North Carolina Demography

Members: Rachel Richards, Raza Lamb, and Sarwari Das

Introduction

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.[1] The Census Bureau carries out this responsibility by attempting to survey every person in the country, an estimated 330 million.[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, and second, it 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 is critical 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] With so many

[1] U. S. Const. Art. I, § 2.

[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.

[3] Bureau, US Census. "Our Censuses." Census.gov. Accessed September 19, 2022. https://www.census.gov/programs-surveys/censuses.html.

decisions built on Census counts and estimates, it is easy to see that 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 many gaps exist at 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] 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] Unlike the PES, this method is not survey-based but instead 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 significant assumptions about the underlying data. The 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 undercounted or overcounted.

[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.
[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.

Project Goals

For this project, we worked with the North Carolina Office of State Budget and Management (OSBM) with the goal of achieving 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 occurred and how they correlate with various demographic attributes.
2. Utilize various datasets to aid in the development of population estimates independent of Census methods and data 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 creating accurate and unbiased estimates is 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.

Methods

Data

For this project, there were two broad categories of data required. First, we needed datasets that count and estimate the number of housing units and people in North Carolina and provide estimates of demographic and economic characteristics. These datasets were utilized in the first phase of the project on behalf of Goal 1, which was focused on identifying where undercounts and overcounts may occur and how they correlate with various demographic attributes. Secondly, we needed sources of data that are independent of survey methodology, which directly and indirectly provide information about housing, population, population flows, and demographics. These datasets were utilized in the second phase of the project on behalf of Goal 2, which is aimed at supporting the development of population estimates.

Table 1 in the appendix displays the datasets used for Goal 1, with their source, the information extracted, and the level of detail of their available geographies. The 2020 Census redistricting data file provides the actual counts of housing and population obtained by Census surveyors, while the county population totals contain estimates of population and housing, based on previous Census counts and estimates of population flows. These two datasets enabled our analytical approach: identifying large differences between the 2020 counts and the expected values from previous estimates. The remaining datasets provide demographic and economic attributes, which were used to identify correlates and potential predictors of large differences between estimates and counts.

Table 2 in the appendix displays datasets that were investigated with the potential of providing support to independent estimates. This table does not include all datasets that were investigated, but only those identified as having the highest potential impact and relevance.

Comparing Counts and Estimates

As mentioned previously, there are no existing measures of Census accuracy at the sub-state level. In order to create a proxy metric for accuracy, we compared the counts (housing and population) to the population estimates for the same time period.[6] The estimates are based on previous Census counts, adjusted for population inflows and outflows (i.e., births, deaths, and migration). For the purpose of this analysis, we treated the estimates as a “ground truth.” While this metric has potential biases, it allowed us to generate measurement error at the county level, using the formula below.

$$\%\ Measurement\ Error\ = \frac{(Census\ Counts - Census\ Estimates)}{Census\ Estimates} \cdot 100$$

The bias in this method mostly stems from uncertainty: are the Census counts or estimates more likely to be accurate? However, the concern at this stage is minimal, because regardless of which is more likely to be accurate, large relative differences are important in identifying coverage gaps.

After creating a metric for measurement error, the next step was to identify whether specific demographic or economic characteristics correlated with larger (or smaller) differences between estimates and counts. This was initially done in two ways:

1. We identified counties as undercounted or overcounted using a ±5% error threshold.[7] Then, we measured the difference in demographic groups between undercounted, overcounted, and non-undercounted or overcounted counties. Two-tailed t-tests were conducted to measure significance.
2. For each demographic group, we separated counties into low, medium, or high prevalence (separated by one standard deviation below/above mean). Then, within each group, we measured average measurement error.

[6] Census counts and estimates are both for April 1, 2020.
[7] This threshold was recommended by North Carolina State Demographer Micahel Cline.

In addition to the above methods, we also investigated methods that would allow us to simultaneously examine all the demographic and economic characteristics of counties simultaneously. In other words, looking at a specific demographic group's correlation with measurement error is a relatively small piece of the puzzle, given that many demographic features are highly correlated themselves. To approach this problem, we utilized dimensionality reduction to answer the general question: does overall variation within the data correlate with measurement error? Two distinct approaches were taken, and in both the data was appropriately standardized).

