Investigating the Impact of Search Query Data on Forecasting Housing Prices and Future Fluctuations

William Corkery, B.A. Economics, Haverford College
April 27, 2017

Abstract

This study follows and adds to existing literature using search query data as a predictor of housing prices and future fluctuations. Based on the premise that homebuyers reveal their intention to purchase or sell their house on the internet, this study aggregates positive and negative real estate related queries into positive and negative search query indices. Using these indices in housing price models serves as a sentiment-like indicator for the U.S. housing market. The study employs both fixed effects and ordinary least squared (OLS) regression analysis to determine the ability of the models to use search query data to predict housing prices and future fluctuations. The results reveal that the inclusion of search query data mitigates error and improves forecasting of housing prices. The study finds that the addition of search query data improves models at the national and MSA-level, especially after accounting for spatial heterogeneity via fixed effects. Notably, this study's main contribution to the literature is the positive sentiment index (PSI) that appears to be helpful in predicting future housing prices, especially at the national level.

Acknowledgements

First, I would like to thank my advisor, Tim Lambie-Hanson, for all of his help throughout this process – it was truly a team effort. I would also like to thank Carola Binder for all of the help that she provided me with as I embarked upon this journey. Next I would like to thank Grayton Downing for being there every step of the way, I could not have done it without you. I certainly need to thank The Council for these indescribable four years – it's pretty crazy to say but we made it. The world has yet to see the best of us. Lastly, and most importantly, I want to thank my parents for the sacrifices and opportunities they have presented me with. I cannot possibly put into words how much it means to me but thank you, for everything.

Table of Contents

1. Introduction

The U.S. housing market is extremely important to the broader economy and thus the health of the housing market is a topic of much discussion and research. As was highlighted by the housing crisis in the mid-to-late 2000s, few forecasts adequately predicted the housing downturn that was to come. Notably, forecasts failed to capture the souring of consumer sentiment amongst homeowners and potential homebuyers. Economists have placed great emphasis, in recent years, on using the American public's increasing reliance on the internet and social media to help with economic predictions. A growing mass of literature focuses on real estate economics. My study investigates whether incorporating search query data into models improves prediction of housing prices and housing price fluctuations. Specifically, I construct basic housing price models from macroeconomic and housing data to predict future housing prices, while integrating search query data into these models.

Incorporating this search query data to serve as a sentiment-like variable may improve forecasting of housing prices, because it could serve as a leading indicator for an increase or decrease in fear in the housing market. This study builds off existing literature that studies search query data signaling negative, fearful attitudes and fills a gap in the literature by using positive search queries that show positive sentiment in the housing market. Predicting large upticks or downturns in housing prices has been a point of focus for researchers since the Sub-Prime Mortgage Crisis in the United States (Bennöhr and Oestmann (2014)). This study aims to find a more comprehensive and predictive model of future housing prices by using both positive and negative search query data.

The paper is organized as follows. We begin in Section 2 with a comprehensive review of the literature pertaining to housing price models, highlighting those that incorporate search query data into models. This section also delves into the extensions and methodology for my study based on the previous works of other economists. Section 3 provides a description of the data used to create the

housing price models and where that data can be obtained. Sub-section 3.3 provides an in-depth description of Google Trends data and how I create the positive and negative search query indexes. Building off that, Section 4 describes the methodology of my particular study and the types of regressions I will be using for the baseline of my analysis. Section 5 provides discussion and analysis of this study's fixed effects regressions, in and out-of-sample forecasting performance, and individual MSA analysis. I close with concluding remarks, limitations to the study and a path for future research in Section 6.

2. Literature Review

Given the importance of housing prices and housing sales to the overall economy there has always been interest from economists in researching the housing market. This is especially true since the 2007-2008 Financial Crisis, which was, in large part, the result of a massive housing downturn in the United States. There are a number of approaches that economists and researchers alike have experimented with to find an adequate model to predict future housing prices but none have yielded satisfactory results.

Case & Shiller (1990) famously explore forecasting prices and excess returns in the housing market using a weighted repeat sales (WRS) method. They find a momentum-like phenomenon in real estate prices. If home prices go up in one year, they will likely go up again the next year but only by one-third as much (Case and Shiller (1990)). Their model shows a positive serial correlation of home price changes in the short term and a negative serial correlation in the longer term (Case and Shiller (1990)). Case and Shiller also find that construction costs, income growth, and adult population increases are all positively linked to home price fluctuations in the next year (Case and Shiller (1990)). They find incorporating these variables into a model improves prediction of housing prices. Similarly,

Case & Shiller (1988) find that there are distinct differences in people's expectations in booming housing markets and post-boom markets. Through the use of a survey, they observe that homebuyers in booming cities have higher future housing price expectations than those in post-boom or non-boom markets (Case & Shiller (1988)). In addition to this, they find that expectations are based on consumer sentiment and that some price changes are grounded in society and not rationality (Case & Shiller (1988)). The consumer sentiment aspect of forecasting real estate prices is extremely difficult to quantify, which is a major hurdle that we must overcome in order to forecast housing prices with some measure of accuracy. Following Chauvet et al. (2013), I use Google Trends data to gauge consumer sentiment of homebuyers and homeowners.

Since 2004, Google has made its search query data available for public use and it has quickly become an area of research and investigation for academics. One of the major advantages of Google Trends data is that it is updated daily. This addresses the issue of using data that comes out on a one or two-month lag, like the Case-Shiller Housing Price Index where, for example, the January housing price report doesn't come out until March. Moreover, Google is, often times, the first place people will go to acquire information and some contend it as a transparent process that sends a more honest signal of the intention to purchase a house (Hohenstatt and Kaesbauer (2014), 253-254). For these reasons, Google Trends data is a potentially powerful tool in the attempt to quantify the consumer sentiment aspect of the home buying process.

The most prominent study that incorporates search query data is similar in methodology to this study but not in subject matter. Ginsberg et al. (2009) uses search query data to predict and combat Flu epidemics before they breakout. They use search query data from Google Trends along with data on influenza-like illness (ILI) physician visit data. The Center for Disease Control (CDC) maintains readily accessible data on ILI visits. Ginsberg et al. (2009) use a log-log model to predict ILI visits

from search indices on the national and state level to investigate if dynamics change on a state by state basis. Their study finds that using up-to-date search query indices provides one of the most accurate and timely predictions for ILI visits available. The use of Google Trends improves predictions and eliminates the difficulties of using lagged data. The Ginsberg et al. (2009) study provides a framework for future research on how to use Google Trends in a study.

Recently, a number of real estate economics papers have utilized Google Trends data. Wu and Brynjolfsson (2013) employ a simple empirical framework by investigating whether the incorporation of search query data results in smaller mean absolute errors (MAE) than a baseline model (Wu and Brynjolfsson (2013)). They do find that the model including search query data beat predictions made by experts from the National Association of Realtors by 23.6% for future housing sales in the U.S. (Wu and Brynjolfsson (2013), 26). Additionally, Beracha and Wintoki (2012) research how much future real estate prices can be predicted by search engine intensity. Their findings suggest that "abnormal" search intensity for real estate in a particular city can help predict the city's future abnormal housing price change (Beracha and Wintoki (2012)). Also, the results show that home prices are more sensitive to an increase in search intensity rather than a fall in intensity, which may suggest a "stickiness" – or resistance to change - in housing prices (Beracha and Wintoki (2012), 21).

Furthermore, Webb (2009) tackles foreclosure volume in the real estate market using search query data and finds that the query "foreclosure" is highly correlated with actual foreclosure volumes and can serve as a leading indicator of future foreclosures in the housing market. This could prove beneficial for policy makers who closely watch foreclosure figures to gauge the health of the economy.

As it relates to my study, I include the unique search query "foreclosure" for all my local HDI's to work around Google's privacy filter issues.[1]

Various economists have researched Google Trends' relationship with housing markets in European countries. Hohenstatt and Kaesbauer (2014) examine the U.K. housing market using a vector auto regression (VAR) model. They corroborate findings from U.S. based studies that the Google Trends subcategory "real estate agency" has predictive power on real estate transaction volume. Previous studies, like Hohenstatt, Kaesbauer and Schaefers (2011), find empirically confounding results about the subcategory "home financing." However, Hohenstatt and Kaesbauer (2014) find that the subcategory can serve as a potential leading indicator of stress in the housing market, after controlling for mortgage approvals (Hohenstatt and Kaesbauer (2014), 271). Importantly, they find that transaction volumes are twice as sensitive to search query fluctuations compared to housing prices. Bennöhr and Oestmann (2014) expand upon this analysis by investigating 14 EU-countries through fixed effect regressions to explore the sentiment-based factors effecting real estate prices (Bennöhr and Oestmann (2014)). In line with prior research, they find that incorporating Google Trends data has a substantial impact on predicting housing prices, even across their multinational framework. Though my study focuses exclusively on the U.S. real estate market, the findings from these European studies proved valuable in creating my methodology and comparing my results to previous literature.

Almost all of the research using Google Trends data finds that it adds explanatory power to models. Marian Alexander Dietzel (2016) uses Google Trends data to serve as a sentiment indicator to predict turning points in the housing market. The author believes it is the first study to use search query data in a binary model that is designed to predict turning points in the housing market. Notably, the study finds that the model predicts the "direction of monthly price changes correctly, with over 89%

[1] Refer to Section 3.3 for more information on Google's privacy filters and the reasoning behind including "foreclosure" in the local HDI's

in-sample and just above 88% in one to four-month out-of-sample forecasts" (Marian Alexander Dietzel (2016), 108). In my study, I consider both in-sample and out-of-sample forecasting. However, instead of predicting the direction of the price change, I examine the percent error reduced with the inclusion of search query data.

The paper most closely aligned with my methodological framework is Chauvet, Gabriel and Lutz (2013), who create a housing distress index (HDI) that gauges the sentiment in the U.S. real estate market. In a way, it is similar to the Chicago Board Options Exchange (CBOE) Volatility Index (VIX) for equity markets because the HDI can serve as a barometer for homebuyers and homeowners levels of fear in a certain region and time period (Chauvet et al. (2013), 1). Their HDI is comprised of search query data that combines a distress signal, like the word "help", and a real-estate related search term, like "foreclosure" or "home financing" (Chauvet et al. (2013)). They control for potential seasonality concerns by seasonally adjusting the index via the x12 algorithm from the U.S. Census Bureau. Chauvet et al. (2013) standardize their HDI in order to give it zero mean and unit variance. I follow this methodological approach of standardization.

Chauvet et al. (2013) run OLS and fixed effects regressions to test the benefit of the search query data. As expected, they find that when the HDI increases, home prices fall. In their study, the HDI appears to be a good tracker of "anecdotal accounts of fear over the sample period as it remains low throughout the housing expansion but then spikes at the height of the crisis" (Chauvet et al. (2013), 26). Thus, the HDI is more predictive during crisis periods, providing more accurate predictions of future prices in times of heightened distress.

In addition to the entire U.S., my study considers the geographical implications of using search query data and evaluates whether there is a particular metropolitan statistical area (MSA) which could serve as a leading indicator for rising housing distress and forthcoming housing price fluctuations.

Google Trends data is available at the city and MSA level. This allows me to investigate whether major cities with typically volatile real estate prices are leading indicators. Chauvet et al. (2013) create local HDIs for all of the 20 MSA's that make up the Case-Shiller 20-City Composite Home Price Index (Chauvet et al. (2013)). In order to create the local HDIs, I need to compensate for the smaller search query volume for a particular MSA. This can negatively impact my data since a query too few searches will trigger Google's privacy filter. Chauvet et al. (2013) work around this issue by including the query "foreclosure" into their local HDI's because it is searched more frequently than the more specific distress queries. I follow their methodological lead in this case and create the local HDIs in the same fashion.

