University of Connecticut

Department of Economics Working Paper Series

Information Value of Property Description: A Machine Learning Approach

by

Lily Shen
Clemson University

Stephen Ross
University of Connecticut

Working Paper 2019-20R
December 2019, Revised September 2020

365 Fairfield Way, Unit 1063
Storrs, CT 06269-1063
Phone: (860) 486-3022
Fax: (860) 486-4463
http://www.econ.uconn.edu/

Information Value of Property Description: A Machine Learning Approach *

Lily Shen
Clemson University

Stephen Ross
University of Connecticut

September 14, 2020

Abstract

This paper employs machine learning to quantify the value of "soft" information contained in real estate property descriptions. Textual descriptions contain information that traditional hedonic attributes cannot capture. A one standard deviation increase in the uniqueness of a property based on this "soft" information leads to a 15% increase in property sale price in a hedonic price model and a 10% increase in a repeat sales price model. The effects in the hedonic model appear to arise through two channels: the unobserved quality of the housing unit, and the market power of the housing unit relative to competing properties. The effects in the repeat sales model appear to be driven entirely by the market power of the unit. Further, an annual hedonic price index ignoring our measure of unobserved quality overstates real estate prices by between 10% to 23% and mistimes the stabilization of housing prices following the Great Recession. Similar, but smaller effects, are observed for the repeat sales price index.

Key Words: *Natural Language Processing, Unsupervised Machine Learning, Soft Information, Housing Prices, Price indexes, Property Descriptions*

JEL codes: *R31, G12, G14, C45*

*Lily Shen, yannans@g.clemson.edu, Clemson University Finance department. Stephen Ross, stephen.l.ross@uconn.edu, University of Connecticut. For helpful comments and discussions we would like to thank seminar attendees at University of Connecticut, Clemson University/Finance, the Federal Reserve Bank of Philadelphia, the Federal Reserve Bank of Cleveland, and the Federal Reserve Bank of Atlanta. We would also like to thank session participants and discussants at the 2019 Homer Hoyt Meeting, the 2019 AREUEA/ASSA Meeting, and the American Real Estate Society's 35th Annual Meeting. We are especially grateful for comments and suggestions from Brent Ambrose, James Conklin, Chris Cunningham, Bruce Fallick, Kris Gerardi, Roberto Pinheiro, Rubn Hernndez-Murillo, Stuart Rosenthal, and Vincent Intintoli. All errors are our own.

1 Introduction

Real estate is a large and essential component of the U.S economy,[1] and valuing real estate assets has always been challenging given the limited information available in traditional assessment and real estate databases. For example, while the number of bedrooms might be viewed as a "hard" attribute of a housing unit, the description of a housing unit may represent unreported "soft" features that capture information about the unit's value (Liberti and Petersen, 2018). According to a 2018 Wall Street Journal article, institutional investors who buy and sell hundreds of houses on a daily basis utilize newly developed artificial intelligence technologies to extract "soft" information from unstructured real estate data.[2]

Up to this point, academic studies on the use of "soft" information from property descriptions to value real estate have been primarily restricted to analyses of keywords, and those results have been mixed. On the one hand, Levitt and Syverson (2008), Rutherford and Yavas (2005), Liu et al. (2020), and Lawani et al. (2018) find that the inclusion of indicator variables for positive/negative words and short phrases in real estate advertisements can reduce bias from omitted variables. On the other hand, Goodwin (2014) and Pryce (2008) point out that the effects of positive/negative words on real estate prices are not consistent across different word classes.

In this paper, we use a natural language processing algorithm to convert real estate property descriptions into numeric vectors of housing attributes. Unlike keyword approaches where the researcher creates a list of key words, these attribute vectors are created by a relatively unstructured, data driven process. We then use these quantitative measures for each description to assess the uniqueness of each property relative to its neighbors. Similarly, Ahlfeldt and Holman (2018) characterize neighborhood architectural quality based on distinctiveness, arguing that distinctiveness minimizes the influence of normative judgments

[1] In 2015, the real estate industry generated $3 trillion of revenue, which accounted for 17.3% of GDP. 2016 National Association of Realtors report: Economic Impact of Real Estate Activity

[2] Dezember, Ryan, "How to Buy a House the Wall Street Way" The Wall Street Journal, September 16, 2018.

and personal taste, unlike measures of quality or beauty. Our resulting measure of uniqueness is then included in both hedonic and repeat sales models of housing sales price. We examine both the impact of uniqueness on sales price and the effect of including controls for uniqueness on estimated price indexes.

To our knowledge, our study is the first paper to apply natural language processing to characterize the uniqueness of real estate properties and to apply such a measure of uniqueness to models of housing prices. Our paper is similar to Aubry et al. (2019) and Lindenthal (2017) who use machine learning techniques to characterize the attributes of artwork and real estate photos and then estimate hedonic models of the sales price. While we are the first to measure real estate uniqueness using textual data, Haurin (1988) and Haurin et al. (2010) model real estate "atypicality" using the observable, quantitative house features, and Ahlfeldt and Holman (2018) characterize neighborhood architectural distinctiveness using survey data.

The uniqueness of a property may influence prices through a few mechanisms. First, real estate agents use advertisements to emphasize positive features of a housing unit and are unlikely to mention negative aspects of the property in the description. As a result, our uniqueness measure likely captures the extent to which a housing unit has positive features that agents view as adding value. Even in a repeat sales model, housing quality may differ across transactions due to renovations or differences in maintenance (Harding et al., 2007).[3] Second, in differentiated product markets, products that are unique in product attribute space have greater market power and so tend to sell for higher prices (Bayer et al., 2007 and Berry et al., 1995). Similarly, Haurin et al. (2010) shows that in a search model more "atypical" properties will have a greater variance of offers and a higher final sales price.[4]

[3]Over 10 percent of transactions in our repeat sales sample involved a renovated housing unit. Further, Harding et al. (2007) document that housing depreciates in value about 2.5 percent per year and maintenance expenditures to address deprecation vary widely, e.g. the 25th percentile of maintenance spending per year is near zero at 0.15 percent, and the 75th percentile is 1.77 percent.

[4]To the extent that uniqueness captures seller market power, including controls for uniqueness is similar to Goetzmann and Peng (2006) where they adjust price indexes for the effect of asset liquidity at the time of sale.

Finally, in principle, some agents may be able to craft advertisements that are so unique and appealing that they influence the sales prices, perhaps by generating foot traffic, beyond the actual desirable attributes associated with the housing unit itself.

We analyze a data set that encompasses more than 40,000 single-family houses that were advertised and then sold in the metropolitan housing market of Atlanta, GA from 2010 to 2017. The analysis results suggest that houses with unique property descriptions are associated with higher sales prices than those units with less unique descriptions. Comparisons among houses located in geographical proximity and advertised in the same year show that a one standard deviation increase in description uniqueness leads to a 15% increase in the property sales price. Similarly, when comparing housing units that sold more than once, we find that a one standard deviation increase in uniqueness increases the sales price of the same unit by 10%. Even in the hedonic model, the information provided by the measure of uniqueness is, at most, weakly correlated with traditional hedonic attributes like square feet of living space or the number of bathrooms. On the other hand, the inclusion of uniqueness in the hedonic regressions reduces the importance of less well-defined hedonic variables like an indicator for whether a unit was newly renovated and an indicator for whether the house has special features. Further, our measure of uniqueness captures information that is independent of traditional controls for key words. Including advertisement keywords only reduces the impact of uniqueness on prices from 15% to just under 14% in the hedonic model and from 10% to 8% in the repeat sales model.

We estimate additional models to separate the price premium of uniqueness into two channels: the unobserved quality of the housing unit, and the influence of market power in a differentiated product market. Rather than calculating uniqueness by geographic area and year, we start with our sample of all transactions within a geographic area and then for each unique transaction we divide all other transactions in the same area between transactions that occur within six weeks before or after the transaction and those that occur outside of that window. The competing unit measure of uniqueness is intended to isolate the effects

of market power, while eliminating competing units from the measure of uniqueness should tend to focus the measure more on unobserved attributes of the housing unit.

While the two new measures are highly correlated, the findings suggest that the effects of uniqueness in the hedonic model are due to both capturing unobserved attributes and market power, but the effects in the repeat sales model arise primarily from market power. The magnitude of the hedonic estimates on both measures are similar in magnitude to each other whether included independently or together in a horse race specification. In the repeat sales model, the horse race specification including both uniqueness measures yields significant estimates on uniqueness based on competing units and a near zero estimate based on non-competing units. The estimated effects of market power on housing price are comparable to effects of bargaining differentials on price estimated in Harding et al., 2003.

The third proposed mechanism was the possible role of agents in writing the advertisements and influencing the final sales price of the housing unit. The inclusion of agent fixed effects does erode the estimates by one and one-half percentage points in the hedonic model and by three percentage points in the repeat sales. However, while the ability of agent fixed effects to explain part of the uniqueness effect could be driven by real estate agent behavior, this ability also could arise because real estate agents differ in the type, quality and uniqueness of the housing units they represent. If the agents were truly influencing the sales price beyond the intrinsic value or market power of the housing unit, presumably these effects would be also reflected in the spread between the list price and the sales price (Haurin et al., 2010, Merlo and Ortalo-Magne, 2004), but we find no relationship between the list-sales price spread and uniqueness.

Finally, we use standard hedonic and repeat sales models to estimate price indexes (Silver, 2016 and Hill, 2013). We find changes in our transaction sample over time in terms of our uniqueness measure of almost 1/2 a standard deviation increase in uniqueness. This compositional change arises beginning in 2011, relatively early in the recovery from the Great Recession. The change in the composition of the housing stock biases the traditional

hedonic price index post-2010 upwards by between 10% and 23%. We observe a similar, but smaller, upward bias in the repeat sales model of between 6% to 9%. Further, the biased price indexes suggest stabilization of the housing market in 2011, with prices falling by less than 6% in the hedonic model and flat in the repeat sales model. On the other hand, the index that is conditioned on uniqueness falls by 16% and 7% in 2011 based on the hedonic and repeat sales price indexes, suggesting that the housing market did not stabilize until 2012. These findings are consistent with the Lovo et al. (2014) model of art sales that predicts changes in the composition of sales over market cycles, which lead to bias in price indexes.

