# Predicting Real Estate Sales Using Machine Learning and Spatial Dependence

Boosting ML Predictive Accuracy Using Spatial Lags

## Contents

# Introduction

In this paper, we explore a technique for more accurately predicting real estate transactions, both their occurrence (probability of sale) and their dollar amount (sale price per square foot). We explain how this predictive technique may be applied to combat Economic Exclusion, a precursor to Income Inequality. The technique marries the use of machine-learning predictive models (Random Forrest) with "spatial-lag" features typically seen in geographically-weighted regressions (GWR). We find that, while the addition of many new variables to a modeling data set can inhibit the models' ability to generalize into the future, spatial-lag features 1) consistently outperform zip-code level aggregation features, and 2) outperform all models for specific types of properties in specific areas. We conclude that spatial-lag features, while computationally expensive, can be used to significantly increase the predictive accuracy of spatial predictive models.

## What is Economic Exclusion?

Income inequality may be a defining challenge of our time, yet the causes of inequality remain unclear. As discussed by Zuk (2015), "Neighborhoods change slowly, but over time [they] are becoming more segregated by income, due in part to macro-level increases in income inequality". Researchers at the Urban Institute (Solomon Greene and Lei 2016) recently identified the socio-economic phenomenon of "Economic Exclusion" as an explanation for income inequality in the US. Economic Exclusion can be explained as follows: vulnerable populations–disproportionately communities of color, immigrants, refugees, and women–who are physically displaced by local economic prosperity can enter into a gradual cycle of diminished access to good jobs, good schools, health care facilities, public spaces, etc. Diminished access leads to more poverty, which leads to more displacement. Such self-reinforcing poverty gradually exacerbates income inequality over the course years and even generations.

## Predicting Gentrification

One practical way to combat Economic Exclusion is to focus on preventing displacement, i.e., the physical relocation of populations away from economic resources. As argued by Clay (1979), displacement is the negative consequence of gentrification. Predicting gentrification at an early stage, however, has proven to be a difficult task historically. When an area experiences economic growth, increased housing demands and subsequent affordability pressures can lead to voluntary or involuntary relocation of low-income families and small businesses. Government agencies and nonprofits tend to intervene once displacement is already underway, and after-the-fact interventions can be costly and ineffective. As explained by Solomon Greene and Lei (2016), there are several preemptive actions that can be deployed to stem divestment and ensure that existing residents benefit from new investments. Not unlike medical treatment, early detection is the key to success. Consequently, in 2016, the Urban

Institute put forth a call for research into the creation of "neighborhood-level early warning and response systems that can help city leaders and community advocates get ahead of neighborhood changes" (Solomon Greene and Lei 2016). This paper explores a technique to answer that call in part.

# Literature Review

We review Economic Displacement as it has been addressed in academia, primarily in relation to the study of gentrification. We also examine "mass appraisal techniques", which are automated analytical techniques used for valuing large numbers of real estate properties. Finally, we will briefly examine machine learning as it relates to the problem of predicting gentrification and/or Economic Displacement.

## How Has Economic Displacement Been Addressed in the Past?

Economic Displacement has been intertwined with the study of gentrification since shortly after the latter became academically relevant in the 1960's. The term "gentrification" was first used by Ruth Glass in 1964 to described the "gentry" in low income neighborhoods in London. Gentrification was originally understood as a "tool of revitalization for declining neighborhoods" (Zuk 2015), however, in 1979 Phillip Clay made the distinction between two types of revitalization: "incumbent upgrading" and "gentrification", noting that Economic Displacement was the negative consequence of the latter (Clay 1979). Today, the term has evolved to describe "a spatial organization and re-organization of human dwelling and activity" (Zuk 2015). Specific to cities, gentrification is thought of as "the transformation of a working-class or vacant area of the central city into middle-class residential or commercial use" (Lees 2008).

Studies of gentrification and displacement generally take two approaches in the literature: supply-side and demand-side, or "the flows of capital versus flows of people to neighborhoods", respectively (Zuk 2015). Supply side arguments for gentrification tend to focus on "private capital investment, public policies, and public investments" (Zuk 2015). Smith (1979) argued that the return of capital from the suburbs to the city drives gentrification. He describes a "political economy of capital flows into urban areas" as largely responsible for both the positive and negative consequences of gentrification. According to Dreier (2004), public policies that have been linked to increased Economic Displacement have been, among others, automobile-oriented transportation infrastructure spending and mortgage interest tax deductions for home owners.

More recently, income inequality has been explored as a consequence of Economic Displacement, defined as "higher compensation in the top quintile and the lack of jobs for the bottom quintile" (Reardon 2011); (Watson 2009). The concentration of wealth allows "certain households to sort themselves according to their preferences – and control local political processes that continue exclusion" (Reardon 2011). This results in a self-reinforcing feedback loop where wealthier households influence public policy toward their self interest. Gentrification prediction tools could be used to help break such feedback loops through early identification and intervention.

Many studies conclude that gentrification in most forms leads to Economic Displacement, however, Zuk (2015) characterizes the results of many recent studies as "mixed, due in part to methodological shortcomings". In this paper, we attempt to further the understanding of gentrification prediction by demonstrating a technique to better predict real estate sales in New York City.

## A Review of Mass Appraisal Techniques

Much of the research on predicting real estate transaction has been in service of "mass appraisal" models, or models that are used to value large numbers of buildings automatically. Such techniques are commonly used by governments for the purposes of collecting taxes from property owners. Mass appraisal models share many characteristics with predictive machine learning models (and are not mutually exclusive), in that they are data-driven, standardized methods that employ statistical testing (Eckert 1990). A variation on mass appraisal models are the "automated valuation models" (AVM), which use "often the same methodological framework of mass appraisal… a statistical model and a large amount of property data to estimate the market value of an individual property or portfolio of properties" (d'Amato 2017).

Scientific mass appraisal models date back to 1936 with the reappraisal of St. Paul, Minnesota (Silverherz 1936). Since that time, and accelerated with the advent of computers, much statistical research has been done relating property values and rent prices to various characteristics of those properties, including characteristics of their surrounding areas. Multiple regression analysis (MRA) has been the most common set of statistical tools used in mass appraisal, including Maximum Likelihood, Weighted Least Squares, and the most popular, Ordinary Least Squares (OLS) (d'Amato 2017). The primary drawbacks of MRA techniques are "excessive multicollinearity among attributes" and "spatial autocorrelation among residuals" (d'Amato 2017). Another group of models that seek to correct for spatial dependence are known as Spatial Auto Regressive (SAR) models, chief among them the Spatial Lag Model, which aggregates weighted summaries of nearby properties in order to create independent regression variables (d'Amato 2017). Geographic Weighted Regressions frequently employ spatial lags.

Hedonic regression modeling is the practice of defining and quantifying the components of a price of a good based on the intrinsic and extrinsic characteristics. Koschinsky (2012) is a recent and thorough discussion of parametric hedonic regression techniques. Some of the variables included in Koschinsky's models are derived from nearby properties, similar to the technique used in this paper, and these variables were found to be predictive. The real estate hedonic model as defined by Koschinsky describes the price of a property as:

$$P_i = P(S_i, N_i, L_i)$$

Where $P_i$ represents the price of house $i$, which is a composite good comprised of a vector of structural characteristics $S$, a vector of social and neighborhood characteristics $N$, and a vector of locational characteristics $L$. The model calculates spatial lags for properties of interest using neighboring properties within 1,000 feet of a sale. The derived variables include characteristics such as average age, quantity of poor condition homes nearby, percent of homes with electric heating nearby, construction grade, etc. Koschinsky found that in all cases, "the relation between a home's price and the average price of its neighboring homes is characterized by positive spatial autocorrelation" meaning that homes near each other were typically similar to each other and priced accordingly. Koschinsky concluded that locational characteristics should be valued at least as much "if not more" than structural characteristics.

