INTERNATIONAL REAL ESTATE REVIEW

2020 Vol. 23 No. 2: pp. 893 – 934

Can Google Search Data be Used as a Housing Bubble Indicator?

Are Oust[*]
NTNU Business School, Norwegian University of Science and Technology, 7491 Trondheim, Norway. Email: Are.oust@ntnu.no

Ole Martin Eidjord
Department of Industrial Economics and Technology Management, Norwegian University of Science and Technology, 7491 Trondheim, Norway

The aim of this paper is to test whether Google search volume indices can be used to predict house prices and identify bubbles in the housing market. We analyze the data that pertain to the 2006−2007 U.S. housing bubble, taking advantage of the heterogeneous house price development in both bubble and non-bubble states in the U.S. Using 204 housing-related keywords, we test both single search terms and indices that comprise search term sets to see whether they can be used as housing bubble indicators. We find that several keywords perform very well as bubble indicators. Among all of the keywords and indices tested, the Google search volume for “Housing Bubble” and “Real Estate Agent”, and a constructed index that contains the twelve best-performing search terms score the highest at both detecting bubbles and not erroneously detecting non-bubble states as bubbles. A new housing bubble indicator may help households, investors, and policy makers receive advanced warning about future housing bubbles. Moreover, we show that the Google search outperforms the well-established consumer confidence index in the U.S. as a leading indicator of the housing market.

Keywords

Google Trends, Housing, Housing Bubble Indicator, Housing Bubble, Real Estate Agent

[*] Corresponding author

1. Introduction

Asset-price bubbles have been the cause of some of the largest economic downturns in history. Housing price bubbles, in particular, have had a significant impact on the economy and tend to have much more prolonged effects than other types of bubbles. As residential property comprises the majority of the wealth of many households, its wealth effect on consumption is significant and apparently greater than that of financial assets (see, for example, Case et al., 2001; Benjamin et al., 2004; Campbell and Cocco, 2004). In addition, spillover effects from a housing bubble can be significant due to the large share of housing debt in bank portfolios. When there is a financial crisis, amplification mechanisms are an important factor that amplifies the impact. These mechanisms can be direct or indirect. The former is "caused by direct contractual links" whereas the latter is "caused by spillovers or externalities that are due to common exposures or the endogenous response of various market participants" (Brunnermeier and Oehmke, 2012).

Stiglitz (1990, p. 13) defines a bubble as follows: "…the basic intuition is straightforward: if the reason that the price is high today is only because investors believe that the selling price will be high tomorrow—when 'fundamental' factors do not seem to justify such a price—then a bubble exists". Lind (2009) argues that the most important aspect of bubble theories is their ability to predict future bubbles from a policy point of view. Good indicators should help separate bubbles from normal business cycles. Also, Robert Shiller, a renowned American economist, proposes a bubble checklist to determine if there is indeed a bubble (Ewing, 2010)[1]. The aim of this paper is to examine whether Google search data could be a part of such an indicator system to predict future housing price bubbles, and identify which search keywords have the best performance.

The hypothesis proposed in this work is that the Google search volume can capture/measure general public interest around a given topic. Case and Shiller (2004) use newspaper articles related to housing to measure the extent of media related housing frenzy. In recent decades, rapid developments in information technology and the increasingly widespread use of search engines have enabled new ways of predicting the future (see, for example, Ettredge et al., 2005; Horrigan, 2008; Kuruzovich et al., 2008; Choi and Varian, 2012; Challet and Ayed, 2013). Pentland (2010) finds that as Google searches often precede purchase decisions, they are in many cases a more "honest signal" of actual interests and preferences because no bargaining, gaming, or strategic signaling is involved, in contrast to many market-based transactions or other types of data gathering, such as surveys. Other authors are more skeptical of the use of web

[1] Shiller writes about the psychological factors behind bubbles, for example, in Irrational Exuberance (Shiller, 2005) and the need to look at different indicators which indicate that a bubble exists. Shiller is generally interested in the interest that an asset can generate in both the media and among the public.

searches for prediction. For example, Goel et al. (2010) point out that even though search data are easy to acquire and often helpful in forecasting work, they might not provide dramatic improvements in prediction accuracy. Since Google started providing search volume data in 2004, the information has become increasingly popular as an economic indicator (see for e.g., Bijl et al. 2016; Preis et al., 2010, 2013). Wu and Brynjolfsson (2015) find evidence that queries submitted to the Google search engine are correlated with the volume of housing sales, as well as a house price index—specifically the Case-Shiller index—released by the Federal Housing Finance Agency (FHFA). They further observe that, while search queries can reveal the current housing trends, Google search is especially well-suited for predicting the future unit sales in the housing sector.

We focus on the 2006−2007 U.S. housing bubble, as it was characterized by different housing price developments across the U.S. states. In particular, California, Nevada, Arizona, and Florida experienced a *real* bubble, whereas six states, Michigan, Rhode Island, Maryland, Idaho, Oregon, and Washington are denoted as *minor* bubble states in terms of the size of the boom-bust cycle of the house prices[2]. These bubble states, along with the ten states that experienced the lowest decrease in house prices, serve as the benchmark states in an in-sample bubble identification test. Based on 204 housing market related search terms (Appendix B), we test whether they are leading and correlated with the house prices. We then propose a housing bubble identification approach based on the differences in the Google search volume index (henceforth the GSVI) levels in the housing bubble period compared to a non-bubble period. Subsequently, we test whether individual GSVIs were leading, coincident, or lagging compared to the house prices in the different states.

We find that single search terms such as "Housing Bubble" and "Real Estate Agent" perform the best in the in-sample predictions and also outperform the self-created indices that consist of the average GSVI for the three, six, twelve and twenty best-performing search terms. "Housing Bubble" performs especially well as a housing price bubble indicator, but so do several other search terms. When optimizing the results yielded by the identification of our ten bubble states, the GSVI for "Housing Bubble" correctly identifies all of the bubble states and erroneously indicated a bubble in only one non-bubble state. When the objective is changed to not erroneously detecting non-bubble states as bubbles, the GSVI for "Housing Bubble" indicates bubbles in all four real bubble states, as well as four out of the six minor bubble states. We continue to focus on the two search terms "Real Estate Agent" and "Housing Bubble", as well as the best performing of the indices, Index12 (based on the 12 best performing search terms that performed best among the indices).

In terms of the ability to predict the house prices in the U.S., the GSVI for "Real Estate Agent" outperforms that for "Housing Bubble" and "Index12". The

[2] Our ranking of all of the US bubble states is presented in Appendix A

GSVI for “Real Estate Agent” shows the highest correlation with the “House Price Index (HPI)”, especially pertaining to the non-bubble period. This correlation is the highest when lagged values are used for the Google searches, thus implying that the GSVI for “Real Estate Agent” is leading the changes in the house prices. Furthermore, we find that the two-time series are cointegrated, and observe a long-term effect that runs from the GSVI for “Real Estate Agent” to the HPI. This effect is the strongest in the states that were experiencing a real bubble, followed by those experiencing a minor bubble, and finally the non-bubble states. The GSVIs for “Real Estate Agent” show good in-sample predictive abilities at the state level, using simple linear models that include only the GSVI, and lead the house prices during both the bubble and non-bubble periods. We also find that including the GSVI for “Real Estate Agent” in our error correction model (ECM) for house prices improved with respect to all of the evaluation criteria compared to the ECM, while the well-established consumer confidence index (CCI) provides the least accurate results against all of the criteria. The results are valid for the *real bubble*, *minor bubble* and *non-bubble* states, as well as for the remaining thirty U.S. states that are not defined as either bubble or non-bubble states.

Based on the aforementioned results, we conclude that the GSVI for “Housing Bubble” can be a strong housing bubble indicator, while the GSVI for “Real Estate Agent” can predict housing trends and should thus be included in price models to improve their predictive accuracy at the state level.

The remainder of the paper is organized as follows. First, we present our data in Section 2, while the empirical approach adopted in this work is described in Section 3. Section 4 is designated for the main findings, and the paper concludes with Section 5.

2. Data

2.1 House Prices

We use the all-transactions house price index (HPI) published by the FHFA on a quarterly basis as a housing market indicator. The all-transactions HPI is a broad measure of the development of house prices for each geographic area (i.e., state or district). The prices are estimated by using repeated observations of the market value of individual single-family residential properties on which at least two mortgages are originated and subsequently purchased by either Freddie Mac or Fannie Mae. The data are adjusted for seasonality and inflation. We use house prices at the state level.

2.2 Google Search Volume Indices

Google publishes Google search volume data that date back to Q1 2004 at www.google.com/trend. The data are presented as GSVIs, which range from 0 to 100, where 100 equals the point in time where the use of a specific search term peaks in relative terms. Thus, this reflects the moment when the relative interest in a search term is the highest. All of the other GSVI values denote individual levels relative to the maximum. High values indicate that interest for the search term is high, while low values indicate little interest in the search term. An important aspect of the construction of the GSVI is that the total number of searches at any given moment must be above the threshold set by Google for the GSVI to be published. We have not been able to find the exact threshold. Nevertheless, we find that it is reasonable to interpret GSVI=0 as very low interest in the search term if the specific state has a relatively high population and disregards the result if the population in that state is relatively low. The indices are adjusted for the total use of Google search. We use Google search data at the state level.

2.3 Other Exponential Variables

The remaining exponential data are obtained from the DataStream database and presented in Table 1. The data, where applicable, are adjusted for seasonality effects by using the centered moving average (CMA) method, and for inflation, the consumer price index (CPI).

Table 1 Exponential Variables

#	Variable Name	Abbreviation	Available at	Data adjusted for
1	Housing Price Index	$HPI_{s,t}$	State Level	Seasonality & Inflation
2	Disposable Personal Income	DPI_t	Country Level	Seasonality & Inflation
3	Housing Permits Authorised	$HPA_{s,t}$	State Level	Seasonality effects
4	Unemployment Rate	$UR_{s,t}$	State Level	Seasonality effects
5	Interest Rate	IR	Country Level	
6	Population	$PO_{s,t}$	State Level	Dummy of Population
7	Google Search Volume Index	$GSVI_{w,s,t}$	State Level	Seasonality effects
8	Consumer Confidence Index	CCI_t	Country Level	Seasonality effects

Note: The eight variables used in the different ECMs.

3. Empirical Approach

3.1 Bubble Identification and Ranking

There are many definitions of a "bubble", and most of them are normative, i.e., Palgrave (1926), Flood and Hodrick (1990), Stiglitz (1990), Shiller (2005) and Cochrane (2010). To implement more formal testing, we adopt the descriptive definition of a bubble given by Lind (2009) and Oust and Hrafnkelsson (2017), which is a dramatic price increase quickly followed by a dramatic fall in prices. We first use the algorithm in Harding and Pagan (2002) to identify housing price peaks and troughs in different U.S. states, following Bracke (2013)[3]. We then use the peak with the highest value and the corresponding date (quarter/year) in the calculations to identify the housing price three and five years before the peak and calculate the changes. Next, we find the point with the lowest housing price value after the peak and use the point to calculate the price decline, as per the bubble definition.

We then identify the bubble states and rank all of the states by the total price decrease. As the aim is to compare bubble states to non-bubble states, we include the same number of non-bubble states as the number of identified bubble states, as these serve as the benchmark states. The non-bubble states are associated with the smallest price reduction (see Appendix A). Conversely, Nevada, Arizona, Florida, and California are defined as real bubble states (Table 2), based on price increase. Berkovec et al. (2012) conclude that California, Arizona, Florida and Nevada are the four states with the highest bubble level. They also argue that these four states are the ones most closely associated with the housing bubble. To compare the effects in the states that have experienced a real housing bubble with those that only experienced a large correction, we further add six states (Michigan, Rhode Island, Maryland, Idaho, Oregon, and Washington) based on their total price fall, as well as ten states that have experienced the smallest correction in house prices during the 2006−2007 housing bubble.

3.2 Selection of Search Terms

In order to establish if the Google search data can serve as a housing bubble indicator, it is necessary to first identify potential search terms that might indicate public interest in housing as an asset class. However, we exclude local search terms, such as the name of a real estate agent company, as well as those that appear to be time-specific. We test 204 search terms, which are provided in Appendix B. We download Google search data for our ten bubble states and our 10 non-bubble states. In addition to testing single search terms, we construct

[3] There are several methods that can be used to identify asset bubbles (Gürkaynak 2008). Our choice is mainly based on our interest in the peaks and troughs in our further analyses. Bracke (2013) uses this method to identify house price cycles and Oust and Hrafnkelsson (2017) use the methods to identify house price bubbles.

indices based on the three, six, twelve and twenty best-performing search terms, henceforth denoted as Index3, Index6, Index12, and Index20, respectively.

Table 2 Bubble and Non-Bubble States

Real bubble state	Minor bubble state	Benchmark state (non-bubble state)	
Nevada	Maryland	Kansas	Texas
Arizona	Oregon	Nebraska	Iowa
Florida	Washington	Wyoming	South Dakota
California	New Jersey	Louisiana	Oklahoma
	Virginia	Alaska	North Dakota
	Connecticut		

Note: The states are grouped into: real bubble states (RBS), minor bubble states (MBS) and non-bubble states (our benchmark group). Full state list can be found in Appendix A.

3.3 Are the GSVIs Leading, Coincident or Lagging the HPI?

To look at how the different GSVIs co-vary with the house prices, we use the same method to identify peaks and troughs for the GSVIs as for the house prices (Harding and Pagan, 2002 and Bracke, 2013). We find peaks and troughs for the GSVIs in our four RBSs, six minor bubble states (MBS) and ten control states or the non-bubble states (NBS).

