Assuming Everyone Voted, Trudeau Would Win 34% of the Popular Vote in the 2019 Canadian Federal Election*

Xinyi Zhang

16 December 2020

Abstract

In the 2019 Canadian Federal Election, incumbent Prime Minister Justin Trudeau claimed a narrow victory over the Conversative Party amid 62% voter turnout. In this paper, we develop a multi-level regression model with post-stratification by training a multinomial logistic regression model using voter survey results and predicting the outcome of the popular vote using large-scale demographic data for the Canadian population. Our model predicts that assuming 100% voter turnout, Trudeau would have won the popular vote of the 2019 Canadian Federal Election with a larger margin. Our prediction speaks to the importance of encouraging voter turnout and our breakdown of votes by demographic groups provides political parties with information on how to target voters during future elections.

Keywords: forecasting, 2019 Canadian Federal Election, Justin Trudeau, Andrew Scheer, multilevel regression with post-stratification, voter turnout

1 Introduction

In 2019, Trudeau defied election predictions and narrowly won enough seats to win reelection. This raises the question, what would the popular vote have been if everyone voted? By choosing 5 explanatory demographic variables closely associated with voting and political affiliation (age, gender, state, race, and education), we develop a multi-level regression model with post-stratification in order to predict the outcome of the popular vote.

In our analysis, we modeled the relationship between our selected demographic variables and a person's likelihood to vote for either Trudeau or Scheer. We analyzed the significance of our model, and the importance of using multilevel regression with post-stratification because the training data is not proportional to the Canadian population. We found that we were able to make predictions with approximately 70% accuracy and the strongest indicators of 2019 voting were education and gender.

This paper discusses the 2 datasets we used, how they were collected and key highlights of these datasets, followed by visualizations of the data. Next we explained the construction of our model and the positives and negatives of extrapolating information from a smaller voter survey to the Canadian population using a post-stratification dataset. Finally, we present our results and discuss how our results should inform future political strategies.

-vs riding -actual number of popular vote -general info about election -where dataset is from

*Code and data supporting this analysis are available at: https://github.com/cindyzhang99/sta304_ps5.

2 Data

To train our model to predict voting on an individual level for the 2019 Canadian Federal Election, we used data from the 2019 Canadian Election Study (CES). To make predictions on the outcome of the 2019 Canadian Federal Election through post-stratification, we used data from the 2017 General Social Survey (GSS).

In the following subsections (Individual-level Survey Dataset, Post-stratification Dataset), we will discuss how each dataset was collected and highlight their key features. Then, in the Data Visualization subsection, we'll graph the distribution of our variables of interest. We will use this data in the multilevel regression with post-stratification (MRP) technique that we will describe in the Model section.

2.1 Individual-level Survey Dataset

From September 13, 2019 to October 21, 2019, the Consortium on Electoral Democracy conducted online surveys on political views and voting intent prior to the 2019 Canadian Federal Election. Their target population was all Canadian citizens and permanent residents, aged 18 or older.

The online platform Qualtrics was used to conduct sampling. Qualtrics used several panels as the sampling frame for this survey. Sampling frames are lists of the individuals that will be selected for the survey sample, meaning that the members of panels aggregated by Qualtrics form a list of a subset of the target population. Then, a sample was selected from the frame using a purposive sampling method. This is a non-probability sampling method where the researcher decides which samples are most representative of the target population. According to the CES Codebook, demographic targets for province, gender (50% male and 50% female) and age (28% aged 18-34, 33% aged 35-54, and 39% aged 55 or older) were set. More specific information about the sampling method was not provided.

Of the 74,548 individuals contacted to take the survey, about 26% did not complete the questionnaire. Another 13% exceeded demographic quotas. Lastly, approximately 10% of responses were removed for speeding (spending less than 500 seconds completing the survey) or for "straight-lining" answers (selecting the same response for all questions), resulting in a final sample size of 37,822 respondents. To ensure results were representative of the Canadian population, survey responses were weighted using data from the 2016 Census. This ensures that the discrepancy between the target population and survey responses is minimized.

