Volume Title: Economic Analysis of the Digital Economy

Volume Author/Editor: Avi Goldfarb, Shane M. Greenstein, and Catherine E. Tucker, editors

Volume Publisher: University of Chicago Press

Volume ISBN: 0-226-20684-X; 978-0-226-20684-4

Volume URL: http://www.nber.org/books/gree13-1

Conference Date: June 6–7, 2013

Publication Date: April 2015

Chapter Title: The Future of Prediction: How Google Searches Foreshadow Housing Prices and Sales

Chapter Author(s): Lynn Wu, Erik Brynjolfsson

Chapter URL: http://www.nber.org/chapters/c12994

Chapter pages in book: (p. 89 – 118)

3

The Future of Prediction
How Google Searches Foreshadow Housing Prices and Sales

Lynn Wu and Erik Brynjolfsson

It's difficult to make predictions, especially about the future.
—Attributed to Niels Bohr

3.1 Introduction

Traditional economic and business forecasting has relied on statistics gathered by government agencies, annual reports, and financial statements. Invariably, these are published after significant delay and are aggregated into a relatively small number of prespecified categories. This limits their usefulness for predictions, especially for addressing time-sensitive issues or novel questions. However, the widespread adoption of search engines and related information technologies facilitates the near-real-time collection of highly disaggregated data on literally hundreds of billions[1] of economic decisions. Recently, query technology has made it possible to obtain such information at nearly zero cost, virtually instantaneously and at a fine-grained level of disaggregation. Each time a consumer or business decision maker searches for a product via the Internet, valuable information is revealed about that individual's intentions to make a future economic transaction. In turn, knowledge of these intentions can be used to predict future demand and

Lynn Wu is assistant professor of operations and information management at The Wharton School, University of Pennsylvania. Erik Brynjolfsson is the Schussel Family Professor at the MIT Sloan School of Management, director of the MIT Center for Digital Business, and a research associate of the National Bureau of Economic Research.

We thank Karl Case, Avi Goldfarb, Andrea Meyer, Dana Meyer, Shachar Reichman, Lu Han, and Hal Varian as well as seminar participants at the NBER, MIT, the Workshop on Information Systems and Economics, and the International Conference on Information Systems for valuable comments on this research. The MIT Center for Digital Business provided generous funding. For acknowledgments, sources of research support, and disclosure of the authors' material financial relationships, if any, please see http://www.nber.org/chapters/c12994.ack.

1. Americans performed 14.3 billion Internet searches in March 2009, which is an annualized rate of over 170 billion searches per year. Worldwide searches grew by 41 percent between 2008 and 2009.

supply. This revolution in information and information technology is well underway, and it portends a concomitant revolution in our ability to make business predictions and, ultimately, a sea change in business and policy decision making. This new use of technology is not a mere difference in degree, but a fundamental transformation of how much is known about the present and what can be known about the future.

Assisting with predictions has always been a central contribution of social science research. In the past several decades, much of social science research has focused on ever more complex mathematical models for many types of important business and economic predictions. However, the latest recession has shown that none of the models was sophisticated enough to foresee the biggest economic downturn in our recent history (Krugman 2009). Perhaps instead of honing techniques to extract information out of noisy and error-prone data, social science research should focus on inventing tools to observe phenomenon at a higher resolution (Simon 1984). Search engine technology delivers such a tool by effectively aggregating consumers' digital traces and improving data quality by several orders of magnitude. This technology can transform the ways we solve the problem of predicting the future. By observing billions of consumers and business intentions as revealed by online search, researchers can significantly improve the accuracy, granularity, and timeliness of predictions about future economic activities.

In this chapter, we demonstrate how data on Internet queries could be used to make reliable predictions about changes in both market prices and sales volumes literally months before they actually change in the marketplace. We use the housing market as our case example. We started making housing market predictions in January of 2009 and showed they outperformed both the baseline model as well as those of experts like the National Association of Realtors. As of September 2011, almost three years after we released our first set of real estate predictions, search queries continue to provide a significant improvement in forecasting real estate trends and outperform predictions from the National Association of Realtors. This suggests the persistence of the economic value derived from search.

Economic predictions from search data can be applied to almost any market where Internet search often precedes the transaction, which is to say, an increasingly large share of the economy. Our techniques can be focused on particular regions or specific cities or the nation as a whole, and can look at broad or narrow product categories. Search not only precedes purchase decisions, but in many cases is a more "honest signal" (Pentland 2010) of actual interests and preferences because no bargaining, gaming, or strategic signaling is involved, in contrast to many market-based transactions or other types of data gathering such as surveys. As a result, consumers' digital traces can be compiled to reveal their likely underlying economic intentions and activities. Using aggregated query data collected from the Internet has the

potential to make accurate predictions about areas as diverse as the eventual winners of standard wars or the potential success of product introductions.

3.2 The Real Estate Market

We use the real estate market to demonstrate how online search can be used to reveal the present economic activities and predict future economic trends. Studying the real estate market is especially important in the wake of the recent bursting of the real estate bubble that triggered an economic downturn in the United States and the rest of the world. In turn, the recovery of the housing market may signal the recovery of the economy as well. Economists, politicians, and investors alike pore over government data released every month to assess the current housing market and predict its recovery and, subsequently, the revival of economic growth. However, as noted above, government data arrives with a lag of months or more, delaying assessment of the current economic conditions. By analyzing consumers' interests, as revealed by their online behaviors, we are able to uncover trends before they appear in published data.

By using the Internet as a research tool, consumers can find critical information to make purchase decisions (Horrigan 2008; Brynjolfsson, Hu, and Rahman 2013). As the Web becomes ubiquitous, more shoppers are using the Internet to gather product information and refine their purchasing choices, especially for products that require a high level of financial commitment, such as buying a home. According to the 2012 Profile of Home Buyers and Sellers by the National Association of Realtors (NAR), 90 percent of home buyers used the Internet to search for a home in 2012 (NAR 2012). Similarly, a report written by the California Association of Realtors in 2008 shows that 63 percent of homebuyers find their real estate agent using a search engine (Appleton-Young 2008). To explore the link between search and actual sales, we analyze individual searches from eight years of data in the Google Web Search portal[2] to predict housing sales and housing prices. Using these fine-grained data on individual consumer behaviors, we built a comprehensive model to predict housing market trends.

We find evidence that queries submitted to Google's search engine are correlated with both the volume of housing sales as well as a house price index—specifically the Case-Shiller index—released by the Federal Housing Finance Agency. The Case-Shiller index is a popular housing index and is widely used in government reports. Search frequencies can reveal the current housing trends, but search is especially well suited for predicting the *future* unit sales of housing. Specifically, we find that a 1-percentage point increase in search frequency about real estate agents is associated with selling an additional 3,520 future quarterly housing sales in the average US state. We

2. See http://www.google.com/insights/search/#.

also compared our predictions with the prediction released by the NAR and our simple linear prediction model using search frequencies outperforms NAR's predictions by 23 percent.

Similarly, we also examine the relationship between housing prices and housing-related searches online. Using the house price index (HPI) from the Federal Housing Finance Agency,[3] we find a positive relationship between housing-related online queries and the future house price index, though the predictive power is not as strong as it is for home sales. Perhaps, predicting HPI is intrinsically more difficult than predicting sales volume because the effects of search volume on HPI are theoretically ambiguous. On one hand, if the search volume reflects changes in demand, as when potential buyers look for houses, then HPI will increase with searches. On the other hand, if the search volume reflects the supply side, as when sellers look at comparable homes and assess the market, then HPI might decrease with increased searches. Thus, aggregated search indices on general real estate categories may be well suited to predict sales volume but not as effective for predicting HPI. However, less aggregated and more fine-grained search categories could be created to differentiate the shifts on the demand side from the supply side.

We also find evidence that the total volume of houses sold is correlated with consumers' intention to purchase home appliances. We use the search frequency of home appliances to approximate consumers' interests (Moe and Fader 2004). We find that every thousand houses sold six months earlier are correlated with a 1.14 percentage point increase in the frequency of search terms that are related to home appliances. This highlights the linkages between home sales and other parts of the economy that complement home sales.

3.3 Literature Review

In the past decades, much of the social science research focused on refining increasingly complex mathematical models to predict social and economic trends. However, in recent years, the availability of fine-grained digital data opens up new options. Specifically, advances in information technologies such as the Internet search technologies, mobile phones, e-mail, and social media offer remarkably detailed records of human behaviors. Recently, researchers have started to take advantage of real-time data collected from these new technologies. For example, deploying sociometric badges to measure moment-to-moment interactions among a group of IT workers, Wu et al. (2008) uncovered new social network dynamics that are only possible

3. Historical HPI data can be downloaded at http://www.fhfa.gov/Default.aspx?Page=87.

by accessing accurate data at the microlevel. Lazer et al. (2009) provided various examples of how high-quality data produced by novel technologies are transforming the landscape of social network research. Similarly, firms have also used the massive amounts of data collected online to make predictions about consumer preferences, supplies, and demands for various goods, as well as basic operational parameters such as inventory level and turnover rate. The ability to collect and efficiently analyze the enormous amount of data made available by information technology has enabled firms such as Amazon, Caesar's Entertainment, and Capital One to hone their business strategies and to achieve significant gains in profitability and market shares (McAfee and Brynjolfsson 2012; Davenport 2006).