1. We conducted principal component analysis (PCA). We measured the proportion of variation contained in each component and visualized the correlation between the first two components and measurement error.
2. We grouped the data using both K-means and spectral clustering, varying the number of cluster groups from two to eight. Then, we compared the mean measurement error between groups.

Investigating Datasets for Building Estimates

After measuring the difference between estimates and counts and investigating correlation with demographic estimates, we began to investigate non-survey-based datasets (see Table 2). The purpose of investigating these datasets was to identify how they could fit into population estimates. To standardize the information collected, we explored each dataset using the following guidelines:

- How is the data accessed (freely available or by special request, requires special status, etc.)?
- How frequent is the data (daily/monthly/annually/other)? What time-period does one observation represent?
- How soon (compared to the present) is data available (i.e., what is the data lag)?
- How often (if at all) is the data revised? If revised, when is the data considered final?
- What is the source of the data (i.e., administrative, sample-survey)?
- What portion of the population does this cover (i.e., 65+, licensed drivers)?
- What demographic (or other) characteristics are provided in the data?

In addition to the above, each dataset was examined for potential missing or biased data, limitations, or other concerns.

Results

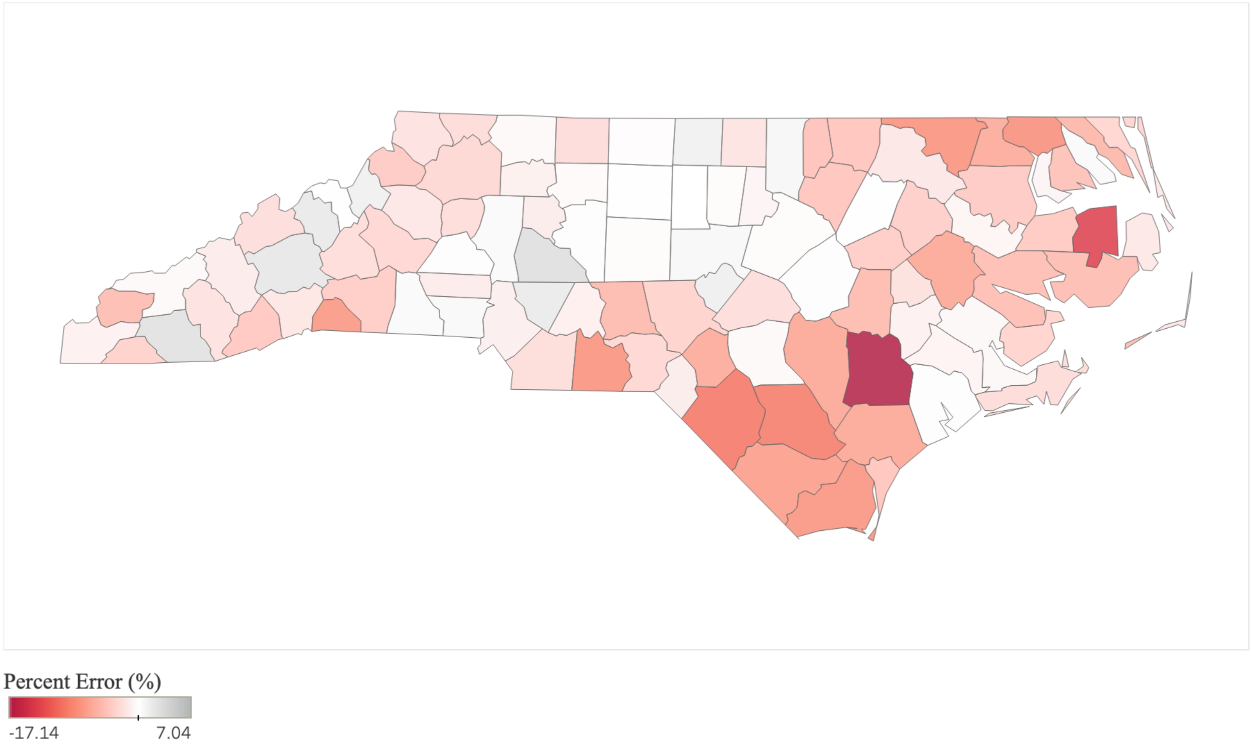
Comparing Counts and Estimates

First we identified undercounted or overcounted areas by generating measurement errors at the county level. Table 3 in the appendix shows this error across all 100 counties of North Carolina. By flagging counties with ±5% error threshold, there were 21 counties that were undercounted, and 79 that were not undercounted. No counties were overcounted by this definition. For the purpose of analyses, we focused on undercounted counties and non-undercounted counties. [8]

[8] On repeating this process with housing counts, we have 36 counties that are undercounted while 64 are not undercounted. Now, one county is overcounted. (Table 4 in appendix)

The errors were visualized geographically in Figure 1. We see that counties in eastern NC are more likely to be undercounted, with Duplin County showing the highest undercounts (17.14% below estimates). Additionally, it's apparent that some groups of counties that are geographically closer show comparable measurement errors. These regions may be hard to count for a myriad of reasons, including the demographics of the population, the socio-economic makeup, or particular geographic features that make it difficult to contact residents. These factors were investigated in the section below.

Fig 1: Percent Difference Between Census Counts and Population Estimates



Correlating Measurement Errors to County Specific Information

We first explored attributes like race, age, components of change and covid fatalities (Table 5). For each attribute, we compared means across the undercounted and non-undercounted counties and performed two tailed t-tests to evaluate if the difference in means was significant.

