**WORKING PAPER  
June 12, 2020**

**The Urbanization Perceptions Small Area Index: An Application of Machine Learning and Small Area Estimation to Household Survey Data**

Shawn Bucholtz\*, U.S. Department of Housing and Urban Development

Emily Molfino, U.S. Census Bureau

Jed Kolko, Indeed

**Abstract**

Definitions of urban and rural are abundant in government, academic literature, and data-driven journalism. Equally abundant are debates about what is urban or rural and which factors should be used to define these terms. Absent from most of this discussion is evidence about how people perceive or describe their neighborhood. Moreover, as several housing and demographic researchers have noted, the lack of an official or unofficial definition of *suburban* obscures the stylized fact that a majority of Americans live in a suburban setting. In 2017, the U.S. Department of Housing and Urban Development added a simple question to the 2017 American Housing Survey (AHS) asking respondents to describe their neighborhood as urban, suburban, or rural. This paper illustrated how the AHS “neighborhood description” data was used to create the Urbanization Perceptions Small Area Index (UPSAI). To create the UPSAI, we first applied machine learning techniques to the AHS neighborhood description question to build a model that predicts how out-of-sample households would describe their neighborhood (urban, suburban, or rural), given regional and neighborhood characteristics. We then applied the model to the American Community Survey (ACS) aggregate tract-level regional and neighborhood measures, thereby creating a predicted likelihood the average household in a census tract would describe their neighborhood as urban, suburban, and rural. This last step is commonly referred to as small area estimation. Our approach is an example of the use of existing federal data to create innovative new data products of substantial interest to researchers and policy makers alike.

Disclaimer: The views expressed are those of the authors and not necessarily those of the U.S. Census Bureau or the U.S. Department of Housing and Urban Development.

DRB Approval Numbers CBDRB-FY19-565, CBDRB-FY20-200, CBDRB-FY20-230

\*Contact: shawn.j.bucholtz@hud.gov

**1. Introduction**

Spend a few months following major housing think tanks or prominent housing researchers on Twitter or the blogosphere and you will inevitably come across the ongoing debate over whether an area of the country is “urban”, “suburban”, and “rural.” In the social sciences, this spirited debate exists for at least three reasons. First, where a person is from is an important part of their *identity* - there is an extensive body of research in geography and other social sciences supporting this claim (Taylor S. , 2010; Massey, 2005; Cresswell, 2004; Rose, 1995). Second, where a person is from or where they currently live is an important part of their *reality*. There are clear statistical differences among Americans living in urban, suburban, and rural parts of America when it comes to voting patterns (Scala & Johnson, 2017), attitudes on social issues (Parker, et al., 2018), intergenerational mobility (Chetty, Hendren, Kline, & Saez, 2014), and health outcomes (Anderson, Saman, Lipsky, & Lutfiyya, 2015).

A third reason this debate exists is because the distinction between urban and rural matters to the Federal Government (National Academies of Sciences, Engineering, and Medicine, 2016). Several official “urbanization” definitions (definitions of urban, suburban, or rural, or similar concepts) are in use by the Federal Government (Cromartie & Bucholtz, 2008; Economic Research Service, 2019). Among the most important uses of the federal definitions is to allocate billions of dollars of federal funding to rural areas (Reamer, 2018).

Existing concepts or definitions of urbanization generally fall into three categories: administrative, land-use, and economic (Cromartie & Bucholtz, 2008). Administrative concepts are typically defined along jurisdictional boundaries such as cities or counties and are often used by government entities to determine eligibility for programs (National Academies of Sciences, Engineering, and Medicine, 2016). Land-use concepts are typically tied to some measure of the land, such as population density, impervious surface area, or tree canopy cover. Perhaps the most well-known land use measure is the U.S. Bureau of the Census’s Urbanized Areas (U.S. Census Bureau, 2010). Economic concepts are typically tied to measures of economic activity or economic relationships. A widely-used measure by the Federal government and many other entities is the U.S. Office of Budget and Management’s Metropolitan and Micropolitan Areas, which incorporate measures of economic relationships (commuting) between counties to determine if an “outlying” county is economic tied to a “core” county (U.S. Census Bureau, 2013).

More recently, a fourth concept, *perceptual*, has gained attention. The perceptual concept captures how people perceive or describe their environment. The concept is motivated by the lack of consistent definitions of urbanization in the ecology literature (Short Gianotti, Getson, Hutyra, & Kittredge, 2016).

We believe there are at least two issues with current federal urbanization definitions. First, there is little direct empirical evidence that federal urbanization definitions align with how Americans describe their neighborhood. Some federal urbanization definitions are based on what Congress has decided is or is not urban or rural, while other federal urbanization definitions are developed by experts at federal agencies, often using input received from the public. This limits the utility of the definitions.

Second, as several housing and demographic researchers have noted, the lack of an official or unofficial definition of suburban obscures the stylized fact that a majority of Americans live in a suburban setting. The Federal Government’s own housing data shows that more than half of Americans live in single-family homes that are surrounded by other single-family homes - a neighborhood description many people associate with a suburban setting.[[1]](#footnote-1)

Motivated primarily by the second issue, Jed Kolko and colleagues at Trulia conducted a first-of-its-kind national survey in 2015 asking more than 2,000 households throughout the nation to describe their neighborhood as either urban, suburban, or rural. They found that more than half of survey respondents described their neighborhood as suburban (Kolko, 2015). Using the survey data, they created a classification model and then used the model to classify each ZIP Code Tabulation Area (ZCTA) as urban, suburban, or rural. In other words, they produced the first nationwide “small area” urbanization classification product based on people’s description of their neighborhood.

