# Exploring Gentrification in the United State through Machine Learning

## Introduction

Gentrification is the widespread emergence of middle- and upper middle-class enclaves in formerly deteriorated, inner-city neighborhoods. This phenomenon represents a marked reversal of the usual pattern of urban neighborhood change. As such, it has become a major focus of theoretical and practical concern. However, the newness of the phenomenon--most renovation has taken place since 1975-means that the causes and consequences of the gentrification process remain poorly understood. Although there is a burgeoning literature on the topic, the vast majority of this work is based on the observation of one or a few cases. The analysis of gentrification must be taken beyond the case study level if we hope to develop a thorough understanding of the phenomenon and its policy implications.

Currently, most studies start with a particular characteristics of urban communities and use methods such as econometric or spatial analysis to explore its relationship with gentrification, such as the relationship between vacant properties and neighborhood gentrification, The relationship between gentrification and neighborhood crime rates in Chicago, the determinants of gentrification is urban housing renovation, the characteristics of new green space contribute to gentrification, the role of the artistic dividend in neighborhood gentrification and so on.

However, both socio-economic spatial changes and changes in the physical environment of cities arise under the influence of a combination of factors, so the determinants of urban gentrification should be considered from multiple perspectives and in multiple ways. Recent developments in the field of machine learning offer new ways of modelling complex socio- spatial processes, allowing us to make predictions about how and where they might manifest in the future. Reades(2019)used machine learning to analyze socio-economic transition in London neighborhoods (based on 2001 and 2011 Census variables) which is used to predict those areas most likely to demonstrate ‘uplift’ or ‘decline’ by 2021. As a demonstration of the capabilities of machine learning, this paper underlines the continuing value of quantitative approaches in understanding complex urban processes such as gentrification.

Therefore, this paper would like to draw on this article to use machine learning as a tool to explore models of the impact of gentrification across counties in the United States and to compare the differences in the performance of the models as well as identify important influencing factors. After constructing the gentrification scores for each county using Principal Components Analysis (PCA), this paper will use Simple Linear Regression, Multiple Linear Regression, Extremely Random Trees (default), and Extremely Random Trees (tuned) machine learning models to test the explanatory and dependent variables of gentrification for each county in the United States.

## Calculating scores

We can use four variables to measure neighborhood status: per capita income , property sale value , the percentage of the neighborhood’s residents in the relatively ‘top’ classes, and education attainment. Finally, I selected these variables to calculate the scores: Personal Income, Percent of Bachelor Degree or More, Housing Price Index, New Private Housing Construction Start, White Population Ratio, Population Aged between 25-44, Population above Poverty Level.

To train the ML algorithm to predict neighborhood change we need to combine these four variables into a singular measure of ‘socioeconomic status. Since we are working with a long but fairly narrow data matrix, Principal Components Analysis (PCA) is an obvious choice as it will yield just four com-ponents: by taking only the first one we capture the majority of the variation in the input data using a single numeric value.

The construction of these scores necessarily entailed decisions about the re-scaling of variables since differences in magnitude could allow one dimension to dominate. Simple unit scaling is unlikely to address this problem because the existence of ‘heavy tails’ would lead to the bunching of the data at one end of the scale. Equally, since house prices and incomes are also highly skewed, the mean is unlikely to be a robust measure of centrality. Robust standardization using the median and Inter- Quartile Range (IQR) addresses both issues: it preserves outliers while producing comparable scales for the bulk of the data. In our testing, this approach yielded the most consistent performance and was applied to all score dimensions.

Based on PCA's calculations, we obtained the gentrification scores for each county in the U.S. in 2021, and from the figure we can see that counties with high levels of gentrification are clustered on the West Coast, the East Coast, and in the Great Lakes region, with some sporadic distributions. The results show a spatial pattern of high east-west and low central gentrification. At the same time, the results also show more spatial patterns of different levels of gentrification of geographic units in close proximity to each other and mixed distribution.