As recently as 2015, much research has dealt with mitigating the drawbacks of spatial multiple regression analysis through the use of multi-level hierarchical models. Fotheringham (2015) explored the combination of Geographically Weighted Regression (GWR) with time-series forecasting to predict home prices over time. He used "adaptive bandwidths" of local data, i.e., for each estimate, the number of data points included varied was optimized using cross-validation. Adaptive bandwidths are an interesting extension to spatial lag models and are included in the "Future Research" portion of this paper.

Automated valuation modeling got a legal update in the aftermath of the 2008 financial crisis by way of the The Dodd Frank Act. In particular, the Title XIV, subtitle F distinguishes the "appraisal" process from automated valuation modelling, and reorganized both (d'Amato 2017). The Act asserts that appraisal, or valuation conducted by a human being, cannot be replaced by AVM. At current, AVM is "increasingly adaptable in describing real estate market behavior" but has yet to supersede the importance and necessity of local information and human evaluation.

## Has Machine Learning Been Applied to this Problem Before?

Both Mass Appraisal techniques and Automated Valuation Modeling seek to predict real estate prices using data and statistical methods, however, traditional techniques typically fall short of reality. This is because property valuation is inherently a "chaotic" process that does not lend itself to binary or linear analysis (Zuk 2015). The value of any given property is a complex combination of perceived value and speculation. The value of any building or plot of land belongs to a rich network where decisions about and perceptions of neighboring properties influence the final market value. Guan et al. (2014) compared traditional MRA techniques to alternative "data mining techniques" resulting in "mixed results". However, as Helbich (2013) states, hedonic pricing models "can be improved in two ways: (a) Through novel estimation techniques, and (b) by ancillary structural, locational, and neighborhood variables on the basis of Geographic Information System (GIS)". Recent research generally falls into these two buckets: better analysis algorithms and/or better data.

In the "better data" category, researchers have been striving to introduce new independent variables to increase the accuracy of predictive models. Dietzell (2014) successfully used internet search query data provided by Google Trends to serve as a sentiment indicator and improve commercial real estate forecasting models. Pivo and Fisher (2011) examined the effects of walkability on property values and investment returns. They found that on a 100-point scale, a 10-point increase in walkability increased property investment values by up to 9%.

Research into better prediction algorithms do not necessarily happen at the exclusion of "better data". For example, Fu (2014) created a prediction algorithm, called "ClusRanking", for real estate in Beijing, China. ClusRanking first estimates neighborhood characteristics using taxi cab traffic vector data, specifically as they relate to nearby "business areas". Subsequently, the algorithm performs a rank-ordered prediction of investment returns segmented into five categories. Similar to Koschinsky (2012), though less formally stated, Fu (2014) thought of a property's value as a composite of individual, peer and zone characteristics. In the predictive model, Fu includes characteristics of the neighborhood (individual), the values of its nearby properties (peer), and the prosperity of the affiliated latent business area (zone) based on taxi cab data (Fu 2014).

Several other recent studies compare various "advanced" statistical techniques and algorithms either to other advanced techniques or to traditional ones. Most studies conclude that the advanced, non-parametric techniques outperform traditional parametric techniques, while several conclude that the Random Forest algorithm is particularly well-suited to predicting real estate values.

Kontrimasa (2011) compares the accuracy of linear regression against the SVM technique and found the latter to outperform. Schernthanner H. (2016) compared traditional linear regression techniques to several techniques such as krigging (stochastic interpolation) and Random Forest. They concluded that the more advanced techniques, particularly Random Forest, are sound and more accurate when compared to traditional statistical methods. Antipov and Pokryshevskaya (2012) comes to a similar conclusion about the superiority of Random Forest to real estate valuation after comparing 10 algorithms: multiple regression, CHAID, Exhaustive CHAID, CART, 2 types of k-Nearest Neighbors, Multilayer Perceptron neural network (MLP), Radial Basis Function neural network (RBF)), Boosted Trees and finally Random Forest.

Guan et al. (2014) compared three different approaches to defining spatial neighbors: a simple radius technique, a k-nearest neighbors technique (KNN) using only distance and a KNN technique using all attributes. Interestingly, the location-only KNN models performed best, although by a slight margin. Park (2015) developed several housing price prediction models based on machine learning algorithms including C4.5, RIPPER, Naive Bayesian, and AdaBoost. By comparing the models' classification accuracy performance, the experiments demonstrate that the RIPPER algorithm, based on accuracy, consistently outperformed the other models in the performance of housing price prediction. Rafiei (2016) employed a restricted boltzmann machine (neural network with back propagation) to predict the sale price of residential condos in Tehran, Iran. Rather than focusing on predictive performance, their paper focuses on computational efficiency by employing “a non-mating genetic algorithm” for dimensionality reduction. The paper concludes that two primary strategies help in this regard: weighting property sales by temporal proximity (sales which happened closer in time are more important), and also using a learner to accelerate the recognition of important features. The paper compares this technique to several other common neural network approaches and finds that while not necessarily the only way to get the best answer, it is the fastest way to get to the best answer.

Finally, it should be noted that many studies, whether exploring advanced techniques, new data, or both, rely on aggregation of data by some arbitrary boundary. For example, Turner and Snow (2001) predicted gentrification in the Washington, D.C. metro area by ranking census tracts in terms of development. K. Chapple (2009) created a gentrification “early warning system” by identifying low income census tracts in central city locations. Barry Bluestone & Chase Billingham (2010) analyzed 42 census block groups near rail stations in 12 metro areas across the United States, studying changes between 1990 and 2000 for neighborhood socioeconomic and housing characteristics. All of these studies, and many more, relied on aggregation of data at the census-tract or census-block level. In contrast, this paper compares boundary-aggregation techniques (specifically, aggregating by zip codes) to spatial-lag features and finds the spatial lag techniques to consistently outperform.

# Methodology

Predictive modeling using spatial dependence has been employed extensively in recent years, notably in Crime Prediction (Almanie 2015). However, a key deficiency of many spatial models are their use of arbitrarily defined geographic regions, such as zip codes, political districts, police precincts, state lines, neighborhoods, etc., which potentially diminish and obscure valuable insights. Worse yet, many predictive models ignore spatial dependence, violating one of the basic tenets of geography: the direct relationship between distance and likeness (Miller 2007).

Our goal is to compare the use of spatial lags as features in a machine learning predictive model against traditional feature engineering techniques. We will create three modeling data sets:

- Base
- Zip Code
- Spatial Lag

We will then create 2 predictive models for each modeling data set, using a different outcome variable for each:

1) Probability of Sale. The probability that a given property in New York City will sell in a given year

2) Amount of Sale. Given that a property sells, how much is the sale value?

There will be six predictive models built in total, as follows:

Table 1: Six Predictive Models

| # | Model | Model Type | Data | Outcome Var | Outcome Type | Eval Metric |
|---|---|---|---|---|---|---|
| 1 | Probability of Sale | Classification | Base | Building Sold | Binary | AUC |
| 2 | Probability of Sale | Classification | Zip Code | Building Sold | Binary | AUC |
| 3 | Probability of Sale | Classification | Spatial Lag | Building Sold | Binary | AUC |
| 4 | Sale Price | Regression | Base | Sale Price per SF | Continuous | RMSE |
| 5 | Sale Price | Regression | Zip Code | Sale Price per SF | Continuous | RMSE |
| 6 | Sale Price | Regression | Spatial Lag | Sale Price per SF | Continuous | RMSE |

To accomplish this, we combine three open-source data repositories provided by New York City via nyc.gov and data.cityofnewyork.us. Our base modeling data set includes all building records and associated sales information from 2003-2017.

Following the creation of the base modeling data, we create two additional data sets through feature engineering: a “Zip Code features” data set and a “Spatial Lag features” data set. The primary goal of this study is to compare the predictive power of the spatial lags vs. the base and zip code features. We will seek to predict sales one year into the future by training and validating on 2003-2016 data and making predictions on 2017 data.

