



INSTITUTE  
OF ECONOMIC STUDIES  
Faculty of Social Sciences  
Charles University

$$\frac{n!}{(n-1)!} p^{m-1} (1-p)^{n-m} = p \sum_{\ell=0}^{n-1} \frac{\ell+1}{n} \frac{(n-1)!}{(n-1-\ell)! \ell!} p^{\ell} (1-p)^{n-1-\ell}$$
$$= p \frac{n-1}{n} \sum_{\ell=0}^{n-1} \left[ \frac{\ell}{n-1} + \frac{1}{n-1} \right] \frac{(n-1)!}{(n-1-\ell)! \ell!} p^{\ell} (1-p)^{n-1-\ell} = p^2 \frac{n-1}{n} +$$

$$\frac{\ell!}{(n-1)!} p^{m-1} (1-p)^{n-m} = p \sum_{\ell=0}^{n-1} \frac{\ell+1}{n} \frac{(n-1)!}{(n-1-\ell)! \ell!} p^{\ell} (1-p)^{n-1-\ell} = p \frac{n-1}{n} \sum_{\ell=0}^{n-1} \left[ \frac{\ell}{n-1} + \frac{1}{n-1} \right] \frac{(n-1)!}{(n-1-\ell)! \ell!} p^{\ell} (1-p)^{n-1-\ell} = p^2 \frac{n-1}{n} +$$

Institute of Economic Studies,  
Faculty of Social Sciences,  
Charles University in Prague

[UK FSV – IES]

Opletalova 26  
CZ-110 00, Prague  
E-mail : [ies@fsv.cuni.cz](mailto:ies@fsv.cuni.cz)  
<http://ies.fsv.cuni.cz>

Institut ekonomických studií  
Fakulta sociálních věd  
Univerzita Karlova v Praze

Opletalova 26  
110 00 Praha 1

E-mail : [ies@fsv.cuni.cz](mailto:ies@fsv.cuni.cz)  
<http://ies.fsv.cuni.cz>

**Disclaimer:** The IES Working Papers is an online paper series for works by the faculty and students of the Institute of Economic Studies, Faculty of Social Sciences, Charles University in Prague, Czech Republic. The papers are peer reviewed. The views expressed in documents served by this site do not reflect the views of the IES or any other Charles University Department. They are the sole property of the respective authors. Additional info at: [ies@fsv.cuni.cz](mailto:ies@fsv.cuni.cz)

**Copyright Notice:** Although all documents published by the IES are provided without charge, they are licensed for personal, academic or educational use. All rights are reserved by the authors.

**Citations:** All references to documents served by this site must be appropriately cited.

**Bibliographic information:**

Pavlicek J. and Kristoufek L. (2019): "Modeling UK Mortgage Demand Using Online Searches" IES Working Papers 18/2019. IES FSV. Charles University.

This paper can be downloaded at: <http://ies.fsv.cuni.cz>

# Modeling UK Mortgage Demand Using Online Searches

Jaroslav Pavlicek<sup>a</sup>  
Ladislav Kristoufek<sup>a</sup>

<sup>a</sup>Institute of Economic Studies, Faculty of Social Sciences, Charles University  
Opletalova 21, 110 00, Prague, Czech Republic  
Email (corresponding author): [jaroslav.pavlicek@hotmail.com](mailto:jaroslav.pavlicek@hotmail.com)

July 2019

## **Abstract:**

The internet has become the primary source of information for most of the population in modern economies, and as such, it provides an enormous amount of readily available data. Among these are the data on the internet search queries, which have been shown to improve forecasting models for various economic and financial series. In the aftermath of the global financial crisis, modeling and forecasting mortgage demand and subsequent approvals have become a central issue in the banking sector as well as for governments and regulators. Here, we provide new insights into the dynamics of the UK mortgage market, specifically the demand for mortgages measured by new mortgage approvals, and whether or how models of this market can be improved by incorporating the online searches of potential mortgage applicants. Because online searches are expected to be one of the last steps before a customer's actual application for a large share of the population, intuitive utility is an appealing approach. We compare two baseline models – an autoregressive model and a structural model with relevant macroeconomic variables – with their extensions utilizing online searches on Google. We find that the extended models better explain the number of new mortgage approvals and markedly improve their nowcasting and forecasting performance.

**JEL:** C22, C52, C53, C82, E27, E51

**Keywords:** Mortgage, online data, Google Trends, forecasting

**Acknowledgements:** Support from the Charles University PRIMUS program (project PRIMUS/19/HUM/17) and SVV project 260 463 is highly appreciated.

# 1 Introduction

During the last two decades, the internet has become an inevitable part of our lives. For many people, it has replaced books and newspapers as the primary source of readily accessible information. As a result, internet activity reflects people's current behavior and their future needs and desires. Because online searching for information is generally a multistep procedure, utilizing search engines becomes an essential part and the usual starting point of general online activity. Data on internet searches can thus provide real-time insights into the current and future demand for certain goods and services. Since 2004, Google has been providing data on relative search volumes for particular terms through a facility that is presently called Google Trends (previously Google Insights), and the employment of such data has been gaining popularity in the academic literature as researchers have found ways to make use of this new type of data.

At the same time, mortgage lending has become an important macroeconomic indicator of the state of the economy. Excess mortgage lending has started financial crises in the past, and mortgage lending markets are currently closely controlled by the regulators in the wake of the global financial crisis (Bhardwaj & Sengupta, 2012; Mocetti & Viviano, 2017; Agarwal & Zhang, 2018). Mortgages are often the largest liability of an individual or a family, and they form the majority of total lending provided to individuals and families<sup>1</sup>. The growth in total lending to individuals is one of the risk assessment indicators considered by the European Central Bank. Understanding the decision-making process, priorities and timing of the demand for mortgages have thus become essential for governments, central banks and policy-makers in general.

The decision of an individual to take out a mortgage is influenced by various factors. Many factors, such as economic growth, interest rates or house prices, are commonly collected by various institutions and influence aggregate mortgage demand (Basten & Koch, 2015; Kim & Wang, 2018; Adelino et al., 2019). However, some factors, such as an individual's preference for own housing and satisfaction with her current living conditions, are specific to each individual and thus more difficult to observe and aggregate without a further specific survey. With the internet becoming increasingly ubiquitous, more people rely on it when searching for various goods, particularly when a notable financial commitment is involved (Wu & Brynjolfsson, 2015). Because a home purchase is usually the largest liability of an individual and requires a significant amount of funds (Plerhoples Stacy et al., 2018; Badarinza, 2019), the data on the amount of mortgage-related searches could help to explain the demand for new mortgages reflected in the eventual mortgage approvals.