Gabriel and Lutz (2014) expand upon the framework of Chauvet et al. (2013) in their paper investigating the impact of atypical monetary policy, such as quantitative easing, on real estate markets. The major extension this paper adds to the literature is their model tests for the impact of quantitative easing shocks on housing distress (Gabriel and Lutz (2014), 13). An unconventional monetary shock the results in a 25 basis point reduction in the yield of the 10-year Treasury bond will reduce the growth rate of search queries signaling housing distress by 30%, which is nearly 3 standard deviations (Gabriel and Lutz (2014), 22). This result shows that reducing the interest rate calms fears in the real estate market. As yields on the 10-year Treasury fall so too does the HDI, which drives the housing price index (HPI) higher. My study uses the spread between the 10-year Treasury and the Federal Funds rate to investigate this dynamic.

The HDI addresses fear and distress in the U.S. real estate market, while neglecting positive sentiment. I aim to address this aspect of the HDI and fill a gap in the literature by creating a positive sentiment index (PSI). Hohenstatt and Kaesbauer (2014) find that the Google Trends' subcategory "home financing" can serve as an indicator of real estate market stress. They conclude this after

filtering out for the number of mortgage approvals. While this search term could be a negative indicator, I do think it is probable that people will search "home financing" if they are interested in purchasing a home or hoping to gather more information about the home buying and financing process. I believe that this query can serve as a positive search term in my PSI. I also include the Google Trends subcategory "real estate agency" in the PSI because it is also widely used in other studies. Hohenstatt and Kaesbauer (2014) find that using "real estate agency" as a search term has a certain measure of predictive power when forecasting transaction volumes for houses. I believe this can serve as a positive indicator because if there is an uptick in people searching for real estate agencies, then there could certainly be an increase in houses purchased. In addition to these two subcategories, I use a number of other unique queries in my PSI.[2]

3. Data

3.1 Macroeconomic and Housing Data

In this study, the two main sources of data are the Federal Reserve Economic Data (FRED) website maintained by the St. Louis Federal Reserve and Google Inc. The macroeconomic data comes from FRED but the data sets originate from various other sources. One of my independent variables used in this study is the unemployment rate at both the national and MSA level. The unemployment rate in this study is the U-3 measure of underutilization. It is defined as the percentage of people without jobs who have actively looked for a job in the last four weeks divided by the total number of people in the U.S. labor force. This data is reported as a seasonally adjusted percentage on a monthly basis. I gathered the

[2] See Tables 1 and 2 in the Appendix for a complete list of queries used in the HDI and PSI

national unemployment rate as well as the unemployment rate for all 20 MSA's that are in the Case-Shiller 20-City Composite Housing Index.[3]

In line with Chauvet et al. (2013), the interest rate variable that I use in this study is the spread between the U.S. 10-year Treasury bond and the Federal Funds rate. The data from FRED ultimately combines two data sets from the Board of Governors of the Federal Reserve. The first data set is the 10-year Constant Maturity Rate and the second is the Effective Federal Funds rate. Like the unemployment data, this variable is available at a monthly frequency but it is not seasonally adjusted. I use it in my study to examine the impact of interest rates on housing prices.

Another variable that I use in my national OLS model is the delinquency rate of single-family residential mortgages that are booked in domestic offices. Unlike the previous data sets, this data is released quarterly in a seasonally adjusted, percent format. The delinquency rate is the percentage of loans that have missed or been late with a mortgage payment. The delinquency rate is thought to show and potentially foreshadow housing distress. Conversely, housing starts data can show positivity and prosperity in the real estate market. This monthly variable is measured in thousands of houses that are started in a given month. Housing starts data comes from the U.S. Bureau of the Census via FRED. Note that this data is not seasonally adjusted. I expect that an increase in the number of houses is driven by increased demand for houses and therefore prices will increase, although there is a negative relationship at play where greater supply of housing reduces real estate prices, so that may negate the price increase from heightened demand.

Retail sales can also provide a gauge for consumer spending. The U.S. Bureau of the Census provides monthly, seasonally adjusted retail sales figures, in millions of dollars. Retail sales excludes food service sales. Note that e-commerce sales are included in these estimates, which is important

[3] Refer to page 13 for the full list of the 20 cities included in the Case-Shiller 20-City Composite Housing Index.

because that is becoming a bigger portion of sales as the internet grows and becomes ever more popular. I believe there will be a positive relationship between retail sales and housing prices.

3.2 Housing Price Indices

One of my main dependent variables is the Case-Shiller 20-City Composite Home Price Index (HPI) which can be found on the S&P Dow Jones Indices website. The home price index tracks the fluctuations in prices for residential real estate. It uses data from 20 MSAs: the Phoenix metropolitan area, Greater Los Angeles, San Diego County, San Francisco, the Denver-Aurora metropolitan area, the Washington metropolitan area, the South Florida metropolitan area, the Tampa Bay Area, the Atlanta metropolitan area, the Chicago metropolitan area, Greater Boston, Metro Detroit, Minneapolis-Saint Paul, the Charlotte metropolitan area, the Las Vegas metropolitan area, the New York metropolitan area, Greater Cleveland, the Portland metropolitan area, Dallas–Fort Worth Metroplex, and the Seattle metropolitan area. The index is calculated monthly and is seasonally adjusted. January 2000 serves as the benchmark of the index and therefore the index level in January 2000 is equal to 100.[4] As previously mentioned, my study also uses the individual HPI's for all 20 of the MSA's above.

The Case-Shiller 20-City Composite Home Price Index has a mean of 168.48 from January 2004 – December 2016, compared to its baseline of 100 in January 2000. The index grows steadily to over 200 from January 2000 to March 2007 and then tumbles back down towards 140 in late 2008 and early 2009. The index has been growing since it bottomed out in March 2009, but at a slower pace. The other macroeconomic variables share similar trends of increasing until 2007 and falling during the 2007-2008 Financial Crisis. These variables also bounce back but increase at slower rates in this post-

[4] All the information necessary to fully understand how the index is constructed can be found at http://us.spindices.com/index-family/real-estate/sp-corelogic-case-shiller. There is a factsheet and methodology document available that provide great detail about the index and the construction of it.

crisis time period. Unemployment is a counter cyclical variable that shot up during the 2007-2008 Financial Crisis and has been falling ever since, and is now back down below 5%. The delinquency rate is very similar in this regard because delinquencies skyrocketed during the crisis and have been steadily falling since peaking in the first quarter of 2010.

The other dependent variable in this study is the Federal Housing Finance Agency's (FHFA) Housing Price Index (HPI). Like with the Case-Shiller HPI, I use both national and MSA-level HPI's to test the impact of search query data on the entire country and at the regional level. The base period for the FHFA national purchase-only index that appears with a monthly periodicity is January 1991. At the MSA-level the base period for the MSA housing price indices is the first quarter of 1995. There are positives and negatives to using the Case-Shiller and the FHFA HPIs. Both indices employ the same fundamental repeat-valuation approach to create their housing price indices. The Case-Shiller index compiles information obtained from county assessor and recorder offices while the FHFA HPI only accounts for "conforming" mortgages that are provided by Fannie Mae and Freddie Mac, under the jurisdiction of the FHFA (Housing Price Index Frequently Asked Questions). The Case-Shiller index is value-weighted, meaning that price trends for more expensive homes have a greater influence on estimated price changes than other homes. The FHFA index weighs price trends equally for all properties. Notably, the Case-Shiller index uses a three month moving average for their housing price index, so there is a substantial amount of autocorrelation present in those models. Lastly, the FHFA HPI data comes out on a monthly basis at the national level but is only released quarterly for the individual MSA level. The Case-Shiller HPI comes out on a monthly frequency for both the national and MSA levels.

3.3 Google Trends Data and Indices Description

I also use search query data from Google Trends in my analysis. Refer to Table 1 in the Appendix for the list of search queries in the HDI. The HDI used in this study is the same as Chauvet et al.'s (2013) housing distress index. Also, a list of search queries used in the PSI is provided in Table 2.

Google Trends data shows the proportion of Google searches for a specific query – one example would be the query "mortgage help" – over the total number of searches in a certain region over a specified time period. For my data, I use the entire U.S. and the 20 distinct MSA's mentioned in Section 3.2 as my regions. The data spans the time period from January 1, 2004 to December 31, 2016. Google creates a query share based on searches for the keyword over a given time period, then normalizes the data so that the period with the most queries in the time frame will have a query share equal to 100. The query share will fall from there based on the frequency of searches compared with the most searches at any given time period. Therefore, the query share will always be equal to 100 at one point in the specified timeframe because it is normalized for that to be the case. Google Trends data is time series data that can be split into sub-samples based on geographical region. The unit of observation is query share. The data can be gathered as frequently as daily but I chose to gather it monthly to be compatible with the other variables that I have. I seasonally adjust the search query data based on the x12 algorithm, a commonly used algorithm that Chauvet et al. (2013) also use. After seasonally adjusting the data, I sum all of the query shares together to create one housing distress variable. Given that an increase in query share is not easily interpretable and meaningful, I standardize both the HDI and PSI. I standardize both search query indices so they have a mean of 0 and unit variance. This creates an easily interpretable variable for my analysis.

The distressed search query terms show similar patterns to the macroeconomic variables. The search query data tends to begin increasing slightly before the macroeconomic variables do. They

generally increase through the 2007-2008 Financial Crisis and fall rapidly back down, after a brief period of elevated queries. Ideally, the HDI would peak earlier then is seen in Figure 1 in the Appendix and could serve as a leading indicator for housing distress. Since it peaks around March 2009, it is likely too late into the Financial Crisis to be considered a leading indicator. Nonetheless, this relationship is in line with my expectations that HDI will increase when distress is higher.

As mentioned in Section 2, the first two queries used in the PSI are "real estate agency" and "home financing." The first unique query used is "no down payment mortgage." This is because when the housing market is booming and house prices are appreciating, people feel more confident about their finances and are more likely to purchase homes with minimal money down. They may assume that they will be able to refinance as they wish since the value of their house will be continuously increasing. In fact, we see that there is a large increase in search volume in 2004 and 2005 when people were buying homes with little to no money down. For this query, there is a steep drop off in query share in 2007 and 2008.

I include "mortgage adjustable rate," "adjustable rate mortgage," and "option ARM" as queries in the PSI because I hypothesize that people search these terms during times of price appreciation, when they would prefer to purchase an adjustable rate mortgage over a fixed rate mortgage because of the typically lower monthly payments. Adjustable rate mortgages are considered to be riskier because the rate is usually low for the first couple of years and then can spike after that, so it is more typically used when housing prices are appreciating.

Similar to the query "no down payment mortgage," the query "no income mortgage" is included in the PSI because it could capture people who wish to take advantage of the lax lending policies of the time period and purchase a house even if they have no steady income stream. A number of the queries in the PSI are related to the financing side of the home buying process, which may

attract queries from people who are further along in the home buying process than people simply beginning their housing search. However, there are a number of general home buying queries in the PSI to account for this portion of real estate related queries. It is interesting to note that most of these queries peak in April, June or July of 2004 which could signal that they are good indicators of positive sentiment in the housing market as many Americans purchased a house in this time frame. As with the HDI, I seasonally adjusted the PSI query data, summed the query shares and then standardized the index to have zero mean and unit variance.

