Forecasting Existing Home Sales using Google Search Engine Queries

Brian D. Humphrey

Professor James W. Roberts, Faculty Advisor

Honors Thesis submitted in partial fulfillment of the requirements for Graduation with Distinction in Economics in Trinity College of Duke University.

Duke University
Durham, North Carolina
2010

Acknowledgements

I would like to thank Professor James Roberts for his valuable assistance throughout the year. Professor Kent Kimbrough also did a tremendous job leading the Economics Honor Seminar and providing feedback. Additionally, I would like to thank my seminar classmates for their many useful comments. I also appreciate Professor Emma Rasiel's advice. Special thanks to Joel Herndon, Mark Thomas, and Kofi Acquah for their assistance with STATA, data issues, and various research questions. Finally, I would like to thank my parents for their support. These individuals played a critical role in the development of this paper. Any errors are my own.

Abstract

This paper employs OLS regressions to determine whether Google search query data improves national and local existing home sales forecasts. The local dataset features metropolitan statistical area data from Texas. Initially, the national and local regressions are estimated without macroeconomic variables. Macroeconomic variables are subsequently included in order to determine if Google search queries provide information not already present in the macroeconomic variables. The impact of the Google variables is assessed using root mean squared error, p-values, and adjusted r-squared values. Finally, the top models are compared using out-of-sample testing. Both the in-sample and out-of-sample test results suggest that Google search query data improves national and local existing home sales forecasts.

I. Introduction

"If you can look into the seeds of time, and say which grain will grow and which will not, speak then unto me" – William Shakespeare

Shakespeare's words remain true today. Forecasts are used to predict everything from interest rates and the stock market to elections and the weather. Individuals use these predictions to shape expectations and formulate decisions. Yet, despite advances in forecasting techniques, forecasting remains an inexact science. Researchers continue to search for new methods that improve forecasts, enabling better decision making. One new approach incorporates Google search engine query data into econometric models in an attempt to improve forecasts of future events or data releases. The real-time availability of the Google data makes this approach particularly interesting. In this vein, this paper examines the relationship between Google search queries and existing home sales at the national and local level.

Theoretically this approach is related to the field of information search. The information search process has been an important area of economics since 1961 when Nobel Laureate George Stigler published *The Economics of* Information. In his seminal work, Stigler demonstrates that information is a valuable resource that requires a cost, generally time, to acquire. Stigler also notes that "the larger the fraction of the buyer's expenditures on the commodity…the greater the amount of search" (Stigler 219). Thus, individuals are likely to search relatively longer for a home than a less expensive good.

The internet, which has revolutionized the information search process by dramatically lowering the cost of acquiring information, is one likely information source for real estate market participants. In fact, according to the 2009 USC Digital Future

Report, the internet is the most important information source for the 80% of Americans that are internet users. With regards to real estate information, a recent Google search for "real estate" returned around 361 million results indicating the internet is a fertile source for real estate information.

The vast amount of information online necessitated the development of an efficient search tool – the search engine. Search engines, the primary method of locating information online, employ an algorithm to scour the internet for information and return results relevant to a user's query. Analyzing search engine queries provides insights into the type of information individuals are seeking online, and the limited existing literature on the subject suggests that these insights may help predict future behavior and forecast upcoming economic data releases. Search engine queries could help forecast future behaviors or data releases if individuals conduct an online search for information prior to making a decision or taking an action. For example, an individual might research cars for several weeks prior to purchasing a vehicle. In other instances, search engines could be used to predict future data releases but not future demand. For example, increased search queries for an airline's ticketing website might indicate increased ticket sales prior to the airline releasing quarterly financials. However, the majority of airline ticket purchases likely occur the day of the search and not in the future. This is a nuanced delineation, and the underlying takeaway is that search engine queries may help predict future data releases.

Despite having promising forecasting potential, search query data was not publicly available until the launch of Google Insights and Trends in 2008. Google Insights provides data detailing relative changes in the search volume of user specified

Google search queries from 2005 to present. The number of queries is staggering with around 13.6 billion searches occurring in July 2009 alone. Google Insights is especially useful because Google controls 65% of the search engine market (comScore).

Although Google Insights data has not been widely incorporated into economics papers, recent articles by Choi and Varian (2009), Askitas and Zimmerman (2009), and Ginsberg et al (2009) have demonstrated that Google search data can help eliminate lags and improve forecasts for home sales, unemployment, and the spread of the flu. These papers provide the theoretical framework for my research on the predictive value of Google search queries in the housing market. Out of these papers, only Choi and Varian examine the relationship between home sales and Google predictors.

Although Choi and Varian (2009) develop a national home sales prediction model using Google search data, their work has several limitations. First, the study is intentionally simplistic with the goal of encouraging future research. While Choi and Varian accomplish this goal, the duo's model includes no macroeconomic variables other than home price and lagged home sales. Other researchers, namely Dua and Smyth (1995), Dua and Miller (1996), and Dua, Miller, and Smyth (1999), demonstrate that macroeconomic variables can be used to forecast existing home sales. Given that Google search queries are assumed to be impacted by macroeconomic events, it is possible that the Google search query variables may simply include information that is contained in publicly released macroeconomic indicators.[1] Additionally, Choi and Varian only focus on national home sales. This ignores the local nature of the housing market.

Improving prediction models for the housing sector has significant implications. The BEA estimates the housing sector represented 11.5% of U.S. GDP in 2007.

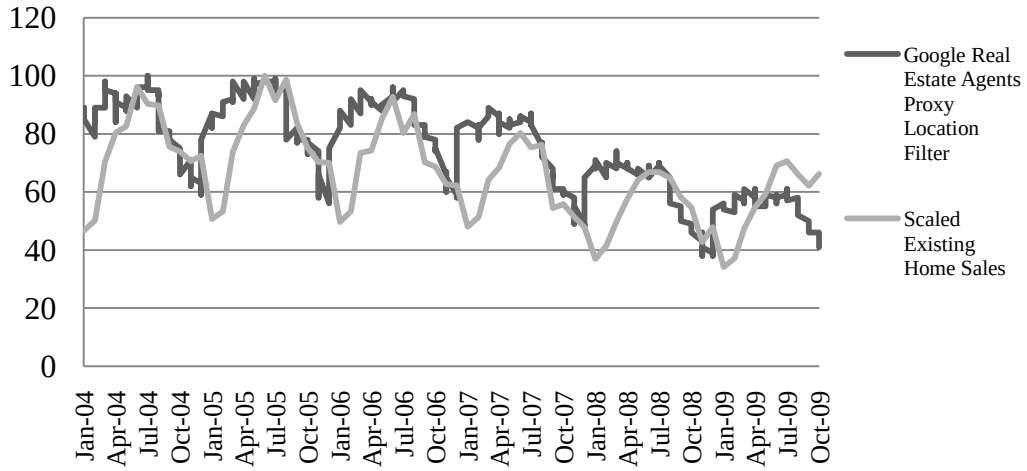
[1] For example, increased unemployment might lead to more unemployment related queries.

Additionally, the collapse of the housing bubble and subprime mortgage crisis played a substantial role in the recent global financial crisis and recession. Improving national home sales forecasts could help policy makers, lenders, and individuals better respond to future adjustments in the housing market. While improving national forecasts could have substantial implications, housing's stationary nature makes it a local market in many respects. Improving local housing forecasts could enable local lenders, realtors, and builders to better respond to future changes. Additionally, the home is the largest individual investment for most American homeowners and improved forecasts would provide individuals with more information, hopefully leading to better decisions. In sum, both the local and national housing markets are important to Americans, the American economy, and the global economy. Finally, this paper also tests the theory that Google search query data is a useful forecasting tool for macroeconomic indicators. This has implications for a wide range of organizations interested in forecasting indicators or events such as unemployment, sales, or even elections.

This paper employs OLS regressions to determine whether Google search query data improves national and local existing home sales forecasts. The local analysis is conducted using metropolitan statistical areas in Texas because the Texas A&M Real Estate Center has compiled an extensive set of historical local real estate data from Texas. Comparable historical data from other states is less available and often controlled by local real estate associations. Initially, the regressions are tested without macroeconomic variables. Later, macroeconomic variables are included to determine if Google search queries provide information not accounted for by the macroeconomic variables. The impact of the Google variables is assessed using root mean squared error,

p-values, and adjusted r-squared values. Finally, the models are compared using out-of-sample testing. The hypothesis is that Google search queries will improve existing home sales forecasts. Figure 1 graphically illustrates the correlation between existing home sales and a Google search query variable at the national level.

Figure 1 U.S. Existing Home Sales vs. Google Variable



Sources: National Association of Realtors (NAR) and Google Insights

Figure 1 indicates the two variables are correlated throughout the period. This supports the hypothesis.

This paper's results overwhelmingly support the inclusion of the Google variables in the forecasting models. The Google variables improve the RMSE and adjusted r-squared of models with and without macroeconomic variables at both the local and national level. Additionally, many of the Google variables are statistically significant. Perhaps most importantly, the model incorporating the Google terms outperforms the non-Google models in out-of-sample testing at both geographic levels. These results suggest national and local existing home sales forecasts can be improved by incorporating Google variables.

Section II will analyze statistics on the internet and review the existing literature in order to support the theoretical framework. Section III will develop the theoretical framework, while Section IV will discuss the data. Section V will discuss the empirical specifications and results. Section VI will conclude the paper. Additional results are in the Appendix.

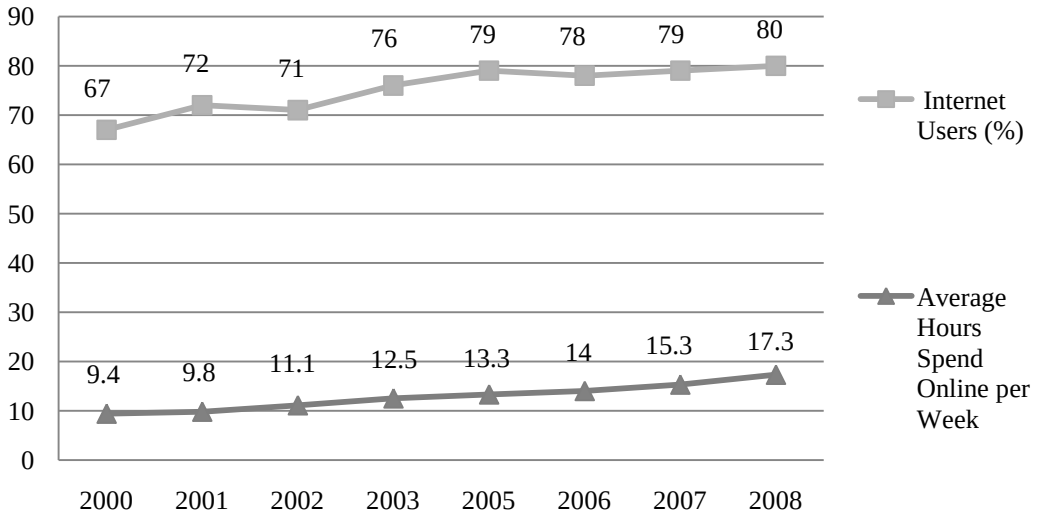
II. Background and Literature Review

A. Background

This section provides background information on the internet, the reliability of internet data, and Google's role in internet search. The goal of this section is to support the theoretical framework by demonstrating that the internet is a trusted and widely used information source. The section also seeks to show that Google search queries are a good proxy for online information search. Additionally, this section examines how my empirical results could potentially be affected by changes in internet usage from 2004 to 2009. Unfortunately, this section primarily relies on national statistics because there is limited data available on Texas internet usage.

The internet has transformed society and revolutionized information search by allowing individuals to efficiently access large quantities of information. Figure 2 illustrates that in the United States the number of internet users and the time spent online has increased since 2000.

Figure 2 United States Internet Usage



Source: 2009 USC Digital Future Survey

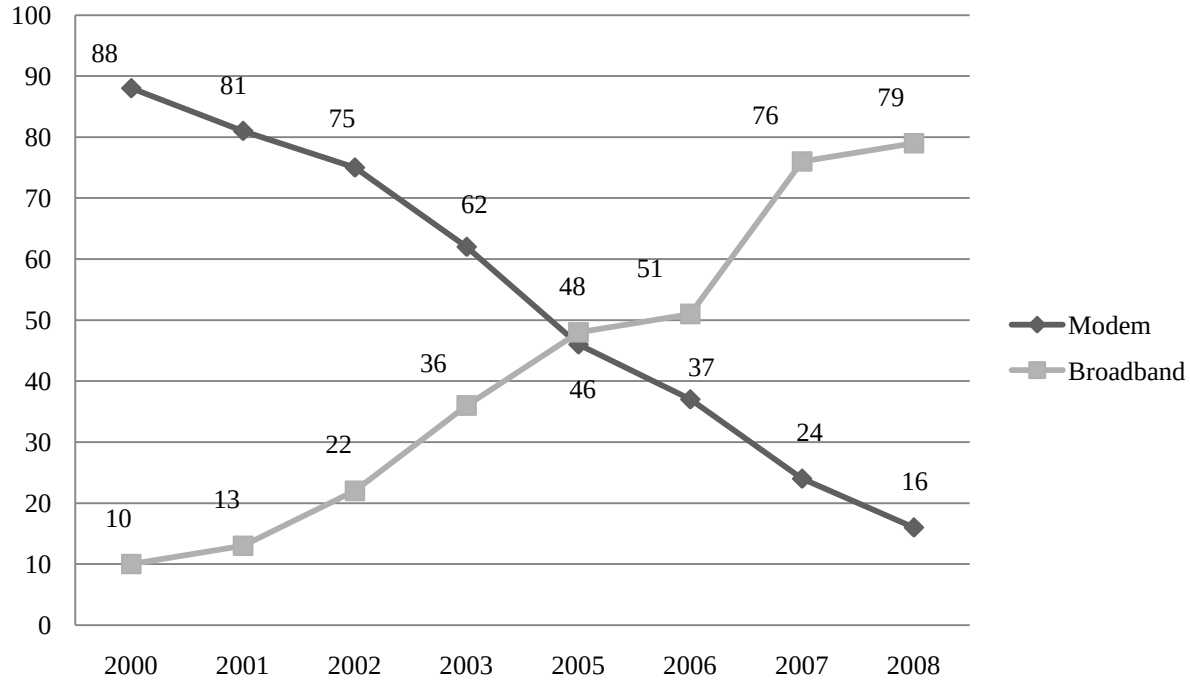
According to the annual Digital Future Survey conducted by the University of Southern California, 76% of Americans were internet users in 2003 – spending an average of 12.5 hours per week online.[2] The percentage of American internet users reached 80% in the 2008 study with the average time online climbing to 17.3 hours per week. According to the U.S. Census Bureau, 54.7% and 67.1% of U.S. homes had the internet in 2004 and 2007 respectively (U.S. Census). Texas lagged the national average, but displayed a similar pattern with the household figure increasing from 51.8% to 60.3% over the same period. The changes in internet usage should not significantly impact the empirical results for several reasons. The U.S. Census data shows that from 2003 to 2008 the number of national households with internet access increased by 12.4%, but the number of national users only increased by 4%. This suggests most of the household adopters already had internet access through other sources by 2003. Furthermore, the 4% change in national

[2] The Digital Survey does not list any data from 2004.

users is quite small. These changes are further mitigated because the Google data has been scaled and normalized to track relative changes in search queries over time.[3]

The Digital Future Survey also found a dramatic change in internet connection types from 2000 to 2008. Figure 3 shows internet connections shifted from phone modems to high-speed broadband.

Figure 3 U.S. Internet Connection Type (%)



Source: 2009 USC Digital Future Report

In 2003, 62% of users connected via a telephone modem while 36% used broadband. In 2008, broadband accounted for 79% of internet connections, while telephone modems accounted for only 16% of connections. This change has had two significant impacts. First, broadband is generally always on eliminating the slow process of dialing-up and connecting via a modem. This makes internet usage more convenient and information search less costly in terms of time. Second, the proliferation of high-speed internet has

[3] The data section discusses this in detail.

