FORECASTING THE 2020 US PRESIDENTIAL ELECTIONS

Cesar Y. Villarreal Guzman

02 November 2020

Abstract

In this paper develop two models with the purpose of forecasting the 2020 United States presidential election. Multilevel regression with postratification (MRP) and Stacked regression with postratification models where constructed using the Nationscape survey and postsratified using the ACS census dataset. We conclude by forecasting the victory of Joe Biden in the popular vote and commenting on the high likelihood of winning the electoral college process.

Keywords: Forecasting; U.S. 2020 Election; Trump; Biden; Multilevel Regression with Poststratification

1 Introduction

It is a common theme in the social sciences and statistics literature to attempt constructing a model that accurately predicts the outcome of a presidential election. Often the purpose of these experiments ranges from finding a novel statistical model to a commentary of the impact at the economical, social or political level that the expected winner would cause. All such experiments, however, share the same underlying problem, predicting the outcome of the elections.

As Pablo Hiriat writes in his article for "El Financiero", the future of world's politics will be decided on November 3rd in the 2020 United States presidential elections. Isidore (2020) discusses how the U.S. elections are now the most bet-on event in history. As of Wednesday morning, a week prior to the elections, about $284 million USD had already been wagered on the Betfair Exchange by British bettors. Betfair Exchange is one of the largest betting exchanges in the world based on London. The author continues by predicting this figure to grow close to $500 million USD.

It seems now more than ever this years elections carry with them greater level of importance, and in this paper we attempt to forecast the percentage of popular votes that will appear in the screens of many Americans on November 3rd. To accomplish this two methods are used, a multilevel regression and poststratification (MRP) model and a stacked regression and poststratification model (SRP). To fit the models we used the Democracy Fund + UCLA

Nationscape survey data set and, since this survey is a non-probability sample, to obtain accurate estimates we poststratify our predictions using the ACS U.S. census data set.

This paper is structured in the following manner. Section two contains commentary on both data sets, what they are and where they come from, here we also explain what is meant by a non-probability sample. This is followed by a brief discussion on our data cleaning methodology. In section three we elaborate on how MRP and SRP work and how they are implemented to generate predictions. Section four is where we present our estimates for both models and finally in section five we comment on this estimates and comment on the most likely winner of the U.S. presidential election, Joe Biden.

2 Data

2.1 Democracy Fund + UCLA Nationscape Survey

Nationscape (Tausanovitch and Vavreck 2020) is a survey conducting 500,000 interviews of Americans from July 2019 through December 2020, covering the 2020 presidential election. The survey includes online interviews with roughly 6,250 people per week starting July 10, 2019.

As in almost all contemporary survey research, the Nationscape survey is not a random sample of the population of interest. In particular, the Nationscape survey is a convenience sample selected on a set of demographic criteria by a market research platform that runs an online exchange for survey respondents. Such samples where provided by the company Lucid.

With this description we can classify this survey as an online non-probability sample. At its core, what this means is that the sample is obtained by non-random selection, or a selection without the use of probability. This contrasts with probability samples, where, a full list of the target population is available from which respondents are selected uniformly at random, although the exact method may vary depending on the study. In this case the population would be Americans 18 years or older.

The survey is divided into two phases. Phase 1 of the data, released in January of 2020, includes approximately 156,000 cases collected over 24 weeks, beginning from the week of July 18, 2019 and concluded with the week of December 26, 2019. Phase 2 of the data, released in September of 2020, includes a re-release of Phase 1 data and new data from January 2020 to July 2020 (Phase 2 data). Each weekly survey is released as its own data set, and combining all data set results in 318,697 cases. To access the survey data one must request access by means of the Voter Study Group website.[1]

2.2 ACS Census Data Set

The American Community Surveys (ACS) (Steven Ruggles and Sobek 2020) is a project of the U.S. Census Bureau that has replaced the decennial census as the key source of information about American population and housing characteristics. This survey has been conducted

[1]Data: https://www.voterstudygroup.org/publication/nationscape-data-set

since 2000 and the most recent on 2018. An important distinction is that the ACS is a sample and not a full census data set.

