Who Voted in 2016?
Using Fuzzy Forests to Understand Voter Turnout

Seo-young Silvia Kim*† R. Michael Alvarez† Christina M. Ramirez‡

†*California Institute of Technology*
‡*University of California, Los Angeles*

January 9, 2020

Abstract

Objective: What can machine learning tell us about who voted in 2016? There are numerous competing voter turnout theories, and a large number of covariates are required to assess which theory best explains turnout. This paper is a proof-of-concept that machine learning can help overcome this curse of dimensionality and reveal important insights in studies of political phenomena.

Methods: We use Fuzzy Forests, an extension of Random Forests, to screen variables for a parsimonious but accurate prediction. Fuzzy Forests achieve accurate variable importance measures in the face of high dimensional and highly correlated data. The data that we use is the 2016 Cooperative Congressional Election Study.

Results: Fuzzy Forests chose only a small number of covariates as major correlates of 2016 turnout and still boasted high predictive performance.

Conclusion: Our analysis provides three important conclusions about turnout in 2016: registration and voting procedures were important, political issues were important (especially Obamacare, climate change, and fiscal policy), but few demographic variables other than age were strongly associated with turnout. We conclude that Fuzzy Forests is an important methodology for studying over-determined questions in social sciences.

*Kim is a Ph.D. candidate in the Division of Humanities and Social Science, Caltech, and is the corresponding author (email: sskim@caltech.edu, phone: +1 (626) 554-8088, Twitter: @sysilviakim); Alvarez is Professor of Political Science, Division of Humanities and Social Sciences, Caltech (Twitter: @rmichaelalvarez); Ramirez is Professor in the Department of Biostatistics, UCLA School of Public Health. An earlier version of this paper was presented as a poster at the Polmeth 2018 conference; we thank conference participants for their comments and suggestions.

1 Introduction

Despite the surprising outcome of the 2016 presidential election, there has been little research on voter turnout in 2016. What factors are most closely associated with who voted in 2016? There is a sizable literature on voter participation in U.S. federal elections that presents many different theories of what determines turnout. The dataset that we use—the 2016 Cooperative Congressional Election Survey (CCES)—is rich in width (the number of variables) and length (the number of respondents). The richness of theory and data can present a problem for quantitative analysis: how do we parsimoniously determine which variables of the data are most predictive of *actual* turnout, while retaining flexibility and imposing the fewest restrictions?

As, *a priori*, we do not know which of these many variables are important for predicting turnout in the 2016 election, in this paper we examine the utility of machine learning for studying over-specified, over-determined social science questions like this. Specifically, we use the 2016 election as a proof-of-concept for a machine learning method that we argue is useful for this type of quantitative problem: "Fuzzy Forests" (Conn et al., 2018; Conn and Ramirez, 2016; Ramirez et al., 2019). This novel machine learning approach is an extension of Random Forests (Breiman, 2001). The Fuzzy Forests methodology has attributes that make it useful for studying political behavior—most importantly, it filters variables by their importance after it creates a weighted correlation network, and identifies "modules" or clusters of variables that are closely connected. These modules take into account the correlation among potential predictors. Hence it functions well in data (like survey data on voter participation and behavior) where the variables are highly inter-correlated, which can lead to biased variable importance measures of Random Forests (Strobl et al., 2007). Fuzzy Forests helps us determine which variables are most important in understanding turnout, and how they are correlated into clusters.

Using this methodology, we learn three important lessons about voter turnout in 2016. First, despite a long debate in the literature about whether issues "matter" for understanding why people participate in elections, we find that a number of political issues, in particular Obamacare, climate change, and fiscal policy, were associated with whether registered voters participated in the election. That said, while we find that political issues are important variables in their association with turnout, in our overall counterfactual analyses based on our model, we do not find that these political issues reshaped the 2016 electorate.

Second, and related to the lengthy literature on turnout, we also do not find that many demographic variables were closely associated with turnout in 2016. Finally, consistent with many past studies, we find that registration and voting administrative procedures remain important correlates of turnout in American national elections. Thus, machine learning methods like the Fuzzy Forests can shed new light on longstanding debates in the literature on political participation, and help scholars better understand the 2016 presidential election.

