EMBRY-RIDDLE Aeronautical University
SCHOLARLY COMMONS

Beyond: Undergraduate Research Journal

Volume 7 Article 1

July 2023

REU-DEIM Classification of Hispanic Voters in Hispanic Groups Using Name and Zip Code Data in Palm Beach, Florida

Kamila Soto-Ortiz
Embry-Riddle Aeronautical University, sotoortk@my.erau.edu

Follow this and additional works at: https://commons.erau.edu/beyond

 Part of the Applied Statistics Commons, Data Science Commons, and the Probability Commons

Recommended Citation

Soto-Ortiz, Kamila (2023) "REU-DEIM Classification of Hispanic Voters in Hispanic Groups Using Name and Zip Code Data in Palm Beach, Florida," *Beyond: Undergraduate Research Journal*: Vol. 7 , Article 1.
Available at: https://commons.erau.edu/beyond/vol7/iss1/1

This Article is brought to you for free and open access by the Journals at Scholarly Commons. It has been accepted for inclusion in Beyond: Undergraduate Research Journal by an authorized administrator of Scholarly Commons. For more information, please contact commons@erau.edu.

REU-DEIM Classification of Hispanic Voters in Hispanic Groups Using Name and Zip Code Data in Palm Beach, Florida

Cover Page Footnote

I would like to thank Dr. Mihhail Berezovski for their mentorship during this project. Support for this project has been provided by the National Science Foundation (NSF) through the REU Award Number DMS – 2050754.

This article is available in Beyond: Undergraduate Research Journal: https://commons.erau.edu/beyond/vol7/iss1/1

UNDERGRADUATE RESEARCH JOURNAL
Beyond

REU-DEIM Classification of Hispanic Voters in Hispanic Groups Using Name and Zip Code Data in Palm Beach, Florida

Kamila Soto-Ortiz, Mihhail Berezovski Ph.D.

Abstract

When it comes to registering to vote, Hispanic voters can only register as "Hispanic" in the "Race/Ethnicity" category, causing difficulties when analyzing voting trends amongst the Hispanic community. Upon the recent idea that not all Hispanic Groups vote the same, the goal is to create a model that can possibly identify a voter's Hispanic Group with the information provided on the public Florida voter file. This is accomplished using name and zip code data for all voters in Palm Beach, Florida. This paper will explore the model implemented, its findings and limitations. Palm Beach, Florida, is met with low confidence in classification, leaving the final sample of highly confident active Hispanic voters with 15% of its original sample. Further analysis on other counties will be needed to gauge how impactful this limitation might be on the rest of the state.

Introduction

The Hispanic community is a key demographic vote in Florida, given they make 26% of the state's population [9]. Over the last decade, Hispanics have had a population growth of 52% in the United States of America [10]. These facts make understanding the Hispanic vote crucial for future elections. Recently, the idea that not all Hispanic Groups vote similarly [1][2][7] has brought an important obstacle when analyzing voter behavior from the Florida public voter file. When Hispanic Floridians register to vote, the only option in the "Race/Ethnicity" category is "Hispanic", which limits the research done on the publicly available data. A process is needed to classify voters in a Hispanic Group. The voter's name and zip code are taken from the public voter file and paired with additional data containing the most popular names at the country the Hispanic Group originates from and the breakdown of Hispanic Groups per zip code. Only 15% of the Active Hispanic Voters (AHV) yield high confidence results (>=75%).

To test the model, the investigation focuses on a single county. The chosen county is Palm Beach County. Out of 260,003 Hispanics in this county [3], 66% are registered to vote. Out of all registered Hispanic voters, 31% are considered passive voters –having registered but never voted–, while 69% are considered active voters. Palm Beach County has a total of 118,926 AHV, 12% of all active voters. Its population of AHV has a 1.26 female to male ratio, while the party affiliations are distributed as: 43% with the Florida Democratic Party, 37 % with the Non-Party Affiliated, 19% with the Republican Party of Florida and 1% with other parties.

Most of the data is extracted from the public Florida voter file. Voter registration and voting history information is public record under Florida law [3]. There are three different motives for the use of the following data points. County Code, Voter ID and Race are used for identification purposes; Gender, Party Affiliation and Date of Birth are used for demographic purposes; and Last Name, First Name, Middle Name and Residential Zip Code are used for the classification model. For AHV, 95% have data available within the model implemented in this paper, meaning their voter file data is matched to the additional data and analyzed through the model. The remaining 5% are classified as

False Negatives, which are voters that identify as Hispanic yet match no name data and are thus unclassifiable. The opposite of False Negatives is False Positives, Non-Hispanic voters that match Hispanic name data. These voters pose little to no problem, as the model disregards any voter that is not Hispanic.

The additional data is used to match the voters' name and zip code to their voter file. The name data [4] is composed of the one thousand most popular names in each Hispanic Group's country of origin. The zip code data [5] breaks down the Hispanic Group's percentages of population at a zip code. These are the main elements in the model.