Unfortunately, the non-probability sampling method employed by Qualtrics is a major weakness of the CES surveying methodology. Although the non-response rate is relatively low, this is clearly because the sampling frame consists of individuals who on survey panels who regularly take online questionnaires. Even so, several key features of the 2019 CES methodology should be highlighted as strengths. First off, the specificity of answer choices for many questions was very detailed. In fact, when matching CES variable levels to GSS variable levels, we often had to combine levels in the CES data because the GSS data was not as specific. Additionally, the 2019 CES recontacted some participants after the election to participate in a post-election survey. Although we will not be using the post-election data, there are a multitude of applications where a longitudinal survey about the Canadian Federal Election provides valuable insights. Lastly, as we briefly touched upon when discussing the non-response rate, the 2019 CES survey results met stringent data processing standards prior to being publicly released. Therefore, we can trust that the responses represent Canadian voter sentiments to a certain extent when working with the data.

Although the non-probability sampling method is a major weakness of the 2019 CES methodology, we will ensure that our prediction of the election popular vote will be representative of the Canadian population by conducting post-stratification using data from the 2017 General Social Survey. Therefore, we can still make valid conclusions using the CES dataset even though the responses were drawn through non-probability sampling.

2.2 Post-stratification Dataset

From February 1, 2017 to November 30, 2017, Statistics Canada gathered data on the Canadian family unit by conducting voluntary telephone interviews for Cycle 31 of the General Social Survey. Their target population was all non-institutionalized individuals living in Canada, aged 15 or older.

Cross-sectional sampling was conducted in a two-stage design. The stratified simple random sampling method was used in the first stage. Here, the sampling frame consisted of telephone numbers from the Census grouped as households using data from Statistic Canada's dwelling frame. Strata were formed at the census metropolitan area (CMA) level and at the province level (i.e., large CMAs formed their own stratum, smaller CMAs were grouped together, and the non-CMA regions of each province were grouped together), forming a total of 27 non-overlapping strata. Finally, households were sampled randomly from each stratum such that the number sampled units from each stratum corresponded to the population sizes of each stratum. To reiterate, the sampled population for this first stage was the chosen households from each stratum. The stratified simple random sampling method was also used in the second stage. Here, the sampling frame was a list of household members, aged 15 and older, from the households selected in the first stage. Then, one individual was randomly selected from each household, forming the sampled population. Approximately 43,000 individuals were contacted to participate in the survey.

Statistics Canada reported that the non-response rate was 52.4%. This presents problems for data analysis based on survey data if respondents differ significantly from non-respondents. To reduce the effects of non-response bias, survey estimates were adjusted based on the demographic characteristics of households that were non-responsive (by pulling their information from the 2016 Census). Another source of non-sampling error is imperfect coverage. For example, households without telephones are excluded from the sampling frame. Again, survey estimates were adjusted by weighing responses to represent all individuals in the target population. Lastly, another weakness of the survey methodology is the exclusion of the Canadian population residing in the Northwest Territories, Nunavut, and the Yukon Territory. As we found when trying to match variable levels in the CES dataset to the GSS dataset, we had to drop responses from the territories in the survey data because of the lack of information available for the territories in the post-stratification data. Due to this limitation in the GSS data coverage, our prediction of the popular vote of the 2019 Canadian Federal Election excludes the voters in the territories of Canada.

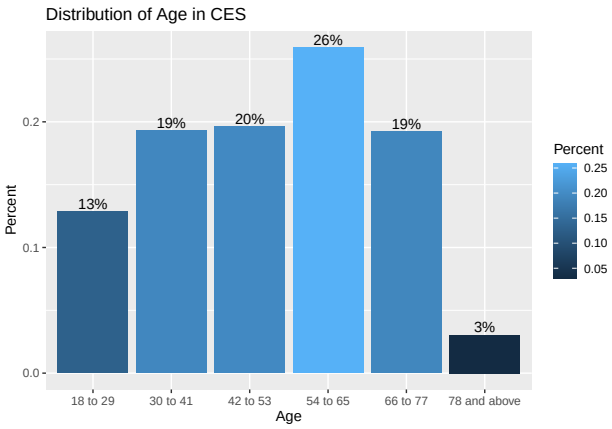
On the other hand, several key features of the GSS stand out as strengths of their surveying methodology. A major strength of the questionnaire is that it contains focused questions that comprehensively and extensively capture the subject of interest (the Canadian family). Extensive research and testing was conducted when designing the questionnaire. Upon reading through the questionnaire made available by Statistics Canada, the wording of each question is precise and clear, leaving little room for ambiguity. Additionally, another strength of the survey is that a vast majority of questions were objective (dates, events, counts) removing potential response biases that occur with subjective questions.