2 Machine Learning and Voter Turnout

There is an extensive literature that has offered many theories for why eligible voters participate in U.S. national elections—theories that have been tested with observational and experimental data (e.g., Leighley and Nagler (2013)). There are theories that assert that there are legal and institutional factors associated with voter turnout, that the socioeconomic and political context of a particular election can be an important component shaping the nature of the electorate, and that other factors like demographics, issues, uncertainty, and habitual behavior, all may account for why some eligible voters decide to participate in a particular election while others stay home.

With such an expansive literature, it is not easy to determine the important factors associated with 2016 voter turnout. With an array of theoretical models, some focusing on policy or institutional factors, others on attributes of the voters themselves, and still others on issues and the context of the election, there are quite literally many dozens of variables that can be drawn from a micro-level survey data, all of which are highly inter-correlated. And in the 2016 election, it can easily be argued that many of these potential predictors may have mattered, but it is unclear which mattered most, in which direction they operated, and which help tell the descriptive story of "what happened."

Furthermore, there could be lesser known variables that do not have a strong theory component, that are election-specific or contextual, that may be important to prediction. Perhaps, indeed, the dynamics of voter turnout in the 2016 election were different than in past presidential elections. Our theories tell us which variables we may wish to include in our models of voter turnout, but not how particular variables might matter more or less in a particular election, nor exactly how particular variables might benefit one candidate or campaign over another. Given the array of different theories of voter turnout, and the high dimensionality of the potential variable space, it is also quite difficult to present a parsimonious set of hypotheses for testing. Given these issues, we focus not on hypothesis testing, but rather on variable selection and developing a descriptive portrait of who voted in the 2016 presidential election.

As the specific correlates of voter participation in this election are not clear, and as we do not have strong theory about which factors might be more important in a presidential election such as 2016, we need a principled way of screening potential predictors and summarizing who voted in a succinct and parsimonious model, which is why we turn to machine learning and Fuzzy Forests.

3 Machine Learning and Fuzzy Forests

There has been considerable interest in the social sciences in the opportunities provided by new data and computational tools (Alvarez, 2016). A number of new studies in political science have

used one important type of machine learning methodology—regression trees (Green and Kern, 2012; Imai and Strauss, 2011; Kastellec, 2010; Montgomery and Olivella, 2018). This literature has noted that regression trees can be very useful methodological techniques for studying political phenomena. In large and complex datasets these techniques can effectively identify the important variables that best predict the outcome of interest (Montgomery and Olivella, 2018). Many have also noted that single trees can be unstable and the usage of ensembles of trees such as Random Forests (RF) usually produce a better predictive model. These ensembles of trees have been discussed in recent political science research, many citing their advantage of allowing the data to "speak for themselves" instead of exercising overly high discretion in areas without strong theory (Green and Kern, 2012; Hill and Jones, 2014; Kastellec, 2010; Levin et al., 2016; Montgomery et al., 2015; Muchlinski et al., 2016).

Fuzzy Forests are an extension of Random Forests, a popular machine learning ensemble algorithm. Fuzzy Forests were developed for situations where researchers seek to determine the most important variables from a large list of p potentially correlated parameters. That is, they wish to know which variables are implicated in driving the signal so that further mechanistic analysis can be performed. It is well-known that for CART (Classification and Regression Trees, the base learner in Random Forests), variable selection is biased when potential predictors are correlated, have differing numbers of categories, or are coded on different scales (Conn and Ramirez, 2016; Conn et al., 2018; Kim and Loh, 2001; Nicodemus and Malley, 1890; Strobl et al., 2007). Fuzzy Forests overcome many of these issues by constructing "modules" based on a weighted correlation network, where the potential predictors in each module are more correlated within the module than between the modules—that is, they cluster together.

The algorithm is implemented in three stages. The first stage utilizes unsupervised learning in the form of a weighted correlation network (Langfelder and Horvath, 2008). In brief, a similarity function is created using a similarity measure, such as Pearson correlations between each pair of potential predictors in the dataset. We weight the correlations by raising the correlation to the power β (parameter *power*) to limit spurious correlation, where the value of β yields a network with approximate scale-free topology. This essentially emphasizes strong correlations and punishes weak ones, yielding what we call the adjacency matrix. These are then summed to yield the connection strength or connectivity between the variables, which in turn tell us how strongly related each variable is to the other variables in the dataset. Next, we construct a topological overlap matrix to find which set of variables are strongly connected. The clusters are then determined by a hierarchical clustering tree algorithm. It is important to note that this is unsupervised learning—that is, the algorithm does not know the outcome variable for this part of the analysis. Thus the modules are created from the data predictors alone and not subject to any unconscious bias as to what "should" be clustered together.

