The Impact of COVID-19 in 2020 US Presidential Election

Victor Fuentes and Fernando Garcia

Introduction

The 2020 United States presidential election was the 59th quadrennial presidential election, held on Tuesday, November 3, 2020. The Democratic Party of former vice president Joe Biden and the U.S. senator from California Kamala Harris defeated the incumbent Republican president Donald Trump and vice president Mike Pence. Trump became the first U.S. president since George H. W. Bush in 1992 and the eleventh incumbent president in the country's history to lose a bid for a second term. This project attempts to shed light on why former President Trump lost the election, despite many views indicating the contrary.

We can't begin to explain this result without considering the current COVID-19 pandemic and the response, or lack thereof, by the former President. Did voters act irrationally and blamed Trump for the pandemic, or did voters penalize Trump for the lack of response and leadership that's expected under such a crisis? It is no trivial question, and it's beyond the scope of this project. However, we will explore previous trends in U.S. elections, use sentiment analysis to understand Trump's behavior before and during the pandemic and utilize econometric tools to find correlations between vote margins and socio-demographic characteristics.

Part II introduces the data and methodology used in this study to perform the tasks mentioned above. Part III presents the analysis and econometric results. Part IV summarizes findings, and part V presents the conclusions of this study.

Data & Methodology

The `code` folder in our project repository contains six files allowing total reproducibility of our work. The first one, `0_download_data`, portraits the retrieving process of the ten datasets used in this project. All files created or downloaded are stored in the `input` folder mainly using a `csv` format.

Historical overview

We started calling historical electoral data. Hence, we pulled out information from the MIT Elections Lab through URL codes about the Presidential election results from 1976 to 2020 at the state level. Complementary, we scraped the Electoral College website to obtain electoral votes numbers using the `rvest` library. In the `1_wrangling_data` file, we performed some data cleaning and reshaping steps to merge the two mentioned datasets properly. The output data frame containing 9 columns and 2,457 rows is available in the `intermediate` folder for use as the primary input for an interactive plot using the `shiny` library. We constructed the user interface (front-end) and the server (back-end) in the app file. Overall, this interactive map shows popular and electoral votes by candidates and states during 44 years.

Analyzing 2020 results by county

We also downloaded the 2020 Presidential election results by county from a Github repo fed with information from the prominent U.S. newspapers (i.e., New York Times). We retrieved socioeconomic and demographic data from the Census Bureau using the `tidycensus` library. From the American Community Survey 5-Year Data, we downloaded information about total population, gender, age groups, educational attainment, race and ethnicity, place of birth, median household income, and Gini index. We also downloaded employed and unemployed people for 2019 and 2020 by month and county from the Local Area Unemployment Bureau (BLS) using the `BLSapi` library. Furthermore, COVID-19 statistics were available at the Coronavirus Resource Center at the John Hopkins University. We pulled out daily confirmed cases and deaths by county through their Github page. Additionally, we downloaded data from the Community Mobility Report at Google to depict the impact of spread diseases prevention measures taken across the country.

With the downloaded data, we used regular expressions to decode information identifiers through the `1_wrangling_data` file. For instance, the Census data contained very disaggregated information, thus requiring reshaping to achieve a one-state one-row structure. On the other hand, BLS information included a single variable including county code, month, and type of employment status. While the COVID-19 disease data offered daily data using a column for each day, the community mobility data included each geographical unit across the time using one for each day. After wrangling all these data, we successfully constructed a dataset containing 24 columns and 2456 rows stored in the `intermediate` folder for use in this project's plotting step.

Using the `2_plotting` file, we created four choropleth maps at a county level assessing the distribution of election results favorable to the Republican party, the change in the unemployment rate between 2020 and 2019 for the first 10 months, the number of COVID-19 confirmed cases, and the difference in the community mobility index between the peak of deaths by Coronavirus and a baseline prior the spread of disease. Additionally, based on the two first spatial visualization pieces, we developed a scatter plot between the electoral margin results and the unemployment rate change. As the last step of the project's scope analysis, in the file `4_analysis`, we delivered three linear regression models and included a draft of a tree regression model.

