Regional Heterogeneity and U.S. Presidential Elections: Real-Time 2020 Forecasts and Evaluation*

Rashad Ahmed† M. Hashem Pesaran‡

June 23, 2021
First Version: October 3, 2020
Forecasts Formed: October 14, 2020

Abstract

This paper exploits cross-sectional variation at the level of U.S. counties to generate real-time forecasts for the 2020 U.S. presidential election. The forecasting models are trained on data covering the period 2000-2016, using high-dimensional variable selection techniques. Our county-based approach contrasts the literature that focuses on national and state level data but uses longer time periods to train their models. The paper reports forecasts of popular and electoral college vote outcomes and provides a detailed *ex post* evaluation of the forecasts released in real time prior to the election. It is shown that all of these forecasts outperform autoregressive benchmarks, with a pooled national model using One-Covariate-at-a-time-Multiple-Testing (OCMT) variable selection significantly outperforming all models in forecasting both the U.S. mainland national vote share and electoral college outcomes (forecasting 236 electoral votes for the Republican party compared to 232 realized). This paper also shows that key determinants of voting outcomes at the county level include incumbency effects, unemployment, poverty, educational attainment, house price changes, and international competitiveness. The results are also supportive of myopic voting: economic fluctuations realized a few months before the election tend to be more powerful predictors of voting outcomes than their long-horizon analogues.

Keywords: Real-time Forecasts, Popular and Electoral College Votes, Simultaneity, High Dimensional Forecasting Models, Lasso, One Covariate at a time Multiple Testing, OCMT.

JEL Classifications: C53, C55, D72

---

[*] Authors are grateful to Ron Smith and two anonymous reviewers for their detailed and constructive comments. Rashad Ahmed acknowledges partial financial support from USC Dornsife INET. The first draft of this paper was distributed as a CESifo working paper on October 3rd, 2020. Forecasts reported in this paper were formed on October 14th, 2020, and made available as revised CESifo WP 8615 at https://www.cesifo.org/en/publikationen/2020/working-paper/regional-heterogeneity-and-us-presidential-elections.

[†] University of Southern California, USA, e-mail: rashadah@usc.edu.

[‡] University of Southern California, USA, and Trinity College, Cambridge, UK, e-mail: pesaran@usc.edu.

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 potentially related to rising political polarization (Sides et al. [2017], Gelman and Azari [2017]). 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. In doing so, we produce real-time forecasts made in October 2020 for the November 2020 election and also evaluate the 2016 election. 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 to 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

---

[1] In all four cases the Republican candidate lost the popular vote but won the electoral college vote.

mixed.

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 may be subject to structural breaks. Certainly the factors influencing voting behavior and the make-up of the voting body have changed since the 1950's and continue to evolve. In this paper we provide an alternative to time-series national models or state-level panel data models, and use county level electoral data which allow us to exploit the rich cross-sectional variation that exists in electoral and socioeconomic outcomes. But it is important to note that data on county-level election outcomes are only publicly available from 2000, and we are therefore constrained to four time-series observations – the period November 2000 to October 2020 covers 5 election cycles, but the first cycle is used up as initial values.

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 generated forecasts for the 2020 election using two sets of models: a single pooled-county level panel data model, and eight regional models to allow for possible regional heterogeneity by estimating separate county-level panel regressions for the eight BEA regions. 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 fundamentals-based election modeling. We consider economic variables such as local unemployment, income, house prices, government employment and healthcare expenditures. We further distinguish between the effects of short-term and longer term economic conditions, allowing for the possibility of voter myopia. There is a sizable literature arguing that changes in economic conditions closer to the election bear more influence on voting behavior compared to longer-term economic conditions. See, for example, Kramer [1971], Alesina et al. [1995], Wlezien and Erikson [1996], Achen and Bartels [2004], and Wlezien [2015]. 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. [2020] and Jensen et al. [2017], we also consider 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.

To generate real-time forecasts we begun by estimating the pooled and regional models over the 2000-2016 training sample. Our analysis corroborated the usefulness of several variables identified in the literature as important in explaining voting outcomes. At the same time, we investigated the extent of regional heterogeneity and their effects on forecasts and associated forecast errors over the training sample. We found that important factors explaining voting behavior included voter turnout, local economic performance, unemployment, poverty rates, education, house price changes, and urban-rural mix. Our results also corroborated evidence supporting incumbency effects and voter myopia. Specifically, we find that economic fluctuations realized a few months prior to the election are more effective predictors of voting outcomes than their long-horizon analogues.

Based on data available at the time of forecasting (October 14, 2020), and following the variable selection strategy used over 2000-2016 sample, we generated forecasts for the November 2020 U.S. presidential election under different pooled and regional model specifications. We generated four main forecasts, based on models selected using pooled-OCMT, regional-OCMT, pooled-Lasso and regional-Lasso procedures. We also consider average Lasso-OCMT pooled and regional forecasts for a total of 6 forecasts. The final count for the November 2020 Presidential election resulted in a Democratic victory with Democrat Joe Biden winning the popular vote along with 306 electoral votes, and Donald Trump winning the remaining 232 electoral votes. Interestingly, all models forecasted that the two-party U.S. mainland national Republican vote share (similar to the popular vote share excluding Hawaii and Alaska and third party candidates) would be won by the Democratic party. The pooled-OCMT model produced the most accurate two-party national Republican vote share forecast of 47.6% compared to the realized outcome of 47.7%. Moreover, all but one model predicted less than 270 electoral votes for the Republican party implying a Democrat candidate electoral victory, although individual models varied substantially in their forecasts of the exact number of electoral votes. Only one of models predicted a Republican victory with 270 electoral votes, winning by a razor thin margin of a single electoral vote! These results starkly contrast forecasts from naive random walk and autoregressive model specifications which predicted a Republican victory with 329 electoral votes.

In terms of total electoral votes, the pooled-OCMT model forecasts performed best, forecasting 236 Republican electoral votes compared to the 232 realized. Statistical analysis of the forecasts suggest that all pooled and regional models significantly outperformed naive forecasts for the 2020 election and that the pooled-OCMT forecasts outperformed the other model forecasts: pooled-Lasso, regional-Lasso, regional-OCMT. The pooled-Lasso specification predicted state-level winners best, mis-predicting 2 out of the 48 mainland states plus D.C. We also provide forecast performance and evaluation of the 2016 election using models trained over the 2000-2012 sample. Here, we find that regional models would have correctly predicted a 2016 Republican victory.

The rest of this paper is organized as follows. Section 2 presents our modeling approach and its relation to the literature. Section 3 characterizes the two-stage model of voter turnout and election outcomes. Section 4 discusses our identification procedure to consistently estimate the model. Section 5 provides an overview of the data used in the analysis. Section 6 discusses the choice of the variables in 'active sets', and Section 7 describes the variable selection techniques and their application to our forecasting problem. Section 8 describes the U.S. Electoral College process from which we generate election forecasts using county level predictions. Section 9 investigates key determinants of U.S. election cycles over the

2000-2016 period. Section 10 then uses models trained over the 2000-2016 sample to generate forecasts for the 2020 election along with forecast evaluations. Section 11 concludes. Detail of data sources, variable selection algorithms, and a weighted version of the Diebold-Mariano test used in forecast evaluation exercise are provided in the Appendix. In an online supplement we provide additional results primarily dealing with the 2016 election forecasts obtained using 2000-2012 period as the training sample. We also report forecast results for 2020 election using an alternative regional classification and an extended set of covariates. These *ex post* results are generally in line with our *ex ante* forecasts.[2]

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.[3] See, for example, Kou and Sobel [2004], Biemer [2010], Shirani-Mehr et al. [2018], Graefe [2018].[4] 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], Jérôme et al. [2020], Zandi et al. [2020] and Enns and Lagodny [2021] 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

---

[2] This *ex post* analyses were carried out on recommendation of one of the reviewers.

[3] 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..

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

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.

As already explained in the Introduction, in order to deal with the heterogeneity and possible model instability, we exploit variations in electoral and socioeconomic outcomes across the 3,107 mainland U.S. counties instead of using national or state-level models that rely mainly on long time-series data. But currently there is an important drawback to using county level election data, since such data are publicly available only from 2000 (Bush-Gore election). This data limitation also prevents us from state-level modeling that allows the effects of socioeconomic factors to differ across all the 48 mainland states.[5] 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. While we reduce the risk of structural breaks with the short time dimension, a limitation of this approach is its inability to take advantage of time-series variations in the national-level covariates, such as output growth and stock returns. We partially address this issue by employing several economic indicators which vary at the county level. Because the variation in these measures can be decomposed into a national and county-specific component, the model is able to implicitly incorporate national-level fluctuations in economic conditions. However, our modeling approach imposes the restriction that the coefficients on national and county-specific variation are equal.

Furthermore, counties across state borders often share similar features such that estimation could be made more efficient by pooling information from such neighboring states. As noted earlier, we address these challenges by grouping the states into eight regions defined by the BEA, and estimate eight separate regional panel regressions. In this way we hope to strike a balance between allowing for 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.[6] We can, therefore, capture possible regional differences in voting preferences and, more generally, differences in demographic, social, and economic heterogeneity across the United States.

---

[5] 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.

[6] 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.

The pooled and regional models are used to generate predictions for 3,107 counties for a given election year. These predictions are aggregated to generate state level and national level popular vote predictions, as well as electoral college vote predictions. One limitation of our region-based modeling approach is that while information across counties and states within a region are pooled, regions themselves are treated as separate (political) entities. However this assumption may not be satisfied particularly for those counties which are adjacent but lying in different regions. Alternative modeling approaches could potentially allow further for dependencies between regions as well, at the cost of imposing further assumptions. To further address the sensitivity of this issue we examine forecasts generated under a different regional classification as a robustness check, to be reported in the online supplement.

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. [2020] and Jensen et al. [2017]. Scala and Johnson [2017] identify large differences in voting preferences across the urban-rural spectrum in elections from 2000 to 2016. In a cross-sectional study, Kahane [2020] shows that the urban-rural spectrum, poverty rates, education, among several other demographic factors, shaped 2012 and 2016 election outcomes. Like these studies, we exploit 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 OCMT procedure proposed in Chudik et al. [2018], respectively. See Section S2 of the Appendix for further details.

3 Modeling Turnout and Election Outcomes

3.1 Voter turnout

One novel departure of our modeling strategy from the prevailing literature is the joint modeling 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.[7] 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 Campbell et al. [1960], Powell [1986] and Jackman [1987]. The latter two studies look at cross-country voting patterns and uncover a similar theme where countries with greater institutional quality also have higher voter turnouts.[8] 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].[9] The likelihood of voting has even been linked to genetics (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 requirements 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]). Recent research on voter turnout which is particularly

---

[7] 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.

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

[9] 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.

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 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 $t = 2004, 2008, 2012, 2016, 2020$, 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.[10] 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 $V_{cr,t}$.[11] 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 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 $a_{DLRO,r}$ are region-specific fixed effects, $\mathbf{z}_{DLRO,cr}$ denote the vector of variables that vary across counties but not over time, and $\mathbf{x}_{DLRO,cr,t+3}$ are the remaining covariates that vary both across counties and over time, and are dated preceding the election. For the pooled

---

[10] The use of the log-odds ratio (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.

[11] 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 improves the model fit over the vote share, albeit marginally. These tests are done within sample, and details can be found in Section S1 of the online supplement.

models the distinction between the time-varying covariates, $\mathbf{x}_{DLRO,cr,t+3}$, and the covariates that do not vary over our sample (included in $\mathbf{z}_{DLRO,cr}$) is not consequential. But as we shall see the distinction between the types of variables become important when we consider regional models. In what follows, without loss of generality, we refer to $\mathbf{z}_{DLRO,cr}$ as the vector of time-invariant covariates, which includes variables such as education, religiosity, and rural/urban mix.[12] In our application, $t \in \{2000, 2004, 2008, 2012, 2016\}$ 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, $VT_{cr,t+4}$, which is defined by

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

which is equal to the total two-party votes as a proportion of the voting age population ($VAP_{cr,t}$) of county c in region r for election year t. $VAP_{cr,t}$ is reported as a 5-year average. Due to data availability, we use 2012-2016 voting age population 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 vote outcome and turnout equations, time-invariant factors can include socioeconomic and demographic factors exhibiting little or no time variation over the sample like education, migration, religiosity, and urban-rural mix. Time-varying factors include local unemployment rate, poverty rate, household median income, changes in house prices, government and private employment, among others.

Notice that (2) and (4) represent a system of simultaneous equations. Voting outcomes may depend on voter turnout, 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

---

[12] Technically speaking all variables are potentially time-varying, but some are either observed only once during our sample, or vary very slowly such that they can be viewed as time-invaraint.

$\varepsilon_{cr,t+4}$ 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 voter turnout and election outcomes. 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. There are, however, potential caveats. For instance, if election uncertainty is correlated with turnout, the recursive assumption would be violated. As a robustness check we also report and evaluate forecasts obtained (*ex post*) using a reduced form single equation model of $DLRO_{cr,t+4}$ which considers an extended set of covariates obtained as the union of the covariates used to model VT and $DLRO$.

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 $\widehat{VT}_{cr,t+4}$) are used as a regressor in the election outcome equation ($DLRO_{cr,t+4}$).[13] 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

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

In addition to estimating pooled models, we allow for regional heterogeneity in both equations – all coefficients are specific to region r by estimating eight separate regional panel regressions. 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 post variable selection, which will be addressed.

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. Recall that there are no publicly available presidential voting outcome data at the county-level before 2000. 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.[14] Each state (and hence county) falls into to one of the eight BEA regions. The list of states included in these regions is given in Table S.5 of the online supplement. Figures S.8 and S.9 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 U.S., as well as for the eight BEA regions.[15] These histograms provide a visual account of the degree of regional heterogeneity in VT and $DLRO$ variables which, as we

---

[13] A formal proof of consistency is provided in Section S2 of the online supplement.

[14] 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 Appendix.

[15] To save space additional result tables and figures are provided in the online supplement Section S6.

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 vary across counties or states, but either do not vary or show very little variation over the four election cycles that we are considering – we treat all such variables as time-invariant and use their time averages if more than one data point is available over our sample. These include measures on county demographics, education, religiosity, migration, population density, urban-rural mix, and vote-by-mail policy of the state. Time-varying measures vary at state or county levels and over time. 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 gross domestic product (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 Appendix 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, inclusive of county c. We consider both individual and local measures for many county-level variables such as migration and education. Local variables are denoted with a '*'. Hence, "Education" and "Education*" correspond to individual and local education rates, respectively. County house prices and unemployment rate variables 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, referred to as voter myopia, where voters typically care only about the past year's economic performance when evaluating the incumbent party's overall performance, rather than performance over the entire term. 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 voting behavior. 1-year (L1) and 3-month average (M3) unemployment rates will be denoted by "Unemployment (L1)" and "Unemployment (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 are the 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 Appendix.

6 Active Sets for Voter Turnout (VT) and Changes in Log Republican Odds ($DLRO$) Panel Regressions

As is clear from the above account, there are many covariates that can be considered as potential predictors of voter turnout (VT) and changes in the log-odds ratio of Republican-to-Democrat votes ($DLRO$) 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 model. The choice of potential covariates is largely inspired by the literature. We also account for temporal effects, again inspired by the literature, documenting myopia or 'short-memory' among voters.

Table 1: Summary Statistics for the Covariates in the Active Set for Voter Turnout (VT) Panel Regressions over the Period 2000-2016

Covariate	Description	Mean	St. Dev.	Regional Coverage
Incumbent party	indicator taking 1 if incumbent party is Republican, -1 if incumbent party is Democrat	0.000	1.000	National
Incumbent president	indicator taking 1 if Republican re-election, -1 if Democratic re-election, 0 if no re-election	0.000	0.707	National
Lagged voter turnout (VT)	voter turnout proportion from the preceding election	0.564	0.097	County
Lagged VT × incumbent party	Lagged VT interacted with incumbent party indicator	0.015	0.583	County
Healthcare costs (L1)	change in log healthcare expenditures, year preceding election	0.046	0.016	State
Government employment (L1)	change in log government employment, year preceding election	-0.012	0.015	State
Unemployment (L1)	unemployment rate avg., year preceding election	0.061	0.020	County
House price (L1)	change in log house prices avg., year preceding election	0.022	0.043	County
Rent price (L1)	change in log rental expenditure, year preceding election	0.032	0.012	State
Religiosity	religiosity rate	0.511	0.170	County
Religiosity × incumbent party	religion interacted with incumbent party indicator	0.000	0.539	County
Migration	net migration (time-invariant)	0.005	0.009	County
Migration × incumbent party	migrate interacted with incumbent party indicator	0.000	0.010	County
Education	proportion with bachelor's degree or higher (time-invariant)	0.165	0.078	County
Education × incumbent party	education interacted with incumbent party indicator	0.000	0.183	County
ln(Median income)	log median household income	10.633	0.254	County
ln(Median income) × incumbent party	ln(median income) interacted with incumbent party indicator	-0.075	10.636	County
Poverty	poverty rate	0.155	0.062	County
Poverty × incumbent party	poverty interacted with incumbent party indicator	-0.013	0.167	County
Rural	urban-rural score (-4 to 4, time-invariant)	0.111	2.680	County
Rural × incumbent party	rural interacted with incumbent party indicator	0.111	2.680	County
Mail-in voting	indicator whether state mandates mail-in voting (1), optional (0), no mail-in voting (-1)	-0.301	0.564	State

Additional detail on variables can be found in Section S1 of the Appendix.