While Conditional Inference Forests can also take into account correlation in Random Forest variable importance, they are computationally intensive and not feasible for large or even mid-sized

data sets (Conn et al., 2018; Conn and Ramirez, 2016). Note that simpler pairwise clusterings and alternative clustering algorithms that can be effortlessly incorporated into the `fuzzyforest` function call. However weighted correlation networks can yield valuable information about the strength of the connections between variables in the dataset allowing for a much more mechanistic interpretation of the variables.

The next step is recursive variable elimination Random Forests (RFE-RF) (Diaz-Uriarte and De Andres, 2006) on each of the modules. In each module, a Random Forest is created, and the variables with the lowest variable importance are dropped by a user-specified drop fraction (parameter *drop_fraction*). Until a user-specified percentage of variables remain in each module (parameter *keep_fraction*), a new Random Forest is run on the surviving variables, and the dropping process continues.

In the last step, all the survivors from each module are put into a last RFE-RF for a final filtering of user-specified number of variables (*number_selected*). Note that we allow users to pre-specify how many variables they would like, such as the top 20 important predictors. We also allow "favorite" variables to skip the screening step and to be placed in the final RFE-RF, although we did not do that for this analysis; we caution users about skipping the screening step, as that can potentially bias the results. Once the selected variables are finalized, a final Random Forest is produced.

Last, Fuzzy Forests is implemented by the `fuzzyforest` package in R, available at the Comprehensive R Archive Network (CRAN). There are many tuning parameters available at user discretion, including aforementioned *power*, *drop_fraction*, *keep_fraction*, and *number_selected*. In Appendix A (page 3) we discuss key tuning parameters in Fuzzy Forests, and the final set of parameters chosen. Note that the variable importance measures must be taken with a grain of salt, as discussed in Conn and Ramirez (2016).

4 Data

We study voter turnout in the 2016 presidential election, using the Cooperative Congressional Election Study (CCES). The CCES is a national stratified sample survey administered every two election years.[1] In 2016, there were 64,600 individuals surveyed. We limit our sample to respondents who answered both the pre- and post-waves of the survey, and who have been either (1) validated as having voted, (2) validated as having not voted, or (3) whose vote cannot be validated but who self-reported not being registered to vote or to have not voted in the general election. This reflects the belief that while social desirability bias may make some respondents

[1]Note that there are certain limitations with our use of the CCES data. Most importantly, the Common Content is liable to change between elections, especially with changes to the relevant issues at hand. This may make it difficult to compare our results to similar studies using CCES data from other presidential elections.

lie as if they have voted when they have not, there is no incentive to lie as if they have not voted, when they have.[2] This yields a total sample size of 47,782 respondents available for analysis.

Using the various theories of voter turnout discussed above, we choose 97 variables from the CCES "Common Content" (i.e., questions that were asked for all respondents). All key variables discussed in the literature are included, such as age, race, education, and income; additionally, we include a variety of political interest/behavior questions such as the respondent's ideological placement. We also include an array of issue questions, such as approval/disapproval of the Affordable Care Act (ACA).[3] Next, we one-hot encode them, resulting in 392 variables, including non-response variables.[4] This yields a reasonable sized dataset, with a large number of intercorrelated variables, and a sizable number of validated voters.

Table 1 shows the summary of the dependent variable (voter turnout), tabulated by key demographic variables (vote preference, gender, race, and age). In our sample, 74.4% of respondents reported voting in the 2016 presidential election, while 25.6% did not.

Table 1: Summary of Turnout and Demographics in CCES 2016 Data (Subset)

Turnout	Obs.	Voted/Preferred Candidate		Gender		Race		Mean Age
Yes	31,841 (74.4%)	Clinton	41.4%	Male	45.1%	White	80.0%	47.2
		Trump	48.9%	Female	54.9%	Black	8.4%	
		Others	9.3%			Hispanic	5.4%	
		Non-response	0.5%			Others	6.2%	
No	10,941 (25.6%)	Clinton	32.0%	Male	35.0%	White	69.3%	55.7
		Trump	38.2%	Female	65.0%	Black	10.5%	
		Others	3.5%			Hispanic	5.7%	
		Non-response	26.3%			Others	10.2%	

In Table 1 we see that of those respondents who actually voted, Trump has a decided edge in support. The table shows that 48.9% of respondents who stated Trump support voted while 41.4% of respondents who stated they supported Clinton voted. A similar gap exists among non-voters. We also see some uncertainty or ambivalence for non-voters, as 26.3% of the respondents did not have a preference between the two presidential candidates. The sample of non-voters in these data were more likely to be non-white and older than the sample of voters.

