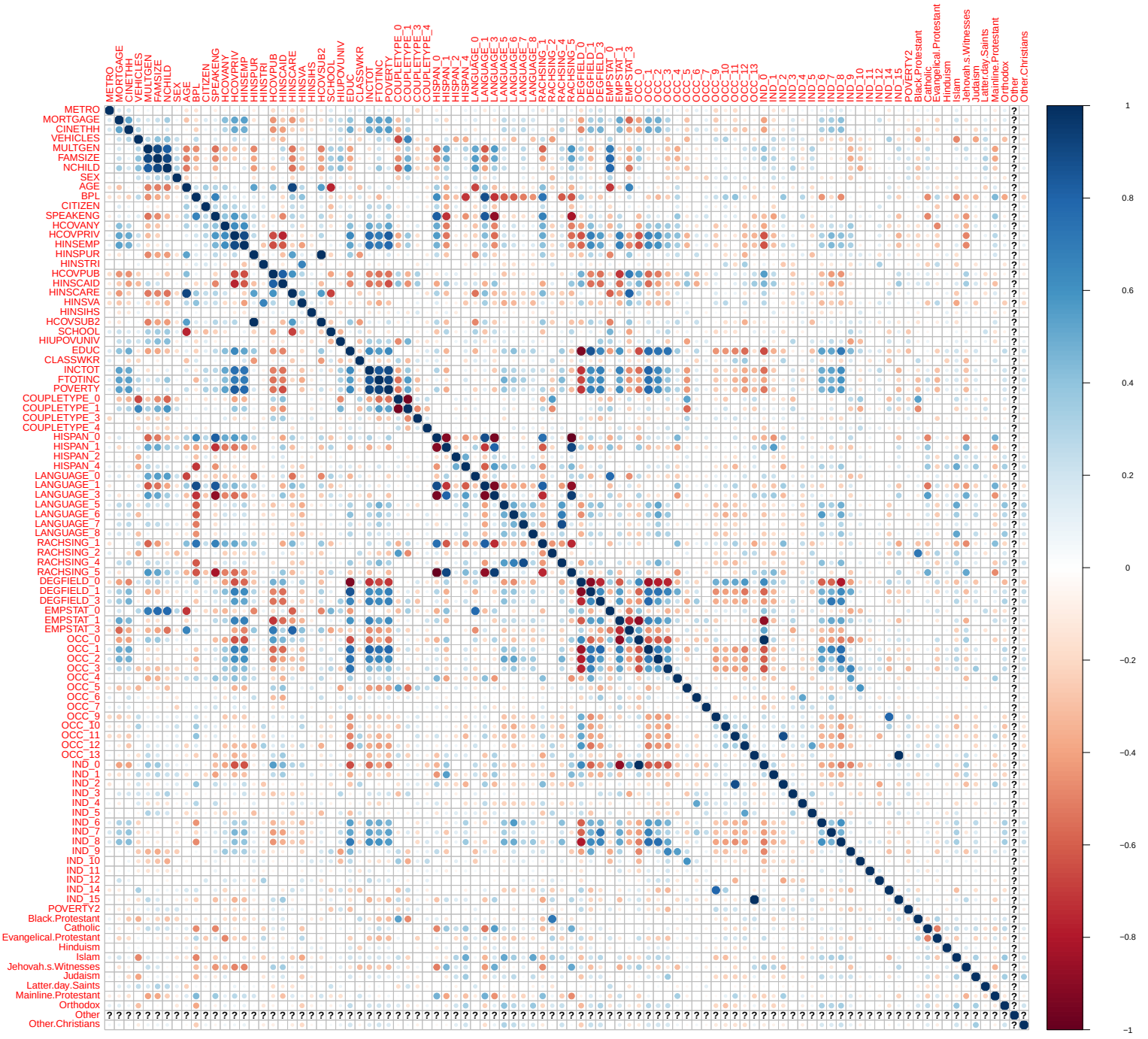
3 Preprocessing

Our dataset, obtained after aggregating at the county level as explained in the previous section, contained some collinear columns that would compromise any statistical analysis.

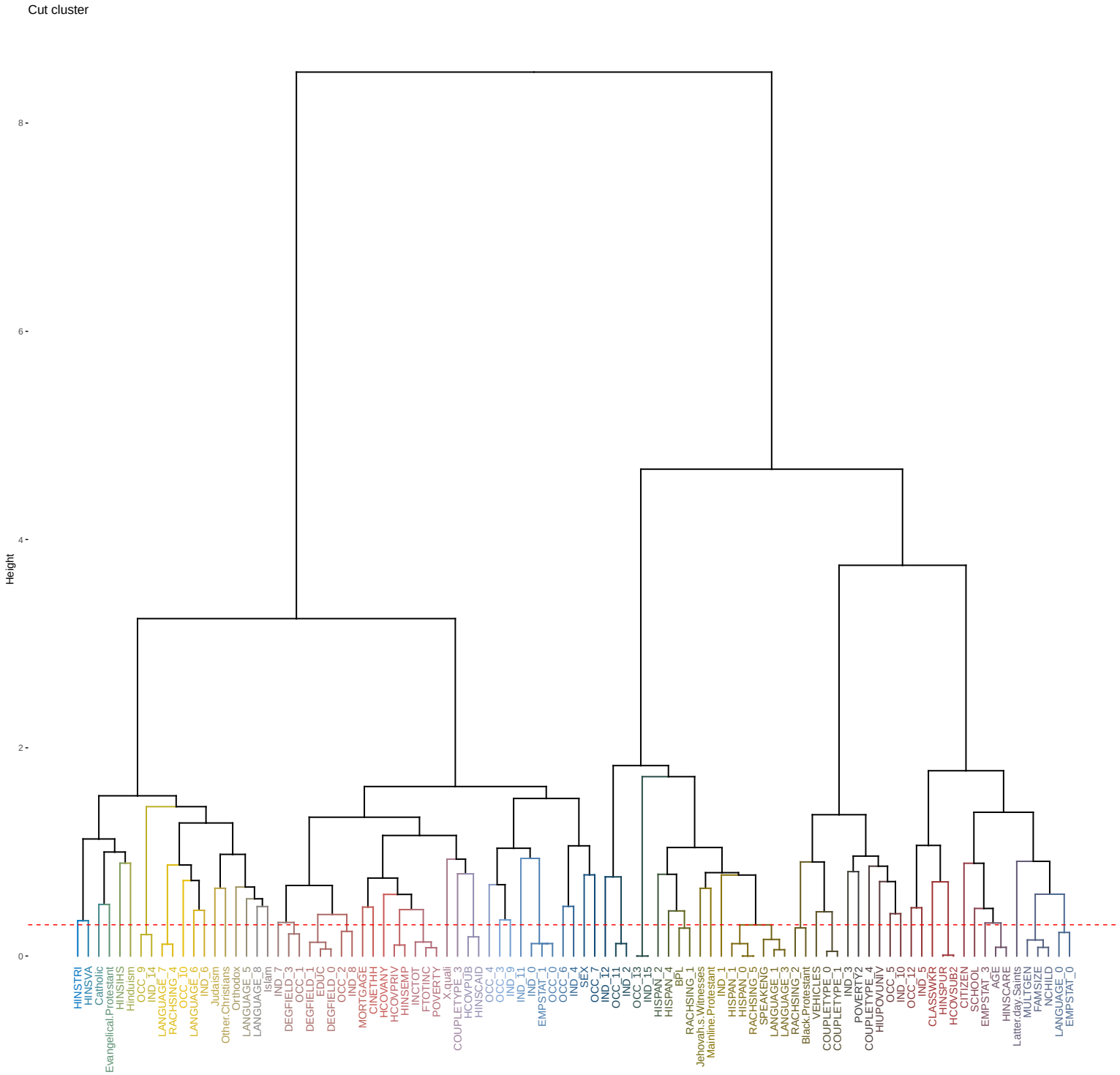
Consequently, the first step was to remove perfectly collinear columns. To accomplish this, it was sufficient to drop one column from every group of structurally collinear columns, such as one of the religion columns and one of the race columns.

