Department of Economics

Working Paper Series

Some Remarks on Real Estate Pricing

Crocker Liu

Adam Nowak

Patrick Smith

Working Paper No. 17-20

This paper can be found at the College of Business and Economics Working Paper Series homepage:

http://business.wvu.edu/graduate-degrees/phd-economics/working-papers

Some Remarks on Real Estate Pricing

Preliminary Draft - Please Do Not Cite

Crocker Liu *
Cornell University

Adam Nowak †
West Virginia University

Patrick Smith ‡
San Diego State University

July 4, 2017

JEL Codes: R00 (General Urban, Rural, and Real Estate Economics), R33 (Nonagricultural and Nonresidential Real Estate Markets)

Key Words: Omitted variable bias, textual analysis, agent-owned, agency problem

*Cornell University; Email: chl62@cornell.edu
†West Virginia University; Email: adam.d.nowak@gmail.com
‡San Diego State University; Email: patrick.smith@sdsu.edu

[1]We thank Brian Chew for providing access to the Georgia MLS data. We are grateful to participants at the AREUEA national and WEAI annual meetings. We also thank Tian Liu and Jenny Schuetz for their discussion and insightful comments. Errors are the responsibility of the authors.

Abstract

This paper develops a framework for addressing the omitted variable bias that plagues most real estate research. We incorporate qualitative information from text to control for property attributes that are generally unobserved. The textual information is entered by real estate agents for every property sold on a Multiple Listing Service (MLS). The agents, who arguably have the most local market and property specific knowledge, use the unstructured text to highlight important information that is not clearly conveyed in other areas of the listing. Although the framework can be applied universally in real estate research, we demonstrate its effectiveness in the estimation of agent-owned sales premiums. Similar to previous studies, we find agent-owned premiums between 2% to 6% when no textual information is included. When we include the textual information the agent-owned premiums dissipate. The results suggest that the market distortions reported in Rutherford et al. [2005] and Levitt and Syverson [2008] do not exist.

1 Introduction

Empirical results in real estate research are often reported with a disclaimer that they may suffer from an omitted variable bias. Two of the most commonly cited omitted variables are the condition and quality of the property.[1] Researchers recognize that property condition and quality are likely correlated with other observable variables (e.g. physical characteristics, neighborhood amenities, distressed sale conditions) that are incorporated into hedonic models. The "bias" occurs when the model compensates for the omitted variables by over- or underestimating the effect of the other observable variables. Multiple listing service (MLS) and tax assessor data typically do not include accurate measures of property condition or quality at the time of sale.[2] Techniques to address omitted variable bias, such as including house fixed effects in a repeat sale model, assume that house condition and quality remain constant. Although quality may remain constant over time, condition is expected to change because of deterioration and renovation.[3]

In their examination of preferences for schools and neighborhoods Bayer et al. [2007] note that they do not

> "address the possibility that the higher-income households on the higher test score side of a school boundary might be more likely to make home improvements (install granite countertops, e.g.) unobserved by the researcher, in turn contributing to the higher average house prices on that side of the boundary. That said, we are unaware of any paper in the literature that has been able to deal with this issue."

In this paper we propose a new approach for addressing the omitted variable bias described

[1] Property condition and quality are two different measures. Property condition is a time-varying measure of how well the property has been maintained. Whereas, property quality is a time-invariant measure of the workmanship and materials used in the property's construction.

[2] County tax assessor datasets include a CDU (condition, desirability and utility) grade for each property. The CDU grade is often unreliable because the county's appraiser does not enter the house post-construction.

[3] The constant quality assumption does not always hold as large scale renovations (obsolescence) may increase (decrease) house quality over time.

in Bayer et al. [2007]. We show that the public remarks section of the MLS includes relevant information about the time-varying and time-invariant property features that contribute to omitted variable bias. The public remarks section allows the listing agent, who is the only professional to enter and evaluate the interior of the house, to provide a description of the property beyond what is captured in the standard MLS fields.[4] Given its limited length, listing agents use the public remarks section to highlight important information (i.e. property quality, property condition, seller motivation, purchase incentives) that is not clearly conveyed in other areas of the listing. We incorporate the qualitative information from the public remarks section to remedy the omitted variable bias - thereby addressing the concerns in Bayer et al. [2007].

The approach we present can be employed universally across all real estate research. In this paper we demonstrate its effectiveness in the estimation of agent-owned sales premiums. We examine agent-owned sale premiums because they represent a rare example of a principal-agent conflict that offers a clean identification strategy. When real estate agents list their own house for sale on the MLS they are required by law to notify potential buyers that the principal (i.e. the agent who is selling the house) holds a real estate license.[5] Seminal studies on the topic by Rutherford et al. [2005] and Levitt and Syverson [2008] argue that listing agents, who are better informed than their clients, use their informational advantage to sell their own house for a higher price than their clients' house. Both studies find that agent-owned houses sell for a premium relative to non-agent-owned houses. Rutherford et al. [2005] find a 4.5% premium using data from Texas between 1998 and 2002 and Levitt and Syverson [2008] find a 3.7% premium using data between 1992 and 2002 from Illinois.

[4]Even appraisers who play a critical role in lenders' underwriting decisions rely on MLS listing information. For example, Young [2012] states that "today's appraisers are required to rate property conditions of both subject properties and comparables using a numerical scale from C1 to C6. Where do they get the information needed to make these ratings? Typically from the information that is provided in the MLS listing by the listing agent, including photos, remarks, and descriptions of physical features found in the various fields for listing input. As appraisers rely on the information found in the MLS, the more descriptive and accurate that information is, the better appraisal reports can be."

[5]For example, Rule 520-1-.09 (8) of Georgia's Administrative Code states that "A licensee shall not advertise to sell, buy, exchange, rent, or lease real estate in a manner indicating that the offer to sell, buy, exchange, rent, or lease such real estate is being made by a private party not licensed by the Commission."

Rutherford et al. [2005] and Levitt and Syverson [2008] conclude that real estate agents do, in fact, exploit their informational advantage when selling their own house. However, both studies note that agent-owned houses systematically differ from those of their clients and report their results with the usual caveat that the agent-owned estimates may suffer from an omitted variable bias. For example, Rutherford et al. [2005] note that "another possible explanation is that owner-agents initially buy higher quality properties" and Levitt and Syverson [2008] note that "a particular concern is that agents live in houses that are especially attractive along dimensions that are difficult to observe or quantify." Rutherford et al. [2005] attempt to address this concern by including a quality measure from the county appraisal board for a small subsample of the properties in their study. Whereas, Levitt and Syverson [2008] take an approach similar to the one we advocate in this paper. They add indicator variables for nearly 100 keywords used in the written marketing description of the house.[6] The two studies take different approaches to address the omitted variable bias concern, but in the end both conclude that agents sell their own house for a premium.

We address the omitted variable bias using qualitative information from the public remarks section of the MLS. In addition to including information about property condition, our approach incorporates other salient time-varying (seller motivation, purchase incentives) and time-invariant (quality, amenities) features of the property that may bias agent-owned premium estimates. We apply the approach using MLS data from Atlanta, Georgia and Phoenix, Arizona. Similar to previous studies, we find agent-owned premiums in the 2% to 6% range when the qualitative information is not included. After incorporating the qualitative information from the public remarks the agent-owned premium ceases to exist. Contrary to previous studies, the results suggest that an agency problem does not exist and that real estate agents do not exploit their informational advantage when selling their own house.

