University of Connecticut

Department of Economics Working Paper Series

Housing Prices and Property Descriptions: Using Soft Information to Value Real Assets

by

Lily Shen
Clemson University

Stephen Ross
University of Connecticut

Working Paper 2019-20
December 2019

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

This working paper is indexed in RePEc, http://repec.org

Housing Prices and Property Descriptions: Using Soft Information to Value Real Assets *

Lily Shen
Clemson University

Stephen Ross
University of Connecticut

December 20, 2019

Abstract

Recent research in economics and finance has recognized the potential of utilizing textual "soft" data for valuing heterogeneous assets. 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 unobserved quality based on our "soft" information leads to a 15% increase in property sale price. Further, annual hedonic house price indices ignoring our measure of unobserved quality overstate real estate prices by 11% to 16% and mistime the recovery of housing prices following the Great Recession.

Key Words: *Natural Language Processing, Unsupervised Machine Learning, Soft Information, Housing Prices, Price Indices, 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, Stuart Rosenthal, and Vincent Intintoli. All errors are our own.

1 Introduction

Several recent studies have recognized the potential of soft information, often inferred from textual data, in shedding light on the value of heterogeneous assets. Tetlock (2007), Garcia (2013), Loughran and McDonald (2011), Huang et al. (2019) find words that convey positive/negative meanings used in articles from the popular press and in 10-K reports can explain positive/negative stock returns. Similarly, Goetzmann et al. (2013) uses the text of screenplay sales pitches to predict prices, and Aubry et al. (2019) uses information on the artist, the auction and the artwork including machine processed visual information to predict the results of art auctions.

Real estate is a large and essential component of the U.S economy,[1] and valuing real estate assets has always been problematic 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 as "charming" may represent an unreported "soft" feature that captures information about the unit's value (Liberti and Petersen, 2018). Up to this point, attempts to use "soft" information 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), Nowak and Smith (2017), 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 the only exception in real estate that we know of to the keyword approach for gathering "soft" information, Lindenthal (2017) uses machine learning to compare architectural design similarity using photos from google street view.

According to a 2018 Wall Street Journal article, institutional investors who buy and sell

[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

hundreds of houses on a daily basis utilize newly developed artificial intelligence technologies to extract "soft" information from unstructured real estate data, such as the textual property descriptions.[2] To our knowledge, our study is the first paper to extend traditional keyword-based approaches by applying machine learning techniques from natural language processing to characterize the content of real estate property descriptions. Property descriptions are the written portion of a real estate advertisement that summarizes critical features of the underlying house. Specifically, we train our machine learning algorithm to quantify the semantic meaning of real estate property descriptions, and use these quantitative measures for each description to assess the uniqueness of each house relative to its neighbors. Our use of property description uniqueness is similar to Ahlfeldt and Holman (2018)'s characterization of neighborhood architectural quality by surveying residents on the distinctiveness of the neighborhood.[3]

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. Second, in differentiated product markets, products that are unique in product attribute space have greater market power and so tend to sell for more (see for example Bayer et al., 2007; Berry et al., 1995).[4] Finally, in principle, some agents may be able to craft advertisements that are so unique and appealing that they influence the sale prices beyond the actual desirable attributes associated with the housing unit itself.

We combine this novel application of machine learning methods with hedonic pricing (hereafter referred to as the "ML–Hedonic approach") to estimate the price impact of hous-

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

[3]Ahlfeldt and Holman (2018) argue that asking about distinctiveness rather than quality or beauty minimizes the influence of normative judgements and personal tastes of respondents.

[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 indices for the effect of asset liquidity at the time of sale.

ing unit uniqueness, a type of "soft" information, and to examine the impact of ignoring uniqueness when estimating hedonic based housing price indices. Following Rosen (1974), a massive literature has 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).[5] Further, this hedonic framework has been used extensively throughout the world to measure asset prices over time using price indices.[6] However, houses are heterogeneous goods for which some characteristics (the "soft" information) cannot be easily captured by numerical data, and the accuracy of the widely used hedonic framework may suffer from the omission of these unobserved attributes (Liberti and Petersen, 2018; Garmaise and Moskowitz, 2004) .

It is important to emphasize that the machine learning method used in this study is unsupervised, which is very different from the supervised algorithms discussed in Mullainathan and Spiess (2017). Supervised machine learning methods are often used to generate out–of–sample predictions based on a large sample of training data with observed outcomes, such as housing prices in the case of Mullainathan and Spiess (2017). The goal of supervised learning is to produce an inferred function to map a large number of input variables to predictive output values, often referred to in machine learning as labeled data. However, when the input information is not easily organized into quantitative variables (unlabeled data), such as in natural language processing problems, machine learning analysts often turn to unsupervised learning approaches. Unlike supervised learning, which is trained to perform

[5]Examples of numerical data include, but are not limited to, asking price, sale price, size, age, property type, structural attributes, location, and market condition

[6]The best known price indices 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) indices all use the repeatsales method. However, the OneFamily Houses Price Index of the Census Bureau, MultiFamily 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 Verbund 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).

the task of interest, unsupervised learning is trained on a related task and then performs the desired task as one step in that process. Generally, the goal is to quantify qualitative features of some information set, e.g. visual, textual or contextual information, by using those quantitative features to explain patterns observed within the information.

Specifically, in our case, the advertisement copy or property description of the housing unit is described by numeric values within a high dimensional space intended to capture abstract attributes or features of the paragraph that best predict the occurrence of words in certain contexts within the text. Although specific meanings are not assigned to each dimension, the relative distance between two documents in feature space indicates the relative semantic distance between the corresponding descriptions. Therefore, uniqueness is measured by comparing the distance between a housing unit's description and the other descriptions for units in the same geography and at the same time. In this sense, our paper is similar to Aubry et al. (2019) who use machine learning techniques to characterize the attributes of artwork and then predict the value of art sales based on that characterization. To our knowledge, we are the first to measure real estate uniqueness using textual data. However, Haurin (1988) models real estate "atypicality" using the observable, quantitative house features and sale price.

Using a data set that encompass 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 in the Multiple Listings Service (MLS) are associated with higher sale prices than those units with less unique descriptions. Comparisons among houses located in geographical proximity, the MLS market area, and advertised in the same year show that a one standard deviation increase in description uniqueness leads to a 15% increase in property sale price. The information provided by the measure of uniqueness is at most weakly correlated with traditional hedonic attributes like square feet of living space or 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 whether a unit was renovated and an indicator for whether the house has special features. Further, our measure of uniqueness captures information that is independent of the traditional controls for key words, and so including key words only reduces the impact of uniqueness on prices from 15% to 14%. On the other hand, a one standard deviation increase in uniqueness is only associated with a four-day delay in days on the market.

We also estimate several additional models to show that the measure of uniqueness is primarily picking up information on the quality of the housing unit, rather than a price premium arising from market power in a differentiated product market. First, we use uniqueness among all housing units in our sample for the same MLS area, rather than just housing units advertised in the same year. This alternative measure of uniqueness should contain less information on market power because it is not based on housing units that were directly competing against this unit at the time of sale. The influence of uniqueness is still quite large, only falling from 15% to 13%. Next, we estimate repeat sales models where the inclusion of a housing unit fixed effect would be expected to eliminate many of the unobservables that uniqueness captures, and as expected, the effect of uniqueness declines substantially from 15% to 7.5%. Again, shifting to uniqueness relative to MLS area regardless of year on market only modestly lowers repeat sales estimates on uniqueness from 7.5% to 5.5% suggesting that most of the effect of uniqueness in the repeat sales model arises from changes in housing quality over time, changes that would not be captured by repeat sales price indices. The effect of uniqueness on days on the market is somewhat less stable to basing uniqueness on all units advertised in the MLS area with the estimate falling from 4.3 to 2.7 days, consistent with direct competitors having a larger impact on days on the market. Finally, we examine the effect of adding controls for agent fixed effects, which might represent factors associated with the agent written property description rather than the housing unit, and these controls reduce the influence of uniqueness by less that 2 percentage points.

Finally, we use a similar hedonic model to estimate hedonic price indices using the stan-

dard time dummy approach where the price level is captured by dummy variables associated with each time period (Silver (2016); Hill (2013)). We find changes in our transaction sample over time in terms of our uniqueness measure, almost 1/2 a standard deviation increase in uniqueness. This compositional change arises relatively early in the recovery from the great recession, starting in 2011 and prior to the recovery in housing prices. As a result, the change in the composition of the housing stock biases the estimated price index post-2010 upwards by between 11-16% and leads to a mistiming of the recovery in housing prices when using hedonic models that do not control for the uniqueness of the properties on the market at that time. Looking at individual counties, these differences are primarily located within Fulton County, which contains the city of Atlanta and a majority of the housing stock in the metropolitan region. 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 indices.

More broadly, this study provides several theoretical and empirical insights in response to the newly available machine learning tools for research in economics and finance. First, our algorithm defines the meanings of words within their contexts, addressing criticism for the keyword-based studies raised by Larcker and Zakolyukina (2012): "simply counting words (bag–of–words) ignores important context and background knowledge." Second, the ML–Hedonic approach used in this study provides an example of the integration of unsupervised learning methods into quantitative empirical analyses, similar to Aubry et al. (2019), and demonstrates that the information generated by these learning methods has substantial explanatory power for real estate asset prices. In summary, our context-based ML algorithm can be applied to extract information from a wide range of textual documents to study a wide variety of economic and financial phenomena.