However, perfectly collinear columns were not the only obstacle to statistical analysis: our dataset also included collinear columns that were not perfectly collinear and whose collinearity was not eliminated after that process. This collinearity was caused either by structural reasons or by high correlation. We use the former expression to indicate cases where dropping one column was not effective in curbing collinearity (for example, dropping 'Other religion' column leaves a quasi-collinear structure) We use the latter expression to indicate cases of features from different domains that were very highly correlated (for example, percentage of people born outside the US and percentage of people who do not speak English as first language).

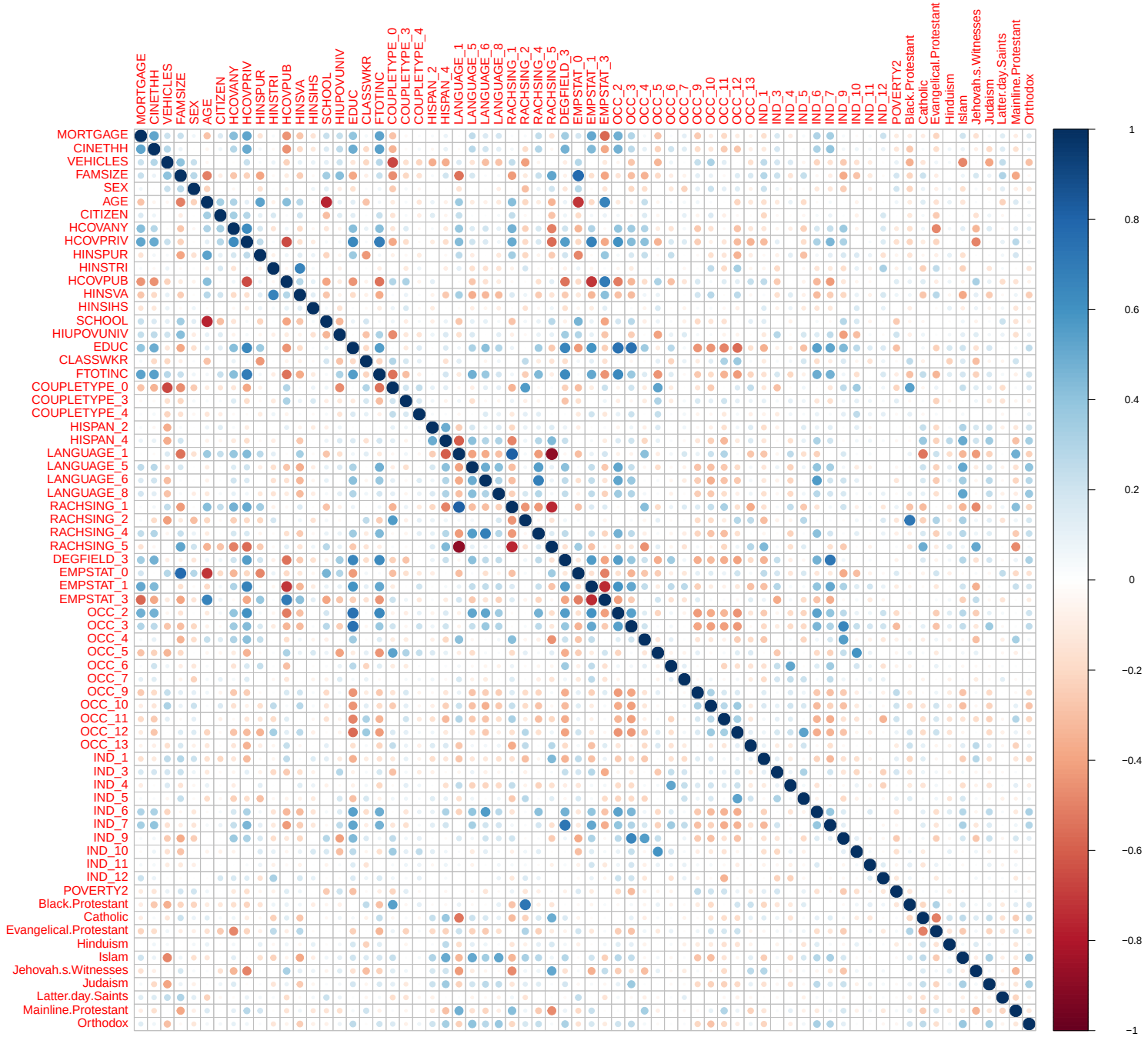
The following matrix shows correlation between variables after dropping structurally collinear columns.



To address these collinearities, we have analysed clusters of variables using hierarchical clustering and studying the emerging dendrogram.



We have maintained a large number of clusters of variables, because selecting variables was the goal of our statistical analysis. However, columns that were too highly correlated would have biased our analysis. Consequently, for each cluster we have chosen one representative variable to reduce dimensionality while preserving important information.

