

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

RACHEL RICHARDSON^{1,*}, DAMON LEACH¹, ANASTASIYA PRYMOLENN², DAVID DEGNAN¹,
NATALIE WINANS¹, AND LISA BRAMER¹

¹902 Battelle Boulevard, Richland, WA, 99354, Biological Sciences Division, Pacific Northwest National Laboratory, USA

²902 Battelle Boulevard, Richland, WA, 99354, Environmental Molecular Sciences Division, Pacific Northwest National Laboratory, USA

Abstract

Homeownership, an indicator of financial security, is impacted by discrimination towards minority populations. Factors that contribute to unequal homeownership between white and minority population have not been fully characterized since unequal homeownership persists. In order to alleviate the issue, policymakers need a better understanding of how risk factors affect homeownership on a per race level. Here, we utilized several publicly available surveys, such as the American Community Survey and United States Census, and statistical learning models to investigate potential factors related to homeownership in White, Black, Hispanic, and Asian populations, with a focus on how risk factors vary per race. We incorporated both known risk factors and potential factors like job availability per job class and commute times to work.

Keywords *Disparity; Homeownership; Survey; Census; Economics; United States*

1 Introduction

The purchasing of a home benefits both the family, as a catalyst for growing financial assets, and the nation at large, as an increase in the overall wealth and economy (Turner and Luea, 2009). In the United States, the impacts of historic prohibitive practices (e.g. slavery, Jim Crow laws, redlining, restrictions to loans, etc.) have created a racial homeownership gap (Ray et al., 2021). Current risk factors contributing to homeownership disparity can be traced back to these events. For example, the impacts of redlining help explain the lack of intergenerational wealth in minority populations today (Ray et al., 2021). Several risk factors are well characterized; however, there may be persistent undetected factors contributing to homeownership disparity.

Published studies have identified race-specific risk factors for this inequity, including a lower percentage of underbanking in white populations, a higher median credit score within the same zip code between white and black populations, and gaps in educational attainment (Ray et al., 2021). Other factors such as gender, marital status, presence of children, and occupation have also been identified as potential race-specific factors (Kuebler and Rugh, 2013). What is still not well understood is how these factors vary per race, and whether some risk factors are more pertinent to one race than they are to another.

This work aims to provide further insight into factors related to homeownership. Multiple publicly available datasets were leveraged with random forest regression models to evaluate and

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

identify potential potential risk factors for homeownership inequity in specific racial groups. Herein, observed trends in homeownership and model predictive performance by racial group are presented. Further, key race-agnostic and race-specific predictive factors for homeownership are evaluated with the intent to promote more equitable homeownership policies and data collection strategies in the United States.

2 Methods

2.1 Data Sources

Data from 2015 to 2019 and the four races reported in the Homeownership Demographics Data: Asian, Black, Hispanic, and White were considered for modeling. Data sources used are given below, and Table 1 provides a list of variables referenced in the body of this manuscript. All datasets used and 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.

2.1.1 Census Population Estimates

Census population estimates were included in the model to accurately describe the racial population in each county. Data were downloaded from the county-level report of “Annual County Resident Population Estimates by Age, Sex, Race, and Hispanic Origin” for years included from 2015 through 2019. Population estimates were restricted to individuals with the plausible homeownership eligibility of age 20 and above ([US Census Bureau, 2019a](#)). Average population values were computed by county and race within a county.

2.1.2 Census Population Change

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 (U.S.) Census Bureau’s “Annual Resident Population Estimates” report. Variables included group quarter 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 populations residing in shared spaces, such as correctional facilities, nursing homes, college dormitories, or shelters.

2.1.3 Per Diem Rates

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 downloaded and available at a city or county level, depending on the area, and for the years 2015 through 2019 ([Department of Defense, 2014](#)). The annual maximum lodging per diem rates were averaged over that timeline for each county.

2.1.4 Education: High School Completion

Educational data was included as a metric of socio-economic success of youth 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).

2.1.5 Homeownership

The “Homeowner’s Demographic 5-year” dataset from the American Community Survey released by the US Census Bureau summarized estimated homeownership 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 non-white 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 people of the respective race(s) of age 20 or greater. 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. The averages of these variables were then calculated over this timeline.

2.1.6 Job Availability

Data on job availability were included as indicators of financial opportunity in each county. Counts of job availability per age (≤ 29 , 30-54, ≥ 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 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.

2.1.7 Commute Demographics

Commute data was included as an indicator of affluence to afford travel or the relative inaffordability of living closer to work opportunities. 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 opportunity. Job access for 1,913 counties were then estimated as described elsewhere (Urban Institute, 2021b). To calculate 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).

Table 1: List of variables with high importance in at least one model, with their corresponding descriptions. A complete list of all variables used can be found in Table S1 of the Supplemental Material. All variables were averaged to the county level.

Variable	Description
job30to54	Proportion of jobs that are available for ages 30 to 54
jobIncomeL	Proportion of jobs that are available with monthly earnings less than 1250
jobAgr	Proportion of jobs that are available for the agriculture industry
jobMining	Proportion of jobs that are available for the mining industry
jobConst	Proportion of jobs that are available for the construction industry
jobWholeS	Proportion of jobs that are available for the wholesale industry
jobScience	Proportion of jobs that are available for the science industry
jobWM	Proportion of jobs that are available for the waste management industry
jobPubAd	Proportion of jobs that are available for the public administration industry
joblessHS	Proportion of jobs that are available for those with less than a high school education
jobHS	Proportion of jobs that are available for those with some college education
jobBach	Proportion of jobs that are available for those with a bachelor's education or higher
jobWhite	Proportion of jobs that are available for those who are white
jobBlack	Proportion of jobs that are available for those who are black
jobNative	Proportion of jobs that are available for those who are native
jobAsian	Proportion of jobs that are available for those who are asian
jobAAPI	Proportion of jobs that are available for those who are pacific islander
jobMRacial	Proportion of jobs that are available for those who are multiracial
jobNotHisp	Proportion of jobs that are available for those who are not hispanic
jobHisp	Proportion of jobs that are available for those who are hispanic
jobHS	Proportion of jobs that are available for those with a high school education
jobMgmt	Proportion of jobs that are available for the management industry
jobRealEst	Proportion of jobs that are available for the real estate industry
jobEntmt	Proportion of jobs that are available for the entertainment industry
jobOther	Proportion of jobs that are available for the other services industry
jobMale	Proportion of jobs that are available for those who identify as male
com10less	Proportion of people with job access with a commute time of less than 10 minutes
com30to59	Proportion of people with job access with a commute time of 30 to 59 minutes
com60plus	Proportion of people with job access with a commute time of 60 minutes or more
hsRate	Proportion of individuals with at least a high school degree
gqest	Population counts of those living in shared spaces, such as correctional facilities, nursing homes, college dormitories, or shelters
rBirth	Rate of births
rDomestic	Rate of net domestic migration, which is the rate of US citizens moving into a county subtracted by the rate of US citizens leaving a county

