HIGH SCHOOL EDITION
Journal of Student Research

Exposing Undercounts in the Census through Regression Modeling

Tarun Shah

Lane Technical High School

ABSTRACT

Undercounts in the Census are notoriously difficult to measure but are necessary to understanding the extent of systemic divides within America. Although many community leaders have proposed that language barriers pose significant obstacles to Census outreach, this paper explores the viability of using predictive models to quantify the extent the role language plays. By using a multivariable regression model trained on student data from the Civil Rights Data Collection, we concluded that Limited English Proficient (LEP) Asian populations were undercounted by over 98,000 people, LEP Native Hawaiian/Pacific Islander populations were undercounted by over 73,000 people, and LEP White populations were undercounted by over 164,000 people. However, biases in the dataset made the results for other ethnic groups unreliable, indicating that regression modeling should be used as a tool for identifying areas of improvement in the Census rather than producing an exact estimate of the population. Our data suggests a strong correlation between language proficiency and Census undercounts, and the Census Bureau could ideally use the results to create targeted language outreach programs for specific states.

Introduction

Undercounts in the United States Census have existed for decades, and have disproportionately affected minority populations, leading to a lack of representation and funding for disadvantaged communities. Following the 2020 Census, the Census Bureau conducted a Post Enumeration Survey (PES) that attempted to determine the accuracy of the recent Census; the results revealed large undercounts for Hispanic, Native American, and Black communities (Census Bureau 2022). Census data is used to allocate funding for programs such as Medicaid, highway and public utility construction, the National School Lunch program, and even redistricting in states, meaning systemic undercounts for minority populations have a very real impact on perpetuating existing inequalities by preventing resources from being distributed to those who need them.

Language barriers are considered one the leading causes of undercounts in the Census, especially among Asian Americans, but few studies have attempted to quantify the magnitude of role language plays in the accuracy of Census data. Given the United States diverse population, it would be extremely difficult to find a source distinct from the Census that measures all the different languages of the American population. However, the Census Bureau does measure individuals classified as Limited English Proficiency (LEP), which describes those who have difficulty communicating, reading, or writing English. In order to assist those who aren't proficient in English, the Federal Government requires certain institutions, such as schools and healthcare providers, to also independently measure LEP populations in order to provide adequate language assistance. Therefore, we proposed that independent datasets that measure LEP populations could be used to accurately measure the magnitude of undercounts for different ethnic groups. We used data from the Civil Rights Data Collection, which measures LEP counts from schools over multiple years, to create a linear regression model that used growth trends to predict the results of data from the Census Bureau. This paper explores how supervised linear regression can be used to predict the populations of LEP ethnic groups, evaluates the reliability and legitimacy of those predictions, and compares those results to Census data to determine the extent of undercounts and overcounts.

Background

Undercounts within the Census have already been well documented, whether it be from the Census Bureau itself in the form of the Post Enumeration Survey (PES), or through independent studies, such as the work of Jennifer Lee, Karthick Ramakrishnan, and Janelle Wong, who documented the specific effects of undercounts on the Asian American population. Undercounts for Black, Hispanic, Indigenous populations have been reported across the board, whereas undercounts for Asians vary heavily depending on socio-economic class and language. However, overcounts are possible as well; the Census Bureau reported that White populations were actually overcounted in the PES, and as will be shown later in the paper, the regression model also predicted overcounts. Overcounts usually occur due to people having multiple or conflicting places of residence, such as students in college who received a form in both their dorm and where their family lives.

Most studies that attempt to calculate errors in the Census do so in the form of more surveys in a much smaller representative sample. Researchers will record data in certain areas and, through thorough analysis, will scale it to a national estimate. This approach is often labor-intensive, expensive, and time-consuming; there's a reason that the results of the PES were published over two years after the Census came out. Additionally, the inherent flaw in using survey data is that it is the same method used in the original Census; therefore, it's entirely possible that the limiting factors of the Census would also apply directly to the new collected data. Alternatively, sources like schools and hospitals already collect population data on a national scale, that, despite potential biases, should emulate the national population to a decent extent. This enables predictive modeling to estimate Census data, which, while not as direct as survey sampling, provides a unique and distinct approach that doesn't require years of research and a large budget, which enables more rapid and diversified tests of the Census. This method also allows us to explore underlying causes of undercounts, as even though explanations have been proposed, modeling the data actually provides data to corroborate the claim that language barriers play a significant role in the accuracy of the Census.