Saxa (2014) was the first to relate the amount of new mortgages to online searches. Using monthly data from the Czech Republic for 2007-2014, he compared simple autoregressive models of month-on-month growth rates in the number of new mortgages to models extended by the inclusion of a mortgage search variable. The models extended to include search variables showed both better in-sample fit (an increase of up to 34 percentage points

---

<sup>1</sup>According to Bankstats tables published by the Bank of England, mortgage lending accounted for approximately 87% of total lending to individuals as of December 2018.

as measured by the adjusted  $R^2$ ) and better out-of-sample forecasting performance (a decrease in the root mean squared error of 23% for the model augmented by Google searches). Oehler (2019) examined the usefulness of Google searches to explain and predict new mortgages in Germany during the period 2004-2018. He evaluated several search queries and added unemployment and the interest rate to the model to control for other drivers of mortgage demand. Using a stepwise regression procedure, he selected the best predictors and the best models both with and without the Google search variables. In his settings, the models augmented by the Google search variables showed the smallest mean absolute errors and root mean squared errors.

Following the researchers who previously examined this topic, we assume that the supply of new mortgages is not limited and directly relate the demand to the volume provided. We assume that an internet search for mortgage-related terms is involved (as one of the last steps) in an individual's decision-making process of buying a home and obtaining a mortgage. We extend the work of McLaren & Shanbhogue (2011), who suggested the usefulness of using Google Trends data as economic indicators for the United Kingdom based on unemployment and house prices. We analyze the series of net monthly mortgage approvals in the UK provided by the Bank of England. Four competing models are compared and selected among based on stepwise regression and  $k$ -fold cross-validation. The competing models include the standard autoregressive model as the baseline and a structural model built on interest rates, unemployment rates, GDP growth and the housing price index.

The paper is structured as follows. Section 2 summarizes the literature on the use of internet-based data in economics. Section 3 provides an overview of the data used in our paper. Section 4 covers the baseline and Google search-based models and presents the results. Section 5 concludes the paper. We show that the inclusion of Google searches in the models markedly improves their performance. Both in-sample and out-of-sample, the Google-based models improve the adjusted  $R^2$  by almost 0.3 compared to the levels below 0.05 for the baseline models. Furthermore, in a situation in which the stepwise procedure can choose among multiple variables, search queries are preferred over the macroeconomic indicators. The mortgage-related Google searches thus provide additional information about the current and future dynamics of mortgage demand above the information provided by the standard macroeconomic indicators and do not merely serve as a good substitute for them when their data are absent or reported with a lag.

## 2 Internet data in the economics and finance literature

To the best of our knowledge, Ettredge et al. (2005) were the first to publish a paper that employed web search data in the econometric literature. At that time, Google search data were not available, and the authors used the WordTracker's Top 500 Keyword Report published by Rivergold Associates, LTD – which offered weekly statistics of terms that were searched for the most with their relative frequencies. They found a positive significant relationship between the job search variables and official US unemployment data. However,

due data limitations, they did not perform time-series analysis and relied only on cross-sectional methods.

In 2008, Google launched the first version of the Google Trends facility (at that time Google Insights) that allowed researchers to collect data on the volume of searches for particular terms in particular locations and a given period of time. The seminal work of Choi & Varian (2009b) laid the foundation for all future econometric research concerning Google search data obtained through this service. They examined the usefulness of Google search data in predicting US retail sales, automotive sales, home sales and travel destinations. Choi & Varian (2009a) extended the analysis to the US initial claims for unemployment benefits, and Choi & Varian (2012) added consumer confidence. In their papers, they used simple autoregressive models as baselines and compared them to models augmented with the search data to produce near-term predictions of the previously mentioned economic indicators. Based on a comparison of the mean absolute errors, the models with Google variables performed better, with improvements ranging between 5% and 20% compared to the baseline models.

A number of studies that addressed the plausibility of internet data in making predictions about various economic variables followed. Bughin (2011) used error-correction models to forecast Belgian unemployment and retail sales. D'Amuri & Marcucci (2010, 2017) performed extensive research examining the relationship between the US unemployment rate and Google search data. Their work focused on predictive power by comparing 520 different models with different variables and their transformations. They showed that models that included Google data performed better in terms of a smaller mean squared error. The best models even outperformed the Survey of Professional Forecasters conducted by the Philadelphia Fed. Kholodilin et al. (2009) nowcasted US private consumption using a number of autoregressive models with aggregated search indices. Based on the root mean squared error, they compared the forecasting performance during two distinct periods (recession/non-recession) and arrived at the conclusion that during normal times, both the models with and without Google search data perform quite similarly, but during turbulent periods, the models using the Google search data tend to outperform the baseline models. Schmidt & Vosen (2011, 2012a) introduced nowcasting models for private consumption in the US and Germany that outperformed the survey-based indicators in the out-of-sample exercises. Schmidt & Vosen (2012b) showed how Google search data can account for unusual events and enhance predictions of private consumption. Some of the more recent papers adopted Bayesian methods. By combining structural time-series models with the Bayesian approach, Scott & Varian (2014) introduced a new framework for automatic model selection using Google Trends and Google Correlate (another Google facility). They illustrated their approach using weekly US initial claims for unemployment benefits and monthly retail sales. They concluded that the Google data were useful predominantly for identifying turning points.

Several papers have also confirmed the suitability of internet search data in connection with financial variables. Mondria et al. (2010) examined the attention allocation and home bias in international investment, Drake et al. (2012) studied the volume of searches with respect to earnings announcements and tested how the demand for information influences

the effect of such earnings announcements. Preis et al. (2013) confirmed that internet-based data offer insights into the behavior of market participants and found patterns that can be interpreted as early warning signs of stock market moves and reversals. Kristoufek (2013a,b) first found that portfolios that account for the popularity of stocks measured by volume of their Google searches tend to be better diversified, and second, he described the relationship between the price of the digital currency Bitcoin and Google search volumes for the term and its Wikipedia page views, revealing interesting dynamics with positive feedback loops. The latter results were confirmed by Garcia et al. (2014) and Kristoufek (2015). Using the sentiment indicators constructed from Google search data and using the Bayesian moving average approach, Kapounek et al. (2016) examined the determinants of the foreign currency savings and arrived at results questioning the international Fisher effect.

Summarizing the state-of-the art research, internet data have been increasingly gaining popularity, and scholars have recently demonstrated their benefits in a growing number of fields, including economics and finance. Our analysis builds on and extends the results showing that Google search data can enhance predictions of house sales and house prices (McLaren & Shanbhogue, 2011; Choi & Varian, 2009b; Wu & Brynjolfsson, 2015) as well as the volume of new (Saxa, 2014; Oehler, 2019) and the performance of existing mortgages (Askitas & Zimmermann, 2011).

