INSTITUTE FOR HOUSING STUDIES at DePaul University

Autumn, 2024

Technical Documentation for Breaking Up the NOAH Monolith

Clustering algorithms bring together multiple variables to organize items based on their overall similarities. In the context of community-level data, clustering provides a mechanism for the comparison of geographically distinct areas that may share common characteristics. The goal of this study is to apply clustering methods to analyze rental housing markets across the top 50 metropolitan areas. By incorporating data on the rental housing stock, rental affordability, and renter demographics, the model helps to identify communities with comparable rental housing conditions and challenges.

This analysis brings together data products from the American Community Survey (ACS) 1-Year Public Use Microdata Sample (PUMS) data for the years 2012 through 2019 and 2021 as well as the Picture of Subsidized Households from the US Department of Housing and Urban Development (HUD) for years 2017 through 2019 and 2021. The project was conducted on the Metropolitan and Micropolitan Statistical Areas (CBSA) level which correspond to county boundaries. Please note that the estimates for 2020 are omitted due to data collection issues experienced during the COVID-19 pandemic.

For more about the data and definitions, see the section ‘About the Data’. For more on the specific clustering algorithm used in this project, see the section ‘Spectral Clustering Technique’.

About the Data:

American Community Survey

The American Community Survey (ACS) is a survey conducted every year by the Census Bureau based on a geographically stratified sample of about 1 percent of the United States population. PUMS data is a publicly available package of the original survey responses. PUMS data are provided to the public at the Public Use Microdata Area (PUMA) level, which is a special geography that generally contains between 40,000 and 100,000 households covering at least 100,000 people. Depending on the size of the geography that is being analyzed, ACS data are available in 1-year and 5-year varieties. To produce statistically valid estimates at smaller geographies, the Census Bureau combines multiple years of survey responses to increase the sample size. Data were extracted at the PUMA level, with the corresponding CBSA labels added via crosswalk, and finally filtered by the top 50 CBSA’s according to population.

- ***Data Timeframe -*** For this analysis, the PUMS data was processed using Python and the standard weights provided by the United States Census Bureau. All estimates are based on 1-year PUMS data. Any analyses related to ‘Housing Affordability’ are based on averages from

2012 through 2014, 2017 through 2019, and 2021 using 1-year PUMS data. The key data points include the average for 2017 through 2019 and the amount of change from the average for 2012 through 2014 to the average for 2017 through 2019. These years were chosen to highlight conditions pre-COVID-19 pandemic and to mitigate any volatility from pandemic-era data reporting. Post-analysis using data from 2021 was completed to confirm the consistency of conditions in each cluster. Income levels were calculated using household median income for each of the respective CBSAs based on, ACS 1-Year estimate Table B19013.

Housing Affordability - Housing affordability is marked to the federal poverty threshold for a four-person household in a given year. An affordable unit is a unit with gross rent less than or equal to 30 percent of the income of a household earning 150 percent of the poverty level. A household that demands affordable housing is any household with income less than or equal to 150 percent of the poverty level, or a household paying gross rent that is already affordable. A low-income renter living in an unaffordable unit is a renter earning less than 150 percent of the poverty level living in a rental unit with monthly gross rent that is more than 30 percent of the income of a household earning 150 percent of the poverty level.

Picture of Subsidized Households

Subsidized housing program datasets include HUD or predecessor programs, including Project-Based Section 8, Housing Choice Vouchers, and Public Housing. Picture of Subsidized Households data were accessed for census tracts and aggregated to the PUMA level. Following the methodology described for ACS data, CBSA labels were added via crosswalk and filtered to include the top 50 CBSAs according to population. Data was calculated as the average value of 2017 through 2019 for each HUD program.

Data Indicators

More than 60 variables were used in the clustering model. In addition to these data indicators, additional data on population-level race and ethnicity was used in post-analysis as well as data for all indicators in 2021. Data used in the clustering model included both point-in-time and change variables. For more on how data were cleaned and processed before modeling, see the section 'Data Preprocessing'. An abbreviated list of data categories follows:

- **Rental Rate** – The share of households that are renters
- **Age of Head of Household (Renter)** – The age of the head of household binned as follows – 15 to 24, 25 to 34, 35 to 44, 45 to 54, 55 to 64, 65 or older
- **Renter Household Income Level** – The number of renter households by income level binned as follows – less than 30 percent of the Area Median Income (AMI), 30 to 50 percent of AMI, 50 to 80 percent of AMI, 80 to 120 percent of AMI, 120 to 200 percent of AMI, 200 percent or more of AMI