Accordingly, I create two variables from all of this search query data: the housing distress index (HDI) and the positive sentiment index (PSI), both at the national level. I employ the same approach for the local HDI's and PSI's. I create a unique HDI and PSI for all 20 of the MSA's included in the Case-Shiller 20-City Composite Housing Price Index. Note that Google has a privacy filter restricting access to data if there are not sufficient searches for a particular query in a given geographic region at a specific time. To address this issue, I added the query term "foreclosure" to all of my local HDI variables. The "foreclosure" query is the lone query term used to create my local HDI's that appears in all twenty of the local HDI's. The terms that are included in the local HDI's vary by MSA. This method of incorporating the "foreclosure" query into local HDI's follows Chauvet et al.'s (2013) methodological approach to local queries. Google's privacy filter is not triggered as frequently for the positive search terms at the local level so a similar approach to the local PSI's is not necessary.

4. Methodology

As was mentioned in Section 3, both macroeconomic and housing specific data is necessary to model and explain fluctuations in housing prices. These two elements capture how the economy is functioning and how the real estate market is performing, but they do not tell the whole story. Macroeconomic and housing data is not available at the periodicity of search query data and is made available on a lagged basis. Also, macroeconomic and real estate data does not fully capture the consumer sentiment for potential homebuyers and the fear or optimism present in the market. This is why the use of search query data is hypothesized to improve predictions of housing prices. Current home prices contribute greatly to future housing prices and therefore, I incorporate a number of lags of the dependent housing price variables into the models. Hohenstatt and Kaesbauer (2014) utilize a lagged relationship up to 13 months to investigate the dependency of future home prices on previous months housing prices. I base the number of lags included in my models on AIC. The basic model for all of the analysis is:

$$r_{it} = \boldsymbol{\beta}_0 + \sum^{k}_{j=1} \boldsymbol{\beta}_j HDI_{i,t-k} + \sum^{n}_{m=1} \boldsymbol{\beta}_m HDI_{i,t-n}{}^{crisis} + \sum^{r}_{p=1} \boldsymbol{\beta}_p PSI_{i,t-r} + \sum^{s}_{q=1} \boldsymbol{\beta}_q PSI_{i,t-s}{}^{crisis} + \boldsymbol{\beta}_5 Controls_{it} + \alpha_i + \boldsymbol{\varepsilon}_{it} \quad (1)$$

The dependent variable, r_{it}, is either the return on the Case-Shiller HPI or the return on the FHFA HPI. As with Chauvet et al. (2013) the housing return dependent variable is the first difference between the current period and the previous period. This allows for interpretation of the impact of the HDI and PSI on housing prices. Both the HDI and PSI variables are lagged one period. The amount of time represented by one period varies depending on whether the data is monthly or collapsed into quarterly data. For the Case-Shiller regressions, I examine the data from both monthly and quarterly perspectives. For the FHFA models, I investigate a monthly model at the national level and employ quarterly regressions at the MSA-level because MSA-level data is only available quarterly. The crisis interaction terms take on the value of $\sum^{k}_{j=1} \boldsymbol{\beta}_j HDI_{i,t-k}$ and $\sum^{r}_{p=1} \boldsymbol{\beta}_p PSI_{i,t-r}$ for the period between

January 2007 to March 2009, which is in line with Chauvet et al. (2013), and is equal to 0 in all other times. The controls vary depending on the type of model. For the fixed effects regressions, the controls include a number of lags of the housing price index, MSA fixed effects, and MSA unemployment rates. The MSA fixed effects are denoted by the α_i in (1). The standard errors used in the fixed effects models are Driscoll-Kraay standard errors that account for heteroscedasticity and are robust to spatial heterogeneity as well as general temporal dependence. The lags on r_{it} are determined by AIC and serve to mitigate the autocorrelation that is present in housing price indices. These fixed effects allow for increased power in the model and account for unobserved differences in housing markets across the United States.

My study also uses basic OLS regression on national level data to assess the impact of both positive and negative search query data on housing prices. The model generally follows (1) except I utilize a number of different controls in this model and there are no fixed effects. Again, the dependent variable, r_{it}, is either the return on the Case-Shiller HPI or the return on the FHFA HPI for time *t* and geographic region *i*. In this case the geographic region *i* is the entire U.S. Both the HDI and PSI variables and their crisis interaction terms include two months of lags. These regressions are run at the monthly frequency. The crisis interaction terms take on the value of $\sum^{k}_{j=1}\boldsymbol{\beta}_j \, HDI_{i,t-k}$ and $\sum^{r}_{p=1}\boldsymbol{\beta}_p \, PSI_{i,t-r}$ for the period between January 2007 to March 2009, and is equal to 0 in all other times. For the OLS regressions, the controls include 10 or 13 lags of the dependent variable, housing starts, the interest rate variable of the difference between the 10-year Treasury rate and the Federal Funds rate, the unemployment rate for the U.S., mortgage delinquency rate and retail sales. All control variables are lagged one month. The number of lags of the housing price index are determined based on AIC. The Case-Shiller regressions utilize 10 months of lags to control for autocorrelation, while the FHFA model uses 13 months' lag, following the methods of Hohenstatt and Kaesbauer (2014). The

delinquency rate variable is originally a variable with a quarterly frequency but I convert it to a monthly variable. To convert the variable to a monthly periodicity, I use an averaging process where the Q1 and Q2 delinquency rates are averaged to obtain delinquency rates for the second and third months of the year (February and March respectively).[5] I apply this technique for all four quarters of each year from January 2004 to December 2016 to obtain a delinquency rate for each month during that time period.

In addition to the fixed effects and OLS regressions on the national level, this study investigates the impact of local HDI's and PSI's on housing price indices at the MSA-level. In order to get a holistic view of the impact of search query data on predicting housing price fluctuations, it is helpful to observe individual MSA's. As was mentioned in Section 3, I create a local HDI and PSI for all 20 MSA's included in the Case-Shiller 20-City Composite Housing Index. The OLS models run on the individual MSA's follow the model in (1). As is the case for the other regressions, r_{it} is the housing return, either for the Case-Shiller MSA HPI or the FHFA MSA HPI, for MSA *i* at time *t*. The crisis interaction terms follow the same methodology as the previous models. The controls include a number of lags of the housing price index and the particular MSA's unemployment rate data. The dependent variable is lagged four periods in this regression. The lags on r_{it} are determined by AIC. The four lags amount to one full year. For the MSA analysis, the study analyzes the impact of search query data on all 20 of the MSA's included in the Case-Shiller 20-City Composite Home Price Index.

Last, I forecast future housing prices to determine whether search query data improves both in

[5] An example of how the monthly variable is created is: To obtain the monthly delinquency rate for the second month (February) of the year, I take 2/3 of the delinquency rate from Q1, which is released in January, and 1/3 of the delinquency rate from Q2, which is released in the fourth month (April) of the year. I then sum them together to get the February value. Similarly, to get the delinquency rate for the third month of the year (March) I take 1/3 of the delinquency rate from Q1, which is released in January, and 2/3 of the delinquency rate from Q2, which is released in the fourth month (April) of the year. I repeat this process until every month has an averaged delinquency rate attached to it.

and out-of-sample fit. Forecasting is used for both the MSA regressions as well as the fixed effect regressions. For the fixed effect regressions, the model makes a prediction for every quarter from Q1 2004 to Q4 2016 for each of the 20 MSA's in the study. In all, this totals 940 observations. The two primary metrics in this analysis are the percent reduction in mean squared error (MSE) and the overall MSE for the forecasts.

To test for the impact of including search query data into a model, this study runs regressions with and without the HDI and PSI. The regressions used in each forecast are identical, aside from the inclusion or exclusion of the HDI and PSI variables. In forecasting it is common practice to restrict one's sample period to create a "pseudo" sub-sample to simulate out-of-sample forecasting analysis. This is because in-sample forecasting can often times be overly optimistic about the ability of a model to predict future housing prices. Thus the out-of-sample forecast on the sub-sample of the data is beneficial. A true out-of-sample analysis would take into account all data up until today, construct a forecast of next month's value and wait for next month's housing price value to come out. Then, repeat this process for a number of months. When this is impractical due to time constraints, a sub-sample is withheld from the overall data set to test for out-of-sample error. The sample of data that I use to construct my forecasts is from January 2004 to December 2013, because withholding ~20% of data is standard in forecasting analysis. Therefore, the out-of-sample forecasts include 3 years' worth of data (2014 - 2016), totaling 12 quarters of housing price values.

The first step in forecasting analysis is to forecast housing prices based on regressions without search query data, using the constricted sample through 2013. Next, forecast housing prices using those same regressions except include the HDI, PSI and their crisis interaction terms. I then calculate the difference between the predicted housing price index value for the two regressions and the actual value for the housing price index. To get the squared errors, I square the difference between the

predicted and the actual housing price index value and sum all of the squared errors. This allows one to get the mean squared error (MSE) for the regression. Next, given the MSE for both forecasting regressions, I calculate the percent change in MSE that comes with the inclusion of the search query data into the forecasting regressions. This provides insight into whether or not the models are better at forecasting in-sample or out-of-sample error. I use both metrics in this analysis.

The amount of MSE present in regressions can be useful in providing supplementary information to the percent reduction in MSE. Less error in forecasting signifies predictions that are closer to the actual housing price index values. My study examines absolute MSE at the MSA-level. This sheds light on the discrepancies between forecasting for more volatile real estate markets compared to more stable real estate markets. It may show whether or not search query data has more impact on forecasts in the more speculative real estate markets in the U.S.

5. Results

Figure 1 depicts the graphical relationship between the constructed search query indices and the Case-Shiller 20-City National Housing Price Index. It is immediately apparent, when looking at the plot, that the PSI is positively correlated with the Case-Shiller HPI, while the HDI is negatively correlated. The HDI appears to bottom out when the housing price index reaches its peak, in late 2006. Similarly, the HDI begins steadily rising as the Case-Shiller HPI starts to plummet. The HDI peaks in early 2009, which is when the housing market hit its trough. It is also evident that in recent years, as the housing market in the U.S. has rebounded, the HDI has remained depressed and even fallen as the Case-Shiller 20-City National Housing Price Index has been progressively rising back up to near-2006 levels.

Conversely, when investigating the relationship between the positive sentiment index (PSI) and the Case-Shiller 20-City National Housing Price Index the correlation appears weaker. It's interesting to note that the PSI peaks in November 2004 while the Case-Shiller HPI doesn't peak until April 2006. The PSI begins falling more than 18 months before the HPI, suggesting it could serve as a leading indicator for housing price fluctuations. The PSI does not bounce back as much as the Case-Shiller HPI in the aftermath of the 2007-2008 Financial Crisis which may hurt the predictability of the index.

Table 3 serves to strengthen the assessments made from Figure 1. Table 3 displays correlation coefficients between the two search query indices and the two housing price indices used in this study. As noted, the HDI has a negative correlation to the housing price indices and the PSI has a positive correlation to the housing price indices. The table includes correlations for the full sample period, January 2004 – December 2016, and for the crisis period, defined as January 2007 – March 2009. For the housing distress index (HDI), the magnitude of the correlation between the HDI and the housing price indices increases during the crisis period. For example, the correlation coefficient between the HDI and the FHFA Housing Price Index is -0.448 for the full sample and is -0.922 during the crisis period. The same phenomenon is true for the PSI and the housing price indices.[6] The correlation also increases in magnitude when the search query indices are lagged by one month. Both the full sample and crisis period correlation coefficients increase in magnitude when lagged one month, compared to the non-lagged correlations. This indicates that not only are housing indices and search query indices linked but they are more closely linked in times of crisis and when a lag is employed to the search query indices.