[2] See CCES Guide 2016, page 126.

[3] Lists of the variables that we use from CCES are available in the Appendix B, in Table 1.

[4] One-hot encoding refers to the practice of constructing binary indicators from categorical variables. This is a process to ensure better regression and classification performance, common in machine learning applications, and to ensure that categorical responses to survey questions—such as (1) for, (2) against, and (3) skipped, are treated as factors and not ordinal variables.

5 Turnout Model

5.1 Modules and variable Selection

Figure 1 shows the output of the weighted correlation network analysis on the variables, with their module labels. We use colors to describe our labels for brevity; other researchers might use more descriptive labels. Our preference is to label the modules here for descriptive purposes only, and to avoid ascribing meaning to the modules, except for the 'grey' module, which encompasses all variables without module assignment, i.e., variables without any particular strong connectivity detected. Considering our mid-sized dataset, we have set our minimum module size to have at least three variables.

The components of the modules are interesting in themselves, because Fuzzy Forests is able to detect blocks of similar responses to a similar set of questions. Many modules are clusters of item non-responses to surveys, either in the form of 'unsure' or 'skipped.' For instance, the 'black' module is a series of variables indicating that the respondent has skipped answering questions on which type of spending should be cut to cover the federal deficit. The 'turquoise' module is comprised of variables indicating that the respondent has answered 'unsure' when asked about the ideological positions of key political figures and groups, such as Obama, Clinton, Trump, the Democratic and the Republican parties, the Supreme Court, and the respondent themselves. It is interesting to see that Fuzzy Forests seem to be clustering these item non-responses with a similar topic, perhaps suggesting that voters are uncertain about these issues.

Indeed, the only module that does not consist of item non-responses is the 'green' module, which is an eclectic mix of abortion and LGBTQ policy questions. The abortion policy related questions ask respectively whether the respondent supports (1) "allowing a woman to obtain an abortion as a matter of choice", (2) "allowing employers to decline coverage of abortions in insurance plans", and (3) "prohibiting expenditure of funds authorized or appropriated by federal law for any abortion." While these are linked to support/oppose question for gay marriage, three other abortion policy questions are not included in the same module; these are whether the respondent supports (1) "permitting abortion only in case of rape, incest, or when the woman's life is in danger", (2) "prohibiting all abortions after the 20th week of pregnancy", and (3) "make abortions illegal in all circumstances." We do not include all of the 'grey' module variables in the interest of brevity. See Appendix C for full module membership.

After the module detection stage is completed, Fuzzy Forests then uses recursive variable elimination within each module until a pre-specified user-stopping criterion is met. In Figure 1, the listed variables in bold are the ones that are selected as being "important" and surviving the screening step. Since the dataset in question is mid-sized and minimum module size is 3, we keep 25% of variables in each module after recursive variable elimination. The variables chosen are marked with a darker font in each module in Figure 1. Thus, from Figure 2 we can ascertain