Dataset

In order to measure Census undercounts of American adults with Limited English Proficiency in different ethnic groups, we required a (a) reliable, (b)nationally-scalable, and (c) census-independent population dataset. While there are many datasets that estimate population and LEP status, like insurance databases and Facebook friendship networks, school records were the most resistant to the sampling pitfalls of the census. Although language assistance is provided by the Census in the form of online surveys, households may not be properly informed that such assistance exists, how to access that assistance, or even of their obligation to complete the Census at all. Additionally, there has no enforced repercussion to not filling out the Census since the 1970s, nor is there any way to confirm the accuracy of the surveys, which, given the government's historical abuse of private records to oppress minority populations (Núñez 2018), could reasonably incentivize many households to provide inaccurate information about their ethnicity, language, or immigration status. Data from K-12 educational institutions, on the other hand, is largely resistant to these flaws. Unlike participating in the Census, school attendance is strictly enforced, and school administrators are legally obligated to determine students' LEP status and provide language assistance. Due to the level of rigor and course load that schools require, most LEP Students would struggle to hide a language barrier, especially those in higher grades. Given all the aforementioned factors, school data seemed to be the best for identifying undercounts attributable to language barriers.

Our dataset was sourced from The Civil Rights Data Collection (CRDC), which publishes a national aggregate of every school's LEP reports. The CRDC publishes their metrics annually, which is critical for creating a predictive model, as it is impossible to identify a trend with a single year of data. The datasets for each year contained raw counts of all the LEP Students, broken into six distinct ethnic groups: Hispanic, Native American/Alaskan Native, Asian, Black, White, and Native Hawaiian/Pacific Islander. In order to compare the data directly against Census data, we made a few modifications. First, we used information only from high schools, because schools often could not publish identifying information about younger students. Additionally, we summed all the districts by state, so the aggregate broke down by state, mirroring the census data.

Table 1. Public availability of CRDC and Census Bureau Data by year.

	CRDC LEP Data	**Census Bureau LEP Data**
2011	Publicly Released	Not Publicly Released
2012	Publicly Released	Publicly Released
2013	Publicly Released	Not Publicly Released
2014	Publicly Released	Publicly Released
2015	Publicly Released	Publicly Released
2016	Publicly Released	Not Publicly Released
2017	Publicly Released	Not Publicly Released
2018	Publicly Released	Not Publicly Released
2019	Not Publicly Released	Not Publicly Released
2020	Not Publicly Released	Publicly Released

The primary challenge in collecting data was aligning CRDC data with Census data of the corresponding year. The latest datasets made publicly available by the CRDC only dated up to 2018, meaning that attempting to accurately predict results to compare with the 2020 Census would be impossible. However, the Census Bureau does release an annual, independent measurement of national LEP rates. Given that this data collection effort uses similar methods as the decennial Census, it still provides the opportunity to find flaws in the Census and test our hypothesis. Even so, those datasets that were made publicly available pre-dated even the CRDC data, with the most recent being from 2015. Therefore, we selected three years for which both CRDC and Census data existed: 2012 and 2014, and 2015. We used the data from 2012 and 2014 to predict the data from 2015. In total, our dataset contained two sets of raw counts of LEP ethnic populations broken down by state for three years, with one set sourced from the CRDC, the other sourced from the Census Bureau. The graphs below depict the LEP population counts from each source for 2015.

Commented [1]: again, what's the hypothesis!

