CESifo WORKING PAPERS

8615
2020

October 2020

Regional Heterogeneity and U.S. Presidential Elections

Rashad Ahmed, M. Hashem Pesaran

CESifo

Impressum:

CESifo Working Papers
ISSN 2364-1428 (electronic version)
Publisher and distributor: Munich Society for the Promotion of Economic Research - CESifo GmbH
The international platform of Ludwigs-Maximilians University's Center for Economic Studies and the ifo Institute
Poschingerstr. 5, 81679 Munich, Germany
Telephone +49 (0)89 2180-2740, Telefax +49 (0)89 2180-17845, email office@cesifo.de
Editor: Clemens Fuest
https://www.cesifo.org/en/wp
An electronic version of the paper may be downloaded
· from the SSRN website: www.SSRN.com
· from the RePEc website: www.RePEc.org
· from the CESifo website: https://www.cesifo.org/en/wp

CESifo *Working Paper No. 8615*

Regional Heterogeneity and U.S. Presidential Elections

Abstract

This paper develops a recursive model of voter turnout and voting outcomes at U.S. county level to investigate the socioeconomic determinants of recent U.S. presidential elections. It is shown that the relationship between many socioeconomic variables and voting outcomes is not uniform across U.S. regions. By allowing for regional heterogeneity and using high-dimensional variable selection algorithms, we can explain and correctly predict the unexpected 2016 Republican victory. Key factors explaining voting outcomes include incumbency effects, voter turnout, local economic performance, unemployment, poverty, educational attainment, house price changes, urban-rural scores, and international competitiveness. Our results corroborate evidence of 'short-memory' among voters: economic fluctuations realized a few months prior to the election are indeed powerful predictors of voting outcomes as compared to their longer- term analogues. The paper then presents real time forecasts for the 2020 U.S. Presidential Election based on data available at the end of July 2020 which are then updated based on data available as of mid-October.

JEL-Codes: C530, C550, D720.

Keywords: voter turnout, popular and electoral college votes, simultaneity and recursive identification, high dimensional forecasting models, Lasso, OCMT.

Rashad Ahmed
University of Southern California
Los Angeles / CA / USA
rashadah@usc.edu

M. Hashem Pesaran
University of Southern California
Los Angeles / CA / USA
pesaran@usc.edu

First Version: October 3, 2020
Forecasts Updated: October 14, 2020
Authors are grateful to Ron Smith for helpful comments. Rashad Ahmed acknowledges partial financial support from USC Dornsife INET. The first working paper version (CESifo WP 8615) can be found at:
https://www.cesifo.org/en/publikationen/2020/working-paper/regional-heterogeneity-and-us-presidential-elections
The latest version can be found at:
http://www.econ.cam.ac.uk/people-files/emeritus/mhp1/Regional_Heterogeneity_and_U.S._Presidential_Elections.pdf

1 Introduction

The U.S. Presidential election of 2016 caught many by surprise. Most models and polls predicted a victory for the Democratic candidate, Hillary Clinton. She lost to Donald Trump, the Republican candidate, who won an overwhelming majority of electoral votes (304 out of 538) despite coming short on popular votes by around 2.9 million votes. Not only did many come to realize the inherent unpredictability of elections, it revealed that investigating the drivers of election cycles remains an open and important area of research.

The 2016 election highlighted one important reason why popular and electoral vote outcomes may not align – namely voter heterogeneity resulting from increased regional political polarization. In U.S. elections since 1828, there have been only four (out of forty eight) election cycles where the popular votes did not align with the electoral college outcomes. These were: 1876 (Rutherford versus Tilden), 1888 (Harrison versus Cleveland), 2000 (Bush versus Gore), 2016 (Trump versus Clinton).[1] The 1876 and 1888 elections occurred soon after the American Civil War when the country was still highly divided politically. It is particularly interesting that two out of four non-aligned election outcomes have occurred during the past five election cycles, partly reflecting the heightened divisions in the U.S. political landscape in the 21st century.

In the presence of growing political polarization, incorporating heterogeneity in presidential election models becomes even more necessary than ever for better understanding regional disparities in election outcomes, and for more reliable forecasting. This paper studies the determinants of election outcomes and their predictive content at the level of U.S. counties in a model which admits such heterogeneity. We rely on high-dimensional statistical modeling and consider many socioeconomic and demographic indicators at national, state and county levels, and in particular do not make use of polling data that are likely to be volatile and subject to sudden change. We build upon the earlier work of Fair [1978], and more recent developments of Zandi et al. [2020], also referred to as Moody's election model. While an advantage of the polling approach is that it theoretically elicits current electoral preferences directly, it is subject to a variety of sampling issues with survey outcomes contributing to significant total survey error (Kou and Sobel [2004], Biemer [2010], Shirani-Mehr et al. [2018], Graefe [2018]). In the presence of increased political polarization, polling approaches may become even less reliable due increased voter heterogeneity and the added difficulties of eliciting true voter intentions due to "socially desirable responding". Hence, forecasting performance based on polls has been mixed.

[1] 1960 (Kennedy versus Nixon) was very close but did not produce conflicting outcomes. In all four cases the Republican candidate lost the popular vote but won the electoral college vote.

Most statistical/econometric models of U.S. Presidential elections rely on relatively long time series data and primarily use aggregate socioeconomic and demographic indicators as potential predictors. However, time series models estimated over long time periods are generally subject to structural breaks. Certainly the factors influencing voting behavior and the make-up of the voting body changed since the 1950's and continues to evolve. We focus on more recent election cycles and consider the five completed election cycles since 2000. To compensate for lack of time series variations we use county level data and rely on cross-sectional variation to identify the key determinants of voter turnout and election outcomes. Variation at the level of U.S. counties admits an additional novel feature – it allows for modeling regional heterogeneity. If factors influencing voting behavior differ geographically across the U.S., then heterogeneity will capture this crucial feature of the data. Surprisingly, regional heterogeneity has received limited attention in the literature. Zandi et al. [2020] does allow for fixed effects in a state-level model, but assumes that all time-varying determinants of election outcomes have equal effects across states. The implicit assumption of such pooled models is that over time, voters across the U.S. are similarly affected by socioeconomic and political factors. Recent history suggests that this assumption could be too restrictive.

In view of the above considerations, our model allows for heterogeneity in the effects of socioeconomic and demographic factors on voter turnout and election outcomes across the eight U.S. regions, as defined by the Bureau of Economic Analysis (BEA). With county-level data we could have allowed for a greater degree of heterogeneity, allowing the socioeconomic indicators to have differential effects even at the individual state level. But such a fully heterogeneous approach is subject to its own drawbacks. First, some states do not have enough counties to consistently estimate state-specific models. To compensate, one could increase the time dimension by collecting historical data on states with a small number of counties, but this would increase the risk of structural breaks, and require county-level data to be available further back in time, which is not so in the case of many socioeconomic factors. Second, counties across state borders tend to share similar features, and pooling their data into regions is likely to result in more efficient estimates.

In addition to allowing for heterogeneity, we also address the issue of simultaneous determination of voter turnout and election outcomes, by modeling them together at the level of counties. A large and growing literature on voter turnout tends to study the phenomenon separately to voting, despite the intimate link that exists between the two choices. Zandi et al. [2020] cites that ignoring unexpected voter turnout was a key contributor to their incorrect 2016 election prediction. We adopt a recursive approach to deal with this simultaneity by first modeling voter turnout, and then condition the election outcomes on the fitted (predicted) values of voter turnout. We allow for regional heterogeneity by estimating

separate county-level panel regressions for the eight BEA regions, and compare the results to the estimates and predictions we obtain from pooled homogeneous models. We also apply high-dimensional variable selection algorithms to guide our selection and estimation procedure over a large set of potential covariates. We consider both penalized regression and high-dimensional variable selection techniques, and use the 'Least Absolute Shrinkage and Selection Operator' (Lasso, Tibshirani [1996]) as an example of the former, and 'One Covariate at a time Multiple Testing' (OCMT, Chudik et al. [2018]) as an example of the latter. Our collection of socioeconomic and demographic data across states and counties is largely motivated by the literature on election modeling. We consider economic variables such as local unemployment, income, house prices, government employment and healthcare expenditures. We also consider demographic and geographic indicators such as population density, urban-rural classification, poverty rates, education and religiosity. Inspired by recent evidence from Autor et al. [2016] and Jensen et al. [2017], we also test the effects of being economically 'left behind' and international competition on voting outcomes. In addition, our model is sufficiently flexible to allow for interactions intended to capture presidential and party incumbency effects on voter turnout and election outcomes.

We show that the relationship between many economic variables and voting outcomes are not uniform across U.S. regions. First using only data available prior to the 2016 Presidential election, we estimate a model allowing for such regional heterogeneity and show that it forecasts correctly the unexpected 2016 Republican candidate victory. By contrast, we find that a standard model which pools information across counties at the national level would have predicted a presidential victory for the Democratic candidate – in line with a majority of 2016 presidential forecasts leading up to the election. The two regional models we estimated using Lasso and OCMT, respectively, forecast 308 and 307 electoral votes for the Republican candidate, compared to the actual 304 won by Donald Trump in 2016. These results support the view that political polarization across regions contributed to the surprise 2016 presidential election (Sides et al. [2017], Gelman and Azari [2017]). Moreover, models incorporating regional heterogeneity and variable selection vastly improves electoral predictions among key swing states which drive the resulting Republican victory in 2016. Meanwhile, models which pool all of the data more accurately predicted overall popular vote outcomes, which the Democratic party won in 2016.

We then further investigate regional heterogeneity in the determinants of election cycles by estimating the recursive model over the full sample from 2000 to 2016. Our analysis corroborates the usefulness of several variables identified in the literature as important in explaining voting outcomes. At the same time, we highlight the extent of geographical variation in the estimates and their importance for our forecasting analysis. Important factors

explaining voting behavior include voter turnout, local economic performance, unemployment, poverty rates, education, house price changes, urban-rural scores. Our results also corroborate evidence supporting incumbency effects and evidence of 'short-memory' among voters: economic fluctuations realized a few months prior to the election are indeed more powerful predictors of voting outcomes as compared to their longer-run analogues.

Based on available data at the time of forecasting (early September 2020) we have also updated the 2016 model specification to generate predictions for the 2020 U.S. Presidential election under different regional model specifications. Our predictions suggest the outcomes to be very close. The Lasso-regional model forecasts Republicans winning 260 electoral votes, while the OCMT-regional model forecasts Republicans winning 290 electoral votes, recalling that 270 electoral votes are need for a win. Averaging the county-level predictions of these two models we predict 269 electoral votes for the Republican candidate. All models point to a popular vote favoring the Democratic candidate.

The rest of this paper is organized as follows: Section 2 presents our modeling approach and its relation to the literature. Section 3 characterizes our two-stage model of voter turnouts and election outcomes. Section 4 discusses our identification procedure to consistently estimate the model. Section 5 goes over the data used in the analysis. Section 6 discusses how we consolidate the data into 'active sets' prior to estimation and Section 7 covers variable selection techniques applied during estimation. Section 8 describes the U.S. Electoral College process from which we generate election forecasts using county level predictions. Section 9 evaluates the 2016 U.S. presidential election outcome under our framework, generating 2016 election forecasts only using data available prior to the election. Then Section 10 more broadly investigates key determinants of U.S. election cycles using data over the full sample from 2000 to 2016. Section 11 provides and discusses the forecasts for the 2020 U.S. Presidential election. Section 12 concludes.

2 Our Modeling Approach and its Relation to the Literature

Generally speaking, two approaches are considered in modeling and predicting U.S. presidential elections: statistical (econometric/machine learning) and polling, or a combination of the two (Leigh and Wolfers [2006]). Political opinion polls exclusively rely on survey responses and aim to elicit the voting intentions of respondents (Wang et al. [2015]). Opinion polls provide timely information on possible election outcomes, but have a number of well known shortcomings, including sample selection bias which tends to become accentuated

due to voter heterogeneity, and the phenomenon known as socially desirable responding, which is believed to have biased the polling outcomes in favor of Hillary Clinton during the 2016 election.[2] See, for example, Kou and Sobel [2004], Biemer [2010], Shirani-Mehr et al. [2018], Graefe [2018].[3] The statistical approach primarily relies on demographic and socioeconomic indicators to predict election outcomes believing that voting intentions are formed largely by voters' personal experiences and their counterfactual evaluation of socioeconomic outcomes under alternative candidates. Among the statistical approach, time-series models have historically dominated, starting from the seminal work of Kramer [1971], Fair [1978], Fair [1996], and Arcelus and Meltzer [1975]. More recently, Kahane [2009], Hummel and Rothschild [2014]. Zandi et al. [2020] extend time-series models using panel data, estimating state-level models for U.S. elections. Zandi et al. [2020] employs fixed effects panel regressions which allow for some state-level heterogeneity through the intercepts, but otherwise all time-varying determinants of election outcomes are assumed to have homogenous effects across all states. The aggregate time series and the state-level panel data models both rely on time series dimension of the panel, T, to be sufficiently large to obtain reasonably precise estimates of the relationship between socioeconomic variables and the election outcomes. This in turn requires model stability which is unlikely to hold over long time spans, particularly considering that the socioeconomic determinants of election cycles in the 1950's are unlikely to apply in the 21st century.

To deal with the heterogeneity and possible model instability, we exploit variations across 3,107 mainland U.S. counties and consider only the last five election cycles starting with 2000 (Bush-Gore election) to avoid possible model instability. In principle, we could allow for the effects of socioeconomic factors to differ across all the 48 mainland states.[4] However, some states have only a few counties, and with the time dimension being quite small (with $T = 4$, noting the data for the 2000 election must be used for construction of lagged values), the state level estimates are unlikely to be reliable and could introduce unexpectedly large sampling errors into the analysis. Furthermore, counties across state borders often share similar features such that estimation could be made more efficient by pooling information from such neighboring states. We address these challenges by grouping the states into eight regions defined by the Bureau of Economic Analysis (BEA), and estimate eight separate regional panel regressions. In this way we hope to strike a balance between allowing for

[2]Stratified sampling is required for reliable polling which could be quite costly to implement properly, especially in a vast country with sizeable political heterogeneity such as the U.S..

[3]Opinion polls are to be distinguished from exit polls that are a kind of "nowcasting" and are not of concern in this paper.

[4]We do not model turnout and election outcomes for Alaska and Hawaii, and with some justification assume that the election results for these states in 2012 carry over to the 2016 and 2020 elections.

heterogeneity and achieving reasonable estimation precision. A pre-determined regional classification ensures against data mining and provides a level of heterogeneity suitable for the data.[5] We can, therefore, capture possible regional differences in voting preferences and, more generally, differences in demographic, social, and economic heterogeneity across the United States. Our modeling framework allows for intercept and slope heterogeneity across regions, while assuming homogeneity within regions. Our model generates predictions for 3,107 counties for a given election year. We further aggregate these predictions to generate state level and national level popular vote predictions, as well as electoral college vote predictions.

Several recent papers have studied the geographical determinants of election outcomes, focusing on cross-county variation. Economic performance linked to international competitiveness has been shown to influence county-level voting preferences in Autor et al. [2016] and Jensen et al. [2017]. Scala and Johnson [2017] identify large differences in voting preferences across the rural-urban spectrum in elections from 2000 to 2016. In a cross-sectional study, Kahane [2020] shows that the rural-urban spectrum, poverty rates, education, among several other demographic factors, shaped 2012 and 2016 election outcomes. Like these studies, we investigate U.S. election cycles, specifically exploiting variation at the U.S. county level while also allowing for regional heterogeneity. However, the scope of our work not only allows for ex-post evaluation, it can also be used for forecasting election outcomes, as we show by reporting predictions for the 2020 U.S. Presidential Election. Moreover, we rely on recent advances in high-dimensional data analytic techniques to guide our analysis both for selecting important determinants of voting outcomes and also for evaluating elections. Modeling elections is a high dimensional, mixed-frequency problem. Many potential economic and demographic explanatory variables have been documented in the literature. These variables are observed at different frequencies, and their long-term versus short-term impact on voting outcomes is not necessarily the same. We consider both penalized regression and variable selection adjusting for multiple testing. Specifically, we apply Lasso (Tibshirani [1996]) and the One, One-Covariate-at-a-time-Multiple-Testing (OCMT) procedure proposed in Chudik et al. [2018], respectively. See Section S3 of the Online Supplement for further details.

[5] We do not follow the alternative statistical grouping strategy whereby the number and the membership of the groups are determined by machine learning techniques. This could be the subject of future research.

3 Modeling Turnout and Election Outcomes

3.1 Voter turnout

One novel departure of our modeling strategy from the prevailing literature is the joint consideration of voter turnout and election outcomes. Voter turnout and election outcomes have traditionally been studied separately. Zandi et al. [2020] discusses election scenarios based on low, medium and high turnouts, but does not explicitly model the turnout process.[6] By contrast, we impose a recursive strategy to consistently model the simultaneous voter turnout and election outcomes.

Understanding voter turnout, like voting behavior itself, is a topic of interest among many political scientists and economists. Despite its importance, there is no consensus on what best explains, causes, and/or predicts turnout. As a result, researchers have approached the question from several different angles. Early research on understanding voter turnout can be traced back to Powell [1986] and Jackman [1987]. Both studies look at cross-country voting patterns and uncover a similar theme where countries with greater institutional quality also have higher voter turnouts.[7] More recent research, however, argues that the role of institutional quality is much less clear-cut (see Blais [2006]), highlighting the challenges faced by researchers attempting to understand voter turnout.

Given its long and active history, a wide variety of theories and research approaches have led to many interesting findings. For example, survey-based approaches - where survey-takers are simply asked whether they will vote - have been used for predicting voter turnout. Despite their drawbacks (e.g. Social Desirability Bias) survey data used directly or fed into a statistical model have both been shown to predict turnouts with mixed results (Rogers and Aida [2014], Keeter et al. [2016]). Alternatively, several empirical studies show significant associations between voter turnout and socioeconomic factors, including campaign spending, voting history, contact with campaign workers, sector of employment, marital status, education, gender, age and income. See, for example, (Wolfinger and Rosenstone [1980], Matsusaka [1995], Rogers and Aida [2014]).[8] The likelihood of voting has even been linked to genetics. See (Fowler and Dawes [2008] and Fowler et al. [2008]).

Cancela and Geys [2016] conduct a meta-analysis of 185 articles focused on voter turnout in the U.S., finding that campaign expenditures, election closeness and registration require-

[6] Zandi et al. [2020] find that their predictions errors for 2016 are largely explained by unexpected turnout, and their 2020 election prediction crucially depends on which scenario is adopted for turnout.

[7] These qualities include: competitive districts, electoral disproportionality, multipartyism, unicameralism, and compulsory voting.

[8] In contrast, Matsusaka and Palda [1999] show that, despite statistical significance, explanatory power for predictive purposes is not much better than if one were to guess randomly.

ments have more explanatory power in national elections, whereas population size and composition, concurrent elections, and the electoral system play a more important role for explaining turnout at subnational elections. More recently, machine learning methods, trained on individual-level socio-demographic data have been applied by campaigns to micro-target potential voters (Rusch et al. [2013]). A recent research on voter turnout which is particularly relevant to our analysis is the paper by Biesiada [2018],who analyzes county-level voter turnout and finds that inequality, education, past voter turnout, gender proportion and median age are significantly associated with turnout at the county-level. We shall make use of these insights in arriving at the set of potential covariates that we will be using for our regional models of voter turnout.

3.2 Log-odds ratio of Republican to Democrat votes

Consider county c located in region r for the election years $t = 2000, 2004, 2008, 2012, 2016$, and denote the log-odds ratio of Republican to Democrat votes for this county by $LRO_{cr,t}$. Specifically, let

$$LRO_{cr,t} = \ln\left(\frac{R_{cr,t}}{D_{cr,t}}\right) = \ln\left(\frac{V_{cr,t}}{1 - V_{cr,t}}\right), \tag{1}$$

where $R_{cr,t}$ and $D_{cr,t}$ denote Republican and Democratic votes, respectively, and $V_{cr,t} = R_{cr,t}/(R_{cr,t} + D_{cr,t})$ is the Republican vote share in year t.[9] The BEA regional classification groups the 48 mainland states and the District of Columbia into eight regions: New England, Mideast, Southeast, Great Lakes, Plains, Rocky Mountain, Southwest, and Far West.

While the literature tends to study the two-party vote share, $V_{cr,t}$, we have chosen to consider the log-odds ratio variable, $LRO_{cr,t}$. Our preference for the log-odds ratio is its wider range of variations $(-\infty, +\infty)$ as compared to $(0, 1)$ for $V_{ct,t}$, and the fact that its use as the dependent variable universally provides better in-sample fits as compared to using $V_{cr,t}$.[10] The use of $LRO_{cr,t}$ is also more likely to support the linearity assumption made in the panel regressions specified below. Also to deal with the highly persistent nature of the LRO variable we use the transformation $DLRO_{cr,t+4} = LRO_{cr,t+4} - LRO_{cr,t}$, namely the change in the log-odds ratio from one election cycle to the next, for county c in region r. For

[9]The use of LRO as a measure of election outcome assumes that the effect of third party independent candidate(s) on the two-party race outcome is negligible. This assumption seems reasonable for the election cycles 2016 and 2020 that are the focus of this paper.

[10]We empirically validate our choice of the functional form by comparing model fit across both dependent variables: the log-odds ratio and the traditional vote share measure. We find that the log-odds ratio indeed improves the model fit over the vote share. Details can be found in Section S2 of the Online Supplement.

each region $r = 1, 2, ..., 8$ we consider the following separate panel regressions

$$DLRO_{cr,t+4} = a_{DLRO,r} + \phi_r' \mathbf{z}_{DLRO,cr} + \beta_r VT_{cr,t+4} + \gamma_r' \mathbf{x}_{DLRO,cr,t+3} + \varepsilon_{cr,t+4}, \tag{2}$$

where $a_{DLRO,r}$ is the region-specific fixed effects, time-invariant county-specific covariates are represented by $\mathbf{z}_{DLRO,cr}$, and state or county level time-varying covariates from the year preceding the election are included in $\mathbf{x}_{DLRO,cr,t+3}$. In our application, $t \in \{2000, 2004, 2012, 2016\}$ and therefore $t + 4$ denotes the upcoming election (four years after the year t election, and $t+3$ denotes the year preceding the upcoming election. The voting outcome is also a function of the voter turnout variable, $VT_{cr,t+4}$.

We define voter turnout of county c in region r in election year t as

$$VT_{cr,t} = \frac{R_{cr,t} + D_{cr,t}}{VAP_{cr}}, \tag{3}$$

which is equal to the total two-party votes as a proportion of the voting age population (VAP_{cr}) of county c in region r for election year t. VAP is considered time-invariant due to its persistent, slow-moving nature. Specifically, our measure of VAP is reported as a 5-year average. Due to data availability, we use 2012-2016 VAP estimates for 2016, 2008-2012 estimates for 2012, and 2005-2009 estimates for 2008 and 2004 elections.