rMigration	Rate of net migration, which is the rate of internationals moving into a county subtracted by the rate of internationals leaving a county
rForeclose	Rate of foreclosures
nw100.150ami	Proportion of non-white homeowners earning between 100 and 150 of area-median-income
w100.150ami	Proportion of white homeowners earning between 100 and 150 of area-median-income
mort100ami	Count of homeowners with an outstanding mortgage earning less than 100 percent of area-median-income
mort150ami	Count of homeowners with an outstanding mortgage earning less than 150 percent of area-median-income
costBurdSev	Proportion of homeowners with a mortgage who are severely cost-burdened (spending 50 percent or more of annual income on housing costs)
popDensity	Count of people per square kilometer
annualTot	Count of total jobs
region	Four census regions encompassing each quadrant of the United States (West, Midwest, South, Northeast)
division	Each region is divided into two or more census divisions

2.2 Data Analysis

All analyses were done in R version 4.2.2 (R Core Team, 2020). Data processing was facilitated using 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). The Spearman correlation (Spearman, 1904) between all pairs of predictor variables to identify key relationships with the *stats* package in base R (R Core Team, 2020). The *urbanmapr* (Strochak et al., 2022) R package was used for all geographical plots. Partial dependence plots from the R package *randomForest* are used for visualization of marginal effects of highly important features on the response variable.

2.3 Modeling Approach

A metric of homeownership equity was constructed as the response variable of interest in subsequent models. This metric, defined as HEI , was calculated as:

$$HEI_c = \frac{h_{r,c}/h_{t,c}}{p_{r,c}/p_{t,c}}, \quad (1)$$

where $h_{r,c}$ is the number of homeowners of race r in county c , $h_{r,t}$ is the total number of homeowners in county c , $p_{r,c}$ is the population of racial group r in a county c of age 20 or greater, and $p_{t,c}$ is the total population in county c of age 20 or greater.

A random forest regression model (Breiman, 2001) was used to accommodate potential non-linear relationships (Biau and Scornet, 2016). A separate model was fit for each of the four most populous races in the homeownership data: Asian, Black, Hispanic, and White, and HEI_c was used as the response variable of interest. A total of 69 variables (Table S1) were used as potential explanatory variables. Random forest regression models were fit separately to allow for a clearer picture of race-specific predictors of homeownership equity. Models were fit using the *randomForest* (Liaw and Wiener, 2013) package.

2.3.1 Performance Metrics

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:

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

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

$$\widetilde{R}^2 = \frac{1 - MSE}{s^2}, \quad (3)$$

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

The importance of variables to the model was measured using the increase in MSE as calculated by the *randomForest* package. Briefly, values are shuffled for the predictor variable of interest and the change in MSE compared to the model using the original predictor variable is calculated. Typically, larger increases in MSE are indicative of more important variables to the overall model performance.

2.3.2 Population Requirements

The racial populations in each county are unequal with some counties containing few individuals of a specific race. Thus, we evaluated our model 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 (the "elbow" of the curve). This value was determined to be 500.

2.3.3 Parameter Tuning

Initial parameter tuning was conducted over potential parameter values for the number of trees ($n_{tree} = 500, 1000, 1500, 2000, 2500, 3000$), number of randomly selected predictors considered at each split ($m_{try} = 10, 12, 15, 20, 31, 35, 30, 35, 40, 45, 50$), and the minimum terminal node size ($n_{node} = 3, 5, 6, 9, 12$) (Figure S2-S4). Ten random forest models were fit for each combination of tuning parameters. Evaluation of tuning parameters was done using randomly selected counties for training and model validation (70% and 30%, respectively). These tuning parameters were selected based on iterative testing (Fig. SXX) for combinations that minimized the MSE and maximized the \widetilde{R}^2 . Parameter values were set to $n_{tree} = 1000$, $m_{try} = 35$, and $n_{node} = 9$.

2.3.4 Model Validation

Evaluation of model performance was performed using cross-validation by holding out the counties from each geographical division, as defined by the US Census Bureau, per fold (Figure 1A) as the testing 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 race model are given in Table S2. A total of 36 random forest regression models were fit, nine cross-validation folds for each of the four race models.

2.3.5 Interaction Effects

In cases of high dimensional data, random forest models struggle to automatically account for variable interactions, unless strong marginal effects exist `interact1`, `interact2`. Therefore, potential variable interactions were explicitly evaluated for the purpose of specification and inclusion as explanatory variables. The interpretable machine learning (Molnar et al., 2018) package was then used to generate interaction variables for all pairwise variable interactions between the variables with a variable importance of at least 5% increase in MSE. Interactions with variable importance metrics in the top 1% of MSE increase scores, for that race’s model, were retained as explanatory variables.

3 Results

3.1 Exploratory Data Analysis

3.1.1 Predictor and Response Variable Overview

Included predictor variables were predominantly proportions of individuals within a race with a particular trait, such as the number of individuals graduating from high school over the total number of individuals within that county. Other variables included count data, factors (e.g., the level of poverty), or continuous data (e.g., average yearly income) per race. Examination of the distributions across all variables in the final dataset demonstrated that many of these metrics contained outliers that often were located in urban hubs (Figure 1C).

