Random Forest modeling for mass appraisal:

Predicting Manhattan real estate prices from 2017 through 2022

Lyle S. Prockop (lp2974)

Quantitative Methods in the Social Sciences

Columbia University

Professor Michael Parrott

May 10, 2023

Contents

Abstract **3**
Introduction **3**
Literature Review **4**
Background and Significance 4
Hedonic Modeling 5
Alternative Modeling Methods 6
Spatial Parametric Models 6
Nonparametric Models 7
Significance of the Setting 9
Data **10**
Explanatory Variables 10
Data Sources 12
Data Cleaning and Processing 14
Methodology **18**
Methods 19
OLS Model 19
Random Forest Model 20
Evaluation Metrics 20
Results **22**
Descriptive Statistics 22
Model Accuracy 25
Random Forest Feature Importances 27
OLS Coefficients 29
Discussion **30**
Limitations 30
Conclusion **31**
References **33**

Abstract

While mass real estate valuation is traditionally conducted using hedonic modeling via Ordinary Least Squares regression, research suggests that this modeling approach is not appropriate and produces biased predictions (Hong et al. 2019). Researchers have suggested nonparametric models as alternatives; specifically, Random Forest models have been shown to perform well on spatial data and in a real estate setting specifically. This study compares the performance of Ordinary Least Squares (OLS) and Random Forest (RF) models in predicting real estate transaction prices in Manhattan from 2017 through 2022. Variables describing each property's structure, location, and neighborhood are used to predict the natural logarithm of sale price for properties that were sold during the study period. We examine each model's performance over this period of macroeconomic volatility and find that the Random Forest model performs slightly better on average and significantly better at extremes than Ordinary Least Squares, suggesting that Random Forest models might yield important insights during volatile periods or in particularly heterogeneous settings.

Introduction

Prior research suggests that the most common model for mass real estate valuation, hedonic modeling using Ordinary Least Squares (OLS) regression, is mis-specified (Hong et al. 2019). Recent advances in non-parametric models and increased computing power have unlocked new model architectures that might be more appropriate for the task. Specifically, research has found that Random Forest (RF) models are more suitable for mass real estate appraisal than traditional hedonic models: in addition to overcoming the specification errors of OLS, they yield more accurate predictions (Hong et al. 2019). However, these findings have been limited to settings with relatively stable housing markets. This research compares the approaches during a tumultuous economic period to assess each model's robustness to volatility. The hypothesis is that RF models continue to outperform OLS models even in this setting with heterogeneous properties and volatile macroeconomic conditions. The scope of the study is real estate transactions in Manhattan, New York City, New York from the beginning of 2017 to the end of 2022.

Accurate real estate valuation is a critical step in protecting housing stability (Wallace and Wallace 2020; Zuk et al. 2015; Kiely and Bastian 2020). The Covid-19 pandemic showed that housing is deeply connected to intersectional equity disparities, from health to education and employment (Robbins 2022). Thus, to implement policies that protect residents and foster healthy communities in both stable and unstable times, citywide valuation models must be accurate, reliable, and robust to volatility. This research aims to assess whether RF is a viable option.

To achieve this, we fit OLS models and RF models to predict the natural log of sale prices for properties that were sold between January 1, 2017 and December 31, 2022. We assess model accuracy using the Mean Absolute Percent Error (MAPE), R-squared, and Coefficient of Dispersion (COD) (Hong et al. 2019). The data is obtained from New York City Open Data (NYCOD) (Table 1).[1] The dependent variable is the natural log of the sale price for a property as recorded in NYCOD Rolling Sales Data, and the independent variables are characteristics of each property's structure, location, and neighborhood that are made available through NYCOD (Table 2, Table 3).

Literature Review

Background and Significance

Mass real estate appraisal is an important step in systematic property valuation, which has applications in tax computation and housing policy (Hong et al. 2019). Although it has important political and economic consequences, mass real estate valuation remains more art than science (Zuk et al., 2015). Thus, "a stable, accurate, and fast tool for appraisal is needed" (Hong et al. 2019:140). Methods used for mass appraisal are automated valuation methods (AVMs) and are dominated by hedonic modeling (Hong et al. 2019; Kiely and Bastian 2020). Hedonic modeling uses parametric linear models; the most common technique is ordinary least squares (OLS) regression, but other methods include multiple regression, maximum likelihood regression, and weighted least squares regression (Kiely and Bastian 2020). Researchers argue that real estate

[1] https://opendata.cityofnewyork.us/

data violates key assumptions of OLS, which implies that hedonic models using OLS generate biased and unreliable predictions (Kiely and Bastian 2020). Thus, while the modeling task has significant real-world implications, the predominant method is flawed.

Hedonic Modeling

The theoretical foundation of hedonic modeling is the idea that a good can be treated as composed of many individual components or characteristics that independently contribute value toward the good's total value (Hong et al. 2019). This model is based on Lancaster's consumer theory and was applied by Rosen (Lancaster 1966; Rosen 1974). The former holds that consumers derive utility not from goods themselves but directly from the characteristics that comprise the good; Lancaster treated consumption of a good as consumption of the individual attributes (Hong et al. 2019). Rosen extended this theory of consumption to valuation, claiming that the value of a good can be divided into the value of each of its characteristics (Hong et al. 2019). If each attribute has its own implicit price, the total price of the good can be regressed on its characteristics (Hong et al. 2019).

Hedonic models dominate real estate appraisal and property value modeling (Potrawa and Tetereva 2022:50; Hong et al. 2019). Among the proposed frameworks for breaking real estate prices into individual characteristics, Chin and Chau's (2003) is commonly used, which designates features as "locational", "structural", or "neighborhood" (Chin and Chau 2003).

While the hedonic model is theoretically relevant to real estate because it allows a property's value to be attributed to individual characteristics, the application - through linear parametric models - often isn't appropriate (Potrawa and Tetereva 2022:51). As Zurada et al. (2011) state: "[OLS has] failures that would result in untenable or imprecise coefficients caused by functional form misspecification, interaction among variables, multicollinearity, and non-linearity problems". Hong et al. (2019) echo this statement, pointing out that the OLS assumption that attributes are separable and constant are often not met in practice. In summary, "there are some factors of the value determination process that cannot be fully explained in the simplified assumptions of the conventional hedonic pricing model", and thus, "the stability and

accuracy of [hedonic modeling] remain questionable" (Hong et al. 2019:140). Next, alternative models are introduced that overcome these limitations.

Alternative Modeling Methods

Spatial Parametric Models

OLS models rely on the assumption that residuals are uncorrelated and normally distributed; however, property values often inherently exhibit spatial autocorrelation, meaning that observations are affected by their neighbors (d'Amato and Kauko 2017). In the presence of spatial autocorrelation, OLS coefficient estimates are biased and lead to unreliable predictions (d'Amato and Kauko 2017).

Researchers have explored methods for handling spatial autocorrelation in real estate data to make linear parametric models suitable. So-called "spatial autoregressive models" include spatial lag models, spatial error models, and Geographically Weighted Regression (GWR) models (Kiely and Bastian 2020). Applying spatial autoregressive models enables researchers to still employ linear parametric modeling and interpretation in the presence of spatial autocorrelation. Spatial lag models include weighted summaries of nearby data points as independent regression variables (Kiely and Bastian 2020; d'Amato and Kauko 2017). Spatial error models function similarly, but include nearby residual values as an independent variable rather than nearby dependent variable values (Kiely and Bastian 2020). GWR, on the other hand, considers varying effects of independent variables across space, meaning that spatial effects can vary across a study area (d'Amato and Kauko 2017). Among available weighting schemes for GWR, a Gaussian kernel with a distance decay function has been shown to perform the best in real estate settings, which suggests that factors affecting real estate valuation vary locally (d'Amato and Kauko 2017). These models enable researchers to explore regionally-specific trends that can be correlated with other localized phenomena (Del Giudice and De Paola 2017). Hedonic models aim to use this localized information to attribute values to individual characteristics of each property (Kochinsky et al. 2012).

However, spatial models are an imperfect solution for real estate valuation: while they overcome spatial autocorrelation, they still rely on the assumptions that independent variables are constant and separable (Potrawa and Tetereva 2022). Assuming constancy may require transformations that would introduce functional form bias, and assuming separability might require a significant number of manually specified interaction terms among explanatory variables, which can drastically increase the total number of explanatory variables and lead to overfitting in models (Potrawa and Tetereva 2022:52). Thus, when spatial characteristics are considered in hedonic models, the level of predictability and model generalizability is often low (Hwang and Quigley 2004).

Nonparametric Models

Alternatives to spatial parametric models are nonparametric models. This is where machine learning has been employed in a variety of capacities and had a significant impact (Kiely and Bastian 2020; Park and Bae 2015). Machine learning-enabled techniques "are capable of identifying nonlinear relations between the variables that (from the economic perspective) would reflect agents' behavior when interacting in the market" (Rico-Juan and de La Paz 2021:2). Nonparametric techniques enabled by machine learning have been found to consistently outperform linear regression and other parametric techniques (Breiman 2001).

Among the nonparametric techniques studied, there is consensus that tree-based models are particularly well-suited to real estate value prediction (Breiman 2001; Schernthanner et al. 2016; Antipov and Pokryshevskaya 2012). One benefit of tree-based models is their ability to capture interaction effects and nonlinearities, which is an advantage over both traditional linear and spatial methods (Potrawa and Tetereva 2022:53). Another reason for the prevalence of tree-based models is their interpretability: while many machine learning models are relatively successful in prediction, they do not all allow the researcher to ascertain the impact of specific property characteristics in a way that makes them viable alternatives to hedonic models (Potrawa and Tetereva 2022:53). For example, neural network-based models have been effectively rejected by academics for real estate modeling because the results cannot be easily interpreted and explained (Rico-Juan and de La Paz 2021).