Modeling Hispanic Voter Classification

The following are concepts to be used throughout the writing.

Let

S_v = sum of data in all Hispanic Groups for Last Name 1

P_v (Hispanic Group) = probability of Last Name 1 for a Hispanic Group

S_w = sum of data in all Hispanic Groups for First Name

P_w (Hispanic Group) = probability of First Name for a Hispanic Group

S_x = sum of data in all Hispanic Groups for Last Name 2

P_x (Hispanic Group) = probability of Last Name 2 for a Hispanic Group

S_y = sum of data in all Hispanic Groups for Middle Name

P_y (Hispanic Group) = probability of Middle Name for a Hispanic Group

P_z (Hispanic Group) = probability of Zip Code for a Hispanic Group

V (Hispanic Group) = final probability calculation for Last Name 1

W (Hispanic Group) = final probability calculation for First Name

X (Hispanic Group) = final probability calculation for Last Name 2

Y (Hispanic Group) = final probability calculation for Middle Name

Z (Hispanic Group) = final probability calculation for Zip Code

p (Hispanic Group) = final probability for each Hispanic Group per voter

To calculate the final probabilities for all categories (V, W, X, Y), the data must be examined first. While it is commonplace for Hispanics to share multiple last names, this is not always the case. Of the voters, only 14% did, with 10% of those having available data within the system of names [4]. In that scenario, the data for voters in the category of Last Name 2 is sparse, limiting the operations utilized. To account for this, a controlled sum is made.

$$V(\text{Hispanic Group}) = \begin{cases} S_v > 0 \rightarrow P_v(\text{Hispanic Group}) \\ S_v = 0 \rightarrow 1 \end{cases} \quad (1)$$

Equation 1 is recreated for all different categories, V, W, X and Y.

Using the probability of independent events that a voter will have all categories lead to a Hispanic Group, the categories, now adding Z, are multiplied.

$$p(\text{Hispanic Group}) = V * W * X * Y * Z \quad (2)$$

Hispanic Group Classification: A Case Study

The subgroup classification process requires two main components to classify a voter on their Most Probable Spectrum (MPS). These are their full name and their zip code, which are combined using Equation 2. An example for a single voter shows all datasets that influence their resulting group.

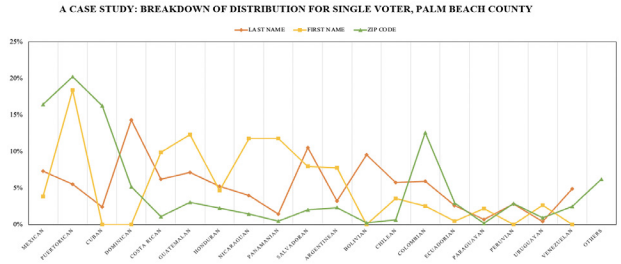

Figure 1: Different categories such as zip code, last name and first name for single voter

The biggest peaks for the single voter distribution are Mexican, Puerto Rican, Dominican and Colombian. Because there is a high percentage of Mexicans in the Palm Beach zip code, this will most likely bring up their likelihood, although the peaks for first and last name are not as high compared to other Hispanic Groups. For zip code 33467, the biggest group is Puerto Ricans, while also having a high first name peak, despite their low last name peak. Their last name is most popular in the Dominican Republic, although their first name is not at all. The percentage of Dominicans in this zip code is not quite high, although the Dominican peak will most likely still be noticed. Finally, Colombians have a somewhat noticeable peak in first and last name along with enough Colombians in the zip code.

All probabilities and ratios are brought together with Equation 2.

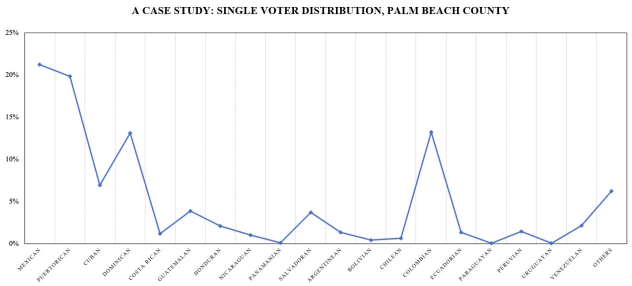

Figure 2: Single Voter Distribution

Their MPS, or the highest peak, is Mexican. The voter's MPS has a low confidence in its calculation, 21%, while also providing other high probability groups. Their probability of being Puerto Rican or having Puerto Rican ancestry is 20%, which is a 1% difference from their MPS. This scenario of high uncertainty happens often, leaving the voter unclassifiable.

To address the occurrence, low confidence voters are filtered out of the sample at the threshold of 75%. A voter whose MPS is lower than 75% won't be considered as a highly confident classification. Thus, the voter represented in Figure 2 is not considered in the sample for the county. The filtering of low-confidence voters leaves the sample with 18,298 AHV, 15% of the original AHV sample in Palm Beach.

