Estimating Residential Property Values in NYC

Problem Statement:

Understanding and estimating property values is a critical task in New York City that has extensive implications for a variety of industries and use cases. Accurate property valuations are necessary for urban planning, real estate development, investment decisions, mortgage lending, insurance assessments, risk assessments, and more.

In NYC, there are three main methods for estimating property values: the appraisal approach, market analysis, and assessment rolls. While the appraisal approach is considered the most accurate, it can be time-consuming and expensive to conduct on an individual property basis. Market analysis is a popular alternative that involves analyzing trends and patterns in the real estate market to estimate property values and is a technique that is used by Zillow, Property Shark, and other real estate databases. The NYC Department of Finance uses an assessed value to determine property taxes, but these values are often outdated and not reflective of current market conditions.

Methodological Need

Estimating property values requires spatial analysis because property values are inherently spatially autocorrelated. The value of a property is determined by its own characteristics, the characteristics of surrounding properties and the neighborhood characteristics.

For example, in New York City and elsewhere, properties located in proximity to the subway or in certain desirable neighborhoods may have higher value than a building with the same characteristics located in other areas.

Spatial analysis is used to incorporate the spatial relationships between properties and their values into a model. A model can then be used to predict property values for properties where data is not available, or to estimate the impact of different factors on property values. Overall, spatial analysis is essential for accurately estimating property values and understanding the spatial patterns of the real estate market.

Spatial Analysis Methods:

Spatial analysis is an essential tool for accurately estimating property values, particularly in areas with high spatial autocorrelation like New York City. New York City property values are highly dependent on the characteristics of surrounding properties and the neighborhood in which it is located.

To estimate property values and account for spatial autocorrelation in property values, spatial analysis methods such as k-means clustering and kriging can be used. K-means clustering is a machine learning algorithm that groups properties based on shared characteristics such as size, building class, and age. These clusters can then be used as input variables in a prediction model. This method can be useful in identifying different groups of properties within an area and determining how they relate to property values.

Kriging is a geostatistical interpolation method that uses nearby observations to estimate values at unsampled locations. This method produces more accurate estimates than simpler interpolation methods because it accounts for spatial autocorrelation. When examining property values, kriging can be used to estimate the value of a property at a specific location based on the values of nearby properties.

By using K-means and kriging, it is possible to develop a model that accurately estimates property values in New York City. This model could be updated regularly using the latest available data and would allow for a more dynamic understanding of the real estate market.

Extant Science

In the study entitled “Estimating Residential Property Values on the Basis of Clustering and Geostatistics,” researchers used a “two-stage” model to first incorporate structural building features into a k-means clustering model based on floor area, the number of rooms, additional rooms, story, construction year, and the type of market. Researchers found the optimal number of clusters should be determined using an agglomeration model. The second stage involved developing a spatial model using ordinary kriging to predict property values based solely on geographic location. One benefit of using kriging is that it computes the interpolation error. Cross validation was then used to analyze the accuracy of the model based on the chosen semi-variogram model. Error was then estimated using Mean Absolute Error and Mean Absolute Percentage Error calculations and found this process allows property values to be estimate without error exceeding ten percent.

The study entitled “Automated valuation models for real estate portfolios : A method for the value updates of the property assets” utilized additional explanatory variables in an analysis of property values in Milan, including walking distance to a subway stop and walking distance to an urban park. The analysis included extensive information about building condition, including date of last renovation and maintenance condition. Researchers employed an Evolutionary Polynomial Regression which was found to maximize accuracy because it is an iterative algorithm.

Data Statement

The study utilized publicly available data from the City of New York. Primary datasets and variables will include:

1. MapPLUTO: tax lot data and features from the Department of Finance’s Digital Tax Map. This is a comprehensive and reliable dataset of all tax lots in New York City. The data are in shapefile format. The following attributes were used in this analysis: Borough Block Lot : 'bbl', Total Lot Area: 'lotarea', Total Commercial Area: 'comarea', Total Residential Area : 'resarea', Total Office Area: 'officearea', Total Retail Area: 'retailarea', Total Garage Area: 'garagearea', Number of Buildings: 'numbldgs', Number of Floors: 'numfloors', Units Residential : 'unitsres', Property Lot Width: 'lotfront', Property Lot Depth : 'lotdepth', Year Built : 'yearbuilt', Landmark Status : 'landmark', Year Altered: 'yearalter1', Building Class 'bldgclass'
2. 2020, 2021 and 2022 Property Sales data from the NYC Department of Finance. The data are separated by borough in excel spreadsheets. The following fields from the dataset were used in the analysis: Borough Block, Address, Residential Units, Sale Price.