[6] Levitt and Syverson [2008] do not describe their variable selection process, so it is unclear how the keywords were chosen. The authors also note that they selected several keywords that are either "superficially positive", redundant, or "do not describe particular characteristics of the house".

2 Theory and Estimation

2.1 Hedonic Pricing Model

Consider the hedonic model in equation 1 where price of house n at time t is linear in date of sale, t, observable time-varying attributes, x_{nt}, observable time-invariant attributes, z_n, unobserved time-varying attributes, ψ_{nt}, unobserved time-invariant attributes, μ_n, and a disturbance term, e_{nt}.

$$p_{nt} = w_{nt}\delta + x_{nt}\beta + z_n\theta + \mu_n + \psi_{nt} + e_{nt} \tag{1}$$

Here, $w_{nt} \in \mathbb{R}^T$ is a basis vector of 0s except for a 1 in the t position, $x_{nt} \in \mathbb{R}^{K_x}$ includes time-varying attributes such as square footage, age, bedrooms, and bathrooms, $z_n \in \mathbb{R}^{K_z}$ includes zip code or census tract fixed effects that capture time-invariant property and neighborhood attributes such as local schools, parks, or water access. In Equation 1, $\delta = (\delta_1, ..., \delta_T)'$ is the vector of the time-varying market-wide value of housing. Although Wallace and Meese [1997] provide evidence that β and θ are time-varying, for simplicity, we assume they are time-invariant parameters. Throughout, without loss of generality, we also assume e_{nt} is a zero mean, independent random variable. The unobserved time-varying attributes, ψ_{nt}, include the condition, seller motivation, and other features of the house that change over time. Whereas, the unobserved time-invariant attributes, μ_n, include features of the house, such as its quality, that remain constant over time. By definition, only observable attributes are included in the hedonic regression estimated as shown in Equation 1.

$$p_{nt} = w_{nt}\delta + x_{nt}\beta + z_n\theta + u_{nt} \tag{2}$$

The term $u_{nt} = \mu_n + \psi_{nt} + e_{nt}$ is a composite term including relevant unobserved attributes and the error term. Agent-owned premiums are estimated by including an indicator variable for agent-owned transactions in x_{nt}. Specifically, $g_{nt} = 1$ if the sale is

an agent-owned transactions and $g_{nt} = 0$ otherwise. Under the orthogonality conditions $E[x_{nt}u_{nt}] = E[z_n u_{nt}] = E[w_t u_{nt}] = 0$, the coefficient estimate for the agent-owned transaction is unbiased. However, when one or all of these conditions are not met, the estimating equation suffers from an omitted variable bias, and the coefficient for agent-owned transactions is biased.[7] As mentioned above, these conditions are violated if agent-owned transactions are in superior condition relative to non-agent-owned houses $E[g_{nt}\psi_{nt}] > 0$ or if agents are more likely to purchase and subsequently sell high quality properties $E[g_{nt}\mu_{nt}] > 0$.

The omitted variable bias may not be resolved when a repeat-sales estimator approach is taken a la Mayer (1998).[8] Differencing Equation 1 gives us:

$$\Delta p_{nt} = p_{nt} - p_{ns} = \Delta w_{nt}\delta + \Delta x_{nt}\beta + \Delta\psi_{nt} + \delta e_{nt} \tag{3}$$

Similar to the hedonic model, an unbiased estimate of the agent-owned premium requires $E[\Delta x_{nt}\psi_{nt}] = 0$.[9] Unlike the hedonic model, an unbiased agent-owned premium in the repeat-sales estimator does not require any assumptions about the correlation between agent-owned transactions and quality. In any event, agent-owned properties with superior maintenance (i.e. excellent condition) will still bias agent-owned premium estimates.

2.2 Textual Analysis

We address the omitted variable bias concern using qualitative information from the public remarks section of the MLS. Augmenting the standard attribute fields in the MLS with information from the public remarks section is not entirely new. Levitt and Syverson [2008] use "nearly 100 indicators for keywords included in the written description of the home (such as spacious, amazing, granite, youthful)" in their examination of agent-owned real

[7] Even when $E[g_{nt}u_{nt}] = 0$, the coefficient for is still biased unless it is uncorrelated with the other variables in the regression.

[8] The use of the three-stage generalized least squares (GLS) methodology proposed by Case and Shiller (1987) addresses some concerns. However, Case and Shiller use the methodology to create a repeat sales index, so they are only interested in estimating the time coefficients.

[9] In addition, we also require $E[\Delta w_{nt}\psi_{nt}] = 0$ but this is not the focus of the paper.

estate transactions. Ben-David [2011] identifies houses with inflated prices by examining "the textual description of the properties for clues", although he does not include the variables in his empirical analysis. In both studies, the keywords were chosen by the researchers.

Nowak and Smith [2017] show that the MLS remarks section contains indicators of both time-invariant and time-varying property quality and condition. This is not surprising because real estate agents likely use the remarks section to convey important information that is not conveyed by the standard attributes of the property (e.g. square feet of living area, number of bedrooms, number of bathrooms, etc.) Nowak and Smith [2017] use a tokenization and penalized regression approach in order to (i) select and (ii) identify the implicit price for keywords in the public remarks section. The tokenization approach views the remarks as an exchangeable collection of tokens where the tokens are either words or phrases. It is common to refer to single words as *unigrams* and two-word phrases as *bigrams*.

Table 1 includes a sample of property listings for three bedroom, two bathroom houses in zip code 30043 along with the sale prices and remarks. Although the properties are otherwise identical in terms of location, bedrooms and bathrooms, there is considerable variation in the sales price. Several tokens in the remarks can be used to explain the variation in sales price. For example, the most expensive house has a *marble master bath*. In contrast, the least expensive house requires some *sweat equity*. Still, the property is in a *sought after* school district. Following Nowak and Smith [2017] we assume that indicator variables for specific words or phrases in the remarks can approximate the unobserved attributes. For example, a listing where the tokens *luxury* and *beautifully renovated* appear in the remarks is likely to have $\psi_{nt}, \mu_n > 0$. In contrast, a listing with the token *sweat equity* in the remarks is likely to have $\psi_{nt} < 0$. However, the same listing might also include *sought after* in which case $\mu_n > 0$.

To incorporate the qualitative information we define $r_{nt} \in \mathbb{R}^{K_r}$ as a vector of indicator variables for the presence of specific tokens in the remarks. We assume that $r_{nt}\gamma = \mu_n + \psi_{nt} + \epsilon_{nt}$ where γ is the vector of implicit prices for the tokens and v_{nt} reflects the approximation

error of the tokens for the true, unobserved attributes. Using this, the price can then be written as

$$p_{nt} = g_{nt}\tau + w_{nt}\delta + x_{nt}\beta + z_n\theta + r_{nt}\gamma + v_{nt} + \epsilon_{nt} = h_{nt}\alpha + v_{nt} + \epsilon_{nt} \tag{4}$$

Here, we separate g_{nt} from x_{nt} for expositional purposes, collect all explanatory variables in the vector $h_{nt} \in \mathbb{R}^K$[10], and use $\alpha \in \mathbb{R}^K$ for the vector of K parameters.