Previous research[9] has shown that areas with a higher percentage of black residents usually have a self-response rate lower than the state average, which makes them susceptible to being undercounted. (For reference, an Urban Institute report on the 2020 census[10] projected estimates for Black residents in North Carolina to have a net undercount of 2.35 percent). In our data, we find practically and statistically significant differences in the percent of black population attributed across undercounted/non-undercounted groups. On average, undercounted counties are 27.8% black, while non-undercounted counties are 18.1% black.

We also subset for counties where the college or military population was greater than zero and find average percent error in counts between counties with populations and counties without populations. We found significant differences in college dorm counts. (Table 6)[11]

To confirm our results, we decided to identify counties with relatively high shares of certain populations or characteristics, and then measure the average measurement error in those counties. We do this by separating counties into low, medium, or high prevalence (separated by one standard deviation below/above mean). Table 8 shows the results for this analysis for counties that have high prevalence of a certain county-level trait.

[9] Kramer, Melody. “Getting Everyone Counted in the 2020 Census – Carolina Demography.” https://www.ncdemography.org/2020/06/24/getting-everyone-counted-in-the-2020-census

[10] Simulating the 2020 Census: Miscounts and the Fairness of Outcomes (Washington, DC: Urban Institute, 2021).

[11] On repeating this process for housing counts, we find significant differences across undercounted and non-undercounted counties for % of black population, % of over 65 population, growth rate, deaths, natural increase, migration (both domestic and international), house vacancies and college dorms. (Table 5 and 7 in appendix)

From this table, it is visible that there is still a noticeably higher undercount in counties with higher Black populations. Continually, there are new insights to extract from this analysis. Most notable are counties with high populations of American Indian/Alaska Natives and counties with large populations in military housing. While the counts of these counties are relatively small (6 and 3, respectively), the difference in error is noticeable.

Separate from demographic makeup, an area's economic and social characteristics have significant effects on its development and need for various types of public programs. Our next set of analysis focused on a range of economic and social characteristics of a county. We started with county-level typology codes provided by the USDA Economic Research Survey (ERS)[12], which classify counties into six mutually exclusive categories of economic dependence including farming, mining, manufacturing, Federal/State government, recreation, and non specialized counties. We chose specific codes that would be of interest for the state of North Carolina and created indicator variables for whether a country belonged to one of the following categories: Manufacturing or Farming, Recreation, Retirement and Metro, and Retirement and Nonmetro.