Our investigation indicates that textual property descriptions provide important information for quantitatively explaining housing prices. The uniqueness measure's ability to explaining housing prices appears driven by two factors, housing attributes that are typically not captured in traditional property databases and the ability of the seller to extract surplus based on the housing unit's uniqueness compared to other available housing. In the case of repeat sales models, the importance of uniqueness appears to arise almost entirely due to market power. These findings on market power provide unique evidence that housing markets are relatively non-competitive operating as differentiated product markets, as in Bayer et al. (2007) and Wong (2013), and that not accounting for market power may lead to errors in estimated price indexes. The market power findings also complement the literature on bargaining in housing markets (Han and Strange, 2015, Harding et al., 2003, and Merlo and Ortalo-Magne, 2004).

2 Methodology

Our research design follows three steps. First, we use our natural language processing (NLP) algorithm to quantify the semantic meaning of real estate property descriptions. Each description is represented as as numeric values in a high-dimensional vector space based on its

contents. These vectors arise during the estimation process and are not predefined by the researcher to represent any specific housing attributes.[5] Second, we calculate the average pairwise difference between every house i in our data and its neighboring houses to identify the uniqueness of house i. Finally, we estimate the impact of description uniqueness on real estate sale prices using hedonic and repeat sales price models, and then use similar models to construct hedonic and repeat sales price indexes. We introduce the NLP model in Section 2.1, document the calculation of uniqueness in section 2.2, and in Section 2.3 we describe our hedonic and repeat sales model specifications.

2.1 The Machine Learning Semantic Analysis Model

Natural language processing (NLP) algorithms use mathematical and statistical methods to help computers learn and process human language. In this paper, we use a large number of real estate property descriptions in order to estimate latent variables associated with each property description. We estimate these latent/unobserved attributes of real estate descriptions by means of the paragraph vectorization (PV) method first introduced in Le and Mikolov (2014).

PV can be represented as a multinomial choice model that is estimated using maximum likelihood estimation where the goal is to select the identity of a specific word in a given document. Below, we sometimes refer to this word that the procedure is trying to select as the "target" word. The first step is to create a list of all unique words appearing in the entire data set of real estate property descriptions. This list of words becomes the choice set for the multinomial choice problem where the total number of unique words K appearing in the real estate description data is the number of choices.

Next, we create the regression sample that will be used to estimate the multinomial choice model. The property descriptions in our data must be converted into a traditional

[5]These types of algorithms are typically referred to as unsupervised machine learning, as opposed to the supervised algorithms discussed in Mullainathan and Spiess (2017). Unsupervised methods are typically used to process non-quantitative data like text, while supervised machine learning methods are applied to high dimensional quantitative samples, such as housing prices in the case of Mullainathan and Spiess (2017).

flat data set of observations with a dependent variable and many independent variables for each observation. To create observations, we use every word i from every property description j in our data. Each location i represents an observation within a property description j. The dependent and independent variables are essentially variables identifying the words occurring in specific locations. The identity of the actual word in location i is the "target" word w_{ij}^{out} that we will attempt to select from the choice set and so represents the dependent variable in our multinomial choice model.

The first set of independent variables that will be used to select the identity of the word in location i and property description j is the list of words nearby the target word. Specifically, we define a bandwidth of length L around location i and use all words within the bandwidth to select the identity of the word in location i. This resulting set of "input" words is a vector of length L denoted as w_{ij}^{in}. The intuition behind this selection approach is that the identity of a word in a given location is best suggested by the identity of the nearby words. For example, if the target word in location i is "southern", then one might expect that words like "charm" or "graceful" are likely to appear nearby, i.e. within the bandwidth, and so if those words occur nearby, then one would assign a higher likelihood of the word in location i being "southern".

In practice, PV successfully selects the words in each location by assigning similar latent attributes to words that tend to appear together in the text. This process is entirely data driven, and we do not apply any structure or initial meaning to attributes that are being estimated. We define Z_k as the vector of latent attributes for the k-th word from the list of words drawn from the entire set of property descriptions. Next, we define w_{lij} as the lth element of the input word vector w_{ij}^{in}. We then define an aggregate latent variable vector $\overline{Z}_{ij}$ by averaging the latent attribute vectors for each word appearing within the bandwidth around location i in paragraph j.

$$\overline{Z}_{ij} = \frac{1}{L} \sum_{l=1}^{L} Z_{w_{lij}} \tag{1}$$

where $\overline{Z}_{ij}$ captures the average attributes of the words that will be used to select the identity of the target word w_{ij}^{out}.

In addition to the nearby words, the selection of the target word also depends upon the property description itself. Certain words may be more likely to arise in some advertisements describing certain types of properties, and not in other advertisements. The latent attributes of the property description adjust to maximize the likelihood of observing the actual target words in that description conditional on the neighboring words surrounding each target word. So, for example, if the word "southern" in location i has a low probability because other words related to "southern" do not appear nearby within property description j, then the low likelihood of the word arising based on the neighboring words must be offset in the model by characterizing this property description as one in which the word "southern" is more likely to arise. We define the vector of latent description attributes as X_j. Essentially, the attributes of a property description are based on explaining the presence of words that were unlikely to arise given the actual words nearby. These estimated property description attributes represent the final goal of this entire procedure, which is to characterize the text and housing unit that is described by that text.

In summary, each word in each property description forms a triplet $(w_{ij}^{out}, j, w_{ij}^{in})$ where w_{ij}^{out} is the dependent variable, i.e. the word selected or the outcome from the choice set, j is the property description, and w_{ij}^{in} is the vector of input words based on the location of the word within the property description. Using this triplet, we can calculate the index x_{ijk} that describes the likelihood that the target word in location i in description j is the k-th word in the choice set, given the set of L input words around location i in property description j.

$$x_{ijk} = \beta_k{}' X_j + \gamma_k{}' \overline{Z}_{ij} \tag{2}$$

where β_k and γ_k are the parameter vectors estimated for each choice/word in the multinomial logit model. As in any choice model, these parameters relate the attributes of the observation, i.e. the attributes of the property description and the attributes of nearby words, to the

propensity of observing a given word k in that location.[6] Correspondingly, the correct selection for the actual output word w_{ij}^{out} should be

$$x_{ijw_{ij}^{out}} = \beta_{w_{ij}^{out}}{}' X_j + \gamma_{w_{ij}^{out}}{}' \overline{Z}_{ij} \tag{3}$$

Given the multinomial choice framework, we can describe the circumstance where the target word w_{ij}^{out} arose in that location as

$$x_{ijw_{ij}^{out}} + \epsilon_{ijw_{ij}^{out}} = Max_k[x_{ijk} + \epsilon_{ijk}] \tag{4}$$

which is the standard condition for determining the outcome of multinomial choice problems. If we assume that ϵ_{ijk} follows an extreme value distribution, then the equation above yields a multinomial logit model, and the conditional probability of the target word occurring can be written as:

$$Pr\left[w_{ij}^{out}|i,j\right] = \frac{e^{x_{ijw_{ij}^{out}}}}{\sum_{k=1}^{K} e^{x_{ijk}}} \tag{5}$$

where x_{ijk} is a function of the paragraph j and the location i of the word within the paragraph.

Finally, defining I_j as the number of words in property description j and J as the number of property descriptions, the log likelihood problem can be written as

$$\underset{\beta_k, \gamma_k, X_j, Z_{ij}}{Min} \sum_{j=1}^{J} \sum_{i=1}^{I_j} -log\left(Pr\left[w_{ij}^{out}|w_{ij}^{in}\right]\right) \tag{6}$$

[6]The symbol $'$ represents the transpose of the vector.

where

$$
\begin{aligned}
log\left(Pr\left[w_{ij}^{out}|w_{ij}^{in}\right]\right) &= x_{ijw_{ij}^{out}} - \log\sum\nolimits_k \exp\left(x_{ijk}\right) \\
&= \left(\beta_{w_{ij}^{out}}{}'X_j + \gamma_{w_{ij}^{out}}{}'\left(\sum_{l=1}^{L} Z_{w_{lij}}\right)\right) - \log\sum\nolimits_k \exp\left(\beta_k{}'X_j + \gamma_k{}'\left(\sum_{l=1}^{L} Z_{w_{lij}}\right)\right)
\end{aligned} \tag{7}
$$

Unlike a traditional multinomial logit, this model not only requires estimation of the standard parameter vectors β_k and γ_k, but also estimation of the property description and word latent variable vectors X_j for all property description j and Z_k for all words k.[7]

The parameter estimates of the property description attributes or latent variables are the final output of the PV approach, and these estimated attributes are used in the next section to measure the uniqueness of housing units based on the property description text.[8]

2.2 Construction of the Uniqueness Measure

From the previous subsection, we obtained vector representations to quantify the information content/latent attributes of the property descriptions. For every pair of properties, we calculate the pairwise distance between the two vectors of latent description attributes to capture the content differences between these property descriptions. This distance is measured using the cosine angle between a pair of vectors obtained during the vectorization process (X_j in

[7]The resulting optimization problem is extremely high dimensional, and the estimated parameters are interactive. Therefore, like many machine learning applications, the objective function is being optimized in an environment where data is relatively sparse, and so traditional optimization approaches are unlikely to perform well. Accordingly, we follow Mikolov et al. (2013) and Le and Mikolov (2014) and minimize the log-likelihood function using a optimization technique called stochastic gradient descent with backpropagation, which is often used in machine learning applications.

[8]The standard multinomial logit parameters that map from the attributes to the choice propensities can be viewed as incidental parameters. In fact, we empirically verify that the mean and standard deviation of paragraph weights for each latent attribute across all target words are very similar. Therefore, these parameters do not have any meaningful impact on the relative importance of each latent attribute in selecting target words.

Equation 2), shown in the equation below.

$$Distance(\boldsymbol{X}_1, \boldsymbol{X}_2) = 1 - cos(\boldsymbol{X}_1, \boldsymbol{X}_2) = 1 - \frac{\boldsymbol{X}_1 \cdot \boldsymbol{X}_2}{||\boldsymbol{X}_1|| \cdot ||\boldsymbol{X}_2||} \tag{8}$$

Note that 0 implies two identical descriptions with 0 semantic distance in between. This measure is mathematically bounded between 0 and 1.