HIGH SCHOOL EDITION
Journal of Student Research

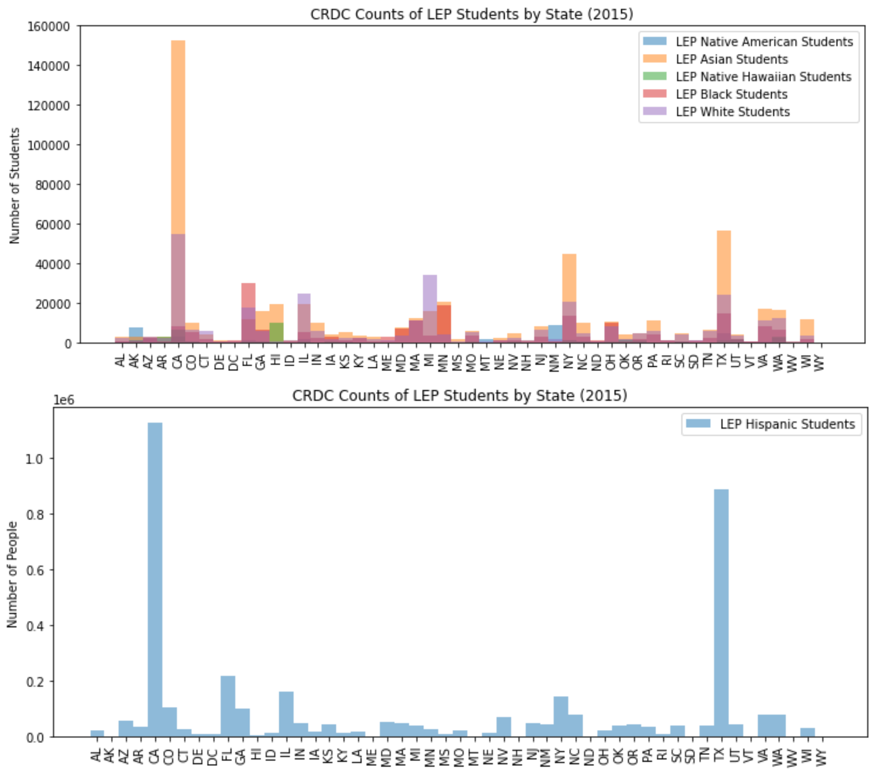
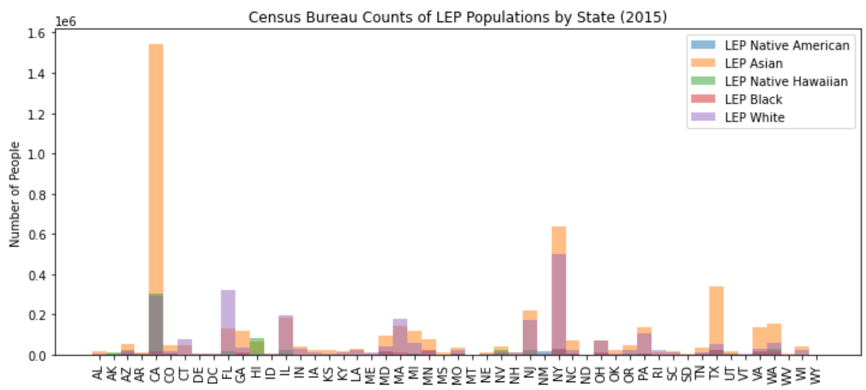


Figure 1. CRDC breakdowns of Students by State in 2015.



HIGH SCHOOL EDITION
Journal of Student Research

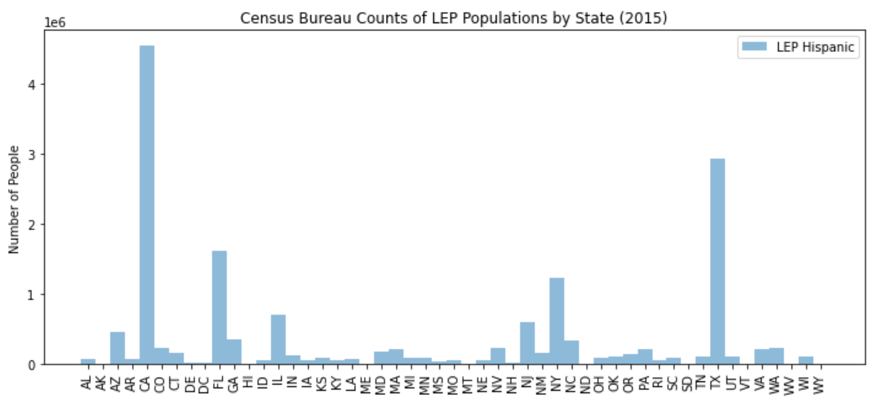


Figure 2. Census Bureau breakdowns of Students by State in 2015.

Methodology/Models

In order to analyze the trends from each dataset to estimate Census undercounts, we used a regression algorithm to predict the populations of LEP ethnic groups. Regression algorithms analyze the correlations between independent and dependent variables to predict the value of dependent variables given a set of inputs. Since linear regression is a supervised machine learning model, it is provided with labeled training data to create a line of best-fit to describe the relationship between variables, with the goal of predicting on new data. In our case, there were five dependent variables, or desired outputs: the LEP counts for Hispanic, Black, White, Native American/Alaskan, Native Hawaiian/Pacific Islander, and Asian populations for the year 2015. The independent variables, or input features, were LEP counts of the same ethnic groups for the years 2012 and 2014. Our dataset contained the counts for each state, each serving as a point of data to model the overall trend. Although it was possible to use smaller regions of data (as the CRDC initially broke down by district), which would have provided more sample data, we decided to train the model on state aggregates so the model would focus on the more macro-level trends for LEP populations. There would have been too much variation on the district level, which could have possibly hurt the model's ability to generalize, but, more importantly, the Census data only broke down to the state level, and since our goal was to test the accuracy of the Census data, using the district data would have introduced nuances not necessary to testing our hypothesis. It is important to emphasize that although each training sample consisted of population counts for an individual state, the actual identification of which state was not given to the model to prevent overfitting. Therefore, once the model was trained on data from all fifty states, it could predict the ethnic breakdown of the LEP population in 2015.

