Projecting the 2020 Census: A Study of Hard to Count Communities

**Abstract**

This paper utilized Census Bureau data to examine the relationship between demographic factors (i.e., race and ethnicity) and Census response rate for communities in Texas labeled as “Hard to Count.” These areas were broken down by Census tract and were considered hard to count if the mail return rate was 73% or lower (CUNY, 2017). It was hypothesized that there would be a positive correlation for response rate and persons identifying as white and a negative correlation for response rates and people of color, since people of color are more likely to be undercounted in the Census. The results of this project found no correlation between 2010 Census response rates and Texas residents broken down by race and ethnicity.

**Background**

The purpose of this project is to use US Census data and trends to analyze if there is any similarity in demographic information (e.g., race and ethnicity.) for areas of Texas (e.g., Census tracts) that are labelled as “Hard to Count Communities” in. This project examined publicly available data from the Census 2020 Hard to Count map by the City University of New York. This data was originally provided by U.S. Census Bureau from the previous decennial Census in 2010 and from American Community Survey (ACS) population estimates from 2013-2017. Specific data includes mail-in response rates from the 2010 Census and demographic population information and internet access from the 2013-2017 ACS.

With the approach of the 2020 Census and the controversy about a potential citizenship question being added to the Census, there is a large amount of concern over obtaining an accurate and clear Census count. This research question is important as Census data has far reaching implications. For example, federal funding for states and local governments is allocated based on Census data (CUNY, 2017). Census data is also used in planning for population growth and development, like investments in infrastructure, transportation, and education. In Texas specifically, Census data is incorporated into determining state legislative and Congressional districts and representation, a practice that has been questioned in relation to ethics (CUNY, 2019).

Populations that have historically been undercounted in the US Census tend to include communities of color, immigrant households, households with limited English proficiency, households of lower socioeconomic status, children, and rural communities (CUNY, 2017). This undercounting then goes on the disproportionally impact these groups in a myriad of ways. As a result of this power imbalance, a number of transformational theoretical frameworks that specifically discuss power dynamics and sociological factors came to mind in throughout this project. Mainly, Critical Race Theory, Intersectional Theory, Systems Theory, the Advocacy Coalition Framework in policy studies.

**Research Question**

The research question for this project is what relationship do demographic factors have to Census 2010 response rates in Hard to Count Communities? For the purposes of this study, only the demographic variables of race and ethnicity for the total population will be measured. Race and ethnicity were studied as the independent variable while mail-in response rate to the 2010 Census was the dependent variable.

The hypothesis for this question was that response rates would have a positive correlation with white communities and a negative correlation with populations of color based on literature about Hard to Count Communities and who usually is at risk of being undercounted in the Census.

**Methodology**

In its initial form, this project intended to merge and compare the previously mentioned Texas hard to count data with another dataset that contained information about Texas general elections. To acquire the data needed for this project, a search for this data on the *Texas Secretary of State*’s website occurred, however, the available data was not disaggregated according to demographic information or was extremely outdated.

After failing to find an adequate dataset on from the Texas Secretary of State, the *United States Elections Project* (McDonald) site was then searched for a data source that contained voter turnout information. However, the Voting Age Population data for “State Turnout Rates” was not disaggregated by race and ethnicity, or by Census tract, which was unhelpful for the context of this research project.

The Integrated Public Use Microdata Series (IPUMS) website was also browsed for panel data that may have contained demographic voting information form wither the US Census, American Community Survey, or Current Population Survey. This data options on this site, unfortunately, were also not sufficient for the confines of the original research question.

At this point, since so much time had been spent in the data acquisition phase without luck, a return to the project planning phase was necessary. After consulting with the course instructor and teaching assistant, it was decided to restructure the project focusing on the data source from CUNY’s Center for Urban Research. After a return to the data planning phase, a new research question and hypothesis were formed. Instead, the project would look at different categorical variables (e.g., population demographic information) and examine how they were correlated with mail-in response rates for the 2010 Census.