Results for Classification Process

After doing the same calculation for all voters that fall in the sample, the distribution for the county is plotted.

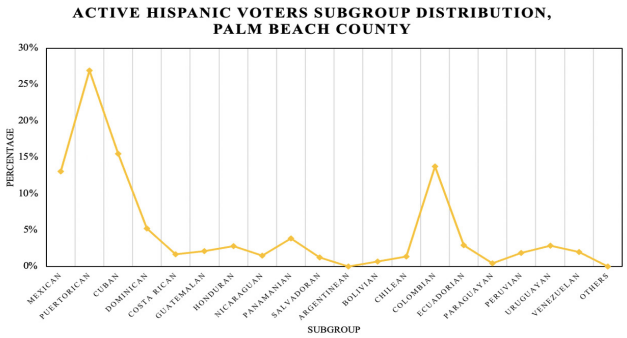

Figure 3: General Distribution of AHV in Palm Beach County

The largest group of high-confidence AHV in Palm Beach County are Puerto Ricans at 27%, followed by Cubans with 15% and Colombians with 14%. Following the model for the subgroup classification process, 97% out of all AHV in Palm Beach have name data available for the classification.

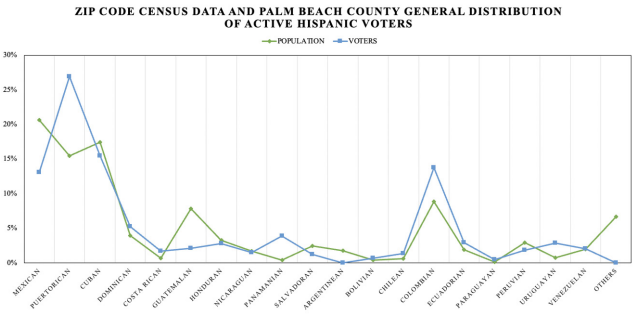

Figure 4: General Distribution of AHV compared with the Hispanic population distribution at Palm Beach County

In Figure 4, a comparison between the zip code census data distribution and the results in the classification is made to compare their similarities. Despite not all groups following the exact distribution of Palm Beach Hispanics, it can be expected that certain groups register to vote more often than others.

Conclusion

The high confidence voter distribution of Palm Beach, Florida presented a high percentage of Puerto Rican and Colombian voters. The

distribution calculated follows the population closely, with exceptions that could be more thoroughly investigated in a future project. Its model is deemed successful with limitations, although the scope and severity of these could be better understood if applied to a wider dataset, such as the state of Florida instead of a county.

Acknowledgements

Support for this research has been provided by the National Science Foundation (NSF) through REU Award Number DMS - 2050754.

References

[1] Boryga, A. (2020, October 24). Democrats push Puerto Rican voters To outmuscle CUBAN Republicans in Florida. Retrieved April 18, 2021, from https://www.sun- sentinel.com/news/politics/fl-ne-democrat-push-puerto-rican-voters-florida-20201024- 6m6rwvuf6vbg5jl5mnq2cqaoxa-story.html

[2] Busette, C., & Shiro, A. (2020, November 06). The importance of Understanding Latino voters in battleground states. Retrieved April 18, 2021, from https://www.brookings.edu/blog/how-we-rise/2020/11/03/the-importance-of-understanding-latino-voters-in-battleground-states/

[3] Florida Department of State. (2022). Voter extract disk request. Florida Division of Elections. Retrieved May 25, 2022, from https://dos.myflorida.com/elections/data-statistics/voter-registration-statistics/voter-extract-disk-request/

[4] Forebears. (2012, June 20). Retrieved April 1, 2021, from https://forebears.io

[5] Hispanic Latino population by specific origin by county. (2010) Retrieved February 26, 2021, from http://proximityone.com/hispanic_origin.htm

[6] Instituto Cervantes. (2020). El Español: Una Lengua Viva: Informe 2020. Retrieved April 17, 2021, from https://cvc.cervantes.es/lengua/espanol_lengua_viva/pdf/espanol_lengua_ viva_2020.pdf

[7] Krogstad, J. (2020, October 02). Most Cuban American voters identify as Republican in 2020. Retrieved April 18, 2021, from https://www.pewresearch.org/fact-tank/2020/10/02/most- cuban-american-voters-identify-as-republican-in-2020/

[8] Lopez, M., Krogstad, J., & Passel, J. (2020, September 22). Who is hispanic? Retrieved April 18, 2021, from https://www.pewresearch.org/fact-tank/2020/09/15/who-is-hispanic/

[9] U.S. census bureau QUICKFACTS: Florida. (n.d.). Retrieved February 26, 2021, from https://www.census.gov/quickfacts/FL

[10] Vogel, M. (2020, August 26). Florida's Hispanic population boom. Retrieved February 26, 2021, from https://www.floridatrend.com/article/29770/floridas-hispanic- population-boom