Other outliers were not so easily explained; for example, the predicted home foreclosure rate were observed to be very low in the entire state of Kentucky for unknown reasons (Figure 1D). Further research and investigation is needed to examine data collection methods or data values that might be influenced by county, state or regional laws.

Additional geographical maps of variable distribution are available in a public GitHub repository: https://github.com/rarichardson92/Homeownership_disparity

Response Variable While a strong outlier was observed in Petroleum County, MT for the response variable, HEI, in the Asian population, the majority of the response values trended well below equitable representation of for non-white populations (Figure 2).

Variables across all datasets tended to correlate strongly, especially in the case of percentages of the population labeled by the predictor variable or related rates. The largest similarly performing correlation block among the variables implied consistent relationships between type of employment, income, education, and migration into the county (Figure S5).

3.1.2 Urban and Rural Representation

Counties with a population density of greater than 1000 people per square kilometer are considered to be urban and counties with a population density of less than 500 people per square kilometer are considered to be rural. The representative homeownership distribution shown in Figure 3, is representative of both rural and urban counties, where we see a higher median of White homeownership across both rural and urban counties. Unlike rural counties, urban Asian homeownership median is greater in Urban counties.

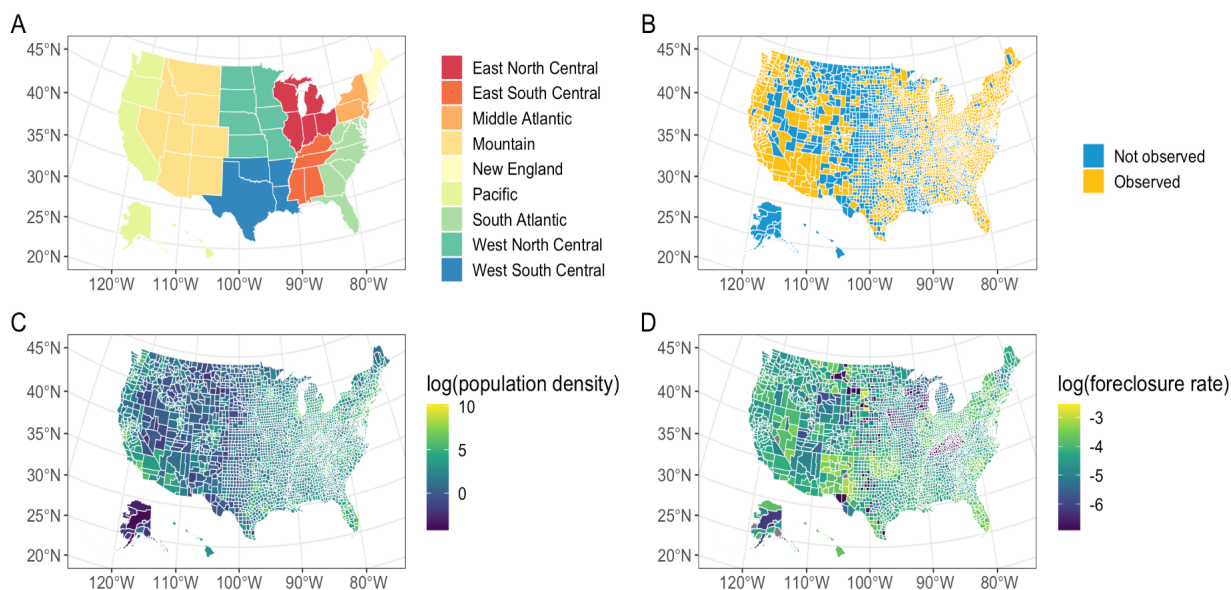


Figure 1: Geographic maps of counties and variables. A) Divisions in the United States used for cross validation (US Census Bureau 2013). B) Counties represented in the final dataset based on consensus observations in all datasets. C) Log of the population density across U.S. counties. D) Log of the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

3.2 Population Requirements

Populations of each race and ethnicity were not present in equal numbers across counties. Where too few of any single race or ethnicity exists, cannot be a reasonable estimation for a community of that race or ethnicity. These estimations can skew the results of a model - iterative testing was performed to determine reasonable minimums without losing too many counties for each model. The final model used required a minimum of 500 members of a race or ethnicity based on diminishing returns in performance and regional representation (Figure S6). After removing counties below this threshold, the number of counties used per race and region are given in Table 2.

Table 2: Number of counties per race and region with ≥ 500 individuals of each race

Race	Midwest Region	Northeast Region	South Region	West Region
Asian	219	139	343	157
Black	320	160	704	148
Hispanic	376	172	648	223
Asian	523	203	800	224

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 1,750 (Fig. 1B).

