Deciphering the Digital Pulse: Interpreting the Impact of Social Media Sentiment on 2020 U.S. Presidential Election Outcomes

Stroilă Alexandru-Mihail

January 9, 2024

Abstract

This research presents a nuanced exploration of the 2020 U.S. Presidential Election, aiming to interpret the influence of social media sentiment alongside traditional socio-demographic and economic factors on the state-wise election outcomes. Focusing on the major candidates, Donald Trump and Joe Biden, the study employs a spatial econometric model to understand how various factors correlate with their respective vote percentages across states.

A pivotal aspect of our analysis is the incorporation of state-specific Twitter sentiment scores for each candidate and a range of traditional electoral determinants, including gender distribution, education, unemployment, household income, poverty levels, and racial demographics. Additionally, the study incorporates the impact of COVID-19, acknowledging its unprecedented role in the 2020 election.

1 Introduction

The 2020 U.S. Presidential Election was a pivotal event in American political history, marked by unprecedented circumstances. Amidst the backdrop of a global pandemic, economic turbulence, and social unrest, the election not only reflected the nation's current state but also offered a lens through which the interplay of various factors influencing voter behavior could be examined. This study aims to dissect these influences, focusing particularly on the impact of social media sentiment and its interaction with traditional electoral determinants on the state-wise election outcomes.

The emergence of social media as a powerful tool in shaping public opinion has been increasingly recognized in political discourse. Platforms like Twitter have become arenas for political engagement, providing real-time reflections of public sentiment. In this context, analyzing the sentiment expressed towards the major candidates, Donald Trump and Joe Biden, offers a unique perspective on how digital discourse may correlate with electoral preferences. Studies have shown the influence of social media platforms, particularly Twitter, on election outcomes, indicating a potential shift in voter preferences due to the platform's content and user engagement [1].

However, understanding an election's dynamics necessitates more than examining digital sentiment. It requires a holistic approach that considers a range of socio-demographic and economic factors. This study therefore includes an analysis of traditional electoral determinants such as gender distribution, education levels, unemployment rates, household income, poverty levels, and racial demographics. These factors have been established in previous research as influential in shaping electoral outcomes. Moreover, the 2020 election was significantly influenced by the COVID-19 pandemic, a factor that cannot be overlooked. The pandemic's impact on voter behavior and preferences added a unique dimension to the election, intersecting with and possibly altering the influence of other variables.

Employing a spatial econometric model, this study aims to unravel the complex interdependencies and correlations between these diverse factors and their collective impact on the election results in each state. By doing so, it seeks to contribute to the broader discourse on electoral analysis, highlighting the evolving nature of political science in an age where both traditional factors and digital platforms play crucial roles. This assertion is supported by research indicating the growing impact of social media on political polarization and voter behavior, demonstrating significant shifts in public discourse and sentiment on platforms like Twitter during the election period [2] [3].

The following sections will detail the methodology employed, the data sources and variables used, the results of the analysis, and the interpretation of these results, culminating in a discussion that situates our findings within the wider context of electoral studies and political science.

2 Methodology & Data

2.1 Sentiment Score Analysis

The primary data for sentiment analysis in this study was extracted from a curated subset of the "US Election 2020 Tweets" dataset, available on Kaggle [4]. This dataset, encompassing a wide range of tweets related to the 2020 U.S. Presidential Election, provides a foundation for examining public sentiment towards the candidates across various states.

Our sentiment analysis was anchored by the "Twitter-roBERTa-base for Sentiment Analysis" model sourced from Hugging Face [5]. This model, a specialized version of the RoBERTa-base model, was trained on an extensive dataset comprising approximately 124 million tweets from a period extending from January 2018 to December 2021. It was further refined for sentiment analysis through the TweetEval benchmark. The choice of this model was driven by its robust training regimen, specifically tailored to decode the complexities of sentiment in Twitter's English-language content.

The derivation of sentiment scores was a critical component of our methodology, involving several steps to ensure the reflection of true public sentiment towards each candidate within each state, which can be summarized as follows:

- **Step 1.** Normalization of engagement metrics:
  We began by normalizing the likes and retweets for each tweet on a state-wise basis. This normalization aimed to balance the engagement metrics across states, ensuring that tweets from states with varying levels of Twitter activity were equitably represented. In this process, we calculated the total likes and retweets for all tweets in each state. Each tweet's likes and retweets were then represented as a percentage of these totals, thereby achieving a normalized engagement measure within the 0-100 range.

- **Step 2.** Weighted Sentiment Score calculation:
  The sentiment of each tweet, categorized as positive, negative, or neutral by the Twitter-roBERTa-base model, was assigned a corresponding numerical value (+1 for positive, -1 for negative, and 0 for neutral). These values were then weighted by the normalized likes and retweets, with likes receiving a 60% weight and retweets 40%, reflecting their respective importance in gauging public opinion. The final step involved averaging these weighted scores for each candidate across all tweets within a state, culminating in a composite sentiment score for each candidate in each state.

2.2 Socio-Demographic Data Collection

In conjunction with the sentiment analysis derived from Twitter data, our study incorporated a comprehensive array of socio-demographic variables, pivotal in painting a more holistic picture of the electoral landscape during the 2020 U.S. Presidential Election. This socio-demographic data was meticulously gathered from authoritative sources, namely the U.S. Census Bureau and the U.S. Department of Labor. These institutions are renowned for their comprehensive and reliable datasets, which offer detailed insights into the socio-economic fabric of the United States.