Our work follows a similar stream in demonstrating the power of using fine-grained data to predict underlying social and economic trends. Unlike previous research and businesses that have primarily used proprietary data, we leverage free and publicly available data from Google to accurately forecast economic trends. Research has shown that online behaviors can be used to reveal consumers' intentions and predict purchase outcomes (e.g., Kuruzovich et al. 2008). We believe that we can rely on digital traces left by trillions of online searches to reveal consumers' intentions and examine their power to predict underlying social and economic trends. The study of individual buying or selling decisions or transactions has been called nanoeconomics (Arrow 1987).

We believe that we are only at the beginning of the data revolution. Newer and more fine-grained data are becoming available every day from various search, social media, and microblogging platforms. These data are made available instantaneously, allowing consumers, business managers, researchers, and policymakers to tap into the pulse of economic activities as they are happening. However, predicting medium or longer-term trends, such as movements in the real estate market, could be easier because they are less prone to short-term manipulations, such as fake Twitter feeds that go viral quickly but die down shortly after they are revealed to be false.

Our methodologies are similar to a recent analysis of flu outbreaks using Google Flu Trends (Ginsberg et al. 2009) and also to parallel research by Choi and Varian (2009) where the authors also correlate housing trends in the United States using search frequencies. Similarly, Scott and Varian (chapter 4, this volume) applied Bayesian variable selection techniques to forecast some present economic trends such as the current consumer sentiment and the current gun sales. Whereas Choi and Varian (2009) and Scott and Varian (chapter 4, this volume) mainly focus on using search frequencies to reveal the current economic statistics, our work attempts to predict *future* economic trends, such as forecasting the price and quantity of houses sold in the future. Within the real estate setting, at least, we show that using search is especially beneficial for predicting the future when compared to

existing models that do not use search data. Furthermore, our work also uses more fine-grained data at the state level, instead of at the level of the whole nation, to provide a more nuanced prediction of the real estate market, which often varies greatly depending on geographical location. In future work, we intend to expand the analysis to the metropolitan statistical areas and other products and services.

3.3.1 Economics of Real Estate

Our work also contributes to the literature on real estate economics. There are two general methodologies for forecasting real estate market trends. The first is technical analysis, similar to techniques used to predict stock market trends. The main assumption for this type of analysis is that the key statistical regularities of changes in housing market trends do not change. Price-trending behaviors might appear to exhibit short-term momentum, but also long-term reversion to the mean (e.g., Case and Shiller 1987, 1989). Glaeser and Gyourko (2006) found evidence of long-term mean reversion in housing prices. They found that, ceteris paribus, if regional prices go up by an extra dollar over one five-year period, they would also drop by thirty-two cents on average over the next five years. The second methodology for predicting housing market trends is to focus on the underlying economic fundamentals. Housing prices should depend on the cost of construction, interest rates available to finance housing purchases, regional income, and even the January temperature (Glaeser 2008). In principle, this suggests that regions with steady building costs and relatively stable income levels should have steady housing prices. However, these economic variables do not seem to fully capture housing price trends. For instance, Dallas is a region with steady fundamentals, but housing prices have been increasing despite the predictions of fundamental analysis.

Some dynamic housing demand models try to incorporate both approaches to predict housing trends (Glaeser and Gyourko 2006; Han 2010). Using dynamic rational expectations to model housing price, Glaeser and Gyourko (2006) detect a mean-reverting mechanism but they cannot explain serial correlation or price changes in most volatile markets. Glaeser (2008) suggests this may reflect sentiment or even "irrational exuberance" in some housing markets, generating a bigger boom and bust cycle than what is predicted by the model (Glaeser 2008).

With the ability to gather billions of search queries over time, Google Trends is essentially aggregating signals of decision makers' intentions to capture some of this overall level of "sentiment." This provides an opportunity to improve predictions in housing markets. Using very simple regression models, we demonstrate that Google search frequencies can be used as a reliable predictor for the underlying housing market trends both in the present and in the future.

3.4 Data Sources

3.4.1 Google Search Data

We collected the volume of Internet search queries related to real estate from Google Trends, which provides weekly and monthly reports on query statistics for various industries. It allows users to obtain a query index pertaining to a specific phrase, such as "housing price." Google Trends has also systematically captured online queries and categorized them into several predefined categories such as "computer and electronics," "finance and business," and "real estate." As Nielsen NetRatings has consistently placed Google to be the top search engine, which processed more than 66.7 percent of all the online queries in the world in December 2012 (comScore 2012), the volume of queries submitted to Google reflects a large fraction of Americans' interests over time.

Google Trends provides a search index for the volume of queries based on geographic locations and time. The search index is a compilation of all Internet queries submitted to Google's search engine since 2004. The index for each query phrase is not the absolute number of queries submitted. Instead, it reports a query index measured by query share, which is calculated as the search volume for the query in a given geographical location divided by the total number of queries in that region at a given point in time.[4] Thus, the reported index is always a number between 0 and 100. The reports on search indices are also much more finely grained than most government reports. Typically, Google calculates the query index on a weekly or a monthly basis, and the index can be disaggregated down to country, state/province, and city levels around the world. For example, in the United States, a query index can be calculated at the state level. A more detailed query index at the metropolitan statistical area (MSA) level can also be computed by specifying the appropriate subregions within a state. Figure 3.1 shows the overall interest in the search category "real estate" using online searches in the United States, using the quarterly averages of the search index. From the graph, interests in housing peaked in 2005 at the height of the recent real estate bubble and fell through 2009 amid the housing market collapse and the onset of the Great Recession.

Our analysis uses a predefined category in Google Trends, "real estate agencies" and "real estate listings" to approximate the overall interest for housing.[5] We also compiled our own sets of phrases related to vari-

4. For details, please refer to http://www.google.com/support/insights//bin/answer.py ?answer=87285.

5. We explored various predefined categories on Google Trends: "apartments and residential rentals," "commercial and investment real estate," "property management," "property inspection and appraisals," "property development," "real estate agencies," "real estate listings," and "timeshares and vacation properties."

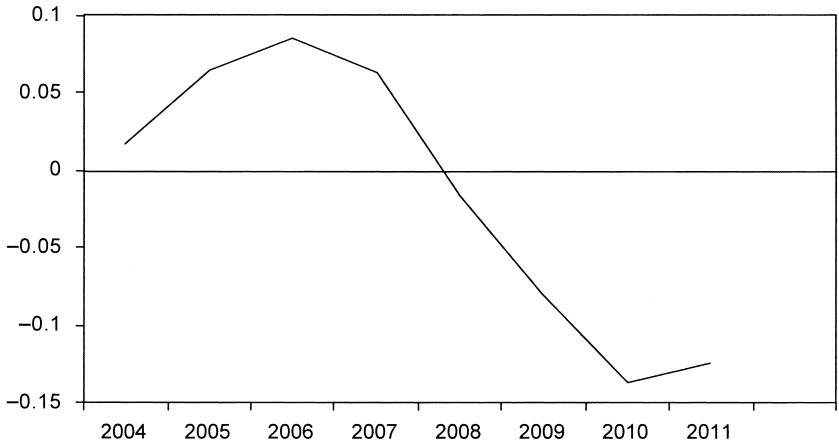


Fig. 3.1 Quarterly search index for "real estate" normalized to total search volume ranging from 0 to 100

ous housing-related transactions such as "housing sales," "home staging," and "home inspection." We hypothesized that these housing-related search indices are correlated with the underlying conditions of the US housing market. To test this hypothesis, we gathered housing market indicators such as the volume of houses sold and the house price index in each US state, all from publicly available sources.

3.4.2 Housing Market Indicators

We collected data on the volume of sales of existing single-family housing units from the National Association of Realtors for all fifty states in the United States and the District of Columbia from the first quarter of 2006 to the third quarter of 2011.[6] This date range coincides with published expert predictions from the National Association of Realtors (NAR). The NAR started publishing their predictions in 2005, but stopped publishing them after the third quarter of 2011. We also obtained the house price index (HPI) for the same period at the Federal Housing Finance Agency, which collects housing prices for nine Census Bureau divisions.[7] The Federal and Finance Agency has calculated the HPI for each state in the United States and the District of Columbia on a quarterly basis since 1975.[8] Because search engine data is only available after 2004 and data on NAR's prediction is only available before the third quarter of 2011, we were able to match the real estate market data with the Google Trend data from the first quarter of 2004 to the third quarter of 2011 for fifty states in the United States and the District of

6. See http://www.realtor.org/research.
7. See http://www.fhfa.gov.
8. See http://www.fhfa.gov/Default.aspx?Page=81.

Columbia. We use roughly half of the sample as training data and use the rest to test our prediction models.

