Journal of Data Science 22 (4), 591–604
October 2024
DOI: 10.6339/23-JDS1116
Data Science in Action

Race-Specific Risk Factors for Homeownership Disparity in the Continental United States

Rachel E. Richardson[1,*], Damon T. Leach[1], Natalie M. Winans[1], David J. Degnan[1], Anastasiya V. Prymolenna[2], and Lisa M. Bramer[1]

[1]*Biological Sciences Division, Pacific Northwest National Laboratory, Richland, WA, 99354, USA*
[2]*Environmental Molecular Sciences Division, Pacific Northwest National Laboratory, Richland, WA, 99354, USA*

Abstract

The United States has a racial homeownership gap due to a legacy of historic inequality and discriminatory policies, but factors that contribute to the racial disparity in homeownership rates between White Americans and people of color have not been fully characterized. In order to alleviate this issue, policymakers need a better understanding of how risk factors affect the homeownership rates of racial and ethnic groups differently. In this study, data from several publicly available surveys, including the American Community Survey and United States Census, were leveraged in combination with statistical learning models to investigate potential factors related to homeownership rates across racial and ethnic categories, with a focus on how risk factors vary by race or ethnicity. Our models indicated that job availability for specific demographics, and specific regions of the United States were factors that affect homeownership rates in Black, Hispanic, and Asian populations in different ways. Based on the results of this study, it is recommended policymakers promote strategies to increase access to jobs for people of color (POC), such as vocational training and programs to reduce implicit bias in hiring practices. These interventions could ultimately increase homeownership rates for POC and be a step toward reducing the racial wealth gap.

Keywords *census; economics; random forest; survey*

1 Introduction

Purchasing a home provides a benefit to both the purchaser, as a catalyst for growing financial assets, and the nation at large, as an increase in overall wealth and a stimulus to the economy (Turner and Luea, 2009). In the United States, there is a gap between White and non-White homeownership rates, with the gap between White and Black homeownership being the largest (McCargo et al., 2019; Choi et al., 2019). The legacy of racial discrimination (slavery, Jim Crow laws, redlining, restrictions to loans, etc.) has perpetuated the homeownership discrepancy between White and Black populations (Ray et al., 2021; McCargo et al., 2019; Choi et al., 2019), triggering long-term effects that continue racial and financial inequality. This inequality can be explained by gaps in educational attainment, inter-generational wealth, credit scores, income, and access to affordable interest rates on mortgages, among other variables (Choi et al., 2019; Hilber and Liu, 2008; McCargo et al., 2019; Gabriel and Rosenthal, 2005; Ray et al.,

*Corresponding author. Email: rachel.richardson@pnnl.gov.

Received December 10, 2022; Accepted October 29, 2023

2021). Even after several years of studies, an estimated 17% of the variance in the White-Black homeownership gap has yet to be explained (Choi et al., 2019).

Though not as wide as the gap between White and Black populations, there are also homeownership gaps between White and Asian populations, and White and Hispanic populations (Choi et al., 2019; Kuebler and Rugh, 2013). Variables such as gender, marital status, presence of children, and occupation have been identified as potential race-specific factors for which the effect on homeownership differs between White and non-White populations, though all risk factors have yet to be explored and fully characterized (Kuebler and Rugh, 2013). Any efforts to alleviate these inequities will benefit from an understanding of how the impact of these factors varies across racial groups.

This work aims to provide further insight into factors that contribute to discrepancies in homeownership rates between White and Black, Hispanic, and Asian populations. Multiple publicly available datasets with both untested, potential risk factors, such as commute times, and more intuitive risk factors, such as educational attainment and job availability, were used with random forest regression models to evaluate and identify potential risk factors for homeownership inequity among racial groups. To control for region-specific effects, multiple models were run with regions of the United States held out. Variables that explained a large portion of the variance across models, regardless of each iteration of region holdout, were compiled from each of these models, and unique risk factors were identified for each racial group to identify potential avenues to close the homeownership gap in the United States.

2 Data

The datasets used in this study are described below. All datasets used, the years represented in each data source (Figure S1), and the complete list of variables used for modeling and their original dataset source (Table S1) are provided in Supplemental Material. All datasets were averaged to the county level and across any years within the 2015 to 2019 range, when applicable. Four racial-ethnic categories for which data was consistently available across data sources were investigated. Hereafter, the following terms are used to refer to racial-ethnic categories: Asian refers to Asian Americans and others of Asian descent, Black refers to African Americans and others of African descent, Hispanic refers to those who identify as Hispanic or Latino in survey responses, and White refers to European Americans or those who do not identify with other racial or ethnic categories.

Census population estimates were included in the model to accurately describe the racial population demographics in each county. Data were downloaded from the county-level report "Annual County Resident Population Estimates by Age, Sex, Race, and Hispanic Origin" for years 2015 through 2019. Population estimates were restricted to individuals with plausible homeownership eligibility being of age 20 and above (US Census Bureau, 2019a). Average population values were calculated by county and by race within a county. A variable was added to identify rural and urban counties, where any counties with more than 1000 people per square kilometer were labeled "urban" and all others labeled "rural".