In the year of the election, $VT_{cr,t+4}$, voter turnout, like $DLRO_{cr,t+4}$, is determined by a variety of demographic and economic factors:

$$VT_{cr,t+4} = a_{VT,r} + \psi_r' \mathbf{z}_{VT,cr} + \lambda_r VT_{cr,t} + \delta_r DLRO_{cr,t+4} + \theta_r' \mathbf{x}_{VT,cr,t+3}, + v_{cr,t+4}, \tag{4}$$

such that turnout is a function of time-invariant and time-varying variables, along with the turnout from the previous election, and also the change in the log-odds ratio, $DLRO_{cr,t+4}$. We allow the innovations to the $DLRO_{cr,t+4}$ and $VT_{cr,t+4}$ equations to be correlated, $cov(\varepsilon_{cr,t+4}, v_{cr,t+4}) \neq 0$, which reflects the simultaneity of the decision to vote and for which candidate to cast one's vote.

In both voter and turnout equations, time-invariant factors can include slow-moving socioeconomic and demographic factors like education, migration, religiosity, and urban-rural classification. Time-varying factors include local unemployment rate, poverty rate, household median income, changes in house prices, government and private employment, among others.

Notice that Equations 2 and 4 represent a system of simultaneous equations. Voting is a function of voter turnout (one can only vote if one shows up), and voter turnout is (in general) a function of the voting outcome. This introduces endogeneity into the voting process and

biases the least squares estimates of β_r and δ_r when $\varepsilon_{cr,t+4}$ and $v_{cr,t+4}$ are correlated. Non-zero correlations between $\varepsilon_{cr,t+4}$ and $v_{cr,t+4}$ could arise due to common beliefs about the election outcomes. For example, strongly held beliefs about the election outcome in a given state might adversely impact the decision to vote, whilst the decision to vote clearly does affect election outcomes no matter which way the voter decides to cast his/her vote.

4 Recursive Identification

The estimation of $DLRO_{cr,t+4}$ and $VT_{cr,t+4}$ equations clearly encounters an identification problem very much akin to the identification of demand and supply shocks in standard supply-demand models in economics. However, if one is concerned with prediction, a reduced form model of $DLRO_{cr,t+4}$ can be used where the turnout variable $VT_{cr,t+4}$ is solved out and $DLRO_{cr,t+4}$ is defined only in terms of the union of predetermined variables included in the two equations. Such an approach ignores the possible contemporaneous effect of voter turnout on election outcomes and could lead to inefficient predictions. In this paper we follow the alternative structural approach, and identify the model by imposing a triangular restriction on the contemporaneous dependence between $DLRO_{cr,t+4}$ and $VT_{cr,t+4}$, namely by setting $\delta_r = 0$. The intuition behind this restriction is that the individual decision to vote is not affected by his/her expected state-level collective outcome. This type of restriction is inspired by the pioneering work of Wold [1960], and is known as recursive causal ordering and often adopted in empirical macroeconomic analysis of simultaneous equation systems. But note that we do allow for contemporaneous dependence between the innovations to the voter turnout and the election outcome. In this sense the identification scheme adopted can be viewed causal with VT causing $DLRO$ and not *vice versa*.

We believe the recursive ordering, with $VT_{cr,t+4}$ included first, is a plausible *a priori* restriction, especially in the U.S. context where presidential elections are held simultaneously with other local and state-level elections, covering the election for the Senate and all the House seats. These additional elections influence turnout regardless of expected presidential ballot outcome. Second, the data and the literature suggest that turnout is highly persistent. Moreover, the existence of 'blue' states and 'red' states – states which consistently and predictably vote for one of two parties – suggests that turnout does not collapse when collectively there are strong expectations for a particular party to win the state.

Subject to the identifying restriction, $\delta_r = 0$, consistent estimation of the remaining parameters of the $VT_{cr,t+4}$ and $DLRO_{cr,t+4}$ equations can be carried out recursively using a two-stage estimation procedure. In the first step the turnout equation ($VT_{cr,t+4}$) is estimated by least squares, and then the *fitted* values of voter turnout (denoted by $\widehat{VT}_{cr,t+4}$) is used as

a regressor in the election outcome equation ($DLRO_{cr,t+4}$).[11] The estimating equations can now be written as

$$\widehat{VT}_{cr,t+4} = \hat{a}_{VT,r} + \hat{\psi}_r' \mathbf{z}_{VT,cr} + \hat{\lambda}_r VT_{cr,t} + \hat{\theta}_r' \mathbf{x}_{VT,cr,t+3}, \tag{5}$$

and

$$\widehat{DLRO}_{cr,t+4} = \hat{a}_{DLRO,r} + \hat{\phi}_r' \mathbf{z}_{DLRO,cr} + \hat{\beta}_r \widehat{VT}_{cr,t+4} + \hat{\gamma}_r' \mathbf{x}_{DLRO,cr,t+3}. \tag{6}$$

We allow for regional heterogeneity in both equations – all coefficients are specific to region r. In addition to the eight region-specific panel regressions, we also consider a pooled model for comparison purposes. The pooled model is a restricted version of the heterogeneous model such that all coefficients in the turnout and voting equations are restricted to be the same across all the regions, namely $a_{TO,r} = a_{TO}$, $\lambda_r = \lambda$, $a_{DLRO,r} = a_{DLRO}$, $\beta_r = \beta$, and so on. The regional and pooled models are estimated by least squares, subject to the variable selection problem that will be addressed below.

5 Electoral and Socioeconomic Data and Their Sources

We use data from county-level presidential votes and turnouts for five U.S. elections: 2000, 2004, 2008, 2012, and 2016. Because we model the *change* in the log-odds ratio of Republican vote, our regression estimates are based on four election cycles: 2000-2004, 2004-2008, 2008-2012, and 2012-2016. Our data set is composed of a total of 3,107 counties over the mainland 48 states plus Washington D.C. for a total of 12,428 county-election cycles.[12] Each state, and therefore each county, falls into to one of the eight BEA regions. The list of states included in these regions is given in Table 1. Figures S.2 and S.3 of the Online Supplement show the histograms of the voter turnout variable, VT, and the change in Republican log-odds ratio, $DLRO$, respectively, both for the mainland US, as well as for the eight BEA regions.[13] These histograms provide a visual account of the degree of regional heterogeneity in VT and $DLRO$ variables which, as we shall see, play an important role in understanding and predicting U.S. presidential election outcomes.

As predictors of voter turnout and election outcomes we consider two categories of covariates: time-invariant and time-varying. Data on time-invariant covariates tend to be collected at low frequencies and either do not vary or show very little variation over the four

[11] A formal proof of consistency is provided in Section S4 of the Online Supplement.

[12] The number and composition of some of the counties have undergone some changes over the past two decades. The procedure we followed to deal with these changes is explained in the Online Supplement.

[13] To save space additional summary and result tables, and and figures are provided in the Online Supplement and designated with the prefix letter S.

Table 1: Bureau of Economic Analysis regional classification with Swing States designated in bold

	BEA Region	States
1	New England	ME, **NH**, VM, MA, RI, CT
2	Mideast	NY, NJ, **PA**, DE, MD, DC
3	Southeast	**VA**, **NC**, SC, GA, **FL**, KY, TN, AL, MS, AR, LA, WV
4	Great Lakes	**MI**, **OH**, IN, IL, **WI**
5	Plains	**MN**, MO, KS, NE, **IA**, SD, ND
6	Rocky Mountains	MT, ID, WY, UT, **CO**
7	Southwest	TX, OK, NM, AZ
8	Far West	CA, **NV**, WA, OR, AK, HI

election cycles that we are considering - we treat all such variables as time-invariant and use their time averages if needed. These include measures on county demographics, education, religiosity, migration, population density, urban-rural classification scores, and vote-by-mail policy of the state. Time-varying measures vary at state or county levels. These include economic data on unemployment rates, house prices, poverty rates, and median incomes. Moreover, we consider data on export-weighted real exchange rates by U.S. state (as a proxy for international competition) , government size, healthcare costs, inflation, and Midterm elections, that vary across states but do not vary across counties within a given state.

The choice of the covariates is guided by the literature. But we also include a new covariate that measures relative economic performance to gauge the degree a county has been 'left-behind'. This is measured as county c's annual real GDP growth relative to the national and/or the regional average real GDP growth. We find that being economically left behind over the past several years is significantly correlated with changes in the Republican vote share, and we therefore incorporate this novel measure as a covariate to explore its implications further. See Section S1 of the Online Supplement for further details.

To capture spatial effects, we compute and incorporate local average measures of several county-level covariates. The local variables corresponding to county c are the average of individual county measures of all counties within 100 miles of county c. We consider both individual and local measures for many county-level variables including migration and education, while unemployment and house prices are computed as local measures only. Local variables are denoted with a '*'. Hence, "edu2000" and "edu2000*" correspond to individual and local education rates, respectively. County house prices and unemployment rates are always local averages.

The dynamic nature of election cycles admits additional complexity into the prediction problem. Dynamics matter, and voters may place differential weight on determinants of

their vote depending on not just what was realized, but *when* it was realized relative to the election. The literature, for example, documents a strong short-lived memory among voters, who typically consider only the past year's economic performance when evaluating the incumbent party's overall performance. To embed these features in our model, we take a mixed-frequency approach and include both short-term and longer-term measures of our time-varying covariates which have data reported at high (monthly) frequencies. This includes three variables: county house price changes, county unemployment rates, and state export-weighted real effective exchange rates. For example, we include annual average house price changes as well as house price changes three months in the election year but prior to the election held in November. We do similarly for unemployment rates and exchange rates, to capture both shorter-term and longer-term effects of economic conditions on the voting behavior. 1-year and 3-month average unemployment rates will be denoted by "ump_L1" and "ump_M3", respectively. The 1-year average is computed over the 12 months from June in the election year to July of the previous year, and the 3-month average is computed using data for July, August and September of the election year.

Finally, to capture the incumbency effects on voter turnout and election outcome we consider two types of indicators, and distinguish between presidential and party incumbency indicators. The "incumbent party indicator" takes the value of 1 if on the election day the president in power is Republican and -1 if he/she is a Democrat. The "incumbent president indicator" takes the value of 1 if the president who is running for re-election is a Republican, takes the value of -1 if he/she is a Democrat, and takes the value of 0 if neither of the two candidates is incumbent. These indicators are considered on their own, as well as interacted with a number of other covariates. In this way we allow for a wide variety of incumbency effects (positive or negative) discussed in the literature, without biasing the results in favor of or against the incumbent president or party.

Additional information on data sources, the transformations used to construct the covariates and data cleaning carried out to deal with changes in county boundaries and other variable definitions, are provided in Section S1 of the Online Supplement.

6 Active Sets for VT and $DLRO$ Panel Regressions

As is clear from the above account, there are many covariates that can be considered as potential predictors of VT and $DLRO$ variables, and some variable selection is required to avoid over-fitting. Variables for the voter turnout regression, $\mathbf{z}_{VT,cr}$ and $\mathbf{x}_{VT,cr,t+3}$, are taken from a set of covariates designated to turnout. Similarly, covariates for the voting odds ratio regression, $\mathbf{z}_{DLRO,cr}$ and $\mathbf{x}_{DLRO,cr,t+3}$, are selected from a different set designated to the

voting equation. We refer to these sets as 'Active Sets' for VT and $DLRO$, respectively.

First, we construct a single data set which includes many individual and local measures, temporal lags, incumbency indicators and their interactions. The result is a large set of potential predictors which reflect changes in social, economic, or demographic conditions across both space and time. Many of these variables are highly correlated with each other. Therefore, to discipline our estimation procedure, active sets contain exclusively the set of covariates to be considered by the regression model. The choice of potential covariates is largely inspired by the literature. We also account for the temporal effects, again inspired by the literature, documenting a strong short-memory among voters such that they tend to disproportionately overweight economic progress or deterioration made within the several months preceding the election (and ignore longer-lived developments over the entire 4-year election cycle) .

Table 2: Summary statistics for the covariates in the active set for VT panel regressions over the period 2000-2016

Covariate	Description	Mean	St. Dev.	Regional Coverage
r_incu_pa	indicator taking 1 if incumbent party is Republican, -1 if incumbent party is Democrat	0.000	1.000	National
r_incu_pr	indicator taking 1 if Republican re-election, -1 if Democratic re-election, 0 if no re-election	0.000	0.707	National
VT_L1	voter turnout proportion	0.564	0.097	County
VT_L1 x r_incu_pa.	Lagged VT interacted with incumbency indicator	0.015	0.583	County
hlt_L1	change in log healthcare expenditures, year preceding election	0.046	0.016	State
gov_L1	change in log government employment, year preceding election	−0.012	0.015	State
ump_L1	unemployment rate avg., year preceding election	0.061	0.020	County
hpret_L1	change in log house prices avg., year preceding election	0.022	0.043	County
rp_L1	change in log rental expenditure, year preceding election	0.032	0.012	State
religion	religiosity rate	0.511	0.170	County
religion x r_incu_pa.	religion interacted with incumbency indicator	0.000	0.539	County
migrate	net migration (time-invariant)	0.005	0.009	County
migrate x r_incu_pa.	migrate interacted with incumbency indicator	0.000	0.010	County
edu2000	proportion with bachelor's degree or higher (time-invariant)	0.165	0.078	County
edu2000 x r_incu_pa.	edu2000 interacted with incumbency indicator	0.000	0.183	County
ln(m.inc)	log median household income	10.633	0.254	County
ln(m.inc) x r_incu_pa.	ln(m.inc) interacted with incumbency dummy	−0.075	10.636	County
povr	poverty rate	0.155	0.062	County
povr x r_incu_pa.	povr interacted with incumbency dummy	−0.013	0.167	County
rural	urban-rural score (-4 to 4, time-invariant)	0.111	2.680	County
rural x r_incu_pa.	rural interacted with incumbency dummy	0.111	2.680	County
vmail_d	indicator whether state mandates mail-in voting (1), optional (0), no mail-in voting (-1)	−0.301	0.564	State

Among time-varying factors ($\mathbf{x}_{DLRO,cr,t+3}$ and , $\mathbf{x}_{VT,cr,t+3}$) we include both short-run (3-months before the election) and medium-term (1-year preceding the election) changes in those measures which are observed at high frequency, like house price changes and local

unemployment rates. This allows economic changes which occur just prior to an election to have a different, potentially more powerful, impact on voting behavior compared to longer term changes in economic conditions. Time-invariant covariates ($\mathbf{z}_{VT,cr}$ and $\mathbf{z}_{DLRO,cr}$) include slow-evolving socioeconomic and demographic factors like migration, urban-rural score, education and religiosity.

Table 2 lists and describes the active set for the voter turnout (VT) variable. The active set contains a variety of national, county, and state-varying covariates. Voter turnout is a highly persistent process, and as such lagged turnout is also included in the active set. To account for covariates having effects which are party-agnostic, and rather go in favor or against incumbent parties, we interact several variables with an incumbent party indicator which indicates whether the current president is Democratic or Republican.

Table 3 lists and describes the active set for the change in log-odds ($DLRO$) variable. As with the model for voter turnout, this active set contains national, state, and county-level covariates. The number of regressors in the active set exceeds 30. Time-invariant active set regressors include population density, rural-urban score, education rates and migration rates. Covariates which vary over time include house election results, economic 'left-behind' variable (not included in the voter turnout regressions), healthcare costs, government employment share, export-weighted state-level real exchange rate changes, local unemployment, house price changes, rent costs and inflation. Notice also that this active set includes the fitted values of voter turnout variable, $\widehat{VT}$, which is obtained from the application of variable selection algorithms to the VT panel regressions. As a result the particular fitted values, $\widehat{VT}$, included in the active set for the $DLRO$ variable will depend on the outcome of the variable selection algorithm applied to the panel regressions for the VT variable (which mimic the recursive nature of our identification scheme). In a sense high-dimensional variable selection algorithms are applied twice, but recursively. With this in mind the summary statistics given for the VT variable in Table 3 refer to the realized voter turnout values, and not the fitted ones used for variable selection in the case of $DLRO$ regressions.

Finally, in the case of the regional models, we exclude state-level covariates (that do not vary across counties within a given state) listed in the active set because they do not provide sufficient variation and become collinear. The national or pooled model includes state-level covariates listed in the active sets as well.

7 Estimation and Variable Selection Algorithms

Given the high-dimensional nature of the problem, we consider two estimation/selection algorithms that address the over-fitting problem, namely cross-validated Least Absolute

Table 3: Summary statistics for the covariates in the active set for $DLRO$ panel regressions over the period 2000-2016

Covariate	Description	Mean	St. Dev.	Regional Coverage
r_incu_pa	indicator taking 1 if incumbent party is Republican, -1 if incumbent party is Democrat	0.000	1.000	National
dLRO_hous	change in log Republican odds from preceding House election	0.087	0.346	State
VT	voter turnout proportion from the first-stage VT regression	0.576	0.090	County
VT x r_incu_pa.	VT interacted with incumbency indicator	0.015	0.583	County
LBCG_L1	county 'Left-Behind' measure, year preceding election	−0.005	0.087	County
LBCG_L1 x r_incu_pa.	LBCG_L1 interacted with incumbency indicator	−0.002	0.087	County
hlt_L1	change in log healthcare expenditures, year preceding election	0.046	0.016	State
gov_L1	change in log government employment, year preceding election	−0.012	0.015	State
rusd_L1	change in log real effective USD, year preceding election	0.009	0.055	State
rusd_L1 x r_incu_pa.	rusd_L1 interacted with incumbency indicator	−0.047	0.031	State
rusd_M3	Change in log real effective USD, 3 months preceding election	−0.012	0.114	State
rusd_M3 x r_incu_pa.	rusd_M3 interacted with incumbency indicator	0.046	0.105	State
ump_L1	unemployment rate avg., year preceding election	0.061	0.020	County
ump_L1 x r_incu_pa.	ump_L1 interacted with incumbency indicator	−0.007	0.064	County
ump_M3	unemployment rate avg., 3 months preceding election	0.060	0.019	County
ump_M3 x r_incu_pa.	ump_M3 interacted with incumbency indicator	−0.004	0.063	County
hpret_L1	change in log house prices avg., year preceding election	0.022	0.043	County
hpret_L1 x r_incu_pa.	hpret_L1 interacted with incumbency indicator	0.001	0.048	County
hpret_M3	change in log house prices avg., 3 months preceding election	0.025	0.055	County
hpret_M3 x r_incu_pa.	hpret_M3 interacted with incumbency indicator	−0.007	0.060	County
rp_L1	change in log rental expenditure, year preceding election	0.032	0.012	State
inf_L1	inflation, year preceding election	0.021	0.022	State
migrate	net migration (time-invariant)	0.005	0.009	County
migrate*	local net migration (time-invariant)	0.010	0.006	County
edu2000	proportion with bachelor's degree or higher (time-invariant)	0.165	0.078	County
edu2000*	local proportion with bachelor's degree or higher (time-invariant)	0.165	0.040	County
ln(popdens)	log population density (time-invariant)	3.727	1.668	County
ln(m.inc)	log median household income	10.633	0.254	County
ln(m.inc) x r_incu_pa.	ln(m.inc) interacted with incumbency indicator	−0.075	10.636	County
povr	poverty rate	0.155	0.062	County
rural	urban-rural score (-4 to 4, time-invariant)	0.111	2.680	County

Mean and standard deviation for actual, not model-fitted voter turnout "VT" reported. In actual model estimation the active set for $DLRO$ contains $\widehat{VT}$, the fitted value of VT obtained from estimating Equation 5. Because $\widehat{VT}$ is model-specific, the mean and standard deviation of the fitted voter turnout $\widehat{VT}$ differs from actual VT and also varies across models.

Shrinkage and Selection Operator (Lasso) originally introduced by Tibshirani [1996]), and the One Covariate at a Time (OCMT) recently proposed by Chudik et al. [2018]. We estimate both nationally pooled and regional models, the latter allowing for heterogeneity across BEA regions. At the regional level, Lasso and OCMT is applied to the region-specific covariates, by pooling the observations over the four election cycles under consideration. The main difference between Lasso and OCMT is in the way they deal with the over-fitting problem. Lasso introduces a penalty term in the minimand used for estimation, and calibrates the extent of penalization by cross-validation (typically 10-fold cross-validation). The use of cross-validation is supported by Monte Carlo evidence for standard models with homoscedastic and cross-sectionally independent errors. But both of these assumptions are

likely to be violated in the case of the panel regressions on U.S. counties.

By contrast, OCMT is a multi-step algorithm which allows for multiple testing in variable selection. In the first stage, OCMT runs univariate regressions, one at a time, selecting significant covariates after adjusting the critical value of for multiple testing. In subsequent stages, OCMT includes all selected variables in the first stage in a multiple regression, and then re-tests those covariates which were not selected in the first stage, and so on. The critical values adjusted for multiple-testing given by $c_p(k,\delta) = \Phi^{-1}\left(1 - \frac{p}{2k^{\delta}}\right)$, where $\Phi^{-1}(.)$ is the inverse of the cumulative distribution function of the standard normal, p is the nominal size of the individual tests (not allowing for multiple testing), k is the number of covariates in the active set, δ captures the degree to which the critical values are adjusted for multiple testing. Extensive Monte Carlo experiments carried out by Chudik et al. [2018], suggest setting $\delta = 1$ in the first stage of OCMT and $\delta = 2$ in subsequent stages. We set $p = 0.05$ and note that the results are reasonably robust to setting $p = 0.01$ or 0.10.

We also adjust the standard errors of the individual tests used in the OCMT procedure for possible error variance heterogeneity and spatial dependence across counties, assuming that equation errors within the same state are correlated due to political boundaries and the state-level governing nature of the U.S., but rule out residual serial correlation. Accordingly, we base our computation of individual t-tests using standard errors clustered by state-year for the pooled model. This yields a reasonably large number of 196 clusters (49 states $\times$ 4 years). For the regional model, we cluster standard errors by state.[14]

Details of the selection and estimation procedures for Lasso and OCMT are provided in Section S3 of the Online Supplement.

8 U.S. Electoral College

U.S. elections are determined by the number of Electoral College votes obtained. The Electoral College consists of 538 electors and an absolute majority of 270 electoral votes is required to win the election. Each state is assigned a fraction of total delegates for the electoral vote. For example, the share of California in 2016 was 55/538. This share is to be compared to the share of popular votes by state, given by$w_{st} = (R_{st} + D_{st}) / (R_t + D_t)$, where R_{st} is the number of Republican votes in state s, and R_t is the total number of Republican votes across all states (including DC): $R_t = \sum_{s=1}^{51} R_{st}$. Similarly, for D_{st} and D_t. Let $V_{st} = R_{st}/(R_{st} + D_{st})$ and $V_t = R_t/(R_t + D_t)$, denote state-specific and national level shares of Republican votes, respectively. Then $V_t = \sum_{s=1}^{51} w_{st} V_{st}$, where w_{st} is defined above.

[14]Similar results are obtained if clustering is done at either the state-year or state level.