As shown in figure 3.2, panel (a), the number of houses sold in the United States peaked at around 2005 and then declined precipitously soon after, reaching a historical low at the beginning of 2009, and has since recuperated slightly after 2011. The HPI also increased gradually and reached a peak in 2007, two years after the housing sales peak (figure 3.2, panel [b]), and began to fall shortly after. Comparing housing market indicators (figure 3.2) to their associated online search indices (figure 3.1) shows that they appear to be correlated. As shown in figure 3.1, housing-related search peaked in 2005 and gradually declined to its lowest point in early 2009, mirroring the volume of houses sold in figure 3.2, panel (a) and the HPI in figure 3.2, panel (b). This provides some evidence that the search indices are related to underlying housing trends and they could be used to predict both the contemporaneous and future housing market trends.

3.5 Empirical Methods

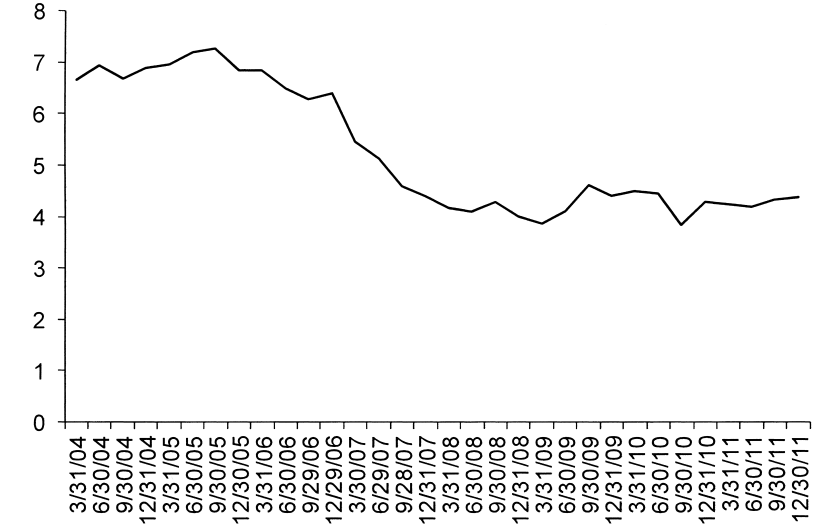
First, we show that search indices are highly correlated with the underlying housing trends. We use a simple seasonal autoregressive (AR) model to estimate the relationship between search indices and housing market indicators—the volume of housing sales and the house price index (HPI). A single class of explanatory variable is studied: search indices for housing-related queries for each state in the United States and the District of Columbia. In this chapter we primarily focus on a simple and consistent set of models to highlight the power of the new data, rather than the sophistication of our modeling techniques, although we found simple linear regression to perform just as well as or even better than more sophisticated nonlinear models. We first estimate the baseline model to predict the current housing sales using only home sales and HPI in the past. Then, we add the search indices to see if they improve predicting the contemporaneous home sales.

$$(1)\quad \text{HomeSales}_{it} = \alpha + \beta_1 \text{HomeSales}_{i,t-1} + \beta_2 \text{HPI}_{it,-1} + \beta_3 \text{Population}_{it} + \Sigma S_i + \Sigma R_j + \Sigma T_t + \varepsilon_{it}.$$

$$(2)\quad \text{HomeSales}_{it} = \alpha + \beta_1 \text{HomeSales}_{i,t-1} + \beta_2 \text{HPI}_{i,t-1} + \beta_3 \text{SearchFreq}_{it} + \beta_4 \text{SearchFreq}_{i,t-1} + \beta_5 \text{Population}_{it} + \Sigma S_i + \Sigma R_j + \Sigma T_t + \varepsilon_{it}.$$

We then examine whether housing-related search indices could forecast future home sales. We only use the past housing statistics to predict the future housing trends because the present housing sales and HPI are not available. Essentially, we are using a two-period lag to predict the future as opposed to a one-period lag to predict the present. Although the govern-

(a) Number of Existing Houses Sold Quarterly.



(b) Quarterly House Price Index

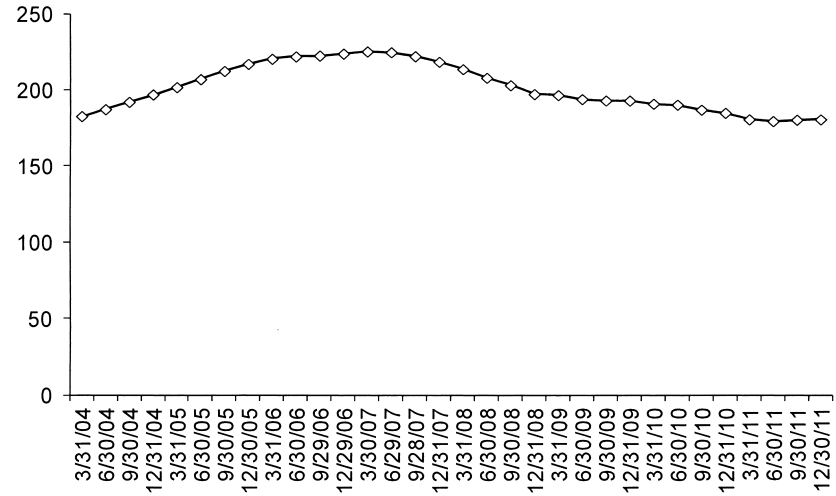


Fig. 3.2 Prices and volumes of existing houses sold in the United States

Note: Panel (a), number of existing houses sold quarterly; panel (b), quarterly house price index.

ment statistics are released with a lag, search frequencies on housing-related inquiries are available in real time and instantaneously down to the daily level. We can thus use both the present and the past search indices to predict future housing sales. Specifically, we use both one-period and two-period lags in the model because they are the most relevant for predictions. Third-order lags can sometimes improve predictions, but in general, higher-order lags fail to have much predictive power. Presumably, housing searches nine months or one year earlier are too early to predict the present and future housing trends because most of these searches likely have already resulted in purchase decisions.

$$(3)\quad \text{HomeSales}_{it+1} = \alpha + \beta_1\text{HomeSales}_{i,t-1} + \beta_2\text{HPI}_{i,t-1} + \beta_3\text{SearchFreq}_{it} + \beta_4\text{SearchFreq}_{i,t-1} + \beta_5\text{SearchFreq}_{i,t-2} + \beta_6\text{Population}_{it} + \Sigma S_i + \Sigma R_j + \Sigma T_t + \varepsilon_{it}.$$

Similarly, we use the same approach to predict the current and future HPI. In the baseline model, we only use the past HPI and the past housing sales to predict the current HPI. We then incorporate the current and past search indices into the baseline model.

$$(4)\quad \text{HPI}_{it} = \alpha + \beta_1\text{HPI}_{i,t-1} + \beta_2\text{HomeSales}_{i,t-1} + \beta_3\text{Population}_{it} + \Sigma S_i + \Sigma R_j + \Sigma T_t + \varepsilon_{it}$$

$$(5)\quad \text{HPI}_{it} = \alpha + \beta_1\text{HPI}_{i,t-1} + \beta_2\text{HomeSales}_{i,t-1} + \beta_3\text{SearchFreq}_{it} + \beta_4\text{SearchFreq}_{i,t-1} + \beta_5\text{Population}_{it} + \Sigma S_i + \Sigma R_j + \Sigma T_t + \varepsilon_{it}.$$

Lastly, we predict the future HPI by adding the present and past search indices into the model. In addition to exploring various lags, we also explored nonlinear functions of the search indices to see if they improve model fit and predictions.

$$(6)\quad \text{HPI}_{it+1} = \alpha + \beta_1\text{HomeSales}_{i,t-1} + \text{HPI}_{i,t-1} + \beta_2\text{SearchFreq}_{it} + \beta_3\text{SearchFreq}_{i,t-1} + \beta_5\text{Population}_{it} + \Sigma S_i + \Sigma R_j + \Sigma T_t + \varepsilon_{it}.$$

For all the models above, we apply state- and region-level dummies in order to control for any time-invariant influences, such as the demographics of a state/region, and any statewide/region-wide policies that may affect real estate purchase decisions. We then train these models using data between the first quarter of 2006 and the fourth quarter of 2008[9] to find a set of

9. We chose this period for training because it roughly divides the data in half. We also tested a tenfold cross-validation approach that randomly partitions the sample into ten equal sizes regardless of the timing of the data. Nine sets are then used to train the model and the tenth set is used to test the model. While we were able to improve the predictive accuracy using cross validation, we chose to train the model only using the past data because it is a more conservative estimate. It also reflects the reality that we should not know anything about the future in the training data to make predictions about the future.