The variables sourced from these repositories included key demographic indicators such as gender distribution, educational attainment levels, and racial composition. Economic variables were also considered, including metrics like unemployment rates, real median household income, and poverty levels. Furthermore, due to the unique context of the 2020 election, we also integrated data on the COVID-19 pandemic's impact (gathered from the Centers for Disease Control and Prevention), specifically the death rates, to understand its potential influence on voting behavior.

	VOTES_B	VOTES_T	SENTIMENT_B	SENTIMENT_T	COVID_19	M_HOUSEHOLD_IN	POVER	CH_UNEMP	EDUCATION	BL_OR_AF_AMER	HISP_OR_LAT	ASI
Mean	49%	49%	-0.01	-0.08	98.21	77396.86	12%	111%	11%	12%	12%	5%
Median	49%	49%	0.00	-0.03	103.10	75980.00	11%	100%	11%	8%	10%	3%
Maximum	92%	70%	0.43	0.09	158.80	106900.00	19%	368%	25%	46%	49%	38%
Minimum	27%	5%	-0.31	-0.59	21.80	50880.00	7%	39%	6%	1%	2%	1%
Std. Dev.	12%	12%	0.11	0.11	35.66	13330.10	3%	56%	4%	11%	10%	6%
Skewness	79%	-87%	0.35	-2.36	-0.18	0.15	57%	211%	123%	126%	179%	461%
Kurtosis	466%	485%	8.79	9.86	2.07	2.24	299%	1011%	540%	405%	598%	2772%
Jarque-Bera	11.13	13.69	72.24	147.25	2.11	1.43	2.77	145.32	25.09	15.77	46.18	1478.79
Probability	0%	0%	0.00	0.00	0.35	0.49	25%	0%	0%	0%	0%	0%

Figure 1: Descriptive statistics

3 Spatial Statistics

In this section, we will use spatial statistics to enable us to identify patterns such as clustering, dispersion, and regional trends that traditional statistical methods might overlook. These patterns provide insights into the spatial dynamics of electoral preferences, revealing how factors like socio-economic conditions, demographic composition, and even sentiment scores, as captured through social media, vary across space and potentially influence voting behavior.

The following subsections will present and interpret a series of maps generated using GeoDa. These maps will include spatial distributions of key variables, such as vote shares and sentiment scores, and spatial autocorrelation analyses, like Moran's I, which indicate the degree of clustering in the data. By examining these spatial representations, we can gain a deeper understanding of the complex interplay between geography and electoral outcomes in the 2020 U.S. Presidential Election.

3.1 Choice of weight matrix

In our spatial econometric analysis, particularly given the geographical diversity of the United States, the selection of a weight matrix is a critical step. Our study, which includes states with unique geographical challenges such as Hawaii and Alaska, required a thoughtful approach to defining spatial relationships. After evaluating various options, we adopted a two-fold strategy: employing a **K-Nearest Neighbors (KNN) approach with 8 neighbors** and complementing it with an **inverse distance weighting, raised to the power of 1**.

The KNN approach, with 8 neighbors, was chosen to ensure that every state, including geographically isolated ones like Hawaii and Alaska, is adequately represented in the spatial framework. This method guarantees a fixed number of neighbors for each state, fostering a consistent spatial interaction structure.

To further refine this spatial relationship, we incorporated an inverse distance weighting. This method adjusts the influence of each neighbor based on the distance: closer neighbors have more influence, and those farther away have less. The use of an inverse distance weighting with a power of 1 effectively scales the impact of spatial relationships. It ensures that states which are relatively far apart can still exert an influence on each other, albeit reduced in proportion to their distance. This aspect is particularly important for integrating distant states like Hawaii and Alaska into the overall spatial analysis in a meaningful way.

The inverse distance weighting complements the KNN approach by adding a nuanced layer of spatial interaction. It acknowledges that spatial influence does not abruptly end at the closest neighbors but rather diminishes gradually with distance. This combination of KNN and inverse distance weighting captures both the immediate spatial interactions and the more subtle, distance-based influences, providing a comprehensive and balanced framework for analyzing the spatial dynamics of the 2020 U.S. Presidential Election data.

3.2 Spatial clustering for the sentiment analysis and COVID-19 death rate

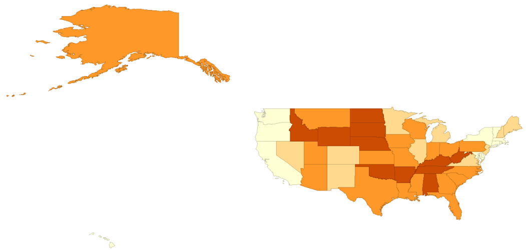

Figure 2: Distribution of Trump voters

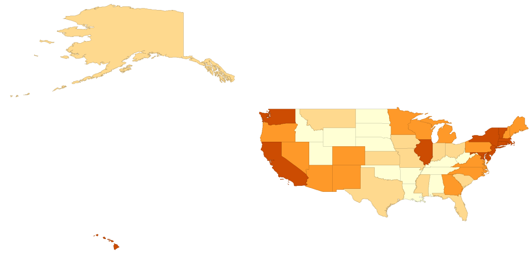

Figure 3: Distribution of Biden voters

The side-by-side comparison of the quantile maps for Trump and Biden voters in the 2020 U.S. Presidential Election reveals distinct geographical patterns and spatial clusters that speak to the political landscape of the United States.

