Googling For a New Home: Forecasting the Dynamics of the Dutch Housing Market using Google Search Data

Eric Schaap [1] & Stephan Smeekes [2]

- Working Paper -

Maastricht Centre of Real Estate

Version August 2019

[1] Corresponding Author, Email: e.schaap@maastrichtuniversity.nl.

Department of Marketing and Supply Chain Management, School of Business and Economics, Maastricht University, Maastricht, The Netherlands

[2] Department of Quantitative Economics, School of Business and Economics, Maastricht University, Maastricht, The Netherlands

Abstract

This paper studies whether Google Trends data contains additional information not captured by common macroeconomic variables and, hence, is able to improve the forecasting performance of prediction models for two different indices of the Dutch housing market, namely the house price index and the flow of residential mortgages. Real estate related search terms are selected based on a novel approach using the output of Google Correlate in combination with pages of the web directory *Startpagina.nl*. Evaluating post-lasso OLS estimators for different regularisation parameters, the empirical findings suggest that there is not enough evidence to support the hypothesis that the inclusion of Google search queries significantly increases the forecasting accuracy across short and long forecast horizons. However, the augmented models prove to perform better than the models comprising the search index of a predefined Google Trends category when predicting the residential mortgage flows. Therefore, it is concluded that observing specific search terms is of particular interest to government officials as they give insights into the the search behaviour of Dutch citizens in response to policy changes affecting the housing market.

Contents

List of Figures

List of Tables

1 Introduction

The root causes of the global financial and economic crisis of 2007/2008 are partly traced back to the United States housing bubble which pulled the national economy into a deep recession. At the time, many experts, researchers, and government officials failed to predict the collapse of the American real estate market despite consistent observations and measurements of key macroeconomic indicators and affecting factors (Kouwenberg and Zwinkels 2014). Hence, the availability of additional information which is not captured by macroeconomic variables is of great importance to policymakers as it supports a proactive decision-making process.

At this point, private sector companies and social media platforms, e.g. Google, Facebook, Twitter, MasterCard etc., provide new sources of data which potentially hold information on real-time economic activity. This new type of data is generally referred to as big data. The fast-growing hype around big data entails the development of technological tools which enable the extraction of the information or even knowledge from this kind of data (Gandomi and Haider 2015). In this paper, two examples of such tools serve as the primary sources of data: Google Correlate and Google Trends. Both are based on the same query data but return two different outputs: Google Correlate lists those queries which are most highly correlated with a given economic time series, while Google Trends supplies an index of search frequencies over time for different predefined query categories or individual queries (Stephens-Davidowitz and Varian 2014). The availability of these tools represents an opportunity to gather new intelligence on the level and growth of economic indicators by exploiting the real-time data provided. It should come as no surprise that a large academic literature has arisen on the use of big data sources in the context of economic predictions and contemporaneous forecasting, also known as *nowcasting* (Choi and Varian 2011). The aim of this paper is to verify whether the incorporation of Google query data in predictive models improves the short- and long-term forecasting of selected Dutch housing market indices in comparison to a benchmark model which is based only on macroeconomic indicators. The results of the analysis show whether Google search queries hold additional information not captured by the macroeconomic variables.

In the Netherlands, Google dominates the search engine market as the company has maintained a market share of above 90% for the past decade. In addition, Google is often used as a first directory to the website of interest because internet users commonly google a website first and then click on the appropriate link in the search results, rather than directly typing the URL in the address bar (Askitas and Zimmermann 2009). As argued for in the following sections, this observation also holds for Dutch housing websites, which justifies the approach taken in this paper. It seems reasonable to assume that Google searches related to the housing market are driven by both the supply and demand side of the market. Consequently, it is likely that there is a high number of distinct queries which could be used as input for the predictive models developed in this paper. Using many more variables can lead to the issue of overfitting as there might be not enough observations in the data set. To overcome this problem, the lasso method is applied, a shrinkage method which simultaneously selects the appropriate model and estimates the coefficients of the most relevant regressors (Tibshirani 1996).

The focus of this paper is on two distinct indices on the Dutch housing market, namely the house price index of sold dwellings and the flow of residential mortgages extended by Dutch monetary financial institutions (henceforth MFIs). The intention behind using these two economic indicators is to detect potential differences between the supply and demand side of the housing market, based on the assumption that residential mortgages is a proxy for the search behaviour related to the demand for new dwellings.

The remainder of this paper is structured as follows. First, recent academic literature on the introduction of Google data in the practice of economic forecasting is discussed. Second, the approach taken in this paper is motivated by arguing for the evident relationship between the real estate market and the use of the Google search engine. Third, three different types of data used in the empirical analysis are introduced: (i) indicators of the Dutch housing market, (ii) Google search indices obtained by means of Google Correlate and Google Trends, and (iii) macroeconomic data which enter the models as control variables. Fourth, the actual data analysis and computation of the forecasts is carried out. Next, the forecasting results and potential limitations regarding the data and methodology of this study are discussed. The conclusion summarises the main takeaways of this paper and gives suggestions for further research.

2 Literature review

2.1 Forecasting with Google data

In recent academic literature, Google search data has been increasingly employed as a new source of real-time data. In particular, Google Trends, a real-time index of the volume of queries entered into Google, has allowed researchers, academics and economists to overcome the reporting lag of indicators on economic activity.

Choi and Varian (2011) are among the first academics to realise the potential of Google search data to contemporaneously forecast macroeconomic variables such as automobile sales, unemployment claims, travel destination planning and consumer confidence. From an analytical point of view, their approach is rather basic using simple seasonal AR models and including relevant Google search queries. Nonetheless, the authors conclude that these models outperform those that exclude the corresponding queries by 5% to 20%. Although, explicitly focusing on short-term predictions, Choi and Varian (2011, p. 1) find "that queries can be useful leading indicators for subsequent consumer purchases in situations where consumers start planning purchases significantly in advance of their actual purchase decision". This argument certainly holds for the purchase of a new home.

Götz and Knetsch (2018) integrate Google data in bridge equation models to fore-, now- and backcast German GDP. In their work, the authors assess whether Google search data can be used in addition to or as a replacement for survey data in models predicting German GDP growth. In contrast to Choi and Varian (2011), the German economists experiment with and evaluate different variable selection methods to select Google Trends queries and categories before developing prediction models. In particular, ad-hoc methods including subjective selection and Google Correlate (subsection 4.2), factor methods using principal component analysis and partial least squares regressions, and shrinkage methods such as the lasso method are deployed and evaluated. Furthermore, based on preliminary results, Götz and Knetsch (2018) implement a targeted selection of regressors by using the optimal selection method for each GDP component to build an augmented model. The outcomes of their analysis suggest that forecast improvements depend on the variable selection method applied. In addition, forecast accuracy can be enhanced by replacing survey variables with Google indicators, in particular for long-horizon nowcasts and forecasts. This finding is in line with the aforementioned usefulness of Google search queries in high-involvement purchases and supports the application of Google data in the case of the housing market.

Similar to the variable selection methods employed by Götz and Knetsch (2018), L. Yu et al. (2017) follow a two-step framework when analysing the relationship between selected Google Trends queries and global oil consumption. In a first stage, the authors investigate the relationship between the Google search terms and oil consumption using co-integration tests and Granger causality tests. Next, based on the findings from the first stage, several directional and level prediction models are developed by including the relevant search queries, and the results of these models are evaluated and compared. The authors conclude that significant improvements in forecasting performance can be achieved by including Google Trends data, for both directional and level predictions.

A large part of the academic literature on the introduction of Google search data in economic forecasting focuses on predicting national unemployment rates. The research in this field is based on the assumption that individuals who are currently unemployed or who are at the risk of losing their job google queries such as "jobs", "unemployment benefits", or "unemployment office" (Choi and Varian 2011). Indeed, D'Amuri and Marcucci (2017) find empirical evidence that the inclusion of the Google search index of the term "jobs" enhances the forecasting accuracy of developed AR models. In particular, their analysis shows that the model performance increases with the length of the forecast horizon due to the fact that individuals look for an employment long before losing their current job or being reclassified as unemployed. Furthermore, the authors claim that the model including the search data predicts particularly well at the beginning of the Great Recession. The general finding that Google search data improves the forecasting accuracy with respect to the unemployment rate is in line with the studies of Askitas and Zimmermann (2009) and McLaren and Shanbhogue (2011) on the German and British unemployment rates respectively. Moreover, Fondeur and Karamé (2013) prove that this result also holds when studying the unemployment rate of a particular demographic group such as French youth aged 15 to 24 years old. When carrying out the same forecasting exercise for the 25- to 49-years old group and those aged 50 and older, the authors observe a selection bias of the youth using the Internet as primary means for the job search.

2.2 Forecasting in the housing market

Soon after the emergence of new data sources provided by Google, the early work of academics such as Choi and Varian (2011) has been applied to the housing market. Nonetheless, forecasting in the context of the this market induces certain challenges due to the particularities of the housing dynamics which need to be taken into account.

One of the key stylised facts of the housing market is the mean reversion of housing prices in the long-run due to slow construction responses and mean-reverting income shocks (Glaeser and Gyourko 2006). The mean-reverting process starts with an increase in demand following a minor change in a fundamental factor such as the interest rate or income, which is aggravated by the multiplier effect. The positive demand shift cannot be offset by an increase in supply as the latter is rigid in the short run. As a consequence, housing prices grow while developers increase new constructions which will enter the market with a time lag. At the same time, growing prices attract further demand since any type of real estate is defined as a consumer and investment good. Eventually, demand returns to its long-run level, and excess supply and falling prices characterise the market (Augustyniak et al. 2014). Besides the mean reversion, Glaeser and Gyourko (2006) have identified a strong positive serial correlation in house price changes at an annual frequency caused by the positive serial correlation of labour demand shocks. When forecasting housing dynamics "understanding what drives real estate values is no less important than is understanding the pricing dynamics of other asset classes, such as stocks, bonds, commodities, and currencies" (Ghysels et al. 2013, p. 1). Nonetheless, the housing market presents further particularities which differentiates it from other financial markets, e.g. extreme heterogeneity due to location and house attributes, large search and transaction costs, tax considerations, and illiquidity. Consequently, forecasting real estate prices is complicated by the fact that some predicting factors are time-invariant and, hence, cannot be the source of price volatility over time. Furthermore, the choice of measurement index is essential to the task of forecasting as certain indices by construction exhibit serial correlation, and, despite measuring the same pricing dynamics, different indices potentially display different time series (Ghysels et al. 2013).

Wu and Brynjolfsson (2015) took these particularities into consideration when integrating selected Google search queries into their regression models to forecast present and future housing prices in the United States. They find evidence for a strong positive relationship between Google search queries and the future unit sales of housing, and a more moderate positive relationship between search queries and future house prices. Based on these findings, the authors argue that predicting house prices is more challenging than predicting the volume of sales because of the ambiguous effects of Google searches on the chosen house price index. More searches generated by the demand side of the housing market are linked to price increases, whereas a higher search volume of the supply side potentially indicates a decrease in house prices. Consequently, aggregated search queries might be relatively less effective for predicting price changes as compared to predicting sales volume. Furthermore, the scholars also find evidence for a two-way causality as their results suggest that the house sales volume has a lagged positive effect on the search frequency of home appliances.

In a similar way as Wu and Brynjolfsson (2015), Humphrey (2010) applies ordinary least squares (henceforth OLS) regression models to predict local and national home sales in Texas and the United States by integrating selected Google search queries in addition to different macroeconomic variables. The author argues that the impact of Google on the housing market can be represented in a supply and demand model. With Google as a new source of information, increases in the search volume on either side of the market lead to an outward shift of the corresponding curve. Humphrey (2010) hypothesises that, under the assumption of an impact of macroeconomic events on Google searches, the queries may capture information that is contained in macroeconomic indicators. Hence, the author tests the regression models with and without macroeconomic variables to determine whether Google search queries provide any additional information not accounted for by these variables. Indeed, he finds evidence to support his hypothesis as the inclusion of Google search variables improves different measures of fit of the models including and excluding macroeconomic indicators at both local and national level.