Having gathered raw counts of LEP populations from the Civil Rights Data Collection to compare against the Census Bureau, we created two distinct predictive models to test our hypothesis. The first was a regression model trained on the CRDC data, which would then predict the overall LEP population estimates for 2015 using the Census input data. The second was also a regression model, but instead trained on Census Data, which would be used to predict the student LEP counts for 2015 using the CRDC input data. The underlying assumption behind comparing the two datasets was that the trends in student data should mirror the trends for the overall population; therefore, if the 2015 Census data contained undercounts, the model trained on CRDC data should predict values nominally larger than those in the official Census. Even accounting for disparities that could exist between LEP concentrations of students and adults, it is unlikely that using the CRDC would lead to overestimations, because although it is not uncommon to have LEP adults with English proficient children, there are considerably fewer LEP children with English-proficient parents. The rationale behind using two models was to account for possible biases within the model and to cross check the results. If only the CRDC model was used to test the hypothesis, it would also be entirely possible for the model to be poorly trained and not capture the trends of the CRDC data accurately; consequently, the model could predict undercounts in the Census data simply due to poor analysis. If the Census truly had undercounts in their data, then when given the CRDC input data, a regression model trained on the Census data should theoretically predict nominally lower values for 2015 than the CRDC. Therefore, any undercounts predicted by the CRDC model should also ideally be reciprocated by the Census Bureau data.

Due to small number of samples in the dataset, instead of doing a standard train-test split, which would substantially limit the breadth of the model, or testing on the training set, which could have inflated results due to the model memorizing the data, we used a K-linear testing model, which iterated through each of the fifty states, omitting each from the training data. It would then create a temporary regression model trained on the remaining forty-nine states and test it on the omitted sample. Once every state had been tested, it averaged the coefficient of determination, or R^2, value for each model, producing a more accurate assessment of the model's performance on new data. Initially, for the regression model trained on the CRDC data, the K-linear produced a model with a R^2 of 77%, which, while not terrible, isn't accurate enough to use to claim undercounts on government data. On closer examination, the model's overall R^2 was being weighed down by its performance in Arizona, which had produced a R^2 of -7.9. The model performed so poorly on Arizona's data because Arizona's LEP student counts were completely out of ordinary: the CRDC reported a 20,000 student decrease in LEP Latino Students from 2014 to 2015, which the model intuitively didn't predict. Once Arizona had been removed from the K-Linear algorithm, the average R^2 shot up to 97% on the testing data, signaling the regression model had found a strong relationship in the data. However, this anomaly introduced concerns about the reliability of the CRDC data. We initially proposed that there was a decrease in the number of schools that the CRDC surveyed in Arizona, but the dataset revealed that the CRDC actually surveyed around thirty more schools from 2014 to 2015. Our next prediction was that the literacy rate improved among Hispanic youth in Arizona, and although studies show that English proficiency has improved substantially over the past decade (Pew Research Center 2016), the literacy rate was increasing nowhere near fast enough to account for a 25% decrease in LEP students, especially only in a span of year. The standards of classification for LEP students also did not change in 2015, leaving no plausible explanation for such a massive drop. Since the remainder of the dataset didn't seem to contain any more severe, obvious errors, we proceeded our research without the Arizona data; however, once we achieved our results, we made sure to account for any and all biases, which will be discussed later in the paper. The regression models trained on the Census model had no such blemishes in the data, and returned an average R^2 of 99%, which was sufficient for testing. It should be noted that in our case study, the R^2 of each model does not speak to reliability of the data, nor does it prove whether or not those predicted numbers are accurate; a higher R^2 value merely proves that the model was able to create a more well-defined relationship between the variables and can make estimations that closely reflect the trends in each dataset.

Once the models were created, the CRDC model was fed the inputs of the Census data, while the Census model was fed the inputs of CRDC data. Since we were assuming that the CRDC data was the more accurate of the two datasets, and therefore the CRDC model produced a more accurate estimation of LEP ethnic concentrations, we calculated the percentage of each ethnic population the Census data had either over or underestimated compared to the predictions. We defined an error threshold of 10%, meaning that Census reports that had a calculated percent error of under 10% could be attributed to standard model biases rather than a systemic undercount. Finally, we cross-checked those error estimates with predictions from the Census regression model, which should have estimated lower counts of the LEP student population than the CRDC for every state that the CRDC model estimated higher counts than the Census Bureau.