The map for Trump voters exhibits strong support in the central regions of the country, with the highest quantiles predominantly situated in the Midwest and South. This pattern aligns with historical voting trends, where these areas have often leaned Republican. The spatial clustering in these regions suggests a strong regional identity that correlates with conservative political preferences.

In contrast, the Biden voter map shows higher percentages of votes in the coastal regions, particularly along the West Coast and the Northeastern states. These areas, traditionally Democratic strongholds, are characterized by higher urbanization levels, greater racial diversity, and different socio-economic dynamics compared to the central U.S. The spatial clustering here indicates that these factors may significantly influence the voting patterns in these regions.

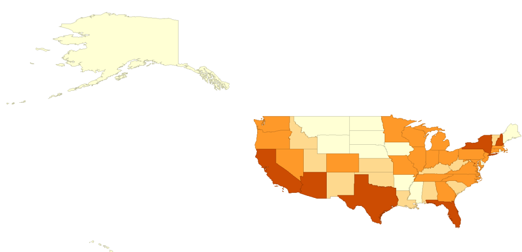

Figure 4: Sentiment analysis for Trump

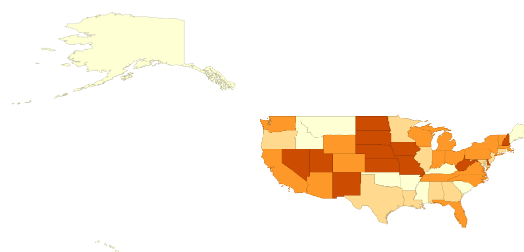

Figure 5: Sentiment analysis for Biden

Analyzing the two Twitter sentiment score maps for Biden and Trump in relation to the distribution of voters shows us that the Twitter sentiment score map for Biden appears to show that higher sentiment scores **are not confined to the coasts, where we saw higher voter percentages for him**. Instead, the sentiment seems to be more evenly distributed across the country. This could suggest that while positive sentiment for Biden was widespread, it **did not always translate into a majority of votes in many states**, possibly due to the influence of other overriding local factors or the distribution of the electorate. **The same can be observed in the case of Trump**.

In order to reinforce this, we have also plotted the Bivariate Moran's I between each candidate vote percentage and his sentiment score, thus, when analyzing the distribution of voters from the earlier maps, it is **clear that while there is a geographical pattern to the voting, the correlation to Twitter sentiment is not straightforward**. The sentiment maps and the Moran's scatterplots together suggest that while social media sentiment provides an interesting dimension of public opinion, it does not directly translate into voting patterns. This could be due to the complex nature of sentiment expression online, the presence of vocal minorities, or the influence of various other factors on voter decision-making that sentiment analysis alone cannot capture.

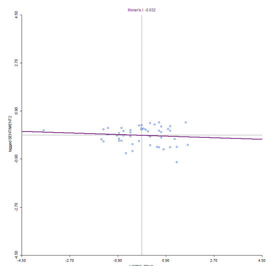

Figure 6: Bivariate Moran's I for Trump

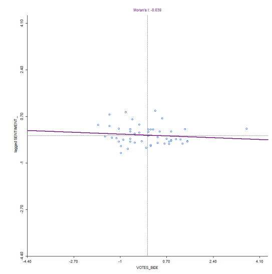

Figure 7: Bivariate Moran's I for Biden

Next, we will analyze the Moran's I bivariate scatterplots for COVID-19 death rates and the percentage of votes for Trump and Biden, and as we can see below, for Trump we have a Moran's I value of 0.200 suggesting a moderate positive spatial autocorrelation. This indicates that **states with higher COVID-19 death rates may have had a tendency to have a higher percentage of votes for Trump**. This relationship suggests that factors such as political loyalty, the **perceived effectiveness of state versus federal pandemic responses**, and differing public health policies may have influenced voting behavior more significantly than the death rates alone.

While in the case of Biden, we have a **negative correlation** as regions with higher COVID-19 death rates seeking change, **potentially benefiting the challenger**.

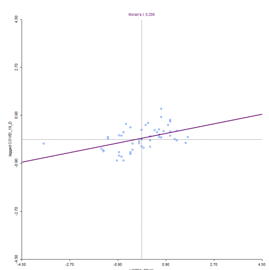

Figure 8: Bivariate Moran's I for Trump

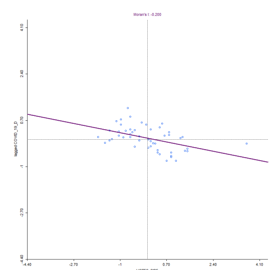

Figure 9: Bivariate Moran's I for Biden

3.3 Spatial clustering for socio-demographic variables

Having reviewed the maps reflecting socio-demographic variables—education levels, changes in unemployment, and real median household income—in relation to the voting percentages for Trump and Biden, we can draw several conclusions.