2 Methodology

The ML–Hedonic approach follows three steps. First, we train our Machine Learning semantic analysis algorithm to understand the semantic meaning of real estate descriptions. Each description is represented as a numerical vector in a high–dimensional vector space based on its contents and semantic meaning. The distance between two vectors represents the pairwise difference between two houses. 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 linear hedonic models, and then use similar models to estimate hedonic price indices. We introduce the ML model in Sections 2.1 and 2.2, and in Section 2.3 we describe our hedonic 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. Applications of NLP include, for example, language translation, speech recognition, automatic summarization, and language understanding and quantification, whereas our study focuses on this last task.

In this study, we implement the paragraph vector (PV) method introduced in Le and Mikolov (2014), a Neural Network approach to obtain vector representations of text, in our case, real estate property descriptions. This method is an extension of the Continuous Bag of Words (CBOW) method described in Mikolov et al. (2013). CBOW preserves word context by characterizing words based on the words that they are nearby within the text, and PV also characterizes paragraphs exploiting the fact that the patterns arising in the proximity of different words can systematically vary across paragraphs. Dai et al. (2015) compared the PV method against other textual analysis algorithms, including the widely used simple Bag-of-Words method, on the analysis of 4,490,000 Wikipedia articles and 886,000 technical

research papers, and concluded the PV method strictly outperformed the other methods.

Essentially, the paragraph vector is a generalization of approaches that infer a word's meaning by its context by identifying words that are usually found near that word. The following example demonstrates how the algorithm defines the meaning of "Southern" by its contexts in the house descriptions:

- This home is a graceful Southern beauty with rare stately double–front porches.
- Southern elegance in the Georgian style renovated for today.
- Graceful Southern charm!
- Exquisite Southern living, backyard w/stunning granite pool.

Semantically and syntactically, "Southern" is related to "elegance", "exquisite", "beauty", "graceful" etc. In this case, NLP algorithms tend to assign similar features to words like "elegance", "exquisite", "beauty", and "graceful" because all four of these words tend to predict the presence of the same nearby words, like "southern". Paragragh vector simultaneously estimates both word features and paragraph features where the estimated paragraph features capture the information arising because some property descriptions tend to contain different combinations of words than other descriptions.

This approach offers the following advantages in analyzing property descriptions data:

First, our algorithm is more suitable for detecting nuances of human language compared to sentiment analysis methods based on word polarity. False-positive words are often used to glorify negative features of houses in the descriptions. For example, "good," "potential," "cozy," "cute," and "original" are all positive words in daily uses. However, in real estate descriptions "good potential" is often used to describe houses that require extensive renovation, "cute" and "cozy" are used to describe small houses, and "original" is used to describe old houses, which might explain why the literature has had mixed results when including advertisement key words into hedonic models.

Second, our algorithm is more suitable for understanding abbreviations and typos, compared to sentiment analysis methods based on counting word frequency. Unlike formal documents such as 10-K reports and newspaper articles that are carefully polished before being released to the public, typos are often found in property descriptions (e.g. "morgage" vs. "mortgage"). In addition, the MLS system imposes a 250-word limit on description length, and thus the full spelling of a word might be replaced with an unstandardized abbreviation to save space. For example, "tender love and care" is a common expression in descriptions to describe old houses that need renovation. Depending on space availability, it can be written as "tender loving care," "tender love care" or "TLC". Unstandardized abbreviations and typos would have been dropped by previous algorithms based on counting word frequency. Since our algorithm defines the meaning of textual data within their contexts, it is able to understand that all four expressions have the same meaning.

We train the Neural Network model iteratively to get a vector representation of each description. To generalize the idea, we define w_{ij}^{out} as the ith output word (target) randomly selected from paragraph j, and $\underline{w}_{ij}^{in}$ as a vector of input words from its context. We source the context words within a distance of L from the target word w_{ij}^{out}. The distance of L can also be regarded as the size of a sliding window, which defines the extent of the context that we would like to include in the word vector analysis.[7] For an arbitrary element l in the $\underline{w}_{ij}^{in}$, we use w_{ijl}^{in} to denote the lth specific element in the vector between 1 and L. Our goal will be to simultaneously estimate a numeric paragraph feature vector (v_j^p) in an A–dimensional paragraph attribute space for each paragraph of advertisement text and a numeric word feature vector (v_l^w) in a second A–dimensional space (word attribute space) for each word occurring within our population of advertisements to best predict the identities of each randomly selected target word given the word's paragraph and list of input words.[8]

[7]The window size or bandwidth is sometimes selected by cross-validation. However, in our example, we verify that the results are very robust to bandwidth choice, and we find nearly identical estimates for bandwidths between 10 and 50 words.

[8]Note that the dimensionality of the paragraph and word feature spaces need not be equal, but in this example they have both been set to A.

First, we tokenize the entire pool of house descriptions into a vocabulary list of length N. Each unique word (also called a token) in the list will have a unique vector (also called a one-hot vector) to represent its position k in the list (i.e., the one-hot vector is of length N and contains a list of many zero's and only one non-zero value "1" at the position of the word).

Second, for each advertisement/paragraph j, we randomly draw a sample of target words i. Each observation for estimation will involve one target word, L input words that fall within the established window and the identity of the paragraph j from which the word was selected. Therefore, each observation is quantitatively characterized by the location k of the word w_{ij}^{out} within the tokenized list of words, L word feature vectors (v_l^w) where l indexes the elements of the vector of input words $\underline{w}_{ij}^{in}$, and one paragraph feature vector (v_j^p).

Third, we specify two additional sets of vectors of parameters to be estimated. These vectors represent weights that map between attributes/features of either paragraphs or input words and an index that will be used to predict the likelihood of observing an output word in a given location, i.e. the paragraph and surrounding input words. Specifically, we define N word weight vectors and N paragraph weight vectors, one for each word in the tokenized word list. Each vector is of length A representing the number of dimensions in both the paragraph and word feature spaces. The entire collection of weight vectors can be regarded as a transformation matrix between what is typically referred to as the input layer containing the word and paragraph feature vectors and the hidden layer that lies between those vectors and the predicted probability of words arising in a given location.[9] These weight vectors define how important each feature is for predicting the presence of a specific target word. We use μ_k^w and μ_k^p to represent the weight vectors for words and paragraphs, respectively, for word k.

Now, we can define the index that describes the likelihood that an output word is the kth word in the population given the set of input words for the ith randomly selected word

[9]In NLP applications, the weight vectors are usually combined into matrices and the one-hot vector is used to select the appropriate column for word k.

from the jth paragraph.[10]

$$x_{ijk} = \mu_k^{p\prime} v_j^p + \sum_{l=1}^{L} \mu_k^{w\prime} v_{w_{ijl}^{in}}^w \tag{1}$$

Correspondingly, the correct prediction for the actual output word w_{ij}^{out} should be:

$$x_{ijw_{ij}^{out}} = \mu_{w_{ij}^{out}}^{p\prime} v_j^p + \sum_{l=1}^{L} \mu_{w_{ij}^{out}}^{w\prime} v_{w_{ijl}^{in}}^w \tag{2}$$

Figure 1 illustrates this process using a window that is four words wide with the bottom row representing the paragraph feature vector and four input word feature vectors associated with a specific output or target word. These vectors are multiplied by the appropriate paragraph and word weight vectors illustrated by the middle row of the figure and summed to create an index. The final/top step applies a transformation to calculate the probability of observing the output word given the observed index.

If we assume a standard multinomial logit maximization framework so that the true word arose in that location because

$$x_{ijw_{ij}^{out}} + \epsilon_{ijw_{ij}^{out}} = Max_k[x_{ijk} + \epsilon_{ijk}] \tag{3}$$

where ϵ_{ijk} follows an extreme value distribution, then the conditional probability of the target word occurring can be written as:

$$Pr\left[w_{ij}^{out} | \underline{w}_{ij}^{in}\right] = \frac{e^{x_{ijw_{ij}^{out}}}}{\sum_{k=1}^{K} e^{x_{ijk}}} \tag{4}$$

with probabilities that are initialized to sum to one for any word i from paragraph j over all possible words k in the tokenized list. This multinomial probability calcuation is typically referred to as the log–linear Softmax function in Artificial Neural Network applications.

Finally, assuming that I words are drawn randomly from J paragraphs, the log likelihood

[10]Where $'$ represents the transpose of the vector.

problem can be written as

$$\underset{\mu_k^p, \mu_k^w, v_j^p, v_k^w}{Min} \sum_{j=1}^{J} \sum_{i=1}^{I} -log\left(Pr\left[w_{ij}^{out} | \underline{w}_{ij}^{in}\right]\right) \tag{5}$$

Iteratively, we maximize the probability of getting the correct outputs through the fine tuning of θ=$\{$ μ_k^p, μ_k^w, v_j^p, v_k^w $\}$ by minimizing the log-likelihood function ϵ below:

$$\begin{aligned} \epsilon &= \log \sum\nolimits_k \exp\left(x_{ijk}\right) - x_{ijw_{ij}^{out}} \\ &= \log \sum\nolimits_k \exp\left(\mu_k^{p\prime} v_j^p + \sum_{l=1}^{L} \mu_k^{w\prime} v_{w_{ijl}^{in}}^w\right) - \left(\mu_{w_{ij}^{out}}^{p\prime} v_j^p + \sum_{l=1}^{L} \mu_{w_{ij}^{out}}^{w\prime} v_{w_{ijl}^{in}}^w\right) \end{aligned} \tag{6}$$

The resulting optimization problem is extremely high dimensional, and the parameters map into the indices x_{ijk} following a non-linear, interactive process. Therefore, traditional optimization approaches are not feasible, and we follow Mikolov et al. (2013) and Le and Mikolov (2014) using stochastic gradient descent and backpropagation to minimize the log-likelihood function. Weighting vectors are initialized prior to optimization with random numbers.