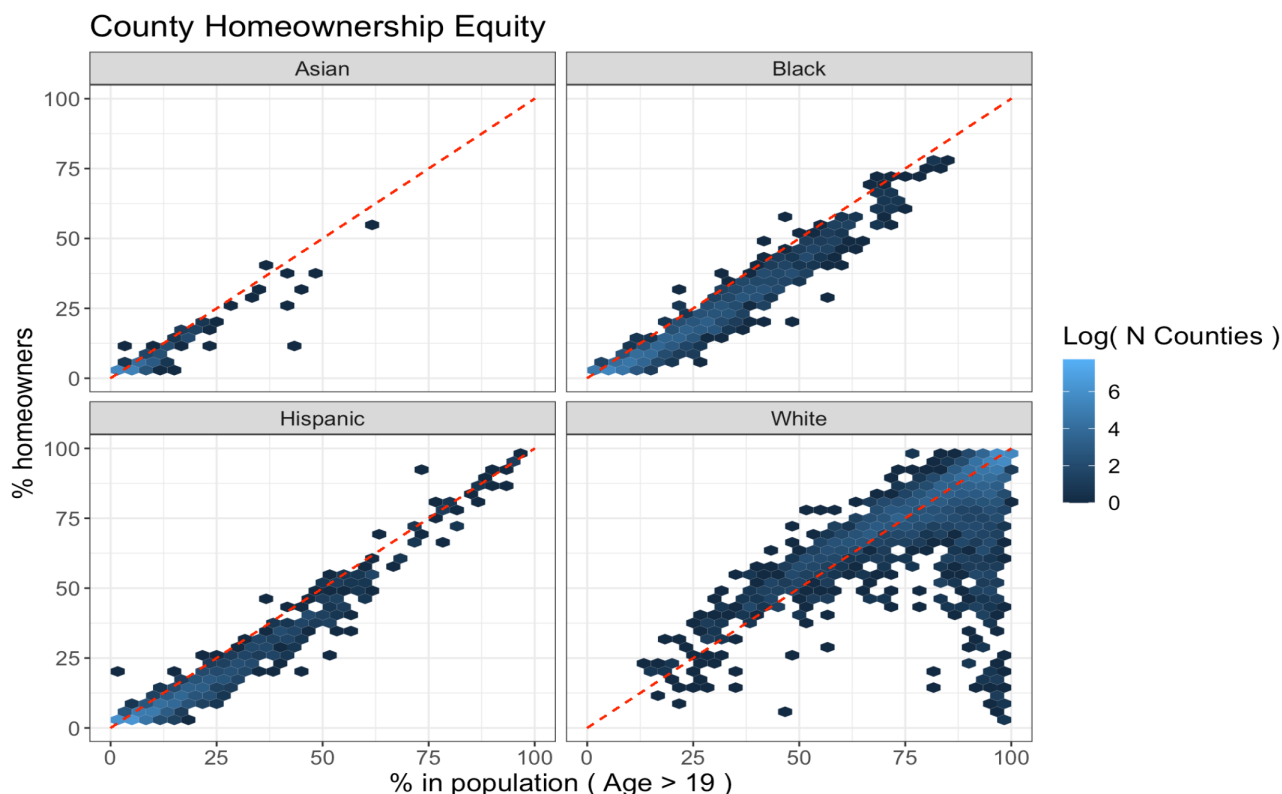


Figure 2: Equity of homeownership across US counties, where trending along the dotted red line indicates homeownership equity. Hexes below the redline depict racial underrepresentation of homeowners, while hexes above the line depicts racial overrepresentation of homeowners. Population representation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

3.3 Interaction Effects

Model performance at different variable importance cutoffs, where only variables above the threshold increase MSE were included (Figure S7). The strongest interaction effects were observed in the two models (Black and White) with the lowest out of bag errors (Figure 4). The top interactions per race were not consistent. Within the Asian and Hispanic models, interactions between commute time and job availability per job family, as well as education were observed to have high importance. In the Black model, many interactions were region-specific. Within the White population, the strongest interaction was between highly educated individuals living within 10 minutes of work, which was not as strong in other models. Features with mean interactions across divisions above the 99th quantile for each race were tabulated. The unique set of features from the 99th percentiles were added manually as interactions to data for generating the final random forest model.

Applying interaction effects to the random forest models adjusted the out-of-bag errors for White, Black, Hispanic, and Asian populations by an average of 2.1, -1.7, 3.4, and 1.4% respectively (Figure 5).

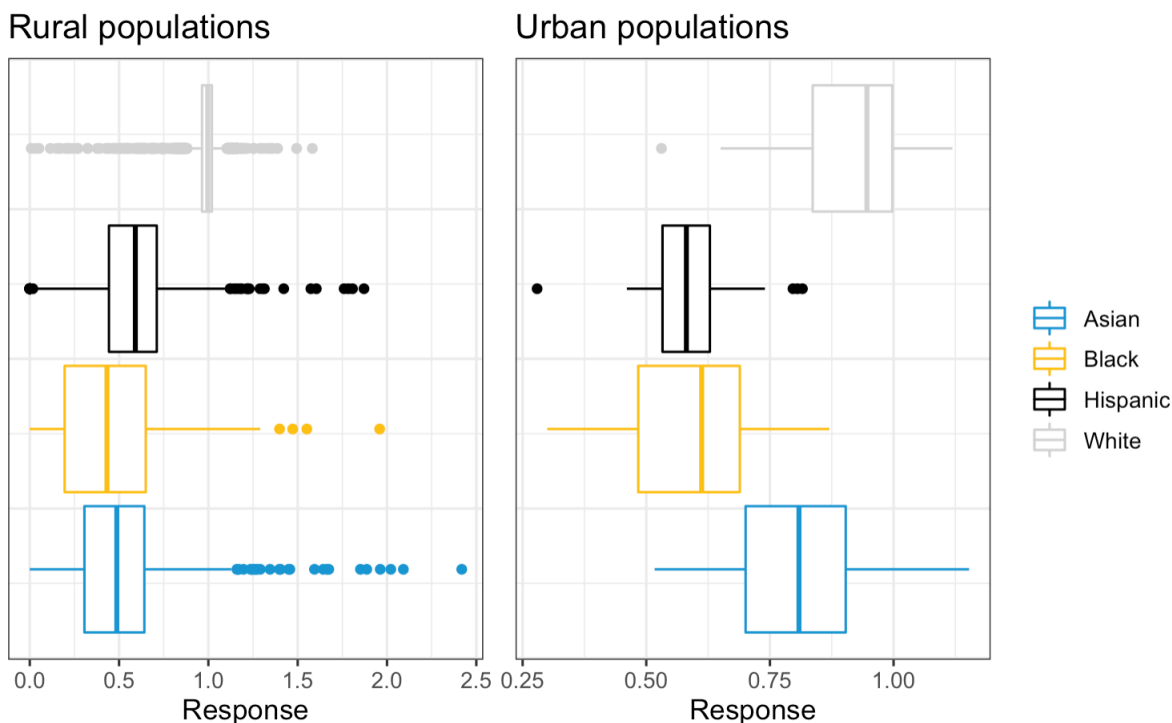


Figure 3: Summary of homeownership distribution within urban and rural populations by race.

3.4 Random Forest Model Performance

The performance of the random forest model changed drastically across races and holdout regions despite having identical predictor variables used in each random forest model. Poor performance was observed in the Hispanic population, where the best model had a \widetilde{R}^2 value of 0.222 (Figure 5). This poor performance was not due to a lack of counties to train the model, as the Hispanic subset contained the second most counties, but might be related to racial and ethnic definitions of this group changing over time. The White population model also had poor performance when predicting homeownership equity in the New England division. This may be due to the small holdout set for this division (3.32% of all counties in the White model). However, other similarly low holdouts in the Asian and Black models did not show the same level of poor performance (Table 1). It is also possible that the demographics in the New England region are particularly unique for the data used in the model.

