EUROPEAN CENTRAL BANK
EUROSYSTEM

Statistics Paper Series

Per Nymand-Andersen, Emmanouil Pantelidis

Google econometrics: nowcasting euro area car sales and big data quality requirements

No 30 / November 2018

Disclaimer: This paper should not be reported as representing the views of the European Central Bank (ECB). The views expressed are those of the authors and do not necessarily reflect those of the ECB.

Contents

Abstract

"Big data" is becoming an increasingly important aspect of our daily lives as the digital sources of information and intelligence that it encompasses become more structured and more publicly available. These sources may enable the generation of new datasets providing high-frequency and timely insights into unconscious digital behaviour and the consequent actions of economic agents, which may, in turn, assist in the generation of early indicators of economic and financial trends and activities. This paper examines the usefulness of Google search data in nowcasting euro area car sales, as a leading macroeconomic indicator, and considers the quality requirements for using these new data sources as a toolkit for sound decision and policy making. The paper finds that, while Google data may have predictive capabilities for nowcasting euro area car sales, further quality improvements in the data source are needed in order to move beyond experimental statistics. If these quality requirements can be met, the resulting advances in theory and knowledge around interpreting big data can be expected to significantly re-shape how we think about and explain both behaviour and complex socio-economic phenomena.

JEL codes: C53, C82, E58, E71.

Keywords: big data, modelling, vector auto regression, nowcasting, statistics, quality, google internet search.

1 Introduction

“Big data” are becoming an increasingly important aspect of our daily lives as the digital sources of information and intelligence that they encompass become more structured and more publicly available. Big data are part of the “data service evolution”. They are borderless and affect the structure and functioning of financial markets, our economies and our societies. The new data service of big data has been identified as having a high growth potential. Big data appear to be a product of the interaction between the “causes” and “effects” of the constantly changing ways in which we live, communicate, socialise and exchange information – our digital trails are available to be explored.

Central banks may benefit from exploring the feasibility of extracting economic signals in near real time and learn from the new sources and methodologies. More specifically, big data may be used to enhance economic forecasts and to provide more timely feedback on the impact of central banking policies – as well as on reaction functions and market and household sentiment – as the effects of policies spread throughout the economy.

The aim of this paper is to experiment with internet search data to provide statistical and econometric evidence of the usefulness of internet search terms in the context of a leading macroeconomic indicator. For example, the paper uses Google internet search data on “Autos & Vehicles” to test its links to new car registrations data (the latter relates to automotive manufacturing as an industry). The automotive industry in Europe has always been a major manufacturing sector, accounting for a large part of the European economy. The European Union is among the world’s largest manufacturers of motor vehicles and its automotive industry sector accounts for 4% of EU GDP. The sector is the largest private investor in research and development in Europe and provides employment for 12 million people. New vehicle registrations can be seen as a leading macroeconomic indicator of economic activity and households’ spending, so the ability to forecast them could provide early signals of potential future economic turning points and directions and would therefore be useful for policy purposes. An increase in household sector expenditure on vehicle purchases could, all other things being equal, indicate that household consumption is increasing and the economy may be expanding. This could alert policymakers to consider a potential tightening of monetary policy, or vice versa. This kind of forecasting is called “nowcasting”, which is defined by Banbura, Giannone, Modugno and Reichlin (2013) as “the prediction of the present, the very near future and the very recent past”. We are therefore interested in the very short-term predictive capabilities of internet search data. Big data can be defined as a source of information and intelligence resulting from the (digital) recording of operations or from the combination of such records. See Nymand-Andersen (2016, 2017).

The fundamental question posed by this paper is therefore whether the digital trail of the relative number of internet searches[1] for vehicles may be useful as a leading indicator of future increases in new vehicle registrations and could therefore be used as a supplemental near real-time indicator to forecast the next release of statistical vehicle registration data. This would require not only a correlation but also a causal connection between “increases in the relative number of internet searches for vehicles” and “increases in the numbers of new vehicle registrations”. Even where both a correlation and a causal connection exist, it cannot be taken at face value.

First, if this causality holds true, what is the lag between the time the aggregated searches for a vehicle are made and the consumer’s actual purchase of a vehicle? Second, how can data be adjusted to distinguish the relative increases of search terms from those searches which are conducted for other purposes, unrelated to a purchase? The volumes of these internet searches must somehow be revised downwards by a certain factor, and this factor may not be stable over time but require adjustment; for example, in the case of an exceptional event such as the recent diesel emissions scandal. These statistical concepts are described in more detail in Chapter 6.

However, the most essential criteria that must be taken into account are whether the data service provider can live up to the mandatory statistical quality standards for use of their data for other than experimental purposes? These quality standards are a prerequisite to consider using big data sources for any policy-related toolkit and will be addressed in Chapter 6. In this context, and for the purposes of this paper, the term “experimental source” is used to denote a source which does not comply with the necessary statistical quality standards, but would, however, enable statisticians, economists and researchers to experiment and analyse datasets and to test methodologies, algorithms and software tools in order to obtain insights from the patterns and behaviours revealed by internet searches. Experimenting with new data sources has its own merits in testing new insights and for research purposes, but for the data source to be regularly used in decision-making, it must comply with these quality requirements.

Internet penetration rates are continually increasing and these days households are making more frequent use of search engines. However, it is worth noting that experimentation with internet search data has been going on for almost a decade. Hal Varian, Google’s Chief Economist, has been forward-looking and transparent in sharing programming codes to explore Google search terms and their ability to nowcast economic activities, mainly in relation to the US economy (Choi and Varian (2009, 2012)). He and his co-author conclude that Google searches may be helpful in providing directional guidance in advance of the publication of official statistics.

The present paper contributes to the relevant literature in five different ways. First, we focus on establishing a leading economic indicator for the euro area by creating an index of euro area new vehicle sales based on the corresponding national internet

[1] The absolute numbers of internet searches are not available in the dataset. These absolute numbers have been normalised and indexed. See Chapter 3 for a description of the dataset and methodology.

search data for vehicle sales. Second, we assess the impact of volume changes in internet search terms to determine their potential effect on future vehicle sales. Third, we examine (i) the relationship and (ii) the length of time between an internet vehicle search carried out today and its appearance in vehicle sales in the future. Fourth, we test the nowcasting ability of using the internet vehicle search term in predicting future euro area vehicle sales. Fifth, we develop a big data analytics quality concept for going beyond experimenting with big data sets.

In Chapter 2 we review the findings of the increasing literature related to the application of Google search data in the field of economics. Chapter 3 provides a description of the datasets, their transformations and the method applied in creating the euro area vehicle sales indicator. In Chapter 4, we demonstrate the short and long-term dynamics underpinning the relationship between the internet car search data and euro area car sales figures. Chapter 5 presents the nowcasting ability of using internet search data in predicting car sales. Finally, Chapter 6 presents the big data analytics quality concepts, and Chapter 7 concludes.

2 Insights from the literature

Google search data have been a rich source and playground for statisticians, economists and researchers over the past decade, ever since the paper by Choi and Varian in 2009 (entitled "Predicting the Present with Google Trends") encouraged users to experiment with Google data. The basic assumption underlying the use of Google data in the field of economics stems from the perception that increases and decreases in the volumes of relevant internet searches may be correlated with economic activity, and could therefore be used to predict the levels of future economic and financial statistical data releases with greater accuracy, at higher frequency and in a more timely fashion than conventional methods.

In their paper, Choi and Varian use an autoregressive model both with, and without, Google Trends data as an explanatory variable, and find that the model that uses Google Trends data generally outperforms the model that does not. This method is then applied to the nowcasting of various economic indicators, for example house sales and tourism, with similar results. The paper also provides the relevant computation codes, thereby facilitating both their use on a transparent basis and the replication of the experiment.

Nevertheless, Google search data have been applied in research endeavours in a variety of ways which can be categorised into four groups.

1. testing the search data for **the ability to predict macroeconomic and financial activity**;
2. replicating similar experiments for **other geographical areas**;
3. conducting similar experiments applying **various linear and non-linear models**;
4. creating **new indicators**.

For example, Askitas and Zimmermann (2009) aim at forecasting the unemployment rate in Germany using Google Trends data and monthly unemployment rates published by the German Federal Employment Agency. The results of the research may have been affected by differences between social structures in the United States and Germany. Unemployment rates tend to be a lagging macroeconomic indicator and therefore may have less of a forward-looking nature.

D'Amuri and Marcucci (2010, 2013) use Google Trends data and seasonally adjusted monthly unemployment rate data for the United States (including the initial claims indicator) published by the Bureau of Labor Statistics and suggest that an indicator based on Google data is best for predicting the US monthly unemployment rate.

Koop and Onorante (2013) present a model that uses Google Trends search data innovatively in forecasting macroeconomic variables. Rather than using the search volume as a variable, the authors add to the literature by nowcasting using dynamic model averaging methods which allow for model switching between time-varying parameter regression models. They allow the model switching to be controlled by the Google search intensity through “Google probabilities”, which determine which nowcasting model should be used at each point in time. In an exercise involving nine major US monthly macroeconomic variables, this approach provides improvements in nowcasting performance.

Guzmán (2011) tests internet search behaviour as an economic forecasting tool for inflation expectations. Her paper is based on the assumption that if volumes of search queries for “inflation” increase, consumers are feeling increasingly concerned about the possibility of rising prices and may be anticipating an increase in inflation. The paper compares higher-frequency measures of inflation expectations with lower-frequency surveys, such as the quarterly Michigan Survey, the quarterly Survey of Professional Forecasters and the semi-annual Livingston Survey. It finds that the higher-frequency measures tend to outperform the lower-frequency surveys and indicates that the Google inflation search index performs best.

In the field of finance, Vlastakis and Markellos (2010) have found, using Google Trends data, that demand for information in relation to the individual stocks that are most traded on the New York Stock Exchange has a significant impact on the trading volumes of those individual stocks. Similarly, Preis, Moat and Stanley (2013) investigate trading behaviour in financial markets (reflected by the Dow Jones Industrial Average) using Google search data. By analysing changes in Google query volumes for search terms related to financial markets, they trace patterns which can be interpreted as “early warning signs” of stock market changes. The same patterns appear to exist in the French stock market, as indicated in the paper of Arouri, Aouadi, Foulquier and Teulon (2013). They use Google Trends data to form multifactor models to identify the determinants of liquidity in the French stock market. Their paper finds that adding search volume to a model of turnover in the French stock market improves out-of-sample forecast performance and that internet research volumes tend to be positively related to market liquidity.

Chamberlin (2010), Choi and Varian (2012) and Du and Kamakura (2012) provide econometric evidence on how Google search data can be related to car sales. They find similar results as Google search data appear to precede changes in official car sales data.

Carrière-Swallow and Labbé (2011) construct a Google Automotive Index and find that, despite the low level of internet usage among the population in Chile, models incorporating their index outperform a benchmark model which does not – in both in and out-of-sample testing.

Barreira, Godinho and Melo (2013) attempt to identify cointegrating relationships between official car sales data for Spain, France, Italy and Portugal, and relevant Google Trends data. They also carry out simple forecasting exercises, but conclude that there is little evidence that search query data can be used to produce improved predictions for the countries under consideration.

Fantazzini and Toktamysova (2015) forecast car sales in Germany using 24 different multivariate models with and without Google variables and other economic variables.

The authors use the monthly sales of ten car brands in Germany for the period from 2001 to 2012, with a forecast horizon of one month to two years. The authors conclude that no single model outperforms the others, although the performance of the Google-based models seems to improve where the data relate to a period of recession and the Google data may explain part of the non-linearity exhibited in car sales data.

Tomczyk and Doligalski (2015) conduct a similar exercise for the Polish market and conclude that, while Google data for small/less popular car brands do not improve the forecasts of official car sales, using Google data for the five major car brands in Poland improves the predictive power of the model, at least in the short term.

Figueiredo (2016) constructs a vehicle search index for different vehicle brands across various provinces and territories in Canada and concludes that these indices contain additional predictive information for vehicle sales, at least in the more populous provinces and territories.

Geva, Oestreicher-Singer, Efron and Shimshoni (2017) study the importance of Google's comprehensive index of internet discussion forums in addition to the Google data themselves. They find that taking such additional sources into account improves the predictive accuracy of the competing models, particularly for inexpensive car brands – and to a lesser extent for "premium" brands.

The literature review seems to confirm the predictive capabilities of using internet search data across several geographical areas.

We will apply a similar methodology to Choi and Varian (2009, 2012) of using several autoregressive models with and without Google data, since the original dataset shows significant seasonality. While most of the literature relating to predicting car sales in European countries, Canada, the United States and Chile seems to indicate that using Google data may increase the predictive capabilities, Barreira, Godinho and Melo (2013) indicate that Google data provide little evidence for Spain, France, Italy and Portugal. We will then examine how Google data perform at the euro area level.

In Chapter 3 we will therefore look more closely at the data sources as part of the preparation for the testing exercise. That chapter provides a description of the datasets, their transformations and the method applied to create the relevant euro area indicators.

3 Description of the datasets

In this chapter we will explore two data sources; official car registrations data and Google search data.

Official data on new passenger car registrations are produced monthly by the European Automobile Manufacturers Association (ACEA).[2] These data are released in the form of a "press release" with monthly new passenger car registrations broken down by EU country. The release of data for a particular month generally occurs approximately 15 calendar days after the end of that month. Releases also include revisions and updates to registrations in previous months. We use the series related to "registrations of vehicles for the carriage of passengers, comprising no more than eight seats in addition to the driver's seat". These datasets are raw data and are neither working-day nor seasonally adjusted. We use the data in logarithmic form rather than absolute values.