2.3 Variable Selection Process

The number of unique tokens in the remarks is extremely large. To minimize the approximation error, one must include indicator variables for hundreds or thousands of tokens. Doing so directly increases the total number of variables in the model. It is well known that a least-squares estimator that includes a large number of variables is intractable at worst or overfits the data at best. Dropping the less frequent tokens from the analysis is not recommended as there is no guarantee that the most frequent tokens will have the most predictive power. Of course, it is possible that some tokens are redundant. For example, the bigrams *sweat equity*, *handyman's delight*, and *fixer upper* are all euphamisms for properties in poor condition. However, without ex ante knowledge of the relationship between these tokens, it is not possible to reduce the number of tokens by using one of these euphamisms as a stand-in for the other two.

Noting the tradeoff between approximation error and overfitting, Nowak and Smith [2017] include a large number of tokens, place an ℓ_1 penalty on the coefficients and choose a to minimize[11]

$$\sum_n (p_{nt} - h_{nt}a)^2 + \lambda \sum_j |a_j| \tag{5}$$

In Equation 5, the first term is the total sum of squares, and the second term is a penalty

[10] $K = T + K_x + K_z + K_r$

[11] The ℓ_1 norm of a vector x is given by $\|x\|_1 = \sum_k |x_k|$.

term for the coefficients. Define the minimizer of Equation 5 as $\widehat{a}_{LASSO}$. Furthermore, define $\widehat{S}_{LASSO} \subset \{1, ..., K\}$ as the support of the non-zero coefficients in $\widehat{a}_{LASSO}$ where $\widehat{S}_{LASSO} = \{k|\widehat{a}_{LASSO} \neq 0\}$. Likewise, define $h_{nt}(\widehat{S}_{LASSO})$ as the set of variables in the support. The variable λ determines the size of the penalty. When $\lambda = 0$, there is no penalty on the coefficients and $\widehat{a}_{LASSO}$ is the least-squares solution. As $\lambda \to \infty$, the penalty on non-zero coefficients increases and $\widehat{a}_{LASSO} \to 0$. The choice of λ is either based on theory or cross-validation. The estimator in Equation 5 is known as the LASSO estimator in the literature, Tibshirani [1996].[12]

The shape of the ℓ_1 norm implies that some coefficients can be set equal to 0 at the optimum for λ between 0 and ∞. When a coefficient is equal to 0, that coefficient does not have predictive power in the model. An alternative interpretation is that the LASSO estimator simultaneously performs variable selection and coefficient estimation. In a real estate setting, Nowak and Smith [2017] find the tokens in r_{nt} have a significant amount of predictive power.

The penalty in Equation 5 biases the $\widehat{a}_{LASSO}$ towards 0. To correct for this bias, Chernozhukov et al. [2015] describe a two-step procedure that exploits the variable selection feature of the LASSO. They suggest the following:

1. Minimize Equation 5 and collect $\widehat{a}_{LASSO}$ with support $\widehat{S}_{LASSO}$.

2. Regress p_{nt} on $h_{nt}(\widehat{S}_{LASSO})$ and g_{nt}.

The two-step procedure identifies variables with significant predictive power; however, the coefficients in $\widehat{a}_{LASSO}$ are biased. Regressing p_{nt} on the variables in $\widehat{S}_{LASSO}$ yields unbiased estimates. Hypothesis testing is performed using heteroskedastically consistent standard errors.

The variables in $\widehat{S}_{LASSO}$ are chosen based on their predictive power for p_{nt} - which is not the focus of this study. Instead, we are interested in using the tokens in r_{nt} to mitigate the

[12]LASSO: Least Angle Selection and Shrinkage Operator

omitted variable bias by proxying for $\mu_n + \psi_{nt}$. Of course, there is no guarantee that the variables in $\widehat{S}_{LASSO}$ will effectively control for the omitted variable bias. With the explicit goal of estimating τ, Belloni et al. [2014] present an auxiliary estimation that can be used to identify variables not included in $\widehat{S}_{LASSO}$ that are relevant for unbiased estimation of τ. Using g_{nt} as the dependent variable, Belloni et al. [2014] specify the auxiliary equation

$$g_{nt} = w_{nt}\delta + x_{nt}^{(-g)}\beta + z_n\theta + r_{nt}\gamma + v_{nt} = h_{nt}^{(-g)}\alpha^{(-g)} + v_{nt} \tag{6}$$

Here, $h_{nt}^{(-g)} \in \mathbb{R}^{K-1}$ is the vector of variables excluding g_{nt}, and $\alpha_{nt}^{(-g)} \in \mathbb{R}^{K-1}$ is the associated vector of parameters.

The procedure described in [Belloni et al., 2014] for unbiased estimation of τ proceeds in the following manner

1. Minimize Equation 5. As above, define the minimizer as $\widehat{a}_{LASSO}$ with support $\widehat{S}_{LASSO}$.

2. Minimize Equation 5 using g_{nt} as the dependent variable and $h_{nt}^{(-g)}$ as the explanatory variables. Define the minimizer as $\widehat{a}_{LASSO}^{(-g)}$ with support $\widehat{S}_{LASSO}^{(-g)}$.

3. Create the intersection of the supports as $\widehat{S}_{BCH} = \widehat{S}_{LASSO}^{(-g)} \cap \widehat{S}_{LASSO}$ and the associated set of explanatory variables $h_{nt}(\widehat{S}_{BCH})$.[13]

4. Regress p_{nt} on $h_{nt}(\widehat{S}_{BCH})$ and g_{nt} and calculate heteroskedastically consistent standard errors.

The multi-step procedure identifies the relevant variables for unbiased estimation of τ based on two predictive equations. The first equation identifies the variables that predict p_{nt}; the second equation identifies the variables that predict the agent-owned indicator, g_{nt}.[14] Thus,

[13] Obviously, the intersection should take into account the change in the number of variables from K to $K-1$.

[14] The multi-step process helps address the omitted variable bias present in the estimation of agent-owned premiums. To ensure the equations do not select redundant factors we remove words and phrases that are perfect substitutes for the agent-owned indicator variable. The words that are removed include *owner is agent*, *agent is owner*, *seller is agent*, *owner is real estate agent*, etc. A complete list is available from the authors.

the multi-step process selects variables that are relevant for 1) pricing and 2) reducing the omitted variable bias. Whereas, the two-step process in Chernozhukov et al. [2015] focuses solely on pricing and does not address the omitted variable bias.

3 Data

We examine agent-owned transactions using MLS data from Atlanta, Georgia and Phoenix, Arizona. The Atlanta data covers the five counties (Clayton, Cobb, DeKalb, Fulton and Gwinnett) that form the core of metro-Atlanta and was provided by the Georgia Multiple Listing Service (GAMLS). The GAMLS data set includes single-family detached houses that were sold using the services of a real estate agent between January 2000 and September 2016.[15] The Phoenix data includes all transactions in Maricopa County and was provided by the Arizona Multiple Listing Service (ARMLS). Maricopa County covers both the city of Phoenix and the surrounding cities including Glendale, Mesa, Scottsdale, Tempe. The ARMLS data set includes single-family detached houses that were sold using the services of a real estate agent between January 2000 and December 2013.