[6] The correlation between the positive sentiment index (PSI) and the Case-Shiller Composite National Housing Price Returns is the only pairing that does not see an increase in magnitude during the crisis period. Refer to Table 3 for the full list of the correlation coefficients

Given that the correlations substantiate what we expect to find regarding the relationship between search queries and housing price indices, we now shift to the study's regression analysis. Table 4 shows the results of the fixed effects regressions that I perform on both the Case-Shiller Housing Price Index and the FHFA Housing Price Index. The regressions take on the form of model (1) from the Section 4. The model employs MSA fixed effects to compensate for unobserved differences between cities. Model (1) in Table 4 shows the results from the fixed effects regression for the Case-Shiller monthly housing price data. The FHFA does not make its MSA Housing Price Index data available on a monthly basis but Case-Shiller does. Model (1) shows that given a one standard deviation increase in the housing distress index (HDI) in the previous month, there will be, on average, a 0.064% decrease in the Case-Shiller housing price index in the subsequent month. This is significant at the 1% level and economically meaningful because the mean of the first difference in the Case-Shiller HPI is 0.250% and the minimum is -9.86%.

Much like the correlation coefficients, the model predicts a larger magnitude effect during crisis times. In this model, a one standard deviation increase in the previous month's HDI during the crisis predicts an average fall of 0.126% in the Case-Shiller HPI in the successive month. The impact is almost twice as large during crisis periods compared to non-crisis times, and is statistically significant at the 1% level. Moreover, when looking at model (2) in Table 4, we see that a one standard deviation increase in non-crisis times leads to a 0.233%[7] decrease in the Case-Shiller Housing Price Index, whereas a one standard deviation increase in the HDI during a crisis predicts a 0.524% decrease in the Case-Shiller Housing Price Index in the next quarter. This is statistically significant at the 1% level, but the non-crisis prediction is insignificant.

[7] This variable is insignificant but is close to being significant at the 15% level which is the upper bound of Chauvet et al.'s (2013) barometer for significance.

Next, turning to the positive sentiment predictions from models (1) and (2), all predictions of PSI, in crisis and non-crisis times, for both monthly and quarterly periodicities are significant at the 1% level. A one standard deviation increase in the PSI in the previous month is predicted to lead to a 0.093% increase in the Case-Shiller Housing Price Index next month. Moreover, during crisis periods, model (1) predicts a one standard deviation increase in the PSI would lead to a 0.287% increase in the Case-Shiller HPI for the next month. This is an economically meaningful statistic that is ~0.04% higher than the average monthly change for the Case-Shiller HPI over this time period. The quarterly PSI impacts housing prices in a similar statistically significant and economically meaningful manner. Specifically, the model predicts that a one standard deviation increase in the PSI over the previous quarter would lead to, on average, an increase of 0.522% in the Case-Shiller HPI for the current quarter, significant at the 1% level. Like the correlation coefficients in Table 3 and the HDI variables in models (1) and (2) in Table 4, the PSI variable increases in magnitude during the crisis period. In model (2) a one standard deviation increase in the PSI predicts a 1.534% increase in the Case-Shiller Housing Price Index during a crisis, significant at the 1% level.

It may seem counterintuitive to predict that housing price indices would increase during a time of crisis, but given the positive correlation between the PSI and housing price indices it does make sense. It is important to note that the model predicts an increase in housing prices given an increase in positive real estate search queries. During the 2007-2008 Financial Crisis, positive search queries dropped off drastically and thus the model would predict that with a fall in the PSI there would be a corresponding fall in housing price indices, which is the exact phenomenon that occurs.

The last model in Table 4, model (3), is the regression using FHFA quarterly data to assess the impact of HDI and PSI on housing prices. There are some interesting inconsistencies between this regression and models (1) and (2) from Table 4. While model (3) predicts a one standard deviation

increase in the HDI, in the previous quarter, would lead to a statistically significant 0.714% decrease in the FHFA HPI in the current quarter, the crisis interaction term is positive. The interaction term is insignificant, but the sum of the crisis HDI interaction term and the HDI variable is significant at the 1% level.[8] One potential reason for the differences in the sign of the interaction terms could be due to the differences in the type of loans that fall under the FHFA jurisdiction and the loans included in the Case-Shiller HPI. Fannie Mae and Freddie Mac are the lenders of the loans that the FHFA regulates and all loans must be "conforming." [9] The Case-Shiller HPI compiles loan data from county assessor and recorder offices across the country and includes all types of loans – e.g. Alt-A and Jumbo mortgages. There may be something inherently different about families who apply for Fannie Mae and Freddie Mac loans compared to those who do not. This potential difference may be contributing to the discrepancy in signs of the two interaction terms.

Additionally, model (3) predicts that a one standard deviation increase in the PSI for the previous quarter will result in a 0.697% increase in the FHFA HPI for the current quarter. This is statistically significant at the 1% level and similar in magnitude to the Case-Shiller quarterly model. Again the divergence lies in the interaction term. While the Case-Shiller crisis interaction term was positive, the FHFA crisis interaction term is negative and when summed with the PSI term there is a slight decrease in the magnitude of the predicted impact of the PSI on the FHFA HPI. Although this is insignificant, it is confounding that the FHFA and Case-Shiller Housing Price Indices' would move in opposite directions in times of increased distress.

---

[8] The statistical significance of the sum of the crisis interaction term with the non-crisis HDI or PSI is determined based on a F-test examining whether or not the sum of the terms is statistically different from 0. The p-value from this test is listed in parentheses under the coefficients for these sums.

[9] The most important factor in a conforming loan is the size of the loan. As of 2017, the limit on a conforming loan was $424,100 for single family homes in the continental U.S., except for rare exceptions. Other factors to qualify for a conforming loan are loan-to-value ratio, debt-to-income ratio, credit score, etc. The most important thing is far and away the loan amount. Further information on conforming loans can be found at: https://www.fhfa.gov/.

Table 5 shows the results from OLS regression's at the national level to test the impact of the HDI and PSI. However, the results are weak and imprecisely estimated with very little statistical significance. The only search query variables that are statistically significant are the coefficients on the HDI and the sum of the HDI and its crisis interaction term on the FHFA model. These results highlight the importance of the panel approach where increased power and MSA fixed effects help the precision and accuracy of estimates.

Next, I assess the impact of local HDI and PSI's on MSA-level housing price indices. As Tables 6 – 9 show,[10] there is significant variation in the impact of search query data on the different MSA's. Generally, the more volatile housing markets exhibit greater fluctuations in housing prices given changes in search query intensity. The five most volatile housing markets predict both larger declines in housing prices with an increase in distressed queries and a larger increase in housing price indices with an increase in positive real estate queries.[11] More volatile real estate markets may be prone to drastic swings in housing prices and this result strengthens that ideology.

There is much variation in the MSA regressions, both in significance of results as well as impact of the crisis interaction terms. All 20 MSA's have one model predicting Case-Shiller MSA-level Housing Price Index fluctuations based on HDI and PSI and one model predicting FHFA MSA-level Housing Price Index changes. Unlike the fixed effects regressions, in crisis times, the models, on average, predict that the $HDI_{t-1,\ MSA}^{Crisis}$ term will be positive. In some instances, this causes the sum of the two HDI terms – the crisis and non-crisis components – to become positive. This contradicts Chauvet et al.'s (2013) findings as well as this study's fixed effects results in Table 4. The fixed effects regressions show that given an increase in distressed queries the magnitude of the effect on housing

[10] Refer to Tables 6 - 9 in the Appendix to see the full regression results at the MSA level

[11] The volatility of a MSA is determined based on the standard deviation of the Case-Shiller MSA-level Housing Price Index. The top five most volatile MSA's are Miami, Phoenix, Las Vegas, Los Angeles and San Francisco. The bottom five are Cleveland, Charlotte, Boston, Atlanta, and Dallas.

prices increases. However, in the MSA regressions the magnitude decreases in 12 out of the 20 MSA's for the Case-Shiller MSA Housing Price Indices. While the magnitude decreases in 13 of the 20 regressions for the FHFA MSA Housing Price Indices. The MSA-level regressions appear to be less predictive than the panel regressions that utilize fixed effects. This further illuminates the benefits of using a panel and also leads me to believe that there is significant improvement in predictions when accounting for heterogeneity.

Similarly, the predictions for the impact of positive real estate search queries on housing prices is variable based on MSA. The fixed effects models reveal that, on average, an increase in the PSI leads to an increase in the magnitude change in the housing price index during a time of crisis. This increased magnitude occurs in the MSA-level Case-Shiller regressions. In 17 out of the 20 regressions at the MSA-level the magnitude of the impact of the PSI variable on housing prices increases during times of heightened distress. The MSA-level FHFA regressions are much more split. Only half of the 20 regressions predict an increase in magnitude in the positive direction during a time of crisis. It may be the case that there are distinct differences between housing markets across the country and therefore the impact of increased search queries has varying effects on the housing price indices for those MSA's. Chauvet et al. (2013) show in their study that the national HDI is a better predictor of future housing price fluctuations and this study's confounding MSA results seem to corroborate that. The fixed effects regressions are more statistically significant and economically meaningful, which may shed light on the interconnectedness of the U.S. real estate market and the more predictive nature of the national HDI and PSI, compared to local search query indices.

The five least volatile housing markets over the sample period were the Atlanta, Boston, Charlotte, Cleveland, and Dallas MSAs.[12] Looking at the FHFA Dallas regression is Table 7, a one

[12] Refer to Tables 6 - 9 in the Appendix to see the full MSA-level regression results.

standard deviation increase in the Dallas-specific HDI predicts a decrease of 1.227% in the local Dallas FHFA HPI, significant at the 5% level. Interestingly, during times of crisis, that same one standard deviation increase predicts a 0.912% decrease in Dallas FHFA HPI. While this is insignificant at the 10% level, it is notable that the crisis interaction term is positive. Since Dallas has a relatively stable housing environment, compared to other regions of the U.S, it may be true that elevated distress does not have as much of an effect on housing prices. Given the inverse nature of the correlation between the HDI and the housing price indices, it could be true that the concern that is reflected in the increase in search queries is not fully reflected by a similar, but inverse, housing price decrease in times of extreme distress.

Next, this study investigates the ability of using search query data as a forecasting tool. I investigate both individual MSA and panel data. I investigate the amount of error that is reduced, by including search query data in models. This method follows the description provided in the Section 4 of this paper. Table 10 in the Appendix shows the amount of forecasting error that is reduced (or increased) with the inclusion of search query data into the models for all 20 MSA's in the Case-Shiller 20-City Composite Housing Price Index. Negative numbers indicate less forecasting error in the regression with the search query data compared to the regression without the HDI and PSI variables. Table 10 displays the percent change in error for both the FHFA and Case-Shiller regressions. The in-sample error for the top five most volatile real estate markets – Los Angeles, Las Vegas, Miami, Phoenix, San Francisco – is reduced by 34.812% for FHFA data while the five least volatile markets – Atlanta, Boston, Charlotte, Cleveland, Dallas – reduce error by 30.677% for FHFA data when the search query data is included in the models. Therefore, the more volatile cities reduce 13.48% more error compared to the least volatile real estate markets when including the HDI and PSI in forecasting

models. It should be noted that reducing error in any capacity is beneficial to obtain more accurate forecasts, but the more volatile real estate markets reduce more error than the more stable markets.