### 3 Data

We base our research on two main series of data – the data on mortgage lending made publicly available by the Bank of England (BoE) and an index of web search volumes from Google. Furthermore, we employ a series of interest on new mortgages – interest rates, house prices, unemployment, and GDP – to control for some of the other potential drivers of demand and determine the extent to which Google searches have the power to aggregate and/or replace these macroeconomic indicators. While most of the series of macroeconomic variables dates back to pre-2000 and the Google Trends series starts in January 2004, the series of interest on new mortgages starts only in January 2009. To keep the relevant interest rate in our exercise, we proceed by analyzing the period January 2009 - December 2018.

#### Mortgage data

The Bank of England publishes monthly statistics on the number (and the nominal value in pounds sterling) of mortgage approvals. This represents an offer by a lender to provide a mortgage, adjusted for cancellations (i.e., previously made approvals that have not been taken up). Generally, the data are published with a month lag on the 21<sup>st</sup> working day of the following month<sup>2</sup>. The bank differentiates between new mortgages approved for a

---

<sup>2</sup>For a more detailed description of the source of the data and the publication process, please visit <http://www.bankofengland.co.uk/statistics/pages/iadb/notesiadb/LtoI.aspx>

home purchase and those approved for remortgaging or other purposes. We focus on the series of net mortgage approvals for home purchases and use it as the main series in our analysis.

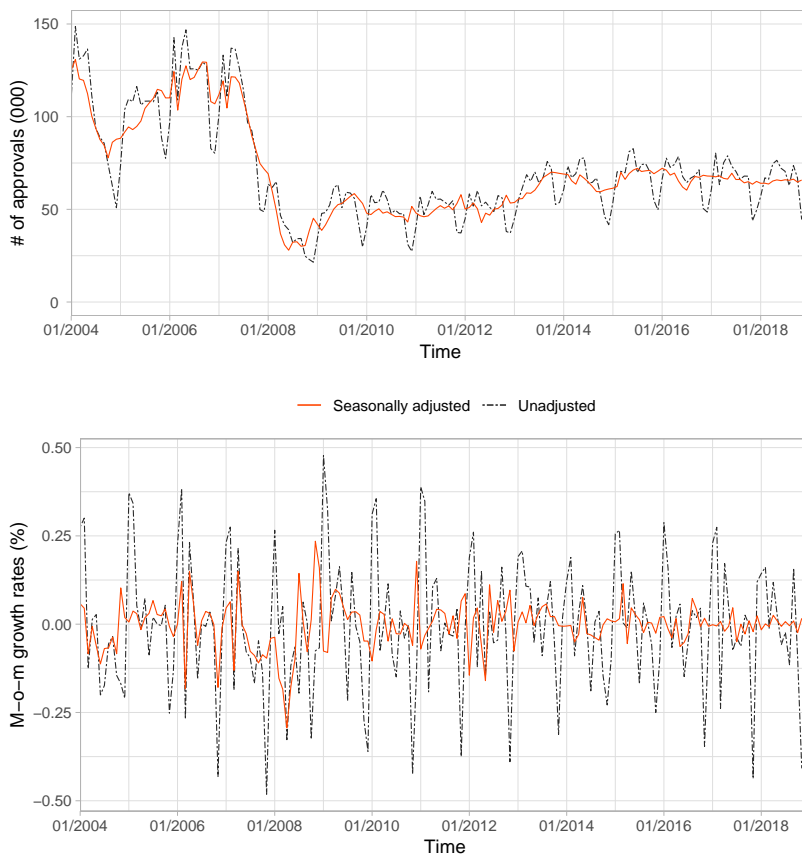


Figure 1: **Mortgage approvals.** Series of net mortgage approvals in UK together with its respective month-on-month (m-o-m) growth rates.

This series is available since April 1993, and in its raw form (Figure 1), the series follows a strong seasonal pattern with significant decreases every January. The series also shows significantly higher values in the period January 2004 - December 2007, consistent with the fact that the UK housing market shrank to almost one fifth of its previous size during the global financial crisis<sup>3</sup>.

---

<sup>3</sup>Specifically, from a peak of 152 thousand houses sold in June 2007 to a low of 33 thousand houses sold in January 2009; see <http://landregistry.data.gov.uk/app/ukhpi/explore> for further detail.



## Google Trends Data

Google is currently the world’s most widely used web search engine<sup>4</sup> and provides data in the form of an index of search volumes for the particular term in a given time period and in a particular geographic area through a facility called Google Trends<sup>5</sup>. The index is based on the volume of the specified search queries relative to the total volume of searches scaled such that the maximum (peak) value in the obtained series is 100. Consequently, if the search volume in the given period is less than 1% popular as the peak, the index value is set to 0. This can lead to some zeros in the index for queries that are not particularly popular or are highly volatile. Furthermore, as a result of this mechanism, even the same nominal amount of searches in every period can lead to smaller index values in periods of increased overall search activity (e.g., holidays).

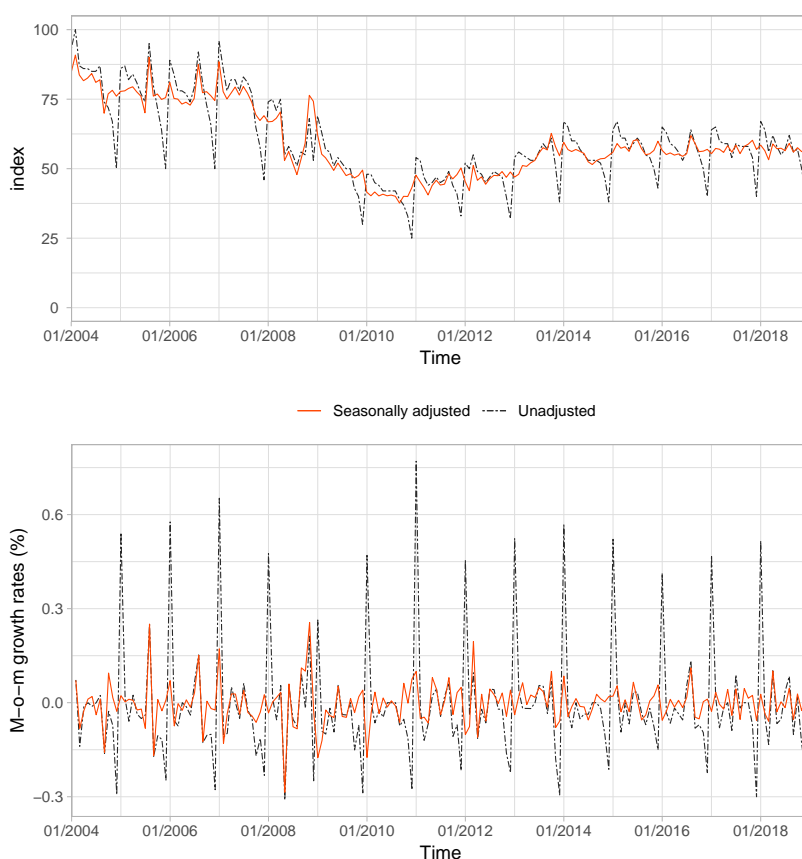


Figure 2: **Google search data.** Series of Google search index for the term “mortgage” and its respective month-on-month (m-o-m) growth rates.