- **Renter Cost Burden** – The number of renters paying 30 percent or more of their incomes towards housing costs
- **Head of Household Educational Attainment (Renter)** – The highest degree level of the head of household focused on two categories – the renter has no high school diploma or a renter has a bachelor's degree or higher
- **Unemployment (Renter)** – The number of households where the head of household is not employed
- **Composition of the Rental Stock** – The number of rental housing units binned as follows – 1 unit, 2 to 4 units, 5 to 49 units, or 50 or more units
- **Rental Stock Vacancy Rate** – The share of rental units that are vacant
- **Age of the Rental Housing Stock** – The age of the rental stock focused on two categories – the rental unit is in a building built before 1950 or after 2000
- **Affordable Rental Housing Supply** – The number of rental units that are affordable to a renter household earning 150 percent of the poverty level
- **Affordable Rental Housing Demand** – The number of low-income renters earning 150 percent of the poverty level as well as the number of renters earning more than 150 percent of the poverty level but who otherwise live in a housing unit with a rent that is affordable to a low-income renter
- **Lower Income Renters Living in Unaffordable Rental Units** – The number of renters earning 150 percent of poverty that live in a unit that is not affordable at 30 percent of their household's income
- **Rental Subsidy Rate** – The share of housing units that are Project Based Section 8, public housing, or connected to a Housing Choice Voucher

Data Preprocessing:

Cleaning

For this analysis, extra procedures were taken before normalization and clustering to achieve more accurate results.

The additional procedures include:

- PUMAs partially in or outside CBSA territories were removed
- PUMAs considered rural, where the 2021 population density was less than 500, were removed
- All non-numerical and categorical columns were removed before normalization

Normalization of Original Data Values

In clustering applications, a typical preprocessing step is to standardize variables so that all data are transformed to a comparable range of values. This is because variables measured at different scales will likely skew an analysis, where a variable with a larger range might outweigh variables with smaller ranges.

To correct this, numerical transformations were applied using SciKit-Learn's MinMaxScaler to scale the values to a range of (0,1). The estimator scales and shifts each feature individually to fit within the specified range, in this case between 0 and 1. MinMaxScaler is often the preferred choice for variables with clear minimum and maximum values as it preserves data distribution through linear transformation, is robust to outliers by scaling the data to a limited range, and helps reduce data bias by ensuring that all features contribute equally to the model fitting process.

Methods for Standard Errors

The Census ACS estimates are based on a sample and as a result, may be affected by high levels of sampling variability. The reliability of each ACS estimate can be analyzed using the published margin of error that is based on a 90-percent confidence level. The margin of error (MOEs) measures the variation in the random samples due to chance.

A commonly used technique to decide whether a certain ACS variable estimate is reliable employs the coefficient of variation (CV) of the sample estimate. The coefficient of variation is defined as the ratio between standard error and estimated value and measures the relative amount of variability associated with the sample estimate. Low CV values indicate more reliable estimates. In line with this criterion, only ACS estimates with CV values below 30 percent were used in this analysis. To include certain ACS variables with CV values exceeding 30 percent, IHS followed Census Bureau protocols to create a new derived variable with a reduced and acceptable margin of error. Then the CV of the aggregated estimate was computed to assess its reliability and the new aggregated variable was used in the analysis if the newly computed CV was below 30 percent.

Spectral Clustering Technique:
About the Method

This analysis employs Spectral Clustering to define clusters of PUMAs with similar characteristics. Spectral Clustering is a graph-based partitioning technique that transforms the data into a lower dimensional space using the eigenvalues of a similarity matrix, which represents pairwise relationships between data points. It identifies clusters in this transformed space by finding groups of data points that are similar based on these relationships.

In this analysis, the similarity matrix was constructed using the radian basis function (RBF) affinity measure, which calculates the pairwise distances in a non-linear manner. The RBF affinity helps capture

complex structures within the data, making it suitable for identifying clusters that may not be linearly separable.

After transforming the data using the eigenvectors of the RBF-based similarity matrix, clustering was performed in the lower-dimensional space. Spectral clustering was selected for this analysis due to its ability to uncover clusters with complex shapes and effectively handle intricacies present in the housing data. The clustering process was implemented using SciKit-Learn's SpectralClustering function. The number of clusters was optimized by evaluating the silhouette score for different configurations.

Choosing the Number of Clusters

One major challenge among clustering methodologies is the need to pre-select an appropriate number of clusters. The intended use of the final clustering results can cause additional complexity. If there are too few clusters the segmentation is coarse, and results in broad, non-specific clusters. With too many clusters, clusters are differentiated by very small differences among variables, and it becomes difficult to characterize the clusters. For spectral clustering, an effective approach to determining the optimal number of clusters is to perform the clustering multiple times, each with a different number of clusters, and evaluate the results using an internal validity metric.

One commonly used metric is the silhouette score, which measures how similar a data point is to its own cluster compared to other clusters, indicating the quality of the clustering. In this analysis, the final choice of six clusters was made based on the results, which indicated a balance between a narrow silhouette score range and a level of cluster granularity suitable for the study's objectives. Spectral clustering's ability to capture complex relationships in the data, enhanced by the use of an RBF affinity measure, made it well-suited for this approach to cluster selection.