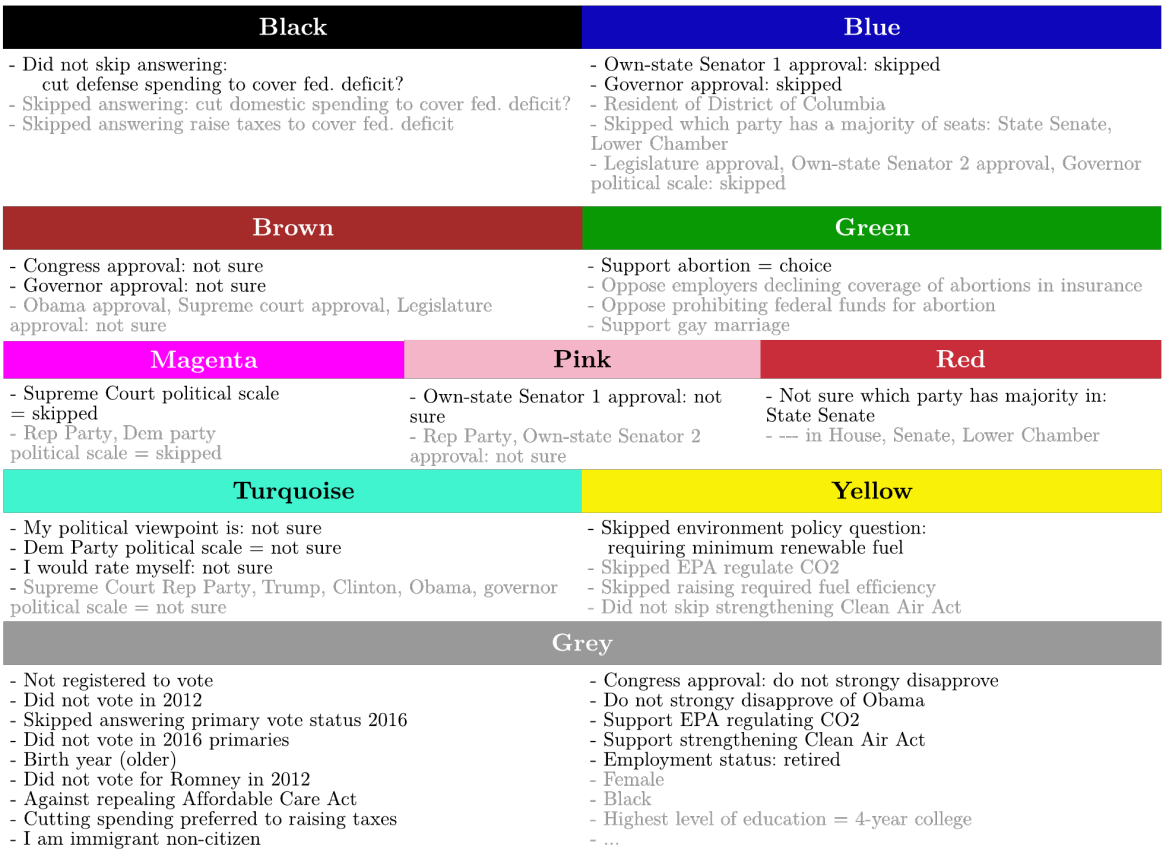

Figure 1: Estimated Modules from the Turnout Model

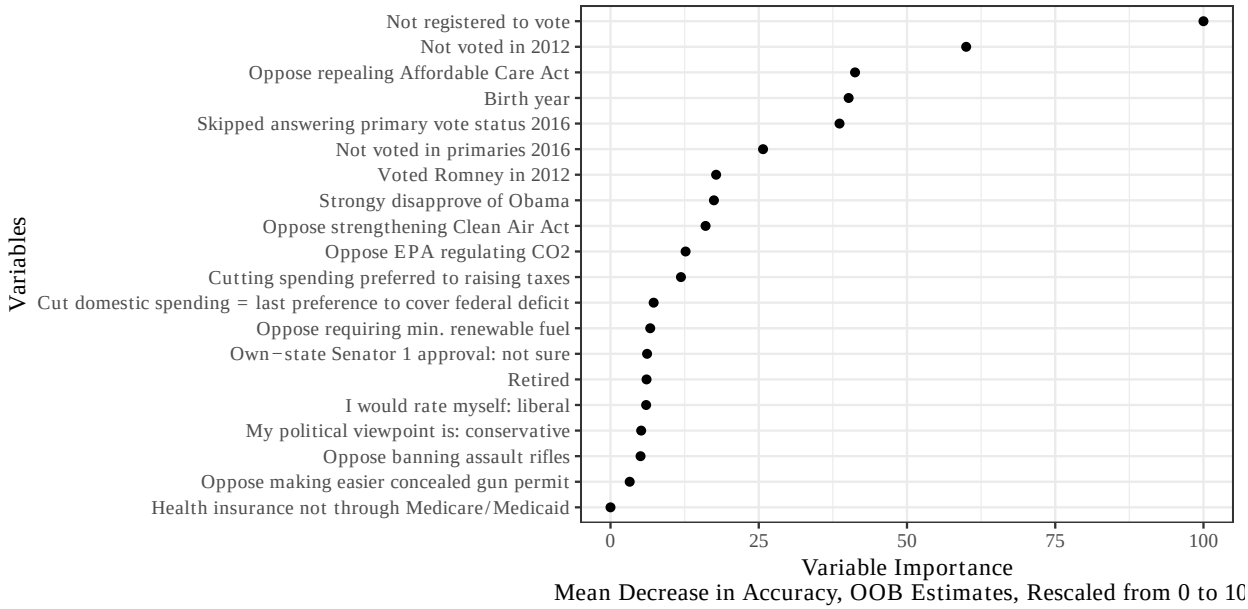

Figure 2: Variable Importance, Top Twenty Variables, Turnout Model

which important predictors are driving that module.

