Literature Review
(The Perceptrons)

Background and Conceptual Overview of Gentrification

Gentrification has a loose, broad definition that generally is related to neighborhood change where there is a shift in demographics from low-income to high-income populations. Studies on neighborhood change and gentrification typically focus on a more narrow definition as defined for each individual study. For example, Stern (2021) defined neighborhood change as falling into one of four categories:

- Gentrification: high displacement risk, reduced low-income population and increase in rents, home values, homeowners, residents with bachelor degrees
- Decline: decrease in total population, increase in vacant addresses and low-income households greater than median for neighborhoods in their CBSA (core-based statistical area)
- Inclusive growth: increase in both low- and high-income households and greater than median for neighborhoods in their CBSA
- Unchanging: when a neighborhood doesn't meet any of the above definitions

Similarly, Reades et al. (2019) defines gentrification as a neighborhood-level phenomenon involving not just an increase in the value of an individual property, but a simultaneous uplift in the values of comparable properties across a given neighborhood. This is similar to the definition used by Knorr (2019), who defines gentrification as a sustained period of disinvestment, followed by an influx of investment and wealthier residents that results in the displacement of existing residents. Knorr (2019) specifies that displacement is the key metric that needs to be observed to separate gentrification from other types of neighborhood change. In general, studies on gentrification appear to focus on increasing home value in correlation with a change in neighborhood to demographic.

Overview of Machine Learning Techniques

Existing studies have used a range of supervised machine learning models to predict gentrification and urban land use, including regression analysis, decision trees, and neural networks. In *Understanding urban gentrification through machine learning*, authors compared a Linear Regression model against a Multiple Linear Regression model, against a Random Forest model, against a Tuned Random Forest model, each of which performed progressively better than the last at predicting neighborhood status. Jain et al. (2021) also used random forest regression to make out-of-sample predictions. An advantage of random forest regression is that it allows us to compute feature importance, i.e., how effective the feature is at reducing uncertainty. Feature importance is calculated using the Mean Decrease Impurity (MDI) score, which represents the percentage of times that a feature is used to split a node weighted by the number of samples it splits. A second approach was decision trees. In *Measuring Neighborhood Change Using Postal and Housing Choice Voucher Data: Results from a Pilot Analysis of Four Metropolitan Areas in Washington, D.C. and Ohio*, authors used a two-step XGBoost model. The first step involved fitting a binary model that predicts the probability that a tract is gentrifying. The second step involved fitting a multi-class model to predict the neighborhood change type. The output of the first model feeds into the second model, which allows the second to gain additional insight about the gentrifying neighborhoods without adding too much complexity and overfitting. A third approach was neural networks. In *A Machine Learning-Based Method For Predicting Urban Land Use*, authors used Python to construct an ANN model based on TensorFlow, which is an open-source ML platform, as its main model framework. The ANN model had just one hidden layer, and the node numbers of the three layers are 40, 80 and 5.

Studies also used a range of unsupervised machine learning models. For instance, Jain et al. (2021) used Latent Dirichlet Allocation (LDA) and representation learning to create socioeconomic indices from Airbnb data as real-time indicators of gentrification. LDA is an unsupervised topic extraction model that uses word frequencies to group text samples into latent topical components. For representation learning, the Doc2Vec method was used. This method maps text to vectors in an n- dimensional space. The output vectors of Doc2Vec preserve semantic information about the input text; in particular, reviews that have similar word frequencies are closer in the n-dimensional vector space. In another paper by Knorr (2019), unsupervised K-means clustering was used to group typologies of neighborhood change. Home and rent values were used to capture investment/disinvestment; built environment changes were captured by % multi-unit dwelling; and income, race, education were included to capture flow of different classes between neighborhoods that could signal displacement. The output from the clustering algorithm was used to label census tracts as gentrifying or not gentrifying in a second stage of the study building a predictive model.

In addition to developing prediction models, the studies used a range of feature engineering and labeling techniques to prepare the data as inputs for the model. *In Measuring Neighborhood Change Using Postal and Housing Choice Voucher Data: Results from a Pilot Analysis of Four Metropolitan Areas in Washington, D.C. and Ohio*, to filter through 100 features on demographic and socio- economic information, authors used ANOVA (analysis of variance) to keep features with a strong correlation with the target variable, as well as Pearson correlations among features to discard similar features to reduce collinearity. After completing this feature selection process, authors annotated the training data using a rule-based approach to assign each census tract one of the four labels for the 2015 and 2016 data. The accuracy of this prediction was verified by using the time-lagged ACS data to predict the labels. In *A Machine Learning-Based Method For Predicting Urban Land Use*, authors combine points of interest and current land use data to obtain a relatively accurate land use grid in a GIS, which is then converted into a matrix and sliced. The data are finally input into the model as 1-D arrays for subsequent training, testing, and optimization processes. To label the data, authors grid the study area at a certain scale, match the land use data to each corresponding grid, calculate the proportion of each of the five categories and obtain a matrix. Here, considering the land use complexity, in order not to lose too much information, authors use the proportion data to represent the entire grid area instead of finding the maximum value to obtain a single dominant category.

