Why Vote in Person in a Pandemic? Using Machine Learning to Predict Voting Methods

Seo-young Silvia Kim[*†], Akhil Bandreddi[†], and R. Michael Alvarez[‡]

[†]American University

[‡]California Institute of Technology

November 24, 2022

Abstract

What spurs voters to vote in person, despite an established universal vote-by-mail (VBM) system and a once-in-a-century pandemic? We explore this question with official voter data from Colorado, a vote-by-mail state since 2013, but where 6% of voters still vote in person. Using multiclass classification, we analyze (1) the choice between voting by mail (VBM), voting in person, and not voting in the 2020 general election, and (2) the choice to switch to in-person voting despite having used VBM in previous cycles. The results suggest that the choice of voting modes is mainly habitual; variables such as registration date, active status, and past choices of turnout and voting mode are among the top most important variables. Local variations of COVID-19 and demographics hardly mattered. Notably, Republican partisanship plays an important role in predicting "switchers" to in-person voting; indeed, the probability of switching to in-person voting was 5.2% conditional on being a Republican as opposed to 1.9% conditional on being a Democrat.

Word Count: 9,985

What spurs voters to vote in person? Despite high confidence in the established vote-by-mail (VBM) system and a raging pandemic, could factors such as nationalized partisan

[*] Kim is an Assistant Professor at American University and is the corresponding author (email: sskim@american.edu). Bandreddi is an undergraduate at American University. Alvarez is a Professor of Political and Computational Social Science at the Division of Humanities and Social Sciences, Caltech. An earlier version of this paper was presented at Election Sciences, Reform, & Administration Conference 2021. We thank the conference participants for their comments.

rhetoric be casting doubt on VBM drive voting behaviors? More fundamentally, can we better predict the choice of voting modes so that elected officials can conduct elections more smoothly? We use supervised machine learning to train models that best predict the choice of voting modes and who switched to voting in person during the 2020 general election in Colorado. By evaluating the prediction performance and the relative importance of variables, we hope to understand the choice of voting modes and whether trust in electoral systems has been systematically undermined. In addition, we assess whether statistical predictions can help efficiently allocate resources in elections.

Understanding voting behaviors around VBM is crucial, especially since many jurisdictions have either recently adopted or are attempting to adopt universal VBM (Gorman, 2021; NCSL, 2022; Chase, 2022). Understanding what predicts decisions to vote in person, even in such a system, will inform both theories of voting and election administration policies. In particular, it is important to understand how partisanship and local variations of the pandemic may have altered voters' choices, given a consistent elite and partisan rhetoric about how VBM might not be trusted. This will help estimate (1) how election officials must allocate resources in a given election and (2) how effective universal VBM will be for other states.

Studying Colorado, in particular, is an important exercise. Although certainly later than Oregon (2000) or Washington (2011) to turn all-mail, Colorado voters have had seven years to get used to voting at home, as Colorado has used all-mail voting since 2013. But despite this transition to all-mail voting, a persistent portion of voters continues to vote in person; 6.0% of voters—amounting to nearly 198,000 voters—still voted in person in the 2020 general election even with the COVID-19 pandemic, compared to 4.7% in 2018, 7.0% in 2016, and 4.9% in 2014. Choosing to vote in person by itself is not something that should be discouraged, as voters are free to choose the mode of voting. However, understanding what determines the decision *not* to adopt a voting mode that is widely perceived to be more convenient has theoretical and practical implications for political participation, especially if it is swayed strongly by partisanship. This also has important ramifications for election administration in Colorado (and other states using all-mail voting), as election officials need to plan for in-person voters and allocate sufficient resources for them in each election cycle.

To analyze voting choices on VBM, we first train a model that predicts voters' actual decisions about how to cast their ballot in the 2020 general elections. We model voting choices between the following three categories: voting by mail, voting in person, or not voting at all. This approach allows for a 'defection,' i.e., not voting, to account for situations in which some voters are reluctant to vote by mail but find in-person options also uninviting. Using supervised machine learning algorithms, we fit a multiclass classification model under the training-test set paradigm (Efron, 2020), using a hold-out set to evaluate out-of-sample predictive performance. Similarly, we investigate which voters have had experience voting by mail since all-mail voting was implemented but have decided to "switch back" to in-person voting during the 2020 general election. We use a machine-learning approach for a number of reasons. First, we want to take advantage of machine learning's training-validation approach, particularly because it is a robust methodology for modeling large and complex datasets (James et al., 2013). Second, the outcome variables are highly imbalanced, and traditional multinomial logistic regressions perform very poorly in such datasets (King and Zeng, 2001).

Overall, the choice of voting modes seems to be mainly habitual. Just as voting is said to be habitual (Gerber et al., 2003; Aldrich et al., 2011), best predictions of the choice of voting modes are highly dependent on data on past choices and voters' registration dates. Other variables, such as demographics and COVID-19 prevalence, are secondary to no importance when predicting the choice of voting modes.

However, partisanship was ranked highly in terms of variable importance when predicting switchers to in-person voting in 2020, although not so much for voting modes overall. Note that in the 2020 general election, coming out to vote in person at a polling place is both a political behavior and a health behavior. And given that the Republican electorate is generally older, placing them at higher risk for COVID-19, this evidence of partisanship altering voters' behavior is concerning. This may indicate some erosion of faith in the electoral system, even in states such as Colorado, where voters are used to the universal vote-by-mail system.

Room for improvement remains due to data limitations. Given (1) an administrative dataset that accurately reflects turnout and voting mode choices but is thin on available covariates and (2) the class imbalance due to a small proportion of voters voting in-person, the performance of predicting in-person voters overall is—although certainly improved from random guessing—relatively poor, even after adjustments such as downsampling. We conclude that more work is needed to determine the habit formation of voting modes to help allocate administrative resources. Moreover, the effect of partisanship in altering voters' choice of voting modes should be closely monitored.

Voting By Mail

Much of the literature on VBM has focused on whether VBM, as a convenience voting measure, expands turnout to nonvoters and makes voting more representative. Consistent with evidence on other cost-reducing measures, empirical evidence is mixed. Some have documented that the more resourced electorate—who are traditional/likely voters—benefit the most, so that voting does not become more representative (or even more unrepresentative) (Berinsky et al., 2001; Karp and Banducci, 2000; Southwell and Burchett, 2000). Kousser and Mullin (2007) find, using a natural experiment, that those assigned to VBM, in fact, turn out at lower rates; however, their findings support Southwell (2009)'s finding that VBM increases turnout in special elections. Bonica et al. (2021) analyzing the 2014 all-mail voting implementation in Colorado find that there was an increase of approximately 8 percentage points in additional ballots cast between 2014 and 2018, especially for lower propensity groups, such as the young, blue-collar workers, voters with less education, and voters of color. However, Thompson et al. (2020) shows that universal vote-by-mail does not increase average turnout rates or either party's vote share, using a staggered rollout of VBM within California, Utah, and Washington.

Then, who chooses to vote by mail as opposed to in-person voting? While pundits have argued that VBM is a Democrat-skewed, highly partisan choice, the evidence on that has been mixed. Southwell (2007), assessing Oregon in 2003, found that it was Republicans and Independents who favored VBM, as well as older and female voters. Studies such as Alvarez et al. (2012) and Kousser et al. (2020) found that there was no significant divide in the choice to vote by mail between Democratic and Republican voters, although policy support to further advance VBM may be partisan divided (Kousser et al., 2020).

However, the evidence has shifted in recent years. Niebler (2020) finds that allegiance to Trump and lack of concern about COVID-19 made people less likely to vote by mail in the 2020 primaries. Examining VBM in Maine, Herron and Smith (2021) find that given that a larger share of VBM ballots was cast by Democrats and unaffiliated voters, they were more vulnerable to postal delivery disruptions than those cast by Republicans. Lockhart et al. (2020) shows that it is not only the support levels for no-excuse absentee ballots that are polarized but also the personal preference to vote by mail. In their experiment, fewer Republicans were willing to choose VBM within just two months' time, while the opposite happened to Democrats. The authors argue that the 2020 election trend may be fundamentally different from historical patterns, which runs counter to the existing scholarly knowledge that interest in VBM is not polarized along partisan lines. In contrast, however, McGhee et al. (2022) has argued that expanding VBM has no robust partisan effects.

Some have offered other explanations for the choice of voting modes other than partisanship. Plescia et al. (2021) find that the preference to vote in person is not ideological but generational. They find that older White voters in the U.S. with little interest in politics prefer VBM the most, while younger Black or Latino voters with a high interest in politics prefer it the least. Examining varying support for absentee ballots, Dominguez et al. (2020) find that women are more likely to support absentee ballots compared to more restrictive voting methods. Menger and Stein (2020) find that the strongest determinant of a voter's voting choice is their trust in the United States Postal System as well as prior voting history. Atkeson et al. (2022) finds that the pandemic voting mode decisions were influenced both by partisanship and by age, the latter of which is highly correlated with post-pandemic risk.

This paper extends and fills the gap in the literature in several ways. First, we look specifically at a case where vote-by-mail has been institutionalized and established as part of voters' experience, where voting in person is a much more conscious, rare choice that voters are making. Second, we use official voter registration data at the individual level instead of county-level, state-level, or survey data. This allows us to leverage both its scale as well as accuracy. Third, when modeling the voter's decision of vote choices, we model it as a threeway decision, allowing for defection and a complete picture of the voter's decision. Fourth, we also model the voters' decision to *switch* to in-person voting, despite having experienced VBM, which has not been explored beforehand.