Among time-varying factors ($\mathbf{x}_{DLRO,cr,t+3}$ and $\mathbf{x}_{VT,cr,t+3}$) we include both short-run (the 3-month period before the election) and medium-run (the 1-year period before 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 socioeconomic and demographic factors like migration, urban-rural mix, education and religiosity.

Table 1 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 2: Summary Statistics for the Covariates in the Active Set for Changes in Republican Log Odds ($DLRO$) Panel Regressions over the Period 2000-2016

Covariate	Description	Mean	St. Dev.	Regional Coverage
Incumbent party	indicator taking 1 if incumbent party is Republican, -1 if incumbent party is Democrat	0.000	1.000	National
House $DLRO$	change in log Republican odds from preceding House election	0.087	0.346	State
Voter turnout (VT)	voter turnout proportion from the first-stage VT regression	0.576	0.090	County
VT × incumbent party	VT interacted with incumbent party indicator	0.015	0.583	County
Left-behind (L1)	county 'Left-Behind' measure, year preceding election	−0.005	0.087	County
Left-behind (L1) × incumbent party	Left-behind (L1) interacted with incumbent party indicator	−0.002	0.087	County
Healthcare costs (L1)	change in log healthcare expenditures, year preceding election	0.046	0.016	State
Government employment (L1)	change in log government employment, year preceding election	−0.012	0.015	State
USD REER (L1)	change in log real effective USD, year preceding election	0.009	0.055	State
USD REER (L1) × incumbent party	USD REER (L1) interacted with incumbent party indicator	−0.047	0.031	State
USD REER (M3)	Change in log real effective USD, 3 months preceding election	−0.012	0.114	State
USD REER (M3) × incumbent party	USD REER (M3) interacted with incumbent party indicator	0.046	0.105	State
Unemployment (L1)	unemployment rate avg., year preceding election	0.061	0.020	County
Unemployment (L1) × incumbent party	unemployment (L1) interacted with incumbent party indicator	−0.007	0.064	County
Unemployment (M3)	unemployment rate avg., 3 months preceding election	0.060	0.019	County
Unemployment (M3) × incumbent party	unemployment (M3) interacted with incumbent party indicator	−0.004	0.063	County
House price (L1)	change in log house prices avg., year preceding election	0.022	0.043	County
House price (L1) × inumbent party	house price (L1) interacted with incumbent party indicator	0.001	0.048	County
House price (M3)	change in log house prices avg., 3 months preceding election	0.025	0.055	County
House price (M3) × incumbent party	house price (M3) interacted with incumbent party indicator	−0.007	0.060	County
Rent price (L1)	change in log rental expenditure, year preceding election	0.032	0.012	State
Inflation (L1)	inflation, year preceding election	0.021	0.022	State
Migration	net migration (time-invariant)	0.005	0.009	County
Migration*	local net migration (time-invariant)	0.010	0.006	County
Education	proportion with bachelor's degree or higher (time-invariant)	0.165	0.078	County
Education*	local proportion with bachelor's degree or higher (time-invariant)	0.165	0.040	County
ln(Population density)	log population density (time-invariant)	3.727	1.668	County
ln(Median income)	log median household income	10.633	0.254	County
ln(Median income) × incumbent party	ln(median income) interacted with incumbent party indicator	−0.075	10.636	County
Poverty	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 turnout (VT), not model-fitted voter turnout ($\widehat{VT}$) 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. Additional detail on variables can be found in Section S1 of the Appendix.

Table 2 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, urban-rural mix, 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 2 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 Shrinkage and Selection Operator (Lasso) originally introduced by Tibshirani [1996], and the One Covariate at a time Multiple Testing (OCMT) procedure 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 are 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 for multiple testing. In subsequent stages, OCMT includes all selected variables in the previous stages in a multiple regression, and then re-tests those covariates which were not selected in the previous stages, and so on. The critical values adjusted for multiple-testing are 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, and 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 (48 mainland states plus D.C. × 4 election cycles). For the regional model, we cluster standard errors by state.[16] Details of the selection and estimation procedures for Lasso and OCMT are provided in Section S2 of the Appendix.

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 (plus Washington D.C.): $R_t = \sum_{s=1}^{51} R_{st}$. 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.

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 U.S. 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

---

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

$\sum_{s=1}^{51} w(d_s) = 1$. 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.[17]

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 $VT_{cr,t+4} = \left(R_{cr,t+4} + D_{cr,t+4}\right)/VAP_{cr,t+4}$, where $VAP_{cr,t+4}$ is the citizen voting age population in county c of region r in the election year $t+4$.[18] Also recall that $LRO_{cr,t+4} = DLRO_{cr,t+4} + LRO_{cr,t}$, and $\ln\left(R_{cr,t+4}/D_{cr,t+4}\right) = LRO_{cr,t+4}$. Suppose that we have forecasts for $VT_{cr,t+4}$ and $LRO_{cr,t+4}$. Then using these identities we have

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

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

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

---

[17] 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.

[18] Voting age population may differ from voting eligible population in that the former does not remove ineligible felons.

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 Key Determinants of U.S. Presidential Elections Using 2000-2016 as the Training Sample

In this section, we present estimates of the model estimated on the 2000-2016 training sample, presenting both pooled and regional estimates to further understand the key factors explaining voting outcomes. We begin with pooled estimates. The pooled model estimates for voter turnout and the Republican log-odds ratio equations are summarized in Tables 8 and 9, respectively.[19]

Several time-invariant covariates are statistically significant, regardless of whether estimated using OCMT or Lasso algorithms. These include urban-rural mix, 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 myopic voting behavior. Specifically, changes in the real effective U.S. Dollar (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 10 and

---

[19] 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 ordinary least squares (OLS) estimates for the covariates selected by Lasso together with their state-year clustered standard errors.

11 summarize the estimates for voter turnout VT under the Lasso and OCMT estimation algorithms, respectively. Similarly, Tables 12 and 13 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 different inference. The estimates also show how heterogeneous U.S. regions can be. Consider Table 12, the regional-Lasso estimates for $DLRO$. The education variable (Education) was selected for 8 out of 8 regions, hence this variable was 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 size of 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 (over the 3 months preceding the election), denoted by 'house price (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 'left-behind × incumbent party'). 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.

10 2020 Presidential Election: Forecasts and evaluation

We generate real-time forecasts of the 2020 presidential election using the active sets tabulated above, and the Lasso and OCMT selection algorithms. Training the models using data from 2000 to 2016, we recursively estimate the panel regressions (4) and (2) subject to the identifying restriction, $\delta_r = 0$, applying variable selection at each stage. The estimated models were then used to generate out-of-sample 2020 election forecasts at the county level formed on October 14, 2020 using data available as of that date. 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 models/forecasts, respectively.

In addition to our four main models (pooled-OCMT, regional-OCMT, pooled-Lasso, regional-Lasso), we also consider average Lasso-OCMT pooled and regional forecasts, along with two naive forecasts that were generated *ex post*: one forecast from a random walk model (RW) where the change in log Republican odds ratio is regressed on an intercept and a second forecast from an autoregressive model (AR) where the log Republican odds ratio level is regressed on its value from the previous election.[20] We refer to these as naive forecasts or forecasts from the naive models. Comparing predicted state and national vote shares and electoral votes against actual outcomes is a natural way to evaluate the forecasting performance of our models. In the following subsection we also provide a more formal statistical analysis of the forecasts across the models.

Perhaps it should be made clear that 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. Fortunately, both Alaska and Hawaii are non-swing states, historically voting Republican and Democrat, respectively. Specifically, Hawaii and Alaska have not changed their party winner in the last 9 and 14 elections, respectively. Both states received statehood in 1959. Alaska has voted for the Republican candidate in every election since 1960 except for the 1964 election when the state voted for Lyndon B. Johnson (Democrat) over Barry Goldwater (Republican). Similarly, Hawaii voted for the Democratic candidate in every election since 1960 except for two: 1972 and 1984 when the state voted for the Republican party. Therefore, in our electoral and national predictions we assumed in October 2020 that Alaska votes Republican and Hawaii votes Democrat.[21]

Two-party mainland national vote share forecasts

Table 3 provides forecasts for the two-party mainland national vote based on the 48 mainland states plus D.C. defined earlier as $V_t = R_t/(R_t + D_t)$. The mainland national vote is similar

---

[20] The two naive forecasts were generated *ex post*, after the 2020 election occurred, but these forecasts would remain unchanged if generated *ex ante* because they rely solely on data from past elections which were fully available prior to the 2020 election.

[21] It is worth noting that we include Utah in our model although large third-party vote shares may potentially induce high prediction error for the state (e.g. in 2016 a third-party candidate (Evan McMullin) won 21.54% of the state vote).

to the popular vote but excludes Hawaii and Alaska and only calculates the two-party vote share. All four models predicted a Democratic mainland national victory. Forecasts ranged from 45.3% (pooled-Lasso) to 49.4% (regional-OCMT). Pooled-OCMT was closest, forecasting 47.6% of the mainland national vote going for the Republican candidate, as compared to the realized value of 47.7%.[22] These results are favorably comparable to the 2020 two-party Republican popular vote share forecasts from national and state-level models: 44.6% predicted by Erikson and Wlezien [2020] who consider a national model with a leading economic indicator plus polling data, 45.6% predicted by the Economist[23], 48.31% by Jérôme et al. [2020], and 45.5% by Enns and Lagodny [2021]. The latter three forecasts are from models using state-level data.

Table 3: 2020 Two-Party Republican US. Mainland Vote Share and National Electoral College Forecasts

		Pooled Forecasts		Regional Forecasts	
	Realized	Lasso	OCMT	Lasso	OCMT
Vote Share (V_s)	0.477	0.453	0.476	0.474	0.494
Electoral College Votes	232	188	236	249	270

Realized U.S. mainland vote refers to 2020 Republican share of two-party votes across mainland U.S. states plus Washington D.C. To produce U.S. mainland 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. Electoral college votes refer to realized and predicted national Republican electoral college votes, and assumes Hawaii casts her electoral votes for the Democratic candidate and Alaska casts her electoral votes for the Republican candidate. Electoral college forecasts determined following Equation 7. Forecasts formed on October 14, 2020.

Electoral college vote and state-level forecasts

Table 14 reports state-by-state pooled and regional forecasts using Lasso, OCMT, and average Lasso-OCMT, all made in October 2020 before the election. A candidate requires 270 out of 538 electoral votes to win the national election. The final count for the November 2020 presidential election resulted in a Democratic victory with Joe Biden winning 306 electoral votes, and Donald Trump winning 232 electoral votes. Five out of six of the forecasts that we released in October 2020 predicted less than 270 electoral votes for the Republican party implying a Democrat candidate victory, although individual models varied substantially in the actual number of electoral vote forecasts. One forecast, regional-OCMT was right on the margin, forecasting 270 electoral votes for the Republican candidate – also reported in Table 3. These results starkly contrast forecasts from naive random walk and autoregressive

---

[22] Note that our forecasts exclude Hawaii and Alaska in the two-party mainland national vote calculation.

[23] The Economist forecasts are found here: https://projects.economist.com/us-2020-forecast/president.

model benchmarks which predicted a Republican victory with 329 electoral votes (Table 15). Meanwhile, the pooled-OCMT model forecasts performed best, forecasting 236 Republican electoral votes compared to the 232 realized. In comparison, Jérôme et al. [2020] predicted 230 Republican electoral votes, The Economist published a forecast of 182 Republican electoral votes, while 259 Republican electoral votes were predicted in Zandi et al. [2020] and 248 Republican electoral votes were predicted in Enns and Lagodny [2021], all state-level models.

10.1 Statistical evaluation of real-time 2020 election forecasts

We consider two types of forecast evaluations. The first is based on predictive accuracy of state-level binary outcomes. However, models may incorrectly predict the binary outcome by mere chance while forecasting well the state Republican vote share (e.g. cases where the vote share comes close to 50%). To allow for quantitative differences in forecast accuracy, we also evaluate forecasts on their ability to forecast Republican vote shares. All forecast evaluations are based on 49 forecasts for the 48 mainland states plus Washington D.C., and do not include Alaska and Hawaii since we did not explicitly model their 2020 outcomes.

Table 4: 2020 Republican Vote Share and State Winner Root Mean Square Forecast Errors

	Naive Forecasts			Pooled Forecasts		Regional Forecasts	
	Coin Flip	RW	AR	Lasso	OCMT	Lasso	OCMT
Vote Share	NA	0.037	0.034	0.029	0.015	0.032	0.034
State Winner	0.704	0.427	0.427	0.288	0.342	0.445	0.425

Root mean square forecast errors (RMSFE) computed as in Equation 13 reported for 2020 Republican vote share forecasts against realized state vote shares for the 48 U.S. mainland states plus Washington D.C. ($S = 49$) in the top row, and RMSFE for 2020 state winner forecasts against realized state winners, where Republican state vote shares greater than 0.50 are assigned a Republican win and state vote shares less than 0.50 assigned a Democrat win. Coin Flip refers to the average RMSFE comparing realized state winners against a benchmark averaging 1,000 simulations of a random {0,1} draw with probability 0.50. Forecasts formed on October 14, 2020.

Table 4 reports root mean square forecast errors (RMSFE) across the 48 mainland states plus D.C. for Republican 2020 vote shares (V_s) computed as:

$$RMSFE = \sqrt{\sum_{s=1}^{S} w_s (V_s - \widehat{V}_s)^2}, \tag{13}$$

across model generated and naive forecasts. Each state's squared forecast error is weighted by electoral vote share as a proportion of the total 531 electoral votes of the 48 mainland states plus D.C., given by w_s, where the weights sum to 1. The second row of Table 4 reports the RMSFE of state Republican wins, namely by converting the vote share to a win/loss indicator. For state winners, we also provide a random benchmark, "coin flip", which randomly draws a Republican win or loss for each state with probability 0.50.[24] Consistent with the overall presidential forecasts, pooled-OCMT which reported the most precise vote share and electoral college predictions also exhibited the lowest RMSFE for state vote shares. Pooled-Lasso, which only mis-predicted 2 states, had the lowest RMSFE for state winners, and naive forecasts generally had substantially higher RMSFEs.

Skill-based tests for forecasting state-level winners

When it comes to predicting the winning candidates across each state, the models varied in accuracy with the pooled models, specifically pooled-Lasso performing best. The pooled-Lasso mis-predicted 2 out of 48 states plus D.C.: Florida and North Carolina, both which the Republican candidate won and the pooled-OCMT model mis-predicted 4 states. Both naive models mis-predicted 9 states.[25]

We evaluate first whether any of the state binary forecasts exhibit significant 'skill', meaning whether they predicted the binary state-level outcomes better than random following the regression approach of Pesaran and Timmermann [1992] referred to as the PT test. Significant coefficients on the ordinary least squares (OLS) estimate from a linear regression of the actual binary state election outcomes on the predicted state election outcomes indicate forecasting 'skill' in that the predictive accuracy is significantly better than random.

Table 5: PT Test Statistics

Naive Forecasts		Pooled Forecasts		Regional Forecasts	
RW	AR	Lasso	OCMT	Lasso	OCMT
6.40	6.40	16.24	10.60	6.42	8.17

Test statistics reported from the PT test of forecasting skill (Pesaran and Timmermann [1992]). Under the null hypothesis of no forecasting skill, the test statistic follows a standard normal distribution. Forecasts formed on October 14, 2020.

Table 5 reports PT test statistics, of which all are statistically significant, including the naive models. The naive models exhibit the least skill in forecasting binary state elec-

---

[24] To minimize the effects of sampling errors we average the coin flip results across 1,000 replications. Note that for each replication the same realized outcomes are used, and only the random outcomes will vary across replications.

[25] For the 2016 election, the naive models mis-predicted 5 states, and all 5 predicted a Democratic win when a Republican win was realized, again suggesting systematic forecast error but this time consistent with a 'Trump effect' which was largely unexpected. 2016 naive model forecasts are omitted for brevity, but available upon request.

tion outcomes. The regional-Lasso model performs marginally better than the naive models, while both pooled forecasts outperform naive models substantially. Pooled-Lasso and pooled-OCMT having the highest PT statistics, consistent with their favorable forecasting performance.

Forecast accuracy tests for state Republican vote share

Evaluating the accuracy of Republican vote share, the target variable V_s, is crucial – and possibly a better gauge of model performance – especially when the prediction results are close to 50% as it is usually the case for swing states. Then the binary outcome uncertainty rises and the chance of predicting the actual winning candidate may be due to random chance despite the model predicting accurately a vote share close to 50%.

We consider Diebold-Mariano (DM) tests of forecast accuracy (Diebold and Mariano [2002]) modified for a single cross-section (e.g. 2020 forecasts across states) and allowing for state results to carry different weights. The DM test statistic is computed as,

$$Z_{DM}(a:b) = \frac{\sum_{s=1}^{S} w_s(|e_{sa}| - |e_{sb}|)}{\hat{\sigma}_\eta(\sum_{s=1}^{S} w_s^2)^{\frac{1}{2}}}, \qquad \hat{\sigma}_\eta^2 = \frac{1}{S}\sum_{s=1}^{S}(|e_{sa}| - |e_{sb}|)^2, \tag{14}$$

where $|e_{sa}| - |e_{sb}|$ defines the absolute loss differential computed as the difference in absolute forecast errors of model a and model b for state s. We allow for state loss differentials to be weighted by electoral vote share as a proportion of the total 531 electoral votes of the 48 mainland states plus D.C., given by w_s where the weights sum to 1. In addition to absolute loss differentials, we consider another common criteria, squared loss differentials, by replacing the expression $|e_{sa}| - |e_{sb}|$ with $e_{sa}^2 - e_{sb}^2$. Further detail on the modified DM test can be found in Section S3 of the Appendix along with the equation for the DM test with squared loss differentials (S.5).