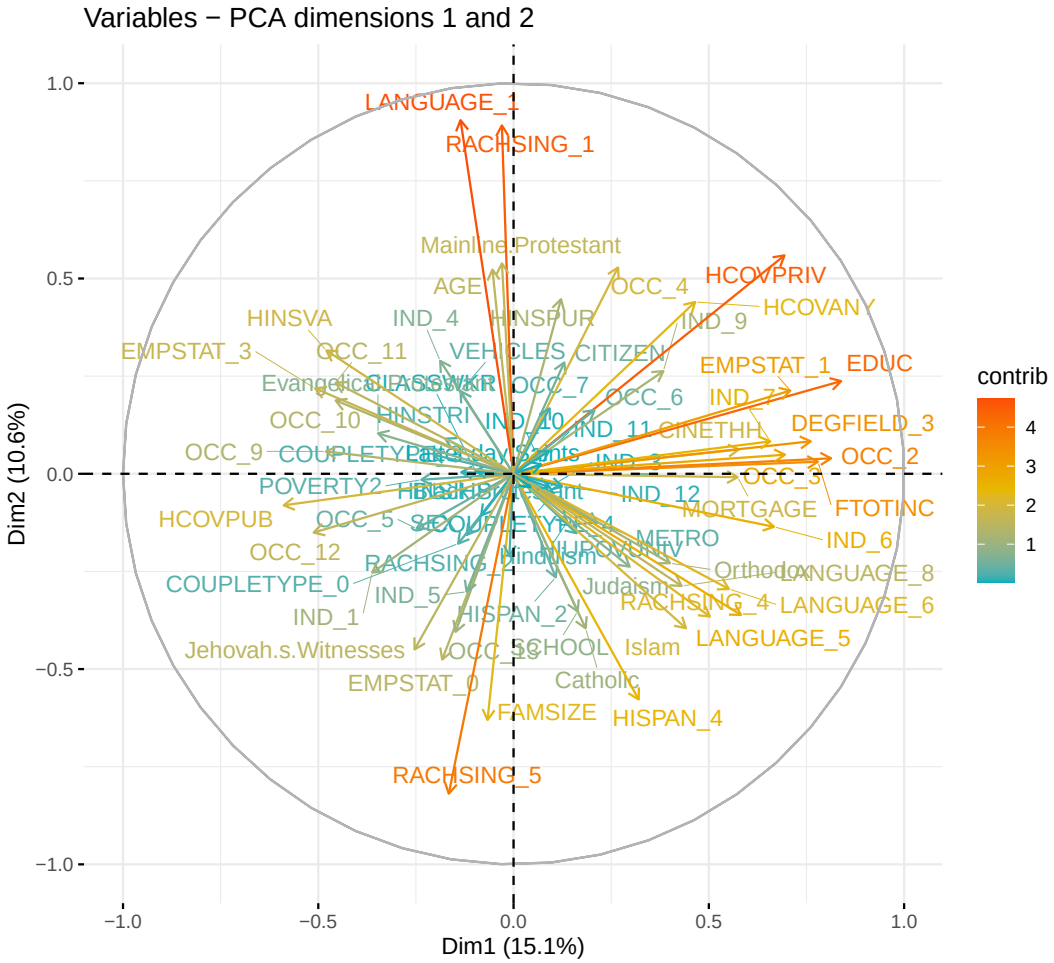

MORTGAGE
CINETHH
VEHICLES
FAMSIZE
SEX
AGE
CITIZEN
HCOVANY
HCOVPRIV
HINSPUR
HINSTRI
HCOVPUB
HINSVA
HINSIHS
SCHOOL
HIUPOVUNIV
EDUC
CLASSWKR
FTOTINC
COUPLETYPE_0
COUPLETYPE_3
COUPLETYPE_4
HISPAN_2
HISPAN_4
LANGUAGE_1
LANGUAGE_5
LANGUAGE_6
LANGUAGE_8
RACHSING_1
RACHSING_2
RACHSING_4
RACHSING_5
DEGFIELD_3
EMPSTAT_0
EMPSTAT_1
EMPSTAT_3
OCC_2
OCC_3
OCC_4
OCC_5
OCC_6
OCC_7
OCC_9
OCC_10
OCC_11
OCC_12
OCC_13
IND_1
IND_3
IND_4
IND_5
IND_6
IND_7
IND_9
IND_10
IND_11
IND_12
POVERTY2
Black.Protestant
Catholic
Evangelical.Protestant
Hinduism
Islam
Jehovah.s.Witnesses
Judaism
Latter.day.Saints
Mainline.Protestant
Orthodox
1
0.8
0.6
0.4
0.2
0
−0.2
−0.4
−0.6
−0.8
−1

4 Unsupervised learning

Before delving into the study of what has determined the partisanship of a county in 2020 presidential election, we have made some unsupervised observations on our dataset, once pre-processed and scaled.

4.1 PCA

The first step was to perform Principal Component Analysis (PCA). PCA is a dimension reduction technique that considers linear combinations of the original feature space and finds the combined dimensions where data vary the most.



We can observe that data vary the most along features that are connected to economic conditions, occupational status and education.

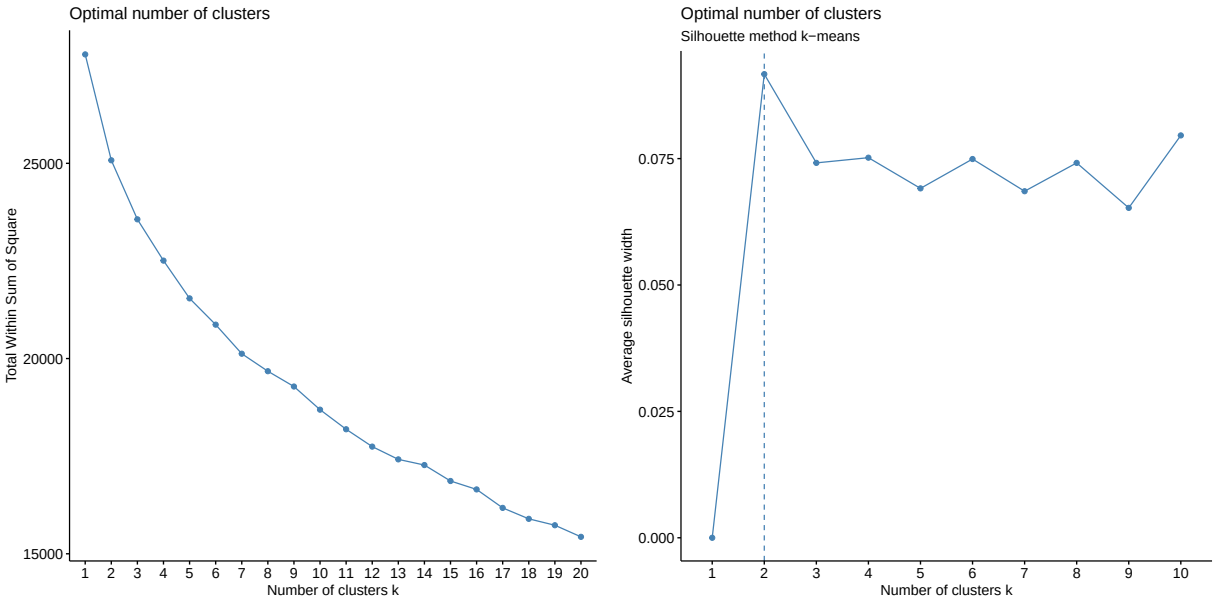
The second principal component, instead, consists mainly of features that describe ethnic composition: LANGUAGE_1 indicates the share of people whose native language is English, RACHSING_1 the share of white people; conversely, RACHSING_5 indicates the share of Latinx and FAMSIZE the average number of people in family, which is partially correlated to the

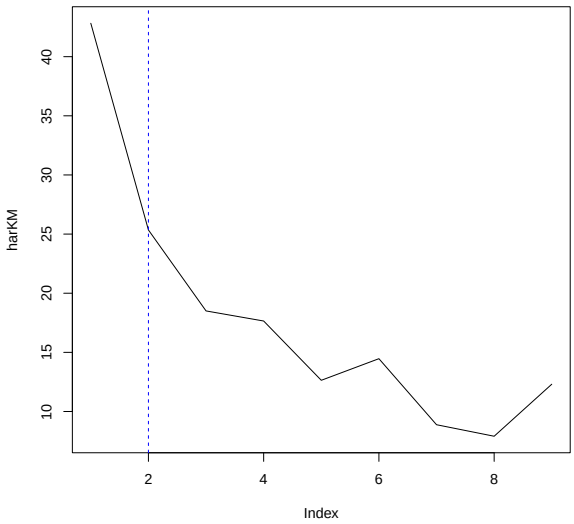
percentage of Latinx as Latinx families tend to be more numerous.

4.2 Clustering

Subsequently, we have performed clustering in order to identify groups of counties with similar characteristics.

We have used k-means clustering choosing the number of clusters according to criteria on total within sum of square, Hartigan Index and average silhouette width.

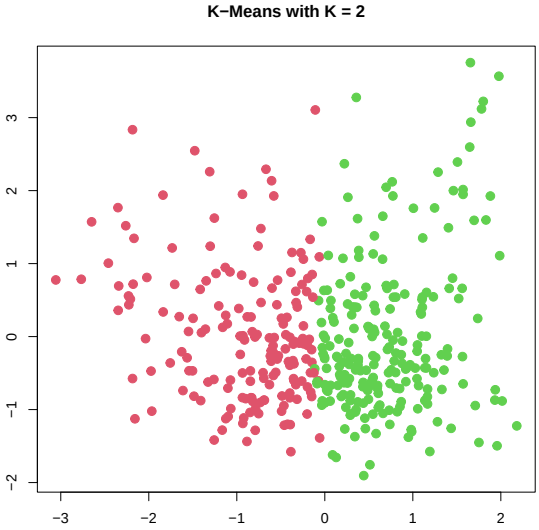




Based on these criteria, we divided the data into two clusters, although the optimal number was not entirely clear. We also experimented with more clusters, but we chose not to present those results as they were not particularly insightful.

We have projected clusters on pairs of principal components in order to visualize divisions among them. In fact, along some principal components we could trace division lines between

clusters; however, clustering is made on the entire space feature, so that visualisation on principal components is a simplification of the underlying dividing mechanism.