We can distinguish between an aggregate predictor of V_t and then declare the Republican candidate as the winner if $V_t > 0.5$. But if we follow the US Electoral College rule, we can only declare the Republican candidate as the winner if:

$$\sum_{s=1}^{51} w(d_s)\mathbb{1}(V_{st} - 0.5) > 0.5 \tag{7}$$

where $\mathbb{1}(a) = 1$ if $a > 0$, and zero otherwise, and $w(d_s) = d_s/d$, with d_s the number of delegates allocated to state s, and $d = 538$ is the total number of delegates. Clearly $\sum_{s=1}^{51} w(d_s) = 1$. Hence the aggregate (popular) and delegate outcomes need not coincide. Note that $V_t > 0.5$ can also be written equivalently as

$$\sum_{s=1}^{51} w_{st} V_{st} > 0.5. \tag{8}$$

Clearly, (8) does not necessarily imply (7). The key assumption here is that all electoral votes go towards the party that wins the state's popular vote. Looking at recent history, this holds generally as many states have implicit commitments to allocate electoral votes to the candidate who wins the state by the popular vote. In 2016, all but seven electors followed this rule.[15]

8.1 Forecasting turnout and election outcomes

From the previous section it is clear that we require state level Republican (Democratic) vote shares to predict the overall outcome of the election. To this end we first note that $VT_{cr,t+4} = (R_{cr,t+4} + D_{cr,t+4})/VAP_{cr,t+4}$, where $VAP_{cr,t+4}$ is the eligible voting population in county c of region r in the election year $t+4$. Also recall that $LRO_{cr,t+4} = DLRO_{cr,t+4} + LRO_{cr,t}$, and $\ln(R_{cr,t+4}/D_{cr,t+4}) = LRO_{cr,t+4}$. Suppose that we have forecasts for $VT_{cr,t+4}$ and $LRO_{cr,t+4}$. Then using these identities we have

$$R_{cr,t+4} = \frac{VAP_{cr,t+4}VT_{cr,t+4}}{1 + exp(-LRO_{cr,t+4})} = VAP_{cr,t+4}VT_{cr,t+4}\left(\frac{exp(LRO_{cr,t+4})}{1 + exp(LRO_{cr,t+4})}\right). \tag{9}$$

[15] In Maine, the popular vote was won by the Democratic candidate. Three of the four electoral votes were given to the Democratic candidate, while one electoral vote was cast for the Republican candidate. In Washington State, four out of eight electoral votes were cast in favor of candidates other than the popular vote winner (which was the Democratic candidate). In Texas, despite the popular vote favoring Republicans, two electoral votes were cast for non-Republican candidates.

Similarly

$$D_{cr,t+4} = VAP_{cr,t+4} VT_{cr,t+4} \left(\frac{1}{1 + exp\left(LRO_{cr,t+4}\right)} \right). \tag{10}$$

These county-specific votes can now be aggregated to the state level. Let $\mathcal{C}_s$ denote the set of all counties in state s. Then state popular votes are computed as

$$R_{s,t+4} = \sum_{cr \in \mathcal{C}_s} R_{cr,t+4}, \text{ and } D_{s,t+4} = \sum_{cr \in \mathcal{C}_s} D_{cr,t+4}, \tag{11}$$

with $R_{cr,t+4}$ and $D_{cr,t+4}$ given by (9) and (10), respectively. Hence the Republican vote share for state s is given by

$$V_{s,t+4} = \frac{\sum_{cr \in \mathcal{C}_s} R_{cr,t+4}}{\sum_{cr \in \mathcal{C}_s} \left(R_{cr,t+4} + D_{cr,t+4}\right)} = \frac{\sum_{cr \in \mathcal{C}_s} VAP_{cr,t+4} VT_{cr,t+4} \left(\frac{exp(LRO_{cr,t+4})}{1+exp(LRO_{cr,t+4})} \right)}{\sum_{cr \in \mathcal{C}_s} VAP_{cr,t+4} VT_{cr,t+4}}. \tag{12}$$

With state-level Republican vote shares in hand, state-level popular vote outcomes, Electoral College vote outcomes, and national popular vote outcomes can be predicted.

9 2016 Presidential Election: Prediction and evaluation

We first generate *ex ante* forecasts of the 2016 Presidential Election using the active sets tabulated above, and the Lasso and OCMT selection algorithms. Using data from 2000 through 2012 only, we recursively estimate the panel regressions (4) and (2) subject to the identifying restrictions, $\delta_r = 0$ and after variable selection. These selected regressions are then used to generate out-of-sample 2016 election forecasts at the county level. We consider both a national pooled model and a model which allows for heterogeneity across BEA regions. We refer to these as pooled and regional model/forecasts, respectively. Importantly, we only model the 48 U.S. mainland states plus the District of Columbia. We do not model Hawaii or Alaska. There are multiple reasons for this. The first reason is because the two states are not in close geographical proximity to other states, hence they are likely to be comprised of relatively unique characteristics such that a regional model would be inadequate. Moreover, the two states cannot be modeled individually because of the relatively small number of counties within each state. Hawaii has five counties and Alaska has 19 boroughs. That Alaska is composed of boroughs rather than counties further complicates modeling county-level voting outcomes for the state. Fortunately, both Alaska and Hawaii are non-swing states, historically voting Republican and Democrat, respectively. Therefore, in our electoral

and national predictions we assume Alaska votes Republican and Hawaii votes Democrat.

Comparing predicted state and national popular vote and electoral votes with actual outcomes is a natural way to evaluate the forecasting performance of our models. Alternatively, we also provide evaluations of state and overall predictions based on traditional statistical measures. We compute state-specific and national level root mean squared forecast errors (RMSFE). State-specific RMSFE are defined by

$$RMSFE_s = \sqrt{\sum_{cr \in \mathcal{C}_s} w_{cs,t} \left(DLRO_{cr,t+4} - \widehat{DLRO}_{cr,t+4} \right)^2}, \tag{13}$$

where $w_{cs,t} = \left(R_{cs,t+4} + D_{cs,t+4} \right) / \left(R_{s,t+4} + D_{s,t+4} \right)$, with $R_{s,t+4}$ and $D_{s,t+4}$ computed as in (11). The national RMSFE measure is given by

$$RMSFE = \sqrt{\sum_{s=1}^{49} w_{s,t} RMSFE_s^2}, \tag{14}$$

where $w_{s,t} = \left(R_{s,t+4} + D_{s,t+4} \right) / \left(R_{t+4} + D_{t+4} \right)$, with $R_{t+4} = \sum_{s=1}^{49} R_{s,t+4}$, and $D_{t+4} = \sum_{s=1}^{49} D_{s,t+4}$.

Our out-of-sample forecast and corresponding evaluations correspond to the 2016 election.

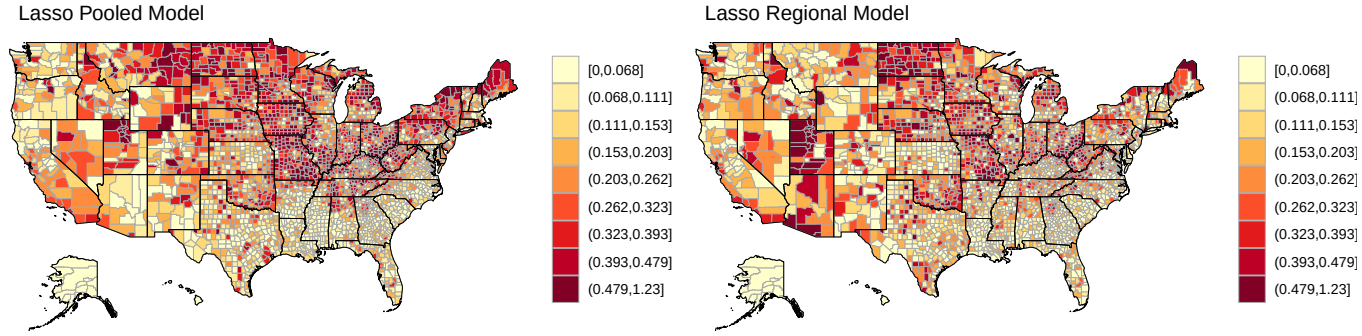
9.1 Pooled and regional forecasts

To produce 2016 out-of-sample forecasts, we use data from 2000 up to but preceding the November election of 2016. The contenders were Democratic candidate Hillary Clinton and Republican candidate Donald Trump. Forecast results are provided for: state-level popular votes, electoral votes, and the overall national popular votes. Tables with electoral outcomes for a subset of notable swing states are also included.

State level forecast results for 2016 are reported in Tables 7 and 8. These include state election outcomes and forecasts for the Republican vote share, V_s $s = 1, 2, ..., 49$, along with the forecasts of Electoral College votes for the Republican candidate. Both tables report the pooled and regional forecasts, with Table 7 giving the results using Lasso and Table 8 giving the results for OCMT.

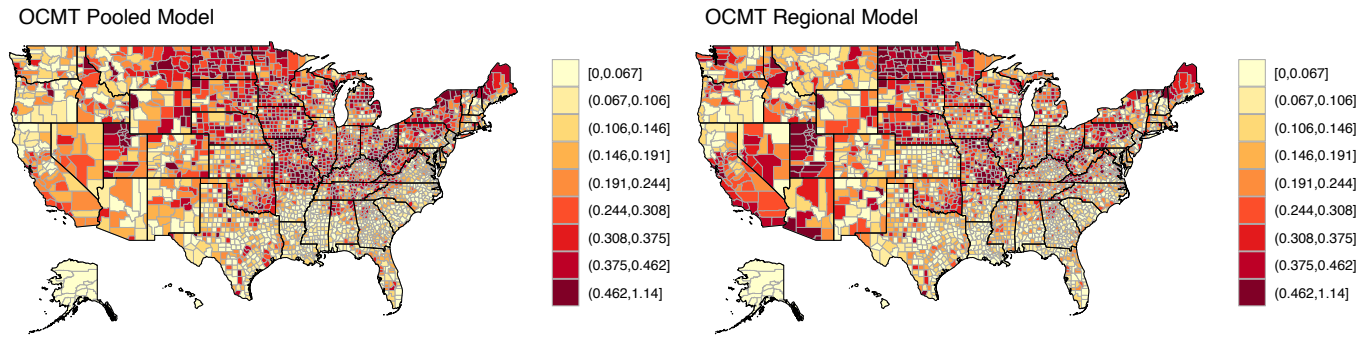
It is clear that, irrespective of which algorithm is used for variable selection, the primary difference between the forecasts is whether we allow for regional heterogeneity or not. Pooled forecasts predict a Democratic victory whilst the regional forecasts correctly predict a Republican victory. For example, the pooled model using Lasso algorithm predicts Republican winning 253 electoral college votes, whilst if we allow for regional heterogeneity the number of electoral votes won by the Republican candidate is predicted to be 308. Based on

Figure 1: Absolute Prediction Errors for changes in 2016 Log Republican Odds ($DLRO_{cr,2016}$) across Counties using the Lasso Estimation Algorithm



Absolute prediction errors for changes in log Republican odds by county, computed as $|DLRO_{cr,2016} - \widehat{DLRO}_{cr,2016}|$ (See Equation 13).

Figure 2: Absolute Prediction Errors for changes in 2016 Log Republican Odds ($DLRO_{cr,2016}$) across Counties using the OCMT Estimation Algorithm



Absolute prediction errors for changes in log Republican odds by county, computed as $|DLRO_{cr,2016} - \widehat{DLRO}_{cr,2016}|$ (See Equation 13).

the realized vote shares, Trump would have won 305 electoral college votes - although as it turned out he received 304 electoral votes since some electors did not follow the state level popular vote outcomes.[16] A very similar conclusion emerges if we use OCMT algorithm. Pooled OCMT would have predicted 265 electoral votes for Trump, as compared to 307 electoral votes under if we allow for regional heterogeneity. These results clearly highlight the importance of heterogeneity and could explain the failure of many professional forecasters to correctly predict the outcome of the 2016 election.

Statistical forecast comparisons based on county-level forecasts provide a similar picture. Figures 1 and 2 present the spatial distribution of absolute prediction errors across mainland U.S. counties for the change in the Republican log-odds ratio, namely $|DLRO_{cr,2016} - \widehat{DLRO}_{cr,2016}|$. Clearly, some counties, regions and states were more difficult to forecast than others. The Midwest exhibits particularly high prediction errors as seen by its generally darker shade. However, the reduction in forecast errors is noticeable when comparing the pooled forecasts against the regional forecasts. On average across counties, absolute prediction errors are about 10 percent lower under the regional model for both Lasso and OCMT. It is worth noting, however, that some county predictions fare better under the pooled model, specifically those located in the southwestern part of the U.S.

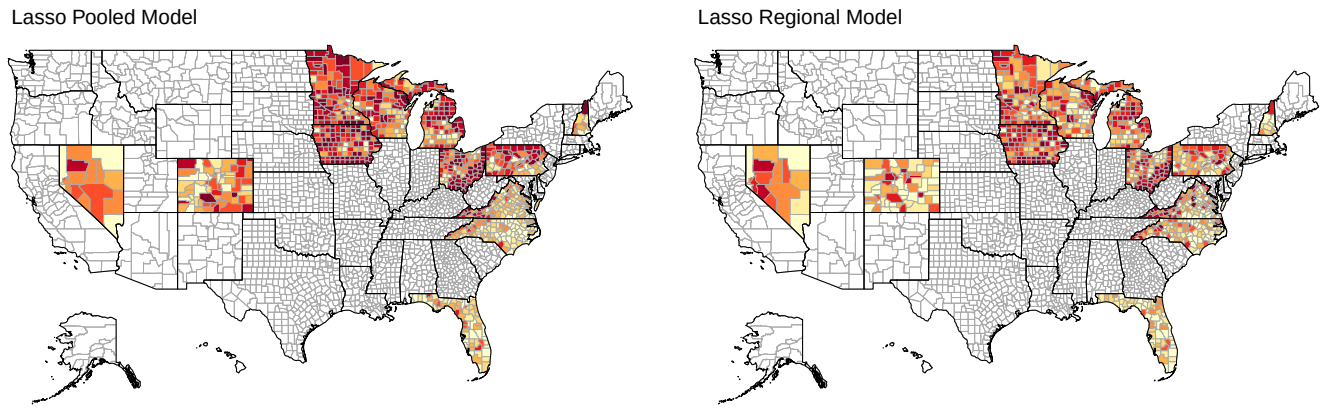
9.2 Swing state forecasts

U.S. presidential elections usually come down to the results from key swing states. Therefore a model that predicts the swing states well is likely to go a long way in correctly forecasting the election. We consider the following 12 states as key swing states: Colorado, Florida, Iowa, Michigan, Minnesota, Nevada, New Hampshire, North Carolina, Ohio, Pennsylvania, Virginia, and Wisconsin. Figures 3 and 4 focus on the county-level prediction errors for these swing states. Both Lasso and OCMT regional models improve upon Lasso and OCMT pooled predictions across swing states broadly noted by the visually apparent reduction in absolute prediction errors.

The improvement in county-level predictions also have important implications for the national outcomes. Table 4 shows the realized and predicted electoral college votes among the key swing states. The Republican candidate won 114 electoral votes from the swing states in 2016 out of the possible number of 156. Comparing the pooled and regional models, the regional models markedly outperform the pooled models in terms of swing state forecasts. The Lasso-regional and OCMT-regional models predicted the Republican candidate winning 117 and 109 electoral votes in the swing states, respectively. By contrast, the pooled Lasso and

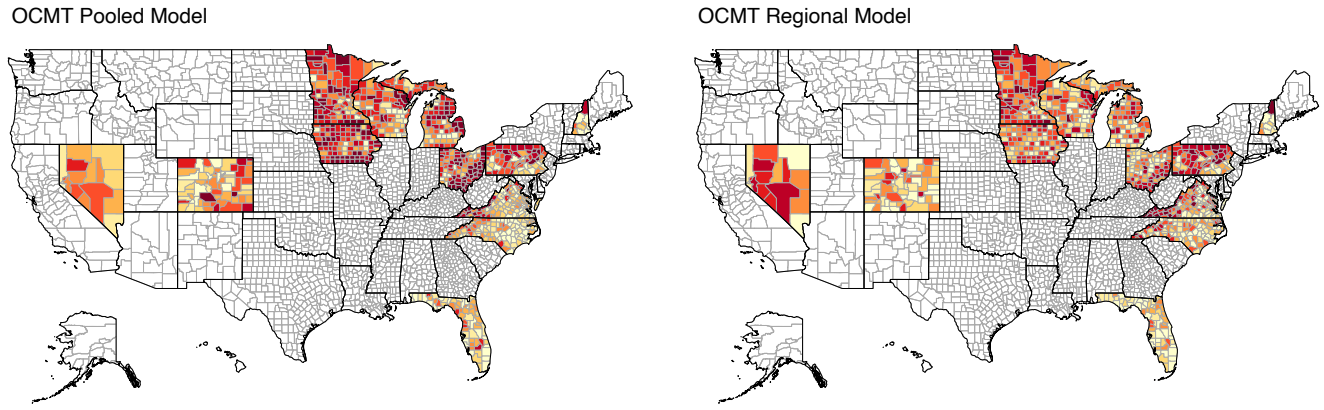
[16] See https://en.wikipedia.org/wiki/2016_United_States_presidential_election

Figure 3: Absolute Prediction Errors for changes in 2016 Log Republican Odds ($DLRO_{cr,2016}$) across Counties in Swing States using the Lasso Estimation Algorithm



Absolute prediction errors for changes in log Republican odds by county, computed as $|DLRO_{cr,2016} - \widehat{DLRO}_{cr,2016}|$ (See Equation 13).

Figure 4: Absolute Prediction Errors for changes in 2016 Log Republican Odds ($DLRO_{cr,2016}$) across Counties in Swing States using the OCMT Estimation Algorithm



Absolute prediction errors for changes in log Republican odds by county, computed as $|DLRO_{cr,2016} - \widehat{DLRO}_{cr,2016}|$ (See Equation 13).

OCMT models predicted 62 and 74 Republican electoral votes, respectively, which resulted the pooled models to forecast an overall presidential victory for the Democratic candidate in 2016.

Table 4: 2016 Swing State Pooled and Regional Republican Electoral College Vote Forecasts

			Pooled Forecasts		Regional Forecasts	
State	d_s	Realized	Lasso	OCMT	Lasso	OCMT
CO	9	0	0	0	9	9
FL	29	29	29	29	29	29
IA	6	6	0	6	6	6
MI	16	16	0	0	0	16
MN	10	0	0	0	10	0
NC	15	15	15	15	15	15
NH	4	0	0	0	0	0
NV	6	0	0	6	0	6
OH	18	18	18	18	18	18
PA	20	20	0	0	20	0
VA	13	0	0	0	0	0
WI	10	10	0	0	10	10
All Swing Votes	156	114	62	74	117	109

Column d_s refers to total number of electoral votes per state (Equation 7). Forecasts are the model implied number of Republican electoral college votes. Regional forecasts are generated using the eight separate panel regressions for the eight BEA regions.

Figure 5 compares swing state predicted Republican vote shares (V_s) obtained using the Lasso algorithm. The Lasso-regional model correctly predicted 9 of the 12 swing states outcomes, namely Florida, Iowa, Nevada, New Hampshire, North Carolina, Ohio, Pennsylvania, Virginia and Wisconsin. The OCMT-regional model also correctly predicted 9 of 12 swing states, namely Florida, Iowa, Michigan, Minnesota, New Hampshire, North Carolina, Ohio, Virginia, Wisconsin (see Figure 6). One swing state mis-predicted by both Lasso and OCMT regional models but correctly predicted by both pooled models was Colorado. Meanwhile the most noticeable improvement from using the regional models over pooled models can be seen with Wisconsin, a Midwest swing state. The state voted Republican in 2016, allocating 10 electoral votes to the Republican candidate. Both the Lasso and OCMT pooled models predicted a Democratic winner in Wisconsin. By contrast, both regional Lasso and OCMT models predicted a Republican win in Wisconsin.

The pooled models also failed to correctly predict Pennsylvania, a major swing state with 20 electoral votes. The Lasso-regional model correctly predicted the Republican win in Pennsylvania. The Republican victory in Michigan was also mis-predicted under both

Figure 5: Swing State Forecasts and Realized Republican Vote Share (V_s) for 2016 using the Lasso Estimation Algorithm

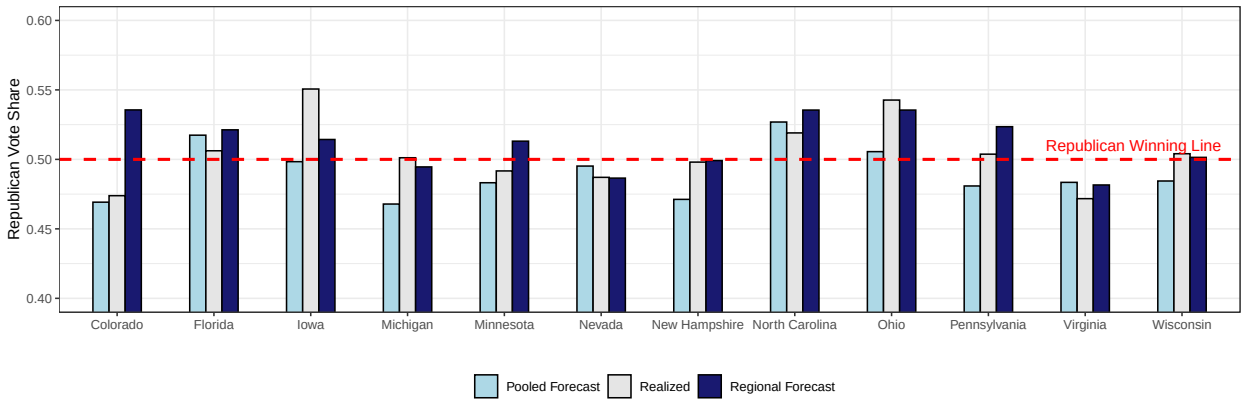
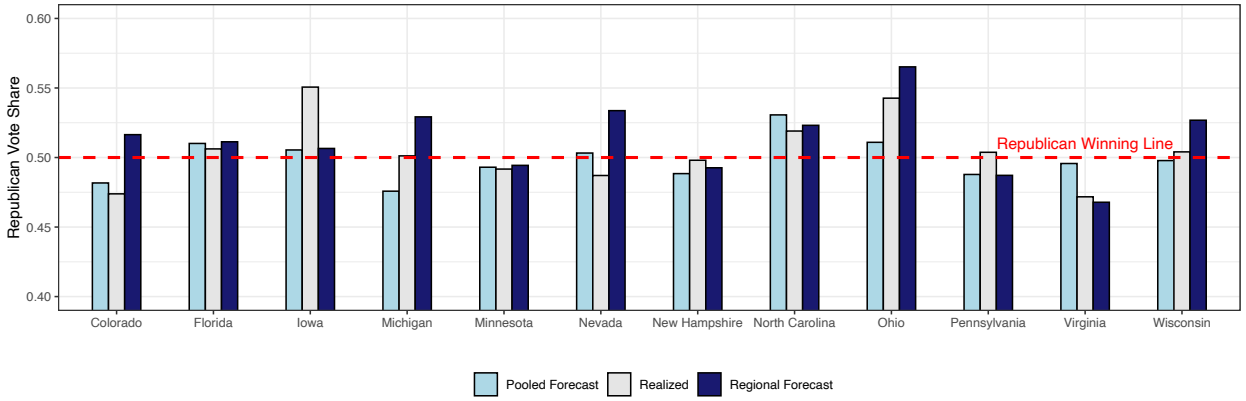


Figure 6: Swing State Forecasts and Realized Republican Vote Share (V_s) for 2016 using the OCMT Estimation Algorithm



pooled model specifications, but correctly predicted by the OCMT-regional model.