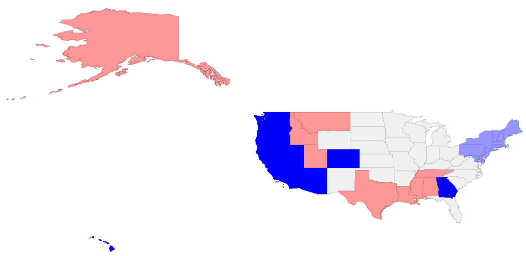

Figure 10: Bivariate Moran's I for Trump - Education

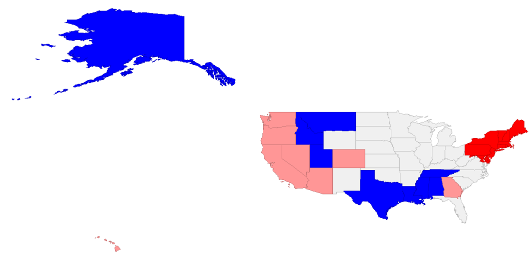

Figure 11: Bivariate Moran's I for Biden - Education

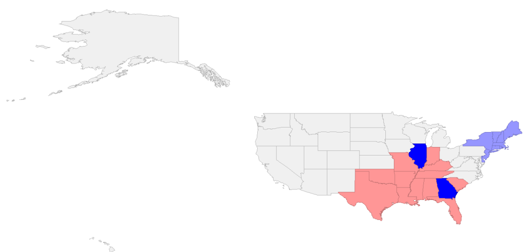

Figure 12: Bivariate Moran's I for Trump - Real Median Income Household

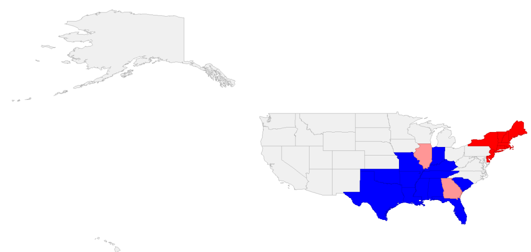

Figure 13: Bivariate Moran's I for Biden - Real Median Income Household

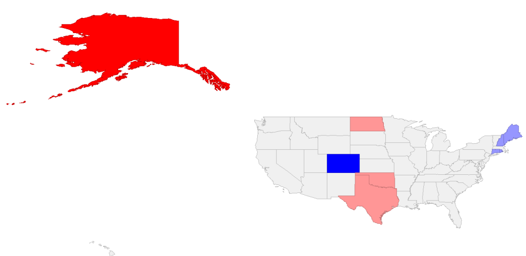

Figure 14: Bivariate Moran's I for Trump - Change in unemployment

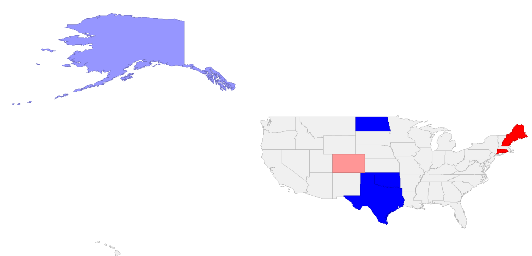

Figure 15: Bivariate Moran's I for Biden - Change in unemployment

Education Levels: The maps suggest a correlation between higher education levels and voting patterns. Regions with higher levels of education have a **positive influence over Biden's vote percentage**, while the opposite effect is present in Trump's case (i.e., **the higher the education level, the lower the percentage of votes obtained**).

Change in Unemployment: Changes in unemployment rates from 2019 to 2020 appear to have had an impact on voting behavior. The maps may reflect that areas with **increased unemployment were more likely to seek change**, perhaps swaying towards Biden, while **regions with stable or declining unemployment** may have been more inclined to vote for Trump, potentially viewing the incumbent's economic policies favorably.

Real Median Household Income: The relationship between median household income and voting preferences suggests a pattern where **wealthier states are more inclined to vote for Biden**, while **regions experiencing economic challenges show a tendency to support Trump**. This could reflect differing perceptions of each candidate's economic policies, with affluent areas possibly favoring Biden's stance on issues like taxation and social programs, and economically strained areas supporting Trump's economic approach and promises of job creation and industry support.

