Politecnico di Milano
M.Sc. in Mathematical Engineering
Nonparametric Statistics
AY 2020/2021

Nonparametric Analysis of 2020 United States Presidential Elections

Ajroldi Niccolò*
Lurani Cernuschi Agostino†
Marchionni Edoardo‡

Abstract

This is a project report of *Nonparametric Statistics* course, held by Professors F. Ieva and S. Vantini, with the support of Dr. M. Fontana, at Politecnico di Milano during academic year 2020/2021. The aim of the project is to give an insight on 2020 US presidential election, in particular in understanding which factors may have influenced the election. The analysis will be devoted both to making some preliminary inference on different possible socio-economic and demographic factors at our disposal and to fit a prediction model using a part of our data as training set. We will compare it with respect to other possible models through a validation set, and, in the end, we will evaluate its predictive performances through the remaining observations, used as test set.

*Mathematical Engineering Student at Politecnico di Milano.
†Mathematical Engineering Student at Politecnico di Milano.
‡Mathematical Engineering Student at Politecnico di Milano.

Contents

1 Introduction

1.1 Dataset

The data set used in the following analysis is taken from kaggle.com [7] and was first skimmed and then enriched with US Census Bureau data [8], according to our purposes. At the end of this preliminary preprocessing step, we have at our disposal the following variables, observed in 3044 counties across the United States:

1. `Percentage 2016 of Hillary Clinton's votes and Donald Trump's votes`
2. `Percentage 2020 of Joe Biden's votes and Donald Trump's votes`
3. `Total number of voters in 2016 and 2020`
4. `Male and female population`
5. `Asian, Black, Hispanic, Native and White population`
6. `Average per capita income`
7. `Population living under poverty conditions`
8. `RUCC code`
9. `Population having completed at least a Bachelor's degree`
10. `Population density`
11. `Covid-19 cases and deaths`
12. `Tertiary sector workforce employed in professional, service, office, construction production sectors`

In particular, the RUCC code is a categorical variable that distinguishes metropolitan counties from more rural entities (see [10]). We have at our disposal also the 2020 polls by state, but we remark that we are of course able also to aggregate our data by states.

1.2 Why Nonparametric?

The motivation behind using nonparametric statistical tools is certainly first and foremost academic, since the following analysis is a part of the *Nonparametric Statistics* course held by Professors F. Ieva and S. Vantini at Politecnico di Milano during the academic year 2020/2021.
Nevertheless, we firmly believe that this coerced choice can be motivated by real evidences, which suggest that a nonparametric setting is more appropriate in the following analysis than the more usual parametric one. In particular, due to the intrinsic nature of our data (e.g. many of them are percentages), they are far from being Gaussian and also applying the most used transformation to overcome this problem, we still struggle with the normality assumption needed in the parametric setting. The numerosity of our data set may help us exploiting asymptotic results, but we would be limited in using several covariates in preliminary inference, since data are not enough to proprerly fill the complete space. Furthermore, we do not a priori believe that the relationship between the election outcome and the other covariates is linear, making us resort to more sophisticated tools that are able to spot nonlinear relationships and that rely on nonparametric or semiparametric methods.

2 Research Questions

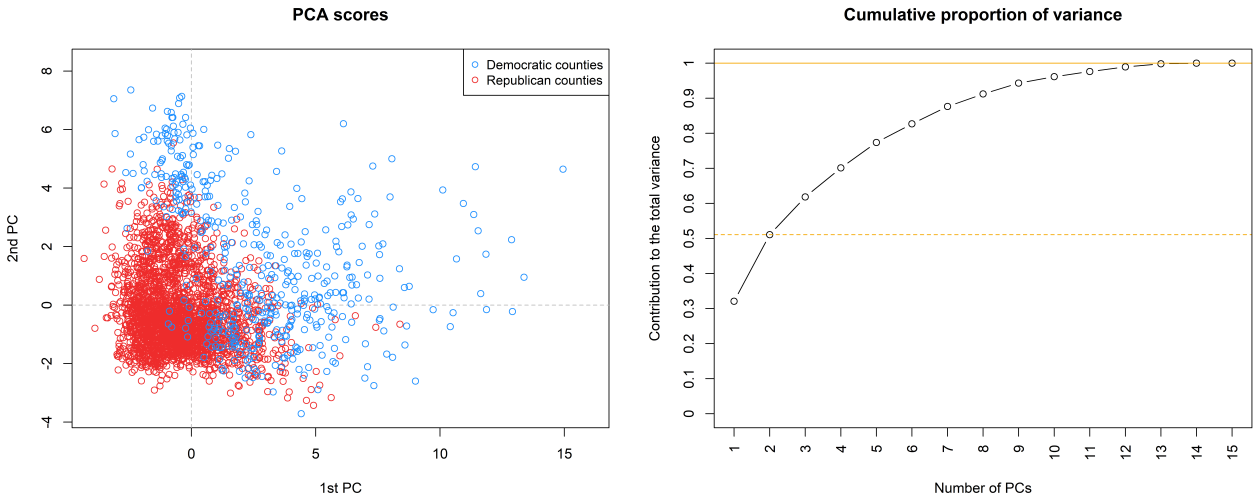
Our primary intention is to investigate factors that may have influenced the election results. In particular, we are interested in describing how voting trends differ with respect to various socio-economic and demographic indicators. A first part of the analysis, namely Section 4, is therefore devoted to testing the independence between the percentage difference between Joe Biden and Donald Trump and other variables.

