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

Suqi Chen
Giovanni Stivella

Statical Learning and Large Data, Scuola Superiore Sant'Anna di Pisa

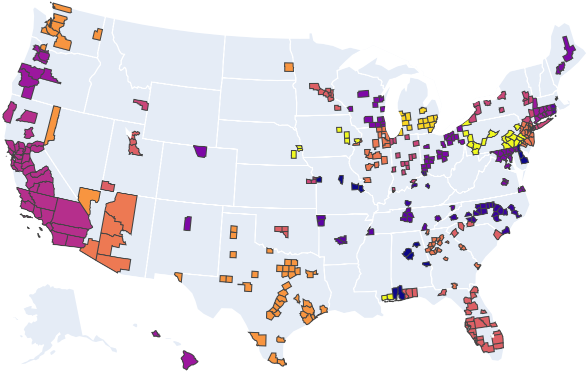
May 9, 2024

1. Introduction
2. Preprocessing
3. Unsupervised learning
   - PCA
   - Clustering
4. Supervised Learning
   - LASSO Regression
   - Screening
   - OLS after LASSO
5. An informed review of unsupervised learning
   - PCA
   - Clustering revisited
6. Limits and further directions

Introduction

- US election results are often investigated through a demographic lens
  - Specialised news outlets such as FiveThirtyEight have emerged in last years
- Statistical analysis has permitted to better study demographic and electoral patterns
- Using data from IPUMS and United States Religion Census we tried to identify these patterns among different counties
  - We collected individual data on multiple variables from IPUMS, county data on electoral results from MEDSL and county data on religious affiliation from United States Religion Census
  - Subsequently, we aggregated individual data on a county level
  - Then we performed statistical analysis having counties as unities of observation

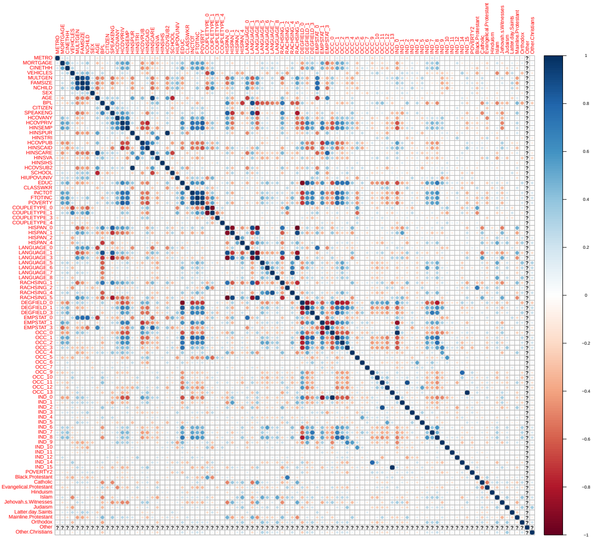
Map of considered counties



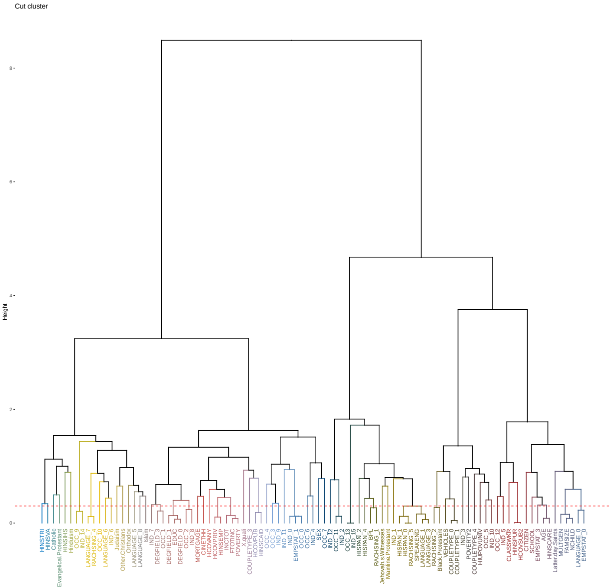
Preprocessing

- There were structurally collinear columns: partitions of population obviously sum up to 1
  - One column from each group had to be eliminated in order to perform any statistical analysis
- Highly correlated columns had the potential to spoil any statistical analysis, too
  - We studied them through a correlation matrix and a dendrogram that represented clusters of variables
- Once we had found clusters, we have chosen one feature for every cluster reducing the number of considered variables