Among tree-based models, Random Forest (RF) models have been found to perform best in a real estate setting (Antipov and Pokryshevskaya 2012). RF is an ensemble method consisting of simple regression trees, which generates predictions based on averaging predictions made by each of its individual trees (Antipov and Pokryshevskaya 2012). Each regression tree consists of a series of branches and nodes where decisions are made based on a feature and threshold value (Hong et al. 2019:145). Tree-based models do not make assumptions about the underlying distribution of the data and can capture nonlinear patterns well (Potrawa and Tetereva 2022:58). In addition, they can capture interactions between covariates well and can handle both numeric and categorical variables (Potrawa and Tetereva 2022:58). According to Hong et al. (2019), the specific advantages of RF for real estate analysis are:

1. RF can successfully manage categorical data without the use of dummy variables, which would increase the number of trainable parameters and potentially lead to overfitting.
2. RF allows for nonlinear relationships between variables, and interactions between independent variables do not need to be manually specified.
3. RF doesn't require detailed model specification, which could cause overfitting; they are defined primarily by the number of trees and depth of each tree.
4. RF has been shown to predict real estate prices accurately in multiple settings.

These findings are echoed throughout other literature (Antipov and Pokryshevskaya 2012). Antipov and Pokryshevskaya (2012) "believe RFs may become one of the most appropriate techniques for mass appraisal … it is expected to avoid the fallacies of many other methods commonly used for mass appraisal". Hu et al. (2019) agree, stating that tree-based bagging algorithms (such as RF) are "effective and robust on spatial prediction" and were found to outperform multilayer perceptrons (neural networks), K-Nearest Neighbors, and Support Vector Regression (Hu et al. 2019:670). As Hong et al. (2019) conclude, "the RF algorithm constructs the data-driven hierarchical structure of the model without the modeler explicitly describing it. Therefore, if the data set sufficiently covers the characteristics of the property, the RF model is expected to more sensitively replicate the complex structure of the house price determination process" (Hong et al. 2019:142). Essentially, Hong et al. (2019) argue that where OLS and other models aim to replicate observed end-results (e.g., sale prices), RF is able to more closely replicate the real-world decision-making process to achieve a more consistently-accurate

prediction (Hong et al. 2019). Because of these advantages, RF is the model employed in this study and compared with OLS for real estate price prediction.

Significance of the Setting

Even before the Covid-19 pandemic, New York City exhibited a uniquely localized and neighborhood-focused real estate market (Mironova, 2019). Wallace and Wallace (2020) mapped premature mortality disparities across New York City communities, concluding that good housing and residential stability can determine whether neighborhoods are predisposed to foster healthy, intergenerational communities or to be "caught in the vicious cycles of poverty, unemployment, and discrimination" (Wallace and Wallace 2020:15).

During the Covid-19 pandemic, economic turmoil and temporary citywide policies created an unprecedented real estate market in New York City (Robbins, 2022). Many residents chose to leave the city for the suburbs, but the effect was varied across the city (Cohen et al. 2022). The pandemic illustrated that hardship, as Canelas and Baptista (2021) point out, was unequally distributed across different segments of the population; for example, lockdown policies had drastically different effects for those who would work remotely than it did for those who did not rely on formal employment and could not satisfy their basic needs locally (Canelas and Baptista 2021:280). Thus, the pandemic illustrated a significant relationship between housing disparities and intersectional equity disparities related to health (Daher-Nashif 2021). Robbins (2022) states that "the pandemic highlighted the deep-rooted links between poor housing and poor health" (Robbins 2022:610). Canelas and Baptista (2021) credit the pandemic with creating a new, hyper-localized conceptualization of neighborhoods in New York City (Canelas and Baptista 2021).

Migration trends were complicated by rent strikes and New York City effectively banning evictions (Kelly 2020; Haag 2021). At the most localized level, individuals with the highest and lowest income levels had the greatest increases in housing demand (Zhao 2020). At a higher level of abstraction, the pandemic widened the housing wealth gap in the city; home sale prices

fell in low-income areas where Covid cases were also high, where more affluent, less densely populated neighborhoods didn't experience such price drops (Cohen et al. 2022).

While prior studies have shown that RF models overcome the assumptions of OLS and generate accurate predictions in a real estate setting, many studies have been limited to relatively homogeneous settings (Hong et al. 2019; Hu et al. 2019; Rico-Juan and de La Paz 2021). In fact, Hong et al. (2019) attribute their RF model accuracy partly to the structural similarity between properties in their study area (Hong et al. 2019). At the time of writing, RF's stability and accuracy over an unstable or volatile housing market has not been studied.

Data

Explanatory Variables

Each observation is a single property transaction. The dependent variable is the natural logarithm of the sale price. In order to ensure model comparability, the same dependent variables are used in RF and OLS models. These features fall into the traditional hedonic modeling categories of structural, locational, and neighborhood characteristics; all variables used in this analysis are adapted from Hong et al.'s 2019 study using RF modeling to predict urban residential property prices (Table 1).

Table 1: Explanatory variables used to predict sale price by Hong et al. (2019)

Category	Variable	Available from NYCOD?	Source (see Table 2)	Variable name	Units
Structural attributes	Elapsed year (transaction year - construction year)	Yes	RSD	elapsed_year	Years
	Area	Yes	PLUTO	area	Square feet
	Number of commercial units*	Yes	RSD	units_commercial	Count
	Number of residential units*	Yes	RSD	units_residential	Count
	Second most recent alteration year*	Yes	PLUTO	yearalter1	Year
	Most recent alteration year*	Yes	PLUTO	yearalter2	Year
	Floor level of a property	No			
	Number of buildings*	Yes	PLUTO	numbldgs	Count

	Number of floors*	Yes	PLUTO	numfloors	Count
	Heating system	No			
	Apartment brand	No			
	Floor area ratio (floor area/lot area)	Yes	PLUTO	FAR	Ratio
	Building coverage ratio (building footprint/lot area)	Yes	PLUTO	BCR	Ratio
	Top floor of an apartment	No			
	Bottom floor of an apartment	No			
Neighborhood attributes	Apartment brand	No			
	Number of units in the apartment complex	No			
	Number of buildings in the apartment complex	No			
	Historic district (binary)*	Yes	PLUTO	bin_histdist	1 if in historic district, 0 if not
	Limited height district (binary)*	Yes	PLUTO	bin_ltdheight	1 if in limited height district, 0 if not
	Split zone (binary)*	Yes	PLUTO	bin_splitzone	1 if in split zone area, 0 if not
	Landmarked (binary)*	Yes	PLUTO	bin_landmark	1 if in landmark area, 0 if not
Locational attributes	Latitude	Yes	PLUTO	latitude	Degrees
	Longitude	Yes	PLUTO	longitude	Degrees
	Distance to national park	Yes	Parks Properties	dist_park	Miles
	Distance to high school	Yes	Public Schools	dist_school	Miles
	Distance to redevelopment area	Yes	Affordable Housing	dist_redev	Miles
	Distance to university	Yes	Colleges and Universities	dist_university	Miles
	Distance to general hospital	Yes	Health + Hospital Facilities	dist_hospital	Miles
	Distance to museum	Yes	Museums	dist_museum	Miles
	Distance to subway station	Yes	Subway Stations	dist_subway	Miles
Macro variable	Transaction period	Yes	Rolling Sales Data	year_sold	Year
	US Gross Domestic Product per capita	Yes	World Bank	GDPpcc	Current dollars, not seasonally

					adjusted
	Regional Consumer Price Index	Yes	Bureau of Labor Statistics	CPI	Index value
	Regional Housing Price Index	Yes	Federal Housing and Finance Authority	HPI	Index value

*denotes variables that were not included in Hong et al.'s 2019 study but are relevant for this analysis

Data Sources

Data was compiled from numerous New York City Open Data (NYCOD) sources. The primary dataset is Rolling Sales Data (RSD), which lists all real estate transactions in the city of New York with sale date, sale price, and structural features such as year built and number of residential and commercial units (Table 1). These properties were geocoded (assigned cartesian coordinates) using the NYCOD Primary Land Use and Tax Lot Output (PLUTO) dataset; RSD observations were matched on either Borough-Block-Lot code or Address (Table 1). Once geographic coordinates were assigned to each property, distances were measured for distance-based independent variables: distance to nearest park, distance to nearest subway station, distance to nearest hospital, distance to nearest redevelopment area, distance to nearest museum, and distance to nearest college or university (Table 1). Finally, macroeconomic variables were joined to the dataset by month or month and year: national annual real Gross Domestic Product (GDP) per capita, regional monthly Consumer Price Index (CPI), and regional monthly Housing Price Index (HPI) (Table 1). Table 2 provides more detail on each data source.

Any sale registered in the city in a given period is included in the primary dataset, so is assumed to be exhaustive. The biggest issue is that data wasn't entered correctly or was omitted in the first place, which is explored further in the Data Description.