---

<sup>4</sup>Google does not regularly report these figures, and these numbers are thus estimates; see <http://searchengineland.com/google-now-handles-2-999-trillion-searches-per-year-250247>.

<sup>5</sup><https://trends.google.com/trends/>

People who are seeking new mortgages might search for various terms. However, we assume that the vast majority of the queries would include some form of the term *mortgage*. For our analysis, we thus use the index for the term “mortgage,” which includes all queries with *mortgage* in both its singular or plural form and other variations (basically any query that includes the sequence *mortgage*). The index was obtained for the entire UK to match the data on mortgage approvals. For the time frame from January 2004 to December 2018, the index comes at a monthly frequency and is free of any zeros. The series exhibits clear elements of seasonality because the index decreases in December and peaks in January (see Figure 2). The index showed higher values before the outbreak of the global financial crisis but has exhibited a tendency toward recovery thereafter.

## Control variables

To control for other potential drivers of mortgage demand, we include other relevant variables available at a monthly frequency. We collect the series of interest on new mortgages, house prices, unemployment, interest rates, and GDP.

Specifically, we utilize the monthly series of the weighted average interest rates on new mortgages to households (provided by the Bank of England), which is the interest rate directly related to the series of new mortgages. The interest rate is in fact the price of the mortgage and, in accordance with the standard economic theory, should negatively affect the demand for new mortgages and the mortgages provided.

House price data are obtained from the HM Land Registry and represent the monthly average prices in the UK, which are used to calculate the UK House Price Index. House prices, as the price of the underlying in the mortgage contract, should also have a negative influence on the amount of mortgages provided.

Unemployment and GDP are used as proxies for the state of the UK economy. Unemployment was obtained from the Office for National Statistics (ONS) in the form of the number of unemployed in the working age population in the UK. GDP is used in the form of monthly series of the year-to-year growth in gross value added obtained from the ONS. In better states of the economy as measured by these variables, the demand for new homes and thus issued mortgages should show higher values.

## Seasonality issues

Both main series in our analysis exhibit strong seasonal patterns throughout the year. As discussed above, the series of Google searches tends to drop in December and subsequently peak in January (this could be attributed to the way the index is calculated as discussed earlier). Similarly, the series of mortgage approvals tends to peak in December and then drop at the beginning of the new year because people probably attempt to complete the process of obtaining a mortgage before the year end and do not wish to start a new process in December. As the process can take some time, this probably leads to fewer mortgages

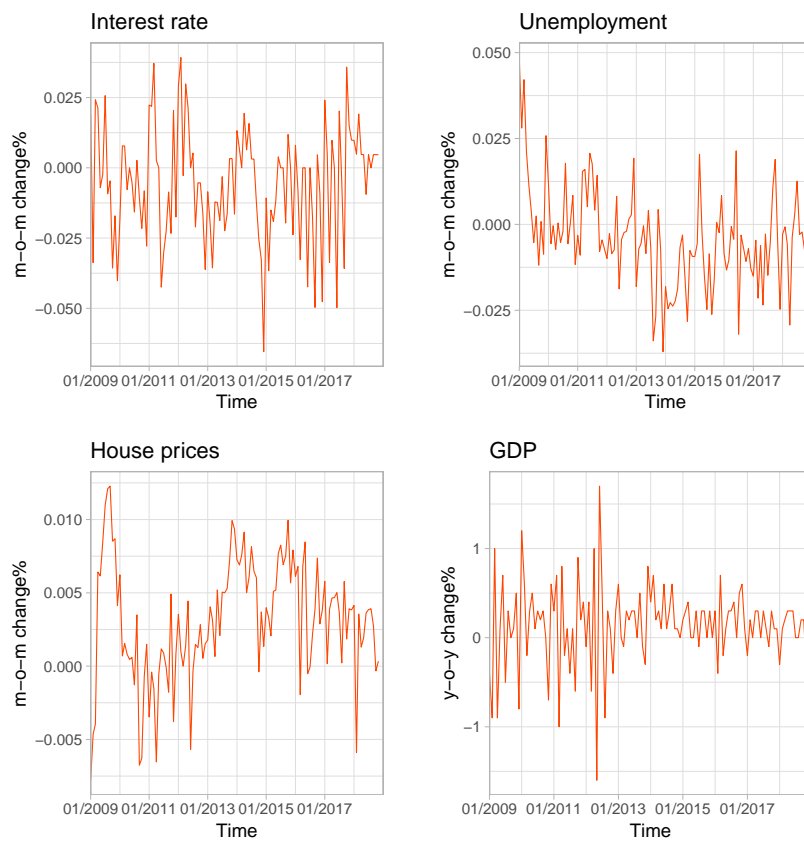


Figure 3: **Control variables.** The figure shows plots of four series used as control variables in our analysis.

generally being approved in January. The series of house prices is not seasonally adjusted, and theoretically, there could also be some seasonality present. To have comparable time series and to ensure that the Google searches explain the actual mortgages rather than seasonality in mortgages, we apply the standard X-11 method of seasonality adjustment. The series of GDP and unemployment already come seasonally adjusted. The series of interest rates on mortgages does not show seasonal patterns, and we keep it in its raw form.

According to the KPSS (Kwiatkowski et al., 1992) and ADF (Dickey & Fuller, 1979) tests, all the series with the exception of GDP exhibit some degree of non-stationarity in their level form (Table 1). The pair of tests mentioned above suggests that the logarithmically differenced series do not contain a unit root and the null of level stationarity is not rejected, and we proceed in our analysis with the logarithmic growth rates of all series with the exception of GDP, which is already in the form of the year-on-year growth rates, i.e., stationary.

	KPSS	<i>p</i> -value	ADF	<i>p</i> -value
<b>Nr. of net approvals</b>				
-levels	2.061	<0.01	-2.438	0.392
-log growth rate	0.050	>0.1	-6.086	<0.01
<b>Google search index</b>				
-levels	1.748	<0.01	-1.520	0.777
-log growth rate	0.164	>0.1	-7.506	<0.01
<b>Interest</b>				
-levels	2.403	<0.01	-2.934	0.189
-log growth rate	0.083	>0.1	-5.573	<0.01
<b>HPI</b>				
-levels	2.665	<0.01	-2.266	0.465
-log growth rate	0.208	>0.1	-3.739	0.024
<b>Unemployment</b>				
-levels	0.871	<0.01	-1.214	0.902
-log growth rate	1.196	<0.01	-3.771	0.022
<b>GDP growth rate</b>				
-raw	0.016	>0.1	-3.283	0.076

Table 1: **Unit-root and stationarity.** Test statistics and *p*-values for the ADF and KPSS tests for unit-root/stationarity.