Thereafter, we want to build a prediction model in order to forecast future elections, or at least to obtain reference plausible projections, based on covariates at our disposal. In Section 5 we build different models, comparing them by the length of conformal prediction intervals and other metrics on a common validation set.

Finally, after having selected the best model, we proceed by testing its performances on a previously selected test set, in which we have included counties that Donald Trump charged for fraudulent voting, in order to spot possible suspicious results in those places according to our model.

3 Data Exploration

In order to first visualize data, we rely on Principal Component Analysis. The first two principal components are able to explain only 51% of total variability, nevertheless, as reported in Figure 1a, we can already see that there is a rough separation between counties won by Democratic party and those won by Republicans.
Moreover, by further investigating the loadings of the first two PC's, we can identify variables with a positive score on such components. Such preliminary analysis already points out some interesting patterns that we will study and test in the following sections.



(a) Counties scores in the first two PC's

(b) Cumulative proportion of explained variance

Figure 1: PCA

4 Preliminary inference

The aim of this first analysis is to explore data and to understand which factors could be relevant for modelling the election outcome across United States.

From now on, the target variable of the whole analysis will be the percentage difference between Joe Biden and Donald Trump. We will call y_i the observed value of the target variable in county i, with $i = 1, ...n$, and $n = 3044$ being the number of counties. Let Y denotes the (absolutely continuous) random variable supported in $[-1, 1]$ such that $y_1, ..., y_n$ are its independent realizations.

In our preliminary inference analysis we will rely on different nonparametric tools, for the motivation behind this choice refer to the dedicated section (subsection 1.2).

4.1 Distance Correlation Index and Permutational Independence Test

The distance correlation index (dCorr) is a generalization of the most known linear correlation coefficient ρ (see [5]). Distance correlation presents more interesting properties for studying the dependence between random variables with respect to the linear correlation coefficient, still preserving a compact representation. Indeed, dCorr takes still values in $[0, 1]$, but unlike the linear correlation coefficient, it is defined for two (real) random vectors in arbitrary dimensions. Morereover, this new index is able to spot nonlinear and nonmonotonic dependence among random vectors X and Z and it is null if and only if $X \perp\!\!\!\perp Z$.
Thanks to those nice properties, dCorr can be used to build some test on independence of random vectors.
In particular, let $X \in \mathbb{R}^p$ and $Z \in \mathbb{R}^q$

$$H_0 : X \perp\!\!\!\perp Z$$
$$H_1 : X \not\perp\!\!\!\perp Z$$

is reformulated using dCorr index as follows

$$H_0 : \mathrm{dCorr}(X, Z) = 0$$
$$H_1 : \mathrm{dCorr}(X, Z) \neq 0$$

Given n realizations of X and Z, it is possible to build a permutational test for the above hypothesis, relying as test statistic on the sample version of dCorr coefficient (see [5]) and using as likelihood invariant transformation under H_0 the permutation of X and Z separately. The cardinality of the set of permutations is hence $n! \cdot n!$. It is noteworthy that the so-built test presents good empirical power on simulated samples comparable with the most-known parametric tests.
The test has been performed setting as Z our target variable Y described above and as X the different factors we think might have had an influence on the election outcome. We compute the p-values of the permutational tests simulating though Monte Carlo the distribution of the test statistic under H_0, setting the number of permutations to $\lfloor 200 + \frac{5000}{n} \rfloor$, as in the simulated examples in [5]. As factors we tested

- Ethnicity: X is a vector of $\mathbb{R}^5$ representing the percentage of the population belonging to the groups Black, White, Native, Hispanic, Asian, and Pacific
- Education: X represents the percentage of the population having at least a bachelor's degree
- Sex: X represents the percentage of females
- Population density: X represents the population density
- Income: X represents the average income per cap.

- Poverty percentage: X represents the percentage of population living under poverty conditions
- Tertiary sector: X represents the percentage of tertiary work force employed in the sectors professional, service, office, construction and production
- Difference in 2020 polls: X represents the percentage difference in votes between Joe Biden and Donald Trump according to 2020 polls
- Difference in 2016 results: X represents the percentage difference in votes between Hillary Clinton and Donald Trump in 2016 elections
- Number of voters: X represents the number of voters in 2020

