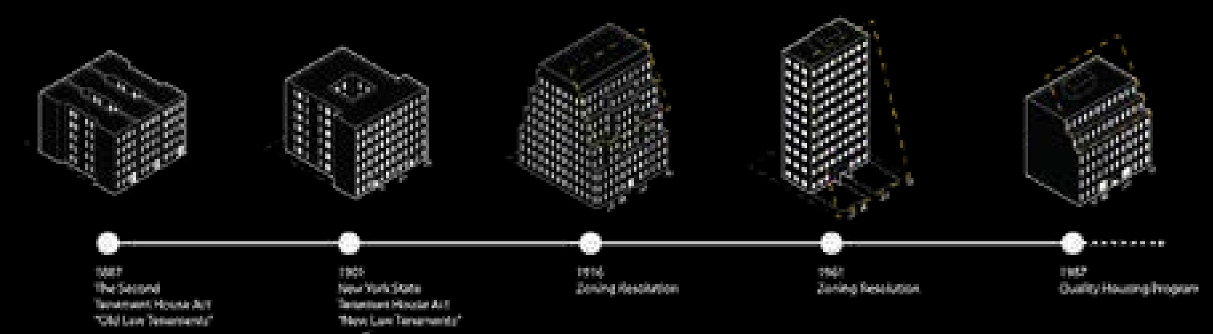
Investigating NYC buildings

Columbia GSAPP
Exploring Urban Data with Machine Learning
Spring 2021
Will Cao, Sori Han, and Jin Hong Kim

TABLE OF CONTENTS



Identifying 20th Century Brooklyn Housing Typologies
through Clustering methods

Introduction

Background

“

The way buildings in New York City look and feel today is a direct reflection of the implemented policies.

The New Law Tenement Act of 1901, subsequent Zoning Resolution amendments, and more recent housing requirements are policies that have direct consequences in the overall form and style of buildings.
For example, exterior facing windows were mandated for every bedroom in the 1879 Tenement House Act, requirements for exterior fire escapes were implemented for older tenement buildings, and building setbacks were required through the 1916 Zoning Resolution as taller buildings were built.

This project will serve as a visual investigation of the changes in building characteristics as a result of these policies through machine learning techniques.

”

Research Question

Do **architectural typologies of multi-family residential buildings** in Brooklyn’s R6, R6A, and R6B districts emerge through the **clustering of location** or **zoning attributes** throughout time?

The Scope

Spatial & Categorical
Multifamily Residential Buildings in Brooklyn, NY

Temporal
Built between 1870 - 2020

Literature Review

Furman Center. "Housing Policy in New York City: A Brief History". 2006. <https://furmancenter.org/files/publications/AHistoryofHousingPolicycombined0601_000.pdf>

This paper tells the story of housing policy in New York City over the past 30 years. The report describes New York City's current housing environment and policy challenges. This paper is particularly helpful in navigating building features for validation.

Li, Yan, et al. "Estimating Building Age from Google Street View IMages Using Deep Learning." GIScience. 2018. <https://www.semanticscholar.org/paper/Estimating-Building-Age-from-Google-Street-View-Li-Chen/2a1fbafdc6e0dc688fae63fa6fc749ff51a31de8>

Li, et al. describes extracting features from Google Street View images using publicly available pre-trained Convolutional Neural Network (CNN) models AlexNet, ResNet, and DenseNet. The research then used two regression models, mean absolute error (MAE) and root mean squared error (RMSE), to evaluate accuracy of the CNN models used.

Lazar Ilic, M. Sawada, Amaury Zarzelli. "Deep mapping gentrification in a large Canadian city using deep learning and Google Street View." PLoS ONE 14(3): e0212814. <https://doi.org/10.1371/journal.pone.0212814>

This paper presents a Siamese convolutional neural network (SCNN) that automatically detects gentrification-like visual changes in temporal sequences of Google Street View (GSV) images. SCNN achieves 95.6% test accuracy and is subsequently applied to GSV sequences at 86,110 individual properties over a 9-year period in Ottawa, Canada. This paper is particularly helpful in seeking relationships worthy of further interpretation of our project.

Datasets

Google Maps Platform. Google Static Street View
<https://developers.google.com/maps/documentation/streetview/overview>
Based on geo-coordinates of the subset data and the respective headings, Google Street View (GSV) API will be used to obtain raw images.