The parameter estimates of the paragraph attribute vectors are the final output of this step in the estimation process, and these estimated attributes will be used in the next section to measure the distance between housing units based on the advertisement text.[11]

2.2 Construction of the Uniqueness Measure

From the previous subsection, we obtained vector representations to quantify the information contents of house descriptions. We define the pairwise distance between two vectors to represent the relative semantic distance between the corresponding property descriptions. This distance is measured using the angle between a pair of vectors obtained during the

[11]The parameters in the weight matrix can be viewed as incidental. In fact, we empirically verify that the mean and standard deviation of paragraph weights for each feature across all target words are the very similar. Therefore, the weights do not have any meaningful impact on the relative important of each feature in predicting target words.

vectorization process (v_j^p in Equation 1), shown in the equation below.

$$Distance(\boldsymbol{v}_1, \boldsymbol{v}_2) = 1 - cos(\boldsymbol{v}_1, \boldsymbol{v}_2) = 1 - \frac{\boldsymbol{v}_1 \cdot \boldsymbol{v}_2}{||\boldsymbol{v}_1|| \cdot ||\boldsymbol{v}_2||} \tag{7}$$

Notice that the distance defined in Equation 7 is the cosine distance between two vectors, where 0 means two identical descriptions with 0 semantic distance in between. This measure is mathematically bounded between 0 and 1.

Figure 3 provides a visual demonstration of the effectiveness of our ML algorithm. In the top text box, the query is "Lenox Mall", a shopping center in northern Atlanta. The blue pins on the map are houses related to the query. 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 ML algorithm via Equation 7. This figure shows our algorithm can successfully sort houses based on descriptions similarity/difference. In this particular example, all the similar houses are near the Lenox mall although the name of the mall does not directly show up in some of the descriptions.[12]

Table 1 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 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–wood–floors," etc. The relationship between the paragraphs cannot be properly analyzed via a simple method based on keywords or word frequencies.

To assess the uniqueness of a description compared to its cohorts (within the same market area and year in this paper), we compute the average pairwise distances from the house of

[12]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.

interest to other houses, as shown in Equation 8 and Figure 5.

$$Unique_i = \frac{\sum_1^{N-1} (pairwise\, distance)}{(N-1)\ pair\, of\, houses} \tag{8}$$

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

2.3 Hedonic Pricing Model

We employ a classic log-linear hedonic model to estimate the impact of unique property descriptions on home prices. The full empirical specification takes the following form:

$$Ln(Price_i) = \alpha + \theta Unique_i + X_i'\beta + \mu_{ct} + \eta_z + \varepsilon, \tag{9}$$

$Ln(Price_i)$ is the natural log of the sold price of house i. $Unique_i$ is the description uniqueness score derived from our machine learning semantic analysis 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), 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. Standard errors are clustered at the MLS market area by year level.

For our hedonic price indices, we estimate a slightly different model. We follow the

time dummy approach to estimating time fixed effects within the hedonic price indices. However, we exclude controls for geography within our metropolitan area because unlike hedonic regressions used for inference purposes, most price indices are based on hedonic regressions that do not include sub-geography fixed effects.

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

The vector X_i 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 indices. 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 at a geography by year level as in the fixed effect models above.

Finally, the price index for a given year t can be calculated relative to a base year 0 as:

$$I_t = \frac{exp(\delta_t)}{exp(\delta_0)} \tag{11}$$

3 Data and Descriptive Statistics

The data we use encompass more than 40,000 single-family home sales in the metropolitan real estate market of Atlanta, GA, from 2010 to 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, a wide range of house characteristics, critical dates regarding the transaction, unique IDs of the listing and buying agent, and, most importantly, the written property description.

We impose three restrictions on the property description data: First, we only include

houses for which the property descriptions are longer than 9 characters. Second, we limit our sample to areas with more than three sales in each year during the sample period to avoid areas of very thin or inactive housing stock. Finally, we only include sold properties because descriptions of unsold houses are often deleted when a house was taken off the market. We calculate uniqueness comparing the 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 final data used in this study consist of 40,918 transactions: 37,124 unique sales and 3,794 repeat sales. We use the unique sales to deliver our baseline results and the repeat sales data for further investigation of our results. Table 2 panel A displays a set of basic descriptive statistics for the data used in this study. 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 on average three and half months later. Since we only focus on the MLS defined market area (shown in Figure 4), most of the houses sold in this area sit on small lots.[13] Only 2.5 percent of the homes in our data are built on lots that are greater than one acre.

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.

[13]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.

4 Empirical Results

In this section, we present empirical results reported by our hedonic models to explore the effects of $Unique_i$ on real estate sale prices. Table 3 displays the estimates from our hedonic price models and for equivalent models for days on the market before sale, both with and without our control for uniqueness. Columns 2 and 4 present the models that include uniqueness showing a large effect on price with a one standard deviation increase in uniqueness being associated with a 15% increase in the sale price and a modest increase in time on market of 4 days.[14] Notably, the estimates on traditional hedonic attributes like the number of bedrooms and baths or square feet are relatively stable, indicating that our measure of uniqueness is capturing information that is typically not captured by those attributes. However, not surprisingly, less well-defined attributes like whether units were renovated or have unique features, are correlated with our measure of uniqueness, and so those attributes' estimated effects on price are eroded more substantially, by 11 and 28 percent respectively, when the control for uniqueness is added. [15]

Next, in Table 4, we examine alternatives to the model to investigate whether uniqueness is truly capturing the price effect of housing unit unobservables. First, we replace our MLS market area by year measure of uniqueness with a measure based on all sales in the area during the entire sample period. If the effect on price is driven primarily by market power arising from the unit's isolated location in housing unit attribute space, then the relevant units for comparison are the competing units on sale in the same year. However, if uniqueness captures the quality of housing, then the correct comparison is to all housing units in the market area. Column 2 presents estimates using the new MLS area measure of uniqueness, and the standardized effect only falls by 2 percentage points from 15 to 13%, suggesting that most of the effect of uniqueness on prices is due to the ability of uniqueness to capture

[14]The days of the market model includes a control for the log of housing price to be conservative, but all results below are very similar for a model that omits the log of housing price.

[15]We confirm the low correlation by directly regressing uniqueness on the hedonic attributes and only the feature and renovation variables have substantial power to explain uniqueness.

unobserved quality.

In columns 3 and 4, we present estimates from a repeat sales model. This model should difference out time-invariant unobserved attributes of the housing unit. If, as we believe, uniqueness is capturing the unobserved quality, we would expect that the ability of uniqueness to explain quality should fall substantially in the repeat sales model. Comparing column 3 to column 1, the estimated effect falls by half from 15 to 7.5%. Further, as in column 2, the shift from using MLS area by year uniqueness to MLS area uniqueness again decreases the effect of uniqueness by 2 percentage points with the estimate falling between columns 3 and 4 from 7.5% to 5.5% suggesting that most of the effect of uniqueness in the repeat sales model is due to changes in the quality of the housing units between sales.

The last two columns present the estimates of the effect on days on the market for both the MLS area by year and the MLS area uniqueness measures. The effect of uniqueness on days on the market falls from 4.3 to 2.7 days when we replace the area by year uniqueness with the area uniqueness. Not surprisingly, the effect of time on the market is more sensitive to whether the housing unit is unique relative to the other housing units that were on sale at the same time.

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 1 presents estimates for the hedonic price regression, and panel 2 presents estimates for days on the market. The first column presents the baseline estimates from Table 3. Column 2 adds transaction characteristics,[16] following Levitt and Syverson (2008) and Rutherford and Yavas (2005) column 3 adds key words drawn from the real estate property description,[17] and column 4 includes agent fixed effects.

Our estimates are relatively stable, remaining at 15% when transaction attributes are

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

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

included and falling to 14% and 12.4% as first keywords and then both keywords and agent fixed effects, respectively, are added. The stability of the coefficient estimate when comparing columns 1 and 2 to column 3 with key words demonstrates that simply focusing on common and well-understood words is not sufficient to capture the information contained in real estate property descriptions. Similarly, the modest influence of including agent fixed effects suggests that agent specific language is unlikely to be playing a major role in explaining the effect of uniqueness of sales price. Further, the influence of agent fixed effects might also arise because the most successful agents tend to represent the best properties that in turn have many unique attributes, and if so the agent specific premia arising from uniqueness would also be attributable to actual unobserved attributes of the property.