To be specific the ACS survey is sent to a sample of addresses (about 3.5 million) in the 50 states, District of Columbia, and Puerto Rico and it is conducted every month, every year. The Master Address File (MAF) is the Census Bureau's official inventory of known housing units in the United States and Puerto Rico. It serves as the source of addresses and hence sampling frame for the ACS. Their sampling process is a complicated 2 phase process but in summary first they assign addresses to sixteen sampling strata, then determine base rate and calculate stratum sampling rates and finally systematically select sample. In conclusion the ACS is a probability sample.

To access the ACS surveys an account from IPUMS USA website is required[2]. The database allows for the creation of a customized data set. In particular we chose the 2018 ACS survey and selected the following variables: sex, age, state, race, and Hispanic. Justification for the choosing of these variables can be found in the "Model" section. Automatically ten other variables are appended to the selection. Out of these ten there is one in particular that was also used. According to the ACS code book, "PERWT indicates how many persons in the U.S. population are represented by a given person in an IPUMS sample." We use this variable in the poststratification step to obtain better population estimates.

2.3 Methodology

To train our models we used the Nationscape surveys conducted in the month of June 2020, since this are the closest surveys to the elections. We do not consider surveys conducted in 2019 and early 2020 since it is well known that political views have shifted because of the U.S. treatment of the pandemic. The resulting data set contained $20,157$ entries. Next we selected the columns for: sex, age, state, race, Hispanic origin and vote choice. To match our ACS selection. This resulted in a rectangular data set of $20,157$ rows and 6 columns.

To clean the data we first removed rows with missing values. Then, we applied the common method of data binning to the age column. We created seven age groups that range from 18 to 93. A similar procedure was used for the race category where we reduced the number of categories. The Hispanic origin column was transformed into a binary feature, either 'Hispanic' or 'Not Hispanic'. The same procedure was done on the vote choice column where we used only data which contained 'Joe Biden' or 'Donald Trump' as values.

The ACS data set followed an almost identical cleaning procedure. First we selected the study columns, created and categorized age into age groups, reduced the number the race categories, and transformed the Hispanic origin column to a binary variable. The only difference is that we performed this cleaning procedure while being consistent with the category values used in the Nationscape data set.

Figure 1 and Figure 2 are visualization of the cleaned data we worked with. The important thing to notice is that the sample appears to be fairly representative of the population.