Census population change estimates were included in the model as a metric of residency changes within each county. Information on the population flux of each county was extracted from the United States Census Bureau's "Annual Resident Population Estimates" report. Variables included group quarters population estimates, and the rates of birth, death, and international and domestic net migration, all at the county level (US Census Bureau, 2019b). Data was

averaged across the years from 2015 to 2019. Group quarters were defined as shared residential spaces, such as correctional facilities, nursing homes, college dormitories, or shelters.

Per diem rates across each county were included to approximate the cost of living across the United States. Department of Defense per diem rates for lodging were available at a city or county level, depending on the area, for the years 2015 through 2019 (Department of Defense, 2014). The county-level annual maximum lodging per diem rates were averaged over that timeline.

Educational data were included as a metric of the socio-economic potential of communities in each county. The average proportion of adults 25 and older that finished high school was available at the county level between 2015 and 2016 (US Census Bureau, 2020b).

The "Homeowner's Demographic 5-year" dataset from the American Community Survey released by the US Census Bureau summarizes estimated homeownership rates between 2015 and 2019 within counties (Urban Institute, 2021a). Average values were calculated for predicted foreclosure rates, cost burden of homeowners, and ownership expenses. For each county, the proportion of White and people of color (POC) homeowners with income between 100% to 150% of the area median income (AMI) was calculated as the number of homeowners with income between 100% and 150% divided by the total number of individuals of the respective race category of age 20 or over. Additionally, the racial proportion of homeownership was calculated as the number of homeowners divided by the population of age 20 or greater for each race category. The averages of these variables were then calculated over this timeline.

Data on job availability were included as indicators of financial opportunity in each county. Counts of job availability per age group ($\leqslant$29, 30–54, $\geqslant$55), monthly earnings ($<$\$1250, \$1250–\$3333, $>$\$3333), industry (e.g., construction, educational services), race, ethnicity (Hispanic or non-Hispanic), education (e.g. high school, some college), and sex were extracted from 2015 to 2018 for both federal and non-federal employment (Urban Institute, 2022). The estimation of these values is described in detail elsewhere (Urban Institute, 2022). Briefly, these values were estimated by combining the American Community Survey with state unemployment insurance wage records to understand the potential employment opportunities for underrepresented groups within counties. The federal and non-federal datasets were summed together at the tract level (e.g., city block) before averaging at the county level and dividing the counts by the total number of available jobs. In this survey, the Hispanic population (ethnicity category) was not distinguished from the White population (race category). Therefore, the proportion of available jobs for the Hispanic population was estimated by multiplying the proportion of available jobs for the White population by the proportion of jobs available to the Hispanic population. The proportion of jobs available for the White population was also adjusted using an analogous approach.

Commute data was included as an indicator of community affluence, the ability to afford travel, or the lack of work opportunities near areas with affordable housing. Data was obtained from the Urban Institute's Commute Data (Urban Institute, 2021b). This dataset contained the commute time for people of four metropolitan statistical areas (MSAs) to access job opportunities (Lansing, Michigan; Seattle, Washington; Baltimore, Maryland; and Nashville, Tennessee). Each census block represented by a GEOid, is a subdivision of a census tract that generally contains between 600 and 3,000 people. Along with each GEOid, this dataset included the county and state associated with the GEOid, and the total counts of the people within the census block who take less than 10 minutes, 10 to 29 minutes, 30 to 59 minutes, and greater than 60 minutes to commute to their work. Job access for 1,913 counties were then estimated as described elsewhere (Urban Institute, 2021b). To calculate commute time category proportions, the dataset was supplemented with population density data from April 2020 of metropolitan areas measured in number of people per square kilometer (US Census Bureau, 2020a).

3 Methods

3.1 Data Analysis

All analyses were performed in R, version 4.2.2 (R Core Team, 2020). Data processing was completed using the R packages *dplyr* and *purrr* (Wickham et al., 2022; Henry and Wickham, 2022). Distributions of each variable were summarized and visualized using *trelliscopejs* and *ggplot2* (Hafen and Schloerke, 2021; Wickham, 2016). Spearman correlation (Spearman, 1987) was used to calculate the association between all pairs of predictor variables. The *urbnmapr* (Strochak et al., 2022) R package was used for all geographical plots. Partial dependence plots were generated using the R package *randomForest* (Liaw and Wiener, 2013). All code, plots, and results are available at https://github.com/rarichardson92/Homeownership_disparity.

3.2 Modeling Approach

A metric of homeownership equity was defined as the response variable of interest in subsequent models. We denote this metric as HEI calculated per county as:

$$\text{HEI}_r = \frac{h_r/h}{p_r/p}, \tag{1}$$

where h_r is the number of homeowners of race r in a county, h is the total number of homeowners in a county, p_r is the population of racial group r in a county of age 20 or older, and p is the total population in a county of age 20 or older.

A random forest regression model (Breiman, 2001) was used to accommodate potential non-linear relationships (Biau and Scornet, 2016), and HEI_c was specified as the response variable of interest. A total of 69 variables were included as potential explanatory variables. A list of all predictor variables and the data source are given in Supplementary Material (Table S1). Random forest regression models were fit separately for each racial-ethnic category to allow for a clearer picture of race-specific predictors of homeownership equity. Models were fit using the *randomForest* (Liaw and Wiener, 2013) package using default parameters (i.e. ntree = 500 and mtry = 22, as determined by number of predictors).