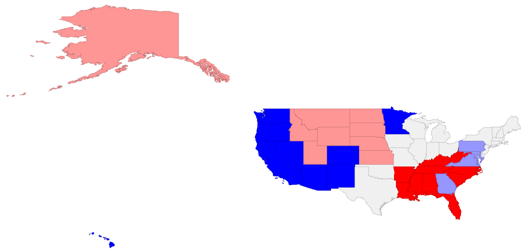

Figure 16: Bivariate Moran's I for Trump - Black or Afro American

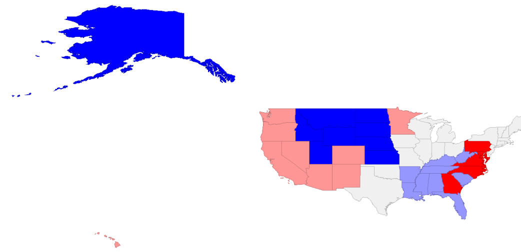

Figure 17: Bivariate Moran's I for Biden - Black or Afro American

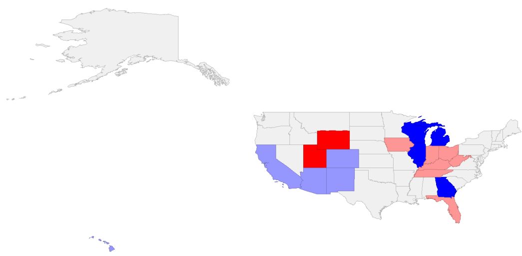

Figure 18: Bivariate Moran's I for Trump - Hispanic or Latino

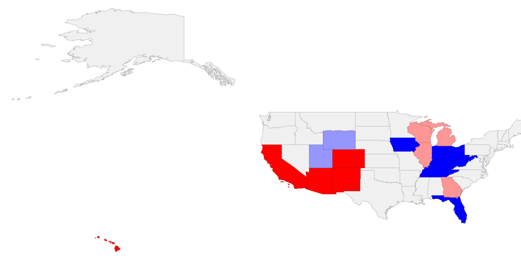

Figure 19: Bivariate Moran's I for Biden - Hispanic or Latino

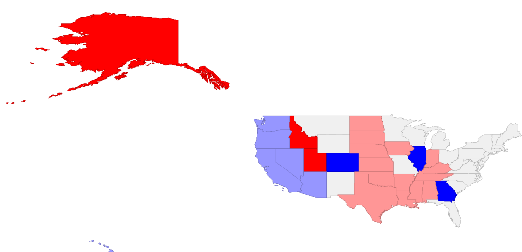

Figure 20: Bivariate Moran's I for Trump - Asian alone

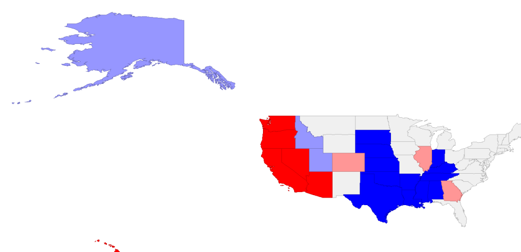

Figure 21: Bivariate Moran's I for Biden - Asian alone

The comparative analysis of maps reflecting voting preferences and racial demographics for both Trump and Biden in the 2020 U.S. Presidential Election reveals distinct patterns that align with national trends and the complex nature of America's diverse electorate.

Black or African American population: For Trump, areas with higher percentages of Black or African American residents tend to show lower support, which is consistent with the strong historical alignment of this demographic with the Democratic Party. Biden's support, indicated by blue clusters in these regions, confirms this trend, likely due to his platform's focus on social justice and equity, which resonated with many Black or African American voters.

Hispanic or Latino population: The relationship between the Hispanic or Latino population and voting preferences for Trump is mixed, suggesting a diversity of political opinion within this demographic. For Biden, blue clusters in areas with significant Hispanic or Latino populations suggest that his policies on immigration, healthcare, and economic opportunities may have appealed to these voters, although the pattern is not uniform across all states, reflecting the diverse concerns and priorities within this group.

Asian Alone population: Trump's maps show a non-uniform pattern of support among the Asian alone population, with some areas indicating preference for him and others leaning away. Biden's maps likely show blue clusters in areas with higher percentages of the Asian alone population, aligning with the recent trend of increasing Democratic support among Asian American voters, driven by issues such as immigration policy, healthcare access, and civil rights.

4 Spatial Regressions

4.1 Ordinary Least Squares

The Ordinary Least Squares (OLS) analysis for Biden's vote share within this research presents intriguing parallels and contrasts when considered alongside the model for Trump's votes. The model's strength is evident, with an R-squared of 0.8134, suggesting a significant proportion of the variation in Biden's vote share is illuminated by the selected predictors. The Adjusted R-squared of 0.7779 indicates a robust fit, adjusting for the number of variables.

```
REGRESSION
----------
SUMMARY OF OUTPUT: ORDINARY LEAST SQUARES
-----------------------------------------
Data set            :us2020election.dbf
Weights matrix      :        None
Dependent Variable  :  VOTES_TRUM                Number of Observations:          51
Mean dependent var  :      0.4922                Number of Variables   :           9
S.D. dependent var  :      0.1203                Degrees of Freedom    :          42
R-squared           :      0.8078
Adjusted R-squared  :      0.7712
Sum squared residual:       0.139                F-statistic           :     22.0631
Sigma-square        :       0.003                Prob(F-statistic)     :   1.004e-12
S.E. of regression  :       0.058                Log likelihood        :      78.203
Sigma-square ML     :       0.003                Akaike info criterion :    -138.406
S.E of regression ML:      0.0522                Schwarz criterion     :    -121.020

White Standard Errors
------------------------------------------------------------------------------------
            Variable     Coefficient       Std.Error     t-Statistic     Probability
------------------------------------------------------------------------------------
            CONSTANT       0.7214715       0.1023353       7.0500765       0.0000000
          ASIAN_ALON      -0.2234271       0.2648073      -0.8437349       0.4035976
          BLACK_OR_A      -0.3682800       0.0871716      -4.2247703       0.0001258
            CH_UNEMP      -0.0251706       0.0302436      -0.8322640       0.4099674
          COVID_19_D       0.0014269       0.0003961       3.6023209       0.0008280
           EDUCATION      -1.1092726       0.3386249      -3.2758155       0.0021167
          HISPANIC_O      -0.4285020       0.0751067      -5.7052405       0.0000011
          REAL_MEDIA      -0.0000013       0.0000010      -1.2863849       0.2053548
          SENTIMENT2       0.0794587       0.0716946       1.1082938       0.2740414
------------------------------------------------------------------------------------

REGRESSION DIAGNOSTICS
MULTICOLLINEARITY CONDITION NUMBER           34.337

TEST ON NORMALITY OF ERRORS
TEST                             DF        VALUE           PROB
Jarque-Bera                       2           0.653          0.7216

DIAGNOSTICS FOR HETEROSKEDASTICITY
RANDOM COEFFICIENTS
TEST                             DF        VALUE           PROB
Breusch-Pagan test                8          14.361          0.0728
Koenker-Bassett test              8          12.772          0.1199

DIAGNOSTICS FOR SPATIAL DEPENDENCE
TEST                           MI/DF       VALUE           PROB
Lagrange Multiplier (lag)         1           4.651          0.0310
Robust LM (lag)                   1           3.968          0.0464
Lagrange Multiplier (error)       1           0.715          0.3979
Robust LM (error)                 1           0.031          0.8600
Lagrange Multiplier (SARMA)       2           4.682          0.0962

================================ END OF REPORT =====================================
```