Table 2: Data sources

Published by (agency)	Data source (agency)	Dataset name	Description	Update frequency (last updated)	Utility	Projection
NYC Open Data[2]	NYC Department of Finance	Rolling Sales Data (RSD)	List of all real estate transactions with sale date, sale price, and	Annually (January 2023)	Sale price, sale date, year built, address, BBL, number of commercial units, number of residential	Not provided

			structural features		units, number of buildings, number of floors, building coverage ratio (BCR); binary variables indicating limited height district, split zone district, historic district	
NYC Department of Planning[3]	NYC Department of Finance, NYC Department of City Planning, NYC Department of Citywide Administrative Service, NYC Landmarks Preservation Committee	Primary Land Use and Tax Lot Output (PLUTO)	List of all unique tax lots with characteristics of the tax lot, the building, and the administrative district	Monthly (April 2020)	Latitude, longitude, year altered 1, year altered 2	Not provided
NYC Open Data[4]	NYC Parks' Planning and Development	Parks Properties	Shapefile of properties managed by NYC Parks' Planning and Development division	As needed (March 2023)	Distance to nearest park	WGS84
NYC Open Data[5]	NYC Department of Education (DOE)	Public Schools	Shapefile of point locations of all public schools	Annually (April 2019)	Distance to nearest school	NAD83; NAD83 Datum, GRS80 Spheroid[6]
NYC Open Data[7]	NYC Metropolitan Transportation Authority (MTA)	Subway Stations	Shapefile of point locations of all subway stations	As needed (August 2019)	Distance to nearest subway station	WGS84
NYC Open Data[8]	NYC Health + Hospitals	Health and Hospital Facilities	Shapefile of point locations of all public health and hospital facilities	As needed (July 2019)	Distance to nearest hospital	Not provided
NYC Open Data[9]	NYC Department of Housing Preservation and Development	Affordable Housing Production by	List of all affordable housing projects	Quarterly (February 2023)	Distance to nearest redevelopment area	Not provided

[2] https://www.nyc.gov/site/finance/taxes/property-rolling-sales-data.page
[3] https://nycplanning.github.io/db-pluto/#/
[4] https://nycopendata.socrata.com/Recreation/Parks-Properties/enfh-gkve
[5] https://data.cityofnewyork.us/Education/School-Point-Locations/jfju-ynrr
[6] Converted to WGS84 using qGIS; accuracy within 1 meter.
[7] https://data.cityofnewyork.us/Transportation/Subway-Stations/arq3-7z49
[8] https://data.cityofnewyork.us/Health/NYC-Health-Hospitals-Facilities-2011/ymhw-9cz9

	(HPD)	Building				
NYC Open Data[10]	NYC Department of Information Technology and Telecommunications (DoITT)	Museums	Shapefile of point locations of all museums	As needed (September 2018)	Distance to nearest museum	WGS84
NYC Open Data[11]	NYC Office of Technology and Innovation (OTI)	Colleges and Universities	Shapefile of point locations of all colleges and universities	As needed (September 2022)	Distance to nearest college or university	WGS84
World Bank[12]	World Bank	GDP by Country	List of GDP per capita for all countries, 1960-present	Annually (December 2021)*	US GDP per capita, aggregated yearly, 2017-2022	N/a
Bureau of Labor Statistics[13]	Bureau of Labor Statistics	New York-New Jersey CPI	List of NY-NJ CPI at monthly intervals	Monthly (March 2023)	NY-NJ CPI, aggregated monthly, 2017-2022	N/a
Federal Housing Finance Agency[14]	Federal Housing Finance Agency	Northeast Housing Price Index (HPI), Monthly Purchase-Only Index	List of regional HPI at monthly intervals	Monthly (March 2023)	Northeast US HPI, aggregated monthly, 2017-2022	N/a

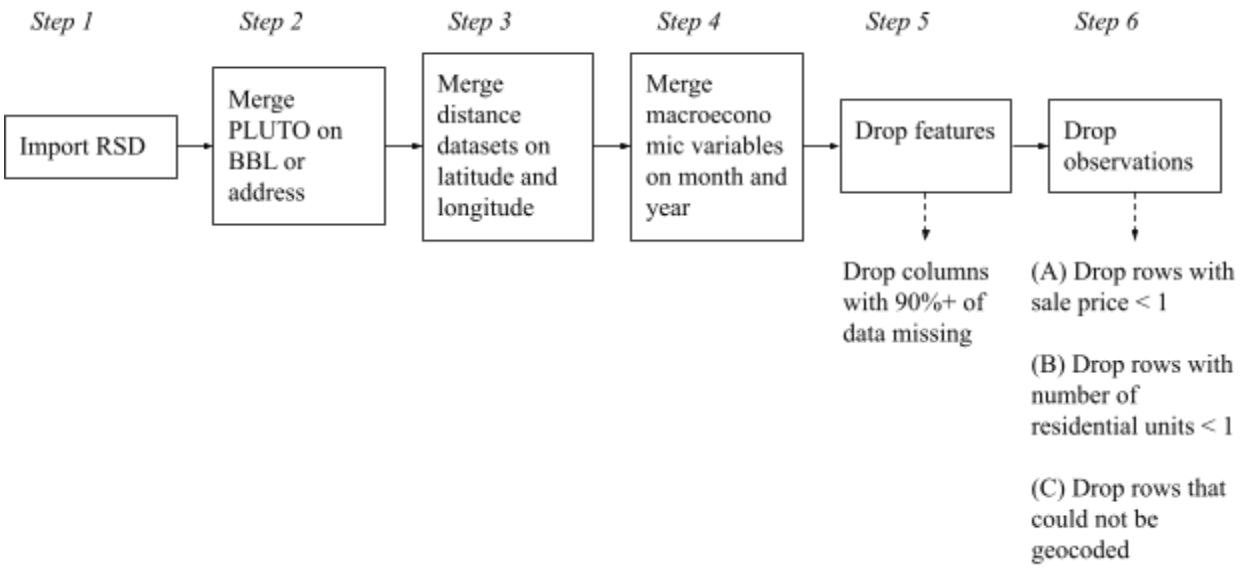
*2021 GDP per capita was used for 2022 sales because the 2022 number has not yet been included in the dataset at the time of this writing

Data Cleaning and Processing

Figure 3 shows the data cleaning process. (Note that the output of these steps is the input data table prior to any preprocessing required by the prediction models.) There was a significant amount of data lost throughout the data cleaning process, as elaborated on below and in the Limitations section at the end of the paper.

Figure 3: Data cleaning process

9 https://data.cityofnewyork.us/Housing-Development/Affordable-Housing-Production-by-Building/hg8x-zxpr
10 https://data.cityofnewyork.us/Recreation/New-York-City-Museums/ekax-ky3z
11 https://data.cityofnewyork.us/Education/Colleges-and-Universities/4kym-4xw5
12 https://data.worldbank.org/indicator/NY.GDP.PCAP.CD?locations=US
13 https://beta.bls.gov/dataViewer/view/timeseries/CUURS12ASA0
14 https://www.fhfa.gov/DataTools/Downloads/Pages/House-Price-Index-Datasets.aspx#mpo

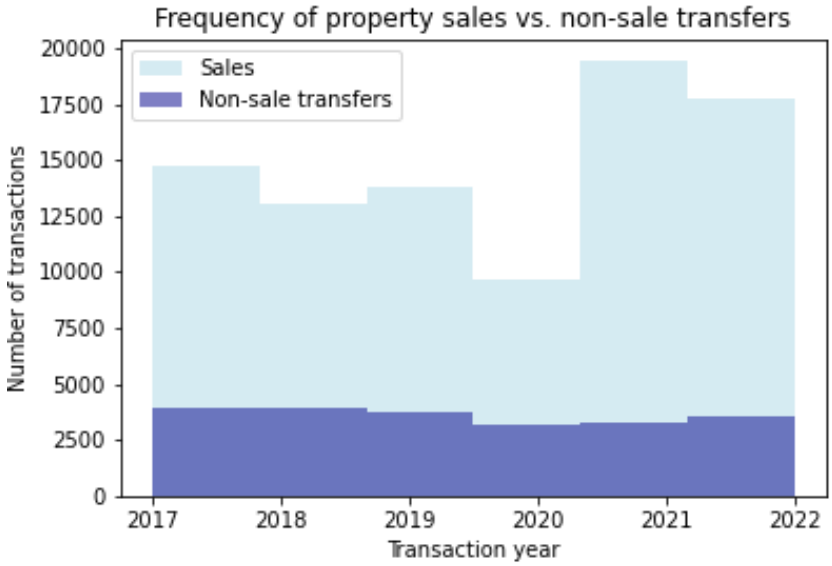


In Step 5, two features were missing 90% or more of their data: floor area ratio (FAR), which was missing 90.458% of data, and year of second-most recent alteration (yearalter2), which was missing 93.83% of data. The columns with the next-most missing data were the number of commercial units (units_commercial), with 58.23% missing, and the year of most recent alteration (yearalter1), with 50.057% of data missing. All other columns had under 45% of data missing. Floor area ratio and year of second-most recent alteration were dropped from the dataset.

Step 6 merits further discussion. In Step 6 (A) in Figure 3, sales were dropped whose listed sale price was less than $1 as these are considered "non-sale transfers" and are not expected to follow the same valuation logic as sales. According to the NYC Open Data data glossary, "[A] $0 sale indicates that there was a transfer of ownership without a cash consideration. There can be a number of reasons for a $0 sale including transfers of ownership from parents to children".[15] The frequency of sales and non-sale transfers is shown in Figure 4.