New York City Department of City Planning. MapPLUTO - Shoreline Clipped (Shapefile) Release 21v1. 2021.
<https://www1.nyc.gov/site/planning/data-maps/open-data/dwn-pluto-mappluto.page>
This dataset will be used to subset tax lots for all multi-family residential buildings in Brooklyn and also be used to train the data based on Built Year.

New York City Department of City Planning. LION Single Line Street Base Map. 2021.
< https://www1.nyc.gov/site/planning/data-maps/open-data/dwn-lion.page?

Methodology

Methodology & Process

To answer our research question, we called in 51,109 views of multi-family residential buildings in Brooklyn through Google's Street View Static API.

While the input parameters for the API were for individual tax lots that were filtered for the scope of this research, the API call resulted in a few images that were not useful for the analysis. These images contained large amounts of scaffolding, trees, and vehicles obstructing much of the building. In order to identify such images, we attempted a color proportion test using k-means clustering.

Images were then normalized in order to make the collected images more consistent as well as making building features more distinct and easy to read for the machine.

Next step was the classification conceptualization through two different concepts: spatiotemporal and morphological relationships by the use of data from NYC's MapPLUTO. K-means clustering and PCA methods were used here.

As the final step, Google's Teachable Machine tool was used to train the algorithm with the normalized data. Teachable Machine is useful in being able to train and define a new library for the purposes of this research rather than having to result in a fixed-library predefined in existing algorithms.

75% of the data from Google Street View was used to train the model in Teachable Machine and through the downloaded model, the remaining 25% of the data was tested.

Analysis & Findings

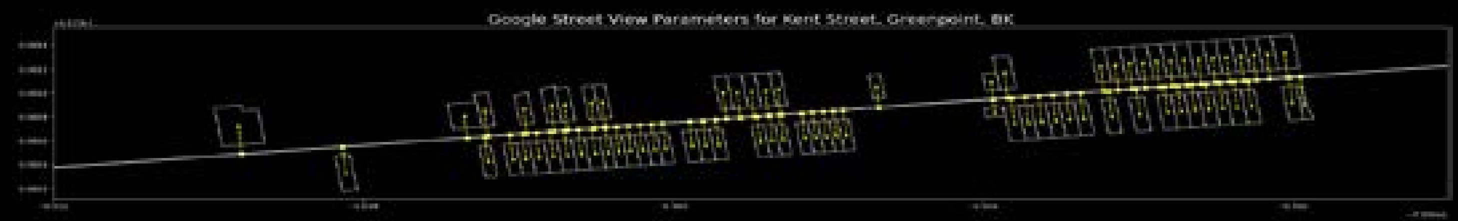
Data Collection

Google's Street View Static API requires 3 inputs as parameters: longitude, latitude, and heading angle, which all have to be projected onto the street network. In order to narrow the scope of this project, R6 multi-family residential buildings in Brooklyn were explored for their historical variance and significant presence over the entire borough. The MapPLUTO dataset was applied the following filters:

borough = ['BK']
zoningDistricts = ['R6', 'R6A', 'R6B']
builtYear >= 1870
numberFloors >= 2 and numberFloors <= 7
numberBuildings > 0

