Department of Economics

Working Paper Series

Textual Analysis in Real Estate

Adam Nowak and Patrick Smith

Working Paper No. 15-34

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

http://be.wvu.edu/phd_economics/working-papers.htm

Textual Analysis in Real Estate

Adam Nowak[1] Patrick Smith [2]

West Virginia University Georgia State University

August 7, 2015

[1]College of Business & Economics, 1601 University Ave., PO Box 6025, Morgantown, WV 26506-6025, USA; Email: adam.d.nowak@gmail.com.

[2]J. Mack Robinson College of Business, 35 Broad St, Atlanta, GA 30303, USA; Email: psmith44@gsu.edu

Abstract

This paper incorporates text data from MLS listings from Atlanta, GA into a hedonic pricing model. Text is found to decrease pricing error by more than 25%. Information from text is incorporated into a linear model using a tokenization approach. By doing so, the implicit prices for various words and phrases are estimated. The estimation focuses on simultaneous variable selection and estimation for linear models in the presence of a large number of variables. The LASSO procedure and variants are shown to outperform least-squares in out-of-sample testing.

JEL Codes: C01, C18, C51, C52, C55, C65, R30.

Key words: textual analysis, big data, real estate valuation

1 Introduction

Real estate is one of the most studied asset classes - and for good reason. Some of the more prominent features of real estate include its incredible market value, the large share of real estate in individual investors portfolios, and the value of mortgages tied to real estate. Even when focusing solely on households, the numbers are staggering. In 2014, household real estate assets were valued at $23.5 triliion USD making up 24.2% of total household assets. Home mortgages on the household balance sheet were $9.4 trillion or 66.2% of total household liabilities [1]. For these reasons alone, researchers, policy makers, investors, homeowners, and bankers all have a significant interest in accurately valuing real estate. Accurately valuing real estate in a hedonic model requires collecting a rich set of property attributes. This study describes methods to create such a set of property attributes from text supplied by listing agents.

Valuation models for real estate can be derived using comparable sales, repeat sales, discounted cash flows, or other means. This study uses a hedonic model where the price of a property is expressed as a linear combination of its attributes.[2] We argue that a useful hedonic model produces coefficients that are easily interpreted and provide pricing accuracy. The contributions of this paper are both methodological and empirical. The methodology described in this paper 1) applies textual analysis methods to real estate listings using a token approach and 2) describes an estimation procedure that yields interpretable results. Empirically, the study finds 1) listing descriptions provided by listing agents contain information that can be used to decrease pricing error when used in conjunction with standard, housing attributes in a hedonic pricing model, 2) the procedure we describe outperforms a least-squares alternative in out-of-sample testing, and 3) a theoretically based tuning parameter outperforms a cross-validated tuning parameter in out-of-sample testing.

The estimation technique described in this paper combines two branches of statistical research: textual analysis and sparse modeling. Textual analysis is a term for techniques that map text - news articles, 10-k's, message board postings, litigation releases, etc. - into variables that can be used in statistical applications including hypothesis testing, prediction, and filtering. This study uses an approach whereby each remark can be expressed as a collection of words or phrases; each word or

[1] http://www.federalreserve.gov/releases/z1/Current/z1.pdf, Table B.101

[2] Early uses of the hedonic model include Rosen (1974). Two helpful literature reviews include Malpezzi (2003) and Kang and Reichert (1991)

phrase is defined as a *token*. Tokens in the remarks can proxy for actual features of the property, seller or neighborhood. We are interested in selecting which tokens are relevant and the implicit prices for the features that they represent. In order to do so, indicator variables for tokens are included along with standard attribute variables in a linear, hedonic model. Because the number of tokens can increase with the total number of observations, the number of indicator variables for the tokens can be large. In such high-dimensional settings, least-squares estimation is prone to over-fit the data producing poor out-of-sample performance. Thus, estimating the parameters requires techniques that are designed for large dimensional parameter spaces.

One approach to high-dimensional data is to transform the data using data reduction methods. Data reduction techniques implicitly or explicitly assume that a large number of variables can be expressed using a much smaller set of observed or unobserved variables. One popular method for dimension reduction is principal components analysis (PCA). PCA creates principal components using linear combinations of a much larger set of variables from a multivariate data set. These principal components can then be used in lieu of or alongside other variables in a regression framework. In the hedonic model, principal components from the covariance matrix of the token indicator variables could be used by themselves or included alongside commonly used hedonic variables such as bedrooms, square footage, etc. In either case, interpreting the coefficients on the principal components requires the researcher to first interpret the principal components. Because all variables have non-zero loadings in each principal component, interpreting the principal components can be a challenge.

An alternative approach to dimension reduction is to assume that the true model is generated by a subset of explanatory variables. Alternatively, the true model for house prices can be sufficiently approximated using only a subset of the available regressors, or that the coefficient vector has some elements equal to 0. In this situation, the coefficient vector is said to be *sparse*. Estimating which coefficients are non-zero is variable selection. Traditional approaches such as AIC and BIC criteria requires estimating all combinations of models. Given the large number of tokens and the combinatorial nature of this approach, this is computationally prohibitive.

The least absolute shrinkage and selection operator (LASSO) described in Tibshirani (1996) provides a feasible alternative. LASSO simultaneously selects both a subset of relevant variables and the coefficient values. Due to a penalty function, the coefficients are biased towards 0 but

are still consistent. Given the large number of observations we have available to us, these biased but consistent coefficeints can improve out-of-sample performance. In short, LASSO 1) identifies which tokens are important, 2) provides easily interpreted coefficients, and 3) and performs well in out-of-sample testing. Three features that are important when valuing real estate.

The remainder of the paper is organized as follows. Section 2 provides a literature review that emphasizes both sparse modeling and textual analysis. Section 3 describes the statistical techniques used, the details of the data source, and the results from the estimation. Section 4 provides a summary of the paper and outlines areas for further research.

2 Literature Review

This study models residential property prices using a hedonic model. An important feature of the hedonic model is that property attributes explicitly impact property prices. Quantitative, qualitative, geographic and municipal attributes have all been found to influence property prices. Brown and Pollakowski (1977), Bond et al. (2002), and Rouwendal et al. (2014) find water access and coastline significantly influence property prices. Benson et al. (1998), Paterson and Boyle (2002), Song and Knaap (2003) and Tu and Eppli (1999) find non-traditional attributes can play a significant role. The running theme in all of these studies is that predicting property prices can be improved by augmenting a simple hedonic pricing model (one that includes bedrooms, bathrooms, square footage, etc.) with non-standard attributes. Unfortunately, these non-standard attributes can be difficult or impossible for the researcher to measure. However, it is quite possible that these non-standard attributes are explicitly mentioned by listing agents in the remarks section of the listing. Hill et al. (1997) were one of the first studies to explicitly acknowledge that the remarks section in MLS data may contain hidden characteristics. When constructing their repeat sales model Hill et al. (1997) use the remarks section to ensure house characteristics remained the same between sales.

Despite the frequent use of MLS data in real estate research, there are a very few studies that examine and include the non-standard attributes available in the MLS remarks section. In the previous studies that utilize the MLS remarks section researchers have manually created indicator variables. Not only is this a time consuming process, but it is prone to human error. Haag et al.

(2000) were the first to include the non-standard attributes available in the MLS remarks section in a hedonic model. They identify a list of keywords and phrases that were prevalent in their dataset (1994-1997) to examine the motivation of the seller, location of the property, physical improvements, or property defects. In a recent follow-up study, Goodwin et al. (2014) extend the Haag et al. (2000) study by including additional keywords and categories. Goodwin et al. (2014) also cover a longer time period (2000-2009) that includes both an up and down real estate market. This is important because a study by Soyeh et al. (2014), that also utilizes the MLS remarks section, finds that incentives offered by sellers are not capitalized into sales price during soft market conditions.

Two approaches have been used when scoring or sorting text for use in financial and economics applications. The first approach pre-specifies positive and negative dictionaries of tokens and scores the text based on the relative frequencies of positive and negative tokens. Tetlock (2007), Loughran and McDonald (2011), and Bollen et al. (2011) find text from the *Wall Street Journal*, 10-k filings and Twitter are all associated with future stock price movements. Garcia (2013) finds that the predictive power of text increases in recessions. In one of the few real estate applications, Pryce and Oates (2008) use a pre-specified dictionary approach and find real estate agents alter their choice of words based on the macroeconomy. It is important to emphasize that the dictionary approach is suitable only when the researcher has ex-ante knowledge of relevant tokens for the application at hand. Loughran and McDonald (2011) emphasize this and show that a customized financial dictionary outperforms the general Harvard-IV-4 TagNeg dictionary when extracting information from 10-k's.

The second approach begins with a scored text and determine which tokens are more likely to increase a text's score. Using the US Congressional record, Gentzkow and Shapiro find tokens can be used to identify Republicans and Democrats. Taddy (2013a) performs a similar analysis using information from Twitter. Taddy (2013b) rigorously studies a token model in a sparse coefficient setting. Mitra and Gilbert (2014) also seeks a sparse coefficient solution when searching for tokens that are associated with successfully funded projects on the crowdfunding website Kickstarter. We follow most closely the last 3 studies and use the sale price of the property as a way to identify a sparse coefficient vector.

Given the large number of potential tokens that can appear, we require a procedure for selecting which tokens are important that will not overfit the data. As mentioned above, PCA is one way

to reduce the dimension of the tokens. Perhaps due to the difficulty in interpreting the results, incorporating PCA into a hedonic model has not been widely used. Kain and Quigley (1970) interpret principal components as measures of housing quality. Maclennan and Tu (1996), Bourassa et al. (1999), and Bourassa et al. (2003) do not explicitly use principal components in a hedonic model but rather use the principal components in order to identify submarkets.

Methods that use principal components as additional variables in a regression in order to improve accuracy similar to factor-augmenter vector autoregression models in the time series literature are a possible solution. In this way, the principal components increase the predictive accuracy of the model while allowing the researcher to easily interpret the coefficients on the remaining variables. However, coefficients on the factor are not easily interpreted unless the factors themselves are easily interpreted. Estimating this augmented hedonic model would (presumably) increase predictive accuracy and allow interpretation of the coefficients on the standard housing attributes but would not allow us to identify the price of any single token. Further complicating things are results in Johnstone and Lu (2009) that find PCA loadings can be inconsistent when the number of variables grows large.

Sparse modeling has been used in engineering, statistics and economics applications; theoretical results draw from all disciplines. The LASSO, proposed in Tibshirani (1996), estimates a linear model subject to a penalty on the l_1 norm of the coefficient vector. This penalty sets less important coefficients to 0, thereby selecting a subset of variables for prediction. In certain situations, the LASSO is able to correctly identify the true non-zero coefficients in deterministic and probabilistic settings, Candès et al. (2006) and Knight and Fu (2000). Knight and Fu (2000) proves consistency and asymptotic normality of the LASSO estimates for a fixed number of variables and a large number of observations. Meinshausen and Bühlmann (2006) and Zhao and Yu (2006) demonstrate the LASSO estimator correctly identifies the true non-zero coefficients with a large number of variables provided the regressors are not too correlated.

An important choice for the researcher when using the LASSO is the parameterization of the penalty function. On approach is to use M-fold cross-validation, Varian (2014). This technique is easy to perform, has been shown to perform well in Monte Carlo studies, but has also been shown to select too many irrelevant variables, Leng et al. (2006), Zou et al. (2007), and Feng and Yu (2013). Alternatively, penalty parameters can be based on theoretical results required

for consistency, Knight and Fu (2000) and Bickel et al. (2009). Belloni and Chernozhukov (2010) provide a feasible procedure for such an estimation. Yet another approach is to use a square root loss function described in Belloni et al. (2011) which results in a parameter free penalty function. In order to more directly compare our results to a least-square procedure, we keep a squared loss function and use both the M-fold cross-validation procedure and the procedure in Belloni et al. (2011).