Results and Discussion

The model's predictions contained mixed results; although it may have exposed certain legitimate errors within the Census, the estimations for certain LEP groups have significantly more evidentiary support than others, and, in the worst cases, the model was overwhelmed by the biases of the CRDC data. The following graph depicts the CRDC Model's predictions of the national LEP population estimates for 2015 compared to the reported Census data. In order to measure inaccuracies in the Census data using the predictions as a baseline, we subtracted the predicted values from the reported values, meaning a negative value in table below indicates that the Census reported less people than the model predicted, whereas a positive value indicates the Census actually reported more people than the model.

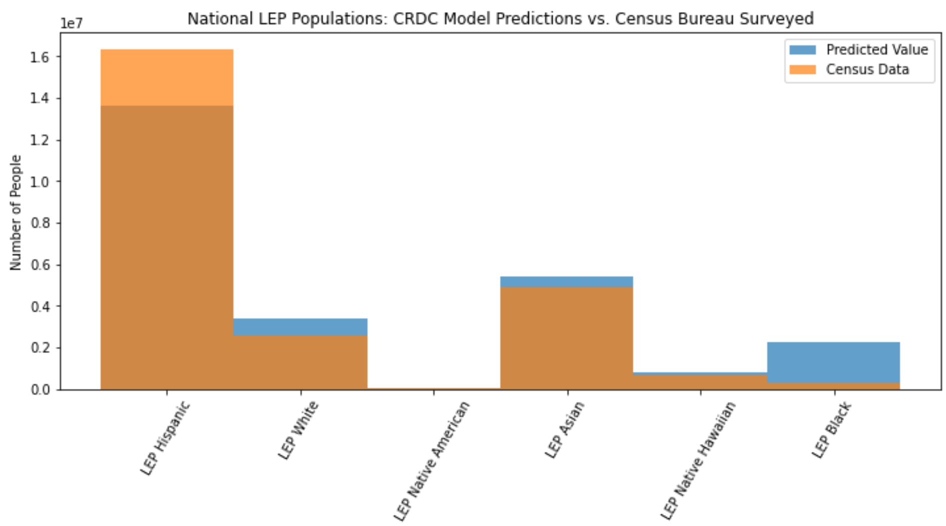


Figure 3. Estimations of national LEP populations vs. surveyed LEP populations by the Census Bureau

Table 1. Differences between Census Bureau Limited English Proficient population counts and predicted populations counts from the regression model.

	Predicted Differences from CRDC Model	Predicted Differences from CRDC Model with Error Threshold	Predicted Differences from CRDC Model cross-referenced with Census Model
LEP Hispanic	+ 2,522,077	+ 2,458,742	+ 1,396,888
LEP White	- 828,877	- 828,877	- 164,220
LEP Asian	- 451,878	- 235,195	- 98,879

LEP Black	- 1,970,568	- 1,970,467	- 1,286,429
LEP Native American	+ 2,954	+ 633	- 300
LEP Native Hawaiian/Pacific Islander	- 119,244	- 117,071	- 73,568

Regardless of the ethnic group, the estimated errors from the CRDC model alone were of remarkably high magnitude, the likes of which seem improbable, especially for LEP Black populations. However, after introducing the error threshold and cross-checking those estimates with the results from the Census model, many of the predicted errors were filtered with more reasonable estimates. Given that the goal of the model is to test the accuracy of the Census data, cross-checking the results with a regression model trained on the Census data ensures that the error predictions are actually due to faulty Census estimates, and not due to CRDC biases. For LEP White, Asian, and Native Hawaiian/Pacific Islander groups, the model predicted sizable, but reasonable estimates for undercounts of each group that are mostly corroborated by other studies done in the field. Although the Census Bureau reported that they actually overcounted Asian Americans in the 2020 Census, experts say that due to the large wealth disparity among Asian Americans, the people who were overcounted were those of a significantly more privileged economic background, while those with significant language barriers were most affected by undercounts (NBC 2022). Therefore, the predicted 100,000 to 250,000 undercounts of LEP Asians (depending on which error threshold is used) is most likely accurate, especially if it is masked by overcounts of more economically privileged Asians, which could possibly affect the resources and outreach to the Asian American community. For LEP Native Hawaiian/Pacific Islander populations, a Post Enumeration Survey (PES) following the 2020 Census showed that Hawaii had a self-response rate 10% below the national average, and leaders in the Hawaiian community specifically cited language barriers and a "profound distrust of the government" as the root cause (Abdul-Hakim and Parks 2020), indicating that the estimated 73,000 to 117,000 undercounts also seems fairly accurate. For LEP White populations, the reliability of the model isn't as clear cut; although many sources indicate that there are overcounts of White populations in America, the economic disparities between white residents across the nation add nuance to the overarching assumption. English proficiency rates are substantially lower in rural areas than in urban centers (U.S. Department of Education 2007), while rural communities are also four times more likely to be classified as Hard to Count (HTC) areas by the Census Bureau (O' Hare 2019). Given that the ethnic makeup of rural America is over 75% Caucasian (Johnson and Lichter 2022), the obstacles of surveying rural areas combined with language barriers make LEP White populations particularly vulnerable to undercounts in the Census. Additionally, the population of white people in America is of high enough magnitude that an undercount of a few hundred thousand LEP white households could go unnoticed, especially if it is being masked by the overcounts in urban areas. Therefore, the model's predicted undercount of 164,000 LEP White people is entirely plausible and could serve as a unique example of the CRDC exposing an undercount not previously recognized by other sources.