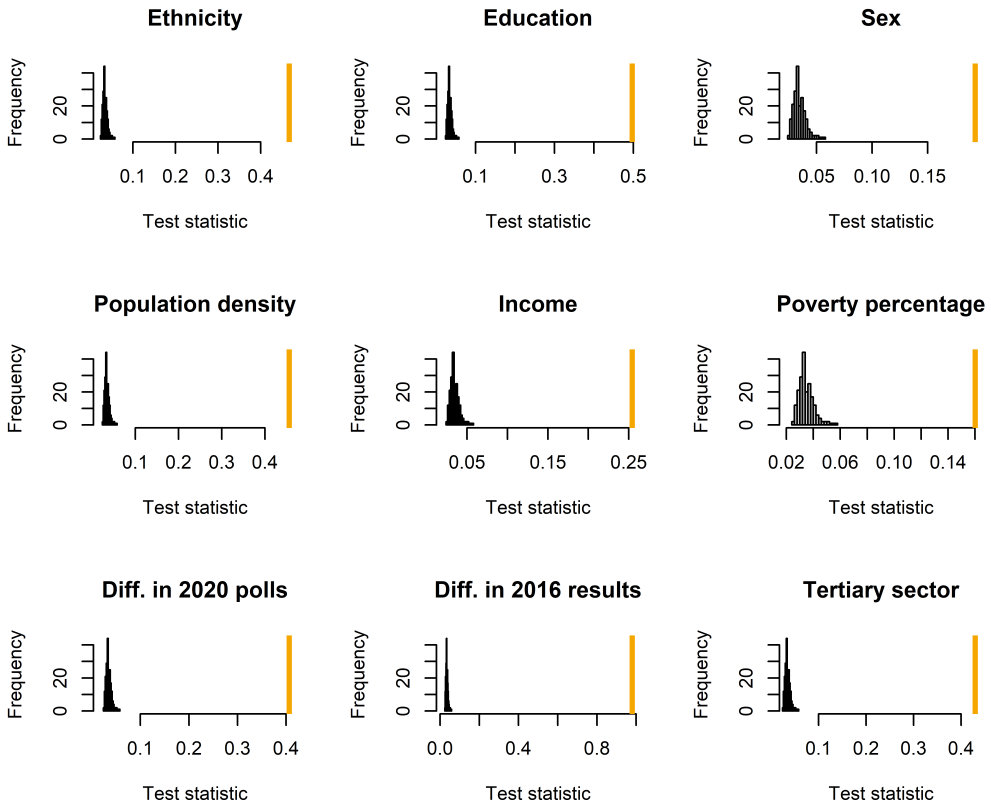


Figure 2: Permutational distribution under H_0 and observed value of the test statistic

Group	Observed Value of test statistic	p-value
Ethnicity	0.47	0
Education	0.50	0
Sex	0.51	0
Population density	0.46	0
Income	0.25	0
Poverty percentage	0.16	0
Tertiary sector	0.43	0
Diff. in 2020 polls	0.41	0
Diff. in 2016 results	0.98	0

Table 1: Results of permutational independence test

In Table 1 we note that for all factors we reject the null hypothesis at every level and hence we have statistical evidence that there is dependence between our target response Y and the considered factors.
It is remarkable that considering as factor X the difference of percentage between Joe Biden and Donald Trump in 2016, the dCorr index takes value close to 1. This, thanks to the property of dCorr index of taking values in the bivariate case always less or equal to the linear correlation coefficient (see [5]), suggests that the dependence between this two variables is close to be linear.

4.2 Permutational ANOVA on Ethnic Groups

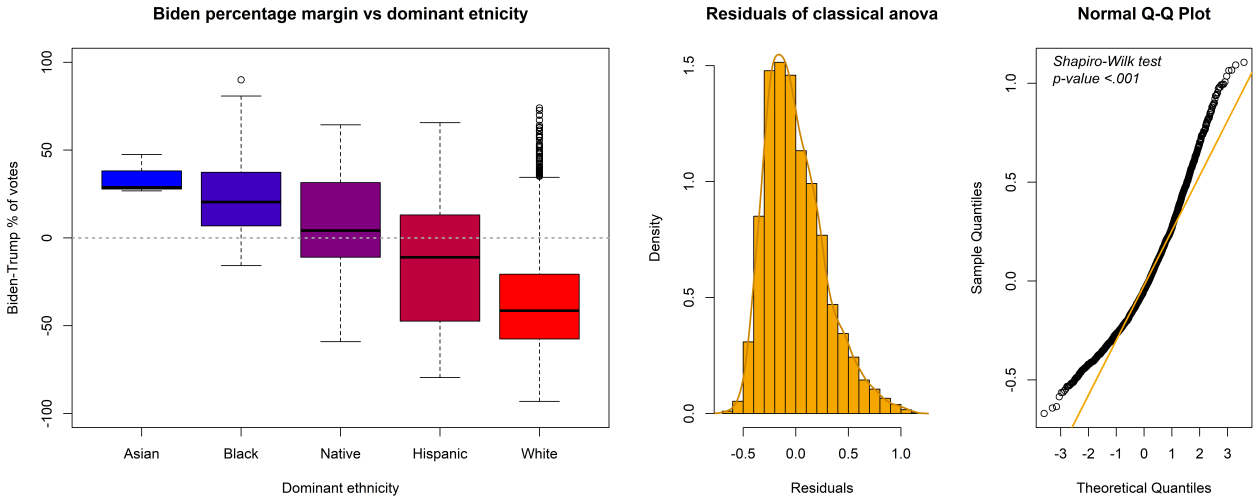
In addition to the previous analysis, we want to go deeper into the study of the role of the different communities on the outcome of the election. The motivation behind this choice lies in the fact that we firmly believe that ethnicity might be a relevant factor for our response variable. For each county, we still analyze data regarding the percentage of the following ethnic groups: Asian, Black, Hispanic, Native and White.
The first step of this analysis consists into dividing counties based on the most present community. In such a way we separate observations in four groups and we can now test if different communities vote in a different way.
Let $X_{i,j}$ be the percentage difference between Biden and Trump in the i-th county of the j-th ethnic group.
We consider first a parametric ANOVA setting, thus assuming:

$$X_{i,j} \overset{iid}{\sim} \mathcal{N}(\mu_j, \sigma^2) \quad i = 1, ...n_j, \quad j = 1, ...5$$

We want to test:

$$H_0 : \mu_1 = \mu_2 = ... = \mu_5$$
$$H_1 : \exists j, k \in \{1, ..., 5\} \text{ s.t. } \mu_j \neq \mu_k$$

Concerning the assumptions, Shapiro-Wilk normality test confirms what was already evident from residuals histogram and qq-plot in Figure 10b: residuals are not normally distributed and thus the validity of the test has to be questioned. Furthermore, we perform the Bartlett test of homogeneity of variances and getting a p-value of 2.099e-06 we have statistical evidence of eteroscedasticity among the different groups. Last, we observe that some groups, namely Asian and Native, have very low numerosity, therefore, we cannot rely on limit theorems. Motivated by the previous considerations, we take into account a Permutational ANOVA. We prefer the Permutational approach than the Bootstrap one, because of the aforementioned small sample sizes of some groups, which undermines the asymptotic validity of the second setting.