[2]Data: https://usa.ipums.org/usa/

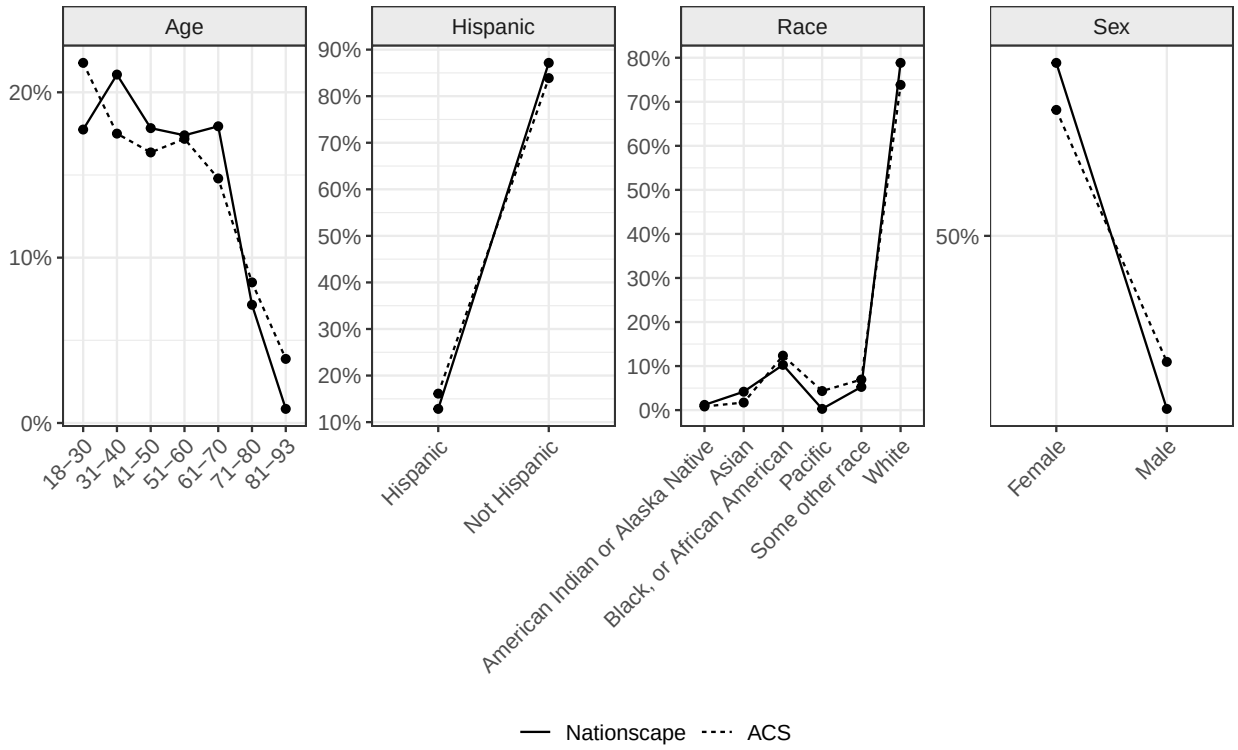

Figure 1: ACS and Nationscape Data Measured by Proportions

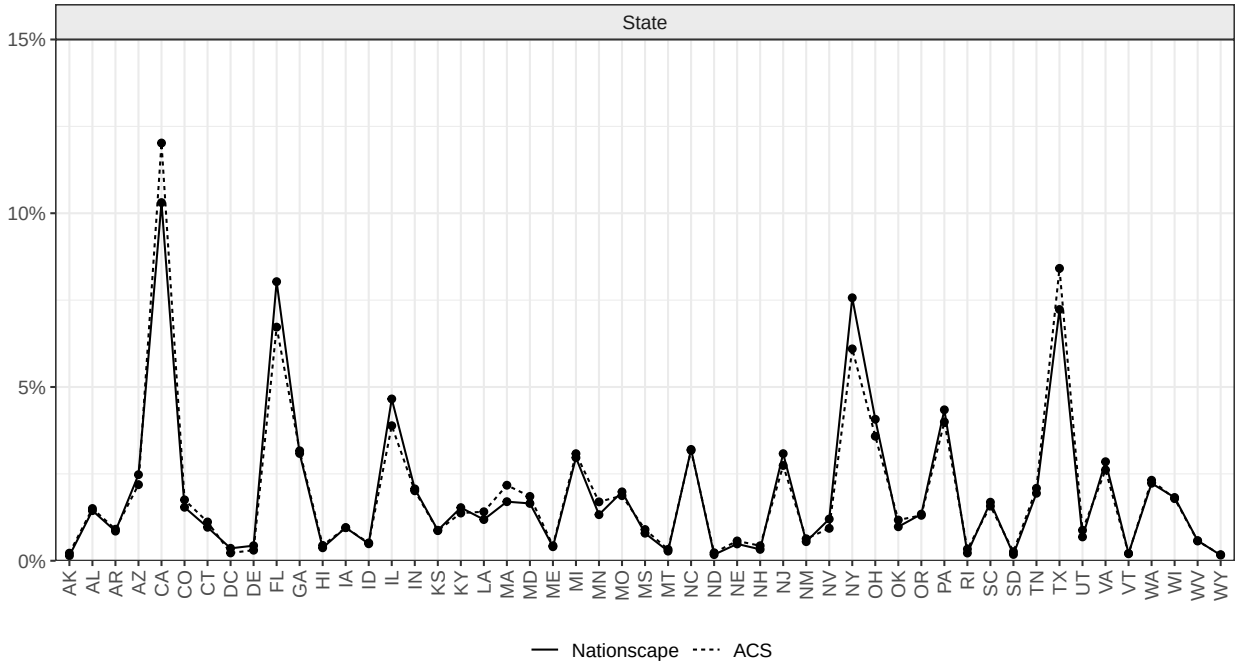

Figure 2: ACS and Nationscape State Proportions

3 Model

3.1 Multilevel Regression and Poststratificatio

The statistical technique known as MRP or multilevel regression and poststratification (Little 1993; Park, Gelman, and Bafumi 2004) allows researchers to infer quantities in the population from a sparse and/or non-representative sample (Wang et al. 2015), accomplished by combining two techniques: small-area estimation (model) and non-response adjustment (poststratification) (Kennedy and Gelman 2020). Furthermore, MRP is widely used in the political science literature and has been proven to be an effective analysis technique for non-probability based samples like the Nationscape survey.