For LEP Native American populations, even without the error-thresholds and cross-checking, the model predicted relatively inconsequential differences between the predictions and the reported Census values, with the percent difference never going above 0.5%. Such results are contrary to most reputable research, with the PES ranking Native Americans among the most underrepresented groups in the Census. However, the obstacle that prevents the Census data from being accurate is its limited access to Native American reservations, which applies equally to the CRDC, as they do not have access to school data on reservations. The model's performance suggests that the CRDC data is not an improvement on the Census data, nor is it an accurate representation of Native American populations.

The CRDC model's predictions for LEP Black and LEP Hispanic populations were not only of the largest magnitude, but also the least reliable. Even after cross-checking with the Census model, the model predicted an undercount of 1.3 million LEP Black people in the Census, which is especially striking considering the Census only reported around 285,000 LEP Black individuals. When examining the breakdowns by state, the model predicted that New York alone undercounted almost 400,000 LEP Black people.

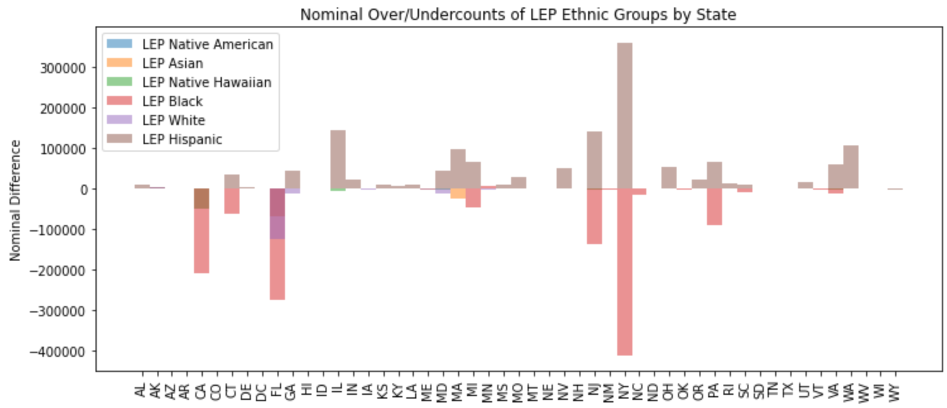


Figure 4. Estimated over/undercounts of LEP populations by the Census Bureau broken down by state.

The PES estimated that Black Americans were underestimated at a national average of 3.3% (Census Bureau 2022), meaning that even if a generous margin of error is given to those estimates, the likelihood of the Census undercounting over 80% of the LEP Black population is extremely unlikely, let alone missing an entire county worth of people in New York. The most likely explanation for the high estimations is the systemic overclassification of black students into LEP programs. Black students have historically been misclassified for Special Education programs and are 2.3 times more likely than white students to be identified as having an intellectual disability (Dekker & et al., 2002). This often fails to give countless black students education suited to their needs, which stunts their academic achievement and limits the resources available to them. Although thorough research hasn't been conducted for LEP programs, the systemic racism that causes overclassification of learning disabilities likely also causes overclassification of black students into LEP programs. Therefore, the CRDC data most likely contained disproportionately high LEP rates for Black students that weren't representative of the Black population, causing abnormally high and misleading predictions. The undercounts that do exist in the Census for LEP Black people cannot be accurately determined by modeling school data.

