DEUTSCHE
BUNDESBANK
EUROSYSTEM

Discussion Paper

Deutsche Bundesbank
No 18/2017

Google data in bridge equation models for German GDP

Thomas B. Götz
Thomas A. Knetsch

Deutsche Bundesbank, Wilhelm-Epstein-Straße 14, 60431 Frankfurt am Main,
Postfach 10 06 02, 60006 Frankfurt am Main

Tel +49 69 9566-0

Please address all orders in writing to: Deutsche Bundesbank,
Press and Public Relations Division, at the above address or via fax +49 69 9566-3077

Internet http://www.bundesbank.de

ISBN 978–3–95729–372–5 (Printversion)
ISBN 978–3–95729–373–2 (Internetversion)

Non-technical summary

Research Question

The internet has become an omnipresent tool in our private and professional lives. When investigating the possibility to generate knowledge about macroeconomic activity from data sets derived from the World Wide Web, nearly all studies focus on specific monthly activity indicators rather than economic activity as a whole. Additionally, systematic analyses of how one should choose among the many available data sets are rather scarce. In this paper, we thus address the following questions: Can we use internet data to improve the forecast accuracy of German GDP growth? How should we choose among the vast amount of data at our disposal?

Contribution

We incorporate data about query searches of Google keyword categories into a bridge equation model for the German macroeconomy, emphasizing the appeal of the underlying model for the integration of such "big data" information. For a set of variable selection approaches (ad-hoc, factor and shrinkage methods) we analyze whether the addition of Google search data improves forecasts of GDP growth, its components and underlying monthly indicators. Hereby, we allow the internet data to appear alongside or instead of survey indicators and inspect the sensitivity of the results to different time periods and estimation schemes.

Results

We find that large forecast accuracy gains are possible, especially when replacing survey by Google variables in equations of the underlying monthly indicators. The evidence for Google data to outperform survey variables appears stronger for more recent observations of the time period under consideration. Estimation on a rolling window even intensifies this effect, at least when selecting the Google variables by specific factor or shrinkage methods.

Nichttechnische Zusammenfassung

Fragestellung

Das Internet ist zu einem festen Bestandteil unserer privaten und beruflichen Leben geworden. Die meisten Studien, in denen untersucht wird ob man aus vom World Wide Web abgeleiteten Datensätzen Rückschlüsse auf die makroökonomische Entwicklung ziehen kann, fokussieren sich allerdings auf spezifische, monatliche Aktivitäts-Indikatoren anstatt auf die gesamte Wirtschaftsleistung. Systematische Analysen verschiedener Methoden zur Auswahl der vielen zur Verfügung stehenden Datensätze sind zudem eher rar. Das vorliegende Papier widmet sich daher folgenden Fragen: Können Internetdaten zur Verbesserung der Prognosegüte des deutschen BIP-Wachstums genutzt werden? Wie soll man aus der großen Anzahl in Frage kommender Daten wählen?

Beitrag

Wir fügen Daten aus Suchanfragen kategorisierter Google-Schlüsselwörter in ein System aus Brückengleichungen für die deutsche Volkswirtschaft ein, wobei die Eignung jener Gleichungen für die Aufnahme solcher "Big Data"-Informationen hervorgehoben wird. Für eine Reihe von Prozeduren zur Variablenauswahl (Ad-hoc-, Faktor- und Schrumpfungsmethoden) analysieren wir, ob das Hinzufügen von Google-Suchanfrage-Daten Prognosen des BIP-Wachstums, seiner Komponenten und der zugehörigen, monatlichen Indikatoren verbessert. Dabei können die Internetdaten sowohl zusätzlich zu, als auch anstatt von Umfrageindikatoren in die Gleichungen einfließen. Außerdem untersuchen wir, inwiefern sich die Ergebnisse für unterschiedliche Zeiträume und Schätzansätze verändern.

Ergebnisse

Es zeigt sich, dass große Verbesserungen der Prognosegüte möglich sind, besonders wenn in den Gleichungen der zugrundeliegenden, monatlichen Indikatoren Umfrage- durch Google-Variablen ersetzt werden. Gerade für die jüngere Vergangenheit besteht also eine gewisse Evidenz dafür, dass Google-Daten als Alternativen für Umfrageindikatoren fungieren können. Eine Schätzung mit rollierenden Fenstern erhöht diese Evidenz noch, zumindest wenn die Google-Indikatoren mittels Faktor- oder Schrumpfungsmethoden ausgewählt werden.

Bundesbank Discussion Paper No 18/2017

Google Data in Bridge Equation Models for German GDP*

Thomas B. Götz† and Thomas A. Knetsch‡

Deutsche Bundesbank

Abstract

There has been increased interest in the use of "big data" when it comes to forecasting macroeconomic time series such as private consumption or unemployment. However, applications on forecasting GDP are rather rare. In this paper we incorporate Google search data into a Bridge Equation Model, a version of which usually belongs to the suite of forecasting models at central banks. We show how to integrate these big data information, emphasizing the appeal of the underlying model in this respect. As the choice of which Google search terms to add to which equation is crucial - for the forecasting performance itself as well as for the economic consistency of the implied relationships - we compare different (ad-hoc, factor and shrinkage) approaches in terms of their pseudo-real time out-of-sample forecast performance for GDP, various GDP components and monthly activity indicators. We find that there are indeed sizeable gains possible from using Google search data, whereby partial least squares and LASSO appear most promising. Also, the forecast potential of Google search terms vis-à-vis survey indicators seems to have increased in recent years, suggesting that their scope in this field of application could increase in the future.

JEL Codes: C22, C32, C53

JEL Keywords: Big Data, Bridge Equation Models, Forecasting, Principal Components Analysis, Partial Least Squares, LASSO, Boosting

*We particularly thank Stephan Smeekes for an extensive review and many insightful remarks on an earlier version of the paper. Furthermore, we thank Etiënne Wijler, Tatjana Dahlhaus, Malte Knüppel, Daniela Rahn, Sandra Paterlini, Klemens Hauzenberger, Alain Hecq, participants at the 2016 CFE in Seville, the Workshop on Advances in Quantitative Economics II in Maastricht and members of the Big Data Project Group of the Deutsche Bundesbank for valuable comments and discussions. The views expressed in this paper are solely ours and should not be interpreted as reflecting the views of the Deutsche Bundesbank.

†Thomas B. Götz, Deutsche Bundesbank, Macroeconomic Analysis and Projection Division, Email: thomas.goetz@bundesbank.de

‡Thomas A. Knetsch, Deutsche Bundesbank, General Economic Statistics Division, Email: thomas.knetsch@bundesbank.de

1 Introduction

The internet, being one of the most influential inventions in recent history, has become a normal part of almost every person's life, surely in developed countries of the world. Gone are the days when people merely used it to send emails; by now the web is being used for buying products, booking hotels, banking, dating, research, reading the news, investing, social interactions and countless more things. The recent advent of the "sharing" culture, be it car or room sharing, only intensified the impact of the world wide web. In Germany, for example, nearly 85% of people above the age of 10 used the internet in 2015 (Destatis, 2015), whereby the rates diminish with age. While nearly everyone below the age of 45 used the internet, still about 90% of people aged 45-64 years and almost 50% of people over the age of 65 browsed the web in 2015 (+2% and +4% on the year, respectively).

Given that the internet is so widely used in our personal and professional lives, the question arises whether we are able to generate knowledge for macroeconomic activity from internet data. Luckily, advances in computer technology now enable researchers at companies or institutions to not only generate vast amounts of data, but also process non-standard, rather unstructured data emerging in business and social activities on the internet and other platforms. For such data we use the term "big data" here.[1] In this paper we investigate whether such big data - more specifically, data derived from them - lead to forecast accuracy improvements as far as macroeconomic quantities are concerned. In light of the omnipresence of the web, we focus on the gross domestic product (GDP hereafter), thereby analyzing the extent to which internet data can help predict macroeconomic activity. To be more precise, we employ Google Search data, a proxy for internet usage behavior, for Germany in this paper.[2]

To the best of our knowledge, almost all related studies (see below) focused on a specific macroeconomic indicator, usually sampled at the monthly frequency. With the exception of Wiermanski and Wilshusen (2015), no one investigated the potential forecast accuracy improvements of Google search data for economic activity as a whole, i.e., GDP growth. In this paper we intend to fill this gap in the literature by incorporating Google search data into a Bridge Equation Model (BEM hereafter), one of the workhorse models used for short-term GDP forecasting in many central banks (see, e.g., ECB, 2008, Bell et al., 2014 or Bundesbank, 2013). Indeed, the model's simplicity, transparency and structure lend themselves eminently to investigate whether Google data improve forecasts of GDP and, if they do, through which channels.

Furthermore, we contribute to the literature by analyzing various ways of choosing among the many available Google search variables. Indeed, the choice of which Google search terms enter a

[1] It is common to characterize big data using the so-called "4 V's definition" (volume, velocity, variety, veracity). Varian (2014) provides an interesting overview of tools to manipulate and analyze big data in general. Furthermore, Nymand-Andersen (2016) gives insights on the use of big data for policy purposes in central banks. See also Einav and Levin (2013) and Diebold (2012), who provide complementary views on the use of big data in econom(etr)ics.

[2] Indeed, many online activities start off with a search engine: people search for a specific product they intend to buy, look for companies they may invest in or collect information on the next potential vacation destination. Among the available internet search engines, Google is clearly the dominant one with a market share of about 95% in Germany (just below 90% worldwide) in 2016 (Destatis, 2015 and Statista, 2016, respectively).

given model often turns out to be crucial for the forecast performance in the end. We investigate different variable selection approaches in terms of their out-of-sample forecast performance: principal components analysis (PCA hereafter), partial least squares (PLS hereafter), the least absolute shrinkage and selection operator (LASSO herafter), Boosting and a couple of subjective (ad-hoc) methods. Apart from the forecasting power of the resulting Google-augmented BEM versions and various robustness checks, we also pay attention to the specific Google search terms actually chosen by the best-performing variable selection methods over time.

Investigating the forecast performance of Google search data for macroeconomic indicators has gained a lot of attention in recent years. While early work on using internet data for forecasting purposes was situated in the field of epidemiology (see, e.g., Ginsberg et al., 2009 or Johnson et al., 2004), more and more work has been devoted to improving the forecast accuracy of different macroeconomic variables. Seminal contributions were made by McLaren and Shanbhogue (2011) examining the use of internet data for the labour and housing markets, Choi and Varian (2012) forecasting automobile sales, unemployment claims, travel destination planning and consumer confidence, and Koop and Onorante (2013) introducing Google probabilities as model switching determinants within a dynamic model selection approach. Focusing on specific applications, Vosen and Schmidt (2011, 2012), Goel et al. (2010) and Toth and Hajdu (2012) dealt with forecasts of consumption, Askitas and Zimmermann (2009), D'Amuri and Marcucci (2012) and Tuhkuri (2016) applied Google search data to unemployment forecasts, the case of inflation (expectations) was analyzed by Guzman (2011) and Seabold and Coppola (2015), Humphrey (2010) considered existing home sales, Kulkarni et al. (2009) looked at housing prices, Pan et al. (2012) investigated forecasts of hotel room demand and Artola et al. (2015) forecasts of tourism inflows.

The remainder of the paper is structured as follows. In Section 2 we introduce the BEM used for our analysis and illustrate how it is augmented by internet search data. The Google data themselves are described in Section 3. In Section 4 the (statistical) methods to determine which Google search terms enter which equation of the BEM are presented. The setup and outcomes of the forecast exercise are discussed extensively in Section 5. Section 6 provides concluding remarks.