McLaren and Shanbhogue (2011) discuss the general applicability of internet search data as economic indicators when they analyse the use of this type of data in the labour and housing markets in the United Kingdom. In particular, the authors include the search volume for "estate agent" when estimating monthly house price growth. Their work is based on the assumption that searches for "estate agent" are dominated by the demand side of the housing market given a positive co-movement between house prices and searches. From the results of their analysis, the authors conclude that the Google search variable has a positive significant effect on monthly house prices and including this variable improves in- and out-of-sample forecasting performance of the model.

With regards to the Dutch housing market, van Dijk and Francke (2018) and Veldhuizen et al. (2016) analyse the relationship between different forms of internet data and indices of the market. In particular, van Dijk and Francke (2018) develop a vector autoregressive (henceforth VAR) model including click data of a leading Dutch housing website called *funda.nl*, real estate prices, and housing market liquidity. The authors use the click numbers to create a market tightness indicator based on the ratio of clicks per house, arguing that the number of clicks is a proxy of the demand within the market, whereas the number of listed houses represents the supply. Their findings suggest evidence for a significant positive relationship between the clicks per house and house prices, and between the click data and the market liquidity measured by the rate of sales. In contrast, Veldhuizen et al. (2016) take advantage of available Google query data as they study the link between the transaction volume in the real estate market and the search term "hypotheek" (i.e. "mortgage"). The authors include up to 11 monhtly lags of the search index in an autoregressive model and compare the explanatory power of the latter to a selected benchmark model. The results of the analysis show that the search indices at no lag, and lags of 1, 6 and 9 months respectively are positively related to the transaction volume. Based on these findings, the authors conclude that buyers of a new home plan their purchase well in advance.

In contrast to previous academic literature, this paper aims to lower the subjectiveness of the choice of Google search queries, which potentially impacts the empirical findings by introducing a novel approach to the selection of relevant search terms using a combination of Google Correlate outputs and hyperlinks from housing-related *Startpagina.nl* pages. *Startpagina.nl* can be defined as a web or link directory, i.e. a website which consists of a catalogue of hyperlinks to listed pages. In addition, the study of two real estate market indices, particularly of the flow of residential mortgages, allows to identify different dynamics between the supply and demand side of the market in terms of online search behaviour. Furthermore, the use of the lasso method enables the inclusion a high number of Google queries without facing the risk of overfitting (Humphrey 2010; McLaren and Shanbhogue 2011; Wu and Brynjolfsson 2015). Besides comparing the developed models to a benchmark model which solely contains macroeconomic variables, the difference in forecasting performance between the use of specific Google search terms and predefined Google Trends categories is analysed.

3 Googling a new home

Before discussing the collection of relevant Google search queries and the applied methodology, this paper motivates the applicability of Google search data to the Dutch housing market.

Figure 1 shows the market share of the four most frequently used search engines on desktop in the Netherlands. From this graph it becomes apparent that Google dominates the market as its share has been almost continuously above 90% for the last decade whereas other search engines including *bing* and *Yahoo!* divide the remaining fraction among each other. Based on this observation, it is concluded that the sampling data from Google searches is representative of the Dutch population.

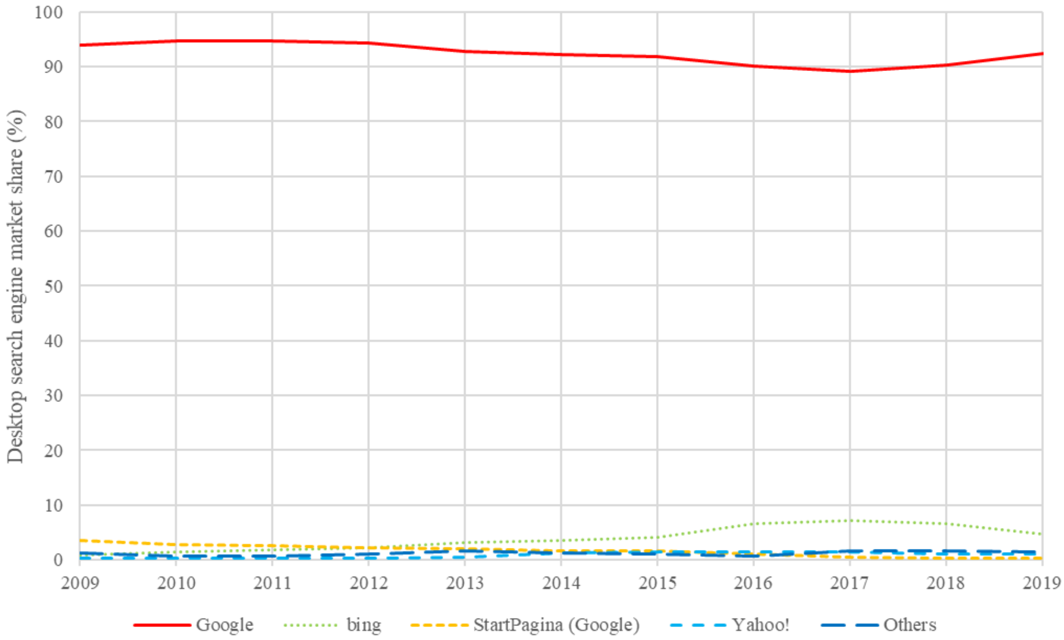


Figure 1: Desktop search engine market share in the Netherlands, 2009-2019. Source: StatCounter.

After having proven the prevalence of Google among search engines, the link between the market leader and the Dutch housing market is established. Table 1 lists some of the most frequently visited Dutch real estate websites, and details the share of their traffic coming from the Netherlands and the percentage of incoming traffic from organic search. The latter is on average 46.19% as values range from as high as 65.94% for *huislijn.nl* to 23.72% for *funda.nl*. In other words, almost half of the traffic on these websites results from (predominantly Dutch) users finding their way to the website through the use of a search engine - most likely Google.

Based on the the evidence provided in Figure 1 and Table 1, the assumption deduced is that for a large share of the Dutch population the search for a new home starts with the Google search engine. This assumption motivates the use Google search queries to predict changes in the housing market.

Website	Share of traffic from NL (%)	Share of traffic from organic search (%)
huislijn.nl	91.98	65.94
huiszoeker.nl	92.31	55.94
makelaarsland.nl	94.23	45.04
jaap.nl	90.99	40.29
funda.nl	92.92	23.72
Average	92.49	46.19

Table 1: Traffic on different housing websites in the Netherlands (desktop only), Jan/2019 - Mar/2019. Source: SimilarWeb.

4 Data

Following the motivation for the intended approach, this section discusses the data required to conduct this approach. In particular, this paper considers different indices on the historical development of the Dutch housing market and examines their applicability to the purpose at hand. Further subsections address the collection and selection of Google search queries.

4.1 Housing market data

Data on the economic and financial indicators related to the dynamics of the housing market are retrieved from the databases of Statistics Netherlands (called

Centraal Bureau voor de Statistiek, henceforth CBS) and the Dutch central bank (called *De Nederlandse Bank*, henceforth DNB). The former provides data on the volume of transactions and the price developments of the market in question, whereas the latter publishes the financial data on residential mortgages.

The data on residential mortgage loans exclusively captures loans which have been issued by banks resident in the Netherlands (excluding the central bank) to Dutch households. The data has been adjusted for securitisations transferred from these banks to special purpose vehicles, and for breaks, errors, and omissions resulting from a variety of factors such as re-classifications of sectors or types of instruments, or takeovers of banks (De Nederlandse Bank 2019). Figure 2 shows the trend of both the monthly stocks and flows of residential mortgages extended by Dutch MFIs from January 2004 to January 2019. The graph indicates that following an initial increase from 2004 until 2011, the stocks have maintained the same level ever since. Consequently, the flows of mortgage loans show a downward trend towards zero for the considered time period. The empirical methods applied in the next section require stationary time series data, which is achieved in a first step by differencing the data to remove the trend. For this reason, the flows of residential mortgages seem to be appropriate for the purpose of this paper as this index potentially requires only first-order differencing in contrast to the second-order differencing needed for the stocks data. Section 5.1 further explains the conditions of stationary time series data and the steps taken to fulfill these conditions.

CBS publishes several indicators on the dynamics of the housing marketing including among others the transaction volume measured by the number of dwellings sold, the average purchase price of residential properties, and the house price

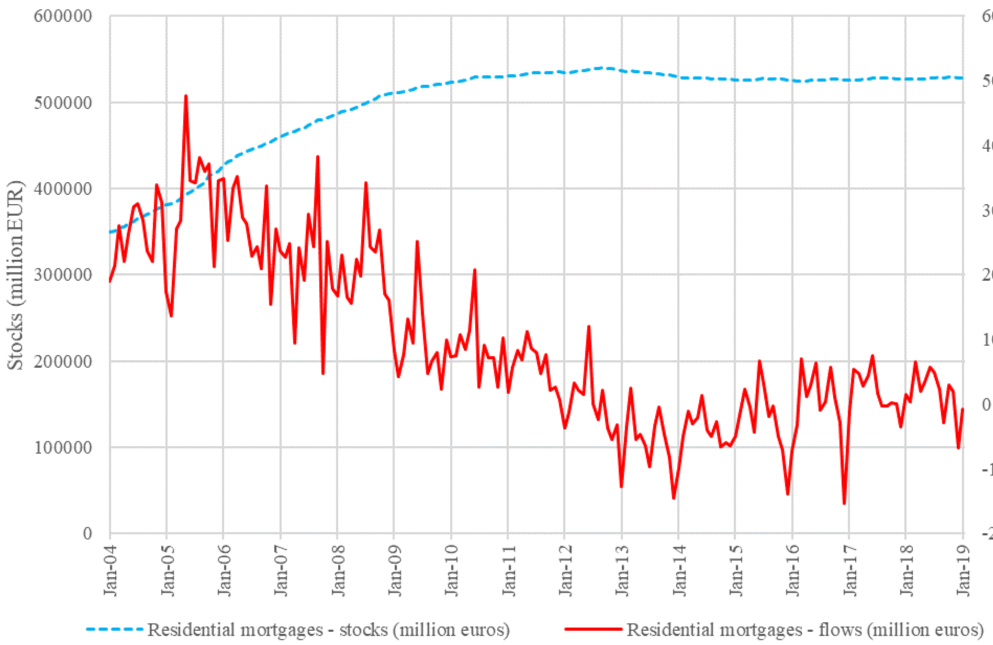


Figure 2: Residential mortgages extended by Dutch MFIs to Dutch households, adjusted for securitisations, Jan/2004 - Jan/2019. Source: DNB.

index (henceforth HPI). Figure 3 depicts the development of two of these indicators, i.e. the average purchase price and the number of dwellings sold. Both indices follow a similar trend over the considered time period: an upward sloping trend preceding the global economic and financial crisis in 2008/2009, a decrease in transactions and prices in the wake of the crisis, and a recovery of the economy observed by an upward trend since 2014. At this point it should be noted that the average purchase price might be a misleading indicator regarding the true price development of the real estate market as it does not capture the monthly change in the characteristics of dwellings sold. Hence, if the quality of dwellings is higher in a particular month than in the previous month, the average price of houses sold is likely to increase. However, it would be wrong to conclude that the actual prices have risen by that same amount. To account for the change in quality of dwellings sold, the HPI is calculated based on the Sale Price Appraisal Ratio (henceforth SPAR) method. The latter allows to capture changes in the relation between the average purchase price and the average quality of dwellings sold measured by the WOZ value which equals the value of a dwelling if the sale would occur on the reference date (van der Wal and Tamminga 2008). Figure 4 shows that both price indices follow the same trend as previously described, except that the development of the HPI is less volatile and noticeably smoother.