Lastly, the final selection step is carried out, i.e., recursive variable elimination of all variables regardless of the module. We set our Fuzzy Forests to select the top twenty variables. These top variables were mostly chosen from the 'grey' module, as can be seen from Figure 2, which shows a variable importance plot.[5]

Fuzzy Forests identifies two variables as the most important.[6] These variables are consistent with past research on voter turnout (Figure 2). If a respondent is not registered to vote, they are less likely to turn out; and if they did not vote in 2012, they are less likely to vote in 2016. In addition, if they skipped answering who they voted for in the primaries, which is highly correlated with not turning out for the 2016 primaries, they are less likely to turn out in the general election. These are related to the theories that suggest that turnout is habitual, and that it is influenced by administrative policies and political institutions (Gerber et al., 2003; Rosenstone and Wolfinger, 1978).

The other selected variables are also interesting. Save for age (birth year) and retirement, no demographics variables or social economic status variables were among the selected set—neither income, education, nor retrospective/prospective economic evaluations. This is interesting as past research has focused on demographic factors like education, finding that they are relatively

[5]Although the original Conn et al. (2018) cautions reading too much into these variable importance measures, we find that running Fuzzy Forests multiple times with different seeds and tuning parameters will produce almost an identical set of top twenty variables.

[6]A robustness check with only validated votes is provided in Appendix E. The majority of conclusions remain the same.

strongly correlated with voter turnout in presidential elections (Rosenstone and Wolfinger, 1978; Nagler, 1991). That part of the voter turnout literature, though, focuses on CPS datasets and heavily restricted model specifications—the CPS does not ask respondents about political identities, political opinions, or policy issues, meaning that demographic variables like education, income, and age are likely surrogates for political and non-political factors in those studies.

In addition, few of the item non-responses, which we argue measure political information and interest, are included in the top twenty. Only the variable indicating that the respondent answered 'unsure' about their approval of their own-state Senator finds its way into the top twenty variables.[7] We do see that a number of ideological and policy variables appear in the top twenty shown in Figure 2: the Affordable Care Act, the Environmental Protection Agency, cutting federal spending, the minimum wage, and gun policies. Immigration issues are, surprisingly given their salience in the election, not selected in the top 20 turnout predictors estimated by Fuzzy Forests. It is important to recall that we are only modeling if a person cast a vote, not for whom the ballot was cast.

It is interesting to note that there is predictive informational content in the non-response item of "Unsure about my own state's Senator's job approval", from the 'pink' module, rather than "I am not sure about my political viewpoint", from the 'turquoise' module, or the "Skip answering my own state's Senator's job approval", from the 'blue' module. This suggests that uncertainty or imperfect information plays an important role in 2016 turnout, but that some care needs to be taken to understand the specific domains where uncertainty is prevalent and how much domain-specific uncertainty deters turnout. Perhaps it may be more socially desirable to admit that you are unsure of the popularity of a particular politician rather than being unsure of your own political viewpoint. If voters are less informed, they can be less informed in a multitude of different ways—are they less informed about the presidential candidates? The Supreme Court? The upper or lower chamber? Their own political stance? Fuzzy Forests has suggested one of them as important, but again, there is no theory that gives any prior to whether one type of uncertainty should matter or be more representative over another.

Note that in Table 1, respondents who stated they favored Trump were more likely to cast a vote in the 2016 general election relative to those who stated a preference for Clinton. Most variables other than voter uncertainty and habitual behavior also hint at this. The variables that rank highest in terms of their predictive importance are those indicating whether the voter approves of the Affordable Care Act, and the ones who opposed to repealing the ACA were likely not to turn out to vote. This party-differential turnout may help explain why the predictions for the 2016 presidential election failed. That is, the candidate choice model of pollsters and scholars may have been accurate, but it may not have been predictive of whether that person would actually come and cast their vote on the day of the election, resulting in a surprise victory of

[7]The Senator here is dubbed 'Senator 1' in CCES, which is respectively displayed as the senior Senator of the respondent's state.