Both MLS data sets contain extremely detailed information including the house's address, physical characteristics (square feet living area, bedrooms, bathrooms, etc.), listing information (list price, agent-owned, vacant, etc.), transaction details (time-on-market, sales price, etc.), and a written description about the house that the real estate agent uses to market the house.[16] We use the written description to create the remarks variable, r_{nt}, in Equation 4.

Prior to running the empirical analysis, we impose a number of restrictions on both MLS data sets. We geocode the data using the property address listed in the MLS to obtain location controls (census tract and zip code) for the empirical analysis. Property addresses

[15]The agent-owned variable was not populated in the GAMLS data until 2008, so the empirical analysis for Atlanta includes every transaction between January 2008 and September 2016.

[16]The GAMLS data does not consistently report the square feet of living area, so we match the properties to county tax assessor records obtained from CoreLogic.

that are not geocoded are dropped. Using the geocoded address we create a unique identifier that allows us to link listing and sales activity on a given property over time. We remove records for which data on variables of interest are missing or contain invalid values. To eliminate outliers and minimize data errors, we exclude houses with less than 1 or more than 6 bedrooms or bathrooms, lot sizes above 5 acres, and sales prices less than $30,000 or greater than $3,000,000. We also remove houses that are less than one year old (i.e. new construction). The filters are comparable to those employed in Levitt and Syverson [2008]. Summary statistics for the cleaned data sets are displayed in Tables 2 and 3.

4 Results

4.1 Variable Selection

Unigram and bigram tokens are both incorporated in the empirical analysis. The results are similar, so we only report on unigrams going forward. Remember, not every token (i.e. unigram) is included. Only tokens selected using the methods in [Belloni et al., 2014] are included in a least-squares regression as control variables. For practical purposes, we begin with a candidate set of the 2,000 most frequent tokens. Similar to the results in Nowak and Smith [2017], our main results and conclusions do not change in a meaningful manner when using a larger set of 3,000 candidate tokens.[17]

In the token selection procedure in Eq 5 and Eq 6, we include quarter fixed effects and zip code fixed effects. We choose zip code fixed effects for three reasons. First, our main conclusions are unaffected when including more granular fixed effects for either census tract or census block. Second, census tract fixed effects have been shown to overfit in-sample [Nowak and Smith, 2017]. Third, we found zip code fixed effects are computationally tractable.[18]

[17] The results when using 3,000 candidate tokens are included in the online appendix.

[18] We used the `hdm` package in `R` to estimate the heteroskedastic LASSO as in Belloni et al. [2014]. Computation time for the heteroskedastic LASSO using a Macbook Pro with 8GB of memory and a 2.7 Ghz Intel Core i5 was approximately 30 minutes using using 2,000 candidate tokens and zip code fixed effects. Computation time on the same machine with additively separable census tract fixed effects was more than

We also include indicator variables for square footage, bedrooms, and bathrooms. Age enters into the hedonic function linearly. We find these specifications allow for possibly important nonlinear relationships in the true hedonic price function while also providing easily interpreted coefficients. Least-squares coefficient estimates for the indicator variables are presented in Table 4 and 5. The second column present the results without tokens and the third column presents coefficient estimates when the tokens are included in the regression.

The primary contribution of this study is the inclusion of the tokens in $\widehat{S}_{BCH}$ in the final hedonic model alongside the standard attributes described above. The tokens in $\widehat{S}_{BCH}$ are estimated by minimizing Eq 5 and Eq 6. In doing so, we do not penalize the time or zip code fixed effects. In total across all subperiods and for both cities, we select approximately XXX of the 2,000 candidate tokens that explain price in $\widehat{S}_{LASSO}$ and XXX tokens that explain the agent-owned indicator in $\widehat{S}^{(-g)}_{LASSO}$. In total, there are X,XXX unique tokens from both of these sets in $\widehat{S}_{BCH}$.

By including the indicator variables for the tokens in the least-squares estimating equation, we are able to estimate implicit prices for each token in $\widehat{S}_{BCH}$. However, similar to others in the machine learning literature, such as Mullainathan and Spiess [2017], we refrain from a strict interpretation of these coefficient estimates as the true price associated with a given token. If anything, we favor an interpretation similar to the inverse regression approach in Taddy [2013] where the likelihood of the appearance of any given token in the remarks is determined by the true condition of the property. More importantly, removing the phrase *fixer upper* from a description while not making any repairs to the property is unlikely to increase its sales price.

For informational purposes the 20 positive and negative coefficients with the largest magnitudes are presented in Figure 1. For interpretation, we present results using the raw tokens in the data.[19] A significant number of negative tokens directly refer to a property

3 hours. Computation using multiplicatively separable census tract fixed effects was infeasible on this same machine.

[19] An alternative is to stem the words in which case *investors*, *investor*, *investment* are all set to *invest*.

in poor condition (*needs*, *investor*, *steal*) as well as tokens includes in euphamisms for poor condition (*fixer upper*, *handymans delight*, and *sweat equity*).

4.2 Agent-Owned Transactions

Results for Altanta are presented in Table 6. Panel A presents the results for the entire time period, 2008-2016. The first column includes the aforementioned standard property attributes and quarterly fixed effects. The agent-owned premium in this model is -2.1%. Column 2 adds zip code fixed effects and the agent-owned premium jumps to 5.5%. Model 3 allows each zip code to have its own quarterly price trend. The agent-owned premium in column 3 (5.5%) is comparable to the specification with homogenous price trends.

Although the estimates in columns 1-3 control for differences in the standard house characteristics, they do not control for the property's condition or quality. Column 4 includes the tokens in $\widehat{S}_{BCH}$ alongside the standard house characteristics and the additively separable quarter and zip code fixed effects used in column 2. In total, we use X,XXX=Y,YYY - ZZZ tokens.[20]. When these controls are included, the agent-owned premium declines to 0.7% and is no longer significant. Results for the heterogenous zip code price trends are reported in column 5. The agent-owned estimate is comparable to case when using additively separable quarter and zip code fixed effects.

Given the large number of tokens we include in columns 4-5, it is reasonable to ask if there is any information in the tokens we did not select. In order to answer this, we include XXX of the 2,000 candidate tokens that are in the complement of $\widehat{S}_{BCH}$ as regressors alongside the standard attributes in column 2. In doing so, we ask the question "does the Belloni et al. [2014] variable selection procedure select enough tokens?" The short answer is yes. The agent-owned premium reported in column 6 is 3.6%. Although the discount is less than the 5.5% reported in column 2, the estimate is hardly comparable to the estimate in columns 4 and 5 when the tokens in $\widehat{S}_{BCH}$ are included. Thus, the Belloni et al. [2014] procedure

[20]This is calculated as the difference in the number of variables included in the regression displayed in Table **??**

identifies a significant amount of pricing information represented in the MLS remarks.

Of course, the results in Belloni et al. [2014] are asymptotic and guaranteed to hold only approximately in a finite sample. Alternatively, a natural criticism is that by including many regressors we are overfitting the data in-sample and reporting a misleading agent-owned estimate. We assuage this critique using a permutation of the remarks. Specifically, we permute the remarks by randomly sampling the remarks without replacement and treat these remarks as the true remarks. We then create token indicators using the $\widehat{S}_{BCH}$ estimated using the true remarks. Results for this experiment are reported in column 7. The agent-owned premium reported in column 7 is nearly identical to the estimate reported in column 2. Thus, it does not appear as though the estimates in columns 4-5 are the result of overfitting. Instead, they are the product of the approach's ability to accurately identify the set of tokens that indicate the true condition and quality of the underlying property.