This analysis did not result in significant results (Table 9). The largest difference was between counties that are both metropolitan and retirement counties and those that are not, but even this difference was relatively muted (-3.84% error and -2.59% error, respectively).

[12] USDA ERS - County Typology Codes. Accessed December 7, 2022.
https://www.ers.usda.gov/data-products/county-typology-codes/

Analyzing Overall Variation in Data

In addition to the above methods, it is also of interest to investigate the demographic and economic characteristics of counties simultaneously. We know that many demographic variables are correlated to each other, which led us to use dimensionality reduction techniques to determine if overall variation within the data can help predict coverage error. We performed Principal Component Analysis (PCA), which is an unsupervised linear transformation technique that helps us identify patterns in data based on the correlation between features. In a nutshell, PCA aims to find the directions of maximum variance in high-dimensional data and projects it onto a new subspace with equal or fewer dimensions than the original one. There are significant advantages from this process, including that it now is possible to visualize the data in a few dimensions, which is not possible using all of the information contained in counties. However, there are also a few disadvantages. Most significantly, the variance is not easily connected back to variables, so we cannot easily determine the "meaning" of the components.

We performed PCA on our dataset of over 40 variables and reduced it into orthogonal components that individually help contribute to the variation within our data. First, we found that while the correlation between variables exists, it still takes a significant number of components to explain 90% of the variance (15 components). However, approximately 40% of all variance is explained in the first two components, which shows potential for reducing dimensions. (Figure 3 in appendix)

In order to approach the same question from a slightly different method, we used clustering to create class labels on unlabeled data. The purpose of using this was to create county categorizations which may help us find groups of counties experiencing larger measurement errors. The advantage, compared to PCA, is that it is easier to examine which factors may cause (or correlate with) groupings.

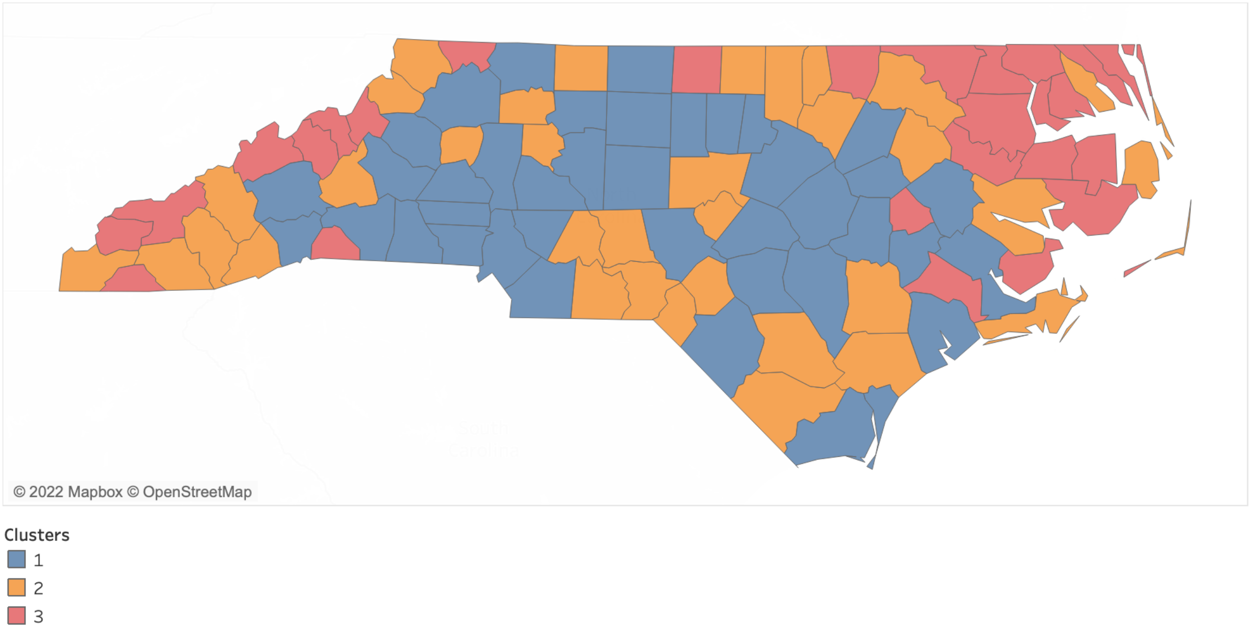
We used both spectral and K-means clustering and found that K-means achieved better groupings (from the perspective of achieving error separation between groups). We varied cluster groups between three to eight to find that three clusters were optimal; they gave good separation, while still being a manageable number. Descriptive statistics of these clusters are listed in a table below.

