Forecasting house prices using online search activity

Stig V. Møller* Thomas Q. Pedersen† Christian M. Schütte‡
Allan Timmermann§

Preliminary version

Abstract

We show that Google search activity is a strong out-of-sample predictor of future growth in U.S. house prices and that it strongly outperforms standard predictive models based on macroeconomic variables as well as autoregressive models. We extract the most important information from a large set of search terms related to different phases of the home search process into a single Google-based factor and then use it to predict movements in future house prices. At the one-month forecast horizon, the Google factor delivers an out-of-sample R^2-statistic of about 50% for the aggregate U.S. market over the period 2009-2018. We show that the strong predictive power of Google search activity holds for longer forecast horizons, for various house price indices, for seasonally unadjusted and adjusted data, and across individual U.S. states.

Keywords: Big data, internet search, forecasting, housing markets.
JEL codes: C10, E17, G10, R3.

*CREATES, Aarhus University, Fuglesangs Allé 4, 8210 Aarhus. E-mail: svm@econ.au.dk.
†CREATES, Aarhus University, Fuglesangs Allé 4, 8210 Aarhus. E-mail: tqpedersen@econ.au.dk.
‡CREATES, Aarhus University, Fuglesangs Allé 4, 8210 Aarhus. E-mail: christianms@econ.au.dk.
§UCSD, CEPR, and CREATES, University of California, San Diego, La Jolla, CA 92093. E-mail: atimmerm@ucsd.edu.

1 Introduction

The data revolution in recent years has given rise to new opportunities that can help opening the black box of household behavior. One of these new opportunities is to use online search data to obtain information about intentions and potential actions of individuals, implying that such data can be used for prediction purposes (Ng, 2018). A recent report of the National Association of Realtors (NAR, 2017) shows that home buyers use the Internet as their main source to acquire information about the housing market. In fact, as much as 95% of home buyers use the Internet to search for a home. Thus, as home buyers typically use housing-related online search prior to their housing transactions, our hypothesis is that the online searching behavior of individuals contains highly useful information about future movements in the housing market. The objective of this paper is therefore to investigate whether it is possible to exploit the information in online search queries of millions of households to forecast housing market dynamics.

House price fluctuations can have profound impact on household welfare, financial stability, and the entire economy. For example, Case et al. (2012) estimate that the decline in U.S. housing wealth from 2005-2009 implied a decline in consumption of about $350 billion annually. Furthermore, in response to the important role played by house prices in macroeconomic stability, the European Commission has included house prices in its recent early warning system for macroeconomic imbalances (the "MIP Scoreboard"). Thus, producing reliable and accurate forecasts of house prices is evidently of great importance and can, for example, be used for early warning of an incipient housing market overvaluation.

Forecasting house prices is highly complex. As a case in point, the crash of the U.S. housing market in 2007 clearly came as a surprise to many observers. However, perhaps somewhat surprisingly, we should actually expect movements in house prices to be somewhat predictable, which has been pointed out by Shiller (2005, page 14),

among others. Due to various market frictions, such as transaction costs, financing constraints, and short sale constraints, new information is only reflected fully in house prices with a time lag, implying that housing markets – unlike other asset markets – are not informationally efficient. At the same time, the predictability cannot necessarily be exploited economically due to the costs of getting in and out of the housing markets because transaction costs and other frictions are too large (Shiller, 2014).

In their recent survey of the literature on house price forecastability, Ghysels et al. (2013) point out that the common way to predict house prices is by linear regression models and highlight that the commonly used predictive variables in the literature include past house price growth rates, valuation ratios such as the price-rent ratio and the price-income ratio, and various economic variables such as income and employment variables. Hence, so far, the main approach in the housing literature has been to use conventional data sources and conventional methods and perhaps the often reported difficulty in predicting house prices can partly be attributed to the limitations of conventional forecasting models. Even if more advanced forecasting methods are used to explore standard macroeconomic data sources, these forecasting models have been shown to involve large forecasting errors during the housing boom of the 2000s and the subsequent crash, see Rapach and Strauss (2009) and Bork and Møller (2018).

Google is strongly dominating the U.S. search engine market.[1] We therefore build our forecasting models based on Google search data, which are freely available on the web. There are several advantages of using Google search data in forecasting models compared to data gathered from government agencies, which so far typically have been used in the housing literature. In particular, economic variables from national accounts are often announced with a substantial time delay and are also subject to subsequent data revisions, which makes e.g. real-time forecasting complicated. In addition, some economic variables are only available at low frequency. In contrast, Google search

[1]The exact market share depends on the company providing the statistics. As of 2018 Google's market share ranges from roughly 60 percent (www.statista.com) to roughly 90 percent (www.statcounter.com).

data are readily available without time-delay, not subject to data revisions, and are available on a high frequency.[2] Another key advantage is that Google search data is available on national-, state- and even city-level, whereas national accounts data often are limited in geographic scope. This fact is particularly important for our purpose because housing markets are local in nature and we should not expect national-level data to capture all of the complexities of local housing market dynamics.

The underlying idea is that we can use aggregate search data across millions of Internet users to identify concerns and interests about the housing market. As Da et al. (2015) point out, search data have the potential to reveal true and objective signals, whereas responses to survey questions may not be fully truthful. In the case of housing, aggregate search volume for, e.g., "mortgage foreclosure" will intuitively reflect a concern about the risk of future mortgage foreclosure, while "homes for sale" reflects an interest in buying a home.

We use a large number of housing-related search indices. However, if some search terms are irrelevant for explaining the object of interest, it may introduce some noise into the model that potentially can reduce its goodness-of-fit. We therefore use pretesting such that only relevant search terms are included in the given model. To do this, we use the elastic net estimator of Zou and Hastie (2005), which is a powerful and increasingly popular tool in machine learning. Zou and Hastie (2005) describe the elastic net procedure as "a stretchable fishing net that retains all the big fish." The elastic net estimator is the ideal solution to the challenge of noisy signals in search data, as it will help us to discard the uninformative search terms, which in turn will enhance the goodness-of-fit as well as the interpretation of our suggested models. In addition, as Varian (2014) points out, the elastic net estimator can be computed quite efficiently.

In an exploratory study, Wu and Brynjolfsson (2015) find some evidence that online

[2] Guo (2009) and Ghysels et al. (2017) show that asset return predictability from macroeconomic data tends to be considerably weaker when using unrevised real-time macroeconomic data as opposed to using revised macroeconomic data.

search data can be useful in capturing movements in house prices. Our paper differs from Wu and Brynjolfsson (2015) in several important ways, as they for example use simple regression models that condition on only a few search indices and they focus on in-sample evidence. Our focus is on out-of-sample forecasting and we make use of state-of-the-art econometric techniques that filter out the noise from the many available signals in the online search data and therefore is capable of handling the complexity of big data. Whereas a conventional regression-based approach is an improper way of aggregating the data because it does not take into account the power of big data, our suggested machine learning approach is designed to parsimoniously identify the relevant information from a large information space and is therefore well-suited for our purpose of building models of housing market dynamics based on online search data.

Our empirical results are as follows. First, we find that there is lots of relevant information about housing markets to be gained from using Google search data. In our main analysis, we forecast the Federal Housing Finance Agency's seasonally adjusted monthly Purchase-Only Index. Recursively-estimated based on real-time data, our forecasting model predicts the price growth rate on the aggreate U.S. housing market with an out-of-sample R^2-statistic (Campbell and Thompson, 2008) of 49.1% at the one-month horizon and 66.3% at the six-month ahead horizon. We also consider S&P/Case-Shiller and Freddie Mac indices as well as seasonally unadjusted house prices and obtain similar results. Second, we find that Google search activity leads to large improvements in forecast performance relative to existing predictor variables from the literature, including various macroeconomic variables, lagged house price growth and survey-based housing market indicators. Finally, we show that Google search activity is a strong predictor also across individual states in the U.S. In general, the results across states are very similar to those obtained at the national level although with some variation. Google search activity performs very well for coastal states such as Florida and New York, where the out-of-sample R^2-statistic is higher than 70% depending on the horizon, while the predictive power decreases slightly for interior

states with large rural areas. Taken together, the results show that our Google-based forecasting models provide a way for obtaining accurate information on expected future housing market conditions

The rest of the paper is structured as follows. Section 2 describes the empirical approach, including the Google data and how we use the elastic net by Zou and Hastie (2005) and the three-pass regression filter by Kelly and Pruitt (2015) to summarize the most important information in this data and construct a Google factor. Section 3 contains the empirical results for the aggregate U.S. market, including a comparison between Google search data and standard predictive models for house prices. In Section 4, we perform a number of different robustness analysis, while Section 5 contains an analysis of the predictive power of Google search data across U.S. states. Section 6 contains some concluding remarks.

2 The Google factor

The purpose is to track the Internet search behavior of households to identify concerns and interests about the housing market. Our hypothesis is that aggregate search volume for, e.g., "mortgage foreclosure" reflects a concern about the risk of future mortgage foreclosure, while "homes for sale" reflects an interest in buying a home. From Google Trends, we can obtain a time series index on the volume of queries for a given search term in a given geographic area.[3] As an illustration, Figure 1 plots the volume of the search term "mortgage foreclosure" in the U.S., which started increasing in early 2007 just around the time of the crash in U.S. house prices. It peaked in February 2009 after which Google search activity for "mortgage foreclosure" steadily declined and by the end of the sample it was historically low.

[3]It is also possible to get search volume series for other services owned by Google such as Image Search, News Search, Google Shopping and YouTube Search. Our analysis focuses on general web search since home buyers are unlikely to search on these other services.