(a) Boxplot of different groups

(b) Parametric ANOVA residuals

Figure 3

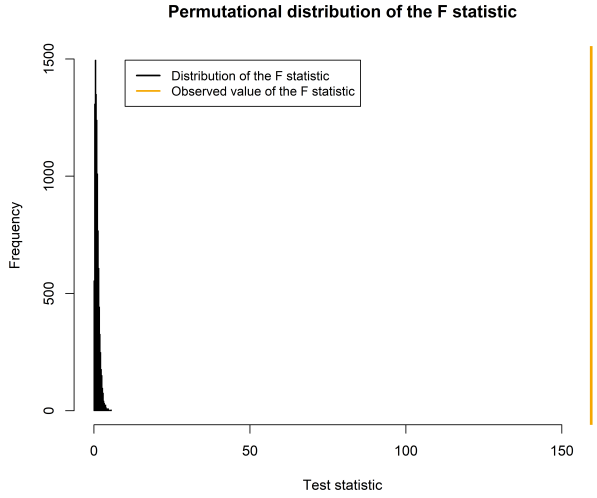


Figure 4

We hence assume

$$X_{i,j} \overset{iid}{\sim} F_j \quad i = 1, ...n_j, \quad j = 1, ...5$$

We want to test:

$$H_0 : X_1 \overset{d}{=} X_2 \overset{d}{=} ... \overset{d}{=} X_5$$

$$H_1 : \exists j, k \in \{1, ..., 5\} \text{ s.t. } X_j \overset{d}{\neq} X_k$$

As test statistic we use the Fisher's F. The set of likelihood invariant transformations of the test statistic is the set of permutations of the categorical labels of data.
In Figure 4 We find strong evidence to reject the null hypothesis and to conclude that different ethnic groups vote in a different way.

4.3 The Role of 2016 Margin

In subsection 4.1 we hypothesised a linear relationship among our target variable, i.e. the democratic percentage margin in 2020, and the one of 2016. This average behaviour is in not surprising in the context of US elections. Indeed, the almost bipolar party system of US leads to sure outcomes in many areas, where electors are fond of the same faction. On the other hand, it is well known that the election results is determined by a minor part of areas where there is no dominant orientation.
To better investigate this believed linear relationship, we fit in the framework of this preliminary analysis a simple linear model, where the response is the democratic margin in 2020, i.e. our target variable, and the democratic margin in 2016 will play the role of covariate.

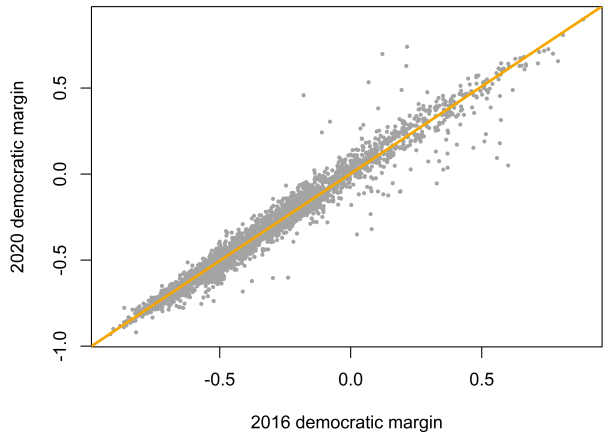


Figure 5: Regression line

In figure 5 the estimated line is plotted, we report an adjusted coefficient of determination of $R^2_{adj} = 0.961$. The qq-plot of residuals highlights the presence of big fat tails and Shapiro-Wilk test provides a p-value $< 2.2e-16$. Hence we perform first a permutational test build on the global F-statistics of the regression, setting as likelihood invariant transformation the permutation of residuals and getting a MC p-value with 10 000 replicates equal to 0, which makes us validate the significance of the regression. As far as the significance of the estimated coefficients is concerned, we estimate the bootstrap distribution of their estimators through a MC resampling from the residuals with still 10 000 replicates.

	OLS estimation	lower bound	center	upper bound
β_0	0.002979	-0.000270	0.002983	0.006236
β_1	1.014329	1.007040	1.014300	1.02156

Table 2: OLS estimation of regression coefficients and their reverse bootstrap percentile confidence intervals at a level of $\alpha = 0.05$

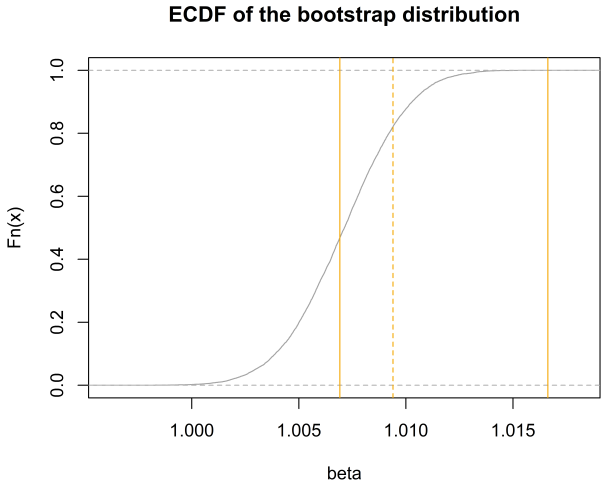
The results in table 2 shows that there is statistical evidence at a level of $\alpha = 0.05$ to affirm that β_1 is different from 0, whereas the intercept estimation is not significant at the same level. We perform a regression through the origin (i.e. intercept-free). We do not report the $R^2_{adj.}$ since in intercept-free regressions it has no interpretable meaning. We still present the lack of normality for the residuals: the qq-plot presents big fat tails and the Shapiro-Wilk test has again a p-value $< 2.2e-16$. To validate the regression, we proceed through a MC estimation with 10 000 replicates of the bootstrap distribution of the estimator of the coefficient β though resampling from the residuals, getting the results reported in table 3.