Motivated by both issues, in 2017, the U.S. Department of Housing and Urban Development (HUD) added the Trulia neighborhood description question to the 2017 American Housing Survey (AHS). The AHS is the nation’s most detailed housing survey and is much larger than the Trulia survey, with a national sample size of over 55,000 household respondents, including large samples (~2,600 households) in the 15 most populated metropolitan areas (U.S. Census Bureau and U.S. Department of Housing and Urban Developemnt, 2019). Perhaps more importantly, the AHS collects more granular geographic information (i.e., address) about respondents.

Initial analysis of the AHS data conducted by Bucholtz and Kolko (2018) revealed that about 52 percent of households in the U.S. describe their neighborhood as suburban, while about 27 percent describe their neighborhood as urban and 21 percent as rural - results consistent with Kolko’s 2015 findings. Bucholtz and Kolko also showed that within the “urban” categories of the two most widely used definitions (OMB’s Metropolitan Areas and the Census Bureau Urban Areas), the majority of people describe their neighborhood as “suburban.”

Since the 2017 AHS, at least one other survey has included the neighborhood description question. The Pew Research Center conducted a survey of 5,000 households, asking each house to describe their neighborhood as urban, suburban, or rural. Analysis of the survey responses revealed that 43 percent of households in the US describe their neighborhood as suburban, while 25 percent describe their neighborhood as urban and 30 percent as rural (Igielnik, Grieco, & Castillo, 2019).

Like Kolko and colleagues at Trulia, the Pew researchers used their survey data to create a classification model and then used the model to classify ZIP Codes as urban, suburban, or rural. In other words, they produced the *second* nationwide small area urbanization classification product based on people’s description of their neighborhood.

Motivated by Trulia’s and Pew’s prior work creating small area urbanization classification products, the purpose of our project was to produce an improved nationwide small area urbanization classification product based on people’s description of their neighborhood. We call our new product the Urbanization Perceptions Small Area Index, or UPSAI.

To create UPSAI, we specified a classification model that predicts how households would describe their neighborhood, given characteristics of the region and neighborhood. We then ran the model using machine learning algorithm (random forest) which produced an *AHS-based classifier*. We then applied the classifier to the American Community Survey (ACS) tract-level aggregate regional and neighborhood measures, thereby creating a predicted likelihood that the “average” ACS household in the tract would describe their neighborhood as urban, suburban, and rural. Finally, we used the predictions to classify each census tract as urban, suburban, or rural.[[2]](#footnote-2)

We hypothesize that our UPSAI tract-level product is an improvement over both Trulia’s and Pew’s a nationwide small area urbanization classification product for two reasons. First, the AHS national sample size (55,000) is an order of magnitude larger than Pew’s sample size (5,000) and Trulia’s sample size (2,000). This larger sample size, combined with machine learning algorithms, allowed use to uncover more complex patterns in the data. Second, the AHS data includes more precise housing unit location information (exact address) compared to Trulia’s and Pew’s original survey (ZIP Code). More precise location information allows us to use more geographically precise explanatory variables (census tract) and to produce a final product that is more geographically precise (census tract) compared to Trulia and Pew (ZCTA and ZIP Code, respectively).

**2. Methodologic Approach**

The goal of this effort was to create a nationwide small area indicator of urbanization perception, whereby small geographic areas are classified as urban, suburban, or rural. To achieve our goal, we proceeded in eight steps. The first step, described in Section 3, was to determine the explanatory variables to use in the classification model. We chose 21 neighborhood-level variables and two regional-level variables.

The second step, described in Section 4, was to choose how to define “neighborhood.” This step was necessary because the concept of neighborhood enters our analysis in a few ways. We identified two options of defining neighborhood: census tracts and ZCTA’s. We elected to define neighborhoods using census tract.

The third step, described in Section 5, was to determine which classification algorithm to use. Specifically, we identify three classification algorithms (decision tree, Ada boosted decision tree, and random forest) that we believed were best suited to our specific goals and we elected to use a random forest.

The fourth step, described in Section 6, was to run our AHS-based model using the random forest classification algorithm, which produced an *AHS-based classifier*.

In the fifth step, described in Section 7, we applied our AHS-based classifier to tract-level ACS aggregate regional and neighborhood measures, thereby predicting how the average ACS household in the tract would describe their neighborhood. Lastly, we “control” the ACS tract-level predictions to ensure the national-level estimates matched AHS national-level estimates. The tract-based product is called UPSAI.

The sixth and final step, described in Section 8, was to compare the performance of the UPSAI product to the products produced by Trulia (Kolko, 2015) and Pew (Igielnik, Grieco, & Castillo, 2019). We demonstrate our product outperforms the Trulia and Pew products.

Section 9 presents our conclusion and suggestions for use of our new data product.

**3. Step 1: Choosing the Explanatory Variables**

There were four important considerations that influenced our choice of explanatory variables.**[[3]](#footnote-3)** First, and most importantly, was the three prior efforts to create small areas estimates using a neighborhood description question (Kolko, 2015; Short Gianotti, Getson, Hutyra, & Kittredge, 2016; Igielnik, Grieco, & Castillo, 2019) as well as the Census Bureau’s Urban Areas framework (U.S. Census Bureau, 2010) and OMB’s Metropolitan Areas framework (U.S. Census Bureau, 2013). Across these five separate efforts, there were more than 30 different explanatory variables, with a common thread across most or all past efforts being population or housing unit density, the absolute size of a city or town, and some measure of population movement, such as commuting.

The second consideration that influenced our choice of explanatory variables was a limitation introduced by the choice to use a small area estimates framework. For a small area estimate framework to work, the explanatory variables in the AHS-based classification model must exactly match the explanatory variables in the ACS small-areas dataset. This meant that our explanatory variables were limited to what was (1) collected in both the AHS and the ACS or (2) could be spatially linked to both the AHS and ACS.

