Original Research Article

Socioeconomic determinants of conservative vote share in the 2016 U.S. presidential election

Xinyao Fu
School of Mathematics and Statistics, University of Glasgow, Glasgow, G128QQ, United Kingdom

Abstract: This study examines the socioeconomic determinants of Conservative voting outcomes in the 2016 U.S. presidential election using county-level data. Combining election returns with indicators of income, educational attainment, poverty, and unemployment, the analysis assesses both linear and nonlinear relationships between local socioeconomic conditions and Conservative vote share. A multi-model framework is employed, including linear and logistic regression, generalized additive models (GAM), and tree-based methods, to balance interpretability and predictive performance. The results show that median household income and educational attainment are the most influential predictors of Conservative support, while poverty and unemployment play more limited roles. Importantly, the effects of income and education are strongly nonlinear, with the largest marginal impacts concentrated among counties with lower socioeconomic levels and diminishing at higher levels. Model comparisons indicate that GAM offers the most informative and interpretable representation of these relationships, whereas random forest models achieve higher predictive accuracy. Overall, the findings suggest that regional political polarization in the United States is closely associated with structural socioeconomic inequality at the county level.

Keywords: county-level analysis; generalized additive models;income and education;model comparison; nonlinear effects;conservative vote share

1. Introduction

The 2016 U.S. presidential election was characterized by pronounced political polarization alongside persistent regional socioeconomic inequality [1]. While national election outcomes highlight deep partisan divisions, county-level returns reveal substantial geographic heterogeneity that is obscured at more aggregated scales [2]. Long-standing disparities in income, educational attainment, poverty, and unemployment vary sharply across U.S. counties and are widely regarded as important correlates of electoral behavior [3,4]. These structural differences provide a critical context for understanding variation in voting outcomes across space.

However, existing explanations of the 2016 election outcome often emphasize cultural, ideological, or identity-based factors, with comparatively less systematic assessment of how measurable local socioeconomic conditions shape county-level variation in Conservative support [3]. Moreover, analyses conducted at the state or national level may mask meaningful within-state heterogeneity arising from uneven economic development and differential access to local opportunities [5]. As a result, important spatial patterns linked to socioeconomic structure may remain underexplored.

Using county-level election returns merged with socioeconomic indicators, this study examines how median household income, educational attainment, poverty, and unemployment are associated with Conservative vote share and county-level Conservative victories in the 2016 election [6]. To assess whether these associations are strictly linear or instead exhibit threshold effects and diminishing marginal impacts, we compare baseline linear and logistic regression models with more flexible nonlinear approaches and tree-based benchmarks [7]. By integrating explanatory and predictive perspectives, this paper provides new empirical evidence on the extent to which structural socioeconomic inequality contributes to spatially patterned political polarization in the United States [8].

2. Data and variables

2.1. Data sources

County-level election results for the 2016 U.S. presidential election were obtained from the MIT Election

Data and Science Lab, including total votes and Conservative votes by county [9]. County socioeconomic indicators were drawn from publicly available U.S. Census-based county profile data, including measures of income, education, poverty, and unemployment [10]. All datasets were merged using standardized five-digit Federal Information Processing Standard (FIPS) county codes to ensure consistent geographic linkage [11].

2.2. Outcomes

The primary outcome is Conservative vote share, defined as the proportion of valid votes cast for the Conservative presidential candidate in each county (continuous outcome). For classification analysis, a binary indicator of Conservative victory is constructed, coded as 1 if the Conservative candidate received more than 50% of county votes and 0 otherwise.

2.3. Key predictors

Core explanatory variables capture major dimensions of county socioeconomic structure: median household income, educational attainment (share of adults aged 25+ with a bachelor's degree or higher), poverty rate, and unemployment rate [10]. These variables are widely used as indicators of local economic capacity, human capital, and material deprivation relevant to political behavior [3].

2.4. Preprocessing and transformations

To enhance comparability across counties and reduce the influence of extreme values, median household income is log-transformed prior to modeling [9]. Percentage variables (education, poverty, unemployment) are converted to proportions bounded between 0 and 1[10]. Categorical fields are recoded into numeric or binary formats to meet regression and classification model requirements [12]. These preprocessing steps improve numerical stability and help align key variables with common assumptions of parametric modeling.

2.5. Descriptive patterns

Exploratory checks indicate substantial heterogeneity in both voting outcomes and socioeconomic conditions across counties, consistent with strong regional stratification. Pairwise associations suggest that income and educational attainment tend to be negatively related to Conservative vote share, while poverty and unemployment show positive associations, and several relationships appear plausibly nonlinear [3,13]. These patterns motivate the use of both parametric models and flexible nonlinear methods in subsequent analysis.