9.3 U.S. Mainland Popular Vote Forecasts

The U.S. mainland popular Republican vote share forecasts (V_t) are reported in Table 9. It is interesting that the pooled forecasts do better than the regional forecasts at predicting the aggregate outcomes, irrespective of whether the OCMT or the Lasso algorithm is used. The RMSFE of pooled forecasts using Lasso (OCMT) is 0.078 (0.077) as compared to 0.090 (0.102) for the regional models. Also the pooled models predicted a Republican vote share of 0.494 (0.499) which is closer to the realized value of 0.489, as compared to 0.510 (0.514) predicted using the regional Lasso (OCMT) model. The main advantage of the regional models lies in their ability to deliver better popular forecasts at state level which matters

for correctly predicting presidential election outcomes. Once again the failure of the pooled models to accurately forecast the outcome of U.S. presidential elections points toward the essential heterogeneity that exists across US states and regions which is responsible for the misalignment of the popular and electoral vote outcomes that occurs, albeit rarely. Our results suggest that political polarization is *not* evenly distributed across the U.S. Rather, voter preferences vary systematically across regions. For example, California may have the largest population, and on average voters within the state share similar preferences, reflecting its historical favor towards Democratic candidates. By contrast, in the Midwest, voters share similar preferences, but those preferences may contrast starkly with those of voters in California, and may change more rapidly at the same time. It is not surprising that voter heterogeneity varies across regions given that industry composition, social values, and demographics are also shown to vary across regions. Taken along with the disproportional electoral vote allocation of some states relative to their population, regional heterogeneity can drive deviations between popular and electoral presidential vote outcomes.

To summarize, allowing for parameter heterogeneity across regions considerably improves 2016 out-of-sample forecasts of both state popular and electoral outcomes when compared to pooling approaches. These results are consistent with regional heterogeneity being an important feature of the U.S. electoral landscape. Homogeneity within regions but heterogeneity across regions can arise when people with similar preferences geographically cluster despite the presence of considerable diversity at the national level. Our findings are consistent with that idea, as our regional model's implicit assumption is that parameters vary across U.S. geographical regions, but are constant within regions. While the regional models help forecast the electoral college victory of the Republican party in 2016, the pooled models are better at forecasting the overall popular vote. Political polarization across regions coupled with disproportionate allocation of electoral votes relative to state populations may be one reason for such deviations. For robustness, we report 2016 forecasts under a Lasso and OCMT averaged model in the Online Supplement Section S4. The averaged model takes Lasso and OCMT county-level predictions of Republican and Democratic votes and averages them together before aggregating to state-level results. The 2016 forecasts remain largely unchanged under this averaging approach.

10 Key Determinants of U.S. Presidential Elections Over the Period 2000-2016

In the previous section, to evaluate the 2016 U.S. Presidential Election we used data up to 2012 to estimate the panel regressions (4) and (2) subject to the recursive order restriction, $\beta_r = 0$, and then generated out-of-sample forecasts for 2016. In this section, we present estimates of the same model based on the full 2000-2016 sample and, to further understand the key factors behind regional heterogeneity, we present both pooled and regional estimates. We begin with pooled estimates. The pooled model estimates for voter turnout and the Republican log-odds ratio equations are summarized in Tables 11 and 12, respectively.[17]

Several time-invariant covariates are statistically significant, regardless of whether estimated using OCMT or Lasso algorithms. These include rural-urban score, migration, and the education covariates. Time-varying covariates are also important. Specifically, short-run economic variables exhibit the strongest overall explanatory power relative to their longer term counterparts. This evidence is consistent with voters having 'short memories'. Specifically, changes in the real effective USD exchange rate (a barometer for international competition), unemployment rates, and house prices over the three months preceding the election are significantly associated with voting outcomes, and their inclusion renders 1-year changes in these variables mostly insignificant. While 3-month house price appreciation unambiguously favors the Republican candidate, higher unemployment rates preceding the election somewhat surprisingly favor the incumbent party. By contrast, real export-weighted USD appreciation 3-months preceding the election significantly punishes the incumbent party. In case of the pooled model we also find that being economically 'left behind' is significantly associated with voting against the incumbent party in the upcoming election.

We now consider estimates that allow for regional differences and discuss the differences in selected covariates and their estimates across the eight BEA regions. Tables 13 and 14 summarize the estimates for voter turnout VT under the Lasso and OCMT estimation algorithms, respectively. Similarly, Tables 15 and 16 report estimates for $DLRO$ using the Lasso and OCMT algorithms. As can be seen, the variation in both the selected covariates and the magnitude of the estimates vary substantially across the BEA regions, and suggest pooling might result in mis-leading inference. The estimates also show how heterogeneous U.S. regions can be. Consider Table 15, the Lasso-regional estimates for $DLRO$. The education variable (edu2000) was selected for 8 out of 8 regions, hence this variable was

[17]For the OCMT estimates we provide standard errors clustered at the state-year level. Lasso estimates that are used for forecasting are computed using cross-validation and there are no associated standard errors to report. However, for completeness we provide OLS estimates for the covariates selected by Lasso together with their state-year clustered standard errors.

identified as informative on a national scale. Moreover, coefficient estimates are negative in all regions suggesting that more educated counties tend to favor the Democratic candidate, regardless of the region in which the county is located. However, the estimates of this variable differ quite a bit regionally: a one percentage point increase in the education rate in the Mideast region (Southwest region) is associated with a change in the Republican odds ratio of -0.246 (-0.845) percent.. Short-run house price appreciation (3 months preceding the election, denoted by hpret_M3) is never associated with greater Democrat vote share – across any BEA region (coefficients are either zero or positive across regional panel regressions).

Most covariates from the active set are not selected across every region. Again, this points to the existence of substantial cross-regional differences in the U.S. Larger voter turnout is associated with votes towards Democrats in 5 of the 8 regions ($\widehat{VT}$). By contrast, Zandi et al. [2020] pools information nationally, which implicitly assumes that greater turnout is unambiguously associated with lower Republican vote share. Being economically left behind tends to punish the incumbent party in 5 of the 8 regions (the covariate LBCG_L1× r_incu_pa). Higher local short-run unemployment favors Democrats in 4 of the 8 regions, has no effect on voting in 3 regions, and favors the Republican candidate in the Plains region.

11 Forecasts of the 2020 U.S. Presidential Election

In the first publicly released version of this paper, distributed as Cambridge Economics and CESifo working papers[18], we reported forecasts of the 2020 election using data available through July 2020. In this version we also provide updated forecasts using more more recent data.[19] We thought it is useful to present both sets of forecasts showing how the results respond to data updates.

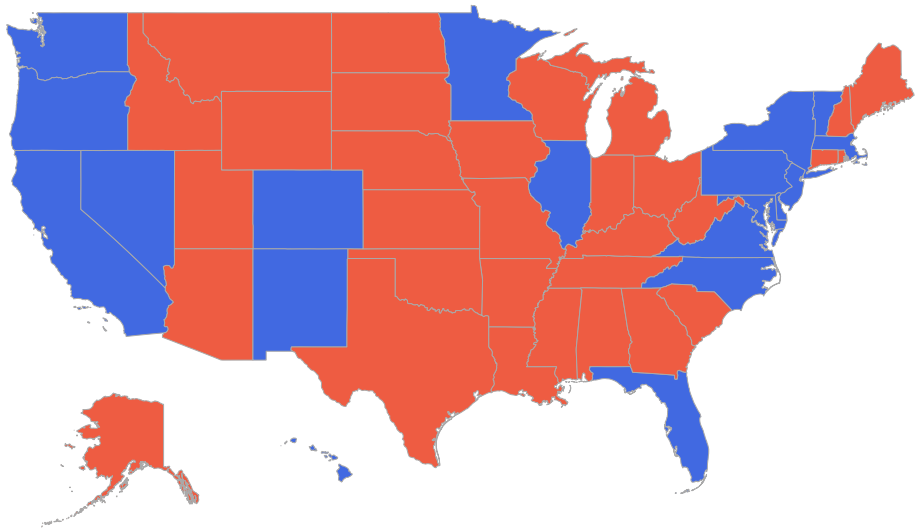
Starting with the forecats based on July 2020 data, we report state-level forecasts of Republican vote share V_s and corresponding electoral college outcomes in Tables S.1 and S.2 for estimates under the Lasso and OCMT algorithms, respectively. U.S. Mainland popular vote predictions are reported in Table 10. Figures 7 and 8 in the Online Supplement chart forecasts of U.S. electoral college outcomes by state.

All pooled models forecast an electoral victory for the Democratic candidate, but we saw in our evaluation of the 2016 election that pooled models ignore crucial regional heterogeneity

[18]See CWPE 2029, http://www.econ.cam.ac.uk/research-files/repec/cam/pdf/cwpe2092.pdf and CESifo WP 8615, https://www.cesifo.org/en/publikationen/2020/working-paper/regional-heterogeneity-and-us-presidential-elections

[19]As of mid July 2020, data on unemployment and house prices were available through June 2020, while real exchange rates were available through May 2020. Data on inflation was available as of 2020Q1 , and the remaining annual frequency data were available through 2018. The missing data were forward-filled with the most recently available values.

Figure 7: 2020 State Electoral College Forecast under the Lasso-Regional Model



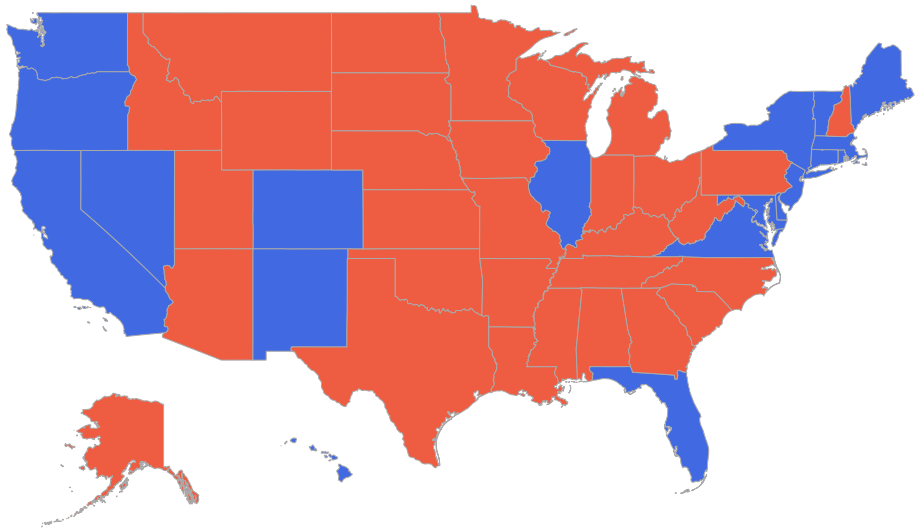
Red indicates Republican electoral victory. Blue indicates Democratic electoral victory.

and could lead to inaccurate forecasts. By contrast, forecasts from the regional models imply a very close electoral college outcome. The Lasso-regional model forecasts a Democratic victory – the Republican candidate is expected to win 260 electoral college votes (recall that 270 is needed to win). Meanwhile, the OCMT-regional model forecasts a Republican candidate victory – the Republican candidate is predicted to win 290 electoral college votes.

Perhaps it is worth noting that the inherent nonlinearity of election outcomes due to the design of the electoral college. Namely, a swing in just one or two state outcomes could swing the entire election. This point re-emphasizes why the aptly named swing states are such crucial political battlegrounds.

Specifically looking towards swing states explains the divergence between the Lasso-regional and OCMT-regional model forecasts (Table 5). The Lasso-regional model forecasts that Republicans take 54 out of 156 electoral votes available across the 12 swing states. Meanwhile, the OCMT-regional model forecasts that the Republican candidate will win 99 electoral votes in the swing states. Both Lasso-regional and OCMT-regional models forecast Iowa, Michigan, New Hampshire, Ohio, and Wisconsin to vote Republican in the electoral college, for a total of 54 Republican electoral votes. In fact, in every swing state that the Lasso-regional model forecasts a Republican victory the OCMT-regional model does also.

Figure 8: 2020 State Electoral College Forecasts under the OCMT-Regional Model



Red indicates Republican electoral victory. Blue indicates Democratic electoral victory.

In addition to those swing states, the OCMT-regional model also forecasts a Republican electoral win in Minnesota, North Carolina, and Pennsylvania. Moreover, the electoral college maps show how winning many states does not imply victory in terms of electoral college votes. This becomes even more apparent at the county level as shown in Figure S.4 and Figure S.5 in the Online Supplement. In 2016, a majority of counties voted for the Republican candidate, yet the Democratic candidate won the popular election. Both models forecast that most counties will vote for the Republican candidate in 2020, yet all four models (Lasso-pooled, OCMT-pooled, Lasso-regional and OCMT-regional) predict the Democratic candidate will win the popular vote. See Table 10.

For robustness, we also report 2020 predictions under a Lasso and OCMT averaged model in Section S4 of the Online Supplement. The average model takes Lasso and OCMT county-level predictions of Republican and Democratic votes and averages them together before aggregating to state-level results. For 2020, the averaged model predicts 269 Republican electoral votes – one shy from winning the presidential election. For comparison, as of September 2020 the Zandi et al. [2020] model is forecasting a Republican victory with the candidate winning between 298 to 351 electoral votes.

Table 5: 2020 Swing State Pooled and Regional Republican Electoral College Vote Forecasts

		Pooled Forecasts		Regional Forecasts	
State	d_s	Lasso	OCMT	Lasso	OCMT
CO	9	0	0	0	0
FL	29	0	0	0	0
IA	6	6	6	6	6
MI	16	0	0	16	16
MN	10	0	0	0	10
NC	15	0	0	0	15
NH	4	0	0	4	4
NV	6	0	6	0	0
OH	18	18	18	18	18
PA	20	0	0	0	20
VA	13	0	0	0	0
WI	10	0	0	10	10
All Swing Votes	156	24	30	54	99

Column d_s refers to total number of electoral votes per state (Equation 7). Forecasts are the model implied number of Republican electoral college votes. Regional forecasts are generated using the eight separate panel regressions for the eight BEA regions.

11.1 Updated forecasts using data available as of October 14, 2020

The updated forecasts are based on the same model specifications which are not affected by data updates. What is changed are the observations on some of the fast moving covariates included in the forecasting models as predictors. Amongst these covariates the most important are the monthly data on the unemployment rate, the rate of change of house prices, changes in the real exchange rate, and the rate of inflation. For these we update the series to June/July of 2020. We have also updated some of the data with annual frequency to 2019 on state-level healthcare expenditures, government employment, and expenditure on rents.

Tables S.5, S.6, and S.7 of the Online Supplement report updated state-by-state forecasts using Lasso, OCMT, and Lasso-OCMT averaging algorithms. Both Lasso-regional and OCMT-regional forecasts move marginally against the Republican candidate. Lasso-regional is now forecasting 249 Republican electoral votes (previously 260). The forecasts based on the OCMT-regional model are now very much borderline, forecasting 270 (previously 290) electoral votes for the republican candidate. The Lasso-OCMT average regional model is now forecasting 265 republican electoral votes (compared to 269 previously). Notable changes are for the OCMT-regional model, which previously predicted Michigan as a win for the Republican candidate (16 electoral votes), now predicts a Democratic state victory. The same goes for New Hampshire (4 electoral votes now in favor of the Democratic candidate,

Table 6: 2020 Swing State Pooled and Regional Republican Electoral College Vote Forecasts using Data Available as of October 2020

		Pooled Forecasts		Regional Forecasts	
State	d_s	Lasso	OCMT	Lasso	OCMT
CO	9	0	0	0	0
FL	29	0	0	0	0
IA	6	6	6	6	6
MI	16	0	0	16	0
MN	10	0	0	0	10
NC	15	0	15	0	15
NH	4	0	0	4	0
NV	6	0	6	0	0
OH	18	18	18	18	18
PA	20	0	0	0	20
VA	13	0	0	0	0
WI	10	0	0	10	10
All Swing Votes	156	24	45	54	79

Column d_s refers to total number of electoral votes per state (Equation 7). Forecasts are the model implied number of Republican electoral college votes. Regional forecasts are generated using the eight separate panel regressions for the eight BEA regions. Using data available as of October 14, 2020.

previously favoring the Republican candidate). Forecast updates for these two swing states explain the 20 electoral vote shift away from the Republican candidate in the OCMT model. The updated forecasts across swing states are reported in Table 6.

12 Concluding Remarks

An increasingly divided political landscape means that regional heterogeneity is crucial for understanding recent voting behavior and presidential election outcomes. We develop a joint model of voter turnout and voting outcomes and exploit county-level variation and regional heterogeneity to identify factors which explain county-level voting outcomes of the 2016 U.S. Presidential Election. While many forecasts failed to predict the outcome of 2016, our out-of-sample forecasts that allow for regional heterogeneity would have correctly predicted the unexpected Republican victory.

It is worth noting that many of the forecasts that were predicting a Democratic 2016 victory were based on polls. However, in a world of increased political division, the biases inherent in poll-based forecasts may become magnified, requiring highly stratified sampling techniques that are very expensive to implement to ensure such poll-based forecasts are

sufficiently reliable. In contrast, we show that the statistical approach using fundamental socioeconomic and demographic data can take us far in understanding presidential election cycle dynamics. We point out that regional heterogeneity is particularly important for modeling swing states. Variable selection techniques, such as Lasso and OCMT, further improve model performance

Incorporating regional heterogeneity reveals that the extent to which several socioeconomic determinants help explain voter turnout and election outcomes which vary substantially across regions. Significant indicators which help explain voting behavior at the county level include: which party is the incumbent, a county's relative economic performance, local short-run unemployment rate, house price changes, education, poverty rate, among others. Some determinants exhibit consistently robust associations with turnout or voting across regions. For example, house price appreciation generally favors the Republican candidate while counties with higher rates of poverty and educational attainment help the Democratic candidate. The influence of most other variables on turnout and voting outcomes, however, is far from uniform, substantially varying across regions.

We also use the selected models to generate forecasts of 2020 U.S. Presidential Election. We report two sets of forecasts, both based on the same selected models, but with different data updates. For data available through July 2020, the regional models, which predicted a Republican victory in 2016, predict a close electoral college outcome for 2020. The predictions are split: the Lasso-regional model forecasts a Democratic electoral victory (260 electoral votes for the Republican candidate) while the OCMT-regional model forecasts a Republican victory (290 electoral votes for the Republican candidate). However, once we use more recent data, available as of mid October 2020, both Lasso and OCMT regional forecasts shift in favor of the Democratic candidate, with updated forecasts predicting 249, 270, and 265 electoral votes for the Republican candidate using Lasso, OCMT, and Lasso-OCMT average regional model forecasts, respectively. All models point towards the Democratic candidate winning the popular vote. We emphasize, however, that the non-linear nature of the U.S. voting process makes these forecasts fragile and subject to a high degree of uncertainty. In addition, unforeseeable events which cannot be modeled using historical data (e.g. nation-wide protests, pandemics) which have been prevalent in 2020 cast additional uncertainty over our forecasts.

References

Arcelus, F. and A. H. Meltzer (1975). The effect of aggregate economic variables on congressional elections. *The American Political Science Review 69*(4), 1232–1239.

Autor, D., D. Dorn, G. Hanson, and K. Majlesi (2016). Importing political polarization? the electoral consequences of rising trade exposure. Technical report, National Bureau of Economic Research.

Biemer, P. P. (2010). Total survey error: Design, implementation, and evaluation. *Public Opinion Quarterly 74*(5), 817–848.

Biesiada, M. J. (2018). Factors that impact direct democracy and voter turnout: Evidence from a national study on american counties.

Blais, A. (2006). What affects voter turnout? *Annu. Rev. Polit. Sci. 9*, 111–125.

Cancela, J. and B. Geys (2016). Explaining voter turnout: A meta-analysis of national and subnational elections. *Electoral Studies 42*, 264 – 275.

Chudik, A., G. Kapetanios, and M. H. Pesaran (2018). A one covariate at a time, multiple testing approach to variable selection in high-dimensional linear regression models. *Econometrica 86*(4), 1479–1512.

Fair, R. C. (1978). The effect of economic events on votes for president. *The Review of Economics and Statistics 60*(2), 159–173.

Fair, R. C. (1996, Summer). Econometrics and Presidential Elections. *Journal of Economic Perspectives 10*(3), 89–102.

Fowler, J. H., L. A. Baker, and C. T. Dawes (2008). Genetic variation in political participation. *American Political Science Review 102*(2), 233–248.

Fowler, J. H. and C. T. Dawes (2008). Two genes predict voter turnout. *The Journal of Politics 70*(3), 579–594.

Gelman, A. and J. Azari (2017). 19 things we learned from the 2016 election. *Statistics and Public Policy 4*(1), 1–10.

Graefe, A. (2018). Predicting elections: Experts, polls, and fundamentals. *Judgment and Decision Making 13*(4), 334.

Hummel, P. and D. Rothschild (2014). Fundamental models for forecasting elections at the state level. *Electoral Studies 35*, 123–139.

Jackman, R. W. (1987). Political institutions and voter turnout in the industrial democracies. *American Political Science Review 81*(2), 405–423.

Jensen, J. B., D. P. Quinn, and S. Weymouth (2017). Winners and losers in international trade: The effects on us presidential voting. *International Organization 71*(3), 423–457.

Kahane, L. H. (2009, Jun). It's the economy, and then some: modeling the presidential vote with state panel data. *Public Choice 139*(3), 343–356.

Kahane, L. H. (2020). Determinants of county-level voting patterns in the 2012 and 2016 presidential elections. *Applied Economics 0*(0), 1–14.

Keeter, S., R. Igielnik, and R. Weisel (2016). Can likely voter models be improved? evidence from the 2014 us house elections. *Pew Research Center, http://www. pewresearch. org/2016/01/07/can-likely-voter-models-be-improved*.

Kou, S. G. and M. E. Sobel (2004). Forecasting the vote: A theoretical comparison of election markets and public opinion polls. *Political Analysis*, 277–295.

Kramer, G. H. (1971). Short-term fluctuations in us voting behavior, 1896-1964. *The American Political Science Review 65*(1), 131–143.

Leigh, A. and J. Wolfers (2006). Competing approaches to forecasting elections: Economic models, opinion polling and prediction markets. *Economic Record 82*(258), 325–340.

Matsusaka, J. G. (1995). Explaining voter turnout patterns: An information theory. *Public choice 84*(1-2), 91–117.

Matsusaka, J. G. and F. Palda (1999). Voter turnout: How much can we explain? *Public Choice 98*(3/4), 431–446.

Powell, G. B. (1986). American voter turnout in comparative perspective. *American Political Science Review 80*(1), 17–43.

Rogers, T. and M. Aida (2014). Vote self-prediction hardly predicts who will vote, and is (misleadingly) unbiased. *American Politics Research 42*(3), 503–528.

Rusch, T., I. Lee, K. Hornik, W. Jank, A. Zeileis, et al. (2013). Influencing elections with statistics: Targeting voters with logistic regression trees. *The Annals of Applied Statistics 7*(3), 1612–1639.

Scala, D. J. and K. M. Johnson (2017). Political polarization along the rural-urban continuum? the geography of the presidential vote, 2000–2016. *The Annals of the American Academy of Political and Social Science 672*(1), 162–184.

Shirani-Mehr, H., D. Rothschild, S. Goel, and A. Gelman (2018). Disentangling bias and variance in election polls. *Journal of the American Statistical Association 113*(522), 607–614.

Sides, J., M. Tesler, and L. Vavreck (2017). The 2016 us election: How trump lost and won. *Journal of Democracy 28*(2), 34–44.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological) 58*(1), 267–288.