Therefore, to transform the Nationscape convenience sample into accurate estimates of voter intent, we make use of demographic information provided by respondents and poststratify raw responses to mimic a representative sample of likely voters.

Before applying MRP we need to prepare what's referred to in the literature as the post-stratification **cells** which comprises of all possible combinations of the categories found in the model's variable inputs. In our case we will be considering the following variables for our model: gender, age, race, Hispanic origin and state. The reason we use these is because of their proven efficacy in predicting political ideology. There are many examples in the literature but some of the ones we take inspiration from are Ghitza and Gelman (2013), Wang et al. (2015) and Ghitza and Gelman (2020).

To generate the cells we consider all possible combinations of gender (2 categories), Hispanic origin (2 categories), age (7 categories), race (6 categories) and state (51 categories), thus partitioning the data into 8568 cells. Applying MRP in our setting comprises of two steps: first, we fit a Bayesian hierarchical model relating Joe Biden support with age, race, Hispanic origin and state. Symbolically, for an individual i we denote $y_i = 1$ if the respondent support Joe Biden in the upcoming elections, and 0 for Donald Trump. The non-hierarchical part of the model can be written as

$$P(y_i = 1) = \text{logit}^{-1}(\beta X_i)$$

where X_i contains indicator variables for gender and Hispanic origin and interaction terms for age and race. Adding the varying intercept (state), the final model can be written as

$$P(y_i = 1) = \text{logit}^{-1}\left(\beta X_i + \alpha^{\text{state}}_{j[i]}\right)$$

$$\alpha^{\text{state}} \sim N(0, \sigma_{\text{state}})$$

We use this model to obtain estimates for our poststratification cells. In the second step, we average over the cells, weighting the values by its relative proportion to the population.

To implement this model we use the **brms** (Bürkner 2018) package for the R (R Core Team 2020) language. One important aspect of the Bayesian framework is the selection of priors. By default brms normalizes and re-scales the data and sets priors that reflect this transformation, this default was our prior of choice.

3.2 Stacked Regression and Poststratification

As discussed in the previous section MRP has enabled a flowering of new research in the social sciences, however the method has its own downsides. An example is from Buttice and Highton (2017) where they argue that the model performs poorly in a number of empirical applications, specifically when the first-stage model is a poor fit for the public opinion of interest. The authors discuss that MRP performs better when there is a greater geographic heterogeneity in opinion.

Table 1: Distribution of Vote Intention in Nationscape Survey

Candidate	Proportion (%)
Joe Biden	52.6
Donald Trump	47.4

Table 1 is included in this section to better point out the limitations of using MRP. Notice that the distribution is very uniform, almost 50-50, which potentially could cause a poorer fit of our MRP model. The researcher Ornstein (2019) recently proposed a solution to this limitation. In his paper, he introduces a technique called Stacked Regression and Poststratification (SRP). In this technique, rather than estimating using a single multilevel regression model, SRP generates predictions from a "stacked" ensemble of models, including regularized regression, k-nearest neighbors, random forest, and gradient boosting. All popular machine learning algorithms. Then, poststratify by weighting these predictions as one would do in MRP.

To understand stacked ensembling one must recall that the majority of classification models are functions that do not explicitly output discrete variable. Instead they output a probability that a given input belongs to a certain category, then a decision function translates this probability into an actual discrete prediction. What stacked ensembling does is to combine multiple learning algorithms, called base classifiers, and use the outputs, or probabilities, as inputs to train another algorithm called the meta-classifier (Breiman 1996; Laan, Polley, and Hubbard, n.d.).

For our purpose we used the following base classifiers: k-nearest neighbors, random forest, support vector machine; and regularized logistic regression (LASSO) as our meta-classifier. This is very similar to what Ornstein (2019) used in his paper, however we do not use gradient boosting to try an minimize computational time. Instead, our method to try and break linearity is by training a support vector machine. If the reader is unfamiliar with any of these learning algorithms we recommend consulting Hastie, Friedman, and Tisbshirani (2017).