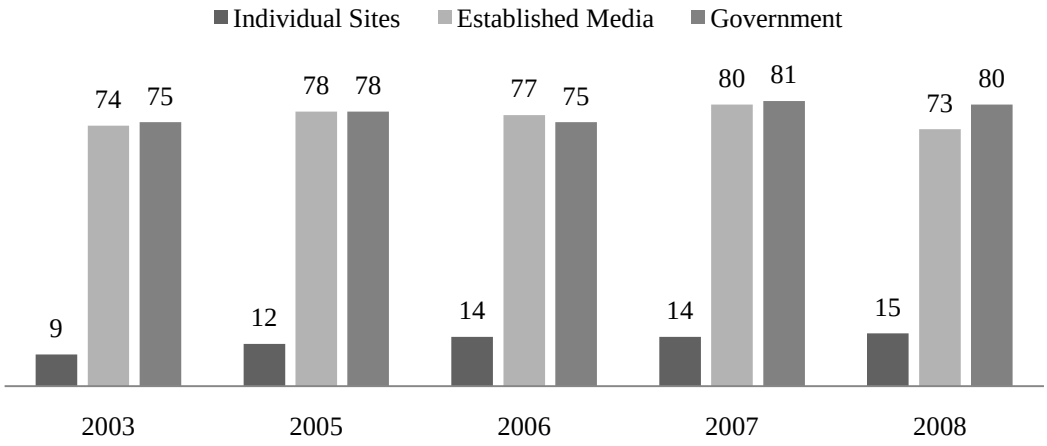
revolutionized available content leading to streaming videos and large image files. These capabilities would be particularly useful to individuals looking to research and sell homes online.

One concern is the rise of broadband might weaken the forecasting power of Google search queries if increased real estate searches are not the result of growing demand but instead are the result of better content and lower search costs arising from the faster broadband connection. Given the relative nature of the Google Insights data, these changes would have to be specific to the real estate sector – in effect the internet becomes a more valuable information source specifically for the real estate sector. However, this weakness is mitigated by the fact that other areas of information search likely experienced similar transformations.[4]

The internet contains massive amounts of unverified information from a variety of sources. Rational users will only use the internet as an information source if they believe the data is trustworthy; thus, it is important to assess the perceived reliability of the internet as an information source. Figure 4 shows the perceived reliability of different categories of websites.

[4] For example, assume that prior to the widespread adoption of broadband, 1 out of 10 searches involves real estate information. If the adoption of broadband increases the quality of online real estate information and reduces the cost of acquiring the information, then searches would be expected to rise even without a change in demand. If no other areas experienced a similar transformation then real estate searches might now represent 2 out of 11 searches – a significant increase. However, it is more likely that the quantity of searches in other areas also rises from broadband adoption for similar reasons. In this case, assume there are now 20 searches with 2 focusing on real estate. Since the Google search data measures relative changes in search queries there is no impact.

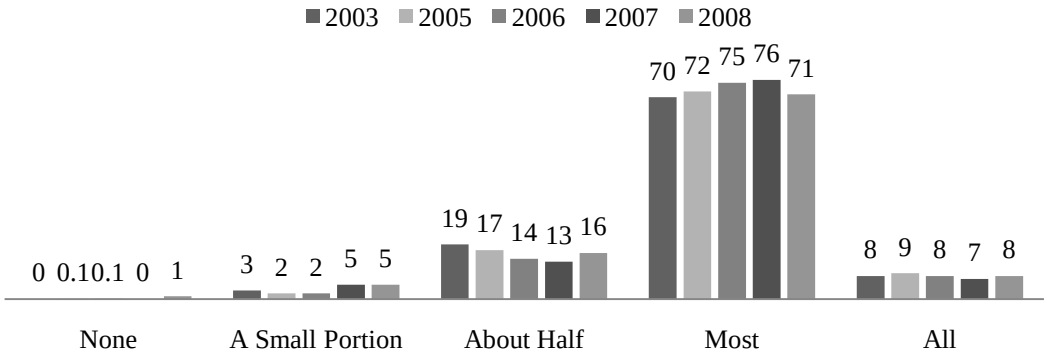
Figure 4 Percieved Web Page Reliability (%)



Source: USC Digital Future Report 2009

Although in 2008 only 15% of internet users believed that most or all of the information on the websites of individuals was accurate, 73% and 80% of users trusted the information on established media and government sites respectively. Historically, these numbers have remained fairly constant. This data indicates that people generally trust established web sites. Figure 5 shows this is especially true for sites that users regularly visit.

Figure 5 Reliability of Regularly Visited Websites (% of Users)



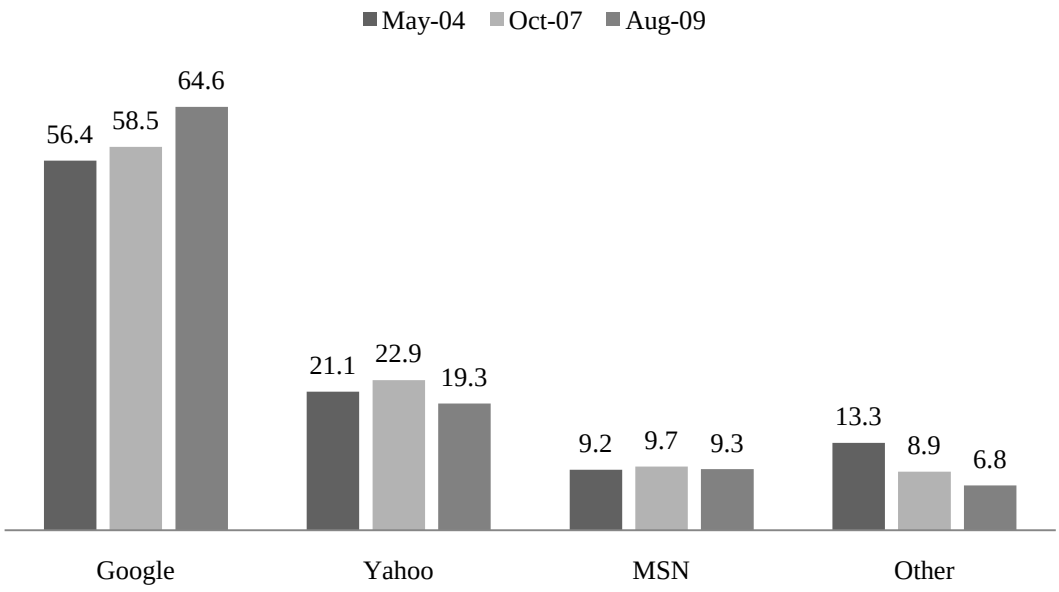
Source: USC Digital Future Report 2009

In 2008, 79% of internet users believed that “most” or “all” of their regularly visited websites were reliable. Only 6% felt that “none” or “a small portion” of these websites

were reliable. These surveys suggest that while individuals realize the internet contains large amounts of unverified and potentially inaccurate information, generally individuals trust established sites and visit sites they feel are reliable. Given that most real estate information is hosted on corporate or established sites, this data suggests individuals would consider online real estate data reliable.

By 2008, the internet had become the top information source for internet users (Digital Report 2004). Search engines, such as Google, are the primary method of locating information on the internet. Search engines aims to improve search efficiency (reduce the time cost) by taking a user submitted search term and applying an algorithm to return relevant search results. The volume of search queries is enormous with Americans conducting 13.6 billion search engine queries in July 2009 alone (comScore). From 2004 to 2009, Google remained the dominant leader in the search market, and the top 3 search engines remained the same with a combined share that increased from 86.7% in May 2004 to 93.2% in August 2009. Figure 6 indicates that Google's market share has steadily increased since 2004.

Figure 6 U.S. Search Engine Market Share (%)



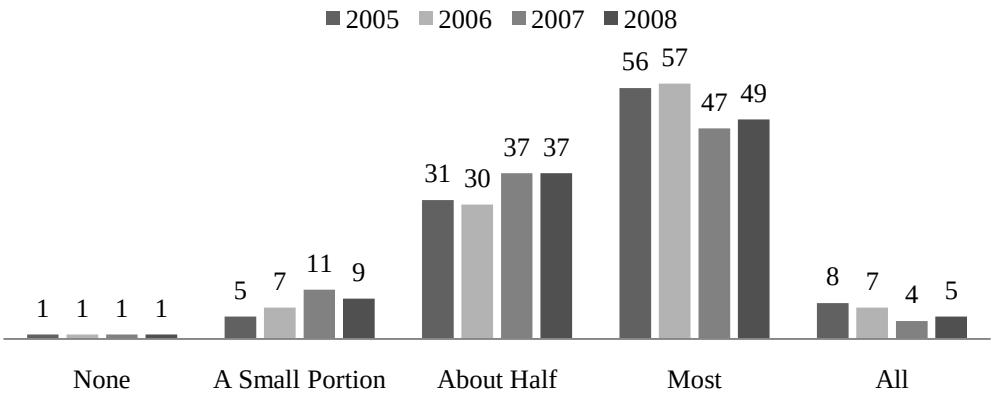
Source: OneStat and comScore

Google had a 56.4% market share in May 2004, while Yahoo had a 21.1% share and MSN controlled 9.2% of the market (OneStat). Google's share in May 2004 was up slightly from its 56.1% share in November of 2003. According to comScore, Google possessed a 58.5% share in October 2007, while Yahoo controlled 22.9% of the market and Microsoft had a 9.7% share. Google's share increased to 64.6% in August 2009.[5] In August 2009, Yahoo had 19.3% share and Microsoft controlled 9.3% of search share. Google's increase in market share suggests that the predictive power of the Google Insights data may have also increased over time because Google accounts for a larger percentage of searches and thus is more representative of behavior. All in all, Google's dominant market share suggests it is a good proxy for national information search. Unfortunately, no Texas specific search engine market share data is available; however, there is nothing that suggests Google is not the most popular search engine in Texas.

[5] Data is from comScore press releases. The press releases are available online at http://www.comscore.com/Press_Events/Press_Releases.

Google's market share coupled with the large number of search engine queries suggests that users trust search engine results – especially those from Google. Figure 7 supports this theory.

Figure 7 Reliability and Accuracy of Search Engine Results



Source: USC Digital Future Report 2009

In 2008, only 10% of Americans felt that "none" or "a small portion" of search engine results were reliable; this figure is up slightly from 6% in 2005. Thus, in addition to being efficient, search engines are viewed as a generally reliable method of gathering trustworthy information online. This section supports the theoretical model developed in Section III by showing that the internet is a popular information source trusted by most users. Additionally, the statistics show Google is the dominant search engine and thus a good proxy for levels of online information search.

B. Literature Review

Search query data was not made publicly available until 2008; thus, the literature on search engine predictors is very limited. However, the internet has proven a valuable research medium in other areas of economic research such as auction theory. Houser and Wooders (2006) examine the impact of seller reputation on auction prices using the

online auction site eBay. Lucking-Reiley, Bryan, Prasad, and Reeves (1999) examine eBay auctions to analyze the determinants of coin prices, while Bajari and Hortascsu (2000) study eBay profit margins, bidding behavior, and the winner's curse. These papers show the internet is a fertile information source for economic research.

Research involving search engine queries became possible in 2008 when Google released Google Insights, a service that tracks relative changes in Google searches dating to 2004 (Google Blog). While no other search engines have released search query data, Google's dominant market share, estimated at 65%, makes it the best single source for query data. Although Texas specific search engine market share data is not publicly available, there is nothing to indicate that Google's share in the state would vary significantly from the national average.

Ginsberg et al (2009) demonstrate that search query data lacks the lags typical in other statistics in *Detecting Influenza Epidemics Using Search Engine Query Data*; the implication is that search data may allow more informed health responses. Choi and Varian (2009) confirm the validity of this theoretical framework in their work *Predicting the Present with Google Trends*. The authors suggest that Google queries "may be correlated with the current level of economic activity in given industries and thus may be helpful in predicting subsequent data releases" (Choi and Varian ii). In effect, searches for unemployment in June may help predict June unemployment data that is not released until the July. Additionally, search query data may also predict future behavior. For example, an individual might research cars before making a purchase.

This theoretical framework forms the basis of several other works in the Google prediction field. Askitas and Zimmerman's (2009) work *Google Econometrics and*

Unemployment Forecasting finds that the Google search data for certain keywords is strongly correlated to German unemployment. Askitas and Zimmerman call this relationship a "Google Predictor". However, Askitas and Zimmerman do not confirm the validity of their results with out-of-sample tests. In *Predicting Initial Claims for Unemployment Benefits*, Choi and Varian (2009) conduct a similar analysis for U.S. unemployment. They find that including Google Trends in a basic regression model reduces the mean absolute error by 15.74%. Similarly, Suhoy (2009) studies the relationship between search queries and macroeconomic cycles. These papers provide additional support for the theoretical framework by showing that Google search data may improve forecasts of economic data.

There has been little analysis of Google search queries and the housing market. Choi and Varian (2009) conduct a brief analysis of home sales using Google data. They find that real estate sales are best predicted by queries within the "Real Estate Agencies" category of Google Trends. The authors construct a simple regression model that features the previous period's home sales (t-1), the Google Trends index for "Rental Listings and Referrals", the Google Trends index for "Real Estate Agencies", and the average home price. Choi and Varian determine that the Google "Real Estate Agencies" index is positively correlated with home sales and the Google "Rental Listings and Referrals" index is negatively correlated with home sales. Additionally, they find that including Google Trends data in the regression model reduces mean absolute error by approximately 12%. Choi and Varian demonstrate that Google Trends can improve existing home sales prediction models. This paper will employ Choi and Varian's model

as a guide and expand the regressions to include additional macroeconomic variables, other Google terms, and Texas MSA data.

Other papers examine the housing market using more traditional approaches. Dua and Miller (1996) suggest that the Connecticut housing market is regional because national economic factors were not significant predictors of home sales. Other economics papers including Blackley and Follain (1990) and Clapp and Giaccotto (2002) analyze housing at the metropolitan level, and the literature suggests that economists generally view housing as a local issue. Thus, this paper will examine whether including search query data improves the existing home sales prediction models at the MSA level.

The more traditional literature also models home sales using macroeconomic factors. Dua and Smyth (1995) develop a Bayesian vector autoregressive (BVAR) model to predict home sales. The model includes mortgage rates, unemployment rates, real income, home prices, and buyer attitudes. Dua and Miller (1996) develop a similar BVAR model for Connecticut that includes "coincident and leading employment indexes" (Dua and Miller 220). The coincident index contains "total unemployment rate, the insured unemployment rate, nonfarm employment, and total employment", while the leading index includes workweek data, short-term unemployment stats, initial unemployment insurance claims, help-wanted advertising levels, and homebuilding permits (Dua and Miller 220). The indices are designed to measure the current and future state of the economy. These studies demonstrate macroeconomic data is useful for predicting home sales.

This study is the first paper to comprehensively examine the predictive relationship between Google search queries and existing home sales. In doing so, this

work expands upon the existing literature in several important ways. First, this paper combines Google search query data and a range of macroeconomic data into a single regression model. Second, this study analyzes both local and national data – this is especially important because housing is a local market. Third, this work tests more Google variables and takes a more systematic approach to empirical specification than the existing literature. Finally, this study employs extensive out-of-sample tests, the gold standard of forecasting accuracy, to validate the results.