The third consideration was whether to ignore multicollinearity. It is well-known that multicollinearity poses problems for modeling efforts where statistical inference is the goal. Multicollinearity can reduce the statistical significance of an explanatory variable. In our case, our goal is prediction, not inference, and multicollinearity does not impact predictions. As such, our final set of explanatory variables includes variables expected to be collinear, including population density with housing unit density and employee density with business density.

The final consideration was whether to include additional demographic, socioeconomic, and housing structure aggregate characteristics. There is empirical evidence supporting the theory that individual differences (e.g., age, race, class, gender) lead to difference in how people describe their neighborhood (Coulton, Korbin, Chan, & Su, 2001; Coulton C. J., 2012). Given the existing empirical evidence support individual differences, we elected to include demographic (race and age), socioeconomic (income), and housing structure characteristics (structure age, share of single-family homes, and presence of skyscrapers).

Our final set of explanatory variables included 21 neighborhood-level variables and two regional-level variables. Table 3.1 shows the 21 aggregate neighborhood features that were spatially linked to both the AHS and ACS household microdata. Table 3.2 shows the regional-level explanatory variables.

Table 3.1: Aggregate Neighborhood Explanatory Variables

|  |  |
| --- | --- |
| **Census Tract-level Explanatory Variable** | **Source** |
| 1. Population density (persons per square mile of land) | 2017 ACS 5-year |
| 2. Housing unit density (units per square mile of land) | 2017 ACS 5-year |
| 3. Average household size (persons per housing unit) | 2017 ACS 5-year |
| 4. Land area (square miles) | 2017 TIGER files[[4]](#footnote-4) |
| 5. Median household income | 2017 ACS 5-year |
| 6. Share of housing units that are single family detached | 2017 ACS 5-year |
| 7. Median year built of all housing units | 2017 ACS 5-year |
| 8. Share of commuters not using a car | 2017 ACS 5-year |
| 9. Share of population less than 18 years old | 2017 ACS 5-year |
| 10. Share of population 18-24 years old | 2017 ACS 5-year |
| 11 Share of population 25-39 years old | 2017 ACS 5-year |
| 12. Share of population 40-54 years old | 2017 ACS 5-year |
| 13. Share of population 55 years old and older | 2017 ACS 5-year |
| 14. Share of population that is non-Hispanic Black | 2017 ACS 5-year |
| 15. Share of population that is non-Hispanic Asian | 2017 ACS 5-year |
| 16. Share of population that is Hispanic | 2017 ACS 5-year |
| 17. Share of population that is all other race and ethnicity categories | 2017 ACS 5-year |
| 18. Employment density (employees per square mile of land) | 2015 LODES WAC[[5]](#footnote-5) |
| 19. Business density (businesses per square mile of land) | Dunn and Bradstreet, 2017 |
| 20. Employee density (employees per square mile of land) | Dunn and Bradstreet, 2017 |
| 21. Flag indicating presence of skyscrapers over 185 feet (Skyscrapers) | SkyScraper Center, 2017 |

Table 3.2: Regional-level features

|  |
| --- |
| **Explanatory Variable** |
| 22. Incorporated Place/Census Designated Place population (2017) |
| 23. Census Division |

**4. Step 2: Choosing the Representation of Neighborhood**

Twenty-one of our explanatory variables are aggregate neighborhood-level explanatory variables. As such, we faced a key decision about how to define neighborhood for these variables. The AHS question asks respondents to describe their neighborhood but does not define neighborhood for the respondent.

The issue of how to define neighborhood has received scholarly attention for at least 75 years (Taylor R. B., 2012). The concept of neighborhood appears in a wide range of disciplines, including criminology (Buslik, 2012); sociology (Nicotera, 2007; Coulton, Korbin, Chan, & Su, 2001); economic geography (Donaldson, 2013); and ecology (Short Gianotti, Getson, Hutyra, & Kittredge, 2016).

We considered three options for defining neighborhood. The first option was to define neighborhoods using small areas defined by the Census Bureau, and our candidates were block groups, tracts, and ZCTA’s. The second option was to use the “buffer” approach whereby we define a neighborhood as some distance (e.g., one-quarter mile, one mile) from the AHS respondent’s home (Donaldson, 2013). A third option was to use some combination of the first two.

We chose to define a neighborhood using small areas defined by the Census Bureau. In making this decision, we were influenced by three factors. First, and most importantly, we were influenced by the discussion in Hipp (2007), as summarized by Taylor (2012):

*No single layer of neighborhood is correct for research or policy purposes. Rather, the spatial scale chosen to represent a neighborhood layer should match the spatial scale of the dynamics considered from a policy or research perspective.*

We believe the spatial scale of our research topic (how people describe their neighborhoods) is larger than a hyper-local scale (urban block or suburban subdivision) but smaller than a city. At the same time, we acknowledge that neighborhood description may be impacted by the absolute size of a city, especially for smaller cities.

The second factor was computational ease and reproducibility. Our neighborhood-level explanatory variables are aggregate estimates from the 5-year ACS. These estimates are available from the Census Bureau and can easily be linked to the AHS and ACS microdata using geographic identifier codes. Creating similar estimates using the buffer approach would have meant creating one or more sets of unique spatial buffers for over 10 million AHS and ACS households, then using those buffers to create aggregate estimates. While technically feasible, such an effort would require significant computational resources.