Wang, W., D. Rothschild, S. Goel, and A. Gelman (2015). Forecasting elections with non-representative polls. *International Journal of Forecasting 31*(3), 980–991.

Wold, H. O. (1960). A generalization of causal chain models (part iii of a triptych on causal chain systems). *Econometrica: Journal of the Econometric Society*, 443–463.

Wolfinger, R. E. and S. J. Rosenstone (1980). *Who votes?* Yale University Press.

Zandi, M., D. White, and B. Yaros (2020). 2020 Presidential Election Model. *Moody's Analytics*.

Table 7: State Level Forecasts and Realized Republican Vote Shares (V_s) and Electoral Votes using Lasso Algorithm for 2016 Elections

			Pooled Forecasts			Regional Forecasts		
State	d_s	Realized	$\widehat{V}_s$	RMSFE	EC Votes	$\widehat{V}_s$	RMSFE	EC Votes
AK	3	-	-	-	3	-	-	3
AL	9	0.64	0.63	0.16	9	0.65	0.15	9
AR	6	0.64	0.65	0.20	6	0.68	0.21	6
AZ	11	0.52	0.56	0.28	11	0.54	0.43	11
CA	55	0.34	0.39	0.27	0	0.40	0.29	0
CO	9	0.47	0.47	0.14	0	0.54	0.32	9
CT	7	0.43	0.40	0.19	0	0.44	0.22	0
DC	3	0.04	0.08	0.25	0	0.08	0.12	0
DE	3	0.44	0.39	0.33	0	0.43	0.34	0
FL	29	0.51	0.52	0.12	29	0.52	0.12	29
GA	16	0.53	0.56	0.23	16	0.57	0.26	16
HI	4	-	-	-	0	-	-	0
IA	6	0.55	0.50	0.38	0	0.51	0.38	6
ID	4	0.68	0.69	0.28	4	0.73	0.23	4
IL	20	0.41	0.43	0.41	0	0.45	0.42	0
IN	11	0.60	0.58	0.41	11	0.59	0.39	11
KS	6	0.61	0.62	0.21	6	0.66	0.29	6
KY	8	0.66	0.64	0.25	8	0.67	0.20	8
LA	8	0.60	0.60	0.11	8	0.62	0.13	8
MA	11	0.35	0.36	0.37	0	0.43	0.26	0
MD	10	0.36	0.36	0.19	0	0.40	0.20	0
ME	4	0.49	0.43	0.14	0	0.43	0.18	0
MI	16	0.50	0.47	0.22	0	0.49	0.19	0
MN	10	0.49	0.48	0.26	0	0.51	0.29	10
MO	10	0.62	0.58	0.21	10	0.62	0.19	10
MS	6	0.59	0.58	0.47	6	0.59	0.46	6
MT	3	0.61	0.58	0.22	3	0.63	0.18	3
NC	15	0.52	0.53	0.30	15	0.54	0.37	15
ND	3	0.70	0.64	0.17	3	0.63	0.17	3
NE	5	0.64	0.64	0.30	5	0.65	0.21	5
NH	4	0.50	0.47	0.13	0	0.50	0.17	0
NJ	14	0.43	0.40	0.17	0	0.45	0.18	0
NM	5	0.45	0.45	0.11	0	0.44	0.21	0
NV	6	0.49	0.50	0.11	0	0.49	0.14	0
NY	29	0.38	0.34	0.45	0	0.37	0.53	0
OH	18	0.54	0.51	0.28	18	0.54	0.24	18
OK	7	0.69	0.69	0.21	7	0.68	0.21	7
OR	7	0.44	0.46	0.20	0	0.46	0.21	0
PA	20	0.50	0.48	0.22	0	0.52	0.22	20
RI	4	0.42	0.35	0.31	0	0.39	0.13	0
SC	9	0.57	0.58	0.10	9	0.59	0.12	9
SD	3	0.66	0.62	0.23	3	0.64	0.20	3
TN	11	0.64	0.62	0.22	11	0.66	0.24	11
TX	38	0.55	0.60	0.26	38	0.56	0.14	38
UT	6	0.62	0.75	0.68	6	0.80	0.99	6
VA	13	0.47	0.48	0.23	0	0.48	0.35	0
VT	3	0.35	0.32	0.16	0	0.29	0.28	0
WA	12	0.41	0.44	0.21	0	0.45	0.25	0
WI	10	0.50	0.48	0.36	0	0.50	0.36	10
WV	5	0.72	0.67	0.28	5	0.68	0.22	5
WY	3	0.76	0.73	0.25	3	0.76	0.20	3
All Votes	538				253			308

Republican vote shares are calculated as in Equation 12. Column d_s refers to the total number of electoral votes per state (Equation 7). Root mean square forecast error (RMSFE) is calculated as in Equation 13. EC Votes refer to the predicted number of Republican electoral college votes. All Votes accumulates U.S. Mainland electoral college votes, and assumes Hawaii casts her electoral votes for the Democratic candidate and Alaska casts her electoral votes for the Republican candidate. Regional forecasts are generated using the eight separate panel regressions for the eight BEA regions.

Table 8: State Level Forecasts and Realized Republican Vote Shares (V_s) and Electoral Votes using OCMT Algorithm for 2016 Elections

			Pooled Forecasts			Regional Forecasts		
State	d_s	Realized	$\widehat{V}_s$	RMSFE	EC Votes	$\widehat{V}_s$	RMSFE	EC Votes
AK	3	-	-	-	3	-	-	3
AL	9	0.64	0.64	0.17	9	0.64	0.14	9
AR	6	0.64	0.65	0.20	6	0.66	0.22	6
AZ	11	0.52	0.56	0.27	11	0.54	0.44	11
CA	55	0.34	0.39	0.26	0	0.42	0.41	0
CO	9	0.47	0.48	0.15	0	0.52	0.25	9
CT	7	0.43	0.41	0.18	0	0.43	0.19	0
DC	3	0.04	0.07	0.18	0	0.08	0.18	0
DE	3	0.44	0.41	0.29	0	0.41	0.32	0
FL	29	0.51	0.51	0.11	29	0.51	0.11	29
GA	16	0.53	0.56	0.26	16	0.57	0.26	16
HI	4	-	-	-	0	-	-	0
IA	6	0.55	0.51	0.37	6	0.51	0.40	6
ID	4	0.68	0.70	0.28	4	0.71	0.25	4
IL	20	0.41	0.44	0.39	0	0.48	0.30	0
IN	11	0.60	0.58	0.39	11	0.62	0.30	11
KS	6	0.61	0.63	0.22	6	0.67	0.30	6
KY	8	0.66	0.65	0.24	8	0.66	0.18	8
LA	8	0.60	0.61	0.11	8	0.61	0.11	8
MA	11	0.35	0.38	0.32	0	0.42	0.29	0
MD	10	0.36	0.38	0.20	0	0.37	0.22	0
ME	4	0.49	0.44	0.15	0	0.44	0.20	0
MI	16	0.50	0.48	0.21	0	0.53	0.24	16
MN	10	0.49	0.49	0.27	0	0.49	0.27	0
MO	10	0.62	0.59	0.20	10	0.62	0.20	10
MS	6	0.59	0.58	0.46	6	0.58	0.47	6
MT	3	0.61	0.59	0.18	3	0.63	0.17	3
NC	15	0.52	0.53	0.28	15	0.52	0.42	15
ND	3	0.70	0.63	0.15	3	0.62	0.20	3
NE	5	0.64	0.65	0.29	5	0.66	0.21	5
NH	4	0.50	0.49	0.13	0	0.49	0.16	0
NJ	14	0.43	0.41	0.17	0	0.41	0.23	0
NM	5	0.45	0.45	0.12	0	0.44	0.20	0
NV	6	0.49	0.50	0.11	6	0.53	0.08	6
NY	29	0.38	0.35	0.48	0	0.35	0.62	0
OH	18	0.54	0.51	0.27	18	0.57	0.26	18
OK	7	0.69	0.69	0.21	7	0.68	0.19	7
OR	7	0.44	0.46	0.21	0	0.51	0.36	7
PA	20	0.50	0.49	0.21	0	0.49	0.22	0
RI	4	0.42	0.36	0.27	0	0.39	0.15	0
SC	9	0.57	0.58	0.11	9	0.58	0.11	9
SD	3	0.66	0.63	0.21	3	0.64	0.20	3
TN	11	0.64	0.63	0.23	11	0.65	0.22	11
TX	38	0.55	0.60	0.28	38	0.57	0.15	38
UT	6	0.62	0.76	0.76	6	0.78	0.89	6
VA	13	0.47	0.50	0.24	0	0.47	0.36	0
VT	3	0.35	0.33	0.17	0	0.32	0.20	0
WA	12	0.41	0.44	0.21	0	0.50	0.41	0
WI	10	0.50	0.50	0.37	0	0.53	0.29	10
WV	5	0.72	0.66	0.24	5	0.68	0.23	5
WY	3	0.76	0.74	0.24	3	0.75	0.22	3
All Votes	538				265			307

Republican vote shares are calculated as in Equation 12. Column d_s refers to the total number of electoral votes per state (Equation 7). Root mean square forecast error (RMSFE) is calculated as in Equation 13. EC Votes refer to the predicted number of Republican electoral college votes. All Votes accumulates U.S. Mainland electoral college votes, and assumes Hawaii casts her electoral votes for the Democratic candidate and Alaska casts her electoral votes for the Republican candidate. Regional forecasts are generated using the eight separate panel regressions for the eight BEA regions.

Table 9: 2016 U.S. Mainland Republican Vote Share Forecasts

Model	Realized	Pooled	Pooled $RMSFE$	Regional	Regional $RMSFE$
OCMT	0.489	0.499	0.077	0.514	0.102
Lasso	0.489	0.494	0.078	0.510	0.090

To produce popular vote share forecasts, Equation 12 is applied to the sum of predicted Republican and Democrat votes across U.S. mainland states plus Washington D.C. RMSFE calculations based on Equation 14. Regional forecasts are generated using the eight separate panel regressions for the eight BEA regions.

Table 10: 2020 US. Mainland Republican Vote Share Forecasts

Model	Pooled	Regional
OCMT	0.471	0.498
Lasso	0.445	0.477

To produce popular vote share forecasts, Equation 12 is applied to the sum of predicted Republican and Democrat votes across U.S. mainland states plus Washington D.C. Regional forecasts are generated using the eight separate panel regressions for the eight BEA regions.

Table 11: Pooled Panel Regression with Variable Selection for Voter Turnout (VT) as the Dependent Variable Estimated over the 2000-2016 Election Cycles

	Covariate	OCMT	SE-OCMT	Lasso	Lasso(OLS)	SE-Lasso(OLS)
1	(Intercept)	0.0886	(0.1239)	0.1432	0.1412***	(0.0192))
2	r_incu_pa	-0.2179**	(0.1085)			
3	r_incu_pr	0.0314***	(0.0053)	0.0289	0.0337***	(0.0051)
4	VT_L1	0.7977***	(0.0172)	0.7789	0.7973***	(0.0169)
5	VT_L1 x r_incu_pa.	-0.0461***	(0.0166)			
6	hlt_L1			-0.1294	-0.2453	(0.1844)
7	gov_L1	0.1557	(0.187)	0.1449	0.1915	(0.2077)
8	ump_L1	0.4034***	(0.1044)	0.2964	0.3408***	(0.1132)
9	hpret_L1					
10	rp_L1			-0.0576	-0.1889	(0.1898)
11	religion			-0.005	-0.0096	(0.0071)
12	religion x r_incu_pa.	-0.0116*	(0.0066)			
13	migrate10	-0.3548***	(0.1023)	-0.249	-0.4002***	(0.1056)
14	migrate10 x r_incu_pa.	-0.2233**	(0.1025)			
15	edu2000	0.0978***	(0.0163)	0.0955	0.1067***	(0.014)
16	edu2000 x r_incu_pa.	0.0907***	(0.0158)	0.0835	0.0818***	(0.0133)
17	log.m.inc	0.0026	(0.0109)			
18	log.m.inc x r_incu_pa.	0.0231**	(0.0096)			
19	povr	-0.1517***	(0.0495)	-0.1596	-0.1555***	(0.0329)
20	povr x r_incu_pa.	0.0277	(0.0472)			
21	rural	-4e-04	(4e-04)	-4e-04	-7e-04	(4e-04)
22	rural x r_incu_pa.	-8e-04**	(4e-04)	-0.0014	-0.002***	(4e-04)
23	vmail_d	0.0065**	(0.0032)	0.0063	0.0075**	(0.0034)
24						
25	Covariates Selected	19		15	15	
26	Adj. R2	0.8058		0.8034	0.8048	
27	Reg SE	0.0397		0.0399	0.0398	

Reported standard errors are clustered at the State-Year level. For Lasso, 10-fold cross validation is used for model selection, with the random number generator seed is set to: 123. The model selected is the one with CV-MSE 1-SD away from the minimum MSE. Lasso-OLS corresponds to results taking the selected covariates and then subsequently estimating OLS regression in a second-stage. Adjusted R^2 reported for OLS estimates, Deviance ratio reported for Lasso. The list of variables in the active set for VT is given in Table 2.

Table 12: Pooled Panel Regression with Variable Selection for changes in Log Republican Odds ($DLRO$) as the Dependent Variable Estimated over the 2000-2016 Election Cycles

	Covariate	OCMT	SE-OCMT	Lasso	Lasso (OLS)	SE-Lasso (OLS)
1	(Intercept)	0.6955***	(0.0907)	0.6828	0.6763***	(0.0995)
2	r_incu_pa	-0.8364**	(0.3566)	-0.1478	-0.2725***	(0.0645)
3	dLRO_hous	0.025	(0.0281)	0.0186	0.0218	(0.0258)
4	$\widehat{VT}$	-0.3735***	(0.1069)	-0.2839	-0.2894**	(0.1126)
5	$\widehat{VT}$ x r_incu_pa.	-0.1786*	(0.0938)	-0.1094	-0.0166	(0.0921)
6	LBCG_L1			0.0051	0.0235	(0.0375)
7	LBCG_L1 x r_incu_pa.			-0.1118	-0.1119***	(0.0364)
8	hlt_L1					
9	gov_L1			2.4752	2.7807***	(1.0024)
10	rusd_L1	-0.0198	(0.8802)			
11	rusd_L1 x r_incu_pa.	0.0737	(0.8872)			
12	rusd_M3	-0.0389	(0.2468)			
13	rusd_M3 x r_incu_pa.	-0.7329***	(0.2311)	-0.5045	-0.4768***	(0.1479)
14	ump_L1			-0.9088	-0.5727	(0.4429)
15	ump_L1 x r_incu_pa.	-2.6836*	(1.5928)			
16	ump_M3					
17	ump_M3 x r_incu_pa.	4.8527***	(1.6454)	0.9323	2.0594***	(0.4623)
18	hpret_L1	-0.3884	(0.405)			
19	hpret_L1 x r_incu_pa.					
20	hpret_M3	0.7047**	(0.3133)	0.3722	0.4541***	(0.1613)
21	hpret_M3 x r_incu_pa.					
22	rp_L1			-0.8429	-1.4238*	(0.7286)
23	inf_L1			1.0148	1.3794***	(0.5128)
24	migrate10	-1.7827***	(0.4813)	-1.4525	-1.716***	(0.4533)
25	migrateL	0.79	(1.1803)	0.2324	0.872	(1.1734)
26	edu2000	-0.6296***	(0.0962)	-0.6864	-0.7094***	(0.1029)
27	edu2000L	-0.7883***	(0.2045)	-0.7032	-0.7402***	(0.2108)
28	log.popdens.	-0.001	(0.0056)	4e-04	0.0058	(0.0054)
29	log.m.inc					
30	log.m.inc x r_incu_pa.	0.0649*	(0.0336)			
31	povr	-0.6909***	(0.1486)	-0.5167	-0.5939***	(0.1385)
32	rural	0.0089***	(0.0021)	0.005	0.0081***	(0.002)
33						
34	Covariates Selected	21		21	21	
35	Adj. R2	0.5071		0.5185	0.5232	
36	Reg SE	0.1788		0.1768	0.1758	

Reported standard errors are clustered at the State-Year level. For Lasso, 10-fold cross validation is used for model selection, with the random number generator seed is set to: 123. The model selected is the one with CV-MSE 1-SD away from the minimum MSE. Lasso-OLS corresponds to results taking the selected covariates and then subsequently estimating OLS regression in a second-stage. Adjusted R^2 reported for OLS estimates, Deviance ratio reported for Lasso. The list of variables in the active set for $DLRO$ is given in Table 3.

Table 13: Regional Panel Regressions with Dependent Variable as Voter Turnout (VT) over the 2000-2016 Election Cycles using Lasso Algorithm

	Southeast	Southwest	Far West	Rocky Mountain	New England	Mideast	Great Lakes	Plains
(Intercept)	0.063	0.169	0.454	0.201	0.278	0.277	0.121	0.203
r_incu_pa					0.021			
r_incu_pr	0.010	0.030	0.023	0.002		0.021	0.028	0.020
VT_L1	0.796	0.714	0.708	0.677	0.606	0.654	0.761	0.676
VT_L1 x r_incu_pa								
ump_L1		0.100		0.269		-1.081	0.667	0.630
hpret_L1					0.090		0.163	0.268
religion		-0.005	-0.041	-0.026		0.034		
religion x r_incu_pa				0.007				
migrate		-0.175	-0.056					
migrate x r_incu_pa								
edu2000	0.062	0.116	0.103	0.120	0.019	0.069	0.113	0.008
edu2000 x r_incu_pa	0.093	0.061	0.077	0.069		0.011	0.075	0.075
log(m.inc)	0.008		-0.021					
log(m.inc) x r_incu_pa	0.001			0.001				
povr	-0.125	-0.214	-0.345	-0.167	-0.218	-0.327	-0.269	-0.301
povr x r_incu_pa								
rural						-0.001		
rural x r_incu_pa	-0.001	-0.001					-0.001	-

Estimates from recursive voter turnout and voting outcome model. First, Equation 5 is estimated, then $\widehat{VT}$ is used in the active set for estimation of Equation 6. Estimates presented here are for the voter turnout equation, Equation 5. The list of variables in the active set for VT is given in Table 2.

Table 14: Regional Panel Regressions with Dependent Variable as Voter Turnout (VT) Estimated over the 2000-2016 Election Cycles using OCMT Algorithm

	Southeast		Southwest		FW		RM		NE		Mideast		GL		Plains	
(Intercept)	-0.254*	(0.154)	-0.037	(0.236)	0.166***	(0.018)	1.069***	(0.245)	0.072*	(0.043)	0.98***	(0.297)	0.903*	(0.524)	0.195***	(0.04)
r_incu_pa	-0.38*	(0.196)	-0.208	(0.346)	0.963***	(0.314)	-0.315***	(0.12)	0.842**	(0.392)			0.002	(0.179)	-0.023	(0.223)
r_incu_pr	0.02	(0.013)	0.047***	(0.006)			0.003	(0.006)	0.011	(0.016)			0.026***	(0.009)	0.032***	(0.009)
VT_L1	0.857***	(0.02)	0.775***	(0.024)	0.739***	(0.034)	0.704***	(0.022)	0.887***	(0.058)	0.629***	(0.048)	0.805***	(0.054)	0.739***	(0.054)
VT_L1 x r_incu_pa	-0.024	(0.018)	0.041**	(0.018)	-0.114***	(0.036)	-0.026	(0.024)	-0.083	(0.101)			-0.089**	(0.039)	-0.152***	(0.046)
ump_L1	0.163	(0.197)	0.482***	(0.155)							-1.975***	(0.606)				
hpret_L1	0.063	(0.107)	-0.157***	(0.041)					0.117	(0.116)						
religion			-0.022***	(0.006)							0.056**	(0.026)				
religion x r_incu_pa	-0.007	(0.011)	-0.023***	(0.007)	-0.001	(0.012)	0.013**	(0.006)					-0.014	(0.011)	0.005	(0.011)
migrate10			-0.484***	(0.087)	-1.06***	(0.263)							-0.498***	(0.126)		
migrate10 x r_incu_pa	-0.181	(0.151)	-0.015	(0.094)	-0.367	(0.234)							-0.108	(0.152)	-0.157	(0.169)
edu2000	0.047**	(0.022)	0.132***	(0.029)	0.166***	(0.04)	0.227***	(0.029)			0.15***	(0.055)	0.216***	(0.067)		
edu2000 x r_incu_pa	0.073***	(0.02)	0.113***	(0.035)	0.195***	(0.016)	0.062**	(0.025)	0.068	(0.055)			0.129***	(0.024)	0.093***	(0.022)
log(m.inc)	0.033**	(0.013)	0.016	(0.02)			-0.082***	(0.023)			-0.059**	(0.027)	-0.071	(0.045)		
log(m.inc) x r_incu_pa	0.038**	(0.018)	0.016	(0.031)	-0.078***	(0.027)	0.031***	(0.011)	-0.067**	(0.03)			0.006	(0.014)	0.011	(0.02)
povr	-0.07	(0.051)	-0.193***	(0.053)	-0.306***	(0.038)	-0.412***	(0.061)			-0.569***	(0.093)	-0.428**	(0.176)	-0.25***	(0.062)
povr x r_incu_pa	0.05	(0.066)	0.089	(0.087)	-0.464***	(0.168)			-0.435***	(0.167)	-0.013	(0.031)	-0.143	(0.096)	-0.063	(0.052)
rural			0.001	(0.001)							-0.002**	(0.001)				
rural x r_incu_pa	-0.001	(0.001)	-0.002**	(0.001)				(0.001)							-0.001	(0.001)

Estimates from recursive voter turnout and voting outcome model. First, Equation 5 is estimated, then $\widehat{VT}$ is used in the active set for estimation of Equation 6. Estimates presented here are for the voter turnout equation, Equation 5. The list of variables in the active set for VT is given in Table 2. Standard errors are clustered at the state level, in parenthesis to the right of the corresponding column of estimates. FW, NE, RM and GL refer to Far West, New England, Rocky Mountain and Great Lakes regions, respectively.

Table 15: Regional Panel Regressions with Dependent Variable as Changes in Log Republican Odds ($DLRO$) over the 2000-2016 Election Cycles using Lasso Algorithm

	Southeast	Southwest	Far West	Rocky Mountain	New England	Mideast	Great Lakes	Plains
(Intercept)	1.222	3.831	0.454	0.401	-0.272	0.911	0.788	0.436
r_incu_pa	-0.043	-0.020	-0.902	-0.124	-0.524	-0.332	-0.428	-0.108
$\widehat{VT}$	-0.597	0.312	0.116	-0.318	-0.120	-0.607		-0.002
$\widehat{VT}$ x r_incu_pa			-0.189	-0.147		0.215	0.190	-0.087
LBCG_L1	0.008	-0.012	-0.051		-0.309		-0.166	-0.025
LBCG_L1 x r_incu_pa	-0.089		-0.011	-0.154	0.526	-0.243	0.001	-0.086
ump_L1	-1.770	2.100	3.666		-0.890	-1.649	2.253	-1.061
ump_L1 x r_incu_pa	0.001	0.263	-2.168			1.613	3.893	0.380
ump_M3	-0.499	-1.536	-4.135				-1.047	4.108
ump_M3 x r_incu_pa	0.187		3.290		7.901		-1.917	
hpret_L1	-0.976	-0.104			-2.261	0.500	2.071	4.000
hpret_L1 x r_incu_pa	-0.716	-0.696	0.175			-1.306	-2.235	-2.995
hpret_M3	1.689	0.561	0.682	0.128		2.062	1.271	
hpret_M3 x r_incu_pa	0.301	0.717	-0.223	0.921	3.176		1.807	1.055
migrate	-1.335	-3.022	0.994		-0.043	-3.364	-0.008	-0.998
migrate*	1.078		-0.878		0.169		3.976	
edu2000	-0.606	-0.845	-0.646	-0.586	-0.729	-0.295	-0.854	-0.753
edu2000*	-2.301	-0.489			0.204	-0.246	-0.928	-0.442
log(popdens)	-0.010	0.008	-0.009	-0.005	-0.003	0.014	-0.004	
log(m.inc)		-0.322	-0.032		0.055	-0.028	-0.045	-0.025
log(m.inc) x r_incu_pa	-0.004	-0.008	0.079					
povr	-0.943	-1.656	-0.392	-0.239		-0.599	-0.339	-0.592
rural	0.009	0.008	-0.001		-0.014		-0.001	

Estimates from recursive voter turnout and voting outcome model. First, Equation 5 is estimated, then $\widehat{VT}$ is used in the active set for estimation of Equation 6. Estimates presented here are for the log Republican odds equation, Equation 6. The list of variables in the active set for $DLRO$ is given in Table 3.