	OLS estimation	lower bound	center	upper bound
β	1.009398	1.006915	1.011771	1.016628

Table 3: OLS estimation of intercept-free regression coefficient and its reverse bootstrap percentile confidence intervals at a level of $\alpha = 0.05$

We note that our coefficient is significant at a level of $\alpha = 0.05$ and really close to 1, confirming the a priori belief that due to the political framework of United States, the vote on average does not significantly change over subsequent elections, but it is decided over some swinging areas.

(a) Regression line



(b) ECDF of the MC simulated bootstrap distribution of the estimator $\hat{\beta}$

Figure 6: Regression through the origin

Despite this fact, we think that it is still interesting to attempt to characterize the election outcome through the different covariates at our disposal, first being able to find the most significant ones and then being able to

predict the outcome also on areas where the result is not straightforwardly predictable on the basis of the past election.

5 Regression model

In the previous sections we have investigated the relationship between the target variable, i.e. the democratic percentage margin in 2020, and the other covariates. Having selected the most significant regressors, we proceed by fitting a Generalized Additive Model (GAM) and analyze further how such variables characterize influence the target one.

A first crucial step consists in separating data into a training set, which will be used to fit different models, a validation set, useful to provide an unbiased evaluation of performances and to tune hyper-parameters, and a test set, that will be used only once to judge performances of the best selected model. Notice that the split is done by stratified sampling from the target variable, in order to obtain three sets that are representative of the whole population.

5.1 Model Selection

We first build a GAM, considering the whole set of p explanatory variables carefully selected in the previous section, namely $x_{1,i}, ..., x_{p,i}$, observed for each county $i = 1, ..., n_{\text{train}}$. For each term we use cubic a penalized regression spline, here denoted as s, where the smoothing parameter estimation problem is solved by Generalized Cross Validation (GCV) criterion. The model is fit using `mgcv` R-package by penalizing likelihood maximization.

$$y_i = \sum_{j=1}^{p} s(x_{j,i}) + \epsilon_i, \quad i = 1, ...n_{\text{train}}, \quad \epsilon_i \overset{iid}{\sim} \epsilon, \quad \mathbb{E}[\epsilon] = 0$$

As expected, the regressor variable describing the democratic margin in 2016 is a very strong predictor. Indeed, as we have extensively analyzed in subsection 4.3, in most of the counties, 2020 votes are almost indistinguishable from the ones of the previous election. Moreover, we observe a very high adjusted coefficient of determination: $R^2_{adj} = 0.979$. We conclude that the above model is likely overfitting data, predicting for 2020 almost exactly the results of 2016 and certainly not useful at our purpose of characterizing the election outcome of not easily predictable results through demographic and socio-economic indicators. Motivated by such considerations, we remove from the regressors the democratic margin in 2016 and consider the reduced model.

We would like to test the significance of each smooth term. We notice that p-values for the significance of each smooth term, based on Marra and Wood (2012) [3] extension of Nychka's (1988) [7] work, are computed relying on the normality assumption of errors in the model. In Figure 7 we report the distribution of residuals. The Shapiro-Wilk test provides a very low p-value and thus we conclude that gaussianity assumption is not met.

Even though posterior normality of parameters may be justified by large sample size, we recall that p-values are approximate and neglect smoothing parameter uncertainty, thus they are likely to be too low when smoothing parameter estimates are highly uncertain. For all these reasons, we prefer to proceed by permutational inference for the significance of smooth terms.

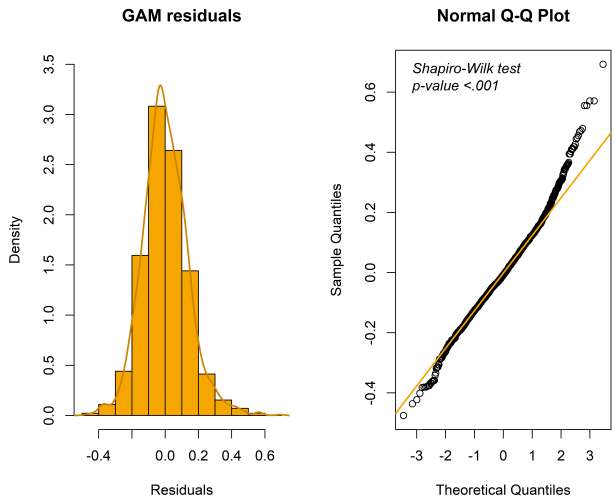


Figure 7: Residuals of Full GAM

5.2 Permutation Test for Significance of Smooth Terms

For each regressor x_j, $j = 1, ..., p$ in the GAM we perform the following test:

$$H_{0j} : s(x_j) \text{ is not significant} \quad vs \quad H_{1j} : s(x_j) \text{ is significant}$$

As justified before, we implement such test in a permutational fashion, employing as test statistic the F_j-statistic related to the significance of the smooth term $s(x_j)$. Under the null hypothesis H_{0j}, the residuals of the reduced model are asymptotically exchangeable, hence, relying on a parallel implementation of the Freedman and Lane scheme, it is straightforward to estimate the distribution of F_j under the null hypothesis.
Notice that we are testing multiple assumptions at the same time and in order to guarantee an overall level of the simultaneous test, we opted for a control of the family-wise error rate for multiple comparisons through the Benjamini-Hochberg FDR correction of p-values [6]. In Table 6 we report p-values for the significance of each smooth term, corrected as explained before.