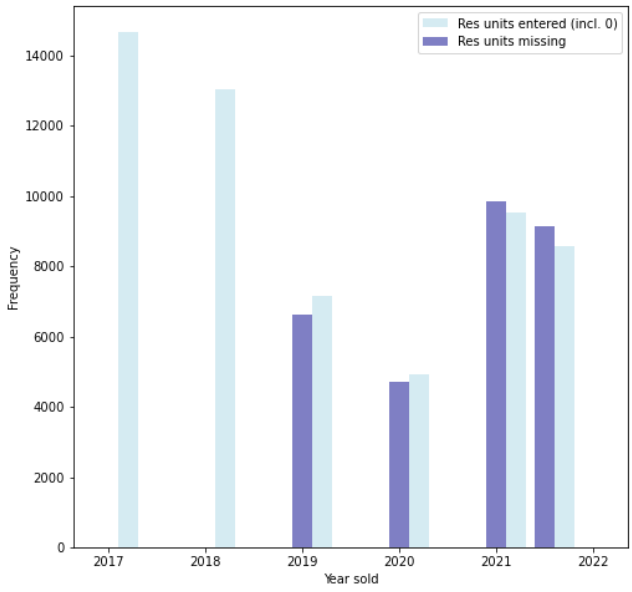
Figure 4: Frequency of sales vs. non-sale transfers

[15] www.nyc.gov/site/finance/taxes/glossary-property-sales.page



Step 6 (B) in Figure 3 was to exclude sales without any residential units. This step was unsuccessful because, upon inspecting the data, there were periods of time where the number of residential units was not entered (Figure 5). It cannot be assumed that missing values denote zero-values.

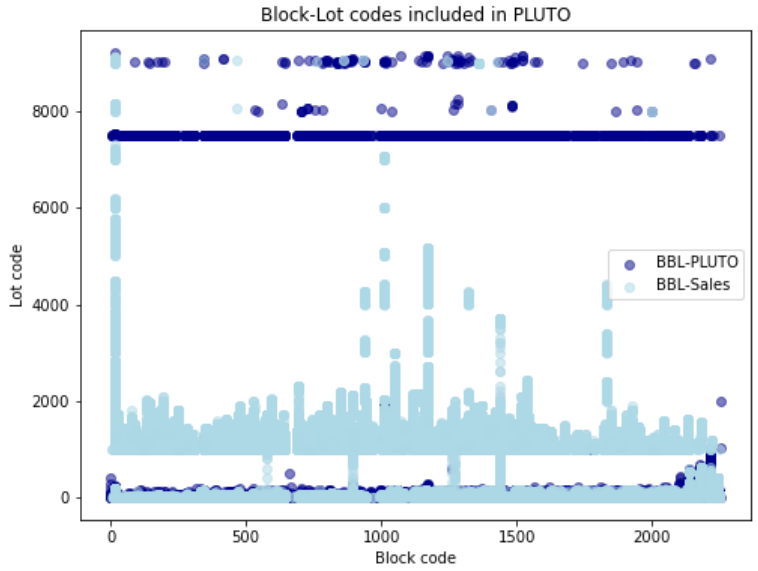
Figure 5: Data entry for residential units



Step 6 (C) in Figure 3 represents the most significant loss of data. In Step 2, properties were attempted to be joined from RSD to PLUTO to obtain cartesian coordinates (Figure 3). This was first attempted using each property's unique Borough-Block-Lot (BBL) code (as done in Hong et al. 2019). However, because not all BBL codes were included in the PLUTO database (Figure 6), properties that could not be joined by BBL were attempted to be joined by street addresses. Ultimately, the success of Step 2 was as follows:

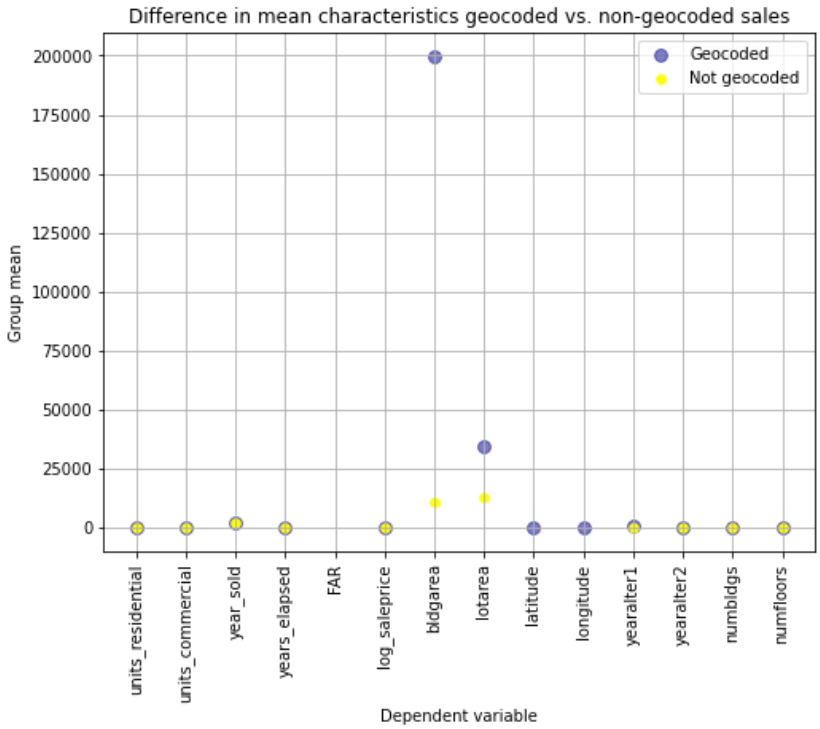
```
88224 total sales
43666 matched on BBL
18471 matched on address
26087 unmatched (29.569% of total)
```

Figure 6: BBL coverage in PLUTO vs. RSD



For the 29.6% of properties that could not be geocoded, no distance-based measures were able to be calculated. Thus, these observations were not able to be used. Figure 7 shows the difference in mean characteristics for properties that could and could not be geocoded.

Figure 7: Difference in mean characteristics for properties that could not be geocoded



The only apparent systematic differences between the two groups are building area and lot area. Properties that could be geocoded had significantly larger building areas and lot sizes on average than those that couldn't. The results of this study could be biased as a result of dropping these observations. This is discussed further in the Limitations section.

After compiling and cleaning the data, there were 62,137 observations and 24 features (as described in Table 9 and used by Hong et al. (2019)).

Methodology

The research question under investigation is how well OLS and RF models predict property sale prices in a period of macroeconomic volatility. The specific aim is to assess whether RF models are more consistently accurate in volatile periods than OLS and therefore whether they are a more useful tool for policymakers and those looking to conduct mass property valuation on an ongoing basis.

Methods

Overall, the methodological approach is to train OLS and RF models on the compiled dataset and compare their relative performance (also used by Hong et al. 2019 and Hu et al. 2019). Because the purpose of this study is to assess which architecture is more accurate as a foundation for policy measures, generalizability is important. Thus, we split the dataset chronologically so the first 80% of observations are used to train the models and the last 20% of observations are used to calculate model accuracy. Further, to avoid overfitting models to the training datasets, we run and evaluate two iterations of each model: one with all explanatory variables included, and one with the 10 most important explanatory variables as measured by RF feature importances (a similar approach is used in Hong et al. 2019). For RF regression, feature importance is defined as the average decrease in Mean Squared Error (MSE) that the feature is responsible for across all individual decision trees (Hong et al. 2019).

Preprocessing steps for each modeling approach are shown in Table 8. Preprocessors are fit on the training set and used to transform both the training set and the test set.

Table 8: Preprocessing steps required for each modeling approach

Data type	Pipeline step	OLS	RF	Strategy
Categorical	OHE	Yes		
Categorical	Impute missing	Yes	Yes	Most frequent
Continuous	Standardize	Yes		Z-score transformation (StandardScaler)
Continuous	Impute missing	Yes	Yes	Median*

*Median was chosen rather than mean to impute missing values for continuous variables because it is less sensitive to outliers; some values in the dataset were entered incorrectly so present as outliers even though their "true" values may not be (see Appendix).

OLS Model

The "conventional" form of the hedonic pricing model is OLS, which takes the form $y = \beta\chi + \varepsilon$, where y is a vector of the natural log of prices, χ is a matrix of explanatory variables, β is a matrix of coefficients corresponding to the explanatory variables, and ε is the vector of "white noise error" (Hong et al. 2019). Per the coefficient matrix β, OLS assumes that parameters' effects can be meaningfully captured linearly. It also assumes that the effects of

independent variables are separable and constant (Hong et al. 2019). OLS coefficients are chosen to minimize the distance between predicted and actual values; the loss function used to compute coefficients is effectively the sum of squared residuals, or $\sum \left(y_i - \widehat{y_i}\right)^2$.

Random Forest Model

A RF model is an ensemble of independently built decision trees and generates predictions by averaging the predictions of the individual trees (Hong et al. 2019; Potrawa and Tetereva 2022:58). Out of the entire collection of explanatory variables, only a random subset is used to grow each individual tree (Hong et al. 2019). RF employs bootstrap aggregation to simulate several new datasets sampled from the original dataset (with replacement), and then independent regression trees are created from each new dataset (Potrawa and Tetereva 2022:58). Each tree is constructed to minimize the impurity in each terminal node ("leaf"), as measured by mean squared error in the case of regression trees (Hong et al. 2019).

