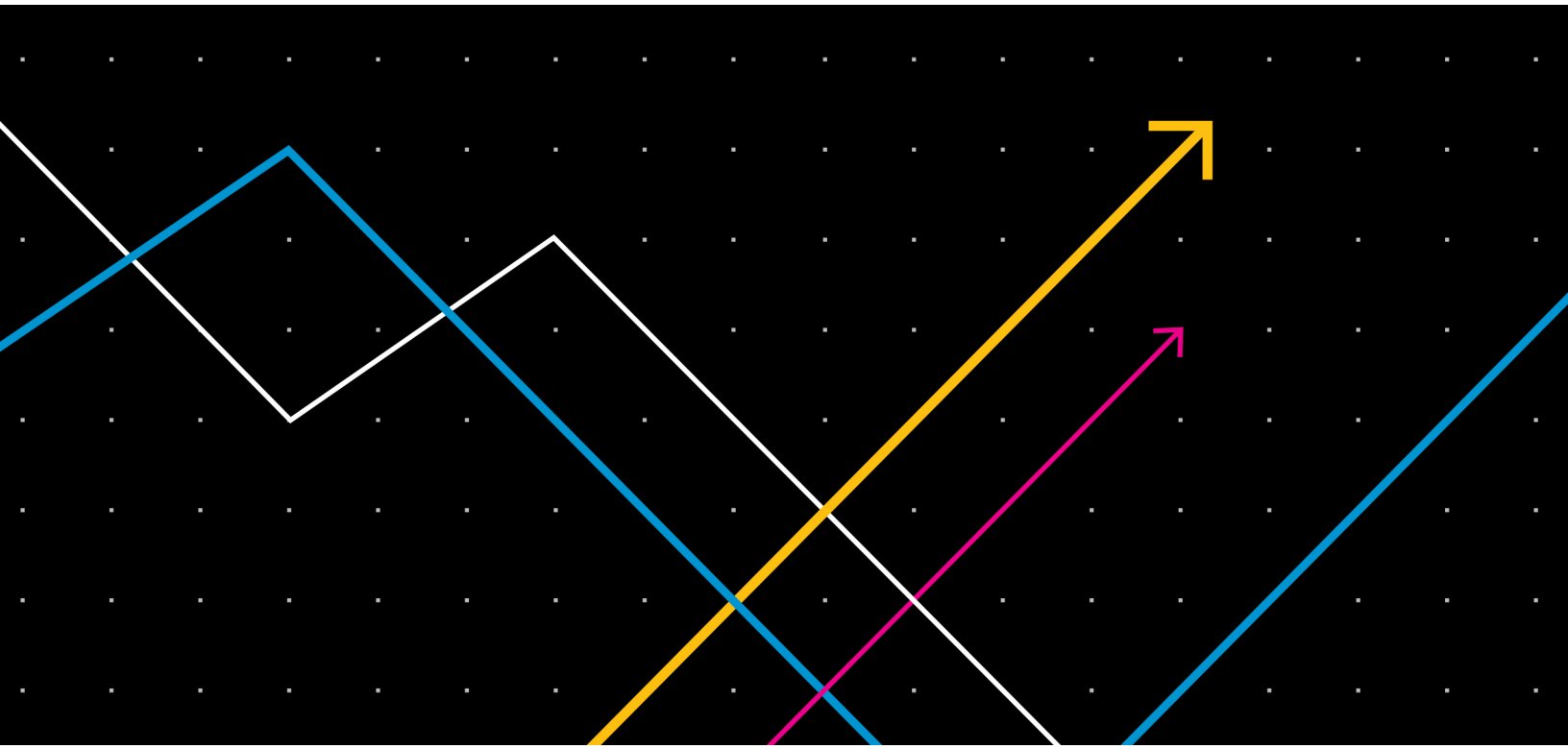


Final Spatial Dataset Codebook

86614623C00008 – Neighborhood Change Indicators

Alena Stern
URBAN INSTITUTE

Manuel Alcalá Kovalski
URBAN INSTITUTE

Claudia D. Solari
URBAN INSTITUTE

September 2024

URBAN INSTITUTE · ELEVATE · THE · DEBATE

ABOUT THE URBAN INSTITUTE

The Urban Institute is a nonprofit research organization that provides data and evidence to help advance upward mobility and equity. We are a trusted source for changemakers who seek to strengthen decision making, create inclusive economic growth, and improve the well-being of families and communities. For more than 50 years, Urban has delivered facts that inspire solutions—and this remains our charge today.

Contents

Acknowledgments

This codebook was funded by the US Department of Housing and Urban Development. We are grateful to them and to all our funders, who make it possible for Urban to advance its mission.

The views expressed are those of the authors and should not be attributed to the Urban Institute, its trustees, or its funders. Funders do not determine research findings or the insights and recommendations of Urban experts. Further information on the Urban Institute's funding principles is available at urban.org/fundingprinciples

Final Spatial Dataset Codebook

Overview of Spatial Dataset Structure

The Spatial Dataset presents the outputs of the neighborhood change predictive modeling. The dataset is at the tract-year level and contains yearly predictions of neighborhood change for 2018-2022. The dataset includes the following groups of variables:

- **Model prediction variables:** These variables present the outputs of the selected predictive model. This includes the prediction year (year that prediction is being made), change year (year of neighborhood change that is being predicted), true neighborhood change outcome, predicted neighborhood change outcome, and the predicted probabilities of each type of neighborhood change.
- **Important features:** These variables present the normalized values for the 20 most important features for the best model for the rural and urban subgroups. Some features may be most important for one model or both.
- **Neighborhood Change Definition Indicators:** These variables present the indicator variables that need to all be true for a tract to be assigned to each neighborhood change type.

Feature Data Sources

American Community Survey (ACS)

Years Covered: 2013-2022

Description: Data on a range of demographic characteristics for all census tracts nationally including race/ethnicity, education status, household median income, housing characteristics, and public benefits received.

Data Sourcing and Notes: Data were pulled using the tidycensus R package, which enables easy downloading of ACS 5-year survey data from the Census Bureau.

TIGER/Line Shapefiles

Years Covered: 2020

Description: Data on the geographic borders of census tracts. Used to identify neighboring tracts and calculate.

Data Sourcing and Notes: We obtained the shapefiles for every tract in the United States (not including US Territories) using the tigris R package, which enables easy downloading of the TIGER/Line shapefiles.

Home Mortgage Disclosure Act (HMDA)

Years Covered: 2011-2022

Description: Data on mortgage loan applications characteristics aggregated at the tract level. Includes summary indicators such as median loan amount, median borrower income, and a set of indicators describing loans made by purpose.

Data Sourcing and Notes: We obtained loan-level data cleaned by researchers at Urban's Housing Finance Policy Center and originally downloaded from the Consumer Financial Protection Bureau's website.

HUD USPS Vacancy Data

Years Covered: 2008-2023

Description: The HUD USPS data provides aggregate vacancy and no-stat counts of residential and business addresses that are collected by postal workers and submitted to HUD on a quarterly basis.

Data Sourcing and Notes: These data were provided directly to us by our HUD COR via secure file transfer protocol. The data were provided in quarterly extracts, aggregated to the 2020 census tract level.

HUD Administrative Data

Years Covered: 2012-2023

Description: The HUD Administrative data provides aggregate counts of housing choice voucher tenants and projects by census tract from HUD's administration of these programs.