Table 6 reports DM test statistics between each pair of models under the absolute loss differential between predicted and realized Republican vote shares V_s. First point to note is that the naive RW model underperforms all other models including the naive AR model. Meanwhile pooled-OCMT is the best performing model, significantly forecasting better than all other models at least at the 5% significance level. Both pooled models also outperformed regional-Lasso and OCMT models. While the pooled-Lasso model performed best with binary state election outcomes, pooled-OCMT predicted better realized Republican vote shares by states. Using squared loss differentials does not alter these conclusions (Table 7).

Forecast errors across models also exhibit different degrees of bias: pooled-Lasso, despite having the best binary prediction record, had a tendency to under-predict Republican vote

Table 6: Diebold-Mariano Test Statistics (Z_{DM}) under Absolute Loss Function

	Naive AR	Pooled Lasso	Pooled OCMT	Reg. Lasso	Reg. OCMT
Naive RW	**4.199**	1.434	**3.514**	1.609	1.567
Naive AR		0.825	**3.135**	0.887	0.723
Pooled Lasso			**3.577**	-0.008	-0.290
Pooled OCMT				**-2.261**	**-2.454**
Reg. Lasso					-0.386

DM statistics (Equation 14) correspond to absolute loss differential between model on x-axis and model on y-axis. Predicted outcome variable is state Republican vote share, V_s. Statistics in bold are significant at least at the 10% level. Negative values indicate that x-axis model outperformed y-axis model, and positive values indicate x-axis model underperformed y-axis model.

Table 7: Diebold-Mariano Test Statistics (Z_{DM}) under Squared Loss Function

	Naive AR	Pooled Lasso	Pooled OCMT	Reg. Lasso	Reg. OCMT
Naive RW	**2.845**	1.308	**3.180**	0.653	0.554
Naive AR		0.792	**2.947**	0.155	-0.046
Pooled Lasso			**4.461**	-0.453	-0.643
Pooled OCMT				**-1.802**	**-2.105**
Reg. Lasso					-0.225

DM statistics (Equation 14) correspond to squared loss differential between model on x-axis and model on y-axis. Predicted outcome variable is state Republican vote share, V_s. Statistics in bold are significant at least at the 10% level. Negative values indicate that x-axis model outperformed y-axis model, and positive values indicate x-axis model underperformed y-axis model.

share on average by 1.4%, especially for middle-ground states. Meanwhile the regional-OCMT over-predicted Republican vote share on average by 2.2%. Comparatively, the naive models have the largest bias in forecast errors with the random walk and autoregressive models over-predicting Republican vote share by 3.9% and 3.6%, respectively.

10.2 The role of voter turnout

A key assumption underlying our forecasting model is the simultaneous nature of turnout and vote share. We follow a recursive approach, first forecasting turnout and then feeding that forecast into a second-stage model for vote share prediction. One potential scenario which would violate our recursive assumption (voter turnout affects voting outcomes but not *vice versa*) is when election uncertainty is correlated with voter turnout. This subsection evaluates the robustness of this modeling choice using the 2020 election results. Specifically, we compare our two-equation systems based forecast to forecasts that would have been produced by a reduced form approach which did not make the recursive restriction. The reduced form approach refers to a model with a single equation for vote share, $DLRO_{cr,t+4}$. The reduced form model simply takes (2) and replaces voter turnout ($VT_{cr,t+4}$) from the right-hand side with the additional covariates from the voter turnout model in (4). Hence, the reduced form approach estimates a single equation, regressing the change in log Republican odds ($DLRO$) on the union of active set covariates across the turnout and vote share equations found in Tables 1 and 2. When transforming $DLRO_{cr,t+4}$ to predict Republican 2020 vote shares, the reduced form approach uses 2016 realized turnout.

A milestone feature of the 2020 election was its record level of voter turnout. We examine the relationship between 2020 turnout and election uncertainty using our updated 2020 data. We do find that 2020 turnout was significantly higher in states which had close elections in 2016 (our proxy for 2020 election uncertainty).[26] States with close election results were: Maine, Nevada, Minnesota, New Hampshire, Michigan, Pennsylvania, Wisconsin, and Florida. 2020 turnout in these states were estimated to be 74.8% on average, compared to an average of 67.2% across all other states. However, the average 2016-2020 *change* in turnout across these states were not significantly different from the average change in turnout of other states (+11.72 percentage points versus +11.45 percentage points). On average across all states, voter turnout rose 11.5 percentage points from 2016. The rise in turnout occurred regardless of political leaning: states which voted Democrat (Republican) in 2016 saw turnout rise on average 12 (11) percentage points.

In addition to comparing our systems based forecasts to a reduced form alternative, we conduct a "what-if" analysis using realized 2020 voter turnout data. Specifically we use our benchmark two-equation model which is estimated on data through October 2020, and when producing vote share forecasts for 2020, we feed in 2020 *realized* turnouts instead of predicted turnouts. This produces forecasts *conditional* on 2020 realized turnouts, as if *ex post* 2020 turnout was predetermined.

In summary, we compare 2020 forecasts from our benchmark two-equation model ("two-equation forecast") made in October 2020 with a single-equation model ("reduced form forecast") and with the two-equation model feeding in 2020 realized turnouts ("conditional two-equation"). This analysis allows us to assess several issues, including: the robustness of our recursive assumption; forecast error variance attributed to turnout prediction; the importance of turnout for the 2020 election.

Tables 16 and 17 report state level Republican vote share forecasts using Lasso and OCMT, respectively. The correlations between benchmark forecasts and the reduced form forecasts are high, ranging from 0.97 to 0.99 in all cases (also see Figures S.10 and S.11 in Section S6 of the online supplement). Using realized turnouts has little impact on the fore-

---

[26] Close elections being defined as those with Republican vote share between 48.5% and 51.5%.

casts, with the resultant vote shares conditional on knowing realized 2020 turnout hardly differing from the benchmark forecasts, OCMT or Lasso, pooled or regional. The simple correlation between the two sets of vote shares (two-equation forecasts and conditional two-equation forecasts) for all cases lie above 0.99. Overall, varying the voter turnout specification and even using 2020 realized turnout results in very small changes to the vote share forecasts, indicating that unsurprisingly, vote share prediction and not turnout prediction is overwhelmingly responsible for a large majority of forecast error variance in Republican vote shares.

In the case of pooled two-equation forecasts, Lasso gives lower forecasts of Republican vote shares relative to the OCMT forecasts for all states, on average 2% lower, with a similar pattern between pooled reduced form forecasts: the average under Lasso being 2.7% lower than OCMT. Pooled-Lasso vote shares from the conditional two-equation forecasts (knowing 2020 realized turnout) produces Republican vote shares on average 2% lower than the OCMT counterpart. In the case of regional forecasts the Lasso-OCMT difference is less pronounced, with Lasso forecasts of Republican vote share on average being lower by 1.3%, than OCMT, and the direction is mixed – Lasso forecasts are not below OCMT forecasts for all states. Similar results hold under the reduced form and conditional formulations.

It seems reasonable to conclude that for forecasting vote shares, the results are broadly robust to relaxing the recursive assumption imposed in the baseline model, yet at the same time the relatively low influence of voter turnout in the vote share predictions suggests that modeling voter turnout explicitly has not been consequential for predicting the 2020 election outcomes.

11 Concluding Remarks

Exploiting heterogeneity at the U.S. county-level, we develop real-time forecasts for the 2020 U.S. presidential election by augmenting national and region-based models of voter turnout and voting outcomes trained over the 2000-2016 sample with high-dimensional variable selection techniques. These forecasts were formed in October 2020. Five out of six forecasts pointed toward a Democratic national victory while the sixth predicted a marginal Republican victory with 270 electoral votes. These predictions starkly contrast those from naive random walk and autoregressive models which would have predicted a 2020 Republican national victory with 329 electoral votes.

While regional heterogeneity may be important for modeling swing states, pooled models performed relatively well in terms of forecast performance. Variable selection techniques, such as Lasso and OCMT, further improve model performance. Specifically, all models outperformed naive autoregressive benchmarks, with pooled-OCMT performing best on forecasting two-party mainland national Republican vote share (47.6% predicted, 47.7% realized) and total Republican electoral votes (236 predicted, 232 realized). The pooled-Lasso model performed best on forecasting state-by-state candidate winners, mis-predicting 2 out of 48 mainland states plus D.C. In terms of forecasting Republican vote shares, the pooled-OCMT model exhibited the lowest prediction error, significantly outperforming all other models. These results suggest that using fundamental socioeconomic and demographic data – particularly at the county level – can take us far in understanding presidential election cycle dynamics.

We also investigate which socioeconomic and demographic factors help explain voting behavior at the county level over the 2000-2016 election cycles. 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. Region-based models suggest that the influence of most other variables on turnout and voting outcomes substantially vary across regions. Our results also corroborate evidence of voter myopia: economic fluctuations realized a few months prior to the election are generally more potent predictors of voting outcomes compared to their long-horizon analogues (e.g. average unemployment rates just prior to the election versus average unemployment rates over the incumbent's entire term).

The 2020 election was also accompanied by a historic rise in voter turnout, increasing by 11.5 percentage points on average across all states. We assess the role of turnout in generating forecasts, finding that the overall impact was relatively small. This may be due to the uniform rise in turnout across 'red' and 'blue' states alike. Overall, we emphasize that the non-linear nature of the U.S. voting process makes forecasting elections challenging and subject to a high degree of uncertainty. In addition, unforeseeable events which could not be modeled adequately using historical data (e.g. nation-wide protests, pandemics) that were prevalent in 2020 certainly may have influenced the election and our resulting forecast errors.

References

Achen, C. H. and L. M. Bartels (2004). Musical chairs: Pocketbook voting and the limits of democratic accountability. In *Annual Meeting of the American Political Science*

Association, Chicago, pp. 1–5.

Alesina, A., H. Rosenthal, et al. (1995). *Partisan politics, divided government, and the economy.* Cambridge and New York: Cambridge University Press.

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

Autor, D., D. Dorn, G. Hanson, K. Majlesi, et al. (2020). Importing political polarization? The electoral consequences of rising trade exposure. *American Economic Review 110*, 3139–83.

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

Biesiada, M. J. (2018). *Factors that Impact Direct Democracy and Voter Turnout: Evidence from a National Study on American Counties.* Ph. D. thesis, University of Nevada, Las Vegas.

Blais, A. (2006). What affects voter turnout? *Annual Review of Political Science 9*, 111–125.

Campbell, A., P. E. Converse, W. E. Miller, and D. E. Stokes (1960). *The American Voter.* Chicago: University of Chicago Press.

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: Journal of the Econometric Society 86*, 1479–1512.

Diebold, F. X. and R. S. Mariano (2002). Comparing predictive accuracy. *Journal of Business and Economic Statistics 20*, 134–144.

Enns, P. K. and J. Lagodny (2021). Forecasting the 2020 electoral college winner: The state presidential approval/state economy model. *PS: Political Science & Politics 54*, 81–85.

Erikson, R. S. and C. Wlezien (2020). Forecasting the 2020 presidential election: Leading economic indicators, polls, and the vote. *PS: Political Science & Politics 54*, 1–4.

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

Fair, R. C. (1996). Econometrics and presidential elections. *Journal of Economic Perspectives 10*, 89–102.

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

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

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

Graefe, A. (2018). Predicting elections: Experts, polls, and fundamentals. *Judgment and Decision Making 13*, 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*, 405–423.

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

Jérôme, B., V. Jérôme, P. Mongrain, and R. Nadeau (2020). State-level forecasts for the 2020 U.S. presidential election: Tough victory ahead for Biden. *PS: Political Science & Politics 54*, 1–4.

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

Kahane, L. H. (2020). Determinants of county-level voting patterns in the 2012 and 2016 presidential elections. *Applied Economics 52*, 1–14.

Keeter, S., R. Igielnik, and R. Weisel (2016). Can likely voter models be improved? Evidence from the 2014 U.S. 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 U.S. voting behavior, 1896-1964. *The American Political Science Review 65*, 131–143.

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

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

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

Pesaran, M. H. and A. Timmermann (1992). A simple nonparametric test of predictive performance. *Journal of Business & Economic Statistics 10*, 461–465.

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

Rogers, T. and M. Aida (2014). Vote self-prediction hardly predicts who will vote, and is (misleadingly) unbiased. *American Politics Research 42*, 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*, 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*, 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*, 607–614.

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

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

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

Wlezien, C. (2015). The myopic voter? The economy and U.S. presidential elections. *Electoral Studies 39*, 195–204.

Wlezien, C. and R. S. Erikson (1996). Temporal horizons and presidential election forecasts. *American Politics Quarterly 24*, 492–505.

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

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

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

Table 8: 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)
Intercept	0.0886	(0.1239)	0.1432	0.1412***	(0.0192))
Incumbent party	-0.2179**	(0.1085)			
Incumbent president	0.0314***	(0.0053)	0.0289	0.0337***	(0.0051)
VT (L1)	0.7977***	(0.0172)	0.7789	0.7973***	(0.0169)
VT (L1) × incumbent party	-0.0461***	(0.0166)			
Healthcare costs (L1)			-0.1294	-0.2453	(0.1844)
Government employment (L1)	0.1557	(0.187)	0.1449	0.1915	(0.2077)
Unemployment (L1)	0.4034***	(0.1044)	0.2964	0.3408***	(0.1132)
House price (L1)					
Rent price (L1)			-0.0576	-0.1889	(0.1898)
Religiosity			-0.005	-0.0096	(0.0071)
Religiosity × incumbent party	-0.0116*	(0.0066)			
Migration	-0.3548***	(0.1023)	-0.249	-0.4002***	(0.1056)
Migration × incumbent party	-0.2233**	(0.1025)			
Education	0.0978***	(0.0163)	0.0955	0.1067***	(0.014)
Education × incumbent party	0.0907***	(0.0158)	0.0835	0.0818***	(0.0133)
ln(Median income)	0.0026	(0.0109)			
ln(Median income) × incumbent party	0.0231**	(0.0096)			
Poverty	-0.1517***	(0.0495)	-0.1596	-0.1555***	(0.0329)
Poverty × incumbent party	0.0277	(0.0472)			
Rural	-4e-04	(4e-04)	-4e-04	-7e-04	(4e-04)
Rural × incumbent party	-8e-04**	(4e-04)	-0.0014	-0.002***	(4e-04)
Mail-in voting	0.0065**	(0.0032)	0.0063	0.0075**	(0.0034)
Observations	12,438		12,438	12,438	
Covariates Selected	19		15	15	
Adj. R2	0.8058		0.8034	0.8048	
Reg SE	0.0397		0.0399	0.0398	

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. 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 1.

Table 9: 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)
Intercept	0.6955***	(0.0907)	0.6828	0.6763***	(0.0995)
Incumbent party	-0.8364**	(0.3566)	-0.1478	-0.2725***	(0.0645)
House *DLRO*	0.025	(0.0281)	0.0186	0.0218	(0.0258)
$\widehat{VT}$	-0.3735***	(0.1069)	-0.2839	-0.2894**	(0.1126)
$\widehat{VT}$ × incumbent party	-0.1786*	(0.0938)	-0.1094	-0.0166	(0.0921)
Left-behind (L1)			0.0051	0.0235	(0.0375)
Left-behind (L1) × incumbent party			-0.1118	-0.1119***	(0.0364)
Healthcare costs (L1)					
Government employment (L1)			2.4752	2.7807***	(1.0024)
USD REER (L1)	-0.0198	(0.8802)			
USD REER (L1) × incumbent party	0.0737	(0.8872)			
USD REER (M3)	-0.0389	(0.2468)			
USD REER (M3) × incumbent party	-0.7329***	(0.2311)	-0.5045	-0.4768***	(0.1479)
Unemployment (L1)			-0.9088	-0.5727	(0.4429)
Unemployment (L1) × incumbent party	-2.6836*	(1.5928)			
Unemployment (M3)					
Unemployment (M3) × incumbent party	4.8527***	(1.6454)	0.9323	2.0594***	(0.4623)
House price (L1)	-0.3884	(0.405)			
House price (L1) × incumbent party					
House price (M3)	0.7047**	(0.3133)	0.3722	0.4541***	(0.1613)
House price (M3) × incumbent party					
Rent price (L1)			-0.8429	-1.4238*	(0.7286)
Inflation (L1)			1.0148	1.3794***	(0.5128)
Migration	-1.7827***	(0.4813)	-1.4525	-1.716***	(0.4533)
Migration*	0.79	(1.1803)	0.2324	0.872	(1.1734)
Education	-0.6296***	(0.0962)	-0.6864	-0.7094***	(0.1029)
Education*	-0.7883***	(0.2045)	-0.7032	-0.7402***	(0.2108)
ln(Population density)	-0.001	(0.0056)	4e-04	0.0058	(0.0054)
ln(Median income)					
ln(Median income) × incumbent party	0.0649*	(0.0336)			
Poverty	-0.6909***	(0.1486)	-0.5167	-0.5939***	(0.1385)
Rural	0.0089***	(0.0021)	0.005	0.0081***	(0.002)
Observations	12,438		12,438	12,438	
Covariates Selected	21		21	21	
Adj. R2	0.5071		0.5185	0.5232	
Reg SE	0.1788		0.1768	0.1758	

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. 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 2.