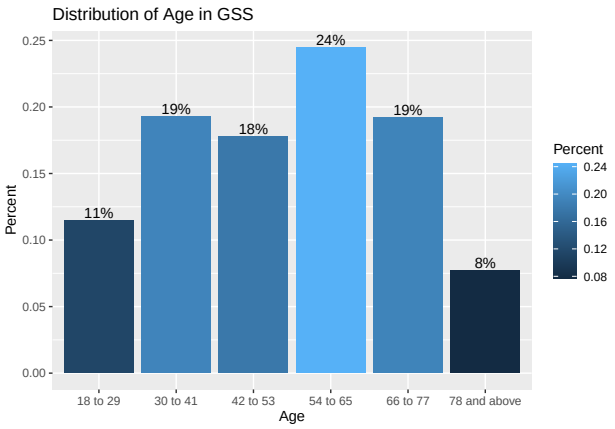
Overall, the GSS surveying method using two-stage simple random stratified sampling is effective in generating a sample that is geographically representative of the population living in the Canadian provinces. Despite limitations in sampling coverage, the demographic information available is much more detailed than in comparable post-stratification datasets (e.g., the 2016 Census).

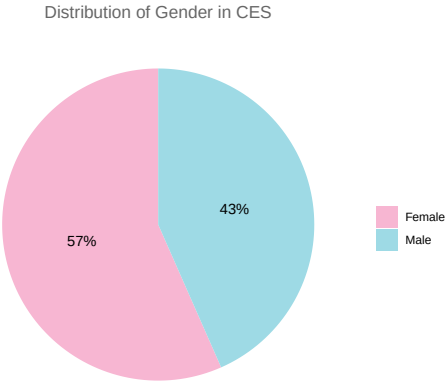
2.3 Data Visualization

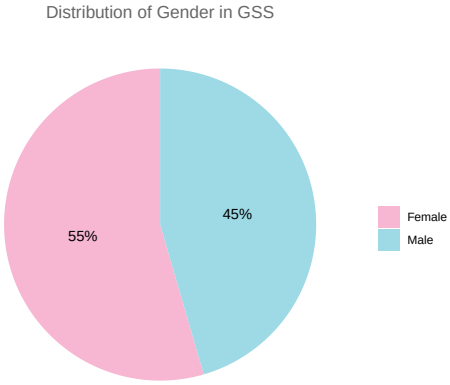
The full dataset of responses to the 2019 CES online survey and the 2017 General Social Survey contains tens of thousands of observations for over 400 variables.

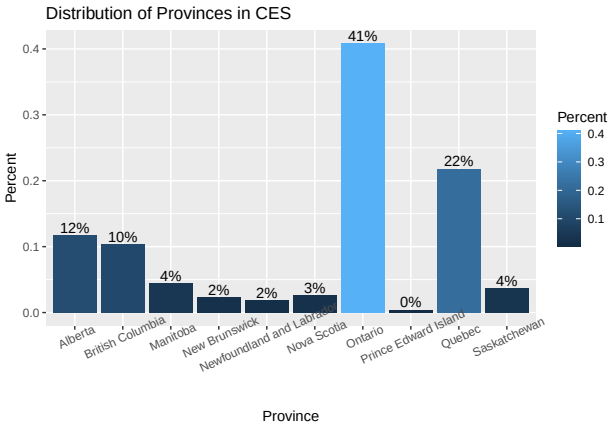
In the interest of space, we will only discuss the variables in the dataset that are relevant to our model. The variable we aim to predict is self_rated_mental_health while the factors that we chose to inform this prediction are age, sex, marital_status, and self_rated_health. We chose these factors based on the demographic information mentioned in mental health statistics (age, sex, and health) and based on what we suspected might contribute to mental health in the context of family composition (marital status, has children). More explicitly, here are the chosen variables we used from the original dataset:

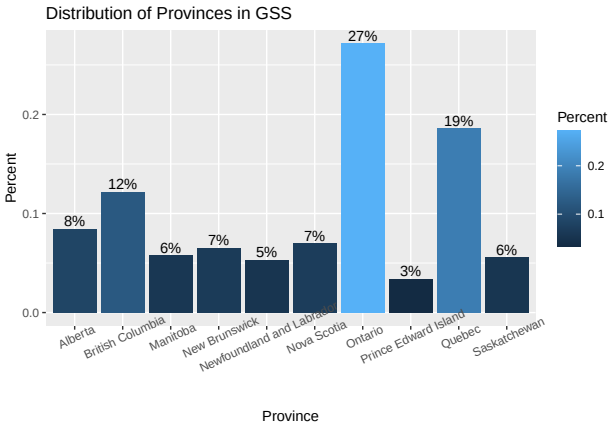

Distribution of Age in CES
Percent
0.0
0.1
0.2
13%
19%
20%
26%
19%
3%
18 to 29
30 to 41
42 to 53
54 to 65
66 to 77
78 and above
Age
Percent
0.25
0.20
0.15
0.10
0.05

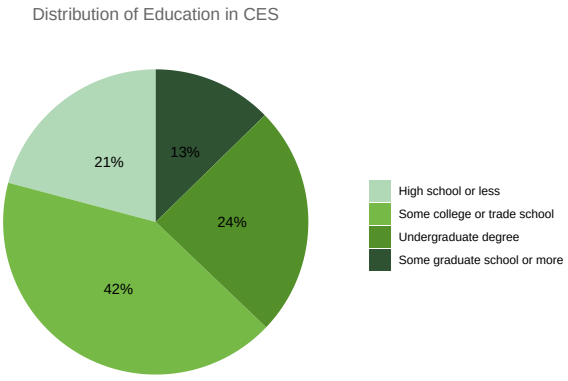

Distribution of Age in GSS
Percent
0.00
0.05
0.10
0.15
0.20
0.25
11%
19%
18%
24%
19%
8%
18 to 29
30 to 41
42 to 53
54 to 65
66 to 77
78 and above
Age
Percent
0.24
0.20
0.16
0.12
0.08