2 The Bridge Equation Model

Bridge Equation Models were introduced by Klein and Sojo (1989) as a regression-based system for GDP growth forecasting, whereby the different GDP components of the National Accounts are modelled individually. The equations for the individual GDP components are then augmented with short-term indicators tailored to the specific equation in question. Thus, intuitively speaking, the information contained in various short-term indicators gets transferred, or "bridged", to the coherent structure implied by the National Accounts (Wohlrabe, 2008). There exist many applications of BEM's in the literature, among which are Angelini et al. (2011), Baffigi et al. (2004), Camacho et al. (2013), Foroni and Marcellino (2014) and Schumacher (2014), the latter comparing MI(xed) DA(ta) S(ampling) (Ghysels et al., 2004) and BEM's as compet-

itive approaches to dealing with the mixed-frequency characteristic of many (macro)economic datasets.

Technically, a BEM is characterized by dynamic linear equations, whereby GDP growth or a component thereof represents the (low-frequency) dependent variable. Apart from low-frequency lags, the regressor set may contain time-aggregated short-term (high-frequency) indicators, e.g., industrial production. Let y_t denote the quarterly growth rate of GDP (or of one of its components) in period $t(= 1, \ldots, T)$, and let x_t^q denote the sole (for explanatory purposes) corresponding stationary short-term monthly indicator, time-averaged to the quarterly frequency (hence the superscript q). Then, the corresponding dynamic linear equation is just an autoregressive distributed lag (ADL hereafter) model:

$$y_t = \mu_y + \rho_y(L)y_{t-1} + \beta(L)x_t^q + \epsilon_t^y, \tag{1}$$

whereby $\rho_y(L) = \sum_{i=0}^{p-1} \rho_{y,i+1}L^i$ and $\beta(L) = \sum_{i=0}^{q} \beta_i L^i$ with L representing the usual lag operator, i.e., $L^i y_t = y_{t-i}$. Note that in case of cointegration between the two series, equation 1 becomes an error-correction model. Time aggregation of the underlying monthly indicator, x_t^m (note the superscript m), is undertaken using a weighting polynomial $w(L^{1/3}) = \sum_{i=0}^{r} w_i L^{i/3}$ with $L^{i/3}$ representing the high-frequency lag operator, i.e., $L^{i/3}x_t^m = x_{t-i/3}^m$, and the weights depending on the stock-flow nature of the indicator in question; for flow variables $w_i = 1/3\ \forall i$ and $r = 2$, for stock variables $w_0 = 1$ and $r = 0$ (see, e.g., Silvestrini and Veredas, 2008 for details). Fractions in the subscripts represent data points within the low-frequency period t, with $i = 0, 3, 6, \ldots$ corresponding to end-of-quarter observations. It follows that $x_{t-1/3}^m$ represents the value of x in the second month of quarter t, $x_{t-2/3}^m$ the one in the first month and $x_{t-3/3}^m \equiv x_{t-1}^m$ the one in the third month of the previous quarter.

Equation (1) can be estimated using ordinary least squares (OLS hereafter), whereafter forecasts of GDP growth, y_{T+h} say, can be computed. To do so, however, we require $x_{T+1}^q, \ldots, x_{T+h}^q$, i.e., forecasts of the time-averaged monthly indicator, which are obtained in two steps: (i) using a model specified at the monthly frequency, usually an autoregressive (AR hereafter) or also an ADL model as in

$$x_t^m = \mu_x + \rho_x(L^{1/3})x_{t-1/3}^m + \delta_x(L^{1/3})z_t^m + \epsilon_t^x \tag{2}$$

with $\rho_x(L^{1/3})$ and $\delta_x(L^{1/3})$ defined similarly to $\rho_y(L)$ and $\beta(L)$, we compute forecasts up to the end of the quarterly forecast period, i.e., up to $\ldots, T+h-1/3, T+h$; then (ii) we time-aggregate the corresponding monthly figures to the quarterly frequency using the appropriate weighting polynomial $w(L^{1/3})$. The variables z are usually survey indicators, which themselves get forecast over the same period using an AR model with straightforward definition of $\rho_z(L^{1/3})$:

$$z_t^m = \mu_z + \rho_z(L^{1/3})z_{t-1/3}^m + \epsilon_t^z \tag{3}$$

The orders of the low- and high-frequency lag polynomials are usually determined via standard information criteria. Furthermore, the forecasts of x and z could just as well be based on any

other model. To keep the BEM simple and transparent, though, ADL or AR models are often applied.

Remark 1 *Depending on whether y referred to GDP growth or a component thereof, one can compute a final, unique GDP growth forecast using either an average of the aggregate forecasts (being based on different monthly indicators) or a weighted average of the various GDP components according to their share in the National Accounts.*

In this paper, we consider an adapted submodel of the full BEM routinely run for short-term forecasting at the Deutsche Bundesbank (see Bundesbank, 2013 for details).[3] As we are interested in the potential benefits of Google data for GDP growth forecasting, we only consider an example-BEM in this paper. In particular, it is a disaggregated BEM covering the production side of the German National Accounts. To be more specific, y in Equation (1) corresponds to 15 different GDP components listed in the left column of Table 1. As far as the choice of the monthly indicators x are concerned, two criteria are considered: first, the indicator must be economically sound and, second, it must have a statistically significant impact on the target variable in question. Based on these considerations and past experience, we chose the indicators listed accordingly in the second column of Table 1. Depending on the GDP component in question, the survey indicator z is taken to be the ifo index assessing the current business situation in trade and industry (ifo ind hereafter) or the purchasing managers index in services (pmi serv hereafter), as depicted in the third column of Table 1. Note that only a survey variable is used for those GDP components without hard indicators. Here, step (2) is skipped by setting $x_t^m = z_t^m$.

Now that the BEM is introduced, let us discuss how we incorporate time series derived from Google search data. Given the structure of the BEM, i.e., using survey indicators to forecast either monthly indicators (if available) in a first step or immediately a quarterly GDP component, we propose to treat the Google data similarly to those survey indicators. Indeed, most papers in the literature focus on a specific macroeconomic indicator, e.g., consumption (Vosen and Schmidt, 2012), that might intuitively be explained by specific Google search terms. After all, users are more likely to search for "jobs", "used car" or "last-minute holiday offers" than, e.g., "GDP". We follow this practice by, on the one hand, augmenting the regression models of the monthly, usually "hard" (x-)indicators, which appear in nearly half of the bridge equations, e.g., Sales Hotel Industry. In other words, equation (2) gets augmented and becomes:

$$x_t^m = \mu_x + \rho_x(L^{1/3})x_{t-1/3}^m + \delta_x(L^{1/3})z_t^m + \gamma_x(L^{1/3})g_t^m + u_t^x, \tag{2a}$$

with g^m representing time-aggregated Google time series, whereby details on the latter are described in the next section. On the other hand, whenever Equation (2) is skipped in the BEM for a specific GDP component, i.e., when a component gets forecast only by a time-aggregated (to

[3]The data set was downloaded from the internal database of the Deutsche Bundesbank and is generally not publicly available; if one wants to replicate the results, though, the vintage of data used in this paper can be provided upon request.

Table 1: The disaggregated production-side Bridge Equation Model

GDP Component (y)	Monthly Indicators (x)	Survey Indicator (z)
Mining	Production Mining	ifo ind
Manufacturing	Industrial Production	ifo ind
Energy & Water Supply	Energy Production	ifo ind
Construction	Production in Construction	ifo ind
Trade (incl. cars)	Real Retail Sales (incl. cars)	ifo ind
Traffic	Toll (Industrial Production)	ifo ind
Hotel Industry	Sales Hotel Industry	ifo ind
Net taxes	Value-added Tax (VAT)	ifo ind
Agriculture & Forestry		ifo ind
Information & Communication		ifo ind
Housing		ifo ind
Financial Services		pmi serv
Corporate Services		pmi serv
Public Services, Health & Education		pmi serv
Other Services		pmi serv

Note: The indicator Toll is only available as of 2007, so if we date back too far within our forecast evaluation (see Section 5) for the estimation to be reliable we use Industrial Production instead. Although not displayed, Equations (2) are often augmented by variables capturing the effects of bridge and vacation days as well as weather conditions, which prove useful for some of the x-variables. These variables are either pre-determined (bridge and vacation days) or extrapolated using historical means (assuming, e.g., a "normal" winter).

match the quarterly frequency) survey indicator – as happens, e.g., for Agriculture & Forestry – we augment the equation with time-aggregated Google data. Indeed, internet users may frequently type in search terms related to Financial Services or Housing such that appropriately selected Google time series may result in forecast accuracy improvements, especially in the absence of a suitable monthly indicator. In this case, equation (1) gets augmented, whereby x_t^q needs to be replaced by z_t^q (and g_t^q) since no "hard" monthly indicator is present:

$$y_t = \mu_y + \rho_y(L)y_{t-1} + \delta_y(L)z_t^q + \gamma_y(L)g_t^q + u_t^y. \tag{1a}$$

Note that the Google search data have an indirect effect on GDP growth through their components and, if available, the respective time-aggregated monthly indicator forecasts. Extending the model in this way might also allow us to track which Google search data are responsible for any alterations in the forecast accuracy of GDP growth. Note that we abstract from adding Google data into the model equations for survey variables implying that in the augmented BEM the respective equations remain unchanged, i.e., (3a)=(3). Indeed, due to the timeliness of these indicators we do not expect significant forecast improvements here.

Schematically, the proposed augmentation of the BEM in Table 1 can be represented by renaming the third column into "Survey & Google Indicators (z and g)" and adding g, which we use as agnostic notation for a Google indicator (more on the Google variable selection in

Section 4), to each equation in the BEM.

3 Google Data

The Google search data we employ in this paper stem from a data set that is provided to the European Central Bank (ECB hereafter) by Google on a weekly basis. The data are available as of 2004 and appear without any publication delay. As in the Google Trends application *Insights for Search*, the data set comprises query searches of keyword categories, i.e., it measures the total amount of searches for a particular category relative to all search queries. Hence, only relative changes in search volumes can be assessed, not absolute search volumes (Koivupalo, 2014). This is crucial, because a search term may have a higher relative search volume in a certain time period, while having a lower absolute number of searches. The data do not get revised, but are based on random samples from all Google search queries during a day. Although the weekly data constitute an average over the corresponding seven consecutive days, the data change slightly whenever an updated data set is considered. The data are normalized to start with one (so the other figures indicate deviations from the starting value) and are greater or equal to zero (the latter representing query numbers falling short of Google's privacy filter).[4]

We time-average the Google data to match them with the monthly frequency of our indicators in the BEM, whereby we assign weeks to the month most of its days fall into. Contrary to the macroeconomic variables the Google data are not seasonally adjusted. We apply the ARIMA-X12 approach to address this issue (instead of relying on year-on-year growth rates as, e.g., in Vosen and Schmidt, 2011) assuming that, by now, the length of the time series should be long enough to compute accurate seasonal factors. As far as the order of integration is concerned, we use the bootstrap sequential quantile test (BSQT) of Smeekes (2015) to check for unit roots in a time series panel,[5] in which many series may be dependent on one another. We use ten equally spaced quantiles for our panel of in total 200 Google search series (see below). The BSQT actually returns zero rejections of a unit root (i.e., a zero proportion of $I(0)$ series) such that we compute first differences of all Google search variables in our dataset.

The data cover 14 different countries of the European Union, whereby we focus on the ones for Germany. Furthermore, the various search terms are allocated into 26 categories (Table 2) and, for a finer distinction, 269 subcategories (Table 4 in the Appendix). However, we a-priori disregarded nine categories we deemed unfitting (Arts & Entertainment, Books & Literature, Games, Hobbies & Leisure, Online Communities, People & Society, Pets & Animals, Reference and Science) as well as several subcategories from Sensitive Subjects as they were particularly prone to outliers or zero-values. In the end, our Google data set consists of 200 series.