Table 9: Error % across Cluster Label

Cluster 1	-1.66%
Cluster 2	-3.34%
Cluster 3	-3.68%

Below, counties are graphed by their cluster number for geographic representation (Figure 2). The groups show some clear distinction—cluster group 1 (with the smallest error) represents high population (metro) counties, while cluster group 2 appears to represent some of the poorest counties, in the flatter parts of the state. Cluster group 3 represents rural western (and beach) counties.

Fig 2: Three county cluster classification



Investigating Datasets for Building Estimates

The next stage of our investigation involved looking into independent, non-survey-based datasets that could inform us of the many alternatives to take into account when developing population estimates. We give a high-level overview of a few of these datasets below, highlighting the drawbacks of each. More information can be found by contacting the authors.

Voter Registration Data : Sourced from the National Voter Registration Act data, this dataset provides current voter-level registration records and snapshots of voter registration records across 15+ years. It includes the most up-to-date publicly available information for individuals registered or formerly registered to vote in North Carolina, as well as individuals who have attempted to register or have steps left uncompleted in the registration process. While it is freely accessible and free of cost, it is relatively large to work with which can make initial computations difficult.

North Carolina Parcels Data: Generated to publish an aggregated set of parcel polygons for counties to serve business needs, this dataset contains significant amounts of information about who owns the parcel, purchase date, parcel purpose, and tax information. However, this too is large to work with (around 6 GB), which makes computation difficult. Secondly, while there is a state standard for reporting information, information across counties may not be standardized.

IRS Migration data: This data is based on year-to-year address changes reported on individual income tax returns filed with the IRS. It presents migration patterns by State or by county for the entire United States and is available for inflows and outflows. However, those who are not required to file United States Federal income tax returns are not included in this file, and so the data under-represents the poor and the elderly. Also excluded is the small percentage of tax returns filed after late September of the filing year.

USPS Vacancy data: Reported by USPS postal workers out on routes, this dataset contains the number of total addresses by Census tract, divided into residential and business. Then, it includes various measures of vacancy like total count of vacant addresses, vacant count by length of vacancy, and so on. The data is split into "Vacancy" counts and "No-Stat" counts, and this distinction could make it complicated to calculate accurate vacancy rates, as in some areas "no-stat" counts may not be appropriate to report as vacancies, while in other areas they may represent vacancies. Continually, areas with high vacancy likely correlate to both areas with large growth and significant decline, making it more challenging to use this dataset, at least individually, to separate areas of population growth and decline.

Conclusion

Our work builds on existing research and Census Bureau methodology to provide estimates of accuracy and potential bias in the 2020 decennial Census at a sub-state level. In our analyses, we found that generally, the state's population is lower in the decennial Census than predicted by previous estimates. These differences are noticeably larger in eastern North Carolina and correlate with certain demographic features. Our work confirms other research suggesting that Black Americans may be undercounted. We found that undercounted counties have a higher black population than non-undercounted counties, and that counties with high Black populations had greater differences between estimates and counts. In addition, we also find that other populations may be undercounted, including persons living in military living quarters and American Indians/Alaska Natives.

In addition to specific demographic groups, our clustering analysis shows that variation between estimates and counts may be predicted by variation within demographic and economic data as a whole. In other words, there may be "types" of counties at high risk for being undercounted (or overcounted). This work can help direct future adjustments to Census estimates and identify areas that may have biased counts.

Next Steps

While significant work towards project goals has been made, there is still additional work to be done before the project is finalized. The immediate next steps are highlighted below.