Panels B and C of Table 6 examine our method across two subperiods (bust and recovery) of the housing cycle in Atlanta. The subperiods were chosen based on a visual inspection of the price trends exhibited in Figure 2. The agent-owned premiums reported in columns 2 and 3 are greater during the bust period of 2008-2011, but remain insignificant once the qualitative information from the public remarks is added as a control in columns 4 and 5.

Similar results are reported for Phoenix in Table 7 for the entire period (2000 - 2013) and several subperiods. The subperiods were chosen based on a visual inspection of the price trends exhibited in Figure 3. They represent pre-boom, boom, bust, and recovery periods in Phoenix. The results in these subperiods are comparable to the results presented in Panel A for the full sample. Although the magnitude of the estimates are smaller for agent-owned premiums in Phoenix, the results are similar to Atlanta. Columns 2 and 3 of Panel A suggest that agents sell their own houses for a 1.0% to 1.4% premium. However, in column 5 when each zip code has its own price trend and the qualitative information from the remarks is included, the premium is no longer economically or statistically significant.

The results in Tables 6 and 7 suggest that the agent owned premiums reported in previous

studies reflect agents purchasing properties of differing quality and condition relative to the average property in the market. This is in contrast to the incentive problems discussed in Rutherford et al. [2005] and Levitt and Syverson [2008]. It is also interesting to note that Levitt and Syverson [2008] select and include indicators for 100 words and phrases in the MLS remarks. Whereas, the Belloni et al. [2014] variable selection procedure finds more than X,XXX words that are important when estimating the agent-owned premiums in our study. The difference in the number of tokens selected highlights the value of machine learning methods, rather than a limitation of the word list used in Levitt and Syverson [2008]. This is not surprising as other researchers have recognized that humans perform poorly when creating word lists from scratch, yet perform well when associating words topics King et al. [2017].[21]

4.3 Robustness Check

As a robustness check and to demonstrate the generalizability of our approach, we also examine vacant house price discounts. Rutherford et al. [2005] include an indicator variable for vacant houses that identifies when "the owner has already moved and thus needs to sell or that an investor is holding the house without a tenant." Although it is not the focus of their study, Rutherford et al. [2005] estimate that vacant houses sell for a 6% to 7% discount.[22] Research on the topic generally attributes the discount to (i) empty houses not showing as well or (ii) motivated sellers who have less bargaining power.

Although we do not doubt the sign and significance of the results reported in previous studies, we suspect that the magnitude of the results may be overestimated due to an omitted

[21]For example, the authors of this study were surprised that a large number of tokens in $\widehat{S}_{BCH}$ clearly indicate property condition ex-post, but would have not been identified as relevant for pricing ex-ante. As a notable example, the token *hates* was identified as a relevant token as real estate agents used variants of the phrase *homeowner hates to sell*. Despite the intentions implicit in this phrase, none of the authors in this study would have identified *hates* as a significant token ex-ante.

[22]Studies that focus primarily on estimating discounts related to vacant houses, such as Harding et al. [2003] and Turnbull and Zahirovic-Herbert [2011], report similar estimates. Levitt and Syverson [2008] do not identify vacant houses in their list of standard attributes or keywords.

variable bias. A similar concern was raised by Turnbull and Zahirovic-Herbert [2011] who note that "vacancy might also signal the presence of an unobservable factor that reduces buyer willingness to pay for the house. The notion here is that vacant houses have undesirable characteristics that are observed by sellers and buyers but are not reported in the data (condition, architecture, etc.)." The undesirable characteristics not only reduce the buyer's willingness to pay, but also contribute to why the property is vacant to begin with. Thus, if they are not properly controlled for the magnitude of the vacant house discount will be overestimated.

We examine whether the "undesirable characteristics" mentioned in Turnbull and Zahirovic-Herbert [2011] can be controlled for using the qualitative information in the remarks section. We estimate the vacant house discount using the same approach as the previous section, except for the use of an indicator for vacant houses in lieu of the indicator for agent-owned houses. The results for Atlanta and Phoenix are reported for the full study period and several subperiods in Tables **??** and **??**. ONE TABLE OF ESTIMATES INSTEAD OF TWO? DESCRIBE RESULTS

5 Conclusion

Real estate agents are experts who hold valuable information. They have the most local market and property specific knowledge, understand what buyers are willing to pay for, and know how to market real estate. Real estate agents use the public remarks section of the MLS to highlight important information, both positive and negative, that is not conveyed in other areas of the listing. If the qualitative information in the public remarks is not included in a model and it is correlated with other variables that are included in the model, then the model will over- or underestimate the effect of the variables that are included.

This paper provides a framework to address the omitted variable bias that plagues most real estate studies. Although we apply the framework to the estimation of agent-owned

premiums, it can be applied in any real estate study that attempts to isolate an effect on house prices. We show that the qualitative information in the remarks section of a MLS listing can be used to mitigate the omitted variable bias associated with agent-owned premiums. Using MLS data from Atlanta, Georgia and Phoenix, Arizona we replicate the findings in Rutherford et al. [2005] and Levitt and Syverson [2008]. The naive estimates suggest that agents sell their own houses for 2% to 6% more than their clients' houses. However, after we incorporate the qualitative information from the public remarks section of the MLS, the agent-owned premiums dissipate. The results suggest that the market distortions reported in previous studies are nonextant.

To demonstrate the generalizability of the approach we also estimate price discounts associated with vacant houses. Similar to previous research, naive estimates suggest that houses sell for X% less when they are vacant. However, after we incorporate the textual information the discount drops to X%. The significance of the approach we present is further augmented by the fact that is applicable to non-real estate assets that are susceptible to an omitted variable bias. For example, there is a rich literature on adverse selection and information asymmetry in the used car sales market (see, for example, Bond [1982] and Genesove [1993]) where the textual description of the used car may include important information, such as damage to the car or title issues, that should be incorporated.

References

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

Steven D Levitt and Chad Syverson. 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, 2008.

Patrick Bayer, Fernando Ferreira, and Robert McMillan. A unified framework for measuring preferences for schools and neighborhoods. *Journal of Political Economy*, 115(4):588–638, 2007.

Miles Young. Property Condition. *Southern Nevada Realtor Magazine*, September 2012. URL http://www.lasvegasrealtor.com/property-condition/.

Nancy E Wallace and Richard A Meese. The construction of residential housing price indices: a comparison of repeat-sales, hedonic-regression, and hybrid approaches. *The Journal of Real Estate Finance and Economics*, 14(1-2):51–73, 1997.

Itzhak Ben-David. Financial constraints and inflated home prices during the real estate boom. *American Economic Journal: Applied Economics*, 3(3):55–87, 2011.

Adam Nowak and Patrick Smith. Textual analysis in real estate. *Journal of Applied Econometrics*, 2017.

Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288, 1996.

Victor Chernozhukov, Christian Hansen, and Martin Spindler. Post-selection and post-regularization inference in linear models with many controls and instruments. *The American Economic Review*, 105(5):486–490, 2015.

Alexandre Belloni, Victor Chernozhukov, and Christian Hansen. Inference on treatment effects after selection among high-dimensional controls. *The Review of Economic Studies*, 81(2):608–650, 2014.