## Data

The New York City government makes available an annual data set which describes all tax lots in the five boroughs. The Primary Land Use and Tax Lot Output data set, known as PLUTO, contains a single record for every tax lot in the city along with a number of building and tax-related attributes such as Year Built, Assessed Value, Square Footage, number of stories, and many more. At the time of this writing, NYC has made this data set available for all years between 2002-2017, excluding 2008. For convenience, we also exclude the 2002 data set from our analysis because sales information is not available for that year. Importantly for our analysis, the latitude and longitude of the tax lots are also made available, allowing us to locate in space each building and to build geospatial features from the data.

Ultimately, we are interested in sales transactions–both frequency, and amount. Sales transactions are also made available by the New York City government, known as NYC Rolling Sales Data. At the time of this writing, sales transactions are available for the years 2003-2017. The sales transactions data contains additional data fields describing time, place, and amount of sale as well as additional building characteristics. Crucially, the sales transaction data does not include geographical coordinates, making it impossible to perform geospatial analysis without first mapping the sales data to PLUTO.

Prior to mapping to PLUTO, the sales data must first be transformed to include the proper mapping key. New York City uses a standard key of Borough-Block-Lot to identify tax lots in the data. For example, 31 West 27th Street is located in Manhattan, on block 829 and lot 16, therefore, its Borough-Block-Lot (BBL) is 1_829_16 (the 1 represents Manhattan). The sales data contain BBL's at the building level, however, the sales transactions data does not appropriately differentiate condos as their own BBL's. Mapping the sales data directly to the PLUTO data results in a mapping error rate of 23.1%. Therefore, the sales transactions data must first be mapped to another data source, the NYC Property Address Directory, or PAD, which contains an exhaustive list of all BBL's in NYC. Once the sales data is combined with PAD, the data can be mapped to PLUTO with an error rate of just 0.291%.

Prior to mapping, the sales data are normalized and filtered so that only BBL's with less than or equal to 1 transactions in a year occur. The final data set is an exhaustive list of all tax lots in NYC for every year between 2003-2017, whether that building was sold, for what amount, and several other variables.

Several building categories are excluded from the data for ease of modeling and subsequent analysis. The following building types are included:

Table 2: Building Types Included in Modeling Data

| Building Type | Description |
|---|---|
| A | ONE FAMILY DWELLINGS |
| B | TWO FAMILY DWELLINGS |
| C | WALK UP APARTMENTS |
| D | ELEVATOR APARTMENTS |
| F | FACTORY AND INDUSTRIAL BUILDINGS |
| G | GARAGES AND GASOLINE STATIONS |
| L | LOFT BUILDINGS |
| O | OFFICES |

The data is further filtered to include only records with equal to or less than 2 buildings per tax lot. The global filtering of the data set reduces the base modeling data from 12,012,780 records down to 8,247,499.

## Feature Engineering

### Base modeling data

The base modeling data–a combination of PLUTO, PAD and Rolling Sales– are enhanced with additional features. A summary table of the additional features are presented below:

Table 3: Base Modeling Data Features

| Feature | Min | Median | Mean | Max |
|---|---|---|---|---|
| has building area | 0 | 1.00 | 1.00 | 1.00 |
| Percent Com | 0 | 0.00 | 0.16 | 1.00 |
| Percent Res | 0 | 1.00 | 0.82 | 1.00 |
| Percent Office | 0 | 0.00 | 0.07 | 1.00 |
| Percent Retail | 0 | 0.00 | 0.04 | 1.00 |
| Percent Garage | 0 | 0.00 | 0.01 | 1.00 |
| Percent Storage | 0 | 0.00 | 0.02 | 1.00 |
| Percent Factory | 0 | 0.00 | 0.00 | 1.00 |
| Percent Other | 0 | 0.00 | 0.00 | 1.00 |
| Last Sale Price | 0 | 312.68 | 531.02 | 62,055.59 |
| Last Sale Price Total | 2 | 2,966,835.00 | 12,844,252.00 | 1,932,900,000.00 |
| Years Since Last Sale | 1 | 4.00 | 5.05 | 14.00 |
| SMA Price 2 year | 0 | 296.92 | 500.89 | 62,055.59 |
| SMA Price 3 year | 0 | 294.94 | 495.29 | 62,055.59 |
| SMA Price 5 year | 0 | 300.12 | 498.82 | 62,055.59 |
| Percent Change SMA 2 | -1 | 0.00 | 685.69 | 15,749,999.50 |
| Percent Change SMA 5 | -1 | 0.00 | 337.77 | 6,299,999.80 |
| EMA Price 2 year | 0 | 288.01 | 482.69 | 62,055.59 |
| EMA Price 3 year | 0 | 283.23 | 471.98 | 62,055.59 |
| EMA Price 5 year | 0 | 278.67 | 454.15 | 62,055.59 |
| Percent Change EMA 2 | -1 | 0.00 | 422.50 | 9,415,128.85 |
| Percent Change EMA 5 | -1 | 0.06 | 308.05 | 5,341,901.60 |

A binary variable is created to indicate whether a tax lot has a building on it (i.e., whether it is an empty plot of land or not). In addition, building types are quantified by what percent of the square footage belongs to the major property types: Commercial, Residential, Office, Retail, Garage, Storage, Factory and Other.

Importantly, two variables are created from the Sales Prices: A price-per-square-foot figure (“Last Sale Price”) and a total Sale Price (“Last Sale Price Total”). Sale Price per Square foot eventually becomes the outcome variable in one of the predictive models, even though it is referred to as Sale Price. Further features are derived which carry forward the previous sale price of a tax lot, if there is one, through successive years. Previous Sale Price is then used to create Simple Moving Averages (SMA), Exponential Moving Averages (SMA), and percent change measurements between the moving averages. In total, 69 variables are input to the feature engineering process and 92 variables are output. The final base modeling data set is 92 variables by 8,247,499 rows.

## Zip Code Modeling Data

The first of the comparative modeling data sets is the Zip code modeling data. Using the base data as a starting point, several features are created to describe characteristics of the zip code where the tax lot resides. A summary table of the Zip code level features is presented below. Note that "bt only" stands for "Building Type only" and refers to aggregations that only include buildings of the same type as the tax lot in question.

Table 4: Zip Code Modeling Data Features (continued below)

| Feature | Min | Median |
|---|---|---|
| Last Year Zip Sold | 0.00 | 27.00 |
| Last Year Zip Sold Percent Ch | -1.00 | 0.00 |
| Last Sale Price zip code average | 0.00 | 440.95 |
| Last Sale Price Total zip code average | 10.00 | 5,312,874.67 |
| Last Sale Date zip code average | 12,066.00 | 13,338.21 |
| Years Since Last Sale zip code average | 1.00 | 4.84 |
| SMA Price 2 year zip code average | 34.31 | 429.26 |
| SMA Price 3 year zip code average | 34.31 | 422.04 |
| SMA Price 5 year zip code average | 39.48 | 467.04 |
| Percent Change SMA 2 zip code average | -0.20 | 0.04 |
| Percent Change SMA 5 zip code average | -0.09 | 0.03 |
| EMA Price 2 year zip code average | 30.77 | 401.43 |
| EMA Price 3 year zip code average | 33.48 | 419.11 |
| EMA Price 5 year zip code average | 29.85 | 431.89 |
| Percent Change EMA 2 zip code average | -0.16 | 0.06 |
| Percent Change EMA 5 zip code average | -0.08 | 0.07 |
| Last Sale Price bt only | 0.00 | 357.71 |
| Last Sale Price Total bt only | 10.00 | 3,797,461.46 |
| Last Sale Date bt only | 12,055.00 | 13,331.92 |
| Years Since Last Sale bt only | 1.00 | 4.78 |
| SMA Price 2 year bt only | 0.00 | 347.59 |
| SMA Price 3 year bt only | 0.00 | 345.40 |
| SMA Price 5 year bt only | 0.00 | 372.30 |
| Percent Change SMA 2 bt only | -0.55 | 0.03 |
| Percent Change SMA 5 bt only | -0.33 | 0.02 |
| EMA Price 2 year bt only | 0.00 | 332.98 |
| EMA Price 3 year bt only | 0.00 | 332.79 |
| EMA Price 5 year bt only | 0.00 | 340.57 |
| Percent Change EMA 2 bt only | -0.47 | 0.06 |
| Percent Change EMA 5 bt only | -0.34 | 0.06 |