**Data Management Plan and Process**

**Workflow**

With the challenges of finding data and ultimate shift in research question from the original project, the workflow for the project detoured from what may be considered the traditional route. The original plan was to find at least two different data sources and compile the data sets before margining them. Raw and cleaned files would be kept and organized by type. A data dictionary would also be created to explain the merged files. However, with the use of only one dataset, this simplified the original data management plan significantly and the structure of the workflow. Many of the previously planned data management steps were no longer applicable to the process since there would not be a merging of data sets. More efforts were placed into data acquisition and analysis, as well as data validation and replication.

**Data Acquisition**

The data source (i.e., "pdb2015tract\_2010MRR\_2017ACS\_TX.xlsx ) used is from the Census 2020 HTC Map Application. This application was created by the Center for Urban Research at the City University of New York Graduate Center's Mapping Service. The dataset contains three datasets that have been merged into one excel file. The original source of the data is from directly from the US Census Bureau and contains 2010 Census mail return rate data and the 2013-2017 American Community Survey (ACS) data. The ACS 2013-2017 data contains both population estimates and internet access estimates for the five year survey. ACS data is available through the American FactFinder website run by the US Census Bureau.

The excel file specifically for Texas (i.e., “TX Excel File”) was utilized for all data analysis in this project. The TX Excel file contains two worksheets: one (titled "pdb2015tract\_2010data\_ACS17\_TX") with Census Bureau data and a second (titled "Fieldnames") with a description of source information and names for columns/fields.

For file organization, three separate folders within a large project folder (“final\_project”) were created. One folder contained the data for the project, another folder held visuals, graphs, charts, and screenshots for the project, while the final folder contained documents for sharing, like the project presentation and final paper. The naming conventions for the files consisted of maintaining original names for data files and utilizing underscores, rather than spaces, for naming files and folders. A Microsoft Word document was also kept that served as a lab notebook for data management plans, processes, steps, and updates.

**Operationalization**

Since the restructuring of the original research question, this project only required the comparison of a few variables. This project was primarily interested in the demographic information for race and ethnicity in Texas. These categorical variables were listed on the second sheet from the ACS 2013-2017 Total Population Estimates (from FIELDNAMES tab on "pdb2015tract\_2010MRR\_2017ACS\_TX.xlsx" sheet). This sheet also served as a data dictionary of sorts later for coding variables. The categorical variables for race and ethnicity were operationalized for the independent variables in this project. These specific variables for Texas Census Tracts are listed below:

* TotPopACS17 – 2013-2017 ACS total population estimate
* WhiteAloneOrCombo – White alone or in combination with one or more other races (ACS 2013-2017 table B02009)
* BlackAloneOrCombo – Black or African American alone or in combination with one or more other races (ACS 2013-2017 table B02009)
* AmerIndAloneOrCombo – People who are American Indian or Alaska Native alone or in combination with one or more other races (ACS 2013-2017 table B02010)
* AsianAloneOrCombo – Asian alone or in combination with one or more other races (ACS 2013-2017 table B02011)
* Hispanic – Hispanic population\* (ACS 2013-2017 table B03002)
* Census Bureau site refers to this table as “Hispanic or Latino origin by race”
* NatHawAloneOrCombo – Native Hawaiian and Other Pacific Islander alone or in combination with one or more other races (ACS 2013-2017 table B02012)

**Data Analysis, Sharing, Preservation, and Reuse**

Since the data was publicly available, it did not need to be analyzed. Additionally, since no merging occurred, edits to code in Google Collaboratory served as the major source of version and quality control for the data. Code in Python was overly noted and steps were thoroughly documented in Google Collaboratory. The source of all code is publicly available through a github repository. The github repository also contains all project files, folders, a README file, a Google Collaboratory Python notebook containing the Python code, lab notebook of STATA commands, and graphs. This repository can be found at <https://github.com/mirandabad/AEM_final_project>. For data security, files were backed up onto an external hard-drive weekly and pushed and committed to github.

**Replication and Validation**

After reviewing the output for the scatter plots in Python, the accuracy of the graphs above was in question, so an attempt to replicate and validate the data source in STATA occurred. The TX excel file was imported into STATA using the "import" command, then used the "summarize", "correlate", "regress", and "scatter" commands to compare the independent and dependent variables. The non-HTC tracts were again filtered out, matching the number of observations (i.e., tracts classified as hard to count communities) in the Python code.