The estimations for the LEP Hispanic population faced the opposite issue, with the model predicting a Census overcount ranging from 1.3 to 2.5 million people. In the PES, the Census Bureau estimated that Hispanic peoples were the most undercounted out of any ethnic group (at an astonishing 4.99%) meaning the predicted overcount is exceedingly improbable, especially one of such a large magnitude. The low predictions of LEP Hispanic populations are likely due to the difference in fluency of English between first- and second-generation Americans. A significant portion of LEP adults in the United States are immigrants who were underrepresented in the CRDC Data. A study done by the Pew Research Center found that second-generation immigrants were comfortable with English at a far greater rate than their parents, especially in Hispanic populations (2013). 93% of second-generation Hispanic immigrants could speak English proficiently compared to 48% percent for first-generation Hispanic immigrants. For Asian populations, the gap was significantly smaller, with 92% for the second-generation versus 77% for the first-generation, and although the report didn't examine other populations, the disproportionate lack of resources and lackluster education system provided for Native American students would strongly suggest that the gap is even smaller for Native Americans and Pacific Islanders (Institute of Education Sciences 2022). Consequently, the CRDC data underestimated rates of LEP Hispanic adults because of the higher English proficiency rates of among the newer generations of students. Even though the disparities didn't affect the remaining ethnic populations to the same extent, they most probably led to slight underestimations in the predictions of the CRDC model.

After interpreting the predictions of the model and accounting for biases, the undercount and overcount estimations still provide valuable data for testing the Census and identifying key areas for improvement. Even if regression cannot be used to determine exact population counts for each of these groups, the model still outlines which ethnic groups require more outreach from the Census Bureau and in which states. Additionally, it suggests that language assistance would be an effective avenue for their outreach. The graph below depicts the percent of each ethnic group that was undercounted or overcounted, and even after accounting for biases, it is clear that a state like Hawaii, for example, needs to target its language assistance towards Native Americans.

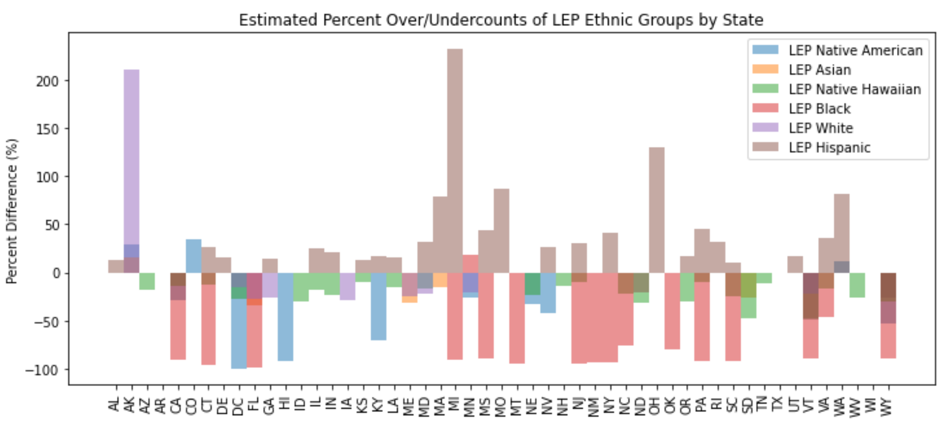


Figure 4. Estimated percent over/undercounts of LEP populations by the Census Bureau broken down by state.

Conclusion

Although the regression model was able to accurately capture the growth trends from LEP population data from schools, the biases of the CRDC dataset made the model unreliable for predicting precise estimations of LEP ethnic populations on a national scale. However, the data still provided strong evidence that language is a key factor in Census undercounts, and, if detailed analysis is used on results of the model to account for biases, the predictions can still be used as a viable method of identifying errors in the Census and creating targeted language outreach programs.

The CRDC regression model produced undercount estimations in the hundreds of thousands for LEP Asian, White, and Native Hawaiian/Pacific Islander populations that were all backed by evidentiary support from similar research. Unfortunately, due to a lack of access of school data from Native American reservations, the systemic overclassification of Black students in LEP programs, and the higher literacy rates of Hispanic children than adults, the CRDC model produced inaccurate estimations for LEP Native American, Black, and Hispanic students. The results indicate that to improve the reliability of the estimations, the regression model should be fitted on data sourced from multiple different institutions. The biases of the population data from schools overwhelmed the model, but including LEP information from health insurance providers or possibly social media platforms could potentially balance out the biases of each dataset, allowing the model to generalize to the national population more accurately.