III. Theoretical Framework

Information is a valuable resource used in the decision making process. Stigler (1961) examines the information search process in his seminal work titled *The Economics of Information*. Stigler begins by noting that information is a valuable resource that requires a cost, generally time, to acquire. He also states that "the larger the fraction of the buyer's expenditures on the commodity, the greater the savings from search and hence the greater the amount of search" (Stigler 219). Additionally, a greater dispersion of prices also leads to increased search. Although Stigler does not explicitly mention the relationship between preferences and price, a related corollary would be that individuals maximize utility by searching longer to ensure that more expensive durable goods meet their preferences. For example, an individual is likely to spend more time searching for a house than a stapler for several reasons including the potential price savings from the housing search and the fact that homes are more heterogeneous than staplers. Moreover, housing almost certainly has a larger impact on the individual's utility. In summary, expenditure on search will increase for goods that are more expensive relative to expenditures or heterogeneous and important in terms of utility.

The internet is a new and efficient search tool that lowers search costs. According to the 2009 USC Digital Future Report, the internet is the most important information source for the 80% of Americans that are internet users, topping television, newspaper, and radio. Search engines such as Google are the primary method of locating information online.

Prior to 2008, search engine firms considered data on search queries proprietary, and search data was publicly unavailable. However, in 2008, Google launched Google Insights, a service that provides a record of relative changes in Google searches from 2004 to present. Ginsberg et al (2009) demonstrates that search engine data lacks the typical reporting lags present when tracking the flu. In this instance, flu-infected individuals use search engines to search for flu related terms; thus, changes in the volume of flu related search queries offers immediate insights into the spread and intensity of the virus. Choi and Varian (2009) apply this theoretical search framework to various economic issues and assert that Google queries "may be correlated with the current level of economic activity in given industries and thus may be helpful in predicting subsequent data releases" (ii). For example, an individual may research a car online in June before making a purchase in July. Stigler's work implies that longer research periods would accompany more expensive goods.

This paper applies these theoretical models to the housing market. Homes are a significant expense relative to income. According to the U.S. Census Bureau and the NAR, the median U.S. home price in June 2009 totaled $181,000, compared to the 2008 U.S. real median income of $50,303. Additionally, housing is a heterogeneous good, and

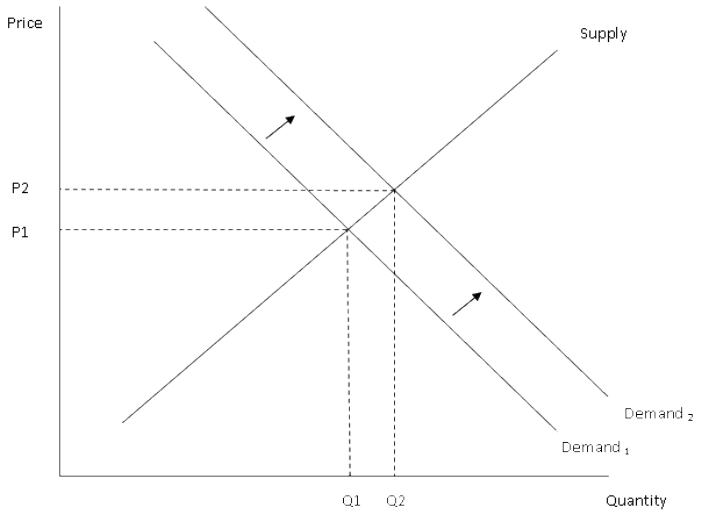
housing choices are often influenced by an individual's preferences.[6] Stigler's work suggests that individuals research housing for longer periods of time than less costly goods.

Given the internet's importance as an information source, it is probable that many individuals conduct online research prior to buying or selling a home. In particular, search engines could be used to gather information on a range of topics such as home prices, realtors, and homes for sale. Aggregating these individual Google search queries provides data on changes in the relative level of housing market information search. The theoretical model assumes that changes in information search indicate future shifts in supply and demand.

The relationship between search queries and home sales can be explained graphically using supply and demand curves. Initially, the housing market is at equilibrium (Q1,P1). Changes in the quantity of existing home sales would be caused by shifts in either the supply or demand curve. This paper examines whether relative changes in search quantities lead to improved predictions of supply and demand curve shifts. Housing market search engine queries could be conducted by either home buyers or sellers attempting to gather information. If home buyers increase the number of searches that are positively correlated to home sales, then this indicates that information search related to housing has increased. It is expected that this indicates a future outward shift in the demand curve leading to an increase in the quantity of housing (existing home sales) as shown in Figure 8.

[6] Examples would include specific styles or locations.

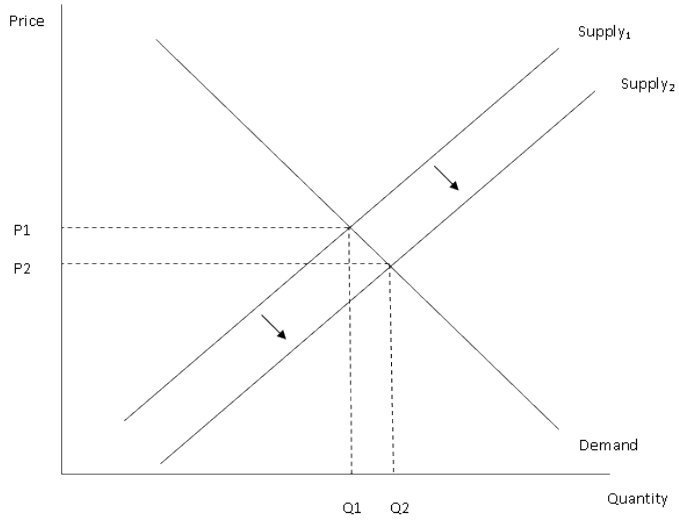
Figure 8



The equilibrium point shifts from (Q1,P1) to (Q2,P2) and the number of existing home sales increases from Q1 to Q2. If relevant searches that are positively correlated to home sales fall, indicating a decline in housing related information search, then the opposite shifts would be expected and future existing homes sales would be expected to decline.

Similarly, if home sellers increase the number of searches that are positively correlated to home sales, this indicates an increase in housing related information search. This increase suggests a future outward shift in the supply curve leading to an increase in the quantity of housing (existing home sales) as shown below in Figure 9.

Figure 9



The equilibrium point shifts from (Q1,P1) to (Q2,P2) and the number of existing home sales increases from Q1 to Q2. If relevant searches that are positively correlated to home sales fall, indicating a decline in housing information search, then the opposite shifts would be expected and existing homes sales would decline.

Thus, an increase in positively correlated search queries would indicate an increase in housing information search; this increase in housing information search would be expected to precede an increase in the quantity of existing home sales regardless of whether the searches were conducted by buyers or sellers.[7] Similarly, a decrease in positively correlated searches would be indicative of a decline in housing information search. This decline would be expected to precede a decrease in the quantity of existing home sales regardless of whether the searches were conducted by buyers or sellers. Therefore, it is not necessary to isolate searches by buyer or seller. The change in buyer

[7] For example, prospective buyers or sellers could both search for real estate agents prior to interacting in the market.

and seller searches on prices has a competing effect; thus, the impact on prices is not as apparent and this topic will be left to future researchers.

IV. Data

This paper requires Google search query data, housing data, and macroeconomic data at both the national and MSA levels. The cities in the Texas MSA dataset are listed in Table 19 in the Appendix.[8]

A. Google Data

The Google search engine query data is publicly available for download on the Google Insights webpage. The data is available in weekly intervals beginning in January 2004. This paper features data collected through November 2009. Google stores, collects, and produces the data. In order to protect the proprietary nature of absolute search query volumes, the data is scaled and normalized. Results can be filtered geographically at the national, state, and local level based on the searcher's IP address.

Google Insights allows users to either enter their own search terms (ex. "Dallas Real Estate") or use Google created categories and subcategories (ex. "Real Estate Category"). These two approaches will be respectively referred to as Google Terms and Google Categories. These two options are substantially different and merit separate discussions.

Google Terms are based on user inputted search terms such as "Dallas Real Estate" or "Houston Realtors". To calculate the Google Terms metric, Google "analyzes a portion of Google web searches to compute how many searches have been done for the terms…entered, relative to the total number of searches done on Google over time".

[8] In a few cases, Google Insights combines several MSA's into a single filter option (ex. Dallas-Fort Worth). This paper deals with this issue by only including the real estate data and macroeconomic factors from the larger city, in this case Dallas.

Google then normalizes the data by dividing by "total traffic from each respective region" (Insights Help). This allows the comparison of data from regions with different levels of absolute search volumes. For example, New York City and Kansas City might show the same value for "rental cars". This does not indicate that the absolute number of searches between the two cities is the same. Instead, individuals in both cities "are equally likely to search for the term" (Insights Help). Google's normalization of the data should account for the increase in search query quantity over time. After normalization, the data is scaled from 0 to 100. Google scales the data by dividing each raw value by the largest normalized number among the set. Thus, if the peak raw value for a search term is 100 and the second highest value is 50% of the maximum, then the second highest value would be assigned a value of 50. Alternatively, if the peak raw value is 80 and the second highest raw value is 40, then the rescaled values would by 100 and 50 respectively. Thus, an increase of one unit of the search variable indicates searches have increased 1% relative to the maximum number of searches conducted over a weekly interval.[9]

For Google Categories, Google automatically groups related search terms into categories and subcategories using a proprietary form of natural language processing. The most relevant example is the "Real Estate" category which has the subcategories of "Home Financing", "Home Inspections and Appraisal", "Home Insurance", "Property Management", "Real Estate Agencies", and "Rental Listings and Referrals". These categories attempt to combine multiple related search terms into a single statistic, and it is the approach used by Choi and Varian (2009) in *Predicting the Present with Google Trends*. When applying a category filter Google provides the data in a percentage form.

[9] For example, assume the two raw values are 40 and 80; thus, the standardized numbers are 50 and 100. If the lower value 40 increases by 1% of 80, then the new raw value is 40.8. To find the scaled Google variable value 40.8 is divided by 80 resulting in a value of 51, a one unit increase in the Google variable.

This statistic, titled "Growth Relative to Category", illustrates the "change over time as a percentage of growth, with respect to the first date on the graph" (Insights Help). The first date is given the value of zero and successive time periods show the change in the number of searches relative to overall search growth. Thus, a positive value of 1% indicates that searches for a selected term are growing 1% faster than searches overall. Changes in the Google variable indicate changes in search volume growth relative to overall growth.

This paper examines a total of 10 different search variables. These variables include the "Real Estate" category and subcategories filtered by location, specific terms ("Dallas real estate"), and specific terms filtered by location ("real estate" searches from Dallas, TX). Multiple terms can also be grouped together to track the aggregate change over time. This paper will use a combination of the approaches discussed above. The complete list of the selected search parameters is listed in Table 1.

Table 1 Google Search Terms

Regression Label	Search Terms or Categories Incorporated	Filter(s)
Google Real Estate Agents Proxy Location Filter	Realtor, Realtors, Real Estate Agents, Real Estate Agencies	Location
Google Location Specific Realtor Proxy	*City* Realtor, *City* Realtors, *City* Real Estate Agents, *City* Real Estate Agencies	None
Google Real Estate Proxy Location Filter	Real Estate, Homes for Sales, Homes	Location
Google Locaton Specific Real Estate Proxy	*City* Real Estate, *City* Homes for Sale, *City* Homes	None
Google Specific Realtor Proxy Location Filter	REMAX, Coldwell Banker, Coldwell, Keller Williams, Keller	Location
Google Rental Proxy Location Filter	Rental, Rentals, Rental Listings	Location
Google Real Estate Category	Real Estate Category	Location
Google Real Estate Agents Category	Real Estate Agencies Subcategory	Location
Google Rental Category	Real Estate Rental Listings Subcategory	Location

In Table 1, the word city is a placeholder for the actual location (ex. Dallas). The first variable aggregates the search terms "Realtor, Realtors, Real Estate Agents, Real Estate Agencies" and filters the search by location; thus, only searches from a specific area's IP addresses will be counted. The second variable is similar to the first but the location is accounted for in the search terms and not with a location filter. As shown in Tables 2 and

3, no national level data was collected for the first or second variables because it makes little sense to search for U.S. real estate or U.S. realtors when gathering information on a local real estate market. The third and fourth search variables follow a similar pattern to the search terms discussed above. The fifth search variable includes the terms "REMAX, Coldwell Banker, Coldwell, Keller Williams, Keller" in an attempt to account for direct searches for the major realtors in Texas. Admittedly, these are large statewide real estate firms and individual cities may have a different realtor that is popular. No national level data was collected for this variable because the realtors are from the leading Texas firms. The next grouping includes the terms "Rental, Rentals, and Rental Listings" filtered by location. The general theory behind this term is that renting is a substitute for buying a home; therefore, this variable is expected to be negatively correlated to existing home sales.

The final three search variables are Google Insights categories and subcategories. Google Insights automatically assigns relevant real estate searches to the real estate category using natural language processing. These variables will be filtered by location. Within the "Real Estate" category, there are several subcategories including "Real Estate Agencies" and "Real Estate Rental Listings". The "Real Estate Rental Listings" subcategory would likely have a negative correlation to existing home sales because renting is a substitute for buying. Google automatically classifies relevant search queries within these subcategories. For example, a search for "Home Rentals in Dallas" would likely be automatically classified in the "Real Estate" category and the "Real Estate Rental Listings" subcategory. To prevent multicollinearity Google Terms and Google Categories are never combined in the same regression model.

Although this set of search terms is admittedly arbitrary, the list expands upon Choi and Varian's category based approach and includes more search terms and different filters. Finally, real estate websites such as Zillow.com and Redfin.com are not explicitly included because the Google search query data on these sites does not span from 2004-2009. Google states this is because "there is not enough search volume" (Insights Help).

The Google data has several weaknesses. First, it is scaled and normalized. Second, the data is not available for every search term at every geographic level throughout the time period. These blank values are excluded from the regressions. Another similar issue is that some terms return a string of zeros for certain periods. Google attributes this to insufficient search data. These zeros have also been excluded from the regressions. Lastly, the Google data is available in weekly intervals, while the housing and macroeconomic data are available in monthly intervals. To account for this, each weekly value was attached to each of the seven days in that particular week.[10] Next, the days in each month were summed and divided by the number of days in the month that contain data. The result is an average monthly value.

B. Real Estate Data

The national level real estate data was provided free of charge by the National Association of Realtors (NAR). The NAR is the national trade group for realtors. The NAR collects housing market data, and the organization provides the data to the U.S. Government for official data releases. The NAR provided monthly data on existing home sales and median home prices from 2000 to present. The existing home sales figures are

[10] For a hypothetical example assume the week January 31- February 6, 2010 was assigned the value of 10. After expanding the week each of the days would separately be assigned a value of 10. Next, January 31st is summed with the other January days, while the days that fall in February are summed with other days in February. These sums are then divided by the days in the month that contain data to calculate the monthly value for the Google variable.

not seasonally adjusted. One weakness is that the data does not include properties sold by owner. However, according to the NAR, realtors were responsible for 88% of home sales in 2007 (Lautz 2008). Additionally, the official data releases are based on NAR data, so this discrepancy should not be an issue.

The MSA level analysis will focus on Texas because it has the best available data. The Real Estate Center at Texas A&M University (TAMUREC) provides extensive data on the Texas market.[11] TAMUREC's dataset features monthly data on home sales and median home prices at the metropolitan and state level from 1979 to present. The TAMUREC data is collected from each metro area's Multiple Listing Service (MLS); thus, the San Antonio data is reported by the local San Antonio Board of Realtors. One limitation is that the data only includes properties sold by realtors. However, according to the NAR only 12% of homes nationally in 2007 were sold without a real estate agent (Lautz 2008). Moreover, homes sold by owner are not included in the monthly data releases; thus this should not be a significant problem. Another issue is the data is not seasonally adjusted; however, the empirical specifications are able to adequately account for this issue using lagged home sales values, logs, and non-seasonally adjusted Google variables.

C. Macroeconomic Data

Mortgages, which are secured by property, are the primary method of financing home purchases in the United States. Monthly mortgage rate data has been collected from Freddie Mac. The dataset contains monthly rates for 15-year fixed rate mortgages. The 71 monthly mortgage rate observations average 5.50% and vary between 4.34% and 6.39%

[11] The data is available on the TAMUREC website located at http://recenter.tamu.edu/data/ datahs.html.

with a standard deviation of .55%. This paper will assume that mortgage rates are constant across the United States.