The proportion of queries for a particular keyword is normalized by the total amount of searches in the selected geographic area and time range. The resulting number is then scaled on a range between 0 and 100 such that the maximum volume for the particular query in the selected time period takes the value 100. Google Trends does not display queries with search volume below a certain undisclosed threshold. With a multi-word search query, Google Trends automatically includes searches that include the keywords in any order. Due to privacy concerns Google Trends does not disclose the actual number of searches on the selected keyword(s). However, for the purpose of forecasting this is not a problem since we are only interested in the time series dynamics of relative search activity.[4]

Another useful feature of Google Trends is that it provides a set of related queries for every main query. This list of related queries (or related terms as we will use interchangeably) includes between 0 and 25 different terms, where the final number depends on the search volume of the main query, i.e. high volume series will usually have 25 related queries and lower volume series will have less. The methodology that Google uses to select related queries is not disclosed. However, the resulting terms are usually intuitively related to the main query. For example, querying for the term "homes" in the geographical region of the U.S. during the period 2004:1 to 2018:10 leads to a list of 25 related queries, of which the top 5 are: "homes for sale", "homes for rent", "new homes", "mobile homes" and "rental homes". From a forecasting perspective this feature is appealing for two reasons. First, each semantically related keyword can provide additional information about the target variable that the original query does not include. Second, since the related terms are likely to be correlated, this results in a natural factor structure between them, which is a feature that we can exploit to our advantage when building the forecasting model.

[4] As noted by Baker and Fradkin (2017), the actual volume of search queries can be estimated using another service from Google called Google Adwords.

2.1 Choice of keywords

Since Google Trends data are only available from 2004 and onwards, our sample spans from 2004:1 to 2018:10 and has a monthly frequency. To construct the main set of predictors, we start by selecting and downloading search volume series for 30 queries: "realtor", "houses for sale", "homes for sale", "real estate", "homes for rent", "mls", "for sale by owner", "land for sale", "homes for sale near me", "mls listings", "condos for sale", "real estate agent", "foreclosed homes", "property", "foreclosure", "homes for sale by owner", "for sale", "buying a house", "realty", "first time home buyer", "mortgage calculator", "mortgage rates", "mortgage", "loans", "home loan calculator", "interest rates", "mortgage payment calculator", "refinance", "moving company", "fha loan". These keywords are selected using the Google Ads keyword planner, which is a service that, for a given category, provides you with highest search volume keywords that a website should include to increase online visibility. We obtain the first 20 search queries from the top twenty queries related to the real state category and the last 10 search queries from the mortgage category. For each of the 30 main queries, we add their related terms and remove duplicates, low volume series and unrelated series.[5] This methodology follows Da et al. (2015), who start with a set of main queries, add related terms to enrich the data set and then remove duplicates, low volume series and economically unrelated series.

As noted by D'Amuri and Marcucci (2017) Google Trends are created based on a sample of queries that change according to the time and IP address of the computer used to download the data. To account for sampling error, we compute the index for all Google Trends queries based on an average over 15 different days. The correlation across different samples is always above 0.99. Hence, the results are, for all practical concerns, not sensitive to this precaution.

[5] We define low volume series as series where more than 10% of the observations are 0. Some of the related terms can be totally unrelated to the housing market. Although unrelated terms are very uncommon, sometimes the related terms which Google provides are clearly not related to the housing market. For example, the related terms for "property" include "distributive property" and "associative property".

This process is performed for the geographical region of the U.S. as a whole and for each of the individual states of the union, resulting in 51 different datasets of predictor time series (i.e one for each target series). We refer to the lists including all predictors as the expanded lists. An appealing feature of our approach is that Google Trends automatically includes geographical idiosyncrasies of home buyer search patterns in each state through the related terms. Hence, each data set will be heavily localized.[6] Since both search volume and the number of related terms vary across geographical regions the number of predictor series for each dataset also varies. For each state we include search activity in neighboring states with the purpose of accounting for potential moves across state borders. The intution is that search activity for individuals residing in a given state count in the overall search volume for that particular state, but the individuals can potentially buy a home in the neighboring state. In forecasting house prices across states, we include both this local search activity and national-level search activity. This is motivated by existing evidence that movements in house prices across the U.S. depend on both local and national factors (e.g. Del Negro and Otrok, 2007 and Hernández-Murillo et al., 2017). Although the use of online search activity provides a different setting compared to existing research on house price movements, we conjecture that national-level search activity can influence local house prices.

Following Da et al. (2011, 2015) and Vozlyublennaia (2014), we start by converting the series to their natural logarithm. Since the Google Trends series appear to be relatively heterogeneous in terms of their order of integration and whether they contain deterministic trends we adopt a sequential testing strategy in the spirit of Ayat and Burridge (2000).[7] The idea is to successively test for stationarity, linear

[6] For example, home buyers in states with large urban agglomerations have a higher tendency to include the names of cities (e.g. New York City or Chicago) or neighborhoods within cities (e.g. Brooklyn or Bronx) in their search queries than home buyers in sparsely populated states.

[7] There is indeed no consensus in the literature as to whether or not Google Trends data is best characterized by stationarity, trend stationarity or a unit root since this appears to be completely dependent on the query in question. Vozlyublennaia (2011), Choi and Varian (2012), Bijl et al. (2016) and D'Amuri and Marcucci (2017) do not perform any differencing or detrending of the series, which suggests that the Google Trends they use are stationary. Yu et al. (2019) use an ADF test on three Google Trends queries: "oil inventory", "oil consumption" and "oil price" and find evidence of stationarity at the 5% level (10% level) in "oil inventory" ("oil consumption"), but are not able to

trend stationarity and quadratic trend stationarity using an augmented Dickey-Fuller (ADF) test. Hence, the first test is an ADF test with a constant term. If the null of non-stationarity is rejected, we stop and use the series without any transformation; conversely, if the null is maintained, we use an ADF test that includes both a constant and a linear time trend. If the null of this second test is rejected, we linearly detrend the series by using the residuals of a regression of the series on a constant and a time trend, otherwise we run a final ADF test that includes a constant, a linear trend and a quadratic trend. If we reject the null of this test, we detrend the series by a similar methodology as before but including a quadratic trend in the regression, otherwise we take first differences.

When the target variable is seasonally adjusted, we remove seasonality by running each series on monthly dummy variables and taking the residual of this regression. To avoid look-ahead bias, we perform the sequential testing for unit roots and deseasonalization on a recursively expanding window, where the smallest window used matches our estimation window for the forecasting model. Hence, we only use information available at time t in both procedures.

2.2 Constructing the Google factor

The forecasting model we rely on is given by:

$$y_{t+h} = \alpha + \beta G_t + \varepsilon_{t+h} \tag{1}$$

where y_{t+h} is the h-period ahead log growth rate in house prices and G_t is the Google factor. To estimate the Google factor, we rely on the three-pass regression filter developed by Kelly and Pruitt (2015).[8] This method is very convenient since it allows us to

reject the null of a unit root for "oil price". Da et al. (2015) take log-differences on the series.

[8]Kelly and Pruitt (2015) show that the three pass regression filter is a generalization of partial least squares (PLS). In particular, the three pass regression filter and the original PLS method developed by Wold (1966) result in identical forecasts if the predictors are studentized and the regressions in step 1 and 2 of their algorithm are run without a constant.

estimate the Google factor and subsequent forecasts of the target variable (y_{t+h}) using only OLS regressions. Furthermore, we can filter out the noise from the many available signals in the Google search data and summarize the most important information in a single common component (Wold, 1966; Kelly and Pruitt, 2015).

More specifically, in constructing the Google factor, we initially consider the search volume for the individual indices in the expanded list. If some of the search indices are irrelevant predictors of house price growth rates, it may introduce some noise into the Google factor that can potentially reduce its forecasting ability. We therefore follow Bai and Ng (2008) and use pre-testing such that only the most relevant search indices are included in the Google factor. To do this, we use the elastic net estimator of Zou and Hastie (2005) because it is known to perform well when there is a high correlation between the predictors. From the elastic net we select the ten most relevant search indices and subsequently apply the three-pass regression filter by Kelly and Pruitt (2015) to summarize the most important information from the ten indices into one common component, i.e. the Google factor.

We construct the Google factor out-of-sample in accordance with our forecasting scheme, where we use a rolling window of 60 observations (5 years) for model estimation. Furthermore, we take into account a two-month publication lag of the house price indices. The Google factor is thus constructed using only information available at the time of the forecast. Applying the elastic net and the three-pass regression filter in this recursive way implies that the both the included search indices and their weight in the Google factor may vary over time.

3 Forecasting U.S. house prices

In our main analysis, we forecast aggregate U.S. house prices using the 30 predetermined queries as well as related queries resulting in 508 predictor variables. We

take into account a two-month publication lag of house price indices and estimate the Google-based forecasting model (1) using a rolling window of 60 observations and hence reserve the period 2004:1 to 2009:3 for initial estimation of the model giving us an out-of-sample forecast period from 2009:4 to 2018:10. As our target in the main analysis, we use the nominal and seasonally adjusted Purchase-Only Index from the Federal Housing Finance Agency (FHFA).[9] Later, we show that our main findings are robust to using seasonally unadjusted data as well as to using other house price indices. We also show that our results are robust to the size of the rolling window as well as the use of an expanding instead of rolling window.

3.1 Predictive power of Google search activity

Table 1 shows the results for five different forecast horizons (1, 3, 6, 9 and 12 months). The table reports the Campbell and Thompson (2008) out-of-sample R^2 (R^2_{OoS}) and the Diebold and Mariano (1995) t-statistic (t_{DM}), where the null hypothesis is that the R^2_{OoS} is equal to zero or negative and the alternative hypothesis is that it is positive. In computing R^2_{OoS}, we use a 60-month rolling mean as benchmark.

As Table 1 shows, the Google factor is able to explain almost 50% of next month's growth in U.S. house prices. The predictive power increases with the forecast horizon and reaches its peak for $h = 6$ with an R^2_{OoS} of 66.3% and then drops to 53.1% for $h = 12$. According to t_{DM}, we strongly reject the null hypothesis that $R^2_{OoS} \leq 0$ for all forecast horizons.

The remarkably good predictive power of Google search activity for future house price growth rates is also illustrated in Figure 2, which for $h = 1$ shows that Google search activity in general captures movements in house prices very well. In particular, we note that it captures both the negative growth rates in 2009-2010 following the collapse in

[9] FHFA also publishes an All-Transactions Index that in addition to price changes takes into account refinancings, but this index is only available on a quarterly frequency. Applying this index would thus severely reduce the scope of the out-of-sample analysis.

the housing market, the subsequent recovery and the more stable house price growth in recent years. The predictive power contained in Google search activity is thus not confined to certain market characteristics.