A short stylistic note: to avoid the inflationary use of inverted commas, we use capital letters whenever we refer to a specific (sub)category. We follow the same practice for the GDP

[4] In contrast to the ECB data, Google Trends data start with a value of zero and the maximum query share during a specific time period is normalized to 100. It covers more countries and deeper levels of categories, but, crucially, the random samples the data are based on are much smaller. The ECB data, which we judge to be sufficiently granular, should thus be more accurate.

[5] We use the openly available GAUSS code corresponding to Smeekes (2015).

components and monthly indicators in the first two columns of Table 1.

Table 2: ECB Google Data: Categories

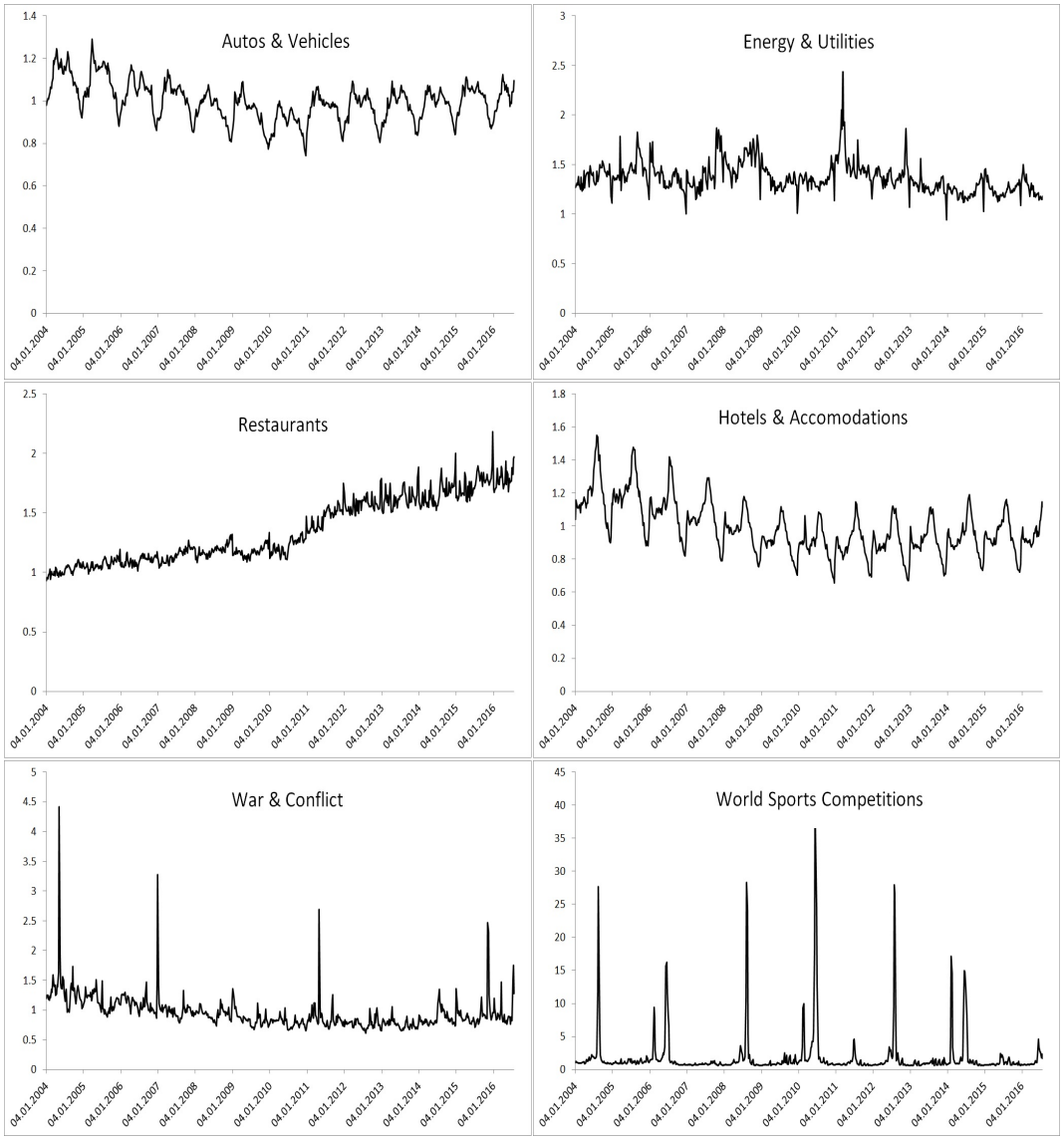
Autos & Vehicles	Beauty & Fitness	Business & Industrial
Computers & Electronics	Finance	Food & Drink
Health	Home & Garden	Internet & Telecom
Jobs & Education	Law & Government	News
Real Estate	Sensitive Subjects	Shopping
Sports	Travel	

To get a feeling for the time series derived from Google Trends we plot various representatives, seasonally unadjusted and at weekly frequency, in Figure 1. The graphs illustrate how diverse the time series implied by Google (sub)categories are: some series possess a seasonal pattern (Autos & Vehicles and Hotels & Accomodations), others show indications of a linear trend (Restaurants) or look rather stationary (Energies & Utilities), and yet others contain outliers or feature jump-like movements (War & Conflict and World Sports Competitions). Putting aside the seasonality, Autos & Vehicles as well as Hotels & Accomodations appear somewhat U-shaped indicating a lower popularity over the medium part of the sample period. The upward trend in Restaurants can be justified by more and more restaurants having an online presence and offering features like booking a table or ordering (for delivery). Energies & Utilities show troughs and peaks around winter periods, whereby the large peak in the middle of March 2011 can be associated with the Fukushima Daiichi nuclear disaster and the discussions about nuclear power phase-out in Germany. World Sports Competitions obviously show huge peaks around football World Cups and Euros as well as Winter and Summer Olympic Games. The major peaks for War & Conflict presumably reflect the wars in Iraq and Syria as well as the Ukrainian crisis.

To assess the potential such Google search terms may have for forecasting, Figure 2 contains two of the macroeconomic indicators (solid lines) whose equations get augmented with Google series, Real Retail Sales (incl. cars) and Sales Hotel Industry, together with two intuitively "fitting" Google time series (dotted lines), Autos & Vehicles and Hotels & Accommodations, respectively. Note that the series shown here are seasonally unadjusted, standardized and represented in monthly frequency, i.e., weekly Google observations are temporally aggregated as explained before. In both cases the development of the Google series is very similar to the respective macroeconomic indicator. In fact, it looks as if the former is leading the latter by one month: especially for the troughs in both series the Google data appear to be a promising leading indicator. Note also, though, that both pairs of time series show some disconnection in the beginning years of the sample period, which may have to do with Google Trends data being available only as of 2004. Overall, the graphs do suggest, though, that internet search data may very well have the ability to improve forecasts of macroeconomic indicators.

Due to the timeliness of the Google data we may expect forecast accuracy gains particularly for short forecast horizons; for now- and one-quarter-ahead forecasting say. For backcasts,

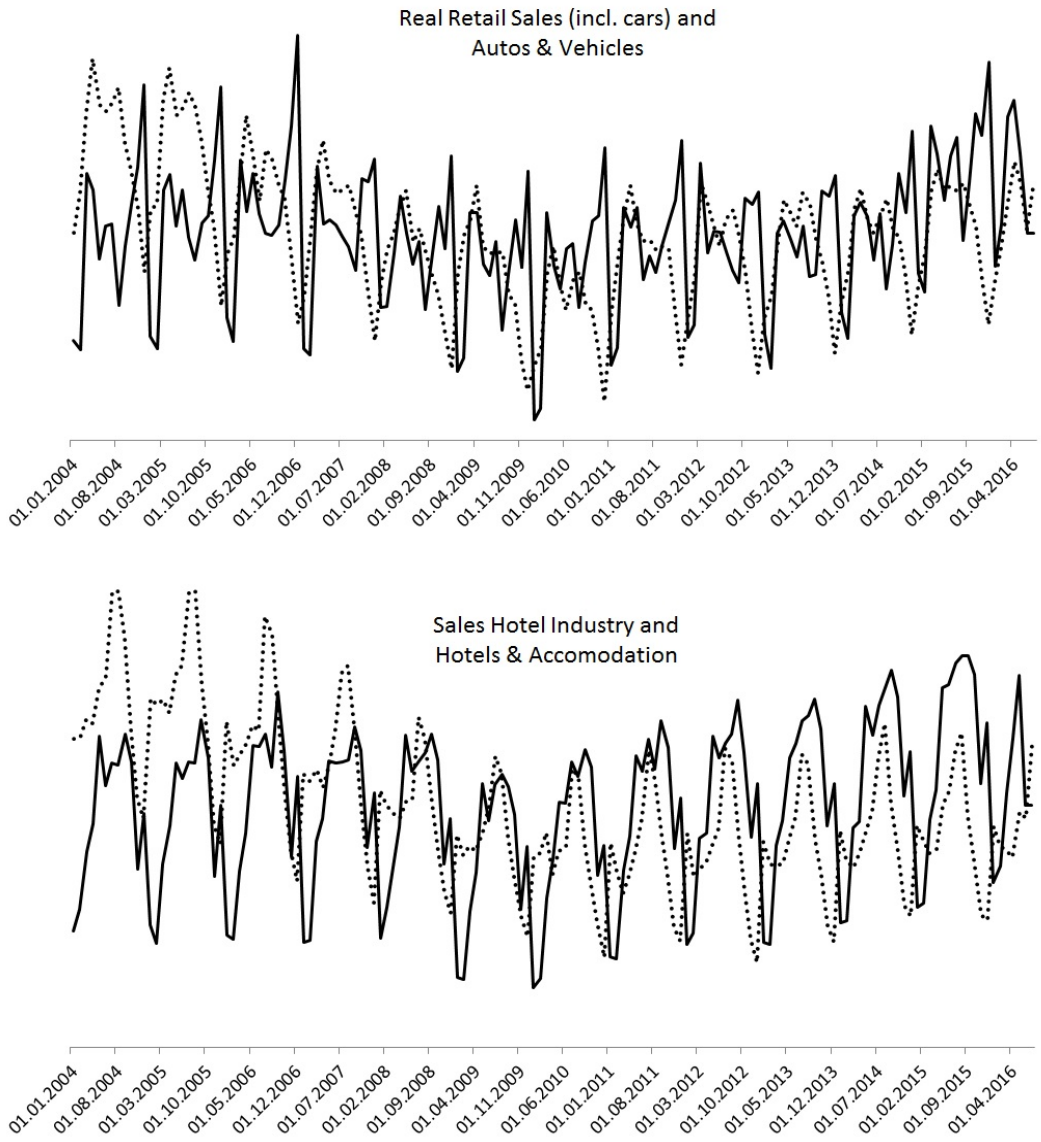
Figure 1: Various Google Time Series Examples



though, the Google series may not add much information as most of the relevant indicators have been published over the reference period. Note that we have to forecast the Google search data as soon as the forecast period under consideration extends to the future. In the spirit of handling g the same way as z, we use an AR model as in (3) to extrapolate the Google series (the lag order also chosen via information criteria).[6] To complete the augmented BEM consisting of equations (1a)-(3a) we thus have to add:

[6]We also considered VAR models (within a category for the subcategories, and across categories for the latter themselves), but the results were less promising. Shimshoni et al. (2009) analyzed the predictability of Google data and found that accounting for seasonality proves very beneficial. Instead of a-priori seasonally adjusting the Google series, we could consider specifications explicitly modelling a seasonal component in the future.