Data Sourcing and Notes: These data were provided directly to us by our HUD COR via secure file transfer protocol. The data were provided in quarterly extracts, aggregated to the 2020 census tract level. In data processing, we calculated yearly averages for each 2020 tract by taking the mean of all of the quarters in the given year for each tract. We do this to address " blips" in the data where there is a great increase or decrease over previous quarters due to reporting lags. Our COR advised us that these blips are smoothed out in yearly data. Note that for 2022, data is only available for a single quarter (December) and for 2023, data is only available for June and September.

Longitudinal Employer-Household Dynamics (LEHD) Origin-Destination Employment Statistics (LODES)

Years Covered: 2013-2022

Description: LODES data provides the number of jobs and workers by income level and number of workers by race (excluding federal jobs).

Data Sourcing and Notes: Pulled using the lehdr R package which allows for users to interact with LODES data. Federal jobs were excluded from the total count because of a change in how they were recorded between 2014 and 2015 that caused an apparent change of 14 percent in the number of federal jobs reported (see more here). Additionally, select states are missing data in certain years (see table below).

Year(s)	Available States	States Without OD and WAC Data[7]	Federal Jobs	Race, Ethnicity, Education, Sex	Firm Age, Firm Size
2002	45	Arkansas, Arizona, DC, Massachusetts, Mississippi, New Hampshire	No	No	No
2003	47	Arizona, DC, Massachusetts, Mississippi	No	No	No
2004-2008	49	DC, Massachusetts	No	No	No
2009	49	DC, Massachusetts	No	Yes	No
2010	50	Massachusetts	Yes	Yes	No
2011-2016	51	(none)	Yes	Yes	Yes
2017-2018	50	Alaska	Yes	Yes	Yes
2019-2021	48	Alaska, Arkansas, Mississippi	Yes	Yes	Yes

The Institute of Museum and Library Services Public Libraries Survey

Years Covered: 2006-2022

Description: The Institute of Museum and Library Services conducts the Public Libraries Survey (PLS) annually in order to examine when, where, and how library services are changing to meet the needs of the public.

Data Sourcing and Notes: These data were downloaded from the IMLS website. The PLS data provides the latitudes and longitudes of library outlet locations. We then performed a spatial join between the library locations and the 2020 census tract polygons to determine the libraries that fall in each tract.

Comprehensive Housing Affordability Strategy (CHAS)

Years Covered: 2005-2009 to 2016-2020

Description: HUD receives custom tabulations of American Community Survey (ACS) data from the U.S. Census Bureau to produce the CHAS measures which demonstrate the extent of housing problems and housing needs, particularly for low-income households.

Data Sourcing and Notes: These data were downloaded from the CHAS website. Before the 2009-2013 data, HUD reported the data at the subtract (080) level. We first aggregated the data to the census tract level for those years. For the other years, the data are provided at the tract level.

National Register of Historic Places

Years Covered: 2008-2022

Description: The National Register of Historic Places geospatial dataset, provided by the Department of the Interior, is intended to be a comprehensive inventory of all cultural resources that are listed on the National Register of Historic Places. However, this dataset excludes all features deemed 'restricted' or 'sensitive', such as sensitive archaeological sites.

Data Sourcing and Notes: These data were downloaded from data.gov. We then performed a spatial join between the historic place location (point or polygon) and the 2020 census tract polygons. We identified a historic place as in a given tract if any of its spatial footprint fell within the census tract. Therefore, a given historic place can span multiple tracts. This dataset also provides the year each site was established. We used this information to calculate the number of historic sites in each 2020 tract in each year.

Federal Emergency Management Agency (FEMA) Disaster Declaration Summaries

Years Covered: 2009-2022

Description: The FEMA Disaster Declarations Summaries is a summarized dataset describing all federally declared disasters. This dataset lists all official FEMA Disaster Declarations from 1953 to present. It includes all three disaster declaration types: major disaster, emergency, and fire management assistance.

Data Sourcing and Notes: We pulled the data from the FEMA API using the rfema package from 2009-2022 at the county level. We then applied the county information to each tract within the county and created binary variables indicating if a severe disaster or moderate disaster occurred by tract and quarter.

Rationale for Data Sources and Variables

Neighborhood Change Factor	Relevant Datasets (Variables Measuring Factors)
Demographic composition	▪ American Community Survey (race, ethnicity, age, foreign born, languages spoken)
Income and education composition	▪ American Community Survey (household income, educational attainment, health insurance coverage)
Land and dwelling use	▪ American Community Survey (tenure, units in structure) ▪ HUD USPS vacancy data (residential and business vacancy, no-stat, and active addresses) ▪ HUD administrative data (housing choice voucher tenants and projects)
Population density	▪ American Community Survey (population) ▪ TIGER/Line shapefiles (land area)
Quality of housing stock	▪ Home Mortgage Disclosure Act (mortgage activity, home prices). Used to identify changes in race, ethnicity, and income of new homebuyers, home values, and the share of investor-owned homes. ▪ American Community Survey (median gross rent, home value, number of bedrooms, ratio income to home value) ▪ Comprehensive Housing Affordability Strategy (housing problems – when a housing unit meets at least one of the following: incomplete kitchen facilities, incomplete plumbing facilities, overcrowded, cost burdened). Used to identify changes in housing problems and housing affordability mismatch. ▪ FEMA Disaster Data. Used to capture disaster shocks that affect housing quality and prices in a given tract.
Economic Investment	▪ LEHD Origin-Destination Employment Statistics (Jobs/workers at different income points). Used to identify changes in the number (in raw and per capita terms) and income of jobs and workers and the race/ethnicity composition of workers. ▪ American Community Survey (broadband access, commute time, employment rate)

Cultural and institutional characteristics	▪ The Institute of Museum and Library Services Public Libraries Survey. Used to identify library openings and closures. ▪ National Register of Historic Places (historic preservation designation)

Process of Modeling Dataset Construction

Sourcing and Cross-walking Data

We sourced each of our main input datasets as outlined in the data sourcing descriptions above. We saved the raw datasets in our secure file storage separately for each source. For the ACS, HMDA, and CHAS data, the pre-2020 tract-level estimates are provided for the 2010 census tracts. For our model, we needed to cross-walk all of the estimates between 2010 and 2020 census tracts to have consistent time-series estimates for 2020 census tracts for modeling. To do this, we created 2010 block group to 2020 tract and 2010 tract to 2020 tract crosswalks from the NHGIS crosswalks that provide a population weighted interpolation measure to transform from 2020 census tracts to 2010 census tracts. We used the 2020-to-2010 NHGIS weights along with 2010 and 2020 tract-level population estimates from the ACS to derive the 2010-2020 weights.

Merging Data Sources

We combined the data sources into two files:

1) Yearly Aggregated Data: Includes data from the data sources (ACS, HMDA, LODES, CHAS, IMLS, NRHP) that provide estimates annually. This dataset is at the census tract – year unit of analysis, with columns for each included variable across the three datasets. Where a given data source does not have coverage for a census tract – year combination (such as the missing years for LODES data described above), the values for the relevant column(s) are stored as NA to be imputed prior to modeling. See Appendix A for the number of missing observations by variable. We include a date column that stores the quarter that an estimate was reported. We store this information as YYYY-12-01 for each year (this is done for ease of analysis though we recognize that the HUD data are better described by the last day of the reporting month).

We choose to capture each annual estimate as December as that aligns with the typical release timing of American Community Survey data.