The European Central Bank and Eurostat[3] collect these datasets from the ACEA and then release them in the form of growth rates for the euro area, the EU, and with the associated national breakdowns – in both seasonally adjusted and working-day-adjusted series – within five working days.[4] Thus, the new passenger car registration statistics for the month of December 2017 were released on 17 January 2018. The data for the two summer months of July and August are released before the third week of September. From a methodological perspective, the category "passenger cars" includes passenger cars registered by either private households or businesses.

[2] See example.

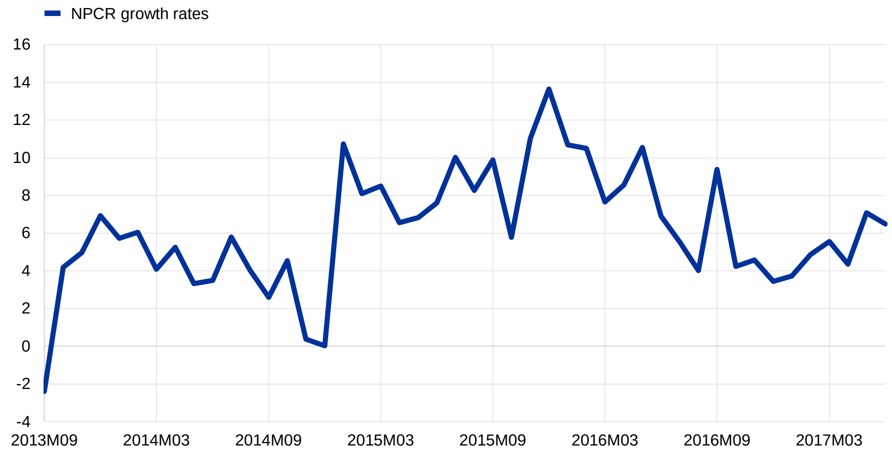
[3] Eurostat and the ECB's Directorate General Statistics are the two European statistics agencies responsible for collecting and releasing EU statistics together with national statistical offices and national central banks, respectively.

[4] The series used can be obtained from the ECB's statistics dissemination platform – the Statistical Data Warehouse (SDW) – using the following reference code: STS.M.I8.W.CREG.PC0000.3.ANR.

Chart 1

Growth rates of new euro area car registrations

(year-on-year percentage changes, working-day adjusted)



Source: Eurostat and ECB

Chart 1 illustrates the year-on-year growth rates of new car registrations in the euro area over the past four years. Changes in the year-on-year observations are a meaningful transformation, particularly when the series displays specific seasonal patterns. While the average growth rate is approximately 6.2% over the period, the series demonstrates volatile annual growth rates throughout. It can be seen that at several points the growth rates reach above 10%, while in 2014 they are much lower than average. This below-average growth in 2014 was anticipated, given that the impact of the earlier financial crisis, in particular on the market for new vehicles, had not yet disappeared. The growth rate of zero for December 2014 means that the absolute number of car registrations in December of that year was the same as in the same month one year earlier (December 2013).

It can also be seen that this average euro area growth rate of 6.2% (for new passenger car registrations) is significantly higher than the average GDP growth rate of 2.8% for the euro area economy (based on market prices) over the same period. In other words, new vehicle sales have been growing at least twice as fast as the economy as a whole. Therefore, “new car registrations” is certainly a variable that should be taken into consideration in order to understand and analyse economic activity and the composition of household spending.

The second dataset provides information on terms searched on the internet using the Google search engine. We obtain Google search data structured by using a taxonomy of 26 different categories[5] and 297 sub-categories for the ten euro area countries shown below:

[5] See Appendix E.

Belgium	Ireland	France	Netherlands	Portugal
Germany	Spain	Italy	Austria	Slovenia

These ten countries account for approximately 96% of euro area new passenger car registrations, according to the ACEA's statistics. We therefore assume that the "euro area" Google search data population from the ten countries represents the euro area as a whole for the purposes of this exercise. While the individual search terms used within each of the ten countries would be classified as "big data" with significant volumes, the individual search data have been standardised and classified into an easy-to-manage weekly dataset. This facilitates the processing and comparability of the dataset and, more importantly, reduces misclassification errors.

We use the standardised Google search category labelled "Autos & Vehicles" as a proxy for new car registrations and receive weekly data from Google on the internet search traffic census relevant to this category (see appendix E). This dataset may deviate slightly from one that uses search terms available to the public via Google Trends[6] data, which is sample-based. The data are obtained weekly (on Tuesdays) and refer to the week two weeks previously. They have been normalised and indexed within each of the respective categories of the taxonomy and by country.

The Google data are indexed with a starting value of 1.00 set for the last week of 2003. The first week's data in 2004 then provided a new value which indicates the deviation of search volumes relative to the start of the index. As a more recent example, the value for the "Autos & Vehicles" category for Austria was 1.47 for the week commencing 2 July 2017. The following week the value was 1.49. The index had therefore increased by two basis points and the absolute search volume had increased by 1.36% in one week (2 July to 9 July). Using the values, we can also see that the search volume for this category had increased 49% since the start of 2004. Since we do not know either the volumes at the starting point or the weekly volumes (i.e. the absolute volumes), the table below provides an example of the how the normalisation calculation is made in respect of three different countries for a four-week period.

[6] See google/trends.

Table 1

Example of Google search data normalisation

Category	Country(i)	Volume(t)	Volume(t+2)	Volume(t+3)	Volume(t+4)
Auto	1,AV	200	250	300	200
	Index(1)	1.00	1.25	1.50	1.00
	% change		25%	20%	-33.3%
	2,AV	400	450	500	400
	Index(2)	1.00	1.125	1.25	1.00
	% change		12.5%	11.1%	-20%
	3,AV	400	500	600	400
	Index(3)	1.00	1.25	1.50	1.00
	% change		25%	20%	-33.3%

Notes: AV = Absolute volumes; the index number in bold is the only known factor.

Although the absolute volume in relation to each of country 1 and country 2 increases by the same amount, e.g. 50 searches in one week, the impact on the index value differs. Similarly, if the index value for each of country 1 and country 3 is the same, e.g. t+1, the absolute increase in volume since time t for each will also differ.

Since the normalisation process is carried out on the basis of the search volumes for each category by country, the weekly volume changes in Google search values for one country are therefore not directly comparable with those for other countries.

In order to create the euro area Google search dataset, we therefore need to weight the national Google datasets using a factor which provides, on the one hand, an indication of the share of the national population with access to the internet (a precondition for internet searches) and, on the other hand, an indication of the size of the national population. Using the official European statistics published by Eurostat, we take the level of households' internet access for a particular country and multiply this by the national population for a calendar year to calculate the relevant national share of the euro area totals per year. As a particular country's national share does not deviate significantly from year to year, we calculate one average weighting factor on the basis of the period 2013-2017 (see Appendix A).

Using these weighting factors and the national Google datasets, the euro area Google search data are calculated as the weighted arithmetic mean of the national Google data for each of the ten countries, as follows:

$$g_t = \frac{1}{n} * \sum_{i=0}^{n} w_i * g_{i,t} \ ,$$

where (gt) is euro area Google search data at time t and (wi) is the relevant country's i weight.

We now have euro area Google search data at a weekly frequency which need to be transformed into a monthly series so as to have the same frequency as the relevant time series. This is done using the same methodology, by applying the arithmetic mean used to calculate the monthly euro area Google dataset. The final result of the data preparation phase is the generation of two symmetric monthly time series, as represented visually in Chart 2.

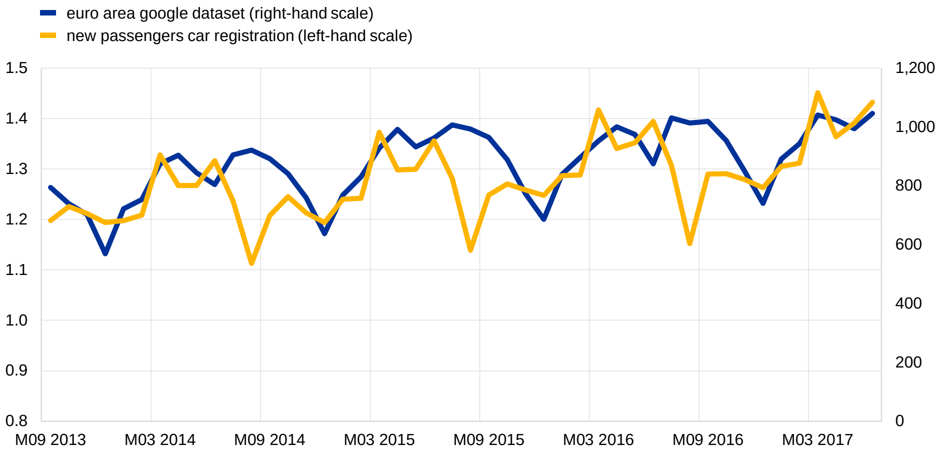
4 Methodology and models

In Chapter 3, we described how the two euro area monthly series were generated using car registration data and Google data. Chart 2 below plots the two datasets.

Chart 2

Visual comparison of new car registrations and the euro area Google dataset for vehicles

(left-hand scale = index; right-hand scale = thousands)



Source: Authors' calculations.

On visual inspection, Chart 2 could show (i) an initial indication of possible co-movements between Google car searches (red line) and the aggregated euro area official car registrations (blue line) (R=0.79), and (ii) an indication that changes in Google searches appear to precede changes in car registrations (although the latter are more volatile). This could be anticipated, as households are likely to search for car-related information prior to the actual purchase of a new vehicle.

In the next sections we will test whether these indications from the visual inspection can also be verified by econometric evidence.

For this econometric testing we apply the augmented Dickey-Fuller test (Dickey and Fuller (1979)) to ascertain whether the series are stationary (i.e. the joint probability distribution is constant over time). Non-stationary series can "distort" regression models and their inferences.

Table 2

Results of augmented Dickey-Fuller test

Variable	P-value (5%)	Order of Integration
Car Registrations	0.0095	I (0)
Google Data	0.0004	I (1)

The test results in Table 2 indicate, first, that euro area car registrations are stationary in level and, second, that euro area Google data are stationary only after first-differencing. Therefore, we can assume that the statistical properties are constant over time, the autoregressive coefficients are not biased and the t-statistics follow a normal distribution.

Box 1
Stationary time series

Broadly speaking, a time series is said to be "stationary" if there is no systematic change in the mean and no systematic change in the variance. In other words, the statistical characteristics of a time series are broadly similar for each period of the time series. This is a useful assumption (although, strictly speaking, stationary time series do not exist), as it suggests that a stationary model can be applied. This is why non-stationary time series are transformed into stationary time series, as done above for the Google dataset in Table 2.

A vector autoregression (VAR) system and an autoregressive distributed lag (ARDL) model can be applied to test the short-term and long-term dynamics and statistical relationships between the Google dataset and the car registration dataset. The test results will guide us in deciding on the number of monthly lags in Google search data to include in the nowcasting model (see Chapter 5).

A VAR system is employed in order to provide evidence for the assumption that Google data can cause and consequently predict future movements in car sales volumes. In such a context, a VAR analysis can reveal:

1. whether a causal relationship exists between the two datasets;
2. what the exact direction of this relationship is (i.e. which dataset affects the values of the other);
3. whether the Google dataset could be an accurate explanatory dataset for nowcasting car sales.

The ARDL model is used to gain further insights into the exact causal relationship. The two models are explained in the next two sections.

Box 2
Vector autoregression (VAR) and autoregressive distributed lag (ARDL)

The vector autoregression (VAR) is a multivariate time series model, based on Gaussian errors (normally distributed and therefore the error terms are uncorrelated) and is frequently used as a description of macroeconomic time series data. It is flexible, easy to estimate, and usually gives a good fit for macroeconomic data. The main advantage is, however, its ability to combine short-term and long-term information (structures and components). The process simultaneously considers several endogenous variables, each of which is explained by its past values. Usually, there are no exogenous variables in the model. The process is useful in combination with, for instance, the

Granger causality test. The Granger causality test assumes that the information relevant to the prediction of the respective variables is contained solely in the time series data. As we are examining two variables (car registrations and Google data), it relates to bilateral causality. This can be extended to multivariate causality through the technique of the VAR. It is reasonable to assume that the effect of a unit change in Google searches on car registrations is distributed over a period of time – and therefore not instantaneous – and the model then becomes an autoregressive distributed lag (ARDL) model. In our particular case, we use the Akaike information criterion (AIC) to select the time distribution effect of Google searches on car registration. The reader wishing to pursue the subject further is advised to consult econometrics handbooks[7].

4.1 Vector autoregression (VAR) system

First, we apply the Akaike information criterion (AIC) to determine the lag length of the VAR system. From the results presented in Appendix B, we find that five lags is the optimal lag length to be included in the VAR system. Applying this criterion, we can then denote the functional form of the VAR system, as follows:

VAR(5) system:

$$\Delta(Gt) = \alpha 1 + \beta 1,1\, \Delta(Gt-1) + \beta 1,2,\, \Delta(Gt-2) + \beta 1,3\, \Delta(Gt-3) + \beta 1,4\, \Delta(Gt-4) + \beta 1,5\, \Delta(Gt-5) + \beta 1,6\, kt + \beta 1,7\, kt-1 + \beta 1,8\, kt-2 + \beta 1,9\, kt-3 + \beta 1,10\, kt-4 + \beta 1,11\, kt-5 + \varepsilon t$$

$$kt = \alpha 2 + \beta 2,1\, \Delta(Gt) + \beta 2,2\, \Delta(Gt-1) + \beta 2,3\, \Delta(Gt-2) + \beta 2,4\, \Delta(Gt-3) + \beta 2,5\, \Delta(Gt-4) + \beta 2,6\, \Delta(Gt-5) + \beta 2,7\, kt-1 + \beta 2,8\, kt-2 + \beta 2,9\, kt-3 + \beta 2,10\, kt-4 + \beta 2,11\, kt-5 + \varepsilon t; \qquad (1)$$

where (Gt) = Google search index at time (t), k(t) = new car registrations at time (t)

and (ε) = error term.