Results

Extensive data cleaning and processing was done to assign property values to properties in the MapPLUTO dataset. Cleaning steps will not be explained in this report but can be found here. A K-means clustering algorithm was trained to the properties with assigned property values. The optimal number of clusters was found to be 5 using the elbow method and silhouette score.

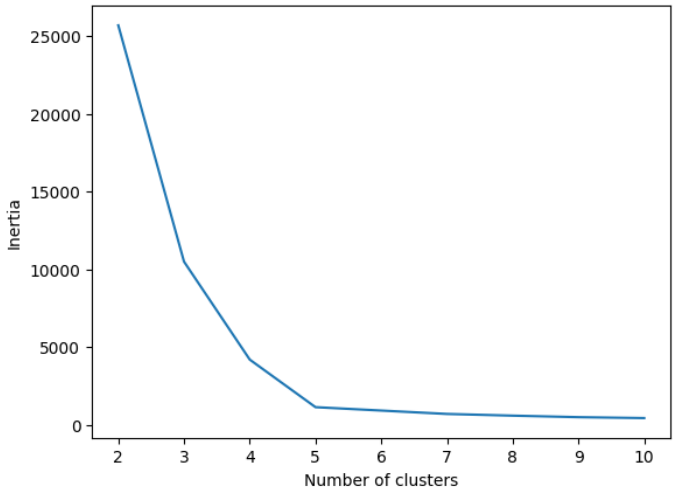

Figure 1:Number of Clusters vs Inertia

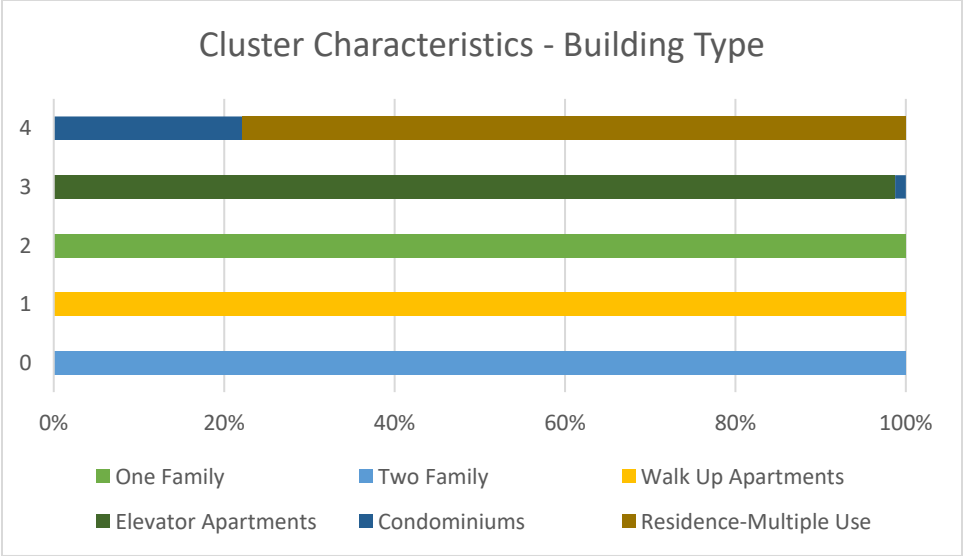

Figure 2: Cluster Characteristics - Building Types

The K-means clustering algorithm discretely separated the residential building types. One family homes are found in cluster 2, and two family homes are found in cluster 0. Walk up apartments are found in

cluster 1. Cluster 3 is comprised of Elevator apartments and cluster 4 is approximately 20% condominiums and 80% multiple use residential buildings.