3.2 Benchmark predictive models

In Table 1 we also compare the predictive power of Google search activity to a set of standard predictive models for house prices. This includes an AR(AIC) model where the optimal number of AR terms are chosen recursively based on the Akaike information criterion (AIC). Case and Shiller (1989) show that housing returns exhibit positive autocorrelation and Crawford and Fratantoni (2003) show that simple ARMA models do well in forecasting house prices in out-of-sample tests.[10] Motivated by the findings of Rapach and Strauss (2009), among others, we also include a set of seasonally adjusted macroeconomic variables: industrial production (differences, logs), employment (differences, logs), the unemployment rate, the all items consumer price index for all urban consumers (differences, logs), average weekly working hours in manufacturing (differences), building permits (differences, logs), housing starts (differences, logs), and commercial and industrial loans outstanding (differences, logs). Furthermore, we include the federal funds rate (differences) and the yield spread computed as the difference between the 10-year treasury constant maturity rate and the federal funds rate. These variables are all obtained from the economic research database at the Federal Reserve Bank of St. Louis (FRED). Finally, we also include the price-rent ratio as predictor, cf. Gallin (2008), Campbell et al. (2009), Plazzi et al. (2010) and Engsted and Pedersen (2015). We compute the price-rent ratio as the log difference between the house price index and rents (both seasonally adjusted), where rents are measured based on the rent of primary residence index available from the FRED.

[10] We have experimented with an ARMA(AIC) model, but the inclusion of MA terms tend to slightly reduce the predictive power. To allow for the strongest possible benchmark, we therefore use an AR(AIC) model. Likewise, we have used the Schwarz information criterion (SIC) instead of AIC. This has no effect on the qualitative conclusions.

To make a methodologically fair comparison of the use of Google search activity and macro variables in forecasting house prices, we also apply PLS on our set of macro variables.

Table 1 shows that the Google factor strongly outperforms the benchmark models across all horizons. Although many of the alternative predictive models do have a positive R^2_{OoS}, they are much smaller than that for the Google factor and only for AR(AIC) and Macro(PLS) do we reject the null hypothesis $R^2_{OoS} \leq 0$ against the alternative that $R^2_{OoS} > 0$. This is, however, only the case for shorter forecast horizons; for longer horizons, only the Google factor remains significant. Figure 2 shows that the two main competing benchmark models, AR(AIC) and Macro(PLS), in general deliver much less volatile forecasts than the Google factor and they both fail in capturing the recovery of the housing market in 2011-2013.

To more formally evaluate if the Google factor improves the accuracy of the forecasts based on AR(AIC) and macro models, Tables 2 and 3 show the results from forecast encompassing tests (Harvey et al., 1998). The idea in forecast encompassing tests is to test if $\lambda = 0$ in the following composite forecast

$$f_{ct} = (1 - \lambda) f_{1t} + \lambda f_{2t}, \quad 0 \leq \lambda \leq 1, \tag{2}$$

where f_{ct} is a weighted average of two individual forecasts, f_{1t} and f_{2t}. Table 2 shows the results where f_{1t} denotes forecasts based on AR(AIC) and macro models, respectively, and f_{2t} denotes forecasts based on the Google factor, i.e. the null hypothesis is that the benchmark model encompasses Google search activity. From Table 2 it is clear that we strongly reject the null hypothesis for all benchmark models and across all horizons. Google search activity thus provides relevant additional information for forecasting house prices not already contained in standard time series models or macroeconomic variables. The weight on the Google factor is close to one for $h > 1$, which indicates that at longer horizons univariate time series models and macro variables do not contain relevant information that can improve forecasts based on Google search

activity. This conclusion is also supported by Table 3, which shows the results where f_{1t} denotes forecasts based on the Google factor and f_{2t} denotes forecasts based on AR(AIC) and macro variables, respectively, i.e. the null hypothesis is now that Google search activity encompasses the benchmark model. According to Table 3, we cannot reject the null hypothesis for $h > 1$. However, with a one-month forecast horizon, λ is in many cases significantly different from zero, which implies that at short horizons the benchmark models do provide relevant information for future house prices beyond what is already contained in the Google factor.

3.3 Survey-based housing market indicators

As Da et al. (2015) point out, search data have the potential to reveal true and objective signals, whereas responses to survey questions may not be fully truthful. In our case, a comparison of search- and survey-based housing market indicators is particularly interesting since the National Association of Home Builders (NAHB) each month publish a survey-based housing market indicator that attracts widespread attention in the media and amongst economists. The Housing Market Index by NAHB is based on a monthly survey of its members designed to provide an outlook for the U.S. housing market. The survey asks respondents to rate market conditions for the sale of new homes at the present time and in the next six months as well as the traffic of prospective buyers of new homes.[11]

Tables 1-3 and Figure 2 show the results using NAHB as predictor for future house price growth. In general, the Google factor strongly outperforms NAHB. For $h = 1$ NAHB generates an R^2_{OoS} of 24.6% compared to 49.1% using the Google factor, while

[11] Fannie Mae also publish a monthly survey-based housing market indicator, the Home Purchase Sentiment Index (HPSI). HPSI differs from the survey by NAHB in the sense that the respondents are consumers and not professional home builders. HPSI is based on six questions that relate to consumers expectations to house prices, mortgage interest rates, job security and household income over the next 12 months as well as their opinion on whether or not now is a good or bad time to buy/sell a house. HPSI was, however, first introduced in 2011, and is, thus, due to data limitations not included in our analysis.

for $h = 6$ the corresponding numbers are 2.8% and 66.3%, cf. Table 1. Irrespective of the forecast horizon, NAHB does not deliver a significantly positive R^2_{OoS} (Table 1) and the Google factor encompasses NAHB (Tables 2 and 3). Figure 2 shows that NAHB delivers even more smooth forecasts than the AR(AIC) and Macro(PLS) models and, hence, the survey-based housing market indicator fails in capturing the short-run fluctuations in house prices. Furthermore, NAHB also fails in capturing (the first part of) the recovery period in 2011-2013 similar to the other benchmark models.

3.4 Interpreting the Google factor

The Google factor is constructed using a large set of search terms. To better understand the content of the Google factor and which search terms that drive the strong predictive power, Table 4 shows the inclusion frequencies of the 15 most frequently used search terms for $h = 1$. In constructing the Google factor, we use the elastic net to select the ten most relevant search terms at each point in time. The selected search terms, thus, potentially varies over time, and we compute the inclusion frequencies by counting the number of times a given search term is selected by the elastic net relative to the total number of out-of-sample observations.

The search terms are in general very intuitive. A group of search terms relate to financing the purchase of a home such as "Loan interest calculator", "Amortization" and "Monthly mortgage payments". Another group relates to defaulting on loans such as "What is foreclosure" and "Mortgage foreclosure". Also the process of finding a new home is represented by search terms such as "Free foreclosed homes", "Foreclosure auction", "Buying a house steps", "Home for sale by owner" and "Houses for sale". Most relevant search terms are not directly related to a specific geographical area, which intuitively makes sense given that we here forecast national house prices. Two of the most frequently selected search terms do, however, contain the name of a specific state: "Tn land for sale" and "Florida homes for sale by owner", where "Tn" is the

abbreviation for Tennessee. The elastic net chooses the most relevant search terms, while at the same time discarding search terms that are highly correlated with already chosen search terms and as such state-specific search terms can be included if they are highly correlated with relevant non state-specific search terms.

In Table 4 we only show the inclusion frequencies for $h = 1$. For other forecast horizons, the inclusion frequencies and the search terms themselves can be different. To conserve space we do not report the results for $h > 1$, but the overall conclusion from Table 4 also applies here. Some words change and the inclusion frequencies are also slightly different, but the relevant search terms still relate to different steps in the home buying process such as financing and finding a home. The only main difference between relevant search terms for $h = 1$ and $h > 1$ is that search terms related to "foreclosure" are only relevant at short horizons.

4 Robustness

4.1 Bootstrap p-values

In the following we carry out a bootstrap simulation experiment to study the robustness of the elastic net/three-pass regression filter technique as a useful approach in incorporating Google search activity into a factor model of expected house price growth rates. We make use of a simulation experiment that is comparable to the "useless" factor tests of Kan and Zhang (1999a,b). In particular, we generate 1,000 bootstrap samples by row-wise resampling from the observed panel of Google series (with replacement). The resampled Google panels have the same length as the original panel of Google series. For each bootstrap sample, we then estimate the Google factor forecasting model and save the R^2_{OoS} statistic. As the resampled Google search data should have no relation to the realized house price growth rates, the Google factor should not be useful in forecasting house price growth rates. Basically, the resampled

Google panels represent random noise (i.e. they are "useless").

We analyze the empirical distribution of the R^2_{OoS} statistic by computing empirical p-values. The simulations show that the share of artificial R^2_{OoS} statistics larger than the original R^2_{OoS} statistic of 49.1% at the $h = 1$ horizon is roughly 0%. Hence, the chance of getting the same goodness-of-fit with random Google data is close to zero. As Table 5 shows, we reach the same conclusion for the other forecast horizons. We find it reassuring that it is not possible to obtain a high explanatory power with the elastic net/three-pass regression filter technique if the panel of Google series is random noise.

4.2 Seasonally unadjusted data

In our main analysis we forecast growth in seasonally adjusted house prices using seasonally adjusted Google search activity and seasonally adjusted macroeconomic variables. As a robustness check of the general applicability of Google search activity in forecasting house prices, we also consider seasonally unadjusted data, both in terms of house prices, Google search activity and macroeconomic variables. For this purpose, we reconstruct the Google factor using seasonally unadjusted data, i.e. we use the approach outlined in Section 2.2 but instead with log house price growth based on the seasonally unadjusted index as target in (1).