3.4.1 Consistently Important Predictors

Including both categorical interactions and numeric interactions, the final model contained 206 predictors. For models performing with a \widetilde{R}^2 above 0.1, the most important predictors were considered using all features with % increase of MSE above the 95 quantile of % increase of MSE in each model. The features found using this metric were designated henceforth as highly important features in this study. Using this metric, 17, 16, 18, and 13 different highly important features were considered for Asian, Black, Hispanic, and White models (Figure 6), respectively.

Of all highly important features identified in each well-performing Asian model (pseudo $R^2 > 0.1$), 23.5% of features were consistently important in all cross-validation datasets, increasing to 35.3% if important in all but one holdout model (Figure 6A). These variables predominantly represented metrics of education, employment of the Asian population, and non-white homeowners with income between 100% to 150% of the area median income; working populations between the age of 30 to 54, employment in waste management, and commutes within 10 minutes also showed up consistently.

Of all highly important features identified in each well-performing Black model (pseudo $R^2 > 0.1$), 31.3% of features were consistently important in all cross-validation datasets, increasing to 50.0% if important in all but one holdout model (Figure 6B). These variables predominantly represented metrics of education, employment of the Black and White population, and group quarter estimates (e.g. population in hospitals, schools, dormitories, or prisons); employment in Public Administration also showed up consistently.

Of all highly important features identified in each well-performing White model (pseudo $R^2 > 0.1$), 61.5% of features were consistently important in all cross-validation datasets, increasing to 77.0% if important in all but one holdout model (Figure 6C). These variables predominantly represented metrics of education, employment of the Black, White, Hispanic, and group quarter estimates (e.g. population in hospitals, schools, dormitories, or prisons); employment in Public Administration also showed up consistently.

Of all highly important features identified in each well-performing Hispanic model (pseudo $R^2 > 0.1$), 27.8% of features were consistently important in all cross-validation datasets, increasing to 50.0% if important in all but one holdout model (Figure 6D). These variables predominantly represented metrics of education, employment of the Hispanic and Native population; group quarter estimates (e.g. population in hospitals, schools, dormitories, or prisons), Average Annual income, and employment in Public Administration, Mining, Waste Management, or Wholesale also showed up consistently.

3.4.2 Important Holdout-dependent Predictors

Some highly important features across the models only appeared when certain regions were excluded from the training data. Population and rate of birth had a strong influence on the Asian model where the East North Central division of the United states was not in the training set. In addition, the interaction between education and employment in the sciences appeared to have an influence when the Mountain region of the US was not present in the training data. For the black populations, commute estimates had a stronger influence when either the East South Central or the South Atlantic regions were excluded from the training data. In the Hispanic models, the categorical variable of Division had increased importance when the Mountain division was excluded from the training data. In addition, populations of white and non-white homeowners with income between 100% to 150% of the area median income appeared as highly important when either Mountain or Pacific divisions were excluded from the Hispanic model - this may be related to the average annual income that appears consistently in these models. No additional highly important variables in the White model appeared across holdouts in well-performing models.

3.4.3 Marginal Effects and Directionality

While the % increase in MSE indirectly indicates the strength of effect each of these variables have on the response variable, it is unable to tell us the directionality of this predictor on the

response variable. As these relationships can be non-linear in nature, partial dependence plots can be a useful depiction of the marginal effect of a predictor on the response. An example using workers with less than high school education is displayed in Figure 7, with all partial dependence plots available in Supplemental Materials.

The highly important predictor variables from the model were visualized to examine how these factors influence HEI. Several of these factors differed greatly depending on model type, but of particular note were important predictors that only identified in the White model over minority models and vice versa. The interaction of jobs available to workers with less than a high school education and jobs available in waste management appeared to be an important predictor across all three minority models, but was not strongly predictive for the White model (Figure 8). Jobs available to minorities was a strong predictor in the white model, both as a univariate predictor and with interactions, however both the univariate predictors of jobs available to the Hispanic population and non-Hispanic population only showed up in the white model. Additional plots for important predictor relationships to the response are available in Supplemental Materials.

4 Discussion

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

Exploratory data analysis of the data collected revealed several interesting trends and identifications of outliers. Though some outliers in the predictor datasets could be explained by rural and urban populations, like the higher job counts of mining jobs in rural centers over urban, others required further explanation and are not well understood. Of particular interest are the reasons why Kentucky had low foreclosure rates in the 2015-2019 time span, with reasons potentially being different foreclosure laws or more generational-owned homes.

The random forest regression models' predictive performance of HEI varied strongly in our analyses across all racial models, with higher performances for the White and Black models over the Asian and Hispanic models. Potential reasons for this difference in performance include inconsistencies with categorizing racial groups both on a survey and social level. For example, Asia is an enormous continent with many diverse subgroups that may identify differently. Some surveys also allowed multi-racial categories which could not be included in these models.

Across all models, the proportion of jobs available and group quarter estimates were observed to influence homeownership. These non-race specific risk factors are not surprising results. Counties with more jobs will have more homeownership, and counties with more apartments, dormitories, and jails will have less homeownership. Of particular interest are the observed differences between white and minority models, most notably Figure 7, where counties with low proportions of jobs available for those with a high school education tended to be filled with mostly white populations. As the proportion of jobs available increased, then the number of jobs available to minorities increased. This demonstrates a specific area of impact for policymakers, where counties with low proportions of jobs that do not require high school diplomas focus on more equitable hiring practices that encourage racial minorities to apply.