- Replicate work done at geographies smaller than county. This includes metropolitan and micropolitan areas, and census tracts (when available). Challenges at this level of detail will be more significant, as differential privacy plays a larger role, and data may be less reliable. However, this work may be crucial in identifying significant demographic or geographic patterns not visible at the county level.
- Continue to investigate datasets for production/alteration of alternative estimates. Current investigation of datasets has focused on general properties. Next, we will incorporate work from the analysis stage into dataset investigation. Specifically, we will work to identify datasets that may be helpful at either identifying specific demographic groups of people (to compare to counts), or in specific regions that may be undercounted.

Appendix

Table 1. Population, Housing, and Demographic Datasets			
Dataset	Source	Information	Geography
2020 Census Redistricting Data[13]	U.S. Census Bureau, Decennial Census API	Population/housing counts, special living populations (college, jail, military, etc.)	State, county, tract, block, place
County Population Totals[14]	U.S. Census Bureau	Population/housing estimates	County
American Community Survey 5-Year Data[15]	U.S. Census Bureau, American Community Survey API	Population by age, race, and sex	State, county, tract, place
Covid-19 United States cases and deaths by county[16]	Johns Hopkins University	Daily Covid-19 deaths and cases	County
County Typology Codes[17]	Economic Research Service, U.S. Department of Agriculture	Metro/non-metro, economic dependence, recreation	County
County Economic Datasets[18]	Economic Research Service, U.S. Department of Agriculture	Unemployment, household income, poverty, education	County

[13] Bureau, US Census. “Decennial Census P.L. 94-171 Redistricting Data.” Census.Gov, https://www.census.gov/programs-surveys/decennial-census/about/rdo/summary-files.html. Accessed October 10, 2022.

[14] Bureau, US Census. “County Population Totals: 2010-2020.” Census.Gov, https://www.census.gov/programs-surveys/popest/technical-documentation/research/evaluation-estimates/2020-evaluation-estimates/2010s-counties-total.html. Accessed 3 Dec. 2022.

[15] Bureau, US Census. “American Community Survey 5-Year Data (2009-2020).” Census.Gov, https://www.census.gov/data/developers/data-sets/acs-5year.html. Accessed October 10, 2022.

[16] Covid-19 United States cases by county. Johns Hopkins Coronavirus Resource Center. Retrieved October 20, 2022, from https://coronavirus.jhu.edu/us-map

[17] USDA ERS - County Typology Codes. https://www.ers.usda.gov/data-products/county-typology-codes/. Accessed November 4, 2022.

[18] USDA ERS - County-Level Data Sets. https://www.ers.usda.gov/data-products/county-level-data-sets/. Accessed November 6, 2022.

Table 2. Survey-Independent Datasets			
Dataset	Source	Information	Geography
North Carolina Voting Registration Data[19]	North Carolina State Board of Elections	Publicly available information on registered voters	Individual address level
North Carolina Statewide Parcels[20]	North Carolina OneMap	Geographic parcel polygons with ownership/value for all public and private land	Individual parcels
SOI Tax Stats Migration Data[21]	Internal Revenue Service	Annual population inflows and outflows of taxpayers	County
Change-of-Address Data[22]	U.S. Postal Service	Total number of address change requests by month.	City, zip code
Administrative Data on Address Vacancies[23]	U.S. Department of Housing and Urban Development (via U.S. Postal Service)	Total addresses and vacant addresses by month, length of vacancy, and type of address	Tract

Table 3: Measurement error across NC counties			
County Name	Error %	County Name	Error %
Alamance County	0.04	Johnston County	-0.11
Alexander County	-2.66	Jones County	-0.84
Alleghany County	-2.73	Lee County	1.49
Anson County	-8.47	Lenoir County	-1.07
Ashe County	-2.17	Lincoln County	-1.46
Avery County	1.34	McDowell County	-2.63
Beaufort County	-5.14	Macon County	2.83

[19] Voter Registration Data | NCSBE. https://www.ncsbe.gov/results-data/voter-registration-data. Accessed November 22, 2022.

[20] Parcels. https://www.nconemap.gov/pages/parcels. Accessed November 22, 2022.