Figure 2: The Potential of Google Series for Prediction: Two Examples



$$g_t^m = \mu_g + \rho_g(L^{1/3})g_{t-1/3}^m + u_t^g. \tag{3b}$$

4 Google Variable Selection

Now that we have introduced the model and the data, it is time to discuss how we choose which Google (sub)categories enter which equation of the BEM. Naturally, including all of the candidate series is neither practical nor feasible given the amount of different (sub)categories. Also, even if there was a statistically significant relationship between a macroeconomic indicator

and a Google search term, it may turn out to be unjustifiable from an economic perspective; something we may label a "spurious relationship". Finally, a Google search term that should intuitively help forecasting a given indicator might, in fact, not have a beneficial effect due to either low popularity of the search term (e.g., industry-related ones) or adverse search behavior (e.g., in recent times, Vehicle Brands may have been looked up by potential car buyers or due to the emission scandal affecting many car brands). Consequently, the issue of which Google variable to select may be a subtle one, that probably has a large impact on the forecast performance of the augmented BEM. In the remainder of this section we describe various out-of-sample Google variable selection procedures we considered; some are rather ad-hoc, others purely data-driven.

(i) **Subjectively:** We choose the Google search data, once on a category- and once on a subcategory level, by hand, i.e., based on "common sense". The corresponding assignment of query terms is presented in Table 5 (Appendix). Note that all Google indicators enter with one lag.

(ii) **Google Correlate:** This tool, embedded into Google Trends, allows a user to search for queries that follow a similar pattern as a specific target series. The latter can be either a user-provided series or a search term itself. In our case, we upload each macroeconomic indicator (seasonally unadjusted here), that is subject to a Google variable augmentation, in turn and let Google Correlate determine the search queries whose time series possess the largest correlation coefficients. By shifting the target series by several time periods, we can inspect the relevance of lagged search terms as well. Subsequently, we manually filter out the search terms that suggest "spurious relationships" and then look for those (sub)categories in our data set corresponding as closely as possible, i.e., based on "common sense" again, to the queries obtained using Google Correlate. The resulting Google variable selection is summarized in Table 6 (Appendix).

(iii) **Principal Components Analysis:** It has become quite common in the literature to address the dimensionality problem confronted with when forecasting economic time series in a data-rich environment by imposing a factor structure to the regressors (Cubadda and Guardabascio, 2012). Summarizing the information present in the usually vast amount of predictors using factor techniques allows a user to balance the trade-off between exploiting as much information as possible while holding the amount of parameters to be estimated at bay. The standard and probably most well-known approach to extract common factors is PCA (see, e.g., Forni et al., 2005 or Stock and Watson, 2002), which has already been used in the context of forecasting with Google time series (Vosen and Schmidt, 2012).

In an attempt to a-priori avoid the factors to load on nonintuitive Google search terms, i.e., "spurious factor loadings" say, we restrict the set of eligible Google variables for a given x-series. To be more precise, for a given monthly indicator or GDP component, we allow the factors to load only on subcategories corresponding to those categories we considered for the subjective approach outlined above (the categories in the medium column of Table

5). Note that all other data-based selection procedures to come (see below) are based on the same pre-selection.[7]

We run two versions of PCA: first, we compute the factor loadings unrestrictedly over all subcategories "surviving" the aforementioned pre-selection; second, we group the eligible subcategories and subsequently draw category-specific factors. The reason to consider the latter is economic interpretability; it may be more intuitive to have, e.g., separate Autos-&-Vehicles- and Business-&-Industrial-factors for Industrial Production than various factors loading on a mixture of all the corresponding subcategories. Additionally, the resulting category-specific factors can be seen as data-driven alternatives to the Google categories, which, in contrast, are obtained using a data-driven criterion, in this case maximizing the variation in the corresponding set of subcategories. The category-specific PCA-version is henceforth labelled PCA-Cat, whereas the usual version is simply denoted PCA.

As far as the amount of factors is concerned, Vosen and Schmidt (2011) used the Kaiser-Guttman criterion, which lead to a comparably large amount of factors. Given the length of their sample period, a re-adjustment of the number of factors was necessary to avoid overfitting of the model. No such issues emerged when using the scree test in Vosen and Schmidt (2012) such that we opted for this criterion. The lag length of the resulting factors is determined via the Schwartz information criterion (SIC hereafter).

(iv) **Partial Least Squares:** Intuitively speaking, PCA extracts factors in such a way as to maximize the variance accounted for within the group of predictors. Technically, each additional factor, i.e., each linear combination of the respective regressors, maximizes the remaining variance within the set of regressors (conditional on being orthogonal to the previous factors).

PLS, originally introduced by Wold (1985) and recently proposed as an alternative to PCA by Groen and Kapetanios (2016), takes the relationship between the regressors and the target variable into account when extracting the factors. Indeed, by extracting factors in such a way as to maximize the variance accounted for within the group of predictors, PCA ignores the relationship with the target variable. Technically speaking, each additional PLS factor is defined as the linear combination that best explains the target variable, conditional on being orthogonal to the previous factors. The weights, w, of the next PLS factor are equal to the covariances between the predictors and a new target variable, which is obtained by removing the linear effects of all previously computed PLS factors:

$$w_{i+1} = \Sigma_{xy} - \Sigma_{xx}\Omega_i(\Omega_i'\Sigma_{xx}\Omega_i)^{-1}\Omega_i'\Sigma_{xy},\ i = 1, \ldots, K-1, \tag{4}$$

with $w_1 = \Sigma_{xy}$, where Σ_{xy} is the sample equivalent of the covariance between predictors and target (Σ_{xx} follows straightforwardly), $\Omega_i = (w_1, \ldots, w_i)$ and K is the number of

[7]Such a pre-selection is quite common in the literature: Vosen and Schmidt (2011, 2012), e.g., select 56 and 41 consumption-relevant categories, before extracting factors from them. See also the discussion in Remark 2 on this matter.

predictors. For details we refer to Cubadda and Guardabascio (2012).[8]

As for PCA we perform both, unrestricted and category-specific, PLS analyses based on the same pre-selection of subcategories (labelled PLS and PLS-Cat, respectively). Also analogously to the situation before, the PLS-Cat-factors can be seen as data-driven alternatives to the Google categories, with the difference of taking into account the co-movement their subcategories have with the target series.

As for the number of PLS factors and respective lag lengths, we experimented with several fixed amounts, information criteria and a cross validation approach, which takes the last two years of available observations as validation sample and the remaining ones as training sample. It turned out that the SIC leads to the stablest and most reliable outcomes.

(v) **LASSO:** The L(east) A(bsolute) S(hrinkage) and S(election) O(perator) was proposed by Tibshirani (1996) and has gained a lot of attention recently due to its documented statistical accuracy for prediction and variable selection, while being computationally feasible (see, e.g., Bühlmann and van de Geer, 2011 or Gasso et al., 2009). The motivation for LASSO is that in a linear regression model, where the amount of predictors far exceeds the time dimension, the OLS estimator is not unique and overfits the model greatly. The idea of LASSO is to regularize the complexity of the model by adding a penalty term:

$$\hat{\beta}(\lambda) = \arg\min_{\beta} \frac{1}{T} \sum_{t=1}^{T} (y_t - X_t\beta)^2 + \lambda \sum_{j=1}^{K} |\beta_j|, \tag{5}$$

where β is a K-dimensional vector, K being the number of predictors. $\lambda \geq 0$ represents the penalty parameter, whereby $\lambda = 0$ corresponds to the OLS estimator and $\lambda \to \infty$ leads to shrinking all parameters to zero. We determine the value of λ using the SIC adapted to the LASSO, i.e., where the degrees of freedom are adjusted based on the framework of Stein's unbiased risk estimation (Zou et al., 2007).[9] To guarantee a large degree of shrinkage among the pre-specified set of candidate Google series, we only consider λ-values leading to at most six non-zero coefficients in the model. Note that lagged Google observations are contained in the set of regressors, i.e., we perform variable and lag selection at the same time.

(vi) **Adaptive LASSO:** To address the potential inconsistency of the usual LASSO estimator, adapted Lasso (AdaLASSO hereafter) versions have been introduced (see, e.g., Zou, 2006

[8] PLS can be interpreted as a middle ground between PCA and canonical correlations analysis (CCA hereafter), where the target variable is usually a vector rather than a single time series (Götz et al., 2016). In CCA, linear combinations on both sides of the equation are determined in such a way as to maximize the covariance between them (again conditional on them being orthogonal to the previous factors). In such systems (often vector autoregressive models) PCA, PLS and CCA are often used to unravel an underlying reduced rank structure in the model (see, e.g., Cubadda et al., 2009).

[9] Another commonly used alternative is time series cross validation (see, e.g., Smeekes and Wijler, 2016). Note that usual k-fold cross validation (see, e.g., Bühlmann and van de Geer, 2011) is not valid in a time series setting. In an extensive Monte Carlo study of several shrinkage (and factor) methods Smeekes and Wijler (2016) show that the SIC seems to have an edge over time series cross validation (and over the Akaike information criterion for that matter).

or Konzen and Ziegelmann, 2016). To be more precise, the parameters in the penalty term are weighted in order to penalize irrelevant variables to a higher degree than relevant ones:

$$\hat{\beta}(\lambda) = \arg\min_{\beta} \frac{1}{T} \sum_{t=1}^{T} (y_t - X_t\beta)^2 + \lambda \sum_{j=1}^{K} \frac{|\beta_j|}{w_j}, \tag{6}$$

where the weights w are determined using cleverly chosen initial estimators, for which the absolute values of OLS or Ridge coefficients are common choices (Smeekes and Wijler, 2016). We follow the former approach to obtain an initial estimator, i.e., $w_j = |\hat{\beta}_j^{OLS}|$. In the second stage equation (6) is estimated, whereby we use a reformulation (see, e.g., Bühlmann and van de Geer, 2011) of the adaptive Lasso into its ordinary counterpart from (5): Set $\tilde{X}^{(j)} = w_j X^{(j)}$ and $\tilde{\beta}_j = \frac{\beta_j}{w_j}$, where $X^{(j)}$ denotes the j-th column of the $T \times K$ data matrix X, then estimating β in (6) boils down to estimating $\tilde{\beta}$ using a conventional Lasso regression. Finally, an estimator for β is obtained by back-transformation, i.e., $\beta^* = w_j \tilde{\beta}^*$, where stars indicate the estimators to be the solutions of the respective optimization problems underlying (5) and (6).

Again, we allow at most six non-zero coefficients, employ SIC to determine the tuning parameter λ and use the same pre-selection as for the standard LASSO (and thereby PCA and PLS).

(vii) **Boosting:** Finally, we propose a variant of the Boosting approach, which originates from the machine learning community (see Schapire, 1990 or Freund, 1995). Recently, the forecast performance of Boosting has been analyzed in several forecast studies (among others Lehmann and Wohlrabe, 2016 or Buchen and Wohlrabe, 2011, 2014), marking it a competitive alternative to existing approaches in data-rich settings. It is an iterative procedure starting off with a simple model, which is sequentially improved, i.e., "boosted", by adding the series with most explanatory power at each step.

1. Initialize $\hat{f}_{t,0} = \bar{y}$ for each t. Set $m = 0$.
2. Increase m by one. For each t compute $u_t = y_t - \hat{f}_{t,m-1}$.
3. For each potential regressor k, regress u_t on $g_{t,k}$ and compute the sum of squared residuals, $SSR_k = \sum_{t=1}^{T} (u_t - g_{t,k}\hat{\theta}_k)^2$. $\hat{\theta}_k$ is just the corresponding regression coefficient. Hence, we implicitly opted for an L_2-loss function and OLS as base learner.
4. Choose g_{t,k_m^*} s.t. $SSR_{k^*} = \arg\min_k SSR_k$ and set $\hat{f}_{t,m} = g_{t,k_m^*}\hat{\theta}_{k_m^*}$.
5. For each t update $\hat{f}_{t,m} = \hat{f}_{t,m-1} + \nu \hat{f}_{t,m}$, where $0 < \nu < 1$.
6. Repeat steps 2 to 5 until $m = M$.