Background: Vote-by-mail in 2020

The state of Colorado, in the chaos of the 2020 election—partisan animosity running high against an incumbent president, COVID-19 running amok, and a highly nationalized Senate election between John Hickenlooper and Cory Gardner—has shown record-breaking turnout at 86.87% of active registered voters, or 73.25% of the voting-age population. Because Colorado has been an all-mail elections state since 2013, most of its ballots in recent elections were cast by mail. There were no particularly significant instances of election administration failures in Colorado, with media praising its VBM implementation as pandemic-proof (Corse, 2020). Still, 6.0% of registrants who cast a ballot in the 2020 elections still voted in person at a voting booth—nearly 198,000 voters.

It may be surprising that 198,000 Colorado voters were willing to vote in person during

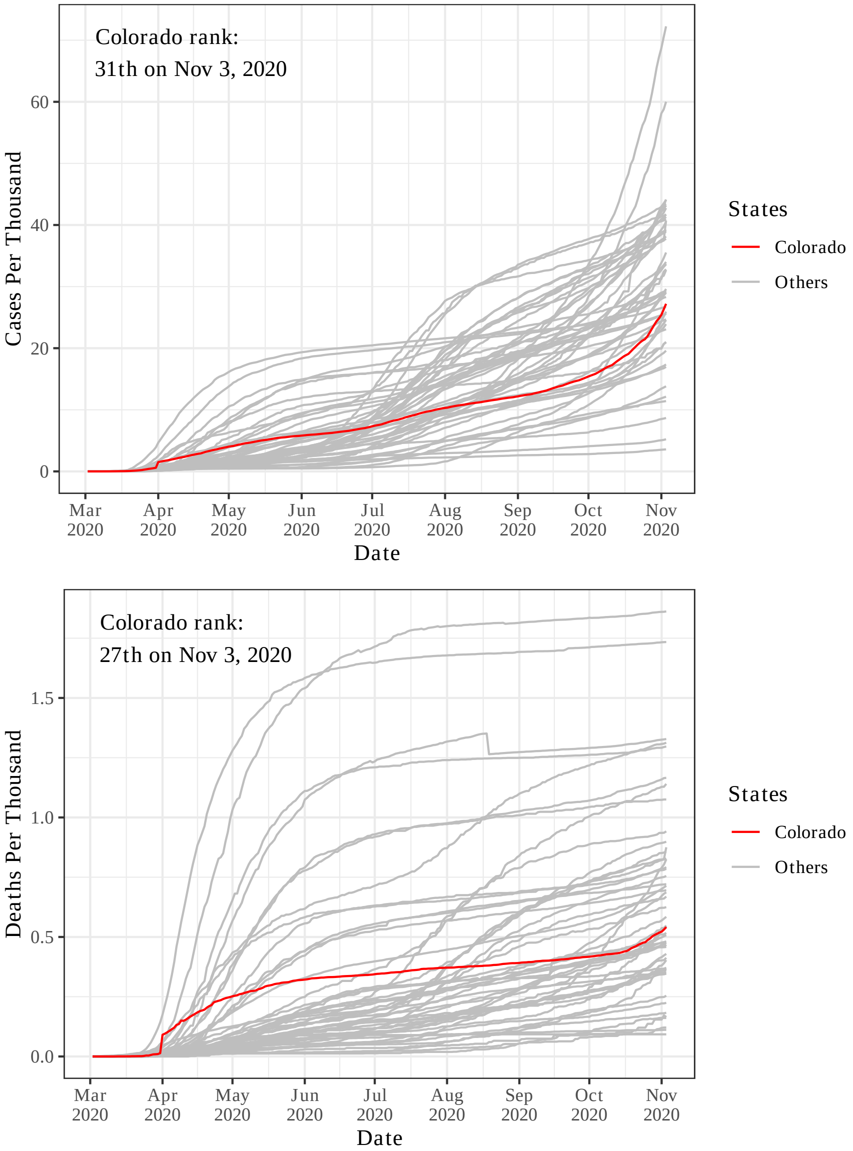


Figure 1: COVID-19 Cases and Deaths Per Thousand, Mar 1 to Nov 3, 2020

the COVID-19 pandemic. On the one hand, in terms of COVID cases per capita, Colorado ranked 40/51 states, including the District of Columbia. In terms of deaths per capita, it ranked 33/51 states. But the number of cases rapidly rose until Election Day across the country, and Colorado was no exception. On Oct 1, 2020, there were 71,218 cases, a 23.3% increase from Sep 1. In the last week leading up to the election, several local newspapers reported that (Wingerter, 2020; Paul, 2020) the continuing trends meant record hospitalizations by early November—which ultimately proved to be true. Governor Polis was pleading for Coloradoans to limit social gatherings, emphasizing the increasing positivity rate and hospitalizations, well before early voting began on Oct 19, 2020.[1] Although paled in the excitement of the general election, there was no doubt that the surging number of cases was a salient issue.

On the national level, elite political communication surrounding COVID-19 has been highly partisan polarized (Green et al., 2020; Grossman et al., 2020), and it was shown that this had created significant partisan differences in health behavior and policy attitudes early in the pandemic (Grossman et al., 2020; Allcott et al., 2020; Gadarian et al., 2021). In addition, then-President Trump strongly pushed inaccurate rhetoric that claimed that voting by mail was a source of fraud (Saul and Epstein, 2020). In September 2020, for example, Trump said in a rally:

> Mail ballots, they cheat ... Mail ballots are very dangerous for this country because of cheaters. They go collect them. They are fraudulent in many cases.

This was only a small component of a larger misinformation campaign claiming that voter fraud was rampant. And there is evidence in the literature that misinformation did change beliefs about the legitimacy of the election (Jurkowitz, 2021; Berlinski et al., 2021), with increasing exposure to the false claims generating an increase in false perceptions that voter fraud is a common (Udani et al., 2018). Such misinformation may also have eroded support for democratic norms (Albertson and Guiler, 2020) unless corrected with factual information (Porter and Wood, 2019). With polarized rhetoric about VBM, it is clear that partisanship may be an important factor driving turnout and the choice of voting modes in states like Colorado.

Data and Methodology

Official Voter Files of Colorado. The data we use comes from the official voter records in Colorado between 2014–2020 from all 64 counties. We use administrative data as opposed to survey data such as in Alvarez et al. (2012), Kousser et al. (2020), or Lockhart et al. (2020) primarily because voters' recall of their past voting method is assumed to be imperfect. In addition, we intend to measure the true choices as opposed to the willingness to choose certain voting methods. Although this means that we cannot access variables typically unmeasured by local election officials, using large-scale administrative data is both highly accurate and extremely valuable.

By merging voting history files across elections and joining them to a master file of

[1] See https://www.colorado.gov/governor/news/3241-governor-polis-public-health-officials-provide-update-covid-response-discusses-plan.

registered voters, we obtain a record of approximately 3.7 million unique voters who have at least one turnout history record in 2014–2020 while being registered for 2020.[2]

Category	Variables and Values
Personal Information	Age, Gender, Year of Birth
Voting Information	Registration Date, County, Congressional District, Party
Turnout History	Primary and General Election Turnout and Voting Mode, 2014–2020
County-level Data	Urban/Rural Designation, 2016 Presidential Vote Share and Winner, COVID-19 Cases and Death on the Day of First VBM

Table 1: Set of Variables Used

Each observation carries an individual's voting history—their voting modes in each election, personal information (such as voter ID, year, gender, and residential information), and administrative variables such as registration date, county, congressional district, and 2018 county designations (urban, rural, and frontier). If missing formal partisan affiliation in the original master voter file or coded as third-party, partisan leaning was further imputed from voters' primary election records. Table 1 shows the variables used.

Voting Mode Choice in the 2020 General Election. While there were multiple unique voting methods in the original data, such as whether the voter voted at a vote center, for brevity and consistency, we binned the records into either *in-person* or *vote-by-mail*. With Colorado being a vote-by-mail state since 2013, there existed a strong class imbalance in the outcome variable–choice of voting mode. Only about 5.3% (197,882) chose to vote in person,[3] with 83.0% (3,082,365) choosing to vote by mail and the remaining 11.6% (432,390) not turning out. Among those in-person voters, 92,485 (46.7%) previously had a mail ballot turned in, who we defined as a 'switcher' who changed voting modes from mail to in-person. Within the set of voters who actually voted, the proportion of switchers was 3.5%.

Table 2 moves on to take a look at the top ten individual voting patterns across the last three election cycles, while Table 3 shows the same table but conditional on having voted in person during the 2020 general elections. Both tables are organized so that the most frequently used combinations of voting modes are in the top rows. Table 2 shows us that

[2] We exclude voters who have never voted during this period of registration due to the lack of certainty that these voters are still residing and eligible to vote in Colorado. The official numbers of registered voters and ballot counts for the 2020 general election are available at the Secretary of State's website (https://www.sos.state.co.us/pubs/elections/Results/Abstract/2020/general/turnout.html). Because Colorado explicitly keeps a snapshot of past elections at the certification point, attrition bias from transitory voter files as pointed out in Kim and Fraga (2022) are expected to be minimal.

[3] Note that this number is smaller than the aforementioned 6.0% because we are no longer restricting our data to those who did vote in the 2020 general.

Gen. 2020	Pri. 2020	Gen. 2018	Pri. 2018	Gen. 2016	Pri. 2016	Obs.	Dem (%)	Rep (%)	Others (%)
Mail	Mail	Mail	Not voted	Mail	Not voted	496,069	30.1	25.0	44.9
Mail	Mail	Mail	Mail	Mail	Not voted	457,086	38.0	20.6	41.4
Mail	Mail	Mail	Mail	Mail	Mail	403,073	43.7	49.6	6.7
Mail	Not voted	Mail	Not voted	Mail	Not voted	317,972	18.9	31.7	49.4
Mail	Not voted	Not voted	Not voted	Mail	Not voted	157,286	23.6	27.4	49.0
Mail	Mail	Mail	Not voted	Not voted	Not voted	104,162	35.4	17.5	47.1
Mail	Not voted	Mail	Not voted	Not voted	Not voted	72,214	22.8	22.4	54.8
Mail	Mail	Mail	Not voted	Mail	Mail	65,543	34.6	53.8	11.6
Mail	Mail	Not voted	Not voted	Mail	Not voted	60,506	28.9	25.5	45.6
Mail	Not voted	Mail	Mail	Mail	Not voted	46,043	21.2	36.3	42.5

Table 2: Top-10 Individual-level Voting History Patterns, Last Three Cycles

all of the top ten combinations of voting modes involved voting by mail or abstention. For example, nearly half of a million Colorado voters voted by mail in the 2016 general election, the 2018 general election, and in the 2020 primary and general elections—but abstained in the 2016 and 2018 primary elections.