The use of LASSO and other penalized procedures becomes a valuable tool when the researcher is faced with a large number of regressors and would like to perform a least-squares estimation. Examples of regression models with a potentially large number of regressors include cross country growth regressions [3], wage equations [4], returns to schooling, [5] and forecasting [6]. In such situations, the researcher must choose which variables to use based on a behavioral model, anecdotal evidence or other results in the literature. The methods discussed in this paper are applicable to these and other applications in finance and economics when the researcher must select relevant variables in a linear model without any such guidance.

3 Modelling and Econometric Analysis

3.1 Penalized Regression

The data is an unbalanced panel. There are $i = 1, ..., I$ houses sold over time periods $t = 1, ..., T$ with some houses selling more than once for a total of N transactions. For each transaction $n = 1, ..., N$, the sale price is given by

$$p_{it} = x_{it}\delta + \epsilon_{it} \tag{1}$$

where p_{it} is the sale price, x_{it} is a $1 \times K$ vector possibly containing an intercept, annual time dummies, control variables, and an indicator for the tokens, δ is a $1 \times K$ vector of implicit prices for the variables and, and ϵ_{it} is an i.i.d $N(0, \sigma^2)$ capturing any variation in house prices not captured by the variables. Details for constructing the indicator variables are described below. The linearity

[3] Sala-i Martin (1997), Liu and Stengos (1999), Fernandez et al. (2001), Rodriguez and Rodrik (2001)
[4] Angrist and Krueger (1991), Buchinsky (1998), Machado and Mata (2005), Angrist et al. (2006), Stewart (2007)
[5] Card and Krueger (1992), Jensen (2010), Dagsvik et al. (2011), Ermisch and Francesconi (2013)
[6] Stock and Watson (2002), Clements and Galvão (2009), Carriero et al. (2011), Fuentes et al. (2014)

assumption in 1 is made for simplicity as the results below hold when the pricing function is not necessarilly linear, $p_{it} = f(x_{it}) + \epsilon_{it}$, and a vector of appropriate transformations $x^*_{it} = g(x_{it})$ are used to approximate f.

When δ is sparse, $Q < K$ elements are non-zero with the remaining $K - Q$ elements are exactly equal to 0. The index of the Q coefficients is denoted by $S \subset \{0, ...K\}$. The number of variables in the model preclude using AIC or BIC methods to determine S; with $1,000 < K$, there will be more than 2^K possible models to estimate. However, Benjamini and Hochberg (1995) and Benjamini and Yekutieli (2001) provide simple procedures for variable selection using p-values from a least-squares regression. Both procedures use an ex-ante false-discovery rate, ρ. After applying either procedure, the expected fraction of irrelevant regressors is equal to ρ. Given the possibly large number of relevant tokens in the model, even conservative values of ρ will result in predicted sale prices based on many irrelevant regressors. Previewing the results, using $\rho = 0.1$ selects more than 1,300 tokens as relevant from a possible 2,000. As a feasible alternative to estimating a sparse δ, we solve the following optimization problem

$$\min_d \frac{1}{2} \sum_{i,t} (p_{it} - x_{it} d)^2 + \lambda \sum_{k=1}^{K} |d_k| \tag{2}$$

Eq 2 is the sum of the sum of squared errors plus a penalty function for all coefficients except the intercept. The penalty function is proportional to the sum of the absolute values of the elements of d. The parameter λ is a tuning parameter or weight for the penalty function that controls the penalty for adding coefficients. When $\lambda = 0$, the objective function in Eq 2 is the least-squares objective function and the minimizer is the least-squares estimator. When $\lambda > 0$, the estimator is the LASSO in Lagrangian form. Define $\widehat{d}$ as the vector that minimizes Eq 2, $\widehat{Q}$ as the number of non-zero coefficients in $\widehat{d}$, and $\widehat{S}$ as the index of these non-zero coefficients in $\widehat{d}$. The least-squares estimator does not provide a sparse solution as almost surely all entries in $\widehat{d}$ are non-zero and $\widehat{Q} = K$.

Eq 2 cannot be solved by taking a derivative as the penalty function $\lambda \sum_{k=1}^{K} |d_k|$ is non-differentiable at $d = 0$. However, the problem can be recast into a Kuhn-Tucker maximization problem ensuring the solution is unique when $K < N$. Due to the penalty function in Eq 2, the estimate $\widehat{d}$ is biased towards when $0 < \lambda$. When $\lambda = 0$ the resulting least-squares coefficient esti-

mates are unbiased. However, due to the large number of variables, the variance of the least-squares coefficients can be large. LASSO makes a bias-variance trade off in order to decrease out-of-sample prediction error. Alternative penalty functions can be used to decrease the variance of the coefficients. When the penalty function uses the sum of squared coefficients, the resulting estimator is a ridge regression. The elastic net described in Zou and Hastie (2005) uses both the sum of the absolute value of the coefficients and the sum of the squared coefficients. However, only the LASSO and elastic net produce $\widehat{Q} < K$ thereby reducing the number of variables the researcher must collect for future forecasts. Results for ridge regression and elastic net are available in the appendix.

Consistency results for $\widehat{d}$ and $\widehat{S}$ rely on specifying the rate at which $\lambda \to \infty$ for $N \to \infty$ and the structure of the covariance matrix of x_{it}. Consistency for $\widehat{S}$ is related to model selection and results when the probability of selecting the true index approaches 1 as $N \to \infty$, Zhao and Yu (2006). In contrast, consistency for $\widehat{d}$ occurs when the probablity limit of $\widehat{d}$ is equal to δ. Consistency of one does not necessarily imply consistency of the other. For the purposes of out-of-sample testing, we focus on consistency of $\widehat{d}$. Although the LASSO estimator has a solution when $N < K$, in order to compate LASSO to least-squares, we focus on asymptotics for fixed K. Knight and Fu (2000) find $\widehat{d}$ is consistent for fixed K if $\lambda = o(N)$. Results for consistency when K grows with N are given in Zhao and Yu (2006) for the linear model and in Bickel et al. (2009) for general functional form. If σ were known, the optimal penalty value is $\lambda = 2\sigma\sqrt{2\ln(2KN)/N}$. Because σ is not known, Belloni and Chernozhukov (2010) provide a procedure for a feasible estimation of λ. First, the researcher chooses $c > 1$, $\alpha = o(N)$, a threshold level η (or a maximum number of steps S), and an initial condition $\widehat{\sigma}^{s=0}$ equal to the sample standard deviation of p_{it}.

The function $\Lambda(1-\alpha|X)$ is equal to a data-dependent quantile based on the matrix of regressors. This function is equivalent in distribution to

$$\Lambda(1-\alpha|X) =_d \max_{1\leq k\leq K} |NE_n[x_{nk}g_n]|, g_n \ N(0,1) \tag{3}$$

where x_{nk} is the value of variable k for transaction n. Eq 3 provides a means to estimate $\Lambda(1-\alpha|X)$: simulate g_n, calculate the maximum over all k, and repeat. The estimate of $\Lambda(1-\alpha|X)$ is the $1-\alpha$ quantile of these simulated values. With this estimated value of $\Lambda(1-\alpha|X)$, a feasible estimate

of λ is found using the following procedure. Using superscripts to indicate a particular iteration $s = 1, ..., S$ the procedure is

1. Set $\lambda^s = 2c\widehat{\sigma}^{s-1}\Lambda(1-\alpha|X)/N$.
2. Estimate Eq 2 using λ^s and store $\widehat{d}^s$ and $\widehat{Q}^s$.
3. Update $\widehat{\sigma}^s$ as $\sqrt{\frac{N}{N-\widehat{Q}^s}\sum_i \left(p_i - x_i\widehat{d}^s\right)^2}$.
4. If $\eta < \left|\widehat{\sigma}^s - \widehat{\sigma}^{s-1}\right|$ or $s = S$, stop.
5. Otherwise, increase s to $s+1$ and repeat steps 1-4.

Following Monte Carlo simulations given in Belloni and Chernozhukov (2010), we set $c = 1.1$ and $\alpha = 0.1$. The above procedure results in a consistent estimate of σ that can be used to create a feasible optimal penalty level $\lambda_F = 2\widehat{\sigma}\sqrt{2\ln(2K/\alpha)}/N$. Belloni and Chernozhukov (2010) recommend the data-dependent penalty level $\lambda_{DD} = 2c\widehat{\sigma}^{s-1}\Lambda(1-\alpha|X)/N$ be used as it adapts to the regressors and is less conservative in that $\lambda_{DD} \leq \lambda_F$. Either penalty level can be used to consistently estimate $\widehat{d}$.

An alternative procedure to estimate λ is to use an M-fold cross-validation procedure. The M-fold procedure uses subset of the observations in order to estimate $\widehat{d}$ and uses the remaining observations to determine out-of-sample performance. This process is repeated M times and the out-of-sample performance is averaged over all M trials. The M-fold value of λ is the value that provides the best out-of-sample performance. The optimal value for the penalty parameter, λ_M, is calculated using the following procedure

1. Sort all the observations into groups $m = 1, ..., M$ groups with an equal number of observations in each group.
2. For a given m and λ, estimate Eq 2 using all observations not in subset m. Store the coefficients as $\widehat{d(\lambda)}^m$.
3. Estimate sale prices for sales in subset m using $\widehat{d(\lambda)}^m$ and store the total sum of squared errors (SSE) for group m as $SSE(\lambda)^m = \sum(p_i - z_i\widehat{d(\lambda)}^m)^2$
4. The average SSE for this choice of λ is then $SSE(\lambda) = \frac{1}{M}\sum SSE(\lambda)^m$

5. The value of λ that minimizes $SSE(\lambda)$ is the M-fold cross-validated, λ_M

As mentioned above, the LASSO procedure results in biased coefficient estimates. The post-LASSO procedure in Belloni and Chernozhukov (2010) is a two-stage procedure that mitigates the bias introduced through LASSO estimation. In the first stage, LASSO is used to select $\widehat{S}$, In the second stage, the $\widehat{Q}$ variables in $\widehat{S}$ are used in a least-squares regression. The post-LASSO procedure is outlined in the following 3 steps

1. For a given λ, estimate Eq 2.
2. Create a $1 \times \widehat{Q}$ vector x_i^{PL} by removing the $K - \widehat{Q}$ variables in x_i that are not in $\widehat{S}$.
3. Set $\lambda = 0$ and estimate Eq 2 using only the variables in x_i^{PL}.

By setting $\lambda = 0$, the second stage estimation procedure becomes least-squares. The value of λ can be estimated using the iterative procedure in Belloni and Chernozhukov (2010). The intuition for this procedure is the following: if LASSO correctly estimates S in the first stage, the model in the second stage is correctly specified, and the resulting least-squares coefficients are unbiased. In practice it is possible that $S \neq \widehat{S}$ and the second stage model is mis-specified. Despite this, Belloni and Chernozhukov (2010) find the variables erroneously included or excluded will have little explanatory power.

3.2 Alternative Pricing Models

We use 5 models as a means to compare the relative and supplemental predictive power contained in the remarks.

$$[BASELINE] : p_i = \alpha_t + x_i\beta + \epsilon_i \tag{4}$$

$$[U1] : p_i = \alpha_t + v_i\theta + \epsilon_i \tag{5}$$

$$[B1] : p_i = \alpha_t + w_i\phi + \epsilon_i \tag{6}$$

$$[U2] : p_i = \alpha_t + x_i\beta + v_i\theta + \epsilon_i \tag{7}$$

$$[B2] : p_i = \alpha_t + x_i\beta + w_i\phi + \epsilon_i \tag{8}$$

Here, α_t is a time fixed-effect for time period t, x_i is a vector controling for bedrooms, bathrooms and an intercept, v_i is a vector if indicator of variables for the unigrams, and w_i is a vector of indicator variables for the bigrams. Construction of the unigram and bigram vectors is described in the following section. The vector β contains the relative prices for bedrooms and bathrooms. The vectors θ and ϕ are the implicit prices of the tokens and are assumed sparse with some elements equal to 0. Finally, ϵ_i is an iid, normally distributed error term $N(0, \sigma^2)$. It is fully acknowledged that ϵ_i includes the effect of any unobserved variable not mentioned in the remarks that can be related to the property attributes or the nature of the transaction.