## 4 Methods and results

### 4.1 Models

We examine the ability of the Google searches to improve models of the number of new mortgage approvals. To assess how much of the variation in new mortgages can be explained by searching for mortgage-related terms and their ability to complement and/or beat other potential predictors, we study the quality of fit of four competing models: (1) a *naive* baseline model in which the series of mortgage approvals is regressed only on its past values, i.e., an autoregressive model; (2) a structural model based on the set of control variables and not the Google search series; (3) mortgage approvals regressed on their past values and series of current and lagged Google searches; and (4) a model with the full set of controls including the Google searches. The following models are thus considered:

**Model 1:**

$$\mathbf{Mortgage} = \boldsymbol{\alpha} + \mathbf{Mortgage} \mathbb{L}^\beta \boldsymbol{\beta} + \boldsymbol{\varepsilon}, \quad (1)$$

**Model 2:**

$$\begin{aligned} \mathbf{Mortgage} = & \boldsymbol{\alpha} + \mathbf{Mortgage} \mathbb{L}^\beta \boldsymbol{\beta} + \mathbf{HPI} \mathbb{L}^\delta \boldsymbol{\delta} + \mathbf{Interest} \mathbb{L}^\theta \boldsymbol{\theta} \\ & + \mathbf{Unemployment} \mathbb{L}^\psi \boldsymbol{\psi} + \mathbf{GDP} \mathbb{L}^\zeta \boldsymbol{\zeta} + \boldsymbol{\varepsilon}, \end{aligned} \quad (2)$$

**Model 3:**

$$\mathbf{Mortgage} = \boldsymbol{\alpha} + \mathbf{Mortgage} \mathbb{L}^\beta \boldsymbol{\beta} + \mathbf{Google} \mathbb{L}^\gamma \boldsymbol{\gamma} + \boldsymbol{\varepsilon}, \quad (3)$$

**Model 4:**

$$\begin{aligned} \mathbf{Mortgage} = & \boldsymbol{\alpha} + \mathbf{Mortgage} \mathbb{L}^\beta \boldsymbol{\beta} + \mathbf{Google} \mathbb{L}^\gamma \boldsymbol{\gamma} + \mathbf{HPI} \mathbb{L}^\delta \boldsymbol{\delta} + \mathbf{Interest} \mathbb{L}^\theta \boldsymbol{\theta} \\ & + \mathbf{Unemployment} \mathbb{L}^\psi \boldsymbol{\psi} + \mathbf{GDP} \mathbb{L}^\zeta \boldsymbol{\zeta} + \boldsymbol{\varepsilon}. \end{aligned} \quad (4)$$

**Mortgage** stands for the column vector of all observations of new mortgage approvals; **Mortgage** is a matrix with observations in the rows and its lagged values up to the 12-month lag in the columns;  $\boldsymbol{\beta}$  is a column vector with coefficients for the individual lags;  $\mathbb{L}^\beta$  is a matrix with ones and zeros on the diagonal that indicate the lags that are allowed to enter the estimation procedure (as explained later); and  $\boldsymbol{\alpha}$  stands for a column vector filled with values of the intercept term. Parallel notation intuitively applies to all used series, i.e., the housing price index (**HPI**), interest rates (**Interest**), unemployment (**Unemployment**), GDP (**GDP**), and Google searches (**Google**).

Most of the macroeconomic variables are published with a lag of several weeks or months. In contrast, information on Google searches is available in real time. To account for this, only the series of Google searches is permitted to have a contemporaneous relationship with mortgage approvals, i.e., with a lag of zero. In our analysis, we allow the model to have up to a 12-month lag. As a result, we effectively end up working with a sample of 107 observations (Feb 2010 - Dec 2018).

## 4.2 In-sample performance

To identify the relevant control variables and their appropriate lags for each of the above-proposed models, we employ a stepwise selection procedure with a sequential replacement algorithm. As opposed to the forward selection used by Oehler (2019), which starts with a plain model and then subsequently adds variables based on their ability to reduce the residual sum of squares, the sequential replacement algorithm goes back at each step and evaluates the potential of other variables by replacing the ones already selected. Using this algorithm, we construct a set of competing models with  $1, \dots, n$  explanatory variables for each model. To limit the models' size, we limit  $n$  to 10. Using this technique, we obtain a set of the best models with  $1, \dots, n$  explanatory variables.

The set of  $n$  proposed models is then evaluated in a semi-out-of-sample setting using the  $k$ -fold cross-validation technique with  $k = 4$ . For each proposed model, the sample is divided into 4 subsamples of approximately equal sizes, with three subsamples being used to estimate the model parameters and the fourth to calculate the implied errors. This process is applied 4 times so that each combination of the training and testing samples is used. For each model, the best performing  $n$  is selected. The final models and their coefficients are summarized in Table 2. The comparison of the models without the Google variables (Model 1 and Model 2) and the enhanced models augmented by the Google search index (Model 3 and Model 4) reveals considerable improvement in the latter two.

The final Model 1 includes the 1-month, 5-month and 10-month lags, with the latter two being statistically insignificant. All three have a negative sign, which is rather counter-intuitive, but it can be explained as a sign of mean reversion and returning to a long-term value. Nevertheless, a weak self-explanatory power of the series is reflected in a very low  $R^2$  of 0.06 and an adjusted  $R^2$  of just 0.03. For the model based on the macroeconomic variables and their lags, the final Model 2 selected the 6-month lag of GDP growth as the only explanatory variable. The coefficient on the variable is estimated to be negative, which goes against our initial hypothesis, and yet again, the explanatory power of this specification is again very low with both  $R^2$  and adjusted  $R^2$  being below 0.05.

The final Model 3 includes five variables, out of which four are the Google searches, specifically the current one and the 1-, 2- and 4-month lags. All of the coefficients on the Google variables are statistically significant and have the expected positive sign according to the initial hypothesis. The inclusion of only the reasonably recent lags supports the hypothesis of Google searching being one of the last steps in the process before obtaining a mortgage, as one would not be expected to be searching too much in advance. The negative effect of previous mortgage approvals remains here as well but only for the first