2) Quarterly Aggregated Data: Includes data from the three data sources (HUD USPS, HUD Administrative, FEMA) that provide quarterly records. This dataset is at the census tract – quarter unit of analysis, with columns for each included variable across the two datasets. Where a given data source does not have coverage for a census tract – quarter combination (such as the HUD Administrative data not having coverage for 2022 and 2023, which are covered by the HUD USPS data), the values for the relevant column(s) are stored as NA to be imputed prior to modeling. See Appendix A for the number of missing observations by variable. We include a date column that stores the quarter that an estimate was reported. We store this information as YYYY-03-01, YYYY-06-01, YYYY-09-01, and YYYY-12-01 for the four quarters of each year (this is done for ease of analysis though we recognize that the HUD data are better described by the last day of the reporting month).

For the HUD USPS data, we produced quarterly estimates of the percentage of total residential addresses and business addresses classified as vacant and no-stat by the data. We also calculated the number and percentage of active addresses by subtracting vacant and no-stat addresses from the total number of addresses.

Data Cleaning

We performed the following data cleaning steps in creating the raw input dataset:

- We adjusted variables reported in nominal dollars to real 2022 dollars to account for inflation. This includes median household income (ACS), median income of mortgage loan borrowers (HMDA), and median mortgage loan amount (HMDA).
- The 2022 ACS shifted from reporting data for Connecticut counties to “county-equivalent” planning regions. This changed the FIPS codes used for Connecticut census tracts for the 2022 ACS. We cross-walked the 2022 ACS data for Connecticut to the 2020 Census geographies using crosswalks created by CT Data and the Census Bureau.
- We filtered the data to the tracts where USPS data are available. The USPS data are available for the 83,985 tracts where mail service is provided out of the 84,414 total 2020 census tracts. We restricted our analysis to just those tracts where the USPS data has coverage. There are three tracts that are included in the USPS data for which some of our other data

sources, including the ACS, do not have coverage: 99999000400, 99999002900, and 99999081403. All of these tracts begin with 99999 indicating they fall outside of a Metropolitan Statistical Area/Metropolitan Division .

Identifying Subgroups for Modeling

We identified two data subgroups for modeling: urban and rural tracts. We assigned each tract to one of these two subgroups using Census Urban Area classifications.[1] The census identifies urban areas as densely concentrated sets of block geographies. Any geography outside of this is considered rural. Because not all census tracts nest perfectly within urban areas, we used a cross-walk from the Missouri Census Data Center's Geocorr tool to identify the portion of the 2020 tract boundaries that was within the urban area versus the rural area using the 2022 Census urban area definitions. Tracts that were then classified as rural or urban based on if the majority of their 2020 population was in the rural or urban portion. The purpose of these subgroups is to improve our modeling results by separating tracts into smaller groups that are likely to exhibit similar types of neighborhood change and for which similar signals of change are likely to be relevant. To perform time-series cross-validation, we needed to keep the modeling subgroups consistent over time. Accordingly, we used the subgroup definitions from a single point in time to separate tracts across the full time series. We recognize that incurs a small loss of precision; for example, a tract that becomes more urbanized over time would be more similar to rural tracts in earlier years and more similar to urban tracts in later years. We offset that loss in precision by including features that capture within-subgroup variation in urbanization, such as address density. Beginning in 2010, census tracts cover the entire United States, therefore, our subgroup classifications will cover the entire country.

Outcome Generation

We focused on three types of neighborhood change: 1) Displacement Due to Price Pressures, 2) Population Loss Due to Economic Disinvestment, and 3) Inclusive Growth.

[1] To qualify as an urban area, the territory identified according to criteria must encompass at least 2,000 housing units or have a population of at least 5,000. See the Census Urban Rural documentation for more information: https://www.census.gov/programs-surveys/geography/guidance/geo-areas/urban-rural.html

We consulted extensively with our project advisory group to inform the neighborhood change types. Based on these consultations, we identified the following goals for the neighborhood change definitions:

- Providing actionable information: Project advisors identified preventing displacement of incumbent residents and businesses as a key goal of having predictions of neighborhood change. They expressed the desire to use this information to help target resources and make the case for interventions to mitigate displacement. They also identified interest in using the data to address the "missing middle" economic core in rural areas caused by outmigration of working age populations.
- Not stereotyping/stigmatizing communities: We and our project advisors felt concern that labels could serve to stigmatize communities. To avoid this, we wanted our change types to focus on the act of change (e.g., displacement/outmigration) rather than assigning a label to a community (e.g., "declining").
- Understandable definitions: We wanted our definitions of change to be easily understood by users of the data. We did not want to aim for comprehensiveness at the expense of understandability.

We measured each type of change over a 5-year period. We needed to measure change over at least 5 years to avoid comparing overlapping ACS 5-year surveys when defining change (e.g., comparing change between the 2013-2017 and 2018-2022 5-year ACS surveys). While other studies have looked at change over a longer period of time (e.g., comparing change over 10 years by comparing decennial census data), because the 5-year ACS data is available beginning in 2013, we chose to look at change over 5-year periods to allow for multiple periods of change for time-series cross-validation as outlined under the model training section below. Our neighborhood change types and definitions are as follows:

Displacement Due to Price Pressures (only measured for urban areas)

A tract is measured as experiencing displacement due to price pressures if all of the following conditions are met:

- Median household income is below county median household income in the starting year, suggesting a population at risk of displacement.
- Monthly median housing costs as a share of mean second quintile income in the start year increases by at least 10 percent, suggesting housing price pressures.

- The proportion of the tract's population using public benefits decreases, suggesting a displacement of the lowest income community members.
- The tract's median household income increases, suggesting a displacement of lower income households by higher income households.

Population Loss Due to Economic Disinvestment

A tract is measured as experiencing population loss due to economic disinvestment if all of the following conditions are met:

- The proportion of the population with a bachelor's degree or higher declines, suggesting a loss of the more highly educated population.
- Median household income declines by at least 5 percent, suggesting a decline in economic outcomes.
- A decrease in the number of households, suggesting population loss.

Inclusive Growth

A tract is measured as experiencing inclusive growth if all of the following conditions are met:

- Median household income is below the county median household income in the starting year, suggesting a sufficient low-income population for inclusive growth to be relevant.
- The inflation adjusted income for the first and second quintile of residents both increase, suggesting that income increases are shared by the lower quintiles of the income distribution.
- Monthly median housing costs as a share of mean second quintile income in the start year increases by less than 5 percent, suggesting limited price pressures.
- The number of tenant-based voucher holders does not decline during the period, suggesting that landlords are not refusing to rent to voucher holders.
- The number of households grows during the period, indicating population growth.

The three types of neighborhood change are mutually exclusive, so a tract can only experience one type of change in a given 5-year period. It is also possible that a tract did not experience any of these three types of change in the time period, which we refer to as “Change Not Measured.” This does not mean that the tract experienced *no change*, but rather that none of these three specific change definitions are met. The proportion of all tracts falling into each change type in a given year (where the year corresponds to the end year of the 5-year period of change) is provided in the table below:

Change Type	2018	2019	2020	2021	2022
Displacement Due to Price Pressures	1.2%	1.9%	2.4%	2.4%	2.2%
Population Loss Due to Disinvestment	4.2%	3.3%	4.0%	3.9%	4.2%
Inclusive Growth	2.4%	2.4%	2.2%	2.0%	2.2%

Some tracts have a NA value for a given neighborhood change type when one of the variables used in the definition cannot be measured. For example, a tract would have a value of NA for all three change types when median household income cannot be measured because a tract has zero households. In such cases, we treat the tract as "change not measured" for the purposes of modeling.