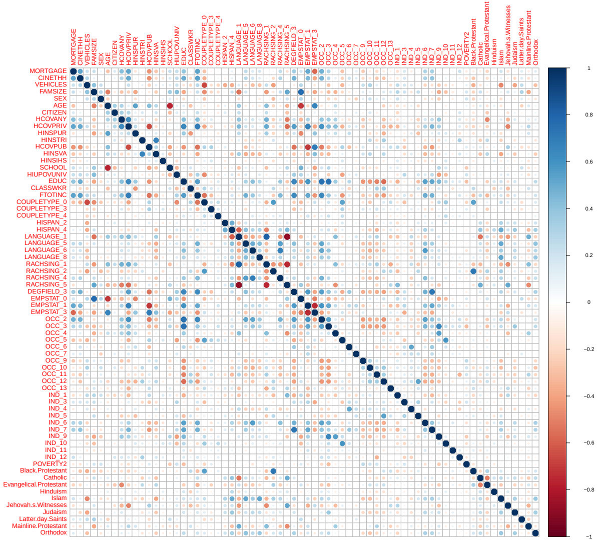
Corplot


METRO
MORTGAGE
CINETHH
VEHICLES
MULTGEN
FAMSIZE
NCHILD
SEX
AGE
BPL
CITIZEN
SPEAKENG
HCOVANY
HCOVPRIV
HINSEMP
HINSPUR
HINSTRI
HCOVPUB
HINSCAID
HINSCARE
HINSVA
HINSIHS
HCOVSUB2
SCHOOL
HIUPOVUNIV
EDUC
CLASSWKR
INCTOT
FTOTINC
POVERTY
COUPLETYPE_0
COUPLETYPE_1
COUPLETYPE_3
COUPLETYPE_4
HISPAN_0
HISPAN_1
HISPAN_2
HISPAN_4
LANGUAGE_0
LANGUAGE_1
LANGUAGE_3
LANGUAGE_5
LANGUAGE_6
LANGUAGE_7
LANGUAGE_8
RACHSING_1
RACHSING_2
RACHSING_4
RACHSING_5
DEGFIELD_0
DEGFIELD_1
DEGFIELD_3
EMPSTAT_0
EMPSTAT_1
EMPSTAT_3
OCC_0
OCC_1
OCC_2
OCC_3
OCC_4
OCC_5
OCC_6
OCC_7
OCC_9
OCC_10
OCC_11
OCC_12
OCC_13
IND_0
IND_1
IND_2
IND_3
IND_4
IND_5
IND_6
IND_7
IND_8
IND_9
IND_10
IND_11
IND_12
IND_14
IND_15
POVERTY2
Black.Protestant
Catholic
Evangelical.Protestant
Hinduism
Islam
Jehovah.s.Witnesses
Judaism
Latter.day.Saints
Mainline.Protestant
Orthodox
Other
Other.Christians
1
0.8
0.6
0.4
0.2
0
−0.2
−0.4
−0.6
−0.8
−1