Table 16: Regional Panel Regressions with Dependent Variable as Changes in Log Republican Odds ($DLRO$) Estimated over the 2000-2016 Election Cycles using OCMT Algorithm

	Southeast		Southwest		FW		RM		NE		Mideast		GL		Plains	
(Intercept)	-2.354***	(0.55)	0.438***	(0.01)	0.128**	(0.062)	0.533***	(0.071)	0.507***	(0.123)	0.346***	(0.036)	0.543***	(0.083)	-0.267	(0.175)
r_incu_pa					-2.545***	(0.392)	-1.534**	(0.752)	-3.68***	(0.235)	-1.494***	(0.508)	-1.539***	(0.383)	-1.022**	(0.409)
$\widehat{VT}$	-0.588***	(0.186)			0.115	(0.129)	-0.383***	(0.098)	-0.566***	(0.185)	-0.308***	(0.11)	-0.033	(0.094)	0.453***	(0.118)
$\widehat{VT}$ x r_incu_pa					-0.285	(0.179)	-0.207	(0.19)	-0.209**	(0.098)	0.417***	(0.094)	0.223*	(0.132)	-0.004	(0.098)
LBCG_L1			-0.078*	(0.042)												
LBCG_L1 x r_incu_pa									0.055	(0.303)	-0.277***	(0.066)				
ump_L1					3.932***	(0.341)										
ump_L1 x r_incu_pa					-1.845**	(0.89)	3.958	(3.645)	14.579***	(4.054)			9.201**	(4.627)	1.094	(4.152)
ump_M3					-4.152***	(1.069)									4.141***	(0.914)
ump_M3 x r_incu_pa					3.432***	(1.226)	-3.961	(4.949)	-10.078***	(3.881)	5.528*	(3.049)	-7.281	(5.613)	2.592	(4.906)
hpret_L1	-0.866	(0.724)			-0.069	(0.529)					0.85	(0.86)	0.41	(1.416)	4.06***	(0.917)
hpret_L1 x r_incu_pa	-1.227***	(0.388)													-4.248***	(1.095)
hpret_M3	2.595***	(0.706)			0.727*	(0.387)	0.748*	(0.41)			1.264	(0.786)	2.14*	(1.158)	-0.026	(0.949)
hpret_M3 x r_incu_pa															2.708**	(1.118)
migrate10	-1.62***	(0.586)	-3.549***	(0.649)	1.359*	(0.82)	-0.812	(1.079)			-4.107***	(0.308)	0.333	(1.016)	-0.807	(0.522)
migrateL					-1.001	(1.653)					2.808*	(1.609)				
edu2000	-0.828***	(0.159)	-1.023***	(0.13)	-0.683***	(0.102)	-0.666***	(0.089)	-0.295	(0.194)	-0.331***	(0.11)	-0.964***	(0.129)	-0.809***	(0.135)
edu2000L	-1.698***	(0.379)			0.041	(0.507)	-0.01	(0.136)			-0.669**	(0.295)	-1.467***	(0.354)	-0.144	(0.48)
log(popdens)	-0.006	(0.005)	0.01	(0.015)	-0.011	(0.01)	-0.016***	(0.003)			0.015***	(0.005)				
log(m.inc)	0.299***	(0.057)														
log(m.inc) x r_incu_pa					0.234***	(0.05)	0.136*	(0.082)	0.316***	(0.023)	0.075*	(0.044)	0.1***	(0.029)	0.065	(0.044)
povr			-0.902***	(0.077)	-0.44*	(0.237)	-0.669***	(0.147)					-0.331***	(0.06)		
rural	0.013***	(0.002)	0.02***	(0.006)	-0.001	(0.004)		(0.004)			-0.001	(0.002)		(0.001)	-0.001	(0.004)

Estimates from recursive voter turnout and voting outcome model. First, Equation 5 is estimated, then $\widehat{VT}$ is used in the active set for estimation of Equation 6. Estimates presented here are for the log Republican odds equation, Equation 6. The list of variables in the active set for $DLRO$ is given in Table 3. Standard errors are clustered at the state level, in parenthesis to the right of the corresponding column of estimates. FW, NE, RM and GL refer to Far West, New England, Rocky Mountain and Great Lakes regions, respectively.

Online Supplement to "Regional Heterogeneity and 2016 and 2020 U.S. Elections"

Rashad Ahmed
University of Southern California, USA

M. Hashem Pesaran
University of Southern California, USA, and Trinity College, Cambridge, UK

October 18, 2020

This Online Supplement is organized in four sections. Section S1 provides detail on relevant data and sources. Section S2 describes selecting the functional form of the election outcome variable. Section S3 gives an account of Lasso and OCMT forecasting algorithms. S4 reports additional results.

S1 Data

Descriptions, Frequency, Sources

Data has been cleaned and merged from several different publicly available sources. County-level voting outcomes are taken from the MIT Election Data and Science Lab. County GDP measures are obtained from the Bureau of Economic Analysis. Education, population, migration, and urban-rural county classifications are from the USDA. Annual median household income and poverty estimates are from the U.S. Census and typically update with a lag ranging from one to two years. Information on religiosity across counties comes from the 2010 survey provided by the Association of Religion Data Archives. Data on voting age population (VAP) are from the American Community 5-year surveys. County-level unemployment rates are provided by the BLS and county-level house price indices are taken from Zillow. State-level inflation is computed from indices reported by the Bureau of Economic Analysis (BEA). State level export-weighted real exchange rates are from the Federal Reserve Bank of Dallas. Government employment growth, healthcare expenditures and rent expenditures at the state level are taken from the BEA. In total, we analyze 3,107 counties from 48 of the U.S. Mainland states plus Washington D.C. The number of counties by state is found in Table S.9.

County classifications change over time, and different data sets rely on different vintage classifications. For these reasons, cleaning and merging the data required manual adjustments for some of the observations. We describe data series and cleaning procedures for the main variables of interest in more detail below.

County FIPS Changes: Some counties changed FIPS codes over the period 2000-2016. For these counties, we made adjustments to ensure different data sets can be merged properly. County 08014 (Colorado) did not exist until 2001 (it was created from 4 other Colorado counties). We add 08014's post-2000 election votes to county 08059, Jefferson County, the largest of the counties which contributed to 08014's creation. The state of Virginia decided to merge County 51515 ("Bedford") into county 51019 ("Bedford County") in 2013, therefore County 51515 no longer existed afterward. 2013. To account for this we allocate votes of County 51515 from 2004, 2008 and 2012 to those of county 51019, effectively combining the two counties over the entire sample. County 46113 (South Dakota) was renamed to Oglala Lakota county in 2015 and given a new FIPs code: 46102.

County U.S. Presidential Votes, every 4 years: Data from the MIT election lab provides election results at the county level for years 2000, 2004, 2008, 2012, and 2016. We focus on two-party vote share, hence rely on Republican and Democrat vote statistics across counties. We also focus on the 48 mainland states, excluding Alaska and Hawaii from our analysis.

Annual County GDP, annual: Data from the U.S. Bureau of Economic Analysis covers annual real (chained 2012 U.S. Dollars) GDP across over 3,000 counties from 2001 to 2018. This yields annual growth rates from 2002 to 2018. We interpolate 1999-2000 and 2000-2001 GDP growth rates with the 2001-2002 growth rate, for all counties. County GDP data has historically been updated with a one-year lag every October. However for 2020, the 2019 GDP data is not expected to be released until December 2020.

Virginia County GDP, annual: For the State of Virginia, the BEA consolidates real GDP data for 52 of the smaller counties into 23 groups of two to three counties each. In order to match GDP to voting data, these consolidated GDP measures need to be matched back to individual counties. To do so, for aggregated GDP assigned to a given group of counties, we assign all counties within that group the GDP values given to the group. Therefore, we assume counties within a group have the same real GDP growth rate.

Voter Turnout, every 4 years: We estimate voter turnout (VT) as the total votes (Republican and Democrat) divided by the VAP, voting age population, which we take from the 5-year ACS. To compute VT, we rely on the 90% upper confidence interval of the VAP estimate. The VAP measure is an estimate over a 5-year period while the number of votes is a single snapshot in time. We use 2012-2016 VAP estimates to compute 2016 voter turnout, 2008-2012 estimates for 2012 voter turnout, and 2005-2009 estimates for 2008 voter turnout. Because we do not have VAP estimates earlier than 2008, we back-fill 2004 turnout values using 2008 turnouts. Four county-year observations (from over 12,000) report VT values of greater than 1, likely because of measurement error. For these cases, we use the average VT of adjacent counties[S1] For counties with a VAP/population ratio larger than 1, we replace VAP for these counties with the product of the county population with the average of VAP/population ratio of surrounding counties (within 100 miles) which have VAP/population ratios less than 1.

Midterm Elections, every 2 years: We collect data on U.S. house votes for biennuial elections by state from MIT election lab. Because the House votes every two years, it may be a useful indicator for political momentum running up to the presidential election, which occurs every four years. For the state of Vermont, where Bernard Sanders (an independent) has received consistent and significant vote share, we consolidate his political affiliation with those of Democrats in order to remain consistent with the two-party framework of this study. In order to merge with the remaining data, we impute vote results of Maryland into Washington DC because the latter does not have voting rights during these elections. We use this data to compute Republican vote share variables using House election data, analogous to county presidential Republican vote share data.

Religiosity, time-invariant: Data is from the Association of Religion Data Archives. Religiosity measures the proportion of county population adhering to a religion. Rates of religious adherence can exceed 1 for some counties because survey participants can report adherence to multiple religions or denominations. While this does not pose any serious issues, in order to keep the rate variable bounded between 0 and 1, for counties with greater than 100 percent religiosity rate, we replace county i's religiosity rate with the local religiosity rate, taken as the average religiosity rate of all counties within 100 miles of i.

County House Prices, monthly: We take monthly house price indices at the county level from Zillow. These go back to the 90's, but not for every county or every year-month.

[S1] The observations are: Harding County, NM 2004/2008/2012, and Hanson County, SD 2012.

We therefore estimate local county house price returns based on the average of counties within 100 miles of county i, inclusive of county i. For counties with no data available, we impute values using the cross-section average of all available local returns over the same time period. For the year 2016, logged annual house price changes are computed as the average monthly change from July 2015 to June 2016. This is then annualized. For each election year, the annualized return is computed similarly. This guarantees that the data used are always available prior to the election. Along with annual house price changes, we also compute short-term averages over the 3-month period of July-September of each election year. The monthly house price data typically update with a two month lag.

State-level Rent Expenditures, annual: We compute annual log growth rates in state-level rents using per capita personal consumption expenditures on housing and utilities. These data are taken from the BEA and are typically updated with a one-year lag every October.

Unemployment Rates, monthly: We take monthly unemployment rates at the county level from the BLS. we also estimate local averages using all counties within 100 miles of county i. For the year 2016, we calculate annual average unemployment based on July 2015 to June 2016. For each election year, the annual average unemployment rate is computed over July of year $t-1$ to June of year t. This guarantees that the data used are available prior to the actual year t election. Along with annual unemployment averages, we also compute short-term averages over the 3-month period of July-September of each election year. The monthly unemployment data typically update with a two month lag.

State-level Inflation, quarterly: From the BEA, we take quarterly real GDP and nominal GDP by state to compute a state-level annual GDP deflator as

$$\text{GDP Deflator} = \frac{\text{Nominal GDP}}{\text{Real GDP}} \times 100.$$

Inflation is calculated as the logged change form the previous quarter's GDP Deflator, by state. Because Elections are held every November, we use state-level inflation rate from year Q3 2015 - Q2 2016 for 2016, and so on to guarantee data availability prior to each election. These data are taken from the BEA and typically released with a 2-quarter lag.

State-level USD Real Effective Exchange Rates, monthly: USD state-level real exchange rates are taken from the Federal Reserve Bank of Dallas. Monthly state-level exchange rates are computed using a trade-weighted average of USD exchange rates via the

primary export partners of the state. We compute logged monthly changes using monthly REERs over July of year $t-1$ to June of year t, averaging monthly changes to compute a monthly average over the year, which is then annualized. Along with annual average exchange rate changes, we also compute short-term averages over the 3-month period of July-September of each election year. The monthly exchange rate data typically updates with a three month lag.

State-level Healthcare Expenditures, annual: We compute annual log growth rates in state-level cost of healthcare using per capita personal consumption expenditures on healthcare by state. These data are taken from the BEA and are typically updated with a one year lag every October.

State-level Government Employment, annual: We compute annual growth rates in the size of local government by state, by computing the share of the state's labor force allocated to the local and state government sector. Annual Growth rates are computed using log changes. These data are taken from the BEA and are typically with a one year lag every September or October.

Population Density, time-invariant: We compute county population densities using 2000 and 2010 population estimates, divided by the total land area (based on 2000) of the county.

State Mail-in Vote Policies, time-invariant: We also collect data at the state level measuring the ease with which one can cast a vote by mail. Policies vary at the state level. In fact, some states, namely Oregon, Utah, Colorado and Hawaii only accept votes by mail. We construct a state-level time-invariant indicator variable which takes values of (1,0,-1) depending on whether mail-in voting is: 1: The default voting method, 0: Optional but open to everyone or -1: An excuse is required to cast a mail-in vote. Underlying source for these data is FiveThirtyEight.com and The National Conference of State Legislatures.

Incumbent Party and Incumbent President indicators, every 4 years: To capture the incumbency effects on voter turnout and election outcome we consider two incumbency indicators, and distinguish between presidential and party incumbency indicators. The "incumbent party indicator" takes the value of 1 if on election day the president in power is Republican, and -1 if he/she is a Democrat. The "incumbent president indicator" takes the value of 1 if the president who is running for re-election is a Republican, takes

the value of -1 if he/she is a Democrat, and takes the value of 0 if neither of the two candidates is incumbent. These indicators are considered on their own, as well as interacted with a number of other covariates. In this way we allow for a wide variety of incumbency effects (positive or negative) discussed in the literature, without biasing the results in favor or against the incumbent president or party.

Being Economically 'Left-Behind'

We take real GDP levels and compute annual log growth rates, denoted by

$$\Delta y_{cr,t} = \ln \frac{Y_{cr,t}}{Y_{cr,t-1}}, \tag{S.1}$$

where $Y_{cr,t}$ is the real GDP of county c in region r during year t. County-level real GDP growth is the main source of data used to construct a new measure representing the degree to which resident of a particular county are, on average, economically 'left behind' (LB). Consider an individual outcome variable of interest, in our case, real GDP $Y_{cr,t}$ for county c in year t and its "local" (or "regional") counterpart, defined by:

$$Y^*_{cr,t} = \sum_{c'=1}^{N} w_{c,c'} Y_{c'r,t}, \tag{S.2}$$

where N denotes the number of counties in the country as a whole, $w_{c,c'} \geq 0$, and $\sum_{c'=1}^{N} w_{c,c'} = 1$. Note that $Y^*_{cr,t}$ is inclusive of c, but we can also compute $Y^*_{cr,t}$ exclusive of c by setting $w_{c,c} = 0$. In practice, $w_{c,c'}$ could be the neighborhood weights, within a given radius around the cth location.

To consider a measure of "Left Behind", an obvious reference measure is to compare $Y_{cr,t}$ or $Y^*_{cr,t}$ to is the national ("global") measure where $w_{c,c'} = w_{c'} \ \forall \ i$. In practice the national measure could be based on population weights. In what follows we denote national (global) reference measure by Y_t, the local/regional measure by $Y^*_{cr,t}$, and the individual county measure by $Y_{cr,t}$.

The extent to which county c is left behind relative to the nation, Y_t, also depends on the time horizon over which the individual/local measure is compared to the reference (national) group. For example, county c can be left behind either individually, or locally, relative to the national group over a period of h years. Accordingly, we consider the change from $\ln(Y_{cr,t-h}/Y_{t-h})$ to $\ln(Y_{cr,t}/Y_t)$, for a given horizon h. The extent to which c is individually "left behind" is measured by

$$G_{cr,t}(h) = \frac{1}{h}\Delta_h \ln(Y_{cr,t}/Y_t) = \frac{1}{h}\Delta_h \ln(Y_{cr,t}) - \frac{1}{h}\Delta_h \ln(Y_t) = \frac{\ln(Y_{cr,t}) - \ln(Y_{cr,t-h})}{h} - \frac{\ln(Y_t) - \ln(Y_{t-h})}{h} \tag{S.3}$$

if $G_{cr,t}(h) < 0$. County c is not left behind if $G_{cr,t}(h) > 0$. A measure of being left behind locally can be similarly defined as

$$G^*_{cr,t}(h) = \frac{1}{h}\Delta_h \ln(Y^*_{cr,t}/Y_t). \tag{S.4}$$

It is clear that c can be left behind relative to the country as a whole, but not at the local level and *vice versa*. Moreover, c could be left behind relative to local as well as national measures.

To study the degree of left-behindedness at a relatively disaggregated level, we consider annual real economic output across U.S. counties (excluding counties in Alaska and Hawaii) as our outcome variable, $Y_{cr,t}$. Our national measure Y_t is simply the aggregate national U.S. real output.[S2] To compute local measures $Y^*_{cr,t}$, we consider a radius of 100 miles around each county c ($R = 100$). In measuring $Y^*_{cr,t}$, all counties outside of 100 miles receive a weight of 0, while the real output measures of all counties within 100 miles are equally weighted, specifically

$$w_{c,c'} = \begin{cases} \frac{1}{N_R}, & \text{if } c' \text{ is within 100 miles of } c \\ 0, & \text{otherwise} \end{cases}$$

where the number of counties within 100 miles of c, inclusive of c, is N_R.[S3]

S2 Functional Form of the Outcome Variable

The standard two-party voting outcome in the literature is given by party vote share

$$V_{cr,t} = \frac{R_{cr,t}}{R_{cr,t} + D_{cr,t}}, \tag{S.5}$$

where $R_{cr,t}$ is the number of Republican votes by county c of region r in election year t, and $D_{cr,t}$ is the number of Democratic votes. The outcome $V_{cr,t}$ is equal to the Republican

[S2] We do not compute Y_t; rather we take the data directly from the BEA.

[S3] Between-county distances are taken from the NBER database, specifically these are great-circle distances calculated using the Haversine formula based on internal points in the geographic area.

share of the two-party vote. However, despite $V_{cr,t}$ being the target variable, whether better predictions are produced using $V_{cr,t}$ or a transformation of $V_{cr,t}$ (which is ultimately re-transformed back) is an issue that needs to be addressed prior to forecasting. In this context, we evaluate three different functional forms of the outcome variable summarized by $V'_{cr,t}$:

$$V'_{c,rt} = \left\{ V_{c,rt},\ \ln(V_{c,rt}),\ \ln\left(\frac{V_{c,rt}}{1-V_{cr,t}}\right) \right\}, \tag{S.6}$$

where the latter term is the main dependent variable we chose to use in our analysis – the Republican log-odds of the two-party vote:

$$LRO_{cr,t} = \ln\left(\frac{V_{cr,t}}{1-V_{cr,t}}\right) = \ln\left(\frac{R_{c,rt}}{D_{c,rt}}\right). \tag{S.7}$$

Despite using $LRO_{cr,t}$ in the regression, the target variable we wish to forecast remains the Republican vote share over an election cycle, $V_{cr,t}$. If we rely on a model with a transformed dependent variable, then its predictions must be re-transformed to match the units of the actual target. While the adjusted R^2 across models may suggest which specification best explains the dependent variable, this accounts for re-transforming the prediction back to the target variable. Therefore, to appropriately compare models under transformed dependent variables, regression standard errors must be adjusted to be comparable across specifications. We follow the likelihood approach discussed in Section 11.7 of Pesaran (2015).

The conventional dependent variable in the political science literature is the (change in) Republican vote share, $V_{cr,t}$, or the dependent variable corresponding to column 2 of Table S.8. To select the best functional form for the dependent variable, standard errors from the active set regression on, say, changes in the standard dependent variable $V_{cs,t}$ can be compared to the adjusted standard errors from the active set regressions under other functional forms (columns 1 and 3). Adjustment factors must be applied for comparison.

For the column 1 dependent variable, $\Delta_4\left(\frac{V_{cr,t}}{1-V_{cr,t}}\right)$, we have the following log adjustment factor:

$$\ln \bar{z}_1 = -\frac{1}{NT}\sum_{t=1}^{T}\sum_{i=1}^{N}\ln V_{cr,t} - \frac{1}{NT}\sum_{t=1}^{T}\sum_{i=1}^{N}\ln(1-V_{cr,t}), \tag{S.8}$$

and for the column 3, with $\Delta_4 \ln V_{cr,t}$, the log adjustment factor is:

$$\ln \bar{z}_3 = -\frac{1}{NT}\sum_{t=1}^{T}\sum_{i=1}^{N}\ln V_{cr,t}. \tag{S.9}$$

The "Adjusted SE" in Table S.8 compares post-adjustment regression SEs. The results show that regression performance under the traditional functional form using simple vote

shares (column 2) may be improved by using instead the change in log odds ratio variable (column 1). The former has a regression standard error of 0.037, compared to the adjusted standard error of 0.036 under the model where we transform the vote share into a log-odds ratio, $\Delta_4 \ln\left(\frac{V_{cs,t}}{1-V_{cr,t}}\right)$. The log vote share, $\Delta_4 \ln V_{cr,t}$ has the largest adjusted SEs. Motivated by these results, we use changes in log-odds ratios as our dependent variable.

S3 Forecasting Algorithms

OCMT

One set of forecasts implement the OCMT algorithm presented in Chudik et al. (2018). We apply OCMT on both the pooled sample and on regional sub-samples, in both turnout and voting regressions on their respective active sets. OCMT selects variables based on multiple-testing corrected statistical significance. We define the critical value threshold as

$$c_p(k,\delta) = \Phi^{-1}\left(1 - \frac{p}{2k^{\delta}}\right),$$

where k is the number of covariate in the active set, $\Phi^{-1}(.)$ is the inverse of the cumulative distribution of the standard normal variate, p is the nominal; size of the test, and δ measures the degree to which multiple testing is taken into account. We set $\delta = 1$ in the first stage, and $\delta^* = 2$ in subsequent stages, and a p-value $p = 0.05$. Under the pooled model, the p-values are derived from state-year clustered standard errors. For the regional model, p-values are derived from state-clustered standard errors. This approach is taken for both regressions of turnout and voting. We refer to the original paper for further technical details.