Unemployment is another major macroeconomic indicator that may influence the housing market. National unemployment figures are available monthly from the Bureau of Labor Statistics (BLS). The 69 national observations average 5.64% and vary between 4.4% and 9.8%. The standard deviation is 1.45%. TAMUREC provides monthly local unemployment numbers for each MSA. There are 1349 total MSA observations with an average of 5.36% and a range between 2.8% and 11.5%. The standard deviation is 1.49%.

The Bureau of Economic Analysis (BEA) releases the Consumer Price Index (CPI) each month. This will be used to proxy for inflation at both the local and national level. MSA specific level inflation is not available in monthly intervals; therefore, the monthly national values are utilized for both the MSA and national regressions. The CPI data includes 72 total monthly observations with an average of 203.82, a standard deviation of 10.0, and a range between 185.2 and 220.0.

The population data has been collected from TAMUREC and the U.S. Census. One weakness is that population estimates are only available in yearly intervals at the national and MSA level through 2008. Thus, the regressions will feature population values from the prior year.[12] Admittedly, this is not ideal, but the variable is designed to capture any large population growth that could increase demand. At the national level there are 71 monthly population observations that range between 290 million and 304 million. The local dataset features 1349 population observations with an average of 837,766.5 and a range between 84,989 and 5,087,127.

[12] For example, all of the 2008 monthly observations in Houston will feature the same 2007 population value.

Additional sample statistics can be found in the tables below. Table 2 contains a complete list of sample statistics for the U.S. level data, and Table 3 contains the complete list of sample statistics for the MSA level data.

Table 2 U.S. Level Data Summary Statistics

Variable	Obs	Mean	Std. Dev.	Min	Max
Existing Home Sales	70	502414.3	119618.1	257000	754000
Existing Home Sales (t-12)	70	519857.1	112822.1	278000	754000
Median Home Price	70	204308.6	19009.8	164800	230300
Google Location Specific Real Estate Proxy	0				
Google Location Specific Realtor Proxy	0				
Google Real Estate Agents Proxy Location Filter	71	74.02832	16.18411	41.09678	98.1
Google Real Estate Location Filter	71	71.06433	15.59896	42.06667	97.54839
Google Real Estate Proxy Location Filter	71	75.29111	13.17016	48.45161	97.70968
Google Rental Proxy Location Filter	71	72.01342	12.4929	49.23333	97.96774
Google Specific Realtor Proxy Location Filter	0				
Google Real Estate Category	71	0.0065222	0.1118856	-0.2306452	0.2
Google Realtor Category	71	-0.0239945	0.154218	-0.3622581	0.216
Google Rental Category	71	0.0074044	0.1292847	-0.2074194	0.2306452
15Y Fixed Mortgage Rate	71	5.501796	0.5482642	4.34	6.39
Populaton (y-1)	71	297000000	4738567	290000000	304000000
CPI	72	203.8248	10.0003	185.2	219.964
Unemployment Rate	69	5.637681	1.44793	4.4	9.8

Existing home sales and median home price data is from the NAR. Google data is from Google Insights. The 15-year fixed mortgage rate is from Freddie Mac. Population data is based on U.S. Census yearly population estimates. CPI data is from the BEA and monthly unemployment data is from the BLS. The three Google variables with zero observations were not collected at the national level because they are designed to capture local searches for a local market. For example, an individual is unlikely to search for U.S. real estate when buying a home in Dallas.

Table 3 MSA Level Data Summary Statistics

Variable	Obs	Mean	Std. Dev.	Min	Max
Existing Home Sales	1315	938.8023	1674.009	34	8628
Existing Home Sales (t-12)	1295	954.6486	1689.484	25	8628
Median Home Price	1314	119279.4	24988.63	60000	196400
Google Location Specific Real Estate Proxy	1321	61.83605	12.5714	30.09677	97
Google Location Specific Realtor Proxy	930	52.28191	13.40469	21.5	100
Google Real Estate Agents Proxy Location Filter	1000	53.44659	14.31962	23.67742	100
Google Real Estate Location Filter	1118	59.37586	13.50786	26.36667	100
Google Real Estate Proxy Location Filter	1310	64.25259	12.94579	26.96667	100
Google Rental Proxy Location Filter	1234	57.42356	13.85168	26.5	100
Google Specific Realtor Proxy Location Filter	979	58.22504	12.84243	27	97.75
Google Real Estate Category	1347	-0.009229	0.2596036	-1	0.7709677
Google Realtor Category	1198	0.0033911	0.3699008	-1	1.527419
Google Rental Category	1257	0.0379287	0.4463345	-1	1.48871
15Y Fixed Mortgage Rate	1349	5.501796	0.5445914	4.34	6.39
Populaton (y-1)	1349	837766.5	1313806	84989	5087127
CPI	1349	203.654	9.898434	185.2	219.964
Unemployment Rate	1349	5.359674	1.493275	2.8	11.5

Existing home sales and median home price data is from TAMUREC. Google data is from Google Insights. The 15-year fixed mortgage rate is from Freddie Mac. Population data is based on U.S. Census yearly population estimates and is available on TAMUREC website. CPI data is from the BEA and monthly unemployment data is from TAMUREC.

V. Empirical Methodology

This section examines the empirical specifications, discusses the regression results, and reviews the out-of-sample tests.

A. Empirical Specifications

While simplistic, Choi and Varian's model offers a blueprint for the regression models that follow:

$$\log(y_t) = \beta + \beta_1 \log(y_{t-1}) + \beta_2 x_{1t} + \beta_3 x_{2t} + \beta_4 x_{3t} + u_t \quad (1)$$

where for each time period *t*, $\log(y_t)$ is existing home sales this period, $\log(y_{t-1})$ is existing home sales last period, x_{1t} is the Google variable for the category "Rental Listing and Referrals", x_{2t} is the Google variable for the category "Real Estate Agencies", x_{3t} is the average home price, and u_t is the error term.

This paper will also utilize the Ordinary Least Squares (OLS) regression method. The MSA level data is organized as a panel data set using location as the case and date as the time period. This allows all of the MSA level regressions to incorporate location fixed effects to account for city-specific variances that stay constant over time. The U.S. level data is analyzed as a time series; thus, there are no fixed effects for these regressions. The dependent variable of the specification is logged existing home sales, and the regression model is:

$$Y_{f,t} = \beta_0 + \sum_{i=1}^{n_1} \beta_i H_{f,t-i} + \sum_{j=1}^{n_2} \gamma_i G_{f,t-i} + \sum_{k=1}^{n_3} \delta_i M_{f,t-i} + u_{f,t} \quad (2)$$

where for each geographic level *f* (national or metropolitan statistical area) and time period *t*, Y is existing home sales, H is a vector of home sales data from the previous period, G is a vector of Google Insights variables from prior periods, M is a vector of macroeconomic variables, and *u* is the error term. After testing the models using logarithmic and linear values for the home sales variables, the logarithmic values appear to improve the fit slightly. Thus, all of the regressions feature logged home sales values for both the dependent variable and the 1 and 12 month lagged home sales terms.

Vector H will contain data on home sales for the specified geographic level *f* from the previous period. This vector will include existing home sales and median home price. The vector will feature a t-12 lagged term to account for the seasonal nature of home sales. An additional lagged variable from period t-1 is also included to account for the prior month's sales. Finally, a 3-month lagged median home price variable is included in the proposed regression because home prices are expected to impact sales.

Vector G will contain the Google Insights variables for selected search terms and categories from the specified geographic level *f* during the specified period. The proposed

regression specifies a 1-month lag for the Google variables. The hypothesis is that incorporating the Google variables will reduce the RMSE, increase the adjusted R-squared values, and improve the out-of-sample forecasts of the models.

The model contains a range of macroeconomic variables. The 3-month lagged 15-year fixed mortgage rate is included because mortgages are the primary method of financing home purchases. Therefore, higher (lower) mortgage rates lead to more costly (less costly) home loans, likely decreasing (increasing) the demand for housing. Unemployment is another major macroeconomic indicator that may influence the housing market. Higher unemployment creates uncertainty and decreases income levels. This is expected to decrease the number of people willing or able to purchase a home. On the other hand, unemployed individuals might be forced to sell their homes leading to a competing effect. The 3-month lagged national consumer price index was included to proxy for inflation. Inflation also increases uncertainty; thus, rising inflation is expected to decrease home sales. The prior year's logged population level is included because changes in population are expected to change the demand for housing. The population is logged because it increases over time and is easier to interpret as a percentage. Additionally, an incremental time trend is included in the national specifications. The time trend begins at 1 in January 2004 and increases by 1 unit each month. The variable is designed to account for an underlying trend that was picked up by the population variable in initial tests.[13]

This paper assumes the Google search variables are proxies for human behavior in the housing markets. Given that this behavior is impacted by macroeconomic factors such

[13] The population variable featured a large negative sign. Theoretically, it makes little sense that an increasing population would significantly decrease home purchases. Including the time trend helped rectify this issue.

as unemployment, it is plausible that the Google search queries provide no prediction advantage over macroeconomic variables. Thus the more pertinent question is whether Google search queries provide information that is not available elsewhere. This paper seeks to address this issue by including the previously discussed macroeconomic factors in a second set of regressions in order to ascertain whether the Google search variables provide additional information. The hypothesis is that, even with the inclusion of the macroeconomic variables, incorporating the Google variables will reduce the RMSE, increase the adjusted R-squared values, and improve the out-of-sample forecasts of the models. The effect is more accurate forecasts for housing sales, which will aid policy makers, corporations, and individuals.

In order to determine whether the Google variables contain new information, it is important to have the best possible macroeconomic variable specification. While the model proposed on the preceding pages is based on theoretical considerations and initial results, Tables 4 and 5 feature a more robust and systematic model specification approach. First, the tables list the results from testing the proposed baseline model. The variables in the baseline model are starred. Next, holding all of the other variables constant the lags were adjusted for one variable at a time. For example, holding all else constant the lag on median home price was changed from 3-months to 2-months and then 1-month. This process was repeated for each of the macroeconomic variables. By comparing the resulting model metrics an improved specification was found for both geographic levels. The variables included in the updated model are in bold and the metrics for the adjusted model are listed last in the table. Table 4 features the results from the macroeconomic variable specification at the national level.

Table 4: USA Macro Model Specification

	AIC	BIC	Adj. R-Squared	RMSE
Proposed Model	-181.22	-161.38	0.9515	0.0588
Median Home Price (t-1)	**-183.31**	**-165.68**	**0.9516**	**0.0588**
Median Home Price (t-2)	-181.09	-161.25	0.9514	0.0589
Median Home Price (t-3)*	-181.22	-161.38	0.9515	0.0588
Median Home Price (t-4)	-178.47	-158.76	0.9519	0.0588
Median Home Price (t-5)	-172.39	-152.82	0.9496	0.0603
Median Home Price (t-6)	-168.82	-149.39	0.9478	0.0607
Median Home Price (excluded)	-181.39	-163.75	0.9502	0.0591
15YR Fixed Mortgage (t-1)	-173.72	-156.08	0.9441	0.0631
15YR Fixed Mortgage (t-2)	-174.87	-155.03	0.9467	0.0617
15YR Fixed Mortgage (t-3)*	**-181.22**	**-161.38**	**0.9515**	**0.0588**
15YR Fixed Mortgage (t-4)	-171.04	-151.33	0.9461	0.0622
15YR Fixed Mortgage (t-5)	-167.71	-148.14	0.9458	0.0625
15YR Fixed Mortgage (t-6)	-165.56	-146.13	0.9450	0.0622
15YR Fixed Mortgage (excluded)	-175.69	-160.26	0.9441	0.0626
Consumer Price Index (t-1)	-180.79	-160.94	0.9512	0.0590
Consumer Price Index (t-2)	-180.95	-161.10	0.9513	0.0589
Consumer Price Index (t-3)*	-181.22	-161.38	0.9515	0.0588
Consumer Price Index (t-4)	**-181.28**	**-163.76**	**0.9525**	**0.0584**
Consumer Price Index(t-5)	-176.90	-157.33	0.9529	0.0582
Consumer Price Index (t-6)	-172.44	-153.01	0.9506	0.0590
Consumer Price Index (excluded)	-182.43	-164.79	0.9509	0.0587
Unemployment (t-1)	-184.42	-164.58	0.9538	0.0574
Unemployment (t-2)	-182.45	-162.60	0.9524	0.0583
Unemployment (t-3)*	-181.22	-161.38	0.9515	0.0588
Unemployment (t-4)	**-183.78**	**-166.27**	**0.9542**	**0.0573**
Unemployment (t-5)	-175.62	-156.05	0.9520	0.0588
Unemployment (t-6)	-174.23	-154.80	0.9520	0.0581
Unemployment (excluded)	-172.77	-155.14	0.9433	0.0630
Population (y-1)*	-181.22	-161.38	0.9515	0.0588
Population (excluded)	**-183.21**	**-165.57**	**0.9515**	**0.0583**
Monthly Time Trend	**-181.22**	**-161.38**	**0.9515**	**0.0588**
Monthly Time Trend (excluded)	-181.12	-163.48	0.9500	0.0592
Revised Model	-183.93	-164.23	0.9557	0.0564
Revised Model (exclude time trend and population)	-176.04	-160.72	0.9469	0.0607
Revised Model (exclude time trend)	-181.35	-163.83	0.9525	0.0579
Revised Model (exclude population)	**-185.89**	**-168.37**	**0.9557**	**0.0559**

Data is from U.S. dataset discussed in Data section. First, the table lists the results from testing the proposed baseline model. The variables in the baseline model are starred. Next, the lags were adjusted for one variable at a time holding all of the other variables constant. For example, the lag on median home price was changed from 3-months to 2-months and then 1-month holding all else constant. This process was repeated for each of the macroeconomic variables and the best lags were included in the revised model.

Shifting the median home price lag to 1-month and the consumer price index and unemployment lags to 4-months improves the national model slightly. Population is excluded because doing so improves the AIC, BIC, adjusted R-squared, and RMSE of the updated model. These adjustments improve the model's AIC from -181.22 to -185.89 and

RMSE from .0588 to .0559. Table 5 takes a similar approach for the MSA macroeconomic specification.