Table 10: Regional Panel Regressions with Dependent Variable as Voter Turnout (VT) Estimated 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
Incumbent party					0.021			
Incumbent president	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								
Unemployment (L1)		0.100		0.269		-1.081	0.667	0.630
House price (L1)					0.090		0.163	0.268
Religiosity		-0.005	-0.041	-0.026		0.034		
Religiosity × incumbent party				0.007				
Migration		-0.175	-0.056					
Migration × incumbent party								
Education	0.062	0.116	0.103	0.120	0.019	0.069	0.113	0.008
Education × incumbent party	0.093	0.061	0.077	0.069		0.011	0.075	0.075
ln(Median income)	0.008		-0.021					
ln(Median income) × incumbent party	0.001			0.001				
Poverty	-0.125	-0.214	-0.345	-0.167	-0.218	-0.327	-0.269	-0.301
Poverty × incumbent party								
Rural						-0.001		
Rural × incumbent party	-0.001	-0.001					-0.001	
Observations	4,252	1,516	600	860	268	712	1,748	2,472

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 1.

Table 11: 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)
Incumbent party	-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)
Incumbent president	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) × incumbent party	-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)
Unemployment (L1)	0.163	(0.197)	0.482***	(0.155)							-1.975***	(0.606)				
House price (L1)	0.063	(0.107)	-0.157***	(0.041)					0.117	(0.116)						
Religiosity			-0.022***	(0.006)							0.056**	(0.026)				
Religiosity × incumbent party	-0.007	(0.011)	-0.023***	(0.007)	-0.001	(0.012)	0.013**	(0.006)					-0.014	(0.011)	0.005	(0.011)
Migration			-0.484***	(0.087)	-1.06***	(0.263)							-0.498***	(0.126)		
Migration × incumbent party	-0.181	(0.151)	-0.015	(0.094)	-0.367	(0.234)							-0.108	(0.152)	-0.157	(0.169)
Education	0.047**	(0.022)	0.132***	(0.029)	0.166***	(0.04)	0.227***	(0.029)			0.15***	(0.055)	0.216***	(0.067)		
Education × incumbent party	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)
ln(Median income)	0.033**	(0.013)	0.016	(0.02)			-0.082***	(0.023)			-0.059**	(0.027)	-0.071	(0.045)		
ln(Median income) × incumbent party	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)
Poverty	-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)
Poverty × incumbent party	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 × incumbent party	-0.001	(0.001)	-0.002**	(0.001)				(0.001)							-0.001	(0.001)
Observations	4,252		1,516		600		860		268		712		1,748		2,472	

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 1. 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 12: 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
Incumbent party	-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
$\widehat{VT}$ × incumbent party			-0.189	-0.147		0.215	0.190	-0.087
Left-behind (L1)	0.008	-0.012	-0.051		-0.309		-0.166	-0.025
Left-behind (L1) × incumbent party	-0.089		-0.011	-0.154	0.526	-0.243	0.001	-0.086
Unemployment (L1)	-1.770	2.100	3.666		-0.890	-1.649	2.253	-1.061
Unemployment (L1) × incumbent party	0.001	0.263	-2.168			1.613	3.893	0.380
Unemployment (M3)	-0.499	-1.536	-4.135				-1.047	4.108
Unemployment (M3) × incumbent party	0.187		3.290		7.901		-1.917	
House price (L1)	-0.976	-0.104			-2.261	0.500	2.071	4.000
House price (L1) × incumbent party	-0.716	-0.696	0.175			-1.306	-2.235	-2.995
House price (M3)	1.689	0.561	0.682	0.128		2.062	1.271	
House price (M3) × incumbent party	0.301	0.717	-0.223	0.921	3.176		1.807	1.055
Migration	-1.335	-3.022	0.994		-0.043	-3.364	-0.008	-0.998
Migration*	1.078		-0.878		0.169		3.976	
Education	-0.606	-0.845	-0.646	-0.586	-0.729	-0.295	-0.854	-0.753
Education*	-2.301	-0.489			0.204	-0.246	-0.928	-0.442
ln(Population density)	-0.010	0.008	-0.009	-0.005	-0.003	0.014	-0.004	
ln(Median income)		-0.322	-0.032		0.055	-0.028	-0.045	-0.025
ln(Median income) × incumbent party	-0.004	-0.008	0.079					
Poverty	-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	
Observations	4,252	1,516	600	860	268	712	1,748	2,472

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 2.

Table 13: 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)
Incumbent party					-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)
$\widehat{VT}$ × incumbent party					-0.285	(0.179)	-0.207	(0.19)	-0.209**	(0.098)	0.417***	(0.094)	0.223*	(0.132)	-0.004	(0.098)
Left-behind (L1)			-0.078*	(0.042)												
Left-behind (L1) × incumbent party									0.055	(0.303)	-0.277***	(0.066)				
Unemployment (L1)					3.932***	(0.341)										
Unemployment (L1) × incumbent party					-1.845**	(0.89)	3.958	(3.645)	14.579***	(4.054)			9.201**	(4.627)	1.094	(4.152)
Unemployment (M3)					-4.152***	(1.069)									4.141***	(0.914)
Unemployment (M3) × incumbent party					3.432***	(1.226)	-3.961	(4.949)	-10.078***	(3.881)	5.528*	(3.049)	-7.281	(5.613)	2.592	(4.906)
House price (L1)	-0.866	(0.724)			-0.069	(0.529)					0.85	(0.86)	0.41	(1.416)	4.06***	(0.917)
House price (L1) × incumbent party	-1.227***	(0.388)													-4.248***	(1.095)
House price (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)
Houes price (M3) × incumbent party															2.708**	(1.118)
Migration	-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)
Migration*					-1.001	(1.653)					2.808*	(1.609)				
Education	-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)
Education*	-1.698***	(0.379)			0.041	(0.507)	-0.01	(0.136)			-0.669**	(0.295)	-1.467***	(0.354)	-0.144	(0.48)
ln(Population density)	-0.006	(0.005)	0.01	(0.015)	-0.011	(0.01)	-0.016***	(0.003)			0.015***	(0.005)				
ln(Median income)	0.299***	(0.057)														
ln(Median income) × incumbent party					0.234***	(0.05)	0.136*	(0.082)	0.316***	(0.023)	0.075*	(0.044)	0.1***	(0.029)	0.065	(0.044)
Poverty			-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)
Observations	4,252		1,516		600		860		268		712		1,748		2,472	

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 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 14: State Level Forecasts of Republican Vote Shares (V_s) and Electoral Votes for 2020 Elections Made in October 2020

				Lasso				OCMT				Lasso-OCMT Average			
	Total EC	2020 Realized		Pooled Forecasts		Regional Forecasts		Pooled Forecasts		Regional Forecasts		Pooled Forecasts		Regional Forecasts	
State	(d_s)	V_s	EC Votes	$\widehat{V}_s$	EC Votes	$\widehat{V}_s$	EC Votes	$\widehat{V}_s$	EC Votes	$\widehat{V}_s$	EC Votes	$\widehat{V}_s$	EC Votes	$\widehat{V}_s$	EC Votes
AK	3	0.553	3	N/A	3	N/A	3	N/A	3	N/A	3	N/A	3	N/A	3
AL	9	0.629	9	0.628	9	0.636	9	0.641	9	0.654	9	0.635	9	0.645	9
AR	6	0.642	6	0.629	6	0.650	6	0.646	6	0.665	6	0.638	6	0.658	6
AZ	11	0.498	0	0.489	0	0.549	11	0.521	11	0.570	11	0.505	11	0.560	11
CA	55	0.351	0	0.302	0	0.313	0	0.337	0	0.340	0	0.320	0	0.326	0
CO	9	0.431	0	0.405	0	0.413	0	0.421	0	0.417	0	0.413	0	0.415	0
CT	7	0.398	0	0.369	0	0.498	0	0.398	0	0.479	0	0.384	0	0.488	0
DC	3	0.055	0	0.032	0	0.036	0	0.033	0	0.040	0	0.032	0	0.038	0
DE	3	0.404	0	0.400	0	0.425	0	0.415	0	0.469	0	0.408	0	0.446	0
FL	29	0.517	29	0.464	0	0.462	0	0.488	0	0.482	0	0.476	0	0.472	0
GA	16	0.499	0	0.491	0	0.506	16	0.511	16	0.525	16	0.501	16	0.516	16
HI	4	0.350	0	N/A	0	N/A	0	N/A	0	N/A	0	N/A	0	N/A	0
IA	6	0.542	6	0.524	6	0.551	6	0.535	6	0.578	6	0.530	6	0.565	6
ID	4	0.659	4	0.665	4	0.648	4	0.677	4	0.655	4	0.671	4	0.652	4
IL	20	0.413	0	0.380	0	0.417	0	0.398	0	0.419	0	0.389	0	0.418	0
IN	11	0.582	11	0.581	11	0.576	11	0.604	11	0.582	11	0.593	11	0.579	11
KS	6	0.576	6	0.577	6	0.576	6	0.585	6	0.592	6	0.581	6	0.584	6
KY	8	0.632	8	0.638	8	0.654	8	0.655	8	0.669	8	0.647	8	0.662	8
LA	8	0.595	8	0.577	8	0.582	8	0.600	8	0.609	8	0.589	8	0.596	8
MA	11	0.329	0	0.292	0	0.408	0	0.318	0	0.391	0	0.305	0	0.400	0
MD	10	0.330	0	0.313	0	0.346	0	0.333	0	0.373	0	0.323	0	0.359	0
ME	4	0.455	0	0.454	0	0.511	4	0.465	0	0.479	0	0.460	0	0.495	0
MI	16	0.486	0	0.474	0	0.508	16	0.497	0	0.500	0	0.486	0	0.504	16
MN	10	0.464	0	0.449	0	0.466	0	0.465	0	0.501	10	0.457	0	0.484	0
MO	10	0.578	10	0.593	10	0.623	10	0.608	10	0.651	10	0.601	10	0.637	10
MS	6	0.584	6	0.581	6	0.582	6	0.602	6	0.610	6	0.591	6	0.596	6
MT	3	0.584	3	0.573	3	0.570	3	0.593	3	0.573	3	0.583	3	0.572	3
NC	15	0.507	15	0.486	0	0.488	0	0.504	15	0.508	15	0.495	0	0.499	0
ND	3	0.672	3	0.661	3	0.693	3	0.688	3	0.734	3	0.674	3	0.713	3
NE	5	0.598	5	0.602	5	0.641	5	0.611	5	0.693	5	0.606	5	0.667	5
NH	4	0.463	0	0.443	0	0.526	4	0.470	0	0.496	0	0.457	0	0.511	4
NJ	14	0.419	0	0.371	0	0.414	0	0.407	0	0.445	0	0.389	0	0.429	0
NM	5	0.445	0	0.410	0	0.465	0	0.442	0	0.479	0	0.426	0	0.472	0
NV	6	0.488	0	0.468	0	0.461	0	0.500	6	0.463	0	0.484	0	0.462	0
NY	29	0.383	0	0.340	0	0.345	0	0.369	0	0.342	0	0.355	0	0.344	0
OH	18	0.541	18	0.520	18	0.542	18	0.540	18	0.541	18	0.530	18	0.542	18
OK	7	0.669	7	0.673	7	0.680	7	0.687	7	0.700	7	0.680	7	0.690	7
OR	7	0.417	0	0.403	0	0.407	0	0.423	0	0.424	0	0.413	0	0.415	0
PA	20	0.494	0	0.468	0	0.497	0	0.499	0	0.555	20	0.484	0	0.524	20
RI	4	0.395	0	0.378	0	0.494	0	0.389	0	0.463	0	0.383	0	0.479	0
SC	9	0.559	9	0.560	9	0.559	9	0.571	9	0.581	9	0.566	9	0.570	9
SD	3	0.634	3	0.641	3	0.632	3	0.652	3	0.640	3	0.647	3	0.636	3
TN	11	0.618	11	0.620	11	0.637	11	0.642	11	0.655	11	0.631	11	0.646	11
TX	38	0.528	38	0.503	38	0.512	38	0.533	38	0.538	38	0.518	38	0.525	38
UT	6	0.607	6	0.603	6	0.598	6	0.615	6	0.606	6	0.609	6	0.602	6
VA	13	0.448	0	0.424	0	0.414	0	0.440	0	0.441	0	0.432	0	0.428	0
VT	3	0.317	0	0.318	0	0.377	0	0.333	0	0.361	0	0.325	0	0.369	0
WA	12	0.401	0	0.370	0	0.384	0	0.399	0	0.411	0	0.385	0	0.397	0
WI	10	0.497	0	0.474	0	0.507	10	0.495	0	0.505	10	0.485	0	0.506	10
WV	5	0.698	5	0.712	5	0.720	5	0.736	5	0.740	5	0.724	5	0.730	5
WY	3	0.725	3	0.726	3	0.728	3	0.750	3	0.736	3	0.738	3	0.732	3
All Votes	538		232		188		249		236		270		215		265

Republican vote shares are calculated as in Equation 12. Column 'Total EC (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. Forecasts made on October 14, 2020 using 2000-2016 as the training sample.

Table 15: State Level Forecasts of Republican Vote Shares (V_s) and Electoral Votes for 2020 using Random Walk and Autoregressive Models

	Random Walk		Autoregressive–AR(1)	
State	$\hat{V}_s$	EC Votes	$\hat{V}_s$	EC Votes
AK	N/A	3	N/A	3
AL	0.663	9	0.664	9
AR	0.663	6	0.664	6
AZ	0.543	11	0.539	11
CA	0.359	0	0.349	0
CO	0.496	0	0.491	0
CT	0.452	0	0.445	0
DC	0.047	0	0.041	0
DE	0.463	0	0.456	0
FL	0.528	29	0.524	29
GA	0.547	16	0.544	16
HI	N/A	0	N/A	0
IA	0.573	6	0.571	6
ID	0.703	4	0.705	4
IL	0.430	0	0.423	0
IN	0.622	11	0.622	11
KS	0.632	6	0.632	6
KY	0.676	8	0.678	8
LA	0.622	8	0.622	8
MA	0.375	0	0.365	0
MD	0.378	0	0.371	0
ME	0.509	4	0.504	4
MI	0.524	16	0.520	16
MN	0.514	10	0.509	10
MO	0.637	10	0.637	10
MS	0.611	6	0.611	6
MT	0.633	3	0.632	3
NC	0.541	15	0.537	15
ND	0.717	3	0.720	3
NE	0.656	5	0.656	5
NH	0.522	4	0.518	4
NJ	0.450	0	0.442	0
NM	0.475	0	0.469	0
NV	0.511	6	0.505	6
NY	0.402	0	0.395	0
OH	0.565	18	0.562	18
OK	0.713	7	0.715	7
OR	0.460	0	0.454	0
PA	0.525	20	0.521	20
RI	0.441	0	0.433	0
SC	0.597	9	0.596	9
SD	0.680	3	0.682	3
TN	0.656	11	0.657	11
TX	0.568	38	0.565	38
UT	0.644	6	0.644	6
VA	0.493	0	0.488	0
VT	0.370	0	0.360	0
WA	0.434	0	0.426	0
WI	0.526	10	0.522	10
WV	0.740	5	0.744	5
WY	0.773	3	0.777	3
All Votes		329		329

Republican vote shares are calculated as in Equation 12. 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. For the Random Walk (RW) model, the log Republican odds ratio is regressed on an intercept. For the Autoregressive model, the log Republican odds ratio is regressed on its value from the previous election. Forecasts are generated *ex post* only using data available as of October 14, 2020.

Table 16: State Level Forecasts of Republican Vote Shares (V_s) for 2020 under Alternative Turnout Specifications using Lasso Algorithm

		Pooled Lasso Forecasts			Regional Lasso Forecasts		
	State	Two-Equation	Reduced Form	Conditional Two-Equation	Two-Equation	Reduced Form	Conditional Two-Equation
1	AL	0.628	0.642	0.631	0.636	0.643	0.639
2	AR	0.629	0.639	0.633	0.650	0.655	0.654
3	AZ	0.489	0.503	0.480	0.549	0.565	0.564
4	CA	0.302	0.314	0.301	0.313	0.312	0.315
5	CO	0.405	0.414	0.401	0.413	0.426	0.404
6	CT	0.369	0.388	0.369	0.498	0.540	0.496
7	DC	0.032	0.035	0.032	0.036	0.040	0.033
8	DE	0.400	0.418	0.404	0.425	0.432	0.422
9	FL	0.464	0.483	0.459	0.462	0.468	0.454
10	GA	0.491	0.505	0.492	0.506	0.514	0.504
11	IA	0.524	0.535	0.521	0.551	0.572	0.548
12	ID	0.665	0.674	0.654	0.648	0.656	0.635
13	IL	0.380	0.394	0.383	0.417	0.438	0.413
14	IN	0.581	0.594	0.582	0.576	0.623	0.577
15	KS	0.577	0.589	0.572	0.576	0.613	0.572
16	KY	0.638	0.652	0.638	0.654	0.662	0.653
17	LA	0.577	0.594	0.581	0.582	0.589	0.585
18	MA	0.292	0.312	0.292	0.408	0.458	0.412
19	MD	0.313	0.325	0.316	0.346	0.354	0.341
20	ME	0.454	0.462	0.449	0.511	0.540	0.518
21	MI	0.474	0.488	0.472	0.508	0.512	0.512
22	MN	0.449	0.464	0.446	0.466	0.495	0.464
23	MO	0.593	0.606	0.593	0.623	0.647	0.616
24	MS	0.581	0.590	0.587	0.582	0.586	0.588
25	MT	0.573	0.576	0.563	0.570	0.573	0.558
26	NC	0.486	0.499	0.482	0.488	0.496	0.482
27	ND	0.661	0.666	0.663	0.693	0.709	0.698
28	NE	0.602	0.613	0.597	0.641	0.647	0.636
29	NH	0.443	0.457	0.441	0.526	0.565	0.526
30	NJ	0.371	0.389	0.372	0.414	0.426	0.408
31	NM	0.410	0.421	0.405	0.465	0.475	0.459
32	NV	0.468	0.481	0.460	0.461	0.460	0.461
33	NY	0.340	0.358	0.346	0.345	0.364	0.360
34	OH	0.520	0.533	0.523	0.542	0.561	0.545
35	OK	0.673	0.685	0.676	0.680	0.697	0.680
36	OR	0.403	0.413	0.399	0.407	0.409	0.408
37	PA	0.468	0.480	0.468	0.497	0.502	0.486
38	RI	0.378	0.399	0.379	0.494	0.529	0.494
39	SC	0.560	0.574	0.558	0.559	0.564	0.553
40	SD	0.641	0.645	0.637	0.632	0.668	0.630
41	TN	0.620	0.634	0.621	0.637	0.645	0.636
42	TX	0.503	0.521	0.496	0.512	0.532	0.516
43	UT	0.603	0.609	0.592	0.598	0.610	0.583
44	VA	0.424	0.440	0.425	0.414	0.422	0.414
45	VT	0.318	0.320	0.313	0.377	0.420	0.367
46	WA	0.370	0.378	0.366	0.384	0.385	0.387
47	WI	0.474	0.486	0.474	0.507	0.537	0.509
48	WV	0.712	0.719	0.713	0.720	0.724	0.721
49	WY	0.726	0.726	0.724	0.728	0.733	0.726

Republican vote shares are calculated as in Equation 12. Two-equation forecast refers to real-time baseline two-equation forecasts. Reduced form forecasts and conditional two-equation are described in Section 10.1. Reduced form forecasts are from a single vote share equation model which includes the union of covariates from both the turnout and vote share active sets. Two-equation and Reduced form forecasts use data available as of October 14, 2020. Conditional two-equation forecasts are from the two-equation baseline model estimated on data through October 14, 2020, but predicted turnouts are replaced with realized 2020 turnouts when calculating 2020 Republican vote share predictions. Two-equation forecasts were formed on October 14, 2020 while reduced form and conditional two-equation forecasts are generated *ex post* only using data available as of October 14, 2020.