Performing text analysis

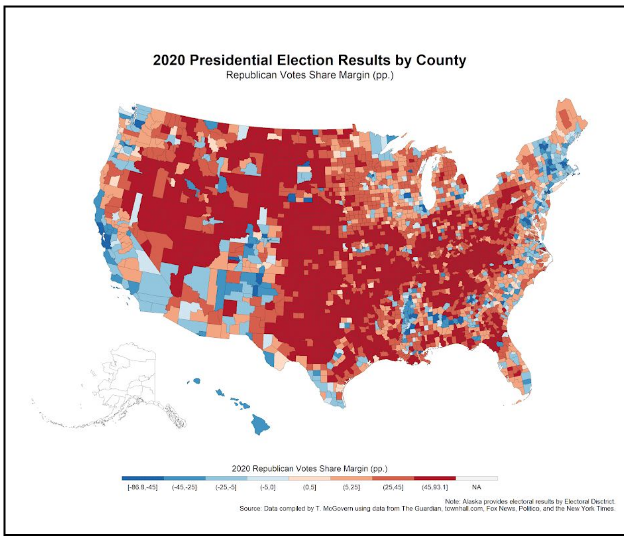
To proceed with the text analysis task, we retrieved a constructed dataset of the news sentiment based on articles from 16 U.S. major newspapers scraped for the period 1980-2020 by a study from the Federal Reserve Bank of San Francisco. We used the news' sentiment as a baseline of the main sentiment in the country.

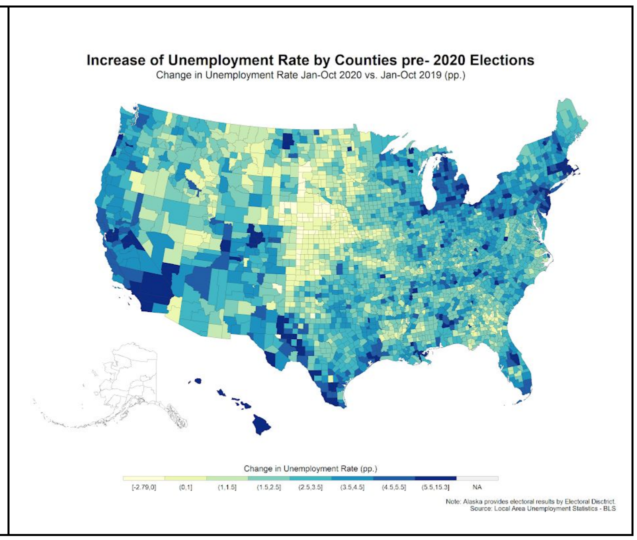
We downloaded all tweets from the former President compiled by TheTrumpArchive for analyzing his sentiment through one of his main social media platforms. We extracted Trump's tweets and performed text prophylaxis for removing `HTML` characters and additional symbols. With the cleaned data, we tokenized each tweet by sentences using the `tidytext` library and parsed them using the `spacyr` library. After lemmatizing and removing stop-words, we used the last version of the AFINN lexicon to calculate his words' sentiment, then aggregating them through a weekly average. Finally, to appropriately compare to the news' sentiment, we scaled both variables with a mean equal to zero and then smoothed them using a six-week rolling average. The piece of visualization is available in the `output` folder.

Empirical Analysis

We begin our analysis by looking at historical trends in U.S. elections. The reader can find U.S. election results for the Republican and Democratic candidates from 1976 to 2020 at the following link: https://nanojgarcia.shinyapps.io/US-Elections-by-State/. We can see how Democratic and Republican candidates alternate in power, showing a healthy use of democratic institutions. Moreover, we can observe that Trump became the first U.S. president since George H. W. Bush in 1992 to lose a bid for a second term. Lastly, Trump's election in 2016 was a ridge election, losing the popular vote, and it shows that the dispute was in a few selected states to win the electoral votes in those states in 2020.