Our model predicts neighborhood change one year ahead. For example, as of 2021, it predicts what change will occur in 2022. To make that prediction, the model uses the most recent input data that would be available on the prediction date. To make a prediction in December 2021, the most recent 5-year ACS data would be the 2016-2020 data. For the purposes of this submission, the 2018-2022 5-year ACS was the most recent ACS data available. Therefore, we used the following periods of change for model training and testing:

Period of Change Predicted	Date Prediction Made	Most Recent ACS 5-year Available for Features
2017-2022	December 2021	2020
2016-2021	December 2020	2019
2015-2020	December 2019	2018
2014-2019	December 2018	2017
2013-2018	December 2017	2016

The most recent data available on a given prediction date depends on the publication lag of the data source. The publication lags for each source used in training are provided in Appendix B. Future analysis could test different prediction periods to see if performance degrades with larger windows of prediction. For example, if a user wanted to predict neighborhood change 2 years ahead, they would predict change in 2017-2022 in December 2020, making the 2014-2019 5-year ACS the most recent data available for prediction.

Feature Generation

We then used the raw input dataset to produce a variety of features, or predictor variables, for modeling. Prior to feature generation, we replaced or imputed missing values of raw variables. This is important to both create the most accurate predictions and because some model algorithms are not robust to missingness and drop observations with any missing feature values. While performing imputation, we took care to avoid data leakage and to tailor our strategy to different statistics. In particular:

- Separate Imputation of Yearly and Quarterly Data: To ensure the integrity of our data, we performed imputation separately for yearly and quarterly datasets. This strategy avoids imputing structurally missing data for quarters that are inherently absent in the yearly data.
- Separate Imputation of Training and Testing Data: We imputed missing values in training and testing datasets independently. This was a key step to prevent data leakage, ensuring that information from the testing dataset does not influence the training process, preserving the model's ability to generalize to unseen data.
- Imputation by Year-Month: We executed imputation at the granularity of the year-month level to avoid any leakage of information over time. This ensures that future data points do not inadvertently influence past imputation, maintaining the integrity of the time-series predictions.
- Imputation Strategy for Each Statistic: We applied median imputation for variables that represent ratios or shares, replaced missing counts and totals with zeros, and utilized k-nearest neighbors (KNN) imputation for medians and indexes using total population, median household income, and share of population with a bachelor's degree as predictors. Normalization was performed prior to KNN imputation to ensure all predictors had equal weight in identifying the nearest neighbor observations.

We created four main types of features:

1. Change in column: measures the change in a given variable over various periods of time prior to the prediction date. These columns can capture early changes in neighborhood change factors prior to the prediction date, which may be predictive of future change.

2. Change in neighbors: measures the change in the average value of a variable in neighboring tracts over various periods of time prior to the prediction date. We define neighboring tracts as tracts that share any contiguity with the given tract (e.g., sharing part of a border).

Changing neighborhood conditions in surrounding areas, such as changes in home prices, can be predictive of future change in a given tract.

3. Change in change in column: measures the difference in the change in a variable over two successive periods of time. For example, the difference in the change between 2015-12-01 and 2016-12-01 and between 2016-12-01 and 2017-12-01. A sharp change in the rate of change in a variable could indicate a shock or a change in neighborhood circumstances that is predictive of future change.

4. Consistency of change in column: measures whether the change in a variable over a given number of consecutive time periods is consistently negative or positive. A consistent change across multiple time periods could indicate a significant trend that is predictive of future neighborhood change.

We created features capturing change over a 3-year period for annually-reported features and over a 1-year and 3-year period for quarterly reported features. For the 3-year change features calculated using ACS data, these changes will be measured with overlapping ACS data. For example, change between the 2021-2017 5-year ACS and the 2018-2014 5-year ACS. Both surveys contain data for 2017 and 2018. Therefore, the observed level of change will likely be artificially low for all observations given the overlap. When we convert the features to z-scores as noted below, we move from looking at the levels of features to the relative values of features across observations. Because the overlap applies to all tracts, we expect that the relative differences in change across tracts will still be accurately captured in features calculated from overlapping 5-year ACS surveys. We also included features for the raw variables for the given prediction year.

Feature Engineering

To prepare the raw variables and derived features described below for modeling, we performed the following feature engineering steps:

- We dropped character variables included in some of the raw datasets such as “CBSA Name.” We converted the logical variables created by the consistency of change features to numeric. The result is a feature dataset of entirely predictor variables.
- We dropped variables with more than 10 percent of missing observations. Given the data imputation performed prior to feature engineering, this only occurred if a given variable is

entirely missing for a year or quarter. In these cases, median imputation would fail to fill NAs since no non-NA median value can be calculated.

- We dropped variables that have zero variance, or where all observations for a given year have the same value. In this case, the variable will not add predictive power because it cannot differentiate between classes.
- We dropped variables that are very highly correlated with other predictor variables, with an absolute correlation of 0.9 or greater. These variables likely don't add predictive power given their strong correlation with other predictor variables and may result in slower training or nonintuitive feature importance results.
- We transformed all-predictors to z-scores by subtracting the variable mean and dividing by the standard deviation. The interpretation of the resulting values is the standard deviations above or below the mean of the given variable. We did this so all variables are on the same scale. This is particularly important for distance-based algorithms like k-nearest neighbors where variables with bigger ranges can have more importance in the prediction simply by virtue of the larger distances between observations created by their larger scales.

At the end of this process, we had 261 (rural) and 249 (urban) of 449 raw feature variables available for modeling. The full list of features used in our final models are listed in Appendix C.

Our approach relied on the implicit feature selection performed by each of our model algorithms. For example, the penalty hyperparameter of multinomial regression determines how much feature selection is performed as part of model fitting. In a decision tree algorithm, the model determines the most effective subset of features and decision points (e.g., median household income greater than $80,000) to use for prediction, implicitly selecting the most predictive subset of features. However, if many features are irrelevant, this implicit feature selection may be insufficient to prevent model performance reduction. Future analysis could try explicit feature selection methods prior to model training to reduce the number of features included in modeling to the most relevant subset.

Neighborhood Change Modeling Methodology

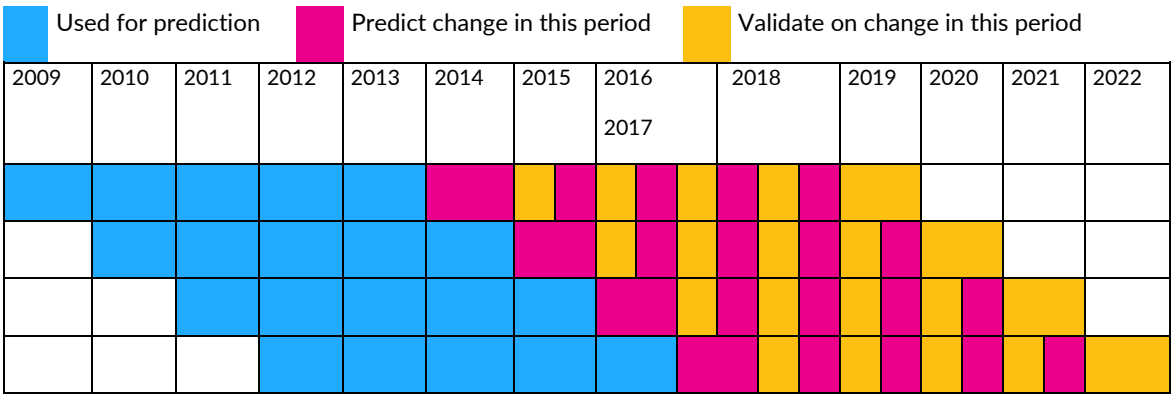
Model Training

We performed model training using the tidymodels package in the R programming language. We used the following machine learning methods to predict neighborhood change: 1) Decision Tree; 2) Random

Forest; 3) Gradient Boosted Tree; 4) Logistic Regression; and 5) K-Nearest Neighbors. We tested both a multiclass modeling approach and a binary modeling approach. Depending on the underlying structure of the data, each of these approaches can be more or less effective than the other. The multiclass approach involved fitting one set of models to predict a single outcome variable with four possible classes: displacement due to price pressures, population decline due to economic disinvestment, inclusive growth, and change not observed. As described above, the "change not observed" class indicates that one of the three neighborhood change types did not occur in the given tract. The binary modeling approach involved fitting three separate binary classification models to predict whether or not each of the change types occurred. In both modeling approaches, we trained separate models for urban and rural tracts. We ultimately determined that the two approaches offered similar results in our initial testing, so we proceeded with the multiclass approach given the computational efficiency of training one model per subgroup instead of three. However, our approach involved using the exact same set of features in the multiclass and binary case. Further analysis could try tailoring the features used in each binary case to the change outcome being predicted, which could improve results.