Sendhil Mullainathan and Jann Spiess. Machine learning: an applied econometric approach. *Journal of Economic Perspectives*, 31(2):87–106, 2017.

Matt Taddy. Multinomial inverse regression for text analysis. *Journal of the American Statistical Association*, 108(503):755–770, 2013.

Gary King, Patrick Lam, and Margaret E Roberts. Computer-assisted keyword and document set discovery from unstructured text. *American Journal of Political Science*, 2017.

John P Harding, John R Knight, and CF Sirmans. Estimating bargaining effects in hedonic models: Evidence from the housing market. *Real estate economics*, 31(4):601–622, 2003.

Geoffrey K Turnbull and Velma Zahirovic-Herbert. Why do vacant houses sell for less: holding costs, bargaining power or stigma? *Real Estate Economics*, 39(1):19–43, 2011.

Eric W Bond. A direct test of the" lemons" model: The market for used pickup trucks. *The American Economic Review*, 72(4):836–840, 1982.

David Genesove. Adverse selection in the wholesale used car market. *Journal of Political Economy*, 101(4):644–665, 1993.

Tables and Figures

Table 1: Sample MLS Listing Zip Code 30043

Zip Code	Beds	Baths	Sale Date	Sale Price	Remarks
30043	3	2	6/7/13	$270,000	back on market!!! located in tranquil neighborhood with sought-after schools close to shopping and i-85. this 3 bedroom 2 bath home is beautifully decorated. new roof was installed 3/20/14. marble master bath is stunning. room for expansion upstairs
30043	3	2	6/16/13	$168,900	wonderful updated one level with vaulted great room w fireplace & gas logs, formal dining room, kitchen with corian, newer stove & microwave, breakfast area overlooks wooded backyard, master bedroom suite w/upgraded master bath with tiled shower & jetted tub
30043	3	2	6/17/13	$150,000	great new listing on 18th fairway of collins hill golf course**on cul de sac too**no hoa**not a short sale and not bank owned**pride of ownership here**new double pane windows**new roof**updated heat and air***gourmet kitchen with double gas oven**ss fridge
30043	3	2	5/1/13	$113,500	adorable fannie mae homepath ranch style home updated and like new with new kitchen appliances, freshly painted, new carpet. large open living room with vaulted ceiling and fireplace, kitchen is spacious with breakfast area, nice master bathroom with tub shower
30043	3	2	6/16/13	$109,000	4 sided brick ranch with full basement. quick access to i85,316,mall of ga.large family room w/fireplace, separate living room and dining room, kitchen w/eat in b'fast room, laundry room, two car carport, deck on back. huge fenced in backyard for kids.
30043	3	2	4/1/13	$96,000	cute 3 bed 2 bath 2-story home in cul-de-sac. great schools & great location. private fenced backyard. needs carpet & paint. short sale. hurry before it's gone. sold as is no repairs..
30043	3	2	4/1/13	$93,000	nice ranch-style home on level,wooded,fenced corner lot!vaulted,sun-filled great room with dining area with wood-laminate floors!master bedroom has full,private bath.single car carport & charming front porch. back yard has large walk-in shed. excellent.
30043	3	2	5/8/13	$86,125	3 bdr 2bth split level home that has tons of potential. great opportunity for investor or first time buyer willing to put in some sweat equity. great location close to shopping and sought after peachtree ridge high school.

Table 2: Atlanta Descriptive Statistics

Statistic	Min	Pctl(25)	Mean	Median	Pctl(75)	Max
Sale Price ($1,000s)	30.0	113.4	204.8	161.7	249.0	3,000.0
Square Feet	506	1,546	2,215.5	2,028	2,693	5,999
Bedrooms	1	3	3.6	4	4	6
Bathrooms	1	2	2.3	2	3	6
Year	2000	2,004	2007.9	2007	2012	2016
REO	0	0	0.165	0	0	1
AGENT	0	0	0.009	0	0	1

NOTE: Descriptive statistics for Atlanta, GA.

SOURCE: Georgia MLS and authors' calculations.

Table 3: Phoenix Descriptive Statistics

Statistic	Min	Pctl(25)	Mean	Median	Pctl(75)	Max
Sale Price ($1,000s)	11.0	125.9	233.7	182.0	272.4	8,800.0
Square Feet	502.000	1,435.0	1,960.5	1,776.0	2,277.0	5,999.0
Bedrooms	1	3	3.363	3	4	6
Bathrooms	0.0	2.0	2.2	2.0	2.5	6.0
Year	2000	2003	2,006.4	2006	2010	2013
REO	0	0	0.192	0	0	1
AGENT	0	0	0.059	0	0	1

NOTE: Descriptive statistics for Phoenix, AZ.

SOURCE: Arizona MLS and authors' calculations.

Table 4: Atlanta REO

Variable	Mean	Basic	Tokens
baths2	0.619	0.224	0.167
baths3	0.203	0.319	0.254
baths4	0.062	0.528	0.420
baths5	0.019	0.677	0.545
baths6	0.003	0.851	0.669
beds2	0.039	0.121	0.097
beds3	0.447	0.148	0.145
beds4	0.362	0.199	0.197
beds5	0.127	0.211	0.214
beds6	0.025	0.169	0.188
sqft500	0.023	-0.104	-0.082
sqft1500	0.264	0.139	0.116
sqft2000	0.202	0.303	0.251
sqft2500	0.140	0.459	0.377
sqft3000	0.079	0.578	0.475
sqft3500	0.043	0.673	0.555
sqft4000	0.024	0.751	0.623
sqft4500	0.014	0.832	0.687
sqft5000	0.008	0.894	0.740
sqft5500	0.005	0.979	0.815

NOTE: Descriptive statistics and implicit prices for hedonic indicator variables in Atlanta, GA. The variables baths2,...,beds6 are indicator variables for the number of bathrooms and bedrooms of the property. The variable sqft500 is an indicator if the property has square footage greater than or equal to 500sqft and less than 1,000sqft. The remaining indicators for square footage are created similarly. All implicit prices are relative to a 1 bed, 1 bath property with 1,000-1,500 square footage.

SOURCE: Georgia MLS and authors' calculations.

Table 5: Phoenix REO

Variable	Mean	Basic	Tokens
baths0.5	0	-0.013	-0.039
baths1	0.04	-0.058	-0.048
baths1.5	0.114	0.069	0.057
baths2	0.524	0.085	0.071
baths2.5	0.152	0.17	0.132
baths3	0.114	0.192	0.149
baths3.5	0.03	0.253	0.195
baths4	0.012	0.287	0.219
baths4.5	0.008	0.36	0.27
baths5	0.002	0.38	0.277
baths5.5	0.001	0.434	0.323
baths6	0.001	0.438	0.317
beds2	0.105	0.161	0.165
beds3	0.504	0.135	0.168
beds4	0.317	0.112	0.162
beds5	0.066	0.052	0.129
beds6	0.007	-0.031	0.075
levels	1.218	-0.18	-0.116
sqft500	0.026	-0.213	-0.177
sqft1500	0.334	0.218	0.179
sqft2000	0.19	0.457	0.373
sqft2500	0.085	0.671	0.552
sqft3000	0.05	0.84	0.695
sqft3500	0.026	0.986	0.82
sqft4000	0.015	1.117	0.931
sqft4500	0.006	1.274	1.064
sqft5000	0.003	1.427	1.19
sqft5500	0.002	1.567	1.31

NOTE: Descriptive statistics and implicit prices for hedonic indicator variables in Phoenix, AZ. The variables baths2,...,beds6 are indicator variables for the number of bathrooms and bedrooms of the property. The variable sqft500 is an indicator if the property has square footage greater than or equal to 500sqft and less than 1,000sqft. The remaining indicators for square footage are created similarly. All implicit prices are relative to a 1 bed, 1 bath property with 1,000-1,500 square footage.

