Predicting Real Estate Sales Using Machine Learning and Spatial Dependence

Boosting Random Forest Predictive Accuracy Using Spatial Lags

Contents

Introduction

What is Economic Exclusion?

Income inequality may be a defining challenge of our time. Researchers at the Urban Institute (Solomon Greene and Lei 2016) recently identified the socio-economic phenomenon of "Economic Exclusion" as one compelling explanation for the recent rise in inequality in the US. As discussed by Zuk (2015), "Neighborhoods change slowly, but over time are becoming more segregated by income, due in part to macro-level increases in income inequality". Vulnerable populations–disproportionately communities of color, immigrants, refugees, and women–who are displaced by localized economic prosperity enter into a gradual cycle of diminished access to good jobs, good schools, health care facilities, public spaces, etc. Such systematic denial causes enduring and self-reinforcing poverty over the course years and even generations, gradually entrenching income inequality and general unrest.

One way to practically combat economic exclusion is to focus on preventing displacement, however, detecting gentrification at an early enough stage can be a daunting task. When an area experiences economic growth, increased housing demands and subsequent affordability pressures can lead to evictions 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. There are a host of 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 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" (2016).

(M. Chapple Karen; Zuk 2016) To be included in the "motivation" section of my thesis. Not about predictive modeling, but is a very recent overview of the application of predictive gentrification models

How Can Machine Learning Help?

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 diminish and obscure potentially 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 Contribution

This paper explores novel techniques to predict gentrification in the pursuit of combating displacement and economic exclusion. Modern techniques of data mining, machine learning and predictive modeling are applied to data sets describing property values and sale prices in New York City. We demonstrate that the incorporation of spatial lags, i.e., variables created from physically proximate observations, can improve the predictive accuracy of machine learning models above and beyond both non-spatial models as well as models which incorporate data aggregated at arbitrary geographic regions such as zip codes.

Literature Review

The literature review for this paper reviews the concept of 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” (Zuk 2015) 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 major contributor to Economic Displacement, specifically, “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. Reardon (2011) also argues that “were income inequality to stop rising, the number of segregated neighborhoods would decline.”

Many studies conclude that gentrification in most forms leads to exclusionary 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 prices has been in service of creating mass appraisal models. Mass appraisal models are most commonly used by local governments for the purpose of collecting taxes from property owners. Mass appraisal models share many characteristics with predictive machine learning models, 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 area. 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).

Hedonic regression models generally seek to break down the price of a good based on the intrinsic and extrinsic components. 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. Specifically, the model calculates spatial lags on properties of interest using neighboring properties within 1,000 feet of a sale. The derived variables include characteristics like average age, quantity of poor condition homes, percent of homes with electric heating, construction grade, etc. within 1,000 feet of the property in question. 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 important structural characteristics.

As recently as 2015, much research has dealt with mitigating the drawbacks of MRA, including 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. GWR is a variation on OLS that allows for "adaptive bandwidths" of local data to be included, i.e., for each estimate, the number of data points included varies and can be optimized using cross-validation.

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. Pivo 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 accessibility to "business areas". Then, 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 using only distance and a k-nearest neighbors 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. A non-mating genetic algorithm is used 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 techniques and finds the spatial lag techniques to generally outperform.

sample citations

Sample Citation: (Antipov and Pokryshevskaya 2012) (see: Antipov and Pokryshevskaya 2012, 33–35; also Antipov and Pokryshevskaya 2012, ch. 1 and *passim*)

A minus sign (-) before the @ will suppress mention of the author in the citation. This can be useful when the author is already mentioned in the text:

Antipov says blah (2012).

You can also write an in-text citation, as follows:

Antipov and Pokryshevskaya (2012) says blah.

Methodology

Building off of previous research (see: Antipov and Pokryshevskaya (2012); also Schernthanner H. (2016)), the goal of this study is to employ the Random Forest algorithm to predict:

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?

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 "zipcode-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 lag data set vs. the base and zipcode level features.

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 Ouput 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 convinience, 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 contains BBL's at the building level, however, the sales transactions data does not appropriately designate condos as their own BBLs. Mapping the sales data directlty 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 BBLs in NYC. Once the sales data is combined with PAD, the data can be mapped to PLUTO with an error rate of 0.291%.

After the Sales Transactions data has been mapped to PAD, it can then be mapped to PLUTO. The sales data is normalized and filtered so that only BBLs 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 additional variables.

Only building categories of significant interest are included in the data. The following building types are included:

Category	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 dataset reduces the base modeling data from 12,012,780 records down to 8,247,499.

Feature Engineering & Data Partitioning

The

Executing base feature engineering… …done. Input 69 variables and output 92 variables. Total rows: 8,247,499

Executing ZIP feature engineering… …Engineering done. Input 92 variables and output 122 variables

```
Running spatial indexing on 514,124 points
```

ID column: bbl Max distance: 500 units=m Number of partions: 7 by 7, 49 total data partitions Parallel: TRUE with number of clusters set to 62

```
...done. Total indexing time: 5 mins
```

Writing radii index to disk… …done Creating radii features… Running RADII feature creation Starting radii feature creation at 2018-06-08 14:30:24 … Building sales features… Creating building features… Building moving average features… Building intensity features… Joining new features to original data… Radii feature creation time: 3.47 hours …done Writing radii features to disk… …done. Writing took 3.99mins …Engineering done. Input 92 variables and output 194 variables

1) Create radii-index & radii features
2) Create zip code features
3) Create 3 modeling data sets:
   a) Base
   b) Zip code features
   c) Radii features

Algorithm

Random Forrest has several advantages over traditional geographic weighted regression, amoung them:

1. Ability to handle large amounts of categorical data without much pre-processing
2. Ability to model in spite of missing values in data
3. Eliminated colinearity as a concern
4. Allows for the introduction of many more variables without requiring penalty for additional predictors
5. Works relatively fast and can be parallelized

Model Diagnostics

Results

Probability of Sale Model

Sale Price Model

Using the Models in Practice

Conclusions and Future Research

Future Research

Conclusion

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.

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.

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.

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."