We used the stacked ensemble algorithm implementation from the **mlxtend** package (Raschka 2018) for the Python version 3 language (Van Rossum and Drake 2009). Similarly we used the implementations for our base classifiers found in the **scikit-learn** package (Pedregosa et al. 2011). To find the optimal parameters we used a grid search algorithm, which is essentially a brute force approach that given a range of parameters it tries them all and finds the optimal combinations. As the original author describes one huge limitation of SRP is

the computational cost. We used a free-tier Google cloud instance to train the model which took approximately 32 minutes to train compared to our MRP model which took less than 5 minutes to fit.

4 Results

Table 2: MRP Forecast of Percentage of Popular Vote at 95% Confidence

Candidate	**Popular Vote (%)**	Lower Quantile (%)	Upper quantile (%)
Joe Biden (Democratic)	52.69	37.44	66.67
Donald Trump (Republican)	47.31	33.33	62.56

Table 3: SRP Forecast of Percentage of Popular Vote at 95% Confidence

Candidate	**Popular Vote (%)**	Lower Quantile (%)	Upper quantile (%)
Joe Biden (Democratic)	54.36	21.55	89.97
Donald Trump (Republican)	45.64	10.03	78.45

As per the titles Tables 2 and 3 display our predictions for popular vote proportions. As can be seen both models forecast a popular vote victory for the democratic party in the U.S. Presidential elections. Similarly 95% confidence intervals are included in the tables. This means that there is a 95% probability that the true popular votes proportion lies between the lower quantile and the upper quantile.

Similarly we estimated the percentage of popular votes per state using both models. Based on these results we created Figure 3, these maps show what we expect to be the states won by the democrats and the republicans. The MRP model estimates a total of 30 states for the democrats and 21 for republicans. Similarly the SRP model estimates 27 states for the democrats and 24 for republicans. A table containing the explicit predictions of popular vote proportions for each state can be found in the Appendix.