<i>Dependent variable:</i>				
Approvals (Net mortgage approvals for home purchases)				
	Model 1	Model 2	Model 3	Model 4
Approvals.1	-0.182* (0.095)		-0.218** (0.094)	
Approvals.5	-0.115 (0.095)			
Approvals.10	-0.116 (0.091)			
Google			0.308*** (0.088)	0.338*** (0.086)
Google.1			0.422*** (0.087)	0.366*** (0.084)
Google.2			0.358*** (0.084)	0.321*** (0.081)
Google.4			0.175** (0.079)	0.224*** (0.076)
GDP.6		-0.024** (0.010)		-0.022** (0.009)
Constant	0.003 (0.005)	0.006 (0.005)	-0.0001 (0.004)	0.003 (0.004)
Observations	107	107	107	107
R <sup>2</sup>	0.060	0.049	0.305	0.307
Adjusted R <sup>2</sup>	0.033	0.040	0.270	0.273
Residual SE	0.047 (df = 103)	0.046 (df = 105)	0.040 (df = 101)	0.040 (df = 101)
F Statistic	2.188* (df = 3; 103)	5.380** (df = 1; 105)	8.859*** (df = 5; 101)	8.958*** (df = 5; 101)

*Note:* SEs in parentheses \* $p < 0.1$ ; \*\* $p < 0.05$ ; \*\*\* $p < 0.01$

Table 2: **Regression outputs.** Regression outputs for the set of models in an in-sample exercise.

lag, which supports the idea of mean reversion in approvals. The quality of the model improves considerably with an  $R^2$  of 0.30 and the adjusted  $R^2$  of 0.27, compared to 0.06 and 0.03, respectively, for the original model without Google searches.

Model 4 is a direct comparison of the Google searches with the other control variables. The stepwise regression procedure selected the same four Google variables as in Model 3 and only one variable from the set of other controls. The coefficients on searches are statistically significant and with the expected positive signs of the effects. The sixth lag of GDP growth is the only one of the macroeconomic variables selected for inclusion in the model. Comparing Model 2 and Model 4, we can see that the Google searches add relevant information about the dynamics of mortgage approvals, but moreover, such information is of much higher quality and different from the information provided by the macroeconomic variables, the usefulness of which is rather questionable for the analyzed market. This is nicely reflected in the (adjusted) coefficient of determination reaching 0.31 (0.27) compared to 0.05 (0.04) of the structural model based solely on the macroeconomic variables.

### 4.3 Out-of-sample performance

We continue our analysis with an out-of-sample exercise. We compare the results of competing models based on the quality of dynamic forecasts using an expanding window estimation. For an initial sample of size  $W$ , the forecasts for time  $t$  are predicted from the model estimated on the sample of  $t-1$  observations dated  $1, \dots, t-1$  for each  $t \in [W+1, T]$ . In this way, we obtain  $T-W$  one-step-ahead forecasts, where  $T$  is the total sample size.

We base our comparison of the competing models on the standard evaluation metrics, such as the mean absolute error (MAE) and the root mean squared error (RMSE), for the series of the one-step-ahead predictions. To further disentangle the differences between the competing models, we evaluate the directional accuracy as measured by the mean directional accuracy (MDA)<sup>6</sup>.

We treat the last 24 observations (January 2016 - December 2018) as if they represented new information, and we re-estimate the four models using the same procedure as in the previous section but on a restricted sample. For our analysis, we set the size of the initial window such that the sample to forecast is 24 months ( $W = 89$ )<sup>7</sup>. The estimated models are of generally similar structure to those derived in the previous section (Table 3), with

---

<sup>6</sup>MDA is defined as  $MDA = \frac{1}{T} \sum_{i=1}^T a_i$ , where

$$a_i = \begin{cases} 1 & \text{if } \text{sign}(y_i) = \text{sign}(\hat{y}_i) \\ 0 & \text{if } \text{sign}(y_i) \neq \text{sign}(\hat{y}_i) \end{cases}$$

and  $\text{sign}(x)$  yields -1 for negative numbers, 1 for positive numbers, and 0 for zero. The statistics are interpreted as the percentage of predictions with the correct direction.

<sup>7</sup>As the set of available data is reduced by the last 24 observations, we might encounter difficulties in estimation during the  $k$ -fold cross-validation step. For this reason, we decrease the number of possible lags included in the model to 6 (no higher lag was identified in the regression analysis in any case). As a result, we effectively ultimately have a sample of 89 observations (Jul 2009 - Dec 2016) for the first estimation (113 in total).



	<i>Dependent variable:</i>			
	Approvals (Net mortgage approvals for home purchases)			
	Model 1	Model 2	Model 3	Model 4
Approvals.1	-0.148 (0.106)		-0.186* (0.105)	
Approvals.3			0.125 (0.096)	
Google			0.318*** (0.098)	0.324*** (0.093)
Google.1			0.403*** (0.100)	0.342*** (0.093)
Google.2			0.364*** (0.102)	0.261*** (0.094)
Google.4			0.162* (0.096)	0.321*** (0.092)
GDP.6		-0.024** (0.011)		-0.023** (0.009)
Unemployment.4				1.002** (0.389)
Unemployment.5				-0.733* (0.373)
Constant	0.003 (0.005)	0.006 (0.006)	0.001 (0.005)	0.006 (0.005)
Observations	89	89	89	89
R <sup>2</sup>	0.022	0.055	0.280	0.353
Adjusted R <sup>2</sup>	0.011	0.044	0.227	0.297
Residual SE	0.051 (df = 87)	0.051 (df = 87)	0.046 (df = 82)	0.043 (df = 81)
F Statistic	1.968 (df = 1; 87)	5.081** (df = 1; 87)	5.315*** (df = 6; 82)	6.307*** (df = 7; 81)

*Note:* SEs in parentheses \* $p < 0.1$ ; \*\* $p < 0.05$ ; \*\*\* $p < 0.01$

Table 3: **Out-of-sample regression outputs.** Regression outputs for the set of models employed for the out-of-sample forecasting exercise.

the difference being that a pair of lags of unemployment is selected for inclusion in Model 4. Overall, the out-of-sample models demonstrate the dominance of the Google search-based models over the baseline models, whether the autoregressive model or the structural model with macroeconomic variables. These results are reflected in the forecasting statistics, which are summarized in Table 4. There, Model 3 emerges as a clear winner with respect to all three measures. Model 4, even though its MDA reaches the same level as that of Model 3, shows signs of over-fitting, with MAE and RMSE being higher or very close to those of Model 1 and Model 2, i.e., of the baseline models. The most significant area of improvement is in directional accuracy, where both models with the Google data perform considerably better<sup>8</sup>.

	Model 1	Model 2	Model 3	Model 4
<b>MAE</b>	0.0146	0.0146	0.0133	0.0157
<b>RMSE</b>	0.0187	0.0205	0.0166	0.0198
<b>MDA</b>	0.5000	0.6667	0.8333	0.8333

Table 4: **Expanding window metrics.** The table summarizes the main metrics used for the evaluation of out-of-sample performance on expanding windows for the individual models.

## 5 Conclusion

The aim of this exercise is to examine whether Google search data can effectively serve as a proxy for mortgage demand in the UK to model and forecast the number of new mortgages and whether they can substitute for or complement macroeconomic indicators and autocorrelation structure in doing so. We base the analysis on the month-on-month logarithmic growth rates of the series of the net mortgage approvals for home purchases and the series of Google search data for the term “mortgage” together with the series of interest rates, house prices, unemployment and GDP in the period from January 2009 to December 2018.

