GENTRIFICATION: IDENTIFICATION AND CAUSATION

Stanford University

Matt Mistele, Kenneth Xu, Soraya Karimi, Ramin Ahmari | (mmistele, kenxu95, skarimi, rahmari)@stanford.edu

Abstract

With gentrification becoming an increasing burden for disadvantaged communities all across the US, our algorithm aims to estimate how susceptible to gentrification a given area is using a satellite image of that area.

An SVM and a logistic regression model were used to predict an 9-factor version of the gentrification susceptibility index established by Chapple et al. (2009)(1) and extended by our team for this project.

Data

We obtained 6479 ZIP codes of major US cities using Social Explorer. We also obtained American Community Survey 2010-2014 5-year estimates for 7 of the 19 Chapple et al. gentrification likelihood factors for each ZIP code. 700 Satellite color images were obtained via the DigitalGlobe Recent Imagery API. Each was labeled with the number of susceptibility factors where it was closer to the mean factor value of gentrifying areas.

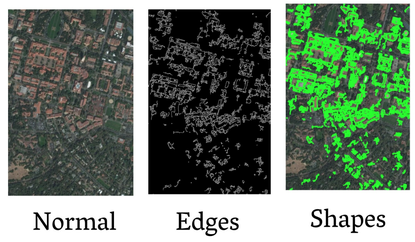
Features & Extraction

7 features from Chapple et al. (2009)(1) for our calculation of the gentrification index label: % Housing Units (5+ Units), % Renter-Occupied Housing, % workers use public transport, Median Gross Rent, Non-Family Households, renters paying >35% of income, Income Diversity.

Extended by 2 features: Population Density, Gini Index. We binarized their values to 1 (above their means) and 0 (otherwise).

Satellite Image Features: Edges (CED), frequency of distinct shapes (shape detection), number of corners (Corner Extractor), color histogram (RGB value bucketing), % black-white via texture filtering, % green (Color Extraction).

Edges, shapes, corners and texture filtering features were extracted as they represent geometric architecture of the respective zip codes while RGB values and percentage of green space indicate how urbanized a certain area is.

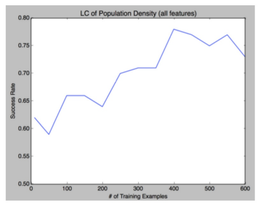


Models

We compared learning curves of a logistic regression model and a support vector machine on our data and observed that the SVM learned less effectively. Thus we chose logistic regression, ftting the sigmoid function on our data:

$$y = \text{sigmoid}(X\beta - \text{offset}) + \epsilon = \frac{1}{1 + \exp(-X\beta + \text{offset})} + \epsilon$$

Our logistic regression utilizes a coordinate descent algorithm.



Logistic Regression Learning Curve

Results

600 images for training set
100 images for testing set
Predictions for all 9 Factors and the Overall Index

FEATURES	TRAINING ERROR	TEST ERROR
% Housing Units (5+ Units)	31.67%	38.00%
% Renter-Occupied Housing	28.00%	33.00%
% Workers taking public transport	27.16%	32.00%
Median Gross Rent	28.16%	44.00%
%Non-Family Households	29.50%	37.00%
% Renters paying >35% of income	31.16%	43.00%
Income Diversity	28.67%	34.00%
Gini Index	30.67%	38.00%
Population Density	26.30%	22.00%
Overall Gentrification Susceptibility	51.16%	76.00%

Discussion & Next Steps

Our model yielded promising error rates on some of the features and did simply better than random on the others. This can be explained retrospectively: Making a connection between, for example, Median Gross Rent and a simple satellite image would obviously yield less accurate results than trying to predict the number of housing units as that is a feature inherent to the images. Furthermore, our overall direct prediction of the gentrification index yielded very inaccurate results.

The overall results can be improved if we do not directly predict the overall index, but build it up from the individual features we predicted. This would be our next future step. Further, we would try to stabilize our error rates across all features. The prediction of high error rate features could be improved by a further analysis of related metadata outside of the image. In general, we ca create more features both within the image and from outside sources. We could also try to find access to data relating to the other 12 features mentioned in Chapple et al. (2009)(1) and also include those to make our model more realistic. This would, furthermore, require finding more data to accurately predict those new features then as well as we would once again want to focus on having accurate predictions across all features.

References

[1] Karen Chapple, "Predicting Susceptibility to Gentrification in the Bay Area," in Mapping Susceptibility to Gentrification: The Early Warning Toolkit, Berkley, CA, University of California | Center for Community Innovation, August 2009, p.6