Perhaps, even if uniqueness captures information that cannot be obtained through a simple keyword approach, keywords may still be valuable in understanding the role that uniqueness plays in explaining prices. We next estimate models where we allow the influence of uniqueness to depend upon the presence of positive or negative keywords. Perhaps, for example, uniqueness may not increase prices when this measure is in part capturing negative keywords. Negative keywords are substantially less common in property descriptions than positive keywords, so we interact our measure of uniqueness with dummies for whether a description has 5 or more positive keywords and for whether a description has 1 or more negative keywords. These results are presented in Table 6. We do not find any evidence that the presence of certain types of keywords explains the content of our uniqueness measure. More unique descriptions are always systematically related to more valuable housing units, regardless of the keywords used. We do observe a slight reduction in the explanatory power of uniqueness when a description contains many positive keywords, perhaps because the most desirable properties will tend to have many unique and positive features leading to many positive keywords. We find no evidence that the use of one or more negative keywords affects the information provided by our uniqueness measure.

5 Hedonic Housing Price Indices

In this section, we estimate a hedonic price model controlling for standard hedonic attributes and year fixed effects for the Atlanta metropolitan real estate market, as well as separate price indices for the two largest counties, Fulton and DeKalb, and a combined sample of the three counties that contain the smallest portion of the Atlanta real estate market: Clayton, Cobb and Gwinnet counties. We estimate these models with and without the controls for uniqueness in order to see if the hedonic price indices are distorted or biased in a substantial way by the omission of the unobserved housing attributes captured by our uniqueness measure.

As discussed above, we are concerned that the composition of housing units on the market may change across the housing market cycle. For our sample, the composition of housing units may change as the housing market recovers from the subprime crisis. Therefore, we first examine the composition of housing sales on both uniqueness and our observed hedonic attributes year by year, see Table 7. Note that as housing prices start to stabilize in Atlanta between 2010 and 2011, see row 2, the composition changes substantially. Uniqueness shown in row 1 increases by almost 1/2 of a standard deviation. Further, 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 stock on the market at the time. The increase in our uniqueness measure suggests that the housing stock is improving on unobservables as well.

Figure 6 graphs the price indices for the entire Atlanta region, while Figure 7 graphs the price indices for the county subsamples. Both Figure 6 and panel A of Figure 7 for Fulton county, which contains the city of Atlanta and the majority of the metropolitan market housing stock, indicate that prices have stabilized between 2010 and 2011 when uniqueness is not included in the hedonic model. However, after controlling 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 appeared to lead the recovery in terms of quality, with housing prices not starting to recover until a year later. We do not observe any evidence of a change in the quality of housing as measured by uniqueness for the more outlying areas of the metropolitan housing market, i.e. panels 2 and 3 of Figure 7 show nearly identical price indices whether or not the hedonic model includes a control for uniqueness.

Table 8 quantifies these changes. In 2011, the naive price index without conditioning on uniqueness shows that prices fell by 5.7% while our alternative index suggests that prices fell by 17.7%, a 12 percentage point difference. In 2012, house prices started to recover from their 2011 low, but the gap in the price indices remains at 12 percentage points. The gap between the two price indices grows to 16 percentage points by 2014 and then declines somewhat falling to a 12 percentage points gap by 2017. These differences are almost all statistically significant. Focusing on Fulton county, the differences in the price index are even larger, ranging between 17 and 26 percentage points. The differences are much smaller in DeKalb and for the three combined counties that both contain much less of the Atlanta real estate market, and typically these differences are not statistically significant. Our estimates suggest substantial bias in the hedonic price indices for the central Atlanta housing market primarily because the quality of the housing units on the market improved a year prior to the stabilization of housing prices. Ignoring this important information on housing uniqueness or quality leads to an overstatement of housing price appreciation during the recovery and also to a mistiming of the year in which housing prices began to recover. Further, the change in the composition of housing units on the market appears to act as a leading indicator for the housing recovery, at least for this particular real estate market.

6 Conclusion

In this study, we investigate the price impact of uniqueness, a type of "soft" information captured in real estate property descriptions. Our contribution is threefold. First, we pro-

pose a Machine Learning algorithm to quantify the semantic uniqueness of textual property descriptions. Second, we estimate the impact of description uniqueness on real estate sale prices using log-linear hedonic pricing models, as well as the impact on time on the market. A one standard deviation increase in description uniqueness is associated with a 15% increase in sale prices while delaying the closing time by only 4 days. A large fraction of the price premium associated with description uniqueness is caused by unique features of the underlying houses, while the impact on days on market might appears more related to the composition of directly competing housing units that were sold in the same year. Finally, we examine the influence of our control for uniqueness on traditional hedonic based housing price indices. During the recovery from the great recession, the composition of the 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 naive house price index that does not consider uniqueness to mistime the start of the recovery of house prices and overstate the strength of that recovery for the metropolitan Atlanta real estate market.

This study also provides several theoretical and empirical insights concerning the use of artificial intelligence technologies. First, our machine learning algorithm defines the meanings of words within their contexts, overcoming a common limitation of the keyword-based textual analysis methods: "simply counting words (bag-of-words) ignores important context and background knowledge (Larcker and Zakolyukina, 2012)." Our machine learning algorithm naturally preserves the meaning of words within their contexts, and therefore, can understand the nuances of marketing language as well as unstandardized abbreviations in property descriptions. Second, the ML–Hedonic approach used in this study provides one of the only examples of the integration of unsupervised learning methods into economic analysis (also see Aubry et al. (2019)).

A recent boom in artificial intelligence and machine learning has had a large impact on academic research. While most of the recent machine learning studies in economics and

finance focus on predictions (supervised techniques), this paper demonstrates that unsupervised techniques can be used to harden "soft" information allowing us in our example to draw economic inferences about the impact of real estate description uniqueness on sale prices. In the future, these context-based machine learning algorithms can be applied to extract information from a wide range of textual documents to study a variety of economic phenomena.

References

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

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.

Dai, A. M., C. Olah, and Q. V. Le (2015). Document embedding with paragraph vectors. *arXiv preprint arXiv:1507.07998*.

Garcia, D. (2013). Sentiment during recessions. *The Journal of Finance 68*(3), 1267–1300.

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.

Goetzmann, W. N., S. A. Ravid, and R. Sverdlove (2013). The pricing of soft and hard information: economic lessons from screenplay sales. *Journal of Cultural Economics 37*(2), 271–307.

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.

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

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

Huang, A. G., H. Tan, and R. Wermers (2019). Institutional trading around corporate news: Evidence from textual analysis. *Review of Financial Studies*.

Larcker, D. F. and A. A. Zakolyukina (2012). Detecting deceptive discussions in conference calls. *Journal of Accounting Research 50*(2), 495–540.

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*.

Loughran, T. and B. McDonald (2011). When is a liability not a liability? textual analysis, dictionaries, and 10-ks. *The Journal of Finance 66*(1), 35–65.

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

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).

Nowak, A. and P. Smith (2017, June). Textual Analysis in Real Estate. *Journal of Applied Econometrics 32*(4), 896–918.

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

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

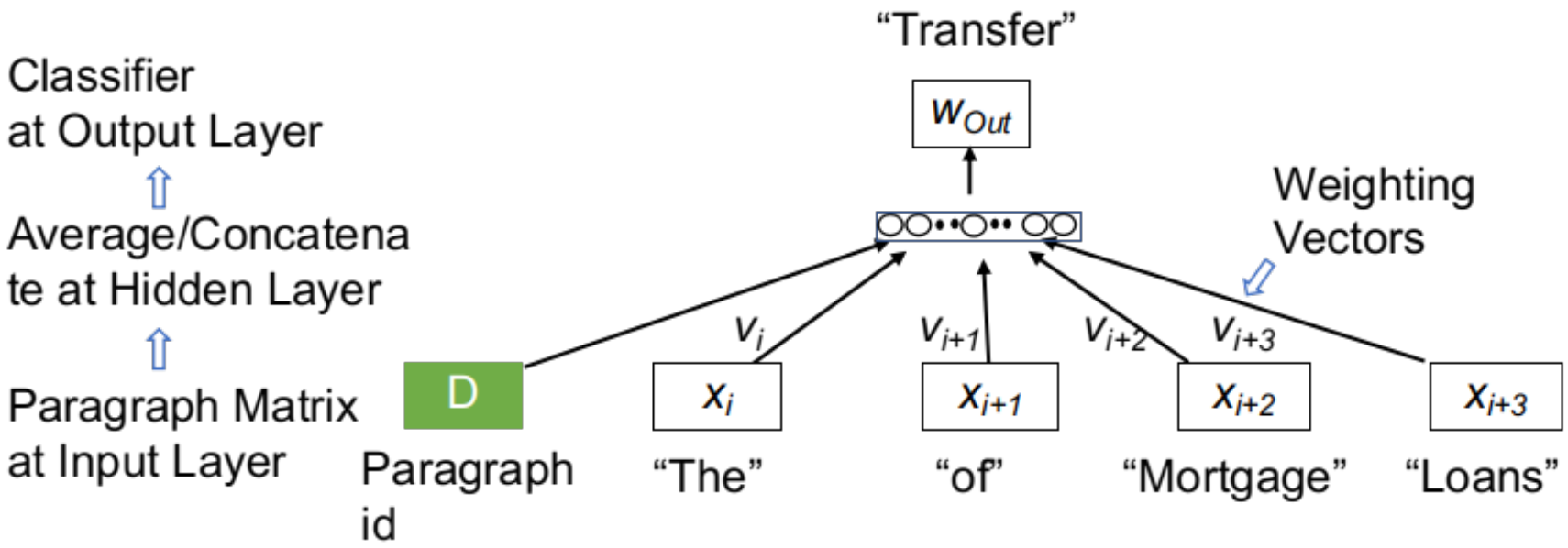
Rosen, S. (1974). Hedonic prices and implicit markets: Product differentiation in pure competition. *Journal of Political Economy 82*(1), 34–55.