For Asian and Black models, the predictive efficacy of the model dropped significantly when approximately <25% of the jobs available require a bachelor's degree or higher education. This trend may be explained similarly to trends regarding minorities in counties with low job availability for those who completed high school. When there are less jobs available for a tier of

educational attainment, those jobs tend to go to the white population over minorities. This may be indicative of a disparity in job availability for minorities at multiple educational levels. Other differences in race models were observed with regard to commute times as well as in the interaction of workers in construction and workers in waste management, which require further examination in future studies. Increased access to well-paying jobs for minorities in all regions is a natural area for future policies targeting improvement of homeownership inequity. Income and employment had high importance for Asian models, Group quarters estimates and employment had high importance for Black models, education and racial representation in employment had strong importance for White models as well as Hispanic models, though certain types of employment representation seemed to have stronger importance in Hispanic models over other racial models.

There are several limitations of this study which should be noted. First, counties with small minority populations were not included in the models from this study. Results of analyses may not represent these counties and any disparity in homeownership. Further, the datasets used in this study do not cover all potential variables which may be related to homeownership. For example, other potential data sources are publicly available on information such as building permits UScensus2019a, household conditions urban2020. However, as noted in the commute data, not all regions are equally surveyed and joining with 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 are hubs for non-White populations, were excluded from this work due to data availability. Additionally, several datasets do not consistently categorize racial groups. In particular, it was noted that data relating to Hispanic populations was not always clearly distinguished from White populations and was instead categorized as an ethnicity in some datasets, such as the employment data, but was treated as a separate race in other data sources.

Continued research is necessary for policymakers to aid in providing equal access to homeownership opportunities and for racial equity to be achieved.

Supplemental Materials

Open-source code, additional visualizations and tables, as well as original datasets are available in a public GitHub repository: https://github.com/rarichardson92/Homeownership_disparity

Tables

- Table S1: Descriptions of variables
- Table S2: Hold-out region percentages per model
- Table S3: Number of counties in each dataset

Plots

- Figure S1: Dataset timeline
- Figure S2: Model tuning of mtry
- Figure S3: Model tuning of nnode
- Figure S4: Model tuning of ntree
- Figure S5: Correlation between predictor variables
- Figure S6: Model performance across population minimums
- Figure S7: Model performance with restricted variables

Acknowledgments

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.
- Breiman L (2001). Random forests. *Machine Learning*, 45(1): 5–32.
- Department of Defense (2014). Per diem rates by location. Retrieved from: <https://www.travel.dod.mil/>.
- 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.
- 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.
- Liaw A, Wiener M (2013). Classification and regression by randomforest. *R News*, 2(3): 18–22.
- Molnar C, Casalicchio G, Bischl B (2018). *iml: An r package for interpretable machine learning*. *JOSS*, 3(26): 786.
- 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 (1904). *The proof and measurement of association between two things*.
- 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.
- Urban Institute (2021a). Homeowner assistance fund county-level targeting data. Retrieved from <https://datacatalog.urban.org/dataset/homeowner-assistance-fund-county-1>. 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 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 (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>.
- Wickham H (2016). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York.
- Wickham H, François R, Henry L, Müller K (2022). *dplyr: A Grammar of Data Manipulation*. R package version 1.0.10.

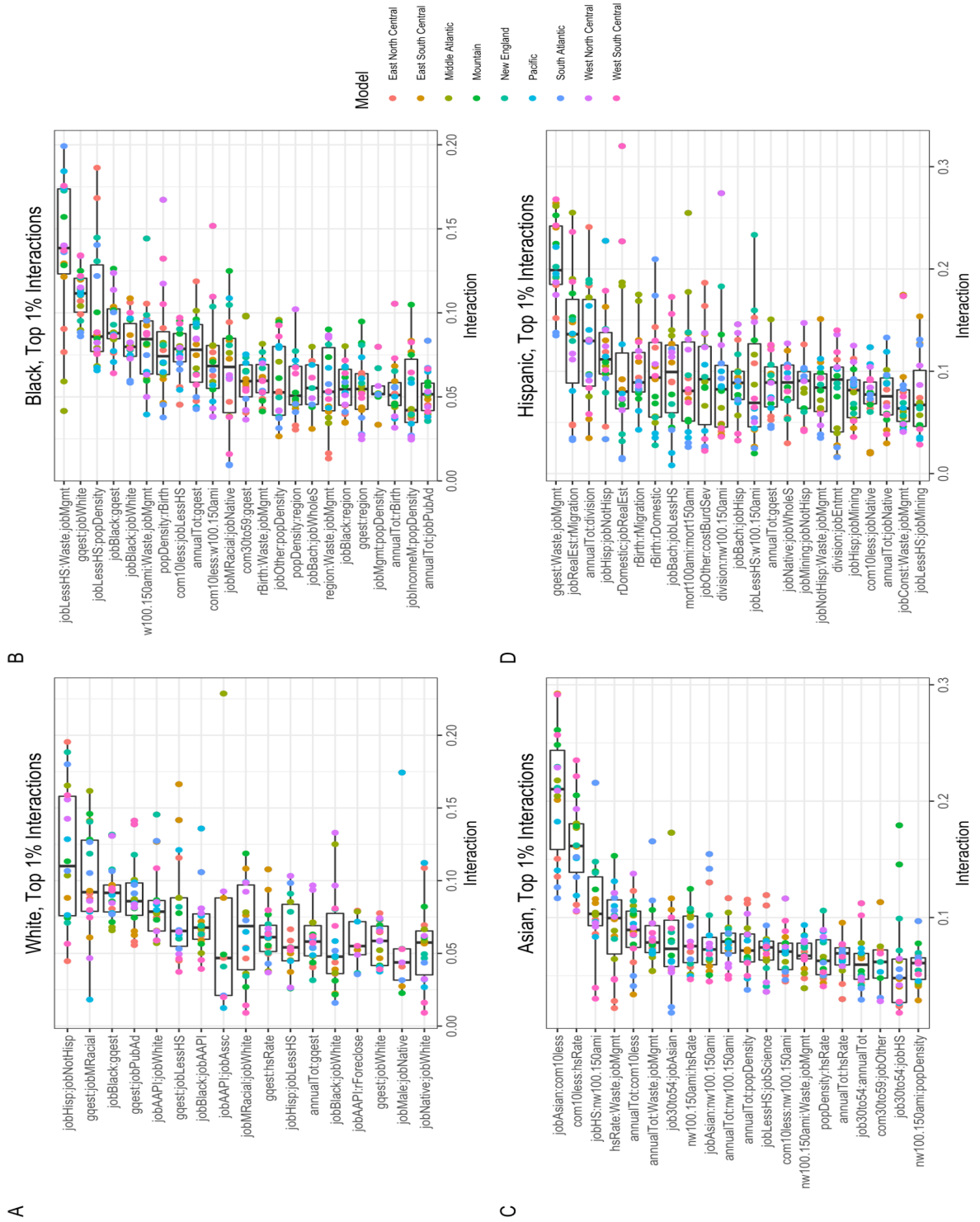


Figure 4: Top 1% of interactions per holdout model per race (A) White, (B) Black, (C) Asian, and (D) Hispanic.