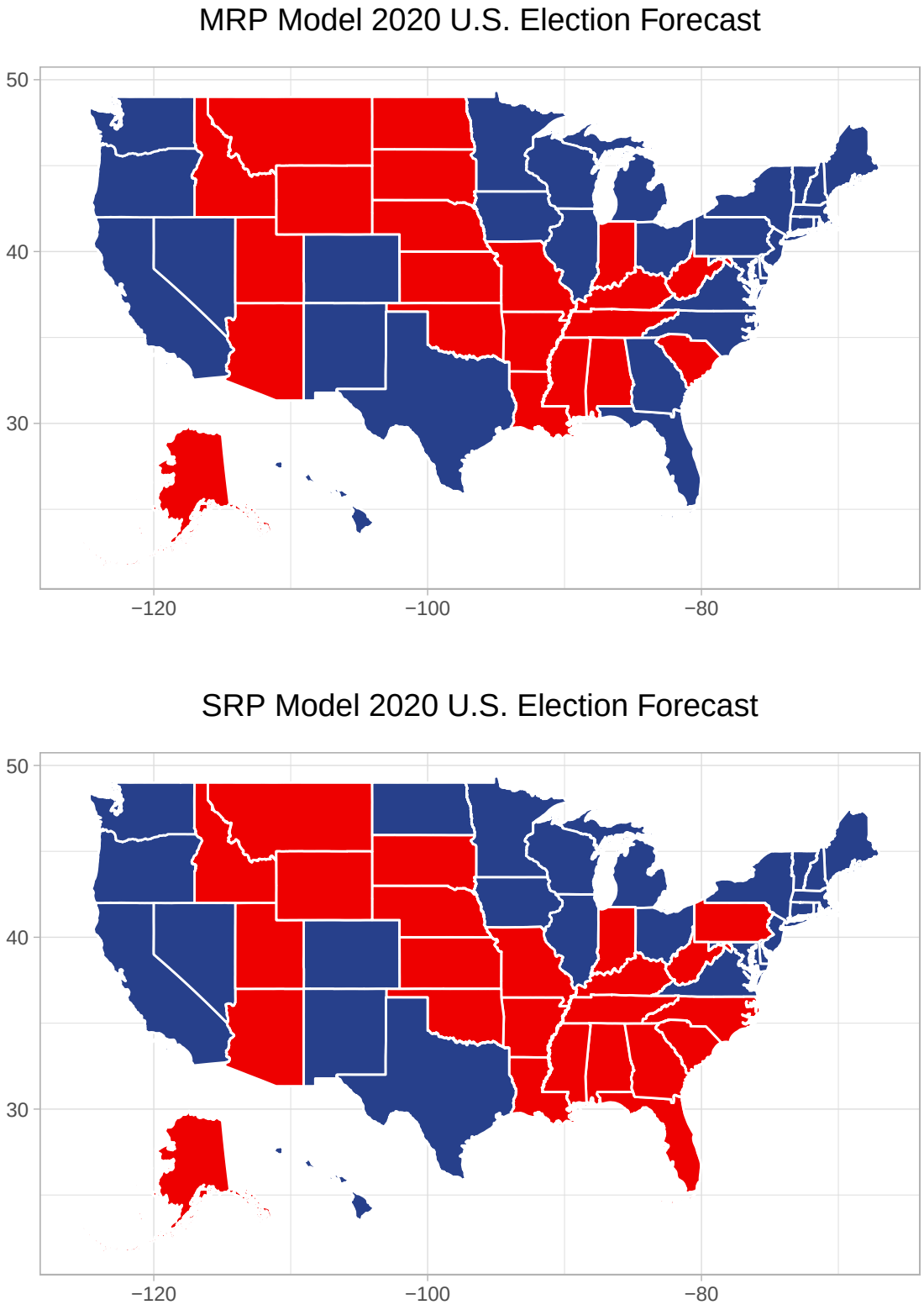

Figure 3: Forecast of State Victories

5 Discussion

Table 4: Estimated Number of Electoral Votes.

Candidate	**MRP Estimate**	**SRP Estimate**	The Economist
Joe Biden (Democrat)	401	324	350
Donald Trump (Republican)	137	214	188

It is necessary, before our discussion, to comment on the U.S. electoral system since for the average non-American this might not be well know information. Contrary to popular belief the president of the United States is not elected by popular vote. On election day, when Americans vote for president they are actually voting for whom their state will vote for. The U.S. is the only country that picks its president using process called the **Electoral College**. The Electoral College process consists of the selection of the electors, the meeting of the electors where they vote for President and Vice President, and the counting of the electoral votes by Congress. There are a total of 538 electors. A majority of 270 electoral votes is required to elect the President.

With this in mind we must clarify that even if our estimates show that Joe Biden will win the popular vote it is perfectly possible for Donald Trump to be re-elected as president. In fact, during the 2016 elections Donald Trump lost the popular vote to Hillary Clinton but won the election by majority of electoral votes.

Looking at Figure 3, perhaps the most noticeable difference between the MRP predictions and SRP are the states near Florida. In the 2016 elections Donald Trump won this state and it is a major cause of uncertainty for our predictions. Looking at other well known pollsters, like "The Economist[3]" and "The New York Times[4]" we noticed that the states that have the greatest uncertainty are Florida, Georgia, South and North Carolina, Texas, Ohio and Iowa.

We agree with these entities that perhaps it is a tossup for Florida, Georgia, South and North Carolina. Furthermore, we question if our models are accurately predicting the outcome for Texas which is another state that Donald Trump won in the 2016 elections. However, we are confident in the predictions for Ohio and Iowa.

Most States have a "winner-take-all" system that awards all electors to the Presidential candidate who wins the state's popular vote. If we make the assumption that all states will follow this system then our model estimates help create an even stronger estimate. In Table 4 our forecast for number electors can be found. We present Table 4 in this section since we are making a big assumption to generate these estimates. Moreover, this forecast is not part of our raw estimates which are presented in the previous section.