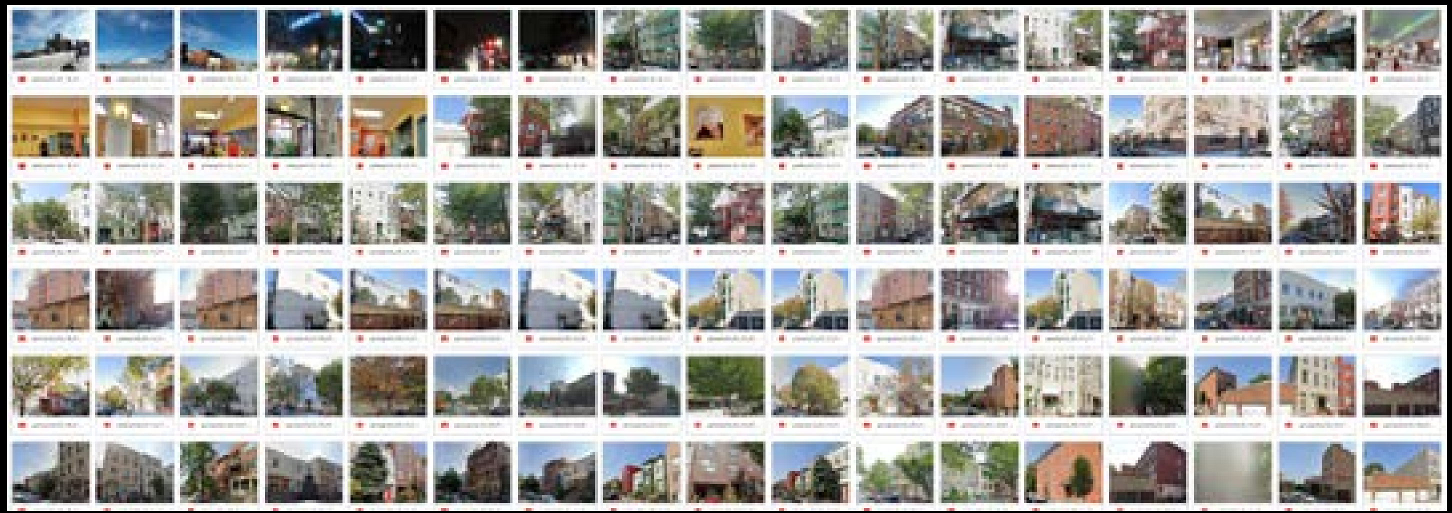
Google's Street View sometimes contains crowdsourced images from inside buildings. In order to avoid these unwanted images, the longitude and latitude of each lot had to be projected onto the street network. Additional processing was required to figure out the heading angle so that the collected data would contain facades of the relevant buildings.

This resulted in a collection of 51,109 images through the API.



Figue 1. Defining Google Street View parameters (Greenpoint's Kent Street visualized here)

Figue 2. Pulling images through Street View Static API



Figue 3. Sample Dataset

Data Filtering

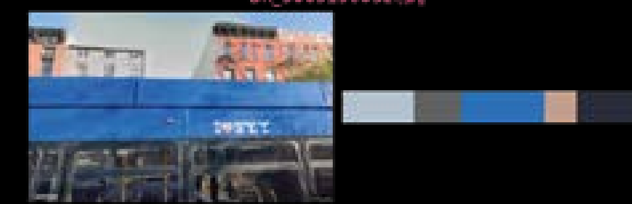
Like most data collection, Google's Street View images were messy and a few of them unusable for the purposes of this research. Images with large amounts of scaffolding, trees, or vehicles obstructing the view of the facade were common. Some corner buildings also posed a challenge as they were referred to from the side streets rather than the front. Partially blurred images, as a result of Google Street View's facial blurring algorithm, sometimes also made the images unusable.

Various techniques were used to try and identify the proportion of the problem to the overall image. While object identification through deep learning methods would be ideal, this method focused on a color proportion test through k-means clustering. This is based on the assumption that images deemed unusable would have a larger proportion of a single color cluster than a usable image. As shown in the performed tests, usable images tend to have better distribution of color clusters.