3.4 Testing GSVI as a Housing Bubble Indicator (Red Flag Test)

To determine whether GSVIs can be used as a housing bubble indicator, we propose a red flag test based on differences in search volume levels during the housing bubble period as compared to the subsequent period. We use the subsequent period as our baseline due to data availability, as GSVIs were not available prior to 2004, and thus no data exist for the pre-bubble period.[4]

The period for the housing bubble is defined as follows:

- BP = Q1.2004 to Q4.2008

Similarly, the following period serves as a proxy for the non-bubble period[5]:

- NBP = Q1.2009 to Q3.2016

In the tests (Figure 1), we use the average of the *GSVI* in the non-bubble period as a benchmark[6]. If the *GSVI* is above M times the average level for the non-

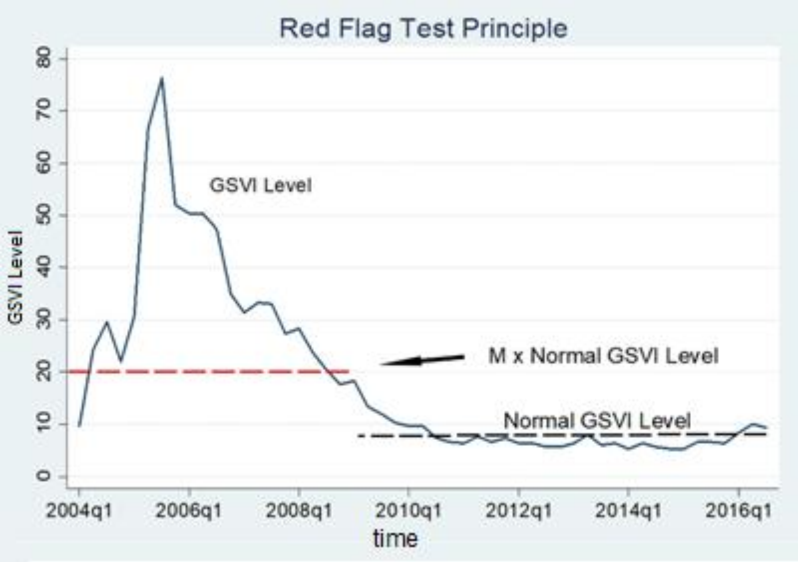
[4] When using GSVIs as a housing bubble indicator in the future, we would use a period of stable search volumes prior to the period to be tested as the normal GSVI period.

[5] A study conducted by Chen et al. (2012) indicates that the crisis was easing in 2009.

[6] Due to data availability, the benchmark period is after the bubble period. For future bubbles, the benchmark period will be before the focus period.

bubble period, it is flagged. The *GSVI* should ideally flag a bubble in all bubble states, but not raise the flag for the non-bubble states. The test names, with short descriptions, are given below, while their general principle is shown in Figure 1. Here, it is not the length of the bubble that is important, but rather the concept that if abnormal search activity associated with a bubble is included in the normal period for the search data, the indicator would appear to be weaker than in reality.

Figure 1 Red Flag Test Principle



Note: The black line (Normal GSVI Level) represents the average value of the GSVI during the normal period, which is defined to run from Q1 2009 to Q3 2016. The red line (M x Normal GSVI Level) represents M times the average level during the bubble period.

The general description of the test is as follows: the "1 in a row test" checks if the *GSVI* is M times higher than normal in at least one quarter; "2 in a row test" checks if the *GSVI* is M times higher than normal in at least two quarters, and so on and so forth. We test with multiples $M = [1.25, 1.5, 1.75, 2, 2.25, 2.5, 2.75, 3, 3.5, 5, 7, \text{and } 10]$. The test increases in stringency by either increasing M or the required number of subsequent periods with high *GSVI* levels.

We rank the performance of the different *GSVI*s for the specific search terms and indices based on two types of error:

- Type I-error: the *GSVI* does not flag a bubble state as a bubble, and
- Type II-error: the *GSVI* flags a non-bubble state as a bubble.

Type I-errors can be deemed as “sub-errors”, as the *GSVI* does not flag a real bubble. If the *GSVI* is not able to detect a real bubble state, this is more problematic than if the *GSVI* does not detect a minor bubble. Based on these rules, we generate a point system, whereby three points are given for detecting a real bubble state, and one point is given for detecting a minor bubble state, while three points are deducted for wrongly detecting a non-bubble state as a bubble state. The maximum amount of points that a search term or an index of search terms might obtain is 18 or 12 points to detect the four real bubble states, respectively, and 6 points to detect all six of the minor bubble states. We conduct four tests and rank the search terms according to the total points scored.

3.5 Johansen Test

Based on the results produced by the red flag test (Table 3), we analyze the causality between the two best-performing GSVIs and the best performing index; namely “Housing Bubble”, “Real Estate Agent”, and Index12, respectively, and the house price. We use the Dickey-Fuller generalized least squares (DF GLS) method at level form and with one lag. We find stationarity with one lag across the U.S. The full test results, including those at the state level and the explanatory variables in the house price model, are presented in Appendix D.

Next, we test for cointegration among the variables by using the Johansen test method, and find that there are one or more cointegrating relationships in all 50 U.S. states at a 5% level of significance (see Appendix E for the full test results).

3.6 Testing for Short- and Long-Term Effects of GSVI

According to Wooldridge (2012), when two variables y_t and x_t are both $I(1)$ and cointegrated, a linear regression of the HPI can be run with the variables in levels and the results interpreted as long-term effects. Therefore, we run the regression on first differenced variables, including the error term from the previous model, and create an ECM. Now, we can interpret the results yielded by the ECM as short-term effects, whereby the coefficient of the error term, also known as the error correction term, denotes the speed of adjustment.

Combining the use of an ordinary least squares (OLS) regression on variables in levels with the ECM to test for both the short- and long-term relationships between HPI and GSVI has several advantages, compared to, for e.g., vector error correction models (VECMs). First, the results obtained by using this method are easier to interpret, especially when a model includes several variables with more than one cointegrating relationship. This becomes increasingly problematic when testing for short- and long-term causalities in the three baseline models (Appendix F), for each of the 50 states, as these models include seven variables. Secondly, VECMs demand the same number of lags on all variables. This is not suitable when only testing the effect from GSVIs with different lags with respect to house prices.

The general regression model used to explain the long-term effect of the GSVIs for “Housing Bubble” and “Real Estate Agent” on the HPI are shown in Eq. (1). Since the GSVI for “Real Estate Agent” in the tests is identified to have the best performance in previous tests, we only test this variable at the state level ($\beta_i = 0$ for the variables that are not included in the specific test). The general regression model that is used to determine the short-term effect of the GSVI for “Housing Bubble”, “Real Estate Agent”, and Index12, on the HPI and the speed of adjustment is shown in Eq. (2) ($\beta_i = 0$ for the variables that are not included in the specific test).

$$\begin{aligned} HPI_t = \alpha + \beta_1 HPI_{t-1} + \beta_2 GSVI_{HB,t} + \ \beta_3 GSVI_{HB,t-2} \\ + \beta_4 GSVI_{REA,t} + \beta_5 GSVI_{REA,t-2} \\ + \beta_6 GSVI_{Index12,t} + \ \beta_7 GSVI_{Index12,t-2} \end{aligned} \tag{1}$$

and

$$\begin{aligned} \Delta HPI_t = \alpha + \beta_1 \Delta HPI_{t-1} + \beta_2 \Delta GSVI_{HB,t} + \ \beta_3 \Delta GSVI_{HB,t-2} \\ + \beta_4 \Delta GSVI_{REA,t} + \beta_5 \Delta GSVI_{REA,t-2} + \ \beta_6 \Delta GSVI_{Index12,t} \\ + \beta_7 \Delta GSVI_{Index12,t-2} + \gamma \epsilon_{HPI,t-1} \end{aligned} \tag{2}$$

where:

$HPI_{s,t}$ = the house price index for state s, at time t

$GSVI_{w,s,t}$ = Google search volume index for search term w, in state s, at time t

We start by regressing the housing prices by using only the GSVI for “Housing Bubble”, followed by the GSVI for “Real Estate Agent”, and finally, Index12. Regressing the house prices with only one variable provides a good indication of both its short- and long-term effects and explanatory power. Next, we regress the house prices by using the GSVI for “Housing Bubble” and apply different lags, and then we do the same for the GSVI for “Real Estate Agent” and Index12.

By including several lags in the independent variable, the aim is to ascertain whether this improves the in-sample prediction results of the model. After testing the GSVIs for the two search terms and Index12 independently, we include both of the search terms to find whether this might improve the result and, if so, by how much. This will give an indication of whether the two search terms capture different information and can thus improve the in-sample prediction results when used jointly. Finally, we include a one-period lag in the housing prices in the different regression models. We expect this to improve the model, in both the short and long term. The goal here is to examine how the explanatory power of the Google searches changes and whether the results concur with those yielded by single search terms/Index.

Regressing the house prices at the state level shows how Google search performs in the states that experienced a bubble compared to those that did not experience one. When moving from the country to the state level, the total volume of Google searches will be lower, likely reducing the data quality. Thus, we expect the GSVIs to have higher explanatory power for the housing prices in states with a large population compared to less-populated states. We start by regressing the house prices with only the GSVI for "Real Estate Agent". Next, we add different lags, and find that more than two lags rarely improve the model. Last, we regress the housing prices with a one-period lag in the house prices and the GSVI for "Real Estate Agent" without any lags. Due to the inclusion of a one-period lag in the dependent variable, we expect the last model to have better in-sample predictive ability. As the aim is to determine how this simple model performs compared to the baseline models, the mean absolute error (MAE) for both the $\overline{HPI}_{s,t}$ and $\Delta HPI_{s,t}$ is calculated.

3.7 Testing Whether GSVI for Real Estate Agent Improves the Baseline Housing Price Model

In determining the short- and long-term dynamics between the GSVIs for "Real Estate Agent" and the HPI, the aim is to establish whether Google searches can improve the baseline model. Due to the existence of cointegration, we first run a linear regression of the HPI with the variables in levels and interpret the results as long-term effects. Next, we run the regression on the first differenced variables, including the error term from the previous model, thus creating an ECM. We interpret the results from the ECM as short-term effects, and the coefficient of the error term as the speed of adjustment.

The general regression model used to model the long-term effect of the independent variables on the HPI is shown in Eq. (3), where $\beta_i = 0$ for the variables that are not included in the specific test. Similarly, the general ECM used to model the short-term effect of the independent variables on the HPI and the speed of adjustment is given in Eq. (4), where $\beta_i = 0$ for the variables that are not included in the specific test.

$$\begin{aligned} HPI_{s,t} = \alpha + \beta_1 HPI_{s,t-1} + \ \beta_2 UR_{s,t} + \beta_3 PO_{s,t} + \beta_4 DPI_t \\ + \beta_5 IR_t + \beta_6 HPA_{s,t} + \ \beta_7 GSVI_{REA,s,t} + \beta_8 CCI_t \end{aligned} \tag{3}$$

and

$$\begin{aligned} \Delta\overline{HPI}_{s,t} = \alpha + \beta_1 \Delta HPI_{s,t-1} + \ \beta_2 \Delta UR_{s,t} + \beta_3 \Delta PO_{s,t} \\ + \beta_4 \Delta DPI_t + \beta_5 \Delta IR_t + \beta_6 \Delta HPA_{s,t} \\ + \beta_7 \Delta GSVI_{REA,s,t} + \ \beta_8 \Delta CCI_t + \gamma \epsilon_{HPI,s,t-1} \end{aligned} \tag{4}$$

where

$HPI_{s,t}$ = The house price index for state s, at time t

DPI_t = disposable personal income at time t

$HPA_{s,t}$ = housing permits authorized for state s, at time t

$UR_{s,t}$ = unemployment rate for state s, at time t

IR_t = interest rate at time t

$PO_{s,t}$ = population in state s, at time t

β_i = corresponding coefficient for the respective variable

$GSVI_{w,s,t}$ = Google search volume index for search term w, in state s, at time t

CCI_t = consumer confidence index at time t

First, we regress the house prices without including either the GSVI or the CCI, setting β_7 and β_8 to zero which allows us to establish how the baseline ECM performs in both the short and long term in all 50 states. Then, we calculate the MAE of the in-sample prediction error of both the $\overline{HPI}_{s,t}$ and $\Delta HPI_{s,t}$ by using Eqs. (3) and (4). We include the GSVI for “Real Estate Agent” by removing the requirement of β_7 = 0, to test whether Google searches improve the baseline model. Finally, we substitute the GSVI with CCI, setting $\beta_7 = 0$ again and removing the requirement of β_8 = 0. Including CCI instead of the GSVI in the baseline model allows us to test the performance of the GSVI compared to a well-established indicator of consumer confidence. The three specific baseline models used to regress the house prices for each of the 50 states are given in Appendix F.