For each model group, we trained a variety of different combinations of machine learning algorithms, hyperparameters, and features using a process called grid search cross-validation to identify the best model. We first split the total set of data into training data and testing data. We performed this split by tract GEOID, assigning 75 percent of the tracts to the training set and 25 percent of the tracts to the testing set. We fitted and evaluated the models on the training data to select the best model, and then evaluated the best model on the testing data to estimate the model performance on unseen (i.e., out-of-sample) data.

To perform cross-validation during the training process, we further split the training data into several subsets, called folds. We used a specific type of cross-validation called time-series or rolling window cross-validation approach to partition our data (Hyndman and Athanasopoulos 2021). This approach splits the training data into multiple subsets or folds over time as shown in the table below. Each fold (represented by a row) is then split into a subset of data for training (blue and pink) and validation (yellow). For example, in fold 1, we train a model to predict change from 2014-2018 and then evaluate how well the model predicts change from 2015-2019. For each combination of algorithm, hyperparameters, and features considered, the cross-validation process trains a separate model on the training data for each fold and then calculates average model performance on the validation sets. This average validation performance is used to select the best model, which we then applied to the unseen test set to estimate out of sample error.



By creating folds using a rolling window, the average validation performance provides a more accurate estimate of how the model will perform over time when applied to new years of data. This helped us better estimate how the model will perform in practice when HUD staff use the model to generate new neighborhood change predictions each year.

We used the common preprocessing workflow described in the feature engineering section above for all models. Then we defined a model specification and hyperparameter tuning grid for each model as described below:

Model	Hyperparameter
Decision Tree	Tree Depth: The maximum number of levels in the decision tree. A deeper tree can capture more complex patterns but may also lead to overfitting.
	Minimum N: The minimum number of data points required to be in a node before the algorithm considers splitting it further. A smaller number of points can capture more complex patterns but may also lead to overfitting.
Random Forest	Trees: The number of trees in the forest. More trees can increase accuracy but with higher computational costs.
	Minimum N: The minimum number of samples required to split a node. A smaller number of points can capture more complex patterns but may also lead to overfitting.
Boosted Trees	Trees: The number of trees to fit. More trees can increase accuracy but with higher computational costs.
	Learning Rate: A number for the rate at which the boosting algorithm adapts from iteration-to-iteration.

	Loss Reduction: the minimum reduction in loss required to make a further partition on a leaf node.
K-Nearest Neighbors	Neighbors: The number of nearest neighbors to consider when making a prediction.
Multinomial Regression	Penalty: The strength of the regularization applied to the model. Higher penalty values perform stronger regularization, which can lead to some features being dropped from the model.
	Mixture: The proportion of L1 (lasso) regularization and L2 (ridge) regularization to use in the model

The specific hyperparameter values tested are outlined in the accompanying code. During model training, we calculated the following metrics on our validation sets during cross-validation:

- Accuracy: How often the model's predictions are correct overall. It measures the percentage of all predictions that are accurate.
- Recall: The proportion of actual positives that are correctly identified by the model. It is particularly useful in situations where the cost of missing a positive instance is high.
- Precision: The proportion of positive identifications that are actually correct. It is important when the cost of false positives is high.
- Receiver Operating Characteristic - Area Under the Curve (ROC-AUC): This score measures the quality of the model's predictions across all possible classification thresholds. It considers both the ability to correctly identify true positives and the rate of false positives. A higher ROC-AUC value indicates a better overall model performance, with 1 being perfect and 0.5 indicating a performance no better than random guessing.

For accuracy, recall, and precision, we used macro averaging to calculate the metric values in the multiclass models. Macro averaging reduces multiclass predictions down to multiple sets of binary predictions, calculates the corresponding metric for each of the binary cases, and then averages the results together.

For each model type, we selected the best model among all the different hyperparameters tried based on the recall metric. We chose the recall metric because it captures what proportion of tracts that truly experience change are predicted to undergo that type of change. In a policy context, we envisioned that the risk of a false negative – or failing to predict change in a neighborhood that does experience change – is the greatest risk that we wanted to optimize for avoiding.

Model Selection

After selecting the best model within each model type, we then selected the overall best model across types. To select the overall best model, we considered both the performance metrics defined above and metrics assessing fairness in model performance by the following groups:

- Year: Model performance by year in the temporal cross-validation training approach. We look at the standard errors of each metric reported across the four annual folds used in training to gauge the amount of variance in performance over time.
- Race and Ethnicity of Tract: Model performance for tracts with predominantly (> 60%) white residents, predominantly (> 60%) residents of color, or similar shares of white residents and residents of color (not falling into either previous group). We compare performance for the predominant groups.
- Rural vs. Urban Tracts: Performance for models trained on rural and urban subgroups. We take the difference between the performance of the subgroup in question and the other subgroup.
- Tenure: Model performance for tracts with predominantly (> 60%) owner-occupied units, predominantly (> 60%) renter-occupied units, or similar shares of renter-occupied and owner-occupied units (not falling into either previous group). We compare performance for the predominant groups.

The specifications and training performance metrics for the best models are as follows:

Best Model for Urban Subgroup

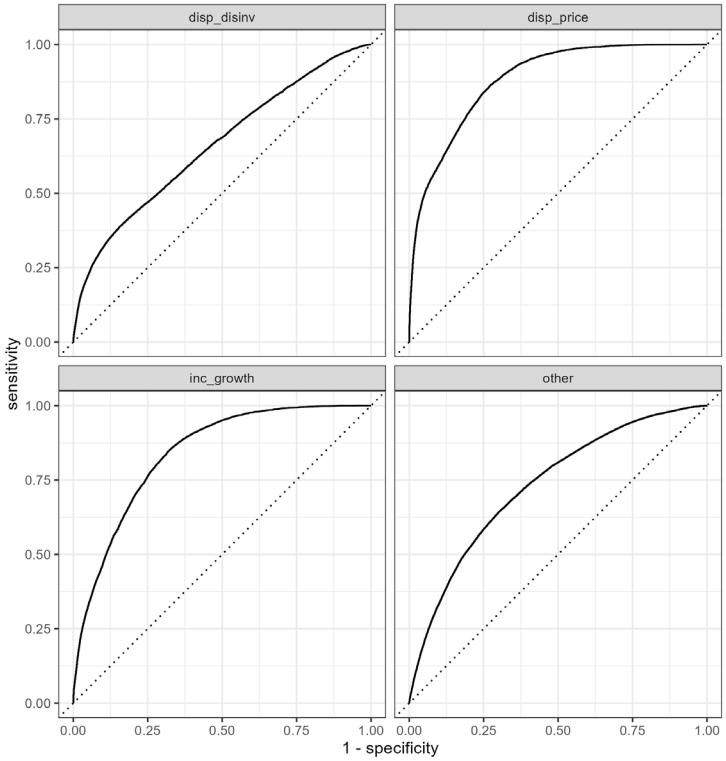
Multiclass boosted tree model with 50 trees, learning rate of 0.1, and loss-reduction of 31.6.

