Identifying Voter Ethnicity by Means of Location and Name Data

Nick Bauman

Dr. Mihhail Berezovski MA395

Florida Democratic Party

I. Introduction

1.1 **Abstract**

The Florida Democrats provide free and public information on voter registration data. They have expressed concern in their voter turnout in 2018, specifically noting the Hispanic community turning out significantly less throughout the state. The task at hand is to note if there is any correlation between a voters' ethnicity and their voter registration. This progress report aims to represent how far the research into identifying a voters' ethnicity, and using those ethnicities along with their party registration to note any significance has come.

1.2 **Historical Review**

The interest in this research began when the Florida Democrats determined that their Hispanic voter turnout in 2018 was not as good as 2016, which was isolated to the South Florida region. Florida does ask a survey of questions for each of their voters and includes whether they identify as Hispanic or not. However, they do not ask the voters for their actual country of origin. By hypothesis, since we know that geographically Hispanic diaspora tends to gravitate towards their similar ethnic origin, we can assume that geography can affect how someone votes dramatically.

II. Methods

2.1 **Datasets**

Voter Registration:

The voting records for all registered Floridians is public information. The records used in this program are updated as of January 7, 2021. The files contain 39 information columns for each voter, but only some of these are of significance. The data used from the voting records are the voters' name, (first and last), zip code, party affiliation, ethnicity, and county.

The column that contains the ethnicity of the voters uses a series of integers. If the integer in the column is 4, the voter self-identifies as Hispanic. This is how the voters will be determined to be Hispanic or not.

Due to the size of this project, there are some limitations of the use of this file. There is some missing information, such as about 8 Hispanic voters that don't have a zip code. Due to the number of Hispanics in Florida being over 4 million, the very few voters with missing

information do not pose any significant limitations to the scope of the project. There are some missing information that weren't compatible with python functionality, such as about 3 voters in the whole state in large counties having quotations somewhere in the row that contains their voter information. These were excluded, and had no real impact on the information provided by these data files.

The voting records are separated into the different counties, named by a "county code" with a date stamp suffix.

Zip Codes:

There is a powerful spreadsheet that contains information on the population, Hispanic population, Hispanic proportion, and the individual proportions of each ethnicity within the "Hispanic" ethnicities, for all 984 zip codes in Florida. This is exceptionally powerful because it contained population information in the same place as the Hispanic population distribution.

From an external source, a list of all of the zip codes in Florida organized by county was used. This was powerful and useful because the voter registration files are organized by county, and having the zip codes also be organized by county adds to the consistency of the provided data.

All of the zip code data is rather complete. The only difference is that the list of zip codes organized by county includes a more complete list. The zip code data with the Hispanic proportions only contains "Non-Unique" zip codes, which are zip codes that contain permanent addresses. The list of zip codes in Florida contains Non-Unique, Unique, and PO Boxes. Unique zip codes are zip codes where the mail is delivered to the office box, and then the mail is distributed internally by a third party. PO Boxes are zip codes that contain post office boxes. Unique and PO Box codes can be disregarded as they don't contain any permanent addresses, and therefore no voter registrations.

Names:

Used to predict ethnicity, there is a list of the most common 1003 Hispanic first and last names in each Hispanic ethnicity. The proportion of the ethnicity which each name is also provided. This database is very complete, and there were no modifications needed to be made in terms of completing any missing information.

2.2 **Statistical and Computational Methodology**

Sub-Ethnicity Scores:

The first half of the scope of the project was to calculate the probability that a voter given the zip code. This was done using methodology adapted from a **journal by Bernard Grofman and Jennifer R. Garcia.** The probability that the voter was of each ethnicity was calculated, using their name, and zip code. The following is an explanation of how this was done.

A python program was developed that filtered all of the provided data. From the inputted county and zip code, a table of all of the Hispanic voters and their information was returned. The information was stripped down only to include their name, county, and zip code since that's the only information needed for the scope of this problem. For each name, the program consults the first name and last name databases to find the proportion of each Hispanic ethnic group that also has those same names.