The third influencing factor was data quality and availability for the explanatory variables. Researchers have noted that both tract-level and block group-level ACS estimates contain a large amount of sampling error (Spielman, Folch, & Nagle, 2014). Our analysis of the 2017 ACS 5-year estimates revealed that for two aggregate estimates we include as explanatory variables (tract median income and median year built), there are missing values for 1 percent of the tract-level estimates and 3 percent of the block group-level estimates. Our analysis also showed that block group-level margin of error for key estimates (tract median income and median year built) were two to three times as large as tract-level margin of error. Due to data quality and availability issues, we removed block group from consideration.

Our choices were then narrowed to Census tract and ZCTA, when then made a qualitative choice to use census tracts, as we cite two reasons for our decision. First, we believe census tract is a better representation of neighborhood than ZCTA due to their smaller geographic size – census tracts are, on average, half as large as ZCTAs. Second, we believe census tract is a better representation of neighborhood because tracts are designed to be roughly equal in size as measured by housing unit counts, whereas ZCTAs vary dramatically in housing unit counts, with ZCTAs being much larger in more densely populated areas.

**5. Step 3: Choosing the Classification Algorithm**

The general form of our classification model can be represented as:

There are several types of classification algorithms available to estimate our model. We dismissed from consideration support vector machines and neural networks as they are not well suited to handle mixed-typed input data (continuous and categorical variables), large training datasets, and multiclass problems. We also dismissed multinomial logistic regression because our model would violate the independence of irrelevant alternatives assumption.

We narrowed our classification algorithm choice to basic decision trees, Ada boosted decision trees, and random forest. We elected to use a random forest classification algorithm, which is a collection of decision trees in which observations and features are randomly selected to build multiple trees. Results are then combined at the end by averaging: each tree chooses the class for a case and the class receiving the most votes is the predicted class for that case. We the felt random forest classification algorithm would mitigate problems seen with basic and Ada boosted decision trees, such as the overfitting and noisy data.

There were two key decisions we made prior to running the classification algorithm. First, the AHS data are survey data and analysis of survey data is carried out using survey weights, which ensure the survey data is representative of the whole population being surveyed. At the time of writing, most random forest classification algorithms could not handle survey weights directly. To get around this, we simply multiplied each AHS household by a ratio of its *survey weight* divided by 500*.* For instance, an AHS household with a weight of 2,000 would be turned into 4 households with equal weight. The final AHS dataset used for the classification model had 242,600 rows with each row representing 500 housing units.[[6]](#footnote-6)

Second, the random forest classification algorithm has several parameter options (e.g., maximum depth, maximum leaf nodes, etc.). With no a priori assumptions for the values of the algorithm’s parameters, we used a hyperparameter optimization approach whereby we searched for the best parameters using 150 random iterations across a grid of parameter values. The best parameters were those resulting in the highest F1-macro score.[[7]](#footnote-7) The random forest classification algorithm was then re-run using those parameters.

**6. Step 4: Model Performance and Results**

*6.1 Technical Aspects of Modeling Performance*

To measure the performance of our classification model, we used a cross-validation approach. Specifically, we designated two-thirds (161,800) of the AHS data for training and one-third (80,800) for cross-validation testing. For the split, we stratified by the description of neighborhood. This ensured we had equal percentages of rural, suburban, and urban AHS cases in the training and testing data.

The output of the random forest classification algorithm is a complex set of rules that we refer to as the AHS-based classifier. The AHS-based classifier is then applied to both the training data and the test data such that each data set includes the original AHS household response and an AHS-based classifier-based predicted response. Both data sets are used to assess model performance.

*6.2 Model Performance Results*

Table 6.1 presents the confusion matrix on the test data. The confusion matrix shows that most of the errors were urban households misclassified as suburban households. With all classification algorithms, there will be a bias to predict the majority class (Nguyen, Bouzerdoum, & Phung, 2009). Our AHS data is imbalanced in terms of how respondents describe their neighborhood: 52 percent say suburban, 27 percent say urban, and 21 percent say rural. As such, it was not surprising to see some bias towards the majority class (suburban). Additionally, some baseline level of error is expected since the AHS can have more than one respondent per census tract, and if two respondents in the same census tract answered the neighborhood description question differently, one of them will be misclassified.

Table 6.1. Classification Model Confusion Matrix (percentage and raw)

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | | **Predicted Class** | | | **Predicted Class** | | |
| Urban | Suburban | Rural | Urban | Suburban | Rural |
| **Respondent** | Urban | 89.2% | 9.4% | 1.4% | 19,000 | 2,000 | 300 |
| Suburban | 2.9% | 94.8% | 2.3% | 1,200 | 39,500 | 950 |
| Rural | 1.4% | 5.8% | 92.8% | 250 | 1,000 | 16,000 |
| Note: Unit of analysis is AHS household. | | | | |  |  |  |

Using the results from the confusion matrix, we can calculate the values of various performance metrics. There are several performance metrics typically used to measure performance of classification model. Precision captures the rate of true and false positives. The F1-score measures how well the classification model performed by incorporating both false positives and false negatives. Accuracy is simply a measure of the number of cases the model predicts correctly in the testing data.

Table 6.2 presents the value of various performance metrics. The first items to note is that all performance metric values were high (maximum equals 1.0), suggesting that the model performed well. The second item to note relates to overfitting. Overfit model will show high precision on the training data, but lower precision on the testing data, meaning the model was fit to the training data but lack ability to predict new (testing) data. Our performance results suggest this is not an issue.[[8]](#footnote-8)

Table 6.2 Classification Model Performance Metrics

|  |  |
| --- | --- |
| **Performance metric** | **Value** |
| F1-Score on testing data | 0.881 |
| Accuracy on testing data | 0.929 |
| Precision on training data | 0.876 |
| Precision on testing data[[9]](#footnote-9) | 0.929 |

*6.2 Explanatory Variable Importance*

Table 6.3 shows the explanatory variable importance for our classification model.[[10]](#footnote-10) Incorporated Place/Census Designated Place population, population density, and housing unit density are the top explanatory variables, meaning they contributed most to classifying household resposnes. It is not surprising that absolute Place population and population density are the most important explanatory variables. These are the two primary components of the Census Bureau’s Urban Areas framework. Moreover, given the collinearity between population density and housing unit density, it can be safely assumed that density is the most important explanatory variables.