Figure 1 provides a simple demonstration of the effectiveness of our ML algorithm. In the top text box, the key word search phrase "Lenox Mall", a shopping center in northern Atlanta, is used to identify a specific property that mentions the key word in the description. We then search for properties that are the most similar based on our distance measure in Equation (8). The blue pins on the map are houses identified based on distance in attribute space including the unit itself. The middle text box displays the description of a selected house on the map. The bottom text box shows the most similar descriptions found in the data by the PV algorithm via Equation (8). This figure shows that our algorithm can successfully sort houses based on description similarity/difference. In this particular example, the top five most similar houses are near the Lenox mall although the name of the mall does not directly show up in some of the descriptions.[9]

Similarly, Table 1 compares the pairwise semantic distances between the property description of a subject house with that of a few comparables in a neighborhood called the Grant Park subdivision. The first column presents the description of the subject house. The second column presents property descriptions for four houses, first the subject house and then three additional houses in the same subdivision. The third column presents the semantic distance. The distance 0 in the first row implies that the description is being compared to itself. The pairwise distance between two descriptions increases as their semantic meanings deviate from each other. Notice that in the house descriptions, there are many abbreviations and typos. For instance, "granite" vs. "granit," "b'ful" vs. 'beautiful," "hrwds" vs. "hard–

[9]We conceal the house ID and the program copyright note to protect data privacy as well as to hide our names during the review process.

wood–floors," etc. The relationship between the paragraphs cannot be properly analyzed via a simple method based on key words or word frequencies unless all of the text is analyzed by hand to detect all the possible variations for the same word or phrase.

To assess the uniqueness of a description compared to a relevant set of comparison descriptions (within the same market area and year in this paper), we compute the average pairwise distances from the house of interest to other houses, as shown in Equation (9) and Figure 2.

$$Unique_i = \frac{\sum_1^{N-1} (pairwise\, distance)}{(N-1)\ pair\, of\, houses} \tag{9}$$

Once the uniqueness scores have been obtained, we include this variable in our pricing models, which will be introduced in detail in Section 2.3.

2.3 Hedonic and Repeat Sales Price Models

Following Rosen (1974), a massive literature developed using quantitative data on real estate transactions to assess the price impacts of key housing characteristics and neighborhood disamenities (Palmquist, 1984, Lindenthal, 2017, Muehlenbachs et al., 2015, Bernstein et al., 2018, Ambrose and Shen, 2020).[10] Further, these approaches have been used extensively throughout the world to measure asset prices over time using either hedonic or repeat sales price indexes.[11] However, houses are heterogeneous goods for which some characteristics,

[10]Examples of numerical data include, but are not limited to, asking price, sales price, size, age, number of bedrooms, number of bath rooms, property type and location

[11]The best known price indexes in the U.S. include the Standard and Poor's/CaseShiller (SPCS) Home Price Indexes, CoreLogic National Home Price Indexes and Office of Federal Housing Oversight (OFHEO) indexes all use the repeat sales method. However, the One Family Houses Price Index of the Census Bureau, Multi Family House Price Index of the Bureau of Economic Analysis (BEA) (see de Leeuw, 1993), and the FNC Residential Price Index all use the hedonic approach. Further, worldwide, the hedonic approach is by far the most common approach. Notable examples are the Halifax Home Price Index in the UK, the permanent tsb index in Ireland, the Conseil Suprieur du Notariat (CSN) and INSEE (the national statistical office of France) index in France, the Zrcher Wohneigentumsindex (ZWEX) in Switzerland, the indexes published by the statistical offices of Finland, Norway and Sweden, and the RPDataRismark indexes in Australia. Other less transparent hedonic indexes include the Verband Deutscher Pfandbriefbanken (VDP) and Hypoport AG indexes in Germany and the Recruit Residential Price, Residential Market and Tokyo Area Condominium Market Indexes in Japan (Hill 2013).

e.g. the condition of floors, windows or cabinets ("soft" information), cannot be easily captured by numerical data, and the accuracy of the widely used real estate pricing frameworks may suffer from the omission of these unobserved attributes (Liberti and Petersen, 2018; Garmaise and Moskowitz, 2004). While repeat sales models will control for time-invariant housing attributes, renovations and differences in maintenance may lead to substantial quality differences between transactions of the same housing unit that could be attributed to changes in housing prices.

We employ classic log-linear hedonic and repeat sales housing price models to estimate the impact of unique property descriptions on home prices. The hedonic specification takes the following form:

$$Ln(Price_i) = \alpha + \theta Unique_i + X_i'\beta + \mu_{ct} + \eta_z + \varepsilon, \tag{10}$$

where $Ln(Price_i)$ is the natural log of the sold price of house i. $Unique_i$ is the description uniqueness score derived from our NLP model. X_i is a vector of physical characteristics and in some specifications transaction circumstances of the sale or advertisement keywords. The physical characteristics include number of bedrooms (Bed), square footage in hundred (Sqft), age (Age), number of fireplaces (Fireplaces), size of lot (Large Lot), whether the house has a pool (Pool), whether the house is recently renovated (Renovated), and whether the house comes with a special recreational feature such as access to a lake or a golf course (Feature). In some specifications, this vector also includes dummy variables that indicate whether a sale has the following transaction circumstances: sold without a repair escrow (Sold–As–Is), sold by an agent who represents both the seller and the buyer (Dual Agent), and listing agent is the seller or is related to the seller (Owner Agent). μ_{ct} is a vector of MLS market area by year fixed effects. Again, in some specifications, a vector η_z of listing agent fixed effects is also included.

The repeat sales model is identical except for the inclusion of housing unit fixed effects and the restriction to a sample of housing units that sold more than once during the sample

period.

$$Ln(Price_i) = \alpha + \theta Unique_i + X_i'\beta + \mu_{ct} + \eta_z + \gamma_i + \varepsilon, \tag{11}$$

where γ_i is the housing unit fixed effect. For both models, standard errors are clustered at the level of the MLS market area by year.

For our price indexes, we estimate a slightly different model. We follow the time dummy approach to estimating time fixed effects within the hedonic price indexes. However, we exclude controls for geography within our metropolitan area because unlike hedonic regressions used for inference purposes, most price indexes are based on hedonic regressions that do not include sub-geography fixed effects.

$$Ln(Price_i) = \alpha + \theta Unique_i + X_i'\beta + \delta_t + \varepsilon, \tag{12}$$

The vector of control variables in these models does not contain additional controls for sale circumstances or advertisement key words since those controls are not standard in the estimation of hedonic price indexes. Recognizing the concern about correlation over space in unobservables for the entire housing stock regardless of when the housing unit was sold or advertised, the standard errors in these models are clustered at the zip code level, rather than by geography by year as in the price models above. The repeat sales price index model is identical except again for the addition of the housing unit fixed effect γ_i

Finally, given the log-log structure, the price index for a given year t can be calculated from the estimates of the year fixed effects relative to a base year 0 as:

$$I_t = \frac{exp(\delta_t)}{exp(\delta_0)} \tag{13}$$

3 Data and Descriptive Statistics

Our data encompasses more than 40,000 single-family home sales in the metropolitan real estate market of Atlanta, GA, from the beginning of 2010 to September of 2017. The source of the data is the Multiple Listing Service (MLS), and the metropolitan area is defined by the MLS, rather than traditional county boundaries, which would include many rural areas. The information provided in the MLS data includes the address of each house identifying the MLS market area (submarket) in which the house is located, the listing and the sales price, a wide range of house characteristics, critical dates regarding the transaction, unique IDs of the listing and buying agents, and, most importantly, the written property description.

Our property description data meet the following two conditions: First, the market areas in our sample are with more than three sales in each year during the sample period. Second, we only include houses for which the property descriptions are longer than nine words. Less than 100 property sale records out of more than 40,000 transactions, had text stating the description text along with other housing attributes data have been deleted. All valid property descriptions with valid housing feature data had more than nine words. For each sale, we calculate uniqueness comparing the property description/advertisement of each housing unit in our sample to all other housing units that are located in the same MLS market area and were sold within the same year. The Atlanta housing market is divided into 25 MLS market areas.

The final data used in this study consist of 40,918 transactions including 7,346 repeat sales and 3,794 houses that involved repeat sales of the same housing unit. We use the entire sample to estimate our hedoninc results and the repeat sales data to estimate repeat sales models. Unlike typical assessor's records, MLS data record the attributes of the housing unit at each listing so that we can observe whether a renovation occurred between sales and any changes in the housing attributes between sales. Table 2 panel A displays a set of basic descriptive statistics for the data used in this study for both the full and repeat sales samples of transactions. The average home in our sample is 46 years old, has 2.7 bedrooms

and 3.6 bathrooms. It is listed for $390,000 and is sold for $373,000 an average three and a half months after listing. Since our housing data are in the metropolitan area of Atlanta, most of the houses sold in this area sit on small lots that are less than an acre.[12] Only 2.5 percent of the homes in our data are built on lots that are greater than one acre. The repeat sales sample is similar in composition with listing prices about $15,000 lower and sales prices only $10,000 dollars lower than the full sample. Consistent with the lower prices the housing units in the repeat sales sample are somewhat smaller, about 10 percent less square feet, and have on average fewer bathrooms and bedrooms. Notably, the units in the repeat sales sample are more likely to have been renovated, and consistent with being in denser, more active local real estate markets are more likely to be on small lots.

Table 2 panel B displays a set of basic descriptive statistics for the property description uniqueness score variable estimated by the machine learning algorithm. The average description in our data uses six sentences and 80 words to describe a house for sale. $Unique_i$ measures the semantic difference between the property description of house i and descriptions of neighboring houses sold during our sample period. This measure is bounded between 0 (a low level of semantic deviation) and 1 (a high level of semantic uniqueness). Neighboring houses are defined as homes sold within the same MLS area in the same year. $Unique_i$ clusters around 0.7 with the minimum value equals to 0.38 and its maximum value is 0.94. The statistics for the repeat sales sample are nearly identical.

4 Empirical Results

In this section, we present empirical results of both our hedonic and repeat sales models to explore the effects of $Unique_i$ on real estate sale prices.[13] Table 3 displays the estimates from these two price models, both with and without our control for uniqueness. The MLS data

[12]The Census defined Atlanta Metropolitan Statistical Area is based on county, and the outlying counties will often contain large rural areas, but the MLS defined metropolitan market area omits most of those rural locations.

[13]Similar results arise using property listing price

contains information on the housing attributes at the time of every transaction, allowing us to identify the effect of hedonic characteristics in the repeat sales model based on changes in those characteristics between sales.