search indices that best predict the present and future housing indicators. We then use these indices and their associated estimates to predict housing trends from the first quarter of 2009 to the third quarter of 2011. For each prediction, we calculate the mean absolute error (MAE)[10] to examine the accuracy of our predictions that use search indices when compared to the predictions from the baseline model, as well as from the National Association of Realtors. The mean absolute error is simply the deviation away from the actual value.

$$MAE = \frac{1}{N}\sum_{t=1}^{N}\left|\frac{y_t - \hat{y}_t}{y_t}\right|. \tag{7}$$

In addition to housing predictions, we also examine whether housing-related search queries can also spur future economic activities in complementary industries. For example, if consumers' intentions can be revealed through online search, we may also expect a surge in Internet queries about home appliances after observing a rise in home sales. Because new homeowners may plan to purchase appliances to furnish their property, tracking their online search behavior allows us to detect their intention to purchase home appliances. Accordingly, we correlate housing sales with the search index for home appliances. If search index for home appliance can translate into actual purchases, we would expect a rise in search frequencies for home appliances, spurred from home sales, to indicate a rise in their future demands as well.

$$\text{HomeApplianceSearch}_{it} = \alpha + \beta_1 \text{HomeSales}_{it} + \beta_2 \text{HomeSales}_{i,t-1} + \varepsilon_{it}. \tag{8}$$

3.6 Empirical Results

First, we compare predictions between the baseline model and the model that uses search indices. We used the model to predict the present home sales and HPI as well as the future home sales and HPI in the next quarter. Although our model can be used to predict even more fine-grained forecasts, such as monthly or even weekly housing trends, we chose to forecast at the quarterly level because the government only releases state-level housing sales and HPI every quarter. To calculate our predictions' accuracy, we aggregated the weekly search data into quarterly data. Furthermore, we also compare our predictions with the forecasts of quarterly housing sales released by the National Association of Realtors (NAR). The NAR does not predict

10. We also use other metrics such as the mean squared errors (MSE) to evaluate the accuracy of our predictions. The results do not qualitatively change when we use MSE. In fact, we find our improvements using MSE are even better than using MAE. Thus, we conservatively reported the MAE values.

Table 3.1 Linear regression to predict the present home sales using search frequency

	Quarterly sales					
Dependent var.	(0)	(1)	(2)	(3)	(4)	(5)
$Sales_{t-1}$	0.864***	0.864***	0.819***	0.842***	0.806***	
	(0.0125)	(0.0125)	(0.0142)	(0.0130)	(0.0144)	
HPI_{t-1}	–0.140***	–0.140***	–0.158***	–0.177***	–0.188***	
	(0.0175)	(0.0175)	(0.0175)	(0.0196)	(0.0195)	
Real estate agencies$_t$		16.55***	17.09***		13.41***	48.47***
		(2.450)	(3.424)		(3.523)	(6.415)
Real estate agencies$_{t-1}$			–0.780		1.170	33.04***
			(3.414)		(3.451)	(6.297)
Real estate listing$_t$				23.36***	18.41***	37.37***
				(4.797)	(4.917)	(9.007)
Real estate listing$_{t-1}$				–8.062	5.503	–13.16
				(4.831)	(4.876)	(8.728)
Obs.	1,561	1,561	1,561	1,561	1,561	1,561
Controls	Quarters, states, regions, population	Quarters, states, regions, population	Quarters, states, regions, population	Quarters, states, regions, population	Quarters, states, regions, population	Quarters, states, regions, population
States	51	51	51	51	51	51
Adjusted R^2	.973	.980	.981	.982	.983	.970

Note: Huber-White robust standard errors are shown in parentheses. Quarterly sales are in 1000s.
***Significant at the 1 percent level.
**Significant at the 5 percent level.
*Significant at the 10 percent level.

future HPI, and thus we cannot compare our model with the NAR's when predicting the future HPI.

3.6.1 Predicting Home Sales Using Online Search

Table 3.1 explores the relationship between housing sales and housing-related search indices that could support the use of search indices in predictions. All models in table 3.1 are based on a seasonal autoregressive (AR) model, which assumes that the sales in the future are related to sales in the past. We see a broad support for the AR model because the lagged sales are strongly correlated with the contemporary sales. We also applied a state-level, fixed-effect specification to eliminate influence from any time-invariant factors, and we use seasonality dummies to control for time-specific changes. In addition, we also included the state population and region dummies to improve the fit of the model. To capture online interests for purchasing real estate properties, we use a search index of a predefined category in Google Trends—"real estate listing"—that contains all queries pertaining to real estate listings and advertisements. We also use the "real estate agencies" category to approximate home buying activities. We assume people who are

looking for real estate agents and real estate listings online are more likely to participate in a real estate transaction than those who search for other related queries such as property management.

First, we estimate the baseline model to predict the present home sales using only the past home sales and the past HPI. As shown in the baseline AR(1) model (model 0), the past home price and sales are highly correlated with the current home sales. We then examine various search indices related to the real estate market[11] and find two categories—"real estate agencies" and "real estate listings"—to best predict the contemporaneous sales. Overall, we find that the contemporaneous search indices for "real estate agencies" and "real estate listings" are statistically significantly correlated with the present home sales. As shown in Model 1, a 1-percentage point increase in the current search index for the category "real estate agencies" is associated with 16,550 additional sales for existing homes in the contemporaneous quarter. The average state-level home sales are 112,037 units per quarter, so 16,550 units of additional sales represent a 14.8 percent increase from the state average. Similarly, a 1-percentage point increase in the search index for the category "real estate listing" is correlated with 23,360 houses sold in the present quarter (Model 3). We explore the effect of using both the present and past search indices for "real estate listing" and "real estate agencies" in Model 4. The present search indices for both categories are again positively correlated with sales, but the past indices are not. However, the adjusted R^2 improved slightly if both the present and the past search indices are included. In Model 5, we only use the search indices, without lagged home sales and HPI, to predict housing sales, and the results are similar to what is shown in Model 4. This suggests that using online search frequencies alone can predict future sales. The adjusted R^2 was just slightly below the baseline model if the past sales and HPI were included. Overall, results in table 3.1 show that online search behaviors are highly correlated with the contemporaneous home sales.

To examine whether our model can actually predict the contemporaneous home sales, we generate a set of one-quarter-ahead predictions. We first create a training set using data from the first quarter of 2006 to the fourth quarter of 2008. Using these eleven quarters of data for fifty states and the District of Columbia, we select a set of features or variables that best predict the contemporaneous sales. We also experimented with various functional forms and the window of data to use that would give the best predictive results in the training set. We find a simple linear model with search terms to consistently provide superior prediction results. For predicting the present sales, using the previous eight quarters of data gives the best consistent

11. We also examined the following predefined categories on Google Trends: "apartments and residential rentals," "commercial and investment real estate," "property development," "property inspection and appraisals and property management," "real estate listings," "real estate agencies," and "timeshares and vacation properties."

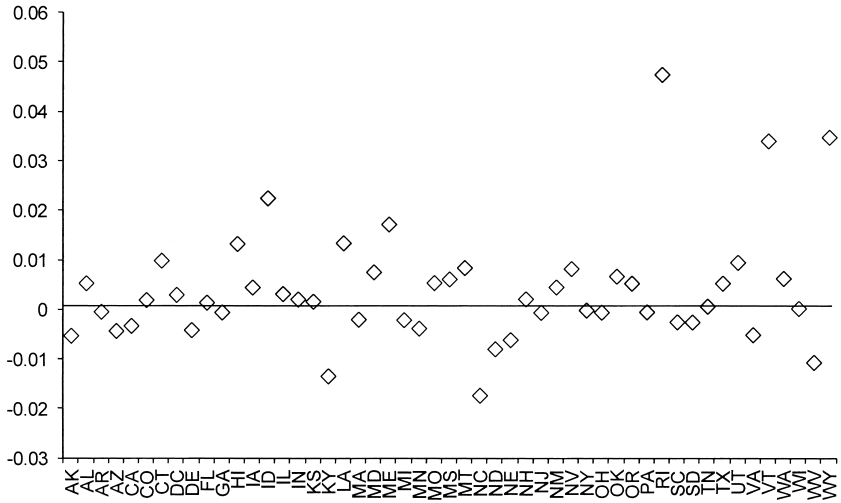


Fig. 3.3 The Y-axis indicates the average difference in MAE between the baseline model (equation [1]) and the model that uses search indices (equation [2])

Note: We use predictions from the first quarter of 2009 to the third quarter of 2011. When the dots are above the zero line, the baseline MAE is worse than the MAE from the model that uses search.

results.[12] In addition to using "real estate agencies" and "real estate listings," we also explored other predefined categories from Google Trends as well as our own set of search phrases. However, we find "real estate agencies" and "real estate listings" are the best features for predicting the present sales in the training set. Next, we use the best-predicted model and estimates to predict sales from the first quarter of 2009 to the third quarter of 2011. To gauge how accurate our predictions are compared to the actual real estate indicators, we use mean absolute error (MAE), as shown in equation (7).