However, once we condition on voting in person in the 2020 general election, we see that most of those voters either voted by mail or abstained in recent past elections. That said, towards the bottom of Table 2, we see that some of the 2020 general election in-person voters have voted in-person in past elections, but still, those voters had some experience voting by mail in at least one recent past election.

Gen. 2020	Pri. 2020	Gen. 2018	Pri. 2018	Gen. 2016	Pri. 2016	Obs.	Dem (%)	Rep (%)	Others (%)
In person	Not voted	Not voted	Not voted	Mail	Not voted	14,087	18.6	36.4	45.0
In person	Not voted	Mail	Not voted	Mail	Not voted	13,234	14.7	42.7	42.6
In person	Mail	Mail	Not voted	Mail	Not voted	11,372	14.2	49.1	36.7
In person	Mail	Mail	Mail	Mail	Not voted	6,972	16.7	48.7	34.6
In person	Not voted	Mail	Not voted	Not voted	Not voted	5,316	17.2	33.3	49.5
In person	Mail	Mail	Mail	Mail	Mail	5,002	14.2	79.0	6.8
In person	Not voted	Mail	Not voted	In person	Not voted	4,556	15.0	44.1	40.8
In person	Not voted	In person	Not voted	Mail	Not voted	3,318	19.3	39.8	40.8
In person	Mail	Mail	Not voted	Not voted	Not voted	3,266	19.6	38.1	42.3
In person	Mail	Mail	Not voted	In person	Not voted	2,681	16.2	51.1	32.7

Table 3: Top-10 Individual-level Voting History Patterns, Last Three Cycles, Conditional on In-person Voting in 2020 General Elections

Gradient Boosting. To predict voters' choices, we use supervised machine learning. This is because when traditional multinomial logit regressions are applied to a hugely imbalanced dataset, the inaccuracies in predicting rare outcomes—the 'minority' class of outcomes that we are interested in—are too great (King and Zeng, 2001). For example, multinomial logit would predict that just 0.15488% of voters would vote in-person as opposed to 5.3%, and for those who did vote, that just 0.00004% of them would switch from VBM to in-person as opposed to 3.5%.[4] As our focus is on how to predict the rare events, we use a machine learning approach instead—indeed, Wang (2019) has shown that gradient-boosted trees perform well in rare event modeling. For an overview of the training/testing paradigm and its comparison against traditional regression methods, see Efron (2020), and for an overview of using machine learning and tree-based methods in social sciences, see Grimmer et al. (2021) and Montgomery and Olivella (2018).

For training, we used the gradient boosting machines (GBM) (Ridgeway, 1999; Montgomery and Olivella, 2018; Ridgeway, 2020; Greenwell et al., 2022).[5] Unlike bagging, boosting builds trees sequentially (James et al., 2013; Montgomery and Olivella, 2018) without bootstrap sampling. The algorithm works, using Friedman's gradient descent algorithm, in iterative steps (Friedman et al., 2000; Friedman, 2001, 2002; Ridgeway, 2020) to fit an additive model to the residuals to improve where the model's predictions are missing the mark. Once the negative gradient of the loss function has been calculated, a new tree fit to predict the negative gradient from the covariates, which is then regularized with a shrinkage factor (i.e., step size reduction). Our model is trained using 10-fold cross-validation (CV) across parameters such as the number of trees and the depth of the interaction. This approach allows robust variable selection—in other words, gradient boosting with cross-validation lets us discern which of the covariates are the best predictors of which mode to be used by Colorado voters in the 2020 general election.

Class Imbalance Correction and Model Metrics. As was seen earlier, our data suffers from a major class imbalance in the variables our models aim to predict. In social sciences, this is better known as the problem of rare events (King and Zeng, 2001). In the machine learning setting, approaches like gradient boosting assume that the distributions of the modeled outcomes are relatively similar (or balanced): for example, when there are three outcomes, that approximately a third of the observations in the dataset fall into each outcome. Class imbalance issues have long been a concern in the machine learning literature, as serious class imbalances erode the ability of machine learning algorithms to predict outcomes accurately (He and Garcia, 2009).

In order to remedy the class imbalance in our data, we resort to a commonly-used solution, which is downsampling the majority class (here, VBM voters when predicting voting modes and non-switchers when predicting switchers). We selected the degree of downsampling by testing different levels from 5 to 25% in increments of 5%. In the

[4] See Appendix.

[5] As we will show in the Appendix, GBM outperformed random forests consistently by a small margin, so we only present GBM results here.

end, we chose the models that maximized the area under the precision-recall curve (PR AUC), which was the training metric when cross-validating.

Our choice to use PR AUC, as opposed to the standard ROC curve, comes from the highly imbalanced nature of the dataset. In the ROC curve, the false positive rate and true positive rate are used as the axes; however, in the case of imbalanced data where the model would accurately classify most of the testing data, the area under this ROC curve would be highly inflated. This gives us a misleading performance estimate, and in particular, it does not represent the model's inability to properly predict the minority class (Saito and Rehmsmeier, 2015). A precision-recall curve on the other hand, while also using the true positive rate, or recall, additionally uses precision: the proportion of data classified as positive that are indeed positive. Precision and the true positive rate serve the purpose of penalizing the model in the form of a reduced area under the curve when it misclassifies the majority of the minority class. Furthermore, Davis and Goadrich (2006) shows that "a curve dominates in ROC space if and only if it dominates in PR space." Therefore, a PR curve would give a more conservative estimate of the model's performance even in a worst-case scenario.

Results

Descriptive Analyses

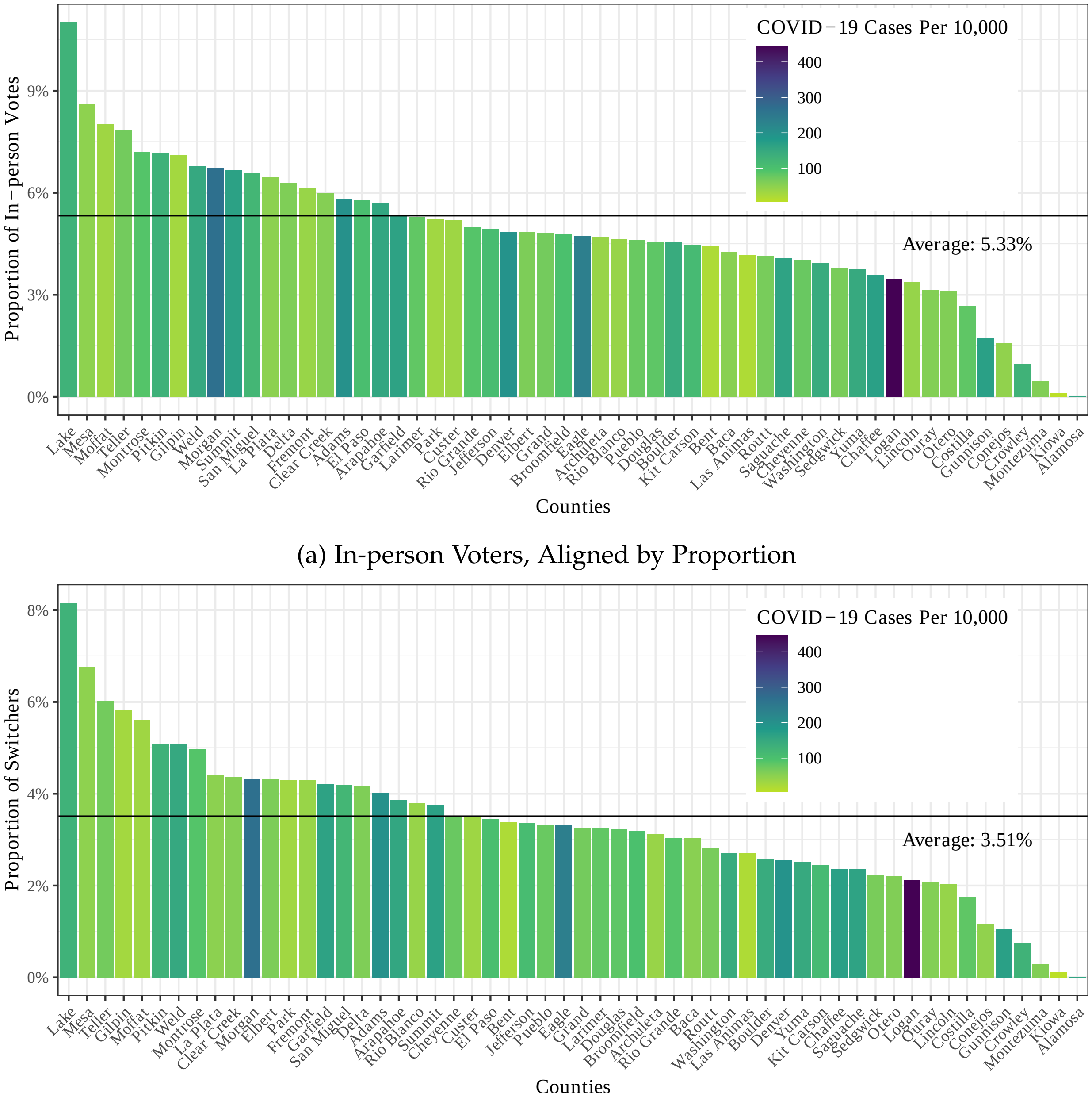
We first perform some descriptive analyses to understand the dataset better.

