U.S. CITY NEIGHBOURHOOD ARCHETYPES

Kriti Agrawal, Vivek Prakash, Tanmayee Kolli, Yuan Jack Yao, Sindhu Panthangi, Nikolaos Kakonas

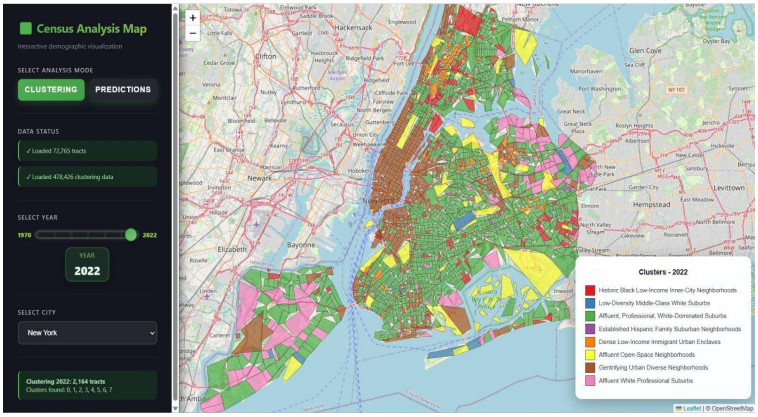

Figure 1: The interactive UI showing the clustered map of New York. There is an option to view "Clustering", which is data from the 1970s to 2022. "Predictions" shows individual features from 2025 to 2070. The year is selected from the slide bar and the city is selected from the drop-down.

INTRODUCTION

Research shows that city planners and policymakers increasingly rely on advanced modeling tools to evaluate housing, zoning, and community development decisions. Current research indicates a lack of resources to analyze these changes. There is a need to deeply understand how U.S. neighborhoods are evolving over time due to socioeconomic and demographic factors. We created our tool, Census Analysis Map, using clustering and predictive modeling to characterize city neighborhoods from 1970 until 2070.

DATA

Our raw data is a temporal dataset sourced from IPUMS NHGIS. It provides data for U.S. census tracts (identified by the unique GISJOIN code) across seven distinct time periods: 1970, 1980, 1990, 2000, 2010, 2015, and 2022. The raw dataset consisted of over 1 million rows with 1.95 GB on disk and spans several critical categories: core demographics (total population, racial composition), housing stock (age of homes, rental rates), socioeconomic status (income, poverty, education), family structures (housing units, homes with children) and commute patterns (public transit use, working from home). For preprocessing, we used crosswalk files to combine pieces of each historical tract into a single weighted 2010 aggregate.

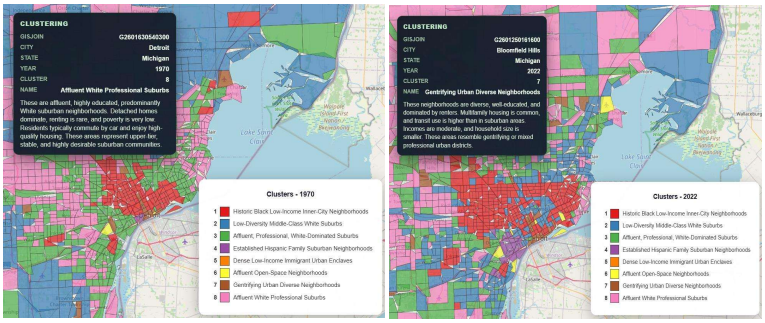

Figure 2: 1970 vs 2022 Detroit - Detroit reflects increased red presence over the years i.e., clear white flight from 1970s onwards, with affluent White residents moving to the suburbs and Black populations increasing in the urban core

EXPERIMENTS & RESULTS

Clustering Model

The clustering patterns align with known demographic and economic shifts across the USA. For instance, over many years, Las Vegas experienced a decline in Black populations with an increase in Hispanic populations. The population of lower-diversity White suburbs shrunk, and affluent White areas remained stable, consistent with the city's broad immigration trends. Detroit reflects clear white flight from 1970s onwards, with affluent White residents moving to the suburbs and Black populations increasing in the urban core. Around Central Park in NYC, increasing diversity aligns with decades of gentrification and reinvestment. In Atlanta, especially around Midtown, clusters show rising gentrification, while many outlying areas display sustained or increasing Black population over time. All these changes have been corroborated via reports and can be seen in our visualization. Unlike K-Means or PCA, SOM preserves the structure of the data on a 2D grid. Points that are similar end up closer together on the map. Additionally, it doesn't force hard cluster boundaries. It naturally shows gradual transitions between neighborhoods or demographic profiles.

Predictive Model

We ran two types of evaluations on our predictive models to determine the most suitable approach. The first involved investigating the alignment of our forecasted data with contemporary research. For instance, we observed patterns like the widening disparity between affluent and low-income neighborhoods, as well as signs of the "white flight" phenomenon, where rising poverty and increasing Black population shares in certain urban cores coincide with projected decline in White population shares. These patterns emerged without being explicitly encoded in the model, indicating that the learned relationships in the data appear realistic. The second method of evaluation involved comparing the predictions of XGBoost, LSTM, and LightGBM models. XGBoost emerged as the most robust and accurate model for neighborhood forecasting, while the poor performance of LSTM highlights the challenges of applying Deep Learning to sparse temporal Census data.

Visualization

We tested our D3 user interface for responsiveness. While we do experience some performance issues during the initial loading of the clustering and prediction data, the interface is consistently responsive once all of the data is cached. Zooming, hovering, and switching views perform smoothly thereafter. The tooltips are informative and relate information associated to cluster labels or predicted feature values, giving the user a contextual understanding about the geographic unit, from the census tract level to the national scale. The buttons, dropdowns and sliders on the UI are self-explanatory and intuitive, allowing users to seamlessly navigate between time periods and cities.