The mean absolute error (MAE) using our model with search indices (equation [2]) is 0.170 (17 percent deviation from the actual value), compared to 0.174, the MAE of the baseline model. Simply adding search terms in the linear model provides a 2.3 percent improvement over the baseline and it is statistically significant at $p < 0.05$ percent (Model 0). We graphed the differences in MAE between the baseline model and the model that uses search indices in figure 3.3, specifically as MAE(baseline)—MAE(search). Dots above the zero line indicate that predictions are better with the added search indices than with the baseline model alone. As shown in figure 3.3, the MAE for the baseline is mostly worse than for our predictions that use

12. We also experimented with using the previous four, six, eight, twelve, twenty-four, and thirty-six quarters to predict the contemporaneous sales and, while there was little difference, using the previous eight quarters appears to produce the most accurate predictions.

search. While the improvement is relatively modest on average, the variation for the improvement among different states is large. In general, predictions using search indices are better for states that have a high volume of sales, possibly indicating that a high volume of sales is also indicative of having more real estate-related online searches. However, because search indices do not indicate the absolute number of searchers, it is difficult to ascertain if more online searches lead to better predictions. We find the correlation between sales and the MAE differences to be negative.

Next, we apply our methods to predict the future housing trends using available data today that include the past housing statistics and the present and past search indices. We only use the housing statistics from the previous quarter because when making a given prediction, the present housing statistics would not be available. Unlike housing statistics, which are always released with a lag, search indices are obtainable almost instantaneously, allowing us to incorporate virtually real-time search behaviors to predict future real estate trends.

We first use the training data to find the best statistical model to predict future sales. The best model we found is a linear prediction model using the past eight quarters of data. After experimenting with various housing-related search terms and predefined search categories from Google Trends, we find the best predictors are the current index for "real estate agencies" as well as its one-quarter and two-quarter lags. Interestingly, the "real estate listings" index no longer adds much predictive power if indices on "real estate agencies" are included. Using only the present and the past indices on "real estate agencies" as well as the past statistics on HPI and home sales, we predict the future home sales and plot the difference between the MAE of the baseline model and the MAE of our predictions in figure 3.4: MAE(baseline)—MAE(search). For most of the states, predictions using search indices outperform the baseline predictions, especially for states where the sales volume is high. For states with lower volumes of real estate transactions, adding search indices does not improve the predictions. Overall, the MAE for predictions using search indices is 0.172 (or 17.2 percent deviation from the true value) whereas the baseline MAE is 0.185. This is a 7.1 percent improvement over the baseline model and it is statistically significant at $p < 0.05$.

Interestingly, this result suggests that search indices are actually better at predicting the future home sales than they are at predicting contemporaneous sales (7.1 percent vs. 2.3 percent over the baseline). Perhaps future sales are more correlated with past search indices because buying and selling a house often takes more than a quarter. For example, while there are many factors affecting the duration of a sale, the average time to sell a home in the United States is ten months in 2011.[13] Thus, search activities on the Internet,

13. Statistics come from the Accredited Seller Agent Council. See http://www.realty101.com/what-is-the-average-time-to-sell-a-home.

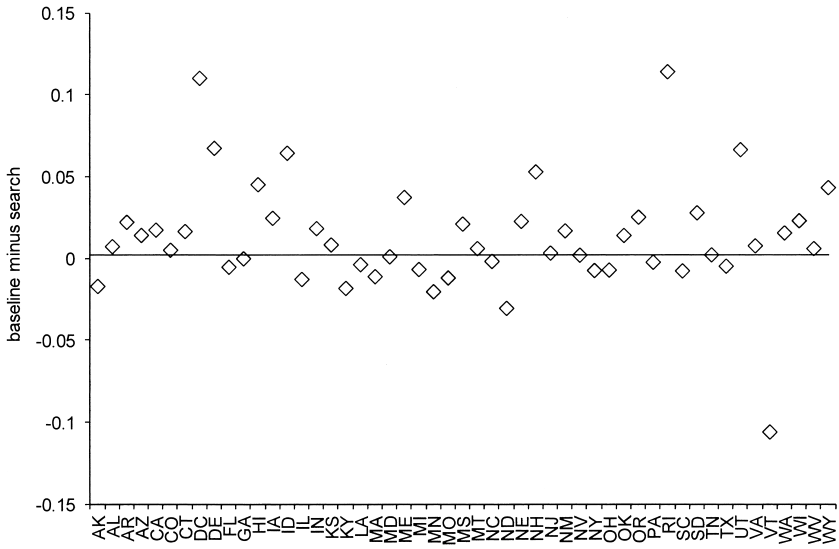


Fig. 3.4 MAE differences between the baseline model and predictions using search indices

at least from the seller side, can potentially forecast home sales ten months in the future. Another reason why using search to predict the future is better could be because the baseline model is not as good at predicting the future as it is with predicting the present. Using housing statistics from the previous quarter can well capture the trends in the contemporaneous quarter better than trends in the future. Thus, behavioral information such as the search indices could be relatively more valuable for these predictions than information provided by the two-period lags of home sales and HPI.

While we find that using search indices can improve prediction outcomes in comparison to the baseline model, it is important to also compare our model with real forecasts from experts in the field. Thus, we collected data from the National Association of Realtors, who release quarterly forecasts for US home sales. To predict home sales for the entire United States, we aggregated the state-level predictions that use search indices in each quarter. We compared NAR's forecasts with our predictions from the second quarter of 2009 to the third quarter of 2011, for a total of ten quarters. For predicting the present home sales, we find that our predictions have been slightly better than NAR's but the difference is not statistically significant. However, our predictions were considerably better than NAR for predicting future home sales. The MAE for the National Association of Realtor's forecast is 0.110 while the MAE for the model that uses search indices is 0.084, a 23.6 percent improvement over the estimates from real estate experts. Results are summarized in the table 3.2. This again shows the power of using search indices for predicting the future. Using a simple linear prediction model

Table 3.2 **Comparing with predictions from the National Association of Realtors for home sales in the United States**

MAE for $sales_{t+1}$	Obs.	Mean	Std. err.	Min.	Max.
Search	10	0.084	0.031	0.012	0.156
NAR	10	0.110	0.026	0.050	0.169
Diff.		23.6%		$p < 0.01$	

with search indices, we are able to outperform predictions from established experts in the field.

One concern is that NAR tends to overpredict the existing homes sales and that is why our predictions are superior. We tested this hypothesis and found that the NAR is indeed more likely to overpredict than to underpredict sales. On average, NAR overpredicted sales in twenty out of the twenty-two quarters from 2006 to 2011, and the error rate was 7.8 percent more than the actual sales.[14] By contrast, our predictions using search overestimated the US home sales in only six out of ten quarters with an average error of 2.4 percent. A reason for the overprediction in both NAR's model and our model could be attributed to the time period of the prediction. Between 2008 and 2010, the US real estate market experienced one of the biggest busts in recent history. In a more stable period, the overprediction could be less severe.

3.6.2 Predicting the House Price Index Using Online Search Data

In table 3.3, we explore the relationship between the housing-related search indices and HPI, which is calculated based on a modified version of the weighted-repeat sales (WRS) methodology proposed by Case and Shiller (1989). All models in table 3.3 use a fixed-effect specification on an AR model with region, population, and seasonality controls. Similar to table 3.1, the purpose of this table is to illustrate that search indices are correlated with HPI and could be used for predictions. As expected from the baseline AR model (Model 0), the lagged HPI and lagged sales are positively correlated with the present HPI. In Model 1, we estimate the correlation between the current search index for "real estate agencies" and the HPI and find that a 1-percentage point increase in the search index is associated with an increase of 5.986 points in HPI. However, the past search index on "real estate agencies" from the previous quarter does not have a statistically significant correlation to the present HPI (Model 2). Next, we introduce both the current and the past indices for "real estate listings" in Model 3. We find that the current search index for "real estate listings" is positively correlated with

14. The error rate is calculated as (actual sales–NAR prediction)/actual sales. This formulation uses the actual error as opposed to MAE that uses the absolute value of the error.