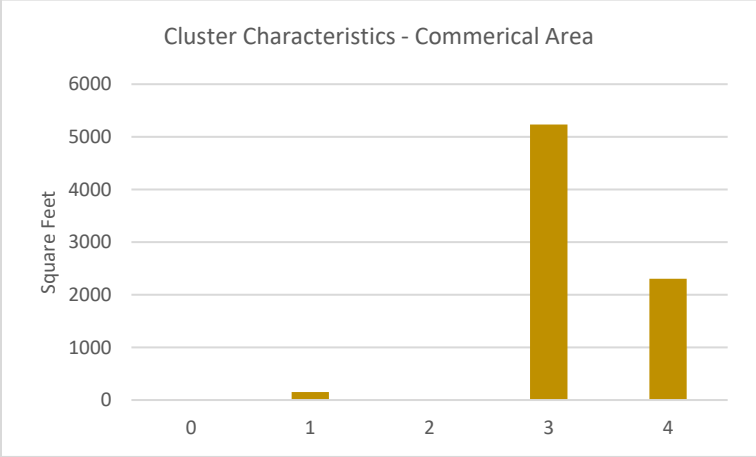

Figure 3: Cluster Characteristics - Commercial Area

The highest average square feet of commercial space correspond to elevator apartments in cluster 3. Condominiums and residential-multiple use properties in cluster 4 have the second highest average of commercial area.

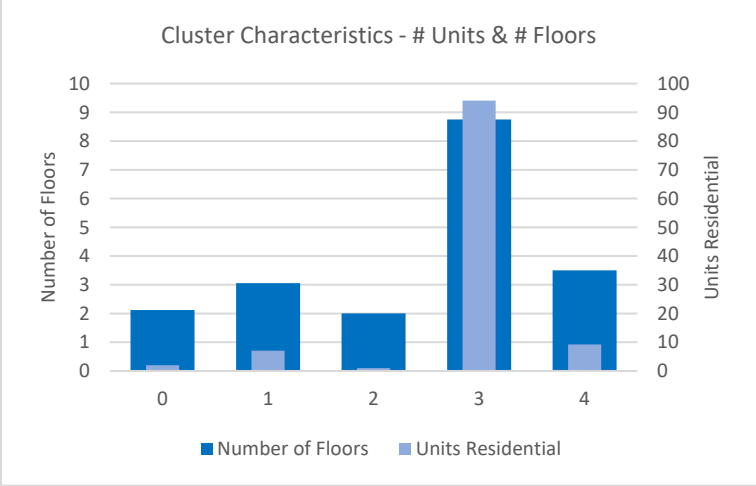

Figure 4: Cluster Characteristics - # Units and # Floors

Elevator buildings in cluster 3 predictably have the highest average number of floors and number of residential units. Condominiums and residential-multiple use properties in cluster 4 have the second highest average number of floors and number of residential units.

Figure 5: Cluster Characteristics - Lot Width & Depth

Lot widths and depth are generally standard across the city, which can be seen in the Figure above. Elevator buildings in cluster 3 have an average higher lot width and lot depth, indicating large elevator buildings are located on larger parcels.

The results of the K-means clustering indicate that properties are well sorted by their respective type. This is an indication that the kriging model will perform well. Buildings with similar properties in proximity are expected to have similar property values.

Kriging analysis

The dataset was separated into labeled and unlabeled data for each cluster. A kriging model was built for each labeled dataset and a prediction was attempted for each dataset using Python and using ArcMap. The model failed due to computing power in multiple instances and multiple systems: using my personal laptop, CUSP RCF remote desktop and on-site computers. For testing purposes, a very small subset of Cluster 3 was selected in lower Manhattan.

A Gaussian variogram model was fitted to the labeled property value data which assumes that the data roughly follows a normal distribution. The figure below shows the output variogram. Further analysis should test multiple variogram models to determine the best fit.