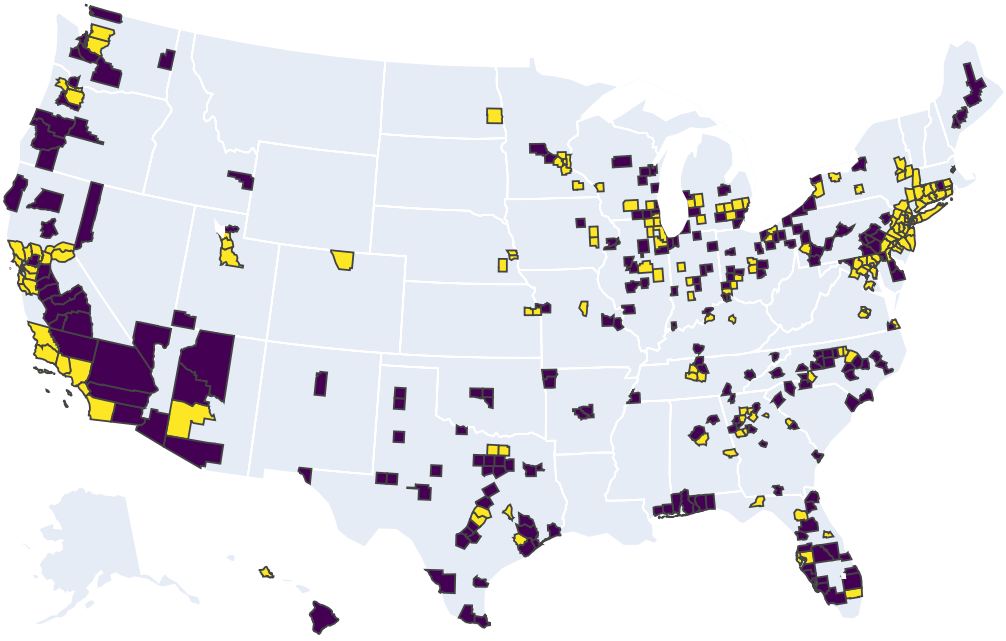


It appears that the division is almost entirely determined by the first principal component, indicating that the dividing mechanism primarily operates on the economic and educational features defining that component. This may be attributed to the abundance of economic features and their high variability, which amplify the influence of economic factors on the division mechanism.

Furthermore, the clusters are not very well-defined; in this case, clustering appears more as a segmentation of a continuous group in feature space rather than a clear division into groups. However, this lack of clarity may again be attributed to the visualization on pairs of dimensions, which does not fully depict the underlying division mechanism.

However, indexes on the right number of clusters indicate that division in cluster based on our features seems a bit facticious.

An useful insight on how clusters are distributed could be borne by a visualisation of clusters on the map:



The main finding is the similarity between counties in coastal California and counties in the Northeast megalopolis. In fact, they are counties with richer and better-educated population.

However, as we have already said, clustering seems a bit facticious and not very informative.

5 Supervised Learning

After unsupervised learning, we moved to the main goal of our research: finding which features have been relevant in determining the partisanship of a county in the 2020 US presidential election.

In order to do so, we considered the difference between the share of votes obtained by the Democratic candidate (i.e., Joe Biden) and the share of votes obtained by the Republican candidate (i.e., Donald Trump).

Given the large number of features and the relatively small size of our sample, a simple regression would have yielded unstable results.

To overcome that problem, we used penalized regressions and feature selection.

5.1 LASSO regression

LASSO is a type of linear regression that adds a penalty term to the ordinary least squares method.

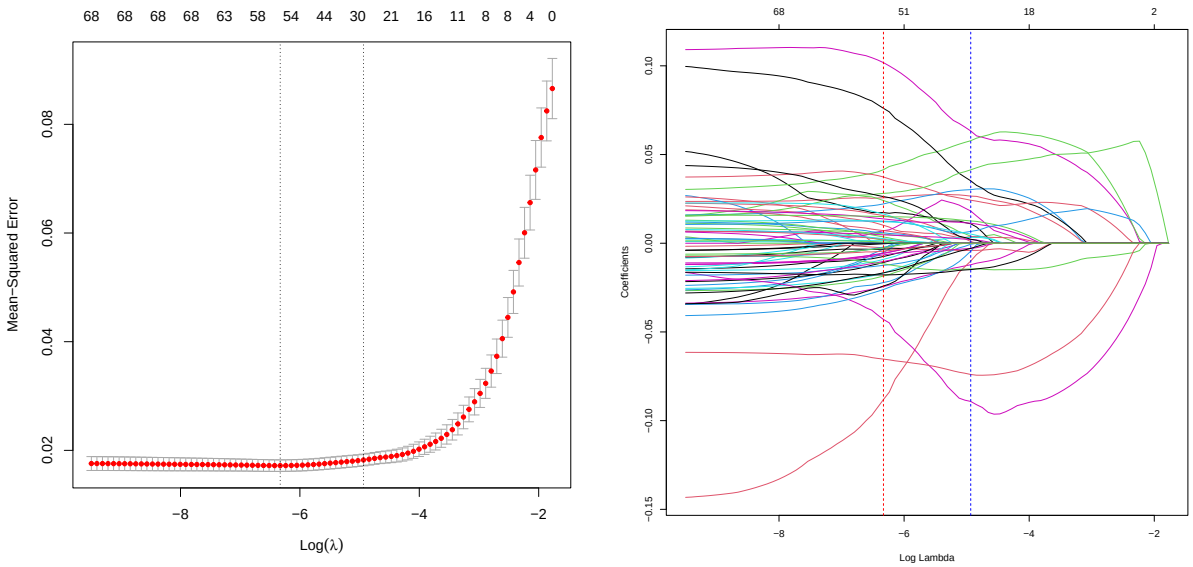
This term penalizes the absolute size of the coefficients, setting some coefficients to zero, thus performing variable selection as well as regularization.

LASSO has been useful to perform feature selection in our high-dimensional dataset, where it has helped identify the most important predictors while shrinking coefficients of others towards zero.

Values of coefficients β, and the quantity of coefficients shrunk to zero, depend on the value of penalisation λ.

In the following graph, we can see different results of LASSO for different values of λ. In particular, we are interested in two values of λ: the value for which estimated Mean Square Error is minimum (min) and the largest value such that estimated Mean Square Error is within 1 standard error of the minimum (1se).

Estimates of Mean Square Error are given by cross-validation.



Choosing λ.1se we get the following coefficients.

(Intercept)	0.011082
MORTGAGE	0.027318
HCOVANY	0.008256
HINSTRI	−0.014072
HCOVPUB	0.012445
HINSVA	−0.003838
SCHOOL	−0.003860
EDUC	0.069241
COUPLETYPE_0	0.054728
COUPLETYPE_3	0.029500
HISPAN_2	−0.001199
LANGUAGE_1	−0.021911
LANGUAGE_5	−0.002855
RACHSING_1	−0.084032
RACHSING_2	0.042581
RACHSING_4	0.003793
DEGFIELD_3	−0.012624
EMPSTAT_1	0.022586

OCC_2	0.026194
OCC_3	0.038185
OCC_4	−0.001300
OCC_5	0.002978
OCC_6	−0.003634
OCC_9	0.005067
OCC_10	−0.014899
OCC_13	0.002845
IND_1	−0.005500
IND_9	0.005696
Black.Protestant	0.011945
Evangelical.Protestant	−0.071974
Hinduism	0.013520
Islam	0.006060
Jehovah.s.Witnesses	0.006908
Latter.day.Saints	−0.015734
Mainline.Protestant	−0.000108

5.2 Feature screening

When dealing with a high-dimensional data with the goal of finding relevant features, we may use techniques of feature screening as well.

In particular, we have compared results obtained with LASSO with results obtained with Sure Independence Screening.