4. Results

4.1 Are the GSVIs Leading, Coincident or Lagging the HPI?

In this section, we examine how different Google search volume indices co-vary with house prices. Due to the large number of GSVIs and the number of states, we have chosen to limit the reporting of the results of the GSVIs for Housing Bubble and Real Estate Agent as well as Index12[7]. These are the three GSVIs that later show the best results. The correlations are reported in Appendix C. The results presented in Table 3 indicate that, on average, the GSVIs for both search terms and Index12 peak before the house prices do, for the real, minor, and non-bubble states. We further find that the GSVI for “Real Estate Agent” peaks before that of “Housing Bubble” and Index12 peaks for all three state groups and seems to have been leading during the bubble period. The GSVI for “Housing Bubble” is not published by Google in nine out of the ten non-bubble states, as the search volume levels are under the minimum threshold. We interpret the low search volume as lack of interest in the housing market and housing bubbles, which is understandable for states that did not experience a sharp increase in house prices. In addition, several of the non-bubble states

[7] It is likely that these two keywords and the index would be among the GSVIs with cycles most similar to the house prices when there is a housing bubble. Results available upon request.

have a relatively low population and are prone to low search volumes for specific queries, such as “Housing Bubble”, which diminishes the data quality.

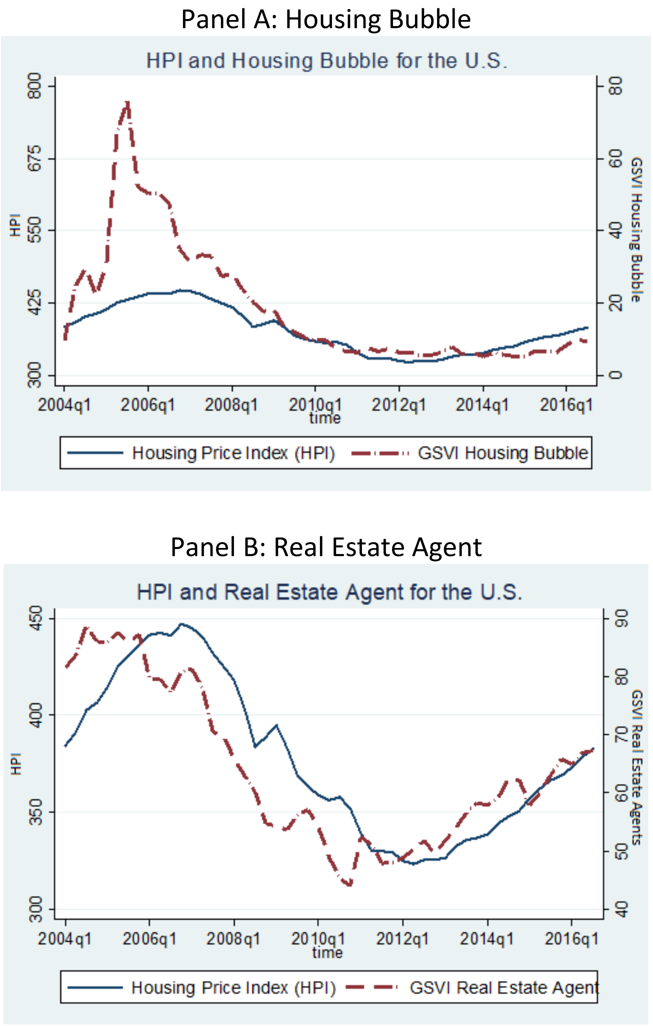
Table 3 House Prices and GSVI Peaks and Troughs

ΔTime	**Housing Bubble**		**Real Estate Agent**		**Index12**	
State	**ΔQ Peak**	**ΔQ Trough**	**ΔQ Peak**	**ΔQ Trough**	**ΔQ Peak**	**ΔQ Trough**
Nevada	1.00	-6.00	1.00	-6.00	2	-10
Arizona	5.00	1.00	9.00	-6.00	5	-13
Florida	5.00	-7.00	8.00	3.00	6	-2
California	3.00	-3.00	5.00	5.00	4	5
Average RBS	**3.50**	**-3.75**	**5.75**	**-1.00**	**4.25**	**-5**
Maryland	5.00	4.00	9.00	6.00	6	-7
Oregon	7.00	-14.00	11.00	-9.00	7	-14
Washington	4.00	-6.00	5.00	2.00	7	-6
New Jersey	5.00	1.00	11.00	8.00	5	-8
Connecticut	2.00	0.00	8.00	18.00	2	5
Virginia	5.00	-11.00	8.00	7.00	5	-11
Average MBS	**4.67**	**-4.33**	**8.67**	**5.33**	**5.3**	**-6.8**
Kansas	N/A	N/A	4.00	-8.00	5	-8
Nebraska	N/A	N/A	5.00	4.00	-1	-12
Wyoming	N/A	N/A	11.00	6.00	10	-15
Louisiana	N/A	N/A	12.00	2.00	6	-7
Alaska	N/A	N/A	11.00	12.00	4	-10
Texas	3.00	-6.00	10.00	5.00	8	-7
Iowa	N/A	N/A	1.00	-18.00	-1	-7
South Dakota	N/A	N/A	12.00	15.00	-2	-9
Oklahoma	N/A	N/A	12.00	-22.00	8	-25
North Dakota	N/A	N/A	9.00	-5.00	7	-25
Average NBS	**3.00**	**-6.00**	**8.70**	**-0.90**	**4.4**	**-12.5**

Note: The number of quarters, ΔQ, that the GSVI for Housing Bubble, Real Estate Agent and a self-created index (Index12) peak and trough before the HPI peaks and troughs for the real, minor and non-bubble states. A positive value of ΔQ indicates that the GSVI for the respective queries peaks/troughs before the HPI peaks/troughs and vice versa. N/A indicates that the GSVI data for the respective state are missing.

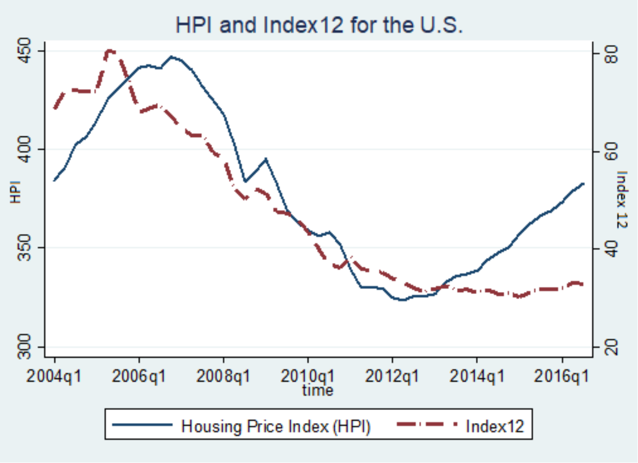
Note:2 shows that the GSVIs for the two search terms and Index12 behave rather differently. The search volume levels for “Housing Bubble” show a bubble in the U.S. housing market. The search term levels seem to be low, and devoid of any trends, both before and after the housing bubble. The graphs in Figure 2 show that the GSVI for “Housing Bubble” demonstrates a dramatic increase in search volume during the boom phase of the bubble, before declining sharply prior to the decline of the house prices. Both house prices and GSVI for “Housing Bubble” seem to reach a minimum in 2012; however, while house prices increase steadily each year, the GSVI for “Housing Bubble” remains at a low level. According to the graph in Figure 2, the search volume levels for “Housing Bubble” are highly correlated during the bubble periods, which decline during normal economic times. Due to the explosive increase in search volume levels during the bubble periods and its capacity to lead house prices, the GSVI for “Housing Bubble” could work as a strong indicator of a housing bubble at both the country and state levels.

Figure 2 House Prices and Keywords for the U.S.



Panel A: Housing Bubble

Panel B: Real Estate Agent

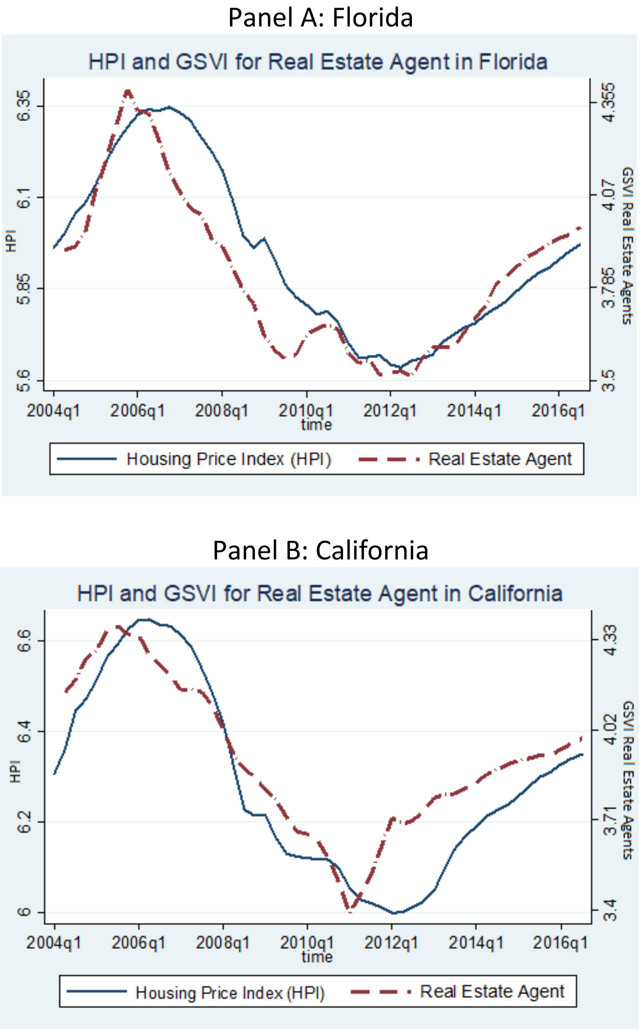


Note: Index12 consists of the average GSVIs for the twelve single best search terms, as determined by an in-sample prediction test.

The search volume levels for both "Real Estate Agent" and Index12 show a downward trend following the peak in 2005, thus indicating that housing prices would decline, but do not show the same explosive increase in search volume levels during the bubble period as is the case with "Housing Bubble". The search volume seems to be at a more normal level, increasing and decreasing before the HPI during the housing bubble. The GSVI for "Real Estate Agent" troughs in 2011, while the graph of the HPI flattens out in 2012. The graph that shows Index12 in Figure 2(c) does not reach minimum before 2015 and, while the Figures 2(a) and 2(b) start to increase annually from the trough, Index12 stays at a low level. Both the GSVI for "Real Estate Agent" and Index12 seem to be leading the house prices during the bubble period. "Real Estate Agent" also leads the house prices in the non-bubble period. Index12 would probably perform better as a house price predictor if it includes fewer typical bubble words.

Figures 3, 4, and 5 show the GSVI for "Real Estate Agent" against the housing prices for two of the real, minor, and non-bubble states. These graphs reveal how the fit between the time series changes in the different groups of states. For the states that are experiencing a real bubble, we find that the GSVI for "Real Estate Agent" fits the housing prices extremely well, which indicates a high correlation between the two-time series for the entire studied period. The graphs that pertain to the minor bubble states indicate that the two-time series are less correlated. For the non-bubble states, while the two-time series converge in the long term, they do not fit as closely as those for the real and minor bubble states. The tendency of higher correlation between Google searches and the housing prices for states that experienced a greater housing bubble is in accordance with the results reported in Appendix C. In general, the GSVI for “Real Estate Agent” leads the housing prices in all six states (Michigan, Rhode Island, Maryland, Idaho, Oregon, and Washington) during the bubble period, but the results are more similar in the non-bubble period.

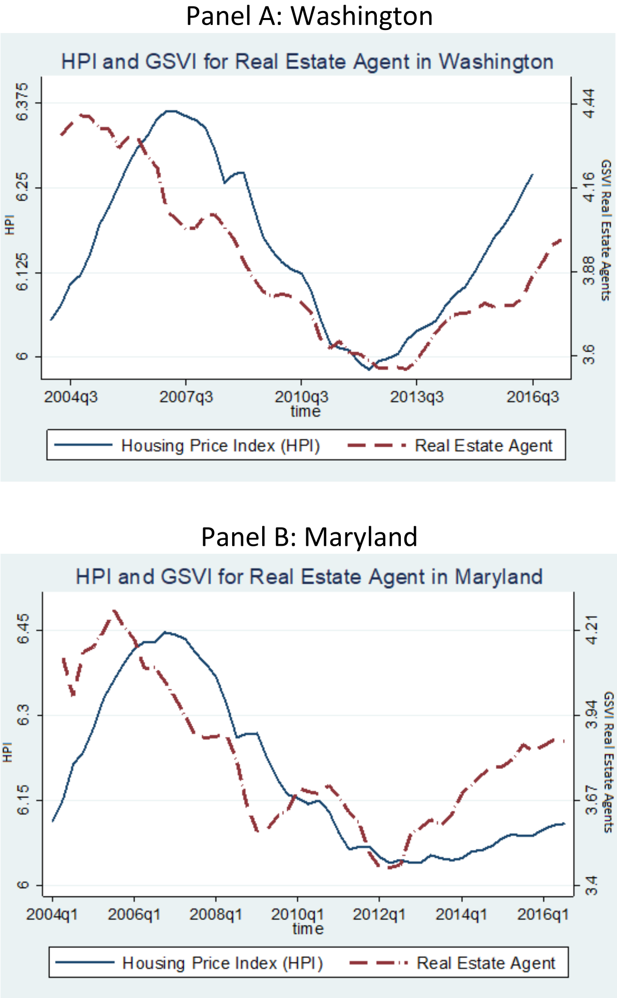
Figure 3 House Prices and GSVI for Real Estate Agent Real Bubble States


Panel A: Florida

Panel B: California

Note: The graphs for two of the states defined as real bubble states. Both time series are transformed to logarithmic form and adjusted for inflation and seasonal effects.