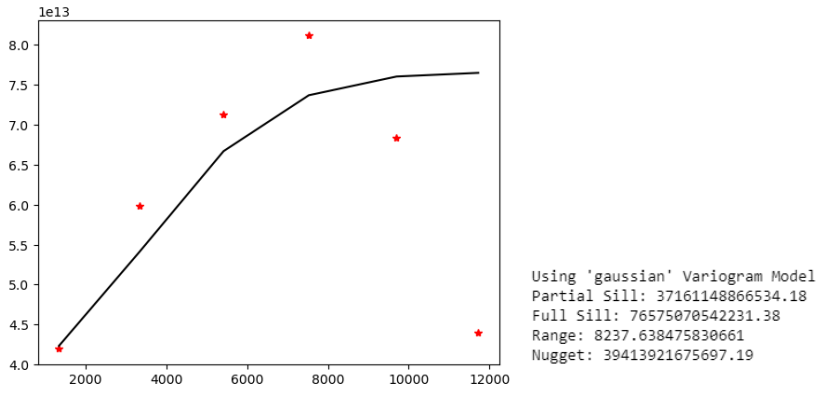

Figure 6:Gaussian Variogram

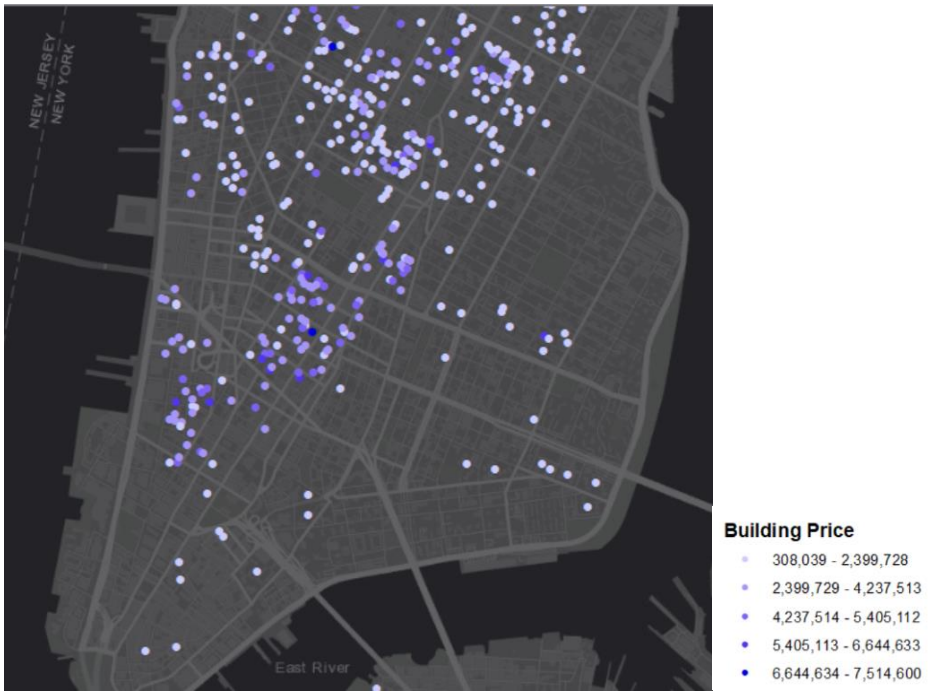

Figure 7: Cluster 3 - Labeled Building Values

Figure 8: Cluster 3 – Predicted Building Values

The spatial distribution of the predicted building prices is not an expected result. Building prices in the West Village are shown to have the lowest property values in the sample size. The results show similar predicted values to the actual values shown in the West Village shown above. The dark blue band is shown to be in an area with few actual values. Two bridges is seen to have relatively low values, which is expected. However, there is no training data available for this area.

Discussion

Additional data prior to 2020 would help the model by filling in some of the spatial gaps for which there is no data. A correction could be added for additional datasets that are older than 2020 to incorporate them into the model while accounting for inflation. With additional property value data, the model will improve.

The cluster assignments improved the model by sorting the data into categories by building type and in future iterations of this project I would utilize the same building characteristics. It is possible that certain characteristics were weighted too heavily, like building type, such that all the same building types were sorted into the same cluster. In a future study, the weighting of attributes in the k-means clustering should be examined.

In the kriging analysis, it was not possible to run the entire model at once with the systems available. For testing purposes, Lower Manhattan was selected for observation. While the analysis was able to run using this subset of data, it is not a precise prediction because labeled properties that are adjacent to the clipping boundary are not included in the kriging analysis. In future analysis, if clipping is necessary for a manageable model size, the borough boundaries would be an appropriate clipping boundary.

An appropriate variogram model can be chosen after observing the outputs for multiple variogram models. Because the data has a large range, it is unlikely that a gaussian model is the best fit. A further study would examine the distribution of the labeled dataset in each cluster.

With further refinement of this model, it can be a reliable estimation of property values in NYC.

Resources

Calka, Beata. 2019. "Estimating Residential Property Values on the Basis of Clustering and Geostatistics" Geosciences 9, no. 3: 143. https://doi.org/10.3390/geosciences9030143

Automated valuation models for real estate portfolios : A method for the value updates of the property assets. Morano, Pierluigi;Ntalianis, Klimis;Tajani, Francesco. Journal of property investment & finance, 02 Jul 2018, Vol. 36, Issue 4, pages 324 - 347