Table 6 shows results similar to Table 1 but with seasonally unadjusted data instead. The Google factor remains the single best predictor of future house price growth with a significantly positive R^2_{OoS} above 50% for all horizons. For $h = 1$, AR(AIC) and Macro(PLS) are the two best alternative predictors with an R^2_{OoS} of 34.7% and 26.6%, respectively.[12] The survey-based housing market indicator by NAHB does not account for seasonality and as such the poor performance with a negative R^2_{OoS}

[12] For a fair comparison, we here use a seasonal AR(AIC) model. A standard AR(AIC) model performs much worse on seasonally adjusted data as expected.

of -13.3% for $h = 1$ is not surprising. Figure 3 shows the actual housing price growth based on the seasonally unadjusted index and the forecasts for $h = 1$ for each of these four predictors. Similar to the case with seasonally adjusted data it is evident that AR(AIC) and Macro(PLS) fail in capturing the recovery period 2011-2013. The figure also clearly visualizes the poor performance of NAHB with respect to seasonally unadjusted house prices. For forecast horizons longer than one month, the strong performance of the Google factor becomes even more evident. The benchmark predictive models only deliver an R^2_{OoS} of more than 10% in three cases across forecast horizons from three to 12 months. Google search activity thus contains highly relevant information for forecasting house price and strongly dominates alternative models, irrespective of the use of seasonally adjusted or unadjusted house prices.

4.3 Other house price indices

There are a number of different price indices for the U.S. housing market. In our main analysis, we focus on the FHFA Purchase-Only Index, which is available on a monthly frequency. Other monthly indices include the Freddie Mac House Price Index (Freddie-Mac) and the S&P/Case-Shiller House Price Index (Case-Shiller). All three indices apply the repeat-sales methodology but differ in various other aspects. For example, Freddie-Mac is the only index of the three that to a certain degree account for appraisal values from refinancing transactions. Another example of how the house price indices differ is in terms of how local house prices are weighted in constructing national indices. Freddie-Mac is constructed using weights based on the estimated property value underlying active Freddie Mac loans, while FHFA and Case-Shiller, respectively, use unit- and value-weighting of the individual homes. These differences (and more) can potentially lead to house price indices that evolve differently over time and as such it is unclear if the strong predictive power of Google search activity for the FHFA Purchase-Only Index carries over to these other indices.

Figure 4 shows the house price growth rates for seasonally adjusted and unadjusted Freddie-Mac and Case-Shiller indices along with the forecasts based on the corresponding Google factor. Similar to our analysis of seasonally unadjusted data in Section 4.2, we reconstruct the Google factor to match the relevant house price index. Compared to FHFA (Figures 2 and 3), Case-Shiller and, especially, Freddie-Mac are more persistent and hence growth rates are much more smooth over time, cf. Figure 4. Google search activity, however, retains its strong predictive power for future house prices. With seasonally adjusted data, the Case-Shiller delivers an R^2_{OoS} of 36.0% for $h = 1$, while Google search activity is able to explain 63.1% of the variation in Freddie-Mac growth rates one-month ahead. Similar to FHFA, the predictive power of Google search activity peaks at $h = 6$ for Case-Shiller and Freddie-Mac with out-of-sample R^2's of 66.0% and 68.8%, respectively. For seasonally unadjusted data and a one-month forecast horizon, Case-Shiller yields an R^2_{OoS} of 40.0%, while it is 60.6% for Freddie-Mac.

4.4 Rolling versus expanding estimation window

Our main results are based on a five-year rolling estimation window, which implies that we always use 60 observations in constructing the Google factor and in estimating the forecasting model. Our main results are, however, robust to the size of the rolling window as well as the use of an expanding instead of a rolling window. We illustrate this result in Figure 5 using a three-year rolling window and an expanding window using three years for initial estimation. Reducing the window-size from five to three years implies that we include the housing-market crash in 2007-2009 in our evaluation period. Existing forecasting models often have difficulties in capturing the negative growth rates in this period (e.g. Rapach and Strauss, 2009, and Bork and Møller, 2018), but despite a small lag for FHFA and Freddie-Mac, the Google factor captures this strong decline in house prices very well. Compared to Figures 2 and 4, we see that a three-year rolling window and an expanding window yield very similar forecast

performance as the five-year rolling window. This is also reflected in the R^2_{OoS}. For $h = 1$ and a three-year rolling window, the R^2_{OoS} is 49.8%, 56.9% and 66.3% for FHFA, Case-Shiller and Freddie-Mac, respectively. For the expanding window, the corresponding numbers are 48.9%, 49.9% and 59.9%. In comparison, the numbers for the five-year rolling window are 49.1%, 36.0% and 63.1%. The predictive power for Case-Shiller improves slightly by reducing the (initial in case of expanding) estimation window, but for FHFA and Freddie-Mac the results remain more or less unchanged. In any case, the overall conclusion that Google search activity is a strong predictor for future house price growth is not restricted to the use of a five-year rolling estimation window.

5 Forecasting state-level house prices

There is compelling evidence in the literature that there are important differences in the degree of forecastability of house price growth across U.S. states. Using data over the period 1995-2006 and a host of different state, regional, and national economic predictor variables, Rapach and Strauss (2009) generally obtain relatively accurate forecasts for interior states, while forecast errors tend to be relatively large for coastal states that experienced strong house price growth in the sample period. Using data over the period 1976-2012, Bork and Møller (2015) show that large gains can be made for coastal states by using more sophisticated forecasting methods that allow for both model and parameter shifts, but still forecast errors are largest for states with highly volatile house price growth.

In light of the existing evidence, we study the forecast power of Google search activity across U.S. states. We use the same 60-months rolling window in creating our out-of-sample forecasts and for each state we use the expanded state-specific and national lists of search terms, cf. Section 2.1. Since FHFA does not provide state-level house prices on a monthly frequency, we use the seasonally adjusted Freddie-Mac state-level

indices as targets. Following Rapach and Strauss (2009), Table 7 shows the results for the 20 largest states as measured by population. For the national Freddie-Mac index, the R^2_{OoS} ranges between 60-70% across forecast horizons from 1-12 months. For many states Google search activity displays a similar degree of predictive power. In most cases the R^2_{OoS} is significantly positive and only for Indiana ($h = 1$) and Virgina ($h = 3$) does the rolling-mean benchmark provide better forecasts for future house prices. In contrast to existing evidence using macroeconomic and time series models, Google search activity performs very well for coastal states and in many cases better than for interior states. For example, for Florida and New York the Google factor explains up to roughly 80% of the variation in house prices depending on the horizon, while for Missouri the R^2_{OoS} is at most roughly 50%. This difference across coastal and interior states is visually illustrated in Figure 6, which shows the R^2_{OoS} for $h = 6$ across all U.S. states. The interior states typically have large rural areas and are more sparsely populated. Our results show that in these states online search activity explains a relatively smaller part of future house price variation compared to states with larger urban areas.

In general, however, Google search activity is a strong predictor of future house prices across most states as seen from Figure 6, where the R^2_{OoS} is negative only for Vermont and it consistently beats the strongest possible benchmark in an AR(AIC) model. This can be seen by comparing Figure 6 to Figure 7, which shows the R^2_{OoS} for $h = 6$ using an AR(AIC) model to forecast house prices.

6 Conclusion

House price fluctuations can have significant impact on household welfare, financial stability, and the entire economy as a whole, and the need for policy makers to closely watch the housing markets is by now obvious. In this paper, we show that there is lots of relevant information about housing markets to be gained from using Google

search data. Our findings imply that Google-based forecasting models can be a particularly valuable tool for obtaining accurate real-time information on the housing market conditions.

Overall, we expect our forecasting models to be a highly valuable tool for market participants as well as regulators and policy makers, who seek timely and accurate information about the outlook for the housing market. Producing reliable and accurate forecasts of house prices is evidently of great importance and can, for example, be used for early warning of an incipient housing market overvaluation.

References

Ayat, L., Burridge, P. (2000). Unit root tests in the presence of uncertainty about the non-stochastic trend. *Journal of Econometrics* 95, 71-96.

Bai, J., Ng, S. (2008). Forecasting economic time series using targeted predictors. *Journal of Econometrics* 146, 304-317.

Baker, S., Fradkin, A. (2017). The impact of unemployment insurance on job search: Evidence from Google search data. *Review of Economics and Statistics* 99, 756-768.

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

Bork, L., Møller, S. (2015). Forecasting house prices in the 50 states using dynamic model averaging and dynamic model selection. *International Journal of Forecasting* 31, 63–78.

Bork, L., Møller, S. (2018).House prices forecastability: A factor analysis. *Real Estate Economics* 46, 582-611.

Campbell, D., Morris, D., Gallin, J., Martin, R. (2009). What moves housing markets: A variance decomposition of the rent-price ratio. *Journal of Urban Economics* 66, 90-102.

Campbell, J., Thompson, S. (2008). Predicting excess stock returns out of sample: Can anything beat the historical average? *Review of Financial Studies* 21, 1509-1531.

Case, K. E., Shiller, R. J. (1989). The efficiency of the market for single family homes. *American Economic Review* 79, 125-137.

Case, K.E., Quigley, J.M., Shiller, R.J. (2012). Wealth effects revisited 1975-2012. Cowles Foundation Discussion Paper no. 1884. Yale University.

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

Crawford, G. W., Fratantoni , M. C. (2003). Assessing the forecasting performance of regime-switching, ARIMA and GARCH models of house prices. *Real Estate Economics* 31, 223-243.

Da, Z., Engelberg, J., Gao, P. (2011). In search of attention. *Journal of Finance* 66, 1461-1499.

Da, Z., Engelberg, J., Gao, P. (2015). The sum of all FEARS: Investor sentiment and asset prices. *Review of Financial Studies* 28, 1-32.

D'Amuri, F., Marcucci, J. (2017). The predictive power of Google searches in forecasting US unemployment. *International Journal of Forecasting* 33, 801-816.

Del Negro, M., Otrok, C. (2007). 99 luftballons: Monetary policy and the house price boom across US states. *Journal of Monetary Economics* 54, 1962-1985.

Diebold, F.X., Mariano, R.S. (1995). Comparing predictive accuracy. *Journal of Business and Economic Statistics* 13, 253-263.