Eq 4 is the baseline model that only controls for time, bedrooms and bathrooms. We refer to variables in x_i as control variables. After experimenting with several configurations for the control variables, we found that using indicator functions for the number of bedrooms and bathrooms performed produced R^2 values that were comparable to R^2 values from other specifications of the control variables including continuous variables for age, bedrooms and square footage as well as interraction terms between these variables. In our estimation procedure, we use a total of 11-15 indicator variables depending on various subsets of the data. We make no claim as to the unbiasedness of the estimates for β but note that the explanatory power of these indicator variables is comparable to the explanatory power of alternative models.

Eq 5 and Eq 6 regress price on tokens in the absence of any control variables. These models are used to assess if information in remarks can substitute for the controls. The performance of these models are of practical importance as important tokens in these models can be used to guide practitioners, assessors and future research in data collection. Eq 7 and Eq 8 are constructed in order to highlight the supplemental information tokens can provide. More elaborate interactions between tokens and control variables are possible but are beyond the focus of this paper.

We also use $ln(p_{it})$. We remain agnostic as to whether the correct model is log or level price as our model specification questions related to variable selection and not the correct power transformation of the dependent variable. Both log and level hedonic models are used in the real estate literature When using log price, the presence of a token will approximately increase the sale price by a given percentage. For example, houses in a *gated community* are 30% more expensive than houses not in a *gated community*.

Further, we also apply our approach to another poplar pricing estimator in the real estate

literature. The *repeat sales regression* regresses differenced sale prices on differenced right-hand side variables. For consecutive sales of the same house i sold at times s and $s \leq t$, the change in price, $\Delta p_{it} = p_{it} - p_{is}$, is given differencing 1

$$\Delta p_{it} = \Delta x_{it}\delta + \Delta \epsilon_{it} \tag{9}$$

Here, Δx_{it} is the difference in right-hand side variables. When x_{it} contains only time-invariant variables and time period fixed effects, Δx_{it} contains 0s, a $+1$ for the time period t variable and a -1 for the time period s variable. Time-invariant variables are treated as nuisance parameters. Such time-invariant variables include location effects and possibly structural effects when quality does not change. Implicit in the repeat sales regression is the assumption that the quality of the underlying property does not change. With this assumption, the coefficients on the time effects are interpreted as a constant-quality price index.

Remarks associated with two different sales of the same house are almost surely different although certain features of the underlying property are time-invariant. When we include tokens in x_{it}, the effects of time-invariant tokens are differenced away. However, certain relevant features of the property are both time-variant and indicated in the remarks. For example, renovating a property would presumably increase the sale price; properties that are recently renovated would have larger changes in prices than non-renovated properties. If macroeconomic factors lead to citywide renovation, the time coefficients are biased and no longer result in a constant-quality index.

The advantages of including tokens in the repeat sales regression are three-fold. First, tokens can be used to mitigate bias in the price index by controlling for time-varying changes in quality. Second, prices of individual tokens can be used to estimate price differential based on listing agent assessments of quality. Third, when included alongside indicator variables for auctions, foreclosures or other events most likely associated with changes in quality, we can obtain unbiased coefficients in the presence of both time-varying and time-invariant. Mayer (1998) uses a repeat sales approach to estimate auction premiums that controls for unobserved time-invariant. Because time-varying controls are not available, the auction premium in Mayer (1998) is presumably biased due to associated time-varying changes in quality associated with auction properties.

[Jerry, Patrick and I would like to control for auction bias due to tim-varying bias in another

paper. We already have the idea fleshed out. It is too presumptuous to mention a working paper we have? We could whip one up by the weekend. I only ask as we would like to make it known we use a token approach to get unbiased auction / foreclosure coefficients ASAP]

3.3 Tokens

Table 1 presents a sample of 8 listings for 3 bedroom 2 bathroom houses in zip code 30043. The sale prices range from $270,000 to $86,125. Based on zip code, bathroom and bedroom it is impossible to explain variation in sale prices. However, the remarks for the property with the largest sale price indicate positive, unobserved features about the location (*located in tranquil neighborhood*) and the property itself (*marble master bath*). These remarks are in contrast to the property with the smallest sale price. There, the remarks indicate the property is not in great condition as the remarks indicate that the buyer must be *willing to put in some sweat equity.*

The public remarks are processed in order to produce a set of variables that indicate certain tokens are present in the remarks. It is possible to create indicator variables for each word in the remarks. In the textual analysis literature, single words are called *unigrams.* Examples of unigrams include *ceiling* and *gated.* In addition to unigrams, this study also examines the use of *bigrams.* A bigram is a two word phrase such as *drop ceiling*, *vaulted ceiling*, *gated windows* or *gated community.*

Before creating the bigrams, *stop words* are removed from the remarks section using a custom set. Stop words are words that are assumed to not convey any information about the property. A list of stop words specific to the remarks section, and real estate at large, is created. The list of stop words is included in the appendix [7].

The token approach models each remark as a collection of tokens. For all unigrams v_j, $j = 1, ..., J$, define the indicator variable $\mathbb{1}(v_j)_i = 1$ if unigram v_j is in remark i and 0 otherwise. The $1 \times J$ vector v_i is then defined as $v_i = (\mathbb{1}(v_1)_i, ..., \mathbb{1}(v_J)_i)$. A similar procedure is used to create the $1 \times J$ vector for bigrams, $w_i = (\mathbb{1}(w_1)_i, ..., \mathbb{1}(w_J)_i)$. Prices for the unigrams and bigrams are contained in the $1 \times J$ vectors $\theta = (\theta(v_1), ..., \theta(v_J))$ and $\phi = (\phi(v_1), ..., \phi(v_J))$, respectively.

Two alternatives to the above approach are also possible. The first approach uses counts and replaces the indicator function with the total number of times the token appears in the remark i;

[7] An additional step called *stemming* is often carried out in textual analysis. In unreported results, we found that stemming did not improve performance or change any of the results in the paper in a substantial manner. Therefore, for the sake of simplicity, the remarks were not stemmed

the second approach uses frequencies rather than counts and replaces the indicator function with the total number of times the token appears in the remark i, divided by the total number of tokens in remark i. In order to facilitate interpretation of the coefficients, we use the indicator function approach but note that in several experiments the results were robust to these two alternative approaches. In the context of Eq's 5 - 8, interpreting the coefficients in θ and ϕ are straightforward. Including u_j in the remarks increases (decreases) the expected price by an amount θ_j if $\theta_j > 0$ ($\theta_j < 0$).

If the researcher is not interested in the prices of tokens but rather aggregating the information contained in the remarks, the inner product $q_i = v_i\theta$ can be used. If we assume that the remarks contain information about house quality, we can interpret q_i as an index of quality. Furthermore, this index of quality can be used as a measure of quality in other regressions. A similar approach using a sufficient reduction paradigm is taken in Taddy (2013b) and Taddy (2013a).

However, it should be emphasized that the tokens are considered exchangeable in that the ordering of the tokens is not important for pricing purposes. For example, when using unigrams, the phrases *gated windows* and *gated community* will be priced as $\theta(\textit{gated}) + \theta(\textit{windows})$ and $\theta(\textit{gated}) + \theta(\textit{community})$, respectively. The difference in price between these two phrases is equal to $\theta(\textit{windows}) - \theta(\textit{community})$. This is counter-intuitive as differences in housing quality indicated by *gated windows* and *gated community* come from the adjective *gated* modifying the nouns *windows* and *community*. Using bigrams alleviates issues associated with unigram exchangeablity by capturing some notion of word ordering. When using bigrams as token, the difference in price between *gated windows* and *gated community* is equal to $\phi(\textit{gated windows}) - \phi(\textit{gated community})$.

Without loss of generality, we use j to indicate the rank of the frequency of the token in the remarks. For example, $j = q$ is the most frequent token, $j = 2$ is the second most frequent token and so on. For practical purposes, it is necessary to truncate the list of total tokens available to use. First, it is hard to rationalize a token that appears in only a single remark is unlikely to appear in future remarks. It is also unlikely that this token can be used to predict future prices. Second, in order to compare the LASSO and its variants to least-squares, we require the matrix of regressors to have full rank which ensures least-squares is feasible. We find the full rank condition is frequently violated when we choose $2000 < J$ using sub periods of the data. We experiment with several alternatives for J including $J \in \{100, 500, 1000, 2000, 3000\}$.

Figure 2 shows the cumulative distribution function for the 4000 most frequent tokens across all listings in the data set. Counts for the less frequent tokens are quite large. The 4000th least frequent unigram (bigram) occurs 65 (220) times in the remarks. In our analysis, we use the 2000 most frequent tokens. The 2000th least frequent unigram (bigram) occurs 249 (453) times in the remarks.

3.4 Out-of-Sample testing

We are interested in both in-sample and out-of-sample performance as measured by the root mean squared error (RMSE). Although the data is an unbalanced panel, due to the large number of houses that transact only once, our data appears to be more cross-sectional. Therefore, we adopt an out-of-sample testing procedure designed for cross-sectional data. For any model, define the RMSE as $RMSE = \sqrt{\frac{1}{N}\sum(p_i - \widehat{p_i})^2}$ where $\widehat{p_i}$ is the predicted price for transaction i. It is well known that the least-squares solution will produce the smallest in-sample $RMSE$. However, due to the large number of regressors, it is possible that the least-squares procedure will overfit the data. Because of this, we perform out-of-sample testing by estimating the models using a subset of observations, the training set, and predicting prices using the remaining observations not in the training set, the testing set.

We use an out-of-sample testing procedure designed to mitigate random sapling error that can come from an arbitrarily chosen testing set using an M-fold procedure similar to the cross-validation procedure, above. Our out-of-sample RMSE is calculated by the following:

1. Randomly sort all of the observations into groups $m = 1, ..., M$ with an equal number of observations per group, N_M. For any m, training set m is the set of observations in groups $1, ..., m-1, m+1, ..., M$ and training set m is the set of observations in group m.

2. For a given λ, $m = 1$, and any specification in Eq's 5-8, solve the optimization procedure in Eq 2 using the obervations in the testing set $m = 1$ the store $\widehat{d(\lambda)^{m=1}}$.

3. Estimate sale prices for sales in testing set $m = 1$ using $\widehat{d(\lambda)^{m=1}}$ and store the root mean squared-error for group $m = 1$: $RMSE(\lambda)^{m=1} = \sqrt{\frac{1}{N_M}\sum(p_i - z_i\widehat{d(\lambda)^{m=1}})^2}$

4. Repeat for $m = 2, ..., M$

5. The average $RMSE$ for this choice of λ is given by $RMSE(\lambda) = \frac{1}{M}\sum RMSE(\lambda)^m$

The above procedure is a means to compare the amount of overfit in each estimation procedure. In our analysis, we set $M = 10$ and fix group assignments across all estimations for a given subperiod. [8] An estimation procedure overfits the data when the in-sample RMSE is small but the out-of-sample RMSE is large. It should be emphasized that in the RMSE calculations, there is no correction for the number of variables included in the model and the resulting degrees of freedom. This is done in order to isolate out-of-sample pricing error due to parameter estimation. Because least-squares uses all variables, the degrees of freedom in least-squares will always be less than or equal to the degrees of freedom when using LASSO. For an analysis of the degrees of freedom in LASSO, see Zou et al. (2007). Therefore, creating RMSE using degrees of freedom instead of N_M in the denominator would decrease the RMSE for LASSO solely based on the number of variables chosen and not the pricing error.