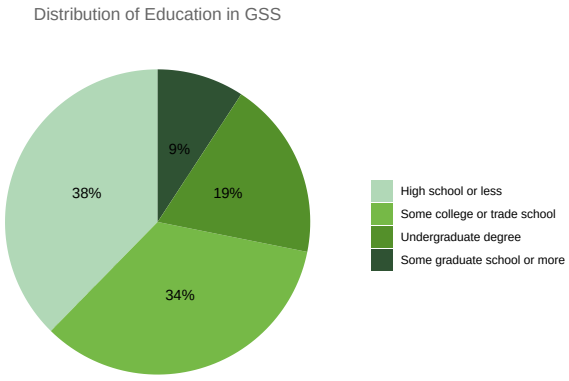

Distribution of Gender in CES
57%
43%
Female
Male

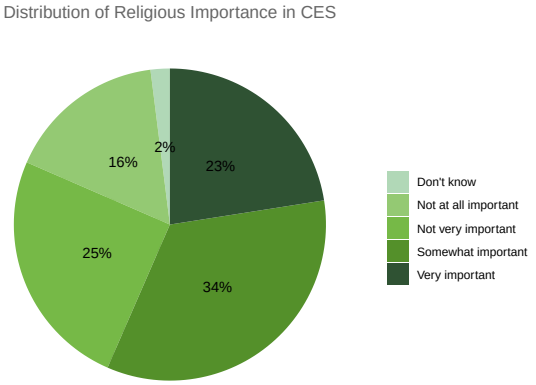

Distribution of Gender in GSS
55%
45%
Female
Male

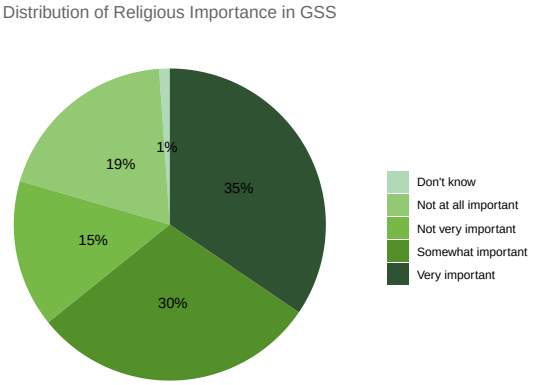

Distribution of Provinces in CES
Percent
0.0
0.1
0.2
0.3
0.4
12%
10%
4%
2%
2%
3%
41%
0%
22%
4%
Alberta
British Columbia
Manitoba
New Brunswick
Newfoundland and Labrador
Nova Scotia
Ontario
Prince Edward Island
Quebec
Saskatchewan
Province
Percent
0.4
0.3
0.2
0.1

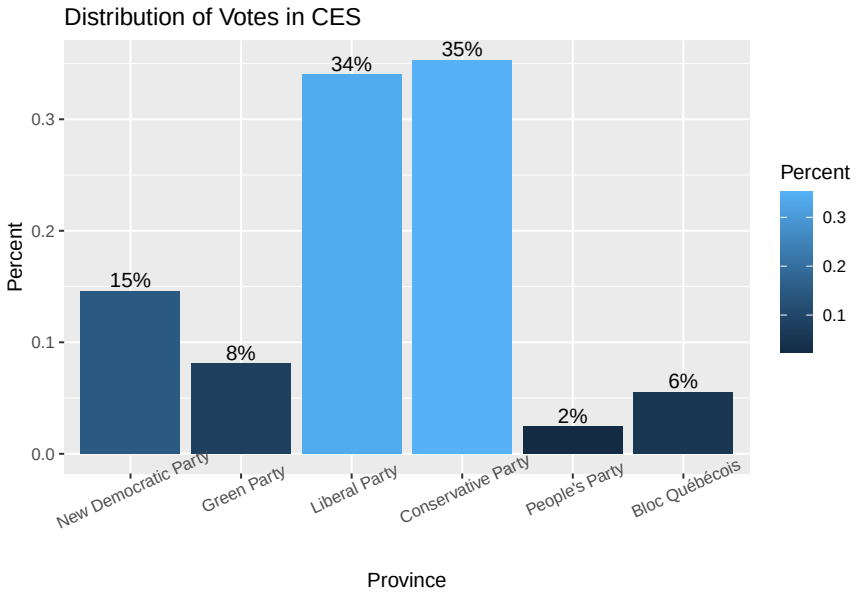

Distribution of Provinces in GSS
Percent
0.0
0.1
0.2
8%
12%
6%
7%
5%
7%
27%
3%
19%
6%
Alberta
British Columbia
Manitoba
New Brunswick
Newfoundland and Labrador
Nova Scotia
Ontario
Prince Edward Island
Quebec
Saskatchewan
Province
Percent
0.2
0.1











3 Model

The purpose of our model is to predict a person's vote in the 2020 US election given their demographic characteristics. The dependent variable is a binary categorical variable, where 0 represents voting for Trump

and 1 represents voting for Biden. The independent variables are the characteristics of the person (age, gender, race, state, and education). To avoid multicollinearity (when there is a relationship between the independent variables) which causes unreliable regression estimates, a few carefully selected independent variables were used for our model. One criteria for the selection was that the variables had to be both present in the Nationscapes dataset and the ACS dataset, so that they be included in post-stratification. Another criteria is that the characteristic should group people that share a similar perspective which may impact on their voting decision. Through literature research (("Factors Associated with Voting" 2015)), we found that some helpful indicators were age, gender, race, state, education, and self-reported income. To finalize the selection, characteristics should not be highly correlated and the data provided should make logical sense. Since the data quality for income was generally poor and it worsened the model performance, income was ultimately excluded from the model. Therefore, the predictors used for the model are age, gender, race, state, and education.