[21] SOI Tax Stats - Migration Data | Internal Revenue Service. https://www.irs.gov/statistics/soi-tax-stats-migration-data. Accessed November 22, 2022.

[22] Change-Of-Address Data. https://about.usps.com/strategic-planning/cs09/CSPO_09_027.htm. Accessed November 22, 2022.

[23] HUD Aggregated USPS Administrative Data On Address Vacancies | HUD USER. https://www.huduser.gov/portal/datasets/usps.html. Accessed November 22, 2022.

Bertie County	-4.16	Madison County	-2.52
Bladen County	-10.04	Martin County	-0.66
Brunswick County	-8.28	Mecklenburg County	-1.19
Buncombe County	2.27	Mitchell County	0.15
Burke County	-3.15	Montgomery County	-5.46
Cabarrus County	1.95	Moore County	-3.51
Caldwell County	-1.76	Nash County	0.12
Camden County	-5.73	New Hanover County	-4.61
Carteret County	-2.69	Northampton County	-8.47
Caswell County	1.31	Onslow County	0.31
Catawba County	0.19	Orange County	-0.26
Chatham County	0.71	Pamlico County	-3.45
Cherokee County	-1.03	Pasquotank County	0.49
Chowan County	-0.77	Pender County	-6.91
Clay County	-3.62	Perquimans County	-4.84
Cleveland County	0.49	Person County	-2.07
Columbus County	-7.54	Pitt County	-6.93
Craven County	-0.51	Polk County	-8.09
Cumberland County	-0.49	Randolph County	-0.27
Currituck County	-3.28	Richmond County	-3.13
Dare County	-1.68	Robeson County	-10.36
Davidson County	-0.18	Rockingham County	-0.21
Davie County	-1.33	Rowan County	3.07
Duplin County	-17.14	Rutherford County	-3.92
Durham County	-0.76	Sampson County	-6.86
Edgecombe County	-3.8	Scotland County	-1.34
Forsyth County	-0.33	Stanly County	-1.16
Franklin County	-4.57	Stokes County	-2.67
Gaston County	0.61	Surry County	-0.45

Gates County	-8.6	Swain County	-0.44
Graham County	-5.24	Transylvania County	-4.38
Granville County	0.84	Tyrrell County	-14.02
Greene County	-2.28	Union County	-2.57
Guilford County	0.14	Vance County	-4.79
Halifax County	-1.73	Wake County	-0.25
Harnett County	-2.55	Warren County	-4.51
Haywood County	-1.4	Washington County	-4.2
Henderson County	-1.83	Watauga County	-4.17
Hertford County	-6.73	Wayne County	-5.35
Hoke County	-6.71	Wilkes County	-3.05
Hyde County	-5.24	Wilson County	-3.9
Iredell County	0.5	Yadkin County	-1.09
Jackson County	-2.1	Yancey County	2.05
Red text marks counties with a greater than ±5% error			

Table 4: Results for grouped mean differences across county level traits

Demographic Group	**Mean (%)**		**P-Value**
	Undercounted	*Non-undercounted*	
Black	0.28	0.18	**0.02**
Hispanic	0.08	0.07	0.3
Age: Over 65	0.20	0.20	0.62
Age: Over 85	0.02	0.02	0.6
Age: Under 20	0.23	0.23	0.82
Ages 0-5	0.05	0.05	0.72
Covid Deaths	0.01	0.01	0.41
Births	0.01	0.01	0.73
Deaths	0.01	0.01	0.81
Natural Increase	0.00	0.00	0.72

Table 4: Results for grouped mean differences across county level traits

	Mean (%)		
Demographic Group	*Undercounted*	*Non-undercounted*	**P-Value**
Black	0.28	0.18	**0.02**
Hispanic	0.08	0.07	0.3
Migration (International)	0.00	0.00	0.69
Migration (Domestic)	0.00	0.01	0.42
High House Vacancy	0.26	0.22	0.39