Dendrogram



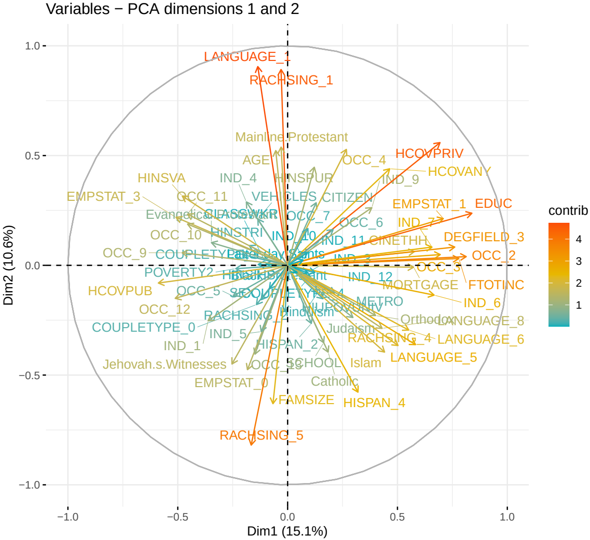
Corplot after Cut



Unsupervised learning

- We started our study by giving a look at distribution of data
  - PCA was performed in order to find directions of higher variability
  - Clustering was performed in order to find natural groups of counties emerging from data distribution
- We visualised clusters both on principal components and on the map

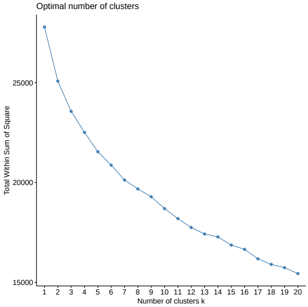
PCA - principal dimension 1 and 2

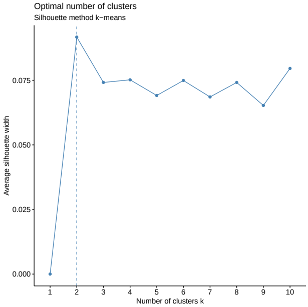


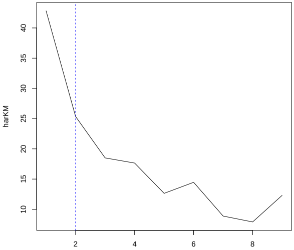
Some insights from PCA

- The main direction of variability combines mainly economic dimensions: income, job, education
- The second principal component is heavily influenced by ethnical and linguistic features, such as the percentage of white and protestant