The mean squared error (MSE) and pseudo R^2, hereafter denoted as $\widetilde{R^2}$, values, as calculated by the *randomForest* package, were used to evaluate model performance. MSE was calculated as:

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2, \tag{2}$$

where y_i and $\hat{y}_i$ are the observed and predicted HEI values, respectively. Pseudo R^2 was calculated as:

$$\widetilde{R^2} = \frac{1 - \text{MSE}}{s^2}, \tag{3}$$

where s^2 is the sample variance of HEI values.

Variable importance was measured by the increase in MSE as calculated by the *randomForest* package. In this process, values of the predictor variable of interest were permuted and the change in MSE compared to the model using the original predictor variable values was calculated. The default scaling for increase in MSE, where MSE is divided by the standard error, was used. Larger increases in MSE indicated variables with a larger impact on overall model accuracy.

Racial populations in each county varied widely with some counties containing few individuals of a specific racial category. To account for this, our model was evaluated for performance at small racial subpopulation sizes ranging from 200 to 2000. An optimum subpopulation size was chosen based on where the range of diminished MSE stabilizes. Based on this, the scope of the model was restricted to counties with at least 500 individuals.

Model performance was evaluated using 10% of counties from each geographical division, as defined by the US Census Bureau, for testing and 90% for model fitting via cross-validation. Cross-validation was performed by holding out the counties from each geographical division per fold (Figure 1A) as the validation dataset. Counties were classified as belonging to one of the following nine divisions: Pacific, Mountain, West North Central, East North Central, West South Central, East North Central, South Atlantic, Middle Atlantic, or Northeast (US Census Bureau, 2013). The percentage of counties in each holdout set (Division) by racial category model are given in Table S2. A total of 36 random forest regression models were fit each with nine cross-validation folds for the four racial category models.

4 Results

4.1 Exploratory Data Analysis

Several predictor variables contained outliers in urban areas (Figure 1C). For example, yearly incomes were higher in urban areas, likely because cities tend to have more higher paying jobs than rural areas. Other outliers were more difficult to explain. For example, the predicted home foreclosure rates were observed to be very low across counties in the state of Kentucky for unknown reasons (Figure 1D).

In the Asian, Black, and Hispanic populations, there was a pattern of below equitable homeownership representation (Figure 2). Some of the markedly low homeownership rates may be explained by small county sizes. For example, the county with the lowest number of Asian homeowners has a population of 519 (Petroleum County, MT). Notably, several counties with a high percentage of White people but low homeownership were observed. Upon further inspection, most of these counties coincided with larger Hispanic communities. This trend in the data for the White population could result from different ways of categorizing Hispanic individuals in the datasets. In general, there is a lack of standardization of how Hispanic communities were captured in data; some surveys code Hispanic as a race while others code it as an ethnicity. These inconsistencies could introduce confounding factors into the data (Perez and Hirschman, 2009). To assess the effect of these counties on the analysis, a sensitivity analysis was conducted. Counties with HEI < 0.50 and White population proportion > 0.90 were excluded from the White random forest models, and results were compared to equivalent models using all counties. Model performance and variable importance did not change appreciably with the exclusion of these counties (Figure S2).

Many of the included predictor variables were highly correlated with each other. Job availability, income, education, and migration into the county (Figure S3) had high correlation values. Notably but not unexpectedly, variables from the same data source were shown in small blocks of highly correlated variables (Figure S3), such as migration rates (rDomestic and rMigration representing the net domestic migration and net migration rates, respectively).