SOURCE: Arizona MLS and authors' calculations.

Table 6: Atlanta Owner Agent

PANEL A: 2008-2016							
	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7
AGENT	-0.021^{*}	0.055^{***}	0.055^{***}	0.007	0.008	0.036^{***}	0.056^{***}
	(0.010)	(0.006)	(0.005)	(0.005)	(0.004)	(0.005)	(0.005)
Num. obs.	154851	154851	154851	154851	154851	154851	154851
R^2	0.49	0.83	0.84	0.879	0.887	0.841	0.831
$\hat{\sigma}$	0.514	0.297	0.287	0.25	0.242	0.286	0.296
K	71	180	3766	1269	4855	1003	1269
PANEL B: 2008-2011							
	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7
AGENT	0.064^{***}	0.088^{***}	0.082^{***}	0.015	0.013	0.062^{***}	0.087^{***}
	(0.017)	(0.010)	(0.010)	(0.008)	(0.008)	(0.009)	(0.010)
Num. obs.	46273	46273	46273	46273	46273	46273	46273
$\hat{\sigma}$	0.464	0.292	0.282	0.246	0.237	0.28	0.29
K	49	158	1717	850	2409	473	850
PANEL C: 2012-2016							
	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7
AGENT	-0.059^{***}	0.048^{***}	0.046^{***}	0.006	0.006	0.031^{***}	0.048^{***}
	(0.011)	(0.007)	(0.006)	(0.005)	(0.005)	(0.006)	(0.007)
Num. obs.	108578	108578	108578	108578	108578	108578	108578
R^2	0.488	0.839	0.845	0.886	0.891	0.85	0.84
$\hat{\sigma}$	0.526	0.295	0.289	0.248	0.243	0.285	0.294
K	55	164	2082	1181	3099	895	1181
FE	Quarter	Quarter+Zip	Quarter×Zip	Quarter+Zip	Quarter×Zip	Quarter+Zip	Quarter+Zip
Tokens				$\hat{S}_{BCH}$	$\hat{S}_{BCH}$	$\hat{S}^{c}_{BCH}$	Permutation

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

NOTE: Models 1-3 do not use any tokens. Models 4 and 5 use tokens in the set $\hat{S}_{BCH}$. Model 6 uses the set of 2,000 most frequent tokens not in $\hat{S}_{BCH}$. Model 6 uses the tokens in $\hat{S}_{BCH}$ but permutes the remarks. All models include the age of the property and indicators for square footage, bedrooms, bathrooms, and levels. All standard errors are two-way clustered at the quarter and zip code level. A further description of the controls is provided in 4.

SOURCE: Atlanta MLS and authors' calculations.

Table 7: Phoenix Owner Agent

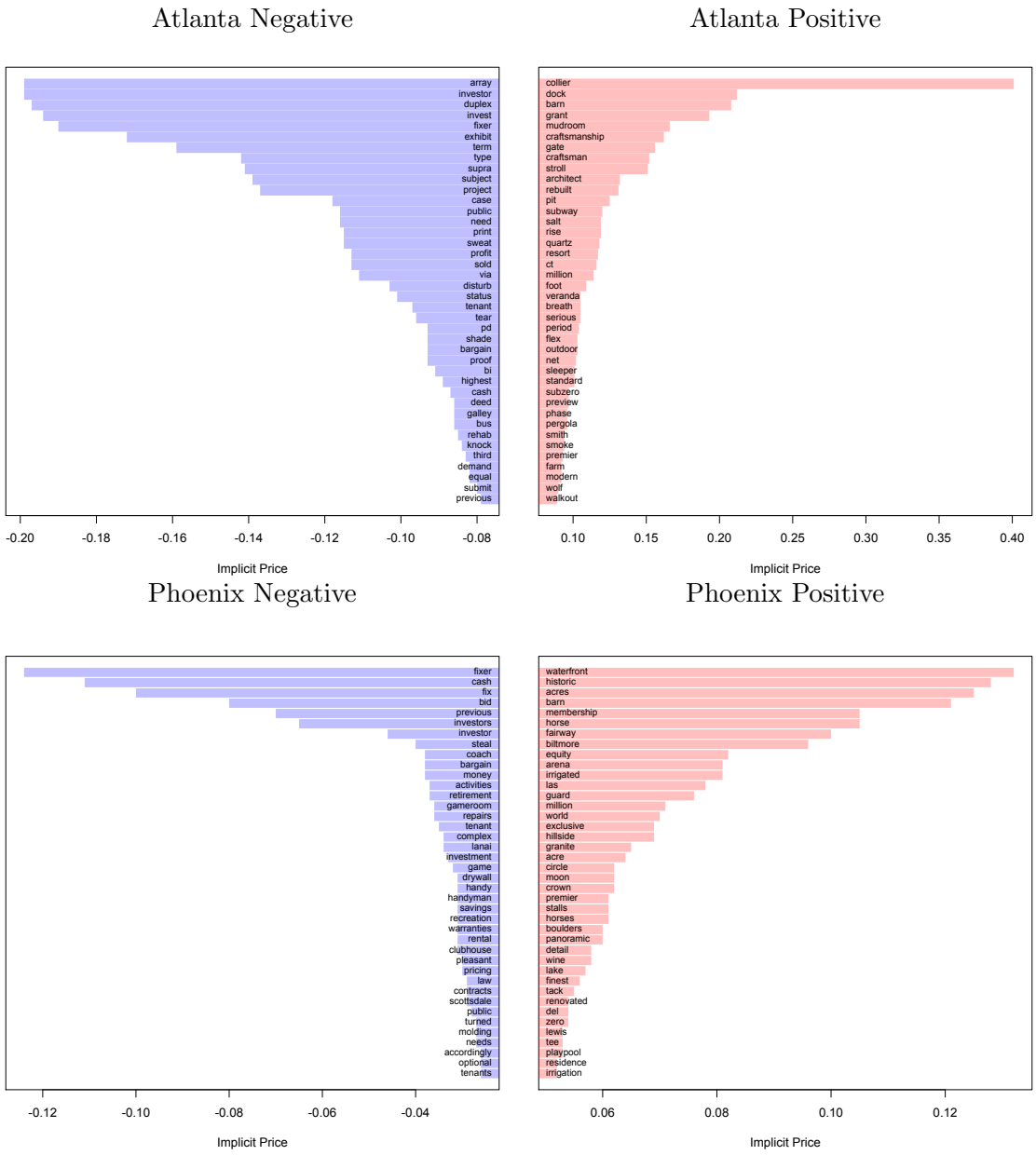
PANEL A: 2000-2013							
	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7
AGENT	0.023***	0.010***	0.014***	−0.003**	0.000	0.000	0.010***
	(0.002)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
Num. obs.	617616	617616	617616	617616	617616	617616	617616
$\hat{\sigma}$	0.298	0.213	0.196	0.18	0.166	0.203	0.212
K	85	232	7349	1251	8368	1002	1251
PANEL B: 2000-2003							
	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7
AGENT	0.036***	0.014***	0.014***	0.005**	0.006**	0.011***	0.014***
	(0.003)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)
Num. obs.	186212	186212	186212	186212	186212	186212	186212
$\hat{\sigma}$	0.244	0.184	0.182	0.16	0.159	0.178	0.183
K	45	160	1850	1143	2833	904	1143
PANEL C: 2004-2006							
	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7
AGENT	0.045***	0.020***	0.020***	0.009***	0.009***	0.015***	0.020***
	(0.003)	(0.002)	(0.002)	(0.002)	(0.001)	(0.002)	(0.002)
Num. obs.	209728	209728	209728	209728	209728	209728	209728
$\hat{\sigma}$	0.25	0.174	0.17	0.149	0.145	0.168	0.174
K	40	164	1514	1169	2519	924	1169
PANEL D: 2007-2009							
	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7
AGENT	0.031***	0.019***	0.021***	−0.008*	−0.006	0.001	0.019***
	(0.006)	(0.004)	(0.004)	(0.003)	(0.003)	(0.004)	(0.004)
Num. obs.	85497	85497	85497	85497	85497	85497	85497
$\hat{\sigma}$	0.345	0.233	0.216	0.197	0.176	0.217	0.232
K	41	170	1564	814	2485	906	814
PANEL E: 20010-2013							
	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7
AGENT	0.006	0.007**	0.009***	−0.004*	−0.003	−0.006**	0.007**
	(0.004)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)
Num. obs.	134156	134156	134156	134156	134156	134156	134156
$\hat{\sigma}$	0.37	0.232	0.225	0.189	0.183	0.217	0.231
K	44	177	2154	1128	3105	810	1128
FE	Quarter	Quarter+Zip	Quarter×Zip	Quarter+Zip	Quarter×Zip	Quarter+Zip	Quarter+Zip
Tokens				$\hat{S}_{BCH}$	$\hat{S}_{BCH}$	$\hat{S}^c_{BCH}$	$\hat{S}_{BCH}$ Permutation