3.5 Data Description

The primary data source used in this study comes from the Georgia Multiple Listings Service (GAMLS). The GAMLS data includes all single-family detached houses that were listed, regardless of whether they sold or not, from January 1, 2000 to December 31, 2014 in the five counties that form the core of the Atlanta metropolitan market (Fulton, Dekalb, Gwinnett, Cobb, and Clayton). In addition to the public remarks field described earlier, the GAMLS dataset includes information on the location and size of the property (address, acreage, etc.), physical characteristics of the house (number of bedrooms, bathrooms, etc.), and details of the transaction (listing date, listing price, sales price, etc.).

All the data in the GAMLS is manually entered by a listing agent. Thus, it is prone to error Levitt and Syverson (2008). It also does not include each houses square feet of living area. We circumvent these potential issues with data obtained from each county's tax assessor office. The tax assessor data includes detailed parcel level information that we use to determine the square feet of living area for each house in our study and validate the information in the GAMLS. The initial GAMLS dataset includes 511,851 listings. We apply several filters to systematically clean

[8] For a given subperiod, $m = 1 + mod(n, 10)$

the data. First, we remove listings with missing or incomplete data. We then winsorize the top and bottom 0.5% of sales price to remove potential outliers. Finally, we exclude houses that were built before 1900, have less than 500 square feet of living area, or have 10 or more bedrooms. We apply these filters to limit the variability in our data and ensure it is reasonably homogeneous as suggested by Butler (1980). The cleaned dataset includes 414,404 unique sales transactions. Descriptive statistics for the entire data are displayed in Table 2.

Our study covers an extended period of time in which the Atlanta real estate market experienced a boom and bust period. Thus, we partition the dataset into two subsets. The first subset includes all listings between January 1st, 2000 and December 31st, 2007 and represents a hot market in which real estate prices rose rapidly. The second subset covers January 1st, 2008 through December 31st, 2014 and represents a cold market in which house prices crashed and subsequently recovered. Descriptive statistics for the two time periods are displayed in Table y. After partitioning the data we rerun the textual analysis separately on both data subsets.

3.6 Results

Figure 1 displays the positive and negative bigram coefficients with the largest magnitudes. The coefficients in Figure 1 were estimated using the cross-validated LASSO procedure in Eq 8 for the entire sample period (2000 2014). Coefficients with a larger magnitude are illustrated in larger font sizes. In Figure 1a, which includes bigrams with positive coefficients, it is clear that the tokens are a mix of qualitative and quantitative variables that describe the property. Figure 1a includes terms such as *shops restaurants*, the *Sandy Springs* neighborhood, and *gated community* that indicate the bigrams capture both location and quality features.[9] In Figure 1b, which includes bigrams with negative coefficients, there are numerous bigrams related to distressed sales and houses in need of repair such as *auction terms*, *investor special*, and *sweat equity*. Overall, the terms in Figure 1 suggest that the tokenization procedure can identify relevant phrases in the MLS remarks section that can be used in pricing models. A detailed list of top fifty unigrams and bigrams sorted by magnitude is available in the appendix in Tables A2 and A3.

[9] We ran the analysis with and without location controls at the zip code and census tract level. The results presented do not include location controls to highlight the breath of textual information available in the MLS remarks section. Including location controls does not have a material impact on the results presented going forward as we present the results of the tokenization procedure in relation to a baseline model. In other words, if we add the location controls to the baseline model they would also be included in the augmented token model.

In the following tables, Panel A presents the in-sample RMSE results, Panel B presents the out-of-sample RMSE results, and Panel C presents the number of variables selected when calculating RMSE. The columns correspond to the RMSE when estimating models in Eq's 5 - 8 using estimated variables from using ordinary least-squares (LS), Benjamini and Yekutieli (2001) false discovery rate (FDR), cross-validated LASSO (CV), Belloni and Chernozhukov feasible LASSO (BC), and Post-LASSO (POST) procedures discussed in Section 3.1. In each model, a maximum of 2,000 tokens are used.

Panels A and B display the RMSE values for each model Eq 5 - 8 divided by the RMSE for the baseline model in Eq 4 in the respective period. By doing so, models and estimation procedures that produce ratios with smaller values are preferred. As mentioned above, each RMSE is not divided by the degrees of freedom in the model. When calculating the RMSE used in Panels A and B, the $\widehat{Q}$ variables in Panel C are used. Only LS uses all $\widehat{Q} = K$ variables. The other 4 procedures select the $\widehat{Q} < K$ variables that are used to calculate RMSE in Panels A and B. By doing so, the RMSE ratios emphasize differences in RMSE due to bias and precision in the coefficient estimates alone and not differences in $\widehat{Q}$. Alternatively, the RMSE ratios do not explicitly reflect . The number of observations and RMSE for the baseline model in each period are listed in Table 3.

We include results for several time frames. The first row in each panel includes data for the entire sample period (2000 to 2014). We then partition the data into pre-crash (2000 to 2007) and post-crash (2008 to 2014) sub periods to examine whether the in- and out-of-sample results are sensitive to the time period selected. Finally, in the last row of each panel we partition the data into a more recent subsample that includes results for 2012 to 2014. We include the smaller, more recent subsample for two reasons. First, while working with the MLS remarks data we noticed that a small percentage of remarks were truncated prior to 2012.[10] Second, we want to ensure that functional obsolescence does not impact the model results. Functional obsolescence in real estate occurs often as the desirability or usefulness of an attribute changes or becomes outdated. Thus, a tokens magnitude and sign may change over time if the attribute becomes functionally obsolete. Given the extended time period of our study, we include the 2012-2014 subsample to

[10]We estimate that the data truncation affected less than 1% of the records prior to 2012. The small percentage of records that were affected were missing less than 16 characters each, which represents less than 7% of their total length. A discussion with our contact at GAMLS revealed that a systems upgrade was performed in the beginning of 2012 and was likely the source of the truncation.

ensure functional obsolescence does not significantly impact the results.

After identifying the relevant tokens, we use tokens alone in the absence of control variables in order to determine the predictive power of text in the absence of any explicit control variables. The results from estimating U! and B1 are displayed in Table 4. Panel A indicates that the tokens have predictive power that is comparable to the control variables. By definition, the LS model has the smallest in-sample RMSE. For the entire sample 2000-2014, the LS RMSE from the unigram model in U1 is smaller than the baseline RMSE by a factor of 0.894. Using the in-sample RMSE from Table 3, this implies RMSE decreases by $11,096. The unigrams perform even better in the sub period 2012-2014 when the RMSE decreases by a factor of 0.815 or a $23,683 decrease in RMSE.

Moving across the columns in Panel A, we find that the RMSE when using unigrams is always less than the RMSE when using bigrams. Further, the RMSE for U1 and B1 are comparable across estimation procedures. Further, the RMSE from the bigram model in B1 is sometimes greater than the RMSE of the baseline model. The bigrams only begin to perform better than the baseline model in the subperiod 2012-2014. Thus, it would appear that the control variables have more predictive power than the bigrams but less predictive power than the unigrams.

Although LS is guaranteed to minimize the in-sample RMSE, we are more intersted in out-of-sampe RMSE. Due to the large number of variables, we would expect LS to overfit the data in-sample resulting in poor out-of-sample performance. The results in Panel B indicate that LS and CV have similar out-of-sample RMSE that is comparable to their in-sample RMSEs. However, given the large number of observations in each subperiod and the consistency of LS, this is not too surprising.

Panel C displays the in-sample $\widehat{Q}$ for each estimation procedure. Each model includes annual fixed effects, and the $\widehat{Q}$ includes the number of non-zero annual fixed effects. Of particular note is the large drop in relevant tokens when using the BC and Post procedures. Despite using roughly half of the tokens, the BC and Post models produce in-sample and out-of-sample RMSEs that are comparable to the other estimation procedures. In particular, estimating U1 using the POST procedure in 2012-2014 selects 673 variables and decreases the in-sample RMSE by a factor of 0.832. This is similar to the 0.815 factor using LS and all 2000 tokens. Thus, the POST procedure suggests there appears to be a large number of tokens that we can discard for the purposes of pricing real estate.

In Table 5, we present the results of U2 and B2 and regress price on both tokens and control variables. Again, the values in Panel A (in-sample) and Panel B (out-of-sample) represent the RMSE divided by the baseline models RMSE. The patterns in Table 5 are comparable to the patterns in Table 4. However, when we augment the standard control variables with tokens, we find a remarkable improvement in both in- and out-of-sample RMSE. For the entire sample, the U1 model estimated using LS decreases in-sample RMSE by a factor of 0.746 or $26,580. Panel B also shows improvement in out-of-sample RMSE as well. Panel C shows a small decrease in $\widehat{Q}$ for the POST procedure when control variables are added to the model. Presumably, tokens in the remarks relating to the quantity of bedrooms and bathrooms become irrelevant after including the control variables. Based on the improvement in pricing, we focus on models U2 and B2 in the remainder of the paper.

In Table 6, we present results for U2 and B2 where we regress the log of price on both the tokens and standard control variables. The log-linear approach is often utilized in empirical real estate research for its easy to interpret coefficients and robustness to outliers and moderate forms of heteroskedasticity. Instead of increasing sale price by a fixed dollar amount, unit increases in the coefficients are interpreted as approximately increasing sale price by a given percentage. Similar to Table 5, we find that including both tokens and the standard control variables improves the performance of all five models relative to the baseline model.[11] Results in Table 6 are comparable to the results in Table 5 and suggest that harnessing the textual information available in MLS remarks can also improve performance when the effect of a token is to increase sale price by a given percentage.

The choice of 2000 tokens was used because it provided a rich set of tokens and permitted computation in a reasonable amount of time.[12] Table 7 shows the results for the period 2000-2014 when using 500, 1000, 2000 and 3000 possible tokens. The results indicate that as few as 100 tokens can improve in-sample and out-of-sample RMSE. For example, using 100 unigrams in U2 and estimating with the LS procedure reduces RMSE by a factor of 0.897.

[11] We also investigated U1 and B1 but do not include the results for the sake of brevity. The results are similar to Table 4 and available upon request.

[12] For example, using using a 2.5 GHz Intel Core i5 processor with 16 Gb of memory and using a paralle program to distribute the program over 4 cores, the results in Table 5 required approximately 3 hours for 2000 tokens. When using 3000 tokens, the results required more than 18 hours for the results in each table. A significant portion of the run time was the LS procedure and the POST procedure. Details of the computing times are available from the authors upon request.

In order to investigate how the tokens perform when the sample is small, we estimate models U2 and B2 for each of the 15 years in our sample. The results are displayed in Table 8. Similar to Panel B in Tables 4 and 5, we find that the out-of-sample results from U2 and B2 perform much better than the baseline model. We also find that the number of tokens chosen in most of the models drops substantially.

All of the hedonic models in Eq's 5 - 8 include both time-invariant and time-variant control variables and tokens. The results in Tables 4 - 8 document the explanatory power associated of tokens associated with both time-invariant and time-varying attributes. In order to separately estimate the explanatory power of tokens associated with time varying attributes, we difference models U2 and B2 using same-property sales. This results in an augmented repeat sales model. The repeat sales model expresses changes in sale prices as changes in indicator variables associated with time. By differencing both sides of U2 and B2 and assuming constant implicit prices over time, we remove the effect of any time-invariant, unobserved variables; because the number of bedrooms and bathrooms does not vary much over time, using U1 and B1 produced nearly identical results. We do not ex-ante identify which tokens are associated with time-invariant variables and instead include all tokens when estimating the repeat sales regression. However, this does not present a problem as the coefficeints on tokens associated with time-invariant attributes should be close to 0. The results for the differenced U2 and B2 using same-property sales are displayed in Table 9. The results in Panel A of Table Table 9 show that the tokens improve the in-sample forecasts in every model. Whereas, the out-of-sample results in Panel B perform better in all but one specification and time period. Overall, these results indicate that the tokens are capturing information associated with time-invariant attributes.