EDUC
COUPLETYPE_0
RACHSING_1
OCC_2
OCC_3
OCC_10
Evangelical.Protestant
Islam

Feature screening selects too few variables: the number of features in our dataset is probably not high enough to justify feature screening. Consequently, we decided to focus on features identified by LASSO.

5.3 Ordinary Least Square after LASSO

In order to correct for the bias given by LASSO regression, we have performed boostrap and OLS.

So, we first performed an Ordinary Least Square regression on regressors obtained by LASSO.

MORTGAGE	0.028***
	(0.010)
HCOVANY	0.026*
	(0.015)
HINSTRI	−0.020**
	(0.010)
HCOVPUB	0.011
	(0.017)
HINSVA	−0.0004
	(0.011)
SCHOOL	−0.018*
	(0.010)
EDUC	0.092***
	(0.016)
COUPLETYPE_0	0.043***
	(0.012)
COUPLETYPE_3	0.026***
	(0.008)
HISPAN_2	−0.013*
	(0.007)
LANGUAGE_1	−0.077***
	(0.025)
LANGUAGE_5	−0.028***
	(0.010)
RACHSING_1	−0.050*
	(0.029)
RACHSING_2	0.075***
	(0.017)
RACHSING_4	0.012
	(0.011)
DEGFIELD_3	−0.035***
	(0.012)
EMPSTAT_1	0.035**
	(0.014)
OCC_2	0.030**
	(0.012)
OCC_3	0.018
	(0.013)
OCC_4	−0.023**
	(0.010)
OCC_5	−0.001
	(0.009)
OCC_6	−0.007
	(0.007)
OCC_9	0.017**
	(0.008)
OCC_10	−0.017**
	(0.008)
OCC_13	0.006
	(0.007)
IND_1	−0.016*
	(0.008)
IND_9	0.032**
	(0.014)
Black.Protestant	0.014
	(0.010)
Evangelical.Protestant	−0.064***
	(0.009)
Hinduism	0.018***
	(0.007)
Islam	0.008
	(0.009)
Jehovah.s.Witnesses	0.012
	(0.009)
Latter.day.Saints	−0.021***
	(0.007)
Mainline.Protestant	−0.003
	(0.008)
Constant	0.011*
	(0.006)

Observations	398
R^2	0.843
Adjusted R^2	0.831
Residual Std. Error	0.121 (df = 367)
F Statistic	65.907*** (df = 30; 367)
Note:	*p<0.1; **p<0.05; ***p<0.01

5.4 Bootstrap with LASSO

Then, we performed residual bootstrap LASSO to find confidence intervals for regression coefficients.

colnames	Beta	min	max
MORTGAGE	0.026562	0.01481	0.048838
HCOVANY	0.003939	−0.01038	0.007878
HINSTRI	−0.012122	−0.02424	−0.005918
HCOVPUB	0.010950	0.01896	0.021900
HINSVA	−0.004009	−0.00802	0.010827
SCHOOL	−0.000569	−0.00130	0.000461
EDUC	0.063287	0.04766	0.100744
COUPLETYPE_0	0.057786	0.03643	0.076627
COUPLETYPE_3	0.030211	0.01875	0.047068
LANGUAGE_1	−0.010043	−0.02009	−0.013852
RACHSING_1	−0.089173	−0.11015	−0.069819
RACHSING_2	0.035182	0.02482	0.067066
RACHSING_4	0.003687	−0.00977	0.007374
DEGFIELD_3	−0.004502	−0.01144	−0.009004
EMPSTAT_1	0.018165	0.01328	0.036330
OCC_2	0.024238	0.00691	0.048476
OCC_3	0.041375	0.01920	0.061224
OCC_5	0.003260	−0.00562	0.006521
OCC_6	−0.002907	−0.00581	0.002342
OCC_9	0.000977	0.00195	0.012617
OCC_10	−0.014513	−0.02903	0.000655
OCC_13	0.002094	−0.00710	0.004188
IND_1	−0.002913	−0.00583	0.002068
IND_9	0.002238	−0.00807	0.004477
Black.Protestant	0.011260	−0.00214	0.022519
Evangelical.Protestant	−0.073977	−0.09530	−0.062027
Hinduism	0.012564	0.00453	0.025127
Islam	0.006657	−0.00967	0.013313
Jehovah.s.Witnesses	0.005174	−0.00266	0.010349
Latter.day.Saints	−0.014862	−0.02972	−0.007734

5.5 Comments on supervised analysis

As you can see the the two methods produced similar and coeherent outputs.
So, we can now give some comments on the results.

Before commenting, it's important to note that the features have been normalised: each coefficient signifies the impact of a standard deviation from the mean on the Democratic advantage. For example, -0.067 associated to Evangelical Protestant means that in counties where Evangelical Protestants are one standard deviation above the average, Republicans are expected to get 6.7% more votes than in an average county.

One important caveat: here, the average county is defined as the average county within our subsample.

Although not highly significant, the constant indicates a slight Democratic favor in the average county.

Observations reveal that Democrats are favored in counties where a higher proportion of individuals live in houses encumbered by a mortgage(MORTGAGE), where individuals do not live with a partner (COUPLETYPE_0) or live with an unmarried partner (COUPLETYPE_3). Furthermore, Democrats tend to garner higher vote shares in areas with higher education levels (EDUC, measured by the percentage of individuals with some form of college degree) and where the proportion of Black individuals is higher (RACHSING_2). Other variables that significantly favor Democrats include higher proportions of employed people (EMPSTAT_1) and people occupied in IT (OCC_2) or care (OCC_3) jobs. Counties with higher proportions of employers in construction and extraction occupations (OCC_9) and Hindu individuals also benefit Democrats, even if less significantly.

Republicans, on the other hand, are favored in counties with higher percentages of native English speakers (LANGUAGE_1) and White people (RACHSING_1), people with business degrees (DEGFIELD_3), Evangelical Protestants and Mormons.

Other variables identified by LASSO are not very significant: the effect of having more individuals with health insurance is ambigous and appears to depend on the provider (HINSTRI indicates military health program, HCOVPUB includes Medicare, Medicaid and Veterans'Administration programs, HINSVA indicates Veterans'Administation programs). Similarly, there are other variables whose effect is ambigous, such as the percentage of people currently at school (SCHOOL), people occupied in service occupations (OCC_5), sales and office occupations (OCC_6) and unemployed individuals (OCC_13); people employed in agriculture, forestry, fishing and hunting (IND_1) and in educational, health care and service assistance (IND_9). Additionally, higher percentages of Asian individuals (RACHSING_4), Black Protestant, Islamic, Jehovah's Witnesses and Mainline Protestant do not seem to significantly affect electoral results.

Our results are quite consistent with usual analysis of demographic determinants of election results.

Indeed, in recent years Democratic voters have been increasingly associated with diversity, both in ethnicity and in familiar composition, as well as higher education.

Conversely, the Republican Party has appealed to White, less educated and more conservative individuals. Moreover, in the last years Republicans have largely dominated votes from the growing share of white Evangelical Protestants.

However, some results are less intutive. Among them, the higher share of Democratic votes in counties where more people are employed in IT countradicts the claim of Republicans who celebrate hard sciences against the perceived emergence of naive social scientists. However, this result may be due to the concentration of IT jobs in extremely progressive and highly educated counties in coastal California or in Northeast metropolis.

Other results requiring explanations concern variables that appear to be not significant. Firstly, the elimination of median income after LASSO regression may be explained by its nonlinear effect and the higher explanatory power held by some correlated variables, such as education. Secondly, the null significance of Islamic individuals or Jehovah's Witnesses is probably due to their small population size: even in counties with a higher share of Islamic

population or Jehovah's Witnesses, their numbers are insufficient to affect electoral results significantly. On the contrary, the null significance of the share of Mainline Protestants may be due to its vastity and diversity: this category includes various religious affiliations with differing identities. The null significance of Black Protestants should also be investigated further: this group is typically strongly associated with the Democratic Party, even if it tends to hold more socially conservative views. However, it is primarily located in conservative and Republican-leaning Southern counties and that may explain its ambigous effect on electoral outcomes.