Generally, two inputs are key to the functioning of Boosting, the step size or shrinkage parameter (usually labeled ν) and the stopping criterion (usually denoted by M). For the former, we take the commonly used value of 0.1 (see, e.g., Bühlmann and Hothorn, 2007), for the latter we choose 250.

Normally, M is chosen by cross validation or information criteria. In order to compare the Boosting-approach more directly to the aforementioned ones, however, we slightly adapt its methodology so as to function solely as a variable selection approach. To be more precise, we follow the steps outlined above and save the chosen Google regressors g_{t,k_m^*}. Then, we select those series that were chosen at least δM times, where $\delta \in [0,1]$ determines the severity with which we select the candidates. We considered a grid of δ-values, i.e., $\delta = 0.05, 0.1, 0.15, 0.2, 0.25, 0.5$ to be precise, and found $\delta = 0.2$ to give the most satisfactory outcomes. We included an additional rule, which – when it applies – often improves the outcomes: whenever no Google regressor surpasses the δM-barrier, we select the one being chosen most of the times. Again we use the same pre-selection of Google variables as before (see above).

Remark 2 *A data-driven alternative to the ad-hoc pre-selection of Google categories is the so-called Group Lasso (Hastie et al., 2015), which performs shrinkage on a group of variables instead of each variable individually and, thus, lends itself to hierarchical or grouped data structures. One can even go a step further in the form of the Sparse Group Lasso and simultaneously apply shrinkage to the members of the selected groups.*

We, however, stick to the manual pre-selection outlined above: note that nothing prevents the Group Lasso to select nonintuitive, i.e., spurious, categories making the selection potentially difficult to interpret economically. Furthermore, using the Group Lasso in a first step and, e.g., a factor approach thereafter seems less natural than immediately employing the Sparse Group Lasso; given that a comparison of various selection procedures is one of the main goals of the paper, though, such a strategy would be more fitting in a follow-up paper that uses the Sparse Group Lasso from the outset.[10]

Remark 3 *We construct PLS factors, perform shrinkage and apply Boosting solely for the Google variables, and not for lags of the respective target variable or the survey indicators as well. In other words, we apply PLS, (Ada)LASSO and Boosting only on the part of the target variable that is not explained by these, say fixed, terms. Note that for PCA this is not necessary since the computation of the factor loadings is not dependent on the target variable in question.*

5 Forecast Exercise

5.1 Setup

Given the Google variables, selected by each of the aforementioned approaches in turn, we perform a forecast exercise to assess whether including Google data improves prediction accuracy for GDP growth, its various components and the underlying monthly indicators. Hence, the outcomes of BEM's augmented with Google series are compared to the benchmark BEM without

[10] Taking the manual pre-selection as given and adding the Sparse Group Lasso as a further selection approach to our "toolbox" will most likely not be much different from the ordinary Lasso, as only a few groups (at most three per target indicator) are pre-selected anyways.

any Google search terms. We conduct the forecast exercise in pseudo-real time, i.e, we mimic the regular routine of a forecaster while abstracting from eventual data revisions.[11] Table 3 provides an overview of the characteristics of the data under consideration: availability, characteristics, transformations and publication delay.

Table 3: Data Features

Variable	Time span	Characteristics & Transformations	Pub. delay (in weeks)
15 GDP components	1991:Q1-...	in chained prices of previous years, 2010=100, sca, log-diff.	6
Production Mining	1991:M1-...	2010=100, sca, log-diff.	5
Industrial Production	1991:M1-...	2010=100, sca, log-diff.	5
Energy Production	1991:M1-...	2010=100, sca, log-diff.	5
Production in Construction	1991:M1-...	2010=100, sca, log-diff.	5
Real Retail Sales (incl. cars)	1994:M1-...	in constant prices, 2010=100, sca, logs	9
Toll	2007:M1-...	in kilometres, sca, log-diff.	7
Sales Hotel Industry	1994:M1-...	in current prices, 2010=100, sca, log-diff.	7
VAT	1991:M1-...	in millions DM/EUR, sa, log-diff.	9
ifo ind	1991:M1-...	2005=100, sca, logs	-1
pmi serv	1997:M6-...	in percentages, sa, logs	-1
All Google series	2004:W1-...	in query shares, 2003=1, sa	0

Note: sca - seasonally and calender-adjusted, sa - seasonally adjusted, Q - quarter, M - month, W - week. Assuming four weeks per month, the publication delay is defined as the amount of weeks between the moment, at which a variable gets published, and the end of the reference period, i.e., Industrial Production in June 2016 gets published five weeks after the end of June, i.e., one week into August. Note that a monotonic transformation is applied to the survey indicators so as to ensure positivity of the entries.

Having downloaded the data on 27^{th} December, 2016, we consider an increasing sequence of estimation samples starting from 1991:M1 - 2013:M6 and ending with 1991:M1 - 2016:M12. Although the evaluation period might appear rather short compared to the length of our sample, it is long enough for our forecast accuracy measures to be based on sufficient observations. Furthermore, as outlined in the Introduction, the impact of the internet on our daily lives has increased over time. Likewise, the use of search engines such as Google has also become more common in recent years: while people may have looked for specific issues, information or products in the past, new tailored applications or online services widened the range of application for Google of late. As also a larger, and more representative, share of the population starts to use the internet regularly every year, it may be that the ability of Google search data to improve

[11] The reason not to consider a real-time data set is that for some of the series the data vintages do not reach far enough into the past. Although we assume the results to be rather robust to the presence of revisions, it is an interesting sensitivity analysis to perform in the future.

macroeconomic forecasts is more visible in the recent past. Also technically, given that Google search data are only available as of 2004 and Google is known for regularly tinkering with its search algorithms, forecasts corresponding to a longer evaluation period could suffer from higher estimation uncertainty and breaks in the data. In light of these arguments we also investigate the robustness of our findings with respect to (i) a longer evaluation period (Section 5.5) and (ii) estimation on a rolling window (Section 5.6).

We follow standard practice at the Deutsche Bundesbank of synchronizing the timing of our forecasts with the publication of two main groups of indicators: "hard" and "soft" ones (Bundesbank, 2013). Consequently, each month we have two forecast dates, after the first and after the third week, which we label "early" and "late". Note that we stick to the publication calendar of the indicators, making sure that we never use data that would have not been available at that time. As our focus is on short-term, i.e., up to the next quarter, GDP forecasting, we determine the forecast horizon for Equations (1) - (3) by the publication dates of GDP. To be more precise, given the publication delay of German GDP (see the final column in Table 3), a new "round" of forecasts always starts late M2, M5, M8 and M11. Let us go through one of these "rounds", the one starting late 2014:M2 say, to illustrate the inherent forecast horizons h as far as y, x and z (or g) are concerned and whether we deal with now-, fore- or backcasts (NC, FC or BC hereafter). Similar to the publication delay in Table 3, we define the forecast horizon as the amount of weeks between the moment we make the forecast and the end of the reference period. For y we have:

- Late 2014:M2 $\Rightarrow$ $\hat{y}_{2014:Q1}$ (NC; $h = 5$), $\hat{y}_{2014:Q2}$ (FC; $h = 17$)
- Early 2014:M3 $\Rightarrow$ $\hat{y}_{2014:Q1}$ (NC; $h = 3$), $\hat{y}_{2014:Q2}$ (FC; $h = 15$)
- Late 2014:M3 $\Rightarrow$ $\hat{y}_{2014:Q1}$ (NC; $h = 1$), $\hat{y}_{2014:Q2}$ (FC; $h = 13$)
- Early 2014:M4 $\Rightarrow$ $\hat{y}_{2014:Q1}$ (BC; $h = -1$), $\hat{y}_{2014:Q2}$ (NC; $h = 11$)
- Late 2014:M4 $\Rightarrow$ $\hat{y}_{2014:Q1}$ (BC; $h = -3$), $\hat{y}_{2014:Q2}$ (NC; $h = 9$)
- Early 2014:M5 $\Rightarrow$ $\hat{y}_{2014:Q1}$ (BC; $h = -5$), $\hat{y}_{2014:Q2}$ (NC; $h = 7$)
- Late 2014:M5 $\Rightarrow$ $y_{2014:Q1}$ got published; the next "round" starts...

For x, z and g the situation is complicated by the ragged edge feature of the dataset. As explained in Section 2, we need to forecast any figures of x, z and g, that are not available over the forecast period, the latter being determined by the availability of y. Hence, (generically) for x we have:

- Late 2014:M2 $\Rightarrow$ $\hat{x}_{2014:M1}, \ldots, \hat{x}_{2014:M6}$ (1 BC, 1 NC, 4 FC; $h = -3, 1, 5, 9, 13, 17$)
- Early 2014:M3 $\Rightarrow$ $\hat{x}_{2014:M1}, \ldots, \hat{x}_{2014:M6}$ (2 BC, 1 NC, 3 FC; $h = -5, -1, 3, 7, 11, 15$)
- Late 2014:M3 $\Rightarrow$ $\hat{x}_{2014:M1}, \ldots, \hat{x}_{2014:M6}$ (2 BC, 1 NC, 3 FC; $h = -7, -3, 1, 5, 9, 13$)

- And so forth...

Actually, given the publication delays of the variables under consideration (see again the final column of Table 3), we obtain no forecast horizons smaller than -7. Obviously, figures that already became available do not need to be forecast; ifo ind and pmi serv, e.g., get published late in the respective quarter, implying that we never compute backcasts for these series ($\hat{z}_t = z_t$ then).

Remark 4 *The outcomes for the monthly and quarterly series and a specific forecast horizon are not directly comparable. Due to temporal aggregation of the monthly series prior to computing quarterly forecasts, monthly forecasts with a specific horizon enter several quarterly forecasts: monthly ($h = 9$)-forecasts, e.g., enter the equations for quarterly ($h = 9$)-, ($h = 13$)- and ($h = 17$)-forecasts. Likewise, quarterly forecasts with a specific horizon depend on several monthly forecasts: Quarterly ($h = 9$)-forecasts, e.g., are obtained using ($h = 1$)-, ($h = 5$)- and ($h = 9$)-forecasts.*

The figures in the following section represent relative root mean squared forecast errors (RMSFE's hereafter) of a Google-augmented BEM compared to the benchmark system. Hence, values larger than one favour the status-quo model, whereas values smaller than one indicate the respective augmentation by Google data to improve forecast accuracy.

5.2 Google Indicators in All Equations

We start off by comparing the BEM summarized by equations (1a)-(3b), "Aug-BEM I" say, with the benchmark model in (1)-(3), i.e., we add Google data to each BEM-equation, whereby the latter already contain survey data. In other words, we investigate whether internet data provides information beyond that contained in surveys, which enhances forecast accuracy. Table 7 in the Appendix contains the outcomes for GDP growth,[12] i.e., the weighted average of our 15 GDP components according to their share in the National Accounts, for the various Google variable selection approaches outlined in Section 4.