Finally, as a robustness check we stratify our dataset by size in order to demonstrate that the tokens improved forecasting results are not solely a product of a heterogenous sample. That is, we would like to know if the increase in predictive power in the tables is driven by larger homes that contain features than smaller homes. In order to determine the predictive power of the tokens across size segments, we stratify the data into quartiles based on each houses square feet of living area. In Tables 10 and 11 we present the results of a hedonic model for houses in the the lower and upper quartile of house size, respectively. Similar to the hedonic results in Table 5, we find that the token augmented models clearly outperform the standard baseline model both in- and

out-of-sample. Surprisingly, we find that the tokens improve performance more for the smaller homes than the larger homes.

4 Conclusions

The linear hedonic model assumes that the price of a property is a linear combination of all of its attributes other factors. By including all of the relevant variables in the model, the researcher or practitioner minimizes the pricing error. This paper makes both empirical contributions when the researcher incorporates text descriptions of the property in order to improve pricing performance.

The paper discusses methods to use when there are a large number of potential variables to choose from. These methods are part model selection part coefficient estimation.

By themselves, unigram and bigram dummy variables have RMSE values that are similar to the observable variables. Augmenting a hedonic model that currently includes unigrams and bigrams can improve performance. This should be the case as the MLS remarks section appears a way for listing agents to provide information about the property that is not observable using only the listed variables in the MLS.

References

Angrist, J., Chernozhukov, V., and Fernández-Val, I. (2006). Quantile regression under misspecification, with an application to the us wage structure. *Econometrica*, 74(2):539–563.

Angrist, J. D. and Krueger, A. B. (1991). Does compulsory school attendance affect schooling and earnings? *Quarterly Journal of Economics*, 106(4):979–1014.

Belloni, A. and Chernozhukov, V. (2010). Post-l1-penalized estimators in high-dimensional linear regression models. *arXiv preprint arXiv:1001.0188*.

Belloni, A., Chernozhukov, V., and Wang, L. (2011). Square-root lasso: pivotal recovery of sparse signals via conic programming. *Biometrika*, 98(4):791–806.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and pow-

erful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 289–300.

Benjamini, Y. and Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. *Annals of statistics*, pages 1165–1188.

Benson, E. D., Hansen, J. L., Schwartz Jr, A. L., and Smersh, G. T. (1998). Pricing residential amenities: the value of a view. *The Journal of Real Estate Finance and Economics*, 16(1):55–73.

Bickel, P. J., Ritov, Y., and Tsybakov, A. B. (2009). Simultaneous analysis of lasso and dantzig selector. *The Annals of Statistics*, pages 1705–1732.

Bollen, J., Mao, H., and Zeng, X. (2011). Twitter mood predicts the stock market. *Journal of Computational Science*, 2(1):1–8.

Bond, M. T., Seiler, V. L., and Seiler, M. J. (2002). Residential real estate prices: a room with a view. *Journal of Real Estate Research*, 23(1/2):129–138.

Bourassa, S. C., Hamelink, F., Hoesli, M., and MacGregor, B. D. (1999). Defining housing submarkets. *Journal of Housing Economics*, 8(2):160–183.

Bourassa, S. C., Hoesli, M., and Peng, V. S. (2003). Do housing submarkets really matter? *Journal of Housing Economics*, 12(1):12–28.

Brown, G. M. and Pollakowski, H. O. (1977). Economic valuation of shoreline. *The review of Economics and Statistics*, pages 272–278.

Buchinsky, M. (1998). The dynamics of changes in the female wage distribution in the usa: a quantile regression approach. *Journal of applied econometrics*, 13(1):1–30.

Butler, R. V. (1980). Cross-sectional variation in the hedonic relationship for urban housing markets*. *Journal of Regional Science*, 20(4):439–453.

Candès, E. J., Romberg, J., and Tao, T. (2006). Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information. *Information Theory, IEEE Transactions on*, 52(2):489–509.

Card, D. and Krueger, A. B. (1992). Does school quality matter? returns to education and the characteristics of public schools in the united states. *Journal of Political Economy*, 100(1):1–40.

Carriero, A., Kapetanios, G., and Marcellino, M. (2011). Forecasting large datasets with bayesian reduced rank multivariate models. *Journal of Applied Econometrics*, 26(5):735–761.

Clements, M. P. and Galvão, A. B. (2009). Forecasting us output growth using leading indicators: An appraisal using midas models. *Journal of Applied Econometrics*, 24(7):1187–1206.

Dagsvik, J. K., Hægeland, T., and Raknerud, A. (2011). Estimating the returns to schooling: a likelihood approach based on normal mixtures. *Journal of Applied Econometrics*, 26(4):613–640.

Ermisch, J. and Francesconi, M. (2013). The effect of parental employment on child schooling. *Journal of Applied Econometrics*, 28(5):796–822.

Feng, Y. and Yu, Y. (2013). Consistent cross-validation for tuning parameter selection in high-dimensional variable selection. *arXiv preprint arXiv:1308.5390*.

Fernandez, C., Ley, E., and Steel, M. F. (2001). Model uncertainty in cross-country growth regressions. *Journal of applied Econometrics*, 16(5):563–576.

Fuentes, J., Poncela, P., and Rodríguez, J. (2014). Sparse partial least squares in time series for macroeconomic forecasting. *Journal of Applied Econometrics*.

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

Gentzkow, M. and Shapiro, J. Vwhat drives media slant? evidence from us newspapers. v forthcoming. *Econometrica*.

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

Haag, J., Rutherford, R., and Thomson, T. (2000). Real estate agent remarks: help or hype? *Journal of Real Estate Research*, 20(1-2):205–215.

Hill, R. C., Knight, J. R., and Sirmans, C. (1997). Estimating capital asset price indexes. *Review of Economics and Statistics*, 79(2):226–233.

Jensen, R. (2010). The (perceived) returns to education and the demand for schooling. *The Quarterly Journal of Economics*, 125(2):515–548.

Johnstone, I. M. and Lu, A. Y. (2009). On consistency and sparsity for principal components analysis in high dimensions. *Journal of the American Statistical Association*, 104(486).

Kain, J. F. and Quigley, J. M. (1970). Measuring the value of housing quality. *Journal of the American statistical association*, 65(330):532–548.

Kang, H.-B. and Reichert, A. K. (1991). An empirical analysis of hedonic regression and grid-adjustment techniques in real estate appraisal. *Real Estate Economics*, 19(1):70–91.

Knight, K. and Fu, W. (2000). Asymptotics for lasso-type estimators. *Annals of statistics*, pages 1356–1378.

Leng, C., Lin, Y., and Wahba, G. (2006). A note on the lasso and related procedures in model selection. *Statistica Sinica*, 16(4):1273.

Levitt, S. D. and Syverson, C. (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.

Liu, Z. and Stengos, T. (1999). Non-linearities in cross-country growth regressions: a semiparametric approach. *Journal of Applied Econometrics*, 14(5):527–538.

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

Machado, J. A. and Mata, J. (2005). Counterfactual decomposition of changes in wage distributions using quantile regression. *Journal of applied Econometrics*, 20(4):445–465.

Maclennan, D. and Tu, Y. (1996). Economic perspectives on the structure of local housing systems. *Housing studies*, 11(3):387–406.

Malpezzi, S. (2003). Hedonic pricing models: a selective and applied review. *Section in Housing Economics and Public Policy: Essays in Honor of Duncan Maclennan*.

Mayer, C. J. (1998). Assessing the performance of real estate auctions. *Real Estate Economics*, 26(1):41–66.

Meinshausen, N. and Bühlmann, P. (2006). High-dimensional graphs and variable selection with the lasso. *The Annals of Statistics*, pages 1436–1462.

Mitra, T. and Gilbert, E. (2014). The language that gets people to give: Phrases that predict success on kickstarter. In *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing*, pages 49–61. ACM.

Paterson, R. W. and Boyle, K. J. (2002). Out of sight, out of mind? using gis to incorporate visibility in hedonic property value models. *Land economics*, 78(3):417–425.

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

Rodriguez, F. and Rodrik, D. (2001). Trade policy and economic growth: a skeptic's guide to the cross-national evidence. In *NBER Macroeconomics Annual 2000, Volume 15*, pages 261–338. MIT Press.

Rosen, S. (1974). Hedonic prices and implicit markets: product differentiation in pure competition. *The journal of political economy*, pages 34–55.

Rouwendal, J., Van Marwijk, R., and Levkovich, O. (2014). The value of proximity to water in residential areas.

Sala-i Martin, X. X. (1997). I just ran two million regressions. *The American Economic Review*, pages 178–183.

Song, Y. and Knaap, G.-J. (2003). New urbanism and housing values: a disaggregate assessment. *Journal of Urban Economics*, 54(2):218–238.

Soyeh, K. W., Wiley, J. A., and Johnson, K. H. (2014). Do buyer incentives work for houses during a real estate downturn? *The Journal of Real Estate Finance and Economics*, 48(2):380–396.

Stewart, M. B. (2007). The interrelated dynamics of unemployment and low-wage employment. *Journal of applied econometrics*, 22(3):511–531.

Stock, J. H. and Watson, M. W. (2002). Macroeconomic forecasting using diffusion indexes. *Journal of Business & Economic Statistics*, 20(2):147–162.

Taddy, M. (2013a). Measuring political sentiment on twitter: Factor optimal design for multinomial inverse regression. *Technometrics*, 55(4):415–425.

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

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.

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

Tu, C. C. and Eppli, M. J. (1999). Valuing new urbanism: The case of kentlands. *Real Estate Economics*, 27(3):425–451.

Varian, H. R. (2014). Big data: New tricks for econometrics. *The Journal of Economic Perspectives*, pages 3–27.

Zhao, P. and Yu, B. (2006). On model selection consistency of lasso. *The Journal of Machine Learning Research*, 7:2541–2563.

Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2):301–320.

Zou, H., Hastie, T., Tibshirani, R., et al. (2007). On the "degrees of freedom" of the lasso. *The Annals of Statistics*, 35(5):2173–2192.

5 Tables and Figures

Table 1: Sample MLS Listing

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: Descriptive Statistics

	Min	Mean	Median	Max	Std Dev
Sale Price ($1,000s)	11.4	193	156.5	1099	143.8
List Price ($1,000s)	1	199	159.9	3400	151.1
Area (1000s ft^2)	506	2220.1	2009	16475	979.4
# of Bedrooms	1	3.6	4	9	0.9
# of Bathrooms	1	2.3	2	12	0.8
Construction Year	1900	1983.5	1989	2014	20.6
Sale Year	2000	2006.9	2007	2014	4.1

Table 3: Observations and Baseline RMSE by Period ($1,000s)

Period	Observations	In-Sample	Out-Of-Sample
2000-2014	414,404	104.648	104.662
2000-2007	234,665	91.102	91.115
2008-2014	179,739	119.908	119.947
2012-2014	76,704	128.019	128.055
2000	22,306	73.935	74.024
2001	24,980	76.344	76.426
2002	24,556	78.219	78.304
2003	29,130	81.732	81.799
2004	33,664	87.554	87.57
2005	36,264	95.954	95.984
2006	35,238	103.432	103.478
2007	28,527	112.041	112.089
2008	26,555	116.266	116.34
2009	26,550	110.692	110.757
2010	23,283	115.655	115.734
2011	26,647	112.261	112.326
2012	27,722	119.016	119.104
2013	26,463	131.255	131.415
2014	22,519	134.099	148.334

Table 4: Hedonic: Price without Control Variables

	LS		FDR=0.1		CV		BC		POST	
Period	(U1)	(B1)	(U1)	(B1)	(U1)	(B1)	(U1)	(B1)	(U1)	(B1)
PANEL A: In-Sample RMSE										
2000-2014	***0.894***	1.034	0.895	1.036	0.895	1.035	0.907	1.050	0.898	1.039
2000-2007	***0.921***	1.045	0.923	1.049	***0.921***	1.045	0.942	1.068	0.928	1.052
2008-2014	***0.849***	0.995	0.852	1.001	0.850	0.996	0.872	1.023	0.857	1.004
2012-2014	***0.815***	0.960	0.824	0.973	0.817	0.963	0.855	1.008	0.832	0.980
PANEL B: Out-Of-Sample RMSE										
2000-2014	***0.900***	1.039	0.901	1.041	***0.900***	1.039	0.911	1.053	0.903	1.043
2000-2007	***0.931***	1.053	0.933	1.058	***0.931***	1.052	0.948	1.073	0.936	1.059
2008-2014	0.862	1.006	0.864	1.011	***0.861***	1.005	0.879	1.029	0.867	1.012
2012-2014	0.843	0.983	0.850	0.994	***0.841***	0.982	0.869	1.019	0.852	0.994
PANEL C: $\widehat{Q}$										
2000-2014	2014	2014	1441	1443	1914	1938	***1002***	1019	***1002***	1019
2000-2007	2014	2014	1441	1443	1914	1938	***1002***	1019	***1002***	1019
2008-2014	2006	2006	1220	1198	1789	1870	842	***840***	842	***840***
2012-2014	2002	2002	976	911	1623	1670	673	***603***	673	***603***

Note: Each value in Panel A and Panel B is the ratio of the in-sample RMSE to the benchmark model for a given period. All models include annual fixed effects. A maximum of 2000 tokens are used. Minimum values in each subperiod are indicated with bold italics.