In addition to RF's limiting the number of original features that can be used in each tree, there are tunable hyperparameters that can be used to improve model performance and limit overfitting (Hong et al. 2019). These parameters - the maximum depth of each individual tree and the number of trees total - were tuned using Grid Search cross-validation and are discussed in the Results section.

Evaluation Metrics

For comparability, the same evaluation metrics were used as in Hong et al. (2019): Mean Absolute Percent Error (MAPE), R-squared, and the Coefficient of Dispersion (COD). Other evaluation metrics used throughout the literature include Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE) (Hu et al. 2019). However, MAPE, R-squared, and COD together provide a thorough picture of model performance (Hong et al. 2019).

MAPE is a standard measure in real estate modeling because it can be easily interpreted; it measures the average percentage of the deviation of predicted values from actual values (Hong et al. 2019):

$$MAPE = \frac{100}{n}\sum_{i=1}^{n}\left|\frac{\widehat{p_i}-p_i}{p_i}\right|$$

R-squared, similarly, is commonly used to quantify regression accuracy; it measures how much of the variation in the dependent variable is explained by the model (Hong et al. 2019):

$$R^2 = 1 - \left(\frac{\sum_{i=1}^{n}\left(p_i-\widehat{p_i}\right)^2}{\sum_{i=1}^{n}\left(p_i-\overline{p_i}\right)^2}\right)$$

COD measures the dispersion of each sales ratio - the predicted price divided by the actual price, for each sale - around the median sales ratio, with a lower COD signaling more uniform predictions and a higher COD showing that there is significant variation in prediction accuracy (Hong et al. 2019):

$$COD = \frac{100}{SR_m}\left(\frac{\sum_{i=1}^{n}\left|SR_i-SR_m\right|}{n}\right), \text{ where } SR_i = \frac{\widehat{p_i}}{p_i}.$$

A lower MAPE and higher R-squared would indicate a relatively more accurate model, while a lower COD would result from a model that predicts more consistently-accurately across all observations. While accuracy and consistency are both important in predicting real estate prices, consistency is of particular interest in assessing a model's robustness to volatility.

Results

Descriptive Statistics

Descriptive statistics for all variables are shown in Table 9, and the distribution of the dependent variable is shown in Figure 10. There is evidence of spatial autocorrelation, as shown in Figure 11. Finally, a correlation matrix of all independent variables is shown in Figure 12.

Table 9: Descriptive statistics

Variable name	Count	Mean	Standard deviation	Min.	25%	50%	75%	Max.
log_saleprice	62137	14.038757	1.487696	0.693147	13.345507	13.95273	14.731801	21.597693
units_residential	34762	3.603159	22.365354	0	0	1	1	899
units_commercial	25951	0.782051	8.905739	0	0	0	0	570
year_sold	62137	2019.854708	1.792338	2017	2018	2020	2021	2022
years_elapsed	59692	98.353598	234.641516	0	54	80	100	2018
latitude	62137	40.763037	0.033522	40.702776	40.737678	40.76319	40.780596	40.878084
longitude	62137	-73.976283	0.021194	-74.018118	-73.993629	-73.978224	-73.959475	-73.909048
yearalter1	31033	1991.630458	13.997397	987	1985	1987	1999	2021
numbldgs	59959	1.349339	2.144255	0	1	1	1	39
numfloors	59280	16.190418	12.555822	1	6	14	20	98
BCR	59898	8.849303	5.140569	0	4.353262	8.112054	12.133492	52.951082
bin_ltdheight	62137	0.019924	0.139739	0	0	0	0	1
bin_splitzone	62137	0.963854	0.186655	0	1	1	1	1
bin_histdist	62137	0.266299	0.442026	0	0	0	1	1
bin_landmark	62137	0.028872	0.167447	0	0	0	0	1
dist_park	62137	0.001567	0.001143	0	0.000629	0.001268	0.002176	0.006403
dist_subway	62137	0.003245	0.001953	0.000155	0.001816	0.002817	0.004148	0.010516
dist_hospital	62137	0.014633	0.008621	0.000209	0.008622	0.012274	0.018683	0.039067
dist_school	62137	0.002965	0.001739	0.000013	0.001674	0.002615	0.003837	0.010841
dist_housingdev	62137	0.003039	0.002137	0.000111	0.001379	0.002769	0.004141	0.012611
dist_college	62137	0.006266	0.003799	0.000001	0.003205	0.005641	0.00887	0.01947
dist_museum	62137	0.004895	0.002905	0.000004	0.002542	0.004476	0.006992	0.018043
CPI	62137	287.81544	15.433149	266.917	274.478	283.624	297.49	315.656

GDPpcc	47028	64957.04877	3823.962857	59907.75426	62823.30944	63530.63348	70248.629	70248.629
HPI	62137	256.602569	42.303634	192.19	218.52	251.62	291.89	320.83

Figure 10: Distribution of sale prices

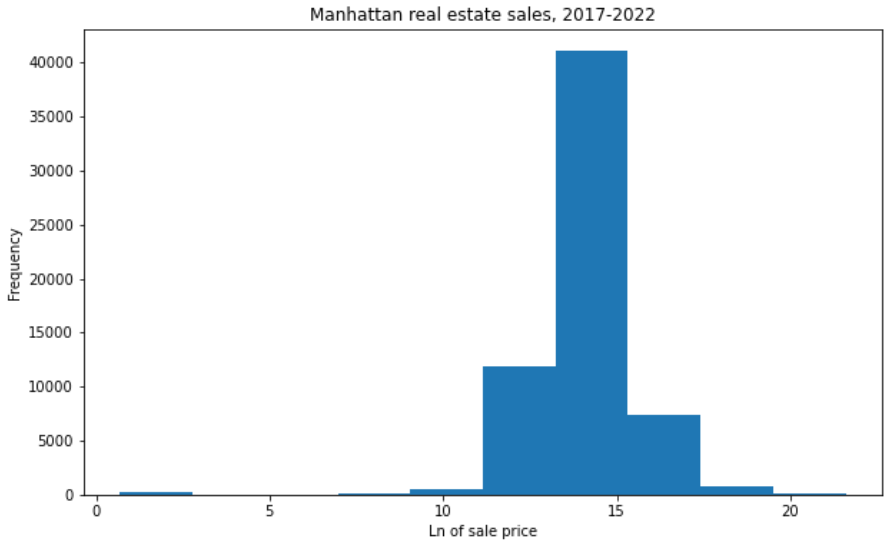
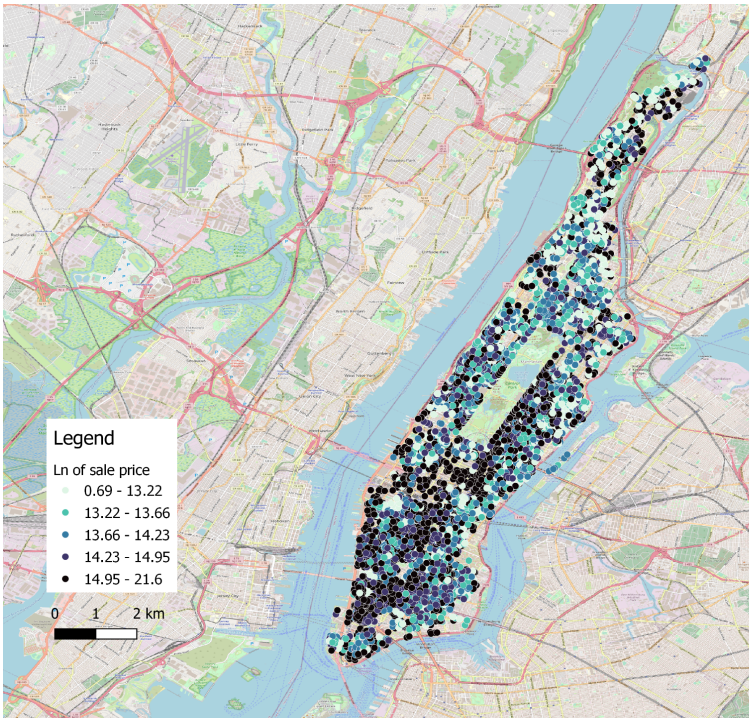
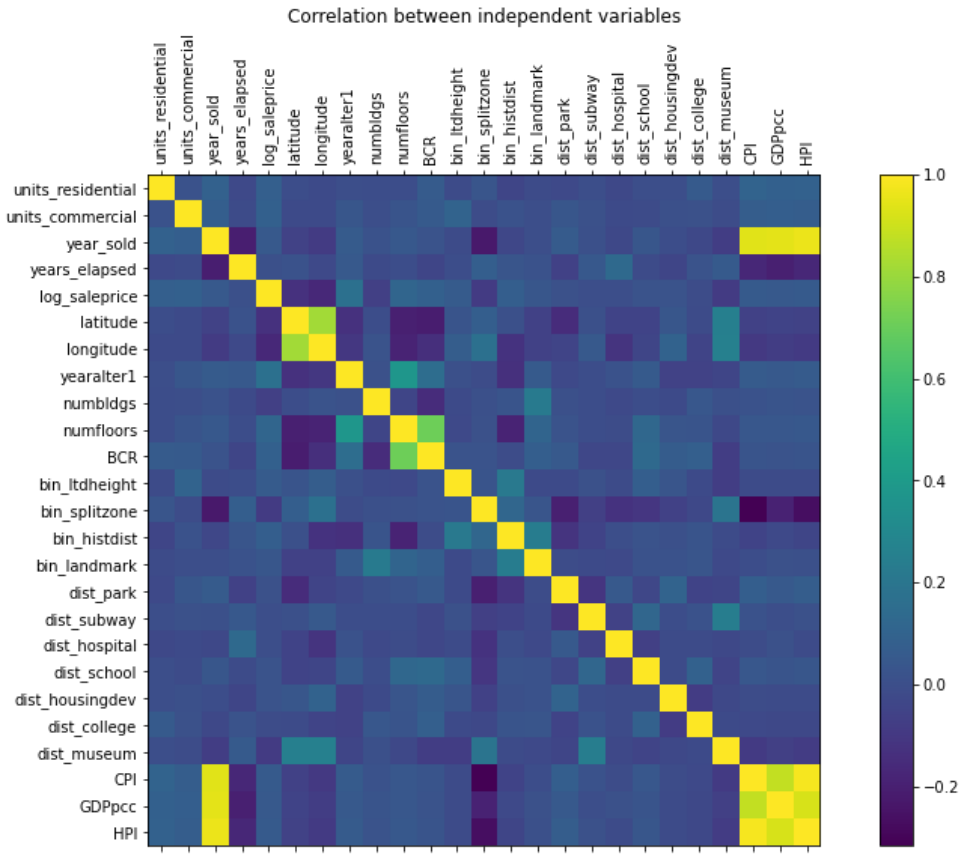