Before starting our quantitative analysis, we performed a data exploration step. We worked on some graphics to understand the potential relationships between the electoral results and the impact of the COVID-19 pandemic. In this regard, the table below includes four maps visualizing (i) electoral results, (ii) change in unemployment, (iii) cumulative COVID-19 confirmed cases, and (iv) the difference in the mobility index[1]. Interestingly, we can notice a spatial correlation between the two first maps, meaning a higher increase in the unemployment rate could be associated with a small or even negative electoral margin to the Republican party.





Based on the previous results, we made a last visualization effort by evaluating the relationship between electoral outcomes and change in the unemployment rate. In the scatter plot graph (review Appendix 2), we observe a slightly negative correlation. However, this relationship is also explained in part by the size of each county's total votes. To un-confound these relationships, we will present econometric outputs.

We explored three different models: a full model with all variables, one that excludes mobility, and the "optimal" model using forward stepwise regression. Results are available in Appendix 3. Column 3 indicates the coefficients associated with the independent variables under the "optimal" regression model. Some coefficients are not surprising. For example, women, non-Hispanic blacks, and counties with higher shares of college-educated people were more likely to vote for the Democratic Party. One unexpected coefficient is the positive relationship between Republican margin and foreign vote share, indicating Trump's cultural win on foreign

[1] The last two maps are available in the Appendix 1

and migration policy. However, one strong predictor of Trump's loss is the change in the unemployment rate, where counties that suffered more unemployment in 2020 relative to 2019 voted against Trump. It is not entirely clear if voters penalize Trump for the lack of reaction under such a crisis or that voters act irrationally blaming Trump for the current pandemic. In this regard, the text analysis of Trump's tweets offers an interesting approach[2]. Comparing his sentiment with that from the 16 U.S. major newspapers as the baseline of the common sense, we found that there is broke relationship coinciding with the spread of the disease. Finally, and more interestingly, we find a positive correlation between COVID cases and mobility on Trump's vote margin. Our interpretation of this result is of reverse causality. Trump supporters were less concerned about the virus, they mobilized more and therefore were more likely to get the virus.

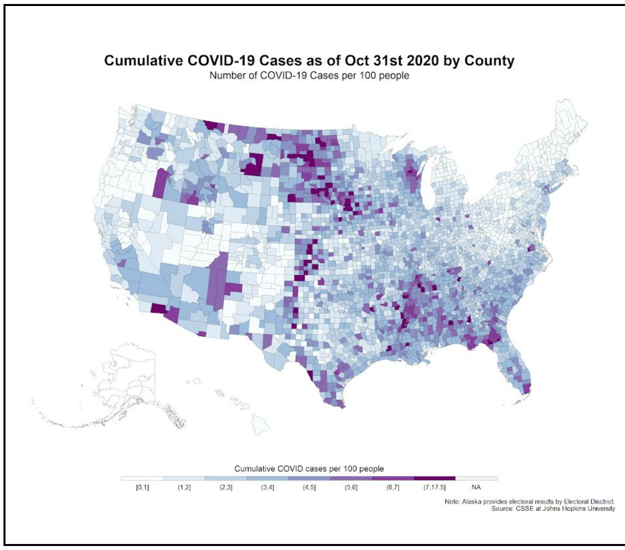
Conclusions

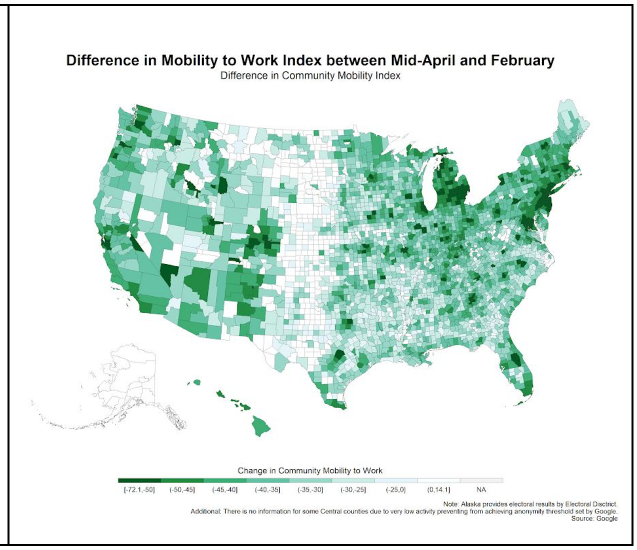
In this project, we explore the determinants of the 2020 election results in the United States. We use socio-demographic data and COVID-19 cases to explain the results in this past election, where President Joe Biden was elected president. Despite the fact that we are not able to determine causal mechanisms that explain Trump's loss, we were able to establish robust correlations between some socio-demographic indicators, such as the change in the unemployment rate, and the increase in COVID-19 cases. Further analysis needs to be done in order to better understand the mechanism underlying the 2020 election. One interesting question, that we are not able to answer here, is whether voters penalized former president Trump for the lack of response and leadership or if they behaved irrationally and blamed him for the current pandemic.