Testing the VAR system against the t-statistics confirms the significance of both the Google search data and car sales variables with five lags (see Appendix B).

We use the Granger causality test (Granger (1969)) to provide evidence of causality between the two variables; i.e. whether the Google searches variable has explanatory power in the movements of car sales, and vice versa.

[7] For instance: Greene, W. (2017). Econometric analysis (8th ed.). Upper Saddle River: Prentice Hall Juselius, K. (2006). Or The cointegrated VAR model: Methodology and applications. Oxford: Oxford University Press or Gujarati, Damodar, & Porter, Dawn C. (2009). Basic econometrics (5th ed., McGraw-Hill Series economics). New York: McGraw-Hill.

Table 3

Granger causality (Ho ≠ dependent)

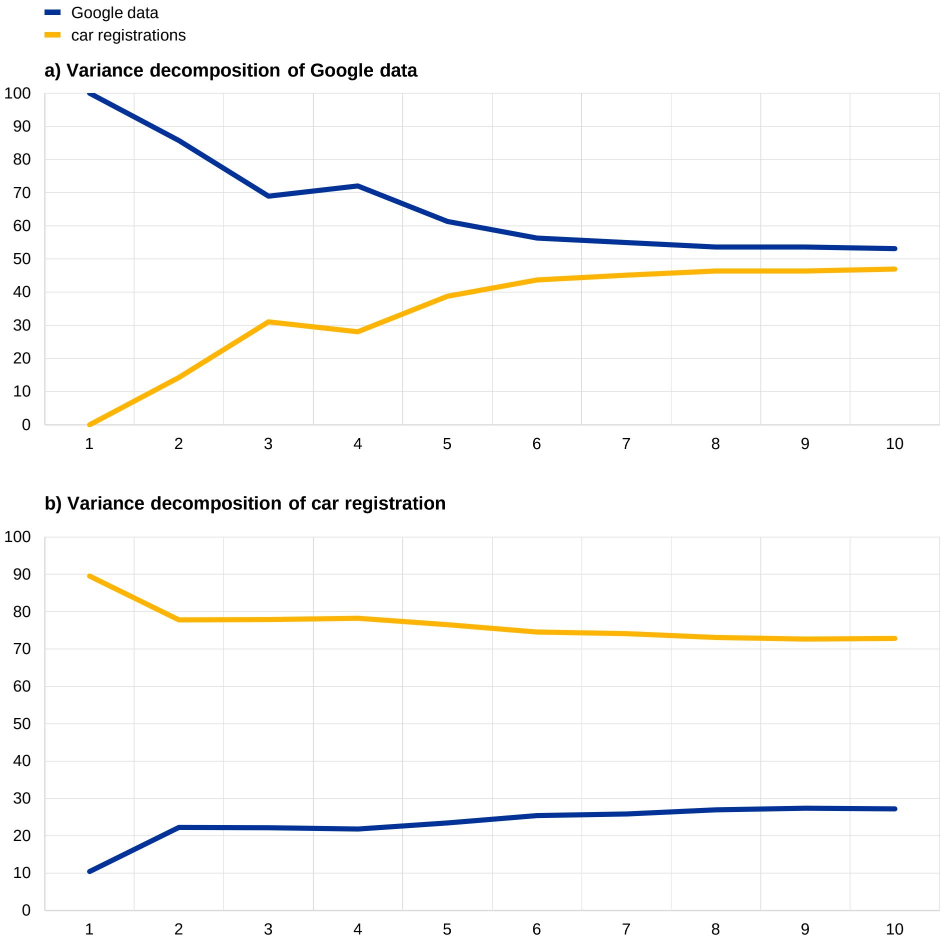
Dependent variable							
Google search data				Car sales			
Response variable	Chi-square	df	Prob.	Response variable	Chi-square	df	Prob.
Car sales	83.33553	5	0.00	Google	37.90325	5	0.00

Note: Df = degree of freedom; Prob. = Probability at the level of 99.5%.

The Chi-square statistics indicate that there is a two-way causal relationship between euro area car registrations and euro area Google search data, i.e. each of the dependent variables helps in predicting the future values of the other variable in the short run. We now check the variance decompositions to obtain further insights into the causality and the direction. For this we use Cholesky ordering (g, k) to compute the variance decompositions taking ten subsequent periods into account.

Chart 3

Ten period variance decomposition of Google data and car registrations



a) Variance decomposition of Google data

b) Variance decomposition of car registration

Source: Authors’ calculations.

The upper panel in Chart 3 shows the variance decomposition of the Google search data. In other words, it shows the explanatory contribution of car registration data to the variance of Google search data over a period of ten months.

One can see from the upper panel that car registration data today and up to the sixth period contribute up to 40% of the variance in the Google search data. In the subsequent period this contribution fades out, reaching its asymptote at around 45%. This means that adding additional periods would contribute only marginally to the variance. In the immediate short term and up to the third period, the car registration data contribute up to 30% of the variance in the Google search data. If we then compare this with the lower panel, we note that the Google search data contribute up to approximately 22% of the variance in the car sales data from as early as the second period onwards. This is important to note, as for nowcasting purposes we are interested in short-term variance.

Let us now turn to the second exercise and apply the ARDL model in order to focus on the long-term variance.

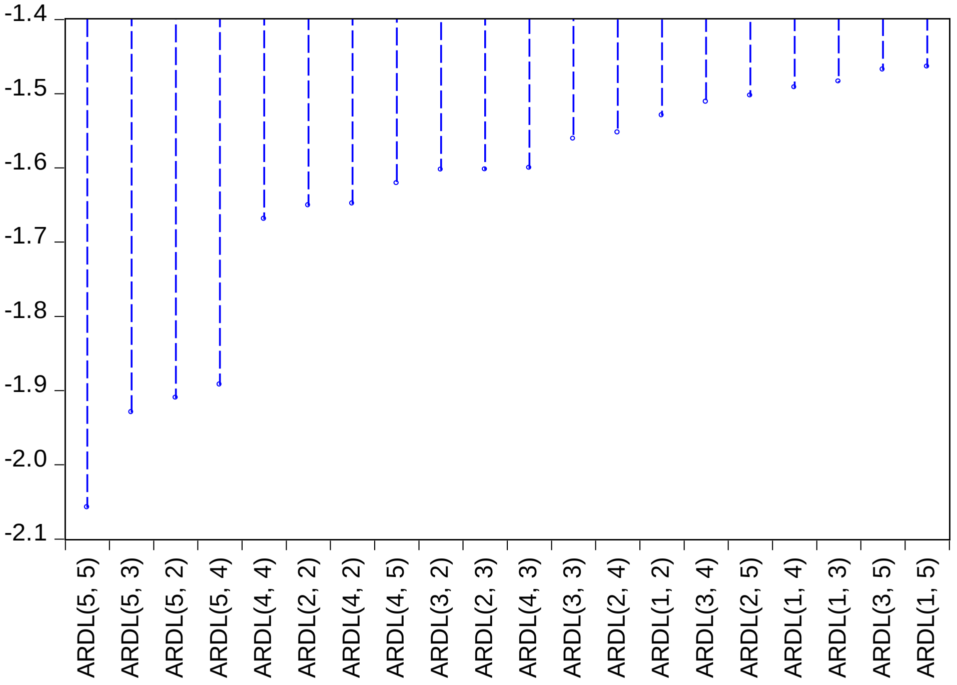
4.2 Autoregressive distributed lag (ARDL) estimation and analysis

By fitting the two variables into an ARDL model, we intend to examine the long-run relationship between Google search data and car registrations. The bounds test, the co-integrating equation and the long-run coefficients will give insights into this long-run relationship, the speed of adjustment at an equilibrium level and whether Google search data provide information for future volumes and movements in car registrations.

The Akaike information criterion is again used to determine the lag length. We find that the ARDL(5,5) model performs best among the 30 options evaluated.

Chart 4

Ranking the top 20 best-performing models



Source: Authors' calculations.

Against this background, we have allowed for a maximum of five lags to be included in the ARDL model. We can then denote the functional form of the ARDL model as follows:

$$kt = \alpha + \beta 1\,\Delta(Gt) + \beta 2\,\Delta(Gt-1) + \beta 3\,\Delta(Gt-2) + \beta 4\,\Delta(Gt-3) + \beta 5\,\Delta(Gt-4) + \beta 6\,\Delta(Gt-5) + \beta 7\,kt-1 + \beta 8\,kt-2 + \beta 9\,kt-3 + \beta 10\,kt-4 + \beta 11\,kt-5 + \delta T + \varepsilon t \tag{2}$$

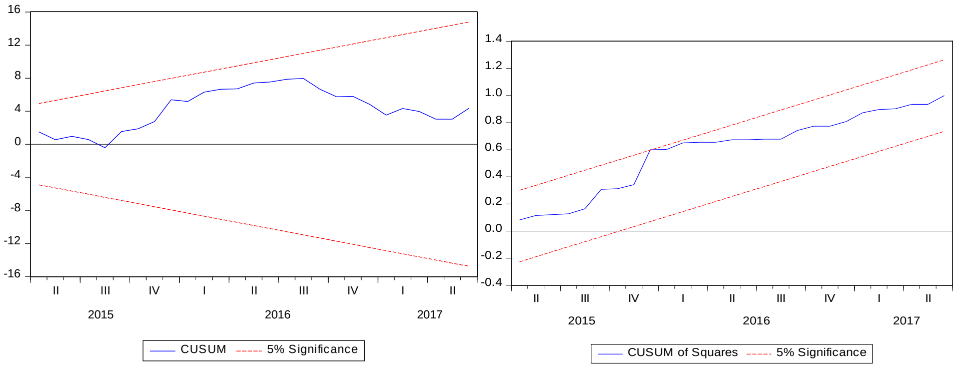
where (kt) = new car registrations at time (t), (Gt) = Google search index at time (t), (T) = linear trend and (ε) = error term.

A linear trend has been added to the equation in order to remove the problem of serial correlation. See Appendix B for the substituted coefficients, t-statistics, standard errors and the serial correlation test.

We now use the cumulative sum (CUSUM/CUSUM of squares) tests (Page (1954)) for breakpoints in the sample.

Chart 5

Testing for breakpoints within the sample



Source: Authors' calculations.

In the chart above we can see that the sample lies within the 5% significance level and thus does not indicate any breakpoints. This, together with the previous test of serial correlation, confirms the accuracy of the results.

We now proceed with the coefficient diagnostics, first applying the bounds test (Pesaran, Shin and Smith (2001)). The test either accepts or rejects the null hypothesis that no long-run relationship exists between the two variables.

Table 4

Results of ARDL bounds test

$H_{(o)}$: No long-run relationships exist		
Test Statistic	Value	k
F-statistic	13.51661	1
Critical Value Bounds		
Significance	I0 Bound	I1 Bound
10%	4.05	4.49
5%	4.68	5.15
2.5%	5.3	5.83
1%	6.1	6.73

The F-statistic value is significantly larger than the critical value bounds at any significance level, indicating a rejection of the hypothesis of "no long-run relationships between the two variables".

The test results in this chapter indicate that both short-term and long-term dynamics exist when examining the statistical relationships between the two series. The result for the ARDL(5,5) model indicates that the Google search data contain information on future car sales. Nevertheless, the two variables seem to converge to the same level in the long run. Therefore, it seems reasonable to conclude that the Google search data could be used as an explanatory variable and could act as an early indicator of new car sales, although from a broader (long-run) perspective the two variables seem identical. This is understandable as they both attempt to measure the same concept. In the next chapter we will test the ability of Google search data to nowcast car sales.

5 Using Google search data to nowcast car sales

In this chapter we will use the Google search data (as an explanatory variable) to test their predictive capacities for nowcasting euro area car sales (the response variable). Using the results from Chapter 4, it seems reasonable to construct a model which includes at least five months of Google search data and the linear trend; the latter to remove serial correlation, if applicable. This is a lag assumption which indicates that there is likely to be a period of up to, and including, five months between the first internet searches for cars by a household and an actual purchase by it of a new car. This seems to be a reasonable assumption, subject to the standard caveats such as variation in people's preferences, culture, ratio between disposable income and savings, wealth, car price, etc. As the Google search data display at least annual seasonality, we further include the twelve month lag.

We start with a baseline forecasting model without the Google search data and use this as the benchmark model for comparing the performance of the model(s) that include the Google search data.

The baseline model aims at predicting car sales using car sales data from the previous month and 12 months ago, as we know that the series displays seasonality and that the month-to-month car sales data are volatile. This model is known in the literature as a seasonal autoregressive model and it is also applied by Choi and Varian (2009, 2012).

The functional form of the model is:

Baseline model (benchmark model)

$$k_t = \alpha + \beta_1 \log k_{t-1} + \beta_2 \log k_{t-12} + \varepsilon_t \quad (3)$$

where (kt) = new car registrations at time (t) and (ε) = error term.

To make the comparisons with this baseline model, we first include the euro area Google dataset and then use four different macroeconomic variables, as explanatory variables, to test, compare and identify the model that performs best with and without the euro area Google data, as compared with the baseline model. First, we use the relevant euro area "Harmonised Index of Consumer Prices" for the category "Motor cars".[8] This index measures the inflation rate of cars. Second, we use the euro area "Industrial Confidence Indicator", which is a leading survey indicator for monitoring industry sentiment.[9] This indicator is calculated as the arithmetic average of the balances of responses on production expectations, the assessment of books and

[8] Series code: "ICP.M.U2.N.071100.4.INX" taken from the ECB's Statistical Data Warehouse.