Figure 4 House Prices and GSVI for Real Estate Agent Minor Bubble States

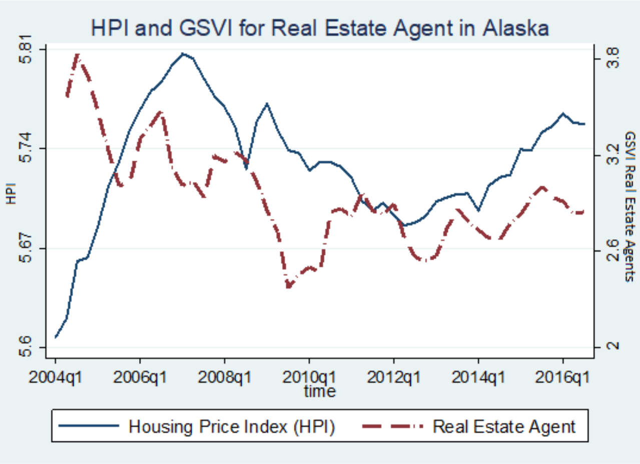

Panel A: Washington

Panel B: Maryland

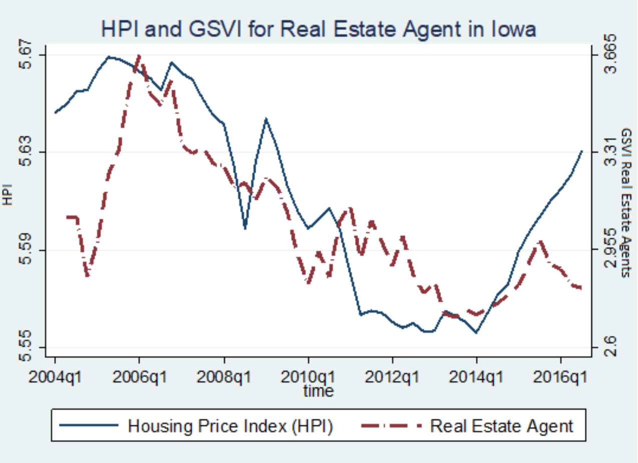
Note: The graphs for two of the states defined as minor bubble states. Both time series are transformed to logarithmic form and adjusted for inflation and seasonal effects.

Figure 5 House Prices and GSVI for Real Estate Agent Non-Bubble States

Panel A: Alaska



Panel B: Iowa



Note: The graphs for two of the states defined as non-bubble states. Both time series are transformed to logarithmic form and adjusted for inflation and seasonal effects.

Appendix C shows the correlations among the GSVIs for “Housing Bubble”, “Real Estate Agent”, and Index12 and the house prices in the bubble period (Q1 2004–Q2 2010), the normal period (Q3 2010–Q3 2016), and the entire period (Q1 2004–Q3 2016). The GSVIs for both search terms and Index12 show significantly higher correlations during the bubble period than in the non-bubble period. In general, the results show the highest correlation for "Real Estate Agent", followed by "Housing Bubble", for the entire period, bubble period, and non-bubble period. For the real and minor bubble states, Index12 shows an even higher correlation than Real Estate Agent during the bubble period at 91.3% and 74.7%, respectively. For the non-bubble period, Index12 shows a negative correlation with the housing prices for all three state groups.

The GSVI for Real Estate Agent shows the highest correlation in the real bubble states with an average correlation of 91%. In the states defined as minor bubble states, the average correlation is slightly lower at 83.4%, and reduced to even lower value of 55.6% in the non-bubble states. In general, the correlation is higher for lagged values of the Google searches for the three groups, thus indicating that the GSVI for "Real Estate Agent" is a good leading indicator for the HPI. At 81.6%, the GSVI for "Housing Bubble" shows a slightly higher correlation in the minor bubble states, compared to 78.8% for the real bubble states. The GSVI for "Housing Bubble", as previously noted, is not recorded/published by Google in nine out of the ten non-bubble states due to search volume levels that fall under a minimum threshold. In comparing the GSVI for "Housing Bubble" with that for "Real Estate Agent" and Index12, we find the former and latter require fewer lags to reach the highest correlation with the house prices.

4.2 The In-Sample Bubble Identification Test (Red Flag Test) Results

In Table 4, the rankings and results of the twenty single search terms (with the highest average correlation, see Appendix C) and four self-created indices are presented, based on their in-sample predictive ability to identify bubble states.

Table 4 shows the rankings and scores obtained from the four in-sample prediction tests based on identifying the states that experienced a bubble for the twenty single search terms and the four self-created indices. We rank the different search terms and indices; the maximum number of points that a search term can obtain based on our point system is eighteen points. We illustrate this process with an example, in which "Housing Bubble" receives sixteen points in all four tests for correctly including all four real bubble states, four out of six minor bubble states, and none of the non-bubble states.

From the results shown in Table 3, it is evident that the GSVIs for the two best-performing search terms, namely "Housing Bubble" and "Real Estate Agent", outperform the self-created indices. We create four different indices that consist of the average GSVI for the three, six, twelve and twenty single best-performing search terms to improve the robustness and the level of information captured. The results, however, indicate that this is not the case. Based on the full test results, it can be deduced that the top two single search terms, in addition to obtaining the highest test score, are robust to changes in the M-values. Taking predictive ability, robustness, and simplicity into account, the GSVIs for single search terms seem to be the most effective as housing bubble indicators. The search term "Housing Bubble" seems particularly suitable as a bubble indicator, as it has the best performance in all four tests. An advantage of using single queries—such as "Housing Bubble" and "Real Estate Agent"—compared to indices, is that they can be combined in order to increase the robustness and level of market information captured by the bubble indicator. Moreover, the GSVIs for single search terms are easier to download and calculate.

Table 4 Results of Red Flag Tests

Rank	Search Term	1 in a row	2 in a row	3 in a row	8 in a row	Total Results
1	Housing Bubble	16	16	16	16	64
2	Real Estate Agent	14	15	16	14	61
3	Real Estate	14	13	15	13	57
4	Housing Market	13	12	12	14	51
5	Realtors	10	10	13	17	50
6	Real Estate Listings	13	13	13	9	48
7	Mortgage	11	11	11	13	46
8	Investment	8	7	11	8	34
9	Real Estate Broker	9	9	9	6	33
10	Real Estate Bubble	8	8	8	8	32
11	Broker	4	5	14	8	31
12	Home Equity	3	3	10	8	24
13	Lending	5	6	7	4	22
14	Real Estate Investment	3	0	3	7	13
15	Property	6	3	1	0	10
16	Apartment	2	0	1	1	4
17	Construction	0	0	0	3	3
18	Bubble	1	0	0	0	1
19	Rent	1	0	0	0	1
20	Flat	0	0	0	0	0
Rank	**Average GSVI of the**	**1 in a row**	**2 in a row**	**3 in a row**	**8 in a row**	**Total Result**
1	Index12	15	15	15	12	57
2	Index6	15	13	13	14	55
3	Index20	15	12	12	13	52
4	Index3	12	12	12	11	47

Notes: The results of the four flag tests, "1, 2, 3 and 8 in a row", and the total result for each of the 20 search terms, in addition to 4 self-created indices. The search terms/indices are given 3 points for correctly indicating a real bubble state, and 1 point for correctly indicating a minor bubble state, while 3 points are deducted for wrongly indicating a non-bubble state as a bubble state. Total results represent the sum of scores obtained in the four tests, where "# in a row" flags a state as a bubble state if the GSVI for the specific search query is above M times the GSVI level during the non-bubble period for # consecutive quarters, where # = {1,2,3 *and* 8}.

4.3 ECM Results for the United States

Table 5 shows the results yielded by the regression of the house prices at level form for the assessment of the long-term effects from Google searches, as well as the results provided by the ECM, to assess the short-term effects and the speed of adjustment from Google searches for the entire U.S.

Table 5 Short and Long Term Effects from GSVI on House Prices for U.S.

Model Variable	Long Term Effects			Speed of Adjustment		Short Term Effects		
	LT C	**P>Z**	**LT R^2**	**SA C**	**P>Z**	**ST C**	**P>Z**	**ST R^2**
HB	0.120	0.000	0.804	0.019	0.645	0.073	0.000	0.314
REA	0.486	0.000	0.896	-0.294	0.000	0.180	0.139	0.567
Index12	0.265	0.000	0.752	-0.046	0.198	0.175	0.013	0.176
HB +	0.195	0.000	0.848	0.015	0.704	0.052	0.010	0.327
L2.HB	-0.079	0.000				0.042	0.004	
REA +	0.051	0.492	0.964	-0.298	0.005	0.252	0.012	0.593
L2.REA	0.453	0.000				0.228	0.002	
Index12 +	0.296	0.006	0.782	-0.026	0.437	0.186	0.05	0.238
L2.Index12	-0.016	0.873				0.129	0.013	
REA +	0.325	0.000	0.956	-0.190	0.026	0.279	0.010	0.516
HB	0.053	0.000				0.043	0.000	
L.HPI +	1.102	0.000	0.974	-0.261	0.371	0.692	0.056	0.442
HB	-0.017	0.004				0.038	0.036	
L.HPI +	0.711	0.000	0.988	-0.814	0.001	0.832	0.000	0.651
REA	0.156	0.000				0.162	0.103	
L.HPI +	0.928	0	0.973	-0.929	0.021	1.441	0.001	0.483
Index12	0.02	0.165				0.124	0.092	
L.HPI +	0.732	0.000	0.988	-0.972		0.991	0.000	0.656
REA +	0.153	0.000			0.001	0.158	0.109	
HB	-0.002	0.616				-0.022	0.150	

Notes: The results of the ECM regression of the HPI by using only the GSVI for Housing Bubble (HB), Real Estate Agent (REA), and a self-created index (Index12) that consist of the twelve best-performing search terms. L2 in front of a variable stands for a two-period lag of the respective variable, while LT R^2 denotes the long-term coefficient of determination, ST R^2 is the short-term coefficient of determination, SA C represents the coefficient for the speed of adjustment, and P>Z is the probability that the respective coefficient is statistically significant.

As can be seen from the results reported in Table 5, the GSVI for "Real Estate Agent" performs significantly better than both "Housing Bubble" and Index12, at all points in both the short- and long-term, for all of the models, when applied on the entire U.S. data sample. Only the models that include the GSVI for "Real Estate Agent" have significant values for the speed of adjustment, which suggests the presence of cointegration and long-term effect running from the

GSVI for only “Real Estate Agent” to the HPI. Index12 shows some signs of a long-term relationship, but this is not significant at the 10% level. The GSVI for “Housing Bubble” is not cointegrated with the HPI and thus does not explain the house price levels in the long term. As “Housing Bubble” is not an everyday search term, we expect the corresponding search volume levels to be relatively low outside the bubble phases, as outlined by Kindleberger and Aliber (2005). Therefore, it is not surprising that the GSVI for “Housing Bubble” and the house prices are not cointegrated. The same argument applies to Index12, albeit to a lesser extent.

In the short term, both the GSVI for Housing Bubble and Index12 show explanatory power that pertains to the house prices. When the GSVIs for both “Housing Bubble” and “Real Estate Agent” are combined, similar results to those produced by using only the GSVI for “Real Estate Agent” are obtained. Substituting “Housing Bubble” with a two-period lag in “Real Estate Agent”, however, yields improved results. This indicates that inclusion of the GSVI for “Housing Bubble” does not capture more of the market information than “Real Estate Agent” does when used in isolation.

“Real Estate Agent” shows good predictive results, which explain for the house prices in both short and long term. We also see that the speed of adjustment is relatively high for all of the models. When only the GSVI for “Real Estate Agent”, without any lags, is used to explain the house prices, the long-term coefficient is 48.6%, and the long-term coefficient of determinations (R^2) is 89.6%. In addition, we obtain a speed of adjustment of -29.4%, short-term coefficient of 18%, and short-term coefficient of determination of 56.7%. The r-squared values are high for both the short- and long-term effects. The speed of adjustment of 29.4% indicates that, in every period/quarter, the error correction term will move by 29.4% towards the long-term equilibrium between the GSVI for “Real Estate Agent” and the HPI. Given that lags in the dependent variable are not included in the model, this confirms the explanatory power of GSVI for “Real Estate Agent” on the HPI. When a two-period lag in the GSVI for “Real Estate Agent” is included, the short- and long-term coefficients of determination increase to 59.3% and 96.4% respectively, while the speed of adjustment remains similar. Substituting the two-period lag in GSVI with a one-period lag in the independent variable HPI produces some marked changes. The short- and long-term coefficients of determination increase to 65.1% and 98.8% respectively, whereby the one-period lag in HPI explains most of the changes in both the short and long term. Nonetheless, the GSVI for “Real Estate Agent” is statistically significant, with short- and long-term coefficients of 15.6% and 16.2%, respectively. The greatest change is observed in the speed of adjustment, which has increased from -29.8% to -81.4%. These results show that even simple linear models, including only the GSVI and a one-period lagged variable of HPI, can explain for the changes in the U.S. house prices.

4.4 ECM Results for all 50 U.S. States Using only Google Searches

Table 6 presents the results yielded by the regression of the house prices at level form for the assessment of the long-term effects of the GSVI for “Real Estate Agent”, along with the results of the ECM, to assess the short-term effects and the speed of adjustment in Google searches for each of the 50 U.S. states.