Next, it is evident that forecasting does an inferior job of reducing out-of-sample error. There are a number of very large increases in error – notably the increase in forecasting error for the FHFA Los Angeles, Phoenix and Cleveland models are all over 100%. The average percent change in out-of-sample error increases with the introduction of the search query data into the models. This is true for the FHFA and Case-Shiller regressions. Excluding the three statistical anomalies in the FHFA forecasting models – Cleveland, Los Angeles, Phoenix – the out-of-sample error is reduced by almost 2%. That is not a large reduction but it does reduce forecasting error nonetheless. In applying the same process to the Case-Shiller data – this time excluding the Phoenix, Washington D.C., Seattle, Chicago percent changes in out-of-sample error – we see that the percent error still increases with the addition search query data.[13] Given that the Case-Shiller housing price indices are generated using three month moving averages for each data point, there is significant autocorrelation in these regressions. This makes observing the impact of incorporating search query data into models difficult and could play a role in the increased error for the out-of-sample forecasting.

The respective MSA's with statistically significant HDI or PSI coefficients in the models appear to do a better job of forecasting future housing prices than models of MSA's without statistically significant coefficients. This is not a finding that holds across all MSA's, however it does seem that MSA's with statistically significant HDI and PSI coefficients are more likely to reduce more out-of-sample error than other MSA models. This finding occurs in both the FHFA and Case-Shiller models. The Boston, New York, Las Vegas, and San Francisco MSA Case-Shiller HPI forecasting model's all have statistically significant coefficients for the HDI and PSI and all of these models

[13] The error would increase by 4.11% after excluding the anomalies, compared to an overall increase of 30.15% otherwise.

reduce out-of-sample error. More precise estimates of the impact of the HDI and PSI in the MSA models leads to a reduction in error compared to the less precise, non-statistically significant estimates where we see an increase in forecasting error with the inclusion of search query data.

In addition to investigating error in forecasting for the MSA models, I also perform the same process for the fixed effects regressions. The results are in Table 11 of the Appendix. For both FHFA and Case-Shiller models, in-sample and out-of-sample error is reduced with the introduction of search query data into the models. Notably, including HDI, PSI and both crisis interaction terms ("All Search Queries" model), yielded a 5.30% reduction of in-sample error and a -10.677% change in out-of-sample error. Forecasting housing prices with a model using only the PSI and its crisis interaction term reduces more error compared to the full model for the Case-Shiller HPI.[14] Interestingly, the panel data forecasting models did a better job, by a significant margin, of reducing out-of-sample error when compared with in-sample error. The MSA forecasting models show the opposite phenomenon. This further strengthens findings that applying fixed effects to account for regional differences in housing markets improves models and forecasts.

It is important to observe the overall mean squared errors (MSE) in these regressions. A regression with more error compared to the true values of the respective housing price indices means that the predictions are less closely fitted to the true values for the particular time period. When viewing this, the average amount of MSE present in the FHFA forecasting models of the five most volatile MSA's has almost eight times more error present in forecasts compared to the bottom five MSA's.[15] This trend holds for in-sample and out-of-sample MSE. Across the board the more stable

[14] The full model reduced in-sample error by 5.30% and out-of-sample error by 10.677% whereas the model with only PSI and its crisis interaction term improved error by 5.703% for in-sample error and by 10.748% for out-of-sample error.

[15] Refer to Table 12 in the Appendix for the full table of mean squared errors for the individual MSA's for the overall sample, in-sample and out-of-sample MSE

markets exhibit less forecasting error. Given the increased volatility in housing price dynamics in MSA's like Miami and Phoenix, it is more difficult to forecast future housing price fluctuations and therefore, more error is involved in those forecasts.

6. Discussion and Conclusion

The results for this study and other literature in this field suggest that incorporating search query data into housing price models improves the predictability of future housing price fluctuations. It may serve as a sentiment-like indicator for the American public, specifically prospective homebuyers and current homeowners who may or may not be concerned about their ability to afford their home and mortgage. The housing distress index (HDI) follows the methodological approach set forth by Chauvet et al. (2013). My study corroborates their results that using negative search query data can improve models and serve as a fear gauge for the level of distress in the U.S. housing market. The study extends the time period of Chauvet et al.'s (2013) study and includes individual MSA models. While there were a number of confounding results, in general the MSA's with more volatile housing markets experience greater fluctuations in housing prices given an increase in negative real estate search queries. Increases in distress queries in more speculative, unstable housing markets likely leads to an increase in levels of fear and panic compared to relatively stable and non-speculative markets.

The main contribution and extension of this research is the creation of the positive sentiment index (PSI). This serves as a positive sentiment indicator to complement the HDI. While the two search query indices are negatively correlated, it is not evident that the HDI and PSI are impacted by the same dynamics. Therefore, it is beneficial to include both indices in housing price models. The PSI captures a new dimension of the home buying process that has not been previously studied in the literature. As is expected, an increase in the number of positive real estate search queries– like "buy a

house" and "real estate agency" – predicts an increase in future housing prices. Additionally, the results show that during crisis times an increase in positive real estate searches leads to a larger positive increase in future housing prices. Given the positive correlation between the PSI and both the FHFA and Case-Shiller HPI's this seems reasonable. This means that during a crisis, a decrease in positive real estate search queries would predict a decrease in future housing prices, which is in line with expectations. As was mentioned in Section 5, including both the HDI and PSI into housing price models reduces in-sample and out-of-sample error at the national level and in certain MSA's. This did not hold true for a number of the MSA's. Overall, the inclusion of search query data into housing price models reduces forecasting error, especially at the national level. This may further strengthen Chauvet et al.'s (2013) assertion that national level search query data is more predictive of housing price fluctuations than MSA-level search query data. Notably, including only the PSI and PSI crisis interaction term reduced error by more than the full model at the national level for the Case-Shiller 20-City Composite Housing Index. While this phenomenon requires further research, but the PSI may serve as a better predictor of future housing price movements than the negative HDI. The two indices provide insight into different sides of the market. The HDI illuminates the side of the market where homeowners are concerned about their ability to afford their mortgages, while the PSI provides information on prospective home buyers and people wishing to purchase a house.

Since Google Trends is relatively new, Google econometrics is still an emerging field for economists to explore. Search query data is only available from 2004 to the present, which means the available data is less extensive than would be optimal. The 2007-2008 Financial Crisis and the recession that followed is the only downturn that Google Trends data encompasses. If Google Trends spanned a longer time period, the results could prove to be more robust and powerful.

The positive sentiment index (PSI) can continue to be improved and elaborated upon. As this is the first study to research and implement the PSI, there is room for continued exploration into the topic. Specifically, many of the queries in the index are related to the financing aspect of the home buying process. This may be a less frequently queried portion of the home buying process and additional queries about real estate agents or searching for a home could prove beneficial. Also, access to Zillow real estate search data from real estate agents and prospective home buyers may be insightful into the two sides of the home buying process. At this present time this is not publically available but in the future it may become available.

Given the rise in social media platforms like Facebook, Twitter, Snapchat and Instagram, there is an opportunity to hone the data embedded in these social media outlets to serve as a complement or replacement to Google Trends. This may, however, present privacy issues by extracting user-level data. This study continues the momentum and increased prevalence of using search query data in real estate economics studies and will be more commonplace as the world continues to globalize and become more reliant on the internet and electronics to conduct business and everyday life. If the use of search query data becomes a widely accepted tool for housing price forecasts, it could benefit policy makers, developers, real estate agencies and the American public as a whole.

References

Bennöhr, L., & Oestmann, M. (2014). *Determinants of House Price Dynamics. What Can We Learn From Search Engine Data?* (SSRN Scholarly Paper No. ID 2513199). Rochester, NY: Social Science Research Network. Retrieved from https://papers.ssrn.com/abstract=2513199

Beracha, E., & Wintoki, B. (2012). Forecasting Residential Real Estate Price Changes from Online Search Activity. *Journal of Real Estate Research*. Retrieved from http://pages.jh.edu/jrer/papers/pdf/forth/accepted/Forecasting%20Residential%20Real%20Estate%20Price%20Changes%20from%20Online%20Search%20Activity.pdf

Case, K. E., & Shiller, R. J. (1988). *The Behavior of Home Buyers in Boom and Post-Boom Markets* (Working Paper No. 2748). National Bureau of Economic Research. Retrieved from http://www.nber.org/papers/w2748

Case, K. E., & Shiller, R. J. (1990). Forecasting Prices and Excess Returns in the Housing Market. *American Real Estate and Urban Economics Association Journal*, *18*(3), 253–273.

Case, K. E., & Shiller, R. J. (2003). Is There a Bubble in the Housing Market? *Brookings Papers on Economic Activity*, (2), 299–362.

Chauvet, M., Gabriel, S. A., & Lutz, C. (2013). *Fear and Loathing in the Housing Market: Evidence from Search Query Data* (SSRN Scholarly Paper No. ID 2148769). Rochester, NY: Social Science Research Network. Retrieved from http://papers.ssrn.com/abstract=2148769

Choi, H., & Varian, H. (2009, April). Predicting the Present with Google Trends. Retrieved September 10, 2016, from http://static.googleusercontent.com/external_content/untrusted_dlcp/www.google.com/en/us/googleblogs/pdfs/google_predicting_the_present.pdf

Das, P., Ziobrowski, A., & Coulson, E. (2014). Online Information Search, Market Fundamentals and Apartment Real Estate. *American Real Estate and Urban Economics Association*

Journal. Retrieved from http://www.academia.edu/7709599/Online_Information_Search_Market_Fundamentals_and_Apartment_Real_Estate

Gabriel, S. A., & Lutz, C. (2014). *The Impact of Unconventional Monetary Policy on Real Estate Markets* (Working Paper). Federal Reserve Bank of San Francisco. Retrieved from http://www.frbsf.org/economic-research/files/MonetaryPolicy_Housing_RealEstate.LUTZ-GABRIELpdf.pdf

Ginsberg, J., Mohebbi, M. H., Patel, R. S., Brammer, L., Smolinski, M. S., & Brilliant, L. (2009). Detecting influenza epidemics using search engine query data. *Nature, 457*(7232), 1012–1014. https://doi.org/10.1038/nature07634

Hohenstatt, R., M. Kaesbauer, and W. Schaefers. (2011). "Geco'' and its Potential for Real Estate Research: Evidence from the U.S. Housing Market. *Journal of Real Estate Research*, 2011, 33:4, 471–506.

Hohenstatt, R., & Kaesbauer, M. (2014). GECO's Weather Forecast for the U.K. Housing Market: To What Extent Can We Rely on Google Econometrics? *Journal of Real Estate Research*, *36*(2), 253–281.

Hott, C. (2012). The Influence of Herding Behavior on House Prices. *Journal of European Real Estate Research*, *5*(3), 177–198. https://doi.org/10.1108/17539261211282046

Housing Price Index Frequently Asked Questions. (n.d.). Retrieved April 18, 2017, from https://www.fhfa.gov/media/publicaffairs/pages/housing-price-index-frequently-asked-questions.aspx

Marian Alexander Dietzel. (2016). Sentiment-based predictions of housing market turning points with Google trends. *International Journal of Housing Markets and Analysis*, *9*(1), 108–136. https://doi.org/10.1108/IJHMA-12-2014-0058

Shiller, R. J. (2007). *Historic Turning Points in Real Estate* (Cowles Foundation Discussion Paper No. 1610). Cowles Foundation for Research in Economics, Yale University. Retrieved from https://ideas.repec.org/p/cwl/cwldpp/1610.html

Webb, G. (2009). Internet Search Statistics as a Source of Business Intelligence: Searches on Foreclosure as an Estimate of Actual Home Foreclosures. *Issues in Information Systems*, 82–87.