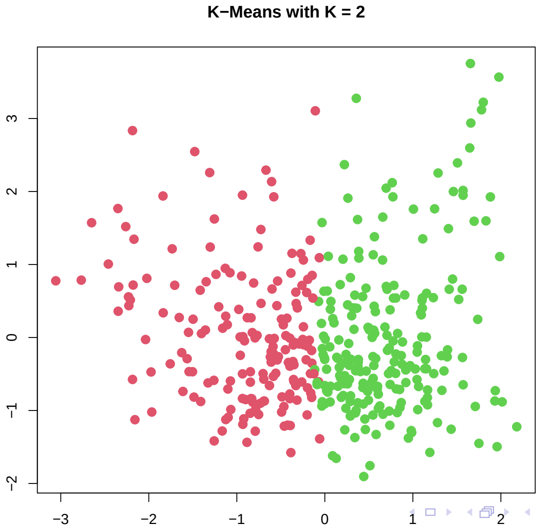
Choosing the right number of clusters



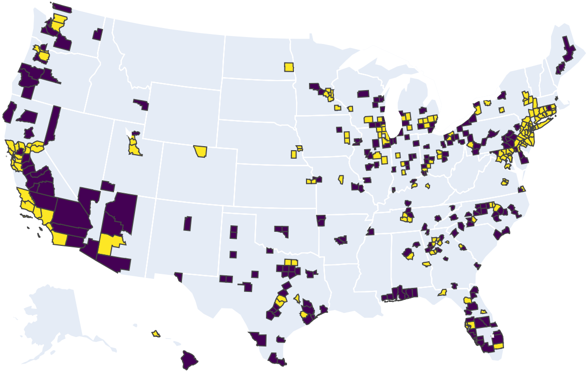




Clustering - k = 2, visualized on the first two principal components



Clustering - k = 2, visualized on the map



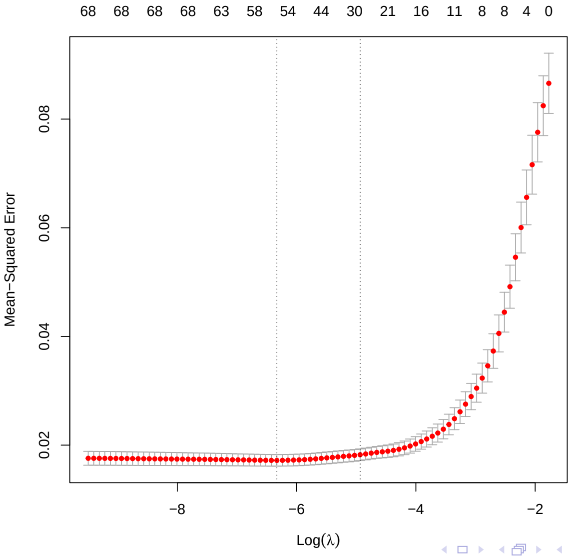
Some insights from clustering

- Clusters are not very well defined and their optimal number is not clear, neither
    - When choosing two clusters, the dividing line appears to lie almost perfectly on the first principal component, meaning that division is on economic characteristics
    - Clustering looks more like a segmentation than a grouping
    - Looking at the map, we observe similarities between coastal California and Northeast metropolis

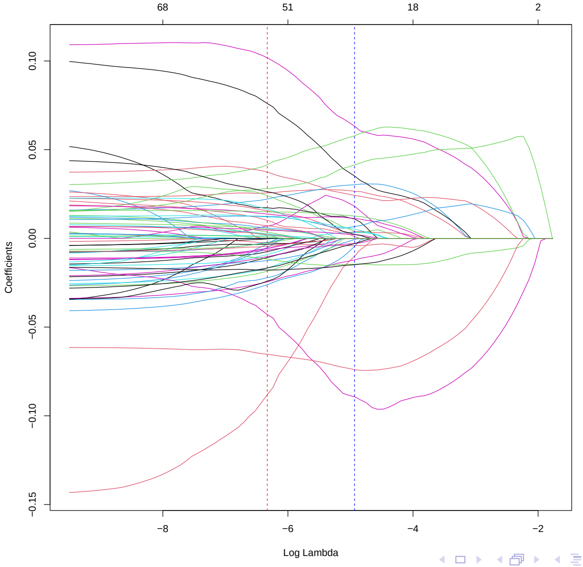
Supervised Learning

- We tried to investigate what determines the partisanship of a county
- In order to manage our high-dimensional dataset, we performed penalised regressions and feature screening which helped identify relevant features
- In the end, we have performed an Ordinary Least Square regression on the relevant features selected by LASSO

LASSO - Estimated Mean square error



LASSO - Coefficients with different λ



LASSO - Coefficients with λ.1se

(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

Screening - Variables after Sure Independence Screening

EDUC
COUPLETYPE_0
RACHSING_1
OCC_2
OCC_3
OCC_10
Evangelical.Protestant
Islam

OLS after LASSO

Variable	Coefficient	(Std. Error)
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:	$^{*}p<0.1$; $^{**}p<0.05$; $^{***}p<0.01$