The R-squared of the models increase modestly from 0.760 to 0.769 in the hedonic model and from 0.956 to 0.958 in the repeat sales model. In both cases the addition of the uniqueness measure absorbs between 4 and 5 percent of the residual variance in the model. In both models, uniqueness has a substantial impact on price with a one standard deviation increase in uniqueness associated with a 14.9% increase in the sales price in the hedonic model and a 9.9% increase in price in the repeat sales model. The smaller 9.9% effect of uniqueness for the repeat sales model suggests that at least part of the effect of uniqueness on prices in the hedonic model is due to unobservable housing quality or attributes that are differenced out in the repeat sales model.[14]

Turning to the estimates on traditional hedonic attributes like the number of bedrooms, baths, and square footage, we should first note significant differences between the hedonic and repeat sales model estimates. The hedonic model implies higher prices for newer units, larger units based on square footage, units with more bathrooms, renovated units. units with pools and units with special features. The hedonic model implies lower prices for ranches and units with more bedrooms. In the repeats sales model, the return for renovating a unit is about 10 percentage points higher than in the hedonic model, and similarly adding a bathroom during a renovation has an 8 percentage point larger effect than the estimated effect of a unit having an additional bathroom in the hedonic model. However, simply adding square footage or bedrooms appear to have minimal additional effects on sales price compared to other renovations that might provide valuable modernization that is not recorded in the hedonic attributes. Further, changes in status as a ranch, adding a pool or adding other special features to the house do not appear to affect sales price unlike in the hedonic model.

[14]In an unreported analysis, we also found results for a model of days of the market finding a small increase of approximately 4 days on the market for a one standard deviation increase in uniqueness. These findings are consistent with real estate agents or owners keeping housing units that are more unique or have more market power on the market longer in order to obtain a higher sales price.

Given the lack of effects in the repeat sales model, the effects of being a ranch, having a pool or other special features in the hedonic model may arise in part due to a cross-sectional correlation between these attributes and the unobserved quality of the housing unit. The effect of age on value is relatively stable between the hedonic and repeat sales.

Comparing Column (1) to Column (2) and Column (3) to (4), the addition of the uniqueness control has minimal effects on the estimates for traditional hedonic variables like age, square feet, bedrooms and bathrooms. However, the hedonic model coefficients on indicator variables for ranch, renovation, and recreational features all decline (between 7 and 29 percent) with the addition of the control for uniqueness, consistent with uniqueness capturing unobserved attributes that are correlated with these variables. On the other hand, the effect of ranch and other features on prices in the repeat sales model remains near zero.

Next, in Table 4, we examine alternatives to the model in an attempt to distinguish between whether uniqueness is capturing the price effect of unobservable housing unit quality or capturing the effect of market power that arises when a housing unit is relatively unique compared to other housing units on the market at the same time. Instead of focusing on housing units sold in the same market area and year, we start with the sample of housing units sold during the entire sample period in the market area. For each housing unit, we divide the sample of sales in the market area into a subsample containing housing units that were available within six weeks before or after the sale date of the housing unit of interest and a second subsample containing all other sales. Uniqueness as measured using the first subsample of sales (potentially competing units) captures the housing unit's uniqueness or market power relative to housing units that had a very good probability of being on the market when this housing unit was also on the market. The second subsample contains housing units that were likely not on the market at the same time (non-competing) as the housing unit of interest, and so potentially captures information on the unobserved quality of the housing unit that would have value regardless of how differentiated the unit is from other units currently on the market.

Columns (1) and (5) repeat the estimates from the previous table for the hedonic and repeat sales models respectively. Columns (2) and (6) present estimates using the new MLS area measure of uniqueness based on competing units, and Columns (3) and (7) present estimates using uniqueness based on non-competing units. Finally, in Columns (4) and (8), we run a horse race between the uniqueness measure based on competing units and based on non-competing units. In Columns (2) and (3), the standardized effect in the hedonic model falls from 14.9 percent to 12.5 and 11.2 percent for competing and non-competing measures, respectively. While the two measures are highly correlated and resulting horse race estimates noisy, we similarly find in Column (4) that the estimates on the competing and non-competing uniqueness measures are similar at 6.9 and 5.9 percent, respectively. Turning to the repeat sales model, both the estimates on the competing and non-competing measures are similar in magnitude at 9.0 and 6.8, but the decline from the initial estimate of 9.9 is three times larger for the non-competing measure. Further, turning to the horse race model in Column (8), we find that the estimate on uniqueness for competing units is statistically significant at the 10 percent level and similar in magnitude to earlier estimates at 8.4 percent, while the estimate on uniqueness based on non-competing units is near zero. While not conclusive due to lack of precision in the horse race estimates, the results are consistent with uniqueness capturing information about both the effects of unobserved housing quality and market power on prices in the hedonic model, but mostly capturing effects of market power on sales price in the repeat sales model.

Next, we run a series of robustness tests for our primary model using the MLS area by year uniqueness measure. These estimates are shown in Table 5. Panel A presents estimates for the hedonic price regression, and Panel B presents estimates for the repeat sales model. The first column presents the baseline estimates from Table 3. Column (2) adds transaction characteristics.[15] Column (3) adds dummy variables associated with the presence of specific

[15]While the estimates are not shown in the table, consistent with findings of previous studies, agent-owned houses are associated with higher sale prices than non-agent owned houses (Levitt and Syverson, 2008 and Rutherford and Yavas, 2005); and dual agency transactions are associated with lower sale prices than sales in which different agents represent the seller and buyer (Han and Hong, 2016 and Brastow and Waller, 2013).

key words in the real estate property description following Levitt and Syverson (2008) and Rutherford and Yavas (2005),[16] and Column (4) includes agent fixed effects.

Comparing Column (1) to Column (2), the inclusion of transaction attributes has little effect on model fit and also leaves the estimates on the uniqueness control relatively unchanged. In Column (3), key words have substantial ability to explain variation in housing prices increasing the R-squared from 0.784 to 0.811 in the hedonic model and 0.961 to 0.968 in the repeat sales model, explaining 13 and 18 percent of the residual variance respectively. However, this substantial increase in R-squared arises from the inclusion of approximately 50 dummy variables associated with the presence of specific key words, and so the fact that a single regressor, uniqueness, explains 4 to 5 percent of the residual variance suggests that uniqueness is substantially more important than most or all of the individual key words identified by Levitt and Syverson (2008). The inclusion of key words modestly decreases the effect of uniqueness from 14.9% to 13.6% in the hedonic model and from 9.8% to 8.0% in the repeat sales model. Therefore, keyword strategies capture only a modest share of the information captured by our uniqueness measure.

Finally, in Column (4), we observe that inclusion of agent fixed effects also erodes the magnitude of our estimates. Specifically, in the hedonic model, the inclusion of agent fixed effects reduces the estimated effect of uniqueness from 13.6% to 12.2%, and in the repeat sales model the fixed effects reduce the effect of uniqueness even more substantially from 8.0% to 5.0%. At the same time, interpretations of these changes is more difficult than interpreting the effects of including key words. While it is possible that some agents both write more unique property descriptions and also bargain harder raising eventual sales price, agents also differ substantially in their ability to attract high quality and high value property listings, and so agent fixed effects could erode the estimated effect of uniqueness simply because unique properties are more valued in the market and some agents are better positioned than others to capture those high value listings.

In addition, houses without repair escrows are sold for lower prices than those with repair escrows.

[16] We include the same words listed in Levitt and Syverson (2008) Table 1.

Additional insights into the role of agents may be provided by examining the difference between the listing price and the sales price. While both the listing and sales price primarily reflect the market value of the property, the listing price may depart from the market value systematically based on agent behavior and perceptions. Haurin et al. (2010) shows that in a search model where agents set a list price, atypical properties not only have a wider array of offers and sell for more on average but also have a greater spread between the list price and the final sales price. Similarly, some agents may be more optimistic or aggressive in pricing units and, thus, possibly extract additional surplus in housing market sales. If those agents write property descriptions differently than other agents when they are being either more aggressive or overly optimistic , then at least part of our estimated effect of uniqueness in both the hedonic and repeat sales models might be the effect of agents on price, rather than the effect of the housing units' market power or unobserved quality. Such behavior is likely to show up as a premium in the list price relative to the final sales price. For example, if the agent simply made a mistake overestimating the value of the unit, then the property is likely to sell near the lower, actual market price. Similarly, an agent who is more aggressive in pricing and marketing may indicate their higher reservation price in the eventual bargaining problem by selecting a higher list price and adjusting the property description.

If bargaining is Nash, one might expect that only half the increase in reservation price would be reflected in the eventual sales price, again leading to a premium in the list price relative to the sales price.[17] Therefore, a positive correlation between the list-sales price spread and uniqueness is consistent with more optimistic or aggressive agents writing more unique property descriptions, while no relationship between the list-sales price spread and uniqueness suggests that uniqueness effects are primarily due to unobserved property attributes and market power.

Column (5) in Table 3 presents the estimates of the conditional correlation between

[17]According to Merlo et al. (2015), variation in listing price for the same property is likely to have at most a limited impact on eventual sales price given the large role that bargaining plays in determining the final price, consistent with agent changes in list price not being reflected in sales prices dollar for dollar.

uniqueness and the list-sales price spread for our baseline specification in Column (1). The estimates of the effect of uniqueness are in the wrong direction, statistically insignificant, and modest in magnitude - less than $1,000 for a one standard deviation change in uniqueness. Therefore, we do not find any evidence to support the idea that our estimated effects of uniqueness are driven by a correlation between how the agent wrote the property description and the impact of agent marketing behavior on the eventual sales price.

5 Hedonic Housing Price indexes

In this section, we estimate both hedonic and repeat sales price models controlling for standard hedonic attributes and dummy variables for each year for the Atlanta real estate market.[18] The coefficient estimates on the year dummies are used to calculate the values of the price indexes. We estimate these models with and without the controls for uniqueness in order to see if the hedonic price indexes are distorted or biased in a substantial way by the omission of the unobserved housing attributes captured by our uniqueness measure.

Uniqueness appears to have a substantial effect on the estimated hedonic price index and a similar, but more modest, effect on the repeat sales price index. The left hand side panel of Figure 3 graphs the hedonic price index for the Atlanta housing market based on models that exclude and include the control for uniqueness. Note that the estimated indexes with and without uniqueness controls are highly correlated so that standard confidence intervals are uninformative in terms of the differences. Therefore, on each side of the index conditional on uniqueness, we plot error bars around the second price index whose width is 1.96 times the standard error on the difference between the price index estimates with and without the control for uniqueness. The error bars represent the 95% confidence interval for the difference. The price index estimates indicate that house prices had stabilized between 2010 and 2011 when uniqueness is not included in the hedonic model. However, after controlling

[18]We only estimate price indexes from 2010 through 2016 because we do not observe a full year of data for 2017

for uniqueness, prices continue to fall in 2011 and do not start to recover until 2012. The supply of housing on the market in metropolitan Atlanta appears to lead the recovery in terms of uniqueness, with housing prices not starting to recover until a year later.