Class	Accuracy	Recall	Precision	ROC-AUC	Year Variation	Race/ Ethnicity	Tenure
Overall Averaged	0.75	0.49	0.34	0.80	0.0065		
Displacement Due to Price Pressures	0.45	0.45	0.25			Predominantly white: 0.39 Predominantly POC: 0.46	Predominantly Owner: 0.34 Predominantly Renter: 0.47
Population Loss Due to Disinvestment	0.21	0.21	0.25			Predominantly white: 0.20 Predominantly POC: 0.22	Predominantly Owner: 0.20 Predominantly Renter: 0.23

Inclusive Growth	0.37	0.37	0.25			Predominantly white: 0.41 Predominantly POC: 0.36	Predominantly Owner: 0.38 Predominantly Renter: 0.40
No Change Measured	0.86	0.86	0.25			Predominantly white: 0.91 Predominantly POC: 0.78	Predominantly Owner: 0.95 Predominantly Renter: 0.70

Notes: ROC-AUC is only measured for the overall multiclass results. The standard error of the accuracy metrics across years is only calculated for the overall multiclass results. The fairness metrics for race/ethnicity and tenure are only calculated for the single-class results.

ROC AUC Curves by Neighborhood Change Type



Best Model for Rural Subgroup

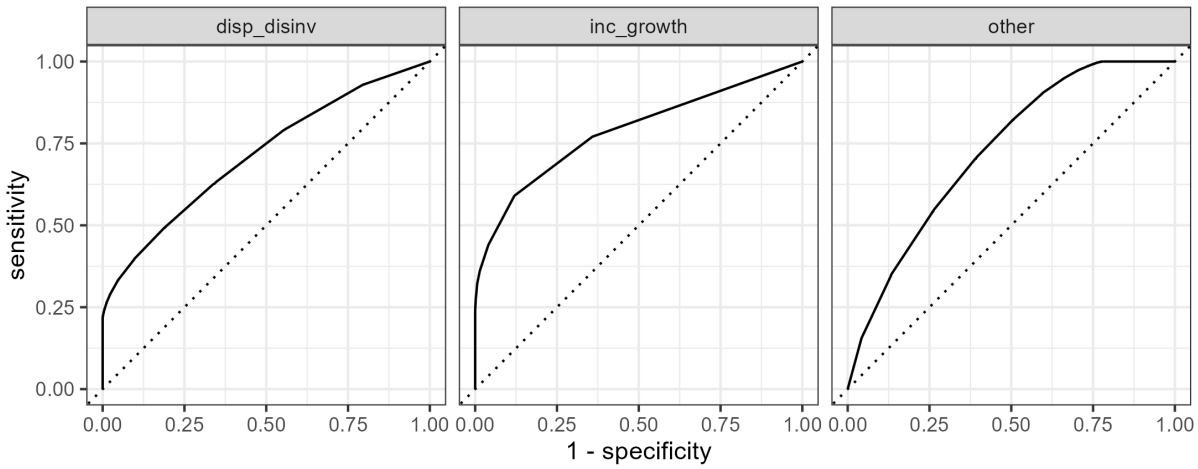
Multiclass random forest model with 25 trees and minimum N of 4.

Class	Accuracy	Recall	Precision	ROC-AUC	Year Variation	Race/ Ethnicity	Tenure

Overall Averaged	0.94	0.35	0.61	0.70	0.0028		
Population Loss Due to Disinvestment	0.22	0.22	0.33			Predominantly white: 0.20 Predominantly POC: 0.22	Predominantly Owner: 0.19 Predominantly Renter: 0.23
Inclusive Growth	0.24	0.23	0.5			Predominantly white: 0.25 Predominantly POC: 0.22	Predominantly Owner: 0.17 Predominantly Renter: 0.22
No Change Measured	1.00	1.00	0.33			Predominantly white: 1.00 Predominantly POC: 1.00	Predominantly Owner: 1.00 Predominantly Renter: 1.00

Notes: ROC-AUC is only measured for the overall multiclass results. The standard error of the accuracy metrics across years is only calculated for the overall multiclass results. The fairness metrics for race/ethnicity and tenure are only calculated for the single-class results.

ROC AUC Curves by Neighborhood Change Type



A list of the model specifications and performance metrics for the best model in each type is provided in Appendix D.

Model Evaluation

We then applied our final model to create predictions on the test data. The model results on this unseen data provide an estimate of how the model will perform on new, out-of-sample data. To maintain the validity of the testing results as an estimate of out-of-sample performance, it is critical to avoid data leakage, or using information from the testing set during the training process. It was therefore important to not review the results on the testing data until the very end of the modeling process when the final model was selected, and no further model training would be performed. We have not included the test results at this time in case further updates are required based on COR feedback. We can update to include the test results once our COR provides feedback.

Discussion

These predictions of neighborhood change are starting point for further exploration. While having the neighborhood change predictions data is valuable, those data would be more useful if users were offered guidance as to how they might make sense of those predictions and to be offered ideas of possible responses to those predictions to ensure equitable and inclusive neighborhood change to strengthen the health of communities (see recommendations for a user guide below).

Our project advisory group offered several suggestions of other indicators of neighborhood change that we were unable to act upon in this project. For example, it can be important to better single out areas planned for development that were once vacant, to better ensure that development embeds equity in its process as early as possible. While we did our best to incorporate signals of institutional shifts that can indicate community-level, and therefore neighborhood-level, change that are especially impactful in rural aeras, we were not able to incorporate all the ideas from the project advisory group. Some ideas include looking at the change in the number of churches, charter schools, hospitals or healthcare facilities, community centers, and nonprofit and service providers. None of our indicators of change reflected issues around safety, which could lead to displacement and would require more information to better tease out reasons behind the change. While this model was able to factor in some amount of environmental hazard risk, the project advisory group encouraged additional pollutants and environmental contaminants are important indicators of neighborhood change.

During the course of this project, we had ideas of several other datasets that would be worth considering to incorporate into our machine model predictions, such as: the National Center on Charitable Statistics, the National Center for Education Statistics Common Core of Data, and the

Homeland Infrastructure Foundation-Level Data. Our advisors indicated that these data sources could be particularly valuable to capture additional dimensions of community change in rural areas. Future developments of these machine learning models should consider these and other datasets that could enhance the predictions of neighborhood change.

Our initial approach focused on predicting neighborhood change one year out as we expected that making nearer-term predictions would give more accurate results. Our advisors suggested that having predictions of change further out could expand the possible policy interventions when change is predicted. Accordingly, future developments of this model should test predictions of change greater numbers of years out. Given data lags, predictions further out in time can better align with the real-time observations of community leaders and policymakers. We also recognize that conceiving of change as a binary variable limits understanding of the extent or severity of change. Future models could aim to capture change as a continuous variable and train regression models to predict the level of change. These efforts could also consider measuring neighborhood of a given tract relative to its county. Future iterations of the model might consider the indicators around the housing stock differently for owner-occupied versus renter-occupied properties. Models could further test different conceptualizations of rural, suburban, and urban areas while also outlining more sensitivities and nuances for tribal areas. One idea from the project advisory group was to offer features in the online tool for users to share feedback on inaccuracies in predictions and explanations so that we could fold that feedback into future iterations of the predictions.