Engsted, T., Pedersen, T.Q. (2015). Predicting returns and rent growth in the housing market using the rent-price ratio: Evidence from the OECD countries. *Journal of International Money and Finance* 53, 257-275.

Gallin, J. (2008). The long-run relationship between house prices and rents. *Real Estate Economics* 36, 635-658.

Ghysels, E., Horan, C., Moench, E. (2017). Forecasting through the rearview mirror: Data revisions and bond return predictability. *Review of Financial Studies* 31, 678-714.

Ghysels, E., Plazzi, A., Valkanov, R., Torous, W. (2013). Forecasting real estate

prices. *Handbook of Economic Forecasting*, G. Elliot and A. Timmermann (Eds.), 509-580.

Guo, H. (2009). Data revisions and out-of-sample stock return predictability. *Economic Inquiry* 47, 81-97.

Harvey, D.I., Leybourne, S.J., Newbold, P. (1998). Tests for forecast encompassing. *Journal of Business and Economic Statistics* 16, 254-259.

Hernández-Murillo, R., Owyang, M.T., Rubio, M. (2017). Clustered housing cycles. *Regional Science and Urban Economics* 66, 185-197.

Kan, R., Zhang, C. (1999a). GMM tests of stochastic discount factor models with useless factors. *Journal of Financial Economics* 54, 103-127.

Kan, R., Zhang, C. (1999b). Two-pass tests of asset pricing models with useless factors. *Journal of Finance* 54, 203-235.

Kelly, B., Pruitt, S. (2015). The three-pass regression filter: A new approach to forecasting using many predictors. *Journal of Econometrics* 186, 294-316.

National Association of Realtors (NAR) (2017). Home Buyers and Sellers Generational Trends Report. Technical Report, National Association of Realtors.

Ng, S. (2018). Opportunities and challenges: Lessons from analyzing terabytes of scanner data. Forthcoming in Eleventh World Congress of the Econometric Society, Cambridge University Press.

Plazzi, A., Torous, W., Valkanov, R. (2010). Expected returns and expected growth in rents of commercial real estate. *Review of Financial Studies* 23, 3469-3519.

Rapach, D., Strauss, J. (2009). Differences in housing price forecastability across US states. *International Journal of Forecasting* 25, 351-372.

Shiller, R. (2005). *Irrational Exuberance*. 2nd Edition. Princeton University Press.

Shiller, R. (2014). Speculative asset prices. *American Economic Review* 104, 1486-1517.

Varian, H. (2014). Big data: New tricks for econometrics. *Journal of Economic Perspectives* 28, 3-28.

Vozlyublennaia, N. (2014). Investor attention, index performance and return predictability. *Journal of Banking and Finance* 41, 17-35.

Wold, H. (1966). Estimation of principal components and related models by iterative least squares. In: Krishnaiaah, P. (Ed.), Multivariate Analysis, Academic Press, New York, 391-420.

Wu, L., Brynjolfsson, E. (2015). The future of prediction: How Google searches foreshadow housing prices and sales. In: A. Goldfarb, S. M. Greenstein, and C. E. Tucker (Eds.), *Economic analysis of the digital economy*, Chicago: University of Chicago Press.

Yu, L., Zhao, Y., Tang, L., Yang, Z. (2019). Online big data-driven oil consumption forecasting with Google Trends. *International Journal of Forecasting* 35, 213-223.

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

Table 1. The out-of-sample predictive power of Google search activity for future house price growth rates and comparison with alternative predictive models. The table reports the Campbell and Thompson (2008) out-of-sample R^2 (R^2_{OoS}) and the Diebold and Mariano (1995) t-statistic (t_{DM}), which is computed using the Newey-West estimator with h lags, where h is forecast horizon in months. The null hypothesis is that the R^2_{OoS} is equal to zero or negative and the alternative hypothesis is that it is positive.

	$h=1$		$h=3$		$h=6$		$h=9$		$h=12$	
	R^2_{OoS}	t_{DM}	R^2_{OoS}	t_{DM}	R^2_{OoS}	t_{DM}	R^2_{OoS}	t_{DM}	R^2_{OoS}	t_{DM}
Google	49.1	3.08	58.4	2.63	66.3	2.60	65.6	2.42	53.1	1.99
AR(SIC)	28.6	2.77	33.8	2.63	20.3	2.10	12.6	1.63	5.9	0.86
Industrial prod.	0.7	0.31	1.8	0.40	0.3	0.07	0.8	0.19	1.02	0.34
Employment	2.7	0.47	8.3	1.13	9.9	1.47	8.3	1.36	7.40	1.11
Unemp. rate	-7.6	-1.44	-8.9	-1.36	-11.3	-1.26	-23.7	-1.51	-36.3	-1.48
Inflation	1.9	1.53	2.0	0.86	1.4	1.00	0.6	0.75	-0.4	-0.43
Hours	-2.1	-1.26	-2.7	-1.40	-0.6	-0.46	-0.2	-0.16	0.9	1.27
Building permits	3.8	0.80	0.3	0.08	1.7	0.45	1.6	0.33	2.7	0.94
Housing starts	-0.2	-0.13	-0.3	-0.15	1.3	0.55	0.5	0.21	0.4	0.29
Price-rent ratio	0.2	0.02	-1.0	-0.06	0.4	0.02	-5.3	-0.22	-16.9	-0.63
Fed funds rate	6.4	1.27	6.4	1.57	10.6	1.84	11.1	1.94	7.8	1.67
Yield spread	-2.0	-0.49	-10.5	-1.54	-16.8	-1.52	-21.3	-1.12	-15.7	-0.75
Loans	-26.3	-2.22	-9.9	-1.13	-11.7	-1.03	-13.2	-1.00	-9.1	-0.77
Macro(PLS)	17.7	2.77	20.0	2.44	14.3	1.40	-4.4	-0.19	-19.3	-0.53
NAHB	24.6	1.75	22.3	0.99	2.8	0.18	-21.3	-1.07	-35.2	-1.18

Table 2. Forecast encompassing tests. The table reports the Harvey et al. (1998) test statistic t_{HLN}, where the null hypothesis is that the weight on the Google factor λ is zero against the alternative hypothesis that it is positive. The test statistic is computed using the Newey-West estimator with h lags, where h is forecast horizon in months.

	$h=1$		$h=3$		$h=6$		$h=9$		$h=12$	
	λ	t_{HLN}	λ	t_{HLN}	λ	t_{HLN}	λ	t_{HLN}	λ	t_{HLN}
AR(SIC)	0.77	3.82	0.90	3.52	1.07	3.16	1.11	2.71	1.05	2.36
Industrial prod.	0.81	4.84	0.92	4.00	0.96	3.56	1.06	3.09	1.02	2.59
Employment	0.82	4.83	0.93	3.93	0.95	3.48	1.06	2.99	1.02	2.51
Unemp. rate	0.90	4.47	0.97	3.74	1.05	3.37	1.16	3.10	1.01	2.71
Inflation	0.80	4.69	0.92	3.84	0.97	3.48	1.04	3.12	1.00	2.65
Hours	0.82	4.88	0.93	4.02	0.96	3.54	1.04	3.12	1.00	2.59
Building permits	0.79	4.88	0.90	4.14	0.94	3.68	1.00	3.27	0.98	2.66
Housing starts	0.81	4.85	0.91	4.01	0.95	3.56	1.02	3.16	0.99	2.65
Price-rent ratio	0.77	5.44	0.88	4.54	0.91	3.83	0.94	3.18	0.99	2.80
Fed funds rate	0.80	5.08	0.89	4.06	0.93	3.60	0.99	3.21	0.96	2.67
Yield spread	0.82	4.69	0.97	3.70	1.01	3.41	1.10	2.97	1.04	2.49
Loans	0.88	4.56	0.95	3.82	1.01	3.46	1.12	2.99	1.07	2.64
Macro(PLS)	0.79	4.32	0.94	3.60	1.04	3.36	1.13	2.64	1.06	2.09
NAHB	0.86	3.68	1.06	3.51	1.51	2.95	1.31	2.37	1.23	2.01

Table 3. Forecast encompassing tests: The reverse hypothesis. The table reports the Harvey et al. (1998) test statistic t_{HLN}, where the null hypothesis is that the weight on benchmark model λ is zero against the alternative hypothesis that it is positive. The test statistic is computed using the Newey-West estimator with h lags, where h is forecast horizon in months**.**

	$h=1$		$h=3$		$h=6$		$h=9$		$h=12$	
	λ	t_{HLN}	λ	t_{HLN}	λ	t_{HLN}	λ	t_{HLN}	λ	t_{HLN}
AR(SIC)	0.23	1.74	0.10	0.69	-0.07	-0.37	-0.11	-0.58	-0.05	-0.29
Industrial prod.	0.19	2.21	0.08	0.86	0.04	0.35	-0.06	-0.40	-0.02	-0.11
Employment	0.18	2.04	0.07	0.69	0.05	0.37	-0.06	-0.36	-0.02	-0.10
Unemp. rate	0.10	1.04	0.03	0.30	-0.05	-0.44	-0.16	-1.54	-0.01	-0.18
Inflation	0.20	2.27	0.08	0.75	0.03	0.27	-0.04	-0.26	0.00	0.01
Hours	0.18	2.19	0.07	0.73	0.04	0.34	-0.04	-0.26	0.00	0.00
Building permits	0.21	2.72	0.10	1.11	0.06	0.52	0.00	0.00	0.02	0.11
Housing starts	0.19	2.23	0.09	0.90	0.05	0.43	-0.02	-0.14	0.01	0.09
Price-rent ratio	0.23	2.40	0.12	1.58	0.09	0.93	0.06	0.67	0.04	0.13
Fed funds rate	0.20	2.23	0.11	1.08	0.07	0.56	0.01	0.05	0.04	0.26
Yield spread	0.18	2.14	0.03	0.33	-0.01	-0.07	-0.10	-0.86	-0.04	-0.29
Loans	0.12	1.74	0.05	0.54	-0.01	-0.12	-0.12	-0.85	-0.07	-0.45
Macro(PLS)	0.21	0.89	0.06	0.50	-0.04	-0.28	-0.13	-1.12	-0.06	-0.63
NAHB	0.14	1.87	-0.06	-0.43	-0.51	-2.19	-0.31	-1.65	-0.23	-1.42

Table 4. Inclusion frequencies. The table reports the inclusion frequencies for the 15 most frequently selected search terms by elastic net for $h = 1$.