Table 6 Linear Regression of HPI Using Only Google Searches: Long Term Effects

Model Variable	L1.HPI	P>Z	GSVI	P>Z	L2.GSVI	P>Z	R^2
Average Results for the Real Bubble States							
Only GSVI			0.734	0.000			0.709
GSVI + L2.GSVI			0.622	0.005	0.198	0.325	0.822
L1.HPI + GSVI	0.836	0.00			0.162	0.001	0.985
Average Results for the Minor Bubble States							
Only GSVI			0.347	0.000			0.345
GSVI + L2.GSVI			0.54	0.089	-0.125	0.325	0.522
L1.HPI + GSVI	0.931	0.00			0.062	0.065	0.978
Average Results for the 30 states not defined as Bubble States or Non-Bubble States							
Only GSVI			0.278	0.029			0.496
GSVI + L2.GSVI			0.652	0.143	0.136	0.243	0.611
L1.HPI + GSVI	0.916	0.00			0.037	0.118	0.971
Average Results for the Non-Bubble States							
Only GSVI			0.059	0.141			0.245
GSVI + L2.GSVI			-0.002	0.346	0.071	0.298	0.241
L1.HPI + GSVI	0.967	0.00			0.003	0.384	0.932

Note: The long-term results of the ECM of the HPI using only the GSVI for Housing Bubble (HB) and Real Estate Agent (REA). L2 in front of a variable stands for a two period lag of the respective variable. LT R^2 is the long-term coefficient of determinations. LT MAE is the MAE between the predicted value and the real value of HPI at level form.

From the results reported in Tables 6 and 7, it can be deduced that the model that uses only the GSVI for “Real Estate” to regress the house prices produces good in-sample predictive results. For the states that experienced a real bubble, the average long-term coefficient is 73.4% and statistically significant, and the average long-term coefficient of determination is 70.9%. The average short-term coefficient is 17.6% and significant at the 10% confidence interval, and the average short-term coefficient of determination is 36.3%. The speed of adjustment is -15.6%. Upon a closer look at the full results, we find the in-sample prediction results to be significantly better for California and Florida than Nevada and Arizona (these results can be obtained upon request). The short-term coefficients of determination are respectively 57.3% and 50.8% for the former two, and 15.2% and 21.9% for the latter two, respectively.

Including a two-period lag in the GSVI for "Real Estate Agent" increases the long-term coefficient of determination to 82.2%, while decreasing the short-term coefficient of determination and speed of adjustment to respectively 34.3% and -10.1%. Substituting the two-period lag with a one-period lag in the dependent variable HPI creates more pronounced changes.

Table 7 Short-Term Results for ECM Real Estate Agent

Model Variable	SA C	P>Z	L1	P>Z	GSVI	P>Z	L2	P>Z	R^2
			HPI				GSVI		
Average Results for the Real Bubble States									
Only GSVI	-0.16	0.003			0.176	0.094			0.36
GSVI + L2.GSVI	-0.10	0.065			0.201	0.106	0.17	0.158	0.34
L1.HPI + GSVI	-0.58	0.009	1.074	0.000			0.12	0.050	0.71
Average Result for the Minor Bubble States									
Only GSVI	-0.08	0.068			0.003	0.515			0.17
GSVI + L2.GSVI	-0.07	0.185			0.062	0.402	0.03	0.344	0.17
L1.HPI + GSVI	-0.69	0.047	1.126	0.004			0.02	0.382	0.53
Average Results for the 30 states not defined as Bubble States or Non-Bubble States									
Only GSVI	-0.09	0.123			0.045	0.319			0.18
GSVI + L2.GSVI	-0.09	0.135			0.043	0.356	0.04	0.268	0.19
L1.HPI + GSVI	-0.84	0.088	1.127	0.018			0.05	0.335	0.38
Average Results for the Non-Bubble States									
Only GSVI	-0.04	0.334			0.015	0.472			0.08
GSVI + L2.GSVI	-0.04	0.352			0.013	0.46	0.02	0.495	0.11
L1.HPI + GSVI	-0.96	0.159	0.928	0.055			0.01	0.538	0.22

Note: The short-term results of an ECM of the HPI by using only the GSVI for Real Estate Agent (REA). L2 in front of a variable stands for a two-period lag in the respective variable. ST R^2 is the short-term coefficient of determination. ST MAE is the MAE between predicted change in HPI and real value. SA C is the coefficient for speed of adjustment and P>Z is the probability that the respective coefficient is significant.

The short- and long-term coefficients of determination increase to respectively 71.4% and 98.5% while the speed of adjustment increases to -58.1%. The same findings apply to all three groups, which are the real, minor, and non-bubble states.

When the results that pertain to the other state groups as shown in Tables 6 and 7 are examined, it is evident that the coefficients of determination for both the short- and long-term are the highest for the real bubble states and the lowest for the non-bubble states. For the minor bubble states and the thirty states not defined as either bubble or non-bubble states, the results are reversed. This observation might be explained by the fact that the housing bubble affected the entire U.S. housing market. Moreover, the size of the population in each state is likely to affect the quality of the corresponding Google trend data.

As a part of this investigation, we also construct a vector ECM (VECM) to investigate the relationship between Google searches and housing prices at the state level. Due to the rigidity of the model and problems with the interpretation of the results of the baseline models, which captured several long-term relationships, we test other models. Nevertheless, it is noteworthy that the results yielded by the VECM coincide with those presented above.

4.5 ECM Results for all 50 States Using the Baseline Variables

In this section, the results yielded by the baseline model with and without the inclusion of the Google searches are examined and compared. To elucidate the utility of using Google search volume in a model to estimate the housing prices, we compare the baseline model not only with a model that includes the Google search volume, but also one that includes the CCI. The CCI is a well-known and widely used leading indicator and should be a good benchmark.

The results reported in Table 8 indicate that all criteria against which the models are assessed are improved when the GSVI for “Real Estate Agent” is included in the baseline model. The adjusted coefficient of determination is increased for both the short- and long-term, and the speed of adjustment is higher as well as more significant. These results apply to the real, minor, and non-bubble states, as well as the thirty states that are not defined as either a bubble or non-bubble state.

For the real bubble states, including the GSVI for Real Estate Agent, the MAE is reduced on average, by 0.007% and 0.054% for the short- and long-term in-sample predictions, respectively. For the minor bubble states, the MAE is reduced by 0.045% and 0.032%. As for the thirty states that are not defined as either a bubble or non-bubble state, these improvements are 0.023% and 0.027%, respectively, while the improvements are 0.023% and 0.008% for the non-bubble states, respectively.

Substituting the Google searches with the CCI yields a significantly inferior performance in all of the aforementioned criteria. The only exception is the short-term MAE for the non-bubble states, which is reduced by 0.0015% on average. Including the CCI in the baseline model improves the MAE in both short- and long-term, but the coefficient of determination and speed of adjustment both decline. Based on these results, we conclude that the GSVI for “Real Estate Agent” improves the fit of the baseline model and reduces the MAE of the in-sample prediction in both the short- and long-term. In addition, the inclusion of the GSVI for “Real Estate Agent” yields significantly better results than the inclusion of CCI.

Table 8 Model Comparison

Model Description	LT R^2	LT MAE	ST R^2	ST MAE	SA C	P>Z
Average results for the Real Bubble States						
Baseline Model	0.992	1.494%	0.816	1.153%	-0.616	0.006
Baseline GSVI Model	0.993	1.440%	0.834	1.146%	-0.664	0.002
Baseline CCI Model	0.992	1.492%	0.824	1.182%	-0.594	0.008
Average results for the Minor Bubble States						
Baseline Model	0.987	1.017%	0.739	0.847%	-0.695	0.004
Baseline GSVI Model	0.988	0.972%	0.760	0.815%	-0.734	0.002
Baseline CCI Model	0.987	1.014%	0.749	0.833%	-0.697	0.002
Average results of the Thirty States not Defined as either Bubble or Non-bubble						
Baseline Model	0.979	0.879%	0.634	0.772%	-0.782	0.003
Baseline GSVI Model	0.980	0.852%	0.660	0.749%	-0.789	0.001
Baseline CCI Model	0.979	0.865%	0.648	0.753%	-0.754	0.007
Average results for the Non-Bubble States						
Baseline Model	0.944	0.715%	0.488	0.661%	-0.858	0.009
Baseline GSVI Model	0.943	0.707%	0.503	0.653%	-0.891	0.007
Baseline CCI Model	0.943	0.712%	0.499	0.652%	-0.856	0.010

Note: The three different versions of a baseline housing price model with Disposable Personal Income, Housing Permits Authorized, Unemployment Rate, Interest Rate and Population as explanatory variables. A one-period lag in the dependent variable is also included. The “Baseline Model” includes the former variables, “Baseline Model Including GSVI” includes the GSVI for Real Estate Agent in addition to the other variables, and “Baseline Model Including CCI” includes the CCI instead of Google searches. The three models are assessed in terms of the following criteria: LT R^2 - the adjusted long run coefficient of determinations, LT MAE - the MAE between predicted and real values for HPI at level forms, ST R^2 - the adjusted short-term coefficient of determination, SR MAE - the MAE between predicted change in HPI and real value, SA C - the coefficient for speed of adjustment, and P>Z - the probability that the coefficient is statistically significant.

As described in the previous section, we also construct a VECM with all the baseline variables. We include the GSVI for “Real Estate Agent” and Index12, separately, for all 50 states. Our findings concur with those provided above. In comparing the results of the model with only the GSVI for “Real Estate Agent” and a one-period lag in the dependent variable with the baseline model, we find the latter to perform better. The former model, however, shows a higher speed of adjustment for the thirty states not defined as either a bubble or non-bubble state, as well as the non-bubble states. Moreover, the long-term coefficient of determination results coincide with the slightly better results for the baseline model. The major difference in performance is seen in the short term, where the baseline model produces a better fit. Nonetheless, the in-sample prediction results of such a simple model are rather good.

5. Conclusion

The aim of this work is to test whether GSVIs can be used to identify bubbles in the housing market. We analyze data that pertain to the 2006-2007 U.S. housing bubble, taking advantage of the heterogeneous development of house prices in different U.S. states with both bubble and non-bubble states. Google publishes search volume data that date back to Q1 2004 at www.google.com/trend. We collect the Google search volume data for 204 housing-related keywords and test both indices with single search terms and indices that comprise search term sets to see whether they can be used as housing bubble indicators.

Taking predictive ability, simplicity, and robustness into consideration, we conclude that the GSVI for “Housing Bubble” is the best candidate as a housing bubble indicator. Optimizing the model so that the model can detect all of the states that have experienced a bubble, the GSVI for “Housing Bubble” erroneously included only one non-bubble state. Conversely, when optimized for not wrongly including any non-bubble states, the model detected all four *real* bubble states and four out of six *minor* bubble states. The model repeatedly produced consistent findings in a wide variety of tests. For the states that have experienced a housing bubble, the GSVI for “Housing Bubble” shows relatively low search volume levels, without any trends both before or after the bubble. However, during the actual bubble period, the search volume levels increased substantially by more than twofolds. The search volume levels for “Housing Bubble” across the entire U.S. show the same characteristics, which led the house prices and strongly suggest a real estate bubble. The extreme characteristics of the GSVI for “Housing Bubble” during a bubble period suggest that there is no need to adjust the data for either seasonal affects or trends, thus simplifying the surveillance of the indicator.

The GSVI for “Real Estate Agent” is found to show the highest correlation with the HPI and obtains the best in-sample predictive results of the house prices in both the short and long term. In addition, the GSVI for “Real Estate Agent” and the HPI are cointegrated in 45 out of 50 states, and the former leads the house prices in both the bubble and the non-bubble periods. When testing the relationship between the GSVI for “Real Estate Agent” and the HPI, we find both short- and long-term effects that run from the former to the latter. These effects are significant in all of the states, regardless of the magnitude of the housing bubble. When a simple linear model that incorporates only the GSVI for “Real Estate Agent” and a one-period lag in the dependent variable is applied, the HPI produces good in-sample predictive results. The model fit and the MAE results are the best for the states that have experienced a *real* bubble, followed by the states that have experienced a *minor* bubble, and finally those that have not experienced any bubble. When the same model is applied to the entire U.S., the model provides even better results than for the states that have experienced a *real* housing bubble.

Including the GSVI for “Real Estate Agent” in our baseline ECM for the house prices improves the performance against all of the criteria. The adjusted coefficient of determination is increased in both short and long term, and the speed of adjustment is more rapid and significant. Substituting the Google searches with the well-established CCI produces inferior results across all of the criteria. These results are valid for the *real*, *minor*, and non-bubble states, as well as the thirty states that are not defined as either a bubble nor non-bubble state.

Based on the results reported in this paper, we conclude that the GSVI for “Housing Bubble” can be a strong housing bubble indicator as a part of a bubble indicator system, while the GSVI for “Real Estate Agent” can predict housing trends and should be included in price models to improve their predictive abilities at the state level. Both policymakers and investors might benefit from this finding.

We also conclude that Google search data have great potential to become housing bubble indicators, with the ability to predict whether a dramatic price increase will be quickly followed by a dramatic fall in the house prices. Google search data might be a part of a bubble indicator system to predict housing price bubbles which would greatly benefit policy makers and investors.

References

Benjamin, J., Chinloy, P. and Jud, D. (2004), Real Estate versus Financial Wealth in Consumption, *Journal of Real Estate Finance and Economics*, 29, 341-354.

Berkovec, J., Chang, Y. and McManus D.A. (2012), Alternative Lending Channels and the Crisis in U.S. Housing Markets, *Real Estate Economics*, 40(1), S8-S31.