Table 17: State Level Forecasts of Republican Vote Shares (V_s) for 2020 under Alternative Turnout Specifications using OCMT Algorithm

		Pooled OCMT Forecasts			Regional OCMT Forecasts		
	State	Two-Equation	Reduced Form	Conditional Two-Equation	Two-Equation	Reduced Form	Conditional Two-Equation
1	AL	0.641	0.667	0.645	0.654	0.642	0.660
2	AR	0.646	0.659	0.651	0.665	0.659	0.672
3	AZ	0.521	0.542	0.509	0.570	0.568	0.570
4	CA	0.337	0.360	0.335	0.340	0.323	0.338
5	CO	0.421	0.447	0.416	0.417	0.420	0.404
6	CT	0.398	0.429	0.400	0.479	0.556	0.468
7	DC	0.033	0.041	0.033	0.040	0.051	0.041
8	DE	0.415	0.442	0.419	0.469	0.482	0.477
9	FL	0.488	0.520	0.480	0.482	0.470	0.479
10	GA	0.511	0.535	0.510	0.525	0.514	0.527
11	IA	0.535	0.554	0.531	0.578	0.585	0.587
12	ID	0.677	0.684	0.663	0.655	0.655	0.636
13	IL	0.398	0.421	0.402	0.419	0.440	0.415
14	IN	0.604	0.618	0.606	0.582	0.628	0.583
15	KS	0.585	0.601	0.581	0.592	0.629	0.590
16	KY	0.655	0.676	0.654	0.669	0.665	0.672
17	LA	0.600	0.632	0.605	0.609	0.596	0.616
18	MA	0.318	0.350	0.318	0.391	0.470	0.382
19	MD	0.333	0.352	0.336	0.373	0.390	0.379
20	ME	0.465	0.480	0.460	0.479	0.561	0.474
21	MI	0.497	0.523	0.495	0.500	0.498	0.506
22	MN	0.465	0.491	0.460	0.501	0.520	0.509
23	MO	0.608	0.627	0.609	0.651	0.662	0.653
24	MS	0.602	0.626	0.609	0.610	0.594	0.618
25	MT	0.593	0.602	0.581	0.573	0.569	0.556
26	NC	0.504	0.531	0.499	0.508	0.496	0.506
27	ND	0.688	0.697	0.691	0.734	0.729	0.742
28	NE	0.611	0.627	0.605	0.693	0.674	0.695
29	NH	0.470	0.488	0.468	0.496	0.589	0.479
30	NJ	0.407	0.436	0.409	0.445	0.476	0.456
31	NM	0.442	0.467	0.436	0.479	0.477	0.476
32	NV	0.500	0.515	0.490	0.463	0.459	0.459
33	NY	0.369	0.393	0.375	0.342	0.390	0.361
34	OH	0.540	0.564	0.544	0.541	0.555	0.545
35	OK	0.687	0.706	0.691	0.700	0.696	0.700
36	OR	0.423	0.443	0.417	0.424	0.413	0.420
37	PA	0.499	0.520	0.499	0.555	0.555	0.556
38	RI	0.389	0.422	0.391	0.463	0.539	0.453
39	SC	0.571	0.593	0.568	0.581	0.568	0.581
40	SD	0.652	0.662	0.649	0.640	0.678	0.644
41	TN	0.642	0.659	0.642	0.655	0.647	0.657
42	TX	0.533	0.559	0.524	0.538	0.530	0.537
43	UT	0.615	0.621	0.599	0.606	0.611	0.585
44	VA	0.440	0.466	0.441	0.441	0.428	0.448
45	VT	0.333	0.334	0.327	0.361	0.438	0.333
46	WA	0.399	0.416	0.393	0.411	0.393	0.407
47	WI	0.495	0.515	0.494	0.505	0.533	0.508
48	WV	0.736	0.744	0.739	0.740	0.734	0.743
49	WY	0.750	0.751	0.749	0.736	0.734	0.734

Republican vote shares are calculated as in Equation 12. Two-equation forecast refers to real-time baseline two-equation forecasts. Reduced form forecasts and conditional two-equation are described in Section 10.1. Reduced form forecasts are from a single vote share equation model which includes the union of covariates from both the turnout and vote share active sets. Two-equation and Reduced form forecasts use data available as of October 14, 2020. Conditional two-equation forecasts are from the two-equation baseline model estimated on data through October 14, 2020, but predicted turnouts are replaced with realized 2020 turnouts when calculating 2020 Republican vote share predictions. Two-equation forecasts were formed on October 14, 2020 while reduced form and conditional two-equation forecasts are generated *ex post* only using data available as of October 14, 2020.

Appendix

The Appendix is organized as follows: Section S1 provides detail on relevant data and sources. Section S2 gives an account of Least Absolute Shrinkage and Selection Operator (Lasso) and One Covariate at a time Multiple Testing (OCMT) variable selection algorithms. Section S3 provides details on the Diebold-Mariano (DM) test for forecast evaluation.

S1 Data

Descriptions, frequency, sources

Data has been cleaned and merged from several different publicly available sources. County-level voting outcomes data are taken from the Massachussetts Institute of Technology (MIT) Election Data and Science Lab.[S1] County gross domestic product (GDP) measures are obtained from the Bureau of Economic Analysis (BEA). Education, population, migration, and urban-rural county classifications are from the United States Department of Agriculture (USDA). For education, we fix values for all years using year 2000 values. We use population levels in year 2000 for election years up through 2008 and year 2010 population levels for election years 2012 to 2020. Migration measures total net international migration from 2010 to 2015, and we use this value for all years. Urban-rural classifications are reported in 2010, and we use these values for all years in our data set. Therefore, data on county education, migration, and urban-rural mix do not vary over time in our sample. 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, and we use these values for all years. Hence, religiosity does not vary over time in our sample. Data on voting age population (VAP) are from the American Community 5-year surveys (ACS). 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 interpolate 2004 turnout values using 2008 turnouts. County-level unemployment rates are provided by the Bureau of Labor Statistics (BLS) and county-level house price indices are taken from Zillow. State-level inflation is computed from indices reported by the BEA. State level export-weighted real exchange rates are from the

---

[S1] The site URL is: https://electionlab.mit.edu/.

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 Federal Information Processing Standard (FIPS) code changes: Some counties changed 5-digit 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: Data from the MIT Data and Science 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 plus Washington D.C., thus excluding Alaska and Hawaii from our analysis. 2020 election results are taken from results published by Fox News, Politico, and the New York Times.[S2]

Annual county GDP: Data from the BEA 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 the 2019 GDP data was not released until December 2020.

Annual Virginia county GDP: 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

---

[S2] 2020 county election data can be found here: https://github.com/tonmcg/US_County_Level_Election_Results_08-20.

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.

County U.S. presidential voter turnout: We estimate voter turnout (VT) as the total two-party votes (Republican and Democrat) divided by the VAP, which we take from the 5-year ACS. To compute VT, we rely on the 90% upper confidence interval of the VAP estimate to avoid turnouts greater than 100%. The VAP measure is an estimate over a 5-year period while the number of votes is a single snapshot in time.[S3] 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 interpolate 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[S4] For counties with a VAP-to-total population ratio larger than 1, we replace VAP for these counties with the product of the county population with the average of VAP-to-population ratio of surrounding counties (within 100 miles) which have VAP-to-population ratios less than 1.

Biennial state U.S. midterm votes: We collect data on U.S. house votes for biennial House elections by state from the MIT Election Data and Science 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 D.C. because the latter does not have voting rights during these House elections. We use this data to compute two-party Republican vote share variables using House election data, in a way that is analogous to county presidential Republican vote share.

Religiosity: Data on religiosity is taken 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,

---

[S3] Alternatively, intercensal and postcensal estimates, which the ACS estimates are based on, could be used which provide snap shots instead of moving average estimates

[S4] The observations are: Harding County, New Mexico in years 2004, 2008, 2012; and Hanson County, South Dakota in 2012.

in order to keep the rate variable bounded between 0 and 1, for counties with greater than 100 percent religiosity rate, we replace county c's religiosity rate with the local religiosity rate, taken as the average religiosity rate of neighboring counties within 100 miles of c.

Monthly county house prices: 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. We therefore estimate local county house price returns based on the average of counties within 100 miles of county c, inclusive of county c. 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 election 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.

Annual state rent expenditures: We compute annual log growth rates in state-level rent expenditure 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.

Monthly county unemployment rates: We take monthly unemployment rates at the county level from the BLS. We estimate local averages using all counties within 100 miles of county c, inclusive of county c. For election 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.

Quarterly state inflation: From the BEA, we take quarterly real GDP and nominal GDP by state to compute a state-level quarterly GDP deflator as GDP Deflator = (Nominal GDP/Real GDP) $\times 100$. Inflation is calculated as the logged change form the previous quarter's GDP deflator for each state. Because elections are held every November, we use state-level inflation rate from year Q3 2015 - Q2 2016 for election year 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.

Monthly state real effective exchange rates: State-level U.S. Dollar (USD) real effective exchange rates (REER) are taken from the Federal Reserve Bank of Dallas. Monthly state-level REERs are computed using a trade-weighted average of USD exchange rates vis-a-vis the primary export partners of that 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. So for election year 2016, the annualized change in the log REER is calculated from July 2015 to June 2016. 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.

Annual state healthcare expenditures: We compute annual log growth rates of 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.

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

Population density: 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 policy: 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 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.

National incumbent party and incumbent president: To capture the incumbency effects on voter turnout and election outcomes we consider two national incumbency indicators, distinguishing between presidential and party incumbency. 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. This way we allow for a wide variety of incumbency effects (positive or negative) considered in the literature, without biasing the forecast 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

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

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.[S5] 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.[S6]

S2 Variable Selection Algorithms

Least absolute shrinkage and selection operator (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.

[S5] We do not compute Y_t; rather we take the data directly from the BEA.

[S6] 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.

One covariate at a time multiple testing (OCMT)

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.

S3 Diebold-Mariano Test for Cross-Section Forecasts

The classic DM test compares time-series forecasts. To formally evaluate forecasting accuracy of our models, we adapt the DM test for assessing a single set of cross-sectional forecasts. Let $L(e_s)$ denote the loss function, where e_s is the forecats error for state s. Examples include $L(e_s) = |e_s|$ or $L(e_s) = e_s^2$, absolute or squared loss functions, respectively. Then, based a cross-sectional sample of forecasts and realizations across 48 states plus D.C., namely with $S = 49$, we have $DM(a:b) = \sum_{s=1}^{S} w_s[L(e_{sa}) - L(e_{sb})]$. Considering the squared loss function, we have $DM(a:b) = \sum_{s=1}^{S} w_s(e_{sa}^2 - e_{sb}^2)$. We suppose the loss-differential of the two forecats follow the simple model $e_{sa}^2 - e_{sb}^2 = \alpha_s + \eta_s$, where α_s is a fixed constant, and η_s is a mean zero random variable distributed indepdently over s. Then averaging over the $s = 1, 2, ..., S$ states with their electoral college vote shares, w_s, we have $\sum_{s=1}^{S} w_s(e_{sa}^2 - e_{sb}^2) = \sum_{s=1}^{S} w_s\alpha_s + \sum_{s=1}^{S} w_s\eta_s = \bar{\alpha}_w + \bar{\eta}_w$. Under the null hypothesis $H_0 : \bar{\alpha}_w = 0$, to be compared to the alternative $H_1 : \bar{\alpha}_w < 0$, when forecasts from model a is preferred to forecast from model b, and *vice versa* if $H_1 : \bar{\alpha}_w > 0$. Assuming η_s are independently distributed across s, and $\sum_{s=1}^{S} w_s^2 = O\left(S^{-1}\right)$, then under H_0, $DM(a:b)$ is approximately normally distributed with zero means and variance $V(DM_s) = \sum_{s=1}^{S} w_s^2 Var(\eta_s)$. Since we have a single cross section we further assume that $Var(\eta_s) = \sigma_\eta^2$, and note that $V(DM_s) = \left(\sum_{s=1}^{S} w_s^2\right)\sigma_\eta^2$, where $\sigma_\eta^2 = E\left(S^{-1}\sum_{s=1}^{S}\eta_s^2\right)$. The DM test statistic can now be written as

$$Z_{DM}(a:b) = \frac{\sum_{s=1}^{S} w_s(e_{sa}^2 - e_{sb}^2)}{\hat{\sigma}_\eta(\sum_{s=1}^{S} w_s^2)^{\frac{1}{2}}}, \qquad \hat{\sigma}_\eta^2 = \frac{1}{S}\sum_{s=1}^{S}(e_{sa}^2 - e_{sb}^2)^2, \tag{S.5}$$

where (S.5) is the squared loss analog of DM test with absolute loss differential given in (14). For absolute loss function, we need to replace e_{sa}^2 and e_{sb}^2 with $|e_{sa}|$ and $|e_{sb}|$, respectively. We can interpret one-sided tests as follows: If $Z_{DM}(a : b) > 0$ and H_0 is rejected, forecast b is preferred to a. If $Z_{DM}(a : b) < 0$ and H_0 is rejected, forecast a is preferred to b.

Online Supplement to

"Regional Heterogeneity and U.S. Presidential Elections: Real-Time 2020 Forecasts and Evaluation"

Rashad Ahmed
University of Southern California, USA

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

June 23, 2021

This online supplement is organized as follows: Section S1 describes selecting the functional form of the election outcome variable. Section S2 derives the proof for consistency of the two-stage estimation of the model of voter turnout and voting outcomes. Section S3 provides *ex post* analysis of forecasts for the 2020 election using a the United States Department of Agriculture – Agriculture Research Service (USDA) regional classification. Section S4 provides *ex post* analysis of forecasts for the 2020 election generated from models considering an extended, larger active set of covariates. Section S5 reports forecasts and evaluation of the 2016 Presidential election using 2000-2012 as the training sample. Section S6 presents additional figures and tables.

S1 Functional Form of the Outcome Variable

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

$$V_{cr,t} = \frac{R_{cr,t}}{R_{cr,t} + D_{cr,t}}, \tag{S.1}$$

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 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 $V'_{cr,t}$:

$$V'_{c,rt} = \left\{ V_{c,rt}, \;\; \ln(V_{c,rt}), \;\; \ln\left(\frac{V_{c,rt}}{1 - V_{cr,t}}\right) \right\}, \tag{S.2}$$

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:

$$LRO_{cr,t} = \ln\left(\frac{V_{cr,t}}{1 - V_{cr,t}}\right) = \ln\left(\frac{R_{c,rt}}{D_{c,rt}}\right). \tag{S.3}$$

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 does not account for re-transforming the prediction back to the target variable. Therefore, to appropriately compare models under transformed dependent variables, the prediction error from the transformed regressions 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.1. 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:

$$\ln \bar{z}_1 = -\frac{1}{NT}\sum_{t=1}^{T}\sum_{i=1}^{N}\ln V_{cr,t} - \frac{1}{NT}\sum_{t=1}^{T}\sum_{i=1}^{N}\ln(1 - V_{cr,t}), \tag{S.4}$$

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

$$\ln \bar{z}_3 = -\frac{1}{NT}\sum_{t=1}^{T}\sum_{i=1}^{N}\ln V_{cr,t}. \tag{S.5}$$

The “Adjusted SE” in Table S.1 compares post-adjustment regression standard errors (SE). 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.

Table S.1: 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

Sample period: 2004-2016. County Republican vote share, $V_{cr,t}$ is defined as in Equation S.1. 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 S1.

S2 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 $T^{\hat{-1}}\mathbf{Z}'\boldsymbol{\xi} \xrightarrow{p} 0$, and hence $\hat{\boldsymbol{\theta}} \xrightarrow{p} \boldsymbol{\theta}$.