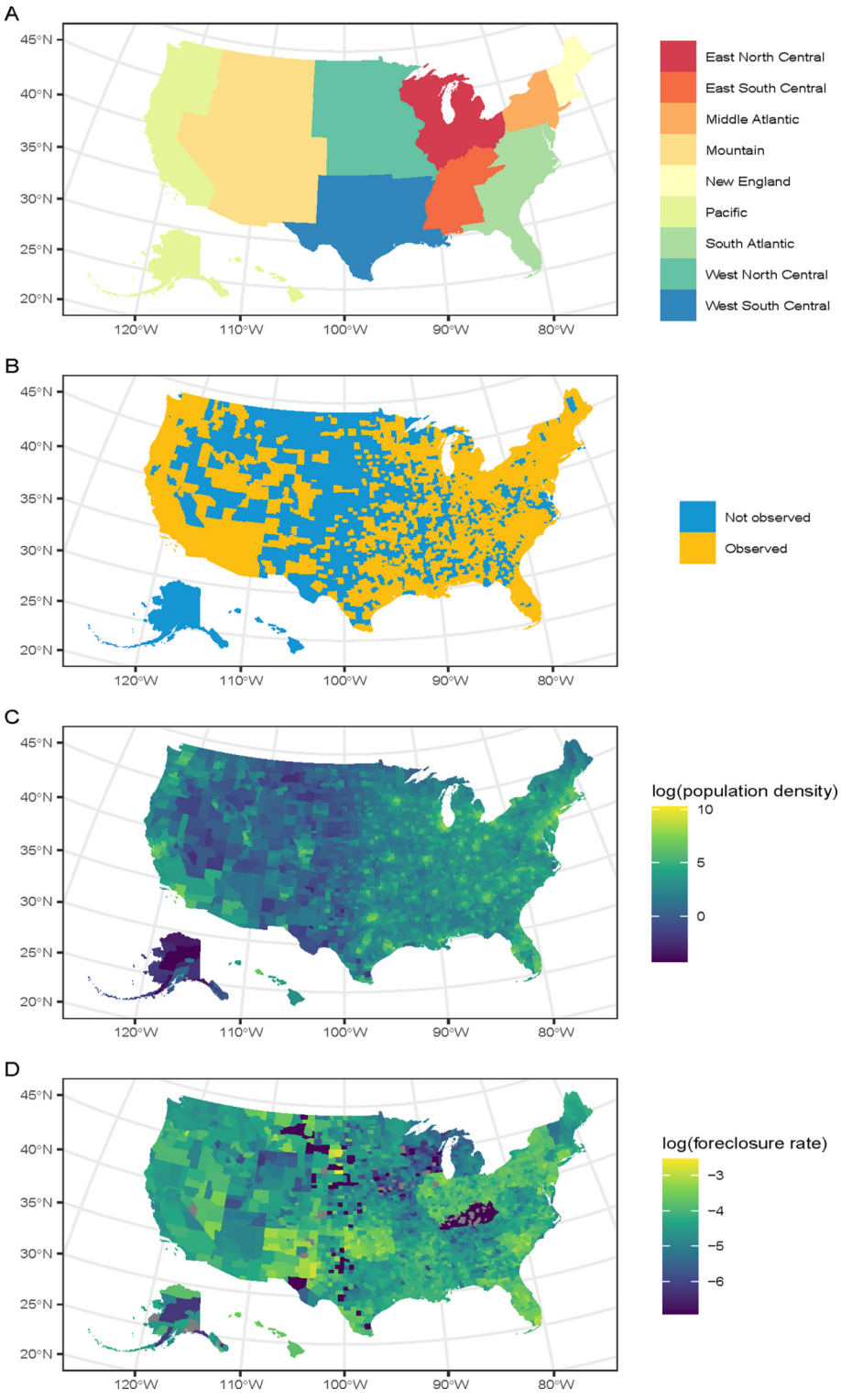

Figure 1: Geographic maps of counties and variables. A) Divisions in the United States used for cross validation (US Census Bureau 2013). B) Counties represented in the final dataset based on consensus observations in all datasets. C) Log of the population density across U.S. counties. D) Log of the predicted home foreclosure rate across U.S. counties.

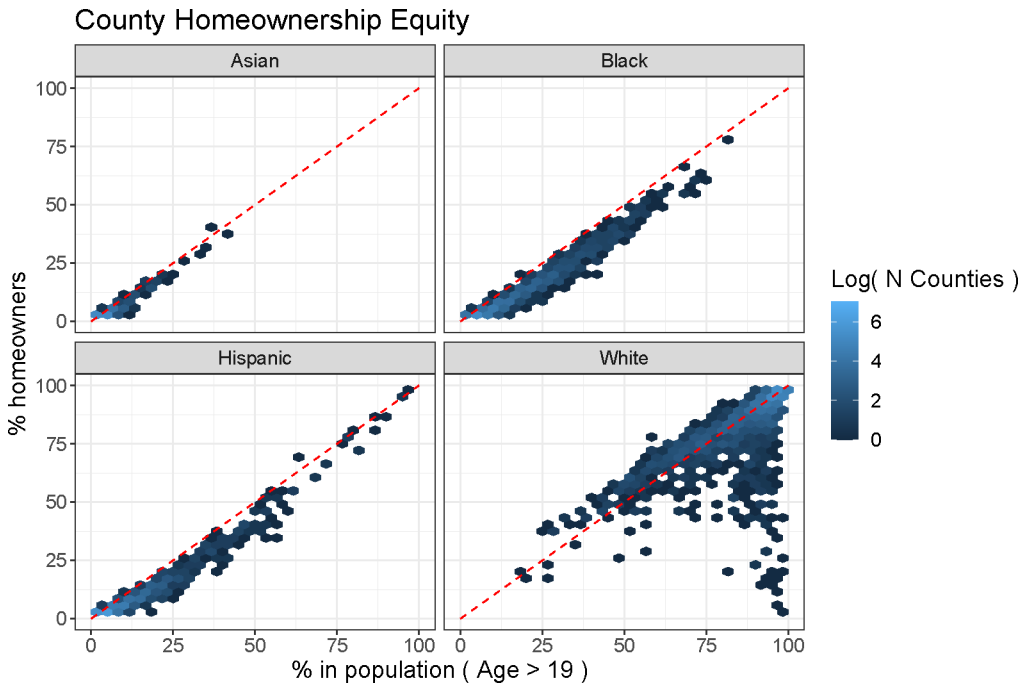

Figure 2: Equity of homeownership across US counties, where trending along the dotted red line indicates homeownership equity. Hexes below the red line depict racial underrepresentation in homeownership rates, while hexes above the line depicts racial overrepresentation in homeownership rates. Population representation is restricted to ages 20 and up to reflect likely homeownership candidates.

4.2 Sample Sizes

Counties with too few residents of any single race were removed from respective race models to avoid skewing the results. The final model required a minimum of 500 members. After removing counties below this threshold, the number of counties used by race and region are given in Table 1.