Since the population estimate data is an estimate, there is an margin of error. Additionally, "that state totals do not reflect the sum for all counties in Texas and are aggregated from congressional district data" (via Census 2020 HTC site).

**Limitations**

Although a number of limitations existed in the process of this project, three major limitations will be discussed. The first limitation of this project is the large potential for of bias in the collection of Census data. Census data is subject to nonresponse, sample, and selection bias due to the number of populations that are undercounted communities that are undercounted don’t receive benefits; non-representative results. In communities with low response rates, the sample may not be representative of the overall population. Furthermore, the next decennial Census is at risk of experiencing more bias and potential undercounting with the possible inclusion of a citizenship question. This question may lower the response rate for communities that have larger immigrant populations or non-US citizens.

The second and final limitations is the is the challenge of validating and replicating Census data. With the Census being administered by a government entity, accessing raw data would be an extremely difficult process. Additionally, the Census Bureau’s website has a large number of links that are broken or inactive, which makes replication difficult. The Census Bureau site is also undergoing a transition where American FactFinder data is being moved to a new page in July 2019. This transition may explain why so many links to data on the Census site are inactive, broken, or could no longer be found. This lack of accessibility should encourage researchers to upload data files directly to their websites, or encourage the Census Bureau to better house their data.

**For Future Studies**

The original purpose of this project was to investigate the relationship between response rates and demographic characteristics for communities labeled as “Hard to Count” in Texas and data on voter turnout (in both midterm and presidential elections). This paper would have discussed the findings upon comparison of this data, speaking on the ethical implications of the results, and examining any limitations of this process, as well as potential solutions to any challenges that may arise or concerning trends found. Any correlations would have been examined to see if there was a connection between reported voter disenfranchisement and suppression in Texas. However, voting data in Texas (to the desired extent) was unattainable within the parameters of this project.

Since the population estimate data is an estimate, some margin of error exists within the data. More time would be spent looking at the data to determine the extent of this margin. Additionally, the Census 2020 HTC Map site notes "that state totals do not reflect the sum for all counties in Texas and are aggregated from congressional district data." This may be something else that needs to be looked into for future study.

**Conclusion – Why Does This Matter?**

The Census is an extremely important tool that impacts a large number of aspects in the US. Census counts determine the amount of federal funding states receive for social services programs, like Medicaid and the Children’s Health Insurance Program (CHIP). Census funds are also allocated foster care and adoption services, public education, businesses, and most notably, the number of US House representatives a state is allowed. In states that have been accused of gerrymandering district lines, its especially important to increase the response rate for communities that are at risk of being undercounted. These populations are less likely to receive adequate representation in legislatures on the state and federal levels, yet are more likely to disproportionally impacted by policy decisions.

The Census 2020 HTC site notes that people of color, immigrants populations, low-income households, young children in multi-generational families, LGBTQ+ individuals, people with disabilities, people with limited-English proficiency, single parents working multiple jobs, people experiencing homelessness, or individuals who are renters are historically at risk of being undercounted. This is why outreach and creation of local and state level Census Complete Count Committees is increasingly important to attain a more accurate count.

References

Center for Urban Research, CUNY Graduate Center. (2017, September). Census 2020 Hard to

Count Map. Retrieved from https://www.censushardtocountmaps2020.us/

Center for Urban Research, CUNY Graduate Center. (2017, September). *Tract Data – TX Excel*

*file* [Data file]. Retrieved from https://www.censushardtocountmaps2020.us/

Center for Urban Research, CUNY Graduate Center. (2019, March 3) Census 2020 Hard to

Count Communities in Texas. Retrieved from

https://www.censushardtocountmaps2020.us/img/mappdfs/Texas.pdf

McDonald, Michael P. (2019). "Voter Turnout: State Turnout Rates." *United States Elections*

*Project*. Retrieved from: http://www.electproject.org/home/voter-turnout/voter-turnout-

data/

US Census Bureau. (n.d.) 2013-2017 American Community Survey 5 Year Estimates. tables B03002, B02009-B02012. Retrieved from

<https://factfinder.census.gov/faces/nav/jsf/pages/searchresults.xhtml?refresh=t>

Appendix

All files mentioned in this paper are available via a github repository at: <https://github.com/mirandabad/AEM_final_project/edit/master/readme.md>.