Table 5: Hedonic: Price with Control Variables

	LS		FDR=0.1		CV		BC		POST	
Period	(U2)	(B2)	(U2)	(B2)	(U2)	(B2)	(U2)	(B2)	(U2)	(B2)
PANEL A: In-Sample RMSE										
2000-2014	***0.746***	0.842	0.747	0.844	***0.746***	0.842	0.757	0.855	0.749	0.846
2000-2007	***0.768***	0.849	0.770	0.855	***0.768***	0.849	0.786	0.868	0.775	0.855
2008-2014	***0.710***	0.814	0.712	0.819	0.711	0.814	0.730	0.837	0.718	0.822
2012-2014	***0.691***	0.800	0.699	0.813	0.693	0.802	0.725	0.839	0.706	0.816
PANEL B: Out-Of-Sample RMSE										
2000-2014	***0.751***	0.846	0.752	0.847	***0.751***	0.846	0.760	0.857	0.753	0.849
2000-2007	***0.776***	0.857	0.778	0.864	***0.776***	0.855	0.790	0.872	0.780	0.860
2008-2014	0.721	0.822	0.722	0.827	***0.720***	0.822	0.736	0.842	0.726	0.828
2012-2014	0.715	0.820	0.721	0.830	***0.713***	0.818	0.737	0.847	0.722	0.828
PANEL C: $\widehat{Q}$										
2000-2014	2029	2029	1432	1370	1916	1929	965	***914***	965	***914***
2000-2007	2029	2029	1432	1370	1916	1929	965	***914***	965	***914***
2008-2014	2021	2021	1200	1132	1786	1832	787	***758***	787	***758***
2012-2014	2015	2015	941	791	1614	1605	623	***528***	623	***528***

Note: Each value in Panel A and Panel B is the ratio of the in-sample RMSE to the benchmark model for a given period. All models include annual fixed effects. A maximum of 2000 tokens are used. Minimum values in each subperiod are indicated with bold italics.

Table 6: Hedonic: Log Price with Control Variables

	LS		FDR=0.1		CV		BC		POST	
Period	(U2)	(B2)	(U2)	(B2)	(U2)	(B2)	(U2)	(B2)	(U2)	(B2)
PANEL A: In-Sample RMSE										
2000-2014	0.727	0.797	0.727	0.798	***0.726***	0.798	0.738	0.809	0.730	0.802
2000-2007	***0.712***	0.792	0.714	0.796	0.714	0.793	0.731	0.811	0.720	0.799
2008-2014	***0.696***	0.766	0.700	0.770	***0.696***	0.767	0.716	0.787	0.702	0.773
2012-2014	***0.668***	0.750	0.677	0.766	0.670	0.752	0.703	0.787	0.681	0.766
PANEL B: Out-Of-Sample RMSE										
2000-2014	***0.731***	0.801	***0.731***	0.803	***0.731***	0.801	0.740	0.812	0.733	0.805
2000-2007	***0.719***	0.817	0.721	0.822	***0.719***	0.799	0.735	0.815	0.725	0.804
2008-2014	***0.704***	0.775	0.707	0.778	***0.704***	0.774	0.721	0.792	0.708	0.780
2012-2014	0.687	0.769	0.694	0.783	***0.685***	0.767	0.712	0.794	0.694	0.776
PANEL C: $\widehat{Q}$										
2000-2014	2029	2029	1509	1365	1964	1920	1014	***923***	1014	***923***
2000-2007	2029	2029	1509	1365	1964	1920	1014	***923***	1014	***923***
2008-2014	2021	2021	1268	1171	1854	1849	886	***809***	886	***809***
2012-2014	2015	2015	987	842	1665	1651	712	***572***	712	***572***

Note: Each value in Panel A and Panel B is the ratio of the in-sample RMSE to the benchmark model for a given period. All models include annual fixed effects. A maximum of 2000 tokens are used. Minimum values in each subperiod are indicated with bold italics.

Table 7: Hedonic: Price with Control Variables Different Tokens

	LS		FDR=0.1		CV		BC		POST	
Tokens	(U2)	(B2)	(U2)	(B2)	(U2)	(B2)	(U2)	(B2)	(U2)	(B2)
PANEL A: In-Sample RMSE										
100	***0.897***	0.926	***0.897***	0.926	***0.897***	0.926	0.900	0.928	0.898	0.926
500	***0.828***	0.884	***0.828***	0.884	***0.828***	0.884	0.833	0.889	0.829	0.885
1000	***0.783***	0.866	***0.783***	0.867	***0.783***	0.867	0.790	0.875	0.785	0.869
2000	***0.746***	0.842	0.747	0.844	***0.746***	0.842	0.757	0.855	0.749	0.846
3000	***0.731***	0.824	0.733	0.826	0.732	0.824	0.745	0.840	0.736	0.830
PANEL B: Out-Of-Sample RMSE										
100	***0.898***	0.926	***0.898***	0.926	***0.898***	0.926	0.900	0.929	***0.898***	0.927
500	***0.829***	0.885	***0.829***	0.885	***0.829***	0.885	0.834	0.890	0.830	0.886
1000	***0.785***	0.868	0.786	0.869	***0.785***	0.868	0.792	0.877	0.787	0.871
2000	***0.751***	0.846	0.752	0.847	***0.751***	0.846	0.760	0.857	0.753	0.849
3000	0.739	0.877	0.740	0.879	***0.738***	0.830	0.750	0.844	0.742	0.834
PANEL C: $\widehat{Q}$										
100	129	129	127	122	128	125	117	***109***	117	***109***
500	529	529	457	435	518	501	376	***359***	376	***359***
1000	1029	1029	815	793	996	982	634	***587***	634	***587***
2000	2029	2029	1432	1370	1916	1929	965	***914***	965	***914***
3000	3029	3029	1874	1819	2729	2797	***1166***	1180	***1166***	1180

Note: Each value in Panel A and Panel B is the ratio of the in-sample RMSE to the benchmark model for a given period. All models include annual fixed effects. Minimum values in each subperiod are indicated with bold italics.

Table 8: Hedonic: Annual, Price with Control Variables

	LS		FDR=0.1		CV		BC		POST	
Year	(U2)	(B2)	(U2)	(B2)	(U2)	(B2)	(U2)	(B2)	(U2)	(B2)
PANEL A: Out-Of-Sample RMSE										
2000	0.863	0.919	0.907	1.051	***0.812***	0.900	0.911	1.009	0.828	0.916
2001	1.097	1.053	1.097	1.107	***0.815***	0.891	0.916	1.006	0.828	0.912
2002	0.875	1.084	0.901	1.062	***0.801***	0.882	0.901	0.995	0.860	0.896
2003	0.840	0.884	0.854	0.928	***0.802***	0.873	0.840	0.917	0.818	0.896
2004	0.796	0.870	0.813	0.921	***0.786***	0.862	0.820	0.902	0.800	0.880
2005	0.791	0.873	0.821	0.911	***0.782***	0.866	0.815	0.907	0.794	0.884
2006	0.810	0.880	0.824	0.908	***0.770***	0.852	0.804	0.889	0.784	0.867
2007	0.774	0.842	0.807	0.918	***0.752***	0.832	0.789	0.877	0.763	0.852
2008	0.739	0.833	0.763	0.879	***0.721***	0.822	0.768	0.888	0.732	0.844
2009	0.765	0.841	0.794	0.897	***0.737***	0.826	0.774	0.881	0.748	0.845
2010	0.749	0.830	0.764	0.857	***0.728***	0.818	0.767	0.870	0.738	0.838
2011	0.772	0.855	0.787	0.883	***0.750***	0.843	0.790	0.897	0.760	0.864
2012	0.758	0.872	0.775	0.905	***0.740***	0.834	0.782	0.885	0.750	0.847
2013	0.764	1.098	0.782	1.089	***0.718***	0.825	0.758	0.873	0.737	0.842
2014	0.936	1.010	0.942	1.059	***0.717***	0.824	0.763	0.877	0.733	0.849
PANEL B: $\widehat{Q}$										
2000	2011	2011	415	***170***	1076	1129	404	288	404	288
2001	1911	1911	422	***259***	1141	1127	432	315	432	315
2002	2011	1911	450	***248***	1125	1177	465	343	465	343
2003	2011	2011	518	370	1210	1307	327	***208***	327	***208***
2004	2011	2011	528	400	1274	1308	367	***259***	367	***259***
2005	2011	2011	556	387	1268	1382	361	***260***	361	***260***
2006	2012	2012	559	377	1297	1292	379	***249***	379	***249***
2007	2011	2011	492	370	1202	1283	364	***240***	364	***240***
2008	2011	2011	624	***376***	1266	1345	608	517	608	517
2009	2013	2013	574	***350***	1191	1230	568	501	568	501
2010	2012	2012	531	***399***	1288	1226	563	483	563	483
2011	2012	2012	531	***399***	1288	1226	563	483	563	483
2012	2012	2012	552	***365***	1295	1281	591	520	591	520
2013	2013	1913	570	378	1343	1265	414	***311***	414	***311***
2014	2013	2013	506	319	1186	1311	372	***264***	372	***264***

Note: Each value in Panel A is the ratio of the in-sample RMSE to the benchmark model for a given year. All models include annual fixed effects. A maximum of 2000 tokens are used. Minimum values in each subperiod are indicated with bold italics. For 2001, 100 tokens were removed due to perfect multicollinearity.

Table 9: Repeat Sales: Price without Control Variables

	LS		FDR=0.1		CV		BC		POST	
Period	(U2)	(B2)	(U2)	(B2)	(U2)	(B2)	(U2)	(B2)	(U2)	(B2)
PANEL A: In-Sample RMSE										
2000-2014	***0.906***	0.956	0.916	0.971	0.909	0.959	0.908	0.958	***0.906***	0.957
2000-2007	***0.892***	0.948	0.934	0.980	0.913	0.962	0.911	0.963	0.899	0.954
2008-2014	***0.794***	0.877	0.875	0.957	0.841	0.904	0.835	0.913	0.809	0.891
2012-2014	***0.636***	0.751	0.898	1.049	0.834	0.892	0.754	0.873	0.666	0.780
PANEL B: Out-Of-Sample RMSE										
2000-2014	0.930	0.975	0.936	0.984	0.928	0.973	***0.927***	0.972	0.930	0.974
2000-2007	0.975	1.151	0.986	1.112	***0.953***	0.988	0.954	0.988	0.965	0.996
2008-2014	0.963	1.264	0.959	1.189	***0.900***	0.962	0.913	0.978	0.925	0.978
2012-2014	1.478	1.645	1.375	1.340	***0.956***	1.047	0.971	1.056	1.081	1.132
PANEL C: $\widehat{Q}$										
2000-2014	2014	2014	699	***503***	1450	1402	1527	1453	1527	1453
2000-2007	2014	2014	699	***503***	1450	1402	1527	1453	1527	1453
2008-2014	2006	1906	227	***173***	637	869	846	798	846	798
2012-2014	2002	1902	81	***23***	164	236	371	334	371	334

Note: Each value in Panel A and Panel B is the ratio of the in-sample RMSE to the benchmark model for a given period. All models include annual fixed effects. A maximum of 2000 tokens are used. Minimum values in each subperiod are indicated with bold italics.