The percentage of counties per race that were withheld per division cross-validation fold, remained relatively consistent (Table S2). The number of total counties that were represented at least once across any model was 1,802 (Figure 1B).

4.3 Random Forest Model Performance

For each region holdout, poor performance was observed in the Hispanic population models, where the best model had a $\widetilde{R^2}$ value of 0.2016 (Figure 3), which may be due to the racial and ethnic definitions of the Hispanic population changing over time (Perez and Hirschman, 2009). Both the Asian and Black models had moderate maximum performances ($\widetilde{R^2}$) of 0.6772 and 0.7214, respectively. The White models had the highest $\widetilde{R^2}$ of 0.9385. The only poorly performing White model was for the New England region, which may be due to unusually high percentage of White people in New England (US Census Bureau, 2022). All cross-validation models performed similarly within Race for the 10% test data (Figure S4).

Table 1: Number of counties per race and division with ⩾500 individuals of each race.

Division	Asian	Black	Hispanic	White
East North Central	133	200	221	322
East South Central	70	187	138	223
Middle Atlantic	97	121	125	141
Mountain	64	64	115	116
New England	42	39	47	62
Pacific	93	84	108	108
South Atlantic	195	350	319	380
West North Central	87	121	156	202
West South Central	104	218	232	248

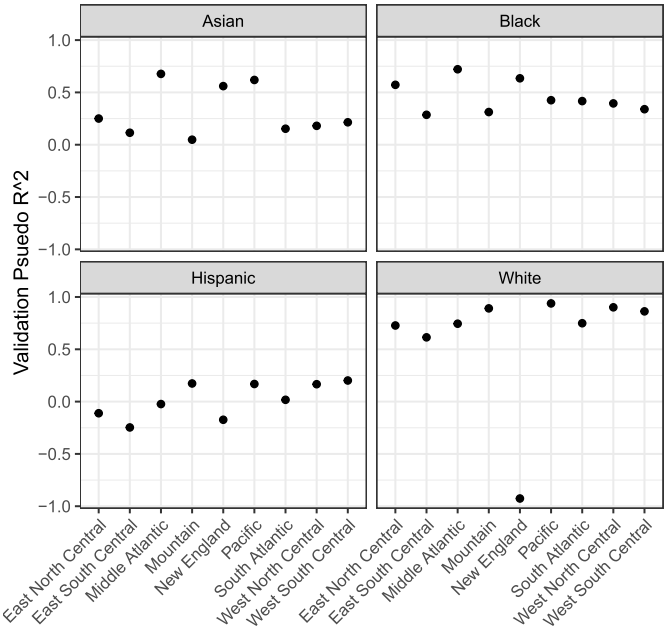

Figure 3: Model performance metric $\widetilde{R^2}$ for each geographical Division across racial categories.

Models in the cross-validation set were considered well-performing if the pseudo R^2 of the model was greater than 0.1. For each well-performing model, highly important features were defined as those with a increase in MSE above the 95th quantile of all predictors in the model. Briefly, the most important risk factors for homeownership in Asian populations were the percentage of POC homeowners earning between 100% and 150% of the area median income, commute times of less than 10 minutes, and job availability for Asian populations, those with a high school education, and those between the ages of 30–54 (Figure 4A). Interestingly, the job availability for White people was also an important predictor, likely due to the high correlation of job availability variables per race.

For the Asian and Black models, the predictive efficacy of the model dropped significantly when fewer than approximately 25% of the jobs available required a bachelor's degree or higher education. Income and employment variables were highly important in Asian models while group

