**North Carolina Demography**

*Members: Rachel Richards, Raza Lamb, and Sarwari Das*

Article 1, Section 2 of the United States Constitution contains a seemingly simple directive: every ten years, the people of the United States must be enumerated (counted).[[1]](#footnote-1) The Census Bureau carries out this responsibility by attempting to survey every person in the country, an estimated 330 million.[[2]](#footnote-2) While this data is useful—to academic institutions, corporations, individuals, and various levels of government—the legal purpose of the Census is to ensure equal and fair representation in the federal government. Knowing precisely where people live is important for this goal in two respects: first, it determines how many representatives each state is allocated, but secondly, it also divides states into relatively equal portions for those representatives. Therefore, much attention must be paid to the accuracy of the Census, as potential errors can introduce bias in how people are represented in government.

Identifying errors and biased procedures are critical for understanding how the Census is inaccurate, and for improving the count in future iterations of the Census. There are also direct implications for the present. As mentioned previously, the Census is not only used for apportionment. Population estimates help local and state governments direct resources and build more effective policies. Private enterprises also use Census data to drive decisions, such as identifying new locations for storefronts or headquarters, building factories, recruiting employees, and conducting market research.[[3]](#footnote-3) With so many decisions built on Census counts and estimates, it is easy to see that a pervasive bias or error in the counts could easily propagate and cause significant harm to undercounted persons and communities. By identifying routinely undercounted areas and/or groups of people, the Census can produce more accurate, equitable, and actionable revised estimates.

There has been a significant body of work on evaluating Census estimates, but the gap this project attempts to fill is that of the sub-state level. There are two methods by which the Census evaluates its estimates. The first is the Post-Enumeration Survey (PES), a representative sample of households surveyed in depth and then matched against records in the Census.[[4]](#footnote-4) From this method, we can determine the net coverage error of the Census, as well as correctly included people, incorrectly-included people, and wholly imputed records. The PES evaluates errors at the national and state level.

The other method used to evaluate the Census is the Demographic Analysis (DA).[[5]](#footnote-5) Unlike the PES, this method is not survey-based but based on birth, death, and migration data. In this way, DA estimates the number of people residing in the United States at the time of the Census. This method has the advantage of not relying on survey participation but still requires a significant number of assumptions about the underlying data. DA is available only at the national level but has coverage estimates by specific demographic attributes (including race, sex, and age).

While both the PES and the DA provide helpful information about the quality of the Census, the granularity is not detailed enough to truly evaluate at a sub-state level. While the PES shows that there was not a significant under or overcount in North Carolina, it may very well be that specific counties and areas were overcounted, especially those with high minority populations.

**Project Goals**

For this project, we will work with the North Carolina Office of State Budget and Management (OSBM) to achieve the following:

1. Compare existing estimates of population and housing in North Carolina at various geographies (city, county, tract) to the 2020 Census counts to identify where undercounts and overcounts occur and if they correlate with various demographic attributes.
2. Utilize various datasets (details below) to develop population estimates independent of Census methods to develop unbiased estimates of undercounts and overcounts.
3. Suggest corrections to current population estimates based on research and findings.

This work will help the OSBM, as accurate and unbiased estimates are their core mission. Indirectly, this work will serve North Carolinians, whether they use population estimates or not. Less biased estimates will ensure equitable distribution of goods and services, both in the public and private sectors.

**Project Milestones**

The following section lists out the goals we hope to achieve by the end of this semester.

**Goal 1: Review current methodology** [Done]

Before conducting any analysis, it is essential to understand how population surveys are traditionally estimated. We review two key benchmarks used by the Census Bureau to evaluate the 2020 Census: the 2020 Demographic Analysis (DA) and the Vintage 2020 Population and Housing Unit Estimates. This process will provide critical information on the current methodology to estimate net coverage error, i.e, a metric to recognize an area as under/overcounted, which helps us develop our own estimation method. Further, given the unique challenges of conducting a population-level survey during the pandemic, we plan to spend time understanding the intricate implications of the stay-at-home orders, altered living arrangements, and general attitude shifts on the Census counts. Apart from operational problems—low response rate, for example—events like rapid in-migration/out-migration or high housing unit vacancy that can introduce potential biases into the current counts will need to be investigated.