Wu, L., & Brynjolfsson, E. (2013). *The Future of Prediction: How Google Searches Foreshadow Housing Prices and Sales* (SSRN Scholarly Paper No. ID 2022293). Rochester, NY: Social Science Research Network. Retrieved from http://papers.ssrn.com/abstract=2022293

Yavas, A. (1992). A Simple Search and Bargaining Model of Real Estate Markets. *American Real Estate and Urban Economics Association Journal*, *20*(4), 533–548.

Data Sources

Board of Governors of the Federal Reserve System (US), Delinquency Rate on Single-Family Residential Mortgages, Booked in Domestic Offices, All Commercial Banks [DRSFRMACBS], retrieved from FRED, Federal Reserve Bank of St. Louis; https://fred.stlouisfed.org/series/DRSFRMACBS, November 19, 2016.

Google. (2016). Google Trends Data [Data file] Retrieved September 11, 2016, from https://trends.google.com/trends/

S&P Dow Jones Indices LLC, S&P/Case-Shiller 20-City Composite Home Price Index© [SPCS20RSA], retrieved from FRED, Federal Reserve Bank of St. Louis; https://fred.stlouisfed.org/series/SPCS20RSA, November 20, 2016.

US. Bureau of the Census, Housing Starts: Total: New Privately Owned Housing Units Started [HOUSTNSA], retrieved from FRED, Federal Reserve Bank of St. Louis; https://fred.stlouisfed.org/series/HOUSTNSA, November 19, 2016.

US. Bureau of the Census, Retail Sales: Total (Excluding Food Services) [RSXFS], retrieved from FRED, Federal Reserve Bank of St. Louis; https://fred.stlouisfed.org/series/RSXFS, November 19, 2016.

Federal Reserve Bank of St. Louis, 10-Year Treasury Constant Maturity Minus Federal Funds Rate [T10YFFM], retrieved from FRED, Federal Reserve Bank of St. Louis; https://fred.stlouisfed.org/series/T10YFFM, November 19, 2016.

US. Bureau of Labor Statistics, Civilian Unemployment Rate [UNRATE], retrieved from FRED, Federal Reserve Bank of St. Louis; https://fred.stlouisfed.org/series/UNRATE, November 19, 2016.

U.S. Federal Housing Finance Agency, All-Transactions House Price Index for the United States [USSTHPI], retrieved from FRED, Federal Reserve Bank of St. Louis; https://fred.stlouisfed.org/series/USSTHPI, November 19, 2016.

Appendices

Table 1. The Search Queries used in the Housing Distress Index (HDI)

Search Query	Date Peaked (Query = 100)	Search Query	Date Peaked (Query =100)
Mortgage Help	March 2009	Help with Mortgage	March 2009
Mortgage Government Help	March 2009	Help for Mortgage	March 2009
Mortgage Foreclosure Help	February 2009	Government Mortgage Help	March 2009
Mortgage Assistance Program	July 2011	Foreclosure Help	February 2009
Mortgage Assistance	March 2009	Foreclosure Assistance	August 2008 & March 2009
Home Mortgage Assistance	March 2009		

Note: These 11 unique search queries are what make up the housing distress index (HDI) used in this study. The query shares of each unique query are summed together and seasonally adjusted using the U.S. Census Bureau's x12 algorithm. Next, I standardized the HDI to have zero mean and unit variance. The data spans January 2004 – December 2016.

Table 2. The Search Queries used in the Positive Sentiment Index (PSI)

Search Query	Date Peaked (query = 100)	Search Query	Date Peaked (query = 100)
Home Financing	July 2004	Adjustable Rate mortgage	April 2004
Real Estate Agency	June 2004	Option ARM	September 2006
No down payment mortgage	July 2004	No Income Mortgage	July 2004
Mortgage Adjustable rate	April 2004	Buying a house	January 2016
Home equity line of credit	March 2004	Mortgage loan	March 2004
Buying a new house	January 2004	Buy a house	January 2012
How to buy a house	January 2012	First time homeowner	February 2009
Starter home	March 2016		

Note: These 15 unique search queries are what make up the positive sentiment index (PSI) used in this study. The query shares of each unique query are summed together and seasonally adjusted using the U.S. Census Bureau's x12 algorithm. Next, I standardized the PSI to have zero mean and unit variance. The data spans January 2004 – December 2016.

Table 3. Correlation coefficients between the search query indices and housing price indices

	Full sample	Crisis Period
Correlation (HDI, FHFA HPI)	-0.448	-0.922
Correlation (HDI_{t-1}, FHFA HPI)	-0.479	-0.954
Correlation (HDI, Case-Shiller HPI)	-0.569	-0.938
Correlation (HDI_{t-1}, Case-Shiller HPI)	-0.604	-0.955
Correlation (PSI, FHFA HPI)	0.170	0.399
Correlation (PSI_{t-1}, FHFA HPI)	0.206	0.472
Correlation (PSI, Case-Shiller HPI)	0.395	0.345
Correlation (PSI_{t-1}, Case-Shiller HPI)	0.425	0.431

Notes: This table shows the correlation coefficients between the two constructed search query indices, the housing distress index (HDI) and the positive sentiment index (PSI), and either the FHFA National Housing Price Returns or the Case-Shiller 20-City Composite National Housing Price Returns. It should be noted that the HDI and PSI used here is the non-standardized version of the variable and the housing price variables are not the first difference housing price variables used in the regressions. The table also includes correlation coefficients between one month lags of the search query indices and the housing price indexes. For each search query and housing price index pair the table shows the correlation coefficient for the full sample (January 2004 – December 2016) and for the crisis period (January 2007 – March 2009).

Table 4. Fixed Effects Regressions of the HDI & PSI's Impact on Housing Price Indices

	(1)	(2)	(3)
VARIABLES	Case-Shiller Monthly	Case-Shiller Quarterly	FHFA Quarterly

HDI_{t-1}	-0.0641***	-0.233	-0.714***
	(0.0247)	(0.154)	(0.175)
HDI_{t-1}^{crisis}	-0.0623	-0.291	0.0411
	(0.0420)	(0.272)	(0.308)
PSI_{t-1}	0.0927***	0.522***	0.697***
	(0.0202)	(0.151)	(0.170)
PSI_{t-1}^{crisis}	0.195***	1.012***	-0.634
	(0.0555)	(0.360)	(0.399)
Constant	-0.125*	-1.165**	-2.806***
	(0.0692)	(0.462)	(0.547)
Observations	3,040	940	940
R-squared	0.800	0.711	0.719
Number of MSA	20	20	20
Controls	Included	Included	Included
MSA Fixed Effects	YES	YES	YES
B1 + B2	-0.1264***	-0.5235**	-0.67262**
	(0.0006)	(0.0355)	(0.0149)
B3 + B4	0.2874***	1.5338***	0.0635
	(0.0000)	(0.000)	(0.8558)

Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Notes: This table displays panel regression results of housing price fluctuations on the HDI, PSI and controls. HDI and PSI are both standardized indices such that the mean is 0 and the variance is 1. These variables are also adjusted in accordance with the x12 algorithm put out by the U.S. Census Bureau. The dependent variable is either the Case-Shiller Housing Price Index or the Federal Housing Funding Agency (FHFA) Housing Price Index. The controls include MSA fixed effects, three or four lags of the dependent variable and a one period lag of the unemployment rate for each respective MSA. The number of lags is determined based on AIC. Crisis interaction terms are included in the regressions. The crisis period is set to be from January 2007 to March 2009 and will only be equal to 1 during that time period. Case-Shiller puts out MSA level housing price data on a monthly basis while FHFA MSA-level HPI data is only available on a quarterly basis. Such is the difference in frequency for the above regressions. The Case-Shiller monthly data is collapsed into quarterly data for model (2) to investigate the quarterly impact of the HDI and PSI. The sum of the coefficients B1 + B2 displays the total effect of the HDI on housing price returns during the crisis period. The sum of B3 + B4 represents the total effect of the PSI on housing returns during the crisis period. The p-value shows the result from the F-statistic that tests the null hypothesis that the sum of the coefficients is equal to zero. This p-value is listed in parentheses.

Table 5. OLS Regressions of the HDI & PSI's Impact on Housing Price Indices at a National Level

	(1)	(2)
VARIABLES	FHFA	Case-Shiller
HDI_{t-1}	-1.145***	-0.0332

	(0.383)	(0.251)
HDI_{t-1}^{crisis}	0.120	-0.0886
	(0.287)	(0.179)
PSI_{t-1}	0.266	0.217
	(0.341)	(0.181)
PSI_{t-1}^{crisis}	-0.243	-0.167
	(0.526)	(0.341)
Constant	-1.044	-4.532**
	(3.394)	(2.129)
Observations	141	145
R-squared	0.707	0.911
Controls	Included	Included
B1+B2	-1.025***	-0.1218
	(0.000)	(0.459)
B3+B4	0.0227	0.0493
	(0.961)	(0.868)

Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Notes: This table displays OLS regression results of housing price fluctuations on the HDI, PSI and controls. HDI and PSI are both standardized indices with a mean of 0 and a variance of 1. These variables are seasonally adjusted with the x12 algorithm from the U.S. Census Bureau. The dependent variable is either the Case-Shiller Housing Price Index or the Federal Housing Funding Agency (FHFA) Housing Price Index. The controls include 10 or 13 lags of the dependent variable (10 lags for the Case-Shiller regressions and 13 lags for the FHFA regressions), housing starts, the difference between the 10-year Treasury rate and the Federal Funds rate, the unemployment rate for the U.S., retail sales, and a two quarter lag of all of the search query variables and their interaction terms. The number of lags is determined based on AIC. Crisis interaction terms are included in the regressions. The crisis period is set to be from January 2007 to March 2009 and will only be equal to 1 during that time period. All of the data in the regressions is monthly. The delinquency rate is converted to a monthly frequency (refer to footnote 3 on page 18 for a full description of this process). The sum of the coefficients B1 + B2 displays the total effect of the HDI on housing price returns during the crisis period. The sum of B3 + B4 represents the total effect of the PSI on housing returns during the crisis period. The p-value shown is the result from the F-statistic that tests the null hypothesis that the sum of the coefficients is equal to zero. This p-value is listed in parentheses.