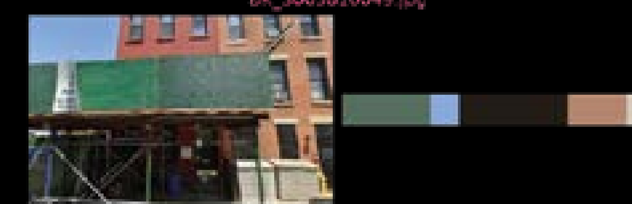
Bad Images vs Good Images

Bad Images

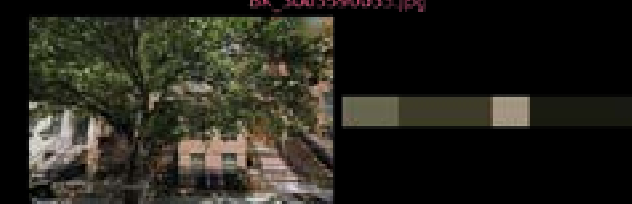
BK_3003190032.jpg



BK_3003810049.jpg



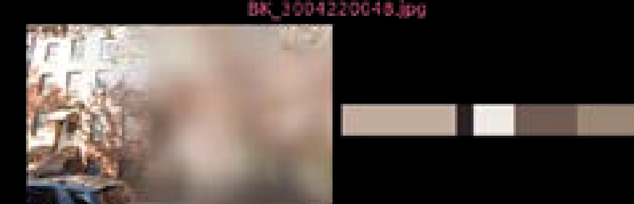
BK_3003590033.jpg



BK_3003300003.jpg

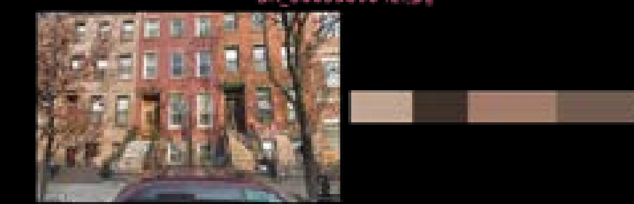
BK_3008910029.jpg

BK_3004220048.jpg



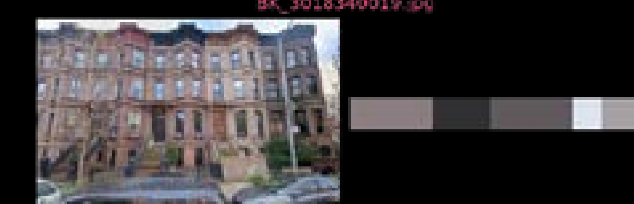
Good Images

BK_3003580042.jpg



BK_3003740046.jpg

BK_3018340019.jpg



BK_3018350026.jpg

BK_3020160068.jpg

BK_3000557504.jpg

A sample set of 5,000 images was tested for this color proportion test, identifying 5 clusters of colors and its proportions. With the threshold set for any single cluster being 40% of the image, 127 images were returned. However, upon closer inspection, this method also brings in useful images, and therefore the technique is not perfect. From the 127 images identified, roughly 60% were manually selected as unusable. Therefore, due to the limitations of this approach, this technique was used to merely flag potentially unusable images but not eliminate them from the dataset.

Image Normalization

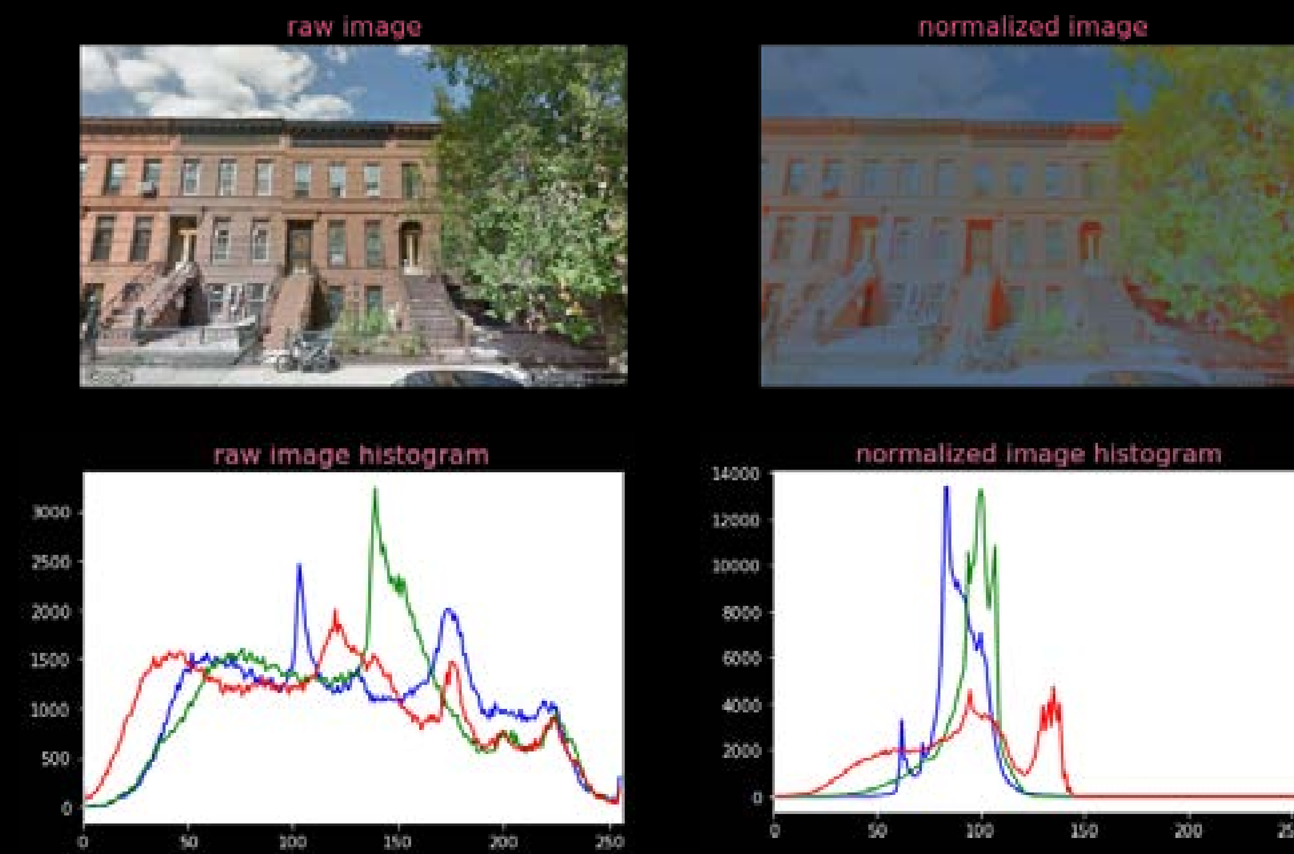
Various normalization techniques were explored in order to standardize the dataset. This included techniques where the mean and standard deviation across all color channels, and maximum and minimum values in order to standardize the images. The most successful method for eliminating highlights, shadows, and overall flattening the image was through normalizing each RGB pixel value. As shown in the formula below, each color channel was proportionally adjusted to add up to 255 for each pixel. When looking at the resulting histogram, this method essentially "squeezes" the RGB channels eliminating many of the lower and higher color values. This step essentially simplifies the image for processing and will yield better training accuracy.