	Incl.freq.
Loan interest calculator	0.417
Free foreclosed homes	0.417
Tn land for sale	0.417
Florida homes for sale by owner	0.304
Foreclosure auction	0.278
Amortization	0.235
Buying a house steps	0.217
Home for sale by owner	0.200
What is foreclosure	0.200
House for rent	0.191
Moving companies	0.191
Monthly mortgage payments	0.183
First time home buyer grant	0.183
Mortgage foreclosure	0.174
Houses for sale	0.165

Table 5. Bootstrap p-values. The table reports bootstrap p-values for the Google factor in forecasting future house price growth.

	$h = 1$	$h = 3$	$h = 6$	$h = 9$	$h = 12$
Actual R^2_{OoS}	49.1%	58.4%	66.3%	65.6%	53.1%
Empirical p-value	0.000	0.000	0.000	0.000	0.000

Table 6. The out-of-sample predictive power of Google search activity and alternative predictive models using seasonally unadjusted data. The table reports the Campbell and Thompson (2008) out-of-sample R^2 (R^2_{OoS}) and the Diebold and Mariano (1995) t-statistic (t_{DM}), which is computed using the Newey-West estimator with h lags where h is forecast horizon in months. The null hypothesis is that the R^2_{OoS} is equal to zero or negative and the alternative hypothesis is that it is positive.

	$h=1$		$h=3$		$h=6$		$h=9$		$h=12$	
	R^2_{OoS}	t_{DM}	R^2_{OoS}	t_{DM}	R^2_{OoS}	t_{DM}	R^2_{OoS}	t_{DM}	R^2_{OoS}	t_{DM}
Google	51.9	3.24	54.3	2.96	50.5	2.06	50.7	2.09	52.1	2.21
AR(SIC)	34.7	2.80	12.7	1.28	-41.6	-2.95	3.6	0.32	20.3	2.56
Industrial prod.	-1.4	-2.13	-0.5	-0.48	0.1	0.16	-0.3	-0.26	-0.3	-0.35
Employment	3.2	0.86	0.0	-0.03	-1.5	-1.26	0.3	0.22	-1.3	-1.17
Unemp. rate	-8.3	-1.17	-7.7	-0.86	-7.8	-0.68	-12.0	-1.07	-31.3	-1.30
Inflation	13.0	2.64	1.9	0.30	-2.3	-1.21	1.3	0.73	-0.8	-0.99
Hours	-0.4	-0.36	1.4	0.89	2.2	1.64	1.0	1.60	-0.5	-1.23
Building permits	12.3	1.85	6.9	1.35	1.4	1.06	0.0	-0.04	-1.3	-0.71
Housing starts	18.3	2.70	8.2	1.35	0.0	0.04	0.2	0.15	-2.4	-1.17
Price-rent ratio	-9.9	-1.24	-2.6	-0.24	-1.2	-0.07	-5.6	-0.24	-14.6	-0.57
Fed funds rate	1.2	0.30	5.7	1.51	6.4	1.34	8.6	1.99	7.9	1.66
Yield spread	-4.5	-1.33	-9.5	-1.69	-9.0	-1.16	-16.4	-1.07	-15.7	-0.75
Loans	-8.0	-1.14	-14.4	-1.33	-9.1	-0.94	-6.6	-0.76	-8.3	-0.81
Macro(PLS)	26.6	3.06	14.2	1.44	4.7	0.31	-9.0	-0.43	-35.3	-0.78
NAHB	-13.3	-0.95	1.2	0.05	7.1	0.42	-26.5	-1.02	-35.6	-1.19

Table 7. The out-of-sample predictive power of Google search activity for the 20 largest states. The table reports the Campbell and Thompson (2008) out-of-sample R^2 (R^2_{OoS}) and the Diebold and Mariano (1995) t-statistic (t_{DM}), which is computed using the Newey-West estimator with h lags, where h is forecast horizon in months. The null hypothesis is that the R^2_{OoS} is equal to zero or negative and the alternative hypothesis is that it is positive.

	$h=1$		$h=3$		$h=6$		$h=9$		$h=12$	
	R^2_{OoS}	t_{DM}	R^2_{OoS}	t_{DM}	R^2_{OoS}	t_{DM}	R^2_{OoS}	t_{DM}	R^2_{OoS}	t_{DM}
Arizona	57.0	2.63	59.7	2.00	63.1	1.64	47.3	1.24	8.8	0.31
California	35.4	1.52	39.4	1.34	56.6	1.56	32.6	0.98	21.3	0.63
Florida	68.3	3.43	72.3	2.59	78.2	2.24	78.2	2.01	74.7	2.18
Georgia	45.5	2.34	61.1	2.23	60.2	2.04	55.4	1.77	37.2	1.23
Illinois	36.5	2.10	52.6	2.07	63.0	2.19	57.8	2.01	46.3	1.68
Indiana	-12.6	-0.34	25.1	1.12	66.0	2.97	69.2	2.95	68.9	3.16
Massachusetts	22.6	1.26	43.8	1.71	57.1	2.04	55.7	2.09	40.6	1.73
Maryland	44.2	2.21	55.5	1.98	53.1	1.53	55.1	1.48	39.7	1.30
Michigan	42.8	2.28	55.8	2.10	70.3	2.09	73.8	1.90	56.8	1.78
Missouri	28.8	1.87	35.4	1.83	42.2	1.67	48.4	1.60	38.7	1.22
North Carolina	64.9	4.64	70.1	4.37	69.2	3.79	69.7	3.85	59.7	2.47
New Jersey	39.8	2.00	22.2	0.84	49.5	1.88	57.6	2.65	54.4	3.46
New York	22.3	1.58	45.2	2.28	68.1	3.40	80.7	4.96	77.0	4.81
Ohio	7.0	0.57	46.9	2.17	67.5	2.89	54.7	1.74	55.3	1.59
Pennsylvania	30.2	2.06	46.1	2.46	59.0	3.80	56.3	3.16	57.5	2.65
Tennessee	59.3	4.67	66.4	3.42	67.8	3.16	65.8	3.06	51.8	1.93
Texas	55.9	3.94	59.3	2.95	54.1	2.42	41.6	1.53	13.8	0.47
Virginia	20.9	1.01	-1.8	-0.01	20.5	0.57	23.7	0.55	51.1	1.28
Washington	67.5	5.02	60.5	3.43	69.1	3.47	56.1	2.62	32.5	1.16
Wisconsin	20.9	1.32	37.5	1.78	57.3	2.70	59.1	3.14	53.6	3.05

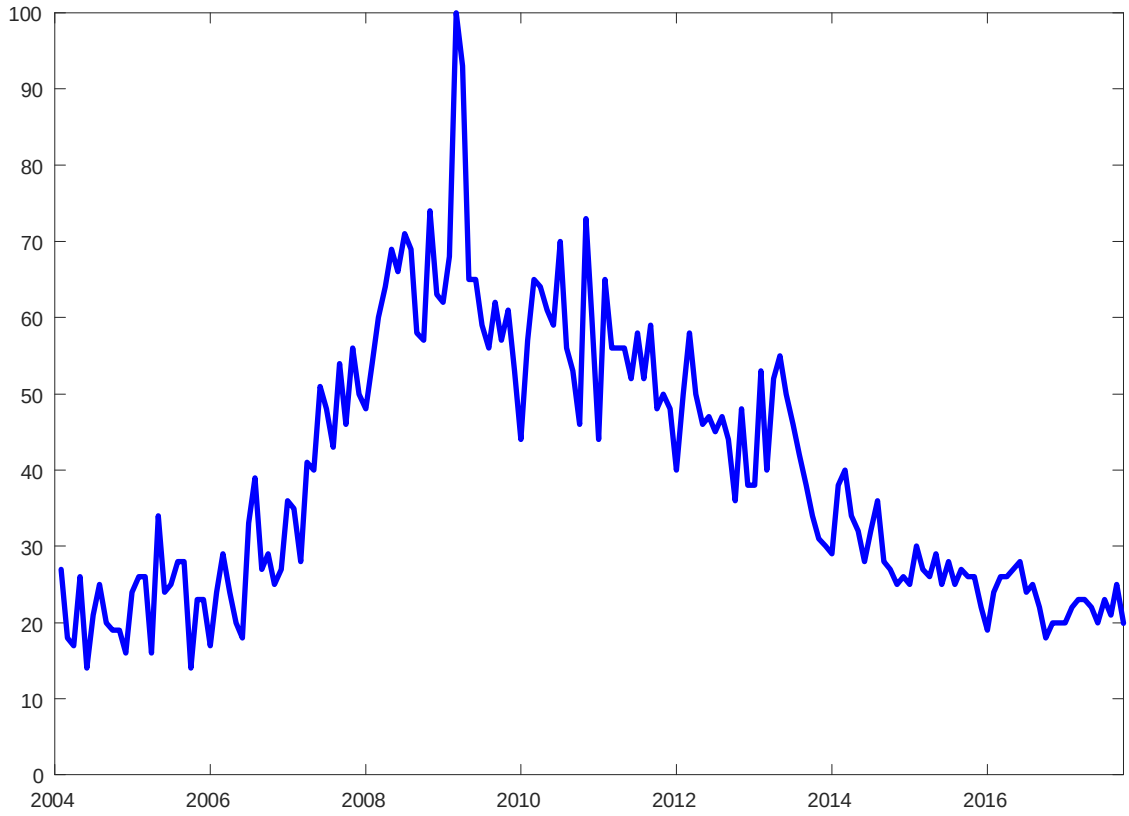

Figure 1. Google search activity for the term "mortgage foreclosure".