Focusing first on the ad-hoc variable selection methods in the first three columns of Table 7 we ascertain that the addition of subjectively chosen Google data yields similar or even slightly better forecasts ($h > 5$). For nowcasts ($0 < h \leq 5$) and backcasts ($h < 0$) the situation is less clear-cut: subjectively chosen categories provide some nowcast potential, but severely harm backcasts, the reverse holds – albeit in a lighter fashion – for Google Correlate, and subjectively chosen subcategories lead to both, worse now- and backcasts. The factor-based approaches show a partly similar picture: now- and backcasts seem to suffer from the addition of Google search data, whereas there is some potential to improve forecasts, at least when PCA and PLS are used. The former (together with Google Correlate maybe) leads to the most robust results in the sense that there are some improvements for back-, now- and forecasts and forecasts deteriorations – if present – or not too severe. The shrinkage methods perform rather disappointingly as they yield

[12] All Tables and Figures in the remainder of this section are contained in the Appendix.

at most equally good results as the benchmark model. All in all, none of the methods presents a selection that leads to an augmented BEM consistently outperforming the one without any Google data.

Let us momentarily zoom into the results of one rather well-performing approach, namely PCA, for illustration purposes. Table 8 shows the relative RMSFE's corresponding to the 15 GDP components. It appears as if the favourable outcomes for PCA stem from better forecasts of some components (presumably Manufacturing, Construction, Trade, Hotel Industry, Net Taxes for nowcasting and a couple of service-components for larger h), whereas other components suffer forecast accuracy losses (most significantly so for various service-components, e.g., Housing). As the results do not look too promising, we refrain from looking at the underlying monthly indicator forecasts at this stage.

5.3 Google Indicators in Monthly Equations only

As can be glimpsed from Table 8, the GDP components without a hard monthly indicator in their model, i.e., the ones depicted in the lower half of the table, not always benefit from the Google augmentation in (1a). While this also holds for the components in the top half of the table, the relative RMSFE's of the service-components very often reach rather extreme values (also for the other variable selection methods not depicted in Table 8), suggesting these components to react very sensitively to the addition of Google series. The absence of an intermediate monthly indicator and the resulting need to add temporally aggregated Google data to the respective component-equation is probably responsible for this. Let us thus construct a new BEM, "Aug-BEM II" say, in which we revert the augmentation of those GDP components that do not get forecast with x-indicators (see Table 1). In other words, equation (1a) is replaced by (1) again, whereby z_t^q enters the model instead of x_t^q for the respective components. Table 9 summarizes the results for GDP growth for the new augmented BEM vis-à-vis the benchmark system.

Compared to the previous augmented BEM version, the results improved in 65% of the cases, whereby the gains acquired (on average about 7%) are much larger than the losses suffered (on average about 2.5%). Noteworthily, and as hoped, refraining from the addition of (quarterly aggregated) Google variables to the equations of the service-components leads to much better results for now- and backcasts. Apart from that, most of the variable selection methods continue to deliver outcomes that are at most slightly better than the benchmark. PLS, though, yields consistently more precise GDP growth estimates except for $h = 1$ and 3. Subjectively choosing Google categories never performs worse than the benchmark for now- and forecasts, whereby the gains are often quite small (except for early, i.e., small-h, nowcasts). As the latter two approaches lead to the most robust results, let us focus on them when considering a more disaggregate level. Table 10 displays the relative RMSFE's of the GDP components with x-indicators in their model specification and of their corresponding monthly indicator forecasts.

Focusing first on the outcomes for the GDP components, it emerges that the good results for GDP growth mainly stem from improved now- and forecasts of Manufacturing, the by far biggest GDP component according to its weight in the National Accounts. There are further

instances of forecast improvements: Hotel Industry, Mining and Trade for late, i.e., large-h, PLS-forecasts as well as Mining and Traffic for late forecasts computed with subjectively chosen categories. Many instances also point to some forecast accuracy losses, though: Net Taxes and (almost always) Construction, for example. Relating the figures for the monthly indicators to their corresponding GDP components sometimes gives the puzzling picture, in which a better (or worse) forecast performance for the former is not accompanied by an equally good (or bad) performance for the latter. Retail Sales and Trade as well as Energy Production and Energy & Water (for now- and especially backcasts) are two such cases. One should keep in mind, though, that the outcomes for the monthly and quarterly series and a specific forecast horizon are not directly comparable (see Remark 4) and that temporal aggregation of the x-variables may impact the quarterly forecasts of the GDP components in a more or less fortunate way. It could, however, also point towards a disentanglement of a GDP component and its assigned monthly indicator.

5.4 Google instead of Survey Indicators

All in all, it seems as if the extent to which Google search data provide information beyond that already contained in survey indicators, and which could be useful for forecasting, is rather limited. While some gains are possible for GDP growth, the situation is more ambiguous for its components. Hence, rather than investigating whether Google search data can add information on top of survey indicators, let us analyze the situation in which we include them instead of survey variables. To this end, we amend equation (2a) to

$$x_t^m = \mu_x + \rho_x(L^{1/3})x_{t-1/3}^m + \gamma_x(L^{1/3})g_t^m + v_t^x \tag{2a*}$$

and compare the new augmented BEM, "Aug-BEM III" say, consisting of equations (1), (2a*), (3a) and (3b) with the benchmark system. Due to the fact that, by and large, the outcomes improved after removing Google variables from equations (1) without x-indicators, we do so here as well. Tables 11 and 12 contain the corresponding set of results in the same spirit as before.

As far as GDP growth is concerned (Table 11) the outcomes improved in 80% of the instances compared to Table 9. Large gains, i.e., on average about 20% and maximally 46%, are possible for forecasting and long-horizon nowcasts (i.e., $h \geq 5$), also backcasts seem to benefit from using Google instead of survey variables, albeit to a lesser degree. Early nowcasts continue to present a challenge for Google search data, though. Indeed, the sometimes large relative RMSFE's point towards Google data missing some crucial information contained in the survey variables.

Even though suitably chosen Google search data seem to constitute a valid alternative to survey variables, especially for fore- and late nowcasts, we do not claim that they should replace survey variables in practice altogether. The validity and usefulness of indicators derived from surveys, that are specifically designed for various sectors of a macroeconomy, is well established and documented for various model specifications (beyond the example BEM we employ here), time periods, applications and so forth. On top of that, survey indicators are available for a longer period of time and are very transparent as to how they are obtained, guaranteeing a

certain level of representativity and reliability. But the outcomes presented thus far at least point towards the potential of internet search data in general, and Google indicators in particular, to contain information that is not embedded in survey variables and which could prove beneficial for macroeconomic forecasting. This potential can be expected to increase in the future, when, on the one hand, the length of Google time series increases making estimation more reliable and, on the other hand, even more people use the internet making the data themselves more reliable.

Investigating the GDP component and monthly indicator forecasts for PLS and LASSO (Table 12) – for they constitute the most robust factor-based and shrinkage method, respectively[13] – unveils that forecast improvements of Manufacturing, Mining and Hotel Industry (as well as Trade for PLS and Traffic for LASSO to some extent) seem to be mainly driving the good results for GDP growth. The results for Construction are clearly inferior to the ones with survey data, the remaining components show more or less no changes in forecast accuracy. As for the monthly indicators, most of them, and especially Industrial Production, benefit from the use of Google instead of survey variables. Note that mismatches between the forecasts of the hard monthly indicators and the respective GDP components prevail.

As a final note, the main analysis presented thus far shows that the way, in which one selects Google data used for forecasting, is crucial for the forecast performance in the end. Some components and indicators (e.g., Construction) generally show much less potential for Google variables to improve their forecasts, such that survey variables should undoubtedly be preferred in these instances. In any case, though, careful and competent monitoring by the researcher is inevitable.

5.5 Longer Evaluation Period

As already mentioned before, we intend to investigate in how far our results are robust to a different, longer evaluation period. To this end, we repeat the forecast exercise outlined above for a time span twice as long as the one under investigation before, i.e., 2010:M1-2016:M12. We investigate both, adding Google data to (Aug-BEM II) or instead of (Aug-BEM III) survey variables, but continue to ignore Google variables in equations (1) when x-indicators are absent. Tables 13 and 14 summarize the results for the comparison between Aug-BEM II and the benchmark system.

As far as GDP growth is concerned the outcomes appear even better than before (i.e., compared to Table 9), with the PLS-variants constituting the best selection methods. For PLS-Cat the addition of Google variables yields accuracy gains between 3 and 18%, whereby they are largest for now- and forecasting. Focusing on both PLS-versions for the GDP components and respective monthly indicator forecasts, the results continue to originate primarily from improved fore-, now- and backcasts of Manufacturing and Industrial Production. With such a long evaluation period, it appears as if the addition of Google to survey variables already leads to noteworthy forecast improvements. Let us nevertheless investigate how a replacement of the

[13]The most robust ad-hoc selection method, subjectively chosen categories, is almost always outperformed by either PLS or LASSO (or both); the only exceptions are $h = 1$ and 3. The results are qualitatively very similar, though.

latter by the former affects the results, i.e., when comparing Aug-BEM III and the benchmark BEM in Tables 15 and 16.

For GDP growth, the situation looks partly similar, partly different from the one we obtained using the short evaluation period. Fore- and late nowcasts seem to benefit from using Google data instead of survey variables, but in a lighter fashion than with the short evaluation period. Backcasts, however, are mostly harmed by removing survey indicators altogether. In the sense that the gains – if any – from using Google series instead of survey variables are smaller than in the previous subsection, the outcomes point towards a larger relevance of the survey variables over the first half of the evaluation period compared to the second half. Hence, the figures support that – over time – Google search data may have the potential to eventually replace survey variables instead of merely adding some information not already embedded in survey variables. For completeness, the results for the GDP components and underlying indicators are, by and large, similar to the ones from before, so we do not discuss them explicitly here.

5.6 Rolling Window Estimation

In view of the intuition that the relevance of Google search data for macroeconomic forecasting increased over time, and in light of the results presented thus far being based on recursive estimations, we repeated the entire analysis with a rolling windows estimation. To be more precise, we consider a window length of six years, which we deem long enough for the estimation to be reliable, but short enough for the windows to gradually adapt to changes in the time series under consideration or in the relationships among them. Hence, we start off with the estimation sample 1997:M7-2013:M6 and keep the length of that sample constant as our forecast exercise progresses such that we finish with the period 2011:M1-2016:M12. Similar to the previous subsection, we compare Aug-BEM II and Aug-BEM III with the benchmark model, the latter also being computed on a rolling window, of course. Tables 17 (Aug-BEM II) and 19 (Aug-BEM III) contain the results for GDP growth, Tables 18 and 20 the ones for the GDP components and monthly indicators, whereby we focus on PLS and LASSO as best-performing methods. Again, we do not go into detail on the latter set of outcomes as they are qualitatively not much different from before.

Clearly, the outcomes are much different compared to the ones using a recursive estimation. Instances with forecast accuracy gains are rather scarce, only PLS, and maybe the LASSO-variants for backcasts, show some improvements. The relative RMSFE's for the other variable selection methods, however, are far higher than before. As far as the benefits of adding Google series to the system are concerned, it seems as if the increased reliability associated with longer estimation samples outweighs eventual structural changes that could affect the data.

Although the conclusions drawn for Aug-BEM II above continue to hold for a couple of selection approaches, especially the ad-hoc ones and also the PCA-versions, the situation is somewhat different for PLS and the shrinkage methods. Here, now- and forecast accuracy gains are recorded in nearly every instance, with gains reaching up to 55% (PLS and $h = 5$) and averaging about 20% (as before). Hence, when it comes to replacing survey by Google variables,

using a rolling window estimation even intensifies the forecast potential of Google search data, at least when being selected with PLS or LASSO.

5.7 Simultaneous Selection of Google and Survey Indicators

Recall from Table 9 or Section 5.3 that the addition of Google to survey variables in the monthly indicator equations lead to a very similar forecasting performance vis-à-vis the benchmark model. In other words, both classes of data seem to explain the same variation in the dependent variable – at least for the short evaluation period – which is why we omitted survey variables in the sequel (Table 11 or Section 5.4). This indeed led to noteworthy forecast accuracy gains for GDP growth, yet continued to deliver a mixed picture for its components and the underlying monthly indicator forecasts. The latter, however, suggests that the two data classes may not contain the exact same information and the differences in informational content influence forecasts on a more disaggregate level.