S3 Alternative Regional Classification

There are several regional classifications for U.S. states with the Bureau of Economic Analysis (BEA) 8-region classification being one of them. In this section, as a robustness check, we re-estimate the regional-based models under an alternative classification. Specifically we consider the United States Department of Agriculture (USDA) five geographic areas to generate region-based forecasts for 2020, again using data available as of October 14, 2020.[S1] However it should be noted that this is an *ex post* analysis conducted after the election took place.

Table S.6 reports vote share and electoral forecasts under the USDA-Lasso and USDA-OCMT regional models, respectively. The USDA-Lasso model predicted 253 Republican electoral votes compared to 249 by the baseline BEA regional-Lasso model (Table 3). Differences in total electoral vote forecasts is attributed to differing winning candidate predictions for Arizona, Maine, Michigan, North Carolina, and Pennsylvania. The USDA-OCMT model predicted 284 Republican electoral votes compared to 270 by the BEA regional-Lasso model

---

[S1] The five regions are Northeast, Midwest, Southeast, Plains, and Pacific West.

resulting from differing winning candidate forecasts for Maine, Michigan, Minnesota, and New Hampshire.

State-level Republican vote share forecasts are overall very similar whether using the 8-region BEA classification or 5-region USDA classification. The correlation between BEA and USDA regional-based vote share forecasts is 0.986 and 0.987 under Lasso and OCMT approaches, respectively. Moreover, the differences between BEA and USDA region-based forecasts are not statistically significant. The electoral-weighted Diebold-Mariano (DM) statistic is -0.013 for absolute loss differentials between the BEA regional-Lasso forecasts and USDA regional-Lasso, and 1.53 when comparing BEA regional-OCMT forecasts with USDA regional-OCMT forecasts.

S4 Extending the Active Set

It is well known that the performance of variable selection algorithms could depend on the number of covariates in the active set relative to the sample size. Here we investigate the robustness of our October 2020 forecasts in an *ex post* exercise where we extend the active set for the change in log Republican odds ($DLRO$) equation, by adding two new categories of covariates to the October 2020 active set.[S2]

First, we consider an indicator variable which takes the value of +1 (respectively -1) for state-year observations which correspond to the running Republican (Democratic) candidates' home state, and zero otherwise. For 2016, we do not include such an indicator because both candidates, Trump and Clinton, had designated home states of New York. We also consider interactions of this indicator variable with unemployment rates (1-year and 3-month), house price changes (1-year and 3-month), and the median income, yielding a total of six additional covariates to be added to the active set.[S3]

Second, we introduce several interactive effects where we interact four covariates with little to no time-series variation, such as education, migration, urban-rural mix, and population density, with other covariates that exhibit large time-series variation such as house price changes (1-year and 3-month), unemployment rates (1-year and 3-month), and median household income. This effectively allows for a greater degree of heterogeneity in the response of county vote share to changes in house prices, unemployment, and income. We then included only those interactive variables whose correlations with their underlying variables were not too high. Specifically, denoting the two types of variables by x and y, we

---

[S2] We do not add any new variables to extend the active set for the voter turnout (VT), which is not the primary object of interest.

[S3] Analysis using an extended active set was carried out on recommendation of one of the reviewers.

then considered the correlations of $x \times y$ with x and y, and only kept $x \times y$ variables if both correlations, $\text{cor}(x \times y, y)$ and $\text{cor}(x \times y, x)$, were less than 0.50 in absolute value. Due to the slow-moving nature of x relative to y, many of the interacted variables had very high correlations with y.

This resulted in three new interacted covariates: migration × house prices (M3); rural × house prices (L1); rural × house prices (M3); plus the six new covariates reflecting candidate home state for a total of nine new covariates to extend the *DLRO* active set. The list of all additional covariates is provided in Table S.2. We re-estimate and generate *ex post* forecasts for the 2020 election using the data available as of October 2020 with the extended *DLRO* active set. We consider both the pooled-OCMT and pooled-Lasso forecasts as the pooled models allow for all variables across different levels of regional coverage to be considered. The extended active set contains a total of 40 covariates. The number of covariates selected by OCMT decreases from 21 to 20. No new covariates were selected, and population density falls out from the selected covariates.[S4] The number of selected covariates by Lasso increases, rising from 21 to 29, selecting seven of the 9 new covariates from the extended set.[S5]

Table S.2: New Covariates Added to the Active Set for Changes in Log Republican Odds (*DLRO*) *ex post* Analysis

Covariate	Description	Mean	St. Dev.	Regional Coverage
Candidate state	indicator taking 1 when for Republican candidate's home state, -1 for Democrat candidate's home state, 0 otherwise	0.005	0.200	State
Candidate state × unemployment (L1)	pstate interacted with unemployment (L1)	0.000	0.013	County
Candidate state × umemployment (M3)	pstate interacted with unemployment (M3)	0.000	0.013	County
Candidate state × house price (L1)	pstate interacted with house price (L1)	0.001	0.009	County
Candidate state × house price (M3)	pstate interacted with house price (M3)	0.001	0.012	County
Candidate state × ln(median income)	pstate interacted with ln(median income)	0.049	2.123	County
Migration × house price (M3)	migration interacted with house price (M3)	0.000	0.001	County
Rural × house price (L1)	rural interacted with house price (L1)	0.010	0.134	County
Rural × house price (M3)	rural interacted with house price (M3)	0.009	0.168	County

New variables extending the *DLRO* active set provided in Table 2, with further detail on variables and descriptions also available in Section S1 of the Appendix.

Table S.7 reports the state-level vote share and electoral college forecasts when using the extended active set for *DLRO*. As can be seen, extending the active set does not affect our October 2020 electoral college forecasts, with 188 Republican by pooled-Lasso, 236 by pooled-OCMT, despite the differences in selected variables. The Republican vote share forecasts also do not change in any substantial way, as the October baseline forecasts share very high correlations with the *expost* forecasts under the extended active set (correlations of

---

[S4] This is due to the fact that as the number of covariates in the active set rises, the critical value threshold also rises because it depends on the number of covariates k, and could end up selecting fewer covariates from the original active set. See Section S2 of the Appendix for further details on the OCMT algorithm.

[S5] Lasso selecting relatively more variables than OCMT is consistent with Monte Carlo studies reported in Chudik et al. [2018].

0.999 for both pooled-Lasso and pooled-OCMT with their respective October baseline state vote share forecasts).

Popular vote forecasts also do not change very much, but do improve slightly. The two-party mainland Republican vote share forecasts with the extended active set are 0.456 and 0.477 using Lasso and OCMT, respectively, as compared to the October baselines of 0.453 and 0.476, respectively, and the 2020 realized share of 0.477.

S5 2016 Presidential Forecasts Based on 2000-2012 as the Training Sample

This section evaluates *ex ante* forecasts of the 2016 Presidential Election using the Lasso and OCMT selection algorithms over the 2000-2012 training sample. We recursively estimate the panel regressions (4) and (2) subject to the identifying restrictions, $\delta_r = 0$ and apply 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. We only model the 48 U.S. mainland states plus the District of Columbia. We do not model Hawaii or Alaska.

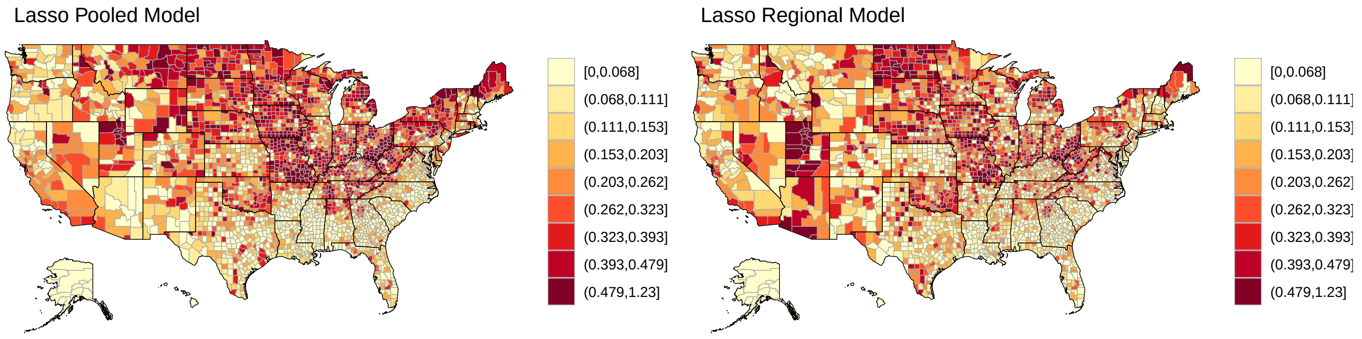
S5.1 Pooled and regional forecasts

To produce 2016 out-of-sample forecasts, we the model trained over the 2000-2012 period, feeding in data 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 two-party: state-level votes, electoral votes, and the overall U.S. Mainland national votes. Tables with electoral outcomes for a subset of notable swing states are also included.

State level forecast results for 2016 are reported in Table S.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. The table reports pooled and regional forecasts along with pooled and regional forecasts for Lasso-OCMT average forecasts.

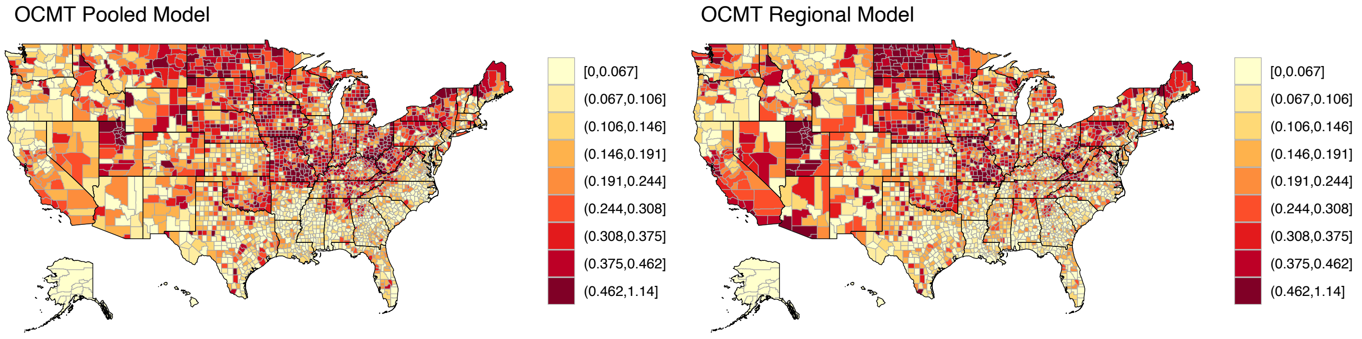
It is clear that, irrespective of which algorithm is used for variable selection, the primary difference between the forecasts is whether we allow for regional heterogeneity or not. Pooled forecasts predict a Democratic victory whilst the regional forecasts correctly predict a Republican victory. For example, the pooled model using Lasso algorithm predicts Re-

Figure S.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 $|DLRO_{cr,2016} - \widehat{DLRO}_{cr,2016}|$.

Figure S.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 $|DLRO_{cr,2016} - \widehat{DLRO}_{cr,2016}|$.

publican 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 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.[S6] 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 S.1 and S.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.

S5.2 Swing state forecasts

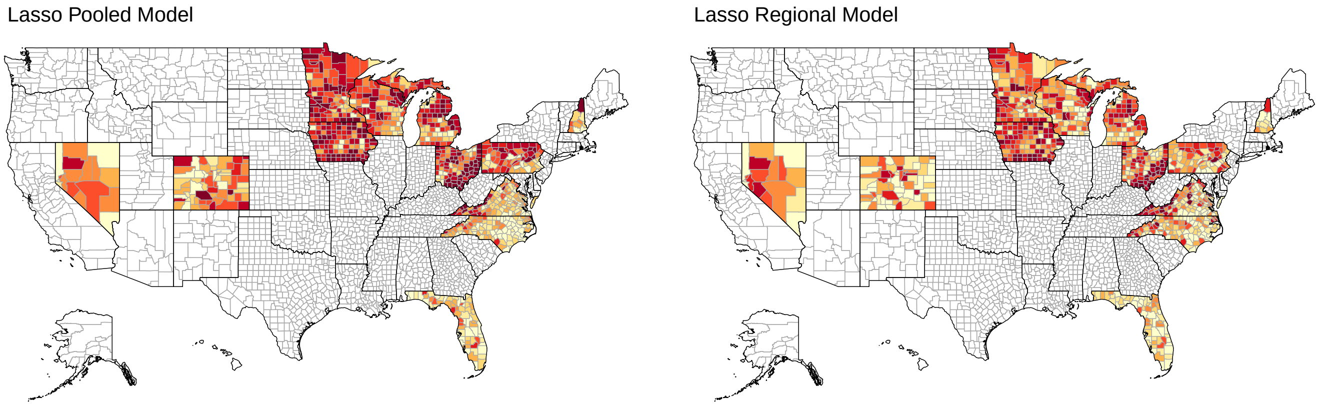
U.S. presidential elections usually come down to the results from key swing states. Therefore a model that predicts the swing states well is likely to go a long way in correctly forecasting the election. We consider the following 12 states as key swing states: Colorado, Florida, Iowa, Michigan, Minnesota, Nevada, New Hampshire, North Carolina, Ohio, Pennsylvania, Virginia, and Wisconsin. Figures S.3 and S.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 S.3 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

---

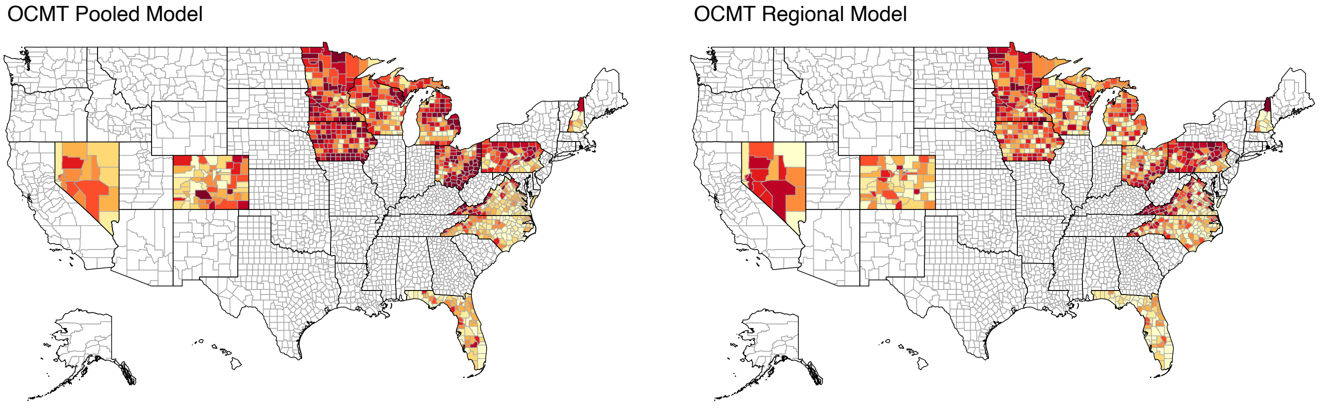
[S6] See https://en.wikipedia.org/wiki/2016_United_States_presidential_election

Figure S.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 $|DLRO_{cr,2016} - \widehat{DLRO}_{cr,2016}|$.

Figure S.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 $|DLRO_{cr,2016} - \widehat{DLRO}_{cr,2016}|$.

regional-Lasso and regional-OCMT models predicted the Republican candidate winning 117 and 109 electoral votes in the swing states, respectively. By contrast, the pooled-Lasso and 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 S.3: 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 S.5 compares swing state predicted Republican vote shares (V_s) obtained using the Lasso algorithm. The regional-Lasso model correctly predicted 9 of the 12 swing states outcomes, namely Florida, Iowa, Nevada, New Hampshire, North Carolina, Ohio, Pennsylvania, Virginia and Wisconsin. The regional-OCMT model also correctly predicted 9 of 12 swing states, namely Florida, Iowa, Michigan, Minnesota, New Hampshire, North Carolina, Ohio, Virginia, Wisconsin (see Figure S.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

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

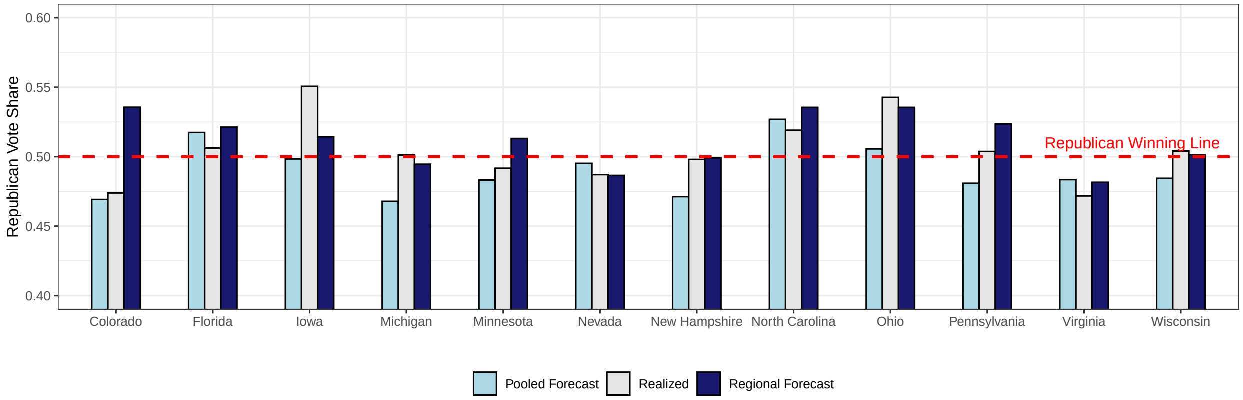
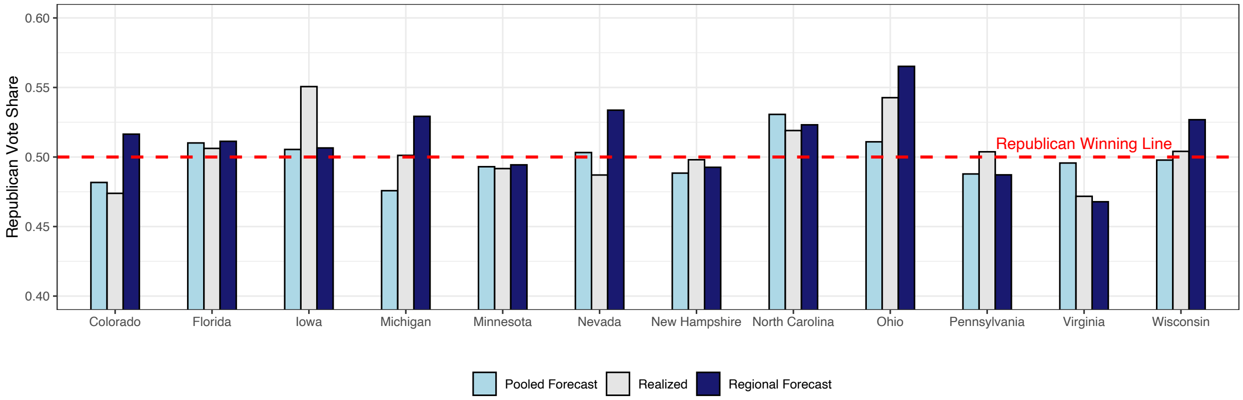


Figure S.6: Swing State Forecasts and Realized Republican Vote Share (V_s) for 2016 using the OCMT Estimation Algorithm



with 20 electoral votes. The regional-Lasso model correctly predicted the Republican win in Pennsylvania. The Republican victory in Michigan was also mis-predicted under both pooled model specifications, but correctly predicted by the regional-OCMT model.