Section 2.2 discusses key stylised facts of real estate markets, including the mean reversion of house prices in the long run, and the positive annual serial correlation in price changes. Although the developments of the HPI and the average purchase price appear to be mean reverting, it can be questioned whether the underlying process is in line with the reasoning described in section 2.2 as the impact of the financial crisis and the rising inflation rates on house prices should not be neglected. Mean reversion is one of the conditions for a time series to be stationary; however, as proven in sections 5.1.1 and 5.1.2, the indices on the Dutch house prices are indeed non-stationary and, hence, appear to not follow a mean-reverting process. With regards to the serial correlation in house price changes at an annual frequency, the autocorrelation estimate for the monthly change in the HPI at lag 12 is indeed equal to 0.203.

In their paper Wu and Brynjolfsson (2015) argue that, based on their empirical findings, there is an ambiguous effect of Google searches on the price indices of dwellings sold. Following economic theory, searches from the demand side of the real estate market equal an outward shift of the demand curve, ultimately increasing house prices. The opposite applies to the searches generated by the supply side (Humphrey 2010). To study this ambiguous effect, this paper focuses on two

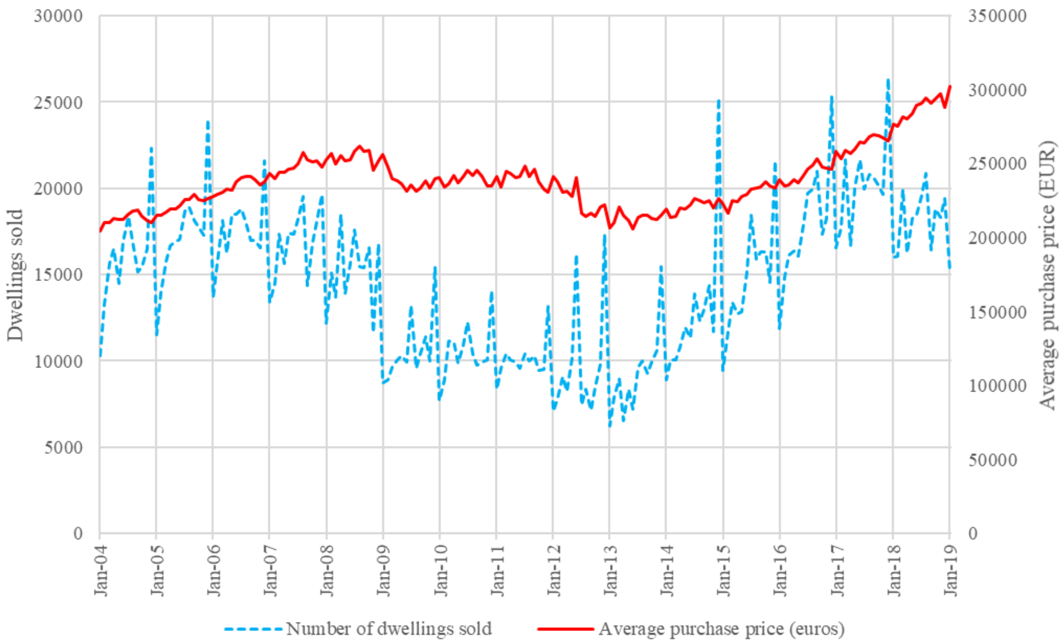


Figure 3: Number of dwellings sold and average purchase price, Jan/2004 - Jan/2019. Source: CBS.

indices of the Dutch housing market: the HPI of dwellings sold, and the flow of residential mortgages extended by Dutch MFIs. The flow of residential mortgages is a proxy of the demand side of the market as mortgage financing is almost university among Dutch households. The tendency of taking on a mortgage results from the fact that the interests can be deducted from taxes up to a maximum period of 30 years (*The Dutch Mortgage Market* 2014).

4.2 Google Correlate

Google Correlate is one of several tools provided by Google which have the potential to introduce new data sources into economic, social and socioeconomic science (Stephens-Davidowitz and Varian 2014). It returns search queries which are most highly correlated with other queries or with user-entered data series (i.e. the target

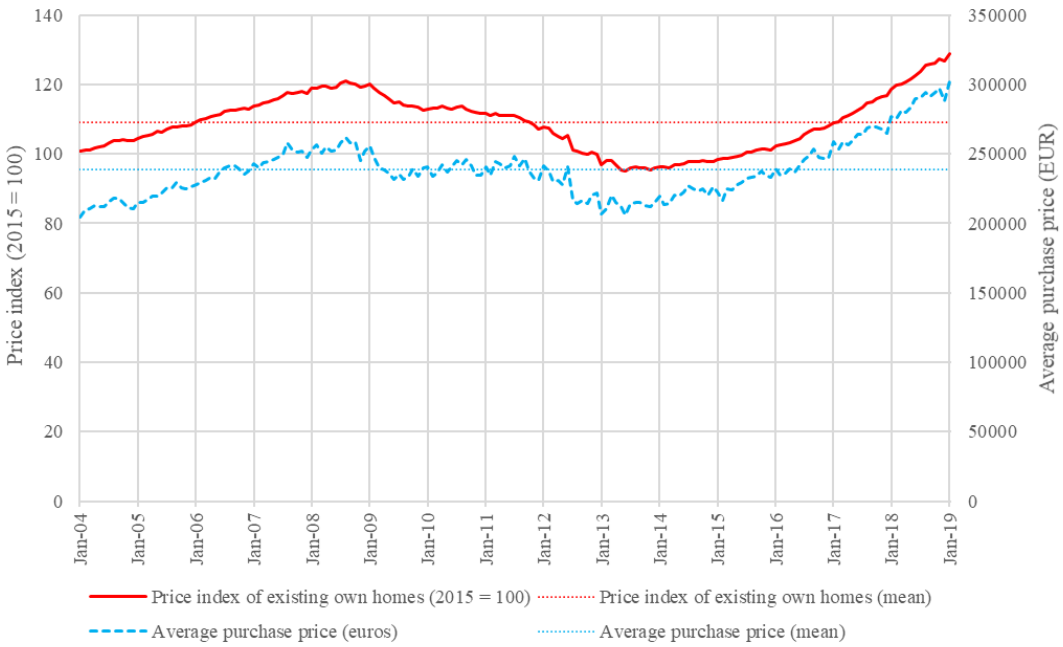


Figure 4: Price index of existing own homes (2015 = 100), Jan/2004 - Jan/2019. Source: CBS.

series) across time for a specific geographical area. The correlation is computed based on the Pearson correlation coefficient between the target series and the frequency time series for every search query in the Google database (Mohebbi et al. 2011).

Despite being a potentially powerful tool, many academics tend to subjectively select search queries which are assumed to be highly correlated with the targeted time series based on economic reasoning rather than relying on the output of Google Correlate (Humphrey 2010; McLaren and Shanbhogue 2011; Wu and Brynjolfsson 2015; L. Yu et al. 2017). To limit the impact of the subjectiveness of the selection on the results of the empirical analysis, this study collects the Google data by entering the aforementioned indices on the Dutch housing market into Google Correlate and gathering the returned correlated queries. Based on Choi and Varian (2011)'s argument about the usefulness of queries as leading indicators for purchases involving extensive decision-making, lagged correlated queries are collected up to a 12-month lag for each index. The intuition behind the inclusion of lagged search queries relates to the "honest signal" (Pentland 2010) of intended purchasing decisions demonstrated by preceding searches (Wu and Brynjolfsson 2015). As the purpose of this paper to perform forecasting, the time series data for the HPI and residential mortgage flows entered into Google Correlate has been limited to the period from January 2004 to December 2015. Hence, the time period between January 2016 until January 2019, which accounts for 20% of the entire data set, is left out and used as a holdout in the empirical analysis of this paper.

Following this approach, 2,600 search queries are collected from the Google Correlate output (including duplicate search queries which are correlated with both indices or over multiple lags). In turn, the list of search queries is cleaned, i.e. the queries are split into single words (e.g. "makelaars Eindhoven" is split into "makelaars" and "Eindhoven"), put into lowercase letters, and punctuation, stopwords, special characters, numbers, and particular words related to domain names (e.g. "www" or "nl"), downloads (e.g. "download" or "downloaden"), popular webpages (e.g. "gmail" or "msn"), and Dutch cities are removed. This first filtering of the search queries results in a list of 1,017 distinct search terms for which the word frequencies are counted. Ranked in decreasing frequency order, the top 5% of the list includes terms such as "bluetooth" (frequency: 35) and "server" (frequency: 16) which are likely to reflect spurious correlations with the housing market indices. Nonetheless, "makelaar" (frequency: 14) also figures in that top section of the list and it seems very plausible that this term reflects a Google search done in the context of a transaction on the real estate market.

Following these lines of reasoning, words which suggest spurious relationships are excluded from the list. To reduce the amount of manual work and the subjectiveness in the approach, real estate related web pages of *Startpagina.nl* which group hyperlinks according to a common topic, are used as filters. On these pages, companies or individuals can customise the hyperlink to their own website, e.g. the real estate agency *Perfect Housing* is listed as "Verhuur via Perfect Housing" rather than "perfecthousing.nl". This example highlights the advantage of this filtering approach: instead of just selecting the agency name, the word "verhuur" is also retained as a relevant search term. For the case at hand, the topics "hypotheek" (i.e. "mortgage"), "huis" (i.e. "house"), "woonwinkels" (i.e. "furniture stores"), "makelaar" (i.e. "estate agent") and "woning-koopwoning" (i.e. "home for sale") are assumed to comprise the majority of real-estate related terms and the listed hyperlinks are stored as input for the filtering of the search queries. Links to the corresponding web pages can be found in Appendix A. The created *Startpagina.nl* lists are cleaned in the same way as the retrieved Google search queries to ensure an accurate matching. Applying this method of refinement results in a list of 36 single search terms including among others "makelaar", "hypotheek", "funda" or "huis". Nonetheless, based on common sense, the list is reduced as some words can either be grouped together because they are similar in meaning (e.g. "hypotheek" and "hypotheken", or "leen" and "bakker"), or can be excluded as they do not seem to be explicitly linked to the housing market, e.g. "je", "limburg", "power", "tip" or "top". Besides, the search terms which have not been selected by the filtering approach have been manually checked and, again based on common sense, added to the list of retained searches, e.g. "studio", "architecten" or "huren". Furthermore, the final list is extended by queries which have been subject of relevant academic literature. The full list of all 33 search terms retained for the empirical analysis of this paper is provided in Table 2. About 50% of all the terms are obtained using the previously described filtering approach.

4.3 Google Trends

After having filtered correlated search terms from the Google Correlate output, the list in Table 2 is used as input for Google Trends. Google Trends is another tool recently developed by Google and uses the same data as Google Correlate, however, it returns a different output: an index of search activity for predefined query categories or individual queries. This index represents the queries which include the search term of interest as a fraction of the total number of queries for a selected geographical location at a particular point or period in time. Hence, the index represents a relative measure and is normalised to a scale between 0 and 100. The relative nature of this index implies that a downward trend for a particular search term over time does not necessarily mean that the absolute search frequency for this term has dropped but rather that it has decreased as a percent of all searches. Furthermore, it should be noted that the data returned by Google Trends is retrieved from a sample of the total Google search corpus and that the data on this sample is cached each day. This means that downloading data from Google Trends on a particular day might result in slightly different values than for a data set which has been downloaded the day before. However, the database of Google is large enough to minimise these differences (Stephens-Davidowitz and Varian 2014). This research retains the full data sample from January 2004 to January 2019 for the Netherlands.