Distribution Across Partisan Affiliations. As can be inferred from Tables 2–3, the descriptive patterns show that partisanship seems to matter significantly regarding which voting mode was used in the 2020 general election. Indeed, conditional on being a Republican, the probability of voting in person was 6.9%, as opposed to 3.4% when conditional on being a Democrat. In addition, the probability of switching to in-person voting was 5.2% conditional on being a Republican as opposed to 1.9% conditional on being a Democrat. Table 4 gives a full cross-tabulation of voters' choices and parties.

	Democrat	Republican	Other
Mail	26.3	33.4	23.3
In person	1.1	2.3	2.0
Not voted	3.4	5.1	3.2

Table 4: Crosstab of 2020 General Turnout/Voting Modes and Partisan Affiliations

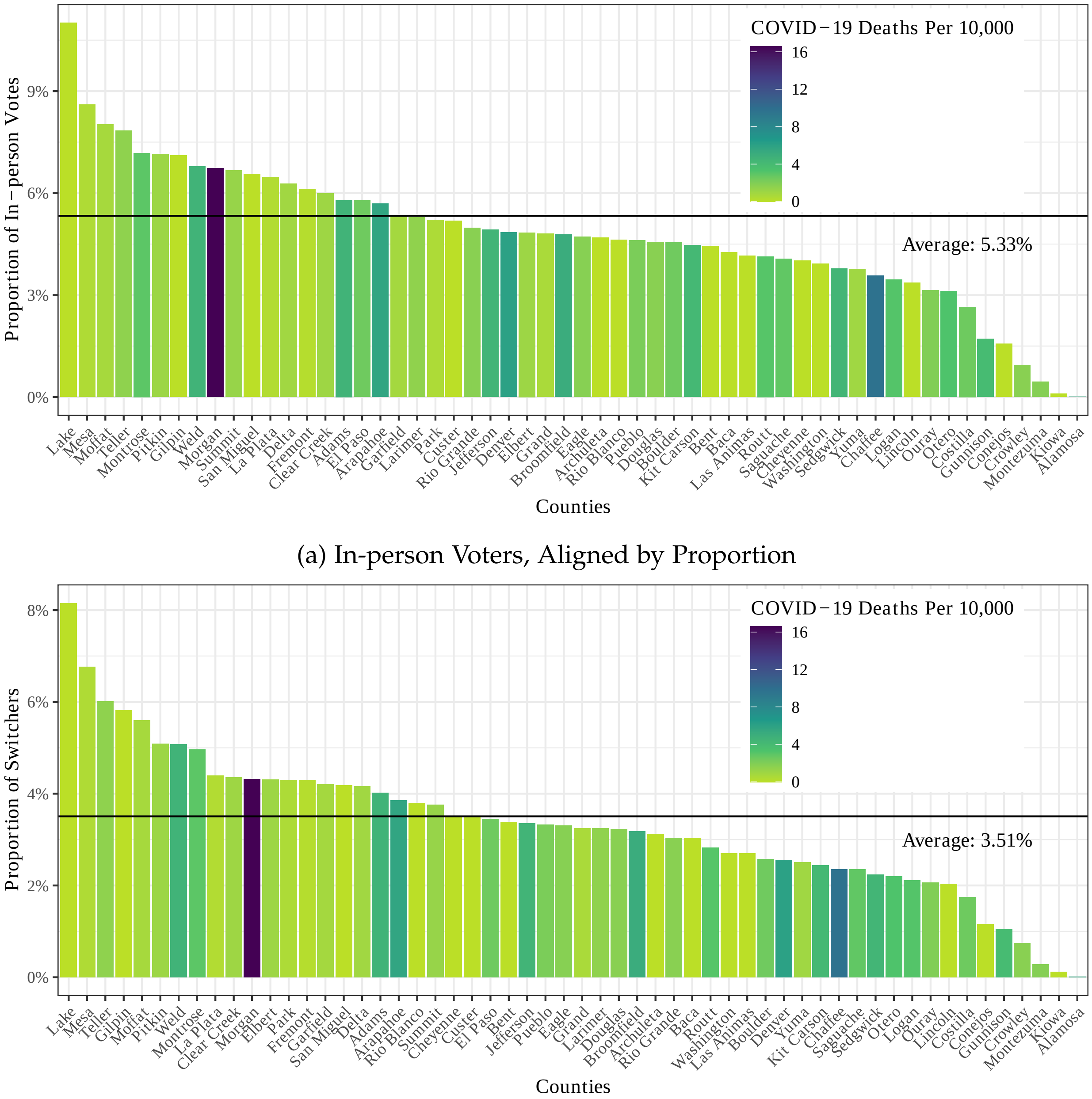
Distribution Across Counties. Next, we examine whether the choice of voting modes varies significantly across Colorado counties and county-level COVID-19 case prevalence. Given our theorizing that COVID-19 may have changed the dynamics in the 2020 general election, Figures 2 and 3 show county-level cases and deaths by the first day of VBM and the outcomes of interest.



(a) In-person Voters, Aligned by Proportion

(b) Switchers, Aligned by Proportion

Figure 2: Proportion of In-person Voters and Switchers by County and COVID-19 Case Prevalence



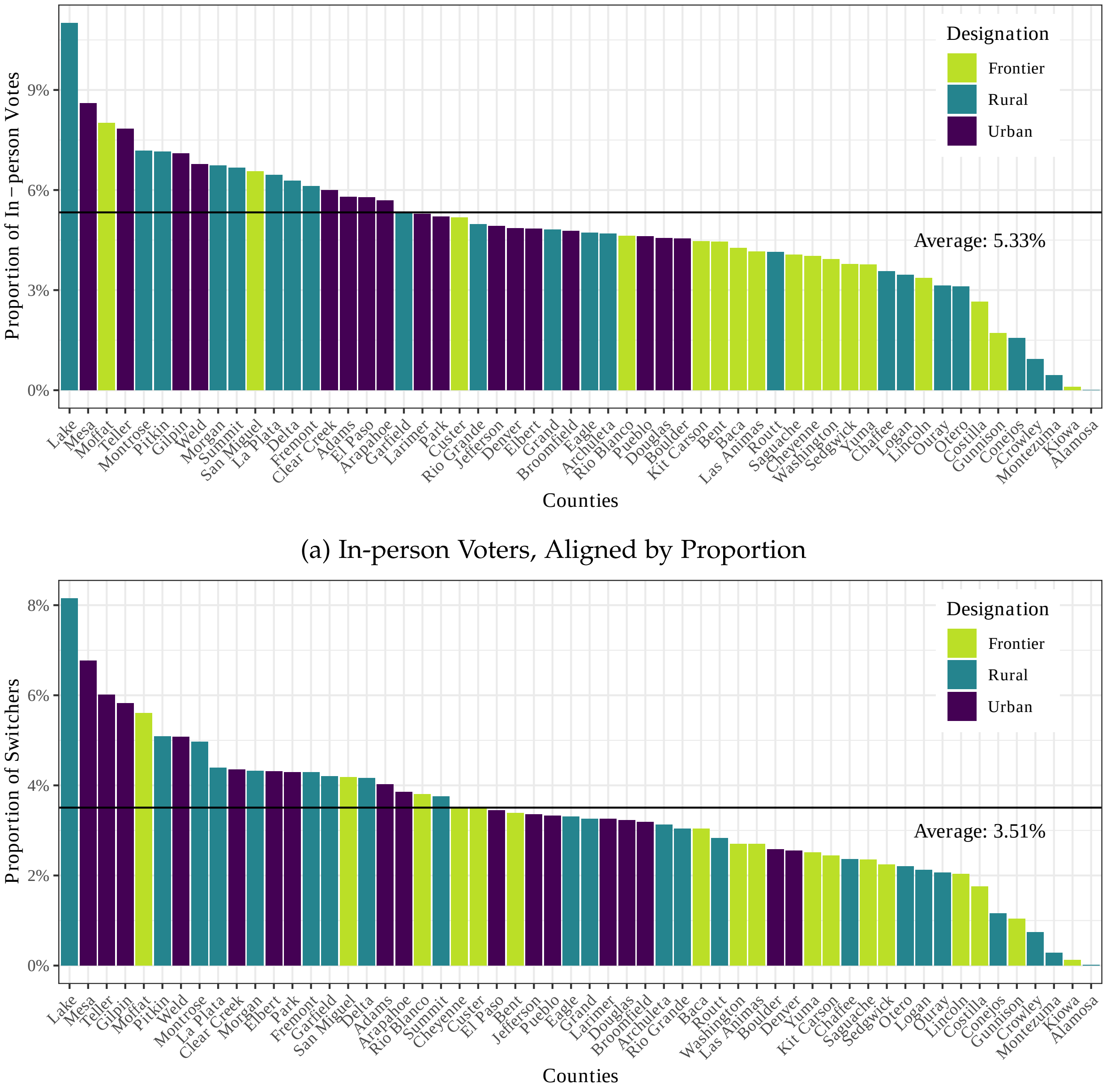
(a) In-person Voters, Aligned by Proportion

(b) Switchers, Aligned by Proportion

Figure 3: Proportion of In-person Voters and Switchers by County and COVID-19 Death Prevalence

The proportion of in-person voters or switchers by county was fairly normally distributed, as we notice in Figure 4.[6] Visually, there does not seem to be a strong relationship between the choice to vote in person and local COVID prevalence or morbidity.

[6] Indeed, when tested with either the Anderson-Darling test for normality or the Cramer-von Mises test, we cannot reject the null that the distribution is normal for both cases.