However, the biases don't compromise the existing undercounts uncovered by the CRDC model; in fact, they suggest that the real undercounts are likely even larger than the estimated undercounts, given the fact that, with the exception of Black students, the model leans more towards overestimating English proficiency rates. These findings indicate that the CRDC model should not be used as an absolute method of calculating the exact population counts, but rather as a tool to fact-check Census reporting. The state breakdown for the model's predictions showed which ethnic groups each state needs to provide better language assistance for. After accounting for the biases in the data, we were still able to glean which undercounts were legitimately prevalent throughout which regions, implying that with thorough analysis, the CRDC model is still a viable method to testing the Census, even if it can't necessarily provide an exact estimate of LEP populations. The goal of this research isn't to replace the Census, but rather to identify areas where the Census needs to improve; the Census Bureau should ideally employ the results of this paper to guide their language outreach programs. These findings also suggest linear regression can also be used to evaluate other areas of the Census; although this paper only examines LEP populations through data from schools, this method can ideally be applied to different metrics of the Census using different sources of data. As it becomes more universally adopted, different studies using a variety of sources and metrics can be combined to create a more comprehensive and accurate evaluation of the Census. In conclusion, regression analysis is a valuable tool for verifying the results of the Census and providing a basis for targeted language outreach in different states.

Acknowledgements

Thank you to Kathryn O'Nell for mentoring me throughout this research project and to the Inspirit AI program for making this paper possible.

References

Abdul-Hakim, Gabriella, and MaryAlice Parks. "Pandemic Shows Need for Native Hawaiians, Pacific Islanders Participation in Census." *ABC News*, ABC News Network, 26 May 2020, https://abcnews.go.com/Politics/pandemic-shows-native-hawaiians-pacific-islanders-participation-census/story?id=70873566.

The American Academy of Political and Social Science, 2018, *Accurately Counting Asian Americans Is a Civil Rights Issue*, https://journals.sagepub.com/doi/abs/10.1177/0002716218765432. Accessed 25 Jan. 2023.

"Census Bureau Releases Estimates of Undercount and Overcount in the 2020 Census." *United States Census Bureau*, Census Bureau, 10 Mar. 2022, https://www.census.gov/newsroom/press-releases/2022/2020-census-estimates-of-undercount-and-overcount.html. Accessed 25 Jan. 2023.

Civil Rights Data Collection. "Entire Database of the Civil Rights Data Collection from 2010-2016." https://ocrdata.ed.gov/resources/downloaddatafile

Dekker, L.M., Krou, C.A., Wright, T.D. & Smith, D.M. (April, 2002). Effective strategies for reducing the overrepresentation of minorities in special education. Paper presented at the Annual Classroom Action Research Conference, South Bend, IN. EDRS Reproduction No. ED464076.

Institute of Education Sciences, 2022,. *English Language Development among American Indian English Learner Students in New Mexico*, https://ies.ed.gov/ncee/rel/regions/southwest/pdf/REL_2022135.pdf.

Johnson, Kenneth, and Daniel Lichter. "Growing Racial Diversity in Rural America: Results from the 2020 Census." *UNH*, University of New Hampshire, 31 May 2022, https://carsey.unh.edu/publication/growing-racial-diversity-in-rural-america#:~:text=Rural%20America%20remains%20predominately%20non,according%20to%20the%202020%20Census.

Núñez, Anna. "Census Data Has Been Misused before - in WWII and after 9/11." America's Voice, 3 Apr. 2018, americasvoice.org/blog/census-data-misuse/.

NBC, “Asian Americans Were Overcounted on the Last Census. Experts Say That Fact Is Misleading.” *NBCNews.com*, NBCUniversal News Group, 29 Apr. 2022, https://www.nbcnews.com/news/asian-america/asian-americans-overcounted-last-census-experts-say-fact-misleading-rcna26509.
O' Hare, William. “2020 Census Faces Challenges in Rural America.” *UNH*, University of New Hampshire, 21 Oct. 2019, https://carsey.unh.edu/publication/2020-census.
Pew Research Center, 2016, *Rise in English Proficiency among U.S. Hispanics Is Driven by the Young*, https://www.pewresearch.org/fact-tank/2016/04/20/rise-in-english-proficiency-among-u-s-hispanics-is-driven-by-the-young/. Accessed 25 Jan. 2023.
Pew Research Center, 2013, *Second Generation Americans*, https://www.pewresearch.org/social-trends/2013/02/07/second-generation-americans/ Accessed 25 Jan. 2023.
U.S. Department of Education, 2007, *Status of Education in Rural America*, https://nces.ed.gov/pubs2007/2007040.pdf. Accessed 25 Jan. 2023.