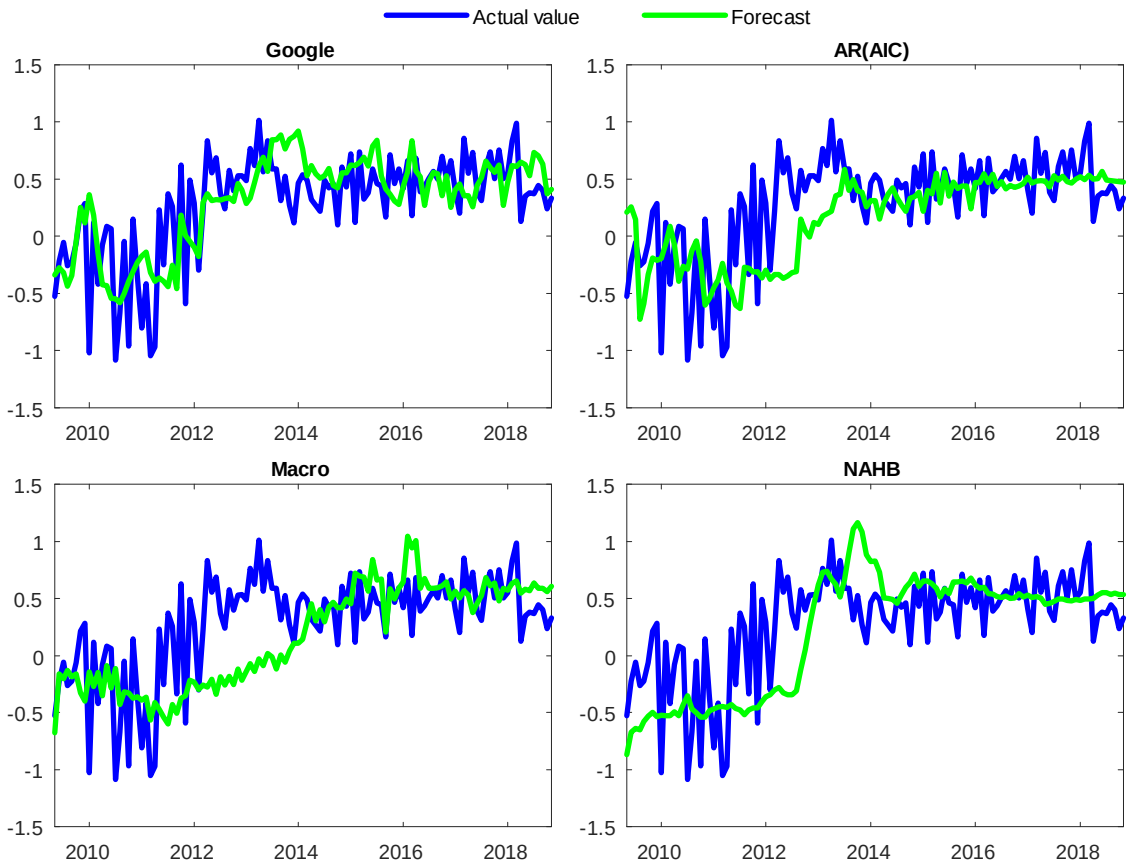

Figure 2. Forecasts of seasonally adjusted house price growth. The figure shows the monthly growth rate in the seasonally adjusted FHFA Purchase-Only Index and forecasts for $h = 1$ using the Google factor, an AR(AIC) model, a common macroeconomic factor constructed using PLS and the survey-based Housing Market Index by NAHB.

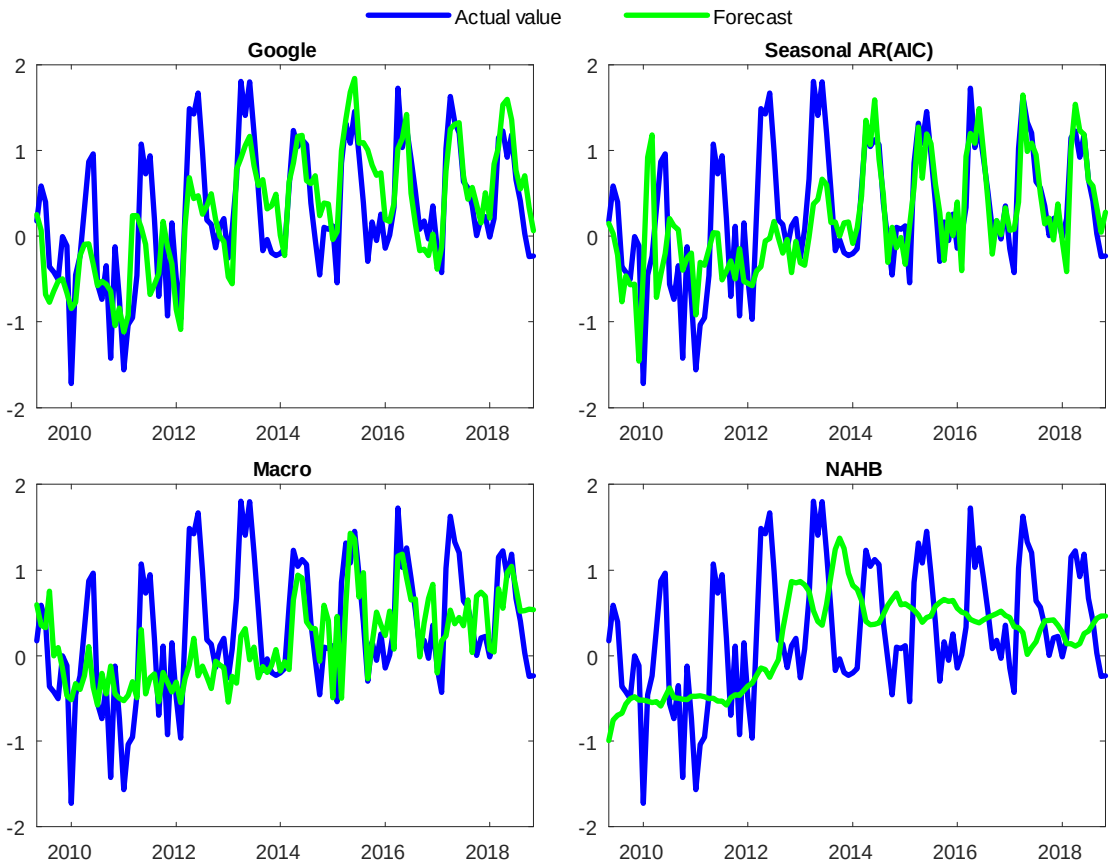

Figure 3. Forecasts of seasonally unadjusted house price growth. The figure shows the monthly growth rate in the seasonally unadjusted FHFA Purchase-Only Index and forecasts for $h = 1$ using the Google factor, an AR(AIC) model, a common macroeconomic factor constructed using PLS and the survey-based Housing Market Index by NAHB.

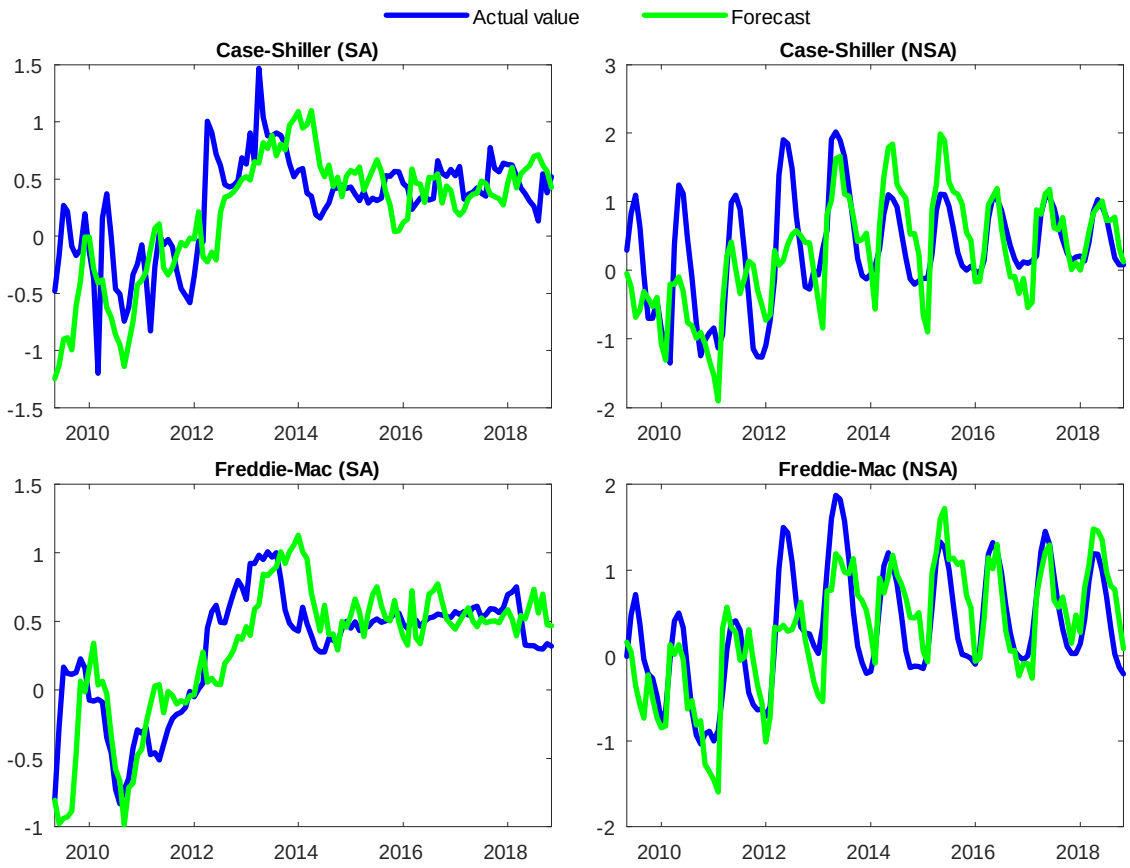

Figure 4. Forecasts of other house prices indices. The figure shows the monthly growth rate in both the seasonally adjusted (SA) and seasonally unadjusted (NSA) Freddie Mac House Price Index (Freddie-Mac) and the S&P/Case-Shiller House Price Index (Case-Shiller) as well as forecasts for $h = 1$ using the Google factor.