3. Methodology

This study adopts a multi-model analytical framework to examine the association between county-level socioeconomic factors and Conservative voting outcomes in the 2016 U.S. presidential election. Given that socioeconomic influences on political behavior may operate in both linear and nonlinear ways, the analysis combines parametric regression models with flexible nonlinear and tree-based methods. This approach allows comparison between interpretability and predictive performance across modeling strategies.

3.1. Linear regression and logistic regression

As baseline parametric specifications, linear and logistic regression models are employed to estimate the associations between county-level socioeconomic conditions and Conservative voting outcomes.

For the continuous outcome, Conservative vote share, ordinary least squares (OLS) regression is used to model the average linear relationship between socioeconomic predictors and voting outcomes:

$$Y_i=\beta_0+\beta_1X_{1i}+\beta_2X_{2i}+...+\beta_kX_{ki}+\varepsilon_i$$

Where Y_i denotes the Conservative vote share in county i, X_{ki} represent socioeconomic predictors (income, educational attainment, poverty, and unemployment), and ε_i represents random error. This specification assumes constant marginal effects across counties and provides interpretable coefficient estimates that serve as a benchmark for more flexible models [13].

For binary outcomes, a generalized linear model (GLM) with a logit link function is applied to model the probability of a Conservative county-level victory:

$$\text{logit}(P(Y_i=1))=\beta_0+\sum_{k=1}^{K}\beta_kX_{ki}$$

Logistic regression estimates how socioeconomic factors are associated with the odds of Conservative electoral success and extends the linear framework to classification settings under standard assumptions of linearity in the log-odds space and conditional independence of predictors [14].

3.2. Generalized additive models

To relax linearity assumptions, generalized additive models (GAMs) are employed to capture potential nonlinear relationships between socioeconomic predictors and voting outcomes. GAMs replace linear terms with smooth functions, allowing marginal effects to vary across the range of each variable:

$$Y_i=\beta_0+s_1(X_{1i})+s_2(X_{2i})+s_3(X_{3i})+s_4(X_{4i})+\varepsilon_i$$

This formulation allows the model to identify curved or threshold effects that linear specifications cannot capture. GAMs are estimated for both continuous outcomes (Conservative vote share) and binary outcomes (Conservative county victories), maintaining interpretability while accommodating nonlinear socioeconomic gradients [15].

3.3. Tree-based models

For comparison with parametric approaches, two nonparametric tree-based methods are implemented: classification and regression trees (CART) and random forests. CART models recursively partition the predictor space using binary splits, yielding transparent decision rules that facilitate interpretation [16]. Random forests extend CART by aggregating predictions from multiple trees constructed on bootstrap samples and random subsets of predictors, improving predictive stability and reducing overfitting. Variable importance measures derived from the random forest are used to identify the most influential socioeconomic predictors [17].

3.4. Model evaluation

Model performance is assessed using both explanatory and predictive criteria. For continuous outcomes, model fit and parsimony are evaluated using the Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC), while predictive accuracy is measured by root mean squared error (RMSE) and mean absolute error (MAE). For binary outcomes, classification accuracy and the area under the receiver operating characteristic curve (AUC) are reported. Together, these metrics provide a unified basis for comparing linear, nonlinear, and machine-learning models in terms of interpretability and predictive performance.

4. Result analysis

This section presents the empirical results on the relationship between county-level socioeconomic conditions and Conservative voting outcomes in the 2016 U.S. presidential election. We first outline the broad spatial patterns, followed by the results from linear, nonlinear, and tree-based models.

At the descriptive level, we observe distinct regional differences. Counties in the South and Midwest tend to have lower median household incomes, higher poverty and unemployment rates, and significantly higher Conservative vote shares. In contrast, counties in the Northeast and on the West Coast generally have higher income and educational attainment, along with lower Conservative support. These spatial patterns align with the urban-rural and regional divides previously documented in U.S. electoral geography.

The linear regression model reveals that both median household income and educational attainment are significantly and negatively associated with Conservative vote share, while poverty and unemployment rates show positive correlations. These findings suggest that counties with lower income and education levels are more likely to support the Conservative candidate. While the linear model explains a substantial portion of the variation in vote share, it assumes constant marginal effects across the entire range of each variable, which may oversimplify the underlying dynamics.

The generalized additive model (GAM) uncovers significant nonlinearities in these relationships. For income and educational attainment, the strongest effects are concentrated in counties with lower socioeconomic status, with diminishing marginal impacts at higher levels. Similarly, while poverty and unemployment exhibit positive nonlinear effects, these impacts become less pronounced as their values increase. These results highlight that the influence of socioeconomic factors on voting behavior varies across counties, rather than remaining constant.