We found when exploring and graphing the data in the section above that the survey data has underrepresented or overrepresented certain characteristics compared to the actual distributions in the population (namely younger individuals were underrepresented and individuals with higher education were overrepresented in the Nationscape survey data). Therefore, if we were to predict the outcome of the popular vote using the demographic proportions found in the Nationscape survey, we would find that the prediction is biased because the survey sample is not representative of the American population. Hence, we use multi-level regression with post-stratification (MRP) to adjust the influence of each subgroup of the respondents to get a better match on the actual population distribution. To do this, a multi-level model is required as well as another resource (ACS dataset in our case) for post-stratification.

3.1 Multi-level logistic regression model

Since we have represented predicting the popular vote as a binary classification problem using multiple explanatory variables, binary logistic regression is a suitable choice for the model. We use logistic regression to predict the probability of a person voting for Biden and determine their vote by rounding to the nearest represented binary categorical variable. Logistic regression takes independent variables (in our case age, gender, race, state, and education, state, and race) as inputs. Based on the assigned weights and a logit function, the output will be a probability [0,1] of voting for Biden/Trump. Equation (1) is the equation of the logistic regression model and this defines the multi-level regression part of MRP:

Equation 1: Logistic regression model

$$Pr(Y_i \in \{Trump, Biden\}) = logit^{-1}(\alpha^{\text{age}}_{a[i]} + \alpha^{\text{gender}}_{g[i]} + \alpha^{\text{edu}}_{e[i]} + \alpha^{\text{state}}_{s[i]} + \alpha^{\text{race}}_{r[i]}) \tag{1}$$

where Y_i represents the probability a respondent is likely to vote for Trump or Biden given various demographic information and the α's are age, gender, education, state, and race and the notation $\alpha_{a[i]}$ refers to the age group the i-th individual belongs to, $\alpha_{g[i]}$ refers to the gender group the i-th individual belongs to, and so forth.

The Nationscape dataset contains voter characteristics and their expected vote for the 2020 US Presidential election, so the model was trained using cross-validation on that dataset. 95% of the dataset was used as the training set to determine the best weights for the model and the remaining 5% of the dataset as the test set to verify the accuracy of the model. The vast majority of the data is used for the training set because we wanted to ensure that the data does not overfit due to a small sample size.

Furthermore, there were many adjustments that had to be done with the inputs to improve model accuracy. Apart from self-reported income, education was another independent variable that we debated whether it should be included in the model. Education severely impacted our cross-validation accuracy due to a wide range of values which were too specific for our purpose. These issues were resolved by summarizing all education levels into two levels: "High School or Less" and "Post Secondary or More". We were able to confirm that there was a significant difference in the voting preference between the two groups, consequently helping our model perform better.

Another variable that initially caused inaccurate results was age. There was no clear trend in having age as a numeric continuous value for predicting votes. As a result, the numerical age values were grouped into bins of size 10 (eg. 18-28, 29-38) in order to better represent different age groups. Another grouping we tried was splitting the data into youth, middle age, and seniors (18-35, 36-55, 55+); however, that split was not able to capture any significant voting pattern in the age groups because we found that the model accuracy actually worsened. So grouping age by bins of size 10 gave us the best results and is the configuration that was included in the model.

After finalizing the inputs that went into our model, we implemented the logistic regression model using the BRMS library written in the programming language R (R Core Team (2020)). We ran our script using the software RStudio. We did not run into any diagnostic issues when running our model. As briefly mentioned, we conducted cross-validation to check our model. The results are fittingly covered in the Results section.

3.2 Post-stratification

As we have previously established, we need to use post-stratification to correct our model estimate of the popular vote because of the discrepancy between the demographic distribution of and the demographics of the American population. We conduct post-stratification by calculating a weighted average of estimates for all combinations of explanatory variables. We found the weights to do so from the ACS dataset. However, we cannot immediately use the ACS distributions as the dataset may not represent American demographics accurately. For example, as we found in our data exploration, there is a discrepancy between the minority distribution found in the ACS dataset and the data reported by the US Census. If minorities tend to democratically and for Biden, we would have a bias against Biden because they are underrepresented in the ACS dataset.