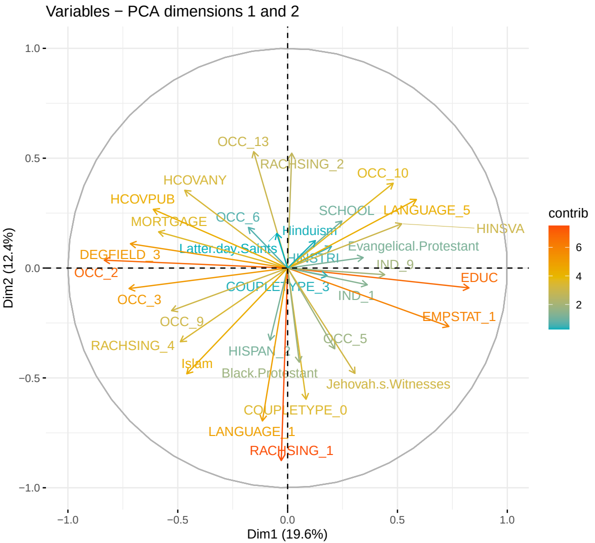
Some remarks on supervised learning

- Counties where Democrats are stronger appear more diverse in ethnicity (RACHSING_2) and familiar composition (COUPLETYPE_0 and COUPLETYPE_3)
- Counties where Republicans are stronger have higher shares of white (RACHSING_1), native English speaker (LANGUAGE_1), people graduated in business (EMPSTAT_1) and Evangelical
- Having higher share of people occupied in IT and care jobs (OCC_2 and OCC_3) appears to favor Democrats

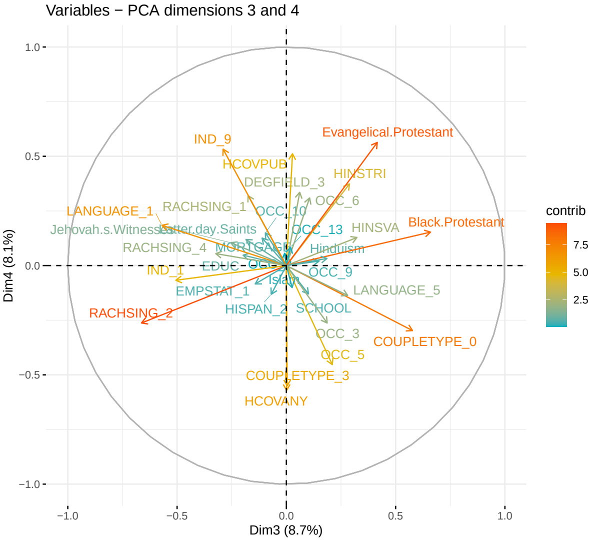
An informed review of unsupervised learning

- In order to better exploit the information given by supervised learning, we used coefficients of OLS to take a new look at counties
- We have taken features selected by LASSO and weighted them according to coefficients found with OLS
- Afterwards, we have performed techniques of unsupervised learning on this newly weighted dataset in order to look at possible differences
  - PCA
  - Clusterisation
- In the end, we have confronted clusters with electoral results

PCA (Revisited) - First two principal components



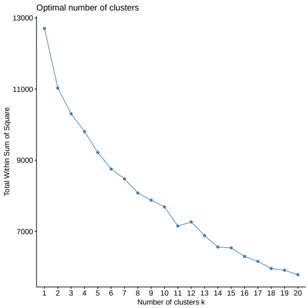
PCA (Revisited) - Third and fourth principal components

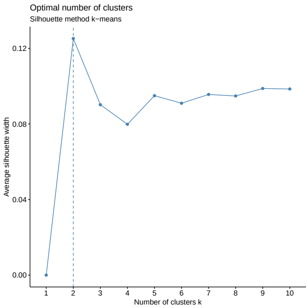


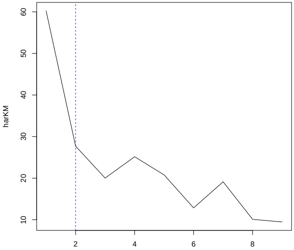
Some remarks on informed PCA

- PCA map is obviously less dense: however, economic and occupational variables are once again the main determinants of first dimension
- Other variables emerge more evidently, such as religious ones, especially on third and fourth principal component