Figure 11: Spatial distribution of sale prices



This shows areas of potential clustering and spatial autocorrelation - specifically, high-high clustering (high values surrounded by other high values) on the west side of lower Manhattan and east side of midtown and upper Manhattan, and low-low clustering (low values surrounded by other low values) in Harlem and areas of upper Manhattan, signaling that residuals are likely not independent, which violates OLS assumptions (Figure 5). In theory, if a linear model is used, spatial diagnostics should be run and a Spatial Lag should likely be included to account for spatial dependence. However, given that OLS is commonly used for mass valuation in the public sector, it is more relevant to use in this comparison (Hong et al. 2019).

Figure 12: Correlation between all independent variables



There is no evidence of perfect multicollinearity, but there is high correlation among the macroeconomic variables (CPI, GDPpcc, and HPI) and year sold. This makes sense as they all describe macroeconomic conditions during the same time period.

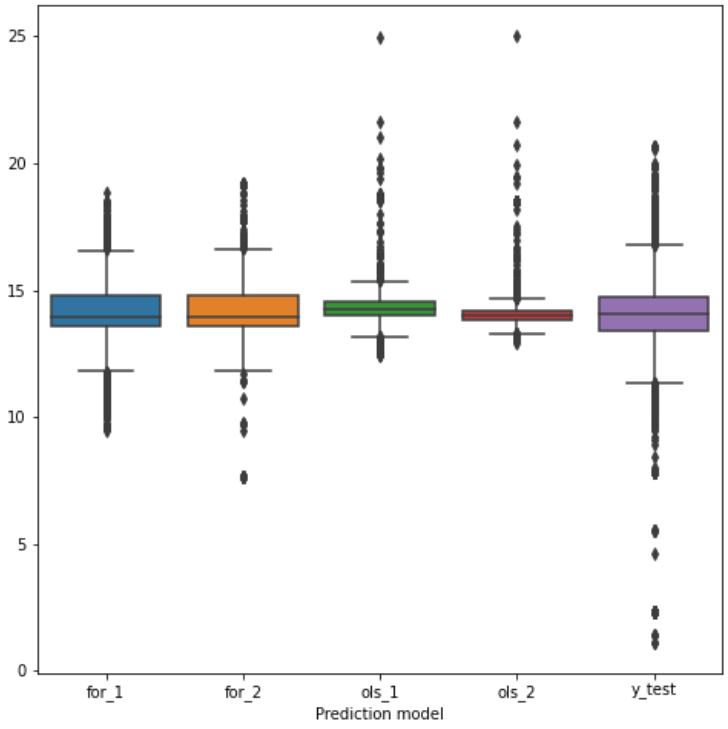
Model Accuracy

Table 13: Model accuracy metrics

Model	R-squared	MAPE	COD
RF, all features	0.189	2.526	7.132
RF, 10 features	0.226	2.752	6.964
OLS, all features	0.006	3.882	7.976
OLS, 10 features	-0.008	2.057	8.306

RF with 10 features has the highest R-squared value of 0.226, meaning that the model explains 22.6% of the variance in the dependent variable. This model also has the lowest COD (6.964), signaling that the predictions are the most consistently-accurate. However, it has the second-highest MAPE, so is not universally the best-performing. Interestingly, OLS with 10 features had the lowest MAPE despite having the lowest R-squared and highest COD. This could be partly explained by the distribution of predictions; OLS with 10 features had the most concentrated predictions around the mean, which could mean most predictions from the subsetted OLS model are lower than those from other models and MAPE would be biased downwards (Figure 14). The second-best-performing model was RF with all features included; it had the second-highest R-squared (0.189), second-highest MAPE (2.526), and second-lowest COD (7.132) (Figure 13). Thus, we can conclude that the RF models overall performed better than OLS models overall.

Figure 14: Distribution of predictions



Beyond strict model accuracy metrics, this boxplot highlights a critical issue of OLS: the bottom 25% and top 25% of sale prices are systematically over-predicted. While the models may perform similarly for the middle 50% of sale prices, the extremes are important to model correctly as they are directly linked to the "housing wealth gap" and often the biggest equity disparities (Cohen et al. 2022). Both RF models capture the *distribution* of sale prices far more accurately than either OLS model does.

Random Forest Feature Importances

Optimal hyperparameters were found for RF models using Grid Search cross-validation. Grid Search tests and selects the set of hyperparameters that optimizes model performance as defined by the user; in this case, 10-fold cross-validation was used and the accuracy metric was R-squared. The hyperparameters found to perform best are shown in Table 15.

Table 15: Best hyperparameters for RF models

Model architecture	Number of features	Parameter	Values tested	Optimal value	Reference
RF	21	max_depth	10, 20, 30, 40, 50, 60, 70	10	GridSearchCV (10-fold CV, r2 scoring)
		n_estimators	10, 20, 30, 40, 50, 60, 70	60	GridSearchCV (10-fold CV, r2 scoring)
RF	10	max_depth	10, 20, 30, 40, 50, 60, 70	10	GridSearchCV (10-fold CV, r2 scoring)
		n_estimators	10, 20, 30, 40, 50, 60, 70	70	GridSearchCV (10-fold CV, r2 scoring)

The most important features as defined by those responsible for the largest average decrease in MSE across all individual decision trees in the initial RF model were years elapsed (the difference between the year of sale and year built), number of commercial units, building coverage ratio, latitude, longitude, year of most recent alteration, number of residential units,

number of floors, distance to nearest hospital, and whether a building was in a historic district (Figure 16).

Figure 16: Random Forest feature importances, all explanatory variables

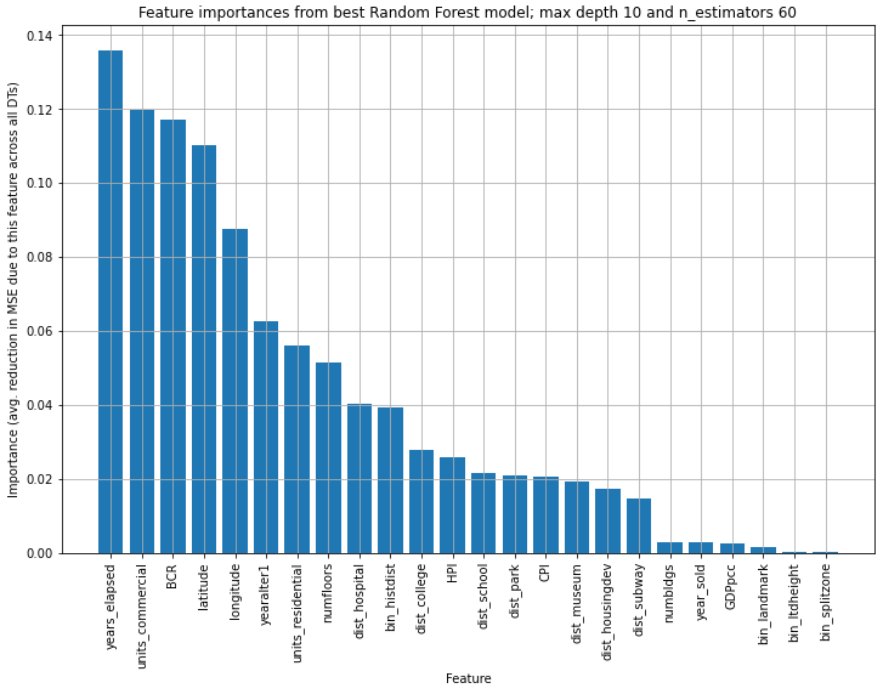
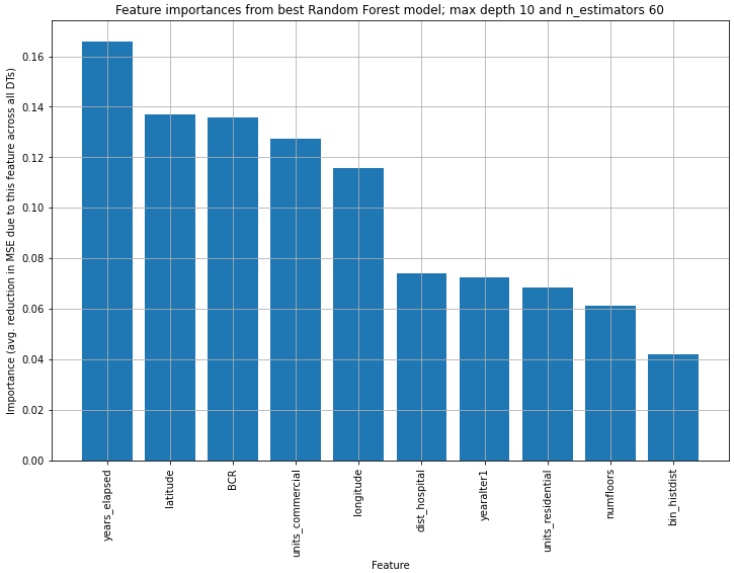