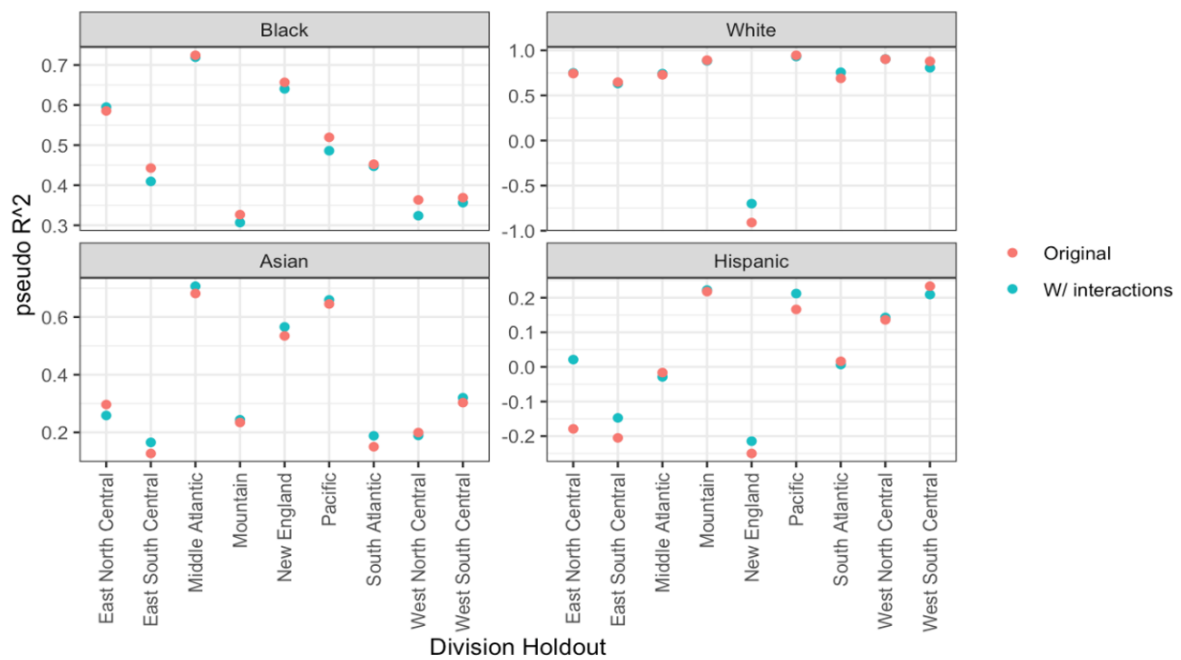


Figure 5: Performance of random forest model with interactions included vs. no interactions on each cross validation holdout model.