| Mean | Max |
|---|---|
| 31.14 | 112.00 |
| Inf | Inf |
| 522.87 | 1,961.21 |
| 11,877,688.55 | 1,246,450,000.00 |
| 13,484.39 | 17,149.00 |
| 4.26 | 11.00 |
| 501.15 | 2,092.41 |
| 496.47 | 2,090.36 |
| 520.86 | 2,090.36 |
| 616.47 | 169,999.90 |
| 341.68 | 113,333.27 |
| 479.38 | 1,883.81 |
| 479.95 | 1,781.38 |
| 472.80 | 1,506.46 |
| 388.90 | 107,368.37 |
| 326.17 | 107,368.38 |
| 485.97 | 6,401.01 |
| 11,745,130.56 | 1,246,450,000.00 |
| 13,497.75 | 17,149.00 |
| 4.30 | 14.00 |
| 462.67 | 5,519.39 |
| 458.50 | 5,104.51 |
| 481.09 | 4,933.05 |
| 600.10 | 425,675.69 |
| 338.15 | 188,888.78 |
| 442.79 | 5,103.51 |
| 443.02 | 4,754.95 |
| 436.70 | 4,270.37 |
| 377.17 | 254,462.97 |
| 335.17 | 178,947.30 |

In general, the base model data features are aggregated to a zip code level and attached to the individual observations, including SMA and EMA calculations. Additionally, a second set of features are added, denoted as “bt only”, which filter only for tax lots of the same building type. In total, the Zip code feature engineering process inputs 92 variables and outputs 122 variables.

### Spatial Lag modeling data

Spatial lags are variables created from physically proximate observations, where the aggregations can be weighted by an arbitrary distance function. For example, taking the average building age from all buildings within 100 meters of the tax lot in question weighted by inverse euclidean distance would constitue a spatial lag. A spatial lag is defined as:

$$(\rho)W_y + X(\beta)$$

Where $W_y$ is a spatially lagged dependent variable with weights matrix $W$, $X$ is a matrix of observations on the explanatory variable, and $\rho$ and $\beta$ are coeficients.

Creating spatial lags presents both advantages and disadvantages in the modeling process. Spatial lags allow for a much more fine-tuned measurement of a building's surrounding area. Knowing the average sale price of all buildings within 500 meters of a building can be much more informative than knowing the sale prices of all buildings in the same zip code (but not always). However, building spatial lags is computationally expensive. Each observation must be compared against every other observation in the data set to determine which points fall within the spatial lag radius and which do not. This can be conceptualized in Big O notation as:

$$O(n^n)$$

To build spatial lags for all 8,247,499 observations in our modeling data employed a novel spatial indexing technique to sped up the process. Since tax lots rarely move from year to year, we reduced the indexing task to 514,124 points (the number of unique tax lots in New York City). We then partitioned the dataset into an arbitrary number of grids and calculated spatial lags in parallel, assigning one grid to each processing unit until the task was complete. By supplying a sufficiently large number of processing units $p$, the computation time can be reduced to:

$$O(n^{\frac{n}{p}})$$

For each point in the dataset, we calculated and cached every other tax lot within 500 meters of that building. The result was an origin-destination relationship graph that related each tax lot to its surrounding tax lots.

Next, we used the spatial index to create spatial lag features (For an illustration, see Figure 1). One advantage to using spatial lags is the rich number of potential features which can be created. Spatial lags can be weighted based on an arbitrary distance function, e.g., physically closer observations can be given more weight. For our modeling purposes, we created two sets of spatially weighted features: distance weighted features (denoted with a "_dist") and simple average features (denoted with"_basic"). SMA and EMA as well as percent changes were also calculated. In total, the spatial lag feature engineering process input 92 variables and output 194 variables. A summary of the Spatial Lag features are presented in Table 6.

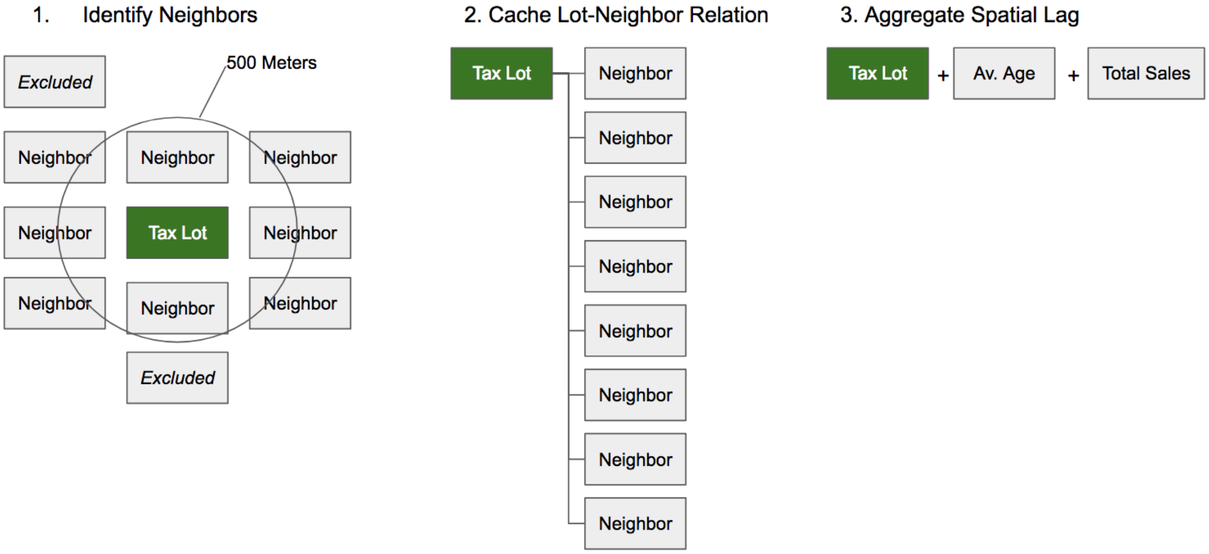

Figure 1: Illustrastion of Spatial Lag Creation

Table 6: Summary of Spatial Lag Features

| Spatial Lag Features |
|---|
| Radius Total Sold In Year |
| Radius Average Years Since Last Sale |
| Radius Res Units Sold In Year |
| Radius All Units Sold In Year |
| Radius SF Sold In Year |
| Radius Total Sold In Year sum over 2 years |
| Radius Average Years Since Last Sale sum over 2 years |
| Radius Res Units Sold In Year sum over 2 years |
| Radius All Units Sold In Year sum over 2 years |
| Radius SF Sold In Year sum over 2 years |
| Radius Total Sold In Year percent change |
| Radius Average Years Since Last Sale percent change |
| Radius Res Units Sold In Year percent change |
| Radius All Units Sold In Year percent change |
| Radius SF Sold In Year percent change |
| Radius Total Sold In Year sum over 2 years percent change |
| Radius Average Years Since Last Sale sum over 2 years percent change |
| Radius Res Units Sold In Year sum over 2 years percent change |
| Radius All Units Sold In Year sum over 2 years percent change |
| Radius SF Sold In Year sum over 2 years percent change |

Temporal and spatial derivatives of the Spatial Lag features presented in Table 6 are also added to the model, including: variables weighted by euclidean distance ("dist"), basic averages of the spatial lag radius ("basic mean"), Simple Moving Averages ("SMA") for 2 years, 3 years and 5 years, exponential moving averages ("EMA") for 2 years, 3 years and 5 years, and year-over-year percent changes for all variables ("perc change"). For a complete list of Spatial Lag features, see Appendix A.