We consider four competing models: (1) a *naive* autoregressive model; (2) a structural model augmented by more classical potential predictors (interest rate, housing price index,

<sup>8</sup>Additionally, we employ the Diebold-Mariano test (Diebold & Mariano, 1995; Diebold, 2015) as an industry standard, although its results in this specific setting should be interpreted with caution as discussed and criticized by Diebold (2015) himself. Based on MAE, RMSE and MDA, Model 3 shows the best out-of-sample performance, but the Diebold-Mariano statistics do not identify the difference as a significant one with a test statistic (and *p*-value) of 0.7394 (0.2336) and 0.9252 (0.1822) for comparing Model 3 with Model 1 and Model 2, respectively.

unemployment, and GDP); (3) a *naive* model augmented by the Google search variables; and (4) a model with the full set of control variables. Using the stepwise regression procedure, we confirm the explanatory power of the Google searches because adding the Google term into the regression improves the models' fit as measured by adjusted  $R^2$  by approximately 23 percentage points, going from a very low level of 0.04 for Model 2 to 0.27 for Model 4. This is further tested and mostly confirmed in the out-of sample exercise. Augmenting the baseline models with the set of Google search data improves their performance in terms of the root mean squared error. The model without the other control variables shows smaller forecasting errors for the one-step-ahead forecasts after the addition of the Google searches to the regression. However, this improvement is not as evident when other variables are controlled for in the regression. A significant increase in the directional accuracy as measured by the mean directional accuracy confirms the ability of the Google search data to identify turning points as suggested by Scott & Varian (2014). As in most of the current literature on the application of Google search data in economics and finance, this paper successfully illustrates the explanatory power of Google search data in a relatively simple empirical framework and these data's ability to complement or replace conventional sources of data. The extended models provide an evident improvement in the models' quality regardless of the relative calmness of the UK mortgage market in the last few years. One would expect that the utility of the model to improve in more dynamic and turbulent markets.

Although the online search data markedly improve the baseline models' fit and forecasting performance, there are certain limitations that need to be noted. First, the availability of the online data combined with the low frequency of reporting and releasing of the macroeconomic indicators and mortgage approval statistics leads to a restricted dataset. If the dataset were longer, it seems likely that the forecasting improvements would be more profound and reflected in statistical significance in the testing procedures rather than (only) in better performance measures. Second, as already noted, specifically the out-of-sample part of the analyzed period is very calm and exhibits dynamics that are apparently not rich. This makes the forecasting results of the search-based models even more notable. Although speculative, the expectation is that such models will perform even better in more turbulent times, as shown in other topical studies (Preis et al., 2013; Curme et al., 2014; Ranco et al., 2016). Third, although the inclusion of the online searches markedly improves the performance of the models, the data on online search queries certainly do not cover the whole demand for mortgages. This leaves us with room for improvement because some of the potential applicants could often directly use the services of financial advisors or banks, potentially without consulting the internet beforehand. However, this would call for detailed microscopic surveys or data from banks and brokers, which might not be available even to regulators. Overall, specifically for such reasons, we have shown that aggregate search data can successfully serve as a high-quality substitute for detailed surveys and microscopic data and help to improve forecasting models for mortgage demand, both at a very low cost.

## References

- Adelino, M., Gerardi, K., & Hartman-Glaser, B. (2019). Are lemons sold first? Dynamic signaling in the mortgage market. *Journal of Financial Economics*, 132(1), 1–25.
- Agarwal, S. & Zhang, Y. (2018). Effects of government bailouts on mortgage modification. *Journal of Banking & Finance*, 93, 54–70.
- Askatas, N. & Zimmermann, K. (2011). Detecting mortgage delinquencies. *IZA Discussion Papers*, (5895).
- Badarinza, C. (2019). Mortgage debt and social externalities. *Review of Economic Dynamics*, 34, 43–60.
- Basten, C. & Koch, C. (2015). The causal effect of house prices on mortgage demand and mortgage supply: Evidence from Switzerland. *Journal of Housing Economics*, 30, 1–22.
- Bhardwaj, G. & Sengupta, R. (2012). Subprime mortgage design. *Journal of Banking & Finance*, 36(5), 1503–1519.
- Bughin, J. R. (2011). Nowcasting the Belgian economy. *Universite Libre de Bruxelles (ULB) Working Papers Series*, (2012/08).
- Choi, H. & Varian, H. R. (2009a). Predicting initial claims for unemployment benefits. Technical report, Google.
- Choi, H. & Varian, H. R. (2009b). Predicting the present with Google Trends. Technical report, Google.
- Choi, H. & Varian, H. R. (2012). Predicting the present with Google Trends. *The Economic Record*, 88(s1), 2–9.
- Curme, C., Preis, T., Stanley, H., & Moat, H. (2014). Quantifying the semantics of search behavior before stock market moves. *Proceedings of the National Academy of Sciences of the United States of America*, 111(32), 11600–11605.
- D’Amuri, F. & Marcucci, J. (2010). “Google it!” Forecasting the US Unemployment Rate with a Google Job Search index. *Fondazione Eni Enrico Mattei, Working papers*, (2010.31).
- D’Amuri, F. & Marcucci, J. (2017). The predictive power of Google searches in forecasting US unemployment. *International Journal of Forecasting*, 33(4), 801–816.
- Dickey, D. A. & Fuller, W. A. (1979). Distribution of the estimators for autoregressive time series with a unit root. *Journal of the American Statistical Association*, 74(366a), 427–431.