Bijl, L., Kringhaug, G., Molnár, P., & Sandvik, K. (2016), Google searches and stock returns, *International Review of Financial Analysis*, 45, 150–156.

Bracke, P. (2013), How Long do Housing Cycles Last? A Duration Analysis for 19 OECD countries, *Journal of Housing Economics*, 22(3), 213-230.

Brunnermeier, M.K. and Oehmke, M. (2012), Bubbles, Financial Crises, and Systemic Risk, National Bureau of Economic Research, Working Paper No. 18398.

Campbell, J. and Cocco, J. (2004), How Do Housing Price Affect Consumption? Evidence from Micro Data, Harvard Institute of Economic Research, Discussion Paper No. 2045.

Case, K.E., Quigley, J. and Shiller R. (2001), Comparing Wealth Effects: The Stock Market Versus the Housing Market, National Bureau of Economic Research, Working Paper No. 8606.

Case, K. E. and Shiller, R.J. (2004), Is there a Bubble in the Housing Market? Cowless Foundation, Paper No. 1089.

Challet, D., and Ayed, A.B.H. (2013). Predicting Financial Markets with Google Trends and not so Random Keywords. arXiv preprint arXiv:1307.4643

Chen, B., Schoeni, R. and Stafford, F. (2012), Mortgage Distress and Financial Liquidity: How U.S. Families are Handling Savings, Mortgages and Other Debts, Technical Series Paper 12-02, Survey Research Center, Institute for Social Research, University of Michigan.

Choi, H., and Varian, H.R. (2012), Predicting the Present with Google Trends, *Economic Record*, 88(1), 2-9.

Cochrane, J.H. (2010), Discount Rates, Working paper, University of Chicago, Booth School of Business, and NBER, Chicago, Illinois, 27 December.

Ettredge, M., Gerdes, J. and Karuga, G. (2005), Using Web-Based Search Data to Predict Macroeconomic Statistics, *Communications of the ACM*, 48(11), 87-92.

Ewing, J. (2010). Shiller's List: How to Diagnose the Next Bubble. The New York Times. January 27. 2010, https://dealbook.nytimes.com/2010/01/27/schillers-list-how-to-diagnose-the-next-bubble/

Flood, R.P. and Hodrick, R.J. (1990), On Testing for Speculative Bubbles, *Journal of Economic Perspectives*, 4(2), 85–101.

Goel, S., Hofman, J.M., Lahaie, S., Pennock, D.M. and Watts D.J. (2010), Predicting Consumer Behavior with Web Search, *Proceedings of the National Academy of Sciences*, 7(41), 17486-17490.

Gürkaynak, R.S. (2008), Econometric Tests of Asset Price Bubbles: Taking Stock, *Journal of Economic Surveys*, 22, 166–86.

Harding, D. and Pagan, A. (2002), Dissecting the Cycle: A Methodological Investigation, *Journal of Monetary Economics*, 49(2), 365-381.

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

Johansen, S. (1991). Estimation and Hypothesis Testing of Cointegration Vectors in Gaussian Vector Autoregressive Models, *Econometrica*. 59(6), 1551–1580.

Kindleberger, C. and Aliber, R. (2005), Manias, Panics and Crashes –A History of Financial Crises, John Wiley and Sons, Hoboken, NJ.

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

Lind, H. (2009), Price Bubbles in Housing Markets: Concept, Theory and Indicators, *International Journal of Housing Markets and Analysis*, 2(1), 78-90.

Oust, A. and Hrafnkelsson, K. (2017), What is a Housing Bubble? *Economics Bulletin*, 37(2), 806-836.

Palgrave, R.H.I. (1926), Palgrave's Dictionary of Political Economy, MacMillan & Co., London, England, p. 181.

Pentland, A.S. (2010), Honest Signals, MIT Press, Cambridge, MA.

Preis, T., Moat, H.S. and Stanley, H.E. (2013), Quantifying Trading Behavior in Financial Markets Using Google Trends, *Scientific Reports*, 3, Article number 1684.

Preis, T., Reith, D. and Stanley, H.E. (2010), Complex Dynamics of Our Economic Life on Different Scales: Insights from Search Engine Query Data, *Philosophical Transactions of the Royal Society of London*, 368(1933), 5707-5719.

Shiller, R.J. (2005). Irrational Exuberance. 3nd. New Jersey: Princeton University Press. ISBN 0-691- 12335-7.

Stiglitz, J.E. (1990). Symposium on Bubbles. *Journal of Economic Perspectives* 4(2), 13-18.

Wooldridge, J.M. (2012). Introductory Econometrics: A Modern Approach. Mason, Ohio: South-Western Cengage Learning.

Wu, L. and Brynjolfsson, E. (2015), The Future of Prediction How Google Searches Foreshadow Housing Prices and Sales, In *NBER book Economic Analysis of the Digital Economy*, edited by Avi Goldfarb, Shane M. Greenstein, and Catherine E. Tucker, pp. 89-118.

Appendices

Table A: Fifty States in United States Sorted After their Total Price Fall

Rank	State	3 years	5 years	Top HPI	Peak	Bottom	Trough	Price Fall
1	Nevada	65.1%	79.5%	491.2	Q1 2006	191.4	Q2 2012	-61.0%
2	Arizona	55.2%	68.8%	506.2	Q4 2006	247.4	Q3 2011	-51.1%
3	Florida	50.2%	78.4%	570.9	Q4 2006	280.4	Q2 2012	-50.9%
4	California	56.2%	84.9%	770.1	Q2 2006	402.7	Q1 2012	-47.7%
5	Michigan	3.0%	9.3%	394.5	Q2 2005	240.1	Q2 2012	-39.1%
6	Rhode Island	36.5%	72.5%	726.0	Q1 2006	448.2	Q4 2013	-38.3%
7	Maryland	42.5%	72.1%	630.2	Q4 2006	420.4	Q1 2013	-33.3%
8	Idaho	35.3%	40.0%	398.5	Q1 2007	266.7	Q2 2011	-33.1%
9	Oregon	34.2%	45.1%	533.6	Q2 2007	357.7	Q2 2012	-33.0%
10	Washington	36.1%	43.7%	580.0	Q1 2007	396.1	Q2 2012	-31.7%
11	Georgia	6.1%	9.8%	382.7	Q4 2006	262.0	Q2 2012	-31.5%
12	New Jersey	26.3%	53.2%	682.8	Q4 2006	469.2	Q4 2013	-31.3%
13	New Hampshire	21.0%	44.0%	561.8	Q1 2006	388.8	Q1 2013	-30.8%
14	Minnesota	14.8%	30.2%	442.7	Q1 2006	306.4	Q2 2012	-30.8%
15	Connecticut	25.6%	43.1%	560.8	Q1 2006	389.8	Q1 2014	-30.5%
16	Illinois	11.9%	21.5%	440.0	Q4 2006	306.1	Q1 2013	-30.4%
17	Delaware	28.7%	47.6%	591.9	Q4 2006	420.2	Q1 2014	-29.0%
18	Massachusetts	24.4%	50.2%	880.5	Q2 2005	628.6	Q4 2012	-28.6%
19	Ohio	2.8%	7.1%	328.2	Q2 2005	241.4	Q1 2014	-26.4%
20	Hawaii	46.5%	78.9%	631.3	Q1 2007	466.4	Q2 2012	-26.1%
21	Virginia	34.9%	54.9%	552.1	Q4 2006	408.0	Q2 2012	-26.1%
22	New Mexico	26.6%	34.1%	382.6	Q1 2007	288.6	Q1 2014	-24.6%
23	Utah	30.4%	30.2%	439.9	Q3 2007	333.4	Q4 2003	-24.2%
24	New York	19.8%	42.0%	760.4	Q4 2006	577.9	Q1 2014	-24.0%
25	Maine	15.6%	34.9%	600.6	Q4 2006	458.3	Q1 2014	-23.7%
26	Wisconsin	12.7%	18.6%	387.0	Q1 2006	297.3	Q1 2014	-23.2%
27	Missouri	6.7%	13.5%	351.3	Q4 2006	275.8	Q1 2014	-21.5%
28	South Carolina	12.7%	15.4%	395.0	Q4 2006	310.5	Q1 2014	-21.4%
29	North Carolina	10.2%	12.7%	387.6	Q2 2007	310.0	Q4 2013	-20.0%
30	Alabama	11.5%	14.5%	349.1	Q2 2007	280.7	Q4 2013	-19.6%
31	Mississippi	11.9%	13.4%	301.6	Q1 2007	243.7	Q4 2013	-19.2%
32	Pennsylvania	19.3%	32.0%	463.4	Q4 2006	375.7	Q1 2014	-18.9%
33	Indiana	1.5%	4.8%	306.5	Q2 2005	249.8	Q1 2014	-18.5%
34	Colorado	14.2%	21.7%	427.9	Q4 2006	349.5	Q1 2012	-18.3%
35	Vermont	23.5%	40.5%	533.9	Q4 2006	440.3	Q1 2014	-17.5%
36	Tennessee	9.7%	12.8%	350.6	Q2 2007	292.1	Q1 2013	-16.7%
37	Montana	20.8%	35.8%	431.5	Q3 2007	363.1	Q2 2012	-15.9%
38	Arkansas	9.3%	13.6%	299.2	Q1 2007	252.1	Q2 2012	-15.7%
39	West Virginia	13.0%	16.8%	259.5	Q4 2006	219.0	Q1 2013	-15.6%
40	Kentucky	3.3%	6.6%	340.4	Q4 2006	292.2	Q1 2014	-14.1%
41	Kansas	2.6%	6.3%	280.6	Q4 2006	241.9	Q1 2014	-13.8%
42	Nebraska	4.7%	7.4%	302.5	Q2 2005	262.2	Q4 2012	-13.3%
43	Wyoming	24.2%	38.6%	323.8	Q3 2007	281.9	Q1 2012	-13.0%
44	Louisiana	15.2%	21.1%	284.2	Q1 2007	251.4	Q1 2013	-11.5%
45	Alaska	22.1%	31.5%	332.6	Q1 2007	294.7	Q2 2012	-11.4%
46	Texas	6.0%	10.0%	257.5	Q2 2007	232.7	Q1 2012	-9.6%
47	Iowa	5.4%	9.6%	289.8	Q2 2005	270.0	Q3 2008	-6.8%
48	South Dakota	3.9%	6.8%	331.1	Q1 2007	309.1	Q3 2012	-6.6%
49	Oklahoma	2.5%	5.2%	231.8	Q1 2007	222.6	Q3 2008	-4.0%
50	North Dakota	12.4%	19.4%	280.0	Q1 2007	271.6	Q3 2008	-3.0%

Notes: The table shows the fifty states sorted according to their price fall from the peak to the trough “3 years” and “5 years” are the percentage price increases in the last three and five years before the price topped out in each respective state. “Top HPI” and “Bottom” are the highest and lowest values for the HPI in each respective state. “Peak” and “Trough” are the quarter and year for the highest and lowest values of HPI. “Price Fall” is the percentage price fall from peak to trough in each respective state.

Appendix B: List of Alphabetically Sorted Search Terms

A	Acres, Acres of Land, Affordable Housing, Analyst
B	Backyard, Beach Front, Broker, Bubble, Building a House, Building Cost, Buying Out
C	CBS Constructed Homes, Consumer Loans, Consumer Credit, Consumer Lending, Condos, Credit
D	Debt, Disposable Income, Down Payment, Duplex Home, Dwelling, Dwellings
E	Equity, Equity Requirement
F	Financial, Financial Analysis, First Time Homebuyer, Future Interest
G	Gated Communities, GDP
H	Home Equity, Home Equity Loan, Homes in up and Coming Communities, House Analysis,
I	Income, Income Change, Income Increase, Income Raise, Increasing Property Prices Increasing Real Estate Prices, Inflation, Installments, Interest Forecast, Interest, Interest Rate
L	Land Price, Land Prices, Leasing, Lending, Lending Standard, Low Down Payment, Low
M	Middle Class Homes, Mortgage, Mortgage Payment, Mortgage Requirements
N	Net Immigration, New Buildings, Newly Renovated, Number of Completed Homes
O	One Story Home, Overpriced, Overvaluation
P	Part Payment, Patio, Peak, Pet Approval, Pool, Pricing, Property Bubble, Property, Property Investment, Property Tax, Property Under Construction, Population
R	Raising Property, Real Estate, Real Estate Advisor, Real Estate Agent, Real Estate Bubble, Real Estate Broker, Realtor, Real Estate Listings
S	Salary Increase, Salary Change, Salary Raise, School District, Second Mortgage
T	Turmoil, Two Storey Home, Two Storey House
U	Unemployment, Unemployment Rate
V	Vacation House, Valuation
W	Wage, Wages, Wage Increase, Wage raise, Waterfront Property
Z	Zero Interest Rate

Note: The table presents the 204 search terms that were originally tested, and sorted alphabetically.