Table 5 MSA Macro Model Specification

	AIC	BIC	Adj. R-Squared	RMSE
Proposed Model	-1129.15	-1088.22	0.9872	0.1536
Median Home Price (t-1)	-1131.67	-1090.73	0.9872	0.1535
Median Home Price (t-2)	**-1133.52**	**-1092.60**	**0.9872**	**0.1533**
Median Home Price (t-3)*	-1129.15	-1088.22	0.9872	0.1536
Median Home Price (t-4)	-1109.16	-1068.34	0.9871	0.1540
Median Home Price (t-5)	-1089.96	-1049.26	0.9871	0.1541
Median Home Price (t-6)	-1069.86	-1029.28	0.9870	0.1544
Median Home Price (excluded)	-1131.92	-1096.10	0.9871	0.1536
15YR Fixed Mortgage (t-1)	-1116.38	-1075.46	0.9870	0.1544
15YR Fixed Mortgage (t-2)	-1116.67	-1075.75	0.9870	0.1544
15YR Fixed Mortgage (t-3)*	**-1129.15**	**-1088.22**	**0.9872**	**0.1536**
15YR Fixed Mortgage (t-4)	-1094.94	-1054.12	0.9869	0.1549
15YR Fixed Mortgage (t-5)	-1085.89	-1045.19	0.9870	0.1544
15YR Fixed Mortgage (t-6)	-1063.18	-1022.60	0.9869	0.1548
15YR Fixed Mortgage (excluded)	-1115.55	-1079.74	0.9870	0.1545
Consumer Price Index (t-1)	-1095.59	-1054.66	0.9868	0.1557
Consumer Price Index (t-2)	-1112.75	-1071.83	0.9870	0.1546
Consumer Price Index (t-3)*	**-1129.15**	**-1088.22**	**0.9872**	**0.1536**
Consumer Price Index (t-4)	-1113.97	-1073.15	0.9871	0.1536
Consumer Price Index(t-5)	-1095.00	-1054.30	0.9871	0.1538
Consumer Price Index (t-6)	-1076.97	-1036.39	0.9871	0.1539
Consumer Price Index (excluded)	-1066.96	-1031.15	0.9865	0.1576
Unemployment (t-1)	-1124.62	-1083.70	0.9871	0.15388
Unemployment (t-2)	**-1129.29**	**-1088.37**	**0.9872**	**0.1536**
Unemployment (t-3)*	-1129.15	-1088.22	0.9872	0.1536
Unemployment (t-4)	-1103.01	-1062.19	0.9870	0.1543
Unemployment (t-5)	-1086.55	-1045.85	0.9870	0.1544
Unemployment (t-6)	-1070.95	-1030.37	0.9870	0.1543
Unemployment (excluded)	-1126.38	-1090.57	0.9871	0.1538
Population (y-1)*	**-1129.15**	**-1088.22**	**0.9872**	**0.1536**
Population (excluded)	-1130.75	-1094.94	0.9872	0.1536
Revised Model	-1133.911	-1092.986	0.9872	0.1533

Data is from MSA dataset discussed in Data section. First, the table lists the results from testing the proposed baseline model. The variables in the baseline model are starred. Next, the lags were adjusted for one variable at a time holding all of the other variables constant. For example, the lag on median home price was changed from 3-months to 2-months and then 1-month holding all else constant. This process was repeated for each of the macroeconomic variables and the best lags were included in the revised model.

In the MSA model, adjusting the median home price and unemployment lags to 2-months improves the specification. These adjustments improve the model's AIC from -1129.15 to -1133.91 and RMSE from .1536 to .1533.

After adjusting the macroeconomic variables, a similar analysis is done to determine the appropriate lag length for the Google variables. Tables 6 and 7 contain the results from this analysis for the national models. The Google Category variables are listed separately in Table 7 because they are never simultaneously tested with the Google Term variables in order to prevent multicollinearity.

Table 6 U.S. Google Variable Model Specification

	AIC	BIC	Adj R-Squared	RMSE
Baseline Model	-192.9571	-168.8709	0.957	0.05201
Google Real Estate Agents Proxy City Filter (t-1)*	-192.9571	-168.8709	0.957	0.05201
Google Real Estate Agents Proxy City Filter (t-2)	-189.202	-165.1158	0.9545	0.05352
Google Real Estate Agents Proxy City Filter (t-3)	-187.9599	-163.8737	0.9536	0.05402
Google Real Estate Agents Proxy City Filter (t-4)	-188.3691	-164.2829	0.9539	0.05385
Google Real Estate Proxy City Filter (t-1)*	-192.9571	-168.8709	0.957	0.05201
Google Real Estate Proxy City Filter (t-2)	-192.4084	-168.3222	0.9567	0.05223
Google Real Estate Proxy City Filter (t-3)	-192.4127	-168.3265	0.9567	0.05223
Google Real Estate Proxy City Filter (t-4)	-193.702	-169.6158	0.9575	0.05172
Google Rental Proxy City Filter (t-1)*	-192.9571	-168.8709	0.957	0.05201
Google Rental Proxy City Filter (t-2)	-191.4953	-167.4091	0.956	0.05259
Google Rental Proxy City Filter (t-3)	-190.6654	-166.5792	0.9555	0.05293
Google Rental Proxy City Filter (t-4)	-190.0541	-165.9679	0.9551	0.05317

Data is from U.S. dataset discussed in Data section. First, the table lists the results from testing the proposed baseline model. The baseline model includes the macroeconomic variables from the revised model in Table 4. The variables in the baseline model are starred. Next, the lags were adjusted for one variable at a time holding all of the other variables constant. For example, the lag on Google Real Estate Proxy City Filter was changed from 1-months to 2-months and then 3-month holding all else constant. This process was repeated for each of the Google variables. The optimal lags are in bold.

Table 7 U.S. Google Category Variable Model Specification

	AIC	BIC	Adj R-Squared	RMSE
Baseline Model	-204.1943	-182.2978	0.9633	0.04806
Google Realtor Category (t-1)*	-204.1943	-182.2978	0.9633	0.04806
Google Realtor Category (t-2)	-193.1268	-171.2302	0.9566	0.05227
Google Realtor Category (t-3)	-185.6091	-163.7125	0.9514	0.05533
Google Realtor Category (t-4)	-182.9949	-161.0984	0.9494	0.05644
Google Rental Category (t-1)*	-204.1943	-182.2978	0.9633	0.04806
Google Rental Category (t-2)	-191.6093	-169.7128	0.9556	0.05287
Google Rental Category (t-3)	-192.492	-170.5954	0.9562	0.05252
Google Rental Category (t-4)	-194.6989	-172.8023	0.9576	0.05165

Data is from U.S. dataset discussed in Data section. First, the table lists the results from testing the proposed baseline model. The baseline model includes the macroeconomic variables from the revised model in Table 4. The variables in the baseline model are starred. Next, the lags were adjusted for one variable at a time holding all of the other variables constant. For example, the lag on Google Realtor Category was changed from 1-months to 2-months and then 3-month holding all else constant. This process was repeated for Google Rental Category. The optimal lags are in bold.

Changing the lag lengths of the variables does not improve any of the statistical metrics in either Table 6 or Table 7. Thus, the national level models will feature Google variables with 1-month lags as initially proposed.

Tables 8 and 9 take a similar approach for the MSA specifications.

Table 8 MSA Google Variable Model Specification

	AIC	BIC	Adj R-Squared	RMSE
Baseline Model	-1040.826	-985.054	0.9922	0.12332
Google City Specific Real Estate Proxy (t-1)*	-1040.826	-985.054	0.9922	0.12332
Google City Specific Real Estate Proxy (t-2)	-1010.16	-954.3874	0.9919	0.12579
Google City Specific Real Estate Proxy (t-3)	-996.5499	-940.7776	0.9918	0.12691
Google City Specific Real Estate Proxy (t-4)	-991.0234	-935.3292	0.9917	0.12683
Google City Specific Realtor Proxy (t-1)*	-1040.826	-985.054	0.9922	0.12332
Google City Specific Realtor Proxy (t-2)	-1039.196	-983.4866	0.9923	0.12301
Google City Specific Realtor Proxy (t-3)	-1032.083	-976.4049	0.9923	0.12337
Google City Specific Realtor Proxy (t-4)	-918.6079	-863.0714	0.9912	0.13192
Google Specific Realtor Proxy City Filter (t-1)*	-1040.826	-985.054	0.9922	0.12332
Google Specific Realtor Proxy City Filter (t-2)	-1034.43	-978.7204	0.9922	0.1234
Google Specific Realtor Proxy City Filter (t-3)	-1000.567	-944.9675	0.9920	0.1254
Google Specific Realtor Proxy City Filter (t-4)	-997.9959	-942.555	0.9921	0.12452
Google Rental Proxy City Filter (t-1)*	-1040.826	-985.054	0.9922	0.12332
Google Rental Proxy City Filter (t-2)	-1041.109	-985.3367	0.9922	0.12329
Google Rental Proxy City Filter (t-3)	-1038.997	-983.2248	0.9922	0.12346
Google Rental Proxy City Filter (t-4)	-1029.954	-974.2753	0.9921	0.12354

Data is from MSA dataset discussed in Data section. First, the table lists the results from testing the proposed baseline model. The baseline model includes the macroeconomic variables from the revised model in Table 4. The variables in the baseline model are starred. Next, the lags were adjusted for one variable at a time holding all of the other variables constant. For example, the lag on Google City Specific Real Estate Proxy was changed from 1-months to 2-months and then 3-month holding all else constant. This process was repeated for each of the Google variables. The optimal lags are in bold.

Table 9 MSA Google Category Variable Model Specification

	AIC	BIC	Adj R-Squared	RMSE
Baseline Model	-1071.824	-1021.922	0.988	0.14822
Google Realtor Category (t-1)*	-1071.824	-1021.922	0.988	0.14822
Google Realtor Category (t-2)	-1050.224	-1000.34	0.9877	0.14957
Google Realtor Category (t-3)	-1040.575	-990.7186	0.9876	0.15011
Google Realtor Category (t-4)	-1018.331	-968.6054	0.9875	0.15074
Google Rental Category (t-1)*				
Google Rental Category (t-2)	-1071.707	-1021.832	0.988	0.14803
Google Rental Category (t-3)	-1064.294	-1014.447	0.9879	0.14834
Google Rental Category (t-4)	-1044.768	-995.0512	0.9879	0.14874

Data is from MSA dataset discussed in Data section. First, the table lists the results from testing the proposed baseline model. The baseline model includes the macroeconomic variables from the revised model in Table 4. The variables in the baseline model are starred. Next, the lags were adjusted for one variable at a time holding all of the other variables constant. For example, the lag on Google Realtor Category was changed from 1-months to 2-months and then 3-month holding all else constant. This process was repeated for Google Rental Category. The optimal lags are in bold.

The statistical metrics suggest that the models are well specified in both cases. In Table 8 the statistical metrics are improved slightly by adjusting the lag on Google Rental

Proxy City Filter to 2-months. However, the improvement in AIC is under .1% and the RMSE did not change; thus, to keep the Google variable lags consistent the Google Rental Proxy City Filter lag was not adjusted. The data in Table 9 shows the 1-month lag is appropriate for both Google Category Variables. Tables 4-9 illustrate that the forecasting models are properly specified.

B. Regression Analysis

After ensuring the proper specification of the models, tests were conducted for heteroskedasticity and autocorrelation. Although heteroskedasticity and autocorrelation do not bias the estimators, the presence of either will impact the standard errors of the regressions. The results of these tests are listed in Tables 10-13 in the rows labeled "Homoskedasticity Test" and "Serial Correlation Test". At the national level, White's Test was used to test for homoskedasticity, and the Breusch-Godfrey test was used to test for higher order serial correlation. At the MSA level, a Likelihood-Ratio test was utilized to test for panel level homoskedasticity.[14] Serial correlation is tested using the Woolridge Serial Correlation Test.[15] At both geographic levels, the null hypothesis of the "Homoskedasticity Test" is that the data is homoskedastic, and the null hypothesis of the "Serial Correlation Test" is that the data is not serially correlated. The models that have heteroskedasticity or serial correlation were recalculated using Newey-West standard errors with a lag length of 12. The Newey-West correction does not alter the coefficients.

[14] This is the approach recommended by STATA. Additional information is available online at http://www.stata.com/support/faqs/stat/panel.html.
[15] Test conducted using xtserial command which must be downloaded.

Furthermore, the standard errors and significance levels are largely unaffected; thus, these results are placed in the Appendix Section.[16]

Table 10 contains the OLS estimates for the baseline U.S. level regressions, with each column representing a different model. The major macroeconomic factors are excluded, and no fixed effects are incorporated.

[16] The U.S. Newey-West corrected regressions are in Table 19, while the MSA Newey-West corrected regressions are in Tables 20 and 21.

Table 10 USA Expanded Log Home Sales Models and Robust Standard Errors

	(1)	(2)	(3)	(4)	(5)	(6)
Log Existing Home Sales (t-12)	0.8143***	0.6739***	0.6535***	0.6549***	0.7310***	0.7710***
	(0.0716)	(0.072)	(0.0524)	(0.0549)	(0.0694)	(0.0921)
Log Existing Home Sales (t-1)	0.4109***	0.3668***	0.3712***	0.3711***	0.3849***	0.3351***
	(0.0589)	(0.0553)	(0.055)	(0.0553)	(0.052)	(0.0949)
Median Home Price (t-1)($10K scaled)	-0.0347***	-0.0453***	-0.0405***	-0.0405***	-0.0455***	-0.0423***
	(0.0072)	(0.0071)	(0.0063)	(0.0064)	(0.0071)	(0.0082)
15Y Fixed Mortgage Rate (t-3)						
Consumer Price Index (t-4)						
Unemployment Rate (t-4)						
Month Trend						
Google Realtor Category		0.6410***				
		(0.1213)				
Google Rental Category		-0.3121***				
		(0.0938)				
Google Real Estate Agents Proxy Location Filter (t-1)			0.0045***	0.0042	0.0026	0.0016
			(0.0009)	(0.0042)	(0.0041)	(0.0042)
Google Real Estate Proxy Location Filter (t-1)				0.0004	0.0032	-0.0021
				(0.0048)	(0.0047)	(0.0083)
Google Rental Proxy Location Filter (t-1)					-0.0025**	-0.0018
					(0.0011)	(0.0013)
Google Real Estate Location Filter (t-1)						0.005
						(0.0065)
Constant	-2.2709***	0.3842	0.142	0.1207	-0.8715	-0.7454
	(0.7425)	(0.7975)	(0.635)	(0.7216)	(0.842)	(0.9078)
Observations	69	69	69	69	69	69
Adjusted R^2	0.873	0.907	0.91	0.908	0.913	0.913
AIC	-134.542	-154.2034	-157.2634	-155.2706	-157.9744	-156.8456
BIC	-125.6055	-140.7987	-146.0929	-141.866	-142.3356	-138.9728
Homoskedasticity Test	0.6646	0.1238	0.2875	0.4592	0.2762	0.6033
Serial Correlation Test	0.0027	0.1163	0.2643	0.2675	0.3518	0.2702
Root MSE	0.0887	0.0759	0.0748	0.0754	0.0734	0.0735

*10% Significance Level, **5% Significance Level, ***1% Significance Level

Data is from U.S. dataset discussed in data section. Dependent variable is logged existing home sales. Observations decrease because of availability of lagged variables. Google variables are explained in Table 1. Month trend is a one unit incremental increasing monthly trend line. Median home price is scaled to show the impact of a $10,000 change in home prices. Homoskedasticity test refers to White's Test. The null hypothesis of this test is that the data is homoskedastic. Serial Correlation Test refers to the Breush-Godfrey test. The null hypothesis of this test is the data is not serially correlated.

The table shows that introducing the Google variables reduces the RMSE of the forecasting model. Column 1 is a baseline model that incorporates no Google terms, and

the RMSE of the model is .0887; additionally, all of the terms are significant at the 1% level and feature the expected signs. Column 2 adds the two Google category variables utilized by Choi and Varian. The terms are both significant at the 1% level and serve to reduce the RMSE by 14.43% to .0759. Columns 3-6 drop the category level terms for more specific search queries. The variables are designed to account for realtor, real estate, and rental searches at the national level. The model in Column 3 features the expected signs and the single Google term is significant at the 1% level. The RMSE decreases slightly from Column 2 to .0748. Columns 4 and 5 continue to expand Column 3 by incorporating additional Google variables. Although only the Google Rental Proxy City Filter variable is significant in Model 5, the RMSE falls to .0734. These results indicate that the Google variables are improving the accuracy of the baseline prediction model. The results support the hypothesis that Google search variables improve prediction models for national level existing home sales. Finally, the adjusted R-squared values are very high indicating that the variables explain nearly all of the variance in home sales. The R-squared values can largely be explained by the seasonal autocorrelation of the data. This trend continues throughout Tables 10-13.

Table 11 shows the OLS estimates for the national level models incorporating macroeconomic variables. Each column represents a different model. The models feature no fixed effects.