$$f(x,y) = (R,G,B)$$

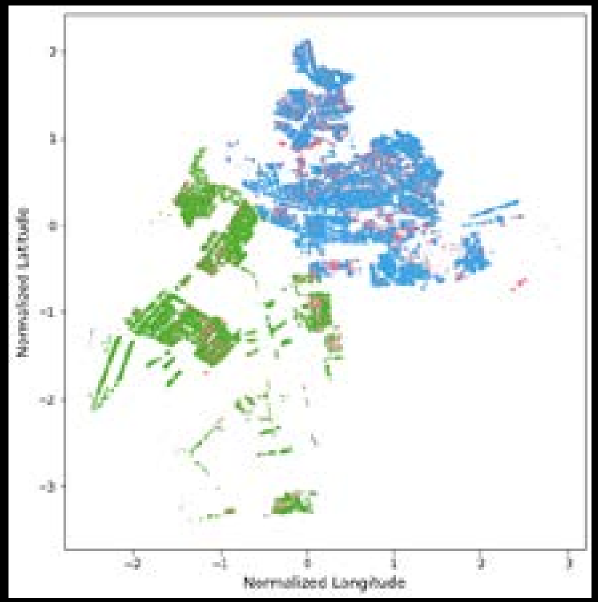
$$total = (R+G+B)$$

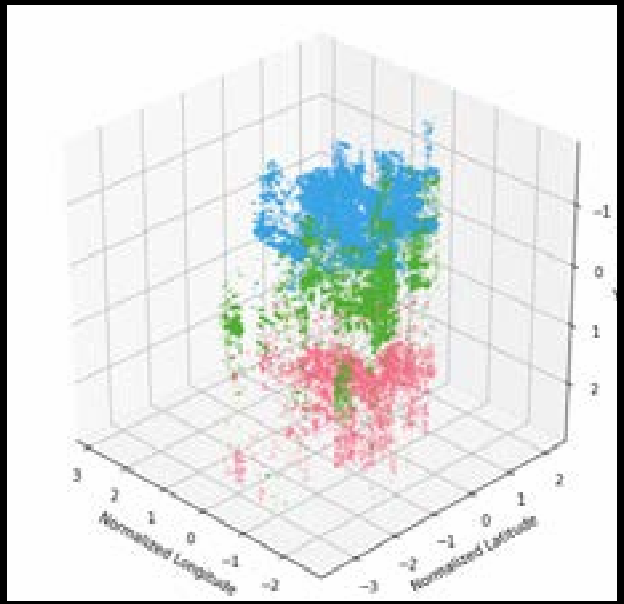
$$R' = \frac{R}{total} \times 255$$
$$G' = \frac{G}{total} \times 255$$
$$B' = \frac{B}{total} \times 255$$