Search terms	Startpagina.nl	Subjective	Literature
abnamro	X		
leen bakker	X		
bouw	X		
directwonen	X		
funda	X		
huis	X		
hypotheek	X		
makelaar	X		
nieuwbouw	X		
plattegrond	X		
rabobank	X		
vastgoed	X		
vbo	X		
verhuisbedrijven	X		
verhuur	X		
wehkamp	X		
architecten		X	
bkr		X	
bouwfonds		X	
haagwonen		X	
herbouwwaarde		X	
huren		X	
meeus		X	
meubelen		X	
remco		X	
stadskamer		X	
stienstra		X	
studio		X	
huis inspectie			X
huizen			X
huizen te koop			X
onroerend goed			X
verkoopstyling			X

Table 2: Retained search terms.

4.4 Macroeconomic data

In addition to the housing market and Google search data, the developed models include selected control variables. In particular, data on the following macroeconomic indicators is collected: inflation, unemployment rate, mortgage rate, long-term interest rate, industrial production, and construction costs.

Empirical evidence proves that the explanatory power of inflation can account for between 50 to 90% of the variation in real estate prices depending on the horizon of the considered time period (Tsatsaronis and Zhu 2004). In this paper, inflation is approximated by the price changes in consumer goods and services, a measure known as the consumer price index (henceforth CPI) for which the monthly time series data is available in the database of CBS. Similarly, research has shown that the unemployment rate represents an indicator on the demand side of the real estate market which significantly affects house prices (Hlaváček and Komarek 2009). Data on the unemployment rate is also retrieved from the CBS database. Given the aforementioned tax deductibility of the mortgage interests in the Netherlands, data on the mortgage rate and the long-term interest rate on government bonds maturing in 10 years are collected from the databases of the DNB and the Organisation for Economic Co-operation and Development (henceforth OECD) respectively. The OECD also provides information on the industrial production of the Netherlands, an indicator which refers to the output of industrial establishments within different sectors such as manufacturing or electricity. The industrial production has been used in previous academic literature including Google search data and the gross domestic product (henceforth GDP) of a country (Götz and Knetsch 2018), and is often used as a proxy of GDP, hence, reflecting the state of a business cycle and household income (Tsatsaronis and Zhu 2004). Lastly, construction costs enter the models as there is empirical evidence on the positive effect of construction costs on house prices. This effect results from the power of house developers to transfer increased costs to potential house buyers (Xu and Tang 2014). The data on this indicator is collected from the Eurostat database.

Besides the aforementioned macroeconomic indicators, the one-month lag of the dependent variable in consideration is added to the prediction models as control variable because both variables of interest show a significant non-zero partial autocorrelation value for lag 1 after taking first-order differences (see section 5.1.1). This study does not consider the inclusion of the respective second dependent variable in consideration as an additional regressor because conducted Granger-causality tests do not provide enough evidence in favour of a two-way relationship between the HPI and the flows of residential mortgages.

5 Methodology

After having collected the data on the Dutch real estate market, Google search queries and selected macroeconomic indicators, a time series investigation is employed on the gathered data set before developing different predictive models using the lasso method.

5.1 Data preparation

5.1.1 First-order differences

As the data at hand is of time series nature, the conditions of stationarity need to be verified before proceeding with the empirical analysis. For a time series to be (weakly) stationary, its first and second moments need to be time invariant; in other words, the time series should represent a mean reverting process. Generally, this condition fails in presence of a unit root (Brockwell et al. 2002). To test for a unit root, the augmented Dickey-Fuller (henceforth ADF) test is applied to the data set of this research (Harris 1992). Appendix B lists the Dickey-Fuller values in the second column. The results show that for the majority of the variables in the data set the null hypothesis of a unit root being present cannot be rejected at standard significance levels. One common solution to this problem is to take first-order differences of the time series data (Hyndman and Athanasopoulos n.d.). After applying this method, the p-values for all the variables drop below 0.01 and, hence, the null hypothesis can be rejected at a significance level of 1%. However, for the HPI the Dickey-Fuller value of the second test equals -1.576 - too high to reject the null hypothesis. The following section argues for the use of the differenced HPI despite this result.

Taking the first-order differences of the time series variables reduces the number of observations within the data set. In addition, as explained in section 4.4, the one-year lag of the dependent variable in consideration is added to models, which further decreases the number of data points. Ultimately, the data set consists of 179 observations covering the time period from March 2004 to January 2019.

5.1.2 First-order difference of HPI

Section 4.1 discusses the difference between the two house price indicators HPI and the average purchase price of dwellings sold, and deduces that a similar trend can be observed for both indicators over the defined period of time. Following this conclusion, it is worth comparing the ADF test results for the two indices. Table 3 shows that the previous finding, i.e. differencing the HPI fails to resolve the issue of not rejecting the null hypothesis of the ADF test, does not hold for the average purchase price.

Variable	ADF test 1	ADF test 2
HPI	-0.893	-1.756
Average purchase price	-0.0112	-4.231***

Table 3: Comparison between the Dickey-Fuller values of the ADF test results for the HPI and the average purchase price.

Note: *** indicates significance at an alpha of 1%.

The time series plot of the differenced values reveals that both indices move closely to each other except towards the end of the observed time period when the HPI slightly deviates upward from the average purchase price (see Figure 5). Hence, it can be argued that the insignificance with regards to the ADF test results of the HPI is likely to be caused by the relatively smooth development of the index as it accounts for the change in the quality of dwellings sold. Based on these observations and for the ease of interpretation of the empirical findings, the first-order differences of the HPI are retained as series of interest rather than taking second-order differences.

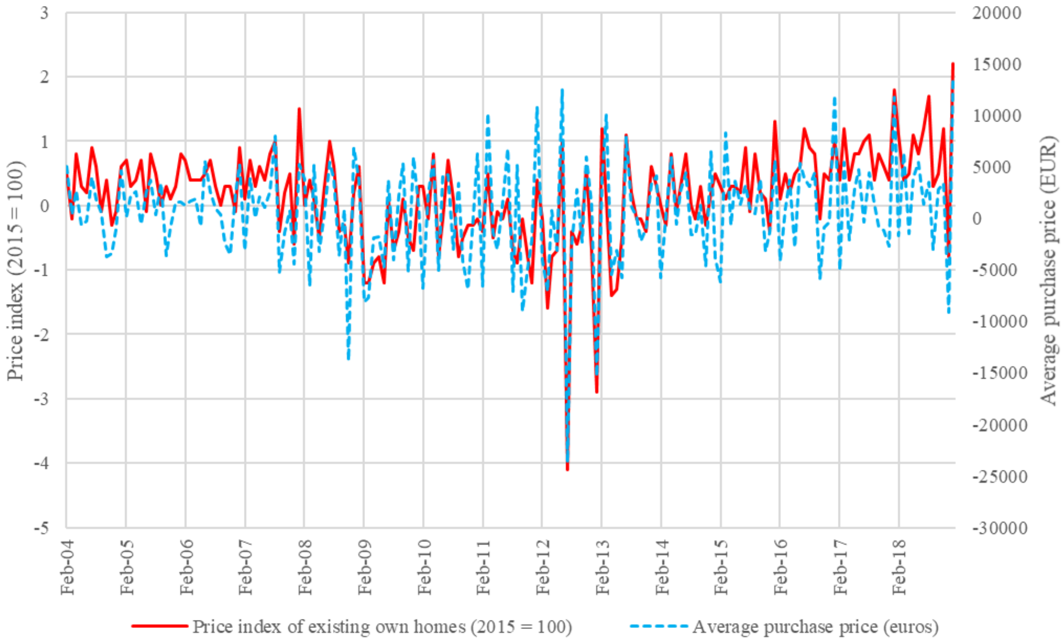
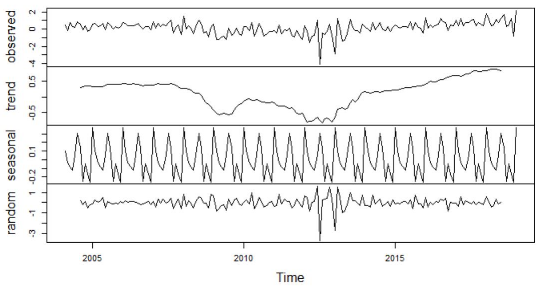


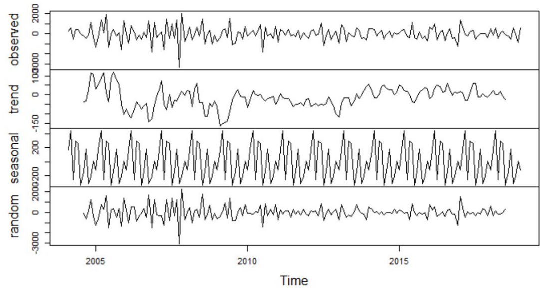
Figure 5: Time series plot of the first-order differences of the HPI and average purchase price, Feb/2004-Jan/2019.

5.1.3 Seasonality

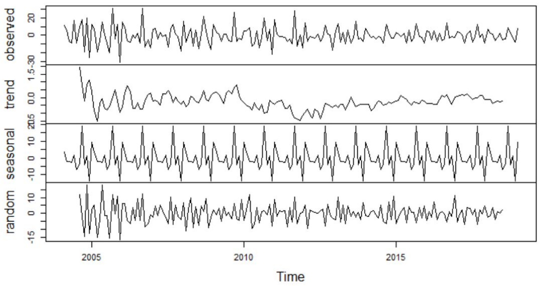
Taking first-order differences has the advantage that it removes the trend from a time series process. However, besides a trend-cycle component, time series data often contains another time-variant component: seasonality. Figure 6 shows that for the HPI, the residential mortgage flows, and selected Google search indices the data is characterised by an apparent additive seasonal component. This observation is supported by the results of the function *nsdiffs()* from the *forecast* package in the statistical software R. The output of the function suggests that based on a measure of seasonal strength higher than 0.64 the number of seasonal differences required for stationarity for many of the retained variables is equal to 1. To account for the observed seasonality, monthly dummy variables are added to the data set with December being withheld as a baseline. Adding dummies rather than transforming the time series into rolling year-on-year differences allows to maintain a high number of observations and, hence, retain a larger part of the observed housing cycle.



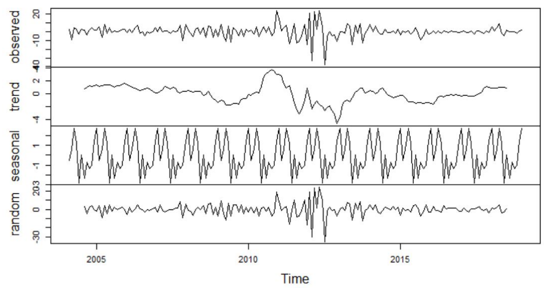
(a) Decomposition of differenced HPI.



(b) Decomposition of differenced mortgage flows.



(c) Decomposition of differenced Google search term "architecten".



(d) Decomposition of differenced Google search term "abnamro".

Figure 6: Decomposition of selected time series variables.

5.2 Models

The academic literature discussed in section 2.2 on the integration of Google search data in forecasting the dynamics of particular real-estate markets is commonly defined by the application of OLS regressions to estimate future house prices or sales. This choice of method is reasonable given the low number of Google search queries retained in these papers (Humphrey 2010; McLaren and Shanbhogue 2011; Wu and Brynjolfsson 2015). However, considering the list of 33 Google search queries in Table 2, following the same method in this paper would expose two types of limitations of OLS regressions: (i) OLS estimates typically have low bias but large variance, therefore a trade-off is to be faced between the two as sacrificing some bias to reduce variance would allow to improve prediction accuracy, and (ii) as all predictors are contained in an OLS model, interpretation of the estimates is more difficult. Two proposed solutions to these limitations are subset selection models, e.g. stepwise and stagewise selection, and the ridge regression. As the former represents a discrete process - regressors are either retained or dropped - it can lead to a high variance, ultimately, reducing prediction accuracy. The latter solution, although defined as a continuous shrinkage model, hardly ever sets any coefficients equal to zero and, hence, fails to ease the interpretation of the estimates (Tibshirani 1996).