Table 11 USA Expanded Log Home Sales Models and Robust Standard Errors

	(1)	(7)	(8)	(9)	(10)	(11)	(12)
Log Existing Home Sales (t-12)	0.8143***	0.8943***	0.7501***	0.7937***	0.7686***	0.7707***	0.7608***
	(0.0716)	(0.0618)	(0.066)	(0.0739)	(0.0777)	(0.0759)	(0.0799)
Log Existing Home Sales (t-1)	0.4109***	0.0436	0.0483	0.0712	0.0674	0.1282*	0.1403*
	(0.0589)	(0.0651)	(0.0589)	(0.0693)	(0.066)	(0.07)	(0.0791)
Median Home Price (t-1)($10K scaled)	-0.0347***	0.0159**	0.0127*	0.0127	0.0133*	0.0072	0.0074
	(0.0072)	(0.0075)	(0.007)	(0.0076)	(0.0078)	(0.0087)	(0.0086)
15Y Fixed Mortgage Rate (t-3)		-0.0878***	-0.1324***	-0.0999***	-0.1067***	-0.1233***	-0.1201***
		(0.0212)	(0.0229)	(0.0214)	(0.0226)	(0.0229)	(0.0228)
Consumer Price Index (t-4)		0.008	0.0153***	0.0102*	0.0106*	0.0120**	0.0123**
		(0.0055)	(0.0051)	(0.0053)	(0.0054)	(0.0053)	(0.0055)
Unemployment Rate (t-4)		0.0661***	0.0696***	0.0621***	0.0625***	0.0559***	0.0603***
		(0.0147)	(0.0123)	(0.015)	(0.0145)	(0.013)	(0.013)
Month Trend		-0.0077***	-0.0087***	-0.0076***	-0.0076***	-0.0068***	-0.0081***
		(0.0028)	(0.0023)	(0.0026)	(0.0026)	(0.0024)	(0.003)
Google Realtor Category			0.7629***				
			(0.149)				
Google Rental Category			-0.4170***				
			(0.1174)				
Google Real Estate Agents Proxy Location Filter (t-1)				0.0026**	0.0070*	0.0078**	0.0075*
				(0.0011)	(0.0037)	(0.0038)	(0.0038)
Google Real Estate Proxy Location Filter (t-1)					-0.0049	-0.003	0.0029
					(0.0038)	(0.004)	(0.0095)
Google Rental Proxy Location Filter (t-1)						-0.0025**	-0.0023*
						(0.0012)	(0.0012)
Google Real Estate Location Filter (t-1)							-0.0059
							(0.009)
Constant	-2.2709***	-0.7581	-0.0574	-0.2973	0.0909	-0.8274	-0.9301
	(0.7425)	(1.3249)	(1.0465)	(1.1971)	(1.2564)	(1.3781)	(1.4258)
Observations	69	66	66	66	66	66	66
Adjusted R^2	0.873	0.95	0.963	0.954	0.954	0.957	0.957
AIC	-134.542	-185.8914	-204.1943	-190.2181	-189.9794	-192.9571	-191.5982
BIC	-125.6055	-168.3742	-182.2978	-170.5113	-168.0829	-168.8709	-165.3223
Homoskedasticity Test	0.6646	0.0516	0.3703	0.1199	0.223	0.4421	0.4421
Serial Correlation Test	0.0027	0.0057	0.1061	0.0279	0.0243	0.0439	0.0457
Root MSE	0.0887	0.0559	0.0481	0.0538	0.0535	0.0520	0.0522

*10% Significance Level, **5% Significance Level, ***1% Significance Level
Data is from U.S. dataset discussed in data section. Dependent variable is logged existing home sales. Observations decrease because of availability of lagged variables. Google variables are explained in Table 1. Month trend is a one unit incremental increasing monthly trend line. Median home price is scaled to show the impact of a $10,000 change in home prices. Homoskedasticity test refers to White's Test. The null hypothesis of this test is that the data is homoskedastic. Serial Correlation Test refers to the Breush-Godfrey test. The null hypothesis of this test is the data is not serially correlated.

Table 11 shows that the Google variables reduce the RMSE of the forecasting model even when macroeconomic variables are included. Furthermore, numerous Google

variables are statistically significant. Column 1 features the initial baseline model that does not include Google or macroeconomic terms; the RMSE of the model is .0887. All of the terms are significant at the 1% level and feature the expected signs. Column 7 incorporates four additional macroeconomic variables to control for mortgage rates (financing costs), inflation, unemployment, and a monthly incremental upward trend. The mortgage rate is negative and significant at the 1% level indicating a rise in mortgage rates decreases home sales. The unemployment rate is unexpectedly positive and significant at the 1% levels; although, the magnitude is small with a 1% increase in unemployment leading to a .0661% increase in home sales. A similar coefficient occurs in Columns 8-12. Conventional theory dictates that an increase in the unemployment rate should decrease home sales. Similarly, the median home price variable unexpectedly has a positive sign; although, again the magnitude of the variable is small.

Column 8 adds the two Google category level variables. These variables are both significant at the 1% level, and the rental variable is negative as expected. Including the variables reduces the RMSE of Model 8 to .0481. Columns 9-12 drop the category level variables for more specific Google search queries that are designed to proxy for real estate agents, real estate, and rental searches. The RMSE of the models decline as each successive Google variable is added. Column 9 features a RMSE of .0538, while the RMSE of Columns 10 and 11 is .0535 and .0520 respectively.

Column 8 features the highest adjusted R-squared, lowest RMSE, and lowest BIC among the models in Table 11. This suggests it is the best prediction model for national home sales. The RMSE is 45.77% lower than Column 1 and 13.95% lower than Column 7. This indicates including the variables substantially improves the model's predictive

ability. The Google Realtor coefficient is significant at the 1% level, and a 1 unit increase in the variable leads to a .7629% increase in existing home sales. The Google Rental Category coefficient is also significant at the 1% level with a unit increase in the variable leading to a .4170% decrease in home sales. All in all, the results in Table 11 support the hypothesis that the Google search variables provide a predictive advantage over macroeconomic variables by providing additional useful information.

Tables 10 and 11 support the hypothesis that the Google search query variables improve the predictive ability of national home sales models. This confirms the work of Choi and Varian. While the results of the national model may be useful to policy makers, housing remains a local market and a stationary good. Given the local nature of housing, most housing searches are likely locally focused. For example, an individual looking for a home in Dallas is more likely to search for "Dallas real estate" than "real estate". Thus, the specific (non-category) Google terms are hypothesized to be more effective at the local level. Homeowners, real estate agencies, and the building industry would all benefit from improved local housing forecasts.

Table 12 shows the OLS estimates for the baseline MSA level specifications, with each column representing a different model. The major macroeconomic factors are excluded, and while the models feature location fixed effects, time fixed effects are not incorporated.

Table 12 MSA Log Home Sales Models and Robust Standard Errors

	(1)	(2)	(3)	(4)	(5)	(6)
Log Existing Home Sales (t-12)	0.4284***	0.3747***	0.3433***	0.3397***	0.3260***	0.3261***
	(0.0313)	(0.037)	(0.0303)	(0.0384)	(0.042)	(0.043)
Log Existing Home Sales (t-1)	0.4273***	0.3806***	0.3698***	0.3391***	0.3531***	0.3532***
	(0.0282)	(0.0309)	(0.0275)	(0.0299)	(0.0239)	(0.024)
Median Home Price (t-2) ($10K units)	-0.0284***	-0.0379***	-0.0279***	-0.0129**	-0.0186***	-0.0186***
	(0.0041)	(0.0044)	(0.0039)	(0.0052)	(0.0055)	(0.0056)
15Y Fixed Mortgage Rate (t-3)						
Consumer Price Index (t-3)						
Unemployment Rate (t-2)						
Log Population (y-1)						
Google Realtor Category		0.1460***				
		(0.0293)				
Google Rental Category		0.0904***				
		(0.0262)				
Google Location Specific Real Estate Proxy (t-1)			0.0061***	0.0055***	0.0049***	0.0049***
			(0.0005)	(0.0007)	(0.0007)	(0.0007)
Google Location Specific Realtor Proxy (t-1)				0.0029***	0.0023***	0.0023***
				(0.0006)	(0.0006)	(0.0006)
Google Specific Realtor Proxy Location Filter (t-1)					0.0014**	0.0014**
					(0.0007)	(0.0007)
Google Rental Proxy Location Filter (t-1)						0.0000
						(0.0005)
Constant	1.1724***	1.8970***	1.6164***	1.6243***	1.7509***	1.7496***
	(0.1561)	(0.2052)	(0.1532)	(0.1834)	(0.1923)	(0.2059)
Location Fixed Effects	Yes	Yes	Yes	Yes	Yes	Yes
Observations	1249	1098	1230	879	777	777
Adjusted R^2	0.986	0.987	0.988	0.991	0.992	0.992
AIC	-1031.82	-974.7831	-1186.051	-1044.858	-1012.273	-1010.273
BIC	-1011.299	-944.7757	-1160.477	-1016.185	-979.6846	-973.0295
Homoskedasticity Test	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
Wooldridge Serial Correlation Test	0.0000	0.0000	0.0000	0.0002	0.0006	0.0006
Root MSE	0.1610	0.1560	0.1502	0.1342	0.1267	0.1267

*10% Significance Level, **5% Significance Level, ***1% Significance Level

Data is from MSA dataset discussed in data section. Dependent variable is logged existing home sales. Observations vary because of availability of Google variables and lagged variables. The Google variable issue is discussed in the Data Section. Google variables are explained in Table 1. Median home price is scaled to show the impact of a $10,000 change in home prices. Homoskedasticity test refers the STATA recommended Likelihood-Ratio test for panel data homoskedasticity. The null hypothesis of this test is that the data is homoskedastic. Serial Correlation Test refers to the Woolridge Serial Correlation test. The null hypothesis of this test is the data is not serially correlated.

The table illustrates that introducing the Google variables reduces the RMSE of the forecasting model. Additionally, many of the Google variables are statistically significant. Column 1 is a baseline model that incorporates no Google terms, and the RMSE of the model is .1610; additionally, all of the terms are significant at the 1% level and feature the expected signs. Column 2 adds the two Google category variables utilized by Choi and Varian. The terms are both significant at the 1% level and serve to reduce the RMSE by 3.11% to .1560. Columns 3-6 drop the category level terms for more specific search queries that are hypothesized to be more effective for predicting the local housing market. The variables are designed to account for generic real estate, realtor, and rental searches, along with queries for specific prominent Texas real estate firms. Column 3 features the expected signs and the single Google term is significant at the 1% level. The RMSE is reduced to .1502. Columns 4 and 5 continue to expand Column 3 by incorporating additional Google variables. These variables are all significant and are positive as expected. The RMSE continues to decline to .1342 in Model 4 and .1267 in Model 5. Thus, the RMSE is reduced by 21.30% from Model 1 to Model 5. These results indicate that the Google variables are improving the accuracy of the baseline prediction model. Column 6 includes a variable for rental searches. However, the value of the coefficient is essentially 0 and the RMSE does not improve from Model 5. The results in Table 12 support the hypothesis that Google search variables improve forecasts for MSA level existing home sales.

Table 13 shows the OLS estimates for the MSA level models that incorporate macroeconomic variables. Each column represents a different specification. The models

include location fixed effects to account for differences between cities that remain constant over time; however, there are no time fixed effects.

Table 13 MSA Log Home Sales Models and Robust Standard Errors

	(1)	(7)	(8)	(9)	(10)	(11)	(12)
Log Existing Home Sales (t-12)	0.4284***	0.4783***	0.3812***	0.3773***	0.3603***	0.3369***	0.3504***
	(0.0313)	(0.0363)	(0.0407)	(0.0366)	(0.0462)	(0.0502)	(0.0521)
Log Existing Home Sales (t-1)	0.4273***	0.3758***	0.3297***	0.3401***	0.3469***	0.3600***	0.3721***
	(0.0282)	(0.0299)	(0.0307)	(0.0294)	(0.0312)	(0.0259)	(0.0263)
Median Home Price (t-2) ($10K units)	-0.0284***	0.0130**	0.0051	0.0054	0.005	-0.0022	-0.0009
	(0.0041)	(0.0053)	(0.0059)	(0.0053)	(0.0066)	(0.007)	(0.0071)
15Y Fixed Mortgage Rate (t-3)		-0.0502***	-0.0679***	-0.0453***	-0.0472***	-0.0543***	-0.0629***
		(0.0125)	(0.0135)	(0.0119)	(0.0134)	(0.0133)	(0.0146)
Consumer Price Index (t-3)		-0.0062***	-0.0065***	-0.0051***	-0.0046***	-0.0043***	-0.0039***
		(0.0007)	(0.0008)	(0.0007)	(0.001)	(0.001)	(0.001)
Unemployment Rate (t-2)		-0.0131**	-0.0186***	-0.0146***	-0.0103	-0.0085	-0.0074
		(0.0059)	(0.0062)	(0.0056)	(0.0073)	(0.0072)	(0.0072)
Log Population (y-1)		-0.1333	0.0197	0.0966	0.2556	0.3656*	0.3156
		(0.183)	(0.188)	(0.1847)	(0.1924)	(0.1888)	(0.1926)
Google Realtor Category			0.1629***				
			(0.0302)				
Google Rental Category			0.0995***				
			(0.0286)				
Google Location Specific Real Estate Proxy (t-1)				0.0052***	0.0056***	0.0048***	0.0050***
				(0.0006)	(0.0007)	(0.0007)	(0.0007)
Google Location Specific Realtor Proxy (t-1)					0.0011	0.0007	0.0008
					(0.0007)	(0.0007)	(0.0007)
Google Specific Realtor Proxy Location Filter (t-1)						0.0020***	0.0024***
						(0.0007)	(0.0007)
Google Rental Proxy Location Filter (t-1)							-0.0012*
							(0.0006)
Constant	1.1724***	4.0042*	3.1719	1.3763	-0.7734	-2.1142	-1.6408
	(0.1561)	(2.2832)	(2.3563)	(2.311)	(2.4492)	(2.4125)	(2.4404)
Location Fixed Effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	1249	1231	1086	1214	873	771	771
Adjusted R^2	0.986	0.987	0.988	0.988	0.991	0.992	0.992
AIC	-1031.82	-1133.911	-1071.824	-1244.346	-1066.772	-1039.51	-1040.826
BIC	-1011.299	-1092.986	-1021.922	-1198.431	-1019.053	-988.3852	-985.054
Homoskedasticity Test	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
Wooldridge Serial Correlation Test	0.0000	0.0000	0.0000	0.0000	0.0000	0.0002	0.0002
Root MSE	0.1610	0.1533	0.1482	0.1455	0.1317	0.1235	0.1233

*10% Significance Level, **5% Significance Level, ***1% Significance Level

Data is from MSA dataset discussed in data section. Dependent variable is logged existing home sales. Observations vary because of availability of Google variables and lagged variables. The Google variable issue is discussed in the Data Section. Google variables are explained in Table 1. Median home price is scaled to show the impact of a $10,000 change in home prices. Homoskedasticity test refers the STATA recommended Likelihood-Ratio test for panel data homoskedasticity. The null hypothesis of this test is that the data is homoskedastic. Serial Correlation Test refers to the Woolridge Serial Correlation test. The null hypothesis of this test is the data is not serially correlated.

Table 13 shows that the Google variables reduce the RMSE of the forecasting model even when macroeconomic variables are included. Additionally, many of the Google variables are statistically significant. Column 1 is the initial baseline model that incorporates no Google or macroeconomic terms; the RMSE of the model is .1610. All of the terms are significant and feature the expected signs. Column 7 adds four macroeconomic variables to control for mortgage rates (financing costs), inflation, unemployment, and population. The mortgage rate and consumer price index coefficients are negative and significant at the 1% level indicating a rise in mortgage rates or inflation decreases home sales. The unemployment variable is negative and significant as expected. Neither the population or median home price variables are significant in Column 7.

Column 8 adds the two Google category level variables. These variables are both significant at the 1% level; although, the rental variable was expected to be negative. Including the Google category variables reduced the RMSE of Model 8 to .1482. Columns 9-12 drop the category level variables for more specific Google search queries that are expected to be more effective at the local level. The RMSE of the models decline as each successive Google variable is added. Column 9 features a RMSE of .1455, while the RMSE of Column 10 is .1317. The RMSE continues to decline to .1235 and .1233 for Columns 11 and 12 respectively. The RMSE in Column 12 is 19.57% lower than Column 7 and 23.42% less than Column 1. The decrease in RMSE suggests that the Google variables lead to more accurate predictions. This result fits the existing literature and theory.