The changing composition of sales over the recovery is illustrated in Table 6. While housing sales prices on average are the same between 2010 and 2011 (see row 3), the composition changes substantially. Uniqueness shown in row 1 increases by almost 1/2 of a standard deviation.[19] Therefore, after conditioning on uniqueness within the hedonic model, adjusted prices appear to fall between 2010 and 2011. Similarly, valuable hedonic attributes like square feet, number of bathrooms, number of fireplaces, whether renovated, and whether the housing unit is not a ranch (ranches are associated with lower housing prices) increase on average as the market recovers consistent with a change in the composition of the housing stock on the market at the time. The increase in our uniqueness measure suggests that the housing stock is improving on unobservables attributes, as well.

Table 7 Columns (1) and (2) quantify these changes by presenting the estimated coefficients on the year dummies in the hedonic log price model. In 2011, the naive price index without conditioning on uniqueness shows that prices fell by only 5.9% perhaps suggesting that the housing market was beginning to stabilize, while our alternative index suggests that prices fell by 16.2%, a 10 percentage point difference in price change. In 2012, house prices started to recover from their 2011 low, but the gap continues to grow reaching a maximum of 23 percentage points by 2014. After that, the gap between the indexes begins to decline falling to an 18 percentage point gap by 2016. These differences are almost all statistically significant.

The right hand side panel of Figure 3 and Table 7 Columns (4) and (5) present the repeat sales price index for the Atlanta housing market. As with the hedonic price index, the repeat sales price index estimates without uniqueness indicate that prices had stabilized

[19]We do not observe any evidence of a change in the quality of housing as measured by uniqueness for the outlying counties within the metropolitan housing market and as a result those counties have nearly identical price indexes whether or not the hedonic model includes a control for uniqueness.

between 2010 and 2011 with a nearly identical price index values in the two years, but after controlling for uniqueness prices continued to fall in 2011 and do not start to recover until 2012. Specifically, the price index after controlling for uniqueness implies a decline in prices between 2010 and 2011 of almost 7%. These differences are relatively steady over time remaining around 9% by 2016. The smaller differences arise in part due to the smaller estimated effect of uniqueness in the repeat sales and also because the changes in uniqueness looking within housing units (housing unit mean differenced values in row 2) in 2011 is smaller than the cross-sectional composition change documented in row 1. This is consistent with uniqueness in the hedonic model in part capturing a change the quality of houses on the market during the recovery. The repeat sales model conditions out that portion of the change in composition.

While uniqueness has strong explanatory power in hedonic price models and its inclusion substantially changes the hedonic price index, the inclusion of uniqueness does not deliver an hedonic price index that is closer to the repeat sales index. In fact, as shown by Figure 4, the hedonic and repeat sales indexes without conditioning on uniqueness (left hand side of the figure) remain relatively close to each other over the sample period, but after controlling for uniquenes (right hand side) the repeat sales index is always greater than the hedonic index. This pattern may be attributable to a commonly cited bias in repeat sales indexes: unobserved upgrading, improvement and renovation of units between sales that may bias repeat sales housing price indexes upwards because the improved property is compared to a previous sale of the unimproved property (OECD, 2013, Haurin and Hendershott, 1991).

Since the MLS data contains information on renovations and changes in hedonic attributes, we can test for this bias by omitting the observed hedonic variables from the repeat sales model. This omission substantially increases the rate of growth in the hedonic price index over time consistent with upwards bias from omitting observable improvements, similar to findings in earlier analyses of this bias by Abraham and Schauman (1991) and Peek and Wilcox (1991). If a substantial share of upgrading and improvement is unobserved in

our data, then the current repeat sales price index is likely also biased upwards compared to the hedonic price index. Once the hedonic price index has been adjusted for composition changes over our sample period, the hedonic index always lies below the repeat sales index consistent with this upwards bias.[20]

6 Conclusion

In this study, we investigate the price impact of property uniqueness. Earlier studies have looked at atypicality (Haurin et al., 2010) using hedonic housing attributes and local neighborhood distinctiveness using survey data (Ahlfeldt and Holman, 2018). We contribute to this literature by demonstrating a similar relationship between uniqueness and price using a new measure based on textual property descriptions and showing that this measure has a large impact on price. Like these earlier studies, being unique in some way has value in the market place. Further, our paper is the first to demonstrate that uniqueness affects prices in a repeat sales framework, where fixed effects have been used to remove time invariant housing and neighborhood attributes.

Further, our paper provides evidence concerning the potential mechanisms behind the impact of uniqueness on price. When estimating a traditional hedonic model, our measure of uniqueness in part captures information related to the unobserved quality of the housing unit, as proposed by Ahlfeldt and Holman (2018). In addition, similar to Haurin et al., 2010, part of the effect of uniqueness appears to arise from the market power held by a property that is relatively unique. Further, in the repeat sales model, the effects of uniqueness are almost entirely driven by market power. This finding provides important evidence supporting

[20]Two other common biases are 1. selection bias associated with estimating models only with properties that sell more than once, and 2. depreciation is ignored (OECD, 2013, Haurin and Hendershott, 1991). We re-estimate the hedonic price index using only a sample of properties that sold more than once, and the resulting hedonic price index is nearly identical to the index based on the full sample. This finding is consistent with earlier work by Abraham and Schauman (1991) and Case et al. (1991) that concluded that selection did not represent a significant bias in repeat sales models. We follow the current literature and address concerns about depreciation in repeat sales models by including controls for age and as noted previously the effect of age on price is very similar in the hedonic and repeat sales models.

papers that propose modeling housing markets as differentiated product markets, like Bayer et al. (2007) or Wong (2013).

To our knowledge, we provide the first application of recent machine learning methods to quantify the uniqueness of textual property descriptions. All previous studies of the relationship between housing prices and property descriptions have specified a list of key words and includes controls for the presence of those words in housing price models, such as Levitt and Syverson (2008). Our findings strongly suggest that less structured approaches to evaluating property descriptions could add substantial information compared to what can be uncovered based on analyses of key words.

We also examine the influence of our control for uniqueness on traditional hedonic and repeat sales based housing price indexes. During the recovery from the great recession, the composition of the Atlanta housing market changes for the better both on observed attributes and on unobserved attributes that are captured by our uniqueness measure, and this composition change leads the recovery in housing prices. These changes in composition cause the house price indexes that do not consider uniqueness to understate price declines in Atlanta near the end of the recovery and overstate the strength of the recovery, especially in the hedonic index. Given known biases in repeat sales models, hedonic price indexes may provide a much more valuable alternative as the industry begins to incorporate qualitative information about housing into the estimation of hedonic indexes.

Finally, most of the recent machine learning studies in economics and finance focus on predictions (supervised algorithms), rather than unsupervised methods like natural language processing that capture qualitative information. Our paper provides a valuable example of how unsupervised machine learning methods can be applied within economics. Further, our translation of the NLP algorithm from the standard neural network framework (the workhorse of unsupervised machine learning) to a maximum likelihood estimation framework should make it much easier for researchers to apply, adapt and explain unsupervised machine learning approaches in future research projects.

References

Abraham, J. M. and W. S. Schauman (1991). New evidence on home prices from freddie mac repeat sales. *Real Estate Economics 19*(3), 333–352.

Ahlfeldt, G. M. and N. Holman (2018). Distinctively different: a new approach to valuing architectural amenities. *The Economic Journal 128*(608), 1–33.

Ambrose, B. W. and L. Shen (2020). Past experiences and investment decisions: Evidence from real estate markets. *Available at SSRN 3589748*.

Aubry, M., R. Krussl, G. Manso, and C. Spaenjers (2019). Machine learning, human experts, and the valuation of real assets. *Working Paper*.

Bayer, P., F. Ferreira, and R. McMillan (2007). A unified framework for measuring preferences for schools and neighborhoods. *Journal of political economy 115*(4), 588–638.

Bernstein, A., M. Gustafson, and R. Lewis (2018). Disaster on the horizon: The price effect of sea level rise. *Journal of Financial Economics*.

Berry, S., J. Levinsohn, and A. Pakes (1995). Automobile prices in market equilibrium. *Econometrica: Journal of the Econometric Society*, 841–890.

Brastow, R. and B. Waller (2013). Dual agency representation: Incentive conflicts or efficiencies? *Journal of Real Estate Research 35*(2), 199–222.

Case, B., H. O. Pollakowski, and S. M. Wachter (1991). On choosing among house price index methodologies. *Real estate economics 19*(3), 286–307.

Garmaise, M. J. and T. J. Moskowitz (2004). Confronting information asymmetries: Evidence from real estate markets. *The Review of Financial Studies 17*(2), 405437.

Goetzmann, W. and L. Peng (2006). Estimating house price indexes in the presence of seller reservation prices. *Review of Economics and statistics 88*(1), 100–112.

Goodwin, K., W. B. W. H. S. (2014). The impact of broker vernacular in residential real estate. *Journal of Housing Research 23*(2), 143–161.

Han, L. and S.-H. Hong (2016). Understanding in-house transactions in the real estate brokerage industry. *The RAND Journal of Economics 47*(4), 1057–1086.

Han, L. and W. C. Strange (2015). The microstructure of housing markets: Search, bargaining, and brokerage. In *Handbook of regional and urban economics*, Volume 5, pp. 813–886. Elsevier.

Harding, J. P., S. S. Rosenthal, and C. Sirmans (2007). Depreciation of housing capital, maintenance, and house price inflation: Estimates from a repeat sales model. *Journal of urban Economics 61*(2), 193–217.

Harding, J. P., S. S. Rosenthal, and C. F. Sirmans (2003). Estimating bargaining power in the market for existing homes. *Review of Economics and statistics 85*(1), 178–188.

Haurin, D. (1988). The duration of marketing time of residential housing. *Real Estate Economics 16*(4), 396–410.

Haurin, D. R., J. L. Haurin, T. Nadauld, and A. Sanders (2010). List prices, sale prices and marketing time: an application to us housing markets. *Real Estate Economics 38*(4), 659–685.