Figure 22: Ordinary least squares for Trump

```
REGRESSION
----------
SUMMARY OF OUTPUT: ORDINARY LEAST SQUARES
-----------------------------------------
Data set            :us2020election.dbf
Weights matrix      :        None
Dependent Variable  :  VOTES_BIDE                Number of Observations:          51
Mean dependent var  :      0.4863                Number of Variables   :           9
S.D. dependent var  :      0.1198                Degrees of Freedom    :          42
R-squared           :      0.8134
Adjusted R-squared  :      0.7779
Sum squared residual:       0.134                F-statistic           :     22.8914
Sigma-square        :       0.003                Prob(F-statistic)     :   5.469e-13
S.E. of regression  :       0.056                Log likelihood        :      79.187
Sigma-square ML     :       0.003                Akaike info criterion :    -140.374
S.E of regression ML:      0.0512                Schwarz criterion     :    -122.988

White Standard Errors
------------------------------------------------------------------------------------
            Variable     Coefficient       Std.Error     t-Statistic     Probability
------------------------------------------------------------------------------------
            CONSTANT       0.2838992       0.0985527       2.8806854       0.0062234
          ASIAN_ALON       0.2602745       0.2434576       1.0690752       0.2911418
          BLACK_OR_A       0.3542140       0.0942110       3.7597963       0.0005196
            CH_UNEMP       0.0225951       0.0278476       0.8113818       0.4217210
          COVID_19_D      -0.0013828       0.0003767      -3.6705537       0.0006773
           EDUCATION       1.2076948       0.3552543       3.3995225       0.0014902
          HISPANIC_O       0.4100297       0.0719468       5.6990651       0.0000011
          REAL_MEDIA       0.0000009       0.0000010       0.8901748       0.3784429
          SENTIMENT_      -0.0520425       0.0708054      -0.7350075       0.4664194
------------------------------------------------------------------------------------

REGRESSION DIAGNOSTICS
MULTICOLLINEARITY CONDITION NUMBER           33.503

TEST ON NORMALITY OF ERRORS
TEST                             DF        VALUE           PROB
Jarque-Bera                       2           0.364          0.8335

DIAGNOSTICS FOR HETEROSKEDASTICITY
RANDOM COEFFICIENTS
TEST                             DF        VALUE           PROB
Breusch-Pagan test                8          13.708          0.0897
Koenker-Bassett test              8          13.981          0.0823

DIAGNOSTICS FOR SPATIAL DEPENDENCE
TEST                           MI/DF       VALUE           PROB
Lagrange Multiplier (lag)         1           5.102          0.0239
Robust LM (lag)                   1           4.924          0.0265
Lagrange Multiplier (error)       1           0.394          0.5303
Robust LM (error)                 1           0.216          0.6418
Lagrange Multiplier (SARMA)       2           5.318          0.0700

================================ END OF REPORT =====================================
```

Figure 23: Ordinary least squares for Biden

In the individual variable analysis, divergences emerge between the candidates' models, particularly in the associations with racial demographics and COVID-19 impacts. For Biden, positive relationships with Black or African American (BLACK_OR_A) and Hispanic or Latino (HISPANIC_O) populations contrast Trump's model, reflecting Biden's stronger support within these groups. Conversely, the negative coefficient for COVID-19 death rates (COVID_19_D) indicates a decrease in Biden's vote share with higher pandemic fatalities, a reversal of the positive association found in Trump's results.

Education (EDUCATION) stands out as a significant predictor for Biden, with a strong positive correlation suggesting that higher education levels were tied to increased support for him. This is in stark contrast to the negative coefficient observed in the counterpart model.

Variables such as ASIAN_ALON, CH_UNEMP, and REAL_MEDIA do not reach statistical significance for Biden, mirroring their non-significant impact in the model for Trump. The sentiment analysis variable (SENTIMENT_), while negative for Biden, does not display statistical significance, echoing the complex and non-linear relationship between social media sentiment and electoral outcomes observed in Trump's model.

The spatial diagnostic tests indicate spatial lag dependence, signaling a similarity between the models in terms of spatial dynamics. The significant Lagrange Multiplier (lag) and Robust LM (lag) tests for Biden suggest that like Trump's vote share, Biden's is also subject to spatial influence, further advocating for the incorporation of a Spatial Lag Model to aptly capture the dependencies in the data.