6 An informed review of unsupervised learning

Supervised learning provided us with insights into how certain features impacted the 2020 presidential election results by county.

Now, we can utilize the coefficients we obtained from supervised learning to weight features. This allows us to employ unsupervised learning techniques on a rescaled feature space where certain features carry more weight.

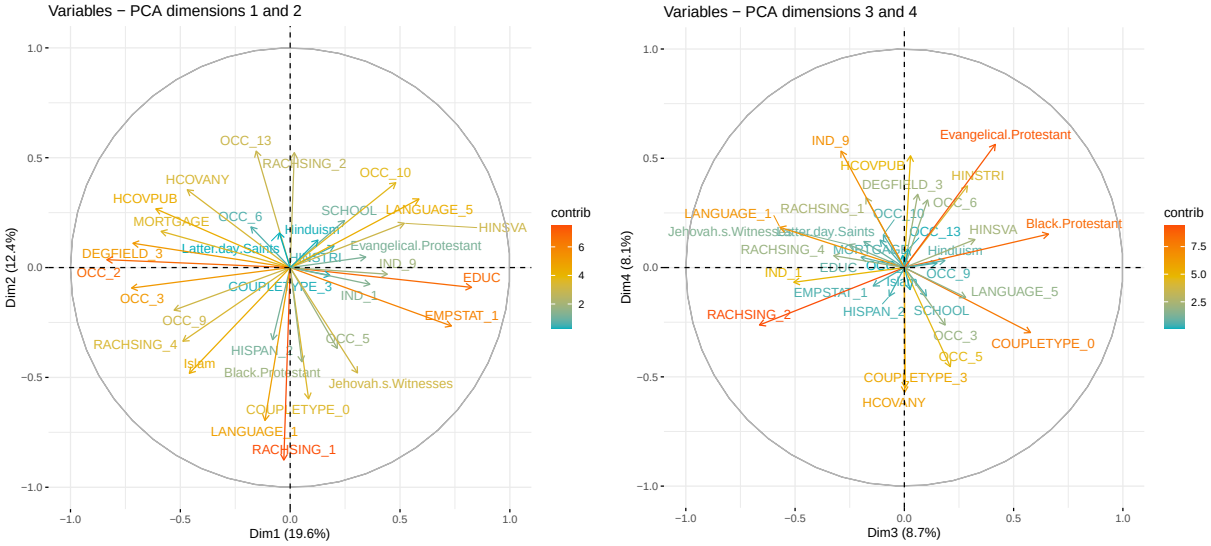
Indeed, we are now employing unsupervised techniques within a 'supervised framework' where we are particularly concerned by features that are most informative on election results.

This enables us to determine whether our initial unsupervised learning was affected by the high number of non-relevant features. However, this carries also some biases.

6.1 PCA

We have performed PCA using axes weighted by their importance in the LASSO regression. There is a caveat: PCA aims to identify the directions where data vary the most. By weighting for their importance, we artificially stretch directions, altering the variability.

In fact, we are not performing a standard PCA on scaled data, but a 'weighted' PCA where we are interested in both the variability and the importance of features.



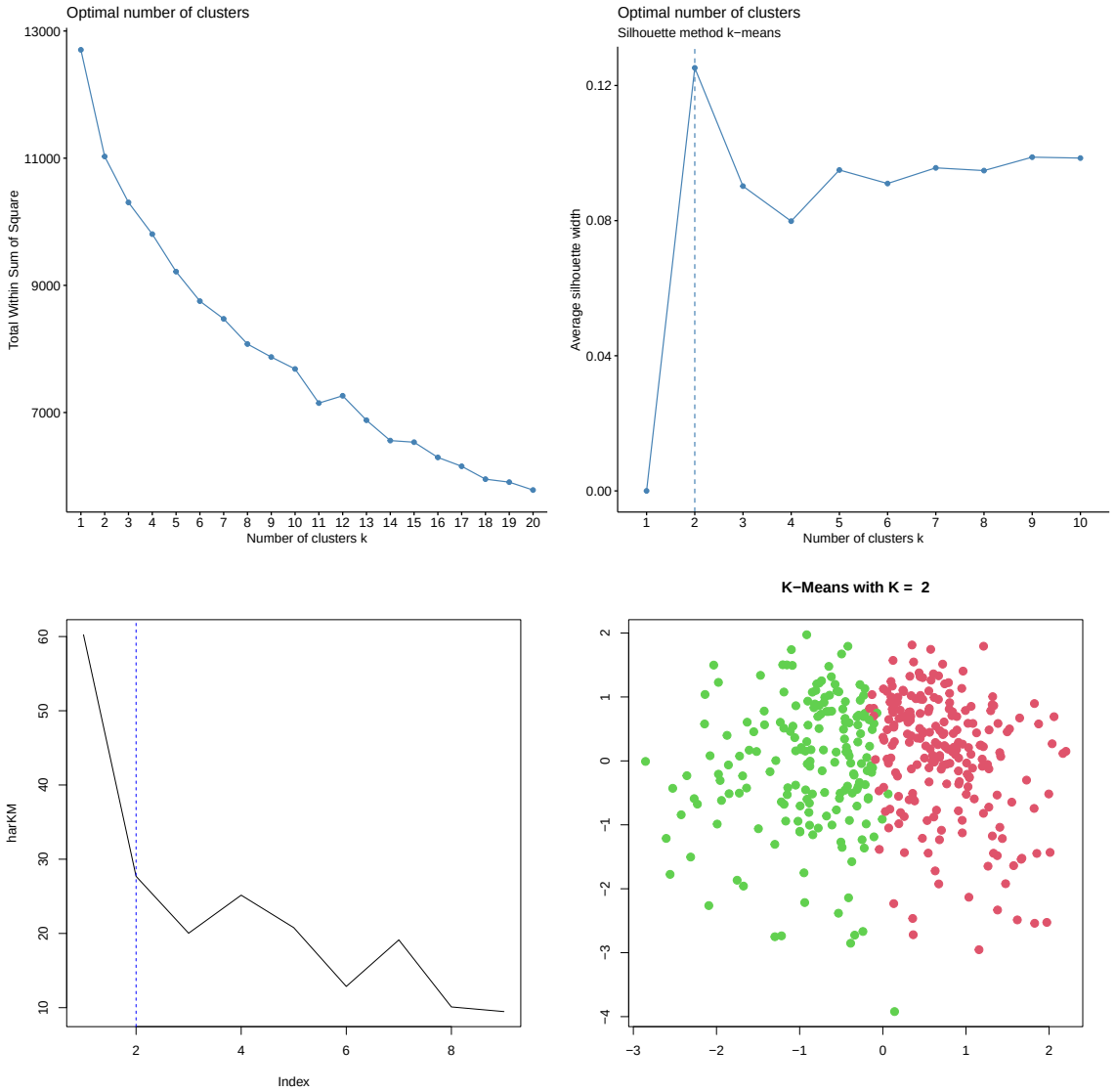
Results hold some similarities with the original PCA results, with education and occupational features being the main contributors to the first principal components and the proportions of White individuals, native English speakers, and unmarried individuals being the main determinants of the second dimension. The third and fourth dimensions give more weight to religious affiliations.