Lasso

Our second set of forecasts are generated using the Lasso algorithm. Because we rely on cross-validation to calibrate the trade-off between fit and parsimony, it is important to set the numeric seed before running simulations - this ensures our results from Lasso algorithm are replicable When running the program, we always set our seed equal to "123". All covariates are standardized to mean zero and unit standard deviation prior to estimation. In n-fold cross-validation, we set $n = 10$ and our loss criteria is based on mean-squared error. The model we select is that which has the smallest regularization penalty parameter yet which still falls within 1-standard deviation of the model yielding the minimum MSE. The Online Supplement of Chudik et al. (2020) contains further technical details providing computer codes for implementation of OCMT and Lasso algorithms used in this paper.

S4 Additional Results

S4.1 Consistency proof of the two-stage estimation

Here we establish consistency of the two-stage estimation of the recursive model, which we write compactly as

$$\begin{aligned}\mathbf{y}_1 &= \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{u}_1,\\ \mathbf{y}_2 &= \gamma\mathbf{y}_1 + \mathbf{X}_2\boldsymbol{\beta}_2 + \mathbf{u}_2,\end{aligned}$$

where $\mathbf{X}_1$ and $\mathbf{X}_2$ are $T \times k_1$ and $T \times k_2$ matrices of exogenous variables, coefficients $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ are $k_1 \times 1$ and $k_2 \times 1$ vectors, and $\mathbf{u}_1$ and $\mathbf{u}_2$ are $T \times 1$ vectors of errors. For instance, let $\mathbf{y}_1$ and $\mathbf{y}_2$ represent voter turnout and the log odds ratio, respectively ($\mathbf{y}_1 = VT$ and $\mathbf{y}_2 = DLRO$). Notice that our recursive structure imposes that $\mathbf{y}_2$ does not enter the $\mathbf{y}_1$ equation. We assume that $\mathbf{X}_1$ and $\mathbf{X}_2$ are weakly exogenous such that

$$\frac{\mathbf{X}_1'\mathbf{u}_1}{T} \xrightarrow{p} \mathbf{0}, \quad \frac{\mathbf{X}_1'\mathbf{u}_2}{T} \xrightarrow{p} \mathbf{0}, \frac{\mathbf{X}_2'\mathbf{u}_1}{T} \xrightarrow{p} \mathbf{0}, \quad \frac{\mathbf{X}_2'\mathbf{u}_2}{T} \xrightarrow{p} \mathbf{0}.$$

It then follows that $\boldsymbol{\beta}_1$ is consistently estimated by $\hat{\boldsymbol{\beta}}_1 = (\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{y}_1$. Using this estimate, we obtained the fitted values, $\hat{\mathbf{y}}_1 = \mathbf{X}_1\hat{\boldsymbol{\beta}}_1$ which can be used in the second stage to consistently estimate $\boldsymbol{\theta} = (\gamma_1, \boldsymbol{\beta}_2')'$ by

$$\hat{\boldsymbol{\theta}} = (\hat{\mathbf{Z}}'\hat{\mathbf{Z}})^{-1}\hat{\mathbf{Z}}'\mathbf{y}_2, \quad \hat{\mathbf{Z}} = (\hat{\mathbf{y}}_1, \mathbf{X}_2).$$

To establish consistency of $\hat{\boldsymbol{\theta}}$, we note that

$$\begin{aligned}\mathbf{y}_2 &= \gamma\hat{\mathbf{y}}_1 + \mathbf{X}_2\boldsymbol{\beta}_2 + \underbrace{\mathbf{u}_2 + \gamma(\mathbf{y}_1 - \hat{\mathbf{y}}_1)}_{\boldsymbol{\xi}}\\ \mathbf{y}_2 &= \hat{\mathbf{Z}}\boldsymbol{\theta} + \boldsymbol{\xi},\end{aligned}$$

such that $\hat{\boldsymbol{\theta}} = (\hat{\mathbf{Z}}'\hat{\mathbf{Z}})^{-1}\hat{\mathbf{Z}}'(\hat{\mathbf{Z}}\boldsymbol{\theta} + \boldsymbol{\xi})$. Hence

$$\hat{\boldsymbol{\theta}} - \boldsymbol{\theta} = \left(\frac{\hat{\mathbf{Z}}'\hat{\mathbf{Z}}}{T}\right)^{-1}\frac{\hat{\mathbf{Z}}'\boldsymbol{\xi}}{T}.$$

But,

$$
\begin{aligned}
\frac{\hat{\mathbf{Z}}'\boldsymbol{\xi}}{T} &= \frac{\hat{\mathbf{Z}}'\mathbf{u}_2}{T} + \gamma\frac{\hat{\mathbf{Z}}'\mathbf{e}_1}{T}, \\
\mathbf{e}_1 = \mathbf{y}_1 - \hat{\mathbf{y}}_1 &= \mathbf{y}_1 - \mathbf{X}_1(\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{y}_1, \\
&= \mathbf{M}_1\mathbf{y}_1, \quad \mathbf{M}_1 = \mathbf{I} - \mathbf{X}_1(\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1',
\end{aligned}
$$

and

$$
\frac{\hat{\mathbf{Z}}'\mathbf{e}_1}{T} = \begin{pmatrix} \hat{\mathbf{y}}_1'\mathbf{e}_1/T \\ \mathbf{X}_2'\mathbf{e}_1/T \end{pmatrix}.
$$

Also, it readily follows that $\hat{\mathbf{y}}_1'\mathbf{e}_1 = \hat{\boldsymbol{\beta}}_1'\mathbf{X}_1'[\mathbf{M}_1\mathbf{y}_1] = 0$, since $\mathbf{X}_1'\mathbf{M}_1 = 0$. Then, we have

$$
\begin{aligned}
T^{-1}\mathbf{X}_2'\mathbf{e}_1 &= T^{-1}\mathbf{X}_2'\mathbf{M}_1(\mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{u}_1), \\
&= T^{-1}\mathbf{X}_2'\mathbf{M}_1\mathbf{u}_1 \\
&= \frac{\mathbf{X}_2'\mathbf{u}_1}{T} - \frac{\mathbf{X}_2'\mathbf{X}_1}{T}\left(\frac{\mathbf{X}_1'\mathbf{X}_1}{T}\right)^{-1}\frac{\mathbf{X}_1'\mathbf{u}_1}{T} \\
&\xrightarrow{p} 0 \quad .
\end{aligned}
$$

Therefore, $\frac{\hat{\mathbf{Z}}'\mathbf{e}_1}{T} \xrightarrow{p} 0$. Also,

$$
\frac{\hat{\mathbf{Z}}'\mathbf{u}_2}{T} = \begin{pmatrix} \hat{\mathbf{y}}_1'\mathbf{u}_2/T \\ \mathbf{X}_2'\mathbf{u}_2/T \end{pmatrix},
$$

and

$$
\frac{\hat{\mathbf{y}}_1'\mathbf{u}_2}{T} = \frac{\hat{\boldsymbol{\beta}}'\mathbf{X}_1'\mathbf{u}_2}{T} \xrightarrow{p} 0, \quad \frac{\mathbf{X}_1'\mathbf{u}_2}{T} \xrightarrow{p} 0.
$$

Hence, overall we have $\hat{T^{-1}}\mathbf{Z}'\boldsymbol{\xi} \xrightarrow{p} 0$, and hence $\hat{\boldsymbol{\theta}} \xrightarrow{p} \boldsymbol{\theta}$.

S4.2 Additional panel regression results

As a robustness test, we average county-level Republican and Democrat predicted votes across OCMT and Lasso approaches to produce model averaged predictions. Table S.3 reports 2016 vote share and electoral college forecasts under the OCMT + Lasso averaging approach. Table S.4 similarly reports 2020 vote share and electoral predictions across states. For 2016, the predicted outcomes largely coincide with outcomes produced from individual

models. Averaging the regional models also predicts a Republican victory in 2016. The regional-average prediction of Republican electoral votes was higher than individual models: 330 (2016 actual was 304). By contrast, individual regional models predicted 308 (Lasso) and 307 (OCMT), for 2016 respectively. The higher vote count of the average model is driven by switched electoral votes for some swing states. For example, OCMT-regional predicted 0 republican electoral votes for Minnesota, 7 from Oregon, and 0 from Pennsylvania. The regional-averaged model flipped these predictions (10 from Minnesota, 0 from Oregon, 20 from Pennsylvania). Hence a difference of 13 electoral votes between the OCMT-regional prediction and the regional-average prediction. For 2020, the regional-averaged model predicts a Democratic electoral victory by a single vote. This reflects the different forecasts under the individual OCMT-regional (which predicts Republican) and Lasso-regional (which predicts Democrat) models.

Figure S.1: Bureau of Economic Analysis Regions

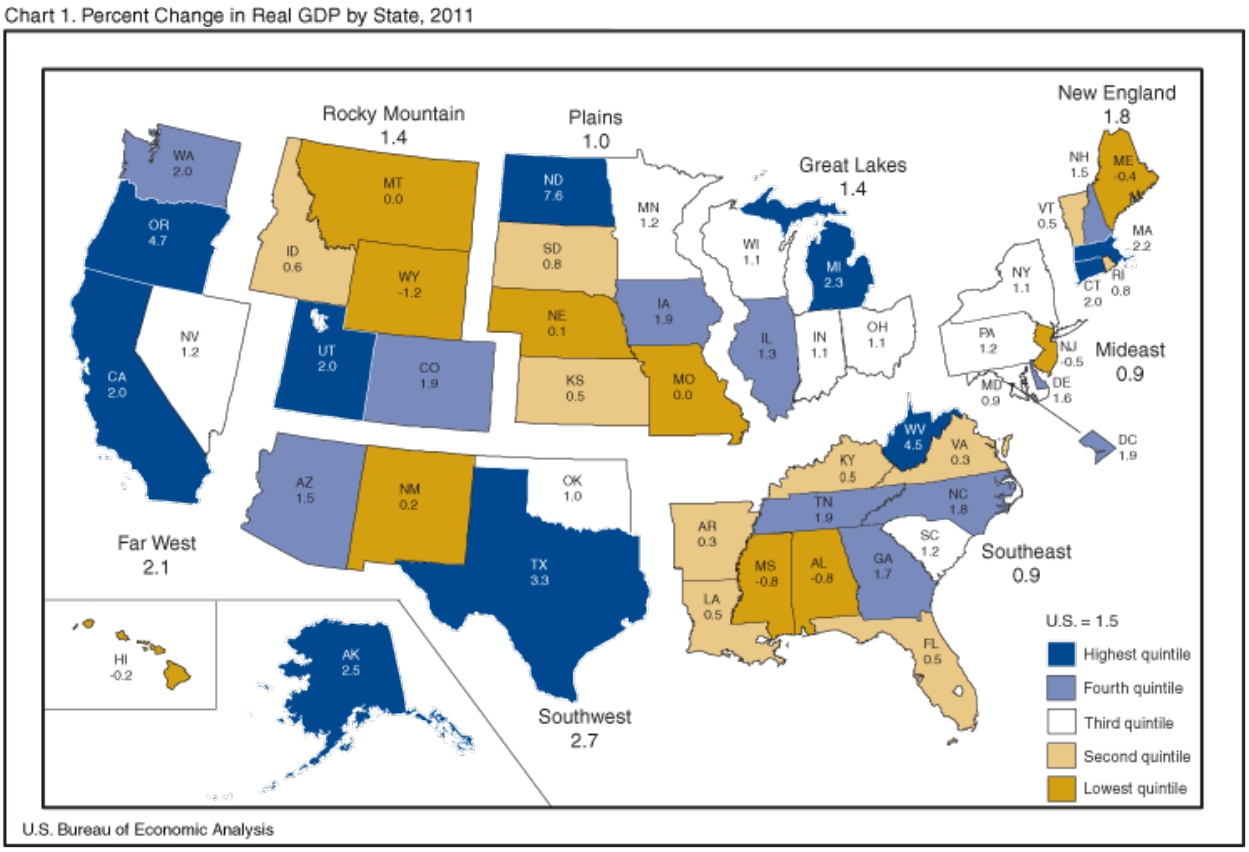


Figure S.2: Histogram of Voter Turnout (VT) over the period 2004-2016 at Mainland U.S. and Regional Levels

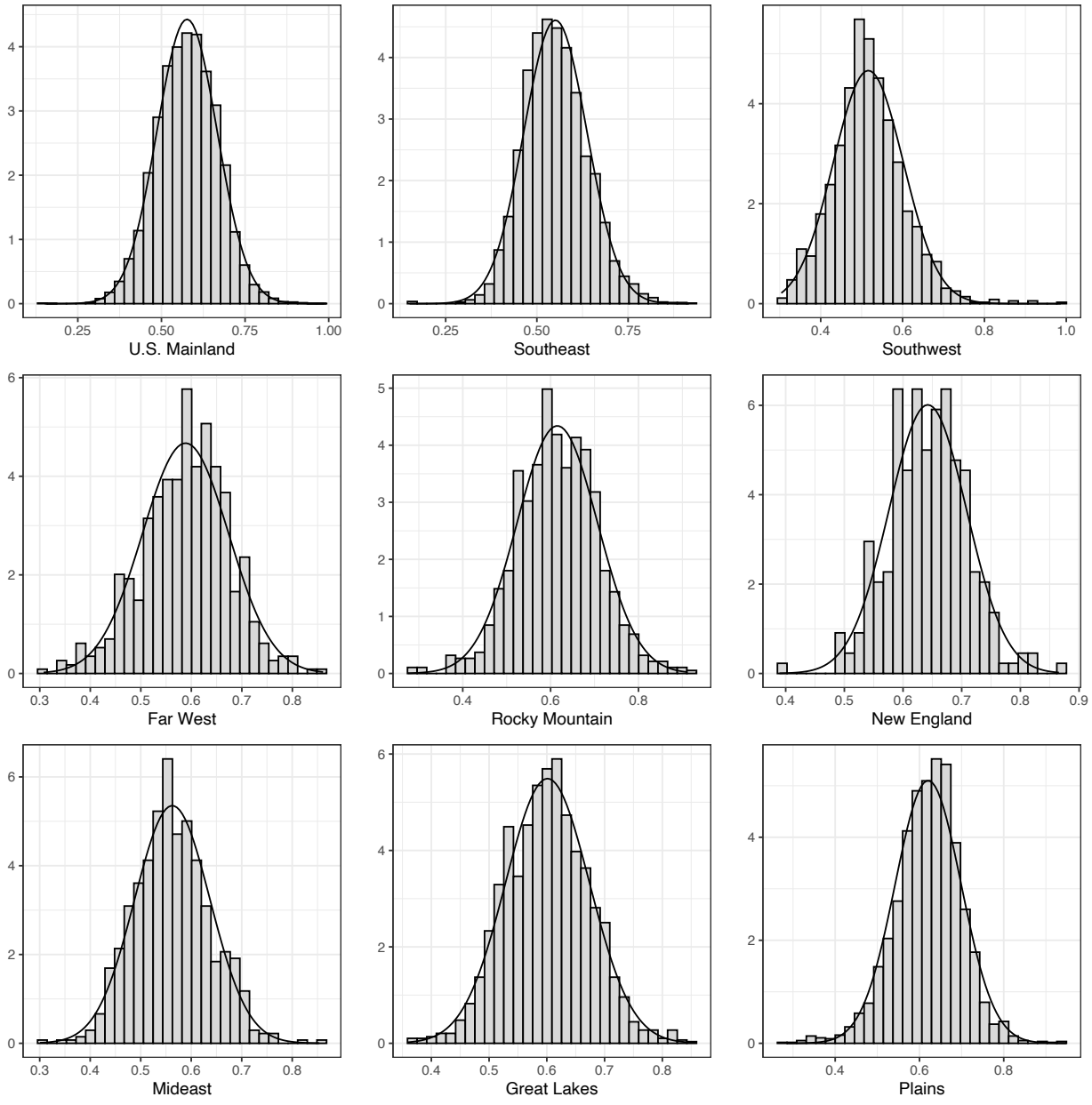


Figure S.3: Histogram of changes in Log Republican Odds Ratio ($DLRO$) over 2004-2016 at Mainland U.S. and Regional Levels

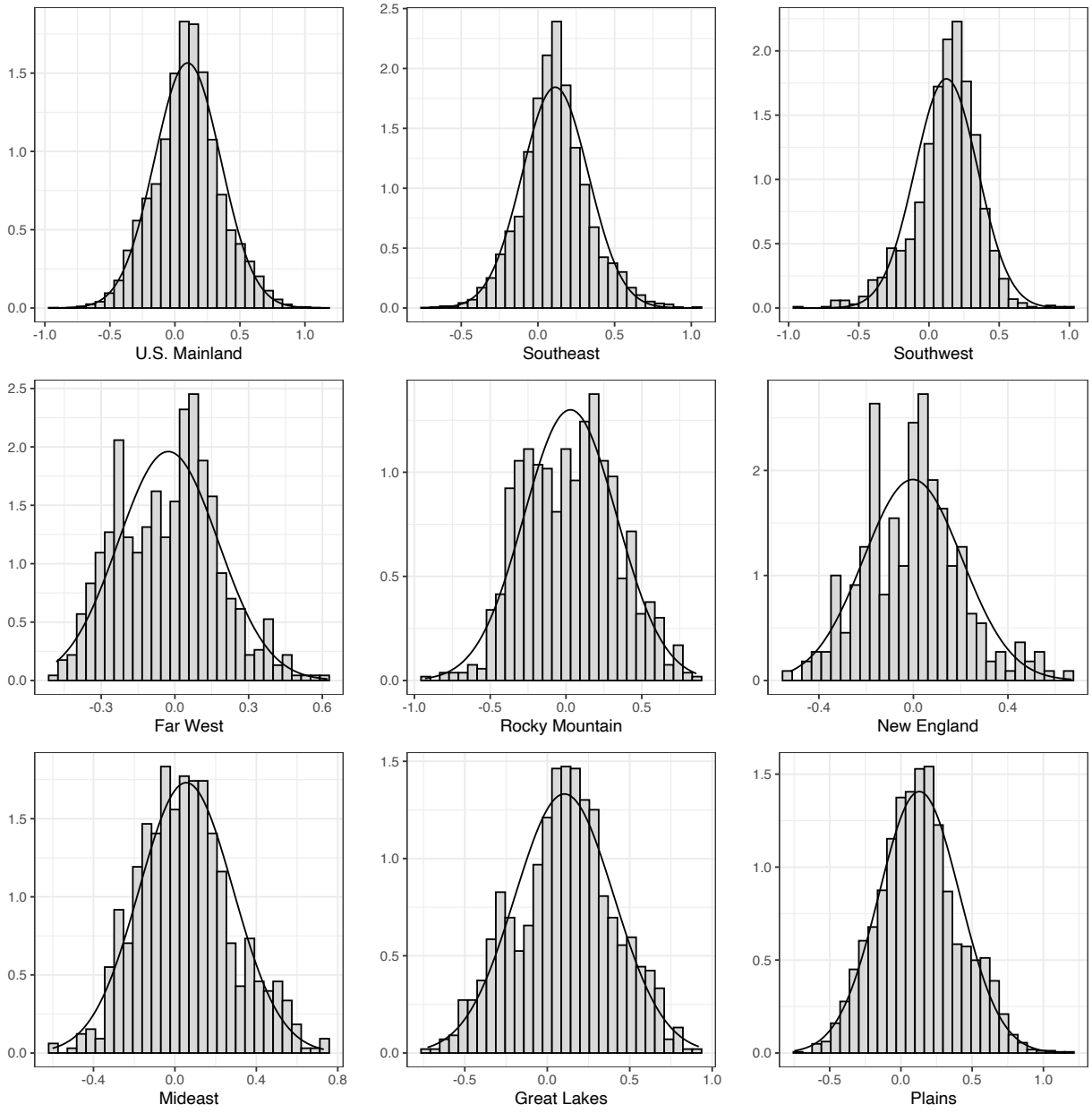
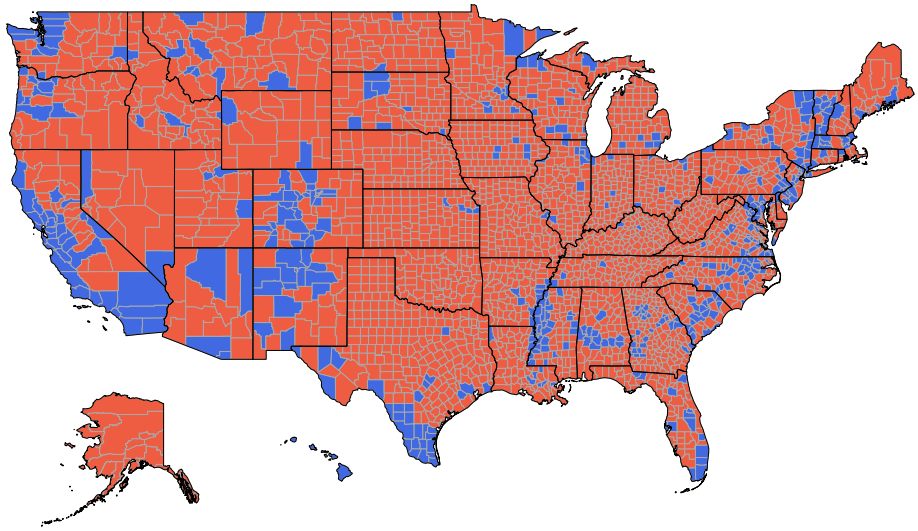
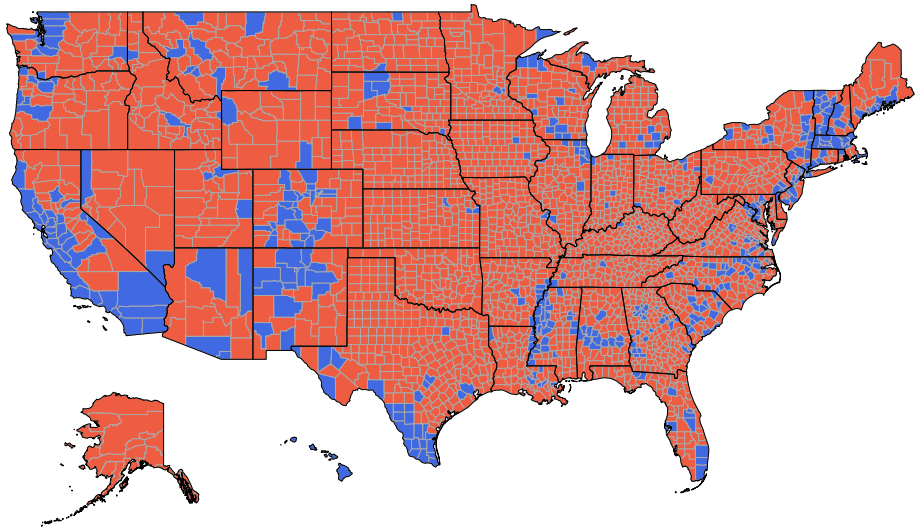


Figure S.4: 2020 Popular Vote Forecasts by County, Lasso-regional



Red indicates Republican electoral victory. Blue indicates Democratic popular victory.