The threshold-based methods and K -means clustering have been proposed to tackle the challenges of the conceptualization and operationalization of gentrification identification, respectively. From the perspective of gentrification theory, a threshold-based method identifies gentrification via changes in socio-economic factors, which are simple, convenient, and theoretically sound. However, the threshold is more or less arbitrary (if logical) . As a possible remedy for the arbitrariness of thresholds, the classification approach is based on a theory of inferential statistics, mainly including supervised and unsupervised classification. The supervised classification is guided by the presence of an outcome variable (e.g., gentrified or not/level of gentrification), whereas the unsupervised classification only concerns the feature(s) (e.g., the socio-economic factors of a census tract), and does not record or address measurements (i.e., interpretations) of the outcome . For supervised gentrification identification, gathering sufficient outcome samples for training relies on a large-scale field survey, which is time-consuming and labor-intensive. Therefore, unsupervised classification is generally preferred. Although statistical methods increase the objectivity in gentrification identification, the interpretation of the results is less intuitive and more challenging, owing to the lack of a direct connection between the data pattern and the gentrification variables. An unsupervised classification algorithm cannot tell which combination of variables is necessary and/or sufficient to define “gentrification” .

By observing the Silhouette Score (an indicator used to evaluate the quality of clustering, which provides a way to measure the separation of data points and the tightness within clusters) calculated under different cluster criteria, we consider the comparability of clustering results. In the case of sex, it is divided into 5 clusters. From the clustering results, we can further confirm that the southern and northern parts of the west coast, the northeastern coast, and the southern coast of the United States are not only areas with a high degree of gentrification, but also have similar gentrification characteristics. Other geographical units with similar gentrification characteristics to these two areas are scattered in the central part.

地图

描述已自动生成

图表, 地图, 散点图

描述已自动生成

## Selecting predictor variables

In line with previous work in this area we attempted to select variables from a range of

categories including: Housing, Households, Work, Travel and Amenity. This set is far from exhaustive, and the use of more built environment and amenity features (e.g. schools) would be one obvious area for improvement; however, these nonetheless encompass the principal areas on which work on gentrification and neighborhood change have focused. We selected 46 explanatory variables.

Table 1 Explanatory Variables Category and description

|  |  |
| --- | --- |
| Category | Description |
| Economic Activities | Total number of job of all industries |
|  | Total number of jobs of information and services industries |
| Tenure | Homeownership |
|  | Owner-occupied housing unit/ renter- occupied housing unit |
| Housing | People live in Subsidized Housing |
|  | Rent/Income above 50% |
|  | Median Gross Rent |
|  | Detached housing unit percent |
|  | Housing with mortgage |
|  | Housing cost/owner income above 30% |
|  | Housing with plumbing facility |
|  | Percent of Household with Severe Problem |
|  | Percent of Household with overcrowding |
|  | Percent of old building that constructed before 1950 |
| Population | Population Density |
|  | Family Household Percent |
|  | Children Under 14 years old percent |
|  | Elderly over 65 years old percent |
|  | Total annual expense |
| Marital Status | Married Household Percent |
| Education | High school enrollment rate |
| Inequity | Gini index |
| Occupation | Work in trade industry percent |
|  | Work in information and service industry percent |
| Commute | Commute by driving percent |
|  | Work at home percent |
|  | Without vehicle percent |
|  | Percent of Long Commute Driving |
| Immigration | Immigrate from the same state |
|  | Immigrate from the other state in the US |
|  | Immigrate from the foreign |
|  | |  | | --- | | Foreign Born Population Percent | |
|  | |  | | --- | | Not citizenship Percent | |
| Vibrancy | Restaurant Index |
|  | Culture Index |
|  | Daytime Population |
| Health and Security | Total Population with health insurance percent |
|  | Total Violent and Property Crime Percent |
|  | Primary Care Physician per 1000sq |
|  | Food Environment Index |
|  | Air Pollution Particulate Matter |
|  | Presence of Drinking Water Violation (Dummy Variable) |
| Accessible to public space | Developed open space area percent |
|  | Grassland area percent |

https://github.com/xg298/4741-or-5741-project-proposal