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.

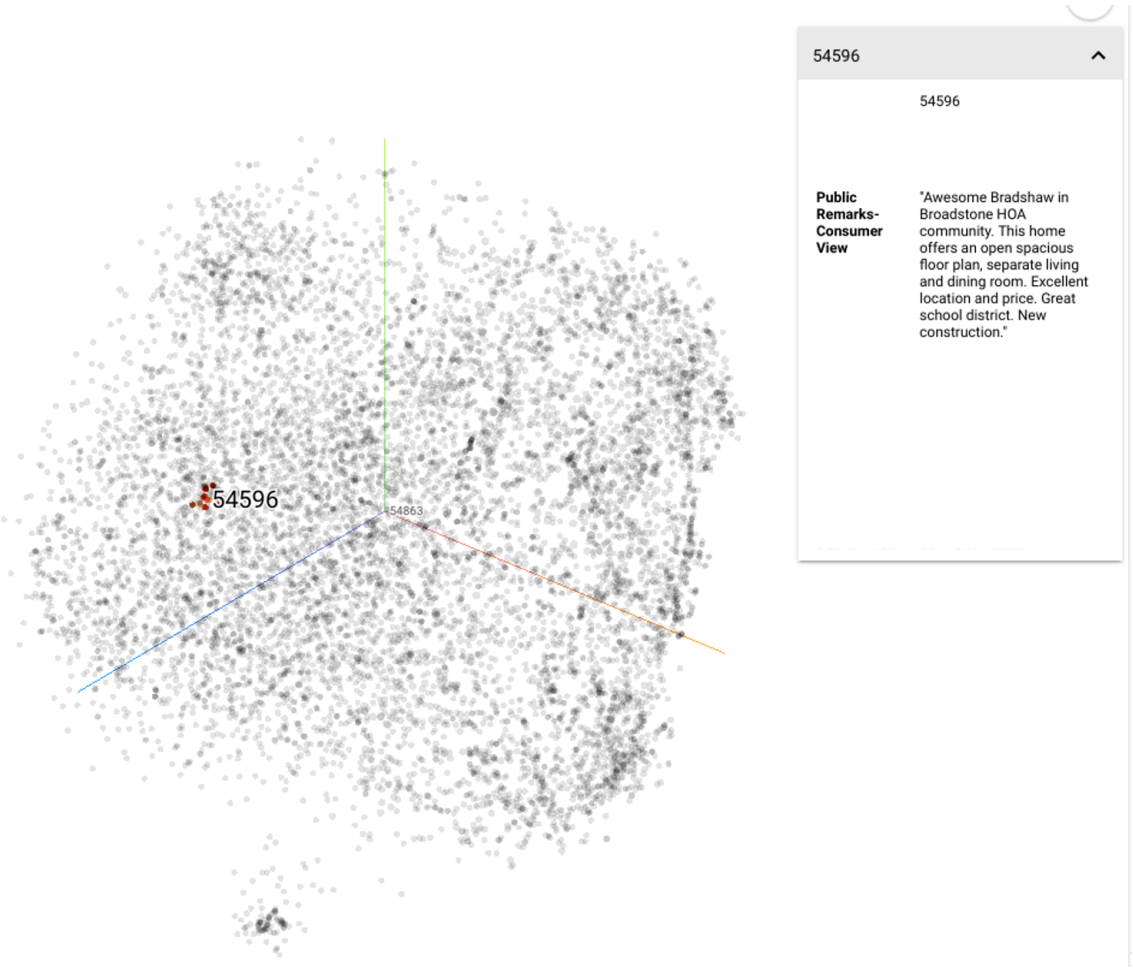
Tetlock, P. C. (2007). Giving content to investor sentiment: The role of media in the stock market. *The Journal of Finance 62*(3), 1139–1168.

Figure 1: Schematic Algorithm



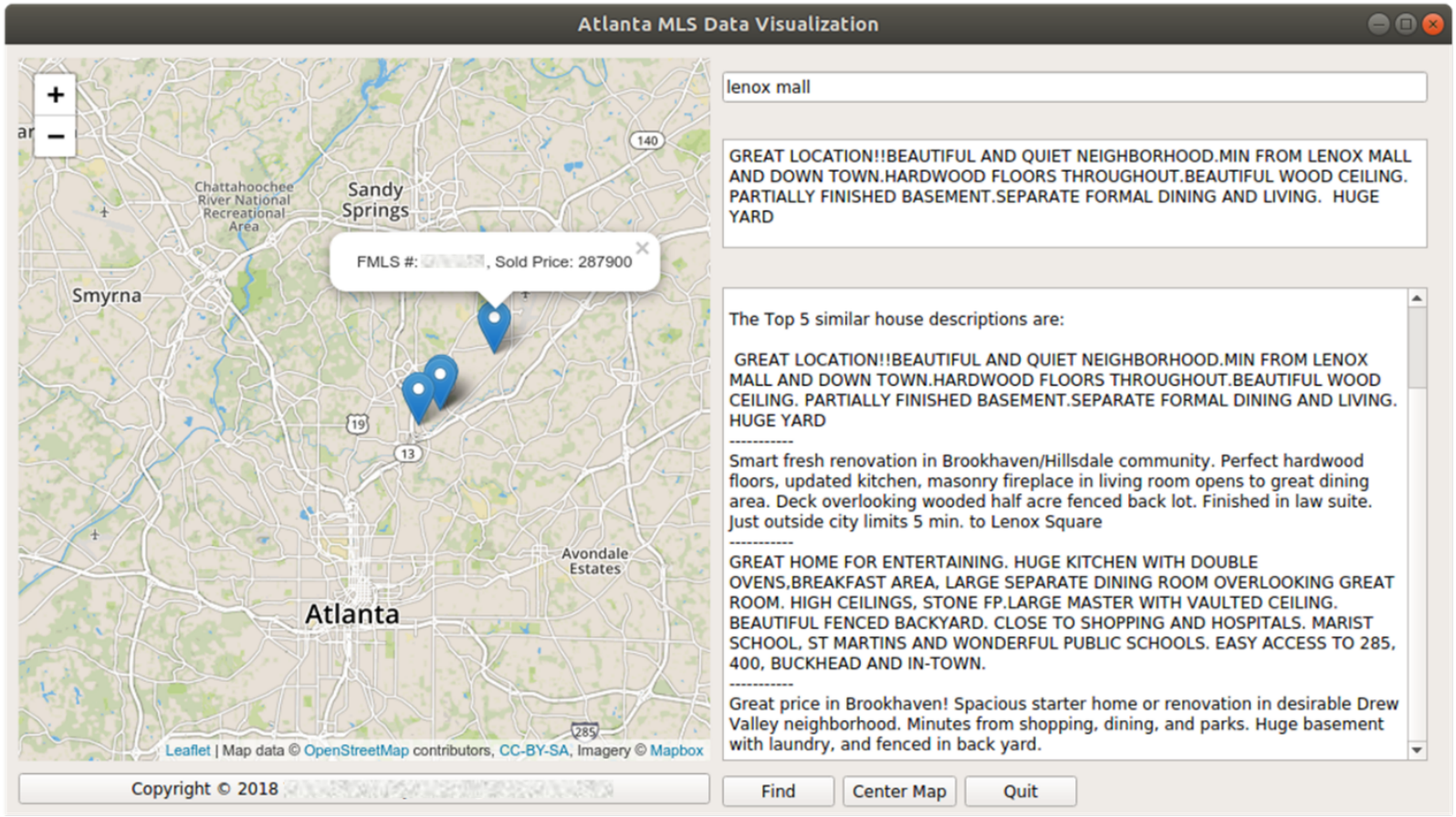
Notes: This figure illustrates the unsupervised learning process with the bottom row representing the paragraph vector and four input word vectors associated with a specific output or target word. These vectors are multiplied by the appropriate paragraph and word weight vectors illustrated by the middle row of the figure and summed to create an index. The final/top step applies a transformation to calculate the probability of observing the output word given the observed index.