Table 10: Hedonic: Lower Quartile of Square Footage

	LS		FDR=0.1		CV		BC		POST	
Period	(U2)	(B2)	(U2)	(B2)	(U2)	(B2)	(U2)	(B2)	(U2)	(B2)
PANEL A: In-Sample RMSE										
2000-2014	***0.673***	0.783	0.679	0.793	0.674	0.785	0.698	0.814	0.683	0.796
2000-2007	***0.686***	0.783	0.699	0.816	0.690	0.787	0.728	0.831	0.706	0.806
2008-2014	***0.620***	0.748	0.638	0.781	0.625	0.753	0.669	0.807	0.642	0.780
2012-2014	***0.577***	0.716	0.625	0.788	0.597	0.740	0.663	0.817	0.623	0.777
PANEL B: Out-Of-Sample RMSE										
2000-2014	0.692	0.824	0.696	0.832	***0.690***	0.797	0.708	0.821	0.696	0.806
2000-2007	0.717	0.829	0.724	0.852	***0.713***	0.806	0.739	0.839	0.722	0.820
2008-2014	0.703	0.815	0.713	0.840	***0.660***	0.781	0.688	0.819	0.668	0.799
2012-2014	0.734	0.816	0.736	0.838	***0.657***	0.790	0.696	0.836	0.670	0.811
PANEL C: $\widehat{Q}$										
2000-2014	2023	2023	927	842	1610	1579	587	***524***	587	***524***
2000-2007	2023	2023	927	842	1610	1579	587	***524***	587	***524***
2008-2014	2014	2014	696	559	1439	1470	465	***334***	465	***334***
2012-2014	2008	2010	385	237	990	976	312	***208***	312	***208***

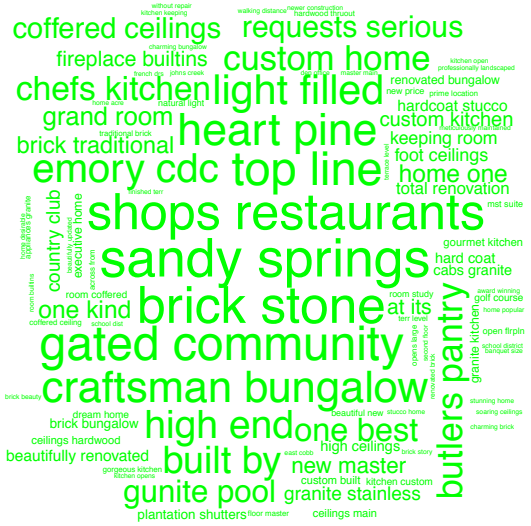
Note: Each value in Panel A and Panel B is the ratio of the in-sample RMSE to the benchmark model for a given period. All models include annual fixed effects. A maximum of 2000 tokens are used. Minimum values in each subperiod are indicated with bold italics.

Table 11: Hedonic: Upper Quartile of Square Footage

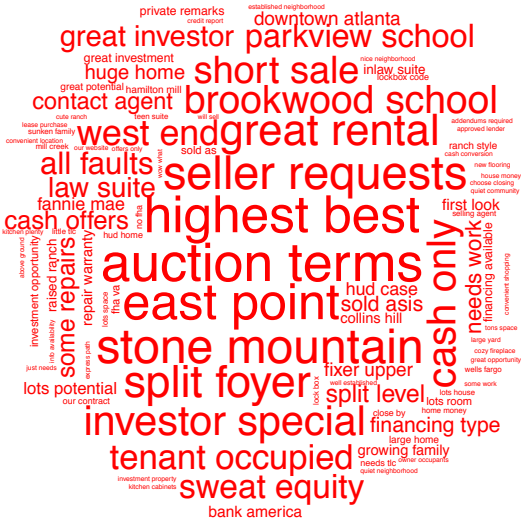
	LS		FDR=0.1		CV		BC		POST	
Period	(U2)	(B2)	(U2)	(B2)	(U2)	(B2)	(U2)	(B2)	(U2)	(B2)
PANEL A: In-Sample RMSE										
2000-2014	***0.759***	0.843	0.780	0.872	0.761	0.844	0.791	0.877	0.772	0.858
2000-2007	***0.776***	0.849	0.806	0.867	0.780	0.852	0.829	0.901	0.803	0.877
2008-2014	***0.718***	0.811	0.732	0.846	0.723	0.816	0.772	0.864	0.745	0.838
2012-2014	***0.686***	0.787	0.852	0.857	0.703	0.804	0.775	0.870	0.738	0.836
PANEL B: Out-Of-Sample RMSE										
2000-2014	0.777	0.859	0.803	0.892	***0.776***	0.857	0.799	0.883	0.784	0.867
2000-2007	0.811	1.073	0.837	1.075	***0.807***	0.873	0.840	0.907	0.821	0.890
2008-2014	0.755	0.843	0.770	0.872	***0.752***	0.839	0.784	0.875	0.765	0.852
2012-2014	0.769	0.883	0.872	1.047	***0.751***	0.843	0.795	0.886	0.771	0.862
PANEL C: $\widehat{Q}$										
2000-2014	2029	2029	1010	874	1631	1689	632	***519***	632	***519***
2000-2007	2029	2029	1010	874	1631	1689	632	***519***	632	***519***
2008-2014	2020	2020	779	578	1532	1486	499	***372***	499	***372***
2012-2014	2015	2015	494	316	1249	1143	338	***248***	338	***248***

Note: Each value in Panel A and Panel B is the ratio of the in-sample RMSE to the benchmark model for a given period. All models include annual fixed effects. A maximum of 2000 tokens are used. Minimum values in each subperiod are indicated with bold italics.

Figure 1: Positive and Negative Bigrams

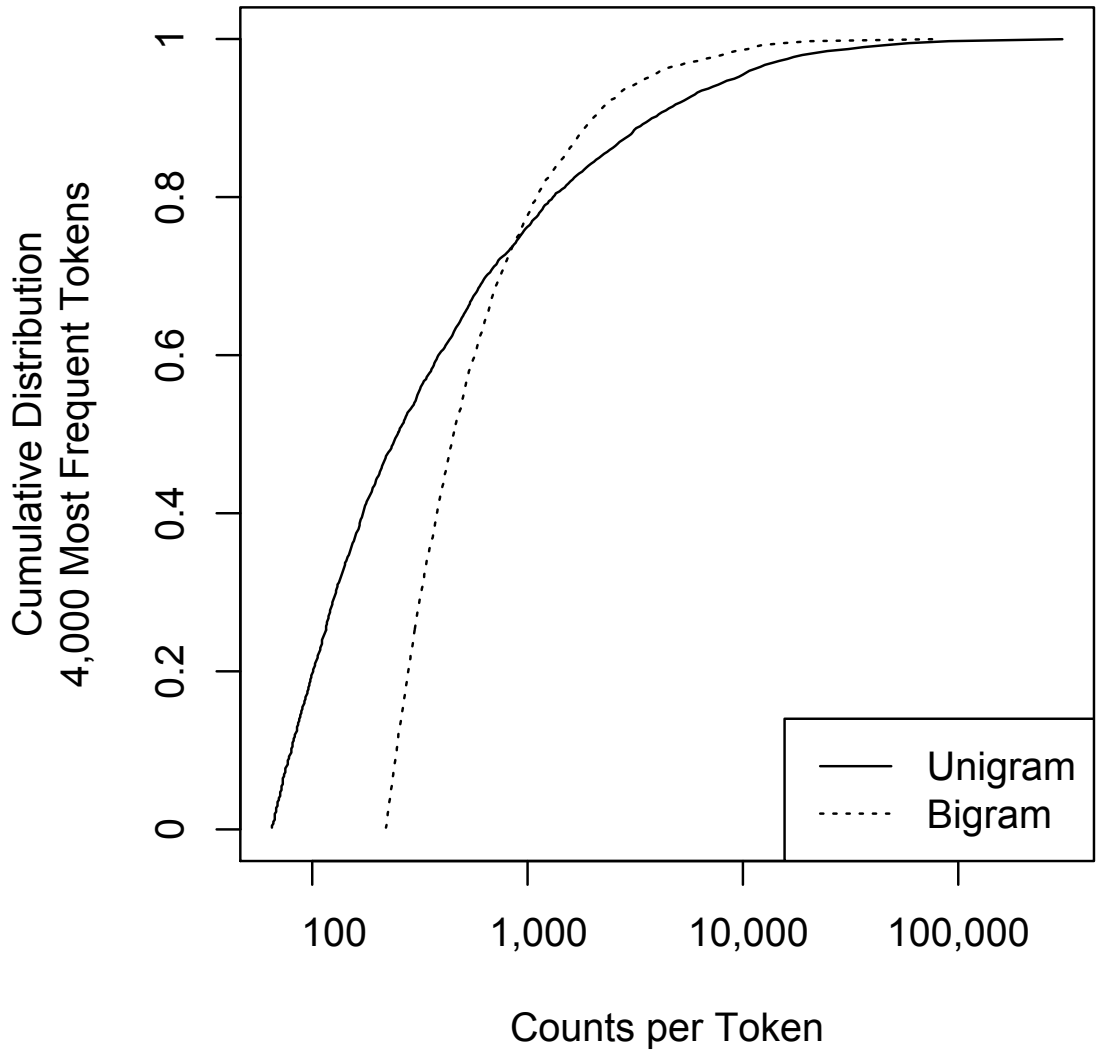


(a) Bigrams With Positive Prices



(b) Bigrams With Negative Prices

Figure 2: Cumulative Distribution for 4,000 Most Frequent Tokens



6 Appendix

Table A1: Repeat Sales: Log Price without Control Variables

	LS		FDR=0.1		CV		BC		POST	
Period	(U1)	(B1)	(U1)	(B1)	(U1)	(B1)	(U1)	(B1)	(U1)	(B1)
PANEL A: In-Sample RMSE										
2000-2014	***0.829***	0.928	***0.829***	0.930	***0.829***	0.929	0.841	0.942	0.832	0.933
2000-2007	***0.840***	0.954	0.841	0.963	***0.840***	0.955	0.860	0.975	0.847	0.961
2008-2014	***0.798***	0.896	0.800	0.898	0.799	0.896	0.821	0.920	0.806	0.904
2012-2014	***0.759***	0.866	0.766	0.876	0.761	0.869	0.799	0.910	0.776	0.883
PANEL B: Out-Of-Sample RMSE										
2000-2014	***0.833***	0.933	0.834	0.935	***0.833***	0.932	0.844	0.945	0.836	0.937
2000-2007	***0.848***	0.992	0.849	1.003	***0.848***	0.961	0.866	0.980	0.854	0.967
2008-2014	0.808	0.905	0.809	0.908	***0.807***	0.905	0.826	0.925	0.813	0.911
2012-2014	0.780	0.887	0.786	0.895	***0.778***	0.885	0.808	0.919	0.789	0.897
PANEL C: $\widehat{Q}$										
2000-2014	2014	2014	1491	1387	1934	1928	1078	***1008***	1078	***1008***
2000-2007	2014	2014	1491	1387	1934	1928	1078	***1008***	1078	***1008***
2008-2014	2006	2006	1291	1225	1852	1793	885	***858***	885	***858***
2012-2014	2002	2002	1062	931	1685	1688	703	***652***	703	***652***

Note: Each value in Panel A and Panel B is the ratio of the in-sample RMSE to the benchmark model for a given period. All models include annual fixed effects. A maximum of 2000 tokens are used. Minimum values in each subperiod are indicated with bold italics.