Table 3.3 **Linear regression of HPI on the search index related to real estate and real estate agencies**

	HPI_t					
Dependent var.	(0)	(1)	(2)	(3)	(4)	(5)
$Sales_{t-1}$	0.959***	0.952***	0.951***	0.952***	0.947***	
	(0.006)	(0.006)	(0.006)	(0.006)	(0.006)	
HPI_{t-1}	0.086***	0.700***	0.069***	0.081***	0.066***	
	(0.004)	(0.0052)	(0.005)	(0.004)	(0.005)	
Real estate $agencies_t$		5.986***	5.069***		3.520***	6.817
		(0.780)	(1.107)		(1.138)	(4.543)
Real estate $agencies_{t-1}$			1.268		2.361**	9.146**
			(1.088)		(1.104)	(4.414)
Real estate $listing_t$				8.951***	7.919***	16.82***
				(1.528)	(1.560)	(6.246)
Real estate $listing_{t-1}$				–5.116***	–4.989***	51.97***
				(1.514)	(1.523)	(5.945)
Obs.	1,561	1,561	1,561	1,561	1,561	1,561
Controls	Quarters, states, regions, population	Quarters, states, regions, population	Quarters, states, regions, population	Quarters, states, regions, population	Quarters, states, regions, population	Quarters, states, regions, population
States	51	51	51	51	51	51
Adjusted R^2	0.987	0.986	0.987	0.987	0.987	0.987

Note: Huber-White robust standard errors are shown in parentheses. Quarterly sales are in 1000s.
***Significant at the 1 percent level.
**Significant at the 5 percent level.
*Significant at the 10 percent level.

the contemporaneous HPI while its one-period lag is negatively correlated with HPI. Finally, we include the present and the past search indices for both "real estate listings" and "real estate agencies" in Model 4 and find that all search indices are correlated with the present HPI. The fit of the model also improves slightly. These results give us confidence that incorporating the present and past search indices from the two search categories can help predict the contemporaneous HPI.

Next, we predict the contemporaneous HPI from the first quarter of 2009 to the third quarter of 2011 after finding the best-fitted model from the training data set. Among various search terms and real estate-related categories, we continue to find the contemporaneous and the one-period lag of search indices on "real estate agencies" and "real estate listings" to best predict the present HPI. In contrast to using the previous eight quarters of data to predict home sales, we find that using data from the past four quarters can best predict the present HPI. Overall, we find that our predictive accuracy improves on the baseline model by 2.54 percent, which is comparable to the results on predicting the present home sales. We show the state-by-state

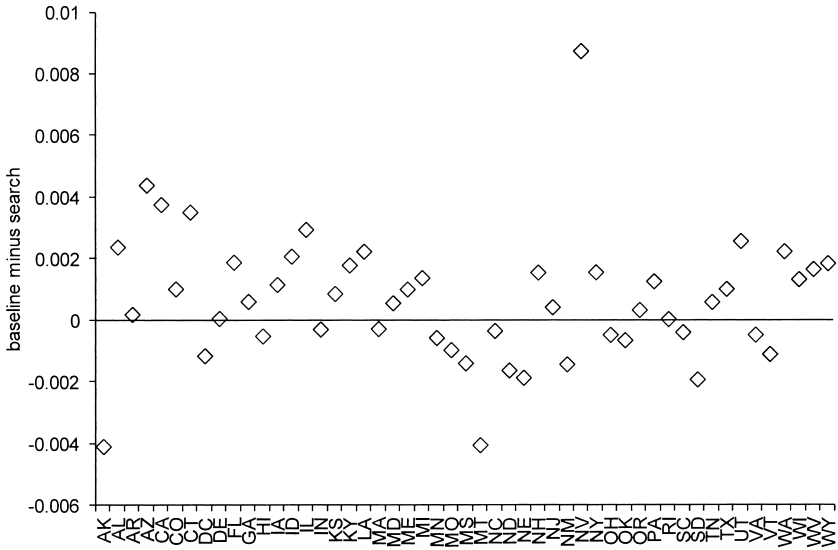


Fig. 3.5 Difference in MAE between the baseline model and the search-based model

scatter plot for the MAE difference between the baseline and the search indices model (figure 3.5). Again, dots above the zero line represent states where the prediction using search outperforms the baseline model, whereas the opposite is true for dots below the zero line.

Overall, using search, we are able to predict thirty-nine states better than the baseline model, but our predictions are particularly bad for a few states, such as Montana and South Dakota. These states tend to have fewer transactions on housing sales than other states. Similar to what we found for home sales, search indices help predictions the most for states where the sales volume is high.

Furthermore, predicting HPI may just be inherently more difficult than predicting home sales. Although home sales can increase when either the housing demand or supply changes, HPI would increase only when the demand for housing is increased without a corresponding increase in supply, and decrease when the supply is increased without a corresponding increase in demand. It is difficult to know whether the search queries in general categories such as "real estate agencies" or "real estate listings" are coming from the demand side or the supply side, and thus it is much harder to predict HPI than the volume of home sales. For example, both sellers and buyers need real estate agents, so an increase in the search index related to real estate agencies could come from both the supply and the demand sides that can either increase or decrease home price. To address this issue, we tentatively aggregated some

search terms relating to buyers' activities—such as home financing, mortgage, and home inspections—and also some search terms related to sellers' activities only, such as home staging. For example, home buyers are more likely to look for loans than sellers, whereas sellers are more likely to hire a staging company to make the property more appealing to the highest number of potential buyers. We would therefore expect that an increase in search frequencies related to financing and loans to shift the demand curve, while a similar increase for searches related to home staging is more likely to shift the supply curve for housing. We see some evidence that home financing is positively correlated with HPI, suggesting it may be shifting the demand outward. Currently, we have not found a set of queries that can consistently identify shifts in the supply curve. However, because of the fine-grained nature of the search terms, we are hopeful that indices can be created to precisely tease out a shift in the demand curve from a shift in the supply curve.

To explore how search indices can be used to predict future HPI in the next quarter, we use the training data to find the best features that can be used to predict the future HPI. In addition to using the present and past search indices of "real estate agencies" and "real estate listings," we also explored some nonlinear forms of the search indices, such as their quadratic terms. Overall, we find the best predictors continue to be the present and past indices for "real estate agencies" and "real estate listings." Interestingly, we find the quadratic terms of "real estate agencies" to also help with the predictive accuracy in the training set. Thus, we include these variables to predict the future HPI from the second quarter of 2009 to the third quarter of 2011. We plot the difference in MAE between the baseline model and the search model for each state of the United States in figure 3.6. For most states, predictions using search were better than the baseline model, though the variance among states is even higher than predicting the present HPI. We predicted eleven quarters for fifty states and the District of Columbia. The baseline MAE is 0.027 and the MAE for the model that uses search is 0.026, about a 2.96 percent improvement in accuracy and statistically significant at the $p = 0.01$ level. Unfortunately, the National Association of Realtors does not forecast HPI, at least from public-available sources, and thus we are not able to compare our HPI predictions with NAR's.

We summarize our results in table 3.4. Whereas using search frequencies can improve the accuracy of prediction for both the present and future home sales as well as HPI, it is actually more effective for predicting the future than predicting the present. Because a housing transaction that often takes months to more than a year to complete, search indices in the present can be particularly useful to forecast future housing indicators. Search frequency data are more effective for predicting sales volume than for predicting HPI, in part because of the difficulty of distinguishing supply and demand shifts, which influence home prices.

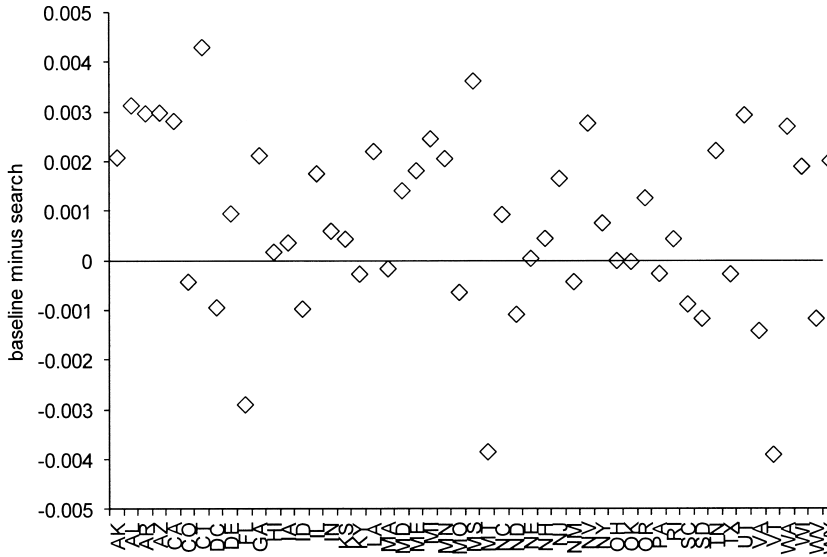


Fig. 3.6 The MAE difference between the baseline model and our prediction model

Table 3.4 Summary of MAE for predicting the present and the future housing trends

	Obs.	MAE search	MAE baseline	Improvement over baseline (%)
$Sales_t$	561	.170	.174	2.3**
HPI_t	561	.026	.027	2.45***
$Sales_{t+1}$	561	.172	.185	7.1**
HPI_{t+1}	561	.026	.027	2.96***

***Significant at the 1 percent level.

**Significant at the 5 percent level.

*Significant at the 10 percent level.