Figure 2: 3–D Dimension Illustration of the MLS Vector Space



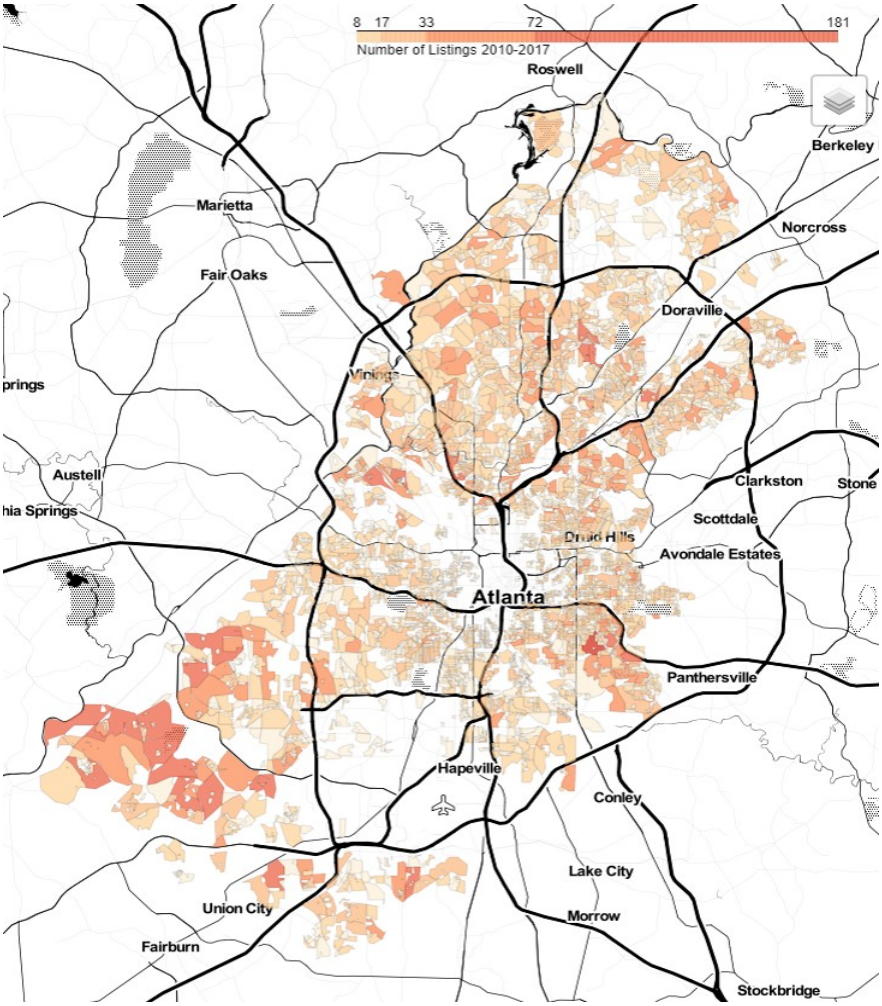
Notes: This figure provides a 3–D demonstration of the high dimensional vector space we constructed of the MLS advertisements based on the textual descriptions. In reality, our vector space has more than 100–dimensions. Every dot represents an individual real estate description. Although specific meanings are not assigned to each dimension, the relative distance between two documents in feature space indicates the relative semantic distance between the corresponding descriptions.

Figure 3: Algorithm Visualization



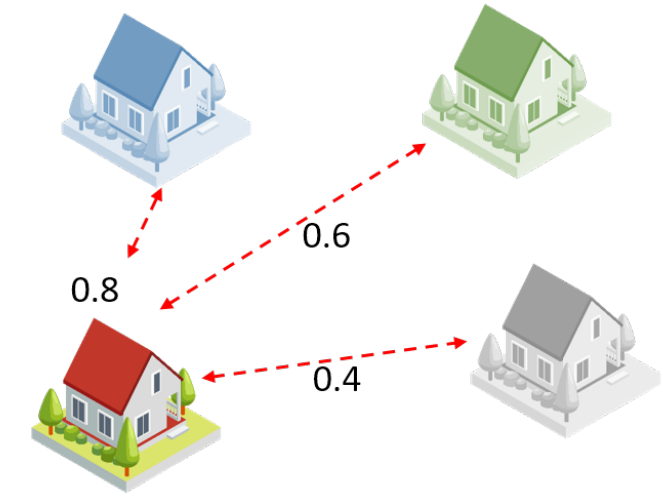
Notes: This figure provides a visual demonstration of our ML algorithm. In the top text box, the query is "Lenox Mall", a shopping center in northern Atlanta. The blue pins on the map are houses that are related to the query. The middle text box displays the description of a selected house on the map. The bottom text box shows the most similar descriptions found in the data by the 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 4: Geographical Distribution of the Real Estate Sales Sample



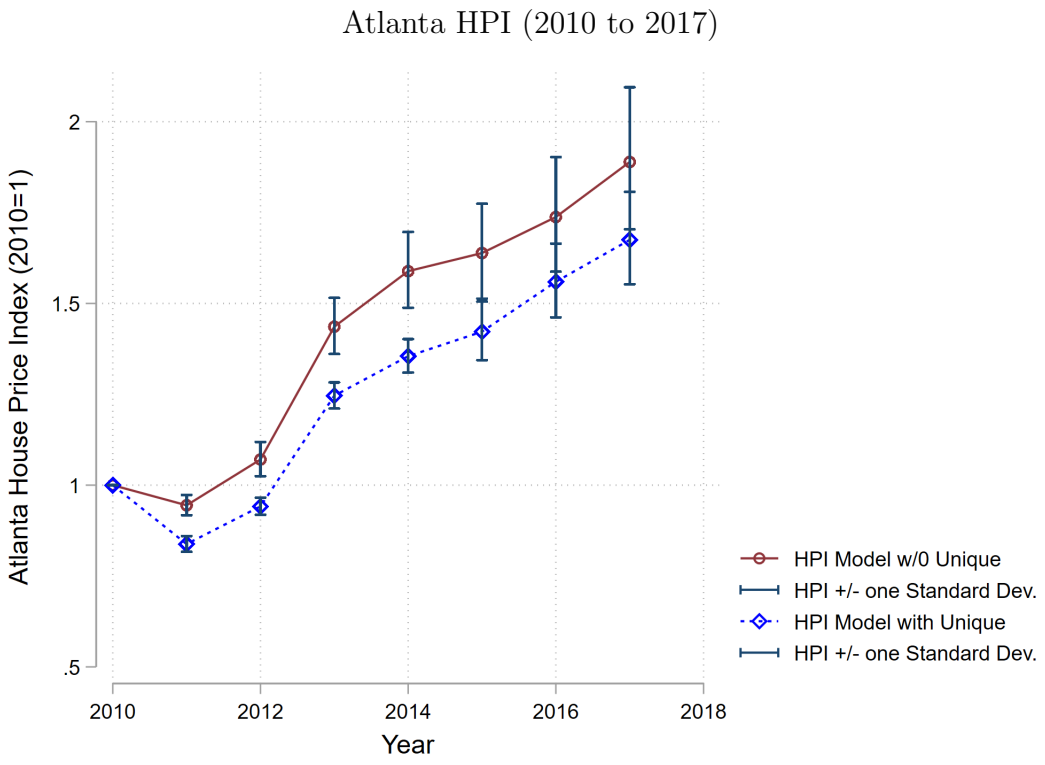
Notes: This figure displays an overview of the geographical distribution of the 40,918 single family houses analyzed in this study. Since we only focus on the city of Atlanta, most of houses sold in this area sit on small lots.

Figure 5: Schematic Unique Score Computation within a Neighborhood



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

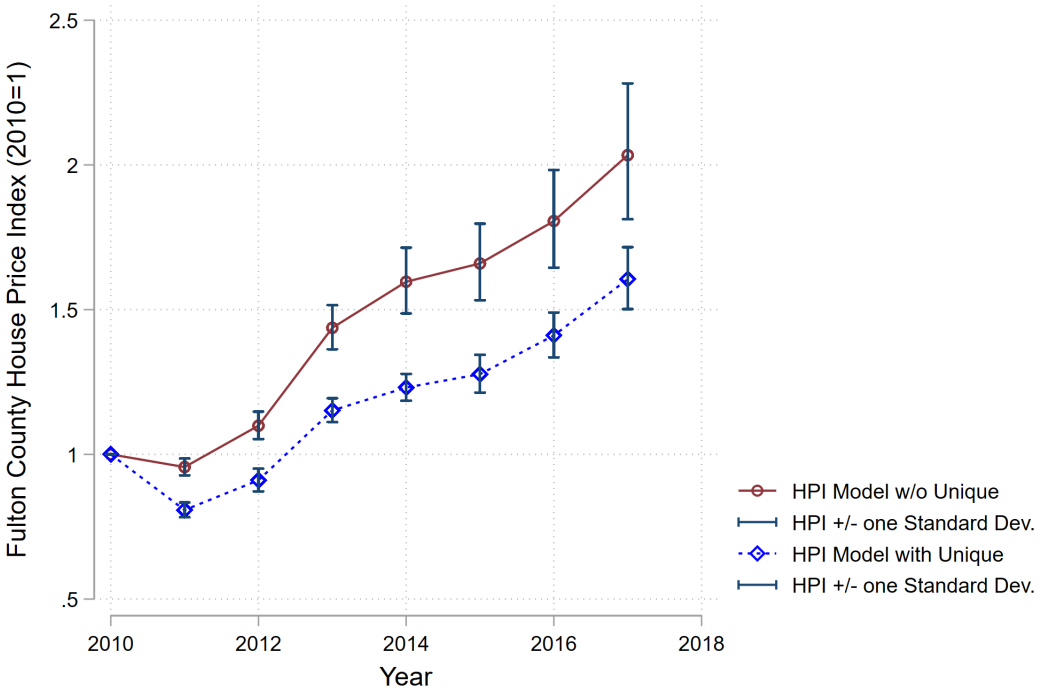
Figure 6: City-Level House Price Index Estimations



Notes: This figure displays HPI estimations using different Hedonic models. The average HPIs are displays with +/- one standard deviation.