Table 7: Distributions for Outcome Variables

| | Sold | Sale Price per SF |
|---|---|---|
| Min. | 0.00 | 0.0 |
| 1st Qu. | 0.00 | 163.5 |
| Median | 0.00 | 375.2 |
| Mean | 0.04 | 644.8 |
| 3rd Qu. | 0.00 | 783.3 |
| Max. | 1.00 | 83,598.7 |

## Algorithm

Previous works (see: Antipov and Pokryshevskaya (2012); also Schernthanner H. (2016)) have found the Random Forest algorithm (Breiman 2001) suitable to prediction tasks involving real estate. While algorithms exist that may marginally outperform Random Forest in terms of predictive accuracy (such as neural networks and functional gradient descent algorithms), Random Forest is highly scalable and parallelizable, and therefore a natural choice for comparing different feature engineering strategies (such as in this paper).

Random Forest can be used for both classification and regression tasks. The Random Forest algorithm works by generating a large number of independent classification or regression decision trees and then employing majority voting (for classification) or averaging (for regression) to generate predictions. Over a data set of N rows by M predictors, a bootstrap sample of the data is chosen ($n < N$) as well as a subset of the predictors ($m < M$). Individual decision/regression trees are built on the n by m sample. Because the trees can be built independently (and not sequentially, as is the case with most functional gradient descent algorithms), the tree building process can be executed in parallel across an arbitrary number of computer cores. With a sufficiently large number of cores, the model training time can be significantly reduced. This provides a highly accurate, robust prediction model that avoids many of the drawbacks of traditional parametric techniques, such as OLS.

The primary advantages to using Random Forest with real estate data are:

1. Can handle an arbitrarily large number of variables while avoiding the curse of dimensionality associated with regression techniques. Increasing the number of predictors in a multiple regression can quickly lead to over-fitting.
2. Can accommodate categorical variables with many levels. Real estate data often contains information describing the location of the property, or the property itself, as one of a large set of possible choices, such as neighborhood, county, census tract, district, property type, and zoning information. Because factors need to be recoded as individual dummy variables in the model building process, factors with many levels will quickly encounter the curse of dimensionality in multiple regression techniques.
3. Appropriately handles missing data. Predictions can be made with the parts of the

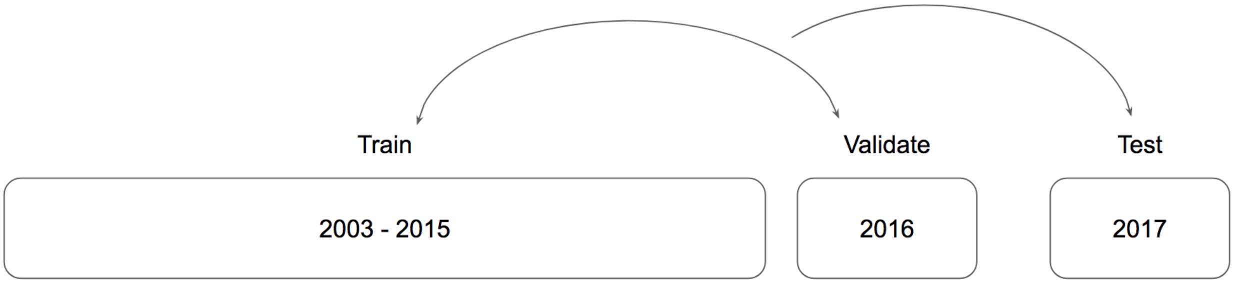

Figure 2: Out-of-time validation

tree which are successfully built, and therefore, there is no need to filter out incomplete observations or impute missing values. Since much real estate data is self reported, incomplete fields are common in the data.

4. Robust against outliers. Because of bootstrap sampling, outliers appear in individual trees less often, and therefore, are reduced in terms of importance. Real estate data, especially with regards to pricing, tends to contain outliers. For example, the dependent variable in one of our models, Sale Price (see: Table 7), shows a clear divergence in median and mean, as well as a maximum significantly higher than the third quartile.
5. Can recognize non-linear relationships in data, which is useful when modeling spatial relationships.
6. Is not affected by co-linearity in the data. This is highly valuable as real estate data can be highly correlated.
7. The algorithm can be parallelized and is relatively fast compared to neural networks and functional gradient descent algorithms.

To run the model, we have chosen the h2o.randomForest function from the h2o R open source library. The h2o implementation of the Random Forest algorithm is particularly well-suited for high parallelization. For more information, see: https://www.h2o.ai/.

## Data Validation

The goal of the predictive models are to be able to successfully predict both the probability and amount of real estate sales into the near future. As such, our models will use out-of-time validation to assess performance. As shown in Figure 2 The models will be trained using data from 2003-2015. 2016 modeling data will be used during the model training process as cross-validation data. Finally, we will score our model using 2017 as a held-out sample. Using out-of-time validation should ensure that the models generalize well into the immediate future.

## Variable Selection

For ease of processing and to improve the ability of the model to generalize into the future, a variable selection step is added to the modeling process. A Random Forest model is first trained on a 1% sub-sample of the modeling data. Variable importance of the resulting model is calculated using the technique proposed by Friedman (2001), i.e., for a collection of decision trees $[T_m]_1^m$:

$$\hat{I}_j^2 = \frac{1}{M} \sum_{m=1}^{M} \hat{I}_j^2(T_m)$$

Where influence $I$ for variable $j$ is calculated as the sum of corresponding improvements in squared-error for node tree $T$. After calculating variable importance for the model data subset, the variables are rank-ordered by descending importance. Variables which account for 80% of the total variable importance are chosen to advance to the model training round on the full modeling data sets.

## Evaluation Metrics

We have chosen evaluation metrics that will allow us to easily compare the performance of the models against other models with the same outcome variable. The classification models (Probability of Sale) will be compared using Area Under the ROC Curve (AUC). The regression models (Sale Price) will be compared using Root Mean Squared Error (RMSE). Both evaluation metrics are common for their respective outcome variable types, and as such will be useful for comparing within model-groups.

### Area Under ROC Curve (AUC)

A classification model typically outputs a probability that a given row in the data belongs to a group. In the case of binary classification, the value falls between 0 and 1. There are many techniques for determining the cut off threshold for classification; a typical method is to assign anything above a 0.5 into the "1" or positive class. An ROC curve (receiver operating characteristic curve) plots the True Positive Rate vs. the False Positive rate at different classification thresholds; it is a measurement of the performance of a classification model across all possible thresholds, and therefore sidesteps the need to arbitrarily assign a cutoff.

Area Under the ROC Curve, or AUC measures the entire two-dimensional area underneath the ROC curve. It is the integration of the curve from (0,0) to (1,1), defined as $AUC = \int_{(0,0)}^{(1,1)} f(x)dx$.

AUC provides a relatively standard measure of performance across all possible classification thresholds, and can be interpreted as the probability that the model ranks a random positive example more highly than a random negative example. A value of 0.5 represents a perfectly random model, while a value of 1.0 represents a model that can perfectly discriminate between the two classes. AUC is useful for comparing classification models against one another because they are both scale and threshold-invariant.

One of the drawbacks to AUC is that is does not describe the trade-offs between false positives and false negatives. In certain circumstances, a false positive might be considerably less desirable than a false negative, or vice-versa. For our purposes, we rank false positives and false negatives as equally undesirable outcomes.