In order to investigate this issue, let us relax Remark 3 and let the data decide whether to include Google or survey indicators or both. To this end, we focus on the two best performing methods from the analysis thus far, i.e., PLS and LASSO, and include survey variables into the selection process. In the case of PLS this boils down to computing factor loadings over all eligible Google subcategories as well as the respective survey indicator. In the case of LASSO we simply add the contemporaneous as well as six lagged observations of the survey variable to the set of eligible Google series (and their lags) and let the operator decide which coefficients are shrunk to zero and which not.[14]

It turns out that LASSO never chooses any of the survey indicators for the evaluation period under consideration, implying that the results are identical to the ones in Tables 11 and 12. Apparently, the selected Google subcategories show a larger degree of commonality with the respective monthly indicator than if survey variables are added to or instead of them. Table 21 below summarizes the outcomes for PLS, whereby we merge the GDP growth outcomes with the ones for the disaggregate quantities.

All in all, the outcomes are not too different, suggesting that the influence of adding survey indicators to the PLS-factor(s) is rather limited. If there are differences, though, they mostly feature small improvements in the figures, implying that survey variables might add some unique information to the model. In a few instances (e.g., Retail Sales or Value-added Tax, VAT hereafter), however, we obtain a pronounced deterioration of the relative performance of the augmented model.

5.8 The Google Series Chosen

Let us have a closer look at the Google series, that actually get selected by the most promising approaches. Instead of going through all the instances considered in the previous subsections,

[14] We could repeat the analysis for PCA, AdaLASSO and Boosting, yet refrain from doing so here to save on space. The categorized versions of PCA and PLS would assign separate factors to survey indicators and thereby force both classes of data into the model. The ad-hoc procedures, i.e., subjectively chosen (sub)categories and Google Correlate, obviously do not qualify for this analysis.

we will – for illustration purposes – focus on the situation, in which Google variables are used instead of survey indicators in monthly indicator equations, i.e., we look at the results of Aug-BEM III in more detail. Since the outcomes are somewhat more convincing for the shorter evaluation period, we zoom into the Google variable selection underlying the bottom part of Table 12. Furthermore, we focus on PLS and LASSO as they were the methods leading to the best results in this instance (recall Table 11). Note that they also serve as representatives of the factor-based procedures (PLS) as well as the shrinkage-methods (LASSO), where a search term is either selected or not.

We start by inspecting the query terms selected by PLS. Recall that under this procedure we determine the loadings unrestrictedly over the eligible subcategories (see Section 4 for details). We consider Industrial Production and Sales Hotel Industry as examples since PLS often leads to forecast accuracy gains for these two x-indicators. We focus on the composition of the first PLS-factors here to save on space; after all it is the one capturing the largest correlations of the eligible Google subcategories and the target variable in question.[15] Figures 3 and 4 (at the end of the Appendix) show the loadings of the first PLS-factors over time. Note that the loadings are scaled such that they add up to 100%. This way we can better compare the relative importance of one search term compared to the other ones in that category.

Starting with the factor for Industrial Production, the largest – in absolute terms – weights are recorded for Vehicle Brands, Vehicle Codes & Driving Laws, Business Education, Vehicle Shopping, Boats & Watercraft and Chemicals Industry. Apparently, search terms related to the production or sale of vehicles show a lot of co-movement with Industrial Production. It is somewhat surprising that the two search terms with the largest weight among the Business-&-Industrial-category are Business Education and Chemicals Industry rather than Manufacturing, but internet users might simply be more inclined to look for such related terms. Another interesting observation is that the majority of subcategories loads negatively on the first PLS-factor. Turning to the factor for Sales Hotel Industry, the five subcategories with the largest average (absolute) weight are Travel Guides & Travelogues, Luggage & Travel Accessories (with a negative loading, though), Car Rental & Taxi Services, Tourist Destinations and Carpooling & Vehicle Sharing. All of these search terms appear logical, especially in light of many online services centering around such issues (e.g., *Tripadvisor*). Note that the Travel-subcategories mostly load positively on the factor, whereas the Food-&-Drinks- as well as Sensitive-Subjects-subcategories receive a negative loading most of the times. When discussing the sign of the factor loadings, though, one should keep in mind that we should normally also inspect the respective PLS-factor-coefficient in the equations for Industrial Production and Sales Hotel Industry. We stick to an illustrative discussion here.

Let us now turn to the Google variables chosen by LASSO and focus exemplarily on Industrial Production, Energy Production and VAT, all three of which achieving improvements in

[15] It turns out that for Industrial Production three factors are chosen almost over the entire evaluation period; the only exception is May 2014 when only two factors enter the model. For Sales Hotel Industry we obtain two factors most of the times; a third factor is added 2013:M7-2013:M11, 2014:M3-2014:M4, 2014:M9-2014:M10 and 2015:M1. Results are available upon request.

forecasting performance using this approach.[16] Figure 5 shows the results and should be read as follows: whenever a color appears as a vertical bar the corresponding subcategory is selected and enters the monthly indicator equation "fully", i.e., if two colors share the vertical space, both Google search terms are selected with a weight of one.

All in all, the outcomes appear quite intuitive and mirror some of the results on the PLS selection above. Vehicle Brands and Boats & Watercraft seem most important for Industrial Production, the "spot-on" subcategory Energy & Utilities proves useful for Energy Production and VAT is dominated by Banking.

6 Conclusion

In this paper we analyzed whether (data derived from) "big data" carry useful information for predictions of economic activity. In particular, we incorporated Google search data, a proxy for internet usage behavior, into a Bridge Equation Model for the German macroeconomy to assess whether they can improve GDP growth forecasts. Treating the Google variables similarly to survey indicators, they affected GDP growth either through its components directly or through underlying monthly indicators, that – when present in a model equation – have an effect on the corresponding GDP component in a preceding step. To address the crucial issue of which Google search terms to choose, we considered several variable selection approaches: Subjectively chosen Google (sub)categories and a selection based on Google Correlate made up the set of ad-hoc methods, variants of principal components analysis and partial least squares constituted factor-based techniques, and, finally, two LASSO-versions as well as an approach based on Boosting represented devices based on shrinkage or machine learning. Subsequently, the performance of accordingly augmented Bridge Equation Models vis-à-vis the benchmark model without internet search data was analyzed in a pseudo-real time out-of-sample forecast exercise.

It emerged that when adding Google data to all equations of the Bridge Equation Model forecast accuracy gains are rather limited. Improvements to this initial setup were detected after refraining from an augmentation of the GDP component equations directly, i.e, when letting Google data only enter underlying monthly indicator equations. In yet another specification of the augmented model, we replaced the survey variables by Google indicators instead of adding the latter to the former. In this case, large forecast accuracy gains were detected, especially for fore- and late nowcasts, providing some evidence for Google data to be a potential alternative to survey variables. This result, however, only partly extended to the underlying GDP components and monthly indicators. These findings were confirmed when considering a longer evaluation period in the sense that, under this scenario, larger improvements were found when Google data were added to survey variables instead of replacing them. In other words, the aforementioned potential of Google indicators to serve as an alternative to those based on surveys only "kicks

[16] We do not show the graphs corresponding to Production Mining and Sales Hotel Industry, indicators for which larger forecast accuracy gains were detected as well, because they are not very illustrative: Production Mining is influenced predominantly by the Automotive Industry and for Sales Hotel Industry the subcategory Travel Guides & Travelogues is the only one selected over the entire evaluation period.

in" when being evaluated on a more recent period. A further robustness analysis with a rolling window instead of a recursive estimation scheme lead to an even intensified forecast potential of Google search data, at least when being selected with PLS or LASSO. The latter two variable selection approaches were the ones leading to the overall largest, and most robust, forecast improvements.

When looking at the results of our analysis on a more disaggregate level, forecast accuracy improvements were possible for some indicators, but not for others. In future work it might be worthy to consider an even tighter Google variable pre-selection than the one presented here to address this issue. A data-driven alternative is the (Sparse) Group Lasso already mentioned before (see Remark 2). Maybe, however, it even pays off more to consider specific, tailored Google search terms instead of categorized versions. Furthermore, Google search data might not be the right representatives of internet data for all indicators in question. By now, many internet platforms or software applications have emerged targeting specific markets or groups. It could be that data derived from *Amazon* or *Autoscout24* would be better fits for predicting Real Retail Sales (incl. cars), *Tripadvisor* or *HRS* for the Hotel Industry and *Immobilienscout24* for Production in Construction. Finally, one should also keep in mind that the Bridge Equation model we considered in this paper is merely an example. An interesting future analysis would be to incorporate Google data into alternative model specifications, e.g., a dynamic factor model. All in all, though, we feel confident in concluding that, although there are still many open issues and pitfalls with using internet search data, they surely show enough potential to improve macroeconomic forecasts.

References

Angelini, E., CambaMendez, G., Giannone, D., Reichlin, L., and Rünstler, G. (2011). Shortterm forecasts of euro area GDP growth. *Econometrics Journal*, 14(1):C25–C44.

Artola, C., Pinto, F., and de Pedraza García, P. (2015). Can internet searches forecast tourism inflows? *International Journal of Manpower*, 36(1):103–116.

Askitas, N. and Zimmermann, K. F. (2009). Google Econometrics and Unemployment Forecasting. *Applied Economics Quarterly*, 55(2):107–120.

Baffigi, A., Golinelli, R., and Parigi, G. (2004). Bridge models to forecast the euro area {GDP}. *International Journal of Forecasting*, 20(3):447 – 460.

Bell, V., Co, L. W., Stone, S., and Wallis, G. (2014). Nowcasting UK GDP growth. *Bank of England Quarterly Bulletin*, 54(1):58–68.

Buchen, T. and Wohlrabe, K. (2011). Forecasting with many predictors: Is boosting a viable alternative? *Economics Letters*, 113(1):16–18.

Buchen, T. and Wohlrabe, K. (2014). Assessing the Macroeconomic Forecasting Performance of Boosting: Evidence for the United States, the Euro Area and Germany. *Journal of Forecasting*, 33(4):231–242.

Bühlmann, P. and Hothorn, T. (2007). Boosting Algorithms: Regularization, Prediction and Model Fitting. *Statistical Science*, 22(4):477–505.

Bühlmann, P. and van de Geer, S. (2011). *Statistics for High-Dimensional Data: Methods, Theory and Applications*. Springer Publishing Company, Incorporated, 1st edition.

Bundesbank (2013). Forecasting Models in short-term business cycle analysis - a workshop report. *Deutsche Bundesbank Monthly Report September 2013*, pages 69–83.

Camacho, M., Perez-Quiros, G., and Poncela, P. (2013). Short-term forecasting for empirical economists: A survey of the recently proposed algorithms. *Foundations and Trends in Econometrics*, 6(2):101–161.

Choi, H. and Varian, H. (2012). Predicting the Present with Google Trends. *The Economic Record*, 88(s1):2–9.

Cubadda, G. and Guardabascio, B. (2012). A medium-n approach to macroeconomic forecasting. *Economic Modelling*, 29(4):1099 – 1105.

Cubadda, G., Hecq, A., and Palm, F. C. (2009). Studying co-movements in large multivariate data prior to multivariate modelling. *Journal of Econometrics*, 148(1):25–35.

D'Amuri, F. and Marcucci, J. (2012). The predictive power of Google searches in forecasting unemployment. Temi di discussione (Economic working papers) 891, Bank of Italy, Economic Research and International Relations Area.