We would like to point out that there is no way to quantify what each state's elector system

[3]Source: https://projects.economist.com/us-2020-forecast/president

[4]Source: https://www.nytimes.com/live/2020/presidential-polls-trump-biden?action=click&module=styln-elections-2020-guide&variant=show&state=default&pgtype=LegacyCollection&region=hub&context=storyline_election_guide

will be and therefore there is no method available to present explicit bounds of error as we did in the previous section. It is for this reason that we have provided the estimates of a well known pollster "The Economist" to compare this forecast to a more statistically backed one. Furthermore, we can comment on Table 4 results as broad estimates since, as stated, most state do follow a winner takes all system. If this is the case, like in the 2016 elections where 48 states and the District of Columbia followed this system then we could confidently forecast that the majority of electoral votes, at least 270, will go to the democratic party candidate, Joe Biden.

Finally recall that in 2016, months of national polls confidently showed Hillary Clinton ahead, and set many Americans up for a shock on Election Nigh. Two particular pollsters Arie Kapteyn and Robert Cahaly who accurately predicted Donald Trump's victory in 2016 theorize that the reason why this happens is because of "shy" Trump voters. They describe them as "... people reluctant to share their opinions for fear of being judged" and continue by stating that "there's a lot of hidden Trump votes out there", and they could make the difference in the upcoming elections (Stanton 2020).

It is the next statement by Cahaly that calls for future work in this research area, "Will Biden win the popular vote? Probably. I'm not even debating that. But I think Trump is likely to have an Electoral College victory." Fundamentally, based on this theory, the problem is that polls are non-representative of Trump's support. A large of portion of our paper was devoted to explain how multilevel modeling and poststratification is a great tool to accurately estimate population parameters when the sample is non-representative of the population. This case is no exception and further work should be devoted to more accurately estimating Trump's support in the United States.

Appendix

All code used to generate this results can be found here:

- LINK

Predictions by State

Table 5: Models Forecast of Popular Vote Proportion by State

State	MRP Estimate (%)	SRP Estimate (%)
AK	45.18	45.28
AL	47.01	30.10
AR	39.54	30.20
AZ	47.69	36.76
CA	63.88	88.72
CO	55.29	66.18
CT	65.22	69.27
DC	70.60	88.25
DE	62.26	70.71
FL	53.34	38.08
GA	50.05	41.73
HI	66.73	83.55
IA	55.62	75.94
ID	37.04	20.86
IL	59.41	73.49
IN	47.87	41.10
KS	48.85	47.17
KY	42.06	27.80
LA	49.73	41.33
MA	66.50	93.28
MD	66.04	76.41
ME	50.02	53.99
MI	55.80	61.70
MN	58.58	56.78
MO	48.85	33.11
MS	48.26	42.71
MT	46.18	49.05
NC	51.78	34.88
ND	46.58	53.46
NE	49.72	47.24
NH	56.72	85.94
NJ	59.33	70.16
NM	56.47	63.18

State	MRP Estimate (%)	SRP Estimate (%)
NV	53.83	60.64
NY	61.63	83.46
OH	53.35	52.00
OK	43.74	34.60
OR	52.90	63.56
PA	50.53	41.74
RI	61.31	84.50
SC	47.62	30.61
SD	44.73	30.98
TN	43.01	20.22
TX	50.18	54.13
UT	41.59	27.95
VA	56.85	53.64
VT	63.55	86.26
WA	62.60	90.39
WI	56.70	61.81
WV	38.64	33.88
WY	36.22	23.61

References

Breiman, Leo. 1996. "Stacked Regressions." *Machine Learning* 24 (1): 49–64. https://doi.org/10.1007/BF00117832.

Buttice, Matthew K., and Benjamin Highton. 2017. "How Does Multilevel Regression and Poststratification Perform with Conventional National Surveys?" *Political Analysis* 21 (4): 449–67. https://doi.org/10.1093/pan/mpt017.

Bürkner, Paul-Christian. 2018. "Advanced Bayesian Multilevel Modeling with the R Package brms." *The R Journal* 10 (1): 395–411. https://doi.org/10.32614/RJ-2018-017.

Ghitza, Yair, and Andrew Gelman. 2013. "Deep Interactions with Mrp: Election Turnout and Voting Patterns Among Small Electoral Subgroups." *American Journal of Political Science* 57 (3): 762–76. https://doi.org/10.1111/ajps.12004.