Qualitative Testing

Clustering seeks to create useful, understandable, and insightful groupings. Considering these goals, qualitative evaluations of cluster quality are also relevant. For this study, mapping, and evaluation of geographic patterns and trends verified that the algorithms produced clusters with merit by assessing whether clusters made sense intuitively and accurately reflected the observed characteristics of the areas.

Clustering results were analyzed using several procedures:

1. Silhouette distances were computed as a quantitative assessment of cluster quality

2. Each variable was then organized by cluster enumeration and analyzed using boxplots, providing a visual representation of each cluster's values and median for deeper interpretation

3. ANOVA (Analysis of Variance) was employed to improve the interpretability of the clusters by providing statistical evidence of the differences among them, thereby strengthening the robustness and credibility of the findings

4. PUMA's and associated clusters were then mapped to determine whether the results were consistent with the observed characteristics in the region

5. Finally, the values for each variable included in the segmentation were compared among clusters to identify significant differences among clusters and to descriptively characterize each cluster

6. The results were further refined through meetings with project partners, resulting in the final housing market segmentation results

Clustering Results:

Cluster	Demand	Supply	Affordability
1	**High:** • Renter households • Renters 55 to 64 • Increasing renters over 55 • Renters earning less than 50% AMI • Unemployment • Renters with no high school diploma • Black population **Low:** • Renters 15 to 34 • Renters earning more than 50% AMI • Renters with bachelor's degree or greater	**High:** • 2 to 4 rental stock • Rental stock built before 1950 • Rental vacancy **Low:** • 5-49 and 50+ unit rental stock • Rental stock built since 2000	**High:** • Cost-burdened renters • Share of units that are "affordable" • Share of renters who are "lower-income" • Housing Choice Vouchers • Project-based section 8 • Public housing **Low:** • Share of lower-income renters in "unaffordable" units
2	**High:** • Renter Households • Renters 25 to 34 • Renters earning over 120% AMI • Increasing renters earning over 120% AMI • Renters with bachelor's degree or greater • Increase in renters with bachelor's or greater **Low:** • Renters over 45 • Renters earning less than 120% AMI • Unemployment • High school degree or less	**High:** • 5 to 49, 50+ unit rental stock • Increase in 50+ rentals • Rental stock built before 1950 • Density • Rental vacancy • Increase in stock built since 2000 **Low:** • Single-family rentals • 2 to 4 unit rentals • Rental stock built after 2000	**High:** • Share of lower-income renters in "unaffordable" units • Project-based section 8 **Low:** • Cost-burdened renters • Share of rental units that are affordable • Share of renters that are lower-income • Public Housing
3	**High:** • Renters over 45-54 years old • Renters over 200% AMI • Renters with no high school diploma • Hispanic/Latino population **Low:** • Renters 15-34	**High:** • 2 to 4 rental stock • 50+ rental stock **Low:** • Rental stock built since 2000 • Rental vacancy	**High:** • Cost-burdened renters • Share of lower-income renters in unaffordable units • Housing Choice Vouchers **Low:** • Share of rental units that are affordable • Share of renters that are lower-income • Public Housing

4	**High:** • Renters under 35 • Renters earning 30% to 80% AMI **Low:** • Renters over 55 • Renters earning over 120% AMI	**High:** • 5 to 49 unit rentals • Mobile home rentals • Vacant rentals • Rental stock built after 2000 **Low:** • 2 to 4 rentals • Rental stock built before 1950 • Density	**High:** • High share affordable rental units • Loss of affordable rental supply • Loss of lower-income renters • Increase in lower-income renters living in unaffordable units **Low:** • Public housing
5	**High:** • Renters 35 to 44 • Renters earning 80% AMI or more • Renters with bachelor's degree or greater • Increase in renters with bachelor's or greater **Low:** • Renter households • Renters under 24 and over 55 • Renters earning less than 50% AMI • Unemployment • Renters with no high school diploma	**High**: • Single-family rentals • Rental stock built after 2000 • Increase in 50+ rentals **Low:** • 2 to 4 unit rentals • Rental stock built before 1950 • Rental vacancy • Density	**High:** • Share of lower-income renters in "unaffordable" units **Low:** • Cost-burdened renters • Share of rental units that are affordable • Share of renters that are lower-income • HCVs, Project-based section 8, public housing
6	**High:** • Renters over 65 • Increase in renters over 65 • Renters 30% to 50% AMI **Low:** • Renter households • Renters under 35	**High:** • Single-family rentals • 2 to 4 rentals • Mobile home rentals **Low:** • 50+ rentals • Rental vacancy • Density	**High:** • Share affordable rental units • Share low-income rentals • Public housing **Low:** • Housing choice vouchers • Public Housing