Appendix C: Correlation between GSVIs and House Prices

Correlation	**Housing Bubble - HPI**			**Real Estate Agent - HPI**			**Index12 - HPI**		
State	**WP**	**BP**	**NP**	**WP**	**BP**	**NP**	**WP**	**BP**	**NP**
Nevada	0.486	0.301	-0.11	0.874	0.794	0.326	0.78	0.923	-0.695
Arizona	0.855	0.701	0.397	0.846	0.902	-0.048	0.73	0.886	-0.718
Florida	0.887	0.776	0.478	0.957	0.955	0.955	0.84	0.922	-0.489
California	0.925	0.938	0.793	0.963	0.968	0.898	0.78	0.920	-0.465
Ave RBS	**0.788**	**0.679**	**0.390**	**0.910**	**0.905**	**0.533**	**0.78**	**0.913**	**-0.592**
Maryland	0.919	0.638	0.633	0.940	0.820	0.736	0.88	0.787	0.182
Oregon	0.697	0.308	-0.22	0.620	0.118	-0.624	0.67	0.750	-0.540
Washington	0.766	0.385	0.617	0.817	0.573	0.433	0.64	0.497	-0.416
New Jersey	0.939	0.686	0.407	0.884	0.746	0.576	0.89	0.797	0.478
Virginia	0.854	0.479	-0.41	0.860	0.921	0.721	0.81	0.834	-0.585
Connecticut	0.723	0.577	0.366	0.880	0.873	-0.285	0.91	0.815	0.722
Ave MBS	**0.816**	**0.512**	**0.231**	**0.833**	**0.675**	**0.260**	**0.80**	**0.747**	**-0.027**
Kansas	N/A	N/A	N/A	0.752	0.729	-0.012	0.66	0.436	-0.296
Nebraska	N/A	N/A	N/A	0.705	0.658	0.444	0.44	0.232	-0.507
Wyoming	N/A	N/A	N/A	0.544	0.647	0.493	0.19	0.231	-0.505
Louisiana	N/A	N/A	N/A	0.675	0.532	-0.140	0.54	0.321	-0.263
Alaska	N/A	N/A	N/A	0.628	0.332	0.141	0.34	0.152	-0.409
Texas	0.045	0.114	0.464	0.271	0.060	0.848	0.25	0.073	-0.576
Iowa	N/A	N/A	N/A	0.753	0.591	0.178	0.62	0.175	-0.212
South Dakota	N/A	N/A	N/A	0.360	0.198	0.350	-0.37	0.123	-0.695
Oklahoma	N/A	N/A	N/A	0.563	0.468	-0.488	0.36	-0.14	-0.402
North Dakota	N/A	N/A	N/A	0.307	-0.09	0.616	-0.23	0.338	-0.726
Average NBS	**N/A**	**N/A**	**N/A**	**0.556**	**0.412**	**0.243**	**0.28**	**0.202**	**-0.459**

Note: The table shows the correlation between: GSVI for Housing Bubble and the HPI, GSVI for Real Estate Agent and HPI, GSVI for Index12 and HPI. The correlation is shown for the whole period (WP), Q1 2004 – Q3 2016, the bubble period (BP), Q1 2004 – Q2 2010, and the normal period (NP), Q3 2010 – Q3 2016. The correlation is calculated for the states defined as real bubble states (RBS), minor bubble states (MBS) and non-bubble states (NBS). Also, the average for each of the three groups is calculated. N/A means there are missing GSVI data for the respective state.

Appendix D

Table D.1.1: Stationarity Test of the Variables at Level Form for U.S.

Country General Variable	ln CCI	ln IR	ln DPI
t-statistics	-1.493	-0.843	-1.234

Note: The table shows the results from the DF GLS unit root test of the following time series at level form. The natural logarithm to CCI, 1 + Interest Rate in percentage (IR) and Disposable Personal Income (DPI). The three time-series are all general for the United States.

Table D.1.2: Stationarity Test of the Variables at Level Form for all 50 States

State Specific Variable	ln HPI	ln UR	ln DPO	ln HPA	ln GSVI
State	t-statistics	t-statistics	t-statistics	t-statistics	t-statistics
Nevada	-1.4	-3.1 *	-2.3	-1.4	-1.2
Arizona	-1.5	-3.9 ***	-2.8	-1.1	-0.9
Florida	-1.3	-2.9	-2	-1	-0.9
California	-1.7	-3.2 **	-2.2	-0.9	-1.4
Maryland	-1.3	-2.3	-2.1	-1.7	-1.4
Idaho	-1.4	-1.3	-1.7	-1.9	-1.7
Oregon	-1.4	-1.3	-1.7	-1.9	-1.4
Washington	-1.4	-3.3 **	-2.1	-1.8	-1.0
Hawaii	-1.3	-2.7	-2.2	-1.5	-5.2 ***
Virginia	-1.3	-2.1	-2.2	-1.2	-1.6
Rhode Island	-0.9	-2.6	-2.5	-1.5	-1.9
Michigan	-0.8	-2.2	-1.6	-1.7	-2.4 **
Georgia	-1.0	-1.8	-1.2	-1.0	-0.6
New Jersey	-1.1	-2.6	-2.1	-1.5	-1.0
New Hampshire	-0.8	-2.5	-2.1	-2.2 *	-4.5 ***
Minnesota	-0.9	-2.2	-3.7 **	-2.4 **	-1.3
Connecticut	-0.5	-2.1	-2.7	-1.8	-1.5
Illinois	-0.7	-1.9	-3.3 **	-1.2	-1.5
Delaware	-1.0	-2.7	-2.7	-1.6	-0.5
Massachusetts	-1.0	-2.8	-2.2	-1.6	-1.6
Ohio	-0.6	-2.1	-2.5	-1.8	-1.1
New Mexico	-1.1	-2.8	-2.5	-1.0	-1.0
Utah	-1.5	-2.1	-3.8 ***	-1.9	-1.9
New York	-1.1	-2.6	-2.8	-2.3 **	-1.6
Maine	-1.0	-2.3	-2.8	-2.5 **	-2.1 *
Wisconsin	-0.8	-2.5	-2.4	-2.3 **	-0.8
Missouri	-0.9	-2.8	-1.8	-1.4	-0.9
South Carolina	-1.3	-2.1	-3.2 **	-1.4	-1.2
Alabama	-1.0	-2.5	-3.2 **	-0.9	-1.0
Mississippi	-0.9	-1.6	-3.4 **	-1.3	-2.5 **
Pennsylvania	-1.2	-2.1	-2.4	-1.7	-2.2 *
Indiana	-0.7	-1.9	-3.0 *	-1.9	-2.0 *

(***Continued...***)

(Table D.1.2 Continued)

State Specific Variable	ln HPI	ln UR	ln DPO	ln HPA	ln GSVI
State	t-statistics	t-statistics	t-statistics	t-statistics	t-statistics
Colorado	-0.6	-2.5	-3.6 **	-1.4	-1.7
Vermont	-1.0	-1.9	-4.2 ***	-3.2 ***	-2.3 **
Tennessee	-1.2	-2.1	-1.7	-1.1	-1.1
Montana	-0.8	-2.1	-3.0 *	-2.9 ***	-4.4 ***
Arkansas	-1.1	-2.2	-3.1 *	-1.9	-1.6
West Virginia	-1.1	-2.3	-3.0 *	-1.8	-3.4 ***
Kentucky	-1.0	-2.1	-3.4 **	-1.6	-2.4 **
Kansas	-1.0	-2.5	-2.8	-1.8	-1.6
Nebraska	-1.1	-2.4	-2.2	-3.1 ***	-2.0 *
Wyoming	-0.8	-2.7	-2.6	-3.8 ***	-1.8
Louisiana	-1.2	-2.7	-2.4	-1.8	-1.3
Alaska	-0.9	-2.5	-2.7	-3.9 ***	-2.5 **
Texas	0.0	-2.6	-2.2	-1.6	-1.0
Iowa	-1.2	-2.8	-2.2	-3.4 ***	-3.2 ***
South Dakota	-0.3	-2.5	-2.0	-4.9 ***	-2.1 *
Oklahoma	-1.3	-3.5 ***	-2.6	-1.7	-0.7
North Dakota	-1.4	-2.6	-2.6	-3.0 ***	-5.1 ***

Note: The table shows the result from the DF-GLS unit root test of the following time series at level form. The natural logarithm to the HPI, 1+Unemployment Rate in percentage (UR), Housing Permits Authorized (HPA), Population (PO) and GSVI. The five time-series are state specific for each of the 50 states. The DF-GLS tests are performed with the no trend option for all variables except Unemployment Rate.

Table D.2.1: Stationarity Test of the First Differenced Variables for U.S

Country General Variable	Δ ln CCI	Δ ln IR	Δ ln DPI
t-statistics	-5.333 ***	-3.557 ***	-4.968 ***

Note: The table shows the results from the DF-GLS unit root test of the following first differenced time series. The natural logarithm to CCI, 1 + Interest Rate in percentage (IR) and Disposable Personal Income (DPI). The three time-series are general for the U.S. The tests are performed with the no trend option for all variables.

Table D.2.2: Stationarity Test of the First Differenced Variables for all 50 States

State Specific Variable	Δ ln HPI	Δ ln UR	Δ ln DPO	Δ ln HPA	Δ ln GSVI
State	t-statistics	t-statistics	t-statistics	t-statistics	t-statistics
Nevada	-1.8	-1.6	-5.1 ***	-6.5 ***	-8.4 ***
Arizona	-2 *	-2 *	-5.6 ***	-4.2 ***	-5.5 ***
Florida	-2.1 *	-1.9 *	-5.0 ***	-3.2 ***	-5.1 ***
California	-1.9 *	-1.6	-2.6 ***	-3.4 ***	-4.9 ***
Maryland	-2.1 *	-2.7 ***	-2.4 ***	-5.8 ***	-3.8 ***
Idaho	-2.9 ***	-2.4 **	-5.0 ***	-5.1 ***	-3.9 ***
Oregon	-2.9 ***	-2.4 **	-5.0 ***	-5.1 ***	-8.0 ***
Washington	-2.3 **	-2.8 ***	-2.5 ***	-5.0 ***	-4.9 ***
Hawaii	-1.9 *	-2.6 ***	-2.4 ***	-6.6 ***	-9.3 ***
Virginia	-2.6 ***	-2.7 ***	-2.4 ***	-5.5 ***	-4.6 ***
Rhode Island	-2.5 **	-2.7 ***	-2.72	-5.0 ***	-6.0 ***
Michigan	-4.4 ***	-3.6 ***	-5.0 ***	-5.9 ***	-1.2
Georgia	-4.0 ***	-2.1 *	-5.2 ***	-6.7 ***	-3.5 ***
New Jersey	-2.6 ***	-2.8 ***	-3.1 ***	-4.6 ***	-1.6
New Hampshire	-3.1 ***	-3.6 ***	-4.6 ***	-5.5 ***	-2.8 ***
Minnesota	-4.6 ***	-2.9 ***	-3.5 ***	-3.9 ***	-6.6 ***
Connecticut	-2.9 ***	-2.5 ***	-3.5 ***	-4.7 ***	-4.1 ***
Illinois	-3.2 ***	-3.1 ***	-4.9 ***	-4.4 ***	-3.7 ***
Delaware	-2.7 ***	-3.0 *	-4.9 ***	-8.3 ***	-0.8
Massachusetts	-3.1 ***	-2.2 *	-2.9 *	-3.0 ***	-3.4 ***
Ohio	-5.4 ***	-3.0 ***	-2.9 *	-4.7 ***	-5.9 ***
New Mexico	-2.9 ***	-3.9 ***	-4.7 ***	-6.4 ***	-7.9 ***
Utah	-3.1 ***	-2.9 ***	-6.6 ***	-5.1 ***	-4.5 ***
New York	-3.1 ***	-2.6 **	-3.2 **	-3.9 ***	-1.4
Maine	-3.0 ***	-3.6 ***	-5.0 ***	-3.8 ***	-2 *
Wisconsin	-3.7 ***	-3.6 ***	-4.7 ***	-5.1 ***	-2.3 **
Missouri	-4.3 ***	-2.4 **	-5.0 ***	-5.6 ***	-4.5 ***
South Carolina	-4.1 ***	-3.2 ***	-2.9 *	-5.2 ***	-2.5 **
Alabama	-4.1 ***	-3.3 ***	-2.9 *	-8.0 ***	-4.0 ***
Mississippi	-4.1 ***	-3.3 ***	-4.9 ***	-9.1 ***	-7.3 ***
Pennsylvania	-3.2 ***	-3.0 ***	-3.0 *	-5.2 ***	-4.5 ***
Indiana	-5.9 ***	-3.3 ***	-2.9 ***	-4.9 ***	-2.2 ***
Colorado	-3.6 ***	-1.8	-3.0 *	-4.8 ***	-7.5 ***

(***Continued...***)

(Table D.2.2 Continued)

State Specific Variable	Δ ln HPI	Δ ln UR	Δ ln DPO	Δ ln HPA	Δ ln GSVI
State	t-statistics	t-statistics	t-statistics	t-statistics	t-statistics
Vermont	-2.9 ***	-4.0 ***	-6.5 ***	-4.4 ***	-7.3 ***
Tennessee	-4.0 ***	-3.8 ***	-5.0 ***	-5.7 ***	-6.1 ***
Montana	-3.0 ***	-2.9 ***	-3.5 ***	-4.6 ***	-5.9 ***
Arkansas	-3.6 ***	-3.8 ***	-7.2 ***	-4.5 ***	-4.6 ***
West Virginia	-4.4 ***	-4.5 ***	-3.3 **	-6.2 ***	-1.9
Kentucky	-5.1 ***	-3.3 ***	-4.9 ***	-7.3 ***	-6.6 ***
Kansas	-4.9 ***	-2.6 ***	-3.5 **	-5.6 ***	-9.7 ***
Nebraska	-4.6 ***	-3.0 ***	-2.4	-3.9 ***	-8.4 ***
Wyoming	-2.6 ***	-3.6 ***	-2.9 *	-5.9 ***	-8.2 ***
Louisiana	-4.0 ***	-5.7 ***	-3.5 ***	-4.1 ***	-4.6 ***
Alaska	-3.4 ***	-3.2 ***	-3.7 ***	-4.2 ***	-3.9 ***
Texas	-3.5 ***	-2.6 ***	-2.6	-4.8 ***	-4.8 ***
Iowa	-5.0 ***	-3.2 ***	-2.4	-4.4 ***	-10.8 ***
South Dakota	-4.2 ***	-3.6 ***	-2.4	-9.3 ***	-4.3 ***
Oklahoma	-5.0 ***	-3.5 ***	-3.2 **	-7.5 ***	-2.0 *
North Dakota	-4.3 ***	-4.9 ***	-2.7	-3.7 ***	-10.0 ***

Note: The table shows the results from the DF-GLS unit root test of the first difference of the following time series. The tests are performed with the no trend option for all variables. The natural logarithm to HPI, 1+Unemployment Rate in percentage (UR). Housing Permits Authorized (HPA), Population (PO) and GSVI. The five time-series are state specific for each of the fifty states. The DF-GLS tests are performed with the no trend option for all variables except Unemployment Rate.