[9] Series code: "RTD.M.S0.S.Y_ISICI.F" taken from the ECB's Statistical Data Warehouse.

stocks of finished products. Third, we use the euro area "Disposable Income of Households" indicator.[10] This measures the income distribution and the income of euro area households available to be spent, for example, on consumer goods such as cars. Fourth we use the euro area "Gross Saving of Households" indicator.[11] We convert the quarterly data into a monthly series using the linear interpolation method.

The three macroeconomic indicators are stationary after first differencing according to the augmented Dickey-Fuller test, which also demonstrates that the series "Harmonised Index of Consumer Prices" should be used with one lag, the series "Industrial Confidence Indicator" with no lags and the series "Disposable Income of Households" with five lags. We use logarithmic values for the series "Disposable Income of Households".

The functional form of the respective models can therefore be written as:

Baseline model with Google Data (Model 1)

$$k_t = \alpha + \beta_1 \log k_{t-1} + \beta_2 \log k_{t-12} + \beta_3 \Delta(G_t) + \beta_4 \Delta(G_{t-1}) + \beta_5 \Delta(G_{t-2}) + \beta_6 \Delta(G_{t-3}) + \beta_7 \Delta(G_{t-4}) + \beta_8 \Delta(G_{t-5}) + \beta_9 \Delta(G_{t-12}) + \varepsilon_t \quad (4)$$

where $\mathrm{k_t}$ *is new car registrations at time (t),* $\mathrm{G_t}$ *is the Google search index at time (t), T is the linear trend and ε is the error term.*

Model with inflation rate for cars (log) including one lag (Model 2)

$$k_t = \alpha + \beta_1 \log k_{t-1} + \beta_2 \log k_{t-12} + \beta_3 \Delta(\log \pi_t) + \beta_4 \Delta(\log \pi_{t-1}) + \varepsilon_t \quad (5)$$

where $\mathrm{k_t}$ *is new car registrations at time (t),* π_t *is the Harmonised Index of Consumer Prices – Motor cars at time (t) and ε is the error term.*

Model with inflation rate for cars (log) including one lag and Google Data (Model 3)

$$k_t = \alpha + \beta_1 \log k_{t-1} + \beta_2 \log k_{t-12} + \beta_3 \Delta(\log \pi_t) + \beta_4 \Delta(\log \pi_{t-1}) + \beta_5 \Delta(G_t) + \beta_6 \Delta(G_{t-1}) + \beta_7 \Delta(G_{t-2}) + \beta_8 \Delta(G_{t-3}) + \beta_9 \Delta(G_{t-4}) + \beta_{10} \Delta(G_{t-5}) + \beta_{11} \Delta(G_{t-12}) + \varepsilon_t \quad (6)$$

where $\mathrm{k_t}$ *is new car registrations at time (t),* $\mathrm{G_t}$ *is the Google search index at time (t),* π_t *is the Harmonised Index of Consumer Prices – Motor cars at time (t) and ε is the error term.*

[10] Series code: "QSA.Q.N.I8.W0.S1M.S1._Z.B.B6G._Z._Z._Z.XDC._T.S.V.N._T" taken from the ECB's Statistical Data Warehouse.

[11] Series code: "QSA.Q.N.I8.W0.S1M.S1._Z.B.B8G._Z._Z._Z.XDC._T.S.V.N._T" taken from the ECB's Statistical Data Warehouse.

Model with the confidence indicator (Model 4)

$$k_t = \alpha + \beta_1 \log k_{t-1} + \beta_2 \log k_{t-12} + \beta_3 \Delta(i_t) + \varepsilon_t \quad (7)$$

where $\mathrm{k_t}$ *is new car registrations at time (t),* $\mathrm{i_t}$ *is the industrial confidence indicator at time (t) and ε is the error term.*

Model with the confidence indicator and Google Data (Model 5)

$$k_t = \alpha + \beta_1 \log k_{t-1} + \beta_2 \log k_{t-12} + \beta_3 \Delta(i_t) + \beta_4 \Delta(G_t) + \beta_5 \Delta(G_{t-1}) + \beta_6 \Delta(G_{t-2}) + \beta_7 \Delta(G_{t-3}) + \beta_8 \Delta(G_{t-4}) + \beta_9 \Delta(G_{t-5}) + \beta_{10} \Delta(G_{t-12}) + \varepsilon_t$$

$$k_t = \alpha + \beta_1 \log k_{t-1} + \beta_2 \log k_{t-12} + \beta_3 \Delta(i_t) + \beta_4 \Delta(G_t) + \beta_5 \Delta(G_{t-1}) + \beta_6 \Delta(G_{t-2}) + \beta_7 \Delta(G_{t-3}) + \beta_8 \Delta(G_{t-4}) + \beta_9 \Delta(G_{t-5}) + \beta_{10} \Delta(G_{t-12}) + \varepsilon_t \quad (8)$$

where $\mathrm{k_t}$ *is new car registrations at time (t),* $\mathrm{G_t}$ *is the Google search index at time (t),* $\mathrm{i_t}$ *is the industrial confidence indicator at time (t) and ε is the error term.*

Model with household income including five lag (Model 6)

$$k_t = \alpha + \beta_1 \log k_{t-1} + \beta_2 \log k_{t-12} + \beta_3 \Delta(\log \sigma_t) + \beta_4 \Delta(\log \sigma_{t-1}) + \beta_5 \Delta(\log \sigma_{t-2}) + \beta_6 \Delta(\log \sigma_{t-3}) + \beta_7 \Delta(\log \sigma_{t-4}) + \beta_8 \Delta(\log \sigma_{t-5}) + \varepsilon_t \quad (9)$$

where $\mathrm{k_t}$ *is new car registrations at time (t),* $\mathrm{\sigma_t}$ *is households' disposable income at time (t) and ε is the error term.*

Model with household income including five lag and Google Data (Model 7)

$$k_t = \alpha + \beta_1 \log k_{t-1} + \beta_2 \log k_{t-12} + \beta_3 \Delta(\log \sigma_t) + \beta_4 \Delta(\log \sigma_{t-1}) + \beta_5 \Delta(\log \sigma_{t-2}) + \beta_6 \Delta(\log \sigma_{t-3}) + \beta_7 \Delta(\log \sigma_{t-4}) + \beta_8 \Delta(\log \sigma_{t-5}) + \beta_9 \Delta(G_t) + \beta_{10} \Delta(G_{t-1}) + \beta_{11} \Delta(G_{t-2}) + \beta_{12} \Delta(G_{t-3}) + \beta_{13} \Delta(G_{t-4}) + \beta_{14} \Delta(G_{t-5}) + \beta_{15} \Delta(G_{t-12}) + \varepsilon_t \quad (10)$$

where $\mathrm{k_t}$ *is new car registrations at time (t),* $\mathrm{G_t}$ *is the Google search index at time (t),* $\mathrm{\sigma_t}$ *is households' disposable income at time (t) and ε is the error term.*

Model with household savings including three lag (Model 8)

$$k_t = \alpha + \beta_1 \log k_{t-1} + \beta_2 \log k_{t-12} + \beta_3 \Delta(\log h_t) + \beta_4 \Delta(\log h_{t-1}) + \beta_5 \Delta(\log h_{t-2}) + \beta_6 \Delta(\log h_{t-3}) + \varepsilon_t \quad (11)$$

where $\mathrm{k_t}$ *is new car registrations at time (t),* $\mathrm{G_t}$ *is the Google search index at time (t),* $\mathrm{h_t}$ *is households' savings at time (t) and ε is the error term.*

Model with household savings including three lag and Google Data (Model 9)

$$k_t = \alpha + \beta_1 \log k_{t-1} + \beta_2 \log k_{t-12} + \beta_3 \Delta(\log h_t) + \beta_4 \Delta(\log h_{t-1}) + \beta_5 \Delta(\log h_{t-2}) + \beta_6 \Delta(\log h_{t-3}) + \beta_7 \Delta(G_t) + \beta_8 \Delta(G_{t-1}) + \beta_9 \Delta(G_{t-2}) + \beta_{10} \Delta(G_{t-3}) + \beta_{11} \Delta(G_{t-4}) + \beta_{12} \Delta(G_{t-5}) + \beta_{13} \Delta(G_{t-12}) + \varepsilon_t \quad (12)$$

where $\mathrm{k_t}$ *is new car registrations at time (t),* $\mathrm{G_t}$ *is the Google search index at time (t),* $\mathrm{h_t}$ *is households' saving at time (t) and ε is the error term.*

Model with all macroeconomic indicators (Model 10)

$$k_t = \alpha + \beta_1 \log k_{t-1} + \beta_2 \log k_{t-12} + \beta_3 \Delta(\log \pi_t) + \beta_4 \Delta(\log \pi_{t-1}) + \beta_5 \Delta(i_t) + \beta_6 \Delta(\log \sigma_t) + \beta_7 \Delta(\log \sigma_{t-1}) + \beta_8 \Delta(\log \sigma_{t-2}) + \beta_9 \Delta(\log \sigma_{t-3}) + \beta_{10} \Delta(\log \sigma_{t-4}) + \beta_{11} \Delta(\log \sigma_{t-5}) + \beta_{12} \Delta(\log h_t) + \beta_{13} \Delta(\log h_{t-1}) + \beta_{14} \Delta(\log h_{t-2}) + \beta_{15} \Delta(\log h_{t-3}) + \varepsilon_t (13)$$

where $\mathrm{k_t}$ *is new car registrations at time (t),* π_t *is the Harmonised Index of Consumer Prices for "Motor cars" at time (t),* $\mathrm{i_t}$ *is the industrial confidence indicator at time (t),* σ_t *is households' disposable income at time (t),* $\mathrm{h_t}$ *is households' savings at time (t) and ε is the error term.*

Model with all macroeconomics indicators and Google Data (Model 11)

$$k_t = \alpha + \beta_1 \log k_{t-1} + \beta_2 \log k_{t-12} + \beta_3 \Delta(\log \pi_t) + \beta_4 \Delta(\log \pi_{t-1}) + \beta_5 \Delta(i_t) + \beta_6 \Delta(\log \sigma_t) + \beta_7 \Delta(\log \sigma_{t-1}) + \beta_8 \Delta(\log \sigma_{t-2}) + \beta_9 \Delta(\log \sigma_{t-3}) + \beta_{10} \Delta(\log \sigma_{t-4}) + \beta_{11} \Delta(\log \sigma_{t-5}) + \beta_{12} \Delta(\log h_t) + \beta_{13} \Delta(\log h_{t-1}) + \beta_{14} \Delta(\log h_{t-2}) + \beta_{15} \Delta(\log h_{t-3}) + \beta_{16} \Delta(G_t) + \beta_{17} \Delta(G_{t-1}) + \beta_{18} \Delta(G_{t-2}) + \beta_{19} \Delta(G_{t-3}) + \beta_{20} \Delta(G_{t-4}) + \beta_{21} \Delta(G_{t-5}) + \beta_{22} \Delta(G_{t-12}) + \varepsilon_t \quad (14)$$

where $\mathrm{k_t}$ *is new car registrations at time (t),* $\mathrm{G_t}$ *is the Google search index at time (t),* π_t *is the Harmonised Index of Consumer Prices – Motor cars at time (t),* $\mathrm{i_t}$ *is the industrial confidence indicator at time (t),* σ_t *is households' disposable income at time (t),* $\mathrm{h_t}$ *is households' savings at time (t) and ε is the error term.*

We use monthly data starting from September 2013 for all models and of October 2013 for models with variables stationary at first difference.

We nowcast car registrations one month ahead using all eleven models. The results and their performance and nowcasting accuracy are presented in Table 5 below,[12] and the nowcasting capabilities of the base model and the best performing model in comparison with actual new euro area car registration data are presented in Chart 6.

[12] See Appendix C for the error terms formulae and the test statistics.

Table 5

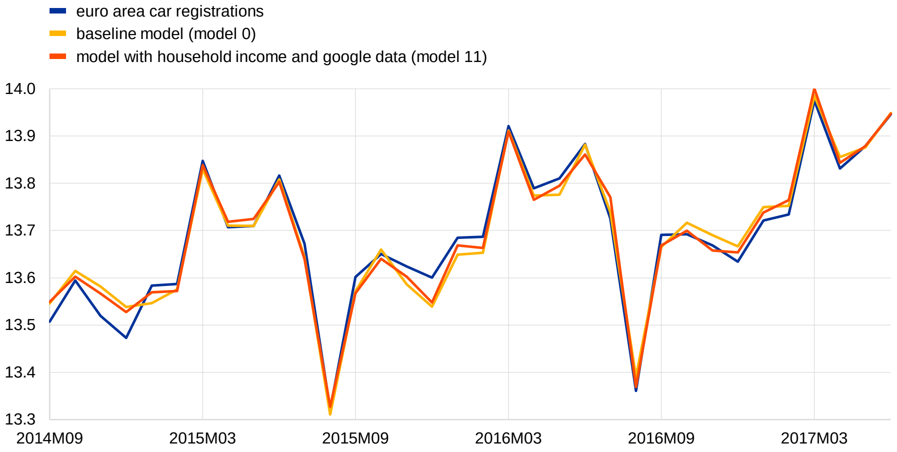
Overview of model performance and nowcasting accuracy

	Model/criteria	RMSE	MAE	MAPE	Imp*	DM# Base	DM# Pair
0	**Baseline**	0.029019	0.023711	0.174652	0	0	0.0331
1	**Baseline & Google**	0.026812	0.020303	0.149357	15.7%	0.0331	
2	**With inflation rate**	0.027585	0.023092	0.169944	3.4%	0.3357	0.0204
3	**With inflation rate & Google**	0.024885	0.019445	0.14288	21.5%	0.0383	
4	**With confidence indicator**	0.028942	0.023738	0.174715	4.6%	0.5289	0.0307
5	**With confidence indicator & Google**	0.026747	0.020334	0.149567	15.6%	0.0347	
6	**With income**	0.027437	0.02231	0.164112	5.8%	0.1791	0.0453
7	**With income & Google**	0.02296	0.018139	0.133189	30.5%	0.0195	
8	**With household savings**	0.027669	0.02122	0.15623	10.9%	0.0383	0.1965
9	**With household savings & Google**	0.025713	0.019338	0.142174	21.4%	0.0218	
Including all explanatory variables							
10	**All indicators**	0.022811	0.017993	0.13253	31.2%	0.0086	0.0019
11	**All indicators & Google**	0.012257	0.010332	0.075806	131%	0.00004	

* Imp: Percentage improvements (in percentage) of error statistics based on a synthetic average of the error measures vis-à-vis the baseline model.
DM#: Diebold and Mariano's test for comparing predictive accuracy. Base: P-values for testing all models against the baseline model. Pair: P-values for testing the model pairwise for the same macroeconomic model with and without the Google search data. For more details see Appendix D.