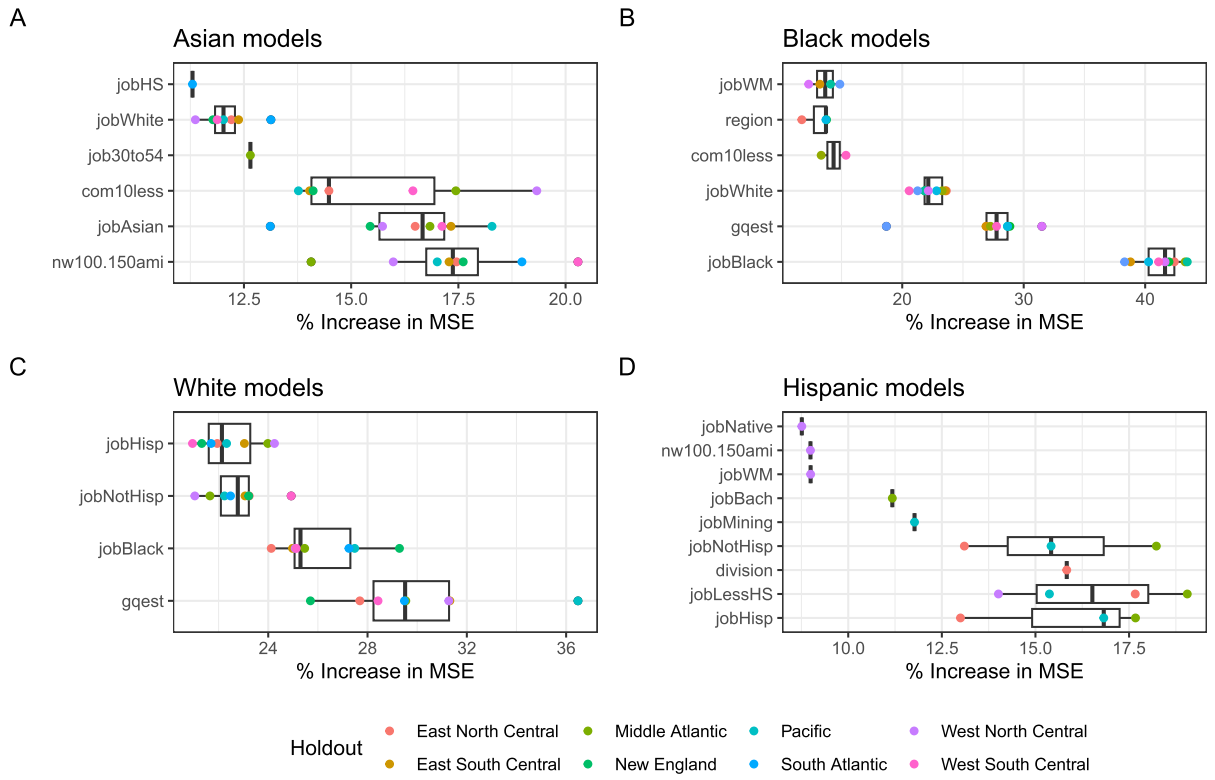

Figure 4: Percent increase in MSE across all cross-validation holdouts. Percent increase in MSE was calculated with the randomForest R package using a permutation method in which larger increases in MSE indicate variables more important to the overall model performance.

quarters estimates and employment variables were highly important in Black models. Education and employment variables were highly important in the White models and Hispanic models, though certain types of employment representation seemed to have stronger importance in Hispanic models compared with other racial categories.

In the Black models, US region, group quarters estimates, commute times of less than 10 minutes, and job availability for the Black populations and White populations were important predictors (Figure 4B). Interestingly, the number of jobs available for waste management positions was also highly important.

In the White models, group quarters estimates and job availability for Hispanic, non-Hispanic, and Black populations were the most important risk factors (Figure 4C). Once again, the job availability per race variables share high correlations which may explain why job availability for other races were important variables in the White models.

Across the Hispanic models, the percentage of POC homeowners earning between 100% and 150% of the area median income, the division of the US, and job availability for Hispanic, non-Hispanic, and Native populations; job availability for mining and waste management positions; and job availability with requirements that are less than high school or at least a bachelor's degree were the most important variables (Figure 4D). Once again, caution must be exercised in the interpretation of these variables, as many job availability variables were highly correlated with each other.

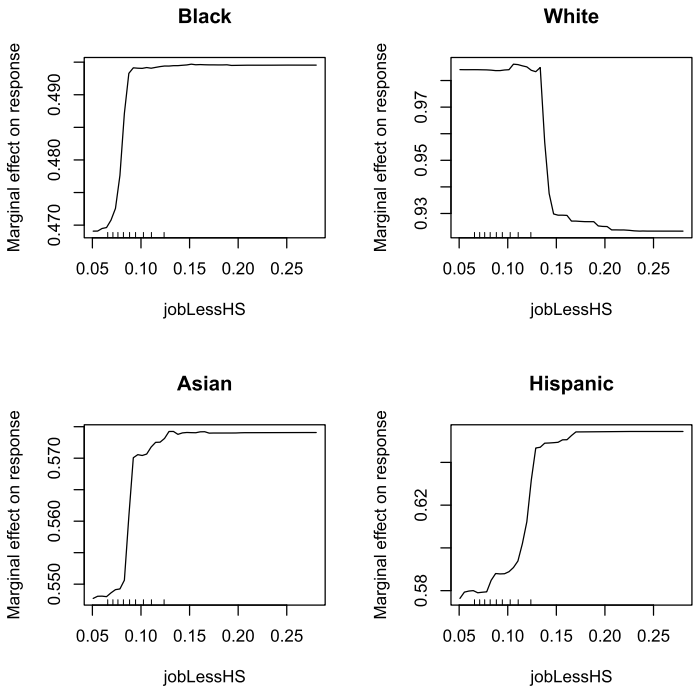

Figure 5: Partial dependence on jobs available to workers with less than a high school education on homeownership equity for Pacific division holdout. Y-axis indicates marginal effects of the predictor variable on the response, while the x-axis indicates the values of the predictor variable.

Factors that were uniquely important to a single racial category include job availability for Asian populations (Asian models), job availability with a high school education (Asian models), job availability for those between 30 and 54 (Asian models), region of the US (Black models), job availability with a high school education (Hispanic models), division of the US (Hispanic models), and job availability with a bachelor's degree (Hispanic models).

Due to the non-directionality of importance metrics and non-linear relationships between the response and variables of interest, partial dependence plots were used to depict the marginal effects of predictors on the response. An example using the proportion of all jobs available for those with less than a high school education (jobLessHS) is displayed in Figure 5. For this example, jobLessHS has a distinctly different effect on the response in the White model as compared to the Asian, Black, and Hispanic models. For the White model, the marginal effect of jobLessHS on HEI was strongest when the job availability was 15% and decreased as the jobLessHS increased. For all other models, the opposite was true – the marginal effects of jobLessHS on HEI increased as jobLessHS increased and plateaued around the 10% threshold. Further, the maximum the marginal effect of jobLessHS on HEI in the White model was nearly double the maximum marginal effect of other races.