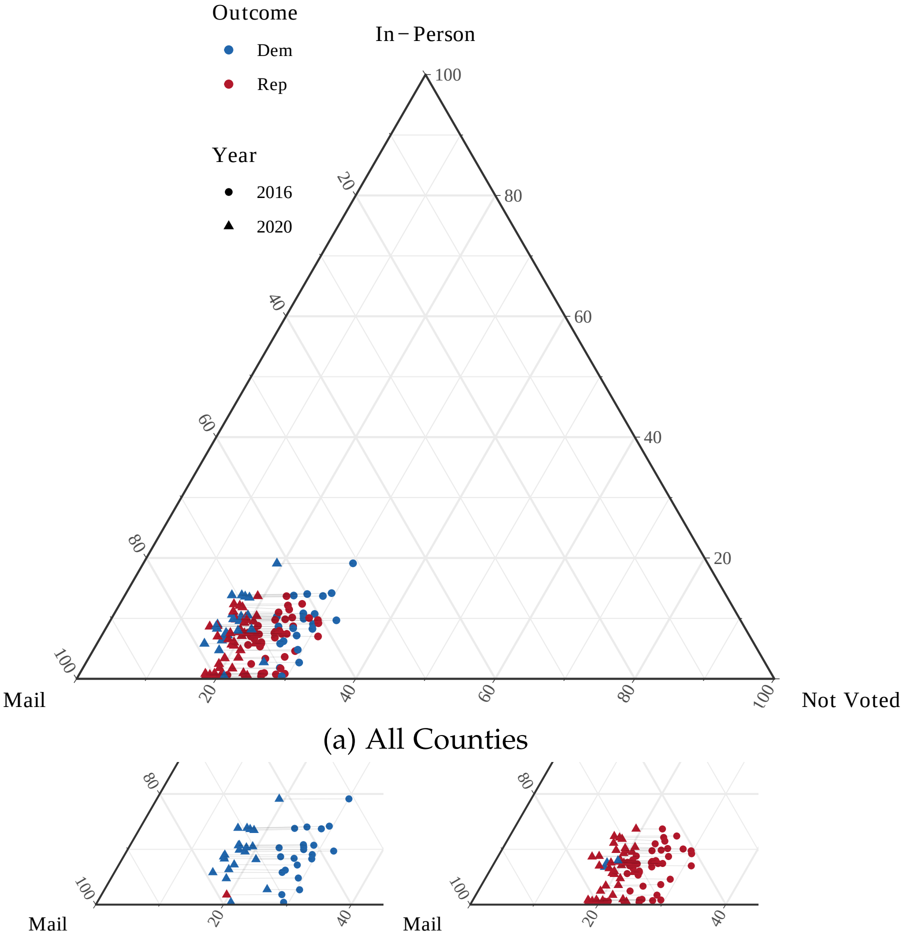


(a) In-person Voters, Aligned by Proportion

(b) Switchers, Aligned by Proportion

Figure 4: Proportion of In-person Voters and Switchers by County/County Designation

We also check for differences in urban/rural/frontier designation given works on group consciousness-based perspectives such as Walsh (2012) that may feed into political behavior in a hyperpartisan election such as the 2020 general. Although the highest proportion for both variables occurred in Lake County, which is rural, overall, no particular pattern by urban-rural designation seems to appear. Indeed, other than frontier counties—counties with very sparse populations—having lower proportions of in-person voters (3.77%) and switchers (2.51%), urban counties had 5.41% in-person voters and 3.54%



(a) All Counties

(b) Dem. Counties

(c) Rep. Counties

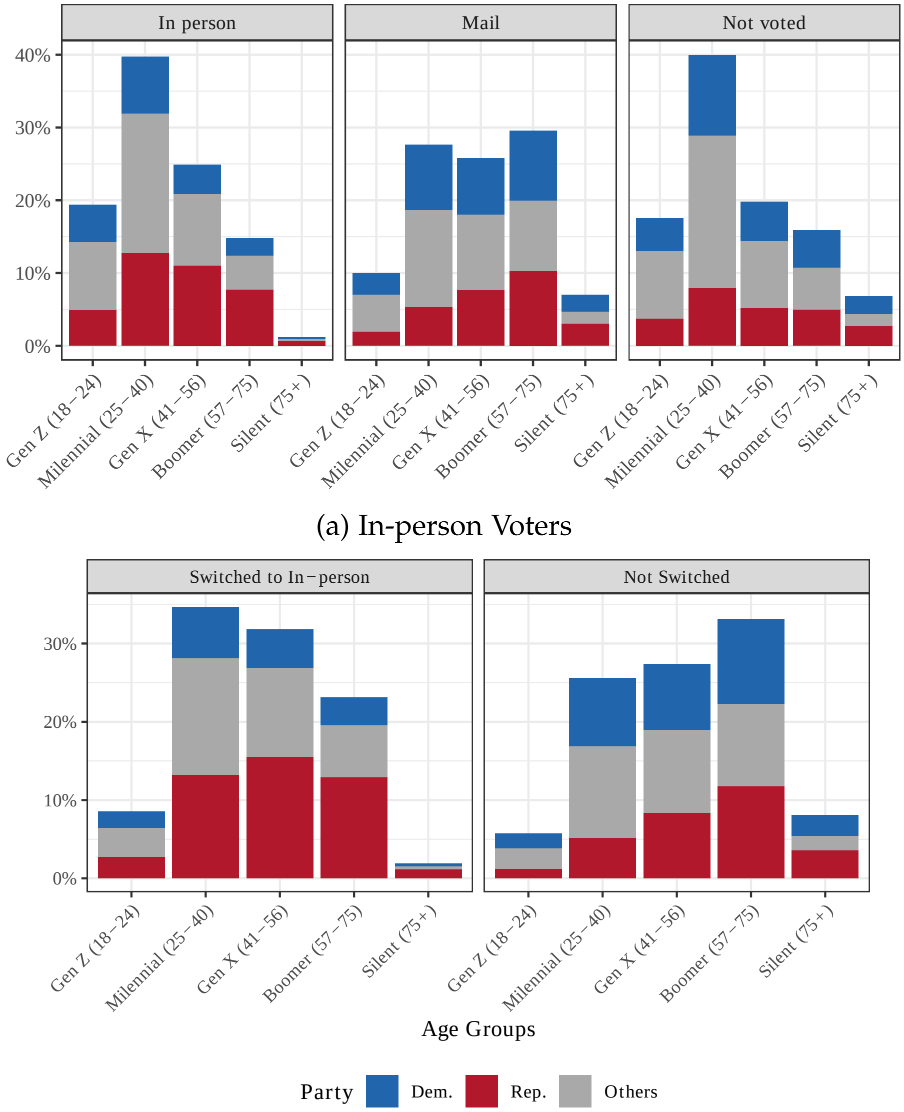
Figure 5: Ternary Plot for County-level Choice of Turnout/Voting Mode, All 64 Counties in Colorado, 2020 General Election

switchers, while rural counties had 5.04% in-person voters and 3.43% switchers, which were quite similar.[7]

We then replicate Persily and Stewart (2021)'s ternary plot with all sixty-four Colorado counties to compare their mix of voting modes between 2016 and 2020. Unlike the original paper, we put early and Election Day in-person voting into a single category and show nonvoting as the final choice. Figure 5 shows this for all registered voters as well as broken down by the presidential winner in 2016 (22 'Democratic' counties, 42 'Republican' counties) and magnified to better show the cluster of data.

[7] Given the large size of the data, they are still statistically significant.

Vote-by-mail predominates all counties in our dataset in both election cycles, but generally speaking, the rate of in-person voters has not shifted within counties. Rather, the movement is 2016 nonvoters shifting to vote by mail in 2020, which is consistent with the overall in-person voting rate in 2016 and 2020, as well as that Colorado's certified turnout as a percentage of registrants was 74.39% in 2016 while 78.16% in 2020.

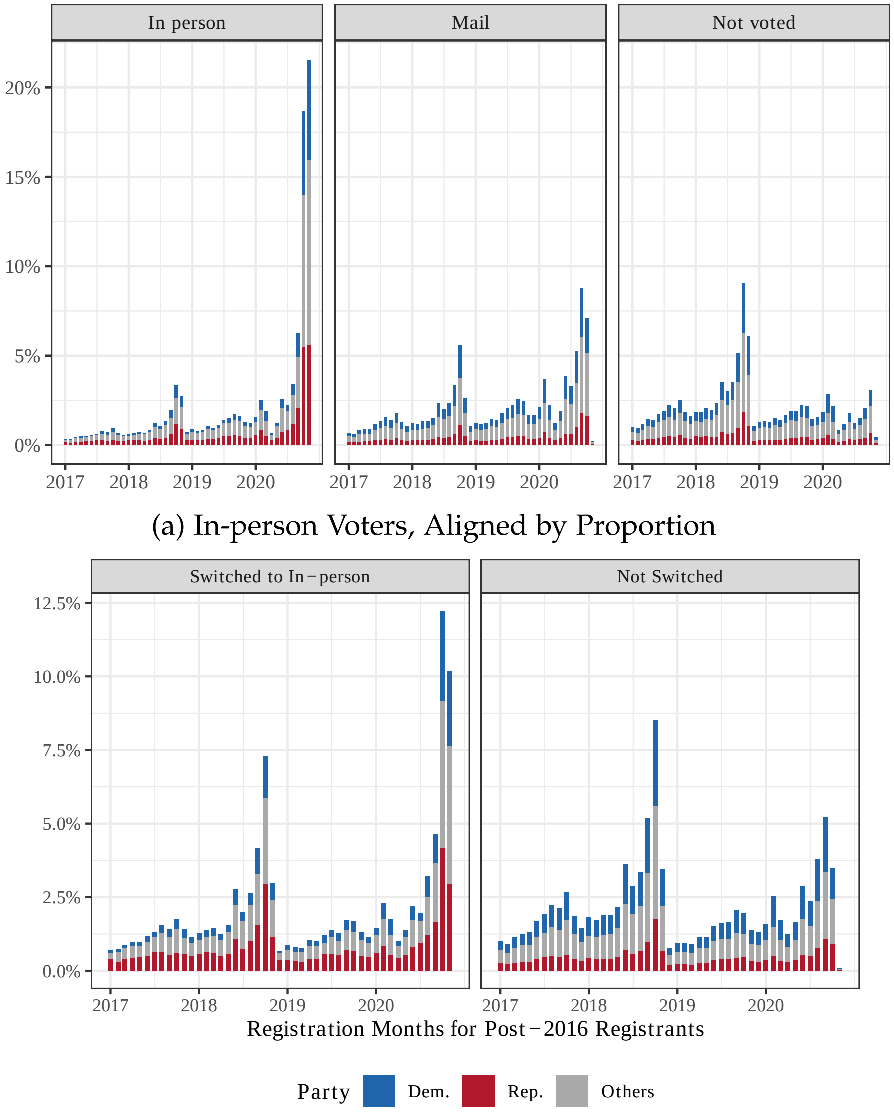


(a) In-person Voters

(b) Switchers

Figure 6: Composition of In-person Voters and Switchers by Party and Age Groups

Distribution Across Age Groups and Registration Dates. Given that Atkeson et al. (2022) have shown that patterns of choosing VBM were correlated with age, we also check for distribution across age groups. Each panel of Figure 6 shows the distribution