### Root Mean Squared Error

The Root Mean Squared Error (RMSE) is a common measurement of the differences between values predicted by a regression model and the observed values. It is formally defined as $RMSE = \sqrt{\frac{\sum_1^T (\hat{y}_t - y_t)^2}{T}}$, where $\hat{y}$ represents the prediction and $y$ represents the observed value at observation $t$.

Lower RMSE scores are typically more desirable. An RMSE value of 0 would indicate a perfect fit to the data. RMSE can be difficult to interpret on its own, however, it is useful for comparing models with similar outcome variables. In our case, the outcome variables (Sales Price per Square Foot) are consistent across modeling data sets, and therefore can be reasonably compared using RMSE.

# Results

## Sale Price Model

Using Root Mean Squared Error (RMSE) as an evaluation metric of predictive power, we find that the Spatial Lag modeling features consistently outperform the Zip Code features, while the Base modeling data tends to outperform both, as shown in the following Table (lower RMSE is better):

Table 8: Sale Price Model RMSE For Validation and Test Hold-out Data

| type | base | zip | spatial lag |
|---|---|---|---|
| Validation | 280.6 | 298 | 286.2 |
| Test | 287.8 | 300.6 | 297.9 |

Interestingly, both the Spatial Lag modeling data and the Zip Code modeling data are extensions of the Base data, yet the Base data tends to generalize to the hold-out data sets better. From this, we make two observations:

1) Building characteristic data, which largely comprises the Base modeling data, are the most important for predicting building sale price per foot
2) The Zip Code and Spatial Lag models are suffering from over-fitting of the data, which is limiting their ability to generalize to the hold-out samples.

It is possible that the over-fitting issue could be corrected through hyper-parameter tuning of the algorithms, as well as implementing stricter variable selection. Despite this, we can still safely conclude that the Spatial Lag features are superior to the Zip Code features in terms of predictive power.

Taking a closer look at the data, we find that the models have varying performance across Building Types and Boroughs. Figure 3 shows RMSE by Model, faceted by Borough across the y-axis and Building Type across the x-axis (See Table 2 for a description of building type codes).

We make the following observations from Figure 3:

- The Spatial Lag modeling data outperforms both Base and Zip Code in 6 cases, notably for Type A buildings (One Family Dwellings) and Type L buildings (Lofts) in Manhattan as well as Type O Buildings (Office) in Queens
- It is generally harder to predict sale prices in Manhattan than other Boroughs
- The "residential" building Types A (One Family Dwellings), B (Two Family Dwellings), C (Walk Up Apartments) and D (Elevator Apartments) have generally lower RMSE scores compared to the non-residential types

If we rank the models by performance for each Borough, Building Type combination, we find that the Spatial Lag models outperform the Zip Code models in 72% of cases. Table 9 shows the average ranking by model type as well as the percent distribution by model type across rankings.

Table 9: Sale Price Model Rankings, RMSE by Borough and Building Type

| Model Rank | 1 | 2 | 3 | Average Rank |
|---|---|---|---|---|
| Base | 22.2% | 9.3% | 1.9% | 1.39 |
| Spatial Lag | 5.6% | 18.5% | 9.3% | 2.11 |
| Zip | 5.6% | 5.6% | 22.2% | 2.5 |

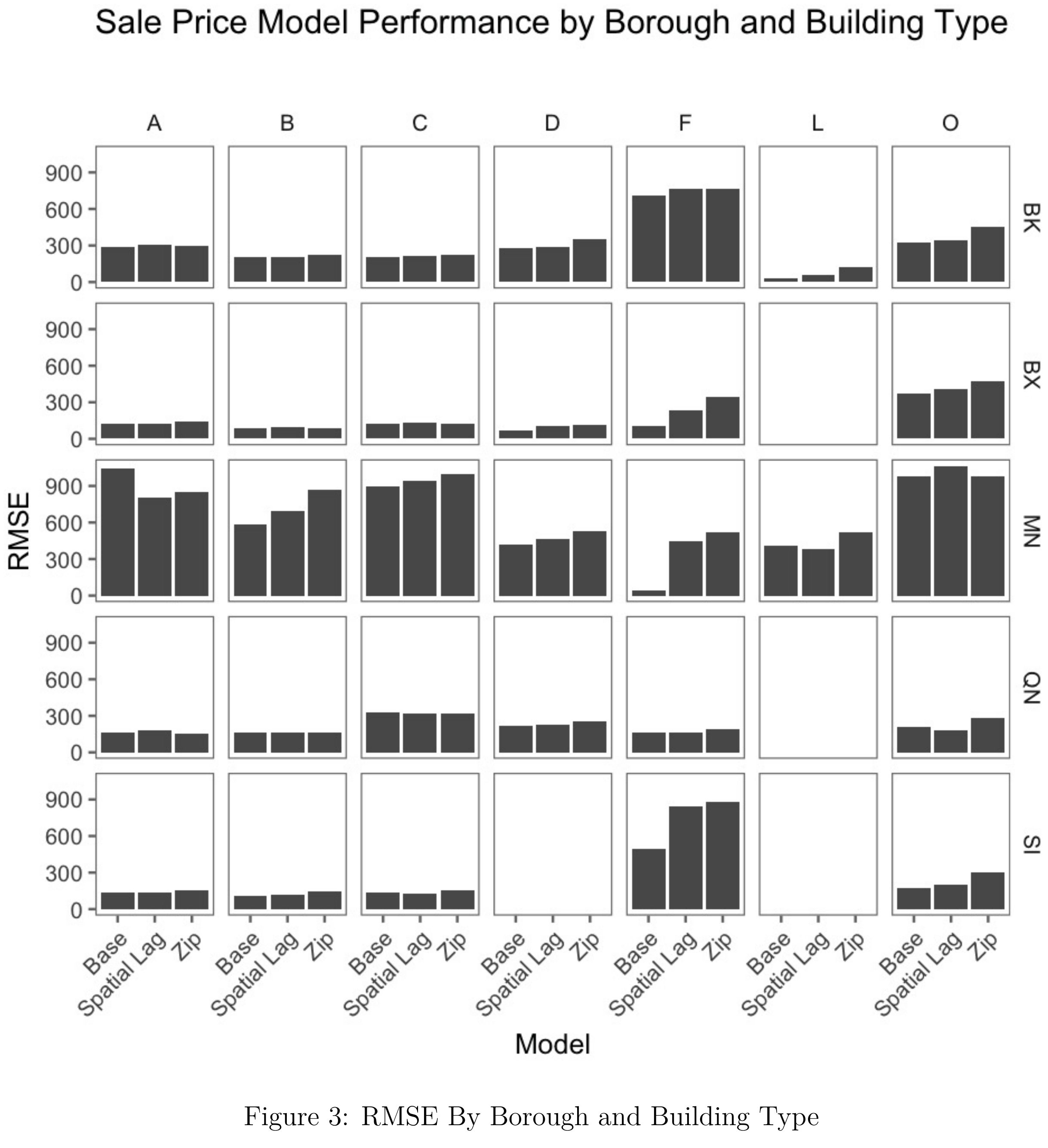

Figure 3: RMSE By Borough and Building Type

## Probability of Sale Model

Similar to the results found in the Sale Price models, using Area Under the ROC Curve (AUC) as an evaluation metric, we find the Spatial Lag model performs better on the hold-out validation data compared to the Zip Code modeling data, as shown in Table 10. The Base Modeling data continues to outperform the Spatial Lag and Zip Code modeling data overall, however, when broken down by Borough and Building Type, some interesting patterns emerge.

Table 10: Probability of Sale Models AUC

| Model AUC | Base | Zip | Spatial Lag |
|---|---|---|---|
| Validation | 0.832 | 0.8292 | 0.8287 |
| Test | 0.83 | 0.8246 | 0.8279 |

Looking at the predictions by the models made on the 2017 validation hold-out data, we see the Spatial Lag model performs best of any model for three out of five Boroughs: Manhattan, Bronx and Staten Island (Table 11).