In his original paper, Tibshirani (1996) proposes the least absolute shrinkage and selection operator, i.e. the lasso, as a solution to these drawbacks. By shrinking some coefficients and setting others equal to zero, the technique produces a model which is stable and continuous as the ridge regression and returns easily interpretable results as with the subset selection. Given these advantages, the lasso method has become a popular tool in economic forecasting in recent years as it is particularly suitable for data sets where the number of regressors approaches or even exceeds the number of observations (Baldacci et al. 2016). This often holds for data sets including Google data as there is an abundance of potential queries stored in the Google search corpus. Hence, the lasso has been applied by numerous academics, scholars and researchers who have integrated Google Trends data into predictive models to estimate future values of different macroeconomic indicators (Götz and Knetsch 2018; Parker et al. 2017; Santillana et al. 2014).

Tibshirani (1996) defines the lasso as follows

$$\text{minimise } (y - \alpha - \sum_{i=1}^{N} x_i \beta_i)^2 \text{ subject to } \sum_{i=1}^{N} |\beta_i| \leq t \tag{1}$$

where y denotes the response variable and x_i the predictor variables. In addition, t is a positive tuning parameter which restricts the sum of absolute values of the coefficients, $\sum_{i=1}^{N} |\beta_i|$, and, hence, controls the degree of shrinkage applied to the estimates. The intercept, α, is not penalised by the parameter t and can take any value.

James et al. (2013) explain equation 1 in a more intuitive manner. When performing the lasso, the objective is to find the set of estimated coefficients leading to the smallest residual sum of squares under the constraint of a *budget*, t, which restricts the size of $\sum_{i=1}^{N} |\beta_i|$. When the budget is nonrestrictive, i.e. t is very large, the sum of absolute values of the coefficient estimates can be large. In fact, if t is extremely large, the least squares solution falls within the budget. However, if t is rather small, then the sum of absolute values must be small too in order to not exceed the budget.

An alternative way to express the lasso as a constrained minimisation problem helps to highlight how the variable selections differ between the lasso and the ridge regression. In fact, the lasso coefficients, $\hat{\beta}_{\lambda}^{L}$, minimise

$$(y - \alpha - \sum_{i=1}^{N} x_i \beta_i)^2 + \lambda \sum_{i=1}^{N} |\beta_i| \tag{2}$$

where the positive regularisation parameter λ has a one-to-one correspondence with t in equation 1. In contrast to the ℓ_2 penalty used by the ridge regression on the sum of squared coefficients, $\sum_{i=1}^{N} \beta_i^2$, the lasso applies an ℓ_1 penalty as represented by the second part of equation 2.

Figure 7 visualises the difference between the penalties used by both shrinkage methods, and how the lasso sets some coefficient estimates equal to zero. The least squares solution is denoted by $\hat{\beta}$, the blue diamond and circle are the lasso and ridge regression constraints, $|\beta_1| + |\beta_2| \leq t$ and $\beta_1^2 + \beta_2^2 \leq t$ derived from equation 2, and the red ellipses centered around $\hat{\beta}$ display the regions of constant residual sum of squares, i.e. the contours of the least suqares error function which is to be minimised subject to the solid blue areas. Ellipses farther away from $\hat{\beta}$ have a higher value of the residual sum of squares than those closer to the center. Equation 2 indicates that the estimated coefficients of the lasso are given by the first point of intersection between the constraint region and an ellipse. The same holds for the ridge regression with the difference that the second part of the equation includes the sum of squared differences. As can be easily observed in Figure 7, the circular constraint of the ridge regression causes this first intersection to generally not be on an axis. Consequently, the coefficients estimated by the ridge regression are non-zero. In contrast, the intersection between an ellipse and the lasso constraint lies on an axis as the corners of the constraint region fall on each of the axes. Hence, one of the estimated coefficients is set to zero. In the case of higher dimensions, multiple coefficient estimates may equal zero at the same time (James et al. 2013).

In their work, James et al. (2013) provide a simple special case for the lasso and ridge regression to demonstrate how the types of shrinkage differ for both methods. Under the assumptions that the number of predictors, p, equals the number of observations, n, and $\mathbf{X}$ is a diagonal matrix with only 1's on the diagonal and

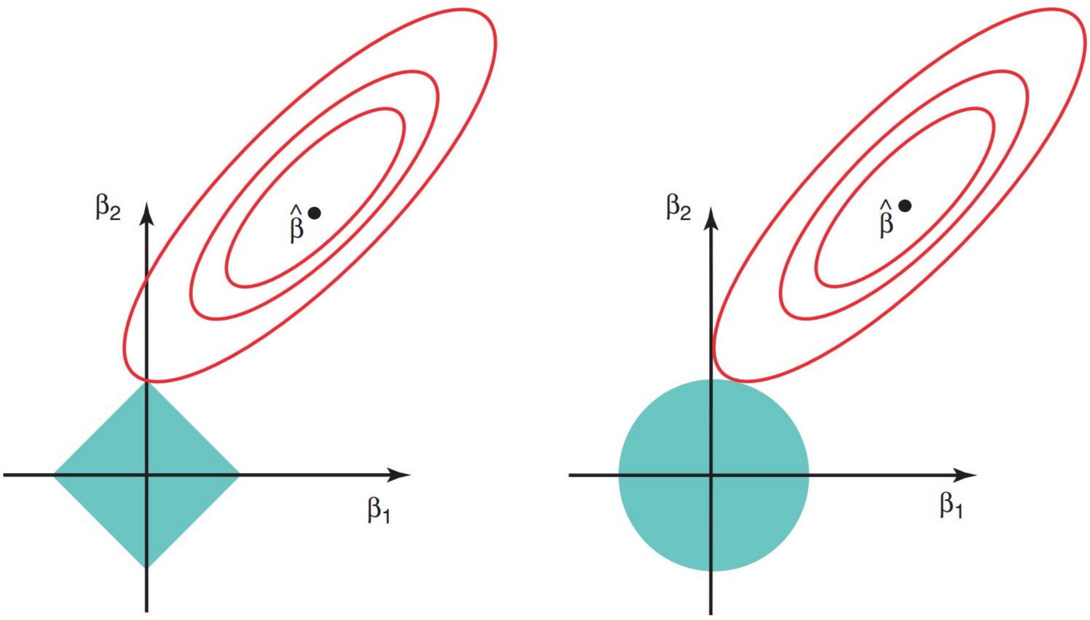


Figure 7: Graphical comparison between the lasso (left) and ridge regression (right) penalties.

0's in the off-diagonal elements, they prove that the lasso performs a so-called soft-thresholding in which all coefficients are shrunken toward zero by a similar proportion, and these coefficients which are small enough are entirely set equal to zero. In contrast, the ridge regression shrinks all coefficient estimates by the same amount (see Figure 8).

5.2.1 Selecting the regularisation parameter

The number of coefficients set to zero by the lasso depends on the value of the regularisation parameter λ which determines how restricted the model will be; in other words, for λ equal to zero, all explanatory variables have nonzero coefficients, which corresponds to the OLS estimates. The optimal value of λ can be determined either based on information criteria or by using cross-validation techniques - both

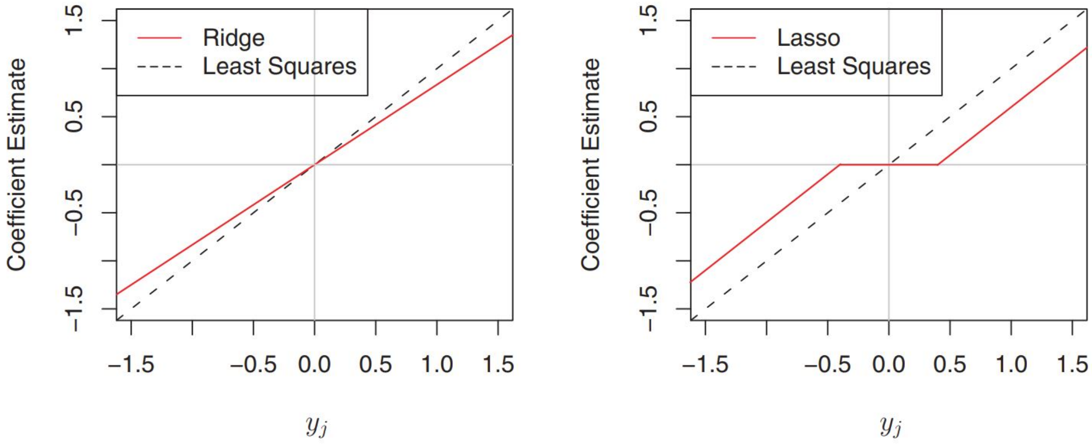


Figure 8: The ridge regression and lasso coefficient estimates with $n = p$ and $\mathbf{X}$ a diagonal matrix with 1's on the diagonal.

methods are used in this paper.

The use of information criteria allows to adjust the training error for the size of the model. When fitting a model to the training data, the coefficient estimates are such that the residual sum of squares of that subdataset is minimised. However, as a consequence, the residual sum of squares of the test data is not necessarily as small as possible. In other words, the mean squared error of the training set generally underestimates the mean squared error of the test data. Several approaches have been developed to account for this error and select the best model among a set of models with different numbers of predictors, including among others the Akaike Information Criterion (henceforth AIC) and the Bayesian Information Criterion (henceforth BIC) (James et al. 2013). In their work, Zhang et al. (2010) conclude that the value of λ which minimises the BIC performs better than cross-validation techniques and consistently selects the true model supposed that the latter is among the candidate models. Furthermore, the authors argue that a regularisation parameter selected based on the AIC is asymptotically loss efficient, a property which does not hold when considering the BIC. Based on these findings, the values of λ which minimise the BIC and AIC (denoted λ_{AIC} and λ_{BIC} respectively) are retained for the model creation process in this paper. In addition, the AICc is considered as a third information criterion when selecting the optimal value for the regularisation parameter (denoted λ_{AICc}) to study potential overfitting with the AIC in case of a small sample size.

Although the choice of λ based on information criteria appears to be a reasonable approach, the true model is unknown for the task at hand. In other words, it is likely that, based on the three aforementioned information criteria, search queries are missing in the final model. For that reason, a k-fold cross-validation is additionally executed. Following this method, the sample is portioned equally over k groups among which one group serves as training data whereas the remaining one are used as validation data. In other words, one group is retained to calculate the lasso estimates for different values of λ, and, in turn, the obtained model is fit to the other groups. At this stage, the standard deviations and errors of the model are computed. This process is iterated k times such that each group serves as training data once, and the average standard deviations and errors per fold are found (Hastie et al. 2015). In this paper, following the example given by James et al. (2013), a ten-fold cross-validation is applied to the lasso fits to determine the value of λ which minimises the mean squared cross-validated errors (denoted λ_{MSE}) and the highest possible value of λ such that the mean squared errors are withing one standard error of the minimum mean squared error (denoted λ_{1SE}). At this point it should be noted that although using cross-validation on the sample at hand has the advantage of using the entire data set, the application of this approach does entail some drawbacks. In particular, as the developed models include the one-month lag of the response variables as a predictor, the training and test data are not independent if randomly chosen. Furthermore, given that the time series on the HPI and residential mortgage flows are generated by process which develops over time, the assumption of independent and identically distributed data is likely to be violated here. Last block evaluation could solve these issues, however, it does not make use of the full data set which is not desired for the case at hand given the small number of observations in the sample (Bergmeir and Benıtez 2012).