$$thus, g(x,y) = (R',G',B')$$



Classification Conceptualization





1. Spatiotemporal Relationship

Our first conceptualization for the presence of unique typologies in Brooklyn R-6 zoned housing was spatiotemporal, comparing location and time. As a premise, 'vernacular' architectural typologies often emerge through similar constraints, such as available building materials and cultural codes, at a specific location. If Brooklyn housing followed a similar trend, it could manifest by clustering location and the built year.

Considering the scaled latitude and longitude on our X- and Y-axis and our scaled Built Year on the Z-axis, we constructed a time-space prism within which the R6-zoned buildings could be plotted as a point cloud. Through a silhouette test, we found the best fit for a K-Means classification of the data when there were 3 and 17 spatiotemporal clusters.

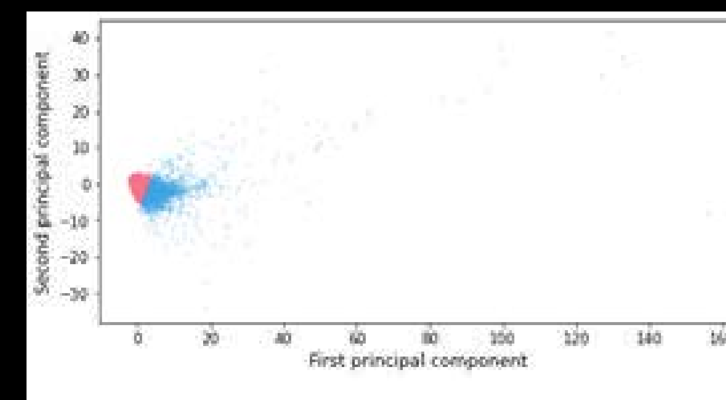
The model of three spatiotemporal clusters appeared to illustrate a pattern of three layered phases of development within the current stock of R6-zoned housing in Brooklyn, with greater stratification through the built year. The oldest phase of R6-zoned housing development, built near the turn of the century, was dispersed as remnants throughout Brooklyn. A second phase of R6-zoned housing emerged in the middle century throughout Southwest Brooklyn, particularly high presence in Brooklyn Heights, Cobble Hill, Sunset Park, and Flatbush. The most recent and densest phase of R6-zoned housing emerged in Northeast Brooklyn bordering Queens, in neighborhoods including Greenpoint, Williamsburg, Bushwick, Bedford-Stuyvesant, and Crown Heights.

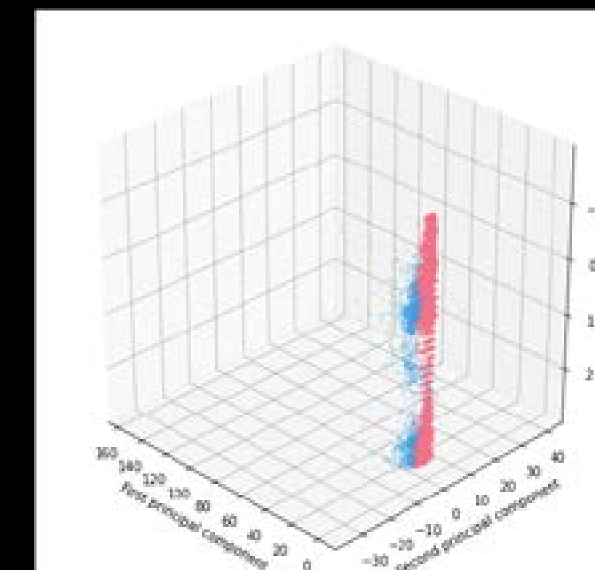
The model of seventeen spatiotemporal clusters appeared to illustrate unique enclaves of housing development presenting more granular geographic distinctions between the clusters. Roughly along neighborhood tabulation boundaries, distinct neighborhood groupings emerge such as Greenpoint and Williamsburg, Bedford-Stuyvesant and Crown Heights, Sunset Park and Bay Ridge, and the many smaller neighborhoods surrounding Downtown Brooklyn. Promisingly, several of these relationships are affirmed by known contextual relationships between these neighborhoods.