Figure 7: County-Level House Price Index Estimations

Panel A: Fulton County HPI (2010 to 2017)



Panel B: Dekalb County HPI (2010 to 2017)

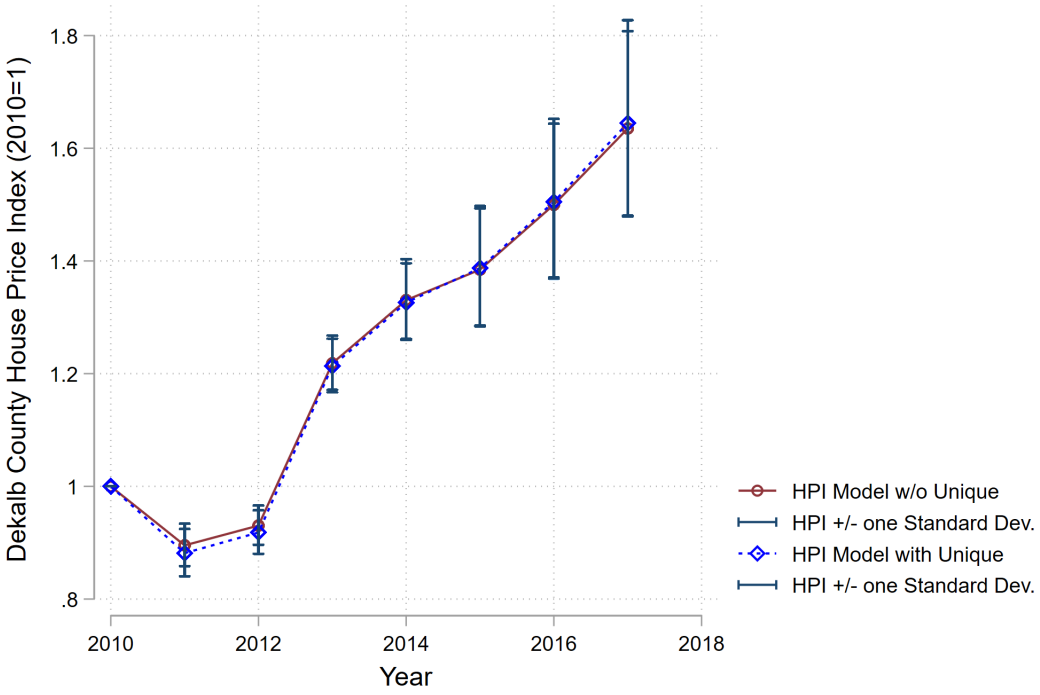
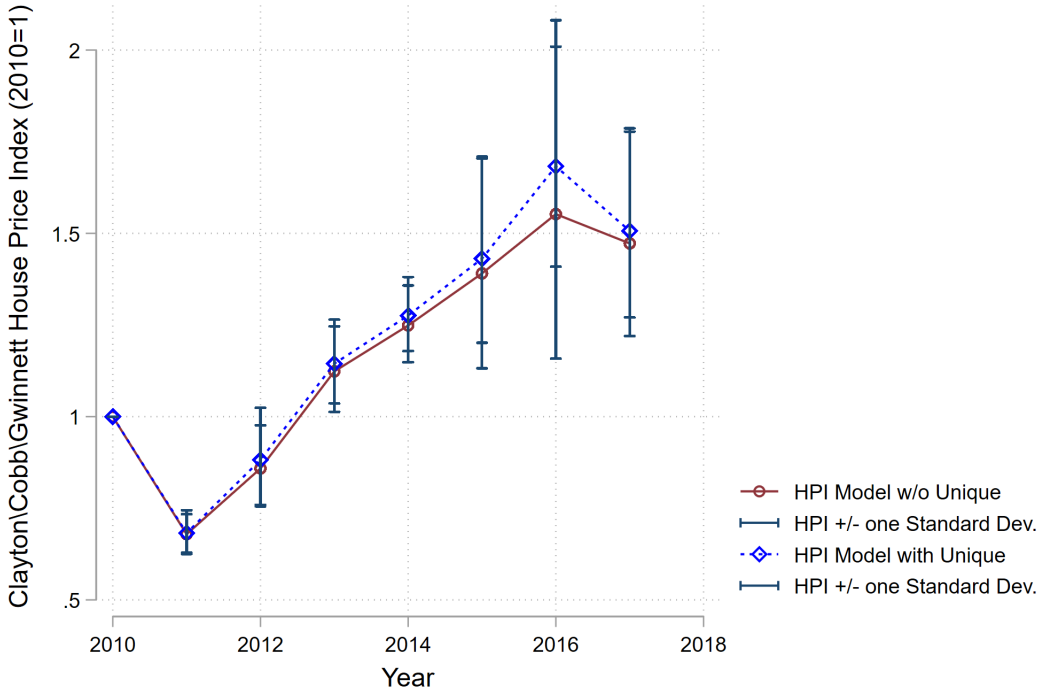


Figure 7: County-Level House Price Index Estimations (Cont.)

Panel C: Clayton–Cobb–Gwinnet HPI (2010 to 2017)



Notes: This figure displays HPI estimations using different Hedonic models. The average HPIs are displays with +/- one 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
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!	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
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!	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
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!	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 keywords or word frequencies.

Table 2: Descriptive Statistics

Panel A: The Real Estate Sale Sample

VARIABLES	(1) Mean	(2) Std. Dev.
DOM (Number of Days on Market)	104.6	94.12
Listing Price ($ Thousand)	390.0	451.9
Ln (Sale Price)	12.26	1.234
Sale Price ($ Thousand)	373.2	411.6
Age	46.20	31.02
Fireplace (Number of Fireplace)	1.043	1.039
Sqft (Hundred)	22.93	13.57
Bath (Number of Bathroom)	2.700	1.285
Bed (Number of Bedroom)	3.606	1.034
Listing Year	2,013	2.289
Sold Year	2,014	2.208
Ranch (d)	0.440	0.496
Pool (d)	0.0491	0.216
Renovated (indicator variable)	0.0708	0.257
Sold-As-Is (indicator variable)	0.118	0.323
Auction (indicator variable if foreclosure auction)	0.0249	0.156
Large Lot (indicator variable if lot>=1 acre)	0.0247	0.155
Feature (indicator variable)	0.0134	0.115
Owner Agent (indicator variable if agent related to owner)	0.0285	0.166
Dual (indicator variable if dual agent)	0.0571	0.232

Panel B: Real Estate Description Sample

VARIABLES	(1) Mean	(2) Min	(3) Max	(4) Std. Dev.
Word (# of words in property description)	80.27	10	164	31.50
Sentence (# of sentences in description)	5.900	1	23	2.778
$Unique\_Area_i$ (MLS Area Unique Score)	0.787	0.595	0.954	0.047
$Unique\_AreaYear_i$ (MLS Area-Year Unique Score)	0.778	0.384	0.940	0.055

Note: This sample contains 40,918 single-family home sales in Atlanta from January 2010 to December 2017. The final data used in this study include 37,124 unique sales and 3,794 repeat sales. We use the unique sales to deliver the baseline results and the repeat sales data are used to conduct robustness analysis.

Table 3: Compare Covariates with and without Unique Measures

Dependent Variable	Ln(Prince)		DOM	
	(1)	(2)	(3)	(4)
MLS Area-Year Unique Score (standardized)		0.152***		4.301***
		(0.035)		(0.952)
Age	-0.013***	-0.012***	-0.522***	-0.510***
	(0.001)	(0.001)	(0.091)	(0.093)
AgexAge	0.000***	0.000***	0.004***	0.004***
	(0.000)	(0.000)	(0.001)	(0.001)
Bed	-0.041**	-0.037*	-0.440	-0.370
	(0.021)	(0.019)	(0.945)	(0.942)
Bath	0.241***	0.240***	6.087***	6.344***
	(0.017)	(0.015)	(1.116)	(1.121)
Ranch	-0.270***	-0.253***	0.228	0.399
	(0.025)	(0.020)	(1.133)	(1.132)
Renovated	0.262***	0.234***	-5.181**	-5.682***
	(0.031)	(0.026)	(2.170)	(2.175)
Sqft	0.009***	0.008***	0.372***	0.366***
	(0.001)	(0.001)	(0.095)	(0.095)
Large Lot	0.020	0.011	21.526***	21.324***
	(0.029)	(0.026)	(4.143)	(4.144)
Pool	0.055***	0.057***	-2.044	-2.006
	(0.018)	(0.017)	(2.913)	(2.902)
Feature	0.127***	0.092***	8.784	7.967
	(0.033)	(0.031)	(6.383)	(6.386)
Constant	11.958***	8.036***	155.179***	130.276***
	(0.063)	(0.401)	(21.350)	(20.821)
Observations	37,124	37,107	37,124	37,107
R-squared	0.762	0.772	0.082	0.083
House Characteristics	Yes	Yes	Yes	Yes
Keywords	No	No	No	No
Transaction Characteristics	No	No	No	No
LocationxYear FE	Yes	Yes	Yes	Yes
Agent FE	No	No	No	No
Cluster	Area-Year	Area-Year	Area-Year	Area-Year

Note: The dependent variable in column(1)–(2) are Ln(sale price), and the dependent variable in column(3)–(4) are number of days on market. We control for the listing price in the DOM models. Robust standard errors are clustered at the MLS Area-Year level, shown in parentheses (*** p<0.01, ** p<0.05, * p<0.1)