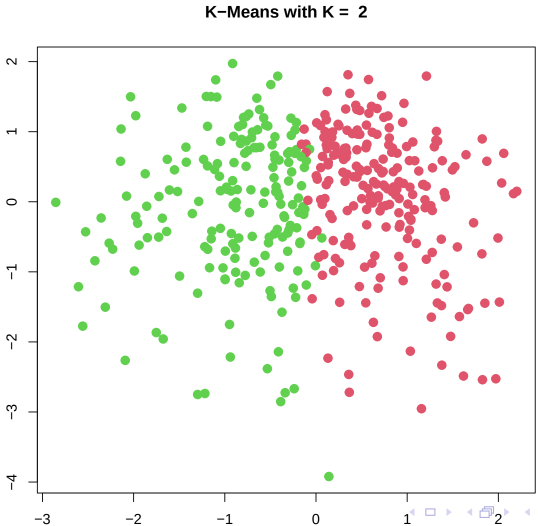
Choosing the right number of clusters



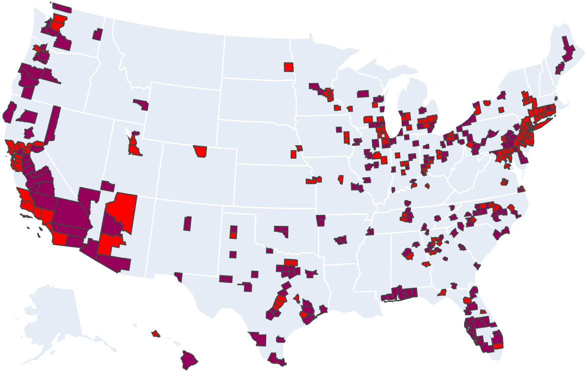




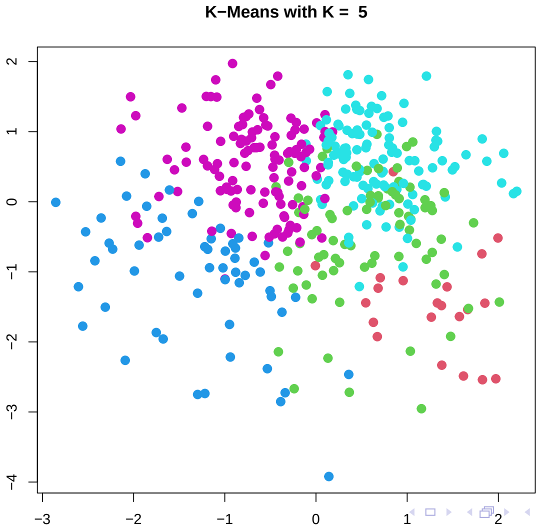
Clustering (Revisited) - k = 2, visualized in the first two principal components



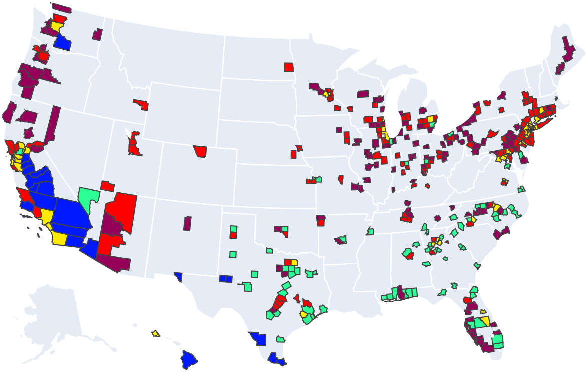
Clustering (Revisited) - k = 2, visualized in the map