To incorporate the aforementioned macroeconomic indicators as exogenous variables in the predictive models, an appropriate penalty factor is defined to ensure that the coefficients of these variables are not set to zero by the lasso. In addition, the initial data set is standardised by centering and normalising the economic variables and Google search indices to lower the potential impact of outliers on the estimations. Consequently, the normalised first-order differences of the HPI lie between -5.6249 and 2.7012, whereas the values of the residential mortgage flows are within a range of -4.6675 to 2.8784.

5.3 Forecasting

To predict future house prices and residential mortgage flows, this study applies the following stepwise approach:

- *Step 1:* A subsample excluding the time period from January 2016 until January 2019 is used to obtain the different regularisation parameters for the lasso model selection.
- *Step 2:* Based on the models selected by the lasso technique, the data set is

filtered by removing the variables for which the coefficients have been set to zero. The obtained subsets are, in turn, used as input for OLS regressions to estimate the post-model selection coefficients. Belloni and Chernozhukov (2013, p. 2) have proven that "OLS post-lasso can perform at least as well as lasso in terms of the rate of convergence, and has the advantage of a smaller bias".

- *Step 3:* The post-model selection estimators are used together with the data for the independent variables for the month which is to be forecasted, e.g. January 2016, to find the predicted value of the dependent variable.

- *Step 4:* Steps 1 to 3 are repeated using a new subsample excluding the time period from February 2016 until January 2019. This process is reiterated until each month of the holdout sample is forecasted. In this way, an expanding window is created as more actual observations enter the subsample used to estimate the lasso regularisation parameters and OLS coefficients (Bulut 2018).

This approach is applied to both dependent variables, HPI and residential mortgage flows, respectively. Furthermore, in order to verify whether search queries are useful for purchases which are planned in advance (see section 2.1), the predicted values are calculated using multiple forecast horizons (denoted h), namely h equal to 1, 3, 6, 9, and 12 months. Following this approach, 25 forecasts are estimated, for example, for the HPI in January 2016 based on the five lasso regularisation parameters in combination with five different forecast horizons.

The "OLS post-lasso" estimators are compared to a benchmark model which is based on an OLS regressions using only the macroeconomic variables introduced in section 4.4 and the one-year lag of the dependent variable in consideration as explanatory variables. The comparison allows to assess whether adding Google search queries retained by the lasso approach improves the forecasting performance at different forecast horizons. The forecasting accuracy of each model is measured by the mean absolute error (henceforth MAE) and the root mean square error (henceforth RMSE). Although both measures are sensitive to the scaling of the variables, this issue is of less importance for the case at hand as the measures are applied to all forecasts made within the same time period (Greene 2018).

6 Empirical findings

Table 4 lists the MAE and RMSE of each model for each dependent variable at the five forecast horizons in consideration. The results indicate that certain models which are obtained by the OLS post-lasso approach and include Google search queries outperform the benchmark model, however, only by a small margin. Furthermore, the accuracy measures of the HPI forecasts transmit different messages. The RMSE of the benchmark model are the lowest whereas the MAE suggest that including specific queries can improve the forecasting performance. In particular, the model obtained with the regularisation parameter λ_{BIC} performs best for 1- and 9-month-ahead forecasts, whereas the model with parameter λ_{MSE} seems best suited for 3- and 12-months-ahead forecasts.

When forecasting 6 months ahead, the model with parameter λ_{AICc} appears to perform best. When studying the post-model selection estimators of these models, three search terms specifically stand out as they appear in every model: "bouwfonds", "hypotheek", and "huizen". Other queries which are included in some models but not all are "abnamro", "onroerend goed", "stadskamer", "rabobank", and "huis".

In contrast to the results for the HPI, the MAE and RMSE for the residential mortgage flows forecasts are somewhat more consistent in terms of the comparative forecasting performance of the models. Both accuracy measures suggest that for 1-, 3- and 6-month-ahead forecasts the model with regularisation parameter λ_{AICc} outperforms all other models. For longer forecast horizons, the measures are contradictory as the MAE are the lowest for the model with parameter λ_{BIC} whereas based on the RMSE the model obtained using λ_{1SE} performs best. With regards to the retained Google search queries, each of these models includes a single search term, namely "herbouwwaarde".

In their work, Zhao and B. Yu (2006) define the so-called Irrepresentable Conditions under which the lasso satisfies the consistency of the selected model and of the estimated coefficients, and discuss sufficient conditions which, when imposed on the models, ensure that the Irrepresentable Conditions are met. These conditions are likely to be violated in presence of high correlation among the variables included in the models. As the latter is potentially the case with the data set at hand, the adaptive lasso (henceforth adalasso) is implemented to check for any improvements in the forecasting performance of the models. However, as no such improvements are achieved, the results for the adalasso models are left out of Table 4.

As previously noted, the margins between the accuracy measures of the benchmark model and the models including Google search queries are rather small. Hence, to measure the statistical significance of these differences, the modified Diebold-Mariano test proposed by Harvey et al. (1997) is employed using the

Dependent variable	Forecast horizon	Benchmark	Category	λ_{AIC}	λ_{AICc}	λ_{BIC}	λ_{MSE}	λ_{1SE}
HPI	1-month-ahead	0.7798	0.7795	0.7948	0.7852	**0.7794**	0.8045	0.8065
		(0.9123)	(0.9341)	(0.9644)	(0.9648)	(0.9545)	(0.98)	(0.9637)
	3-months-ahead	0.7799	0.779	0.7829	0.7858	0.7759	**0.7758**	0.8055
		(0.9142)	(0.935)	(0.9719)	(0.9751)	(0.955)	(0.9607)	(0.935)
	6-months-ahead	0.7873	0.7844	0.8076	**0.7801**	0.7802	0.7978	0.8141
		(0.921)	(0.9404)	(0.9947)	(0.9614)	(0.958)	(0.9812)	(0.9733)
	9-months-ahead	0.7881	0.788	0.8124	0.7852	**0.7819**	0.7967	0.8131
		(0.9235)	(0.9446)	(0.9903)	(0.9667)	(0.9605)	(0.9811)	(0.9746)
	12-months-ahead	0.7901	0.786	0.7914	0.7907	0.7816	**0.7803**	0.8138
		(0.9213)	(0.9389)	(0.9722)	(0.9695)	(0.9552)	(0.9579)	(0.9719)
Residential mortgage flows	1-month-ahead	0.581	0.6	0.6424	**0.5749**	0.581	0.5827	0.585
		(0.7643)	(0.786)	(0.828)	*(0.7554)*	(0.7643)	(0.7608)	(0.763)
	3-months-ahead	0.5831	0.6023	0.6345	**0.5782**	0.5831	0.5851	0.5851
		(0.7619)	(0.7837)	(0.8202)	*(0.7535)*	(0.7619)	(0.7599)	(0.7599)
	6-months-ahead	0.5828	0.6023	0.6581	**0.5782**	0.5852	0.5846	0.5846
		(0.7617)	(0.7821)	(0.8417)	*(0.7531)*	(0.7617)	(0.7586)	(0.7586)
	9-months-ahead	0.5868	0.6077	0.6964	0.5959	**0.5856**	0.5948	0.5948
		(0.7669)	(0.7881)	(0.8869)	(0.7698)	(0.7669)	*(0.7658)*	*(0.7658)*
	12-months-ahead	**0.5819**	0.6019	0.6872	0.5908	**0.5819**	0.5915	0.5888
		(0.7623)	(0.7827)	(0.8827)	(0.7573)	(0.7623)	(0.756)	*(0.7542)*

Table 4: MAE and RMSE (in bracket) values for the predictive models at different forecast horizons.

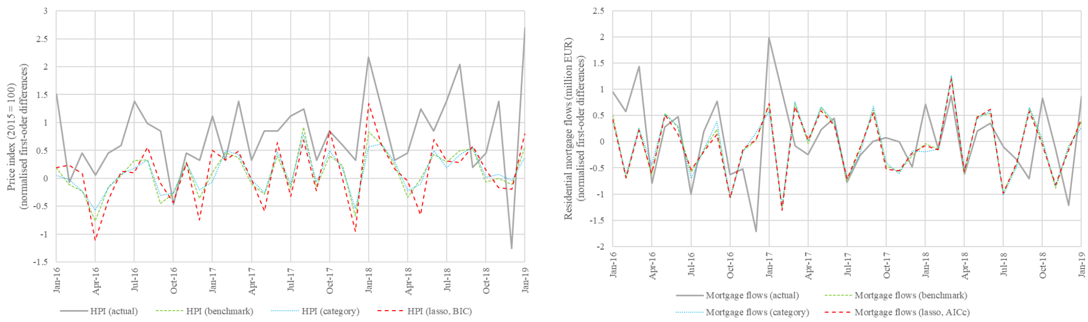
Note: The lowest MAE value is indicated in bold whereas the lowest RMSE value is in italics.

squared error. The null hypothesis, i.e. equal forecast accuracy between the benchmark model and the other models, is tested against the alternative hypothesis that the benchmark model is less accurate than the models augmented by the Google data. The Diebold-Mariano test statistics and the respective p-values are listed in the first column of Table 5. For this test, only the OLS post-lasso estimators with the lowest MAE as reported in Table 4 are retained. The results suggest that there is not enough evidence to reject the null hypothesis, and, therefore, the models including the Google search queries do not significantly outperform the benchmark model. Figure 9 shows that the difference between the predicted values of the benchmark model and the augmented models in consideration is rather small for the HPI forecasts and essentially overlap for the residential mortgage flows.

Dependent variable	Forecast horizon	Benchmark	Category
HPI	1-month-ahead	-1.0146	-1.0757
		(0.8267)	(0.8554)
	3-months-ahead	-0.8289	-1.1401
		(0.7939)	(0.8691)
	6-months-ahead	0.9868	-0.1820
		(0.1652)	(0.5717)
	9-months-ahead	0.6798	0.2721
		(0.2505)	(0.3936)
	12-months-ahead	0.6689	0.7763
		(0.2539)	(0.2213)
Residential mortgage flows	1-month-ahead	-1.2829	-1.4692*
		(0.8961)	(0.0752)
	3-months-ahead	1.0821	-1.3118*
		(0.1432)	(0.0989)
	6-months-ahead	-1.9244	2.2746**
		(0.9689)	(0.0145)
	9-months-ahead	-1.1428	0.4918
		(0.0869)	(0.3129)
	12-months-ahead	-0.3801	1.4046*
		(0.6469)	(0.0844)

Table 5: Diebold-Mariano test statistics and p-values (in bracket) for the best performing models.

Note: * indicates significance at an alpha of 10%, and ** indicates significance at an alpha of 5%.



(a) Best forecasts of the HPI based on MAE.

(b) Best forecasts the residential mortgage flows based on MAE.

Figure 9: Predicted values of the HPI and residential mortgage flows, Jan/2016-Jan/2019.

6.1 Google search terms versus categories

Besides the index on search activity by query, Google Trends also provides data on all searches within a predefined category (Stephens-Davidowitz and Varian 2014). For example, selecting the category *Travel* without typing any search term and restricting the geography to the Netherlands, reveals the index for travel-related search queries on a national level with an apparent yearly peak during the summer holidays. Hence, many academics make use of these categories when integrating Google search data into economic forecasting. For example, Wu and Brynjolfsson (2015) use the categories *Real estate agencies* and *Real estate listings* in their paper on predicting US housing market dynamics. In contrast to their approach, this research relies on search terms rather than categories. Given that all of the search terms listed in Table 2 are either directly (e.g. "huis") or indirectly (e.g. "rabobank") related to the Dutch housing market, it seems worthy to study whether using a predefined category in Google Trends leads to improved forecasting results.

For the case at hand, the category *Property* seems appropriate for two reasons: on the one hand, the category incorporates many housing-related subcategories such as *Real Estate Agencies*, *Real Estate Listings*, *Property Development*, and *Property Management*; on the other hand, the search indices of the previously highlighted search terms "herbouwwaarde", "bouwfonds", "hypotheek", and "huizen" within that category follow an almost identical pattern to the overall search indices obtained without specifying a category (see Appendix C). Hence, an additional model is estimated by adding the search index for the category *Property* to the original benchmark model. The MAE and RMSE of this model are reported in the second column of Table 4. The values show that the model including the category is less accurate than the allegedly best performing model for each dependent variable at the different forecast horizons.