Column 12 exemplifies the predictive value of the Google variables. All of the significant macroeconomic variables in this model feature the expected signs. Additionally, three of the four Google variables are significant. The coefficient on Google City Specific Real Estate Proxy is

significant at the 1% level and indicates that a 1 unit increase in the variable leads to a .0050% increase in home sales. Similarly, a 1 unit increase in the Google Specific Realtor Proxy City Filter, significant at the1% level, leads to a .0024% increase in home sales, while a 1 unit increase in Google Rental Proxy City Filter leads to a .0012% decrease in home sales. The Google Rental Proxy City Filter variable is significant at the 10% level. The magnitude of these coefficients is quite small – especially compared to the optimal national model. Yet, incorporating these variables serves to reduce the RMSE of Column 12 by 23.42% from Column 1 and 19.57% from Column 7. The adjusted R-squared values also increase from .986 in Model 1 to .987 in Model 7 to .992 in Model 12. These metrics further suggest that Google variables improve the prediction abilities of the MSA forecasting specifications.

One interesting result is that the Google Category variables appear to be the best predictors for national level home sales, while the more specific and local Google Term variables are better at the MSA level. One explanation is that the local nature of housing markets makes it unlikely that an individual would simply search for realtors. Instead the individual would likely search for "Dallas realtors"; thus, the national level sales are best predicted by using Google's automatic categorization, while the more specific local terms are better for predicting MSA home sales. The results in Tables 10-13 corroborate the existing literature and support the hypothesis that the Google search variables improve forecasts.

C. Out-of-Sample Testing

While Tables 10-13 support the notion that Google variables improve forecasts, out-of-sample testing remains the gold standard for evaluating forecasting approaches. This paper tests the out-of-sample forecasting accuracy of several U.S. and MSA models from January 2008 to

Fall 2009.[17] This period approximately corresponds to the downturn in the housing sector and the global economic crisis. The rationale for this time period is that if the Google predictors are able to improve forecast accuracy during a period of major upheaval in the housing sector, then the Google variables will also likely be useful in less tumultuous times.

The out-of-sample forecasting model compares the baseline regression model with the macroeconomic regression model and the best model incorporating Google terms for both the U.S. and MSA geographic levels. The forecasting model dynamically updates to include the prior month's data. For example, the January 2008 forecast utilizes a regression that relies on data prior to January 2008. The February 2008 forecast is based on an updated regression model that incorporates the January 2008 data into the coefficients. The model continues to dynamically update each month.

Forecast accuracy is measured using four traditional metrics. The first metric is Mean Forecast Error (MFE). MFE is calculated as:

$$\frac{\sum_{i=1}^{n}(r_i)}{n}$$

Where r is the residual and n is the number of observations. The second metric is Mean Absolute Deviation (MAD). MAD is calculated as:

$$\frac{\sum_{i=1}^{n}|r_i|}{n}$$

Where r is the residual and n is the number of observations. The third metric is the Mean Squared Error (MSE). MSE is calculated as:

$$\frac{\sum_{i=1}^{n}r_i^2}{n}$$

[17] The national out-of-sample tests are conducted through October, and the MSA out-of-sample tests are conducted through November. The discrepancy occurs because less national level data was available when the datasets were constructed.

Where r is the residual and n is the number of observations. The fourth metric is Prediction Absolute Percentage Error. This value is calculated as:

$$\frac{\sum_{i=1}^{n} \left| \frac{predicted_i - actual_i}{actual_i} \right|}{n} * 100$$

Where n is the number of observations. These metrics are designed to measure the forecasting accuracy of the different models.

Tables 14-15 feature the U.S. out-of-sample forecasting metrics. The models can be matched to the earlier regressions by their numbers. Table 14 features the results using the natural log values; the data in Table 15 was converted from logs to levels prior to calculating the out-of-sample metrics in order to show the differences in units of homes. Figure 10 illustrates the accuracy of the different prediction models.

Table 14 USA Out of Sample Test Forecast Error for Log Sales (Jan 2008-Oct 2009)

	Mean Forecast Error (MFE)	Mean Absolute Deviation (MAD)	Mean Squared Error (MSE)	Prediction Absolute Percentage Error (%)
Base Regression Model (1)	-0.0737	0.0896	0.0122	0.6984
Macro Variable Model (7)	0.0191	0.0851	0.0139	0.6572
Google Variable Model (8)	0.0095	0.0733	0.0097	0.5671

Data is from US dataset discussed in the data section. Predictions measured in logs. The model number corresponds to the model number in Table 11. The metrics to measure forecast accuracy are defined on pages 47 and 48. The forecasting model dynamically updates to include the prior month's data. For example, the January 2008 forecast utilizes a regression that relies on data before January 2008. The February 2008 forecast is based on an updated regression model that incorporates the January 2008 data into the coefficients. The model continues to dynamically update each month.

Table 15 USA Out of Sample Test Forecast Error for Log Sales (Jan 2008-Oct 2009)

	Mean Forecast Error (MFE)	Mean Absolute Deviation (MAD)	Mean Squared Error (MSE)	Prediction Absolute Percentage Error (%)
Base Regression Model (1)	-29687.0263	36787.1737	1929464114.5522	9.5463
Macro Variable Model (7)	7733.3876	34642.7925	2130933955.4214	8.1172
Google Variable Model (8)	5408.1587	29480.9654	1523375622.2056	7.1051

Data is from US dataset discussed in the data section. Predictions were converted from logs to levels. The model number corresponds to the model number in Table 11. The metrics to measure forecast accuracy are defined on pages 47 and 48. The forecasting model dynamically updates to include the prior month's data. For example, the January 2008 forecast utilizes a regression that relies on data before January 2008. The February 2008 forecast is based on an updated regression model that incorporates the January 2008 data into the coefficients. The model continues to dynamically update each month.

Figure 10 USA Log Home Sales Out-of-Sample Prediction Models

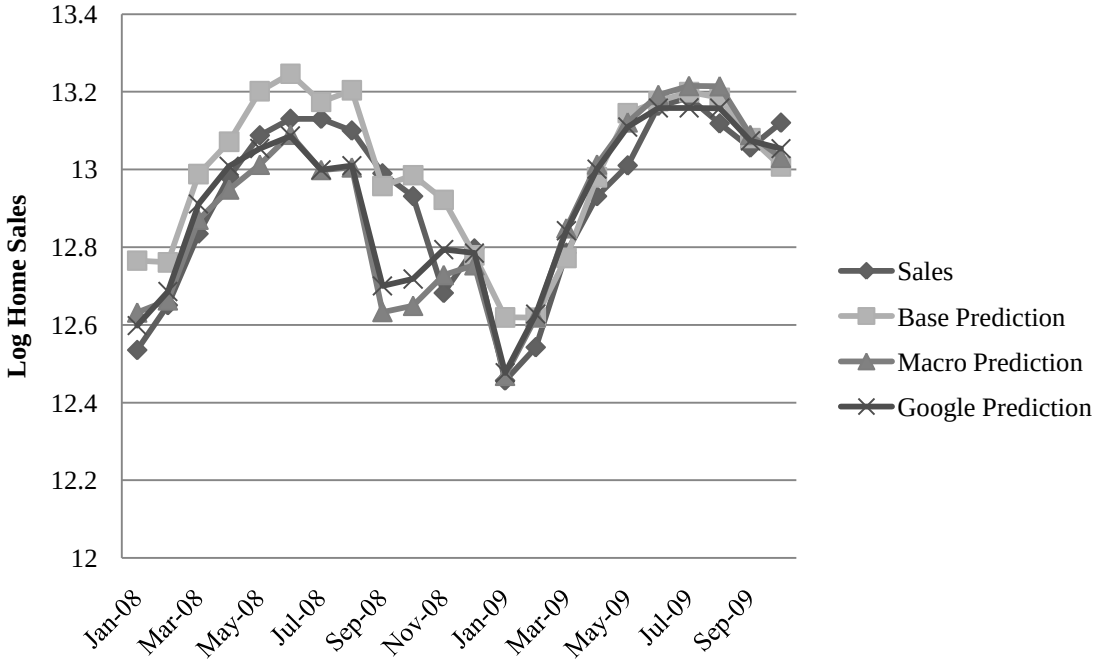


Figure 10 is graphical representation of monthly log existing home sales and out-of-sample predictions for U.S. level data. Data is from U.S. dataset.

Table 14 indicates that Regression 8 featuring the Google variables outperforms the other models across all four metrics. Regression 8 reduces the MFE from .0191 to .0095, a change of 52.26%. The Google variables improve the MAD from .0851 to .0733 a change of 13.861%. In terms of MSE, the Google model reduces the value 20.49% from .0122 to .0097. Percentage Absolute Prediction Error decreases by 13.71% from .6572 to .5671. Table 15 tells the same story using levels. The results in Tables 14 and 15 confirm that including Google variables in the national specification improves forecasting accuracy.

Tables 16 and 17 contain the MSA out-of-sample forecasting metrics, and Figure 11 graphs the predictions of the three models versus the actual sales figures for Houston, TX. The models can be matched to the earlier regressions by their numbers. Table 16 features the results using the natural log values, while Table 17 contains the level data.

Table 16: MSA Out of Sample Test Forecast Error for Log Sales (Jan 2008-Nov 2009)

	Mean Forecast Error (MFE)	Mean Absolute Deviation (MAD)	Mean Squared Error (MSE)	Prediction Absolute Percentage Error (%)
Base Regression Model (1)	-0.0153	0.2629	0.1039	4.5589
Macro Variable Model (7)	0.0917	0.4008	0.2927	6.5439
Google Variable Model (12)	0.0085	0.2233	0.0874	3.8917

Data is from MSA dataset discussed in the data section. Predictions are in logs. The model number corresponds to the model number in Table 13. The metrics to measure forecast accuracy are defined on pages 47 and 48. The forecasting model dynamically updates to include the prior month's data. For example, the January 2008 forecast utilizes a regression that relies on data before January 2008. The February 2008 forecast is based on an updated regression model that incorporates the January 2008 data into the coefficients. The model continues to dynamically update each month.

Table 17: MSA Out of Sample Test Forecast Error for Log Sales (Jan 2008-Nov 2009)

	Mean Forecast Error (MFE)	Mean Absolute Deviation (MAD)	Mean Squared Error (MSE)	Prediction Absolute Percentage Error (%)
Base Regression Model (1)	294.9730	339.9978	518853.0116	27.2935
Macro Variable Model (7)	462.1792	510.9468	1230415.815	36.6150
Google Variable Model (12)	209.4549	255.8354	376529.6239	22.7279

Data is from MSA dataset discussed in the data section. Predictions were converted from logs to levels. The model number corresponds to the model number in Table 13. The metrics to measure forecast accuracy are defined on pages 47 and 48. The forecasting model dynamically updates to include the prior month's data. For example, the January 2008 forecast utilizes a regression that relies on data before January 2008. The February 2008 forecast is based on an updated regression model that incorporates the January 2008 data into the coefficients. The model continues to dynamically update each month.

Figure 11 Houston Log Home Sales Out-of-Sample Prediction Models

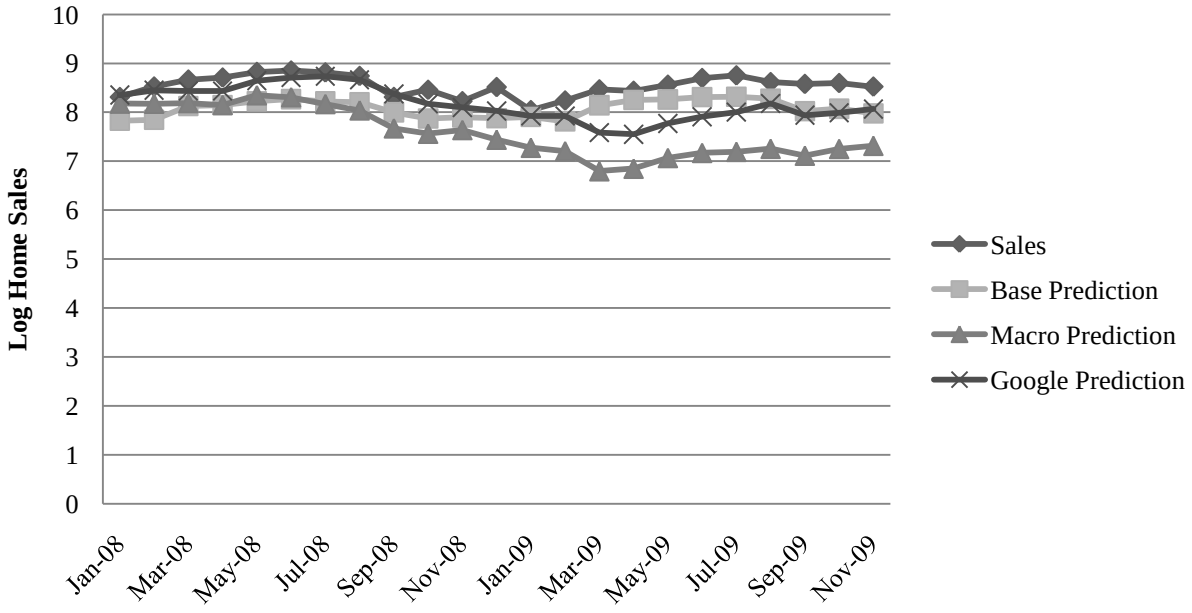


Figure 11 is graphical representation of monthly log existing home sales and out-of-sample predictions for Houston, Texas. Data is from MSA dataset.

Table 16 indicates that Regression 12, which incorporates the Google variables, has the lowest MFE, MAD, MSE, and Prediction Absolute Percentage Error. Incorporating the Google

variables improves the MFE from -.0135 to .0085 an absolute improvement of 32.68%. Similarly, the Google variables improve the MAD from .2629 to .2233 a change of 15.06%. In terms of MSE, the Google model reduces the value 15.88% from .1039 to .0874. Percentage Absolute Prediction Error decreases by 12.72% to 3.98% from 4.56%.

In Table 17, the most interesting result is that the Prediction Absolute Percentage Error is reduced from 27.30% to 20.58%. For comparison, the national Google model in Table 15 reduced Prediction Absolute Percentage Error from 8.11% to 7.11%. Thus, the Google variables were more effective at reducing the local level error. However, the local errors were still much larger on average than the national errors. Tables 16 and 17 support the hypothesis that the Google variables improve the forecasting accuracy of the local specifications.

VI. Conclusion

This paper finds Google search query data can improve national and local existing home sales forecasts. The results show that the Google variables reduce the RMSE of the national and MSA models. Perhaps most importantly, the models incorporating Google terms outperform the non-Google models in all of the out-of-sample tests. Additionally, many of the Google variables are statistically significant. This paper's findings suggest that the Google variables are significant leading indicators that improve housing forecasts. One limitation is the local analysis only focuses on Texas cities because of data availability issues.

Given the large size of the housing sector, this finding is pertinent to policymakers. Improved forecasts could enable the Federal Reserve Bank to better predict future economic conditions, allow firms to better price securitized mortgage bonds, and enable real estate firms to better plan personnel decisions. The results could also allow Congress or state governments to enact tax credits or other policy initiatives in order to stimulate (or repress) housing demand. At

the individual level, the results could enable individuals to better predict housing market trends. Even realtors could find the local level data useful in predicting their future income. Additionally, home sales also impact local real estate firms, construction firms, and individuals.