Table A2: Hedonic: Top 50 Positive and Negative Unigram Coefficients in $1,000s

Positive Words	Coefficient	Counts	Negative Words	Coefficient	Counts
chastain	209.996	327	ideas	-52.336	514
morningside	199.115	600	jefferson	-45.742	303
willis	137.67	272	college	-39.426	540
subzero	132.813	363	auction	-39.005	1242
virginia	126.268	318	inground	-34.995	3137
smith	119.451	312	turner	-33.955	304
buckhead	109.829	991	ground	-33.213	1427
viking	106.788	313	mountain	-32.66	1095
brookhaven	105.355	1198	splitfoyer	-31.25	259
druid	103.549	423	splitlevel	-29.963	739
austin	90.701	263	legacy	-29.215	403
sandy	90.519	619	array	-28.102	306
million	90.229	260	catwalk	-27.842	352
carriage	89.333	267	bowed	-27.767	453
candler	85.236	421	law	-27.346	1021
tudor	85.105	486	investors	-27.171	4296
shake	84.285	328	transportation	-26.19	383
vanderlyn	84.167	260	friday	-26.061	2142
sophisticated	83.328	254	comet	-25.932	540
gated	79.001	1888	snellville	-25.398	375
jackson	78.064	254	investor	-25.164	7148
ashford	77.31	500	investment	-25.067	7701
gracious	75.164	853	prices	-25.025	256
handsome	74.636	338	sliding	-25.013	536
streets	71.602	549	homepath	-24.922	6645
bosch	70.192	371	fixer	-24.88	1647
wine	67.366	851	collins	-24.294	854
grove	67.177	640	highest	-24.061	905
dunwoody	66.135	1235	started	-24.015	411
hilltop	66.013	333	cascade	-23.901	437
enclave	65.442	307	courts	-23.518	258
milton	64.485	390	inlaw	-23.514	5645
finishes	64.441	1118	corners	-22.862	447
walton	62.157	964	shaded	-22.376	468
lightfilled	60.978	304	trilevel	-22.29	503
highlands	60.502	286	ridge	-22.174	694
keep	59.885	508	split	-22.121	20572
sarah	59.181	256	status	-22.03	339
highland	59.14	365	jonesboro	-21.742	324
oakhurst	56.93	726	rent	-21.144	598
cobbs	56.757	255	binding	-21.134	249
paneled	56.527	510	mrtge	-21.051	270
hills	55.102	2477	bargain	-20.856	3631
designed	55.052	1146	apprvd	-20.852	317
manor	54.597	396	receive	-20.619	561
peachtree	54.537	977	table	-20.531	388
classic	54.32	2847	parkview	-20.272	1719
clgs	54.283	340	daily	-20.115	453
coffered	52.85	1641	sunken	-20.06	4034
ptree	51.273	409	mtn	-19.889	556

Table A3: Hedonic: Top 50 Positive and Negative Bigram Coefficients in $1,000s

Positive Words	Coefficient	Counts	Negative Words	Coefficient	Counts
sandy springs	122.744	558	auction terms	-78.539	556
brick stone	82.269	761	highest best	-54.503	766
shops restaurants	81.722	492	east point	-52.688	516
gated community	77.984	770	stone mountain	-52.296	547
top line	77.714	846	seller requests	-49.709	1471
heart pine	71.408	502	split foyer	-49.212	5060
craftsman bungalow	70.772	554	great rental	-48.79	509
light filled	69.203	603	investor special	-47.573	999
emory cdc	69.05	808	cash only	-46.35	655
high end	68.834	764	brookwood school	-46.142	710
built by	65.278	508	short sale	-44.512	8032
butlers pantry	64.142	1003	west end	-43.686	475
custom home	62.892	983	tenant occupied	-42.84	498
chefs kitchen	60.565	2367	sweat equity	-42.717	731
one best	59.177	532	parkview school	-42.195	522
gunite pool	57.818	578	all faults	-39.984	462
requests serious	57.4	1309	appliance package	-39.826	496
coffered ceilings	57.353	460	great investor	-38.189	590
new construction	57.133	2177	law suite	-37.082	755
renovated kitchen	56.165	2579	owner occupant	-37.054	701
grand room	54.706	855	cash offers	-35.441	642
one kind	53.744	565	some repairs	-35.441	482
home one	51.697	498	split level	-35.235	8486
brick traditional	49.511	1600	contact agent	-35.106	468
new master	46.337	787	needs work	-35.012	805
fireplace builtins	46.162	552	sold asis	-34.648	14400
at its	45.128	611	financing type	-34.59	512
custom kitchen	43.577	686	huge home	-34.415	841
country club	42.503	775	fixer upper	-34.07	1519
granite stainless	41.994	2695	downtown atlanta	-33.397	603
keeping room	41.576	5152	fannie mae	-33.241	2760
foot ceilings	40.866	1497	hud case	-33.08	555
total renovation	40.816	795	lots potential	-33.027	1214
beautifully renovated	40.187	949	bank america	-32.776	769
hardcoat stucco	40.071	2325	first look	-32.503	1418
high ceilings	40.003	1990	collins hill	-31.43	734
cabs granite	39.913	573	growing family	-31.423	691
hard coat	38.687	497	repair warranty	-31.214	842
brick bungalow	38.675	1130	raised ranch	-30.795	534
plantation shutters	38.591	1781	inlaw suite	-30.426	4106
executive home	37.916	1177	financing available	-30.423	730
renovated bungalow	36.489	695	great investment	-30.367	4550
granite kitchen	36.439	2114	private remarks	-30.312	578
custom built	36.35	2387	sold as	-30.277	42981
ceilings hardwood	35.564	1026	fha va	-29.843	489
ceilings main	35.4	491	investment opportunity	-29.843	2289
gourmet kitchen	35.168	8064	lots room	-29.784	1190
room coffered	34.959	524	ranch style	-29.667	2039
golf course	34.513	2146	great potential	-29.431	1154
dream home	34.316	847	hamilton mill	-29.39	550

Table A4: Repeat Sales: Top 50 Positive and Negative Unigram Coefficients in $1,000s

Positive Words	Coefficient	Counts	Negative Words	Coefficient	Counts
finishes	41.643	1118	tear	-90.546	312
viking	40.757	313	renovate	-29.956	470
ashford	34.637	500	auction	-19.455	1242
grade	31.938	267	chance	-19.381	355
chastain	30.683	327	array	-17.435	306
sophisticated	29.925	254	sarah	-17.381	256
oakhurst	28.454	726	bones	-17.025	583
highend	27.332	434	updating	-15.624	1147
construction	27.006	3789	binding	-15.195	249
milton	25.941	390	build	-14.186	779
wine	25.608	851	fixerupper	-13.42	532
atlantas	23.145	272	inspections	-13.323	453
detail	23.116	1303	opp	-12.414	299
outdoor	22.784	1338	dec	-12.039	413
coffered	22.62	1641	rough	-11.599	386
tudor	22.6	486	lenox	-11.454	303
subzero	22.032	363	slr	-11.189	624
frontage	21.79	260	hwd	-11.089	366
grove	21.155	640	fixer	-11.051	1647
expanded	20.602	964	homepath	-10.95	6645
austin	20.423	263	prequalif	-10.686	312
designed	19.684	1146	faults	-10.535	495
morningside	19.591	600	soldasis	-10.216	255
shake	19.465	328	hrwds	-10.137	302
properties	19.183	363	diamond	-10.071	462
bosch	19.089	371	statement	-9.636	576
elegance	18.868	624	loans	-9.607	568
lakeside	18.844	581	wful	-9.411	366
const	18.014	355	boa	-9.39	483
glazed	17.731	270	older	-9.359	641
johns	17.377	588	third	-9.136	399
gran	17.347	334	sld	-9.107	458
granite	17.025	29620	potential	-9.08	7057
spaces	16.888	699	opportunity	-8.983	13398
river	16.479	1174	needs	-8.909	11330
chefs	16.223	2951	college	-8.767	540
waterfall	15.988	681	mtg	-8.742	1007
renovation	15.597	8387	shortsale	-8.647	324
candler	15.45	421	rate	-8.56	274
modern	15.433	1637	highest	-8.458	905
brazilian	15.308	305	vanderlyn	-8.343	260
cust	15.244	562	via	-8.3	523
world	15.153	305	former	-8.239	606
pleaseprice	15.105	929	showings	-8.083	399
future	15.074	459	savvy	-8.009	625
virginia	14.999	318	started	-7.815	411
lightfilled	14.947	304	beat	-7.683	275
brookhaven	14.906	1198	considered	-7.538	441
chef	14.747	349	due	-7.534	874
bestassets	14.283	2641	banks	-7.511	489

Table A5: Repeat Sales: Top 50 Positive and Negative Bigram Coefficients in $1,000s

Positive Words	Coefficient	Counts	Negative Words	Coefficient	Counts
new construction	45.867	2177	auction terms	-33.512	556
built by	37.694	508	highest best	-21.887	766
complete renovation	28.428	557	short sale	-21.573	8032
brick stone	26.569	761	use approved	-17.207	548
granite stainless	24.415	2695	great opportunity	-16.566	3108
chefs kitchen	24.092	2367	fixer upper	-16.376	1519
completely renovated	21.6	2107	sold asis	-15.587	14400
custom home	21.465	983	needs work	-15.234	805
custom kitchen	21.391	686	email address	-15.063	459
beautifully renovated	21.257	949	great potential	-15.026	1154
ceilings main	20.497	491	cash offers	-15.008	642
top line	19.848	846	sold as	-13.5	42981
high end	19.767	764	owner home	-13.484	484
total renovation	19.254	795	cash only	-13.417	655
johns creek	17.841	545	all faults	-13.112	462
granite kitchen	16.549	2114	lots potential	-12.965	1214
coffered ceilings	16.509	460	bank america	-12.874	769
room keeping	16.37	474	one story	-12.51	555
shops restaurants	16.081	492	submit offers	-12.441	457
coffered ceiling	15.808	753	brick bungalow	-11.551	1130
fully renovated	14.495	1043	cute bungalow	-11.004	516
custom cabinets	14.414	1011	no fha	-10.82	565
home one	14.236	498	owned home	-10.789	460
beautiful new	14.213	783	sales price	-10.738	495
highly sought	13.984	493	first look	-10.665	1418
granite new	13.88	493	hardwood under	-10.628	708
renovated bungalow	13.641	695	hud case	-10.396	555
everything new	13.372	518	little tlc	-10.369	1429
granite tops	13.316	621	owner occupant	-10.322	701
room coffered	13.261	524	needs tlc	-10.094	1775
completely remodeled	13.167	539	sweat equity	-10.035	731
appliances granite	12.696	1667	fannie mae	-9.962	2760
cabinets granite	12.679	1904	home needs	-9.824	993
custom built	12.023	2387	excellent opportunity	-9.74	2825
totally renovated	12.005	1956	financing type	-9.702	512
bells whistles	11.989	842	investor special	-9.653	999
home sits	11.685	944	seller requests	-9.645	1471
room study	11.603	692	investment opportunity	-9.177	2289
stucco home	11.576	778	home tons	-9.097	528
stainless appliances	11.56	3903	fha va	-9.043	489
covered porch	11.465	1029	cozy fireplace	-9.031	1673
renovated ranch	11.2	823	repair warranty	-9.008	842
above ground	11.189	542	lender letter	-8.964	463
home will	10.787	751	house great	-8.882	678
keeping room	10.709	5152	great deal	-8.788	1626
renovated kitchen	10.7	2579	bank owned	-8.712	7374
gorgeous kitchen	10.685	663	downtown atlanta	-8.654	603
renovated brick	10.644	539	brick story	-8.624	954
beautifully updated	10.585	697	nice lot	-8.551	677
stainless steel	10.555	10986	lowest price	-8.413	467