The Diebold-Mariano test statistics and p-values in the second column of Table 5 suggest that the difference in accuracy between the model including the category *Property* and the best performing models is insignificant for the predictions of the HPI but significant for the residential mortgage flows forecasts at each forecast horizon but the 9-months-ahead forecasts. In particular, the difference is significant at a 10% level for the 1-, 3- and 12-months-ahead forecasts, and significant at an alpha of 5% for $h = 6$.

7 Discussion

The previously discussed findings suggest that there is not enough empirical evidence to support the hypothesis that integrating relevant Google search queries into the forecasting of the dynamics of the Dutch housing market improves the forecasting performance above and beyond the use of macroeconomic indicators. However, the results of this paper still hold some valuable insights which should not be neglected by policymakers and other stakeholders within the real estate market as well as by economic researchers who intend to take advantage of the open access and timely availability of Google search data.

Although the difference in forecasting performance between the benchmark model and the models augmented by Google queries is insignificant, the enhanced models prove to be more accurate than the models including the search index of a category when predicting the flow of residential mortgages. This finding implies that, when working with Internet search data, observing specific key words which are likely to be searched for in the context of the housing market or any other market can prove beneficial for both economic actors within that market and researchers. For example, following the approach described in section 4.2 using Google Correlate in combination with *Startpagina.nl* pages, a company listed on these pages can gain insights into the search behaviour of potential customers and learn whether the terms by which it is defined on *Startpagina.nl* are actually searched for. In turn, the company can optimise its online presence by adopting its hyperlink on the web directory correspondingly. In contrast, researchers who choose to integrate Google search data into forecasting economic indicators might be able to improve the accuracy of their predictive models by focusing on relevant search terms rather than predefined categories in Google Trends.

Indeed, Figure 10 suggests that, whereas the search index for the category *Property* decreases before steadying in 2012, the searches for the word "huizen" follow the trend of the HPI more closely. The correlation values between the respective search indices and the house prices confirm this observation (observed correlation of 0.433 between "huizen" and the HPI, and of 0.191 between the category *Property* and the price index). These findings indicate that a category such as *Property* is likely to comprise search queries which are irrelevant to the study of a particular economic indicator and potentially distort the apparent relationship between the search data and the variable of interest.

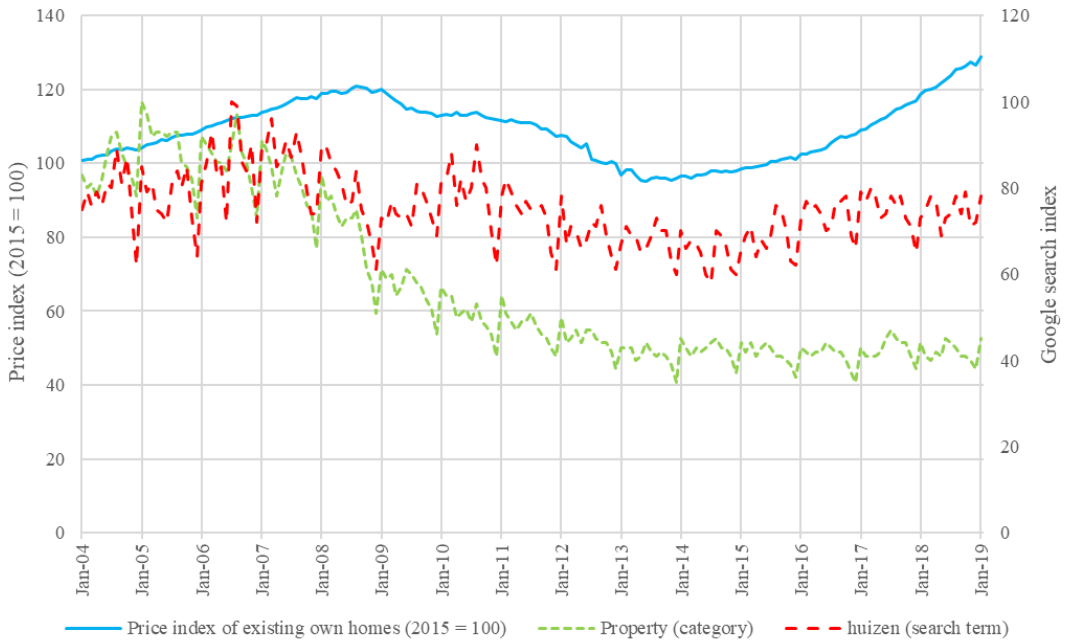


Figure 10: Comparison between the Google search indices for "huizen" and the category *Property*, and the HPI, Jan/2004-Jan/2019.

Furthermore, the empirical results of this study are in line with the findings of Wu and Brynjolfsson (2015) hinting that forecasting changes in house prices using Google Trends data can be challenging. As the authors only find evidence for a rather moderate positive relationship between the search queries and future prices, they conclude that the ambiguous impact of searches from both sides of the market on the price index limit the effectiveness of aggregated queries. To overcome this potential difficulty, this study focuses on the demand side of the real estate market by analysing the relationship between Google searches and the flows of residential mortgages in addition to the HPI. Considering both accuracy measures MAE and RMSE, the models augmented by search data outperform the benchmark model at any forecast horizon except for $h = 12$ when predicting changes in the flow of mortgages, although the differences in accuracy are insignificant.

Finally, the analysis provides insights into the search behaviour of the demand side of the housing market as the term "herbouwwaarde" (i.e. "reconstruction value") is the only search query selected by the lasso approach when forecasting residential mortgage flows. Hence, it can be concluded that this particular search term contains information on the dynamics of the real estate market which is not captured by the selected macroeconomic indicators. The word "herbouwwaarde" can be linked to the construction account (i.e. "bouwdepot") (also called building fund account) - a popular means in the Netherlands to add renovation costs to the initial mortgage and repay them together with the mortgage in monthly instalments. The amount placed on this separate account depends among others on the value of the dwelling after the renovation and the conditions for setting up this type of account require an improvement in the value of the renovated house. Hence, an evaluation of the increased value needs to be made prior to the changes done on the home (Independer n.d.), which is likely to result in online searches for the determination of a reconstruction value, i.e. "herbouwwaarde".

7.1 Limitations

Despite its advantages in terms of availability and vast coverage of respondents and topics, Google search data does entail some difficulties when using it, two of which are important to this study. First, the Google Trends only provides data from January 2004 on - a relatively short period of time when compared to the information available on certain economic indicators. In addition, the number of available observations is further shortened by taking first-order differences in order to transform the data into a stationary time series and including the one-year lag of the dependent variables in consideration. Second, as explained in section 4.3, the data returned by Google Trends is based on a sample of the entire Google search corpus and this sample changes on a daily basis (Stephens-Davidowitz and Varian 2014). This alternation affects the search indices of the queries and, ultimately, the empirical findings of this study as the underlying data changes. Taking the average search index of the search queries over multiple days could mitigate this problem.

Besides these constraints entailed by the characteristics of the Google Trends data, two further limitations of this study should be noted. First, the empirical analysis of this paper is solely based on the employment of the post-lasso OLS estimation. As pointed out by Götz and Knetsch (2018) the results of a research on using Google search queries in forecasting economic indicators depend on the applied variable selection method and the target variable. Hence, the application of other methods such as PCA, PLS, and boosting to select the search queries relevant to the predictions of house prices could be subject of future research. Second, the search terms retained by the approach taken in section 4.2 are solely related to the real estate market. However, it could be of interest to assess the relevance of search terms from other economic fields such as income or consumption. This would allow to further study the claim that Google search queries comprise some of the information contained in macroeconomic indicators as changes in the latter impact search behaviour (Humphrey 2010).

8 Conclusion

Accurate forecasts on the development of the economic situation of a country, region, or city is of significant importance to political and governmental decision-makers to take effective action in response to potential economic and socioeconomic challenges. The rise of big data gives access to novel data sources such as Google or Facebook which represent an alternative for or augmentation of the prevailing practices of data collection on macroeconomic indicators generally relying on surveys. This study attempted to assess whether the inclusion of Google search data could improve the forecasting performance of post-lasso OLS estimators with regards to the dynamics of the Dutch housing market. The objective of the empirical analysis was to determine whether related search queries comprise information above and beyond that contained in common macroeconomic variables and whether this additional information allows to more accurately predict changes of the HPI and the flow of residential mortgages.

In order to carry out the analysis, data on housing-related search queries was collected through a novel approach using Google Correlate in combination with selected *Startpagina.nl* pages. In turn, the retained search terms were used as input for the post-lasso OLS estimations testing different lasso regularisation parameters. Based on the developed models, future values of the target variables were predicted for short- and long-term forecast horizons and compared to the forecast of a benchmark model. In addition, the forecasting performance of models augmented by the search indices of specific terms was measured against a model including the search index of a predefined Google Trends category.

The empirical findings of this study suggest that there is not enough evidence to confirm that the inclusion of relevant Google search queries significantly improves the forecasting accuracy of the post-lasso OLS estimators in comparison to the benchmark model. However, based on the Diebold-Mariano test, it can be concluded that including specific search terms rather than an overarching category leads to more accurate predictions of the flow of residential mortgages. The terms retained by the applied post-lasso OLS approach are of particular interest to government officials and policymakers as they give insights into the search behaviour of citizens with regards to the real estate market. As discussed in this paper, the search term "herbouwwaarde" relates to the common practice of taking on a construction account on top of the residential mortgage in the Netherlands. Hence, the significance of the searches for "herbouwwaarde" can be interpreted as a response of Dutch citizens to implemented policy changes. Following this reasoning, observing relevant search indices can prove beneficial to government representatives during the planning and evaluation stages of the political decision-making process.

The present study provides indications on the potential of the information contained in Google search queries and how the queries can be incorporated into economic modelling. However, as both, the benchmark model and the developed models, include macroeconomic indicators as control variables the real-time availability of the search data has not been fully exploited. In particular, further research needs to be conducted on the capacity of Google Trends queries to add onto or even replace survey data, e.g. the consumer confidence indicator, with regards to nowcasting the dynamics of the real estate market. Moreover, the relationship between the search data and macroeconomic control variables needs additional analysis to examine the added value of a search term over a related economic indicator. Research on the use of Google search data instead of macroeconomic variables in predictive models can be done by following the framework applied in this study. The combination of Google Correlate output with *Startpagina.nl* pages related to different economic fields and the application of the post-lasso OLS approach allow to collect and evaluate a high number of relevant search queries which contain the same or even additional information than corresponding economic indicators. In order to implement this approach, the macroeconomic variables should be excluded from the developed models and only enter the benchmark model as controls. The practical application of such an approach could prove beneficial in cases where economic data is either not (yet) available (e.g. due to a reporting lag) or unreliable (e.g. in developing countries). Lastly, the construction of an aggregated Google search index comprising the most important queries could mitigate potential correlations between search terms and improve the forecasting accuracy of developed models.

A Startpagina.nl web pages

In this appendix, the links to the web pages of *Startpagina.nl* which have been used in the data collection process to filter the search terms obtained through Google Correlate for real-estate related words are listed:

- *Huis*: https://www.startpagina.nl/q=huis
- *Hypotheek*: https://hypotheek.startpagina.nl/
- *Makelaar*: https://www.startpagina.nl/q=makelaar
- *Woning-koopwoning*: https://woning-koopwoning.startpagina.nl/
- *Woonwinkels*: https://www.startpagina.nl/q=woonwinkels

B Augmented Dickey-Fuller test results

Table 6 below lists the Dickey-Fuller values for the ADF test of each variable included in the data set of this research. In the first column, the names of the variables from the third row to the 35th row refer to Google search terms whereas the first two rows specify the indicators of interest and the last 6 the macroeconomic variables. The second and third columns return the Dickey-Fuller values of the ADF tests applied before and after taking first-order differences respectively.