(a) In-person Voters, Aligned by Proportion

(b) Switchers, Aligned by Proportion

Figure 7: Composition of In-person Voters and Switchers, by Party and Month of Registration, 2020 General Election

of age groups and parties within the description.

As suspected, those who are voting in person or switching to in-person voting are younger. Conditional on voting in-person in the 2020 general election, the average age is 40, which is much lower than mail voters (49.6) or even those who did not turn out (42.8). Indeed, between in-person and vote-by-mail voters, there is almost a ten-year difference. What's more, the average age of switchers is 45.8, which is 6.4 years lower than those who did not (52.2).

We also suspect that many in-person voters are last-minute registrants who wish to make sure that their ballots are counted. Figure 7 shows the month of the registration for the subset of voters who registered during the 2020 presidential election cycle. Indeed, conditional on Election Day registration (EDR), 92.9% of voters were in-person voters, and 89.1% of those who did turn out were switchers.

Prediction Results

Given the heavy class imbalance in the dataset, we report results using various levels of downsampling of the majority class. Table 5 shows various performance measures by the level of downsampling.[8] Note that because predicting in-person voting, by-mail voting, and not turning out is multiclass classification, some metrics in Table 5a are calculated as an average across one-versus-rest classification's performance measures. As we have used the PR AUC to train our dataset,[9] we use the outcome from the specification with the highest PR AUC: when predicting voting modes, 25% downsampled training data is used, and when predicting switchers, 15% downsampled data is used.

dp	logLoss	AUC	prAUC	Accuracy	Kappa	Mean_F1	Mean_Sensitivity	Mean_Specificity	Mean_Precision	Mean_Detection_Rate
0.05	0.842	0.878	0.714	0.599	0.284	0.485	0.698	0.832	0.491	0.200
0.1	0.594	0.882	0.719	0.760	0.425	0.588	0.711	0.862	0.549	0.253
0.15	0.489	0.883	0.721	0.833	0.514	0.635	0.688	0.867	0.603	0.278
0.2	0.429	0.885	0.723	0.860	0.552	0.651	0.671	0.864	0.645	0.287
0.25	0.393	0.885	0.723	0.876	0.575	0.662	0.658	0.858	0.681	0.292

(a) Predicting Voting Modes

dp	logLoss	AUC	prAUC	Accuracy	Kappa	F1	Sensitivity	Specificity	Precision	Detection_Rate
0.05	0.463	0.773	0.581	0.789	0.109	0.879	0.796	0.587	0.981	0.768
0.1	0.295	0.772	0.582	0.936	0.187	0.966	0.960	0.258	0.973	0.927
0.15	0.230	0.772	0.583	0.954	0.184	0.976	0.983	0.170	0.970	0.948
0.2	0.197	0.771	0.583	0.960	0.155	0.980	0.991	0.117	0.969	0.956
0.25	0.178	0.768	0.581	0.963	0.135	0.981	0.995	0.091	0.968	0.960

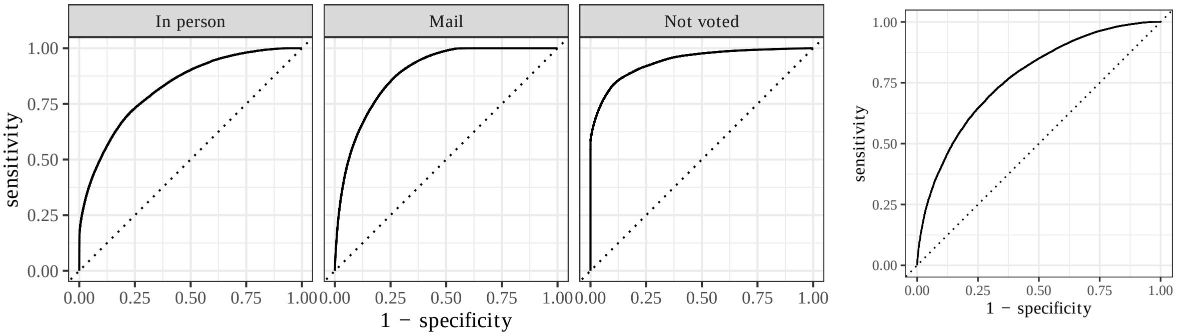
(b) Predicting Switchers

Table 5: Performance Summaries

Overall, the performance for predicting voting modes—especially considering its multiclass nature—was much higher than trying to predict switchers. The reason can be deduced from both the ROC curves (Figure 8) and precision-recall curves (Figure 9). For both figures, three panes are again represented for predicting voting modes due to the all-versus-rest prediction in multiclass classification. Predicting by-mail voters or those

[8] For both outcomes, using the full sample performs poorly than not for almost every metric.

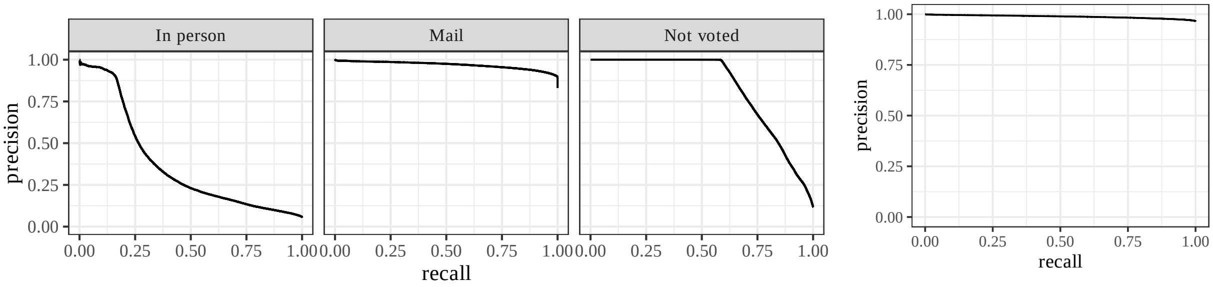
[9] For the multiclass classification, note that metrics such as AUC or PR AUC are also macro-averaged across all-versus-rest binary classification. We choose to macro-average as opposed to micro-averaging, considering that our objective is to predict the in-person voters and switchers.



(a) Predicting Voting Modes, 25% Downsampled

(b) Predicting Switchers, 15% Downsampled

Figure 8: ROC Curves for Classification of Voting Modes and Switchers



(a) Predicting Voting Modes, 25% Downsampled

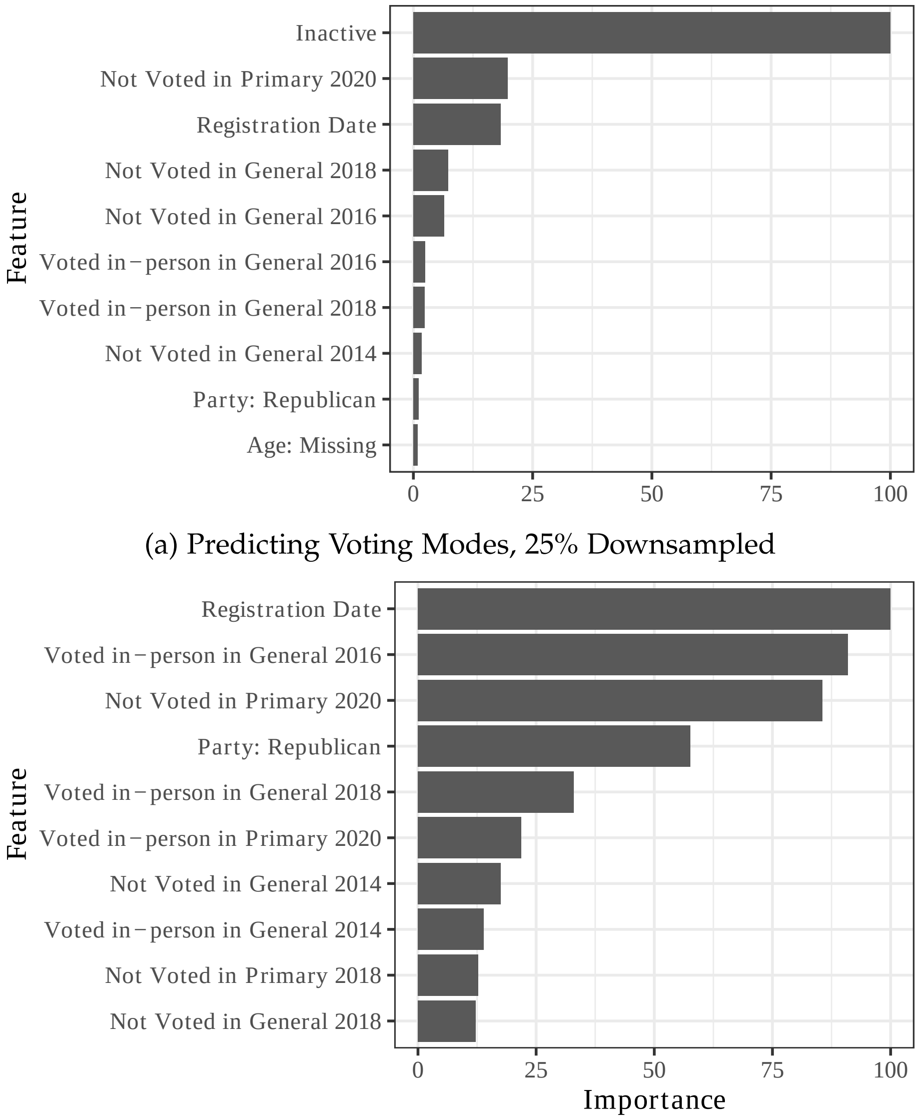
(b) Predicting Switchers, 15% Downsampled