Trump over Clinton. This indicates that the Affordable Care Act may have been more decisive for turnout than other issue or policy—again, a result requiring further study.

Last, are there policy variables that are associated with 2016 turnout? For instance, are environmental advocates less likely to turn out generally, or is this only a relic of the context surrounding the 2016 election? Do they perhaps reflect Green party supporters who did not choose to strategically vote for Clinton? These are, again, variables that theory suggests may be important, but existing theory does not clearly connect to point predictions about voter turnout—underscoring the utility of a methodology like Fuzzy Forests.

Table 2: Counterfactuals for Policy Questions in the Test Set,
Top Twenty Policy variables Selected by Fuzzy Forests
(Baseline Predicted Turnout: 83.31%. True Turnout: 74.43%)

Counterfactual	Prediction	Value (%)	Difference (%p)
Everyone opposes repealing Affordable Care Act	↑	83.51	0.20
Everyone opposes strengthening Clean Air Act	↑	83.80	0.49
Everyone opposes EPA regulating CO2	↑	83.40	0.09
Everyone prefers cutting spending to raising taxes	↓	83.15	-0.16
Everyone thinks cutting domestic spending = last resort for federal deficit	↑	83.74	0.43
Everyone opposes requiring min. renewable fuel	↓	82.89	-0.42
Everyone opposes banning assault rifles	↓	82.83	-0.48
Everyone opposes making easier concealed gun permit	↑	83.33	0.02

We provide in Table 2 counterfactual estimates that show one way of estimating how much the policy variables that Fuzzy Forests identifies might have "mattered" regarding 2016 voter turnout. With all other variables fixed, we change the value of the variable in question to one extreme—e.g. everyone opposes repealing the Affordable Care Act—and predict turnout for each person in the test set, which gives us the third column. The fourth column calculates how it changes relative to the baseline prediction.

We see in Table 2 that even with these relatively extreme counterfactuals (i.e., assuming that *all* of the registered voters in the sample we use take one extreme position on each policy issue), the policy issues do not greatly change the overall predictions for voter turnout. This implies that these policy issues may not have played an overly large substantive influence on voter turnout in 2016, given that all other variables are held constant. That said, the results of this counterfactual analysis on the policy variables shows that there are issue-based strategies that could have altered turnout: for example, opposition to further strengthening of the Clean Air Act could have increased turnout, while opposition to certain types of gun control (here, banning assault rifles) would also have led to a slight increase.

6 Conclusion

In this paper, we used Fuzzy Forests to identify key variables that help predict 2016 presidential election turnout. Substantively, our use of Fuzzy Forests with the 2016 CCES data has shown how different variables cluster into modules, and then which specific variables within modules most "important." Our descriptive analysis found that voter turnout in 2016 was associated with procedural and administrative variables, past voting behavior, and an array of political opinions and policy preferences. Our results also indicated that demographic variables, which have been the focus of past research on voter turnout, do not appear as important variables in our Fuzzy Forests analysis.

Thus, the use of Fuzzy Forests in this application sheds new light on the 2016 presidential election. For example, we find that political issues, especially Obamacare, fiscal policy, and climate change, were associated with whether a registered voter participated in the election. Thus, Fuzzy Forests helps scholars of presidential elections understand the array of political issues that were associated with turnout decisions in 2016. Students of political participation will note that our results provide additional support for the importance of procedural variables in the study of turnout in U.S. presidential elections, but that our analysis does not find that many demographic variables other than age appear to be important for predicting turnout. Our conclusion is that the data is telling us that the demographic variables highlighted in past studies of turnout were likely proxy variables for political issues (as many past studies of turnout rely on survey datasets like the Current Population Survey that do not contain responses about opinions on political issues or other measures of political attitudes).

Methodologically, we find that machine learning is a useful variable screening method for studying over-determined questions like voter turnout in a major presidential election, resulting in a potentially less biased variable importance measure than other methods, while keeping the runtime short and the overall performance reasonable. In particular, Fuzzy Forests outperforms Random Forests in computational costs, while outperforming logit in its screening abilities (See Appendix D). Fuzzy Forests also offers more insights into the data from the module building part of the algorithm.