The next three most important explanatory variables (14, 12, 10) are measures of employment and business density, and they are expected to be highly collinear.

Only one household-level explanatory variable appears in the top 10 (5. household income) while other are of less importance. This implies that household-level explanatory variables do not help classify easy cases. Rather, household-level explanatory variables help classify more difficult cases and those on the border, which then improves overall model scores.

Table 6.3 Explanatory Variables Scores for Random Forest with Household-Level Characteristics

|  |  |
| --- | --- |
| **Explanatory Variables** | **Score** |
| 1. Population density | 0.130 |
| 2. Housing unit density | 0.129 |
| 22. Incorporated Place/Census Designated Place population | 0.099 |
| 4. Land Area | 0.098 |
| 5. Median household income | 0.067 |
| 19. Business density | 0.049 |
| 18. Employment density | 0.037 |
| 7. Median year built of all housing units | 0.036 |
| 8. Share of commuters not using a car | 0.036 |
| 20. Employee density | 0.034 |
| 6. Share of housing units that are single family detached | 0.032 |
| 17. Share of population that is all other race and ethnicity categories | 0.026 |
| 10. Share of population 18-24 years old | 0.025 |
| 11 Share of population 25-39 years old | 0.024 |
| 12. Share of population 40-54 years old | 0.024 |
| 16. Share of population that is Hispanic | 0.024 |
| 14. Share of population that is non-Hispanic Black | 0.024 |
| 3. Average household size | 0.023 |
| 15. Share of population that is non-Hispanic Asian | 0.023 |
| 9. Share of population less than 18 years old | 0.023 |
| 13. Share of population 55 years old and older | 0.022 |
| 23. Census Division | 0.015 |
| 21. Flag indicating presence of skyscrapers over 185 feet | 0.000 |

**7. Step 5: Creating Small Area Estimates**

With our AHS-based classifier in hand, we turned our attention to producing small area estimates. Our goal was to classify each census tract as urban, suburban, or rural based on how our model predicts the average ACS household in the tract would describe their neighborhood.

*7.1 Applying the AHS-based Classifier to the ACS Tract-level Data*

To create our final tract-level product, we first constructed a tract-based data set containing the 21 neighborhood-level characteristics and two regional-level characteristics. This was in fact the same data set that was appended to the AHS data in order to run the classification model.

Second, we applied our AHS-based classifier to the ACS tract-level data set, thereby producing a predicted probability that the average ACS household in the tract would describe their neighborhood as urban, as suburban, and as rural. Then we created an *initial* classification of each ACS tract as urban, suburban, or rural based on whichever predicted probability value was the highest, The ACS output dataset resembles table 7.1.

Table 7.1. Example of the Tract output dataset.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Tract** | **Households** | **P(urban)** | **P(suburban)** | **P(rural)** | **Initial Classification** |
| 09000100002 | 24 | 75% | 15% | 10% | Urban |
| 30405240006 | 65 | 34% | 50% | 16% | Suburban |
| 08770140002 | 19 | 17% | 24% | 59% | Rural |

*7.2 Application of Control Totals to Correct for Bias*

At this stage, we could have stopped the production process and simply published the data set as is. However, we elected to take an additional step of creating a *final* classification that was controlled to national estimates from the AHS. To understand why, recall that in section 6.2, we described a general issue with classification models estimated with unbalanced data: a tendency to bias predictions to the majority class (in our case, suburban). As shown in confusion matrix in Section 6, there was clear, albeit small, bias towards the suburban category. We were worried this could be an issue with our final product.

To determine if our concerns had merit, we performed a simple calculation: We calculated a tract-level weighted estimated (households) based on the initial classification and compared the estimates to the national weighted estimate of AHS households in each of the three categories. The results (Table 7.2) showed that our AHS-based classifier, when applied to the ACS tract-level neighborhood and regional data, exhibited classification bias towards the suburban category. In our initial tract-based classification, nearly 61 percent of households were classified as suburban, whereas only 52 percent of AHS households classified their neighborhood as suburban.

Table 7.2. Pre-controlled household level estimates of urbanization

|  |  |  |
| --- | --- | --- |
| **Class** | **Initial tract-level classification (weighted by households)** | **AHS estimate**  **(weighted by households)** |
| Urban | 21.6% | 26.7% |
| Suburban | 57.4% | 52.0% |
| Rural | 20.9% | 21.4% |

To correct for the bias introduced by the classification algorithm, we implemented a final step: we controlled the share of tracts (weighted by household) classified in each category to the weighted share of AHS households in the same category, at the national level.

To implement the control totals, we did the following four steps:

1. Classify all tracts as suburban.
2. Sort the tracts by the probability of the average household describing their neighborhood as urban in decreasing order, then classify tracts as urban until the sum of urban tracts (weighted by households) equals no more than 26.7 percent of the total households.
3. Sort the remaining tracts by the probability of the average respondent describing their neighborhood as rural in decreasing order, then classify tracts as rural until the cumulative sum of rural tracts (weighted by households ) equals no more than 21 percent of the total households.

This resulted in 26.3% of tracts (weighed by households) being classified as urban, 52.3% of tracts (weighed by households) being classified as suburban, and 21.4% of tracts (weighed by households) being classified as rural. Only tracts with at least one ACS respondent during these years have a category assignment.