D3 is better for visualization because it gives pixel level control making it easier to zoom in on a particular census tract. It is specifically built for the web, so hover, click and animations are configurable. D3 doesn't require any heavy software installations and can stream visuals efficiently – something that GUI-based tools struggle with unless the data is pre-aggregated.

APPROACHES

Clustering Model

We used a Self-Organizing Map (SOM) and K-Means to define stable neighborhood typologies. The SOM is an unsupervised neural network approach that projects high dimensional data, such as our 21 socioeconomic features, onto a lower dimensional grid for easier interpretation. It creates nodes where each is a neighborhood pattern, and neighborhoods with similar patterns are mapped closer together. K-Means then macro-clusters these SOM nodes into cohesive groups that are labeled by the team. SOM is novelly used in conjunction with K-Means across all years as it can cluster high dimensional, non-linearly correlated features, keeping neighborhood relationships intact.

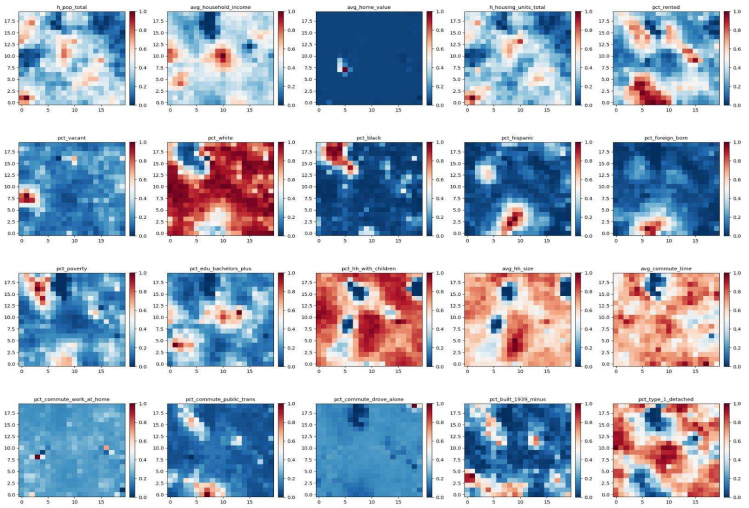

Figure 3: Results from SOM clustering. Every graph represents clusters for each feature, with red areas representing higher values and blue representing lower values. Squares in the graph are nodes that represents neighborhood patterns.

Predictive Model

Prediction involves forecasting neighborhood feature values up to 2070 using a supervised XGBoost model with a First-Order Markov Process, where the next state only depends on the immediately previous state. We used tabular, time-indexed data where every unique GISJOIN is listed 7 times, once for each of the 7 time periods, along with values for all 21 features. XGBoost was chosen as it is optimized for vectorized processes and saves computational time; this means it can make thousands of predictions at once. After the model learns the relationship between historical data from 1970 to 2022, it then predicts values from 2030 to 2070. In this process, the prediction for a time period relies exclusively on the features of the prior period, where data is forecasted for every future year by batch. The input is the current time period's features and lag 1 values, and the output are values for the next time period. The novelty in our implementation comes from the use of a boosting model to predict demographic data over future time periods, something that hasn't been attempted before.

Visualization

The user interface was created using D3 and is composed of two layers - Archetype Layer and Prediction Layer. Both these layers have an Interactive Pane. The Archetype Layer showcases the clustered map of the entire country as well as different cities, which are chosen based on the maximum changes over time. The user can visualize the distribution of the clusters via color as well as by hovering over each cluster to get detailed information. The Prediction Layer displays feature information for a selected city under the "Predictions" tab. Here, users can visualize distributions of individual features over decades. Users can choose between "Clustering" and "Prediction" via the Interactive Pane. The user interface effectively communicates neighborhood clusters and prediction data, enabling users to explore urban phenomena.

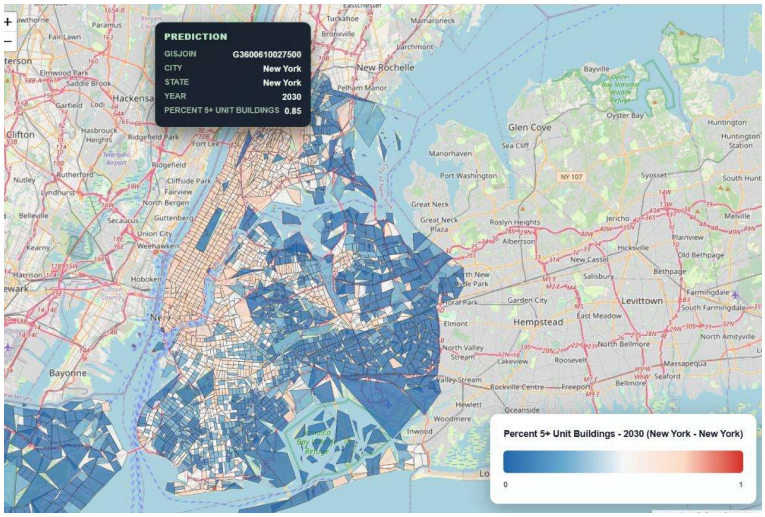

Figure 4: Map of NYC predicting the percent of 5+ Unit Buildings in 2030. This figure shows that housing around Central Park is predicted to have larger amount of 5+ Unit Buildings, compared to the areas further away