Table 11: Probability of Sale Models AUC by Borough

| Model | BK | BX | MN | QN | SI |
|---|---|---|---|---|---|
| Base | 0.8309 | 0.8288 | 0.7926 | 0.8338 | 0.8336 |
| Zip | 0.8234 | 0.8215 | 0.7796 | 0.8283 | 0.8281 |
| Spatial Lag | 0.8257 | 0.8312 | 0.8031 | 0.8327 | 0.8348 |

Figure 4 shows a breakdown of model AUC faceted along the x-axis by Building Type and along the y-axis by Borough. The coloring indicated by how much a model's AUC diverges from the cell average.

We make the following observations about Figure 4:

- The Spatial Lag model outperforms all other models for Elevator Buildings (Type D), particularly in the Bronx
- The Probability of Sale model tends to perform poorly in Manhattan vs. other Boroughs
- However, the Spatial Lag model performs well in Manhattan for the residential building types (A, B, C and D)

If we rank model the probability models' performance for each Borough and Building Type, we see that the Spatial Lag models consistently outperform the Zip Code models, as shown in Table 12.

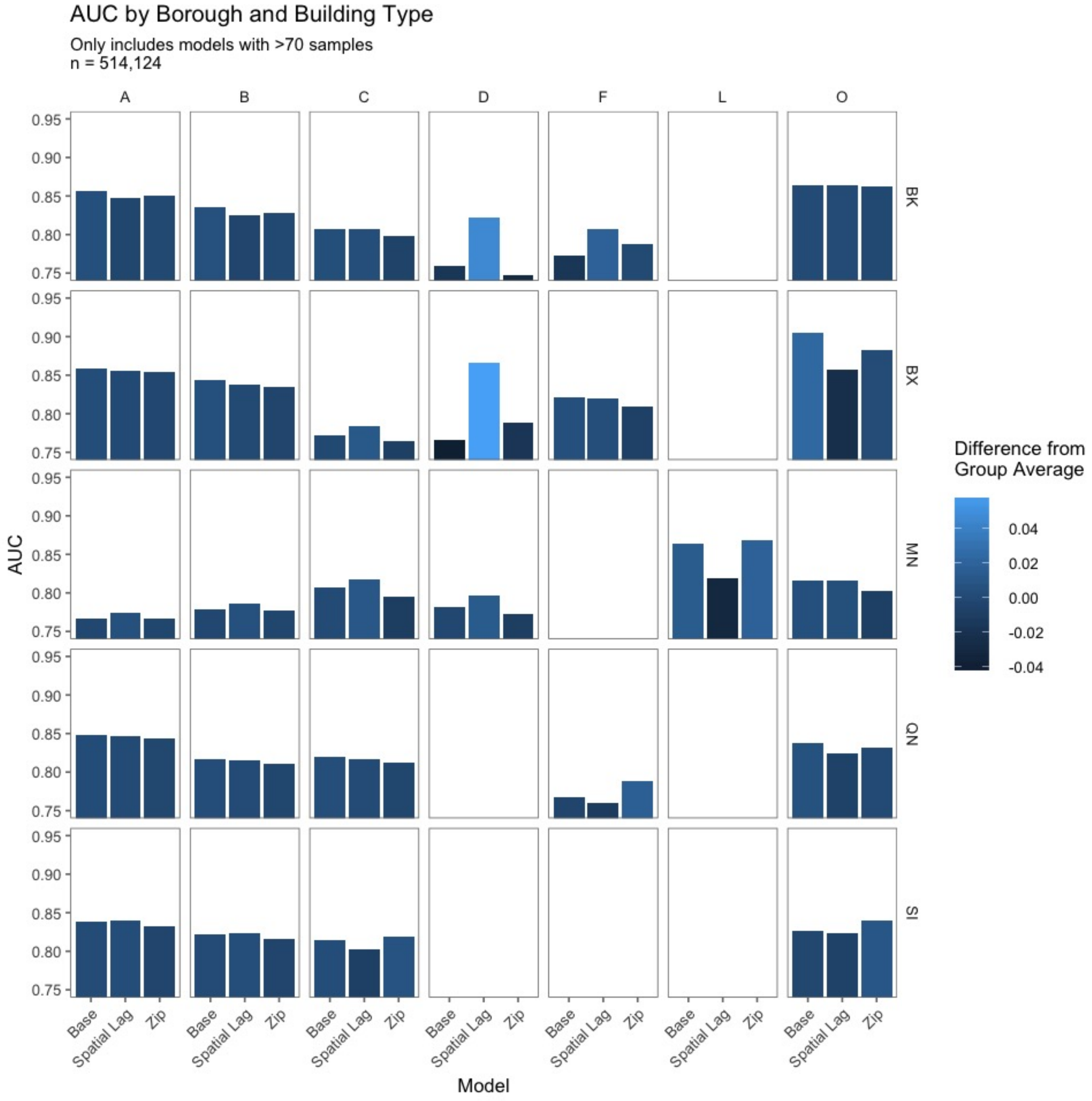

Figure 4: AUC By Borough and Building Type

Table 12: Distribution and Average Model Rank for Probability of Sale by AUC across Borough and Building Types

| Model Rank | 1 | 2 | 3 | Average Rank |
|---|---|---|---|---|
| Base | 16.2% | 12.0% | 5.1% | 2.22 |
| Spatial Lag | 11.1% | 13.7% | 8.5% | 2.09 |
| Zip | 6.0% | 7.7% | 19.7% | 1.69 |

# Conclusions and Future Research

## Future Research

- Adaptive Bandwidth

## Conclusion

# Appendix A: List of Spatial Lag Features

Table 13: Appendix A: All Spatial Lag Features

| Feature |
|---|
| Radius Total Sold In Year |
| Radius Average Years Since Last Sale |
| Radius Res Units Sold In Year |
| Radius All Units Sold In Year |
| Radius SF Sold In Year |
| Radius Total Sold In Year sum over 2 years |
| Radius Average Years Since Last Sale sum over 2 years |
| Radius Res Units Sold In Year sum over 2 years |
| Radius All Units Sold In Year sum over 2 years |
| Radius SF Sold In Year sum over 2 years |
| Radius Total Sold In Year percent change |
| Radius Average Years Since Last Sale percent change |
| Radius Res Units Sold In Year percent change |
| Radius All Units Sold In Year percent change |
| Radius SF Sold In Year percent change |
| Radius Total Sold In Year sum over 2 years percent change |
| Radius Average Years Since Last Sale sum over 2 years percent change |

| Feature |
|---|
| Radius Res Units Sold In Year sum over 2 years percent change |
| Radius All Units Sold In Year sum over 2 years percent change |
| Radius SF Sold In Year sum over 2 years percent change |
| Percent Com dist |
| Percent Res dist |
| Percent Office dist |
| Percent Retail dist |
| Percent Garage dist |
| Percent Storage dist |
| Percent Factory dist |
| Percent Other dist |
| Percent Com basic mean |
| Percent Res basic mean |
| Percent Office basic mean |
| Percent Retail basic mean |
| Percent Garage basic mean |
| Percent Storage basic mean |
| Percent Factory basic mean |
| Percent Other basic mean |
| Percent Com dist perc change |
| Percent Res dist perc change |
| Percent Office dist perc change |
| Percent Retail dist perc change |
| Percent Garage dist perc change |
| Percent Storage dist perc change |
| Percent Factory dist perc change |
| Percent Other dist perc change |
| SMA Price 2 year dist |
| SMA Price 3 year dist |
| SMA Price 5 year dist |
| Percent Change SMA 2 dist |
| Percent Change SMA 5 dist |
| EMA Price 2 year dist |
| EMA Price 3 year dist |
| EMA Price 5 year dist |
| Percent Change EMA 2 dist |
| Percent Change EMA 5 dist |
| SMA Price 2 year basic mean |
| SMA Price 3 year basic mean |
| SMA Price 5 year basic mean |
| Percent Change SMA 2 basic mean |
| Percent Change SMA 5 basic mean |
| EMA Price 2 year basic mean |