As users begin to make decisions and implement interventions based on these predictions, it would be exciting to overlay those activities to see how it impacted neighborhood change. This could take the form of flagging when predictions were inaccurate and investigating if those inaccuracies are due to interventions, which could allow to meaningful case studies and evaluations to determine what works to create equitable and inclusive change.

We could also consider other uses of similar machine learning approaches, such as using these models to better estimate Fair Market Rents (FMRs). FMRs are often criticized as being too low and that voucher-holders cannot afford a unit of their voucher size in the neighborhoods of their choice. Machine learning models could help identify areas where FMRs need to change or be re-evaluated to better reflect the shifts in rental housing costs.

Limitations

The results of this project are a strong first step in developing predictions of change for neighborhoods across the country. However, we face several limitations, including but not limited to the following:

- The analysis used the ACS 5-year estimates which do not include data for individuals living in group quarters. This means that our resulting predictions may be less effective for tracts with a high proportion of individuals living in group quarters, such as college dormitories, nursing homes, prisons, etc. Future analysis could incorporate data from the ACS Group Quarters data and consider how to extend this analysis to provide meaningful data for those tracts.
- We developed our neighborhood change definitions to balance understandability and accuracy. As a result, our streamlined definitions may not capture some nuances or specific types of neighborhood change: absorbing new population without displacement in areas of high vacancy, price pressures due to a decrease in real income while rents stay static, nuances in how renters and owners may experience price pressures differently, and significant relative changes on a local level that were not captured when compared to other tracts in the county. Further analysis could explore other types of neighborhood change or seek to further incorporate these or other dimensions of change.
- Several of our input data sources have significant missingness. The LODES data is missing for several state-year combinations provided in the table above. The HMDA data had about 20% of tracts missing for 2022. And the HUD Administrative data was missing for multiple quarters of 2022 (March, June, and September) and 2023 (March and December). We addressed that missingness using thoughtful imputation approaches as described above. However, this undoubtedly loses precision compared to the true data. Future work could explore options to fill in missing data or incorporate alternate data sources.
- With additional years of data, we could include data on previous year's neighborhood change classifications and the component conditions that have to be met for a tract to be identified as experiencing each type of change during the time-series cross validation. For example, if predicting 2022 change in 2021, we could include data on the neighborhood change classifications as of 2020, the most recent ACS data available at the time of publication. We did not have sufficient years of historical prepared data to make this possible for cross-validation as of this initial analysis. We were able to add these variables when fitting the best

model on the most recent input data. It is possible that adding these variables during cross-validation could result in selecting a set of specifications that improves model performance.

- We faced several limitations as it relates to rural neighborhoods. First, we selected a definition of urban and rural areas that is commonly used in policy decisionmaking, but we could have used several other definitions of rural that might have rendered different predictions. We were unable to treat tribal areas, most of which are in or near rural areas, differently in our predictions even though our indicators may not work well in those areas. We consulted with experts in tribal areas, and determined that the indicators of neighborhood change are nuanced and complex, depending on a mix of state agreements, casino land and local organizing bodies, changing tribal designations, and how some local services may not be permitted to operate within reservations. While we were able to add some indicators of neighborhood change that we know are common in rural areas, they were limited, and did not include factors suggested by our advisors and the literature, like a change in the number of churches, healthcare facilities, charter schools, and more due to project limitations and/or limitations in nation-wide tract-level sources. Future work should take a deeper exploration into sensitivities with different rural definitions, neighborhood change in tribal areas or areas with large populations of American Indians, and whether better capturing changes in community institutions can offer more accurate predictions of neighborhood change in more rural areas.
- Given the nation-wide scope of this project, we could not incorporate many useful state or local data sources that do not have national coverage. Data sources suggested by our project advisors include building permits, AirBnB data, crime statistics, community amenities, etc. We would encourage locally-focused users of the data to supplement the neighborhood change predictions with some of these useful sources for measuring neighborhood change.
- We had limited time to test different permutations of features, models, and parameters to identify the best model. The results present our best models thus far, but we envision that with additional time the modeling results could be improved – potentially significantly so. Future research could test dividing the tracts into other subgroups for modeling, testing other model types and hyperparameters, adding additional data sources and features, testing different features for different subgroups, testing different feature selection approaches – and more- to improve the model results.
- Further research could also explore other measures of model fairness by subgroup. Other dimensions could include the median income of the tract and the housing composition of the

tract (e.g. predominantly single-family homes, mixed single and multi-family, predominantly multi-family homes).

Next Steps

Following the submission of the final spatial dataset, we encourage HUD to consider ways to better contextualize the data, including having a user guide and/or overlaying additional data sources in HUD's tool.

User Guide

We encourage HUD to consider developing a user guide. While having predictions of neighborhood change is powerful, the predictions have some limitations, and as they are based on nationwide data, the predictions are inevitably less sensitive to locally specific context and nuance. The user guide would articulate the importance of using these predictions as the start of a conversation. The predictions should trigger a set of additional questions, both exploring the accuracy of these predictions as well as ideas for next steps. Below is a list of topics to cover in a possible user guide:

- Ideas for community leaders and organizations users might consider reaching out to for deeper discussions, such as a local NNIP partner, or other local partners, such as Data Driven Detroit.
- Intervention and policy ideas relating to various aspects of community that can mitigate the potentially harmful impacts of neighborhood change. Having several ideas will allow users to explore options and determine which mitigating activities might resonate best given the local conditions and circumstances. Topics relevant to consider in the context of neighborhood change include: affordable housing, housing stability, economic inclusion, racial diversity, social capital, transportation access, effective public education, school economic diversity, preparation for college, employment opportunities, jobs paying a living wage, opportunities for income, financial security, wealth-building opportunities, environmental quality, safety from trauma, and more. And example of a list of ideas to address affordable housing includes:
    - » Increasing the overall housing supply, including by reforming zoning and land-use policies, streamlining permitting processes, and creating incentives for developers to build new housing

- Creating more dedicated affordable housing, including by subsidizing affordable housing development, establishing incentives for developers to create affordable units, and exploring ways to build affordable housing on publicly owned land
- Preserving subsidized and unsubsidized affordable housing
- Supporting permanently affordable housing models, such as community land trusts
- Creating affordable homeownership opportunities, such as by providing downpayment or closing-cost assistance and expanding access to financing, including through the use of subsidized or shared appreciation mortgages

These ideas for affordable housing initiatives and policies were pulled from websites, including Results for America and Local Housing Solutions. The user guide could link to these and other websites, such as the Upward Mobility Initiative and Opportunity Insights' Opportunity Atlas, for users to explore several evidence-based strategies and case studies that offer more ideas on equitable approaches to neighborhood change and further inform local discussions and ideas communities could consider based on the neighborhood change predictions.

- A list of literature that users could explore. This list can include examples of equitable and inclusive neighborhood change, such as DC's 11th Street Bridge Park Project (Bogel et al. 2016). It could also reference reports that center on frameworks for sustainable, inclusive, and equitable change (Mallach 2008; Greene and Pettit 2016). And it could also share literature on how to avoid negative change, such as preventing displacement in the face of change (Cohen and Pettit 2019).

This user guide could be a PDF attachment that is stored in or linked to via HUD's tool (such as in the AFFH Tool). Rather than a PDF, the user guide could be in a more interactive html format. A future development of HUD's online tool could use variables in the machine learning model to trigger an output of a suggestion for users to explore one or more topics in a list of compiled local policies and programs designed to spark ideas to mitigate the possible risks that can accompany change, ensuring more equitable, inclusive, and prosperous change. A motivation to integrate this more deeply into the tool is to address the lower visibility of an attached PDF compared to a more integrated format.