We argue that machine learning approaches like Fuzzy Forests offer researchers a relatively fast, parsimonious variable selection model where the researcher loses little in terms of predictive accuracy. They also have the added benefit of an examination of the network structure underlying the data, which can lend insights into the mechanisms driving the predictions. Thus, we suggest that the modules and final selection of variables that Fuzzy Forests provides will be something the researcher will always want to check for when running tree-based models and other methodologies, even in a mid-sized dataset. In our application, we found that the key variables that determined turnout in the 2016 election are not always those that are theoretically deemed important—such as uncertainty of one's Senator—implying that a method like Fuzzy

Forests can offer new perspectives for studying turnout. This is especially true because Fuzzy Forests offers a systematic clustering of variables that are free from the researcher's preconceptions, bringing new insights into what aspects of political behavior may be interrelated. By this, we can find novel insights into why certain variables cluster and how they relate to actual voter turnout, helping us confirm the results of previous studies of turnout on past presidential elections.

References

Alvarez, R. M., editor (2016). *Computational Social Science: Discovery and Prediction*. Cambridge University Press.

Breiman, L. (2001). Random forests. *Machine Learning*, 45(1):5–32.

Conn, D., Ngun, T., Gang, L., and Ramirez, C. M. (2018). Fuzzy forests: Extending random forests for correlated high-dimensional data. *Journal of Statistical Software*.

Conn, D. and Ramirez, C. M. (2016). Random forests and fuzzy forests in biomedical research. In Alvarez, R. M., editor, *Computational Social Science*, chapter 6, pages 168–196. Cambridge University Press.

Diaz-Uriarte, R. and De Andres, S. (2006). Gene selection and classification of microarray data using random forest. *Bioinformatics*, 7(1):3.

Gerber, A. S., Green, D. P., and Shachar, R. (2003). Voting may be habit forming: Evidence from a randomized field experiment. *American Journal of Political Science*, 47:540–550.

Green, D. P. and Kern, H. L. (2012). Modeling heterogeneous treatment effects in survey experiments with bayesian additive regression trees. *Public opinion quarterly*, 76(3):491–511.

Hill, D. and Jones, Z. (2014). An empirical evaluation of explanations for state repression. *American Political Science Review*, 108(3):661–687.

Imai, K. and Strauss, A. (2011). Estimation of heterogeneous treatment effects from randomized experiments, with application to the optimal planning of the get-out-the-vote campaign. *Political Analysis*, 19(1):119.

Kastellec, J. P. (2010). The statistical analysis of judicial decisions and legal rules with classification trees. *Journal of Empirical Legal Studies*, 7(2):202–230.

Kim, H. and Loh, W. (2001). Classification trees with unbiased multiway splits. *Journal of the American Statistical Association*, 96:589–604.

Langfelder, P. and Horvath, S. (2008). Wgcna: an r package for weighted correlation network analysis. *BMC Bioinformatics*, 9:559.

Leighley, J. E. and Nagler, J. (2013). *Who votes now?: Demographics, issues, inequality, and turnout in the United States*. Princeton University Press.

Levin, I., Pomares, J., and Alvarez, R. M. (2016). Using machine learning algorithms to detect election fraud. *Computational Social Science: Discovery and Prediction/ed*, pages 266–294.

Montgomery, J. M. and Olivella, S. (2018). Treebased models for political science data. *American Journal of Political Science*, 62(3):729–744.

Montgomery, J. M., Olivella, S., Potter, J. D., and Crisp, B. F. (2015). An informed forensics approach to detecting vote irregularities. *Political Analysis*, 23(4):488505.

Muchlinski, D., Siroky, D., He, J., and Kocher, M. (2016). Comparing random forest with logistic regression for predicting class-imbalanced civil war onset data. *Political Analysis*, 24(1):87103.

Nagler, J. (1991). The effect of registration laws and education on us voter turnout. *The American Political Science Review*, pages 1393–1405.

Nicodemus, KK, P. and Malley, J. (1890). Predictor correlation impact machine learning algorithms: implications for genomic studies. *Bioinformatics*, 25:1884.

Ramirez, C. M., Abrajano, M. A., and Alvarez, R. M. (2019). Using machine learning to uncover hidden heterogeneities in survey data. *Scientific Reports*, 9.

Rosenstone, S. J. and Wolfinger, R. E. (1978). The effect of registration laws on voter turnout. *American Political Science Review*, 72(1):22–45.

Strobl, C., Boulesteix, A.-L., Zeileis, A., and Hothorn, T. (2007). Bias in random forest variable importance measures: Illustrations, sources and a solution. *BMC Bioinformatics*, 8(25).