[2] Please, review Appendix 4

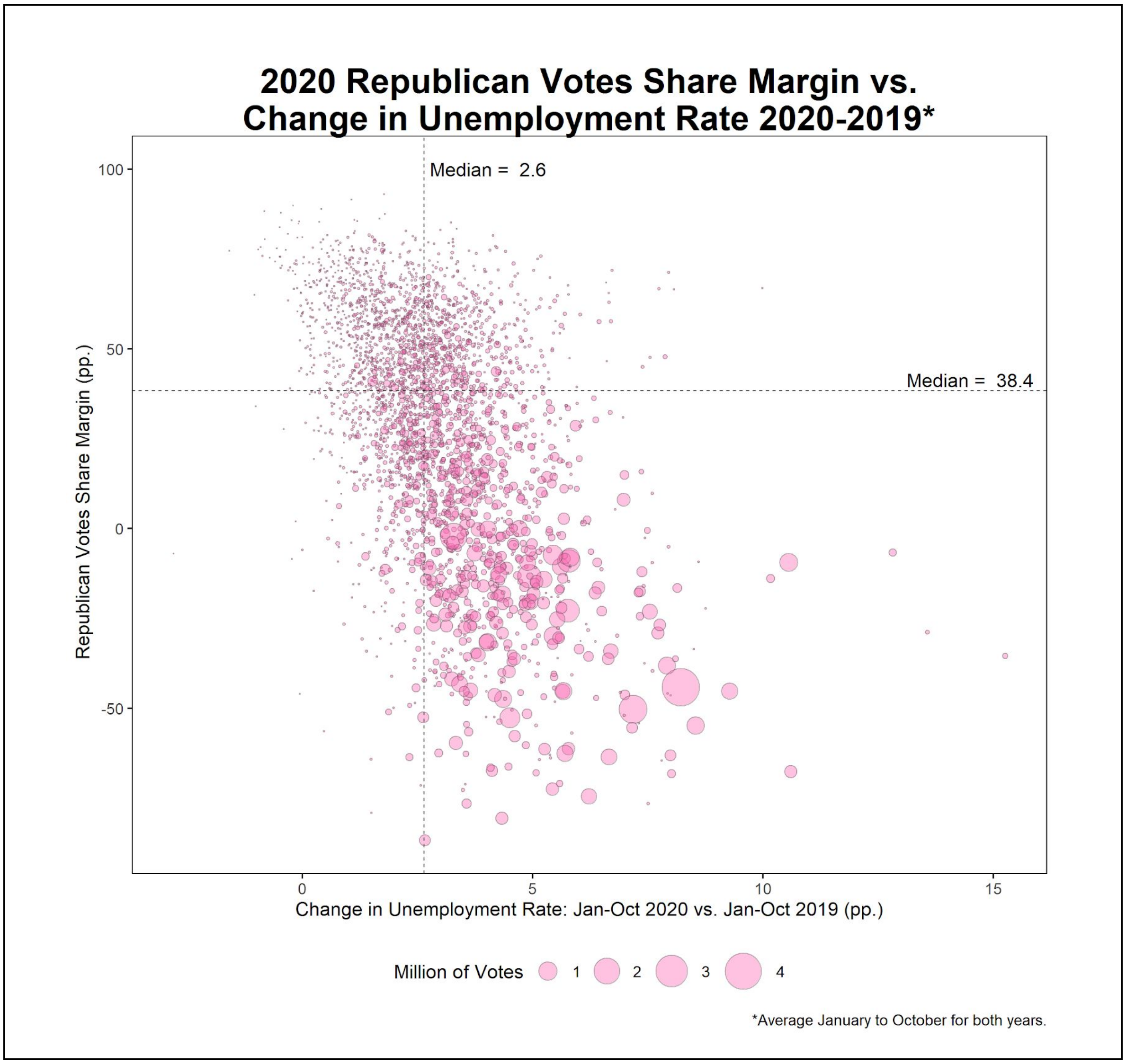
Appendix 1. COVID-10 cases and Difference in Mobility to Work