Model comparisons reveal a trade-off between interpretability and predictive accuracy. GAM strikes a

balance between flexibility and transparency, explaining 70-75% of the variation in Conservative vote share, with highly significant smooth terms for income and education. However, tree-based methods, particularly random forests, offer superior predictive accuracy, as evidenced by lower RMSE and higher AUC scores in classification tasks. Variable importance analysis from the random forest model consistently identifies educational attainment and income as the most significant predictors, followed by poverty, with unemployment playing a relatively minor role.

5. Discussion

The results provide strong evidence that socioeconomic inequality is closely linked to regional variation in Conservative voting outcomes, but in ways that are not fully captured by linear models. The nonlinear patterns identified for income and education suggest that socioeconomic disadvantage exerts its strongest political influence at lower levels of development. Once a county reaches higher income or education thresholds, additional gains appear to have diminishing effects on partisan voting behavior.

These findings are broadly consistent with existing research on U.S. electoral geography, particularly studies emphasizing the political divide between economically distressed, less-educated regions and more affluent, highly educated metropolitan areas. However, by modeling nonlinear effects explicitly, this study refines earlier conclusions by showing that the relationship is not simply monotonic. The concentration of marginal effects among low-income and low-education counties helps explain why political polarization is especially pronounced in economically marginalized regions, while variation among affluent counties is comparatively limited.

Methodologically, the comparison across models underscores the value of flexible statistical approaches in electoral analysis. While machine learning models such as random forests excel in prediction, their limited interpretability constrains substantive inference. GAM offers an effective compromise, allowing researchers to uncover complex, nonlinear relationships while retaining a clear link between socioeconomic theory and empirical estimation. These results suggest that relying exclusively on linear specifications may obscure important features of how structural inequality shapes political behavior.

6. Conclusion

This study examines the socioeconomic foundations of Conservative voting outcomes in the 2016 U.S. presidential election using county-level data and a multi-model statistical framework. The results show that income and educational attainment are the most influential determinants of Conservative vote share, with effects that are strongly nonlinear and concentrated among socioeconomically disadvantaged counties. By integrating interpretable nonlinear models with predictive methods, the analysis demonstrates that regional political polarization in the United States is closely tied to structural socioeconomic inequality. These findings highlight the importance of accounting for heterogeneous and nonlinear effects when analyzing electoral behavior at fine geographic scales.

References

[1] Pew Research Center (2018). The Partisan Divide on Political Values Grows Even Wider. Washington D.C.

[2] Rodden, J. (2019). Why Cities Lose: The Deep Roots of the Urban-Rural Political Divide. Basic Books.

[3] Gelman, A., Kenworthy, L., and Su, Y. (2010). Income Inequality and Partisan Voting in the United States. Social Science Quarterly, 91(5), 1203–1223.

[4] Lipset, S. M. (1959). Some Social Requisites of Democracy: Economic Development and Political Legitimacy. American Political Science Review, 53(1), 69–105.

[5] Autor, D. H., Dorn, D., Hanson, G. H., and Majlesi, K. (2020). Importing Political Polarization? The Electoral Consequences of Rising Trade Exposure. American Economic Review, 110(10), 3139–3183.

[6] Bishop, B. (2008). The Big Sort: Why the Clustering of Like-Minded America Is Tearing Us Apart. Mariner Books.

[7] Glaeser, E. L., and Ward, B. A. (2006). Myths and Realities of American Political Geography. Journal of Economic Perspectives, 20(2), 119–144.

[8] McCarty, N., Poole, K. T., and Rosenthal, H. (2016). Polarized America: The Dance of Ideology and Unequal

Riches. MIT Press.

[9] MIT Election Data and Science Lab (2020). County-Level Presidential Election Returns (2000–2020). Harvard Dataverse.

[10] U.S. Census Bureau (2016). County Facts Database. Washington D.C.

[11] Hastie, T., Tibshirani, R., and Friedman, J. (2009). The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer.

[12] Field, A., Miles, J., and Field, Z. (2012). Discovering Statistics Using R. SAGE Publications.

[13] Fox, J., and Weisberg, S. (2019). An R Companion to Applied Regression (3rd ed.). Sage Publications.

[14] O'Brien, R. M. (2007). A Caution Regarding Rules of Thumb for Variance Inflation Factors. Quality & Quantity, 41(5), 673–690.

[15] Kuhn, M., and Johnson, K. (2013). Applied Predictive Modeling. Springer.

[16] Breiman, L., Friedman, J. H., Olshen, R. A., and Stone, C. J. (1984). Classification and Regression Trees. Wadsworth International Group.

[17] Kuhn, M., and Silge, J. (2022). Tidy Modeling with R. O'Reilly Media.