S5.3 2016 forecasts for U.S. mainland national vote

The U.S. mainland national Republican vote share forecasts (V_t) are reported in Table S.4. 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 pooled Lasso (OCMT) model predicted a Republican vote share of 0.494 (0.499) which are closer to the realized value of 0.489, compared to 0.510 (0.514) predicted using the regional-Lasso (regional-OCMT) model.

Table S.4: 2016 Two-Party Republican US. Mainland Vote Share and National Electoral College Forecasts

		Pooled Forecasts		Regional Forecasts	
	Realized	Lasso	OCMT	Lasso	OCMT
Vote Share (V_s)	0.489	0.494	0.499	0.510	0.514
Electoral College Votes	304	253	265	308	307

Realized U.S. mainland vote refers to 2016 Republican share of two-party votes across mainland U.S. states plus Washington D.C. To produce U.S. mainland 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. Electoral college votes refer to realized and predicted national Republican electoral college votes, and assumes Hawaii casts her electoral votes for the Democratic candidate and Alaska casts her electoral votes for the Republican candidate. Electoral college forecasts determined following Equation 7. Forecasts are formed using 2000-2012 as the training sample.

To summarize, allowing for parameter heterogeneity across regions considerably improves 2016 *ex ante* 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 U.S. mainland 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 table S.8. 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. 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, regional-OCMT 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 regional-OCMT 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 regional-OCMT (which predicts Republican) and regional-Lasso (which predicts Democrat) models.

S6 Additional Figures and Tables

Table S.5: 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

Figure S.7: Bureau of Economic Analysis Regions

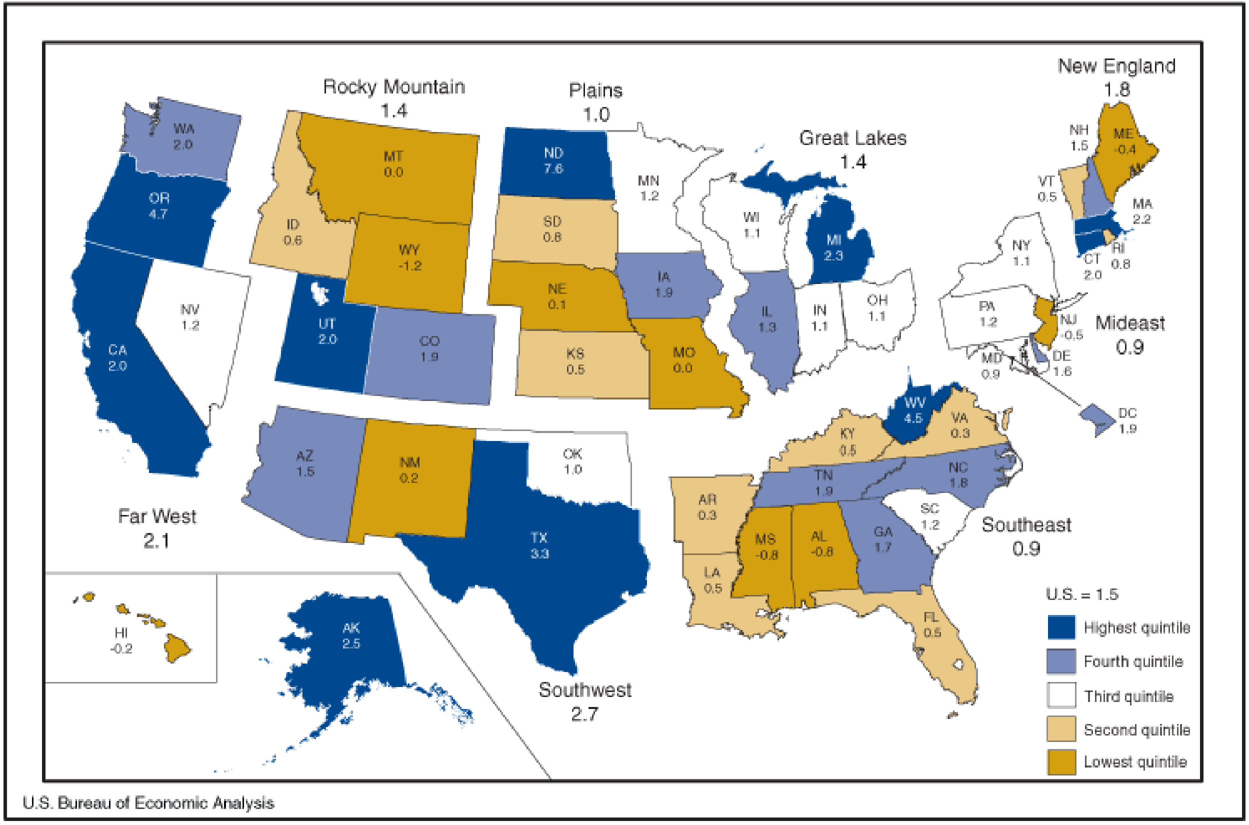


Chart 1. Percent Change in Real GDP by State, 2011

U.S. Bureau of Economic Analysis

Figure S.8: Histogram of Voter Turnout (VT) over the period 2004-2016 at Mainland U.S. and Regional Levels

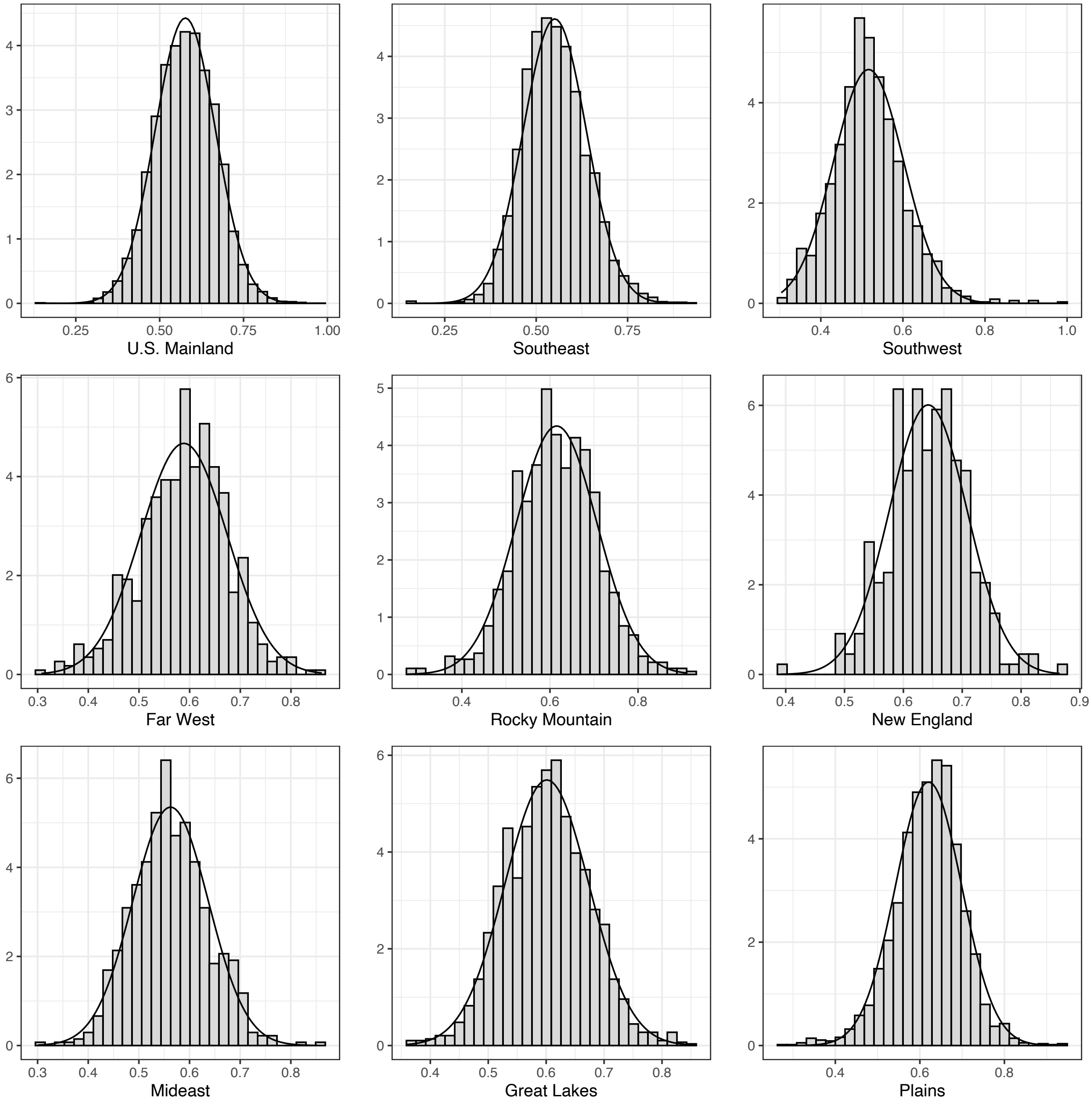


Figure S.9: Histogram of changes in Log Republican Odds Ratio ($DLRO$) over 2004-2016 at Mainland U.S. and Regional Levels

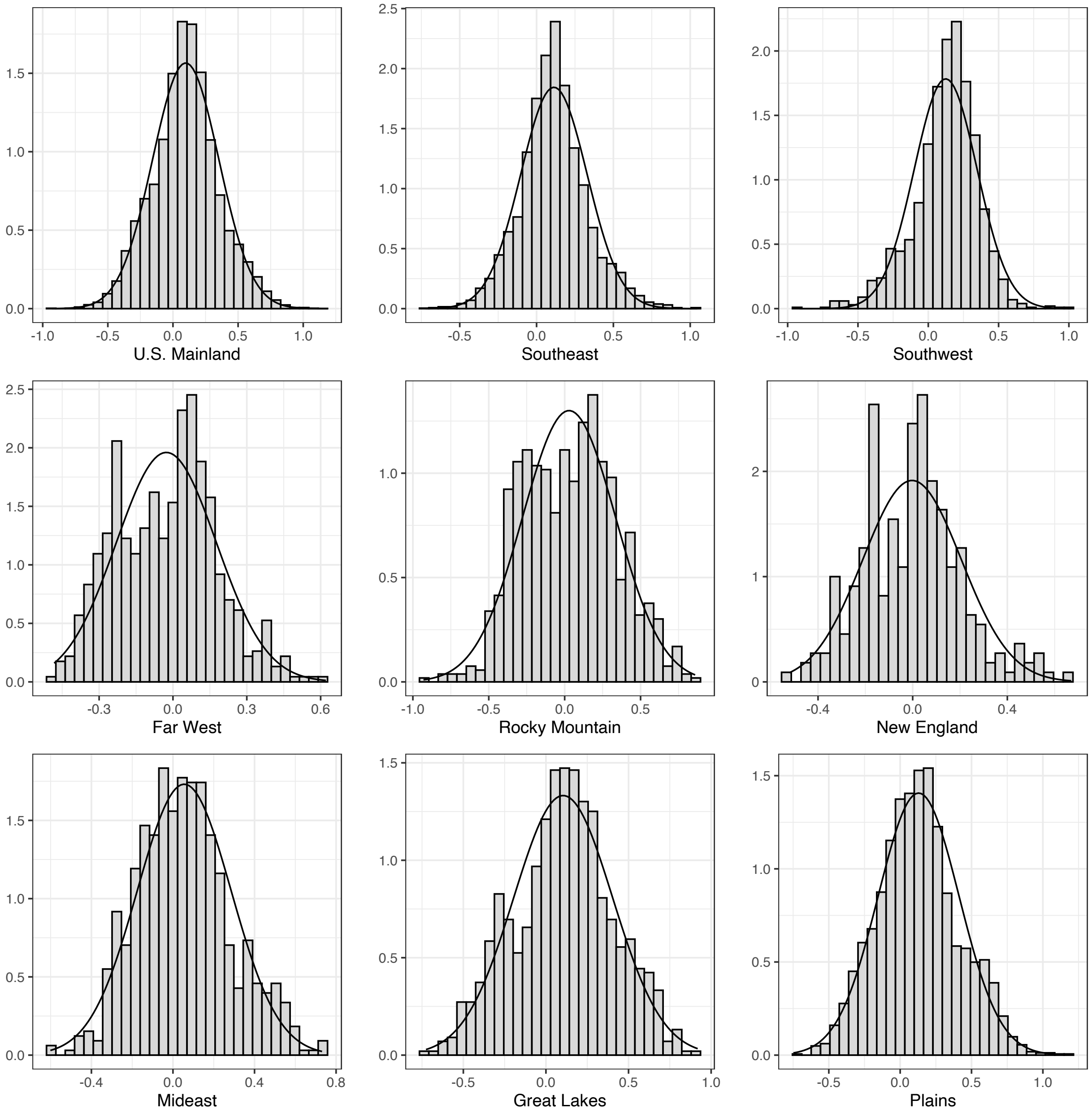
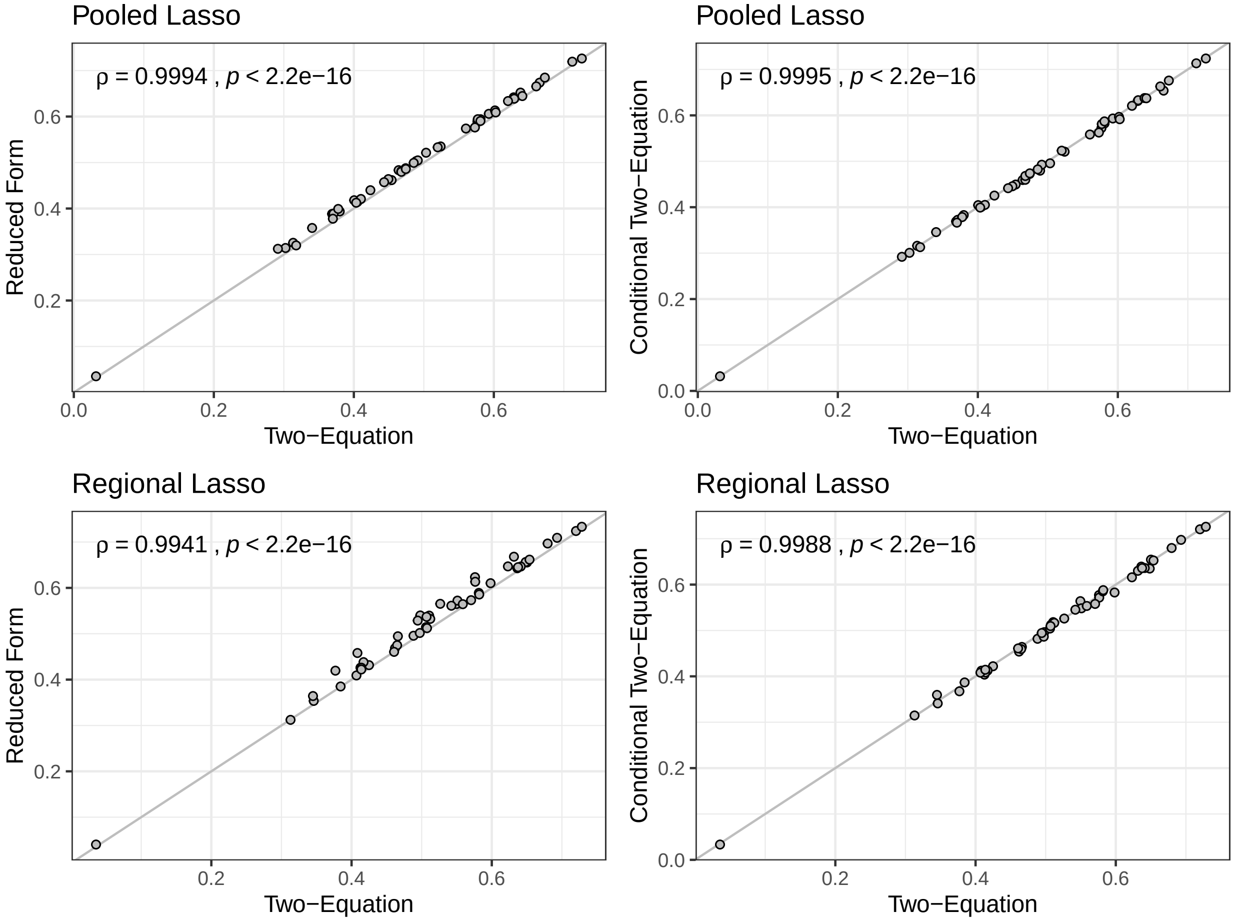
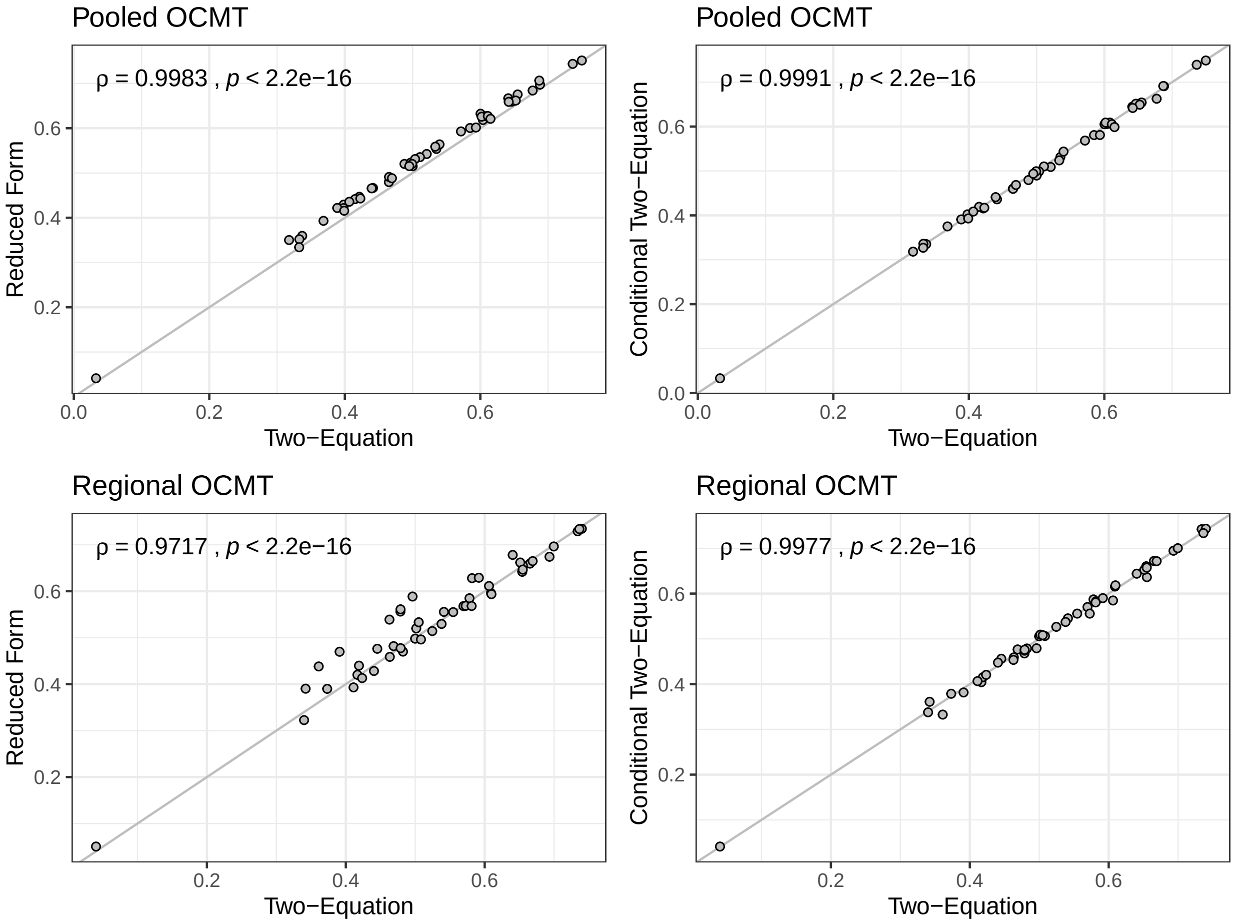