Adding a factor into this, many of those who identify as Hispanic tend to have 2 last names, hyphenated. The program separates these last names into two different names, finds the probability of each, and saves them both.

From these proportions, we get standardized probabilities, using an adaptation of Bayes Theorem. It is explained by **Grofman and Garcia** in their journal on estimating ethnicities in depth. Essentially, all of the proportions retrieved from each ethnicity for the given name are summed together, and each proportion is divided by the sum in order to get *probability.* Probability is treated differently because is not just a count of all people with a name, it is a likelihood that the person with that name is of that ethnic origin. This is of much more interest than just a raw probability.

A similar process of name retrieval also is done when the zip code constraint is considered. The zip code constraint prevents the amount of voters of one of the Hispanic groups to be higher than the amount of Hispanics in the zip code overall. The Hispanic proportions of the zip code are saved, and then standardized as well to get yet another probability.

When the first name probability, last name probability, and zip code probability are multiplied, you get a likelihood that the given voter is a certain ethnicity. If all ethnicities are done with this process, and then the likelihoods are also standardized, the result is a list of probabilities for each ethnicity that the voter is of a certain ethnicity. In many instances, there is one ethnicity that is much higher than the rest, such as 70% for one, and the rest are 10% or less. It is then assumed that the ethnicity with a spike in probability is the voters' ethnicity.

This process is then repeated for all voters in a given zip code, in a given county. This is a far from perfect model for calculating probability of ethnicity, but results can still be used from this model. There are improvements that can be made, such as excluding names that aren't contained in the names databases. These would provide more accurate results for a zip code's voting distribution.

Zip Code Analysis:

A separate approach to this was taken for the second half of the project. More python functionality was created to perform Exploratory Data Analysis on any given zip code in Florida. For the zip code, the function consults the voter registration file to count how many registered democrats, republicans, and "NPAs" are registered for any zip code. Then, the database

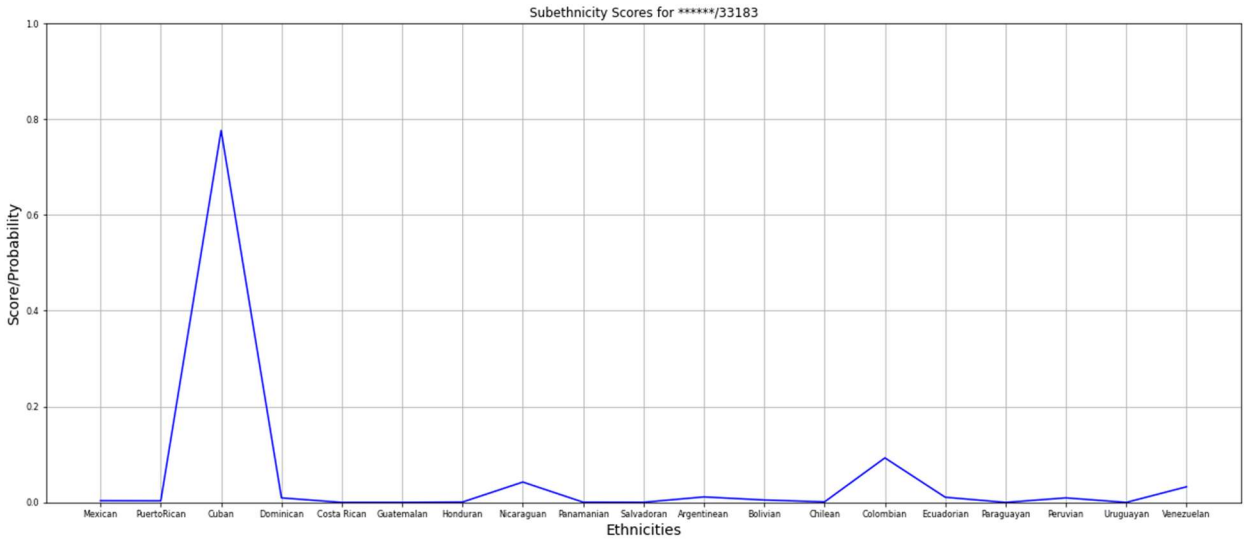
containing the zip codes population data is consulted to find the top 3 ethnicities in the zip code, and their respective proportions. All of this data is saved in a series of variables.