Chart 6

Nowcasting euro area car registrations. Comparing the euro area car registrations with the performance of the baseline model (Model 0) and the best performing model (Model 11 - all variables & Google data).Logarithmic scale



Source: Authors' calculations.

A number of conclusions can be derived from the test results shown in Table 5.

First, it appears that even the seasonal autoregressive baseline model performs quite well, generating low error terms. Nevertheless, all the other selected models reduce the error terms even further, irrespective of the macroeconomic indicator and applied error terms.

Second, the models including the Google search data appear to outperform the models which do not include it. All models which include the Google search data reduce the error terms, as compared with the corresponding models which do not include it.

Third, the model including all macroeconomic indicators and Google search data as explanatory variables performs best in nowcasting car sales, based on the improvement in the error terms.

Fourth, we use the Diebold and Mariano (1995) method to test the null hypothesis of “no difference” in the forecast accuracy between the baseline model and the other eleven models and can conclude from the test results that the null hypothesis is rejected for all models which include the google data. In other words, when using this sample, all models which include the Google search data are statistically significantly better predictors of future car sales than the baseline model. The three models which include inflation rate (Model 2), the confidence indicator (Model 4) and disposable income (Model 6) are not statistically significant better predictors than the baseline model.

Fifth, models 10 and 11 which include all the macro-economic indicators with and without the google data performs best and are statistically significant better predictors than the baseline model.

Sixth, when using the Diebold and Mariano test in a pairwise comparison for the same macroeconomic indicators with and without the Google search data, we find that the models which include the Google search data are statistically significant better predictors of future car sales than the models which do not include the Google search data – with the exception of the models which include household savings.

Seventh, we do not reject the null hypothesis of equal expected errors for the models which include household savings (with and without Google search data), meaning that these models may perform equally well in terms of forecasting ability for the sample used at the 5% significance level. We can only reject the null hypothesis at the 20% significance level.

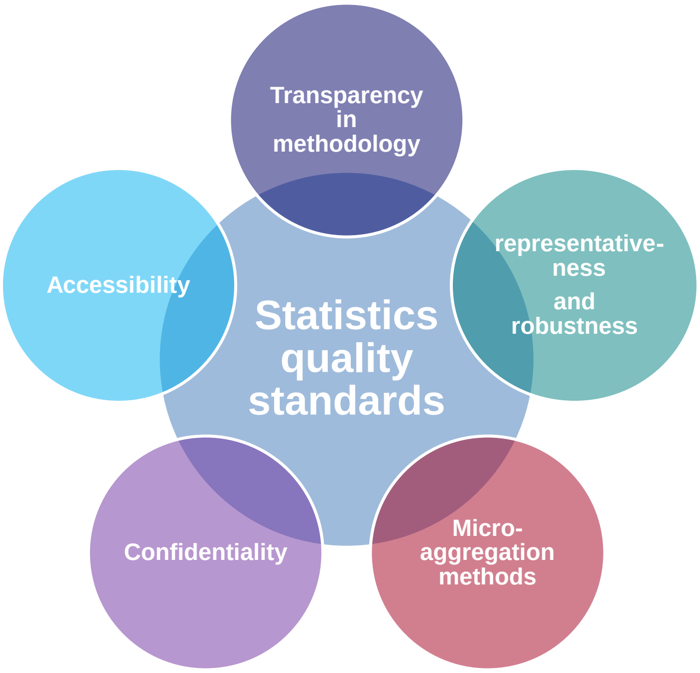
6 Quality assessment

The results described above appear positive and in line with most of the literature and the insights of Choi and Varian (2009, 2012) and Carrière-Swallow and Labbé (2013), who find that including Google search data improves the predictive ability of forecasts of car sales. It therefore seems that the Google search data have predictive capacities for nowcasting next month euro area car sales – they are, in fact, "predicting the present" and useful for experiential purposes.

Let us take a step back and discuss if these results can go beyond experimental purposes and be used for decision and policy making. For this we have develop a big data analytics quality concept (see Chart 7). It is necessary for any (big) data source to live up to these quality standards if these are to be used for decision making and as a policy toolkit.

Chart 7

Big data analytics concept: six statistical quality requirements as part of moving from experimenting to a decision making and policy toolkit


Transparency in methodology
Accessibility
representative-ness and robustness
Statistics quality standards
Confidentiality
Micro-aggregation methods

Source: Authors' designs

Representativeness

An all-too-common misconception about the use of big data is that we do not need to be concerned about representativeness or sample bias, as large volumes of information will supersede standard sampling theory – given that the sources of big data provide de facto census-type information. Representativeness remains crucial for all datasets to ensure that the data are in fact representative – in this case, internet search data on the “household” sector. There are at least six aspects to consider here, in the context of statistical data representativeness.

First, not all households have access to the internet and therefore auxiliary information is needed from households which are purchasing cars whilst not having access to the internet. The assumption of correlation between internet searches and car sales may also be caused by other exogenous elements such as increasing internet speed, increases in car websites, and structural, cultural, social or financial differences. Despite having internet access, it may not necessarily be a household’s primary source of information when contemplating a car purchase and consumers may still visit several car dealerships to obtain technical specifications and printed brochures in order to compare vehicles. Internet searches may likewise be age-dependent, with a higher proportion of users belonging to the “digital native” population.

Second, internet penetration rates may differ from country to country and therefore further country level adjustments may be necessary.

Third, if the objective is to nowcast consumer (household) car sales, we also need to analyse whether adjustments to the data are required to distinguish between internet searches done by households and by businesses. Patterns of car sales to businesses and households may well differ and therefore require adjustments.

Fourth, and importantly, a single search term does not necessarily relate to a single unit measure of an individual. A household engaged in thorough research on a car purchase could contribute to a relative increase in car searches for a given month, although the increase may not relate to more than one individual and therefore would not contribute to an increase in car sales (by more than one). Thus, a method for adjusting for double or multiple counting is warranted.

Fifth, in a similar vein, further adjustments may be required to reflect the fact that households may be using the internet (car) search facility for reasons other than new car purchases. A household may very well be searching for a used car, and such searches would not necessarily impact on new car registrations. A household may include motor vehicle enthusiasts or professionals who use the internet to search for the latest automotive technology, research or motor show. Or a household may search specifically in relation to a particular event, looking for information on whether their own particular make and model of car is affected. An example of such an event is the diesel emissions scandal, and an increase in related searches is therefore likely to be seen in September 2015, and for several months thereafter, owing to checks by households on the impact of the scandal on their specific vehicle model or research into the issue more generally. Such an exceptional event would be classified in statistical terms as an “outlier” or “noise” and would be removed from the sample, or at least the data would be adjusted by a certain factor. However, care should be taken to ensure that such exceptional events really are exceptional and do not represent structural changes. The diesel emissions scandal, for example, could prompt a structural change in demand for diesel cars, with a shift in (future) engine type preferences or a move towards alternative transport options or car sharing.

Sixth, while it is a major and widely used search facility, the Google search engine is only one of several that are available. Others such as Yahoo, Bing and Ecosia (in Germany) or Baidu and Sogou (in China) may be favoured by a particular community of users (for example young people or gamers) because the search engine provider offers related services (images, pictures, etc.) for the specific segment, and other communities may show different concentration rates for certain search engines. Therefore adjustments may be required.

Two final points of a more technical nature can also be made. The first is that the search functionality is borderless, although the allocation of searches may be country based. A household located in the Flemish-speaking area of Belgium may use a website based in the Netherlands to search for cars, in which case their aggregated searches would be allocated to the category of Dutch rather than Belgian households – although it may be possible to use the IP address to adjust for such cases. The second technical point is that adjustments for robots (or bots) may also need to be considered. While bots are generally deployed for other web-related purposes, we cannot rule out the possibility that they may be used to bombard search engines with specific search terms as part of a campaign to influence the economic value of a particular search term or category.

In view of these six considerations (plus the two more technical ones also set out above), representativeness testing and sample adjustments remain crucial, irrespective of the volume and speed of information produced by big data. These representativeness considerations and their potential adjustments have only partially been taken into account in this testing exercise. Further access to the original volume data and patterns over time would be required in order to apply meaningful adjustments to represent the household sector.

Robustness

While much of the related literature tends to test various types or a combination of models for in-sample fitness – which may often lead to the over-fitting dilemma – the touchstone is the model that improves out-of-sample nowcasting. Whether a model can do this remains an important statistical test to ensure the reliability of results. The robustness test ensures that results are replicable and that testing provides similar results over time. Robustness testing becomes even more important when forecasting in event-driven environments with frequently changing topics. We have not been able to perform out-of-sample and robustness tests owing to the relative short period of data availability. However, as more data become available, the opportunity is expected to arise in the foreseeable future.

Transparency in methodology

The structured Google taxonomy of 26 different categories and 297 sub-categories for each of the ten European countries greatly facilitate the exploration and use of Google search data. With roughly 3.5 million Google searches per minute,[13] managing and maintaining a taxonomy of search terms requires a sound and automated process and methodology. There must be transparency in the collection and allocation methods and in the decision structures for handling ambiguous terms, new terms and revisions, if applicable. The requirement for transparency in methodology also applies to calculation methods for the normalisation process, re-basing and indexing, as well as the relevant change calculations within and across categories.

Micro-aggregation methods

Similarly to the requirement for transparency in methodology, there is also a need for clarity in the micro-aggregation methods for metric, ordinal and nominal variables and the applied methods for the recognition and aggregation of text or combinations of texts.

Selecting a methodology requires methods, testing and evaluations and is a fundamental step in the process of obtaining representative, comparable and sustainable search results – the outcome of which may vary for reasons related to methodology and not necessarily to changes in absolute search volumes. It has not been feasible to obtain the methodology documentation for transparency purposes, as the methods may well be an integral part of a future business model.

Confidentiality

What methods and procedures are in place to protect the confidentiality of individual consumers' data? This question is of paramount importance to any actor involved in collecting, processing, exchanging and disseminating data to a broader audience. The relevant methods and procedures relate to documenting compliance with legal regulations, management of IT facilities and staff working with the search data, and monitoring and ensuring compliance.

Accessibility

Results, methods, metadata and documentation (including descriptions of the usage and usage limitations) need to be accessible.

The application of each of the six statistical data quality standards outlined above is mandatory when providing statistics and indicators to the public. Therefore, like any other source, big data sources need to comply with these statistical data quality

[13] Source: Cumulus Media.

standards before being considered for inclusion in any knowledge or policy-related toolkit, although this is often overlooked in the hunt for insight – in particular in social science and behaviour analysis.

Moving from experimental data to a policy toolkit for regular provision of data

Experimenting with big data sources has significant research value in testing datasets, models, hypotheses and/or new theories. However, simply finding new insights is not sufficient for their use in decision-making. To be able to use the insights the data source must transparently comply with all of the six statistical quality requirements. Otherwise it remains useful for experimental purposes, with its own merits, but with marginal impact in practice.

Although we have been able to provide econometric evidence of the existence of a relationship between "new passenger car registrations" and the "Autos & Vehicles" category of Google search data and have shown that the latter has a certain predictive ability in nowcasting car sales, we have used relatively simple models and a small dataset to do so, indicating that our results do not necessarily represent a wider economic relationship. While these results are able to serve as initial evidence, more thorough research and robustness tests are required to determine whether internet search data can be integrated into more dense and complex models and therefore be used to inform policy decisions (Giannone, Lenza and Primiceri (2017)). It is unlikely to become a decision-making tool; nor is that the intention. Instead, the intention would be to use the results as supplementary insights to provide directional and timely information, i.e. as an early indicator.

However, moving from experimental to a policy toolkit with a regular production may be a "game changer". Big data sources are currently available for exploration at (little or) no cost, but as their popularity for this purpose increases they may come to be viewed as a commercial asset and priced accordingly. This then raises questions of data ownership and the ethics of using big data. Should big data sources on individual behaviour and patterns be commercialised or should they become a public commodity? We would favour the latter for research purposes, although this question is beyond the scope of this paper.

However, the true "game changing" will come when we are able to move beyond small experiments to those which provide large-scale and holistic insights. To reach this point, at least four significant changes to current attitudes and approaches, including a regulatory point of view, are necessary.