- Diebold, F. X. (2015). Comparing predictive accuracy, twenty years later: A personal perspective on the use and abuse of diebold-mariano tests. *Journal of Business & Economic Statistics*, 33(1), 1–1.
- Diebold, F. X. & Mariano, R. S. (1995). Comparing Predictive Accuracy. *Journal of Business & Economic Statistics*, 13(3), 253–63.
- Drake, M. S., Roulstone, D. T., & Thornock, J. R. (2012). Investor information demand: Evidence from google searches around earnings announcements. *Journal of Accounting Research*, 50(4), 1001–1040.
- Ettredge, M., Gerdes, J., & Karuga, G. (2005). Using web-based search data to predict macroeconomic statistics. *Communications of the ACM*, 48(11), 87–92.
- Garcia, D., Tessone, C., Mavrodiev, P., & Perony, N. (2014). The digital traces of bubbles: feedback cycles between socio-economic signals in the Bitcoin economy. *Journal of the Royal Society Interface*, 11, 20140623.
- Kapounek, S., Deltuvaite, V., & Korab, P. (2016). Determinants of foreign currency savings: Evidence from google search data. *Procedia - Social and Behavioral Sciences*, 220, 166 – 176. 19th International Conference Enterprise and Competitive Environment 2016.
- Kholodilin, K. A., Podstawski, M., Siliverstovs, B., & Burgi, C. (2009). Google Searches as a Means of Improving the Nowcasts of Key Macroeconomic Variables. *Discussion Papers of DIW Berlin*, (946).
- Kim, J. & Wang, Y. (2018). Macroeconomic and distributional effects of mortgage guarantee programs for the poor. *Journal of Economic Dynamics and Control*, 87, 124–151.
- Kristoufek, L. (2013a). BitCoin meets Google Trends and Wikipedia: Quantifying the relationship between phenomena of the Internet era. *Scientific Reports*, 3, 3415.
- Kristoufek, L. (2013b). Can Google Trends search queries contribute to risk diversification? *Scientific Reports*, 3.
- Kristoufek, L. (2015). What are the main drivers of the Bitcoin price? evidence from wavelet coherence analysis. *PLOS ONE*, 10(4), e0123923.
- Kwiatkowski, D., Phillips, P. C., Schmidt, P., & Shin, Y. (1992). Testing the null hypothesis of stationarity against the alternative of a unit root: How sure are we that economic time series have a unit root? *Journal of Econometrics*, 54(1-3), 159 – 178.
- McLaren, N. & Shanbhogue, R. (2011). Using internet search data as economic indicators. *Bank of England Quarterly Bulletin*, 51(2), 134–140.
- Mocetti, S. & Viviano, E. (2017). Looking behind mortgage delinquencies. *Journal of Banking & Finance*, 75, 53–63.

- Mondria, J., Wu, T., & Zhang, Y. (2010). The determinants of international investment and attention allocation: Using internet search query data. *Journal of International Economics*, 82(1), 85 – 95.
- Oehler, S. (2019). Developments in the residential mortgage market in Germany - what can Google data tell us? In B. for International Settlements (Ed.), *Are post-crisis statistical initiatives completed?*, volume 49 of *IFC Bulletins chapters*. Bank for International Settlements.
- Plerhoples Stacy, C., Theodos, B., & Bai, B. (2018). How to prevent mortgage default without skin in the game: Evidence from an integrated homeownership support nonprofit. *Journal of Housing Economics*, 39, 17–24.
- Preis, T., Moat, H., & Stanley, H. (2013). Quantifying trading behavior in financial markets using Google Trends. *Scientific Reports*, 3(1684), 1–6.
- Ranco, G., Bordino, I., Bormetti, G., Caldarelli, G., Lillo, F., & Treccani, M. (2016). Coupling news sentiment with web browsing data improves prediction of intra-day price dynamics. *PLOS ONE*, 11(1), e0146576.
- Saxa, B. (2014). Forecasting Mortgages: Internet Search Data as a Proxy for Mortgage Credit Demand. *Czech National Bank Working Paper Series*, (2014/14).
- Schmidt, T. & Vosen, S. (2011). Forecasting private consumption: Survey-based indicators vs. google trends. *Journal of Forecasting*, 30(6), 565–578.
- Schmidt, T. & Vosen, S. (2012a). A monthly consumption indicator for germany based on internet search query data. *Applied Economics Letters*, 19(7), 683–687.
- Schmidt, T. & Vosen, S. (2012b). Using internet data to account for special events in economic forecasting. *Ruhr Economic Papers*, (0382).
- Scott, S. L. & Varian, H. R. (2014). Predicting the present with bayesian structural time series. *International Journal of Mathematical Modelling and Numerical Optimisation*, 5(1/2), pp.4–23.
- Wu, L. & Brynjolfsson, E. (2015). The Future of Prediction: How Google Searches Fore-shadow Housing Prices and Sales. In *Economic Analysis of the Digital Economy*, NBER Chapters (pp. 89–118). National Bureau of Economic Research, Inc.

# IES Working Paper Series

2019

1. Davit Maskharashvili: *Duopolistic Competition On a Plane*
2. Petr Hanzlík, Petr Teplý: *Key Determinants of Net Interest Margin of EU Banks in the Zero Lower Bound of Interest Rates*
3. Barbora Mátková: *Bank-Sourced Transition Matrices: Are Banks' Internal Credit Risk Estimates Markovian?*
4. Peter Kudela, Tomas Havranek, Dominik Herman, Zuzana Irsova: *Does Daylight Saving Time Save Electricity? Evidence from Slovakia*
5. Dominika Kolcunová, Simona Malovaná: *The Effect of Higher Capital Requirements on Bank Lending: The Capital Surplus Matters*
6. Jaromír Baxa, Tomáš Šestořád: *The Czech Exchange Rate Floor: Depreciation without Inflation?*
7. Karel Janda, Binyi Zhang: *Renewable Energy Financial Modelling: A China Case Study*
8. Anna Alberini, Olha Khymych, Milan Ščasný: *Estimating Energy Price Elasticities When Salience is High: Residential Natural Gas Demand in Ukraine*
9. Anna Alberini, Olha Khymych, Milan Ščasný: *The Elusive Effects of Residential Energy Efficiency Improvements: Evidence from Ukraine*
10. Jozef Baruník, Matěj Nevrla: *Tail Risks, Asset Prices, and Investment Horizons*
11. Barbora Malinska: *Realized Moments and Bond Pricing*
12. Hamza Bennani, Nicolas Fanta, Pavel Gertler, Roman Horvath: *Does Central Bank Communication Signal Future Monetary Policy? The Case of the ECB*
13. Milan Ščasný, Šarlota Smutná: *Estimation of Price and Income Elasticity of Residential Water Demand in the Czech Republic over Three Decades*
14. Mykola Babiak, Olena Chorna, Barbara Pertold-Gebicka: *Minimum Wage Increase and Firm Profitability: Evidence from Poland*
15. Martin Stepanek: *Sectoral Impacts of International Labour Migration and Population Ageing in the Czech Republic*
16. Milan Ščasný, Iva Zvěřinová, Alistair Hunt: *Nature-Based, Structural, or Soft Measures of Adaptation? Preferences for Climate Change Adaptation Measures to Limit Damages from Droughts*
17. Milan Ščasný, Iva Zvěřinová, Vojtěch Máca: *Consumer Preferences for Sustainable and Healthy Lifestyle: Five-Country Discrete Choice Experiments*
18. Jaroslav Pavlíček, Ladislav Kristoufek: *Modeling UK Mortgage Demand Using Online Searches*

All papers can be downloaded at: <http://ies.fsv.cuni.cz>.

---



Univerzita Karlova v Praze, Fakulta sociálních věd  
Institut ekonomických studií [UK FSV – IES] Praha 1, Opletalova 26  
E-mail : [ies@fsv.cuni.cz](mailto:ies@fsv.cuni.cz) <http://ies.fsv.cuni.cz>