5 Discussion

This study combined multiple publicly available data sources to evaluate factors that are predictive of homeownership rates in the United States. Random forest regression models were fit for four racial categories, and the predictive efficacy of these models was evaluated.

Exploratory data analysis revealed several interesting trends and outliers. Though some outliers in the predictor datasets could be explained by differences between rural and urban populations, like the higher availability of mining jobs in rural areas, others were not so easily explained. For example, Kentucky had unusually low foreclosure rates from 2015–2019 (Figure 1D), which could be the result of unique foreclosure laws, more intergenerationally-owned homes, or simply data collection artifacts.

The random forest regression models' ability to predict HEI varied substantially in our analyses across the racial-category models, with higher predictive performance in the White and Black models compared with the Asian and Hispanic models. Potential reasons for these differences in performance include inconsistencies with categorizing racial and ethnic groups both on a survey and social level. For example, Hispanic/Latino is considered a racial category in some surveys although it is more appropriately categorized as an ethnicity (Perez and Hirschman, 2009). Respondents in some surveys can self-identify as both Black and Hispanic, for example, while in other surveys, only one or the other may be selected. Some surveys also allowed multi-racial categories which could not be included in these models due to a lack of consistent availability across data sources.

The proportion of jobs available for different race categories and group quarters estimates stood out as consistently important variables across all models. It is perhaps not surprising that these non-race-specific factors were important. Intuitively, counties with more jobs will have more homeownership, and counties with more apartments, dormitories, and jails will have less homeownership. However, it is the observed differences between predictor variables' effects on the response that are most interesting. For example, when looking at the change in marginal effect on HEI as the proportion of jobs available to workers in a county with less than a high school education increases, we see a pattern in the White model which is inverted in the Black, Asian, and Hispanic models (Figure 5). While White homeownership equity is most impacted when the proportion of jobs for those with less than a high school education is low, this trend is reversed in the Black, Asian, and Hispanic models, perhaps indicating that when jobs with lower education barriers are plentiful, there is a meaningful effect on POC homeownership equity. This suggests a potential area of impact for policymakers, where supporting vocational training programs or raising the minimum wage may be paths to homeownership equity. Due to the high correlation of variables in our datasets, variable importance of best predictors may confound importance of other highly correlated predictors. This should be taken into consideration in future studies to determine the relationships between best predictors and other correlated predictors.

There are several limitations of this study which should be noted. First, counties with very small Asian, Hispanic, and Black populations were not included in the final random forest models. The results of this study's analyses may not represent any disparity in homeownership rates in these counties. Further, the datasets used in this study do not cover all potential variables which may be related to homeownership rates. For example, other potential data sources are publicly available on information such as building permits (US Census Bureau, 2019c) and household conditions (Urban Institute, 2020). Additionally, as noted in the commute data, not all regions are equally surveyed and joining multiple descriptive datasets can reduce the total amount of data available for a multi-source data analysis. It should be noted that all counties in Louisiana, Hawaii, and Alaska, which historically have large POC populations, were excluded from this work due to data availability. Lastly, the inclusion of commute times in metropolitan areas was only meant to assess the availability and access to jobs in high-density regions. Any observations made about the impact of commute times on homeownership should not be ex-

trapolated to populations in rural communities in this discussion. However, future work could leverage a US Census Bureau report with rural populations (US Census Beaureau, 2019).

Homeownership disparity in the United States is one component of the racial wealth gap. This study aimed to provide research-based information on factors that may impact homeownership rates for POC. Although additional research is required to further elucidate these factors, this work serves as a foundation for policymakers to begin to incorporate study findings into initiatives to support homeownership in POC communities.

Supplementary Material

Open-source code, additional visualizations and tables, as well as original datasets are available in a public GitHub repository:
https://github.com/PNNL-CompBio/HomeownershipDisparity_2015_2019

Tables

- Table 1: Descriptions of variables
- Table 2: Division county percentages per Race dataset
- Table 3: Number of counties in each dataset

Plots

- Figure 1: Dataset timeline
- Figure 2: Important variables for White models with and without outliers
- Figure 3: Correlation between predictor variables
- Figure 4: Model performance on 10% holdout data

Funding

PNNL is a multi-program national laboratory operated for the U.S. Department of Energy (DOE) by Battelle Memorial Institute under Contract No. DE-AC05-76RL01830.

References

Biau G, Scornet E (2016). A random forest guided tour. *TEST*, 25(2): 197–227. https://doi.org/10.1007/s11749-016-0481-7

Breiman L (2001). Random forests. *Machine Learning*, 45(1): 5–32. https://doi.org/10.1023/A:1010933404324

Choi JH, McCargo A, Neal M, Goodman L, Young C (2019). *Explaining the Black-White Homeownership Gap*, volume 25. Urban Institute, Washington, DC. Retrieved: March 25, 2021.

Department of Defense (2014). Per diem rates by location. Retrieved from: https://www.travel.dod.mil/.