**Goal 2: Evaluate and visualize census counts compared to previous and/or alternate sources of population estimates.** [due 28th October]

Success for this goal will be measured with a dataset containing true population counts, estimates, and undercount/overcount flags for counties, sub-counties, cities, and census tracts. This will be supported by plots of North Carolina denoting areas of misrepresentation.

*Sub Goal 1: Data Wrangling* [Done]

The data wrangling process involves collecting ‘true’ sub-state level population and housing unit counts of North Carolina, and comparing them to pre-existing estimates. This comparison defines the error between the two counts, and will be used to flag areas that are undercounted or overcounted.

* Defining true counts: Initially, the U.S Census Bureau’s population and housing unit estimates are treated as ‘ground truth’ to compare all other independent estimates to. They’re accessed using the Census population database.
* Defining estimates: The estimates for population and housing counts are contained in the American Community Survey (ACS)[[6]](#footnote-6) , a demographic survey program conducted by the [U.S. Census Bureau](https://en.wikipedia.org/wiki/U.S._Census_Bureau). This data is accessed using the Python API ‘census’[[7]](#footnote-7).

*Sub Goal 2: Defining Undercounted/Overcounted areas* [Done]

Several methods to estimate the error between counts and estimates are explored. So far, we’ve computed Absolute Error, Error between Census and Estimate, % Error, and Absolute % error. A coverage error of greater than 5% is used to flag undercounted/overcounted areas.

*Sub Goal 3: Visualizing areas of Undercount/Overcount* [Done]

Coverage error across North Carolina are visualized using geographical plots. This helps us hypothesize spatial reasons for being misrepresented in the Census.

*Sub Goal 4: Expanding research to other geographies* [Due 28th October]

All data needs to be collected on several sub-state level geographies, including counties, sub counties, cities and Census tracts. The data collected up to this point has only been at a county level, and we plan to soon expand it to other geographies.

**Goal 3: Determine whether certain demographic characteristics are associated with differences between Census counts and estimates** [Due 4th November]

Success will be measured with a comprehensive dataset that can be used to build accurate population estimates. Additionally, we expect to extract information for specific demographic traits that can help us suggest improvements and corrections to the current estimation methods.

*Sub Goal 1: Collecting demographic variables* [Done]

Once we’ve defined the count/estimate error, we’ll identify various demographic variables to compare this error against. This allows us to study commonalities between areas that are undercounted/overcounted to infer if there are specific characteristics that may lead the area to be undercounted/overcounted. A few of the demographic variables we collect for each sub-state level, with their relevant sources are:

* Growth rate: *U.S. Census Bureau database—estimates*
* Racial/ethnic characteristics: *U.S. Census Bureau database—American Community Survey*
* Age characteristics: [*U.S. Census Bureau*](https://en.wikipedia.org/wiki/U.S._Census_Bureau) *database—American Community Survey*
* Housing unit vacancy rates: *Census API—2020 Census Housing Counts*
* Special populations (College towns, prisons, military bases): *Census API—2020 Census Counts*
* Indicators of Component change (Births, Deaths, Migration): *Census API—estimates*
* COVID deaths: *Johns Hopkins Covid Research Data[[8]](#footnote-8)*

*Sub Goal 2: Identifying significant differences in demographic characteristics across misrepresented counties* [Done]

Once we have demographic traits for an area (counties for example), we group counties based on whether they were undercounted/overcounted and normalize the county-level counts for each variable with respect to the population estimate for the county. For each attribute, we compare means across the two groups and perform t-tests to evaluate if the difference in means is significant. If a particular attribute shows significant differences between undercounted (overcounted) and non-undercounted (non-overcounted) area, we plan to investigate this relationship further with regression models to predict error or likelihood of being undercounted/overcounted based on demographic factors, while accounting for other indicators like economy, geography so on. We’ll explore multiple linear models, logistic regression models and fixed effects/ hierarchical models.