Figure S.10: 2020 State Republican Vote Share Forecasts, Two-Equation versus Reduced Form and Conditional Two-Equation using Lasso Algorithm



Republican vote shares are calculated as in Equation 12. Two-equation forecast refers to real-time baseline two-equation forecasts. Reduced form forecasts and conditional two-equation are described in Section 10.1. Reduced form forecasts are from a single vote share equation model which includes the union of covariates from both the turnout and vote share active sets. Two-equation and Reduced form forecasts use data available as of October 14, 2020. Conditional two-equation forecasts are from the two-equation baseline model estimated on data through October 14, 2020, but predicted turnouts are replaced with realized 2020 turnouts when calculating 2020 Republican vote share predictions. Sample correlations reported as ρ with corresponding p-value.

Figure S.11: 2020 State Republican Vote Share Forecasts, Two-Equation versus Reduced Form and Conditional Two-Equation using OCMT Algorithm



Republican vote shares are calculated as in Equation 12. Two-equation forecast refers to real-time baseline two-equation forecasts. Reduced form forecasts and conditional two-equation are described in Section 10.1. Reduced form forecasts are from a single vote share equation model which includes the union of covariates from both the turnout and vote share active sets. Two-equation and Reduced form forecasts use data available as of October 14, 2020. Conditional two-equation forecasts are from the two-equation baseline model estimated on data through October 14, 2020, but predicted turnouts are replaced with realized 2020 turnouts when calculating 2020 Republican vote share predictions. Sample correlations reported as ρ with corresponding p-value.

Table S.6: State Level Region-based Forecasts of Republican Vote Shares (V_s) for 2020 under the USDA-ARS 5-Region Classification

	USDA-Lasso		USDA-OCMT	
State	$\hat{V}_s$	EC Votes	$\hat{V}_s$	EC Votes
AK	N/A	3	N/A	3
AL	0.651	9	0.657	9
AR	0.659	6	0.664	6
AZ	0.499	0	0.504	11
CA	0.292	0	0.344	0
CO	0.423	0	0.438	0
CT	0.464	0	0.464	0
DC	0.039	0	0.041	0
DE	0.466	0	0.468	0
FL	0.489	0	0.497	0
GA	0.523	16	0.533	16
HI	N/A	0	N/A	0
IA	0.516	6	0.517	6
ID	0.644	4	0.673	4
IL	0.421	0	0.417	0
IN	0.591	11	0.595	11
KS	0.578	6	0.586	6
KY	0.641	8	0.645	8
LA	0.592	8	0.599	8
MA	0.373	0	0.375	0
MD	0.360	0	0.369	0
ME	0.495	0	0.503	4
MI	0.496	0	0.504	16
MN	0.454	0	0.456	0
MO	0.624	10	0.625	10
MS	0.590	6	0.598	6
MT	0.591	3	0.602	3
NC	0.508	15	0.518	15
ND	0.700	3	0.714	3
NE	0.641	5	0.661	5
NH	0.513	4	0.527	4
NJ	0.439	0	0.447	0
NM	0.470	0	0.491	0
NV	0.453	0	0.466	0
NY	0.363	0	0.370	0
OH	0.536	18	0.544	18
OK	0.686	7	0.696	7
OR	0.388	0	0.423	0
PA	0.545	20	0.553	20
RI	0.456	0	0.450	0
SC	0.575	9	0.590	9
SD	0.636	3	0.644	3
TN	0.647	11	0.647	11
TX	0.527	38	0.538	38
UT	0.590	6	0.621	6
VA	0.462	0	0.476	0
VT	0.361	0	0.366	0
WA	0.363	0	0.408	0
WI	0.515	10	0.517	10
WV	0.739	5	0.748	5
WY	0.739	3	0.747	3
All Votes		253		284

Republican vote shares are calculated as in Equation 12. 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 five separate panel regressions for the five USDA-ARS regions. Forecasts are generated *ex post* only using data available as of October 14, 2020.

Table S.7: State Level Forecasts of Republican Vote Shares (V_s) for 2020 under the Extended Active Set for Changes in Log Republican Odds ($DLRO$)

	Pooled-Lasso		Pooled-OCMT	
State	$\hat{V}_s$	EC Votes	$\hat{V}_s$	EC Votes
AK	N/A	3	N/A	3
AL	0.629	9	0.642	9
AR	0.629	6	0.647	6
AZ	0.495	0	0.521	11
CA	0.309	0	0.337	0
CO	0.405	0	0.422	0
CT	0.376	0	0.399	0
DC	0.032	0	0.033	0
DE	0.408	0	0.416	0
FL	0.467	0	0.488	0
GA	0.491	0	0.511	16
HI	N/A	0	N/A	0
IA	0.523	6	0.535	6
ID	0.665	4	0.677	4
IL	0.382	0	0.399	0
IN	0.583	11	0.604	11
KS	0.575	6	0.585	6
KY	0.639	8	0.655	8
LA	0.580	8	0.601	8
MA	0.296	0	0.318	0
MD	0.315	0	0.333	0
ME	0.456	0	0.465	0
MI	0.480	0	0.498	0
MN	0.450	0	0.466	0
MO	0.593	10	0.608	10
MS	0.584	6	0.602	6
MT	0.570	3	0.593	3
NC	0.488	0	0.504	15
ND	0.655	3	0.688	3
NE	0.598	5	0.611	5
NH	0.449	0	0.470	0
NJ	0.379	0	0.407	0
NM	0.412	0	0.442	0
NV	0.472	0	0.500	6
NY	0.346	0	0.369	0
OH	0.523	18	0.540	18
OK	0.671	7	0.687	7
OR	0.407	0	0.423	0
PA	0.473	0	0.500	0
RI	0.384	0	0.389	0
SC	0.561	9	0.572	9
SD	0.638	3	0.652	3
TN	0.621	11	0.642	11
TX	0.502	38	0.534	38
UT	0.600	6	0.615	6
VA	0.425	0	0.440	0
VT	0.320	0	0.333	0
WA	0.374	0	0.399	0
WI	0.477	0	0.496	0
WV	0.714	5	0.737	5
WY	0.722	3	0.750	3
All Votes		188		236

Republican vote shares are calculated as in Equation 12. 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. $DLRO$ equation forecast considers the extended active set which contains the variables found in Table 2 along with those in Table S.2. Forecasts are generated *ex post* only using data available as of October 14, 2020.

Table S.8: State Level Forecasts of Republican Vote Shares (V_s) and Electoral Votes for 2016 Elections

				Lasso				OCMT				Lasso-OCMT Average			
	Total EC	2016 Realized		Pooled Forecasts		Regional Forecasts		Pooled Forecasts		Regional Forecasts		Pooled Forecasts		Regional Forecasts	
State	(d_s)	V_s	EC Votes	$\widehat{V}_s$	EC Votes	$\widehat{V}_s$	EC Votes	$\widehat{V}_s$	EC Votes	$\widehat{V}_s$	EC Votes	$\widehat{V}_s$	EC Votes	$\widehat{V}_s$	EC Votes
AK	3	0.584	3	N/A	3	N/A	3	N/A	3	N/A	3	N/A	3	N/A	3
AL	9	0.644	9	0.635	9	0.649	9	0.635	9	0.642	9	0.635	9	0.646	9
AR	6	0.643	6	0.646	6	0.677	6	0.651	6	0.664	6	0.649	6	0.671	6
AZ	11	0.519	11	0.557	11	0.542	11	0.560	11	0.540	11	0.559	11	0.541	11
CA	55	0.339	0	0.395	0	0.396	0	0.392	0	0.424	0	0.394	0	0.410	0
CO	9	0.474	0	0.469	0	0.536	9	0.482	0	0.516	9	0.475	0	0.526	9
CT	7	0.429	0	0.403	0	0.444	0	0.410	0	0.433	0	0.406	0	0.439	0
DC	3	0.043	0	0.075	0	0.080	0	0.074	0	0.076	0	0.074	0	0.078	0
DE	3	0.440	0	0.392	0	0.434	0	0.410	0	0.413	0	0.401	0	0.424	0
FL	29	0.506	29	0.517	29	0.521	29	0.510	29	0.511	29	0.514	29	0.516	29
GA	16	0.527	16	0.557	16	0.568	16	0.561	16	0.566	16	0.559	16	0.567	16
HI	4	0.326	0	N/A	0	N/A	0	N/A	0	N/A	0	N/A	0	N/A	0
IA	6	0.551	6	0.498	0	0.514	6	0.505	6	0.507	6	0.502	6	0.510	6
ID	4	0.683	4	0.690	4	0.726	4	0.698	4	0.713	4	0.694	4	0.719	4
IL	20	0.410	0	0.429	0	0.448	0	0.439	0	0.484	0	0.434	0	0.466	0
IN	11	0.601	11	0.576	11	0.594	11	0.584	11	0.621	11	0.580	11	0.608	11
KS	6	0.611	6	0.625	6	0.657	6	0.628	6	0.665	6	0.626	6	0.661	6
KY	8	0.657	8	0.637	8	0.668	8	0.645	8	0.660	8	0.641	8	0.664	8
LA	8	0.602	8	0.605	8	0.621	8	0.607	8	0.615	8	0.606	8	0.618	8
MA	11	0.353	0	0.356	0	0.428	0	0.382	0	0.415	0	0.369	0	0.422	0
MD	10	0.360	0	0.365	0	0.395	0	0.377	0	0.371	0	0.371	0	0.383	0
ME	4	0.486	1	0.429	0	0.426	0	0.445	0	0.440	0	0.437	0	0.433	0
MI	16	0.501	16	0.468	0	0.495	0	0.476	0	0.529	16	0.472	0	0.512	16
MN	10	0.492	0	0.483	0	0.513	10	0.493	0	0.494	0	0.488	0	0.504	10
MO	10	0.617	10	0.584	10	0.622	10	0.591	10	0.622	10	0.588	10	0.622	10
MS	6	0.591	6	0.581	6	0.592	6	0.579	6	0.585	6	0.580	6	0.588	6
MT	3	0.611	3	0.577	3	0.626	3	0.593	3	0.625	3	0.585	3	0.626	3
NC	15	0.519	15	0.527	15	0.535	15	0.531	15	0.523	15	0.529	15	0.529	15
ND	3	0.698	3	0.642	3	0.634	3	0.633	3	0.620	3	0.638	3	0.627	3
NE	5	0.635	5	0.639	5	0.653	5	0.647	5	0.655	5	0.643	5	0.654	5
NH	4	0.498	0	0.471	0	0.499	0	0.488	0	0.493	0	0.480	0	0.496	0
NJ	14	0.427	0	0.400	0	0.448	0	0.413	0	0.414	0	0.406	0	0.431	0
NM	5	0.453	0	0.455	0	0.440	0	0.454	0	0.444	0	0.454	0	0.442	0
NV	6	0.487	0	0.495	0	0.487	0	0.503	6	0.534	6	0.499	0	0.509	6
NY	29	0.382	0	0.339	0	0.368	0	0.346	0	0.347	0	0.342	0	0.358	0
OH	18	0.543	18	0.506	18	0.535	18	0.511	18	0.565	18	0.508	18	0.551	18
OK	7	0.693	7	0.686	7	0.675	7	0.686	7	0.676	7	0.686	7	0.675	7
OR	7	0.438	0	0.461	0	0.459	0	0.464	0	0.506	7	0.463	0	0.482	0
PA	20	0.504	20	0.481	0	0.524	20	0.488	0	0.487	0	0.484	0	0.506	20
RI	4	0.417	0	0.349	0	0.395	0	0.360	0	0.391	0	0.355	0	0.393	0
SC	9	0.575	9	0.575	9	0.587	9	0.576	9	0.582	9	0.576	9	0.584	9
SD	3	0.660	3	0.621	3	0.638	3	0.627	3	0.640	3	0.624	3	0.639	3
TN	11	0.636	11	0.625	11	0.659	11	0.633	11	0.653	11	0.629	11	0.656	11
TX	38	0.547	36	0.600	38	0.564	38	0.601	38	0.574	38	0.601	38	0.569	38
UT	6	0.624	6	0.749	6	0.801	6	0.764	6	0.785	6	0.757	6	0.793	6
VA	13	0.472	0	0.483	0	0.482	0	0.496	0	0.468	0	0.490	0	0.475	0
VT	3	0.348	0	0.321	0	0.288	0	0.333	0	0.321	0	0.327	0	0.304	0
WA	12	0.412	0	0.444	0	0.446	0	0.445	0	0.495	0	0.444	0	0.470	0
WI	10	0.504	10	0.484	0	0.501	10	0.498	0	0.527	10	0.491	0	0.514	10
WV	5	0.722	5	0.668	5	0.684	5	0.661	5	0.682	5	0.664	5	0.683	5
WY	3	0.757	3	0.732	3	0.758	3	0.735	3	0.746	3	0.734	3	0.752	3
All Votes	538		304		253		308		265		307		259		330

Republican vote shares are calculated as in Equation 12. Column 'Total EC (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. Models generating 2016 forecasts were trained using the 2000-2012 sample.

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: Journal of the Econometric Society*, 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: Oxford University Press.