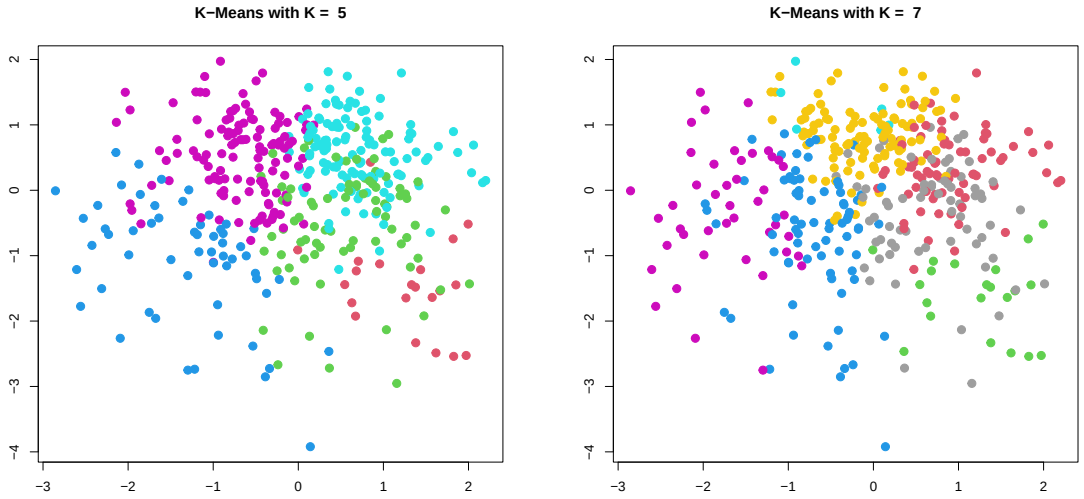
Overall, the principal components are similar to those found in unsupervised learning, but the graph is obviously less dense as fewer features are considered.

6.2 Clustering

We have now performed clustering on weighted data in order to look at how counties can be grouped once we count for the impact of features on the election results. The driving question was: if we focus on features that are more relevant in elections, will counties be grouped differently?

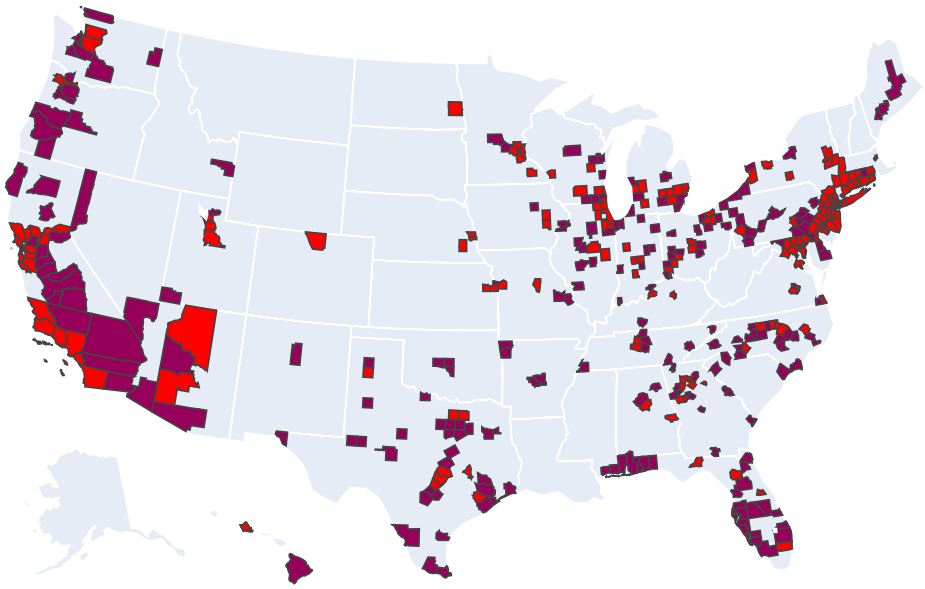
The optimal number of clusters was, once again, not very well defined. We attempted three different numbers of clusters: two, five and seven. We have plotted the results on planes composed of adjusted principal components.