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

NOTE: Models 1-3 do not use any tokens. Models 4 and 5 use tokens in the set $\hat{S}_{BCH}$. Model 6 uses the set of 2,000 most frequent tokens not in $\hat{S}_{BCH}$. Model 6 uses the tokens in $\hat{S}_{BCH}$ but permutes the remarks. All models include the age of the property and indicators for square footage, bedrooms, bathrooms, and levels. A further description of the controls is provided in 5.

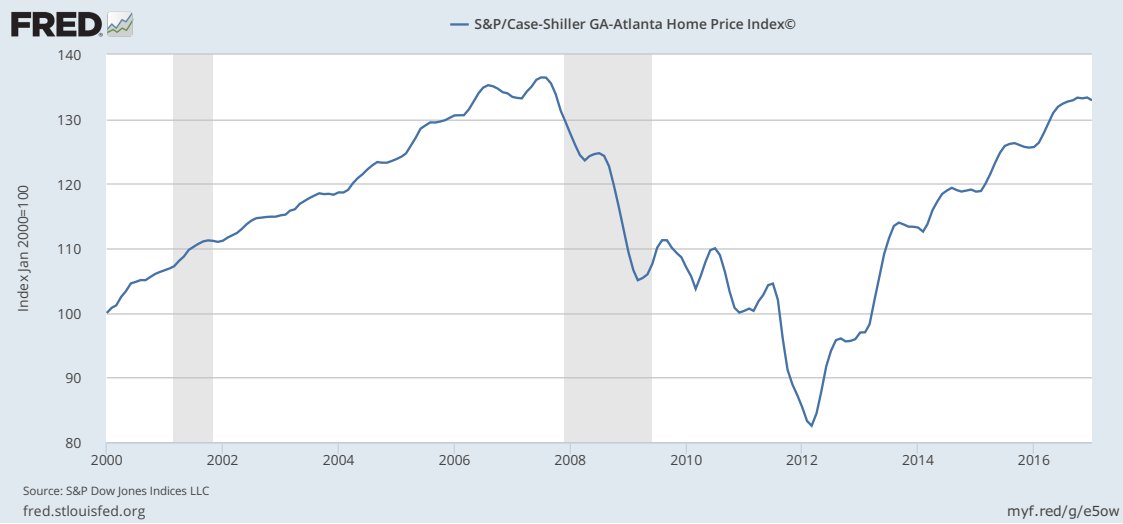
SOURCE: Arizona MLS and authors' calculations.

Figure 1: Implicit Prices for Tokens



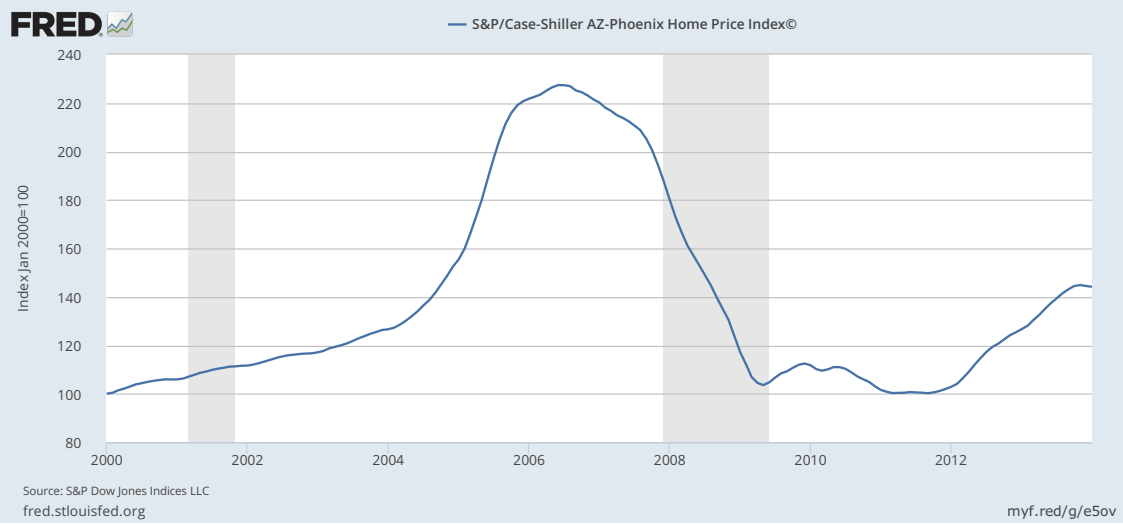
Note: This figure plots the implicit prices for the tokens in both the Phoenix and Atlanta data. The 20 positive and negative tokens with the largest magnitudes are displayed. *SOURCE*: Arizona MLS and authors' calculations.

Figure 2: Atlanta Case-Shiller Repeat Sales Index



SOURCE: St. Louis Federal Reserve Economic Data, *https://fred.stlouisfed.org/series/ATXRNSA*

Figure 3: Phoenix Case-Shiller Repeat Sales Index



SOURCE: St. Louis Federal Reserve Economic Data, *https://fred.stlouisfed.org/series/PHXRNSA*