Table 5: Results for grouped mean differences across county level housing counts

	Mean (%)		
Demographic Group	*Undercounted*	*Non-undercounted*	**P-Value**
Black	0.17	0.26	**0.01**
Hispanic	0.07	0.08	0.44
Growth Rate	0.10	0.10	**0.00**
Over 65	0.21	0.19	**0.02**
Over 85	0.02	0.02	0.43
Under 20	0.17	0.17	0.84
Ages 0-5	0.05	0.05	0.14
Covid Deaths	0.01	0.01	0.42
Births	0.01	0.01	0.08
Deaths	0.01	0.01	**0.00**
Natural Increase	0.00	0.00	**0.00**
Migration (International)	0.00	0.00	**0.05**
Migration (Domestic)	0.01	0.00	**0.01**
High House Vacancy	0.20	0.15	**0.02**

Table 6: Results for grouped mean differences

	% Error among population	% Error among non population	P-Value
College Dorm	-0.02	-0.03	**0.02**
Military Barracks	-0.03	-0.02	0.43

Table 7: Results for grouped mean differences by housing counts

	% Error among population	% Error among non population	P-Value
College Dorm	-2.11	-5.54	0.00
Military Barracks	-1.97	-4.36	0.13

Table 8: Average Error for Counties by High Prevalence of Demographic Features

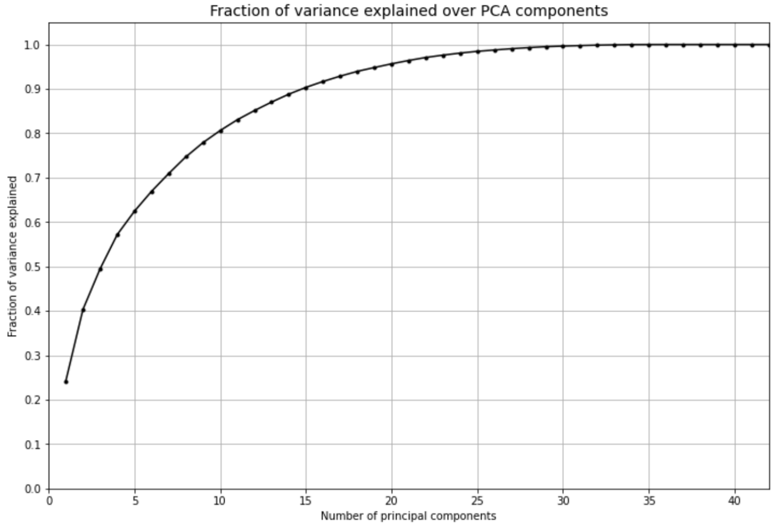
Demographic Group	Count	% Error
Black Population (% of County Population)		
> 1 std above mean	16	-4.26
All others	84	-2.47
Hispanic Population (% of County Population)		
> 1 std above mean	18	-2.90
All others	82	-2.73
American Indian/Alaska Native (% of County Population)		
> 1 std above mean	6	-4.36
All others	94	-2.65
Population over 65 (% of County Population)		
> 1 std above mean	18	-3.11
All others	82	-2.68
Population under 20 (% of County Population)		
> 1 std above mean	14	-3.89

All others	86	-2.57
Population Incarcerated (% of County Population)		
> 1 std above mean	11	-3.25
All others	89	-2.70
Population in College Housing (% of County Population)		
> 1 std above mean	9	-1.67
All others	91	-2.87
Population in Military Housing (% of County Population)		
> 1 std above mean	3	-0.23
All others	97	-2.84
Vacant Housing (% of Total Housing)		
> 1 std above mean	19	-3.18
All others	81	-2.66

Table 9: Identifying Average Coverage Error by Counties, By County Typology Groups

County Typology Group	N (counties)	% Error
Farming or Manufacturing		
In Group	22.00	-3.56
Outside of Group	78.00	-2.53
Recreation		
In Group	22.00	-2.76
Outside of Group	78.00	-2.76
Retirement and Metropolitan		
In Group	19.00	-2.85
Outside of Group	81.00	-2.74
Retirement and Non-Metropolitan		
In Group	13.00	-3.84
Outside of Group	87.00	-2.59

Fig 3: PCA Results


Fraction of variance explained over PCA components
Fraction of variance explained
Number of principal components