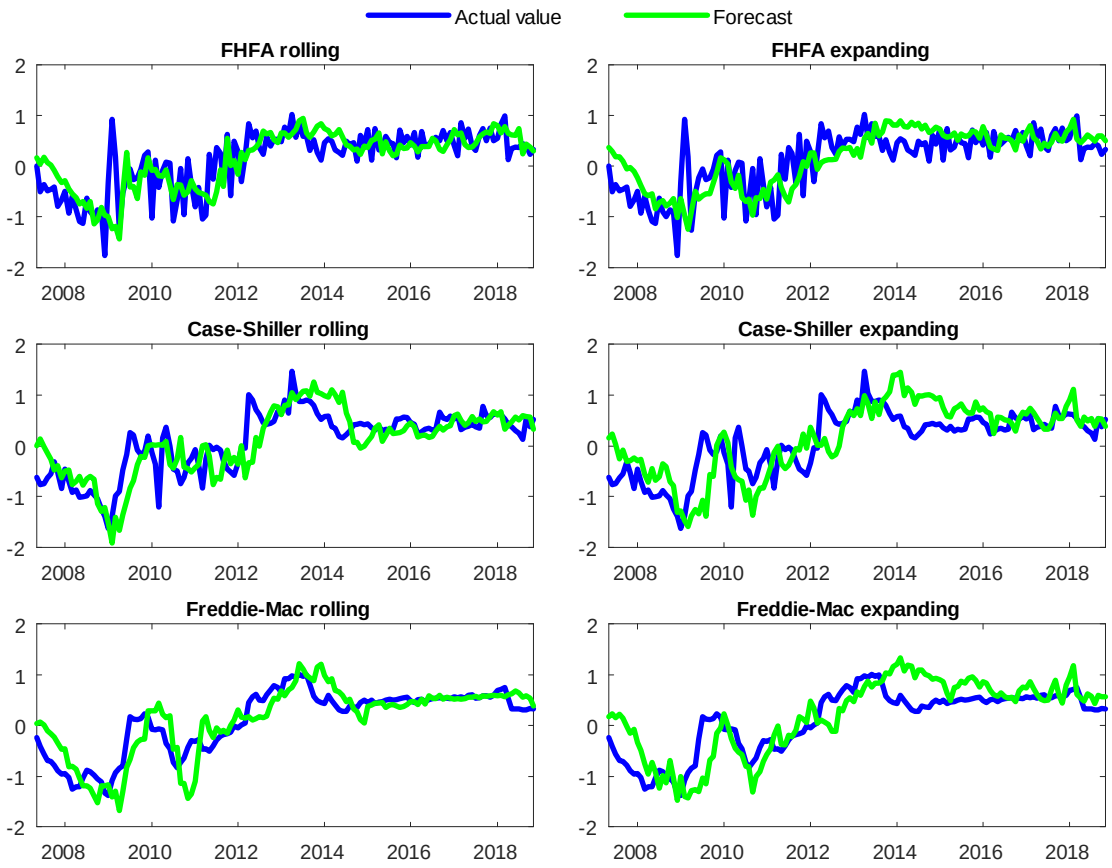

Figure 5. Forecasts using rolling and expanding windows. The figure shows the monthly growth rate in the seasonally adjusted FHFA Purchase-Only Index, Freddie Mac House Price Index (Freddie-Mac) and the S&P/Case-Shiller House Price Index (Case-Shiller) as well as forecasts for $h = 1$ using the Google factor. The figure shows forecasts constructed using a three-year rolling estimation window (rolling) and an expanding estimation window using three years of observations for initial estimation.

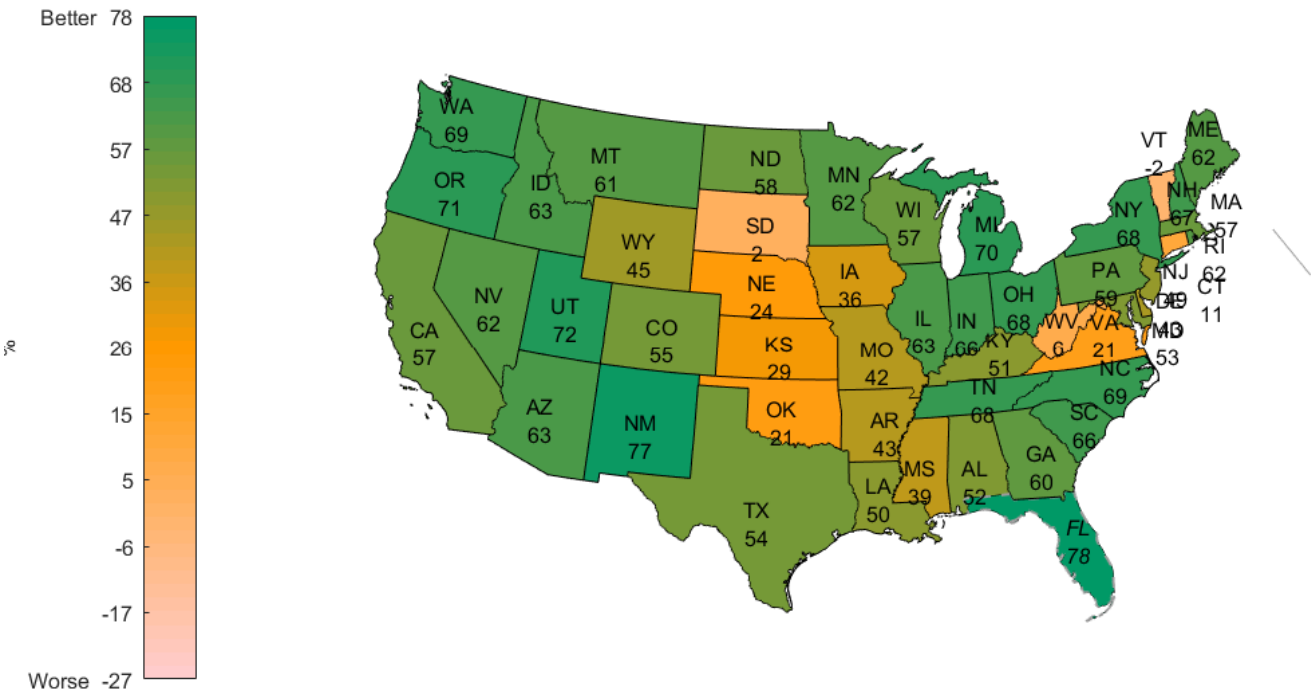

Figure 6. Out-of-sample predictive power of Google search activity across U.S. states. The figure shows the R^2_{OoS} for forecasting growth rates six-months ahead in the seasonally adjusted Freddie Mac House Price Index across U.S. states using the Google factor.

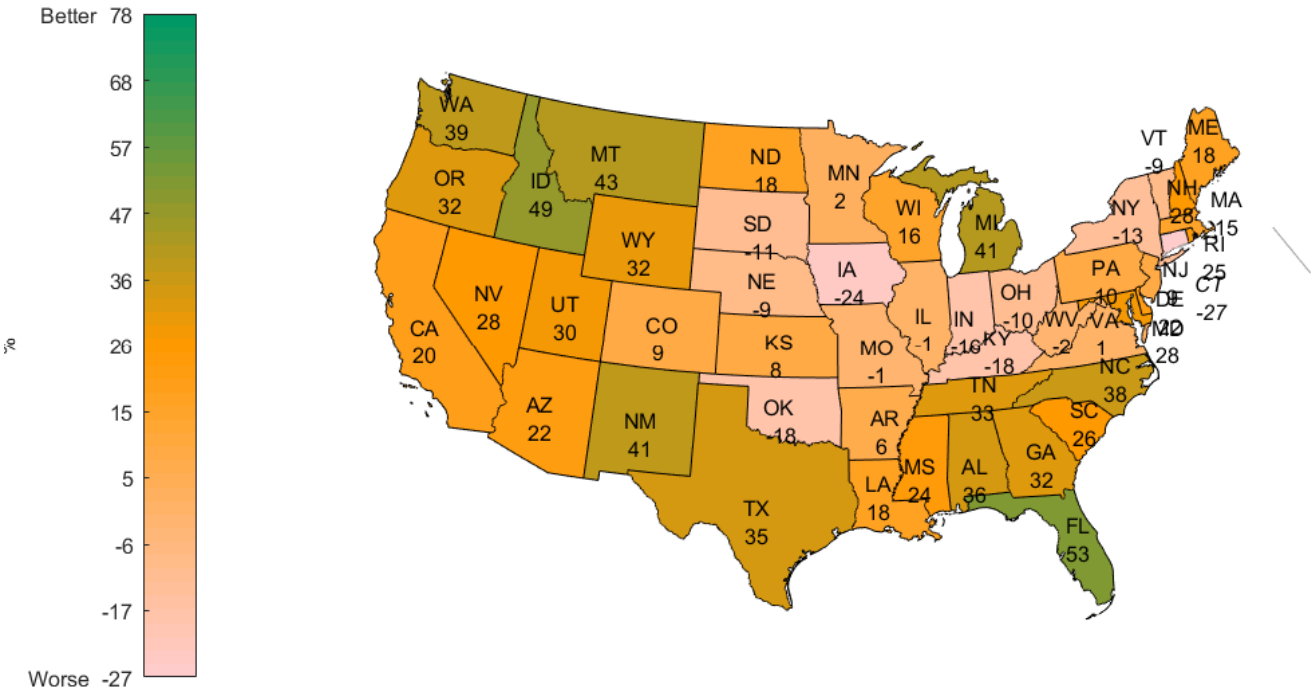

Figure 7. Out-of-sample predictive power of an AR(AIC) model across U.S. states. The figure shows the R^2_{OoS} for forecasting growth rates six-months ahead in the seasonally adjusted Freddie Mac House Price Index across U.S. states using an AR(AIC) model.