| Feature |
|---|
| EMA Price 3 year basic mean |
| EMA Price 5 year basic mean |
| Percent Change EMA 2 basic mean |
| Percent Change EMA 5 basic mean |
| SMA Price 2 year dist perc change |
| SMA Price 3 year dist perc change |
| SMA Price 5 year dist perc change |
| Percent Change SMA 2 dist perc change |
| Percent Change SMA 5 dist perc change |
| EMA Price 2 year dist perc change |
| EMA Price 3 year dist perc change |
| EMA Price 5 year dist perc change |
| Percent Change EMA 2 dist perc change |
| Percent Change EMA 5 dist perc change |
| SMA Price 2 year basic mean perc change |
| SMA Price 3 year basic mean perc change |
| SMA Price 5 year basic mean perc change |
| Percent Change SMA 2 basic mean perc change |
| Percent Change SMA 5 basic mean perc change |
| EMA Price 2 year basic mean perc change |
| EMA Price 3 year basic mean perc change |
| EMA Price 5 year basic mean perc change |
| Percent Change EMA 2 basic mean perc change |
| Percent Change EMA 5 basic mean perc change |

# References

Almanie, R.; Lor, T.; Mirza. 2015. “Crime Prediction Based on Crime Types and Using Spatial and Temporal Criminal Hotspots.” *International Journal of Data Mining & Knowledge Management Process (IJDKP)* 5 (4).

Antipov, Evgeny A., and Elena B. Pokryshevskaya. 2012. “Mass Appraisal of Residential Apartments: An Application of Random Forest for Valuation and a Cart-Based Approach for Model Diagnostics.” *Expert Systems with Applications*.

Barry Bluestone & Chase Billingham, Stephanie Pollack &. 2010. “Maintaining Diversity in America’s Transit-Rich Neighborhoods: Tools for Equitable Neighborhood Change.” *New England Community Developments, Federal Reserve Bank of Boston*, 1–6.

Breiman, L. 2001. “Random Forests.” *Machine Learning* 45 (1): 5–32.

Chapple, Karen. 2009. “Mapping Susceptibility to Gentrification: The Early Warning

Toolkit.” *Berkeley, CA: Center for Community Innovation.*

Chapple, Miriam, Karen; Zuk. 2016. “Forewarned: The Use of Neighborhood Early Warning Systems for Gentrification and Displacement.” *Cityscape: A Journal of Policy Development and Research* 18 (3).

Clay, Phillip L. 1979. *Neighborhood Renewal: Middle-Class Resettlement and Incumbent Upgrading in American Neighborhoods.* Lexington Books.

Dietzell, Nicole; Schäfers, Marian Alexander; Braun. 2014. “Sentiment-Based Commercial Real Estate Forecasting with Google Search Volume Data.” *Journal of Property Investment & Finance,* 32 (6): 540–69.

Dreier, John; Swanstrom, Peter; Mollenkopf. 2004. *Place Matters: Metropolitics for the Twenty-First Century.* University Press of Kansas.

d’Amato, Tom, Maurizio; Kauko, ed. 2017. *Advances in Automated Valuation Modeling.* Springer International Publishing.

Eckert, J. K. 1990. *Property Appraisal and Assessment Administration.* Chicago, IL.: International Association of Assessing Officers.

Fotheringham, R; Yao, A.S.; Crespo. 2015. “Exploring, Modelling and Predicting Spatiotemporal Variations in House Prices.” *The Annals of Regional Science* 54.

Friedman, Jerome H. 2001. “Greedy Function Approximation: A Gradient Boosting Machine.” *The Annals of Statistics* 29 (5): 1189–1232.

Fu, Yanjie; et al. 2014. *Exploiting Geographic Dependencies for Real Estate Appraisal: A Mutual Perspective of Ranking and Clustering.* Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery; data mining.

Geltner, David, and Alex Van de Minne. 2017. “Do Different Price Points Exhibit Different Investment Risk and Return Commercial Real Estate.” Real Estate Research Institute.

Guan, Jian, Donghui Shi, Jozef M. Zurada, and Alan S. Levitan. 2014. “Analyzing Massive Data Sets: An Adaptive Fuzzy Neural Approach for Prediction, with a Real Estate Illustration.” *Journal of Organizational Computing and Electronic Commerce* 24 (1). Taylor & Francis: 94–112. doi:10.1080/10919392.2014.866505.

Helbich, et al., Marco. 2013. “Boosting the Predictive Accuracy of Urban Hedonic House Price Models Through Airborne Laser Scanning.” *Computers, Environment and Urban Systems* 39: 81–92.

Johnson, Ken, Justin Benefield, and Jonathan Wiley. 2007. “The Probability of Sale for Residential Real Estate.” *Journal of Housing Research* 16 (2): 131–42. doi:10.5555/jhor.16.2.0234g75800h5k8x6.

Kontrimasa, Antanas, Vilius; Verikasb. 2011. “The Mass Appraisal of the Real Estate by

Computational Intelligence." *Applied Soft Computing*.

Koschinsky, J. et al. 2012. "The Welfare Benefit of a Home's Location: An Empirical Comparison of Spatial and Non-Spatial Model Estimates." *Journal of Geographical Systems* 10109.

Lees, Tom; Wyly, Loretta; Slater. 2008. "Gentrification." *Growth and Change* 39 (3): 536–39. doi:10.1111/j.1468-2257.2008.00443.x.

Miller, J.; Aspinall, J.; Franklin. 2007. "Incorporating Spatial Dependence in Predictive Vegetation Models." *Ecological Modelling* 202 (3): 225–42.

Park, Jae Kwon, Byeonghwa; Bae. 2015. "Using Machine Learning Algorithms for Housing Price Prediction: The Case of Fairfax County, Virginia Housing Data." *Expert Systems with Applications* 42 (6): 2928–34.

Pivo, Gary, and Jeffrey D. Fisher. 2011. "The Walkability Premium in Commercial Real Estate Investments." *Real Estate Economics* 39 (2): 185–219. doi:10.1111/j.1540-6229.2010.00296.x.

Quintos, Carmela. 2013. "Estimating Latent Effects in Commercial Property Models." *Journal of Property Tax Assessment & Administration* 12 (2).

Rafiei, Hojjat, Mohammad Hossein; Adeli. 2016. "A Novel Machine Learning Model for Estimation of Sale Prices of Real Estate Units." *Journal of Construction Engineering and Management* 142 (2).

Reardon, Kendra, Sean F.; Bischoff. 2011. "Income Inequality and Income Segregation." *American Journal of Sociology*.

Schernthanner H., Gonschorek J., Asche H. 2016. "Spatial Modeling and Geovisualization of Rental Prices for Real Estate Portals." *Computational Science and Its Applications* 9788.

Silverherz, J. D. 1936. "The Assessment of Real Property in the United States." *Albany: J.B. Lyon Co. Printers*.

Smith, Neil. 1979. "Toward a Theory of Gentrification a Back to the City Movement by Capital, Not People." *Journal of the American Planning Association* 45 (4). Routledge: 538–48. doi:10.1080/01944367908977002.

Solomon Greene, Molly Scott, Rolf Pendall, and Serena Lei. 2016. "Open Cities: From Economic Exclusion to Urban Inclusion." *Urban Institue Brief*, June. Urban Institue Brief.

Turner, Margery Austin, and Christopher Snow. 2001. *Leading Indicators of Gentrification in d.C. Neighborhoods*.

Watson, Tara. 2009. "Inequality and the Measurement of Residential Segregation by Income in American Neighborhoods." *Review of Income and Wealth*.

Zuk, Miriam; et al. 2015. "Gentrification, Displacement and the Role of Public Investment: A Literature Review."