*Sub Goal 3: Increasing number of potential predictors* [Due November 11th]

Other variables relevant to a geographical area will be explored to improve our estimates. This can include variables denoting economic conditions, geographical variables and other demographic variables.

*Sub Goal 4: Limiting potential to overfit* [Due November 4th]

Given the large number of features we’re working with, we also plan to use dimensionality reduction methods to explore a higher number of dimensions while limiting potential for overfitting. Methods like PCA will allow us to investigate whether there is information (variance) contained in counties that correlates with the percent of error.

**Goal 4: Develop independent population estimates for counties and other sub-state geographies in North Carolina.** [Due end of semester]

Success will be measured by estimates that are robust to the estimation method, and relatively less biased compared to pre-existing estimation methods. Although there is no specific metric we hope to achieve, our stakeholders will have a final say over the quality of our estimates. Ultimately, our stakeholders should be able to use our estimates to improve future counts in the Census.

*Sub Goal 1: Explore methods to create final population estimates* [early December]

We plan to explore loglinear models, mixture models, mixed-effects models, and Bayesian models with prior estimation. More experimental methods we hope to explore are geographically weighted regressions or GWRs, generating neighborhood “typologies” using classification methods and using sampling to improve the bias-variance tradeoff.

Traditionally, the DA produces a range of estimates — low, middle, and high — to account for uncertainty in the input data and methods used to produce the final estimates. We may attempt a similar range by varying the assumptions in each estimation method.

*Sub Goal 2: Explore range of datasets to support estimation method* [early December]

Our stakeholders have provided a list of datasets that we will explore before attempting the modeling process. They include:

* Individual voter registration data [sourced from NC SBOE Voting Registration Data]
* Births and Deaths data [sourced from US Vital Statistics records]
* County-to-county migration data [sourced from IRS Migration data]

Next semester, we’ll start building our estimates.

1. U. S. Const. Art. I, § 2. [↑](#footnote-ref-1)
2. Bureau, US Census. “2020 Census Apportionment Results Delivered to the President.” Census.gov. Accessed September 19, 2022. https://www.census.gov/newsroom/press-releases/2021/2020-census-apportionment- results.html. [↑](#footnote-ref-2)
3. Bureau, US Census. “Our Censuses.” Census.gov. Accessed September 19, 2022. https://www.census.gov/programs-surveys/censuses.html. [↑](#footnote-ref-3)
4. Bureau, US Census. “Post-Enumeration Surveys.” Census.gov. Accessed September 19, 2022. https://www.census.gov/programs-surveys/decennial-census/about/coverage-measurement/pes.html. [↑](#footnote-ref-4)
5. Bureau, US Census. “Demographic Analysis (DA).” Census.gov. Accessed September 19, 2022. https://www.census.gov/programs-surveys/decennial-census/about/coverage-measurement/da.html. [↑](#footnote-ref-5)
6. US Census Bureau. (2022, September 29). *American Community Survey (ACS)*. Census.gov. Retrieved October 20, 2022, from https://www.census.gov/programs-surveys/acs [↑](#footnote-ref-6)
7. *Census Data API User Guide*. (2021, October 8). Census.gov. Retrieved October 20, 2022, from https://www.census.gov/data/developers/guidance/api-user-guide.html [↑](#footnote-ref-7)
8. *Covid-19 United States cases by county*. Johns Hopkins Coronavirus Resource Center. Retrieved October 20, 2022, from https://coronavirus.jhu.edu/us-map [↑](#footnote-ref-8)