We found a potential positive spatial correlation between electoral results and COVID-19 cases. Nevertheless, using the green-colored map, we observe that among those counties in the central region with the highest disease levels per 100 people, most of them also do not show information about the change to mobility work because there is not enough anonymizing information. It could imply that those counties are very low populated, and the disease's impact could be different.





Appendix 2. Electoral results vs. Change in Unemployment Rate



Appendix 3. Linear regression results

Table 1: Linear Regresion results controllying by state

	2020 Republican Margin Vote (pp.)		
	Without Mobility	Full	Optimal
	(1)	(2)	(3)
Log(total votes)	−0.1841***	−0.1792***	−0.1788***
	(0.0279)	(0.0275)	(0.0273)
Log(total population)	0.1553***	0.1671***	0.1660***
	(0.0274)	(0.0271)	(0.0270)
Urban/Rural	0.0229**	0.0001	
	(0.0079)	(0.0082)	
Female share	−0.9796***	−0.9179***	−0.9104***
	(0.1408)	(0.1391)	(0.1388)
Age 18-24	−2.4799***	−2.3303***	−2.3554***
	(0.1110)	(0.1106)	(0.1099)
Age 25-44	−2.1465***	−2.1068***	−2.1148***
	(0.1726)	(0.1702)	(0.1702)
Age 45-64	−1.9921***	−1.8770***	−1.8968***
	(0.1543)	(0.1527)	(0.1524)
Age 65+	−1.2450***	−1.2844***	−1.3040***
	(0.1173)	(0.1159)	(0.1155)
Foreign born share	0.3685***	0.3096***	0.3020***
	(0.0657)	(0.0651)	(0.0650)
College degree share	−1.2070***	−1.1537***	−1.1948***
	(0.0536)	(0.0532)	(0.0503)
Non-Hispanic White	1.2353***	1.2248***	1.2260***
	(0.0294)	(0.0291)	(0.0289)
Non-Hispanic Black	−0.2812***	−0.3104***	−0.3108***
	(0.0337)	(0.0335)	(0.0332)
Log(household median income)	0.1110***	0.1291***	0.1506***
	(0.0204)	(0.0202)	(0.0182)
Gini index	−0.1754*	−0.1956*	
	(0.0803)	(0.0792)	
Δ unemployment rate	−2.2336***	−1.7883***	−1.8167***
	(0.2111)	(0.2155)	(0.2129)
COVID cases x 10k	0.0001***	0.0001***	0.0001***
	(0.00002)	(0.00002)	(0.00002)
COVID deaths x 100k	0.00001	0.00003	
	(0.00005)	(0.00005)	
Δ Mobility in deaths' peak		0.0027***	0.0026***
		(0.0004)	(0.0003)
Δ Mobility in cases' peak		−0.0003	
		(0.0005)	
Constant	0.9792***	0.6224*	0.3305
	(0.2912)	(0.2900)	(0.2641)
State fixed effects?	Yes	Yes	Yes
N	3,112	3,112	3,112
R^2	0.8731	0.8767	0.8764
Adjusted R^2	0.8703	0.8739	0.8738
Residual Std. Error	0.1156	0.1140	0.1140
F Statistic	317.3913***	318.1342***	337.6280***

Notes: ***Significant at the 0.1 percent level.
**Significant at the 1 percent level.
*Significant at the 5 percent level.

Appendix 4. Sentiment Analysis : Trump vs. Newspapers