3.6.3 Predicting the Demand for Home Appliances

Lastly, we explore trends in home appliance sales. We expect that housing sales would spur interest in buying home appliances, increasing their demand in the future. To gauge the overall interest in home appliances, we use the search index for the "home appliance" category from Google Trends and show its relationship with home sales (table 3.5). We observe that the current home sales are not correlated with the contemporaneous search index for home appliances (Model 1, Model 4). But with a six-month lag, each one thousand houses previously sold is correlated with a 1.14 percentage point increase in the search index for home appliances. Because

Table 3.5 **Linear regression on search terms related to home appliances and the volume of housing sales**

	Search terms on home appliances (quarterly)			
	Fixed effect			
Dependent var. search terms related to home appliances	(1)	(2)	(3)	(4)
Home sale$_t$	–.054			0.188
	(.0001)			(0.0004)
Home sale$_{t-1}$		–.020		–0.627
		(.0001)		(0.393)
Home sale$_{t-2}$			.590**	1.140***
			(.3)	(0.427)
Obs.	254	203	152	152
Controls	Quarters	Quarters	Quarters	Quarters
States	51	51	51	51

Note: Huber-White robust standard errors are shown in parentheses.
***Significant at the 1 percent level.
**Significant at the 5 percent level.
*Significant at the 10 percent level.

buyers move into their new properties first before making major purchases (and often research such purchases), it is natural that the number of online searchers for home appliances would increase after a consumer has already bought a house. Thus, we may expect the online search for home appliances to lag behind housing sales. The actual demand for home appliances may rise after this increase in the appliance search index if some of the online searches translate into future sales. Similarly, we correlated the housing real estate-related search index with the home appliance search index and we find that they are also positively correlated. This highlights the linkages between home sales and other parts of the economy that may complement real estate purchases.

3.7 Implications

Twenty-five years ago, Herbert Simon (1984, 40) observed:

> In the physical sciences, when errors of measurement and other noise are found to be of the same order of magnitude as the phenomena under study, the response is not to try to squeeze more information out of the data by statistical means; it is instead to find techniques for observing the phenomena at a higher level of resolution. The corresponding strategy for economics is obvious: to secure new kinds of data at the micro level.

Today, advances in information technology in general, and in Internet search query data in particular, are making Simon's vision a reality. Who

could have imagined that we would be observing literally billions of consumer and business intentions to buy or sell before the consumer sets foot in a store and transactions occur in the marketplace? Yet that is what search query data enables us to observe. Even more, we can do so at nearly zero cost, virtually instantaneously and at remarkably fine-grained levels of disaggregation. These data are increasingly available to ordinary consumers, business people, and researchers of all types.

We have found that analyzing online search data with relatively simple models can yield more accurate predictions about the housing market than were previously possible. If online search patterns can be construed as a broad indicator of interest within a group, they can also be used as a reliable predictor to forecast economic activities. By analyzing housing market trends, we find evidence that search indices add substantial power to predicting the underlying economic trends, and that predictions using search indices can outperform predictions from experts in the field, such as the National Association of Realtors. This supports the hypothesis that Web search can be used to predict present and future economic activities. For example, housing-related searches might be used to predict turning points in economic cycles.

Currently, we are able to make fairly accurate predictions using simple linear prediction models and a few predefined real estate categories in Google Trends. Because of the fine-grained nature of these data, they can be aggregated in many different ways to predict specific underlying economic shifts. For example, instead of using rough categories such as "real estate agencies" or "real estate listings," we can create our own sets of words specific for gauging changes in demand as well as changes in supply. Distinguishing the search indices of the supply side from those of the demand side can more accurately detect what is driving changes in the real estate market. Similarly, we can test more fine-grained predictions about the real estate market beyond sales and price. For example, search indices can be created to measure the interest of people buying homes as opposed to renting, or whether new construction activities are growing over time or not. Because of the fine-grain nature of individuals' search queries, it is possible to construct different types of indices and quickly test their validity in predicting various real estate trends and beyond.

Timely and accurate predictions about the housing market can benefit a wide array of industries, such as construction and home appliances, as well as individuals, such as homebuyers and sellers. Because buying a home is the single biggest expenditure and one of the biggest financial decisions for most people, obtaining accurate and timely information can help them make informed decisions and potentially save tens of thousands of dollars for the average family. Similarly, businesses that depend on the housing market can benefit from this simple use of Internet search data. Timely and accurate forecasts of housing demand would allow the construction industry to

improve future plans for developments and thus reduce the probability of experiencing the housing boom and bust cycles. Accurate housing market forecasts can also help the home appliances industry to manage its inventory.

Currently, economists, managers, and investors primarily rely on housing data released from the government and trade groups such as the National Association of Realtors to understand the current housing market and forecast future market trends. However, government and trade group data are released with a delay and often with pending revisions. Furthermore, they do not provide fine-grained reports at the town level, which is the level needed for buyers and sellers to make informed decisions. With easy access to billions of online search frequencies, it is now possible to use a simple technology to cheaply collect timely, accurate, and fine-grained data about the housing market. Not only does Google Trends provide weekly reports on the volume of housing-related queries, it also offers a detailed regional analysis at country, state, and city levels. By leveraging microdata collected from Google Trends, investors and policymakers can obtain deeper insight about the housing market in order to make informed decisions.

3.7.1 Other Applications and Future Research

Not only can search data be used to provide better predictions about the housing market using Google Trends, but search data can also be used in many other contexts to predict future economic activities. Scott and Varian (chapter 4, this volume) demonstrate cases of using search indices from Google Trends for "nowcasting." Specifically, they used Bayesian variable selection methods to forecast the current consumer sentiment and the current gun sales. Similarly, Choi and Varian (2009) show that search engine data can be used to forecast other macroeconomic indicators such as retail, car sales, travel, and housing. In addition to predicting the present, we find that Google Trends can be used to predict the outcome of a standards war in the technology sector. We were able to track the progression of the standards war between HD-DVD and Blu-ray. Google Trends and search indices were prescient in predicting that Blu-ray would win in the end. Similarly, we can also use search frequency to predict the market share of an electronic product or an operating system such as Macintosh. Instead of paying a premium for industry reports, Google Trends can be used to predict if a particular technology would gain market shares.

It appears that predicting the future using search engine data can be much better than many existing models, especially for a market that does not change instantaneously, such as real estate and employment. Presumably, finding a job or a place to live often takes many months and thus the signals aggregated from search can be very helpful for predicting the future trends in real estate or the labor market.

Because of the fine-grained nature of the search queries, there are many ways to dissect the data for various prediction purposes. Furthermore, search

data can be combined with other types of nanodata, such as various digital traces from digital and social media. Together, these data allow consumers, managers, researchers, and policymakers to tap into the pulse of economic activities to make more informed decisions.

Many other types of predictions are possible now using Google Trends. For example, instead of waiting for the government to release labor statistics every month, we can use Google Trends to predict the current unemployment rates by using search indices related to job search activities. As job search activities are increasingly done through the Internet, search queries could be far more powerful in predicting the unemployment rates than government surveys. Similarly, instead of waiting for industrial reports to become available, we can use Google Trends to predict sales such as automobiles sales. As purchasing a home, searching for a job, and buying a car can all incur significant search costs and consumers often conduct extensive research online before making purchase decisions, the digital trace left from the searching process can be tremendously valuable for making predictions. We expect research to validate many similar types of predictions in the future.

However, this approach also has important limitations. Precisely because search query data can be easily collected and used to make predictions, they are also prone to be manipulated. For example, searchbots could be used to generate irrelevant search queries to substantially change search indices and consequently influence many economic decision-making processes. Future work should also focus on how to detect data manipulations. Furthermore, when major search engines change their search algorithms or user interface, predictability of search queries could also change significantly. Because some search engines conduct frequent experiments and adjustments to their algorithms, a search query that works well for today's prediction may not work well tomorrow. Thus, it is important to monitor and update the set of keywords used in each search index for prediction purposes. An important focus for research is to improve the methods to generate search keywords and validate them over time. See, for example, the "crowd-squared" approach that draws on a set of users to suggest potential keywords (Brynjolfsson, Geva, and Reichman 2014). If search queries were to have important implications for making important policy and economic decisions, it is also important to ensure key stakeholders, such as the search engine providers, would not be able to manipulate the search data to their own benefits.