Figure 17: Random Forest feature importances, 10 most "important" explanatory variables



Interestingly, the full and subsetted RF models do not agree on the order of feature importances. In the full model, years elapsed was the most important feature, followed by

number of commercial units, building coverage ratio, latitude, and longitude (Figure 16). In the subsetted model, years elapsed was still the most important feature, but latitude was second and building coverage ratio was third (Figure 17). The total numbers are also higher in the subsetted model, with the years elapsed variable accounting for an average of 0.13 reduction in MSE in the full model and about 0.165 in the subsetted model (Figure 16, Figure 17). This could signal that the features that were ultimately dropped are noisy or correlated with other features (Figure 12). It's also interesting that the only distance-based measure within the top 10 most important features is distance to the nearest hospital (Figure 16).

OLS Coefficients

Figure 18: OLS coefficients, all explanatory variables

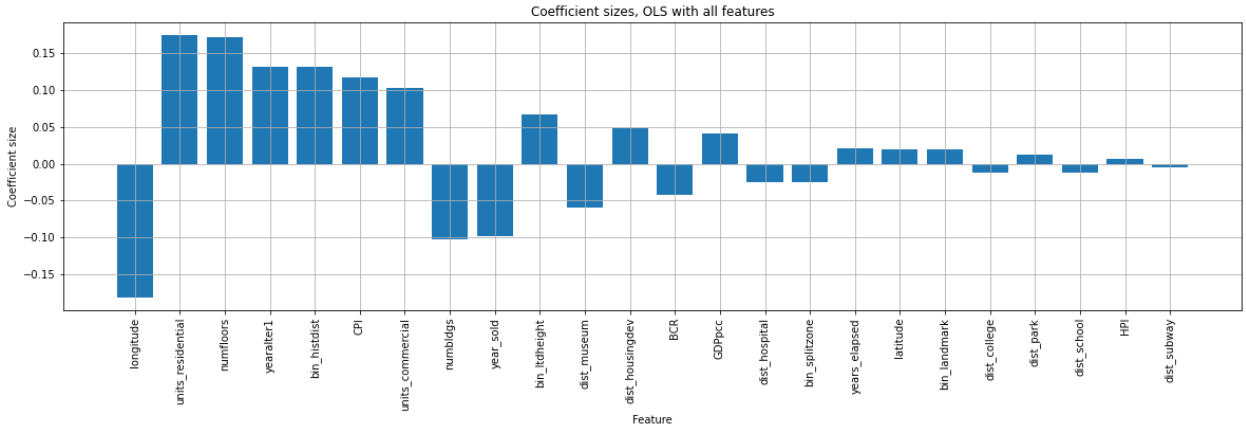
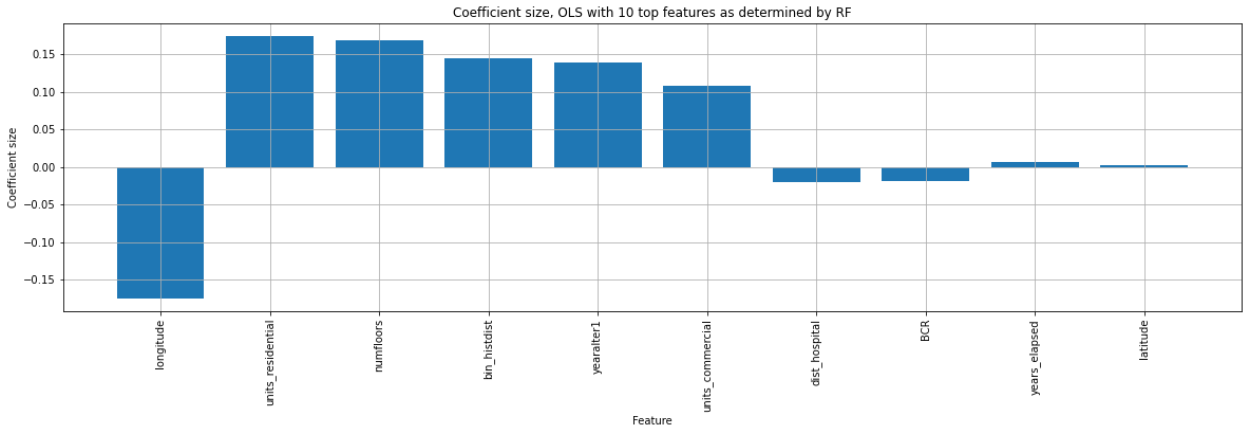


Figure 19: OLS coefficients, 10 most "important" explanatory variables



For the OLS model with all variables, the biggest coefficients (in absolute value) were longitude, number of residential units, number of floors, and year of most recent alteration (Figure 17). In the subsetted model, the biggest coefficients (in absolute value) were longitude, number of residential units, number of floors, and whether a property was in a historic district (Figure 18). It is not surprising that they mostly agree on which coefficients are relatively largest. However, p-values are not represented in this output, so we don't know the statistical significance of these coefficients beyond their sizes.

Discussion

Although we can't conclusively reject or fail to reject the null hypothesis that the models perform similarly, there is strong evidence that RF models overall outperformed OLS models overall (Table 13).

A "false-positive" conclusion would be concluding that RF performs better while it doesn't in practice. There is little at stake in this error; policymakers' using RF would not not change predictive accuracy and therefore policy efficiency would remain constant. A "false-negative", on the other hand, would be concluding that the models perform similarly while in fact RF outperforms OLS. This kind of error could have significant and very specific repercussions. In examining the distribution of RF and OLS predictions compared to true values, we see similar predictions for the middle 50% of sale prices. However, OLS systematically overestimates sale prices for the lowest 25% of sale prices and does not capture the distribution accurately for the highest 25% of sale prices (Figure 14). In implementing policies and allocating resources, it is critical to accurately identify the area of highest need. Using OLS for prediction, which is least accurate for the lowest 25% of prices, does not seem appropriate and could lead to a misallocation of resources.

Limitations

There are numerous substantive and methodological limitations that could affect the generalizability of these findings. First, close to 30% of observations were lost due to geocoding errors in the data cleaning process. While there were no apparent differences between these sales

and the ones that were successfully geocoded beyond property and lot area, there could be systematic differences in characteristics not captured in this dataset; in that case, dropping these observations would produce biased predictions.

Additionally, data entry inconsistencies could limit how applicable these findings are. Sales for commercial buildings were not able to be excluded from the dataset because the number of residential units was not entered for certain years, which could produce systematic errors. There were also a few impossible values resulting from apparent data entry errors (see Appendix); these errors are assumed to be random but it is impossible to know whether they are systematic according to certain property characteristics without identifying all of them.

There could be other omitted variables that could meaningfully impact predictions (or could impact predictions disproportionately for OLS as compared to RF and change the relative model performance). The explanatory variables included here are based on previous literature and availability. This includes distance measurements; using Euclidean distance from sales to relevant features such as schools or subway stations was based on previous research but could be an incomplete measure as linear travel is rarely possible (Hong et al. 2019). A better metric, if available, could be travel time or actual distance required to travel.

Helbich et al. (2013) assert that traditional real estate valuation holds room for improvement through both better data and better algorithms (Helbich et al. 2013). As this is not an exhaustive review of modeling approaches, it is possible that a different model architecture could yield better predictions or be more robust to volatility. An extension of these findings would be comparing additional model architectures beyond these two.

Conclusion

Effects of the Covid-19 pandemic, including both lockdown policies and tumultuous economic conditions, illuminated the connection between housing and health in New York City specifically, highlighting the need for valuation models that are robust to market volatility and provide useful estimates for policymakers in times of crisis (Robbins 2022:610). Incorrect property valuation can render well-intentioned policies inefficient at best and ineffective at

worst; these results show that using OLS for mass appraisal could potentially exacerbate existing equity disparities by failing to capture the true distribution of property sale prices, and that RF could serve as a viable alternative.

References

Abidoye, Rotimi Boluwatife and Albert P.C. Chan. 2017. “Critical review of hedonic pricing model application in property price appraisal: A case of Nigeria.” *International Journal of Sustainable Built Environment* 6(1):250-259. https://doi.org/10.1016/j.ijsbe.2017.02.007.

Ali Kully, Sadef. 2020. “A Look at Rent Strikes in NYC, as Housing Relief Legislation Lags.” *City Limits*, 11/17/2020.

Antipov, E. A. and E. B. Pokryshevskaya. 2012. “Mass appraisal of residential apartments: An application of Random Forest for valuation and a CART-based approach for model diagnostics.” *Expert Syst. Appl.* 39:1772-1778.

Blaha, J. 2017. “Variable selection methods for residential real estate markets: An exploration of random forest trees in spatial economics.” Available from ProQuest Dissertations and Theses Global. (2199540989).