Regressor x_j	Adjusted p-value
Diff. in 2020 Polls	0.006
Number of votes	0.065
Income per cap	0.036
Women%	0.068
White%	0.006
Hispanic%	0.010
Black%	0.000
Bachelor%	0.000
Population density	0.004
Construction	0.008
Poverty%	0.008
RUCC code	0.968

Table 4: Permutation test for regressors significance

5.3 Conformal Prediction and model selection

Having a good predictive model of the election outcome is one of the purposes of this analysis. In particular, we would like to assess whether there is evidence of fraud on some specific area as claimed by Donald Trump by predicting the outcome through the fitted model and compare it with the real one.
We remind that all the contested counties are part of the test set, so that the model has been trained only on the non-contested ones. In particular, through the model previously selected we want to predict the outcome of this year election on the test set that includes counties from Wisconsin, Nevada, Michigan, Pennsylvania and Arizona, i.e. the states believed to be interested of election rigging.

In addition to the previous GAM, we will fit different models and compare their performances on a common validation set in order to obtain an unbaised estimate of their accuracy. As metrics for model evaluation we will consider mean squared error (MSE) and the length of conformal prediction intervals, with coverage apriori fixed at 90%. In particular, having a large number of observations, we opt for split conformal prediction approach, that in this framework should give results similar to full conformal, but with better computational performances. In this setting, absolute value of the regression residuals are used as non-conformity scores and an equal split is employed.

We first consider the GAM built in the previous section. At first, by evaluating the distance between pointwise prediction on the validation set and the true observed value, we are satisfied with the performances of the model. Indeed, the mean distance between the pointwise prediction and the observed value on the test set is little less than 10%, which shows that, on average, we do not make really accurate predictions, neither completely useless ones. On the other hand, measuring prediction intervals length, we observe on average a length of 0.47. Each interval is hence covering almost half of the support of the target variable, becoming practically useless. This might be due first to the limited amount of covariates we have at our disposal, as well as to the choice of the non-conformity measure. Concerning this second point, we do not have the chance to further investigate this issue, since it goes beyond the scopes of this project. Moreover, it is noteworthy how this performance may be improved using 2016 election results, but we refer to Section 5 for the explanation of why this covariate has been dropped.

In the above context, we would like to try different models to approach the problem, hoping in better performances. For such reason, using the set of regressors selected in previous sections, we repeat the analysis using a Random Forest and a Support Vector Machine getting almost the same results in terms of MSE and intervals' length for both methods. The poor performances of the other models confirms our previous conclusions. For this reason and since the GAM is simpler and more interpretable, we select it among the others.

Model	MSE	PI length
GAM	0.0196	0.4787
Random Forest	0.0198	0.4905
SVM	0.0207	0.4852

Table 5: Performances on validation set of different models

Neglecting the problem of the prediction intervals range, we investigate the prediction performance in the incriminated counties. Focusing on the pointwise prediction, it seems that there is no trend in the predicting error of the counties of Arizona, Nevada and the other states. Furthermore, the model is more precise over republican counties, probably due to the high numerosity of those, whereas the biggest predictive errors, which are localized as the outliers of the residuals, are from different states, but all of them are Democratic counties.

State	County	Difference 2020	Predicted	Contested
California	Alpine	0.320	-0.216	0
California	Lake	0.265	-0.336	0
California	Marin	0.670	-0.026	0
California	San Francisco	0.726	0.197	0
California	Sonoma	0.571	0.137	0
DC	Washington D.C.	0.900	0.323	0
Colorado	Elbert	-0.503	-0.089	0
Colorado	Pitkin	0.521	-0.022	0
Hawaii	Honolulu	0.458	0.816	0
Iowa	Des Moines	-0.085	-0.498	0
Iowa	Marshall	-0.077	-0.477	0
Iowa	Muscatine	-0.071	-0.456	0
Illinois	Richland	0.166	-0.207	0

Table 6: Outlier identified from prediction

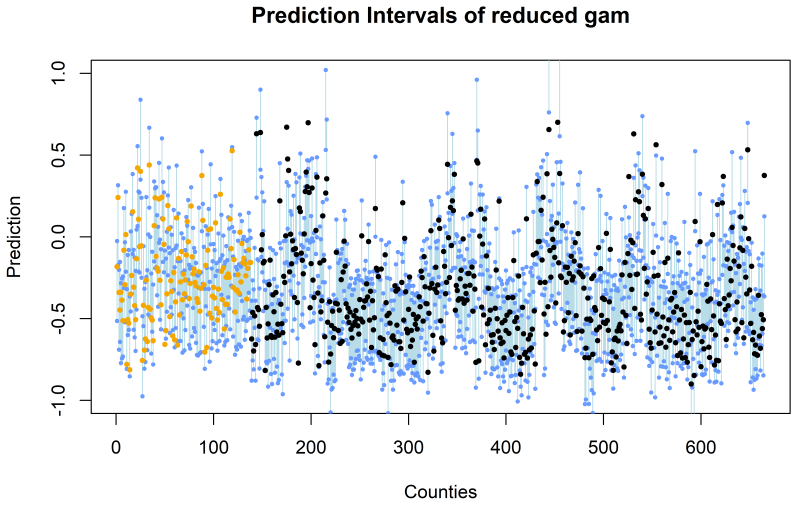


Figure 8: Conformal prediction intervals

It is worth to say that in Alpine the total number of votes was 748 and there is a high presence of Native, more than 20 %, whereas Honolulu has 41% of Asian and 9% of Pacific people. On the other hand, San Francisco and Washington have an incredibly high percentage in favour of J. Biden and we do not expect to perform well in prediction on such extreme observation.