Table 6. The HDI & PSI's Impact on MSA-level Housing Price Indices'

VARIABLES	(1) C.S. PHX	(2) PHX FHFA	(3) MIA C.S.	(4) MIA FHFA	(5) LV C.S.	(6) LV FHFA	(7) LA C.S.	(8) LA FHFA	(9) SF FHFA	(10) SF C.S.
$HDI_{t-1, MSA}$	-2.866**	-4.185*	-3.300	-7.371***	-2.679***	-4.822***	0.626	0.333	-1.195	-1.310*
	(1.254)	(2.077)	(1.976)	(1.774)	(0.963)	(1.061)	(0.733)	(0.989)	(1.432)	(0.760)
$HDI_{t-1, MSA}^{CRISIS}$	1.123	-1.702	4.139**	1.931	-0.0994	-4.283**	-0.323	-0.662	3.397	-2.174
	(1.382)	(2.175)	(1.794)	(2.457)	(1.481)	(1.926)	(0.987)	(1.518)	(3.355)	(2.127)
$PSI_{t-1, MSA}$	2.907**	4.212**	2.951*	5.410***	1.613**	2.418***	1.311	1.734	2.554*	2.280***
	(1.376)	(2.035)	(1.525)	(1.264)	(0.677)	(0.750)	(0.826)	(1.220)	(1.491)	(0.797)
$PSI_{t-1, MSA}^{CRISIS}$	2.690	-14.66**	9.800*	4.029	-2.262	-5.640*	4.281**	3.696	4.629	0.876
	(4.741)	(6.734)	(5.663)	(8.227)	(2.740)	(2.827)	(1.859)	(2.670)	(4.386)	(3.009)
Constant	-8.374**	-7.347	-7.450	-14.04**	-5.300***	-4.906**	1.404	2.382	-4.024	-1.468
	(3.178)	(4.999)	(5.758)	(5.600)	(1.697)	(2.075)	(2.149)	(4.107)	(4.548)	(2.282)
Observations	47	47	47	47	47	47	47	47	47	47
R-squared	0.898	0.804	0.867	0.864	0.898	0.890	0.926	0.882	0.762	0.882
Controls	Included	Included	Included	Included	Included	Included	Included	Included	Included	Included
B1+B2	-1.74	-5.89**	0.838	-5.44*	-2.78	-9.10***	0.303	-0.329	2.202	-3.484
	(0.303)	(0.013)	(0.695)	(0.065)	(0.125)	(0.000)	(0.728)	(0.804)	(0.512)	(0.106)
B3+B4	5.60	-10.45	12.751**	9.44	-0.65	-3.22	5.59***	5.431	7.183	3.155
	(0.215)	(0.139)	(0.021)	(0.258)	(0.802)	(0.248)	(0.001)	(0.013)	(0.115)	(0.303)

Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Notes: This table displays OLS regression results of housing price fluctuations on HDI, PSI and controls at the MSA level. The MSA-level HDI and PSI variables are standardized indices such that the mean is 0 and the variance is 1. These variables are also adjusted in accordance with the x12 algorithm put out by the U.S. Census Bureau. The dependent variable is either the Case-Shiller Housing Price Index or the Federal Housing Funding Agency (FHFA) Housing Price Index at the MSA-level. The controls include 4 quarterly lags of the dependent variable, and a one quarter lag of unemployment rate for that specific MSA. The number of lags is determined based on AIC. Crisis interaction terms are included in the regressions. The crisis period is set to be from January 2007 to March 2009 and will only be equal to 1 during that time period. All of the data in the regressions is quarterly. The p-value shown for B1+B2 and B3+ B4 is the result for the F-statistic testing the null hypothesis that the sum is equal to zero. The p-value is shown in parentheses.

Table 7. The HDI & PSI's Impact on MSA-level Housing Price Indices'

VARIABLES	(1) CLE FHFA	(2) CLE C.S.	(3) CHAR C.S.	(4) CHAR FHFA	(5) DAL FHFA	(6) DAL C.S.	(7) DEN C.S.	(8) DEN FHFA	(9) BOS C.S.	(10) BOS FHFA
$HDI_{t-1,\ MSA}$	-1.228*	-1.384**	-0.601*	-0.746	-1.227**	-1.153***	-1.476***	-2.011**	-1.285	-2.519***
	(0.628)	(0.545)	(0.308)	(0.563)	(0.513)	(0.383)	(0.483)	(0.921)	(0.962)	(0.855)
$HDI_{t-1,\ MSA}^{crisis}$	1.085	1.324*	-0.577	-0.445	0.315	2.650***	0.881	1.561	-0.183	1.415
	(0.737)	(0.673)	(0.778)	(1.218)	(0.984)	(0.684)	(0.616)	(1.207)	(0.848)	(0.867)
$PSI_{t-1,\ MSA}$	0.290	-0.336	0.0858	-0.509	0.820*	0.636*	0.184	0.0191	1.557***	1.737***
	(0.430)	(0.414)	(0.289)	(0.472)	(0.480)	(0.373)	(0.385)	(0.755)	(0.503)	(0.488)
$PSI_{t-1,\ MSA}^{crisis}$	1.133	2.130**	0.709	0.492	-1.472	3.189***	1.140	-0.456	-1.118	-0.883
	(0.944)	(0.923)	(0.993)	(1.749)	(1.524)	(1.163)	(0.784)	(1.418)	(1.016)	(0.990)
Constant	-1.270	-3.511*	0.340	9.968***	0.0632	1.146	0.0146	-2.522	-4.260	-5.365**
	(2.051)	(1.955)	(1.233)	(2.488)	(2.001)	(1.378)	(1.001)	(1.966)	(2.874)	(2.473)
Observations	47	47	47	47	47	47	47	47	47	47
R-squared	0.599	0.398	0.796	0.730	0.727	0.741	0.775	0.734	0.577	0.771
Controls	Included	Included	Included	Included	Included	Included	Included	Included	Included	Included
B1+B2	-0.143	-0.060	-1.178*	-1.190	-0.912	1.497**	-0.595	-0.450	-1.468**	-1.104*
	(0.813)	(0.920)	(0.086)	(0.209)	(0.282)	(0.028)	(0.361)	(0.658)	(0.027)	(0.053)
B3+B4	1.42	1.794**	0.795	-0.0166	-0.652	3.826***	1.325*	-0.437	0.440	0.854
	(0.119)	(0.050)	(0.403)	(0.992)	(0.631)	(0.001)	(0.067)	(0.740)	(0.596)	(0.364)

Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Notes: This table displays OLS regression results of housing price fluctuations on HDI, PSI and controls at the MSA level. The MSA-level HDI and PSI variables are standardized indices such that the mean is 0 and the variance is 1. These variables are also adjusted in accordance with the x12 algorithm put out by the U.S. Census Bureau. The dependent variable is either the Case-Shiller Housing Price Index or the Federal Housing Funding Agency (FHFA) Housing Price Index at the MSA-level. The controls include 4 quarterly lags of the dependent variable, and a one quarter lag of unemployment rate for that specific MSA. The number of lags is determined based on AIC. Crisis interaction terms are included in the regressions. The crisis period is set to be from January 2007 to March 2009 and will only be equal to 1 during that time period. All of the data in the regressions is quarterly. The p-value shown for B1+B2 and B3+B4 is the result for the F-statistic testing the null hypothesis that the sum is equal to zero. The p-value is shown in parentheses.

Table 8. The HDI & PSI's Impact on MSA-level Housing Price Indices'

VARIABLES	(1) NY C.S.	(2) NY FHFA	(3) POR C.S.	(4) POR FHFA	(5) D.C. C.S.	(6) D.C. FHFA	(7) SD FHFA	(8) SD C.S.	(9) SEA FHFA	(10) SEA C.S.
$HDI_{t-1,\ MSA}$	-1.539**	-0.988	-2.650***	-3.468***	0.221	-0.0967	-2.993**	-1.591*	-1.841*	-1.136
	(0.690)	(0.766)	(0.819)	(1.008)	(0.595)	(0.633)	(1.176)	(0.885)	(0.998)	(0.686)
$HDI_{t-1,\ MSA}^{crisis}$	0.409	0.321	0.171	-1.297	-0.582	-4.804***	1.048	0.194	-1.816	0.536
	(0.649)	(0.749)	(1.197)	(1.567)	(1.717)	(1.738)	(1.491)	(1.181)	(1.157)	(0.854)
$PSI_{t-1,\ MSA}$	2.318***	1.177	1.820***	2.212***	2.116**	2.672***	0.739	0.521	-0.0244	-0.164
	(0.752)	(0.900)	(0.634)	(0.780)	(0.838)	(0.965)	(1.109)	(0.820)	(0.597)	(0.413)
$PSI_{t-1,\ MSA}^{crisis}$	-0.262	0.198	5.517	-9.358	0.0293	-7.537***	4.400*	4.573**	1.488	2.579**
	(1.846)	(2.137)	(8.981)	(11.10)	(2.408)	(2.557)	(2.393)	(1.852)	(1.415)	(0.960)
Constant	-3.992**	-1.653	-1.171	-2.425	-6.148**	-5.672*	-3.816	-1.635	4.515	1.168
	(1.886)	(2.382)	(1.999)	(3.080)	(2.396)	(3.069)	(2.760)	(2.100)	(4.038)	(2.274)
Observations	47	47	47	47	47	47	47	47	47	47
R-squared	0.783	0.810	0.861	0.863	0.826	0.829	0.833	0.869	0.899	0.909
Controls	Included	Included	Included	Included	Included	Included	Included	Included	Included	Included
B1+B2	-1.13**	-0.67	-2.48*	-4.77***	-0.36	-4.90***	-1.94	-1.40	-3.66***	-0.60
	(0.024)	(0.199)	(0.061)	(0.002)	(0.816)	(0.003)	(0.189)	(0.264)	(0.002)	(0.477)
B3+B4	2.06	1.38	7.34	-7.15	2.15	4.86*	5.14**	5.09***	1.46	2.42**
	(0.208)	(0.477)	(0.421)	(0.526)	(0.374)	(0.071)	(0.023)	(0.004)	(0.305)	(0.012)

Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Notes: This table displays OLS regression results of housing price fluctuations on HDI, PSI and controls at the MSA level. The MSA-level HDI and PSI variables are standardized indices such that the mean is 0 and the variance is 1. These variables are also adjusted in accordance with the x12 algorithm put out by the U.S. Census Bureau. The dependent variable is either the Case-Shiller Housing Price Index or the Federal Housing Funding Agency (FHFA) Housing Price Index at the MSA-level. The controls include 4 quarterly lags of the dependent variable, and a one quarter lag of unemployment rate for that specific MSA. The number of lags is determined based on AIC. Crisis interaction terms are included in the regressions. The crisis period is set to be from January 2007 to March 2009 and will only be equal to 1 during that time period. All of the data in the regressions is quarterly. The p-value shown for B1+B2 and B3+ B4 is the result for the F-statistic testing the null hypothesis that the sum is equal to zero. The p-value is shown in parentheses.