Figure 9: Precision-Recall Curves for Classification of Voting Modes and Switchers

who did not turn out was a much easier task—predicting in-person voters as opposed to the rest proved to be the harder classification. In particular, the precision-recall AUC for predicting in-person voting is a mere 0.373, compared to 0.968 for by-mail voters and 0.829 for nonvoters. So while there is certainly an improvement from random guessing—the accuracy when randomly guessed taking the class imbalance into account would be 0.705 as opposed to 0.876—there is some room for improvement.

Given the final model, Figure 10 gives the various importance plots for both outcome variables derived by calculating the relative influence. For predicting voting modes, being an inactive voter was ranked as most important, which, in light of Figures 8 and 9, is likely an artifact of successfully classifying nonvoters. In terms of switching, the registration date was ranked as the most important. Surprisingly, Republican status was not ranked very highly in predicting in-person voters and ranked only fourth when predicting switchers.

Most other variables ranked as important are past turnout and voting mode choices.



(a) Predicting Voting Modes, 25% Downsampled

(b) Predicting Switchers, 15% Downsampled

Figure 10: Variable Importance Plots for Predicting Voting Modes and Switchers

These results are in line with past research on habitual voting (Gerber et al., 2003; Aldrich et al., 2011), with more recent choices ranking higher in relative importance. When predicting switching, however, whether the person voted in person in the last presidential general election of 2016 was more informative than the 2020 primary patterns.[10] Overall, primary voting patterns were not found to be highly informative, excluding 2020—it is what voters have chosen to do in general elections that is informative.

Being a Republican was ranked relatively high in its importance in predicting switching

[10] Note that when predicting switchers, a smaller subset is used, excluding 2020 general nonvoters.

to in-person voting. One could say that this was expected given the recent political climate, but again, given that Colorado was relatively experienced in VBM, this finding is worrisome and may be an indicator of a broader erosion of faith in the electoral system. In particular, voting in person is not just political behavior but also a health behavior, given COVID-19.

That voters who have experience using vote-by-mail in past elections have decided to vote in person, which likely poses a higher threat to their health, is not a trivial decision. This is even more so because, on average, Republican voters tend to be slightly older than Democratic voters in the dataset by more than four years. Many studies found that older people are at higher risk of COVID-19 fatalities and side effects (Mueller et al., 2020; Kang and Jung, 2020; Levin et al., 2020), and vaccines were rolled out in multiple stages for different age groups accordingly. Notice that local COVID prevalence and death rates were not informative in either prediction exercise. This might be because we are limited to county-level variations as opposed to measuring individual health outcomes or risk assessments, but also because partisan considerations outweigh any health risk assessment.

What is surprising is the relative uninformativeness of Republican affiliation in predicting voting modes, as opposed to predicting switchers. Given the conditional probabilities we have explored before with our descriptive analyses, it would have been natural to expect Republican status to rank as highly informative for both variables—but it is only for the switching that it is relatively important. Given the model performance above, this might simply indicate that while some Republicans certainly were swayed to vote in person as opposed to by mail, we just do not have a well-performing model of in-person voting in general, as preferences for voting modes form early in voters' experiences.

Age groups, although we hypothesized it to be important, were hardly important and did not appear in the top ten ranks for either outcome variable. Although whether age is missing has been ranked 10th in Figure 10a, this is simply because a voter whose information has been missing and not updated for a long time will likely predict a nonvoter. Instead, it is really the registration dates that seem to matter, signaling that same-day registration or registration close to Election Day—although a small portion of all voters—are highly indicative of in-person voting even for states with an established vote-by-mail system such as Colorado. The direction of the causality is unclear—whether it is because these voters did not have the bandwidth to register early that they are voting in person or whether the partisan framing has shaped the joint decision to register at the last minute and vote in person on the same day.

Conclusion

Although much of the literature surrounding the political behavior of voters is focused on turnout, as can be seen, the choice of voting modes is also an important class of political behavior. In this paper, we analyze (1) the choice between voting by mail (VBM), voting in person, and not voting entirely in the 2020 general election, and (2) the choice to switch to in-person voting despite the experience of voting by mail in previous cycles. We use Colorado, an established universal VBM state since 2013, and build a machine-learning model for both outcome variables and analyze their performance and variable importance. In particular, we allow for defection by voters, i.e., not voting entirely, because there was still a raging pandemic, which could discourage voters from voting in person, and a partisan misinformation campaign waged by former President Trump about how VBM was fraught with voter fraud, which could discourage voters from voting by mail. In doing so, we aim to build, more broadly, a means for election administrators to accurately predict future election demand so that officials can efficiently allocate resources, the first attempt with a large-scale administrative dataset.

Because of the partisan-polarized beliefs about voter fraud that have accelerated in recent years, partisanship was widely believed to play a large role in determining the choice to vote in person. We find that being a Republican was relatively important in predicting who *changed* to in-person voting despite experience with VBM, although we did not find it highly relevant when predicting in-person voting in general. Given that Republican voters are generally older, making them more susceptible to COVID-19 side effects and fatalities, voting in person is both a political behavior and a health behavior. Our findings have strong implications about how elite/partisan communication can shift both. Given the continuous attack on VBM, we might see these shifts grow in future elections.

Other than partisanship, variables such as active/inactive status, registration date, and past choices about turnout and voting mode dominated in terms of variable importance, which we interpret as the choice of voting modes also being habitual. Local COVID-19 prevalence itself was hardly relevant, but this could be both because (1) the local variation that we are leveraging is at the county level and does not capture how individuals assessed the risk of the pandemic and (2) also because COVID-19 was not as widespread in Colorado compared to other states. Surprisingly, age also was not ranked highly in importance, either.

However, note that the classification models using administrative data performed relatively poorly when predicting voting in person, compared to predicting voting by mail or nonvoting. Although less visible when using ROC curves, precision-recall curves clearly shows the model struggling to predict the minority class (in-person voting) against the rest of the data. Although we chose official administrative data given their scale and lack of self-reporting biases typical in surveys, this may signal that the tendency to choose in-person voting despite a universal VBM system may require additional data and other methodological approaches. For example, perhaps metadata from election officials, or their priors about voting mode use in upcoming elections, could be used in predictive models. Future researchers might also explore the use of Bayesian modeling approaches, which have been successfully applied in other election administration contexts (Cao et al., 2022). Clearly, there is room for improvement so that we can build predictive models that can be successfully used by election administrators.

Some limitations should be noted. One is that we cannot parse and capture whether voters filled in their mail ballots and chose to drop them off in polling places or ballot boxes. This may have served as a "hybrid" voting mode that assured both some social distance from other voters and some confidence that the ballot would be received and counted. However, given that Bryant (2020) has shown that the inability to feed one's ballot into a tabulator and get a confirmation of the ballot acceptance is why even voters using dropboxes experience a lower level of confidence, it is likely that this will not substantially change the results. The other caveat is that intuition from using Colorado as a benchmark case may not necessarily carry over to other states, and therefore we need replications across other jurisdictions now testing out universal VBM. One plausible hypothesis is that states with less experience voting universally by mail would have had stronger effects on partisanship—which we leave for future research endeavors.

References

Albertson, Bethany and Kimberly Guiler (2020). Conspiracy theories, election rigging, and support for democratic norms. *Research & Politics 7*(3).

Aldrich, John H., Jacob M. Montgomery, and Wendy Wood (2011). Turnout as a Habit. *Political Behavior 33*(4), 535–563.

Allcott, Hunt, Levi Boxell, Jacob Conway, Matthew Gentzkow, Michael Thaler, and David Yang (2020). Polarization and public health: Partisan differences in social distancing during the coronavirus pandemic. *Journal of Public Economics 191*, 104254.

Alvarez, R. Michael, Ines Levin, and J. Andrew Sinclair (2012). Making Voting Easier:

Convenience Voting in the 2008 Presidential Election. *Political Research Quarterly 65*(2), 248–262.

Atkeson, Lonna Rae, Wendy L. Hansen, Maggie Toulouse Oliver, Cherie D. Maestas, and Eric C. Wiemer (2022). Should I vote-by-mail or in person? The impact of COVID-19 risk factors and partisanship on vote mode decisions in the 2020 presidential election. *PLOS ONE 17*(9), e0274357.

Berinsky, Adam J., Nancy Burns, and Michael W. Traugott (2001). Who Votes by Mail?: A Dynamic Model of the Individual-Level Consequences of Voting-by-Mail Systems. *Public Opinion Quarterly 65*(2), 178–197.