Table 7.3. Post-controlled estimates of urbanization

|  |  |  |
| --- | --- | --- |
| **Class** | **AHS estimate**  **(weighted by households)** | **Final tract-based UPSAI (weighted by households)** |
| Urban | 26.9% | 26.3% |
| Suburban | 52.1% | 52.3% |
| Rural | 21.0% | 21.4% |

*7.3 Final Products*

Our final product is a UPSAI tract-level file where each tract has a final classification as urban, suburban, or rural. The final classification is controlled to AHS national estimates. However, we recognize that some users may (1) prefer to use an uncontrolled classification, or (2) prefer to create more than three categories. To accommodate these uses, our final tract-level output dataset includes the probability an average household would describe their neighborhood as urban, suburban, and rural. These probability values can be used to create an uncontrolled classification or additional categories.

Our final UPSAI files resembles following:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Census Tract** | **Households** | **UPSAI\_Urban** | **UPSAI\_Suburban** | **UPSAI\_Rural** | **UPSAI\_Cat** |
| 09000100002 | 24 | 75% | 15% | 10% | Urban |
| 30405240006 | 65 | 34% | 50% | 16% | Suburban |
| 08770140002 | 19 | 17% | 24% | 59% | Rural |

**8. Step 6: Comparing Product Performance to Existing Products**

As described in the Section 1, we predicted that our product was likely to perform better than the Trulia or Pew products because of our large sample size (55,000) and our additional geographic specificity (tract vs. ZCTA). To test our prediction, we (re)construct product-based confusion matrices for each product, then compare the matrices.

*8.1 Model-based Confusion Matrix vs. Product-based Confusion Matrix*

In Section 6.1 we presented the confusion matrix from the classification model. The confusion matrix’s values represent the AHS original responses vs. AHS predicted responses from the testing data. However, as described in section 7, there were additional steps between the classification model output (i.e., the AHS-based classifier) and the final UPSAI product. Once the UPSAI was created, it was reapplied to the original AHS data, thereby allowing the creation of a *product-based* confusion matrix comparing AHS original responses to the final UPSAI tract-level classification.

Table 8.1 and 8.2 present the model confusion matrix and the product-based confusion matrix, respectively. A comparison of the two matrices illustrates how aggregating household-level classifications to tract-level classification introduces error.

Table 8.1 Model Confusion Matrix

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **AHS Testing Data** | | **Predicted Class** | | |
| Urban | Suburban | Rural |
| **Respondent** | Urban | 89.2% | 9.4% | 1.4% |
| Suburban | 2.9% | 94.8% | 2.3% |
| Rural | 1.4% | 5.8% | 92.8% |

Table 8.2 Product-based Confusion Matrix

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **UPSAI** | | **Predicted Class** | | |
| Urban | Suburban | Rural |
| **Respondent** | Urban | 85.3% | 11.3% | 3.4% |
| Suburban | 9.1% | 85.3% | 5.7% |
| Rural | 3.1% | 12.7% | 84.1% |

*8.2 Product-based Confusion Matrix Comparison for Three Products*

Table 8.3 shows the product-based confusion matrix for UPSAI, Trulia, and Pew. As predicted, the UPSAI tract-level classification product performs significantly better as compared to the other products. The UPSAI classification performs especially well for the urban and rural categories.

Table 8.3. Product-based Confusion Matrix

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **UPSAI Tract-Based** | | | **Predicted Class** | | | | |
| Urban | | Suburban | | Rural |
| **Respondent** | Urban | 85.3% | | 11.3% | | 3.4% | |
| Suburban | 9.1% | | 85.3% | | 5.7% | |
| Rural | 3.1% | | 12.7% | | 84.1% | |

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Trulia Product** | | | **Predicted Class** | | | | | |
| Urban | | Suburban | | Rural | |
| **Respondent** | Urban | 57% | | 38% | | 6% | |
| Suburban | 8% | | 86% | | 6% | |
| Rural | 2% | | 28% | | 70% | |

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Pew Product[[11]](#footnote-11)** | | | **Predicted Class** | | | | | |
| Urban | | Suburban | | Rural | |
| **Respondent** | Urban | 56% | | 34% | | 9% | |
| Suburban | 24% | | 58% | | 17% | |
| Rural | 9% | | 22% | | 66% | |

The better performance of the tract-based UPSAI product could be due to a few factors. First, the underlying classification model for UPSAI may have performed better than Trulia’s or Pew’s underlying classification models.

A second possible explanation is that our process to create UPSAI included a step where we control tract-level classifications to AHS national-level estimates (illustrated in Table 7.2). A numerical assessment of the Trulia and Pew products suggests they did not implement controls in their classification process to ensure their household weighted ZCTA/ZIP Code classification counts matched the national estimates from their respective surveys.

A third possible explanation, and the one we believe is the biggest contributor to our better performance, is that our final product is tract-based, which introduces less aggregation error than a ZCTA or ZIP Code-based final product. In fact, we can demonstrate the improvement of having a tract-based product relative to a ZCTA-based product by replicating the process we used to create the tract-based UPSAI but modifying it to create a *ZCTA-based* UPSAI. We can then use the *ZCTA-based* UPSAI to create a *ZCTA-based* confusion matrix.

Table 8.4 shows the confusion matrix for a ZCTA-based UPSAI product. As the table demonstrates, there is more aggregation error introduced by moving from tract-based product to a ZCTA-based product. For instance, using the tract-based UPSAI product, 16 percent of the respondents who reported urban were misclassified as reporting suburban. However, if we used a ZCTA-based UPSAI product, 27 percent of the respondents who reported urban would be misclassified as suburban.