Table 4: MLS Area Uniqueness vs. MLS Area–Year Uniqueness

Dependent Variables	Ln(Price)		Δ Ln(Prince)		DOM	
	(1)	(2)	(3)	(4)	(5)	(6)
MLS Area-Year Unique Score (standardized)	0.152***				4.295***	
	(0.035)				(0.953)	
MLS Area Unique Score (standardized)		0.130***				2.707***
		(0.030)				(0.802)
Δ MLS Area-Year Unique Score (standardized)			0.075***			
			(0.012)			
Δ MLS Area Unique Score (standardized)				0.055***		
				(0.012)		
Observations	37,107	37,122	3,791	3,794	37,107	37,122
R-squared	0.772	0.770	0.456	0.449	0.083	0.082
House Characteristics	Yes	Yes	Yes	Yes	Yes	Yes
Transaction Characteristics	No	No	Yes	Yes	No	No
LocationxYear FE	Yes	Yes	Yes	Yes	Yes	Yes
Agent FE	No	No	No	No	No	No
Cluster	Area-Year	Area-Year	Area-Year	Area-Year	Area-Year	Area-Year

Note: The dependent variable in column(1)–(2) are Ln(sale price), the dependent variable in column(3)–(4) are change in prices between two repeat sales, and the dependent variable in column(5)–(6) are number of days on market. We control for the listing price in the DOM models. Robust standard errors are clustered at the MLS Area-Year level, shown in parentheses (*** p<0.01, ** p<0.05, * p<0.1)

Table 5: Robustness

Panel A: House Price Analysis

Dependent Variable:	Ln(Price)			
	(1)	(2)	(3)	(4)
MLS Area-Year Unique Score (standardized)	0.152***	0.153***	0.140***	0.124***
	(0.035)	(0.033)	(0.027)	(0.020)
Observations	37,107	37,107	37,107	37,107
R-squared	0.772	0.785	0.812	0.887
House Characteristics	Yes	Yes	Yes	Yes
Keywords	No	No	Yes	Yes
Transaction Characteristics	No	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes
Location FE	Yes	Yes	Yes	Yes
LocationxYear FE	Yes	Yes	Yes	Yes
Agent FE	No	No	No	Yes
Cluster	Area-Year	Area-Year	Area-Year	Area-Year

Panel B: Days on Market Analysis

Dependent Variable:	DOM			
	(1)	(2)	(3)	(4)
MLS Area-Year Unique Score (standardized)	5.076***	5.104***	4.905***	5.975***
	(1.002)	(0.995)	(1.020)	(0.865)
Observations	37,107	37,107	37,107	37,107
R-squared	0.084	0.085	0.093	0.369
House Characteristics	Yes	Yes	Yes	Yes
Keywords	No	No	Yes	Yes
Transaction Characteristics	No	Yes	Yes	Yes
LocationxYear FE	Yes	Yes	Yes	Yes
Agent FE	No	No	No	Yes
Cluster	Area-Year	Area-Year	Area-Year	Area-Year

Note: The dependent variable in all specifications is Ln(sale price). Robust standard errors are clustered at the ZIP Code level, shown in parentheses (*** p<0.01, ** p<0.05, * p<0.1)

Table 6: Good Uniqueness Versus Bad Uniqueness

Dependent Variable:	Ln(Price)		
	(1)	(2)	(3)
MLS Area-Year Unique Score (standardized)	0.159***	0.143***	0.127***
	(0.034)	(0.029)	(0.021)
Good (# of Positive word>=5)	1.023***	0.449*	0.395*
	(0.337)	(0.238)	(0.223)
Good× MLS Area-Year Unique Score	-0.060**	-0.048***	-0.039**
	(0.024)	(0.017)	(0.016)
Bad (# of Negative Word>=1)	0.237	0.255	0.168
	(0.408)	(0.339)	(0.279)
Bad×MLS Area-Year Unique Score	-0.059**	-0.033	-0.023
	(0.030)	(0.024)	(0.021)
Observations	32,500	32,500	32,500
R-squared	0.798	0.815	0.889
House Characteristics	Yes	Yes	Yes
Keywords	No	Yes	Yes
Transaction Characteristics	Yes	Yes	Yes
LocationxYear FE	Yes	Yes	Yes
Agent FE	No	No	Yes
Cluster	Area-Year	Area-Year	Area-Year

Notes: This table displays the estimation results of

$$Ln(Price_i) = \alpha + \theta_1 Unique_i + \theta_2 Good + \theta_3 Good \times Unique_i + \theta_4 Bad + \theta_5 Bad \times Unique_i + X_i'\beta + \eta_z + \mu_c + \delta_t + \mu_c \times \delta_t + \varepsilon.$$

Standard errors are clustered at the ZIP Code Level (*** p<0.01, ** p<0.05, * p<0.1).

Table 7: 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.
Area-Year Unique Score	0.75	0.07	0.78	0.06	0.78	0.05	0.78	0.05
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		2017	
VARIABLES	Mean	Std. Dev.	Mean	Std. Dev.	Mean	Std. Dev.	Mean	Std. Dev.
Area-Year Unique Score	0.79	0.04	0.78	0.05	0.78	0.06	0.78	0.05
Ln(Sale Price)	12.45	1.12	12.47	1.10	12.49	1.05	12.54	1.01
Age	46.85	31.35	47.60	32.02	48.53	31.94	50.44	31.11
Fireplace (d)	1.10	1.06	1.08	1.06	1.03	1.06	0.95	0.98
Pool (d)	0.05	0.22	0.05	0.22	0.05	0.21	0.04	0.20
Bed	3.61	1.05	3.59	1.02	3.60	1.02	3.57	1.02
Ranch (d)	0.44	0.50	0.42	0.49	0.44	0.50	0.47	0.50
Bath	2.73	1.29	2.71	1.28	2.68	1.26	2.64	1.29
Renovated (d)	0.08	0.27	0.09	0.28	0.10	0.30	0.12	0.32
Sqft (100)	23.75	14.51	23.53	13.72	23.06	13.86	22.60	13.74

Notes: This table displays the summary statistics of the hedonic control variables by year.

Table 8: Coefficient of Year FE

Dependent Variable: Ln (Price)												
Region:	Atlanta			Fulton County			Dekalb County			Clayton–Cobb–Gwinnet		
	Model 1 (1)	Model 2 (2)	Diff. M1- M2 (3)	Model 1 (4)	Model 2 (5)	Diff. M1- M2 (6)	Model 1 (7)	Model 2 (8)	Diff. M1- M2 (9)	Model 1 (10)	Model 2 (11)	Diff. M1- M2 (12)
2011 (d)	-0.057*	-0.177***	0.120	-0.045	-0.214***	0.169	-0.111**	-0.126**	0.015	-0.387***	-0.382**	-0.005
	(0.030)	(0.026)		(0.030)	(0.032)		(0.042)	(0.048)		(0.077)	(0.087)	
P(Chi2)			0.000			0.000			0.028			0.803
2012 (d)	0.068	-0.060**	0.128	0.095**	-0.094**	0.189	-0.072*	-0.085*	0.013	-0.152	-0.126	-0.026
	(0.044)	(0.025)		(0.043)	(0.043)		(0.037)	(0.042)		(0.128)	(0.150)	
P(Chi2)			0.004			0.000			0.037			0.327
2013 (d)	0.362***	0.220***	0.142	0.363***	0.141***	0.222	0.198***	0.194***	0.004	0.116	0.135	-0.019
	(0.054)	(0.029)		(0.053)	(0.036)		(0.039)	(0.039)		(0.104)	(0.100)	
P(Chi2)			0.003			0.000			0.378			0.072
2014 (d)	0.463***	0.304***	0.159	0.468***	0.208***	0.260	0.286***	0.282***	0.004	0.222*	0.244**	-0.022
	(0.066)	(0.034)		(0.071)	(0.038)		(0.053)	(0.051)		(0.084)	(0.079)	
P(Chi2)			0.006			0.000			0.617			0.335
2015 (d)	0.494***	0.352***	0.142	0.507***	0.245***	0.262	0.325***	0.327***	-0.002	0.330	0.358	-0.028
	(0.080)	(0.057)		(0.080)	(0.051)		(0.076)	(0.076)		(0.206)	(0.175)	
P(Chi2)			0.019			0.000			0.533			0.443
2016 (d)	0.553***	0.444***	0.109	0.591***	0.344***	0.247	0.405***	0.409***	-0.004	0.440	0.520**	-0.080
	(0.091)	(0.065)		(0.093)	(0.055)		(0.092)	(0.093)		(0.293)	(0.177)	
P(Chi2)			0.115			0.000			0.307			0.555
2017 (d)	0.636***	0.516***	0.120	0.710***	0.473***	0.237	0.492***	0.498***	-0.006	0.387	0.410*	-0.023
	(0.103)	(0.076)		(0.115)	(0.067)		(0.100)	(0.105)		(0.188)	(0.170)	
P(Chi2)			0.038			0.000			0.414			0.272
Control Unique	No	Yes		No	Yes		No	Yes		No	Yes	
Observations	40,918	40,900		26,778	26,770		13,288	13,285		852	845	
R-squared	0.509	0.571		0.568	0.633		0.522	0.527		0.689	0.691	

Notes: This table displays the year indicator coefficient estimates from two seperate hedonic models. The dependent variable is Ln(price). Model 1 does not control for the uniqueness score of each house, and Model 2 includes "Unique" in the hedonic model. Standard errors are clustered at the ZIP Code Level (*** p<0.01, ** p<0.05, * p<0.1).