While the results have obvious implications for the housing sector, the findings also suggest that Google Insights data could improve forecasts in other sectors. Search queries are particularly useful in gauging information search on a near real-time basis. For example, Google Trends could provide up to date insight into web traffic at Amazon.com, which likely corresponds to sales. Given that information search increases with price, Google variables may be most useful for expensive durable goods or other large expenditures that require significant research. Some examples include automobiles, televisions, and medical decisions. Future studies could expand the MSA dataset to other cities or move completely beyond real estate to some of the areas discussed above. All in all, Google search query data is a promising new approach to forecasting.

Appendix

Table 18 Summary of Data

	National				Local			
Variable	**Level**	**Frequency**	**Availability**	**Source**	**Level**	**Frequency**	**Availability**	**Source**
Existing Home Sales	National	Monthly	2001-Present	NAR	MSA	Monthly	Dec 2003-Pres	TAMUREC
Google Search Terms	National	Weekly	2004-Present	Google	City	Weekly	2004-Present	Google
15 Year Fixed Mortgage Rate	National	Monthly	1991-Present	Freddie Mac	National	Monthly	1991-Present	Freddie Mac
Unemployment Rate	National	Monthly	1999-2009	BLS	MSA	Monthly	Dec 2003-Pres	TAMUREC
Inflation /CPI	National	Monthly	1999-Present	BLS	National	Monthly	1999-Present	BLS
Median Home Price	National	Monthly	1999-Present	NAR	MSA	Monthly	Dec 2003-Pres	TAMUREC
Population	National	Yearly	1970-Present	US Census / TAMUREC	MSA	Yearly	1970-Present	US Census / TAMUREC

Table 19 List of Cities in Texas MSA Dataset

City
Abilene
Amarillo
Austin
Beaumont-Port Arthur
Corpus Christi
Dallas-Fort Worth
El Paso
Harlingen
Houston
Laredo
Lubbock
Odessa-Midland
San Angelo
San Antonio
Sherman
Tyler-Longview
Victoria
Waco-Temple-Bryan
Wichita Falls

Table 20 USA Expanded Log Home Sales Models and Newey-West Standard Errors

	(1)C	(7)C	(9)C	(10)C	(11)C	(12)C
Log Existing Home Sales (t-12)	0.8143***	0.8943***	0.7937***	0.7686***	0.7707***	0.7608***
	(0.0345)	(0.059)	(0.0744)	(0.0791)	(0.0784)	(0.0835)
Log Existing Home Sales (t-1)	0.4109***	0.0436	0.0712	0.0674	0.1282*	0.1403*
	(0.0721)	(0.0528)	(0.0587)	(0.0527)	(0.0664)	(0.0814)
Median Home Price (t-1)($10K scaled)	-0.0347***	0.0159***	0.0127***	0.0133**	0.0072	0.0074
	(0.0101)	(0.0041)	(0.0047)	(0.0051)	(0.0056)	(0.0056)
15Y Fixed Mortgage Rate (t-3)		-0.0878***	-0.0999***	-0.1067***	-0.1233***	-0.1201***
		(0.0167)	(0.015)	(0.0133)	(0.0147)	(0.0145)
Consumer Price Index (t-4)		0.008	0.0102*	0.0106*	0.0120**	0.0123**
		(0.0063)	(0.0056)	(0.0056)	(0.0054)	(0.0057)
Unemployment Rate (t-4)		0.0661***	0.0621***	0.0625***	0.0559***	0.0603***
		(0.0155)	(0.0139)	(0.0139)	(0.0128)	(0.0155)
Month Trend		-0.0077**	-0.0076**	-0.0076**	-0.0068**	-0.0081*
		(0.0034)	(0.0029)	(0.003)	(0.0028)	(0.0041)
Google Realtor Category						
Google Rental Category						
Google Real Estate Agents Proxy Location Filter (t-1)			0.0026**	0.0070**	0.0078**	0.0075**
			(0.0012)	(0.003)	(0.0031)	(0.003)
Google Real Estate Proxy Location Filter (t-1)				-0.0049*	-0.003	0.0029
				(0.0027)	(0.0022)	(0.012)
Google Rental Proxy Location Filter (t-1)					-0.0025***	-0.0023**
					(0.0009)	(0.0009)
Google Real Estate Location Filter (t-1)						-0.0059
						(0.0125)
Constant	-2.2709***	-0.7581	-0.2973	0.0909	-0.8274	-0.9301
	(0.8131)	(1.6467)	(1.4173)	(1.4217)	(1.3711)	(1.5273)
Observations	69	66	66	66	66	66

*10% Significance Level, **5% Significance Level, ***1% Significance Level

Data is from U.S. dataset discussed in data section. Dependent variable is logged existing home sales. Regressions feature Newey-West standard errors to correct for heteroskedasticity and autocorrelation, and the regression numbers correspond to the models in Tables 10 and 11. Only the regressions that required the Newey-West error correction are listed. Observations vary because of availability of Google variables and lagged variables. The Google variable issue is discussed in the Data Section. Google variables are explained in Table 1. Median home price is scaled to show the impact of a $10,000 change in home prices.

Table 21 MSA Log Home Sales Models and Newey-West Standard Errors

	(1)C	(2)C	(3)C	(4)C	(5)C	(6)C
Log Existing Home Sales (t-12)	0.4284***	0.3747***	0.3433***	0.3397***	0.3260***	0.3261***
	(0.0346)	(0.0392)	(0.0319)	(0.0403)	(0.044)	(0.0448)
Log Existing Home Sales (t-1)	0.4273***	0.3806***	0.3698***	0.3391***	0.3531***	0.3532***
	(0.0332)	(0.0359)	(0.0326)	(0.0319)	(0.0289)	(0.0287)
Median Home Price (t-2) ($10K units)	-0.0284***	-0.0379***	-0.0279***	-0.0129**	-0.0186***	-0.0186***
	(0.005)	(0.0053)	(0.0045)	(0.0059)	(0.0066)	(0.0067)
15Y Fixed Mortgage Rate (t-3)						
Consumer Price Index (t-3)						
Unemployment Rate (t-2)						
Log Population (y-1)						
Google Realtor Category		0.1460***				
		(0.0364)				
Google Rental Category		0.0904***				
		(0.0271)				
Google Location Specific Real Estate Proxy (t-1)			0.0061***	0.0055***	0.0049***	0.0049***
			(0.0006)	(0.0007)	(0.0008)	(0.0008)
Google Location Specific Realtor Proxy (t-1)				0.0029***	0.0023***	0.0023***
				(0.0007)	(0.0008)	(0.0007)
Google Specific Realtor Proxy Location Filter (t-1)					0.0014**	0.0014**
					(0.0006)	(0.0007)
Google Rental Proxy Location Filter (t-1)						0.0000
						(0.0005)
Constant	0.9963***	1.7010***	1.3212***	1.2199***	1.2591***	1.2580***
	(0.1398)	(0.1892)	(0.127)	(0.1513)	(0.1508)	(0.1598)
Location Fixed Effects	Yes	Yes	Yes	Yes	Yes	Yes
Observations	1249	1098	1230	879	777	777

*10% Significance Level, **5% Significance Level, ***1% Significance Level

Data is from MSA dataset discussed in data section. Dependent variable is logged existing home sales. Regressions feature Newey-West standard errors to correct for heteroskedasticity and autocorrelation, and the regression numbers correspond to the models in Table 12. Observations vary because of availability of Google variables and lagged variables. The Google variable issue is discussed in the Data Section. Google variables are explained in Table 1. Median home price is scaled to show the impact of a $10,000 change in home prices.

Table 22 MSA Log Home Sales Models and Newey-West Standard Errors

	(7)C	(8)C	(9)C	(10)C	(11)C	(12)C
Log Existing Home Sales (t-12)	0.4783***	0.3812***	0.3773***	0.3603***	0.3369***	0.3504***
	(0.0388)	(0.0435)	(0.0387)	(0.0478)	(0.0517)	(0.0529)
Log Existing Home Sales (t-1)	0.3758***	0.3297***	0.3401***	0.3469***	0.3600***	0.3721***
	(0.0339)	(0.0354)	(0.034)	(0.0345)	(0.0312)	(0.0319)
Median Home Price (t-2) ($10K units)	0.0130**	0.0051	0.0054	0.005	-0.0022	-0.0009
	(0.0058)	(0.0065)	(0.0055)	(0.0076)	(0.0084)	(0.0084)
15Y Fixed Mortgage Rate (t-3)	-0.0502***	-0.0679***	-0.0453***	-0.0472***	-0.0543***	-0.0629***
	(0.0152)	(0.0152)	(0.0142)	(0.0149)	(0.0155)	(0.0169)
Consumer Price Index (t-3)	-0.0062***	-0.0065***	-0.0051***	-0.0046***	-0.0043***	-0.0039***
	(0.0007)	(0.0007)	(0.0007)	(0.001)	(0.001)	(0.001)
Unemployment Rate (t-2)	-0.0131**	-0.0186***	-0.0146**	-0.0103	-0.0085	-0.0074
	(0.0066)	(0.0068)	(0.0062)	(0.0077)	(0.0087)	(0.0087)
Log Population (y-1)	-0.1333	0.0197	0.0966	0.2556	0.3656*	0.3156
	(0.2107)	(0.2049)	(0.2062)	(0.2173)	(0.2082)	(0.2074)
Google Realtor Category		0.1629***				
		(0.0308)				
Google Rental Category		0.0995***				
		(0.0285)				
Google City Specific Real Estate Proxy (t-1)			0.0052***	0.0056***	0.0048***	0.0050***
			(0.0006)	(0.0007)	(0.0008)	(0.0008)
Google City Specific Realtor Proxy (t-1)				0.0011	0.0007	0.0008
				(0.0007)	(0.0008)	(0.0008)
Google Specific Realtor Proxy City Filter (t-1)					0.0020***	0.0024***
					(0.0007)	(0.0008)
Google Rental Proxy City Filter (t-1)						-0.0012*
						(0.0006)
Constant	3.762	3.0564	1.2506	-0.7696	-2.0063	-1.5818
	(2.415)	(2.3583)	(2.3762)	(2.5361)	(2.4246)	(2.407)
Location Fixed Effects	Yes	Yes	Yes	Yes	Yes	Yes
Observations	1231	1086	1214	873	771	771

*10% Significance Level, **5% Significance Level, ***1% Significance Level

Data is from MSA dataset discussed in data section. Dependent variable is logged existing home sales. Regressions feature Newey-West standard errors to correct for heteroskedasticity and autocorrelation, and the regression numbers correspond to the models in Table 13. Observations vary because of availability of Google variables and lagged variables. The Google variable issue is discussed in the Data Section. Google variables are explained in Table 1. Median home price is scaled to show the impact of a $10,000 change in home prices.

Literature Sources

Askitas, Nikos, and Zimmermann, Klaus F. "Google Econometrics and Unemployment Forecasting." IZA Discussion Paper No. 4201. (June 2009): 1-22.

Bajari, Patrick, and Hortacsu, Ali. "Winner Curse, Reserve Prices, and Endogenous Entry: Empirical Insights from eBay Auctions." Stanford Institute for Economic Policy Research Discussion Paper (March 2000): 1-53.

Blackley, Dixie M., and Follain, James R. "An Econometric Model of the Metropolitan Housing Market." Journal of Housing Economics 1 (1991): 140-167.

Choi, Hyunyoung, and Varian, Hal. "Predicting the Present with Google Trends." Google (April 2009): 1-20.

Choi, Hyunyoung, and Varian, Hal. "Predicting Initial Claims for Unemployment Benefits." Google (July 2009): 1-5.

Clapp, John M., and Giaccotto, Carmelo. "Evaluating House Price Forecasts." JRER 24:1 (2002): 1-26.

Dua, Pami, and Miller, Stephen M. "Forecasting Connecticut Home Sales in a BVAR Framework Using Coincident and Leading Indexes." Journal of Real Estate Finance and Economics 13 (1996): 219-235.

Dua, Pami; Miller, Stephen M.; and Smyth, David J. "Using Leading Indicators to Forecast U.S. Home Sales in a Bayesian Vector Autoregressive Framework." Journal of Real Estate Finance and Economics 18:2 (1999): 191-205.

Dua, Pami, and Smyth, David J. "Forecasting U.S. Home Sales Using BVAR Models and Survey Data on Households' Buying Attitudes for Homes." Journal of Forecasting 14 (1995): 217-227.

Ginsberg, Jeremy; Mohebbi, Matthew H.; Patel, Rajan S.; Brammer, Lynnette; Smolinski, Mark S.; and Brillant, Larry. "Detecting Influenza Epidemics Using Search Engine Query Data." Nature 457 (February 2009): 1012-1014.

Houser, Daniel, and Wooders, John. "Reputation in Auctions: Theory and Evidence from eBay." Journal of Economics and Management Strategy 15:2 (Summer 2006): 353-369.

Lucking-Reiley, David; Bryan, Doug; Prasad, Naghi; and Reeves, Daniel. "Pennies from eBay: The Determinants of Price in Online Auctions." (January 2000): 1-24.

Roll, Richard, and Ross, Stephen A. "The Arbitrage Pricing Theory Approach to Strategic Portfolio Planning." Financial Analysts Journal (January-February 1995): 122-131.

Stigler, George J. "The Economics of Information." The Journal of Political Economy 69:3 (June 1961): 213-225.

Suhoy, Tanya. "Query Indices and a 2008 Downturn: Israeli Data." Bank of Israel Research Department Discussion Paper No. 2009.06. (July 2009): 1-32.

Additional References

"Announcing Google Insights for Search". Google AdWords Blog, 2009. http://adwords.blogspot.com/2008/08/announcing-google-insights-for-search.html.

Chandra, Shobhana. "Goods Orders, Homes Sales Probably Rose, Signaling U.S. Recovery." Bloomberg, October 28, 2009. http://www.bloomberg.com/apps/news?pid=newsarchive&sid=aS52jB6MF1qQ.

"Comparing New Home Sales and Existing Home Sales." U.S. Census Bureau. http://www.census.gov/const/www/existingvsnewsales.html.

"comScore Press Releases." comScore. http://www.comscore.com/Press_Events/Press_Releases.

"Expanded Tax Break Available for 2009 First-Time Homebuyers". Internal Revenue Service, February 25, 2009. http://www.irs.gov/newsroom/article/0,,id=204672,00.html

"Google Insights for Search." Google, 2009. http://www.google.com/insights/search/#.

"Google Trends." Google, 2009. http://www.google.com/trends.

"Gross Domestic Product by Industry Accounts." BEA, April 28, 2009. http://www.bea.gov/industry/gpotables/gpo_action.cfm?anon =502054&table_id=24753&format_type=0.

"Highlights: The Digital Future Report Year Eight." USC Annenberg School Center for the Digital Future, April 2009. http://www.digitalcenter.org/ pages/site_content.asp?intGlobalId=20.

"Income, Poverty, and Health Insurance Coverage in the United States: 2008." U.S. Census Bureau. http://www.census.gov/Press-Release/ www/releases/ archives /income_wealth/014227.html.

"Insights for Search Help." Google, 2009. http://www.google.com/support/insights// bin/answer.py?hl=en-US&answer=87285.

Lautz, Jessica. "FSBO Sales Decreasing." NAR, March 12, 2008. http://www.realtor.org/ research/economists_outlook/commentaries/commentary_fsbo.

"News and Press Releases." One Stat. http://www.onestat.com/html/aboutus_pressbox.html.

"Schwab Guide to Economic Indicators". Charles Schwab Research, 2009. http://www.schwab.com/public/schwab/research_strategies/market_insight/1/4/schwab_guide_to_economic_indicators_existing_home_sales.html.

"The Digital Future Report: Surveying the Digital Future Year Four." USC Annenberg School Center for the Digital Future, September 2004. http://www.digitalcenter.org/pages/site_content.asp?intGlobalId=20.

"The Digital Future Project: Surveying the Digital Future Year Eight." USC Annenberg School Center for the Digital Future, April 2009.

"Zillow Timeline." Zillow.com. http://www.zillow.com/corp/Timeline.htm