Table 8.4. Product-based Confusion Matrix for ZCTA UPSAI Product

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **UPSAI ZCTA-based** | | | **Predicted Class** | | | | |
| Urban | | Suburban | | Rural |
| **Respondent** | Urban | 65.9% | | 23.4% | | 10.7% | |
| Suburban | 18.6% | | 71.5% | | 9.9% | |
| Rural | 3.2% | | 24.1% | | 72.7% | |

*8.3 Explanatory Variable Levels for UPSAI Tracts by Predicted Class*

Table 6.3 in section 6 showed the importance scores for the explanatory variables included in the final model. These scores illustrate the extent to which each explanatory variable played a role in how households describe their neighborhood. However, the importance scores do not shed light on how levels of an individual explanatory variable vary among households describing their neighborhood as urban, suburban, or rural. Moreover, due to the complex nature of the random forest classification algorithm, it is not feasible to “back out” simple decision rules, such as “households in tracts with population density greater than X are classified as Y.”

One way to illustrate how level of individual explanatory variables vary among the different urbanization classes is to calculate the tract-level median value of each explanatory variable, by class. Table 8.5 provides this information and reveals a few interesting results. For instance, the median population density in tracts classified as urban is more than 100 times greater than the median population density is tracts classified as rural, as the same story is true when measuring business or employment density.

Table 8.5 Median Explanatory

|  |  |  |  |
| --- | --- | --- | --- |
|  | **UPSAI Predicted Class** | | |
| **Explanatory Variable** | Urban | Suburban | Rural |
| Population density | 6,059 | 2,371 | 59 |
| Housing density | 2,596 | 972 | 28 |
| Share of Hispanic | 12.8% | 7.9% | 2.7% |
| Share of Non-Hispanic Black | 10.0% | 3.7% | 0.8% |
| Share of Non-Hispanic Asian | 1.8% | 2.6% | 0.2% |
| Share of Other Race | 42.4% | 76.8% | 93.6% |
| Median income | $40,950 | $65,777 | $52,079 |
| Share 55 and over | 23.3% | 28.6% | 33.1% |
| Share 25 to 39 | 22.2% | 18.9% | 16.0% |
| Share 18 to 24 | 9.7% | 7.9% | 7.3% |
| Median Year Built | 1957 | 1977 | 1980 |
| Share of commuters not using a car | 16.2% | 9.3% | 7.6% |
| Business density | 194 | 67 | 2 |
| Employment density | 1,616 | 513 | 9 |

**9. Conclusion**

We achieved our goal of creating a national wide small area urbanization classification product based on household’s description of their neighborhood. We used confusion matrices to illustrate how our final UPSAI model performs better compared to existing products.

We believe our tract based UPSAI product has several potential uses. First, it can be used to compare how existing federal or other definitions of rural align with how people describe their neighborhoods. HUD has already published tables showing how existing federal definitions align with how people describe their neighborhood, and we hope this new product will allow others to perform similar analysis.

Second, it can be used to aggregate microdata from administrative or other survey sources into national or regional-level urban, suburban, and rural categories. Third, UPSAI can be merged to small area aggregate data (e.g., tract-level data) so analysts can produce national or regional-level aggregate statistics for urban, suburban, and rural areas.

A fourth possible use is to combine the UPSAI product with other two-category urbanization classification datasets for the purposes of adding a third category, suburban. For instance, users may consider combining UPSAI with the Census Bureau’s Urban Areas framework so urban area can be disaggregated into urban and suburban areas.

# **References**

Anderson, T. J., Saman, D. M., Lipsky, M. S., & Lutfiyya, M. N. (2015). A cross-sectional study on health differences between rural and non-rural US counties using the County Health Rankings. *BMC health services research, 15*(1), 441.

Bucholtz, S., & Kolko, J. (2018, November). America Really Is a Nation of Suburbs. *CityLab*. Retrieved from https://www.citylab.com/life/2018/11/data-most-american-neighborhoods-suburban/575602/

Buslik, M. S. (2012). Dynamic Geography: The Changing Definition of Neighborhood. *Cityscape, 14*(2), 237-242.

Chetty, R., Hendren, N., Kline, P., & Saez, E. (2014). Where is the land of Opportunity? The Geography of Intergenerational Mobility in the United States. *The Quarterly Journal of Economics, 129(4)*, 1553-1623.

Coulton, C. J. (2012). Defining Neighborhoods for Research and Policy. *Cityscape, 14*(2), 231-236.

Coulton, C. J., Korbin, J., Chan, T., & Su, M. (2001). Mapping Residents’ Perceptions of Neighborhood Boundaries: A Methodological Note. *American Journal of Community Psychology, 29*(2), 371–383.

Cresswell, T. (2004). *Place: A Short Introduction.* Oxford: Blackwell.

Cromartie, J., & Bucholtz, S. (2008, June 1). Defining the “Rural” in Rural America. *Amber Waves*.

Donaldson, K. (2013). *How Big is Your Neighborhood?* U.S. Census Bureau.

Economic Research Service. (2019). *Rural Classifications*. Retrieved from Economic Research Service: https://www.ers.usda.gov/topics/rural-economy-population/rural-classifications/

Hipp, J. (2007). Block, Tract, and Levels of Aggregation: Neighborhood Structure and Crime and Disorder as a Case in Point. *American Sociological Review, 72*(5), 659-680.