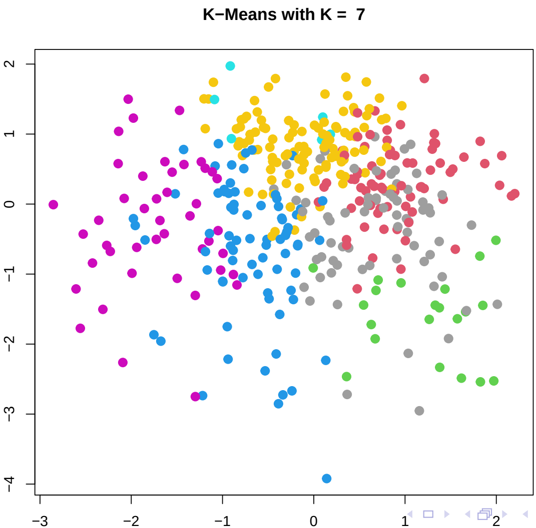
Clustering (Revisited) - k = 5, visualized in the first two principal components



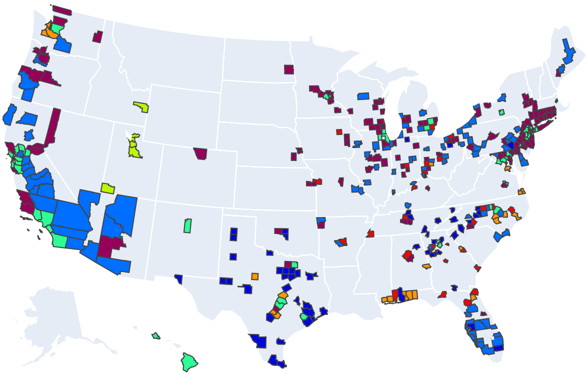
Clustering (Revisited) - k = 5, visualized in the map



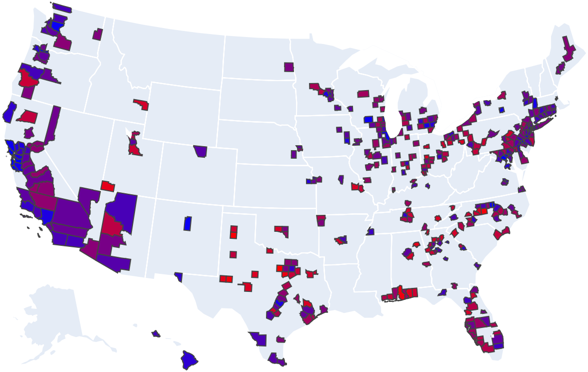
Clustering (Revisited) - k = 7, visualized in the first two principal components



Clustering (Revisited) - k = 7, visualized in the map



Real Election data - visualized in the map



Some remarks on informed clustering

- Once again, the right number of clusters is not clear
  - When looking at division in two clusters, we do not observe dramatic changes with respect to results in totally unsupervised learning
  - When we increase the number of clusters, the segmentation process seems to hold and we do not observe particular geographical patterns
  - The most explanatory pattern is the one which links clusters and the segmentation of differences between parties: that is explained also by the higher weights attributed to more electorally relevant variables

Limits and further directions

- Choice of features
  - We are pretty confident that we have considered all variables that were significant in supervised learning; however, other variables can be more informative to perform clustering
- Sample problems
  - We have assumed that samples were representative not only of the total population, but also of the population of counties **bianchi2020effect**
- Bootstrapping our results
  - We are looking forward to bootstrap our results
- Counties are not individuals and they should be treated accordingly
- Looking at evolution in time

References

- *2020 U.S. Religion Census: Religious Congregations & Membership Study.* (2023). URL: https://www.usreligioncensus.org/node/1639.
- Data, MIT Election and Science Lab (2022). *U.S. Senate Precinct-Level Returns 2020*. Version V1. DOI: 10.7910/DVN/ER9XTV. URL: https://doi.org/10.7910/DVN/ER9XTV.
- *IPUMS USA* (2024). Version Version 15.0. Minneapolis, MN. DOI: 10.18128/D014.V4.0. URL: https://doi.org/10.18128/D014.V4.0.
- James, Gareth et al. (2013). *An Introduction to Statistical Learning: with Applications in R*. Springer. URL: https://faculty.marshall.usc.edu/gareth-james/ISL/.