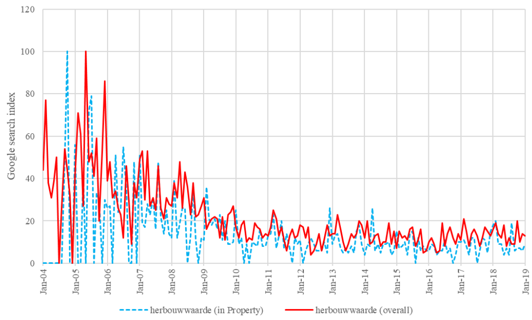
Variable	ADF test 1	ADF test 2
HPI	-0.893	-1.756
Residential mortgage flows	-2.384	-7.623***
abnamro	-2.944	-5.504***
architecten	-3.579**	-9.721***
leen bakker	-3.274**	-6.06***
bkr	-1.842	-7.581***
bouw	-5.184***	-9.946***
bouwfonds	-2.546	-6.332***
directwonen	-4.460***	-9.874***
funda	-2.646	-5.617***
haagwonen	-4.534***	-7.125***
herbouwwaarde	-3.681**	-8.508***
huis	-5.043***	-6.795***
huren	-3.079	-7.153***
hypotheek	-2.751	-6.912***
makelaar	-2.024	-6.257***
meeus	-2.762	-5.334***
meubelen	-3.171*	-6.329***
nieuwbouw	-2.127	-6.007***
plattegrond	-2.293	-7.148***
rabobank	0.13	-5.084***
remco	-3.053	-6.99***
stadskamer	-3.12	-6.821***
stienstra	-2.153	-7.48***
studio	-2.449	-5.945***
vastgoed	-2.135	-8.302***
vbo	-1.792	-7.926***
verhuisbedrijven	-2.8144	-8.275***
verhuur	-2.891	-7.422***
wehkamp	-0.83	-6.978***
huis inspectie	-3.528**	-7.845***
huizen	-2.065	-6.346***
huizen te koop	-1.99	-7.703***
onroerend goed	-0.677	-9.774***
verkoopstyling	-1.641	-7.606***
unemployment rate	-1.142	-6.415***
CPI	-2.344	-6.778***
mortgage rate	-1.501	-4.786***
construction costs	-1.59	-3.935**
long-term interest rate	-1.917	-4.049***
industrial production	-3.25*	-7.131***

Table 6: Dickey-Fuller value of the ADF test results before and after taking first-order differences of each variable in the data set.

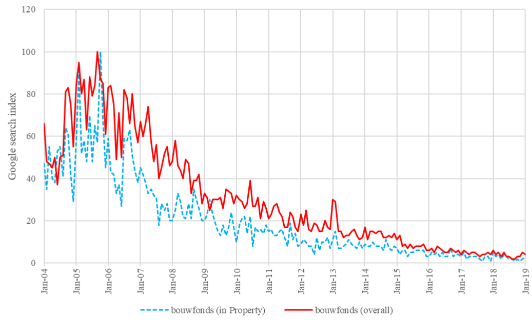
Note: * indicates significance at an alpha of 10%, ** indicates significance at an alpha of 5%, and *** indicates significance at an alpha of 1%. The results for the HPI are discussed in section 5.1.2.

C Google search terms versus categories

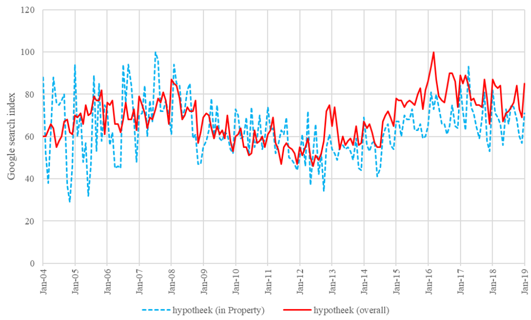
Figure 11 below shows the Google search index for the most relevant search terms from the analysis in section 5 within the predefined category *Property* and the overall search index over the time period in consideration.



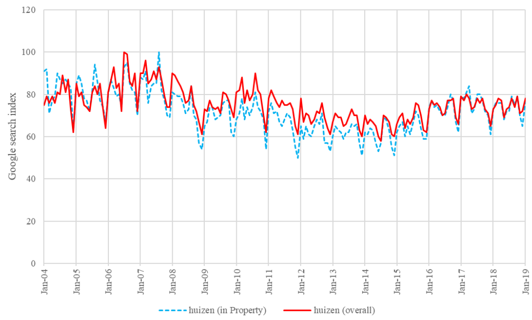
(a) Google search indices for "herbouwwaarde".



(b) Google search indices for "bouwfonds".



(c) Google search indices for "hypotheek".



(d) Google search indices for "huizen".

Figure 11: Google search indices of search terms overall and within category *Property*.

References

Askitas, N. & Zimmermann, K. F. (2009). Google econometrics and unemployment forecasting. *SSRN Electronic Journal.* https://doi.org/10.2139/ssrn.1480251

Augustyniak, H., Leszczyński, R., Łaszek, J., Olszewski, K. & Waszczuk, J. (2014). On the dynamics of the primary housing market and the forecasting of house prices. *Munich Personal RePEc Archive.*

Baldacci, E., Buono, D., Kapetanios, G., Krische, S., Marcellino, M., Mazzi, G. L. & Papailias, F. (2016). *Big data and macroeconomic nowcasting: From data access to modelling.* European Union.

Belloni, A. & Chernozhukov, V. (2013). Least squares after model selection in high-dimensional sparse models. *Bernoulli*, *19*(2), 521–547.

Bergmeir, C. & Benıtez, J. M. (2012). On the use of cross-validation for time series predictor evaluation. *Information Sciences*, *191*, 192–213.

Brockwell, P. J., Davis, R. A. & Calder, M. V. (2002). *Introduction to Time Series and Forecasting* (Vol. 2). Springer.

Bulut, L. (2018). Google trends and the forecasting performance of exchange rate models. *Journal of Forecasting*, *37*(3), 303–315.

Choi, H. & Varian, H. (2011). Predicting the present with Google trends. *Economic Record*, *88*, 2–9.

D'Amuri, F. & Marcucci, J. (2017). The predictive power of Google searches in forecasting us unemployment. *International Journal of Forecasting*, *33*(4), 801–816.

De Nederlandse Bank. (2019). *Background.* https://statistiek.dnb.nl/en/downloads/index.aspx#/details/residential-mortgages-extended-by-dutch-mfis-to-dutch-households-adjusted-for-securitisations-month/dataset/4621251f-8eb9-4c9d-8597-8ae091b789bb/resource/ee07759a-aa91-427d-b8f7-b105671a9a37

The Dutch Mortgage Market (tech. rep.). (2014). Nederlandse Vereniging van Banken.

Fondeur, Y. & Karamé, F. (2013). Can Google data help predict french youth unemployment? *Economic Modelling*, *30*, 117–125.

Gandomi, A. & Haider, M. (2015). Beyond the hype: Big data concepts, methods, and analytics. *International Journal of Information Management*, *35*(2), 137–144.

Ghysels, E., Plazzi, A., Valkanov, R. & Torous, W. (2013). Forecasting real estate prices. *Handbook of Economic Forecasting*, *2*, 509–580.

Glaeser, E. L. & Gyourko, J. (2006). *Housing dynamics* (tech. rep.). National Bureau of Economic Research.

Götz, T. B. & Knetsch, T. A. (2018). Google data in bridge equation models for German GDP. *International Journal of Forecasting*, *35*(1), 45–66.

Greene, W. H. (2018). *Econometric Analysis.* Pearson.

Harris, R. I. (1992). Testing for unit roots using the augmented Dickey-Fuller test: Some issues relating to the size, power and the lag structure of the test. *Economics Letters*, *38*(4), 381–386.

Harvey, D., Leybourne, S. & Newbold, P. (1997). Testing the equality of prediction mean squared errors. *International Journal of Forecasting*, *13*(2), 281–291.

Hastie, T., Tibshirani, R. & Wainwright, M. (2015). *Statistical Learning with Sparsity: The Lasso and Generalizations*. Chapman; Hall/CRC.

Hlaváček, M. & Komarek, L. (2009). Property price determinants in the Czech regions. *CNB Financial Stability Report 2008/2009* (pp. 82–91). Czech National Bank. https://EconPapers.repec.org/RePEc:cnb:ocpubc:fsr0809/2

Humphrey, B. D. (2010). *Forecasting existing home sales using Google search engine queries* (Master's thesis). Duke University.

Hyndman, R. J. & Athanasopoulos, G. (n.d.). Forecasting: Principles and practice. https://otexts.com/fpp2/stationarity.html

Independer. (n.d.). *Bouwdepot in je hypotheek: Voorwaarden en maximale hoogte.* https://www.independer.nl/hypotheek/info/huis-gekocht/bouwdepot

James, G., Witten, D., Hastie, T. & Tibshirani, R. (2013). *An Introduction to Statistical Learning* (Vol. 112). Springer.

Kouwenberg, R. & Zwinkels, R. (2014). Forecasting the US housing market. *International Journal of Forecasting*, *30*(3), 415–425.

McLaren, N. & Shanbhogue, R. (2011). Using internet search data as economic indicators. *Bank of England Quarterly Bulletin*, (2011), Q2.

Mohebbi, M., Vanderkam, D., Kodysh, J., Schonberger, R., Choi, H. & Kumar, S. (2011). Google Correlate Whitepaper.

Parker, J., Cuthbertson, C., Loveridge, S., Skidmore, M. & Dyar, W. (2017). Forecasting state-level premature deaths from alcohol, drugs, and suicides using Google Trends data. *Journal of Affective Disorders*, *213*, 9–15.

Pentland, A. (2010). *Honest signals: How they shape our world.* MIT press.

Santillana, M., Zhang, D. W., Althouse, B. M. & Ayers, J. W. (2014). What can digital disease detection learn from (an external revision to) Google Flu Trends? *American Journal of Preventive Medicine*, *47*(3), 341–347.

Stephens-Davidowitz, S. & Varian, H. (2014). A Hands-on Guide to Google Data.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, *58*(1), 267–288.

Tsatsaronis, K. & Zhu, H. (2004). What drives housing price dynamics: Cross-country evidence. *BIS Quarterly Review*.

van der Wal, E. & Tamminga, W. (2008). *Why the average dwelling purchase price is not an indicator* (tech. rep.). Centraal Bureau voor de Statistiek.

van Dijk, D. W. & Francke, M. K. (2018). Internet search behavior, liquidity and prices in the housing market. *Real Estate Economics*, *46*(2), 368–403.

Veldhuizen, S. v., Vogt, B. & Voogt, B. (2016). Internet searches and transactions on the Dutch housing market. *Applied Economics Letters*, *23*(18), 1321–1324.

Wu, L. & Brynjolfsson, E. (2015). The future of prediction: How Google searches foreshadow housing prices and sales. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.2022293

Xu, L. & Tang, B. (2014). On the determinants of UK house prices. *International Journal of Economic Research*, *5*(2), 57–64.

Yu, L., Zhao, Y., Tang, L. & Yang, Z. (2017). Online big data-driven oil consumption forecasting with Google trends. *International Journal of Forecasting*, *35*(1), 213–223.

Zhang, Y., Li, R. & Tsai, C.-L. (2010). Regularization parameter selections via generalized information criterion. *Journal of the American Statistical Association*, *105*(489), 312–323.

Zhao, P. & Yu, B. (2006). On model selection consistency of lasso. *Journal of Machine learning research*, *7*, 2541–2563.