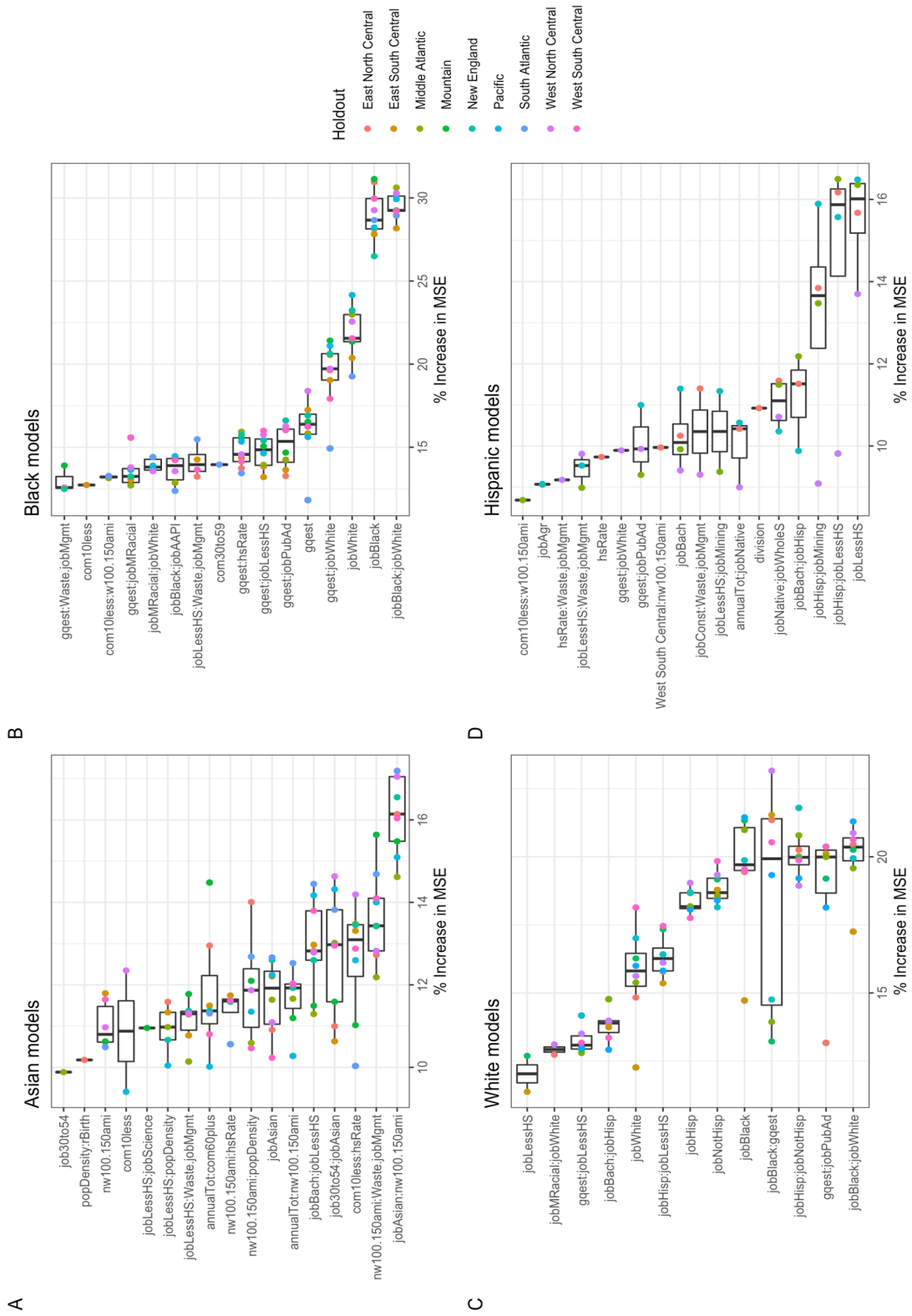


Figure 6: Percent increase in MSE when highly important variables are excluded from the model across all cross-validation holdouts.

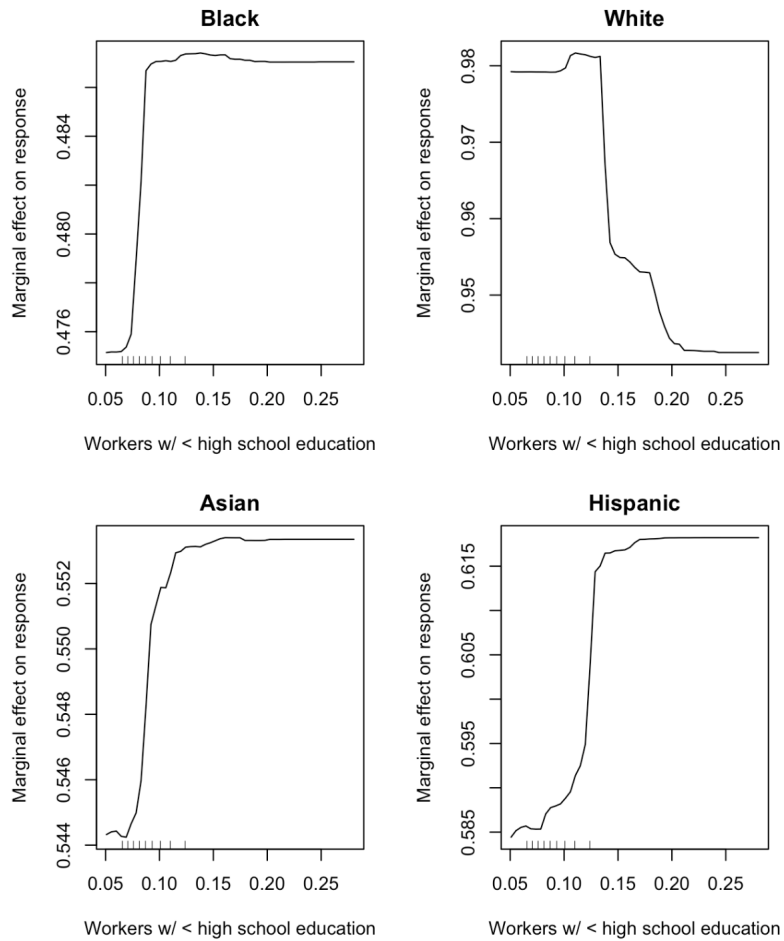


Figure 7: Partial dependence on workers with less than high school education on homeownership equity for Pacific division holdout. Y-axis indicates marginal effects of the predictor variable on the response, while the x-axis indicates the values of the predictor variable.

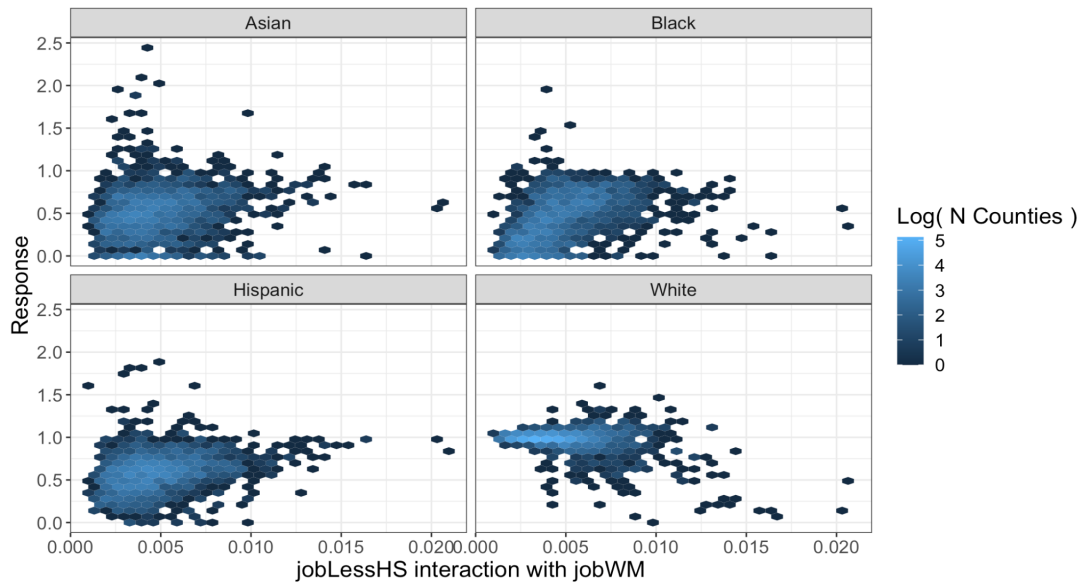


Figure 8: Interaction of jobs available in wastemanagement and jobs available to workers with less than a high school education as compared to the response (HEI).