———. 2020. "Voter Registration Databases and Mrp: Toward the Use of Large-Scale Databases in Public Opinion Research." *Political Analysis* 28 (4): 507–31. https://doi.org/10.1017/pan.2020.3.

Hastie, Trevor, Jerome Friedman, and Robert Tisbshirani. 2017. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer.

Hiriat, Pablo. n.d. "En Una Semana Se Juega Todo: La Democracia Está En Vilo." *El Financiero*. https://www.elfinanciero.com.mx/opinion/pablo-hiriart/en-una-semana-se-juega-todo-la-democracia-esta-en-vilo.

Isidore, Chris. 2020. "The Us Election Is Likely the Most Bet-on Event in History." https://edition.cnn.com/2020/10/30/business/us-presidential-election-wagering-record/index.html.

Kay, Matthew. 2020. *tidybayes: Tidy Data and Geoms for Bayesian Models*. https://doi.org/10.5281/zenodo.1308151.

Kennedy, Lauren, and Andrew Gelman. 2020. "Know Your Population and Know Your Model: Using Model-Based Regression and Poststratification to Generalize Findings Beyond the Observed Sample." http://arxiv.org/abs/1906.11232.

Laan, Mark J. van der, Eric C Polley, and Alan E. Hubbard. n.d. "Super Learner." *Statistical Applications in Genetics and Molecular Biology* 6 (1). https://doi.org/https://doi.org/10.2202/1544-6115.1309.

Little, R. J. A. 1993. "Post-Stratification: A Modeler's Perspective." *Journal of the American Statistical Association* 88 (423): 1001–12. http://www.jstor.org/stable/2290792.

Ornstein, Joseph T. 2019. "Stacked Regression and Poststratification." *Political Analysis* 28 (2): 293–301. https://doi.org/10.1017/pan.2019.43.

Park, David K., Andrew Gelman, and Joseph Bafumi. 2004. "Bayesian Multilevel Estimation with Poststratification: State-Level Estimates from National Polls." *Political Analysis* 12 (4): 375–85. https://doi.org/10.1093/pan/mph024.

Pedregosa, F., G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, et al. 2011. "Scikit-Learn: Machine Learning in Python." *Journal of Machine Learning Research* 12: 2825–30.

"President-Forecasting the Us 2020 Elections." n.d. *The Economist*. The Economist Newspaper. https://projects.economist.com/us-2020-forecast/president.

Raschka, Sebastian. 2018. "MLxtend: Providing Machine Learning and Data Science Utilities and Extensions to Python's Scientific Computing Stack." *The Journal of Open Source Software* 3 (24). https://doi.org/10.21105/joss.00638.

R Core Team. 2020. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Stanton, Zack. 2020. "'People Are Going to Be Shocked': Return of the 'Shy' Trump Voter?" https://www.politico.com/news/magazine/2020/10/29/2020-polls-trump-biden-prediction-accurate-2016-433619.

Steven Ruggles, Ronald Goeken, Sarah Flood, and Matthew Sobek. 2020. "IPUMS Usa: Version 10.0 [American Community Surveys 2018]." https://doi.org/0.18128/D010.V10.0.

Tausanovitch, Chris, and Lynn Vavreck. 2020. "Democracy Fund + Ucla Nationscape, October 10-17, 2019 (Version 20200814)." https://www.voterstudygroup.org/publication/nationscape-data-set.

Van Rossum, Guido, and Fred L. Drake. 2009. *Python 3 Reference Manual*. Scotts Valley, CA: CreateSpace.

Wang, Wei, David Rothschild, Sharad Goel, and Andrew Gelman. 2015. "Forecasting Elections with Non-Representative Polls." *International Journal of Forecasting* 31 (3): 980–91. https://doi.org/10.1016/j.ijforecast.2014.06.001.

Wickham, Hadley. 2011. "The Split-Apply-Combine Strategy for Data Analysis." *Journal of Statistical Software* 40 (1): 1–29. http://www.jstatsoft.org/v40/i01/.

———. 2016. *Ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. https://ggplot2.tidyverse.org.

Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, et al. 2019. "Welcome to the tidyverse." *Journal of Open Source Software* 4 (43): 1686. https://doi.org/10.21105/joss.01686.