Figure S.5: 2020 Popular vote Forecasts by County, OCMT-regional



Red indicates Republican electoral victory. Blue indicates Democratic popular victory.

Table S.1: State Level Forecasts of Republican Vote Shares (V_s) and Electoral Votes using Lasso Algorithm for 2020 Elections

		Pooled Forecasts		Regional Forecasts	
State	d_s	$\widehat{V}_s$	EC Votes	$\widehat{V}_s$	EC Votes
AK	3	-	3	-	3
AL	9	0.62	9	0.63	9
AR	6	0.62	6	0.65	6
AZ	11	0.47	0	0.55	11
CA	55	0.29	0	0.31	0
CO	9	0.40	0	0.41	0
CT	7	0.36	0	0.50	7
DC	3	0.03	0	0.04	0
DE	3	0.39	0	0.42	0
FL	29	0.46	0	0.46	0
GA	16	0.49	0	0.50	16
HI	4	-	0	-	0
IA	6	0.52	6	0.57	6
ID	4	0.66	4	0.65	4
IL	20	0.37	0	0.43	0
IN	11	0.58	11	0.58	11
KS	6	0.57	6	0.58	6
KY	8	0.63	8	0.64	8
LA	8	0.57	8	0.58	8
MA	11	0.29	0	0.43	0
MD	10	0.32	0	0.34	0
ME	4	0.44	0	0.53	4
MI	16	0.46	0	0.54	16
MN	10	0.44	0	0.47	0
MO	10	0.58	10	0.63	10
MS	6	0.57	6	0.58	6
MT	3	0.56	3	0.57	3
NC	15	0.48	0	0.48	0
ND	3	0.66	3	0.70	3
NE	5	0.60	5	0.64	5
NH	4	0.44	0	0.55	4
NJ	14	0.36	0	0.42	0
NM	5	0.40	0	0.47	0
NV	6	0.45	0	0.46	0
NY	29	0.35	0	0.35	0
OH	18	0.51	18	0.56	18
OK	7	0.65	7	0.68	7
OR	7	0.40	0	0.41	0
PA	20	0.46	0	0.50	0
RI	4	0.36	0	0.51	4
SC	9	0.54	9	0.55	9
SD	3	0.63	3	0.64	3
TN	11	0.62	11	0.63	11
TX	38	0.49	0	0.51	38
UT	6	0.59	6	0.60	6
VA	13	0.42	0	0.41	0
VT	3	0.31	0	0.39	0
WA	12	0.37	0	0.39	0
WI	10	0.47	0	0.52	10
WV	5	0.70	5	0.71	5
WY	3	0.71	3	0.73	3
All Votes	538		150		260

Republican vote shares are calculated as in Equation 12. Column d_s refers to the total number of electoral votes per state (Equation 7). EC Votes refer to the predicted number of Republican electoral college votes. All Votes accumulates U.S. Mainland electoral college votes, and assumes Hawaii casts her electoral votes for the Democratic candidate and Alaska casts her electoral votes for the Republican candidate. Regional forecasts are generated using the eight separate panel regressions for the eight BEA regions.

Table S.2: State Level Forecasts of Republican Vote Shares (V_s) and Electoral Votes using OCMT Algorithm for 2020 Elections

		Pooled Forecasts		Regional Forecasts	
State	d_s	$\widehat{V}_s$	EC Votes	$\widehat{V}_s$	EC Votes
AK	3	-	3	-	3
AL	9	0.64	9	0.65	9
AR	6	0.64	6	0.67	6
AZ	11	0.50	11	0.57	11
CA	55	0.33	0	0.34	0
CO	9	0.42	0	0.42	0
CT	7	0.39	0	0.49	0
DC	3	0.03	0	0.04	0
DE	3	0.41	0	0.48	0
FL	29	0.48	0	0.48	0
GA	16	0.51	16	0.52	16
HI	4	-	0	-	0
IA	6	0.53	6	0.61	6
ID	4	0.68	4	0.65	4
IL	20	0.39	0	0.43	0
IN	11	0.60	11	0.59	11
KS	6	0.57	6	0.60	6
KY	8	0.65	8	0.67	8
LA	8	0.59	8	0.61	8
MA	11	0.32	0	0.41	0
MD	10	0.33	0	0.38	0
ME	4	0.46	0	0.50	0
MI	16	0.50	0	0.52	16
MN	10	0.46	0	0.52	10
MO	10	0.60	10	0.66	10
MS	6	0.60	6	0.61	6
MT	3	0.59	3	0.57	3
NC	15	0.50	0	0.51	15
ND	3	0.68	3	0.75	3
NE	5	0.60	5	0.70	5
NH	4	0.47	0	0.51	4
NJ	14	0.40	0	0.46	0
NM	5	0.42	0	0.48	0
NV	6	0.50	6	0.47	0
NY	29	0.36	0	0.35	0
OH	18	0.54	18	0.55	18
OK	7	0.68	7	0.70	7
OR	7	0.42	0	0.43	0
PA	20	0.49	0	0.56	20
RI	4	0.38	0	0.48	0
SC	9	0.57	9	0.58	9
SD	3	0.64	3	0.66	3
TN	11	0.64	11	0.65	11
TX	38	0.52	38	0.54	38
UT	6	0.61	6	0.60	6
VA	13	0.43	0	0.44	0
VT	3	0.33	0	0.37	0
WA	12	0.40	0	0.42	0
WI	10	0.49	0	0.51	10
WV	5	0.73	5	0.74	5
WY	3	0.74	3	0.74	3
All Votes	538		221		290

Republican vote shares are calculated as in Equation 12. Column d_s refers to the total number of electoral votes per state (Equation 7). EC Votes refer to the predicted number of Republican electoral college votes. All Votes accumulates U.S. Mainland electoral college votes, and assumes Hawaii casts her electoral votes for the Democratic candidate and Alaska casts her electoral votes for the Republican candidate. Regional forecasts are generated using the eight separate panel regressions for the eight BEA regions.

Table S.3: State Level Forecasts and Realized Republican Vote Shares (V_s) and Electoral Votes using Lasso-OCMT Average for 2016 Elections

			Pooled Forecasts		Regional Forecasts	
State	d_s	Realized	$\widehat{V}_s$	EC Votes	$\widehat{V}_s$	EC Votes
AK	3	-	-	3	-	3
AL	9	0.64	0.63	9	0.65	9
AR	6	0.64	0.65	6	0.67	6
AZ	11	0.52	0.56	11	0.54	11
CA	55	0.34	0.39	0	0.41	0
CO	9	0.47	0.48	0	0.53	9
CT	7	0.43	0.41	0	0.44	0
DC	3	0.04	0.07	0	0.08	0
DE	3	0.44	0.40	0	0.42	0
FL	29	0.51	0.51	29	0.52	29
GA	16	0.53	0.56	16	0.57	16
HI	4	-	-	0	-	0
IA	6	0.55	0.50	6	0.51	6
ID	4	0.68	0.69	4	0.72	4
IL	20	0.41	0.43	0	0.47	0
IN	11	0.60	0.58	11	0.61	11
KS	6	0.61	0.63	6	0.66	6
KY	8	0.66	0.64	8	0.66	8
LA	8	0.60	0.61	8	0.62	8
MA	11	0.35	0.37	0	0.42	0
MD	10	0.36	0.37	0	0.38	0
ME	4	0.49	0.44	0	0.43	0
MI	16	0.50	0.47	0	0.51	16
MN	10	0.49	0.49	0	0.50	10
MO	10	0.62	0.59	10	0.62	10
MS	6	0.59	0.58	6	0.59	6
MT	3	0.61	0.59	3	0.63	3
NC	15	0.52	0.53	15	0.53	15
ND	3	0.70	0.64	3	0.63	3
NE	5	0.64	0.64	5	0.65	5
NH	4	0.50	0.48	0	0.50	0
NJ	14	0.43	0.41	0	0.43	0
NM	5	0.45	0.45	0	0.44	0
NV	6	0.49	0.50	0	0.51	6
NY	29	0.38	0.34	0	0.36	0
OH	18	0.54	0.51	18	0.55	18
OK	7	0.69	0.69	7	0.68	7
OR	7	0.44	0.46	0	0.48	0
PA	20	0.50	0.48	0	0.51	20
RI	4	0.42	0.35	0	0.39	0
SC	9	0.57	0.58	9	0.58	9
SD	3	0.66	0.62	3	0.64	3
TN	11	0.64	0.63	11	0.66	11
TX	38	0.55	0.60	38	0.57	38
UT	6	0.62	0.76	6	0.79	6
VA	13	0.47	0.49	0	0.47	0
VT	3	0.35	0.33	0	0.30	0
WA	12	0.41	0.44	0	0.47	0
WI	10	0.50	0.49	0	0.51	10
WV	5	0.72	0.66	5	0.68	5
WY	3	0.76	0.73	3	0.75	3
All Votes	538			259		330

The average forecast takes the predicted number of Democrat and Republican votes under OCMT and Lasso for each county and averages them. Republican vote shares are calculated as in Equation 12. Column d_s refers to the total number of electoral votes per state (Equation 7). EC Votes refer to the predicted number of Republican electoral college votes. All Votes accumulates U.S. Mainland electoral college votes, and assumes Hawaii casts her electoral votes for the Democratic candidate and Alaska casts her electoral votes for the Republican candidate. Regional forecasts are generated using the eight separate panel regressions for the eight BEA regions.

Table S.4: State Level Forecasts of Republican Vote Shares (V_s) and Electoral Votes using Lasso-OCMT Average for 2020 Elections

		Pooled Forecasts		Regional Forecasts	
State	d_s	$\widehat{V}_s$	EC Votes	$\widehat{V}_s$	EC Votes
AK	3	-	3	-	3
AL	9	0.63	9	0.64	9
AR	6	0.63	6	0.66	6
AZ	11	0.49	0	0.56	11
CA	55	0.31	0	0.32	0
CO	9	0.41	0	0.41	0
CT	7	0.37	0	0.49	0
DC	3	0.03	0	0.04	0
DE	3	0.40	0	0.45	0
FL	29	0.47	0	0.47	0
GA	16	0.50	0	0.51	16
HI	4	-	0	-	0
IA	6	0.52	6	0.59	6
ID	4	0.67	4	0.65	4
IL	20	0.38	0	0.43	0
IN	11	0.59	11	0.59	11
KS	6	0.57	6	0.59	6
KY	8	0.64	8	0.66	8
LA	8	0.58	8	0.59	8
MA	11	0.30	0	0.42	0
MD	10	0.32	0	0.36	0
ME	4	0.45	0	0.51	4
MI	16	0.48	0	0.53	16
MN	10	0.45	0	0.50	0
MO	10	0.59	10	0.65	10
MS	6	0.58	6	0.59	6
MT	3	0.58	3	0.57	3
NC	15	0.49	0	0.50	0
ND	3	0.67	3	0.72	3
NE	5	0.60	5	0.67	5
NH	4	0.45	0	0.53	4
NJ	14	0.38	0	0.44	0
NM	5	0.41	0	0.47	0
NV	6	0.48	0	0.47	0
NY	29	0.35	0	0.35	0
OH	18	0.52	18	0.55	18
OK	7	0.67	7	0.69	7
OR	7	0.41	0	0.42	0
PA	20	0.48	0	0.53	20
RI	4	0.37	0	0.50	0
SC	9	0.56	9	0.57	9
SD	3	0.64	3	0.65	3
TN	11	0.63	11	0.64	11
TX	38	0.51	38	0.53	38
UT	6	0.60	6	0.60	6
VA	13	0.43	0	0.43	0
VT	3	0.32	0	0.38	0
WA	12	0.38	0	0.40	0
WI	10	0.48	0	0.52	10
WV	5	0.72	5	0.73	5
WY	3	0.73	3	0.73	3
All Votes	538		188		269

The average forecast takes the predicted number of Democrat and Republican votes under OCMT and Lasso for each county and averages them. Republican vote shares are calculated as in Equation 12. Column d_s refers to the total number of electoral votes per state (Equation 7). EC Votes refer to the predicted number of Republican electoral college votes. All Votes accumulates U.S. Mainland electoral college votes, and assumes Hawaii casts her electoral votes for the Democratic candidate and Alaska casts her electoral votes for the Republican candidate. Regional forecasts are generated using the eight separate panel regressions for the eight BEA regions.

Table S.5: State Level Forecasts of Republican Vote Shares (V_s) and Electoral Votes using Lasso Algorithm for 2020 Elections with Data Available as of October 2020

		Pooled Forecasts		Regional Forecasts	
State	d_s	$\widehat{V}_s$	EC Votes	$\widehat{V}_s$	EC Votes
AK	3	-	3	-	3
AL	9	0.63	9	0.64	9
AR	6	0.63	6	0.65	6
AZ	11	0.49	0	0.55	11
CA	55	0.30	0	0.31	0
CO	9	0.41	0	0.41	0
CT	7	0.37	0	0.50	0
DC	3	0.03	0	0.04	0
DE	3	0.40	0	0.42	0
FL	29	0.46	0	0.46	0
GA	16	0.49	0	0.51	16
HI	4	-	0	-	0
IA	6	0.52	6	0.55	6
ID	4	0.67	4	0.65	4
IL	20	0.38	0	0.42	0
IN	11	0.58	11	0.58	11
KS	6	0.58	6	0.58	6
KY	8	0.64	8	0.65	8
LA	8	0.58	8	0.58	8
MA	11	0.29	0	0.41	0
MD	10	0.31	0	0.35	0
ME	4	0.45	0	0.51	4
MI	16	0.47	0	0.51	16
MN	10	0.45	0	0.47	0
MO	10	0.59	10	0.62	10
MS	6	0.58	6	0.58	6
MT	3	0.57	3	0.57	3
NC	15	0.49	0	0.49	0
ND	3	0.66	3	0.69	3
NE	5	0.60	5	0.64	5
NH	4	0.44	0	0.53	4
NJ	14	0.37	0	0.41	0
NM	5	0.41	0	0.47	0
NV	6	0.47	0	0.46	0
NY	29	0.34	0	0.34	0
OH	18	0.52	18	0.54	18
OK	7	0.67	7	0.68	7
OR	7	0.40	0	0.41	0
PA	20	0.47	0	0.50	0
RI	4	0.38	0	0.49	0
SC	9	0.56	9	0.56	9
SD	3	0.64	3	0.63	3
TN	11	0.62	11	0.64	11
TX	38	0.50	38	0.51	38
UT	6	0.60	6	0.60	6
VA	13	0.42	0	0.41	0
VT	3	0.32	0	0.38	0
WA	12	0.37	0	0.38	0
WI	10	0.47	0	0.51	10
WV	5	0.71	5	0.72	5
WY	3	0.73	3	0.73	3
All Votes	538		188		249

Republican vote shares are calculated as in Equation 12. Column d_s refers to the total number of electoral votes per state (Equation 7). EC Votes refer to the predicted number of Republican electoral college votes. All Votes accumulates U.S. Mainland electoral college votes, and assumes Hawaii casts her electoral votes for the Democratic candidate and Alaska casts her electoral votes for the Republican candidate. Regional forecasts are generated using the eight separate panel regressions for the eight BEA regions. Using data available as of October 14, 2020.

Table S.6: State Level Forecasts of Republican Vote Shares (V_s) and Electoral Votes using OCMT Algorithm for 2020 Elections with Data Available as of October 2020

		Pooled Forecasts		Regional Forecasts	
State	d_s	$\widehat{V}_s$	EC Votes	$\widehat{V}_s$	EC Votes
AK	3	-	3	-	3
AL	9	0.64	9	0.65	9
AR	6	0.65	6	0.67	6
AZ	11	0.52	11	0.57	11
CA	55	0.34	0	0.34	0
CO	9	0.42	0	0.42	0
CT	7	0.40	0	0.48	0
DC	3	0.03	0	0.04	0
DE	3	0.42	0	0.47	0
FL	29	0.49	0	0.48	0
GA	16	0.51	16	0.52	16
HI	4	-	0	-	0
IA	6	0.54	6	0.58	6
ID	4	0.68	4	0.66	4
IL	20	0.40	0	0.42	0
IN	11	0.60	11	0.58	11
KS	6	0.58	6	0.59	6
KY	8	0.66	8	0.67	8
LA	8	0.60	8	0.61	8
MA	11	0.32	0	0.39	0
MD	10	0.33	0	0.37	0
ME	4	0.47	0	0.48	0
MI	16	0.50	0	0.50	0
MN	10	0.47	0	0.50	10
MO	10	0.61	10	0.65	10
MS	6	0.60	6	0.61	6
MT	3	0.59	3	0.57	3
NC	15	0.50	15	0.51	15
ND	3	0.69	3	0.73	3
NE	5	0.61	5	0.69	5
NH	4	0.47	0	0.50	0
NJ	14	0.41	0	0.45	0
NM	5	0.44	0	0.48	0
NV	6	0.50	6	0.46	0
NY	29	0.37	0	0.34	0
OH	18	0.54	18	0.54	18
OK	7	0.69	7	0.70	7
OR	7	0.42	0	0.42	0
PA	20	0.50	0	0.55	20
RI	4	0.39	0	0.46	0
SC	9	0.57	9	0.58	9
SD	3	0.65	3	0.64	3
TN	11	0.64	11	0.65	11
TX	38	0.53	38	0.54	38
UT	6	0.61	6	0.61	6
VA	13	0.44	0	0.44	0
VT	3	0.33	0	0.36	0
WA	12	0.40	0	0.41	0
WI	10	0.50	0	0.51	10
WV	5	0.74	5	0.74	5
WY	3	0.75	3	0.74	3
All Votes	538		236		270

Republican vote shares are calculated as in Equation 12. Column d_s refers to the total number of electoral votes per state (Equation 7). EC Votes refer to the predicted number of Republican electoral college votes. All Votes accumulates U.S. Mainland electoral college votes, and assumes Hawaii casts her electoral votes for the Democratic candidate and Alaska casts her electoral votes for the Republican candidate. Regional forecasts are generated using the eight separate panel regressions for the eight BEA regions. Using data available as of October 14, 2020.

Table S.7: State Level Forecasts of Republican Vote Shares (V_s) and Electoral Votes using Lasso-OCMT Average for 2020 Elections using Data Available as of October 2020

		Pooled Forecasts		Regional Forecasts	
State	d_s	$\widehat{V}_s$	EC Votes	$\widehat{V}_s$	EC Votes
AK	3	-	3	-	3
AL	9	0.63	9	0.65	9
AR	6	0.64	6	0.66	6
AZ	11	0.51	11	0.56	11
CA	55	0.32	0	0.33	0
CO	9	0.41	0	0.41	0
CT	7	0.38	0	0.49	0
DC	3	0.03	0	0.04	0
DE	3	0.41	0	0.45	0
FL	29	0.48	0	0.47	0
GA	16	0.50	16	0.52	16
HI	4	-	0	-	0
IA	6	0.53	6	0.56	6
ID	4	0.67	4	0.65	4
IL	20	0.39	0	0.42	0
IN	11	0.59	11	0.58	11
KS	6	0.58	6	0.58	6
KY	8	0.65	8	0.66	8
LA	8	0.59	8	0.60	8
MA	11	0.30	0	0.40	0
MD	10	0.32	0	0.36	0
ME	4	0.46	0	0.49	0
MI	16	0.49	0	0.50	16
MN	10	0.46	0	0.48	0
MO	10	0.60	10	0.64	10
MS	6	0.59	6	0.60	6
MT	3	0.58	3	0.57	3
NC	15	0.49	0	0.50	0
ND	3	0.67	3	0.71	3
NE	5	0.61	5	0.67	5
NH	4	0.46	0	0.51	4
NJ	14	0.39	0	0.43	0
NM	5	0.43	0	0.47	0
NV	6	0.48	0	0.46	0
NY	29	0.35	0	0.34	0
OH	18	0.53	18	0.54	18
OK	7	0.68	7	0.69	7
OR	7	0.41	0	0.42	0
PA	20	0.48	0	0.52	20
RI	4	0.38	0	0.48	0
SC	9	0.57	9	0.57	9
SD	3	0.65	3	0.64	3
TN	11	0.63	11	0.65	11
TX	38	0.52	38	0.53	38
UT	6	0.61	6	0.60	6
VA	13	0.43	0	0.43	0
VT	3	0.33	0	0.37	0
WA	12	0.38	0	0.40	0
WI	10	0.48	0	0.51	10
WV	5	0.72	5	0.73	5
WY	3	0.74	3	0.73	3
All Votes	538		215		265

The average forecast takes the predicted number of Democrat and Republican votes under OCMT and Lasso for each county and averages them. Republican vote shares are calculated as in Equation 12. Column d_s refers to the total number of electoral votes per state (Equation 7). EC Votes refer to the predicted number of Republican electoral college votes. All Votes accumulates U.S. Mainland electoral college votes, and assumes Hawaii casts her electoral votes for the Democratic candidate and Alaska casts her electoral votes for the Republican candidate. Regional forecasts are generated using the eight separate panel regressions for the eight BEA regions. Using data available as of October 14, 2020.

Table S.8: Functional Form of Voting Outcome Variable Regressed on Active Set

	Dependent variable:		
	$\Delta_4 \ln \frac{V_{cr,t}}{1-V_{cr,t}}$	$\Delta_4 V_{cr,t}$	$\Delta_4 \ln V_{cr,t}$
	(1)	(2)	(3)
Adjusted SE	0.036	0.037	0.042
Observations	12,428	12,428	12,428
Adjusted R^2	0.537	0.530	0.492

County Republican vote share, $V_{cr,t}$ is defined as in Equation S.5. Regression fits under different dependent variable transformations are compared using adjusted regression standard errors reported in the row named Adjusted SE. Adjustments made based on different functional forms are described in Section S2.

Table S.9: State and County Sample

	State	Counties
1	AK	-
2	AL	67
3	AR	75
4	AZ	15
5	CA	58
6	CO	63
7	CT	8
8	DC	1
9	DE	3
10	FL	67
11	GA	159
12	HI	-
13	IA	99
14	ID	44
15	IL	102
16	IN	92
17	KS	105
18	KY	120
19	LA	64
20	MA	14
21	MD	24
22	ME	16
23	MI	83
24	MN	87
25	MO	115
26	MS	82
27	MT	56
28	NC	100
29	ND	53
30	NE	93
31	NH	10
32	NJ	21
33	NM	33
34	NV	17
35	NY	62
36	OH	88
37	OK	77
38	OR	36
39	PA	67
40	RI	5
41	SC	46
42	SD	66
43	TN	95
44	TX	254
45	UT	29
46	VA	133
47	VT	14
48	WA	39
49	WI	72
50	WV	55
51	WY	23
	Total	3107

We do not consider Alaska and Hawaii, non U.S. mainland states, in our sample. "DC" refers to Washington D.C.

References

Chudik, A., G. Kapetanios, and M. H. Pesaran (2018). A one covariate at a time, multiple testing approach to variable selection in high-dimensional linear regression models. *Econometrica*, 86,1479-1512.

Chudik, A., M. H. Pesaran, and M. Sharifvaghefi (2020), Variable selection and forecasting in high dimensional linear regressions with structural breaks, Globalization Institute Working Paper 394, Federal Reserve Bank of Dallas.

Pesaran, M. H. (2015) *Time Series and Panel Data Econometrics.* Oxford University Press, Oxford. URL https: //ideas.repec.org/b/oxp/obooks/9780198759980.html.