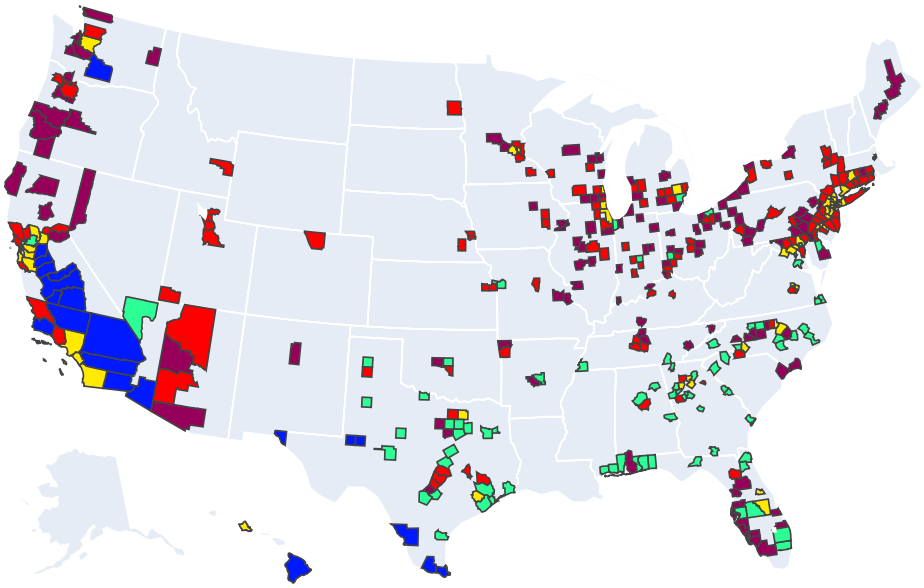


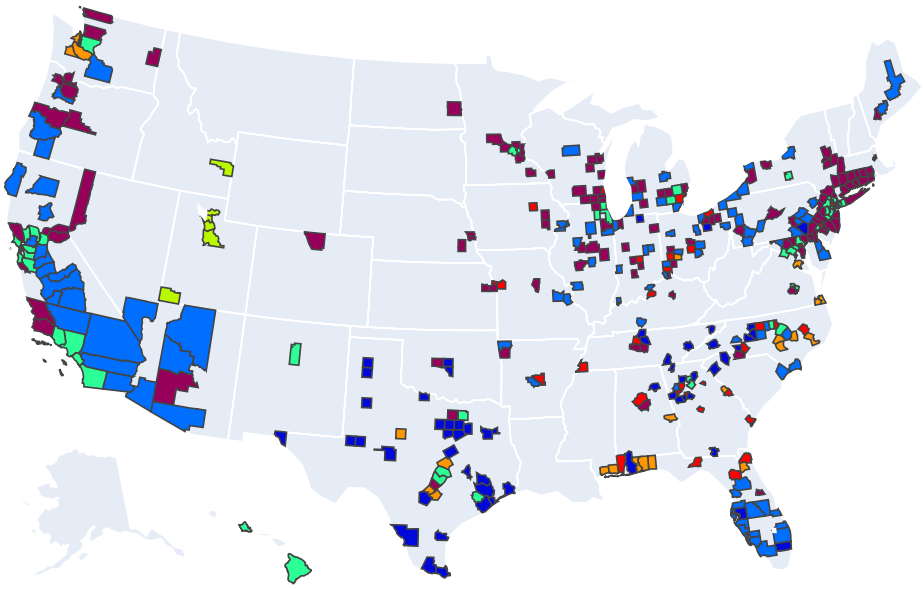


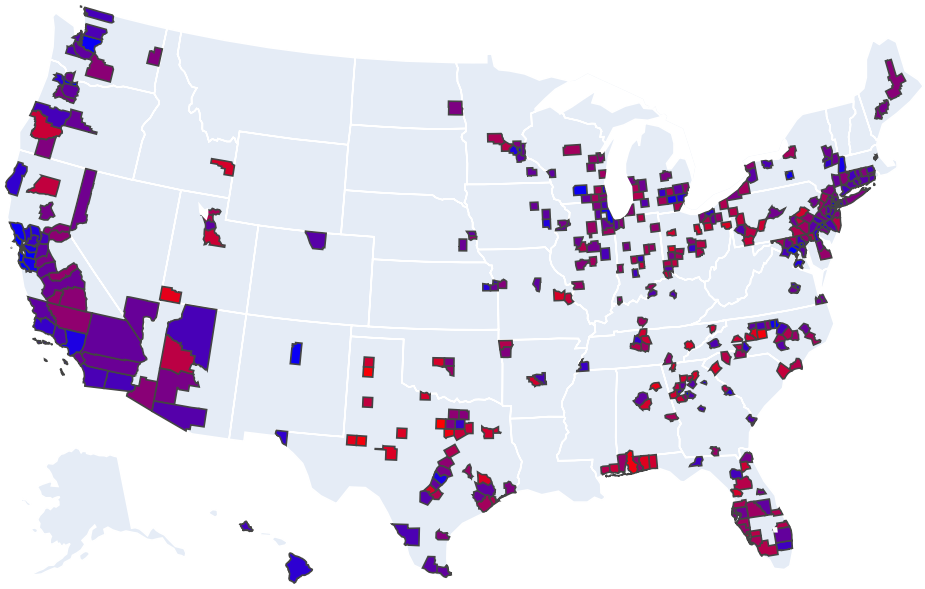
As before, clustering results in segmentation rather than a clear division in well-defined groups.

Subsequently, we have located counties on the map.









Even by looking at the maps, we can observe the segmentation process: there are patterns already present in unsupervised learning (such as similarities between counties in coastal California and counties in Northeast metropolis, as well as similarities between internal Californian counties and southern Texan and Floridian counties). With more clusters, these similarities remain but are better defined.

It appears evident that clustering is a partition of continuity: the feature space seems continously filled without significant gaps.

We can compare the cluster divisions with the percentages of Republican votes shown in the last map. Following the traditional color scheme, red counties are those where Republicans received higher shares of votes, while blue counties are those where Democrats received more votes. We find that counties in the same clusters have similar shares of Republican votes. This is quite expected, as the division into clusters is made in a space where dimensions are weighted for their importance in defining partisanship.

7 Limitations and future directions

In our work we have tried to determine how demographic characteristics, ranging from racial and religious identities to socio-economic conditions, have affected partisanship in 2020 US presidential election.

However, there are various limitations to the validity of our work. We will present some of them, some possibilities to overcome them and other possible extensions to our work.

7.1 Choice of features

The first element one could change is the choice of features: even if in our study we have considered nearly every feature whose effect on elections have been studied (and some more features), one could argue that we did not take in account characteristics that could explain electoral results more effectively. This could be particularly meaningful for unsupervised learning.

The easy solution would be to consider more characterics, and in particular every characheristic that seems useful.

7.2 Representativity of sample

A more important concern is about the selection of our subsample.

First of all, we have assumed that, once weighted according to IPUMS indications, the datased was the result of an independent and identically distributed subsample of the population. That seems a pretty solid assumption, as IPUMS is a widely reknown database and we expect its data to be reliable.

Subsequently, we have assumed that aggregating on county basis would preserve the representativity of our sample. That is a much stronger assumption, and one that can question the validity of the whole study. In fact, an unbiased subsample of a population does not guarantee that every subpopulation is represented in an unbiased way.

If we can expect that large subpopulations are represented well, we may cast some doubts on smaller counties.

If features of counties are not adequately represented, we are estimating links between true electoral results and fake demographic characteristics. The result would be, obviously, badly identified links.

As we have said, that would hurt the validity of every analysis we have made.

A more precise study would require to verify that data are representative for every subpopulation taken into account.

7.3 Counties are not individuals

Our focus has been on counties as the unit of observation, allowing us to discern the impact of higher shares of individuals within specific categories on electoral outcomes.

However, it is crucial to note that if a county with a larger proportion of individuals from a certain category tends to vote more for a particular party, it does not necessarily imply strong support for that party within that category. For instance, counties with higher proportions of Black Protestants may appear only slightly more Democratic than average, despite the

expectation that Black Protestants would be staunchly Democratic. Yet, these counties are often located in the South, where there is also a significant population of conservative white Republicans.

Moreover, the presence of certain demographic groups can also generate a process of opposition as it happens in Europe where the presence of Muslims in certain areas has led to electoral success for far-right parties. While this phenomenon may manifest differently in America, it remains a possibility.

Furthermore, there are group effects to consider, where the influence of an individual's identity on their voting behavior may be less pronounced in a county with few people who share the same identity. For example, as religious communities become more cohesive, they can mobilize more votes and it may be the interest of parties to court these potential electors treating them as a cohesive group.

These are not problems per se, but they have to be considered while looking at our results: our study focuses on county-level effects rather than individual-level effects.

To mitigate some of these complexities, incorporating conditional variables, such as the percentage of Black Protestants conditioned on the number of white Protestants, can provide a more nuanced understanding of the data.

7.4 Further directions

Other possible directions of research include the study of dynamics.

Indeed, our study was focused solely on effects of demographic features in 2020 US presidential election. Studying how these effects have evolved in time or in different electoral cycles could be of great interest. That would require to take in account the power of incumbency, the evolution of demographic patterns, the evolution of generic country partisanship and other dynamic characteristics.

References

[1] *2020 U.S. Religion Census: Religious Congregations & Membership Study*. 2023. URL: https://www.usreligioncensus.org/node/1639.

[2] Thomson-DeVeaux Amelia, Skelley Geoffrey, and Bronner Laura. *What We Know About How White and Latino Americans Voted in 2020*. 2020. URL: https://fivethirtyeight.com/features/what-we-know-about-how-white-and-latino-americans-voted-in-2020/.

[3] MIT Election Data and Science Lab. *U.S. Senate Precinct-Level Returns 2020*. Version V1. 2022. DOI: 10.7910/DVN/ER9XTV. URL: https://doi.org/10.7910/DVN/ER9XTV.

[4] Lauter David. *Why Changing U.S. Demographics Aren't Affecting the Political Balance of Power*. 2024. URL: https://www.latimes.com/politics/newsletter/2024-04-13/why-changing-u-s-demographics-arent-affecting-the-political-balance-of-power-politics.

[5] Morris G. Elliott, Brown Amina, and Aaron Bycoffe. *Swing the Election 2024*. 2024. URL: https://projects.fivethirtyeight.com/2024-swing-the-election/.

[6] Skelley Geoffrey. *How demographic swings could impact the 2024 election*. 2024. URL: https://abcnews.go.com/538/demographic-swings-impact-2024-election/story?id=108700434.

[7] *IPUMS USA*. Version Version 15.0. Minneapolis, MN, 2024. DOI: 10.18128/D014.V4.0. URL: https://doi.org/10.18128/D014.V4.0.

[8] Gareth James et al. *An Introduction to Statistical Learning: with Applications in R*. Springer, 2013. URL: https://faculty.marshall.usc.edu/gareth-james/ISL/.

[9] John B. Judis and Ruy Teixeira. *The Emerging Democratic Majority*. 2002. URL: https://www.ibs.it/emerging-democratic-majority-ebook-inglese-john-b-judis-ruy-teixeira/e/9780743238557.

[10] Amira Karyn and Alexander Abraham. *How the Media Uses the Phrase 'Identity Politics'*. 2022. URL: https://www.cambridge.org/core/journals/ps-political-science-and-politics/article/how-the-media-uses-the-phrase-identity-politics/D7E1DAAF44BC572C0CA38DFC694D3B20.

[11] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria, 2024. URL: https://www.R-project.org/.

[12] The Economist. *Build a voter*. 2024. URL: https://www.economist.com/interactive/us-2024-election/build-a-voter.