Overlaying Additional Data

Another approach to contextualize the data includes incorporating more data to overlay within HUD's online tool. The types of data that could be most appropriate for overlay include data on climate change, such as FEMA's social vulnerability index and community resilience index, data on

transportation, and other data on community assets. For example, one dataset on community assets to overlay could be the Reenvisioning Rural America data that highlights asset clusters that correspond to the over 13,000 rural census tracts. These neighborhood descriptions offer a deeper context around neighborhood assets and can add to the conversation for neighborhoods identified for change. Not only will it add more context that can inform opportunities for locally appropriate mitigating interventions to ensure equity, inclusivity, and prosperity for communities. The tool can incorporate these data on the mapping feature and highlight information on tracts that identify which of the 7 asset clusters that neighborhood falls into. The asset clusters were determined by a combination of approximately 50 asset characteristics, and the seven asset clusters are: The asset clusters were determined by a combination of approximately 50 asset characteristics, and the seven asset clusters are:

- Accessible, Energy-Rich Hubs
- High-Employment Agricultural Areas
- Centers of Wealth and Health
- Diverse, Institution-Rich Hubs
- Remote, Energy-Rich Tracts
- Diverse, Outlying Tracts
- Remote Recreational and Cultural Areas

Making Predictions on New Years of Data

We recommend generating new predictions when the new 5-year ACS survey is released. At the time of submission, the most recent ACS 2018-2022 5-year data was released in December 2023. Therefore, our latest prediction was “as-of” 2023 and use the latest years of our input data sources that are available at that time. Since our final model uses a one-year prediction window, the latest year we can predict neighborhood change is one year ahead of our “as-of” date, or 2024. When the 2019-2023 5-year ACS data is released (likely December 2024), we recommend producing updated predictions “as-of” 2024 of change to occur in 2025.

Users of the code can choose to either generate new neighborhood change predictions using the best fitted model from our model training as described above, or they can re-train the model using the new data. Re-using the pre-trained model is more expedient and allows for more direct comparison of

neighborhood change predictions over time. However, re-training the model with the new data may be more accurate, especially if a significant change in circumstances may result in the previously trained model being less predictive of future neighborhood change. Such changes in circumstances can include policy changes, data measurement changes, or significant shocks, like another pandemic. It is up to the analyst's discretion to determine the best course of action given the circumstances and constraints. The code documentation provides detail on how to generate new predictions via both approaches.

Appendix A: Missing Observations by Variable

See file missing_obs_by_var.csv included with the submission of the Final Input Dataset for the percentage of observations that are missing by raw variable included in the Final Input Dataset. The missingness also applies to all of the features derived from those variables. The denominator for calculating the percentage of missing observations is all 2020 census tracts across the years 2013 to 2022. In some cases, missingness is the result of differences in year coverage across our datasets. In other cases, there are known issues of missingness in specific geographies and years, such as the LODES data discussed above.

Appendix B: Publication Lag by Input Data Source

Data Source	Earliest Year Available	Latest Year Available	Publication Lag (Years)
ACS 5-Year	2013	2022	1
HUD USPS	2008	2023	0
HUD Administrative	2008	2023	0
HMDA	2013	2022	1
FEMA	2009	2022	1
LODES	2011	2021	2
IMLS PLS	2012	2022	1
HUD CHAS	2011	2020	3
National Register of Historic Places	2014	2022	1

This table reflects data availability as of December 2023. The HUD Administrative data is missing data for some quarters of 2022 and 2023. We take the averages of the non-missing quarters of data to produce the 2022 and 2023 estimates.

Appendix C: Feature List

See file features_used_by_subgroup.csv for the full list of features used in the rural and urban subgroup models after applying the feature engineering steps above. A feature would likely appear in one subgroup and not the other because of differences in variance, correlation with other predictors, or missingness across subgroups.

Appendix D: Model Results

Note that these results apply the default probability threshold of .5 for predicting that an observation would experience a given type of change. As discussed above, we recommend that users of these models test different probability thresholds to identify the ideal cutoff that balances precision and recall.

Results for Best Urban Models of Each Type:

Model Type	Specification	Accuracy	Precision	Recall	ROC AUC
Decision Tree	Tree Depth: 15, Minimum N: 2	0.65	0.36	0.40	0.66
Random Forest	Number of Trees: 25, Minimum N: 4	0.91	0.50	0.29	0.74
Boosted Trees	Number of Trees: 50, Learning Rate: 0.1, Loss Reduction: 31.6	0.75	0.32	0.44	0.77
Multinomial Regression	Penalty: 1	0.57	0.29	0.40	0.69

Note: We did not train a nearest neighbors model for the urban subgroup as the relative computational inefficiency of that model on large datasets made it computationally infeasible to run on our urban subgroup within the scope of this project. Future analysis may try to train this model with greater computing power that allows for greater parallelization.

Results for Best Rural Models of Each Type:

Model Type	Specification	Accuracy	Precision	Recall	ROC AUC
Decision Tree	Tree Depth: 15, Minimum N: 2	0.83	0.43	0.42	0.61
Random Forest	Number of Trees: 25, Minimum N: 4	0.94	0.57	0.35	0.69

Boosted Trees	Number of Trees: 25, Learning Rate: 0.00000316, Loss Reduction: 31.6	0.79	0.36	0.42	0.62
K-Nearest Neighbors	Neighbors: 10	0.86	0.43	0.51	0.66
Multinomial Regression	Penalty: 0.0000000001	0.67	0.36	0.47	0.66

About the Authors

Alena Stern is the chief data scientist of the Urban Institute studying policy solutions to advance equity and inclusion. She is a member of Urban's Racial Equity Analytics Lab in the Office of Race and Equity Research. Before joining Urban, she worked with AidData, the Sunlight Foundation, and the Center for Data Science and Public Policy, where she used machine learning, natural language processing, statistical analysis, and geospatial data to inform the design of government policies and international development programs.

Alena holds a BA in economics and international relations from the College of William and Mary and an MS in computational analysis and public policy from the University of Chicago.

Manuel Alcalá Kovalski is a data scientist at the Urban Institute researching policies that promote equity in transportation and housing security. Before Urban, he worked at the Brookings Institution's Hutchins Center on Fiscal & Monetary Policy where he evaluated the fiscal impact of various recovery programs. Manuel holds a BA in mathematical economics with a minor in Statistics from the University of Pennsylvania.

Claudia D. Solari is a senior research associate in the Metropolitan Housing and Communities Policy Center at the Urban Institute. She studies social inequality, with a focus on housing insecurity, homelessness, low-income housing, mixed-income housing, neighborhood inequality and segregation, fair housing, housing crowding, and poverty. Claudia is trained in quantitative and mixed-methods research, as well as survey design, evaluation, and large-scale data collection and analysis. Claudia holds a BA in sociology from Brown University and an MA and PhD in sociology from the University of California, Los Angeles.

Statement of Independence

The Urban Institute strives to meet the highest standards of integrity and quality in its research and analyses and in the evidence-based policy recommendations offered by its researchers and experts. We believe that operating consistent with the values of independence, rigor, and transparency is essential to maintaining those standards. As an organization, the Urban Institute does not take positions on issues, but it does empower and support its experts in sharing their own evidence-based views and policy recommendations that have been shaped by scholarship. Funders do not determine our research findings or the insights and recommendations of our experts. Urban scholars and experts are expected to be objective and follow the evidence wherever it may lead.

500 L'Enfant Plaza SW
Washington, DC 20024

www.urban.org