Appendix E

Table E.1: Test of Cointegration among all Variables for all 50 States

	Johansen Cointegration Test		
Maximum Rank	**Standard Model 5% Critical Value**	**CCI Model 5% Critical Value**	**GSVI Model 5% Critical Value**
0	94.15	124.24	124.24
1	68.52	94.15	94.15
2	47.21	68.52	68.52
3	29.68	47.21	47.21
4	15.41	29.68	29.68
5	3.76	15.41	15.41
6		3.76	3.76

State	**No of CR**	**Trace Statistics**	**No of CR**	**Trace Statistics**	**No of CR**	**Trace Statistics**
Nevada	2	40.3627	3	43.7198	3	41.3831
Arizona	3	23.9541	3	45.1838	4	23.6582
Florida	3	28.8519	3	44.5998	4	28.8787
California	3	25.6202	5	10.3180	5	10.3595
Maryland	2	45.2226	4	24.8237	3	43.2934
Idaho	3	16.3943	3	41.7381	4	16.1368
Oregon	3	16.3943	3	16.3943	4	16.1368
Washington	3	17.6006	4	15.6217	3	40.8641
Hawaii	3	22.5517	3	45.0867	4	22.5375
Virginia	3	23.9379	4	24.3824	3	34.4096
Rhode Island *	3	24.4648	4	19.0211	4	22.7024
Michigan	2	31.6738	3	39.7405	3	31.0343
Georgia	2	43.4937	3	44.6179	3	40.4820
New Jersey	3	23.9767	4	18.7800	4	21.8624
New Hampshire *	2	35.5841	3	41.7453	3	35.5357
Minnesota	2	42.2711	3	33.3372	3	45.0350
Connecticut	2	46.2505	3	46.1145	3	44.0273
Illinois	2	32.5926	3	31.6180	3	28.6395
Delaware	2	37.3162	2	64.4077	3	40.5506
Massachusetts	2	41.3113	3	39.4561	3	29.6629
Ohio	1	66.0071	2	60.5332	2	53.4324
New Mexico	2	40.6647	3	41.9667	3	36.4998
Utah	2	45.7809	3	43.7825	4	22.8082
New York	2	44.4772	3	31.2727	3	43.4184
Maine *	2	39.8359	3	30.2518	3	39.7884
Wisconsin	2	35.5670	3	36.8299	2	61.2751
Missouri	2	42.3427	2	67.6795	3	37.6629
South Carolina	3	22.2126	3	44.4371	4	22.9054
Alabama	2	42.3577	2	67.0684	2	67.5585
Mississippi	2	41.8859	2	62.5591	3	39.9526
Pennsylvania	3	25.7204	4	25.0029	4	25.4845
Indiana	2	41.9070	3	45.1118	3	43.1139
Colorado	2	31.3916	3	38.1430	3	31.0629
Vermont	3	22.4250	3	40.7536	4	22.2068

(***Continued…***)

(Table E.1 Continued)

State	No of CR	Trace Statistics	No of CR	Trace Statistics	No of CR	Trace Statistics
Tennessee	1	68.1641	3	37.7474	2	61.8224
Montana	4	8.8941	4	27.6188	5	8.7579
Arkansas	3	23.3949	3	45.6645	2	63.6456
West Virginia	2	45.6584	3	47.1587	3	42.4379
Kentucky	1	61.5666	1	93.5387	1	93.6420
Kansas	2	45.8627	3	42.8021	3	43.0793
Nebraska	2	26.7690	2	63.7105	2	63.6124
Wyoming	3	20.0871	4	22.6715	4	19.8079
Louisiana	1	56.0549	1	91.7284	2	53.4080
Alaska	2	39.9938	3	42.5329	2	67.8711
Texas	2	33.3084	3	32.9837	2	66.0308
Iowa	2	30.7638	3	32.0965	2	65.6164
South Dakota	1	67.6594	3	42.0215	2	65.7291
Oklahoma *	1	43.2096	2	38.1231	2	38.9640
North Dakota	1	60.7251	1	88.9476	2	62.0380

Notes: The table shows the result from the cointegration test implemented by vecrank in Stata, which is based on Johansen`s method. The test checks if there is one or more cointegrating relationships among variables in the three models Standard, CCI and GSVI. The Standard Model consist of the variables HPI, Unemployment Rate (UR), Interest Rate (IR), Housing Permits Authorized, Population (PO) and Disposable Personal Income (DPI). The CCI model includes the same variables in addition to the CCI. The GSVI model includes the same variables as the Standard Model in addition to the GSVI. All three models are tested with only one lag. The null hypothesis is that there are Maximum Rank (0, 1, 2, …., n-1, where n is number of variables in the model) cointegrating relationships among variables. * Indicates collinearity in the model in the specific state. The Stata function noreduce has been used on these models. Noreduce does not perform checks and corrections for collinearity among lags of dependent variables.

Table E.2: Test of Cointegration among Housing Price Index and Google Search Volume Index for Real Estate Agent in all 50 States

Johansen Cointegration Test of Housing Price Index and Google Search Volume Index					
State	No of	Trace	5% Critical	Max	5% Critical
	CE	Statistics	Value	Statistics	Value
Nevada	1	0.0661 **	3.76	0.0661	3.76
Arizona	1	0.8579 **	3.76	0.8579	3.76
Florida	1	1.3191 **	3.76	1.3191	3.76
California	1	1.2571 **	3.76	1.2571	3.76
Maryland	1	1.2694 **	3.76	1.2694	3.76
Idaho	1	0.6684 **	3.76	0.6684	3.76
Oregon	0	13.1988	15.41	13.1988	15.41
Washington	1	1.1240 **	3.76	1.1240	3.76
Hawaii	1	3.5763 **	3.76	3.5763	3.76
Virginia	1	2.0193 **	3.76	2.0193	3.76
Rhode Island	1	0.5224 **	3.76	0.5224	3.76
Michigan	1	2.5222 **	3.76	2.5222	3.76
Georgia	1	1.2616 **	3.76	1.2616	3.76
New Jersey	1	0.3889 **	3.76	0.3889	3.76
New Hampshire	1	0.4260 **	3.76	0.4260	3.76
Minnesota	1	1.0042 **	3.76	1.0042	3.76
Connecticut	1	0.0656 **	3.76	0.0656	3.76
Illinois	1	0.8502 **	3.76	0.8502	3.76
Delaware	1	0.1084 **	3.76	0.1084	3.76
Massachusetts	1	1.0299 **	3.76	1.0299	3.76
Ohio	1	3.0872 **	3.76	3.0872	3.76
New Mexico	1	0.3967 **	3.76	0.3967	3.76
Utah	1	1.1838 **	3.76	1.1838	3.76
New York	1	1.0557 **	3.76	1.0557	3.76
Maine	1	0.4965 **	3.76	0.4965	3.76
Wisconsin	1	0.8712 **	3.76	0.8712	3.76
Missouri	1	1.2948 **	3.76	1.2948	3.76
South Carolina	1	0.7458 **	3.76	0.7458	3.76
Alabama	1	1.2938 **	3.76	1.2938	3.76
Mississippi	1	0.3842 **	3.76	0.3842	3.76
Pennsylvania	1	1.0007 **	3.76	1.0007	3.76
Indiana	1	2.1225 **	3.76	2.1225	3.76
Colorado	1	1.3479 **	3.76	1.3479	3.76
Vermont	1	1.7154 **	3.76	1.7154	3.76
Tennessee	1	0.5677 **	3.76	0.5677	3.76
Montana	1	3.9230	3.76	3.9230	3.76
Arkansas	0	14.8727	15.41	13.1080	14.07
West Virginia	1	0.7659 **	3.76	0.7659	3.76
Kentucky	1	0.9502 **	3.76	0.9502	3.76
Kansas	1	1.1206 **	3.76	1.1206	3.76
Nebraska	1	0.8089 **	3.76	0.8089	3.76
Wyoming	0	8.1668	3.76	8.1668	3.76
Louisiana	1	1.7536 **	3.76	1.7536	3.76
Alaska	0	7.5477	3.76	7.5477	3.76
Texas	0	8.7813	15.41	5.1980	14.07
Iowa	1	1.0014 **	3.76	1.0014	3.76
South Dakota	1	0.1783 **	3.76	0.1783	3.76
Oklahoma	1	0.7905 **	3.76	0.7905	3.76
North Dakota	1	0.8032 **	3.76	0.8032	3.76

Notes: The table shows the result from the cointegration test implemented by vecrank in Stata, which is based on Johansen`s method. The test check if there is Cointegration between the HPI time-series and the GSVI time-series, individually, in each of the 50 states. The null hypothesis is that there are Maximum Rank (0 or 1) cointegrating relationships among variables. ** = 5% significance level for one cointegrating relationship among variables

Appendix F: The Baseline Error Correction Model for all 50 States

For all the models shown in this paper, the following abbreviations are applicable:

$HPI_{s,t}$	= The house price index for state s, at time t
DPI_t	= Disposable personal income at time t
$HPA_{s,t}$	= Housing permits authorized for state s, at time t
$UR_{s,t}$	= Unemployment rate for state s, at time t
IR_t	= Interest rate at time t
$PO_{s,t}$	= Population in state s, at time t
$GSVI_{w,s,t}$	= Google search volume index for search term w, in state s, at time t
CCI_t	= Consumer confidence index
$\epsilon_{HPI,t-1}$	= Error correction term

The Baseline Model

The long run effect:

$$\begin{aligned} HPI_{s,t} = \alpha + \beta_1 HPI_{s,t-1} + \ \beta_2 UR_{s,t} + \beta_3 PO_{s,t} + \beta_4 DPI_t \\ + \beta_5 IR_t + \beta_6 HPA_{s,t} \end{aligned} \tag{5}$$

The short run effect and the speed of adjustment:

$$\begin{aligned} \Delta HPI_{s,t} = \alpha + \beta_1 \Delta HPI_{s,t-1} + \ \beta_2 \Delta UR_{s,t} + \beta_3 \Delta PO_{s,t} \\ + \beta_4 \Delta DPI_t + \beta_5 \Delta IR_t + \beta_6 \Delta HPA_{s,t} + \gamma \epsilon_{HPI,s,t-1} \end{aligned} \tag{6}$$

The Baseline Model Including GSVI for Real Estate Agent

The long run effect:

$$\begin{aligned} HPI_{s,t} = \alpha + \beta_1 HPI_{s,t-1} + \ \beta_2 UR_{s,t} + \beta_3 PO_{s,t} + \beta_4 DPI_t \\ + \beta_5 IR_t + \beta_6 HPA_{s,t} + \beta_7 GSVI_{REA,s,t} \end{aligned} \tag{7}$$

The short run effect and the speed of adjustment:

$$\begin{aligned} \Delta HPI_{s,t} = \alpha + \beta_1 \Delta HPI_{s,t-1} + \ \beta_2 \Delta UR_{s,t} + \beta_3 \Delta PO_{s,t} \\ + \beta_4 \Delta DPI_t + \beta_5 \Delta IR_t + \beta_6 \Delta HPA_{s,t} \\ + \beta_7 \Delta GSVI_{REA,s,t} + \gamma \epsilon_{HPI,s,t-1} \end{aligned} \tag{8}$$

The Baseline Model Including the Consumer Confidence Index

The long run effect:

$$\begin{aligned} HPI_{s,t} = \alpha + \beta_1 HPI_{s,t-1} + \ \beta_2 UR_{s,t} + \beta_3 PO_{s,t} + \beta_4 DPI_t \\ + \beta_5 IR_t + \beta_6 HPA_{s,t} + \beta_7 CCI_t \end{aligned} \tag{9}$$

The short run effect and the speed of adjustment:

$$\begin{aligned} \Delta HPI_{s,t} = \alpha + \beta_1 \Delta HPI_{s,t-1} + \ \beta_2 \Delta UR_{s,t} + \beta_3 \Delta PO_{s,t} \\ + \beta_4 \Delta DPI_t + \beta_5 \Delta IR_t + \beta_6 \Delta HPA_{s,t} \\ + \beta_7 \Delta CCI_{REA,s,t} + \gamma \epsilon_{HPI,s,t-1} \end{aligned} \tag{10}$$