Potential Data Sources

Socioeconomic indicators from official census data have been widely used in research to predict neighborhood change. Reades et al (2019) used open data from the UK Census of Population and the London Data Store as an important overarching consideration was ensuring that their research was open and replicable. In particular, a common official data source used is the American Community Survey data (Stern 2021) as it provides current social, demographic, economic and housing data of the US population, updated on a yearly basis. Other official data sources that provide important data that can potentially be used as a measure for neighborhood change include US Dept of Housing and Urban Development (HUD) Housing Choice Voucher and United States Postal Server (USPS) data (Stern 2021). However, some of these data sources might not be publicly available and thus is beyond the scope of our study. Beyond official sources, there is a diverse range of publicly available secondary data provided by external research organizations - an example being Zillow Research data used in Stern’s (2021) paper. The study used the Zillow House Value Index and Zillow Observed Rent Index as indicators of gentrification and decline by comparing the average rent and house values in the tract with that of the metro area they are in.

In recent years, there has been rapid growth in the employment of unstructured data, specifically user-generated information from various online social media (e.g. Twitter, Facebook) or commerce platforms (e.g. Zillow, Airbnb). There has been growing interest in using these alternative data sources to predict important socioeconomic outcomes due to the currency of the data and their ability to provide additional insights into human behavior and decision making. Jain et al. (2021) uses a combination of structured and unstructured data from Airbnb listings and reviews in New York City, Los Angeles, and London to nowcast gentrification and found that there was a high correlation between gentrification and Airbnb data. In fact, the study suggested that unstructured data from these review platforms can capture aspects of gentrification beyond those captured by structured data alone. A study conducted by Glaeser et al. (2018) leveraged Yelp data to quantify neighborhood change, highlighting that business listings on Yelp were highly correlated to changes in socioeconomic variables linked to gentrification including age, education, and housing. Additionally, a 2018 study conducted by Chapple et al. used Twitter data to understand mobility and visitor patterns in gentrifying neighborhoods in San Francisco, specifically whether they can help identify areas at risk of change.

Preliminary Research Plan

Identifying, reading, and synthesizing the methods and findings from a series of well-cited research papers that focus on using machine learning to make predictions about gentrification has allowed us to refine our research approach with several changes inspired by the readings. For example, inspired by Stern (2021), we have decided to broaden the focus of our research question from just making binary predictions about gentrification to measuring neighborhood change, which will allow us to make multiclass predictions about the growth, decline, and stagnation of cities. Tentatively, we are keen to explore the research question: How well does structured demographic data or a combination of structured and unstructured data predict neighborhood change in Chicago?

Regarding data sources, we will be drawing on structured data used in the research presented in our literature review and may use some unstructured data. The primary datasets that we are exploring and potentially going to use include:

- Socioeconomic Data from American Community Survey (ACS) and US Census Bureau
- USPS Vacancies Data for Residential and Business Addresses: https://www.huduser.gov/portal/datasets/usps.html
- Zillow Housing Data: https://www.zillow.com/research/data/
- Airbnb Reviews and Listings Data: http://insideairbnb.com/get-the-data

Regarding the methodology, we hope to make predictions on multiple classes of neighborhood change using a series of models, including (but not limited to) simple linear regression, multiple linear regression, random forest, boosted decision trees, and potentially neural networks, drawing on the work presented in Reades et al. (2019) and Xia and Tong (2020). It should be noted that we plan to restrict our models to only supervised learning models, as opposed to an unsupervised, as is used in Knorr (2019), given that the main focus of the class has been on supervised learning models. Nevertheless, we plan to note the potential of an unsupervised learning approach in the section on limitations.

Bibliography

Chapple, K. et al. (2021). Monitoring streets through tweets: Using user-generated geographic information to predict gentrification and displacement. *Environment and Planning B: Urban Analytics and City Science,* https://doi.org/10.1177/23998083211025309

Glaeser, E., Kim, H., & Luca, M. (2018). Nowcasting Gentrification: Using Yelp Data to Quantify Neighborhood Change. In *AEA Papers and Proceedings, Vol. 108. 77–82.*

Jain, S., Proserpio, D., Quattrone, G., & Quercia, D. (2021). Nowcasting Gentrification Using Airbnb Data. *Proceedings of the ACM on Human-Computer Interaction*. https://doi.org/10.1145/3449112

Knorr, D. (2019). Using Machine Learning to Identify and Predict Gentrification in Nashville, Tennessee. https://ir.vanderbilt.edu/bitstream/handle/1803/13285/07242019_Dknorr_Thesis_Final2.pdf?seque nce=1%26isAllowed=y

Stern, A. (2021). Measuring Neighborhood Change Using Postal and Housing Choice Voucher Data: Results from a Pilot Analysis of Four Metropolitan Areas in Washington, D.C. and Ohio. *Cityscape: A Journal of Policy Development and Research. https://www.jstor.org/stable/27039966*

Reades, J., De Souza, J., & Hubbard, P. (2018). Understanding urban gentrification through machine learning. *Urban Studies*. https://doi.org/10.1177/0042098018789054

Xia, X., Tong, Z. (2020). A Machine Learning-Based Method for Predicting Urban Land Use. *Proceedings of the 25th International Conference of the Association for Computer-Aided Architectural Design Research in Asia (CAADRIA)*.

Yesenia, A., Palafox, L. (2019). Gentrification Prediction Using Machine Learning. Mexican Conference of Artificial Intelligence.

Zuk, M., Chapple, K. (2016). Forewarned: The Use of Neighborhood Early Warning Systems for Gentrification and Displacement. *Cityscape: A Journal of Policy Development and Research.*