First, as this paper argues, data owners need to comply with the statistics quality requirements in order for their data to be used beyond experimental purposes as a tool kit for decision-making. Second, public and private data owners will need to be encouraged to become data sharers, enabling borderless linking of micro-level datasets. Third, IT systems able to protect individual privacy and apply data security rules across borders will need to be put in place. Fourth, the possibility of combining structured data from various scientific fields – including statistics, sociology, medicine, political science and psychology – must be fostered. These are not insignificant challenges. But, if they can be overcome, the resulting advances in theory and knowledge around interpreting big data can be expected to significantly re-shape how we think and explain human behaviour and complex socio-economic phenomena.

7 Conclusion

The availability and accessibility of big data is a new and rich field for statisticians, economists, econometricians and forecasters, and is relatively unexploited for central banking purposes. While central banks may not have to be in front of the big data curve, these new digital footprints could potentially contribute to a new generation of high frequency and timely insights into changes in the behaviour of households, trends and turning points within the financial system. These supplementary statistics may therefore provide further insights to support central bankers' decision-making processes and timely assessments of the subsequent impact of these decisions on, and associated risks for, the financial system and the real economy.

The aim of this paper is twofold. First, we test the usefulness of Google internet search data in nowcasting euro area car sales – a leading macroeconomic indicator of economic activity. Second, we establish six data quality requirements to be met before moving beyond "experimental" purposes to the use of big data sources for policy purposes and as an element of central banking toolkits.

The first part of the paper addresses the statistical relationship between the euro area "new passenger car registration" dataset and the Google search category labelled "Autos & Vehicles". We demonstrate, by using a VAR system and an ARDL model, our initial assumption that Google search data can serve as an early indicator of changes in the volume of car sales, both in the short and long-term, and that there is a bi-directional relationship. Employing the Akaike information criterion, we find that using up to five-month lags of Google search data as explanatory variables for modelling new car sales is a reasonable assumption which produces useful results. As the Google search data display at least annual seasonality, we further include the twelve month lag.

We test the predictive capacities of using Google search data in nowcasting new car sales on the basis of a seasonal autoregressive model (baseline model).

We test model performance by using four different explanatory variables separately and as a combined model, each with and without euro area Google data. In addition to using euro area car sales as such, we test using (i) euro area household disposable income, (ii) the euro area industrial confidence indicator, (iii) the euro area harmonised inflation rate for cars and (iv) euro area household savings and (v) all explanatory variables together.

We find that when compared with the baseline model, the model including all macroeconomic indicators with five-month lags of Google data as well as a twelfth lag to account for seasonality provides the best predictive capacity. This model reduces forecasting errors by up to 131%, in comparison with the baseline model. This is an improvement in nowcasting ability, despite the models' simplicity. Furthermore, the Diebold and Mariano (1995) test confirms that the models which include the Google search data are statistically significantly better predictors of future car registrations than the baseline model, using this sample. In addition, we also find that the models which include the Google search data are statistically significant better predictors of future car sales than the equivalent models which do not include the Google search data – with the exception of the models including household savings. These are quite promising experimental results.

The paper then proceeds to a discussion of the statistical data quality requirements to be applied when moving beyond experimental data as a precondition for starting a regular production of supplementary indicators as a tool-kit for policy and/or central banking purposes. Six statistical data quality standards are presented. These standards relate to representativeness, robustness, transparency in methodology, micro-aggregation methods, confidentiality and accessibility of the data source. Compliance with all these six data quality standards is mandatory for any data source that is to be used as a toolkit for policy and central banking purposes.

While still in its infancy, there is no doubt that micro-level data can provide detailed and segregated insights into the undiscovered patterns and behaviour of our digital footprint, whether within the real economy or the financial system.

The big data service evolution is changing our society, and the way we communicate, socialise, date, collaborate, work, use and share data and information. Applying technological enhancements to methods of data collection will also alter the way in which central banks use microdata.

However, the true "game changing" moment will come when we are able to move beyond small experiments to those which provide large-scale and holistic and representative insights. To reach this point, at least four significant changes to current attitudes and approaches, including a regulatory point of view, are necessary. First, as this paper argues, data owners need to comply with the statistics quality requirements in order for their data to be used beyond experimental purposes and for decision-making. Second, public and private data owners will need to be encouraged to become data sharers, enabling borderless linking of micro-level datasets. Third, IT systems able to protect individual privacy and apply data security rules across borders will need to be put in place. Fourth, the possibility of combining structured data from various scientific fields – including statistics, sociology, medicine, political science and psychology – must be fostered. These are not insignificant challenges. But, if they can be overcome, the resulting advances in theory and knowledge around interpreting big data can be expected to significantly re-shape how we think and explain human behaviour and complex socio-economic phenomena.

References

Arouari, M., Aouadi, A., Foulquier, P. and Teulon, F. (2013), “Can Information Demand Help to Predict Stock Market Liquidity? Google it!”, Working Papers, Department of Research, Ipag Business School, 2013(24).

Askitas, N. and Zimmermann, K. (2009), “Google Econometrics and Unemployment Forecasting”, Applied Economics Quarterly, 55(2), pp. 107-120

Barreira, N., Godinho, P. and Melo, P. (2013), “Nowcasting unemployment rate and new car sales in south-western Europe with Google Trends”, *Netnomics*, *14*, pp. 129-165.

Banbura, M., Giannone, D., Modugno, M. and Reichlin, L. (2013), “Nowcasting and the Real-Time Dataflow”, in Elliot, G. and Timmerman, A., Handbook on Economic Forecasting, Elsevier, pp. 195-237.

Carrière-Swallow, Y. and Labbé, F. (2013), “Nowcasting with Google Trends in an Emerging Market”, Journal of Forecasting, 32, pp. 289-298.

Chamberlin G. (2010), “Googling the present”, Economic & Labour Market Review, 4(12), pp. 59-95.

Choi, H. and Varian, H. (2009), “Predicting the Present with Google Trends, Google Technical Report”.

Choi, H. and Varian, H. (2012), “Predicting the Present with Google Trends”, Economic Record, 88, pp. 2-9.

D’Amuri, F. and Marcucci, J. (2010), “Google it! Forecasting the US Unemployment Rate with a Google Job Search Index”, FEEM Working Paper, 31(2010)

D’Amuri, F. and Marcucci, J. (2013), “The Predictive Power of Google Searches in Forecasting Unemployment”.

Dickey, D.A. and Fuller, W.A. (1979), “Distribution of the Estimators for Autoregressive Time Series with a Unit Root”, Journal of the American Statistical Association*,* 74 (366), pp. 427-431.

Diebold, F.X. and Mariano, R. (1995), “Comparing Predictive Accuracy”, Journal of Business and Economic Statistics, 13(3), pp. 253-265.

Du, Rex Yuxing and Kamakura, Wagner A. (2012), “Quantitative Trendspotting”, Journal of Marketing Research, 49, pp. 514-536.

Fantazzini, D. and Toktamysova, Z., “Forecasting German Car Sales Using Google Data and Multivariate Models”, (2015), International Journal of Production Economics, 170, pp. 97-137.

Figueiredo, N. (2016), “Predicting Current Auto Sales in Canada using Google”, (Bachelor thesis). University of Victoria, British Columbia, Canada.

Geva, T., Oestreicher-Singer, G., Efron, N. and Shimshoni, Y. (2017), “Using forum and search data for sales prediction of high-involvement projects”, MIS Quarterly, 41(1), pp. 65-82.

Giannone, D., Lenza, M. and Primiceri, G. E. (2017), “Economic Predictions with Big Data: The Illusion of Sparsity”. Center for Economic Policy Research Discussion Paper No. DP12256.

Granger, C.W.J. (1969), “Investigating Causal Relations by Econometric Models and Cross-spectral Methods”, Econometrica, 37(3), pp. 424-438.

Greene, W. (2017). Econometric analysis (8th ed.). Upper Saddle River: Prentice Hall.

Gujarati, Damodar, & Porter, Dawn C. (2009). Basic econometrics (5th ed., McGraw-Hill Series economics). New York: McGraw-Hill

Guzmán, G. (2011), “Internet Search Behavior as an Economic Forecasting Tool: The Case of Inflation Expectations”, Journal of Economic and Social Measurement, 36, pp. 119-167

Juselius, K. (2006). The cointegrated VAR model: Methodology and applications. Oxford: Oxford University Press.

Koop, G. and Onorante, L. (2013), “Macroeconomic Nowcasting Using Google Probabilities”.

Nymand-Andersen, P. (2016), “Big data: The hunt for timely insights and decision certainty – Central banking reflections on the use of big data for policy purposes”, IFC Working Papers, No 14, Irving Fisher Committee on Central Bank Statistics, February.

Nymand-Andersen, P. (2017), “Making the most of big data”, Central Banking Journal, Vol. XXVIII, No 2, November, pp. 86-91.

Page, E.S. (1954), “Continuous Inspection Schemes”, Biometrika, 41(1-2), pp. 100-115.

Pesaran, M.H., Shin, Y. and Smith, R.J. (2001), “Bounds testing approaches to the analysis of level relationships”, Journal of Applied Econometrics, 16, pp. 289-326.

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

Tomczyk, E. and Doligalski, T., “Predicting New Car Registrations: Nowcasting with Google Search and Macroeconomic Data” (May 25, 2015), in Sł. Partycki (ed.), E-społeczeństwo w Europie Środkowej i Wschodniej. Teraźniejszość i perspektywy rozwoju, Wydawnictwo KUL, Lublin, 2015, pp. 228-236.

Vlastakis, N. and Markellos, R.N. (2010), “Information Demand and Stock Market Volatility”, Journal of Banking & Finance, Elsevier, 36(6), pp. 1808–1821.

Appendices

Appendix A: Calculation of Google weighting factors

Table A1

Population on 1 January – total

GEO/TIME	2013	2014	2015	2016	2017
Austria	8,451,860	8,507,786	8,584,926	8,700,471	8,772,865
Belgium	11,137,974	11,180,840	11,237,274	11,311,117	11,351,727
Germany	80,523,746	80,767,463	81,197,537	82,175,684	82,521,653
Ireland	4,609,779	4,637,852	4,677,627	4,726,286	4,784,383
Spain	46,727,890	46,512,199	46,449,565	46,440,099	46,528,024
France	65,600,350	65,942,267	66,456,279	66,730,453	66,989,083
Italy	59,685,227	60,782,668	60,795,612	60,665,551	60,589,445
Netherlands	16,779,575	16,829,289	16,900,726	16,979,120	17,081,507
GEO/TIME	2013	2014	2015	2016	2017
Austria	8,451,860	8,507,786	8,584,926	8,700,471	8,772,865

Source: Eurostat

Table A2

Households - level of internet access

GEO/TIME	2013	2014	2015	2016	2017
Austria	81	81	82	85	89
Belgium	80	83	82	85	86
Germany	88	89	90	92	93
Ireland	82	82	85	87	88
Spain	70	74	79	82	83
France	82	83	83	86	86
Italy	69	73	75	79	81
Netherlands	95	96	96	97	98
Portugal	62	65	70	74	75
Slovenia	76	77	78	78	82

Source: Eurostat

Table A3
Number of people who has internet access per country

GEO/TIME	2013	2014	2015	2016	2017
Austria	6846007	6891307	7039639	7395400	7807850
Belgium	8910379	9280097	9214565	9614449	9762485
Germany	70860896	71883042	73077783	75601629	76745137
Ireland	3780019	3803039	3975983	4111869	4210257
Spain	32709523	34419027	36695156	38080881	38618260
France	53792287	54732082	55158712	57388190	57610611
Italy	41182807	44371348	45596709	47925785	49077450
Netherlands	15940596	16156117	16224697	16469746	16739877
Portugal	6502119	6777746	7262375	7652584	7732180
Slovenia	1564704	1587035	1609042	1610067	1694034
Total	242089337	249900840	255854661	265850601	269998142

Table A4
Calculation of national weighting factors

GEO/TIME	2013	2014	2015	2016	2017	Average
Austria	0.0283	0.0276	0.0275	0.0278	0.0289	**0.0280**
Belgium	0.0368	0.0371	0.0360	0.0362	0.0362	**0.0365**
Germany	0.2927	0.2876	0.2856	0.2844	0.2842	**0.2869**
Ireland	0.0156	0.0152	0.0155	0.0155	0.0156	**0.0155**
Spain	0.1351	0.1377	0.1434	0.1432	0.1430	**0.1405**
France	0.2222	0.2190	0.2156	0.2159	0.2134	**0.2172**
Italy	0.1701	0.1776	0.1782	0.1803	0.1818	**0.1776**
Netherlands	0.0658	0.0647	0.0634	0.0620	0.0620	**0.0636**
Portugal	0.0269	0.0271	0.0284	0.0288	0.0286	**0.0280**
Slovenia	0.0065	0.0064	0.0063	0.0061	0.0063	**0.0063**
Total	1	1	1	1	1	**1**

Appendix B: Modelling Results

Appendix B1: Results of VAR system

Table B1.1

Results of the VAR Lag Order Selection Criteria

Lag	LogL	LR	FPE	AIC	SC	HQ
0	84.19644	NA	5.63e-05	-4.109822	-4.025378	-4.079290
1	89.34942	9.533028	5.31e-05	-4.167471	-3.914139	-4.075874
2	99.17516	17.19505	3.98e-05	-4.458758	-4.036538	-4.306097
3	109.1627	16.47940	2.96e-05	-4.758134	-4.167026	-4.544408
4	111.4429	3.534313	3.26e-05	-4.672144	-3.912148	-4.397354
5	139.7431	41.03532*	9.79e-06*	-5.887155*	-4.958272*	-5.551300*

* indicates lag order selected by the criterion
LR: sequential modified LR test statistic (each test at 5% level)
FPE: Final prediction error
AIC: Akaike information criterion
SC: Schwarz information criterion
HQ: Hannan-Quinn information criterion