Haurin, D. R. and P. H. Hendershott (1991). House price indexes: issues and results. *Real Estate Economics 19*(3), 259–269.

Hill, R. J. (2013). Hedonic price indexes for residential housing: A survey, evaluation and taxonomy. *Journal of economic surveys 27*(5), 879–914.

Lawani, A., M. M. R. Reed, T. Mark, and Y. Zheng (2018). Reviews and price on online platforms: Evidence from sentiment analysis of airbnb reviews in boston. *Regional Science and Urban Economics*.

Le, Q. and T. Mikolov (2014). Distributed representations of sentences and documents. In *International Conference on Machine Learning*, pp. 1188–1196.

Levitt, S. D. and C. Syverson (2008). Market distortions when agents are better informed: The value of information in real estate transactions. *The Review of Economics and Statistics 90*(4), 599–611.

Liberti, J. M. and M. A. Petersen (2018). Information: Hard and soft. *Working Paper*.

Lindenthal, T. (2017). Beauty in the eye of the home-owner: Aesthetic zoning and residential property values. *Real Estate Economics*.

Liu, C. H., A. D. Nowak, and P. S. Smith (2020). Asymmetric or incomplete information about asset values? *The Review of Financial Studies 33*(7), 2898–2936.

Lovo, S., C. Spaenjers, et al. (2014). A model of trading in unique durable assets. Technical report.

Merlo, A. and F. Ortalo-Magne (2004). Bargaining over residential real estate: evidence from england. *Journal of urban economics 56*(2), 192–216.

Merlo, A., F. Ortalo-Magné, and J. Rust (2015). The home selling problem: Theory and evidence. *International Economic Review 56*(2), 457–484.

Mikolov, T., K. Chen, G. Corrado, and J. Dean (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Muehlenbachs, L., E. Spiller, and C. Timmins (2015). The housing market impacts of shale gas development. *The American Economic Review 105*(12), 3633–3659.

Mullainathan, S. and J. Spiess (2017). Machine learning: An applied econometric approach. *Journal of Economic Perspectives 31*(2).

OECD (2013). *Handbook on Residential Property Prices (RPPIs)*. International Monetary Fund Washington, DC.

Palmquist, R. B. (1984). Estimating the demand for the characteristics of housing. *The Review of Economics and Statistics 66*(3), 394–404.

Peek, J. and J. A. Wilcox (1991). The measurement and determinants of single-family house prices. *Real Estate Economics 19*(3), 353–382.

Pryce, G., . O. S. (2008). Rhetoric in the language of real estate marketing. *Housing Studies 23*(2), 319–348.

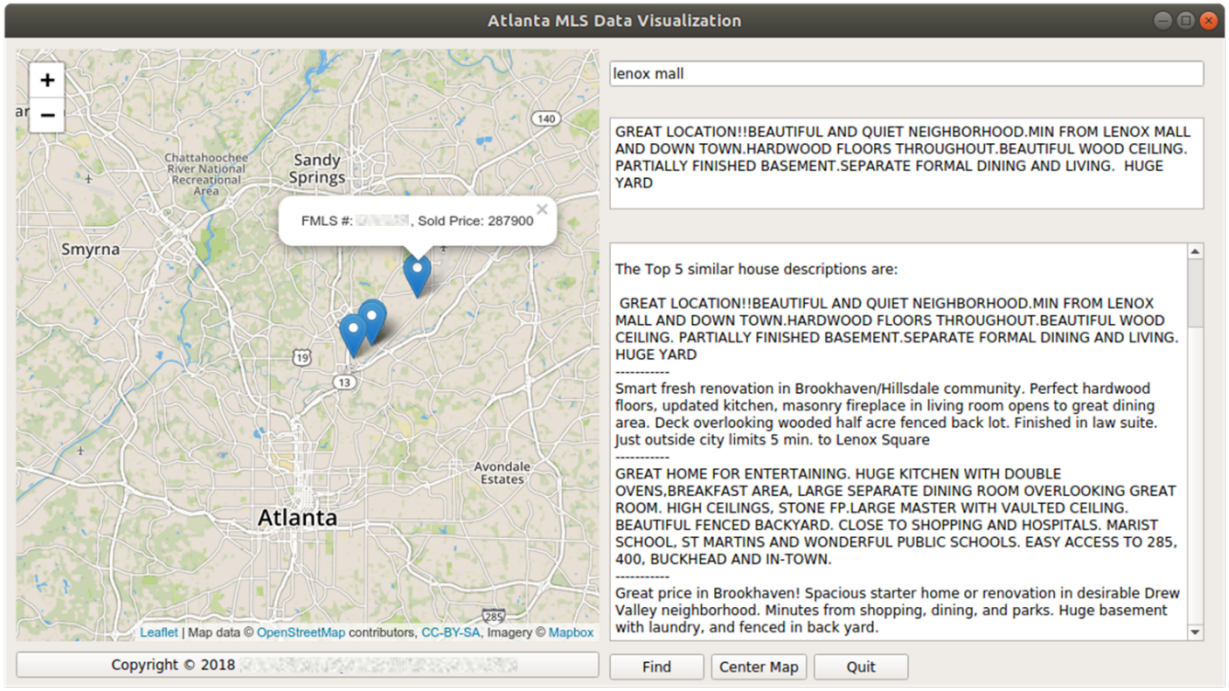
Rosen, S. (1974). Hedonic prices and implicit markets: Product differentiation in pure competition. *Journal of Political Economy 82*(1), 34–55.

Rutherford, R. C., T. M. S. and A. Yavas (2005). Conflicts between principals and agents: evidence from residential brokerage. *Journal of Financial Economics 76*(3), 627 – 665.

Silver, M. S. (2016). *How to better measure hedonic residential property price indexes*. International Monetary Fund.

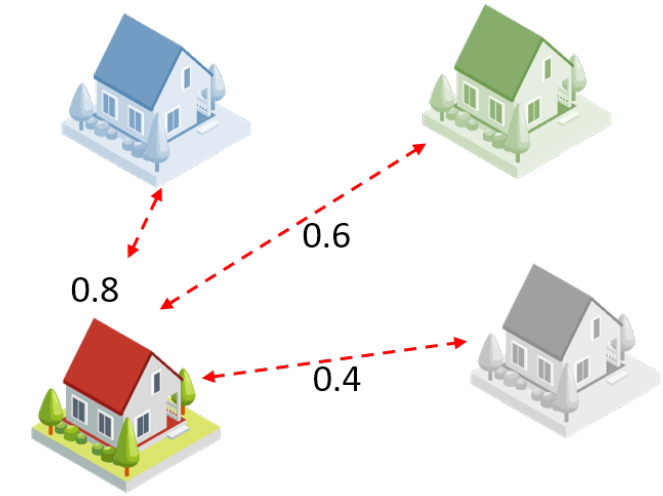
Wong, M. (2013). Estimating ethnic preferences using ethnic housing quotas in singapore. *Review of Economic Studies 80*(3), 1178–1214.

Figure 1: Algorithm Visualization



Notes: This figure provides a visual demonstration of our ML algorithm. In the top text box, the search query is "Lenox Mall", a shopping center in northern Atlanta. In the middle box, we display a selected property description that mentions "Lenox Mall". The blue pins on the map are houses that are related to the housing unit that directly mentions "Lenox Mall" in the property description. The bottom text box shows the most similar descriptions found in the data by the ML algorithm. We conceal the house ID and the program copyright note for data privacy as well as to hide the our names during the review process. In this particular example, all the selected houses are near the Lenox mall although the name of the mall does not directly show up in some of the descriptions.

Figure 2: Schematic Unique Score Computation within a Neighborhood



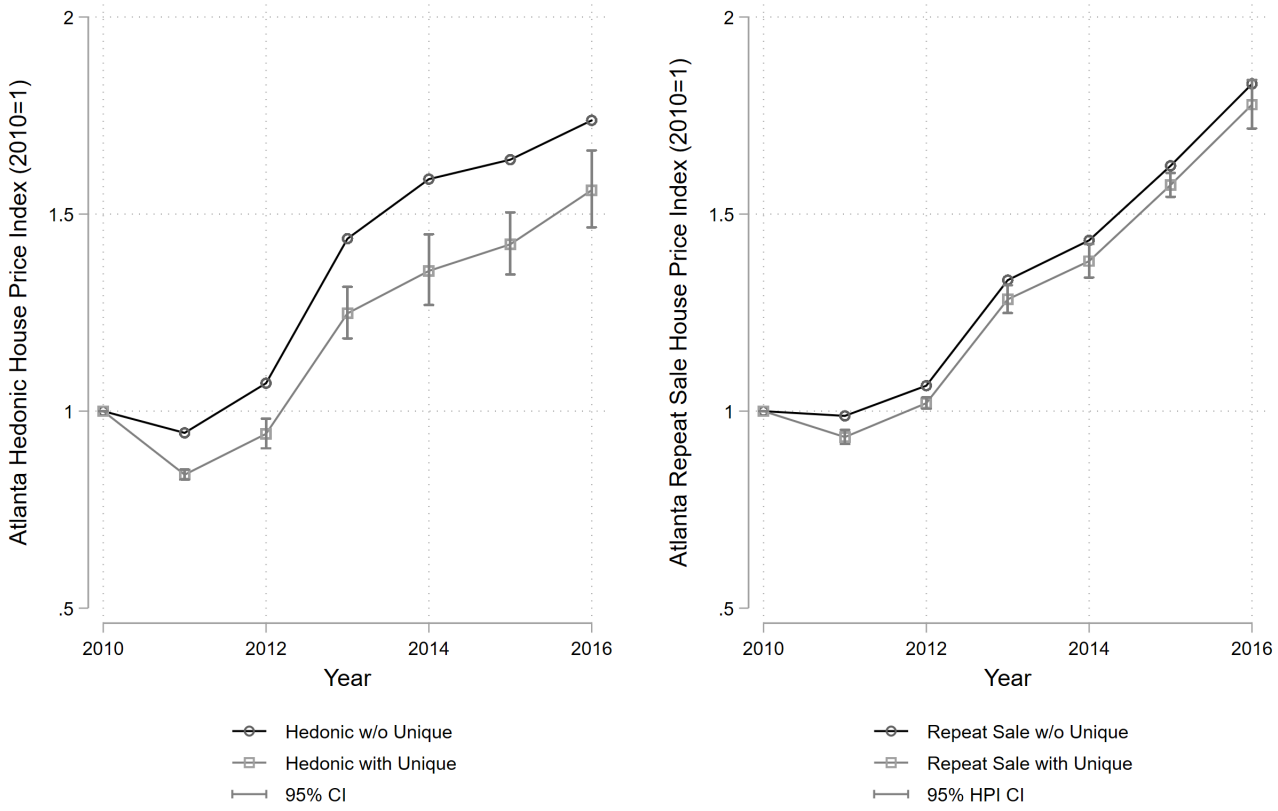
Mean unique score:

$$Unique_i = \frac{\sum_1^{N-1}(pairwise\ distance)}{(N-1)\ pair\ of\ houses} = 0.6$$

for House *i* in neighborhood of *N* houses

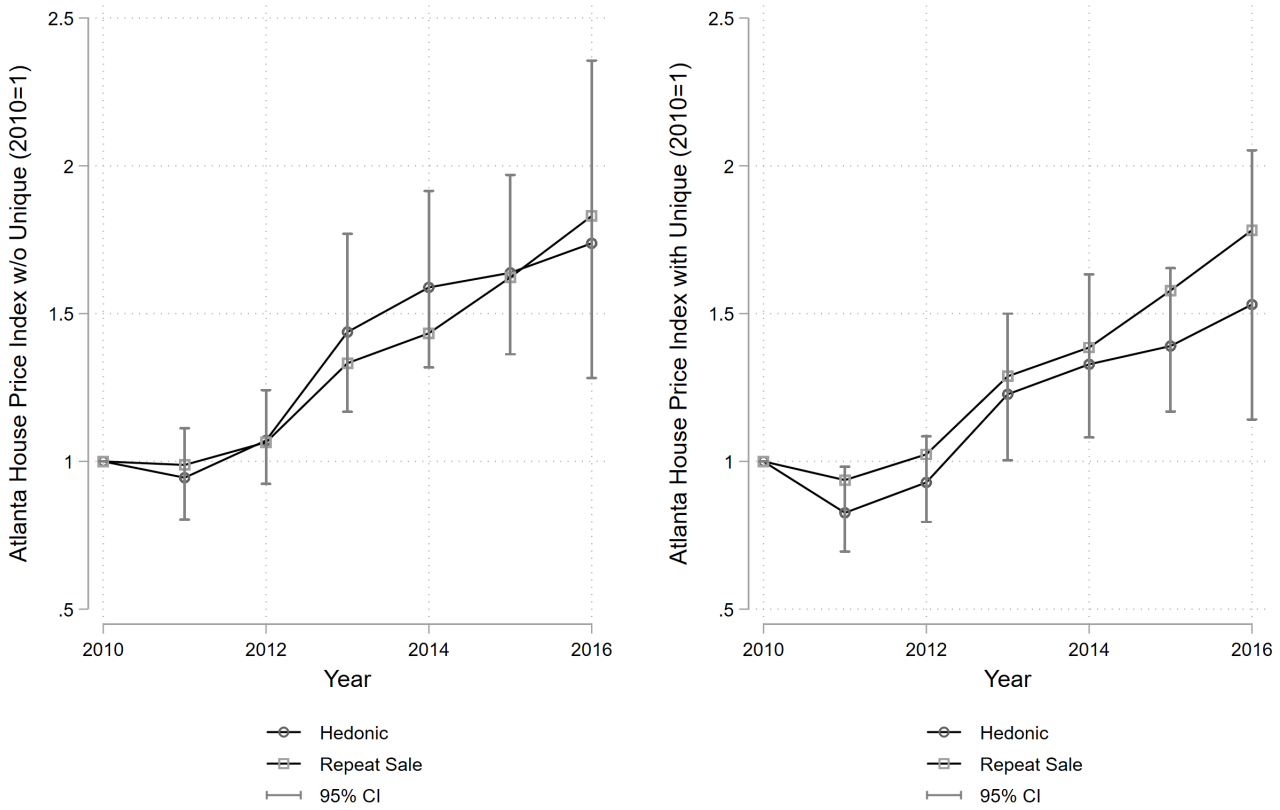
Notes: This figure displays an schematic computation of unique score for a house compared to its cohorts within the same market in the same year. All numbers are provided for illustrative purpose only.

Figure 3: Effect of Uniqueness on House Price Indexes



Notes: This figure displays repeat sale HPI estimations without and with controlling for uniqueness. The dark line traces the HPI that does not control for uniqueness. The gray line traces the HPI that controls for uniqueness. The 95% confidence interval is displayed with bars to indicate +/- 1.96 times the standard deviation.

Figure 4: House Price Indexes Comparison



Notes: This figure displays the effect of controlling for uniqueness on both the hedonic index and the repeat sale index. The left hand side Panel displays the indexes that do not control for uniqueness. The left hand side Panel displays indexes that control for uniqueness. The 95% confidence interval is displayed with bars to indicate +/- 1.96 times the standard deviation.

Table 1: Property Descriptions by Pairwise Distances

Subject Description	Comparable Description	Distance(Subj. vs. Comp.)
all new systems–easy loan qual, no repairs necessary! motivated seller–bring all offers!! welcoming 2 story covered dual porch. shows like a model! b'ful hrdwds extra trim on main. Elegant din rm. french doors to office bonus rm! open kit w granite stainless! lrg mstr w frpl priv deck! fireside liv rm!	all new systems–easy loan qual, no repairs necessary! motivated seller–bring all offers!! welcoming 2 story covered dual porch. shows like a model! b'ful hrdwds extra trim on main. Elegant din rm. french doors to office bonus rm! open kit w granite stainless! lrg mstr w frpl priv deck! fireside liv rm!	0
	newer construction craftsman style charmer in awesome grant park location! absolutely adorable curb appeal! *attractive dbl frt porch beautiful hdwds desirable flr plan huge, light filled fam rm w cozy frpl elegant din rm w detailed moldings gorgeous kit w island bfast rm convenient powder rm on main luxurious mstr ste w frpl balcony access incredible 3rd lvl ideal for rec rm or teen ste det 2–car gar walk to grant pk, zoo atl, turner field stanton elem! purchase for as little as 5% down–apprvd for homepath mortgage renovation financing!	0.412
	grant park's in-town living! this beautiful 2story hm has a back deck off great rm, sep dining, hrdwds, mud rm, lots of windows, granit kit, oversized cabinets, fp wood or gas, great yard, master w huge.	0.597
	foreclosure – two story brick front on bsmnt, prvt fenced yard, hrwds, sep liv din rm, frplc in fam rm and eat-in kitchen. easy showings....	0.639

Notes: This table compares the pairwise semantic distances between the description of a subject houses with that of a few comparables in a neighborhood called the Grant Park subdivision. The distance 0 in the first row implies the subject description is being compared to itself. The pairwise distance between two descriptions increases as their semantic meanings deviate from each other. Notice there are many abbreviations and typos in the descriptions. For instance, "granite" vs. "granit," "bful" vs. "beautiful," "hrwds vs. "hard–wood-floors," etc. The relationship between the paragraphs cannot be properly analyzed via a simple method based on key words or word frequencies.

Table 2: Descriptive Statistics

Panel A: The Real Estate Sale Sample

	Full Sample		**Multiple Transaction**	
VARIABLES	Mean	Std. Dev.	Mean	Std. Dev.
Listing Price ($ 1,000)	390.0	451.9	376.7	379.9
Ln(Sale Price)	12.26	1.234	12.32	1.120
Sale Price ($ 1,000)	373.2	411.6	363.5	361.6
Age (years)	46.20	31.02	51.08	30.76
Fireplace	1.043	1.039	0.949	0.999
Sqft (hundred)	22.93	13.57	21.10	11.70
Bath	2.700	1.285	2.535	1.149
Bed	3.606	1.034	3.461	0.974
Listing Year	2,013	2.289	2,014	2.327
Sold Year	2,014	2.208	2,014	2.237
Ranch	0.440	0.496	0.499	0.500
Pool	0.0491	0.216	0.0393	0.194
Renovated	0.0708	0.257	0.106	0.308
Large Lot (> 1 acre)	0.0247	0.155	0.0155	0.124
Feature	0.0134	0.115	0.0118	0.108
Sold-As-Is	0.118	0.323	0.112	0.315
Auction	0.0249	0.156	0.0246	0.155
Owner Agent	0.0285	0.166	0.0389	0.193
Dual Agent	0.0571	0.232	0.0519	0.222

Panel B: Real Estate Description Sample

	(1)	(2)	(3)	(4)
VARIABLES	Mean	Min	Max	Std. Dev.
Full Sample				
Word	80.27	10	164	31.50
Sentence	5.900	1	23	2.778
MLS Area-Year Unique Score (Orig)	0.777	0.384	0.940	0.0547
MLS Area Competitor Unique Score (Orig)	0.776	0.353	1	0.0555
MLS Area Non-Competitor Unique Score (Orig)	0.788	0.610	0.941	0.0449
Multiple Transaction				
Word	82.47	10	155	30.88
Sentence	6.037	1	21	2.760
MLS Area-Year Unique Score (Orig)	0.782	0.492	0.930	0.0489
MLS Area Competitor Unique Score (Orig)	0.782	0.375	0.932	0.0488
MLS Area Non-Competitor Unique Score (Orig)	0.792	0.616	0.929	0.0429

Note: Panel A presents the mean and standard deviation of the variables available in the assessor's database. The full sample contains 40,851 single-family home sales in the city of Atlanta from January 2010 to December 2017. This sample includes 37,076 houses and 3,532 houses were sold more than once. Panel B presents the descriptive statistics for the attributes of the property descriptions.