The OLS models thus provide a comprehensive view of the socio-demographic underpinnings of the electoral preferences for the two major candidates. With the negative association of COVID-19 death rates and positive relationships with racial demographics and education levels for Biden, in contrast to Trump, the models reveal the distinctive contours of each candidate's electoral support. The noted spatial dependencies underscore the need for spatial econometric models to fully grasp the geographic patterns of voter support.

4.2 Spatial Two Stage Least Squares

The Spatial Two Stage Least Squares (STSLS) regression models for both Biden and Trump provide a more intricate depiction of the voting patterns when incorporating spatial interactions between observations. Notably, the inclusion of a spatially lagged dependent variable (W_VOTES_BIDE for Biden and W_VOTES_TRUM for Trump) captures the influence of neighboring areas' voting behaviors, enhancing the models' explanatory power.

```
REGRESSION
----------
SUMMARY OF OUTPUT: SPATIAL TWO STAGE LEAST SQUARES
--------------------------------------------------
Data set            :us2020election.dbf
Weights matrix      :File: us2020electionsymetric.gal
Dependent Variable  :  VOTES_TRUM                Number of Observations:          51
Mean dependent var  :      0.4922                Number of Variables   :          10
S.D. dependent var  :      0.1203                Degrees of Freedom    :          41
Pseudo R-squared    :      0.8219
Spatial Pseudo R-squared:  0.8225

White Standard Errors
------------------------------------------------------------------------------------
            Variable     Coefficient       Std.Error     z-Statistic     Probability
------------------------------------------------------------------------------------
            CONSTANT       0.5048020       0.1650728       3.0580560       0.0022278
          ASIAN_ALON      -0.2911108       0.2073064      -1.4042537       0.1602433
          BLACK_OR_A      -0.3706198       0.0741185      -5.0003705       0.0000006
            CH_UNEMP      -0.0147356       0.0240211      -0.6134416       0.5395844
          COVID_19_D       0.0013418       0.0003448       3.8916593       0.0000996
           EDUCATION      -0.9788598       0.2914016      -3.3591437       0.0007818
          HISPANIC_O      -0.4164198       0.0628976      -6.6205977       0.0000000
          REAL_MEDIA      -0.0000008       0.0000011      -0.7870812       0.4312343
          SENTIMENT2       0.0621340       0.0535511       1.1602746       0.2459370
        W_VOTES_TRUM       0.3224308       0.1333022       2.4187962       0.0155720
------------------------------------------------------------------------------------
Instrumented: W_VOTES_TRUM
Instruments: W_ASIAN_ALON, W_BLACK_OR_A, W_CH_UNEMP, W_COVID_19_D,
             W_EDUCATION, W_HISPANIC_O, W_REAL_MEDIA, W_SENTIMENT2

DIAGNOSTICS FOR SPATIAL DEPENDENCE
TEST                           MI/DF       VALUE           PROB
Anselin-Kelejian Test             1           0.433          0.5108
================================ END OF REPORT =====================================
```

Figure 24: Spatial Two Stage Least Squares - Trump

```
REGRESSION
----------
SUMMARY OF OUTPUT: SPATIAL TWO STAGE LEAST SQUARES
--------------------------------------------------
Data set            :us2020election.dbf
Weights matrix      :File: us2020electionsymetric.gal
Dependent Variable  :  VOTES_BIDE                Number of Observations:          51
Mean dependent var  :      0.4863                Number of Variables   :          10
S.D. dependent var  :      0.1198                Degrees of Freedom    :          41
Pseudo R-squared    :      0.8288
Spatial Pseudo R-squared:  0.8307

White Standard Errors
------------------------------------------------------------------------------------
            Variable     Coefficient       Std.Error     z-Statistic     Probability
------------------------------------------------------------------------------------
            CONSTANT       0.1786013       0.0861340       2.0735295       0.0381230
          ASIAN_ALON       0.3366121       0.1903334       1.7685393       0.0769708
          BLACK_OR_A       0.3639643       0.0743947       4.8923416       0.0000010
            CH_UNEMP       0.0132089       0.0219001       0.6031444       0.5464127
          COVID_19_D      -0.0012973       0.0003307      -3.9232881       0.0000873
           EDUCATION       1.0362988       0.2998024       3.4566059       0.0005470
          HISPANIC_O       0.4026235       0.0567949       7.0890804       0.0000000
          REAL_MEDIA       0.0000004       0.0000010       0.4307991       0.6666144
          SENTIMENT_      -0.0110856       0.0547427      -0.2025040       0.8395228
        W_VOTES_BIDE       0.3319393       0.1356492       2.4470413       0.0144034
------------------------------------------------------------------------------------
Instrumented: W_VOTES_BIDE
Instruments: W_ASIAN_ALON, W_BLACK_OR_A, W_CH_UNEMP, W_COVID_19_D,
             W_EDUCATION, W_HISPANIC_O, W_REAL_MEDIA, W_SENTIMENT_

DIAGNOSTICS FOR SPATIAL DEPENDENCE
TEST                           MI/DF       VALUE           PROB
Anselin-Kelejian Test             1           0.533          0.4654
================================ END OF REPORT =====================================
```

Figure 25: Spatial Two Stage Least Squares - Biden

For Biden, the STSLS model indicates a Pseudo R-squared of 0.8288, suggesting that the model accounts for a significant portion of the variability in his vote share. The positive coefficient of the spatially lagged variable (W_VOTES_BIDE) confirms the presence of spatial autocorrelation—areas surrounded by regions with high support for Biden also tended to show high support for him.