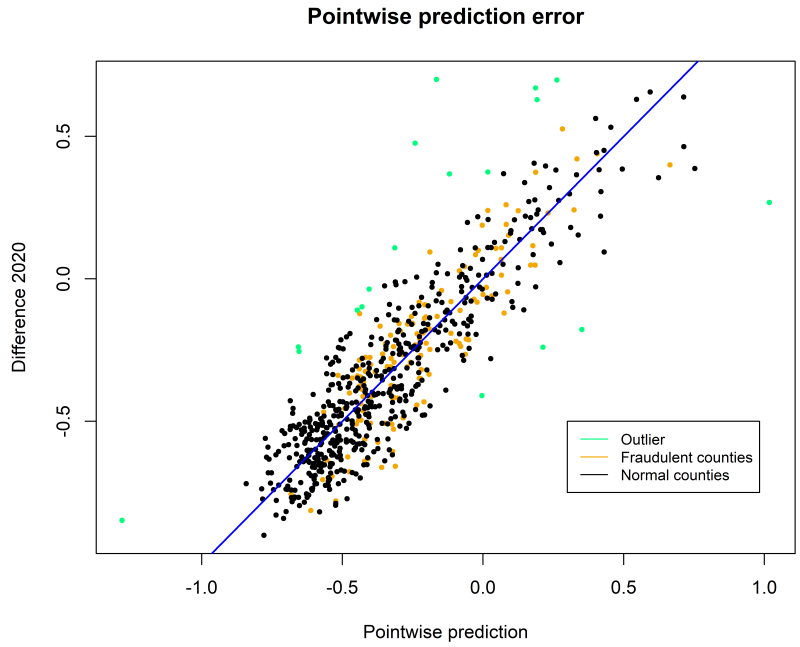


Figure 9: Pointwise prediction vs observed values

Since we are not able to construct useful prediction intervals because of the big range of them, we proceed by building some tests in order to try to make more statistical consideration out of the pointwise prediction. The main idea is that, if there has been a fraud in those counties, and, more importantly, if our model is able to capture it, there must be some difference between the distribution of the two groups errors.
We note that a systematic positive error in contested counties might be a sign of fraud. Since we do not have the above errors, we will use the corresponding residuals. We set those residuals equal to the observed values minus the predicted ones. In particular we proceed by using two different test statistics, the mean and the median of the above residuals in the two groups, i.e. the contested counties and the non-contested ones, and we argue that the test statistics take bigger values in the fraudulent counties. We build two different permutational tests.

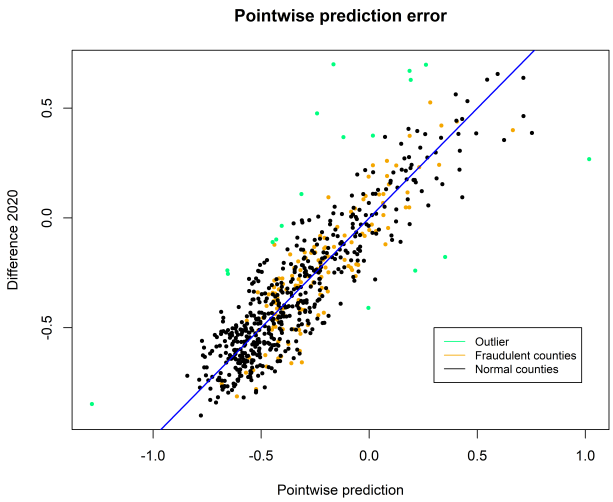
Test 1:

$$H_0 : \varepsilon_{fraud} \stackrel{d}{=} \varepsilon_{normal}$$
$$H_1 : MED(\varepsilon_{fraud}) > MED(\varepsilon_{normal})$$

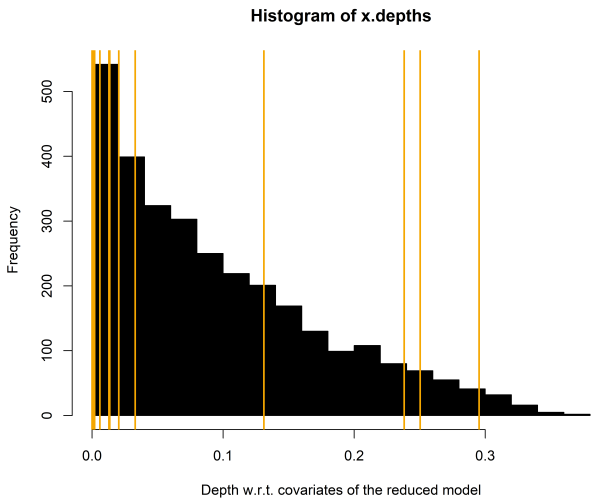
Test 2:

$$H_0 : \varepsilon_{fraud} \stackrel{d}{=} \varepsilon_{normal}$$
$$H_1 : \bar{\varepsilon}_{fraud} > \bar{\varepsilon}_{normal}$$

In both cases we do not have enough evidence to reject the null hypothesis at any level.