2. Morphological Relationship

Our second conceptualization for the presence of typology in Brooklyn housing was morphological, comparing building form and time. In highly urbanized areas such as New York, zoning compliance can factor heavily into the design and development of buildings. Throughout time, changes in these form-based zoning regulations could allow unique typologies to emerge and evolve, or alternatively, could run in parallel with changing tastes and needs or advancements in building technology underlying these stylistic differences.

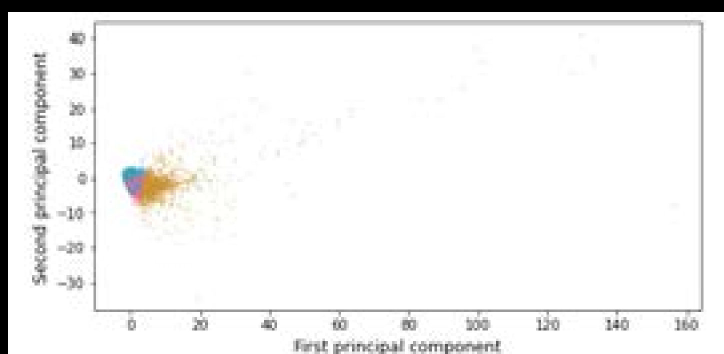
We considered form-based zoning attributes based on Primary Land Use Tax Lot Output (PLUTO™) from the NYC Department of City Planning and Department of Finance. Within the 22 continuous variables in the MapPLUTO data, several variables had pre-existing relationships based on a correlation matrix of all variables. Both to reduce issues of multicollinearity and provide a means of visualizing the data, we used Principal Component Analysis to identify only the significant variation between all of the variables and compress this information into two principal components. The two principal components accounted for 32.8% and 9.5% of the variance in the data, together explaining approximately 42.3% of the variance in the dataset. Although this percentage was not optimal, we accepted that the new data could still account for some of the most significant, if not all, of the formal differences in R6-zoned housing in Brooklyn. Based on the eigenvalues for each variable, the first principal component best captured the variance in lot area, building area, residential area, number of residential units, and number of total units. The second principal component best captures the variance in the number of floors and built FAR.

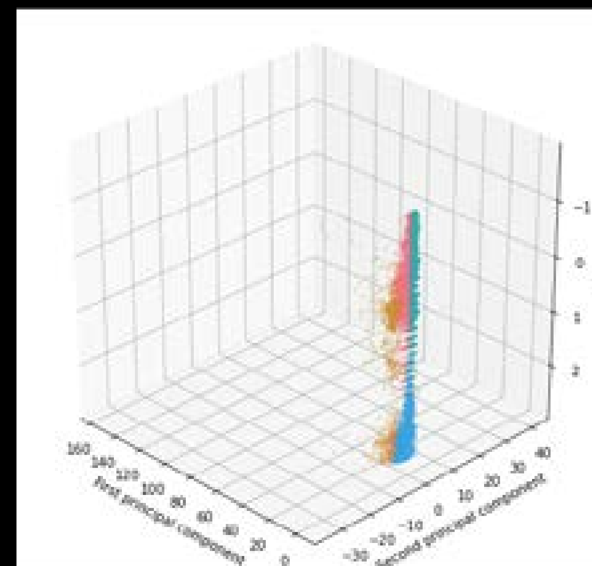




Considering these two scaled principal components on the X- and Y-axis and our scaled Built Year once again the Z-axis, a second morphological prism was created where the R6-zoned buildings could be plotted as a point cloud. Through a silhouette test, we found the best fit for 2, 3, and 6 clusters. Although the silhouette score for two clusters was misleadingly high, in reality the 2-cluster K-means model only separated an overly dominant class of outliers from a few sparse, non-outliers. We opted to consider the 3-cluster and 6-cluster models.

The model of three morphological clusters resulted in a stratification of two classes of building form based on MapPLUTO data, in addition to a class of far outliers. The less varied class is maintained relatively consistently throughout the range of years for the study, while the more varied class begins at the beginning of the century and re-emerges at the middle of the century. As a second test to determine variable significance alternative to the eigenvalues of the principal components, we used a Random Forest classifier to reverse engineer feature importance percentages for the newly produced classes. For the model of three morphological clusters, the total number of units, residential area, and building area were the most important features, followed less substantially by the number of residential units and the lot area.