This process was then generalized to iterate for every county in an inputted zip code. This was done using the list of zip codes organized by county, to iterate for each county. Then, the whole process was repeated for all 67 zip codes in Florida.

III. Results

3.1 **Exploratory Data Analysis (EDA)**

Ethnicity Prediction:



This is an example plot of a Sub-Ethnicity probability score sheet. So far, the model only can calculate sub-ethnicity scores for individual names, and zip code data is hard coded at the moment. However, this information is valuable since it gives us an idea of what types of scores are produced by the model. This specific name and zip code combination has an overwhelmingly high probability of the individual being of Cuban ethnicity. We can infer that because of this, the person is Cuban.

Upon considering this plot, the name and zip code is fitting all of the pieces to a puzzle to seamless prediction.

nan		None	Cuban
nan		None	Colombian
nan		None	Cuban
nan		None	Cuban
nan		None	Cuban
nan		None	Cuban
nan		None	Cuban
nan		None	Dominican
nan		None	Cuban
nan		None	PuertoRican

The chart above is an excerpt from the DataFrame returned from the zip code entry '33139,' which the last row is the calculated ethnicity of the corresponding voters.

Zip Code Analysis:

(Explain zip code EDA script, table produced, etc.)

3.2 **Interpret and Investigate In-Depth**

At this point, it is very likely that the returned ethnicities are highly inaccurate. As it stands, the model does currently calculate the ethnicity using all the factors of first, middle, last, 2nd last if they have one, and their zip code. If any one of these 4 is a 0, then the zip code becomes dominant and the list returned ends up just being more or less random, proportional to the inputted zip code. The model can be improved to only consider the most predictive names that the voter has. Such as, if the voter has a non-Hispanic last name, then maybe their first name and middle name will be used.

(Explain the regressions, and their significances).

IV. Discussion

At this point in the project, there is a list of a voter ethnicity corresponding with their name and zip code in the same table as their voter party registration. This is good, except the model still needs some improvement. However, once this improvement is completed, we will be ready to investigate these trends in voter registration and come up with any of the conclusions we need to.

Once the model is improved, the method for the data analysis will be as follows. 2 map plots will be used, one with how democrat/republican a zip codes' Hispanics voted, and another that represents the zip code's largest Hispanic presence. This will give a visual representation of which communities are voting a certain way, and if there's any significance regarding that.

V. Reference

4.1 References

Grofman, Bernard, and Jennifer R. Garcia. "Using Spanish Surname to Estimate Hispanic Voting Population in Voting Rights Litigation: A Model of Context Effects Using Bayes' Theorem." *Election Law Journal: Rules, Politics, and Policy*, vol. 13, no. 3, Sept. 2014, pp. 375–393, 10.1089/elj.2013.0190. Accessed 22 July 2020.

Kandt, Jens, and Paul A. Longley. "Ethnicity Estimation Using Family Naming Practices." *PLOS ONE*, vol. 13, no. 8, 9 Aug. 2018, p. e0201774, 10.1371/journal.pone.0201774. Accessed 9 Mar. 2021.

Statistical Atlas. *The Demographic Statistical Atlas of the United States*. 2018. *US Census Bureau*, statisticalatlas.com/zip/33183/Ancestry.

List of Zip Codes in Florida (zipdatamaps.com)

VI. Appendix

6.2 **Data Retrieved**