Berlinski, Nicolas, Margaret Doyle, Andrew M. Guess, Gabrielle Levy, Benjamin Lyons, Jacob M. Montgomery, Brendan Nyhan, and Jason Reifler (2021). The Effects of Unsubstantiated Claims of Voter Fraud on Confidence in Elections. *Journal of Experimental Political Science*, 1–16.

Bonica, Adam, Jacob M. Grumbach, Charlotte Hill, and Hakeem Jefferson (2021). All-mail voting in Colorado increases turnout and reduces turnout inequality. *Electoral Studies 72*, 102363.

Bryant, Lisa A. (2020). Seeing Is Believing: An Experiment on Absentee Ballots and Voter Confidence: Part of Special Symposium on Election Sciences. *American Politics Research 48*(6), 700–704.

Cao, Jian, Seo-young Silvia Kim, and R. Michael Alvarez (2022). Bayesian Analysis of State Voter Registration Database Integrity. *Statistics, Politics and Policy 13*(1), 19–40.

Chase, Randall (2022, October). Delaware justices nix vote-by-mail, same-day registration. *Washington Post*.

Corse, Alexa (2020, October). In Colorado, Voting by Mail Was Practiced Well Before Coronavirus. *Wall Street Journal*.

Davis, Jesse and Mark Goadrich (2006). The relationship between Precision-Recall and ROC curves. In *Proceedings of the 23rd international conference on Machine learning - ICML '06*, Pittsburgh, Pennsylvania, pp. 233–240. ACM Press.

Dominguez, Melanie Sayuri, Edward D. Vargas, and Gabriel R. Sanchez (2020). Voting in a Pandemic: Explaining Variation in Support for Absentee Ballots for All. *Politics & Gender 16*(4), 1093–1100.

Efron, Bradley (2020). Prediction, Estimation, and Attribution. *Journal of the American Statistical Association 115*(530), 636–655.

Friedman, Jerome, Trevor Hastie, and Robert Tibshirani (2000). Additive logistic regression: a statistical view of boosting (With discussion and a rejoinder by the authors). *The Annals of Statistics 28*(2), 337–407.

Friedman, Jerome H. (2001). Greedy Function Approximation: A Gradient Boosting Machine. *The Annals of Statistics 29*(5), 1189–1232.

Friedman, Jerome H. (2002). Stochastic gradient boosting. *Computational Statistics & Data Analysis 38*(4), 367–378.

Gadarian, Shana Kushner, Sara Wallace Goodman, and Thomas B. Pepinsky (2021). Partisanship, health behavior, and policy attitudes in the early stages of the COVID-19 pandemic. *PLOS ONE 16*(4), e0249596.

Gerber, Alan S., Donald P. Green, and Ron Shachar (2003). Voting May Be Habit-Forming: Evidence from a Randomized Field Experiment. *American Journal of Political Science 47*(3), 540–550.

Gorman, Steve (2021, September). California becomes 8th U.S. state to make universal mail-in ballots permanent. *Reuters*.

Green, Jon, Jared Edgerton, Daniel Naftel, Kelsey Shoub, and Skyler J. Cranmer (2020). Elusive consensus: Polarization in elite communication on the COVID-19 pandemic. *Science Advances 6*(28), eabc2717.

Greenwell, Brandon, Bradley Boehmke, Jay Cunningham, and GBM Developers (2022). gbm: Generalized boosted regression models. *R package, version 2.1.8.1*.

Grimmer, Justin, Margaret E. Roberts, and Brandon M. Stewart (2021). Machine Learning for Social Science: An Agnostic Approach. *Annual Review of Political Science 24*(1), 395–419.

Grossman, Guy, Soojong Kim, Jonah M. Rexer, and Harsha Thirumurthy (2020). Political partisanship influences behavioral responses to governors' recommendations for COVID-19 prevention in the United States. *Proceedings of the National Academy of Sciences 117*(39), 24144–24153.

He, Haibo and Edwardo A. Garcia (2009, September). Learning from Imbalanced Data. *IEEE Transactions on Knowledge and Data Engineering 21*(9), 1263–1284. Conference Name: IEEE Transactions on Knowledge and Data Engineering.

Herron, Michael C. and Daniel A. Smith (2021). Postal delivery disruptions and the fragility of voting by mail: Lessons from Maine. *Research & Politics 8*(1), 2053168020981434.

James, Gareth, Daniela Witten, Trevor Hastie, and Robert Tibshirani (2013). *An Introduction to Statistical Learning: with Applications in R*. Springer Science & Business Media.

Jurkowitz, Mark (2021, January). Republicans who relied on Trump for news more concerned than other Republicans about election fraud. *Pew Research Center*.

Kang, Seung-Ji and Sook In Jung (2020). Age-Related Morbidity and Mortality among Patients with COVID-19. *Infection & Chemotherapy 52*(2), 154–164.

Karp, Jeffrey A. and Susan A. Banducci (2000). Going Postal: How All-Mail Elections Influence Turnout. *Political Behavior 22*, 223–239.

Kim, Seo-young Silvia and Bernard Fraga (2022). When Do Voter Files Accurately Measure Turnout? How Transitory Voter File Snapshots Impact Research and Representation.

King, Gary and Langche Zeng (2001). Logistic Regression in Rare Events Data. *Political Analysis 9*(2), 137–163.

Kousser, Thad and Megan Mullin (2007). Does Voting by Mail Increase Participation? Using Matching to Analyze a Natural Experiment. *Political Analysis 15*(4), 428–445.

Kousser, Thad, Mindy Romero, Mackenzie Lockhart, Seth Hill, and Jennifer Merolla (2020). Early in the Pandemic, There Was No Partisan Divide over Preferences for Voting by Mail in the 2020 Election. *California Journal of Politics and Policy 12*(1).

Levin, Andrew T., William P. Hanage, Nana Owusu-Boaitey, Kensington B. Cochran, Seamus P. Walsh, and Gideon Meyerowitz-Katz (2020). Assessing the age specificity of infection fatality rates for COVID-19: systematic review, meta-analysis, and public policy implications. *European Journal of Epidemiology 35*(12), 1123–1138.

Lockhart, Mackenzie, Seth J. Hill, Jennifer Merolla, Mindy Romero, and Thad Kousser (2020). America's electorate is increasingly polarized along partisan lines about voting by mail during the COVID-19 crisis. *Proceedings of the National Academy of Sciences 117*(40), 24640–24642.

McGhee, Eric, Jennifer Paluch, and Mindy Romero (2022). Vote-by-mail policy and the 2020 presidential election. *Research & Politics 9*(2), 20531680221089197.

Menger, Andrew and Robert M. Stein (2020). Choosing the Less Convenient Way to Vote: An Anomaly in Vote by Mail Elections. *Political Research Quarterly 73*(1), 196–207.

Montgomery, Jacob M. and Santiago Olivella (2018). Tree-Based Models for Political Science Data. *American Journal of Political Science 62*(3), 729–744.

Mueller, Amber L., Maeve S. McNamara, and David A. Sinclair (2020). Why does COVID-19 disproportionately affect older people? *Aging (Albany NY) 12*(10), 9959–9981.

NCSL (2022). Table 18: States With All-Mail Elections. Last retrieved Oct 8, 2022.

Niebler, Sarah (2020). Vote-by-Mail: COVID-19 and the 2020 Presidential Primaries. *Society 57*(5), 547–553.

Paul, Jesse (2020, October). Colorado health officials warn state could reach record coronavirus hospitalizations by Nov. 10. *The Colorado Sun*.

Persily, Nathaniel and Charles Stewart (2021). The Miracle and Tragedy of the 2020 U.S. Election. *Journal of Democracy 32*(2), 159–178.

Plescia, Carolina, Semra Sevi, and André Blais (2021). Who Likes to Vote by Mail? *American Politics Research 49*(4), 381–385.

Porter, Ethan and Thomas J. Wood (2019). *False Alarm: The Truth about Political Mistruths in the Trump Era*. Cambridge University Press.

Ridgeway, Greg (1999). The state of boosting. *Computing Science and Statistics 31*, 172–181.

Ridgeway, Greg (2020). Generalized Boosted Models: A guide to the gbm package. *R vignette*.

Saito, Takaya and Marc Rehmsmeier (2015). The Precision-Recall Plot Is More Informative than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Datasets. *PLOS ONE 10*(3), e0118432.

Saul, Stephanie and Reid J. Epstein (2020, September). Trump Is Pushing a False Argument on Vote-by-Mail Fraud. Here Are the Facts. *The New York Times*.

Southwell, Priscilla (2007). Vote-by-Mail: Voter Preferences and Self-Reported Voting Behavior in the State of Oregon. *American Review of Politics 28*, 139–146.

Southwell, Priscilla L. (2009). A Panacea for Voter Fatigue? Vote by Mail in the State of Oregon. *Journal of Political and Military Sociology 37*(2), 195–203.

Southwell, Priscilla L. and Justin I. Burchett (2000). The Effect of All-mail Elections on Voter Turnout. *American Politics Quarterly 28*(1), 72–79.

Thompson, Daniel M., Jennifer A. Wu, Jesse Yoder, and Andrew B. Hall (2020). Universal vote-by-mail has no impact on partisan turnout or vote share. *Proceedings of the National Academy of Sciences 117*(25), 14052–14056.

Udani, Adriano, David C. Kimball, and Brian Fogarty (2018). How Local Media Coverage of Voter Fraud Influences Partisan Perceptions in the United States. *State Politics & Policy Quarterly 18*(2), 193–210.

Walsh, Katherine Cramer (2012). Putting Inequality in Its Place: Rural Consciousness and the Power of Perspective. *American Political Science Review 106*(3), 517–532.

Wang, Yu (2019). Comparing Random Forest with Logistic Regression for Predicting Class-Imbalanced Civil War Onset Data: A Comment. *Political Analysis 27*(1), 107–110.

Wingerter, Meg (2020, October). Colorado COVID-19 hospitalizations on track to top April in 11 days if nothing changes. *The Denver Post*.