Igielnik, R., Grieco, E., & Castillo, A. (2019, Noveber 22). *Evaluating what makes a U.S. community urban, suburban or rural*. Retrieved from Pew Research Center: Decoded: https://medium.com/pew-research-center-decoded/evaluating-what-makes-a-u-s-community-urban-suburban-or-rural-159f9d082842

Jain, A. K., & Chandrasekaran, B. (1982). 39 Dimensionality and sample size considerations in pattern recognition practice. *Handbook of statistics, 2*, 835-855.

Kolko, J. (2015, May 5). How Suburban Are Big American Cities? *FiveThirtyEight*. Retrieved from FiveThiryEight: https://fivethirtyeight.com/features/how-suburban-are-big-american-cities/

Massey, D. (2005). *For Space.* London and Thousand Oaks, CA: SAGE.

Montgomery, D. (2018, Nov 20). *Congressional Density Index.* Retrieved from Citylab: https://www.citylab.com/equity/2018/11/citylab-congressional-density-index/575749/

National Academies of Sciences, Engineering, and Medicine. (2016). Rationalizing Rural Area Classifications for the Economic Research Service: A Workshop Summary. Washington, DCC: The National Academies Press.

Nguyen, G. H., Bouzerdoum, A., & Phung, S. L. (2009). Learning pattern classification tasks with imbalanced data sets. *Pattern recognition*, 193-208.

Nicotera, N. (2007). Measuring Neighborhood: A Conundrum for Human Services Researchers and Practitioners. *Americal Journal ofCommunity Psychology, 40*, 26–51.

Parker, K., Horowitz, J., Brown, A., Fry, R., Cohn, D., & Igielnik, R. (2018). *What Unites and Divides Urban, Suburban and Rural Communities.* Washington, DC.: Pew Research Center.

Raudys, S. J., & Jain, A. K. (1991). Small sample size effects in statistical pattern recognition: Recommendations for practitioners. *IEEE Transactions on Pattern Analysis & Machine Intelligence, 3*, 252-264.

Reamer, A. (2018). *Federal Funding for Rural America: The Role of the Decennial Census.* Washington, D.C.: GW Institute of Public Policy.

Rose, G. (1995). Place and Identity: A Sense of Place. In D. Massey, & P. Jess (Eds.), *A Place in the World* (pp. 87-132). Oxford: Oxford University Press.

Scala, D. J., & Johnson, K. M. (2017). Political polarization along the rural-urban continuum? The geography of the presidential vote, 2000–2016. *The ANNALS of the American Academy of Political and Social Science, 672*(1), 162-184.

Short Gianotti, A. G., Getson, J. M., Hutyra, L. R., & Kittredge, D. B. (2016). Defining urban, suburban, and rural: a method to link perceptual definitions with geospatial measures of urbanization in central and eastern Massachusetts. *Urban Ecosystems, 19*, 823–833. doi:10.1007/s11252-016-0535-3

Spielman, S. E., Folch, D., & Nagle, N. (2014). Patterns and causes of uncertainty in the American Community Survey. *Applied Geography, 46*, 147–157.

Taylor, R. B. (2012). Defining Neighborhoods in Space and Time. *Cityscape, 14*(2), 225-230.

Taylor, S. (2010). *Narratives of Identity and Place.* London: Routledge.

U.S. Census Bureau. (2010). *Urban and Rural*. Retrieved 2019, from 2010 Census Urban and Rural Classfication and Urban Area Criteria: https://www.census.gov/programs-surveys/geography/guidance/geo-areas/urban-rural/2010-urban-rural.html

U.S. Census Bureau. (2013). *Metropolitan and Micropolitan*. Retrieved 2019, from https://www.census.gov/programs-surveys/metro-micro.html

U.S. Census Bureau and U.S. Department of Housing and Urban Developemnt. (2019). *Metropolitan Area Oversample Histories: 2015 and Beyond.*

1. Author’s analysis of data from the 2017 American Housing Survey Table Creator, which shows that 56 percent of housing units are classified as single-family detached and have other single-family detached homes within one-half of a block. [↑](#footnote-ref-1)
2. Given national sample size of the 2017 AHS (55,000), it was not feasible to simply aggregate AHS household responses for small areas (e.g., census tracts). [↑](#footnote-ref-2)
3. In the classification model framework, features or attributes are individual independent variables that are inputs into a classification algorithm. Throughout this paper, we will refer to classification model features as explanatory variables for ease. [↑](#footnote-ref-3)
4. TIGER is the Topologically Integrated Geographic Encoding and Referencing GIS files created by the U.S. Census Bureau. [↑](#footnote-ref-4)
5. LODES WAC is the Longitudinal Employer-Household Dynamics (LEHD) Origin-Destination Employment Statistics - Workplace Area Characteristics data file. [↑](#footnote-ref-5)
6. An obvious question is “why not turn each survey respondent into *n* number of rows where n equals the number of housing units represented by the respondent?” Such a strategy would have resulted in more than 125 million rows, which would have posed computational issues. [↑](#footnote-ref-6)
7. We use the F1 macro-average method, which calculates the F1 score for each class and finds their unweighted mean. [↑](#footnote-ref-7)
8. We performed an additional check for overfitting. Our model was re-run using a different 67%-33% splits for the training and testing data. The same cross validation error rates and sensitivity were seen though there were minor fluctuations in the feature importance. This demonstrates that we are not overfitting the model to the specific training dataset. [↑](#footnote-ref-8)
9. We use the precision weighted average method, which calculates the precision for each class and find their weighted means. [↑](#footnote-ref-9)
10. Feature importance is the (normalized) total reduction of the criterion brought by that feature. Feature importance scores sum to approximately 1 with the most important features in the classifier having the highest score. [↑](#footnote-ref-10)
11. This was calculated using results presented in Igielnik, Grieco, & Castillo (2019). The results are not weighted as the information was not available. [↑](#footnote-ref-11)