Similarly, Trump's STSLS model yields a Pseudo R-squared of 0.8219, also indicating substantial explanatory power. The positive coefficient for W_VOTES_TRUM reflects a similar spatial dynamic—Trump's support in a given area is positively correlated with support in neighboring regions.

Examining the coefficients of the socio-demographic variables in Biden's model, we observe significant positive relationships with the Black or African American (BLACK_OR_A) and Hispanic or Latino (HISPANIC_O) populations, and a negative association with COVID-19 death rates (COVID_19_D), aligning with the results from the OLS model. Education (EDUCATION) remains a positive and significant predictor, affirming its role in Biden's electoral support.

In contrast, Trump's STSLS model reveals negative associations with BLACK_OR_A and HISPANIC_O, and a positive correlation with COVID_19_D. Education (EDUCATION) is also significant but exhibits a negative relationship with Trump's vote share, consistent with the OLS findings.

Neither model shows significant coefficients for ASIAN_ALON, CH_UNEMP, REAL_MEDIA, nor for the sentiment analysis variables (SENTIMENT_ for Biden and SENTIMENT2 for Trump), indicating that these factors do not have a statistically significant direct impact on the vote shares within the spatial framework.

The Anselin-Kelejian test results for both models do not indicate significant spatial error autocorrelation, validating the use of spatial lag models over spatial error models.

In synthesis, the spatial regression analyses deepen our understanding of the geographical dimension of the electoral landscape. The models for Biden and Trump reflect the significance of racial demographics and COVID-19 impacts, diverging in their direction of association. Both models confirm the spatial dependence of voting patterns, underlining the influence of regional clusters in presidential support. The findings suggest that while socio-demographic variables play a definitive role in explaining vote shares, they do so within a complex spatial context where the influence of neighboring regions cannot be ignored.

5 Conclusions

In conclusion, this research has embarked on a multifaceted exploration of the 2020 U.S. Presidential Election, investigating the interplay between social media sentiment, socio-demographic variables, and electoral outcomes. Our analyses, spanning from Ordinary Least Squares (OLS) to Spatial Two Stage Least Squares (STSLS) regression models, have unraveled complex patterns and relationships that shape the political landscape of the United States.

One of the core objectives of this study was to understand the impact of Twitter sentiment on voting outcomes. While the sentiment analysis provided valuable insights into public opinion dynamics, its direct correlation with electoral preferences was less pronounced than initially anticipated. This outcome suggests that while social media sentiment is a significant aspect of the modern political discourse, its role as a predictor of voting behavior may require a more nuanced approach or a broader dataset for analysis. Future research might benefit from an expanded Twitter dataset or alternative methodologies in calculating sentiment scores, potentially incorporating sentiment intensity or contextual analysis to capture the multifaceted nature of social media discourse.

Interestingly, the socio-demographic and COVID-19 variables emerged as more definitive factors in electoral outcomes. Our findings underscored the importance of racial demographics, education levels, and the impact of the COVID-19 pandemic in shaping voter preferences. The STSLS models, in particular, highlighted the spatial dimensions of these relationships, revealing how the voting patterns in one region are influenced by those in neighboring regions, thus painting a picture of regional political interconnectedness.

For both candidates, Trump and Biden, distinct patterns emerged. Biden's support correlated positively with higher education levels and racial minority demographics, while Trump's vote share showed an inverse relationship with these variables. Furthermore, the COVID-19 death rates exhibited contrasting associations with each candidate's vote shares, reflecting perhaps the public's varied perception of the pandemic's handling and its broader socio-political implications.

This study's comprehensive approach, integrating traditional socio-demographic analysis with cutting-edge sentiment analysis and spatial econometrics, reveals the intricacies of electoral behavior in the modern era. It highlights that electoral outcomes are the product of a complex interplay of factors, where socio-economic conditions, public health crises, and digital discourse converge.

In essence, the 2020 U.S. Presidential Election serves as a case study in the evolving nature of electoral analyses, where traditional methods meet the emerging influences of social media and spatial dynamics. The insights gained herein not only contribute to a deeper understanding of this particular election but also lay the groundwork for future research in the ever-evolving landscape of political science and electoral studies.

References

[1] Thomas Fujiwara, Karsten Müller, and Carlo Schwarz. The effect of social media on elections: Evidence from the united states. *National Bureau of Economic Research*, Working Paper 28849, 2021.

[2] James Flamino et al. Political polarization of news media and influencers on twitter in the 2016 and 2020 us presidential elections. *Nature Human Behaviour*, Volume 7(904–916), 2023.

[3] Loris Belcastro et al. Analyzing voter behavior on social media during the 2020 us presidential election campaign. *Social Network Analysis and Mining*, 2022.

[4] Man Chun Hui. Us election 2020 tweets, 2020.

[5] Daniel Loureiro et al. Timelms: Diachronic language models from twitter. *arXiv*, arXiv:2202.03829, 2022.