Table 9. The HDI & PSI's Impact on MSA-level Housing Price Indices'

VARIABLES	(1) TAMPA C.S.	(2) TAMPA FHFA	(3) ATL C.S.	(4) ATL FHFA	(5) CHI C.S.	(6) CHI FHFA	(7) DET C.S.	(8) DET FHFA	(9) MIN C.S.	(10) MIN FHFA
$HDI_{t-1,\ MSA}$	-1.538	-6.595***	-0.909	-3.411***	-0.0673	-0.423	-2.345**	-1.253	-0.436	-2.056
	(1.219)	(2.116)	(0.645)	(0.813)	(0.416)	(0.351)	(0.900)	(1.665)	(1.341)	(1.618)
$HDI_{t-1,\ MSA}^{crisis}$	1.348	3.831*	0.363	1.246	-0.280	0.0195	-0.137	0.946	-1.106	0.0570
	(1.263)	(2.058)	(0.792)	(0.902)	(0.744)	(0.616)	(0.766)	(1.441)	(1.250)	(1.303)
$PSI_{t-1,\ MSA}$	3.228***	6.198***	0.268	0.0748	0.633	0.537	0.643	0.955	0.640	1.217
	(0.986)	(1.607)	(0.554)	(0.551)	(0.510)	(0.436)	(0.518)	(0.818)	(0.796)	(0.864)
$PSI_{t-1,\ MSA}^{crisis}$	3.925	-5.782	2.020	-0.388	2.747**	-0.160	0.323	0.841	4.994**	0.858
	(2.984)	(5.116)	(1.845)	(1.939)	(1.015)	(0.913)	(1.054)	(1.775)	(2.120)	(2.181)
Constant	-2.340	-22.90***	0.149	0.692	1.453	-3.393	-5.714**	-3.826	1.838	2.113
	(4.257)	(8.175)	(2.112)	(2.367)	(1.597)	(2.089)	(2.127)	(3.448)	(3.764)	(4.040)
Observations	47	47	47	47	47	47	47	47	47	47
R-squared	0.902	0.783	0.724	0.760	0.731	0.798	0.793	0.613	0.745	0.692
Controls	Included	Included	Included	Included	Included	Included	Included	Included	Included	Included
B1+B2	-0.19	-2.76	-0.55	-2.17***	-0.35	-0.40	-2.48**	-0.31	-1.54	-2.00*
	(0.895)	(0.222)	(0.517)	(0.007)	(0.586)	(0.418)	(0.012)	(0.850)	(0.245)	(0.100)
B3+B4	7.15***	0.42	2.29	-0.31	3.38***	0.38	0.97	1.77	5.63	2.08
	(0.008)	(0.931)	(0.150)	(0.850)	(0.000)	(0.653)	(0.299)	(0.274)	(0.008)	(0.333)

Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Notes: This table displays OLS regression results of housing price fluctuations on HDI, PSI and controls at the MSA level. The MSA-level HDI and PSI variables are standardized indices such that the mean is 0 and the variance is 1. These variables are also adjusted in accordance with the x12 algorithm put out by the U.S. Census Bureau. The dependent variable is either the Case-Shiller Housing Price Index or the Federal Housing Funding Agency (FHFA) Housing Price Index at the MSA-level. The controls include 4 quarterly lags of the dependent variable, and a one quarter lag of unemployment rate for that specific MSA. The number of lags is determined based on AIC. Crisis interaction terms are included in the regressions. The crisis period is set to be from January 2007 to March 2009 and will only be equal to 1 during that time period. All of the data in the regressions is quarterly. The p-value shown for B1+B2 and B3+ B4 is the result for the F-statistic testing the null hypothesis that the sum is equal to zero. The p-value is shown in parentheses.

Table 10. The Percent Error Reduced by Including HDI & PSI into Models

MSA	FHFA error change (%)		Case-Shiller Error Change (%)	
	In Sample	Out of Sample	In Sample	Out of Sample
Phoenix	- 30.605	601.881	-19.724	106.925
Miami	-39.961	9.009	-26.053	-44.011
San Fran.	-17.67	5.995	-31.352	-42.93
Las Vegas	-49.283	23.817	-17.165	-50.82
Los Ang.	-36.542	1210.478	-39.518	89.581
San Diego	-21.32	-72.421	-28.803	-36.544
Wash. D.C.	-37.707	-12.745	-33.331	187.488
Portland	-42.875	28.386	-25.965	38.084
Seattle	-39.94	11.052	-48.341	112.103
Tampa	-35.06	-27.586	-49.67	8.73
Detroit	-6.242	-31.386	-35.934	2.743
New York	-7.252	20.601	-23.568	-24.455
Chicago	-6.053	-23.633	-40.148	130.672
Minneapolis	-25.467	-11.767	-43.064	-39.363
Denver	-21.388	80.152	-33.749	73.584
Cleveland	-30.908	359.927	-32.145	33.254
Charlotte	-17.867	-39.892	-35.485	30.76
Dallas	-20.958	-24.169	-46.917	12.05
Atlanta	-46.011	50.725	-24.248	24.678
Boston	-37.641	-18.601	-32.969	-9.528
Average	**-28.538**	**106.991**	**-33.408**	**30.150**

Notes: This table displays the amount of error reduced (in percent terms) by adding the HDI, PSI and their interaction terms into forecasting regressions. Reducing error signifies that the forecasted housing price index value was closer to the true housing index value. A negative value means the forecasts got better with the inclusion of the search query indices while positive error means the model did a worse job of predicting future housing prices. This forecasting technique follows the methodology outline in Section 4 of the study. Refer back to Section 4 for a complete description of the forecasting process.

Table 11. The Percent Change in Mean Squared Error for the Forecasting Models with the Inclusion of Search Query Indices on the Panel Data

Model	FHFA Error Reduced (%)		Case-Shiller Error Reduced (%)	
	In Sample	Out of Sample	In Sample	Out of Sample
All Search Queries	-0.889	-8.826	-5.300	-10.677
Only HDI	-0.145	-9.056	-0.170	-6.628
Only PSI	-0.908	-1.491	-3.817	-4.781
PSI & PSI*Crisis	-0.667	2.099	-5.703	-10.748
HDI & HDI*Crisis	-0.406	-7.771	-1.561	-5.546

Notes: This table displays the amount of error reduced (in percent terms) by adding various combinations of the HDI, PSI and their crisis interaction terms into forecasting regressions. Reducing error signifies that the forecasted housing price index value is closer to the true housing index value. Negative change in error means the forecasts got better with the inclusion of the search query indices while positive error means the model did a worse job of predicting future housing prices. This forecasting technique follows the methodology outline in Section 4 of the study. Refer back to Section 4 for a complete description of the forecasting process. The terms PSI*Crisis and HDI*Crisis signify that the crisis interaction term was included in the models. The "All Search Queries" Model reflects the utilization of the HDI, PSI as well as both of the crisis interaction terms.

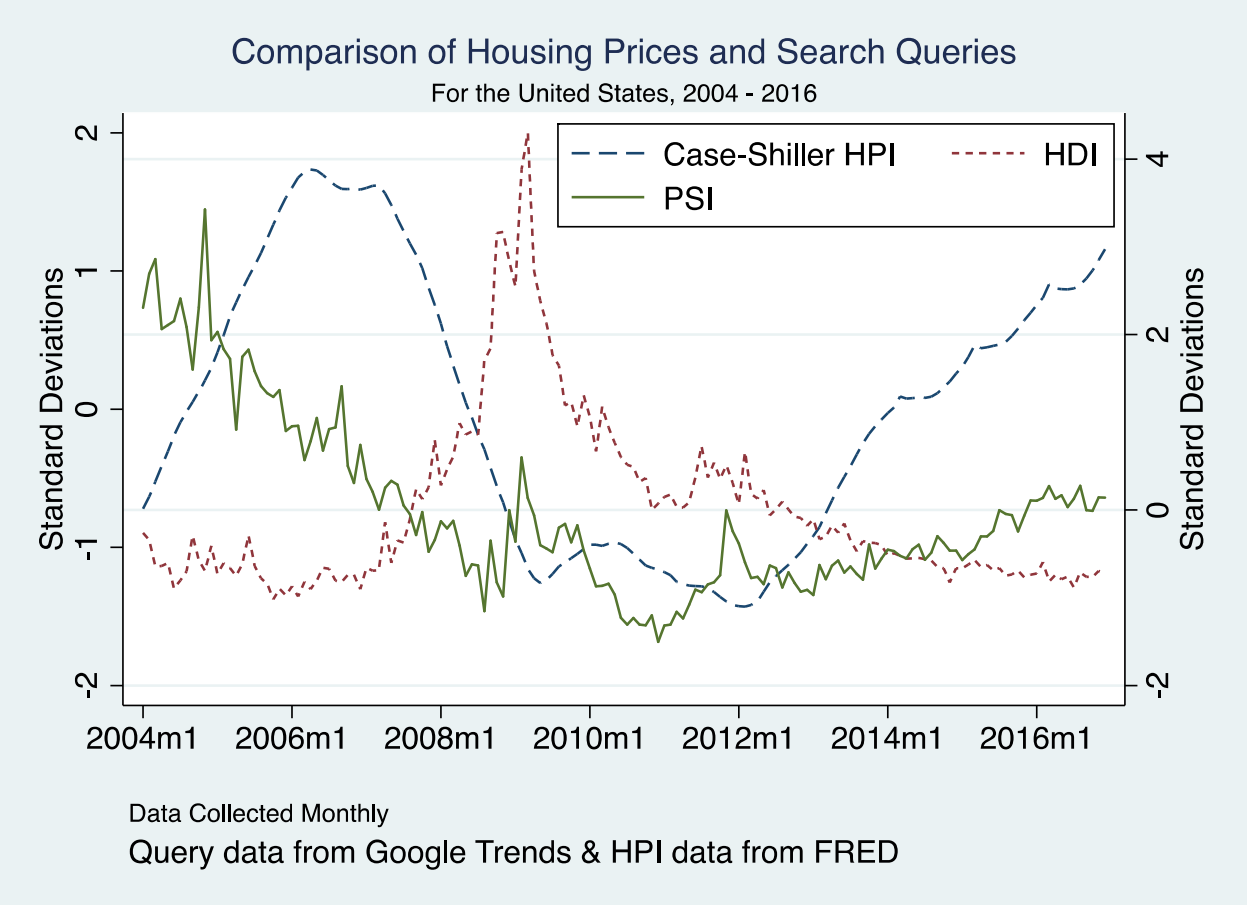
Table 12: Mean Squared Error for the Forecasting Models at the MSA-level

MSA	FHFA Mean Squared Forecasting Errors			Case-Shiller Mean Squared Forecasting Errors		
	Overall MSE	In Sample	Out of Sample	Overall MSE	In Sample	Out of Sample
Phoenix	43.034	39.556	53.176	7.975	8.483	6.493
Miami	15.768	18.688	7.252	10.400	12.884	3.153
San Fran.	13.704	11.292	20.737	5.891	5.010	8.460
Las Vegas	7.097	7.715	5.294	4.761	6.122	0.791
Los Angeles	19.854	9.011	51.478	6.527	3.834	14.384
San Diego	7.680	9.636	1.976	5.224	4.911	6.139
D.C.	5.583	5.739	5.127	9.964	4.663	11.051
Portland	4.870	3.548	8.726	2.327	2.766	3.151
Seattle	3.622	3.242	4.729	2.860	1.214	7.662
Tampa	19.671	15.237	6.253	7.548	4.125	6.149
Detroit	6.302	7.047	1.894	2.190	1.693	3.641
NYC	2.975	2.578	4.134	2.327	2.228	2.613
Chicago	2.211	2.311	1.134	4.326	2.574	9.435
Minneapolis	6.412	5.348	3.691	3.972	4.536	2.327
Denver	4.963	3.226	10.028	1.334	1.122	1.954
Cleveland	2.312	1.579	4.452	1.510	1.566	1.346
Charlotte	1.724	1.863	1.318	0.772	0.550	1.419
Dallas	1.769	1.552	2.403	0.962	1.011	0.820
Atlanta	4.290	2.674	3.555	3.405	2.826	3.062
Boston	2.649	2.590	2.822	2.545	1.805	4.705
Average	**8.824**	**7.722**	**10.009**	**4.341**	**3.696**	**4.938**

Notes: This table displays the amount of overall mean squared error (MSE) present in the MSA forecasting regressions. These regressions contain the HDI, PSI and both crisis interaction terms as independent variables. The smaller the amount of MSE the better because less error means the forecasting predictions were closer to the actual values of the housing price index.

Figures

Figure 1. Comparison of Case-Shiller 20-City Composite Housing Price Index and both Search Query Indices



Notes: The above graph shows the variation over time of the Case-Shiller 20-City Composite Housing Price Index compared to the HDI and PSI. For the purposes of this graph, all three variables are standardized with a mean of 0 and a variance of 1. The left-hand axis shows the variation (in standard deviations) of the Case-Shiller Housing Price Index, while the right-hand side shows the variation over time (in standard deviations) of the two search query indices. The data used spans from January 2004 through December 2016.