Table 3: Compare Covariates with and without Unique Measures

Dependent Variable: Ln(price)	Full Sample		Repeat Sale	
	(1)	(2)	(3)	(4)
MLS Area-Year Unique Score (standard)		0.149***		0.099***
		(0.034)		(0.020)
Age	-0.013***	-0.012***	-0.014***	-0.014***
	(0.001)	(0.001)	(0.002)	(0.002)
AgexAge	0.0001***	0.0001***	0.0001***	0.0001***
	(0.0000)	(0.0000)	(0.0000)	(0.0000)
Bed	-0.047**	-0.042**	-0.014	-0.013
	(0.021)	(0.020)	(0.021)	(0.021)
Bath	0.247***	0.245***	0.320***	0.315***
	(0.018)	(0.016)	(0.035)	(0.033)
Ranch	-0.260***	-0.244***	-0.036	-0.038
	(0.021)	(0.020)	(0.042)	(0.041)
Renovated	0.253***	0.225***	0.359***	0.345***
	(0.032)	(0.027)	(0.042)	(0.038)
Sqft	0.009***	0.009***	-0.001	-0.001
	(0.001)	(0.001)	(0.001)	(0.001)
Large Lot	0.012	0.004	0.066	0.077
	(0.028)	(0.025)	(0.070)	(0.073)
Pool	0.049***	0.051***	0.026	0.010
	(0.018)	(0.016)	(0.056)	(0.057)
Feature	0.126***	0.090***	0.023	-0.004
	(0.031)	(0.030)	(0.083)	(0.078)
Constant	10.338***	8.610***	12.246***	8.696***
	(0.081)	(0.359)	(0.265)	(0.521)
Observations	40,851	40,851	7,317	7,317
R-squared	0.760	0.769	0.956	0.958
House Characteristics	Yes	Yes	Yes	Yes
LocationxYear FE	Yes	Yes	Yes	Yes
Property FE	No	No	Yes	Yes
Cluster	Area-Year	Area-Year	Area-Year	Area-Year

Note: The results displayed in columns (1) and (2) are using the full sample, and results displayed in columns (3) and (4) are based on the repeat sale sample. Columns (2) and (4) include the control for property uniqueness. The unique score coefficients are standardized to show the effect of the increase in one standard deviation. All models include MLS Area by Year fixed effects. Robust standard errors are clustered at the MLS Area by Year level, shown in parentheses (*** p<0.01, ** p<0.05, * p<0.1).

Table 4: House Uniqueness vs. Market Power

Dependent Variables: Ln(Price)	Full Sample				Repeat Sale			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
MLS Area-Year Unique Score (standard)	0.149***				0.099***			
	(0.034)				(0.020)			
MLS Area Competitor Unique Score (standard)		0.125***		0.069		0.090***		0.084*
		(0.032)		(0.071)		(0.020)		(0.047)
MLS Area Non-Competitor Unique Score (standard)			0.112***	0.059			0.068***	0.006
			(0.028)	(0.061)			(0.021)	(0.048)
Observations	40,851	40,851	40,851	40,851	7,317	7,317	7,317	7,317
R-squared	0.769	0.767	0.767	0.767	0.958	0.958	0.957	0.957
House Characteristics	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
LocationxYear FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Property FE	No	No	No	No	Yes	Yes	Yes	Yes
Cluster	Area-Year	Area-Year	Area-Year	Area-Year	Area-Year	Area-Year	Area-Year	Area-Year

Note: The dependent variable this table is Ln(sale price). Columns (1) and (5) repeat the estimates from the previous table for the hedonic and repeat sales models respectively. Columns (2) and (6) present estimates using the new MLS area measure of uniqueness based on competing units, and Columns (3) and (7) present estimates using uniqueness based on non-competing units. Finally, in Columns (4) and (8), we run a horse race between the uniqueness measure based on competing units and based on non-competing units. All of the unique score coefficients are standardized to show the effect of the increase in one standard deviation. All models include MLS area by year fixed effects. Robust standard errors are clustered at the MLS Area by Year level, shown in parentheses (*** p<0.01, ** p<0.05, * p<0.1).

Table 5: Are the Uniqueness Effects Sensitive to Alternative Models?

Panel A: Full Sample Analysis

Dependent Variable:	Ln(Price)				Price Spread
	(1)	(2)	(3)	(4)	(5)
MLS Area-Year Unique Score (standard)	0.149***	0.149***	0.136***	0.122***	-0.761
	(0.034)	(0.032)	(0.026)	(0.019)	(0.616)
Observations	40,851	40,851	40,851	40,851	40,851
R-squared	0.769	0.784	0.811	0.886	0.238
House Characteristics	Yes	Yes	Yes	Yes	Yes
Keywords	No	No	Yes	Yes	No
Transaction Characteristics	No	Yes	Yes	Yes	No
LocationxYear FE	Yes	Yes	Yes	Yes	Yes
Agent FE	No	No	No	Yes	No
Property FE	No	No	No	No	No
Cluster	Area-Year	Area-Year	Area-Year	Area-Year	Area-Year

Panel B: Repeat Sale Analysis

Dependent Variable:	Ln(Price)				Price Spread
	(1)	(2)	(3)	(4)	(5)
MLS Area-Year Unique Score (standard)	0.099***	0.098***	0.080***	0.050***	-0.297
	(0.020)	(0.020)	(0.015)	(0.018)	(0.592)
Observations	7,317	7,317	7,317	4,117	7,317
R-squared	0.958	0.961	0.968	0.989	0.684
House Characteristics	Yes	Yes	Yes	Yes	Yes
Keywords	No	No	Yes	Yes	No
Transaction Characteristics	No	Yes	Yes	Yes	No
LocationxYear FE	Yes	Yes	Yes	Yes	Yes
Agent FE	No	No	No	Yes	No
Property FE	Yes	Yes	Yes	Yes	Yes
Cluster	Area-Year	Area-Year	Area-Year	Area-Year	Area-Year

Note: Panel A shows hedonic estimates using the full sample, and Penal B displays repeat sale results based on the repeat sales sample. The dependent variable in all specifications in Column (1)–(4) is Ln(sale price), and the dependent variable in Column (5) is Listing-Sale Spread in thousand dollars. Column (1) presents the standard hedonic model with just the uniqueness control added, Column (2) presents estimates for a model that also adds transaction characteristics, Column (3) also adds the dummy variables for the presence of specific keywords, and Column (4) adds agent fixed effects. Column (5) presents the model on the listing-sales price spread. The unique score coefficients are standardized to show the effect of the increase in one standard deviation. Robust standard errors are clustered at the Area-Year level, shown in parentheses (*** $p<0.01$, ** $p<0.05$, * $p<0.1$).

Table 6: Summary Statistics of Hedonic Controls by Year

Year:	2010		2011		2012		2013	
VARIABLES	Mean	Std. Dev.	Mean	Std. Dev.	Mean	Std. Dev.	Mean	Std. Dev.
MLS Area-Year Unique Score	0.75	0.07	0.78	0.06	0.78	0.05	0.78	0.05
MLS Area-Year Unique Score (Mean Diff)	0.0013	0.0269	0.0034	0.0262	-0.0003	0.0262	-0.0011	0.0242
Ln(Sale Price)	11.84	1.39	11.85	1.44	12.01	1.32	12.34	1.17
Age	42.62	30.32	42.85	30.20	45.03	30.09	45.33	30.02
Fireplace (d)	0.99	1.05	1.03	1.06	1.05	1.05	1.07	0.99
Pool (d)	0.05	0.21	0.06	0.23	0.05	0.22	0.05	0.23
Bed	3.61	1.02	3.62	1.04	3.61	1.03	3.63	1.06
Ranch (d)	0.47	0.50	0.44	0.50	0.43	0.50	0.42	0.49
Bath	2.64	1.30	2.70	1.33	2.72	1.29	2.75	1.26
Renovated (d)	0.00	0.01	0.03	0.18	0.06	0.24	0.08	0.27
Sqft (100)	20.95	11.61	22.56	12.99	22.95	13.71	23.44	13.58

Year:	2014		2015		2016	
VARIABLES	Mean	Std. Dev.	Mean	Std. Dev.	Mean	Std. Dev.
MLS Area-Year Unique Score	0.79	0.04	0.78	0.05	0.78	0.06
MLS Area-Year Unique Score (Mean Diff)	-0.0021	0.0254	-0.0016	0.0267	-0.0008	0.0272
Ln(Sale Price)	12.45	1.12	12.47	1.10	12.49	1.05
Age	46.85	31.35	47.60	32.02	48.53	31.94
Fireplace (d)	1.10	1.06	1.08	1.06	1.03	1.06
Pool (d)	0.05	0.22	0.05	0.22	0.05	0.21
Bed	3.61	1.05	3.59	1.02	3.60	1.02
Ranch (d)	0.44	0.50	0.42	0.49	0.44	0.50
Bath	2.73	1.29	2.71	1.28	2.68	1.26
Renovated (d)	0.08	0.27	0.09	0.28	0.10	0.30
Sqft (100)	23.75	14.51	23.53	13.72	23.06	13.86

Notes: This table displays the summary statistics of the control variables by transaction year. The means for each variable are shown in separate rows, and means and standard deviation for each year are shown in separate columns.

Table 7: Coefficient of Year FE

Dependent Variable: Ln (Price)						
Model:	Hedonic Index			Repeat Sale Index		
	Model 1 (1)	Model 2 (2)	Diff. M1- M2 (3)	Model 1 (4)	Model 2 (5)	Diff. M1- M2 (6)
2011 (d)	-0.057*	-0.177***	0.120	-0.012	-0.068**	0.056
	(0.030)	(0.026)		(0.033)	(0.028)	
P(Chi2)			0.000			0.009
2012 (d)	0.068	-0.060**	0.008	0.063*	0.020	0.043
	(0.044)	(0.025)		(0.035)	(0.033)	
P(Chi2)			0.004			0.020
2013 (d)	0.362***	0.220***	0.142	0.287***	0.250***	0.037
	(0.054)	(0.029)		(0.047)	(0.039)	
P(Chi2)			0.003			0.057
2014 (d)	0.463***	0.304***	0.159	0.360***	0.323***	0.037
	(0.066)	(0.034)		(0.046)	(0.037)	
P(Chi2)			0.006			0.066
2015 (d)	0.494***	0.352***	0.142	0.484***	0.453***	0.031
	(0.080)	(0.057)		(0.049)	(0.048)	
P(Chi2)			0.019			0.130
2016 (d)	0.553***	0.444***	0.109	0.604***	0.575***	0.029
	(0.091)	(0.065)		(0.072)	(0.063)	
P(Chi2)			0.115			0.206
Control Unique	No	Yes		No	Yes	
R-squared	0.509	0.571		0.947	0.949	

Notes: This table displays the year indicator coefficient estimates from the indexing models first for the hedonic models and then for the repeat sale models. The estimates for each year are presented in separate rows, and each specification is shown in a set of three columns: the first presenting estimates for a model that does not control for the uniqueness, the second from a model that includes uniqueness, and the third column presenting the differences between these two model estimates and the corresponding P–values. Robust standard errors are clustered at the Zip Code level, shown in parentheses (*** p<0.01, ** p<0.05, * p<0.1).