3.8 Conclusions, Limitations, and Future Work

Today, due to advances in IT and IT research, we are gaining the capability to observe microbehaviors online. Rather than rely on costly, time-consuming surveys and census data, predefined metrics and backward-looking financial reports, today's social science researchers can use query data to learn the intentions of buyers, sellers, employers, gamers, engineers,

lovers, travelers, and all manner of other decision makers even before they execute their decisions. It is possible to accurately predict what will happen in the marketplace days, weeks, and even months in the future with this approach. Search technology has revolutionized many markets, and it is now revolutionizing our research.

This is an exploratory study investigating whether online search behavior from Google Search can predict underlying economic activities. Using housing sales data, we find evidence that search terms are correlated with future sales and prices in the housing market. This evidence lends credibility to the hypotheses that Web search can be used to predict future economic activities, for example, when the economy may recover from the recent recession. We are aware of the fact that Google search queries do not represent all the online housing search activities nor do they represent a demographically random sample of all home sellers and home buyers. Some consumers may bypass the search engine all together and go directly to certain websites, such as Realtor.org, when considering buying and selling a home. Others might have a long-standing relationship with a trusted realtor or do not use the Internet. Using Google Search alone would miss these types of consumers. However, despite missing some segments of the population, we can still predict the housing sales and housing price using only online search captured by Google, demonstrating the power of online queries in forecasting economic trends.

Ultimately, microdata collected using Google Trends may prove to be one of the most powerful tools for helping consumers, businesses, and government officials make accurate predictions about the future so that they can make effective and efficient decisions. This data distills the collective intelligence and unfiltered intentions of millions of people and businesses at a point in their decision-making process that precedes actual transactions. Because search is generally not strategic, it provides honest signals of decision-makers' intentions. The breadth of coverage, the level of disaggregation, and the speed of its availability is a radical break from the majority of earlier social science data. Even simple models can thus be used to make predictions that matter.

Of course, there are many obstacles yet to overcome and refinements to be made. For instance, paradoxically, as businesses and consumers come to rely on query data for their decision making, as we expect they will, there will be incentives for opposing parties to try to degrade the value of the data, perhaps by generating billions of false or misleading queries. This will in turn call for countermeasures and perhaps the golden age of simple models using these data will be brief. However, more than four years have passed since we first started using Google Trends to forecast real estate trends. We are encouraged to see that search indices continue to have the power to predict the future, as we have shown in this chapter. Informational value derived from search indices has not been absorbed into economic equilibria, as many

have argued. Instead, its effect, at least for the real estate market, has persisted over time. Meanwhile, new types of nanodata have become available, such as Twitter feeds, social networking data, cell phone location data, and various other digital traces of consumers' daily lives. Along with search, detailed nanodata have continued to proliferate at a pace that has far outgrown our ability to manage and use these data appropriately. For example, a single simple hoax message—claiming that two bombs had exploded at the White House—using a single Twitter feed on April 23rd, 2013, seemingly caused the Dow Jones Industrial Average to drop by 145 points in less than five minutes. Perhaps the instantaneous connectivity of Twitter and the potential short-term nature of stock price fluctuations enable a fake Tweet to quickly go viral and affect the actions of many high-frequency traders. Consequently, the stock market erased $136 billion in equity in a matter of minutes. However, this type of gaming is less likely to happen for markets that are not prone to change so quickly, such as home buying and selling. Because selling a home can take months to complete, a swing in search indices on housing queries in an hour or a day would not make a significant impact on the predictions of future real estate trends. These types of hacking are often quickly discovered using tests for statistical anomalies, making long-term manipulation more difficult. Future research should investigate what types of markets search and other forms of digital trace are most useful for predictions and what types of markets are susceptible to gaming. We have so far identified that the rate of market changes may play a role, but many other factors could also be at play.

Ultimately, the availability of various digital traces[15] has grown so quickly over the years that they have outpaced our ability to understand and use them effectively. It is thus important for future research to investigate how to integrate and use them in a meaningful fashion to understand underlying consumer sentiments and economic consequences. Through improved understanding, we may be able to better distinguish malicious and faulty information from the true economic signals, although it may also be a cat-and-mouse game where malicious attack will always happen on strategic tools that can affect decision making. Through these explorations we will also have a better understanding of which types of markets can benefit from the use of nanodata in predictions and which types of markets are less predictable. Perhaps some markets require higher data quality or are more prone to manipulation, such as the stock market. Markets might vary in the horizon of predictability, depending on the lag between the digital trace presaging the transaction and the transaction itself. There might be some predictions that will always be difficult to do regardless of how fine-grained

15. The antecedents of economic activity have always existed in the form of the daily conversations and wanderings of consumers over the economic landscape. What has changed is the cost of unobtrusive observation of these antecedents.

data have become. However, as more nanodata and methods become more widely used, we can only conclude that the future of prediction is far brighter than it was only a few years ago.

References

Appleton-Young, L. 2008. "State of the California Housing Market 2008–2009." Technical Report, California Association of Realtors.

Arrow, K. 1987. "Chapter Reflections on Essays" In *Arrow and the Foundations of the Theory of Economic Policy*, edited by G. R. Feiwel, pages 727–34. New York: New York University Press.

Brynjolfsson, E., T. Geva, and S. Reichman. 2014. "Crowd-Squared: A New Method for Improving Predictions by Crowd-Sourcing Google Trends Keyword Selection." Working Paper, Center for Digital Business, Massachusetts Institute of Technology.

Brynjolfsson, E., Y. J. Hu, and M. S. Rahman. 2013. "Competing in the Age of Omnichannel Retailing." *MIT Sloan Management Review* 54 (4). http://sloanreview.mit.edu/article/competing-in-the-age-of-omnichannel-retailing/.

Case, K. E., and R. J. Shiller. 1987. "Prices of Single-Family Real Estate Prices." *New England Economic Review* 1:45–56.

———. 1989. "The Efficiency of the Market for Single-Family Homes." *American Economic Review* 79 (1): 125–37.

Choi, H., and H. R. Varian. 2009. "Predicting the Present with Google Trends." Google Research Blog. http://static.googleusercontent.com/media/www.google.com/en/us/googleblogs/pdfs/google_predicting_the_present.pdf.

comScore. 2012. "qSearch: A Comprehensive View of the Search Landscape." Technical Report, comScore. https://www.comscore.com/Products/Audience-Analytics/qSearch.

Davenport, T. H. 2006. "Competing on Analytics." *Harvard Business Review* 84 (1): 98.

Ginsberg, J., M. H. Mohebbi, R. S. Patel, L. Brammer, M. S. Smolinski, and L. Brilliant. 2009. "Detecting Influenza Epidemics Using Search Engine Query Data." *Nature* 457 (7232): 1012–14.

Glaeser, E. L. 2008. "Housing Prices in the Three Americas." *New York Times*, September 30.

Glaeser, E. L., and J. Gyourko. 2006. "Housing Dynamics." NBER Working Paper no. 12787, Cambridge, MA.

Han, L. 2010. "The Effects of Price Risk on Housing Demand: Empirical Evidence from US Markets." *Review of Financial Studies* 2 3(11): 3889–928.

Horrigan, J. B. 2008. "The Internet and Consumer Choice: Online Americans Use Different Search and Purchase Strategies for Different Goods." Technical Report, Pew Internet and American Life Project.

Krugman, P. 2009. "How Did Economists Get It So Wrong?" *New York Times Magazine*, September 2, MM36.

Kuruzovich, J., S. Viswanathan, R. Agarwal, S. Gosain, and S. Weitzman. 2008. "Marketspace or Marketplace? Online Information Search and Channel Outcomes in Auto Retailing." *Information Systems Research* 19 (2): 182–201.

Lazer, D., A. Pentland, L. Adamic, S. Aral, A.-L. Barábasi, D. Brewer, N. Christa-

kis, N. Contractor, J. Fowler, M. Gutmann, T. Jebara, G. King, M. Macy, D. Roy, and M. Van Alstyne. 2009. "Computational Social Science." *Science* 323 (5915): 721–23.

McAfee, A., and E. Brynjolfsson. 2012. "Big Data: The Management Revolution." *Harvard Business Review* 90 (10): 60–66.

Moe, W. W., and P. S. Fader. 2004. "Dynamic Conversion Behavior at E-Commerce Sites." *Management Science* 50 (3): 326–35.

National Association of Realtors (NAR) Research Staff. 2012. "Profile of Home Buyers and Sellers 2012." Technical Report, National Association of Realtors.

Pentland, A. S. 2010. *Honest Signals*. Cambridge, MA: MIT Press.

Simon, H. A. 1984. "On the Behavioral and Rational Foundations of Economic Dynamics." *Journal of Economic Behavior and Organization* 5 (1): 35–55.

Wu, L., B. Waber, S. Aral, E. Brynjolfsson, and A. S. Pentland. 2008. "Mining Face-to-face Interaction Networks using Sociometric Badges: Predicting Productivity in an IT Configuration Task." International Conference on Information Systems 2008 Proceedings, 127.