Table B1.2
Results of the VAR(5) system

	D(Google data)	Log(New cars)
D(Google data(t-1))	-0.008569	-1.252064
	(0.10653)	(0.44091)
	[-0.08043]	[-2.83969]
D(Google data(t-2))	0.106100	1.761641
	(0.10336)	(0.42777)
	[1.02655]	[4.11821]
D(Google data(t-3))	-0.407254	-1.139067
	(0.11452)	(0.47397)
	[-3.55621]	[-2.40324]
D(Google data(t-4))	-0.462786	-0.439098
	(0.12894)	(0.53367)
	[-3.58903]	[-0.82278]
D(Google data(t-5))	0.247583	-0.653996
	(0.13413)	(0.55512)
	[1.84589]	[-1.17811]
Log(New cars(t-1))	0.105321	0.795126
	(0.04063)	(0.16816)
	[2.59212]	[4.72829]
Log(New cars(t-2))	0.059559	-0.262953
	(0.05096)	(0.21092)
	[1.16870]	[-1.24671]
Log(New cars(t-3))	-0.064852	0.387256
	(0.05146)	(0.21300)
	[-1.26013]	[1.81808]
Log(New cars(t-4))	0.211516	0.017683
	(0.04252)	(0.17596)
	[4.97508]	[0.10049]
Log(New cars(t-5))	-0.262233	-0.200873
	(0.03526)	(0.14594)
	[-7.43689]	[-1.37642]
Constant	-0.668141	3.601531
	(0.72618)	(3.00552)
	[-0.92007]	[1.19831]
R-squared		0.791409
Adj. R-squared	0.719480	0.578246
Sum sq. resids	0.018174	0.311310
S.E. equation	0.025034	0.103609
F-statistic	11.00277	6.347107
Log likelihood	97.17546	40.35934
Akaike AIC	-4.308773	-1.467967
Schwarz SC	-3.844331	-1.003525
Mean dependent	0.004266	13.62661
S.D. dependent	0.047265	0.159540

Substituted Coefficients, Standard errors in brackets ( ) and t-statistics results in square brackets [ ]

Appendix B2: Results of the ARDL model

Table B2.1

Results of the ARDL(5,5) model

Variable	Coefficient	Std. Error	t-Statistic	Prob.*
Log(New cars(t-1))	0.387839	0.181545	2.136321	0.0419
Log(New cars(t-2))	-0.606223	0.183165	-3.309712	0.0027
Log(New cars(t-3))	-0.229558	0.197614	-1.161647	0.2555
Log(New cars(t-4))	0.016548	0.186389	0.088783	0.9299
Log(New cars(t-5))	-0.778105	0.189174	-4.113174	0.0003
D(Google data)	0.101221	0.645084	0.156911	0.8765
D(Google data(t-1))	0.602442	0.518798	1.161227	0.2557
D(Google data(t-2))	2.898839	0.385946	7.510997	0.0000
D(Google data(t-3))	-0.044702	0.548384	-0.081515	0.9356
D(Google data(t-4))	0.827420	0.624130	1.325716	0.1960
D(Google data(t-5))	1.486467	0.583173	2.548929	0.0168
Constant	29.67952	6.273015	4.731301	0.0001
Trend	0.013796	0.002997	4.602693	0.0001
R-squared	0.842640	Mean dependent var	13.62661	
Adjusted R-squared	0.772703	S.D. dependent var	0.159540	
S.E. of regression	0.076062	Akaike info criterion	-2.057589	
Sum squared resid	0.156205	Schwarz criterion	-1.508703	
Log likelihood	54.15177	Hannan-Quinn criter.	-1.859129	
F-statistic	12.04846	Durbin-Watson stat	2.166251	
Prob(F-statistic)	0.000000			

*Note: p-values and any subsequent tests do not account for model selection.

Table B2.2

Dickey-Fuller test results

Variable	P-value (5%)	Order of Integration	Lags
Car Registration	0.0095	I(0)	0
Google data	0.0004	I(1)	1
Inflation Rate	0.00	I(1)	1
Confidence Indicator	0.00	I(1)	0
Household Income	0.03	I(1)	5
Household Savings	0.02	I(1)	3

Appendix B3: Results of the Nowcasting Models

Table B3.1

Base Model

Variable	Coefficient	Std. Error	t-Statistic	Prob.
Constant	0.068425	0.530599	0.128958	0.8982
Log(New cars(t-1))	0.007076	0.036932	0.191607	0.8493
Log(New cars(t-12))	0.992699	0.040306	24.62923	0.0000
R-squared	0.962988		Mean dependent var	13.63970
Adjusted R-squared	0.960600		S.D. dependent var	0.153105
S.E. of regression	0.030391		Akaike info criterion	-4.065269
Sum squared resid	0.028631		Schwarz criterion	-3.930590
Log likelihood	72.10957		Hannan-Quinn criter.	-4.019340
F-statistic	403.2800		Durbin-Watson stat	1.012264
Prob(F-statistic)	0.000000			

Table B3.2

Base model with Google Data (Model 1)

Variable	Coefficient	Std. Error	t-Statistic	Prob.*
Constant	0.698736	0.782562	0.892883	0.3812
Log(New cars(t-1))	-0.004202	0.062805	-0.066900	0.9472
Log(New cars(t-12))	0.957636	0.057649	16.61136	0.0000
D(Google data)	0.164438	0.210301	0.781917	0.4422
D(Google data(t-1))	-0.012083	0.152440	-0.079263	0.9375
D(Google data(t-2))	0.136577	0.152913	0.893168	0.3810
D(Google data(t-3))	-0.033768	0.157113	-0.214928	0.8317
D(Google data(t-4))	0.028400	0.143573	0.197810	0.8449
D(Google data(t-5))	0.024029	0.132231	0.181718	0.8574
D(Google data(t-12))	0.040236	0.219969	0.182917	0.8565
R-squared	0.967913		Mean dependent var	13.64522
Adjusted R-squared	0.955357		S.D. dependent var	0.152002
S.E. of regression	0.032116		Akaike info criterion	-3.793858
Sum squared resid	0.023724		Schwarz criterion	-3.340371
Log likelihood	72.59866		Hannan-Quinn criter.	-3.641273
F-statistic	77.08899		Durbin-Watson stat	0.961979
Prob(F-statistic)	0.000000			

*Note: p-values and any subsequent tests do not account for model selection.

Table B3.3

Model: Harmonised Index of Consumer Prices (Model 2)

Variable	Coefficient	Std. Error	t-Statistic	Prob.*
Constant	0.035584	0.524468	0.067848	0.9464
Log(New cars(t-1))	0.019002	0.039954	0.475611	0.6379
Log(New cars(t-12))	0.982763	0.041616	23.61526	0.0000
D(Log(Inflation))	0.036234	0.030427	1.190820	0.2434
D(Log(Inflation(t-1)))	0.034671	0.028140	1.232084	0.2278
R-squared	0.966556		Mean dependent var	13.63970
Adjusted R-squared	0.961943		S.D. dependent var	0.153105
S.E. of regression	0.029868		Akaike info criterion	-4.048994
Sum squared resid	0.025871		Schwarz criterion	-3.824529
Log likelihood	73.83290		Hannan-Quinn criter.	-3.972445
F-statistic	209.5291		Durbin-Watson stat	1.097903
Prob(F-statistic)	0.000000			

*Note: p-values and any subsequent tests do not account for model selection.

Table B3.4

Model: Harmonised Index of Consumer Prices with Google Data (Model 3)

Variable	Coefficient	Std. Error	t-Statistic	Prob.*
Constant	0.76092	0.76743	0.991517	0.3327
Log(New cars(t-1))	0.02220	0.06728	0.329968	0.7447
Log(New cars(t-12))	0.92602	0.06106	15.16545	0.0000
D(Log(Inflation))	0.04711	0.04015	1.173414	0.2538
D(Log(Inflation(t-1)))	0.05338	0.03517	1.517843	0.1440
D(Google data)	0.11540	0.20632	0.559321	0.5819
D(Google data(t-1))	-0.06022	0.15562	-0.386941	0.7027
D(Google data(t-2))	0.27576	0.16789	1.642464	0.1154
D(Google data(t-3))	-0.03282	0.15485	-0.211923	0.8342
D(Google data(t-4))	0.02838	0.14128	0.200878	0.8427
D(Google data(t-5))	0.04491	0.12900	0.348112	0.7312
D(Google data(t-12))	-0.01888	0.22006	-0.085779	0.9325
R-squared	0.972359		Mean dependent var	13.64522
Adjusted R-squared	0.957881		S.D. dependent var	0.152002
S.E. of regression	0.031195		Akaike info criterion	-3.821807
Sum squared resid	0.020436		Schwarz criterion	-3.277622
Log likelihood	75.05981		Hannan-Quinn criter.	-3.638705
F-statistic	67.15894		Durbin-Watson stat	1.052915
Prob(F-statistic)	0.000000			

*Note: p-values and any subsequent tests do not account for model selection.

Table B3.5

Model: Industrial Confidence Indicator (Model 4)

Variable	Coefficient	Std. Error	t-Statistic	Prob.*
Constant	0.118406	0.552335	0.214373	0.8317
Log(New cars(t-1))	0.012265	0.039637	0.309434	0.7591
Log(New cars(t-12))	0.983760	0.046603	21.10955	0.0000
D(Industrial Survey)	0.002560	0.006415	0.399008	0.6927
R-squared	0.963183		Mean dependent var	13.63970
Adjusted R-squared	0.959501		S.D. dependent var	0.153105
S.E. of regression	0.030811		Akaike info criterion	-4.011738
Sum squared resid	0.028480		Schwarz criterion	-3.832167
Log likelihood	72.19955		Hannan-Quinn criter.	-3.950499
F-statistic	261.6145		Durbin-Watson stat	0.964812
Prob(F-statistic)	0.000000			

*Note: p-values and any subsequent tests do not account for model selection.

Table B3.6

Model: Industrial Confidence Indicator with Google Data (Model 5)

Variable	Coefficient	Std. Error	t-Statistic	Prob.*
Constant	0.703969	0.798374	0.881752	0.3874
Log(New cars(t-1))	0.004035	0.068840	0.058619	0.9538
Log(New cars(t-12))	0.948934	0.064550	14.70076	0.0000
D(Industrial Survey)	0.002479	0.007586	0.326835	0.7469
D(Google data)	0.168751	0.214913	0.785207	0.4407
D(Google data(t-1))	-0.011250	0.155510	-0.072344	0.9430
D(Google data(t-2))	0.133575	0.156242	0.854921	0.4018
D(Google data(t-3))	-0.050438	0.168176	-0.299913	0.7671
D(Google data(t-4))	0.034152	0.147498	0.231542	0.8190
D(Google data(t-5))	0.014571	0.137945	0.105631	0.9168
D(Google data(t-12))	0.033364	0.225351	0.148054	0.8836
R-squared	0.968068		Mean dependent var	13.64522
Adjusted R-squared	0.953554		S.D. dependent var	0.152002
S.E. of regression	0.032759		Akaike info criterion	-3.738096
Sum squared resid	0.023609		Schwarz criterion	-3.239260
Log likelihood	72.67858		Hannan-Quinn criter.	-3.570253
F-statistic	66.69647		Durbin-Watson stat	0.904535
Prob(F-statistic)	0.000000			

*Note: p-values and any subsequent tests do not account for model selection.

Table B3.7

Model: Household Disposable Income (Model 6)

Variable	Coefficient	Std. Error	t-Statistic	Prob.*
Constant	0.151317	0.765914	0.197564	0.8450
Log(New cars(t-1))	0.018109	0.051862	0.349181	0.7299
Log(New cars(t-12))	0.975495	0.048814	19.98372	0.0000
D(Log(Income))	0.243967	1.322755	0.184439	0.8552
D(Log(Income(t-1)))	0.954451	1.499309	0.636594	0.5302
D(Log(Income(t-2)))	-1.269655	1.267526	-1.001680	0.3261
D(Log(Income(t-3)))	0.693788	1.253463	0.553497	0.5848
D(Log(Income(t-4)))	0.652807	1.465692	0.445391	0.6599
D(Log(Income(t-5)))	-1.077717	1.182606	-0.911307	0.3708
R-squared	0.966914		Mean dependent var	13.63970
Adjusted R-squared	0.956326		S.D. dependent var	0.153105
S.E. of regression	0.031996		Akaike info criterion	-3.824459
Sum squared resid	0.025594		Schwarz criterion	-3.420422
Log likelihood	74.01580		Hannan-Quinn criter.	-3.686671
F-statistic	91.32505		Durbin-Watson stat	0.943873
Prob(F-statistic)	0.000000			

*Note: p-values and any subsequent tests do not account for model selection.