Breiman, L. 2001. *Random forests.* Mach. Learn. 45:5–32.

Canelas, Patricia, and Idalina Baptista. 2021. "Guerrilla urbanism, guerrilla governance: governing neighborhoods in 'with-COVID' times." *Town Planning Review* 92(3):279.

Chapple, K. 2009. “Mapping susceptibility to gentrification: The early warning toolkit.” *Center for Community Innovation*.

Chen, L., X. Yao, Y. Liu, Y. Zhu, W. Chen, X. Zhao, and T. Chi. 2020. “Measuring impacts of urban environmental elements on housing prices based on multisource data—A case study of Shanghai, China.” *International Journal of Geo-Information* 9:106.

Chin, T. L. and K. W. Chau. 2003. “A critical review of literature on the hedonic price model.” *International Journal for Housing and its Applications* 27(2):145-165.

Cohen, J. P., F. L. Friedt, and J. P. Lautier. 2022. “The impact of the Coronavirus pandemic on New York City real estate: First evidence.” *Journal of Regional Science* 62:858-888.

d'Amato, M. and T. Kauko. 2017. *Advances in automated valuation modeling*, Springer, Berlin.

Daher-Nashif, S. 2021. “In sickness and in health: The politics of public health and their implications during the COVID-19 pandemic.” *Sociology Compass* e12949.

Dietzel, M. Alexander, N. Braun, and W. Schafers. 2014. “Sentiment-based commercial real estate forecasting with google search volume data.” *J. Prop. Invest. Finance* 32(6):540–569.

Fotheringham, S., R. Crespo, and J. Yao. 2015. “Exploring, modeling and predicting spatiotemporal variations in house prices.” *Ann. Reg. Sci.* 54:417– 436.

Fu, Y., H. Xiong, Y. Ge, Z. Yao, Y. Zheng, and Z.-H. Zhou. 2014. "Exploiting geographic dependencies for real estate appraisal: A mutual perspective of ranking and clustering." *Proceedings of the 20th ACMSIGKDD International Conference on Knowledge Discovery and Data Mining, ACM* 1047–1056.

Gelman, Andrew and Zaiying Huang. 2008. "Estimating Incumbency Advantage and Its Variation, as an Example of a Before-After Study." *Journal of the American Statistical Association* 103(482):437-446.

Gelman, Andrew. 2005. "The secret weapon." *Statistical Modeling, Causal Inference, and Social Science*, 03/07/2005, https://statmodeling.stat.columbia.edu/2005/03/07/the_secret_weap/.

Głuszak, M. (2018). "Externalities and house prices: A stated preferences approach." *Entrepreneurial Business and Economics Review* 6(4):181. https://doi.org/10.15678/EBER.2018.060410.

Gokmenoglu, K., and S. Hesami. 2019. "Real estate prices and stock market in germany: Analysis based on hedonic price index." *International Journal of Housing Markets and Analysis* 12(4):687-707. https://doi.org/10.1108/IJHMA-05-2018-0036.

Gröbel, Sören and Lorenz Thomschke. 2018. "Hedonic pricing and the spatial structure of housing data – an application to Berlin." *Journal of Property Research* 35(3):185-208. 10.1080/09599916.2018.1510428.

Guan, J. et al. 2014. "Analyzing massive data sets: An adaptive fuzzy neural approach for prediction, with a real estate illustration." *J. Organ. Comput. Electron. Commerce* 24: 94–112.

Haag, M. 2021. "500,000 New Yorkers Owe Back Rent: What Happens When Evictions Resume." *New York Times*, 07/28/2021.

Helbich, M. et al. 2013. "Boosting the predictive accuracy of urban hedonic house price models through airborne laser scanning." *Comput. Environ. Urban. Syst.* 39:81–92.

Hu, Lirong, Shenjing He, Zixuan Han, He Xiao, Shiliang Su, Min Weng, and Zhongliang Cai. 2019. "Monitoring housing rental prices based on social media:An integrated approach of machine-learning algorithms and hedonic modeling to inform equitable housing policies." *Land Use Policy* 82: 657-673. https://doi.org/10.1016/j.landusepol.2018.12.030.

Hong, J., H. Choi, and K. Woo-sung. 2020. "A house price valuation based on the random forest approach: The mass appraisal of residential property in South Korea." *International Journal of Strategic Property Management* 24(3):140-152. https://doi.org/10.3846/ijspm.2020.11544.

Jim, C. Y. and Wendy Y. Chen. 2007. “Consumption preferences and environmental externalities: A hedonic analysis of the housing market in Guangzhou.” *Geoforum* 38(2):414-431. https://doi.org/10.1016/j.geoforum.2006.10.002.

Joseph, D. S. 1937. “The assessment of real property in the United States.” *Special Report of the State Tax Commission, New York* 10.

Kiely, T. J., and N. D. Bastian. 2020. “The spatially conscious machine learning model.” *Stat Anal Data Min: The ASA Data Sci Journal* 13:31-49.

Kontrimas, V. and A. Verikas. 2011. “The mass appraisal of the real estate by computational intelligence.” *Appl. Soft Comput.* 11(1):443–448.

Koschinsky, J., N. Lozano-Gracia, and G. Piras. 2012. “The welfare benefit of a home's location: An empirical comparison of spatial and non-spatial model estimates.” *J. Geogr. Syst.* 14(3):319–356.

Long, Rebecca G. 2008. “The Crux of the Method: Assumptions in Ordinary Least Squares and Logistic Regression.” *Psychological Reports* 103(2): 431-434. https://doi.org/10.2466/pr0.103.2.431-434.

Mathur, A., Moschis, G.P. and E. Lee. 2008. “A longitudinal study of the effects of life status changes on changes in consumer preferences.” *J. of the Acad. Mark. Sci.* 36:234–246. https://doi-org.ezproxy.cul.columbia.edu/10.1007/s11747-007-0021-9

Mason, C. 1996. “Non‐parametric hedonic housing prices.” Taylor and Francis. doi:10.1080/02673039608720863.

Mironova, O. 2019. “Defensive and Expansionist Struggles for Housing Justice: 120 Years of Community Right in New York City.” *Radical Housing Journal* 1(2):135–152.

Nerlove, Marc. 1995. “Hedonic price functions and the measurement of preferences: The case of Swedish wine consumers.” European Economic Review 39(6):1697-1716. https://doi.org/10.1016/0014-2921(95)00013-5.

Pace, R. K., Zhu, S. 2017. “Implicit Hedonic Pricing Using Mortgage Payment Information.” *J Real Estate Finan Econ*, 54:387–402. https://doi.org/10.1007/s11146-016-9578-8.

Park, B. and J. K. Bae. 2015. “Using machine learning algorithms for housing price prediction: The case of Fairfax County, Virginia housing data.” *Expert Syst. Appl.* 42:2928–2934.

Pivo, G. and J. D. Fisher. 2011. “The walkability premium in commercial real estate investments.” *Real Estate Econ.* 39(2):185–219.

Potrawa, Tomasz and Anastasija Tetereva. 2022. “How much is the view from the window worth? Machine learning-driven hedonic pricing model of the real estate market.” *Journal of Business Research*, 144:50-65.

Rafiei, M. H. and H. Adeli. 2015. “A novel machine learning model for estimation of sale prices of real estate units.” *J. Constr. Eng. Manag.* 142(04015066).

Rajan, K. (2020). Influence of hedonic and utilitarian motivation on impulse and rational buying behavior in online shopping. Journal of Statistics and Management Systems. 23. 419-430. 10.1080/09720510.2020.1736326.

Rico-Juan, Juan Ramón and Paloma Taltavull de La Paz. 2021. ”Machine learning with explainability or spatial hedonics tools? An analysis of the asking prices in the housing market in Alicante, Spain.” *Expert Systems with Applications*, 171. https://doi.org/10.1016/j.eswa.2021.114590.

Robbins, Glyn. 2022. “New York’s housing justice movement: facing the COVID eviction cliff edge.” *City*, 26(4):610-629.

Schernthanner, H. et al. 2016. “Spatial modeling and geo visualization of rental prices for real estate portals.” *Comput. Sci. Appl.* 9788.

Taylor, L. O. 2008. “Theoretical Foundations and Empirical Developments in Hedonic Modeling.” Hedonic Methods in Housing Markets. Springer, New York, NY. https://doi-org.ezproxy.cul.columbia.edu/10.1007/978-0-387-76815-1_2.

Turner, H. 2008. “Gnm: A package for generalized nonlinear models.” Department of Statistics University of Warwick, University of Warwick, UK.

Unel, F. B., et al. "PREFERENCE CHANGES DEPENDING ON AGE GROUPS OF CRITERIA AFFECTING THE REAL ESTATE VALUE." *International Journal of Engineering and Geosciences* 2(2):41.

Wallace, Deborah, and Rodrick Wallace. 2020. “COVID-19 in New York City: An Ecology of Race and Class Oppression.” Springer International Publishing AG.

Wei, C., Fu, M., Wang, L., Yang, H., Tang, F., and Xiong, Y. (2022). The research development of hedonic price model-based real estate appraisal in the era of big data. *Land* 11(3):334. https://doi.org/10.3390/land11030334.

Zhao, Y. 2020. “U.S. housing market during COVID-19: Aggregate and distributional evidence.” *COVID Economics* 50:113–154.

Zuk, M., et al. 2015. “Gentrification, displacement and the role of public investment: A literature review.” *Federal Reserve Bank of San Francisco* 79.