(a) Pointwise prediction error



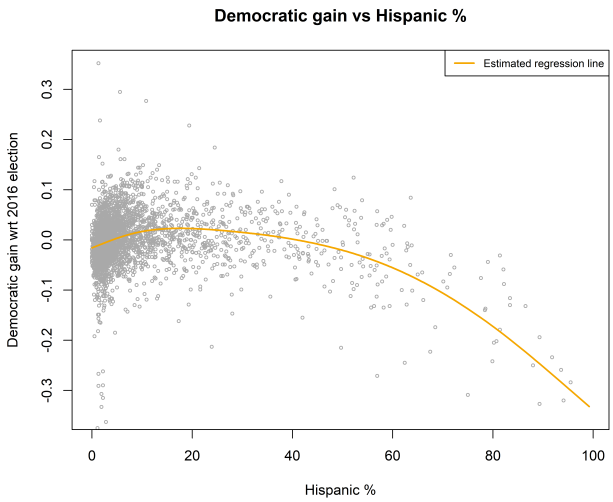
(b) Histogram of the depth of the worst predictive counties

Figure 10: Analysis of prediction

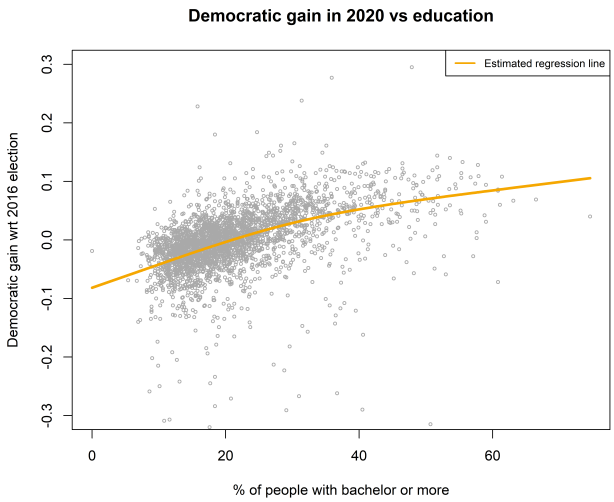
6 Conclusions

This last tests do not really answer our research question and with the model that we have built we are not able to say if what Donald Trump charges are true. This is probably caused by high variability of data, a poor choice of non-conformity measure and probably by the fact that there is an important part of unpredictability in elections output, which is not explicable using only socio-economic and demographic data.
To conclude our work we train a new model to investigate how votes have changed over the last four years. The new response is the difference between the 2016 elections margin between Hilary Clinton and Donald Trump and the one of 2020 between Joe Biden and Donald Trump.

We fit a GAM model that, in addition to the already used covariates, takes into account also the number percentage of Covid-19 cases and deaths. The result is not good in terms of goodness of fit, $R^2 = 0.411$, but reveals some interesting insights. In particular, among all the variables the two significant ones turn out to be the percentage of Hispanic and the percentage of adults with a bachelor or higher. In particular the spline fitted on the percentage of Hispanic shows that as their concentration grows, the Republican favor grows (see fig. 11a). As far as the percentage of adults with a bachelor degree of higher is concerned, as the population gets more educated, this time the favour moves to the Democratic side (see fig. 11b). We point out that the Covid-19 variables as well as all the others are not significant.



(a) Percentage of Hispanic people



(b) Percentage of people with a bachelor or more

Figure 11: GAM on the difference of Democratic margin between 2016 and 2020

7 Code structure

All the analysis were implemented in
`R Core Team (2020). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.`

Codes are publicly available in this Github repository.
The code is structured as follows:

- `data`: folder containing original datasets and RData file from other analysis
- `code`: folder containing original datasets and RData file from other analysis
  - `Models`: scripts for model selection and forecasting, analysis for Section 5
  - `Nations`: scripts for analysis on a national level, not reported in this report
  - `Inference`: scripts for the inference part
- `data clean`: folder containing scripts for data manipulation, cleaning and visualization
- `pics`: folder containing most of the images for this report

The script `install.R` provides automatic installation of required R packages.

8 References

[1] James, G., Witten, D., Hastie, T., and Tibshirani, R. (2013). An Introduction to Statistical Learning with applications in R. Springer.

[2] Wood S.N. (2017). Generalized Additive Models: An Introduction with R (2nd edition). Chapman and Hall/CRC Press.

[3] Marra, G., and Wood, S.N. (2012). Coverage Properties of Confidence Intervals for Generalized Additive Model Components. Scandinavian Journal of Statistics, 39, 53-74.

[4] Pesarin, F., Salmaso, L. (2010). Permutation tests for complex data: Theory, applications, and software. Wiley.

[5] Székely, G. J., Rizzo M. L., and Bakirov, N. K. (2007). Measuring and testing dependence by correlation of distances. The annals of statistics, 35(6), 2769-2794.

[6] Benjamini, Y., and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the Royal Statistical Society, 57(1), 289-300.

[7] Nychka, D. (1988). Bayesian Confidence Intervals for Smoothing Splines. Journal of The American Statistical Association, 83, 1134-1143.

[8] Green, P.J., Silvermann, B.W. (1994) Nonparametric Regression and Generalized Linear Models. A roughness penalty approach. Chapman and Hall.

Data sources:

[7] https://www.kaggle.com/etsc9287/2020-general-election-polls

[8] https://www.census.gov

[9] https://www.ers.usda.gov

[10] https://www.ers.usda.gov/data-products/rural-urban-continuum-codes.aspx