When we perform post-stratification on the ACS data, we find all combinations of our variables and find the weight of each combination representation within the US. Using the PERWT variable provided by the ACS, we can calculate how much of the population each combination represents within the United States. As per the IMPUMS webpage, "PERWT indicates how many persons in the U.S. population are represented by a given person in an IPUMS sample", meaning we can find all the individuals with the same values for the combination of variables we are measuring and add their PERWT values together to estimate how much these specific values would weigh.

Using MRP provides many benefits. As mentioned, we can more accurately estimate the weight of our sample predictions in relation to the population without being heavily affected by any sampling bias. It also allows us to use a small sample for training, and apply the model to a much larger sample that better represents the population. In our example, we can use a small scale election survey from Nationscape and train a model which allows us to predict election results for the respondents in the American Community Survey consisting of over 3 million data. This is crucial, as collecting surveys with a useful sample size with regards to an upcoming election can be extremely expensive and time consuming, as these poll results may be time sensitive. The option to collect from a small sample size allows statisticians to save money and time while providing significant results about the larger population by applying their surveys to general census information. Another noted benefit is computational efficiency. Instead of individually predicting over 3 million people, we summarize all the comparisons and only predict on over 8000 individuals and multiply their proportions, which allows us to compute our predictions significantly faster.

However, there are also cons to using MRP. For instance, we are limited to only use variables that will be found in general census data. If we are interested in looking at religion but the post-stratification dataset does not have that field, we are unable to include religion in our model. This can be challenging as information such as an individual's vote in 2016 may be extremely important as a variable, but since the ACS dataset does not contain such information, we are use it to our advantage. Furthermore, if a variable is broken down to a different granularity, we must group by the common group. For example, if dataset 1 has age grouped in bins of size 3 and dataset 2 has age grouped in bins of 5 (eg. 5-8,9-12 vs 5-10,15-20) then we must map them into bins that are multiples of 15. We cannot use groups of 3, because we cannot accurately break down groups of 5 into groups of 3. Therefore we lose granularity in situations where it may be desired.

In this case, using MRP makes sense as we have voting information provided by Nationscape, however this dataset is not distributionally representative of the target population. We also have access to the ACS which contains a lot of distributional data about the American population but it does not contain any information on who an individual will vote for in the upcoming election. Using MRP allows us to use both datasets in order to predict the outcome of the popular vote of the 2020 US presidential election.

4 Results

5 Discussion

6 Weaknesses

7 Next Steps

8 Appendix

8.1 Post-stratification dataset collection cost

Statistics Canada did not disclose the true cost of conducting the survey but we can make some speculations based on the available information about their field work methodology. Surveying was conducted using Computer Assisted Telephone Interviewing (CATI) wherein interviewers read aloud the computerized questionnaire and immediately record the respondent's answers. Although this allows for a reduction in costs compared to traditional in-person surveying, labor costs still include time spent computerizing the survey, training interviewers, and having interviewers administer the questionnaire. Other labor costs include designing the questionnaire and surveying methodology as well as conducting quality control (data consistency was checked by the CATI system during surveying and unresolved inconsistencies were handled afterwards by support staff). Non-labor costs likely included paying for equipment, phone service, offices, and so forth. Again, although we don't have exact costs, we can conclude that the time and costs associated with conducting the GSS is a clear reason why it is only administered once a year.

9 References

Stephenson, Laura B., Allison Harell, Daniel Rubenson and Peter John Loewen. The 2019 Canadian Election Study – Online Collection.

https://dataverse.harvard.edu/file.xhtml?persistentId=doi:10.7910/DVN/DUS88V/HRZ21G&version=1.0

Rohan, A. & Caetano, S. (2020). 'GSS 2017 Cleaning Code'. MIT License.

https://www23.statcan.gc.ca/imdb/p2SV.pl?Function=getSurvey&SDDS=4501&lang=en&db=imdb&adm=8&dis=2

"Factors Associated with Voting." 2015. *Statistics Canada: Canada's National Statistical Agency / Statistique Canada : Organisme Statistique National Du Canada.* https://www150.statcan.gc.ca/n1/pub/75-001-x/2012001/article/11629-eng.htm.

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

Xie, Yihui. 2020. *Bookdown: Authoring Books and Technical Documents with R Markdown*. https://github.com/rstudio/bookdown.