Table B3.8

Model: Household Disposable Income with Google Data (Model 7)

Variable	Coefficient	Std. Error	t-Statistic	Prob.*
Constant	2.49237	1.43280	1.739510	0.1000
Log(New cars(t-1))	0.03591	0.09162	0.391953	0.7000
Log(New cars(t-12))	0.78391	0.09578	8.184224	0.0000
D(Log(Income))	1.09462	1.79188	0.610882	0.5494
D(Log(Income(t-1)))	5.31552	2.52668	2.103760	0.0506
D(Log(Income(t-2)))	-0.79660	2.49917	-0.318745	0.7538
D(Log(Income(t-3)))	0.25319	1.87025	0.135379	0.8939
D(Log(Income(t-4)))	3.90612	2.26875	1.721701	0.1033
D(Log(Income (t-5)))	0.54643	2.13744	0.255646	0.8013
D(Google data)	0.22993	0.29317	0.784300	0.4437
D(Google data(t-1))	-0.06209	0.25915	-0.239612	0.8135
D(Google data(t-2))	-0.04616	0.23113	-0.199713	0.8441
D(Google data(t-3))	-0.48952	0.31173	-1.570342	0.1348
D(Google data(t-4))	-0.60491	0.35674	-1.695679	0.1082
D(Google data(t-5))	0.13583	0.35662	0.380879	0.7080
D(Google data(t12))	0.14691	0.26322	0.558130	0.5840
R-squared	0.976471		Mean dependent var	13.64522
Adjusted R-squared	0.955710		S.D. dependent var	0.152002
S.E. of regression	0.031989		Akaike info criterion	-3.740442
Sum squared resid	0.017396		Schwarz criterion	-3.014862
Log likelihood	77.71729		Hannan-Quinn criter.	-3.496306
F-statistic	47.03429		Durbin-Watson stat	1.007254
Prob(F-statistic)	0.000000			

*Note: p-values and any subsequent tests do not account for model selection.

Table B3.9

Model: Household Savings (Model 8)

Variable	Coefficient	Std. Error	t-Statistic	Prob.*
Constant	0.38975	0.83163	0.468660	0.6431
Log(New cars(t-1))	-0.00313	0.05039	-0.062114	0.9509
Log(New cars(t-12))	0.97928	0.04647	21.07327	0.0000
D(Log(Savings))	0.05770	0.13066	0.441586	0.6623
D(Log(Savings(t-1)))	0.04580	0.05225	0.876441	0.3885
D(Log(Savings(t-2)))	-0.01292	0.05500	-0.234924	0.8160
D(Log(Savings(t-3)))	0.11716	0.11645	1.006081	0.3233
R-squared	0.966351		Mean dependent var	13.63970
Adjusted R-squared	0.958874		S.D. dependent var	0.153105
S.E. of regression	0.031049		Akaike info criterion	-3.925244
Sum squared resid	0.017396		Schwarz criterion	-3.610993
Log likelihood	73.72914		Hannan-Quinn criter.	-3.818075
F-statistic	129.2339		Durbin-Watson stat	0.955636
Prob(F-statistic)	0.000000			

*Note: p-values and any subsequent tests do not account for model selection.

Table B3.10

Model: Household Savings with Google Data (Model 9)

Variable	Coefficient	Std. Error	t-Statistic	Prob.*
Constant	1.31810	1.15548	1.140742	0.2682
Log(New cars(t-1))	-0.01165	0.07911	-0.147216	0.8845
Log(New cars(t-12))	0.91954	0.07376	12.46741	0.0000
D(Log(Savings))	0.13902	0.29877	0.465299	0.6470
D(Log(Savings(t-1)))	0.10914	0.10251	1.064694	0.3004
D(Log(Savings(t-2)))	-0.04152	0.11839	-0.350733	0.7296
D(Log(Savings(t-3)))	0.21712	0.39315	0.552248	0.5872
D(Google data)	0.04075	0.38170	0.106749	0.9161
D(Google data(t-1))	-0.02292	0.31040	-0.073853	0.9418
D(Google data(t-2))	0.15219	0.18473	0.823866	0.4203
D(Google data(t-3))	-0.04921	0.24690	-0.199328	0.8441
D(Google data(t-4))	-0.21476	0.26627	-0.806550	0.4299
D(Google data(t-5))	0.05731	0.28859	0.198585	0.8447
D(Google data(t12))	-0.12576	0.35335	-0.355916	0.7258
R-squared	0.970489		Mean dependent var	13.64522
Adjusted R-squared	0.950298		S.D. dependent var	0.152002
S.E. of regression	0.033887		Akaike info criterion	-3.635134
Sum squared resid	0.021819		Schwarz criterion	-3.000252
Log likelihood	73.97970		Hannan-Quinn criter.	-3.421515
F-statistic	48.06424		Durbin-Watson stat	0.954262
Prob(F-statistic)	0.000000			

*Note: p-values and any subsequent tests do not account for model selection.

Table B3.11

Model: All Variables (Model 10)

Variable	Coefficient	Std. Error	t-Statistic	Prob.*
Constant	-0.497750	1.192950	-0.417243	0.6814
Log(New cars(t-1))	0.143089	0.082942	1.725179	0.1016
Log(New cars(t-12))	0.898026	0.067848	13.23590	0.0000
D(Log(Inflation))	0.083228	0.047203	1.763199	0.0948
D(Log(Inflation(t-1)))	0.079630	0.043058	1.849390	0.0809
D(Industrial Survey)	0.001528	0.008595	0.177798	0.8609
D(Log(Income))	1.847447	2.387755	0.773717	0.4491
D(Log(Income(t-1)))	-0.065275	2.168176	-0.030106	0.9763
D(Log(Income(t-2)))	-4.312577	2.473982	-1.743172	0.0984
D(Log(Income(t-3)))	-6.010497	4.420895	-1.359565	0.1908
D(Log(Income(t-4)))	0.421232	1.838180	0.229157	0.8213
D(Log(Income (t-5)))	0.558461	1.672930	0.333822	0.7424
D(Log(Savings))	-0.930926	0.566597	-1.643013	0.1177
D(Log(Savings(t-1)))	0.165533	0.390147	0.424284	0.6764
D(Log(Savings(t-2)))	0.637325	0.397437	1.603589	0.1262
D(Log(Savings(t-3)))	0.371296	0.364923	1.017462	0.3224
R-squared	0.977130		Mean dependent var	13.63970
Adjusted R-squared	0.958071		S.D. dependent var	0.153105
S.E. of regression	0.031351		Akaike info criterion	-3.781976
Sum squared resid	0.017692		Schwarz criterion	-3.063689
Log likelihood	80.29360		Hannan-Quinn criter.	-3.537020
F-statistic	51.26993		Durbin-Watson stat	0.808747
Prob(F-statistic)	0.000000			

*Note: p-values and any subsequent tests do not account for model selection.

Table B3.12

Model: All Variables with Google Data (Model 11)

Variable	Coefficient	Std. Error	t-Statistic	Prob.*
Constant	0.156214	1.361406	0.114745	0.9109
Log(New cars(t-1))	0.342955	0.110318	3.108781	0.0111
Log(New cars(t-12))	0.647720	0.091189	7.103093	0.0000
D(Log(Inflation))	0.116579	0.036504	3.193568	0.0096
D(Log(Inflation(t-1)))	0.103804	0.035418	2.930816	0.0150
D(Industrial Survey)	-0.000105	0.007757	-0.013531	0.9895
D(Log(Income))	8.710370	2.863413	3.041954	0.0124
D(Log(Income(t-1)))	5.735553	2.918837	1.965013	0.0778
D(Log(Income(t-2)))	-5.013946	3.036443	-1.651256	0.1297
D(Log(Income(t-3)))	-14.331646	4.971787	-2.882595	0.0163
D(Log(Income(t-4)))	5.438189	2.706843	2.009053	0.0723
D(Log(Income (t-5)))	5.761736	2.284155	2.522480	0.0303
D(Log(Savings))	-2.356333	0.710406	-3.316883	0.0078
D(Log(Savings(t-1)))	0.219324	0.481636	0.455372	0.6586
D(Log(Savings(t-2)))	1.266303	0.483941	2.616647	0.0257
D(Log(Savings(t-3)))	1.054531	0.623363	1.691682	0.1216
D(Google data)	0.002997	0.347991	0.008612	0.9933
D(Google data(t-1))	-0.574455	0.349416	-1.644045	0.1312
D(Google data(t-2))	-0.709587	0.290778	-2.440303	0.0348
D(Google data(t-3))	-1.212511	0.342748	-3.537620	0.0054
D(Google data(t-4))	-1.081213	0.354272	-3.051930	0.0122
D(Google data(t-5))	-0.350872	0.341985	-1.025988	0.3291
D(Google data(t12))	0.079113	0.382285	0.206948	0.8402
R-squared	0.993295		Mean dependent var	13.64522
Adjusted R-squared	0.978543		S.D. dependent var	0.152002
S.E. of regression	0.022265		Akaike info criterion	-4.571547
Sum squared resid	0.004957		Schwarz criterion	-3.528527
Log likelihood	98.43053		Hannan-Quinn criter.	-4.220603
F-statistic	67.33545		Durbin-Watson stat	1.236910
Prob(F-statistic)	0.000000			

*Note: p-values and any subsequent tests do not account for model selection.

Appendix C: Overview of model performance

Appendix C1: Overview of model performance and nowcasting accuracy

Table C1

	Model/criteria	RMSE	MAE	MAPE	Imp*	DM# Base	DM# Pair
0	Baseline	0.029019	0.023711	0.174652	0	0	0.0331
1	Baseline & Google	0.026812	0.020303	0.149357	15.7%	0.0331	
2	With inflation rate	0.027585	0.023092	0.169944	3.4%	0.3357	0.0204
3	With inflation rate & Google	0.024885	0.019445	0.14288	21.5%	0.0383	
4	With confidence indicator	0.028942	0.023738	0.174715	4.6%	0.5289	0.0307
5	With confidence indicator & Google	0.026747	0.020334	0.149567	15.6%	0.0347	
6	With income	0.027437	0.02231	0.164112	5.8%	0.1791	0.0453
7	With income & Google	0.02296	0.018139	0.133189	30.5%	0.0195	
8	With household savings	0.027669	0.02122	0.15623	10.9%	0.0383	0.1965
9	With household savings & Google	0.025713	0.019338	0.142174	21.4%	0.0218	
Including all explanatory variables							
10	All indicators	0.022811	0.017993	0.13253	31.2%	0.0086	0.0019
11	All indicators & Google	0.012257	0.010332	0.075806	131%	0.00004	
Testing if removing income or savings variables will impact model performance (due to strong correlation between income and saving variable (R*R =73%)							
12	All (except for income)	0.02621	0.020658	0.151935	14.4%	0.0648	0.1215
13	All (except for income) & Google	0.02232	0.017842	0.130942	32.9%	0.0136	
14	All (except for savings)	0.02588	0.021348	0.156924	11.4%	0.1336	0.033
15	All (except for savings) & Google	0.02052	0.016800	0.123113	41.7%	0.0086	

Appendix C2: Error Statistics Formulae

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N} \left| y_{i,k} - \hat{y}_{i,k} \right|^2}{N}}$$

RMSE stands for Root Mean Squared Errors

where $y_{i,k}$ is the actual value and $\hat{y}_{i,k}$ the forecasted value on day k.

$$MAE_k = \frac{\sum_{i=1}^{N} \left| y_{i,k} - \hat{y}_{i,k} \right|}{N}$$

MAE stands for Mean Absolute Error

where $y_{i,k}$ is the actual value and $\hat{y}_{i,k}$ the forecasted value on day k.

MAPE= $\frac{1}{N} \sum_{i=1}^{N} \left| \frac{y_{i,k} - \hat{y}_{i,k}}{y_{i,k}} \right| \times 100\%$

MAPE stands for Mean Absolute Percentage Error

where $y_{i,k}$ is the actual value and $\hat{y}_{i,k}$ the forecasted value on day k.

Appendix D: Diebold and Mariano’s test results

Table D1

DM Test: all models as compared to baseline model

Model 1	Model 2	Test Statistic	P-value
Baseline Model	Baseline & Google	1.9016	0.0331
	Model with inflation rate	0.4281	0.3357
	Model with inflation rate & Google	1.8305	0.0383
	Model with confidence indicator	-0.0731	0.5289
	Model with confidence indicator & Google	1.8777	0.0347
	Model with household income	0.9317	0.1791
	Model with household income & Google	2.1528	0.0195
	Model with household savings	1.8282	0.0383
	Model with household savings & Google	2.1014	0.0218

Table D2

DM Test: pairwise comparison with and without Google data

Model 1	Model 2	Test Statistic	P-value
Model with inflation rate	Model with inflation rate & Google	2.1328	0.0204
Model with confidence indicator	Model with confidence indicator & Google	1.9383	0.0307
Model with household income	Model with household income & Google	1.7449	0.0453
Model with household savings	Model with household savings & Google	0.8606	0.1965
Model with all indicators	Model with all indicators & Google	3.1198	0.0019

Appendix E: Google data categories

Table E1

All Categories (Level 1), 26 in total	
Arts & Entertainment	Autos & Vehicles
Beauty & Fitness	Books & Literature
Business & Industrial	Computers & Electronics
Finance	Food & Drink
Games	Health
Hobbies & Leisure	Home & Garden
Internet & Telecom	Jobs & Education
Law & Government	News
Online Communities	People & Society
Pets & Animals	Real Estate
Reference	Science
Sensitive Subjects	Shopping
Sports	Travel

Acknowledgements
We would like to thank Valentina Rapalino, Shervin Sharifi Rad and Jacob Schiele for their useful contribution to the empirical exercise and the ECB's Statistics Paper Series Editorial Board for its valuable input. We would also like to thank Hal Varian for sharing the Google data with the central banking community.

Per Nymand-Andersen
European Central Bank, Frankfurt am Main, Germany; email: per.nymand@ecb.europa.eu

Emmanouil Pantelidis
European Central Bank, Frankfurt am Main, Germany; email: m.pantelid@gmail.com

Postal address 60640 Frankfurt am Main, Germany
Telephone +49 69 1344 0
Website www.ecb.europa.eu

This paper can be downloaded without charge from the ECB website or from RePEc: Research Papers in Economics. Information on all of the papers published in the ECB Statistics Paper Series can be found on the ECB's website.

PDF ISBN 978-92-899-3359-9, ISSN 2314-9248, doi:10.2866/6700, QB-BF-18-004-EN-N