The model of six morphological clusters resulted in a bifurcation in time where there is low variance in the building information with two or three strata on either side, in addition to a class of far outliers. The bifurcation separates the turn of the century with the other years in the study range. Two strata exist at the turn of the century with three after the bifurcation. We again used a Random Forest classifier to reverse engineer feature importances for the six classes. For the model of six morphological clusters, the Built FAR was the most important feature, followed less substantially by the number of floors and building area.

Consideration of Clustering Methods

Between K-Means, Density-Based Spatial Clustering of Applications with Noise (DBSCAN), Agglomerative Clustering, and Gaussian Mixture Model (GMM), we ultimately selected K-Means as the clustering algorithm after identifying some of the important characteristics of our data.

Because the distribution of zoning attributes are determined by people and thus cannot be assumed to be Gaussian or normally distributed, Gaussian Mixture Models could result in more inaccuracy than refinement.

DBSCAN offered an attractive means to identify outliers and noise we knew was present in our data, but we recognised the varying density due to the intervals of the built year data compared to our continuous lat-lon data in our spatiotemporal conceptualization in addition the large number of dimensions we considered in our morphological conceptualization would cause the model to suffer .

Although the hierarchy that Agglomerative Clustering could also potentially be useful for future applications, our current study only accounted for the data at one level.

Finally, because of the large number of datapoints in the MapPLUTO dataset, the processing time would increase substantially for each of these more computationally expensive algorithms. As we were looking to prototype several conceptualizations, we found it would be most informative and efficient to cluster by K-Means at this stage of our study. Nonetheless, DBSCAN and Agglomerative Clustering could provide additional utility for future extensions of the study.

Computer Vision

Because of the specific focus of our computer vision on architectural features, it would be difficult to develop a model to identify similar building features based on pre-packaged computer vision libraries trained on libraries of images we had no control over. Attempts to do so resulted in identifications that were not relevant to the study in addition to several incorrect identifications. To train a custom computer vision model based on user-defined libraries, we used the web-based tool Teachable Machine by Google, which primarily utilizes Tensorflow, a machine learning platform available as a python package.

Based on each of our four classification models, we grouped a training subset of the collected Google Street View images by class into libraries of images through the file management python packages shutil and os. Uploading these images to Teachable Machine, we exported four computer vision models for each of the classification models.

The classes we produced clustering based on input data, independent of the Google Street View images. If the trained computer vision model could relatively reliably predict matching classes based only on information from the Google Street View images to the classes produced by our classification models, then we could assume that there was a connection between the conceptual frameworks we established to cluster building by their spatiotemporal or morphological attributes and the visual presence of the building captured by Google Street View.

Result & Findings

The match between the clustering classification models and computer vision models were the highest for the spatiotemporal or vernacular models, impressively producing accurately matching values 70-80% of the time. The match between models was occasionally correct significantly lower for the morphological or regulatory models, producing matching values 50-60% of the time. For both, the number of clusters did not appear to produce as significant differences in accuracy compared to the conceptualizations.

At this stage of our study, we conclude that there is a stronger relationship between a building's visible architectural features and its spatiotemporal characteristics, as defined by its geographic latitude and longitude and built year, than with its morphological characteristics, as defined by its PLUTO form-based zoning attributes and its built year..

References

Furman Center. "New York City's Multi-family Rental Housing and the Market Downturn." In State Of New York City's Housing & Neighborhoods. 2010. https://furmancenter.org/files/sotc/Multi-family_Rental_Housing_2010.pdf

Li, Yan, et al. "Estimating Building Age from Google Street View IMages Using Deep Learning." GIScience. 2018. https://www.semanticscholar.org/paper/Estimating-Building-Age-from-Google-Street-View-Li-Chen/2a1fbafdc6e0dc688fae63fa6fc749ff51a31de8

Lazar Ilic, M. Sawada, Amaury Zarzelli. "Deep mapping gentrification in a large Canadian city using deep learning and Google Street View." PLoS ONE 14(3): e0212814. https://doi.org/10.1371/journal.pone.0212814

Columbia GSAPP

Exploring Urban Data with Machine Learning

Spring 2021

Will Cao, Jin Hong Kim, and Sori Han