Gabriel SA, Rosenthal SS (2005). Homeownership in the 1980s and 1990s: aggregate trends and racial gaps. *Journal of Urban Economics*, 57(1): 101–127. https://doi.org/10.1016/j.jue.2004.09.001

Hafen R, Schloerke B (2021). *trelliscopejs: Create Interactive Trelliscope Displays*. R package version 0.2.6.

Henry L, Wickham H (2022). *purrr: Functional Programming Tools*. R package version 0.3.5.

Hilber CA, Liu Y (2008). Explaining the Black–White homeownership gap: The role of own wealth, parental externalities and locational preferences. *Journal of Housing Economics*, 17(2): 152–174. https://doi.org/10.1016/j.jhe.2008.02.001

Kuebler M, Rugh JS (2013). New evidence on racial and ethnic disparities in homeownership in the United States from 2001 to 2010. *Social Science Research*, 42(5): 1357–1374. https://doi.org/10.1016/j.ssresearch.2013.06.004

Liaw A, Wiener M (2013). Classification and regression by randomForest. *R News*, 2(3): 18–22.

McCargo A, Choi JH, Golding E (2019). *Building Black Homeownership Bridges: A Five-Point Framework for Reducing the Racial Homeownership Gap*. Urban Institute, Washington, DC.

Perez AD, Hirschman C (2009). The changing racial and ethnic composition of the US population: Emerging American identities. *Population and Development Review*, 35(1): 1–51. https://doi.org/10.1111/j.1728-4457.2009.00260.x

R Core Team (2020). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Ray R, Perry AM, Harshbarger D, Elizondo S, Gibbons A (2021). *Homeownership, racial segregation, and policy solutions to racial wealth equity*.

Spearman C (1987). The proof and measurement of association between two things. *The American Journal of Psychology*, 100(3/4): 441–471. https://doi.org/10.2307/1422689

Strochak S, Ueyama K, Williams A (2022). *urbnmapr: State and county shapefiles in sf and tibble format*. R package version 0.0.0.9002.

Turner TM, Luea H (2009). Homeownership, wealth accumulation and income status. *Journal of Housing Economics*, 18(2): 104–114. https://doi.org/10.1016/j.jhe.2009.04.005

Urban Institute (2020). Household conditions by geographical school district. Retrieved from: https://datacatalog.urban.org/dataset/household-conditions-geographic-school-district. Data originally sourced from NHGIS. developed at the Urban Institute, and made available under the ODC-BY 1.0 Attribution License.

Urban Institute (2021a). Homeowner assistance fund county-level targeting data. Retrieved from: https://datacatalog.urban.org/dataset/homeowner-assistance-fund-county-l. Data originally sourced from NHGIS, developed at the Urban Institute, and made available under the ODC-BY 1.0 Attribution License.

Urban Institute (2021b). Unequal commute data. Retrieved from: https://datacatalog.urban.org/dataset/unequal-commute-data. Data originally sourced from US Census Bureau's 2017 LEHD Origin-Destination Employment Statistics, 2014–18 American Community Survey five-year estimates, Transitland repository, OpenStreetMap, and INRIX's 2019 Global Traffic Scorecard, developed at the Urban Institute, and made available under the ODC-BY 1.0 Attribution License.

Urban Institute (2022). Longitudinal Employer-household Dynamics origin-destination Employment Statistics (LODES) summary files – census tract level. Retrieved from: https://datacatalog.urban.org/dataset/longitudinal-employer-household-dynamics-origin-destination-employment-statistics-lodes. Data originally sourced from the US Census Bureau, developed at the Urban Institute, and made available under the ODC-BY 1.0 Attribution License.

US Census Beaureau (2019). Travel time to work in the United States. Retrieved from: https://www.census.gov/content/dam/Census/library/publications/2021/acs/acs-47.pdf.

US Census Bureau (2013). Census bureau regions and divisions with state FIPS codes. Retrieved from: https://www2.census.gov/geo/pdfs/maps-data/maps/reference/.

US Census Bureau (2019a). Annual county resident population estimates by age, sex, race, and Hispanic origin: April 1, 2010 to July 1, 2019. Retrieved from: https://www.census.gov/data/tables/time-series/demo/popest/2010s-counties-detail.html.

US Census Bureau (2019b). Annual resident population estimates, estimated components of resident population change, and rates of the components of resident population change for states and counties: April 1, 2010 to July 1, 2019. Retrieved from: https://www.census.gov/data/tables/time-series/demo/popest/2010s-counties-total.html.

US Census Bureau (2019c). Building permits survey. Retrieved from: https://www.census.gov/construction/bps/.

US Census Bureau (2020a). Average household size and population density. Retrieved from: https://covid19.census.gov/datasets/USCensus::average-household-size-and-population-density-county.

US Census Bureau (2020b). Highest level of educational attainment. Retrieved from: https://data.ers.usda.gov/reports.aspx?ID=17829.

US Census Bureau (2022). American Community Survey 1-year estimates: New England Division. Retrieved from: https://censusreporter.org/profiles/03000US1-new-england-division/.

Wickham H (2016). *ggplot2: Elegant Graphics for Data Analysis*.

Wickham H, François Henry L R, Müller K (2022). *dplyr: A Grammar of Data Manipulation*. R package version 1.0.10.