Destatis (2015). IT-Nutzung. `https://www.destatis.de/DE/PresseService/Presse/Pressemitteilungen/2015/12/PD15_466_63931pdf.pdf?__blob=publicationFile`. Retrieved on 05-August-2016.

Diebold, F. (2012). A personal perspective on the origin(s) and development of 'big data': The phenomenon, the term, and the discipline. Working Paper 13-003, PIER.

ECB (2008). Short-term forecasts of economic activity in the euro area. *ECB Monthly Bulletin April 2008*, pages 69–74.

Einav, L. and Levin, J. (2013). The data revolution and economic analysis. Discussion Papers 12-017, Stanford Institute for Economic Policy Research.

Forni, M., Hallin, M., Lippi, M., and Reichlin, L. (2005). The Generalized Dynamic Factor Model: One-Sided Estimation and Forecasting. *Journal of the American Statistical Association*, 100:830–840.

Foroni, C. and Marcellino, M. (2014). A comparison of mixed frequency approaches for nowcasting Euro area macroeconomic aggregates. *International Journal of Forecasting*, 30(3):554–568.

Freund, Y. (1995). Boosting a weak learning algorithm by majority. *Information and Computation*, 121(2):256 – 285.

Gasso, G., Rakotomamonjy, A., and Canu, S. (2009). Recovering sparse signals with a certain family of nonconvex penalties and dc programming. *Transactions on Signal Processing*, 57(12):4686–4698.

Ghysels, E., Santa-Clara, P., and Valkanov, R. (2004). The MIDAS Touch: Mixed Data Sampling Regression Models. CIRANO Working Papers 2004s-20, CIRANO.

Ginsberg, J., Mohebbi, M. H., Patel, R. S., Brammer, L., Smolinksi, M. S., and Brilliant, L. (2009). Detecting influenza epidemics using search engine query data. *Nature*, 457(7232):1012–1014.

Goel, S., Hofmann, J. M., L. S., Pennock, D. M., and Watts, D. J. (2010). Predicting consumer behavior with Web Search. *Proceedings of the National Academy of Sciences of the United States of America*, 1077(41):107486–17490.

Götz, T. B., Hecq, A., and Smeekes, S. (2016). Testing for Granger causality in large mixed-frequency VARs. *Journal of Econometrics*, 193(2):418–432.

Groen, J. J. and Kapetanios, G. (2016). Revisiting useful approaches to data-rich macroeconomic forecasting. *Computational Statistics & Data Analysis*, 100:221 – 239.

Guzman, G. (2011). Internet Search Behavior as an Economic Forecasting Tool: The Case Of Inflation Expectations. *The Journal of Economic and Social Measurement*, 36(3).

Hastie, T. R., Tibshirani, R., and Wainwright, M. (2015). *Statistical Learning with Sparsity: The Lasso and Generalizations*. CRC Press.

Humphrey, B. D. (2010). Forecasting existing home sales using Google seach engine queries. Honors thesis, Duke University.

Johnson, H. A., Wagner, M. M., Hogan, W. R., Chapman, W., Olszewski, R. T., Dowling, J., and Barnas, G. (2004). Analysis of Web access logs for surveillance and influenza. *Studies in Health Technology and Informatics*, 107:1202–1206.

Klein, L. R. and Sojo, E. (1989). *Combinations of High and Low Frequency Data in Macroeconometric Models*, volume 17 of *Advanced Studies in Theoretical and Applied Econometrics*, chapter 1, pages 3–16. Springer Netherlands.

Koivupalo, H. (2014). Google data. mimeo, European Central Bank.

Konzen, E. and Ziegelmann, F. A. (2016). Lasso-type penalties for covariate selection and forecasting in time series. *Journal of Forecasting*, 35:592–612.

Koop, G. and Onorante, L. (2013). Macroeconomic Nowcasting using Google Probabilities. Working paper, European Central Bank.

Kulkarni, R., Haynes, K. E., Stough, R. R., and Paelinck, J. H. P. (2009). Forecasting housing prices with Google econometrics. Worksing paper no. 2009-10, George Mason University.

Lehmann, R. and Wohlrabe, K. (2016). Looking into the black box of boosting: the case of Germany. *Applied Economics Letters*, forthcoming.

McLaren, N. and Shanbhogue, R. (2011). Using internet search data as economic indicators. *Bank of England Quarterly Bulletin*, 51(2):134–140.

Nymand-Andersen, P. (2016). Big data: the hunt for timely insights and decision certainty. Ifc working papers no 14, Bank for International Settlements.

Pan, B., Wu, D. C., and Song, H. (2012). Forecasting hotel room demand using search engine data. *Journal of Hospitality and Tourism Technology*, 3(3):196–210.

Schapire, R. E. (1990). The strength of weak learnability. *Machine Learning*, 5(2):197–227.

Schumacher, C. (2014). MIDAS and bridge equations. Discussion Papers 26/2014, Deutsche Bundesbank, Research Centre.

Seabold, S. and Coppola, A. (2015). Nowcasting prices using Google trends: an application to Central America. Policy Research Working Paper Series 7398, The World Bank.

Shimshoni, Y., Efron, N., and Matias, Y. (2009). On the Predictability of Search Trends. Technical report, Google, Israel Labs.

Silvestrini, A. and Veredas, D. (2008). Temporal aggregation of univariate and multivariate time series models: A survey. *Journal of Economic Surveys*, 22(3):458–497.

Smeekes, S. (2015). Bootstrap sequential tests to determine the order of integration of individual units in a time series panel. *Journal of Time Series Analysis*, 36(3):398–415.

Smeekes, S. and Wijler, E. (2016). Macroeconomic Forecasting Using Penalized Regression Methods. Research Memorandum 039, Maastricht University, Graduate School of Business and Economics (GSBE).

Statista (2016). Worldwide desktop market share of leading search engines. `http://www.statista.com/statistics/216573/worldwide-market-share-of-search-engines/`. Retrieved on 05-August-2008.

Stock, J. and Watson, M. (2002). Forecasting Using Principal Components From a Large Number of Predictors. *Journal of the American Statistical Association*, 97:1167–1179.

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

Toth, I. J. and Hajdu, M. (2012). Google as a tool for nowcasting householod consumption: estimations on Hungarian data. Working paper, Institute for Economic and Enterprise Research.

Tuhkuri, J. (2016). Forecasting Unemployment with Google Searches. ETLA Working Papers 35, The Research Institute of the Finnish Economy.

Varian, H. R. (2014). Big Data: New Tricks for Econometrics. *Journal of Economic Perspectives*, 28(2):3–28.

Vosen, S. and Schmidt, T. (2011). Forecasting private consumption: surveybased indicators vs. Google trends. *Journal of Forecasting*, 30(6):565–578.

Vosen, S. and Schmidt, T. (2012). A monthly consumption indicator for Germany based on Internet search query data. *Applied Economics Letters*, 19(7):683–687.

Wiermanski, C. and Wilshusen, S. M. (2015). Exploring the use of anonymized consumer credit infromation to estimate economic conditions: An application of Big Data. Discussion paper, Payment Cards Center.

Wohlrabe, K. (2008). *Forecasting with mixed-frequency time series models*. PhD thesis, Ludwig-Maximilians-Universität München.

Wold, H. (1985). *Partial Least Squares*. John Wiley & Sons, Inc.

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, 101(476):1418–1429.

Zou, H., Hastie, T., and Tibshirani, R. (2007). On the degrees of freedom of the lasso. *The Annals of Statistics*, 35(5):2173–2192.

Appendix

Table 4: ECB Google Data: Subcategories

	Autos & Vehicles	
Bicycles & Accessories	Boats & Watercraft	Campers & RVs
Classic Vehicles	Commercial Vehicles	Custom & Performance Vehicles
Hybrid & Alternative Vehicles	Microcars & City Cars	Motorcycles
Off-Road Vehicles	Personal Aircraft	Scooters & Mopeds
Trucks and SUVs	Vehicle Brands	Vehicle Codes & Driving Laws
Vehicle Maintenance	Vehicle Parts & Accessories	Vehicle Shopping
Vehicle Shows		
	Beauty & Fitness	
Beauty Pageants	Body Art	Cosmetic Procedures
Cosmetology & Beauty Professionals	Face & Body Care	Fashion & Style
Fitness	Hair Care	Spas & Beauty Services
Weight Loss		
	Business & Industrial	
Advertising & Marketing	Aerospace & Defence	Agriculture & Forestry
Automotive Industry	Business Education	Business Finance
Business Operations	Business Services	Chemicals Industry
Construction & Maintenance	Energy & Utilities	Hospitality Industry
Industrial Materials & Equipment	Manufacturing	Metals & Mining
Pharmaceuticals & Biotech	Printing & Publishing	Professional & Trade Associations
Retail Trade	Small Business	Textiles & Nonwovens
Transportation & Logistics		
	Computers & Electronics	
CAD & RAM	Computer Hardware	Computer Security
Consumer Electronics	Electronics & Electrical	Enterprise Technology
Networking	Programming	Software
	Finance	
Accounting & Auditing	Banking	Credit & Lending
Financial Planning & Management	Grants, Scholarships & Financial Aid	Insurance
Investing		
	Food & Drink	
Beverages	Cooking & Recipes	Food & Grocery Retailers
Restaurants		
	Health	
Ageing & Geriatrics	Alternative & Natural Medicine	Health Conditions
Health Education & Medical Training	Health Foundations & Medical Research	Medical Devices & Equipment

Medical Facilities & Services	Medical Literature	Men's Health
Mental Health	Nursing	Nutrition
Oral & Dental Care	Pediatrics	Pharmacy
Public Health	Reproductive Health	Substance Abuse
Vision Care	Women's Health	

Home & Garden

Bed & Bath	Domestic Services	Gardening & Landscaping
Home Appliances	Home Furnishing	Home Improvement
Homemaking & Interior Decor	Home Storage & Shelving	HVAC & Climate Control
Kitchen & Dining	Laundry	Nursery & Playroom
Pest Control	Swimming Pools & Spas	Yard & Patio

Internet & Telecom

Communications Equipment	Email & Messaging	Mobile & Wireless
Search Engines	Service Providers	Teleconferencing
Web Apps & Online Tools	Web Portals	Web Services

Jobs & Education

Education	Jobs

Law & Government

Government	Legal	Military
Public Safety	Social Services	

News

Broadcast & Network News	Business News	Gossip & Tabloid News
Health News	Journalism & News Industry	Local News
Newspapers	Politics	Sports News
Technology News	Weather	World News

Real Estate

Apartments & Residential Rentals	Commercial & Investment Real Estate	Property Development
Property Inspections & Appraisals	Property Management	Real Estate Agencies
Real Estate Listings	Timeshares & Vacation Properties	

Sensitive Subjects

Accidents & Disasters	Death & Tragedy	War & Conflict

Shopping

Antiques & Collectibles	Apparel	Auctions
Classifieds	Consumer Resources	Entertainment Media
Gifts & Special Event Items	Luxury Goods	Mass Merchants & Department Stores
Photo & Video Services	Shopping Portals & Search Engines	Swap Meets & Outdoor Markets
Tobacco Products	Toys	Wholesalers & Liquidators

	Sports	
College Sports	Combat Sports	Extreme Sports
Fantasy Sports	Individual Sports	Motor Sports
Sporting Goods	Sports Coaching & Training	Team Sports
Water Sports	Winter Sports	World Sports Competitions
	Travel	
Air Travel	Bus & Rail	Carpooling & Vehicle Sharing
Car Rental & Taxi Services	Cruises & Charters	Hotels & Accomodations
Luggage & Travel Accessories	Specialty Travel	Tourist Destinations
Travel Agencies & Services	Travel Guides & Travelogues	

Table 5: Google Variable Selection: Subjectively

Monthly Indicator / GDP component	Category-Level	Subcategory-Level
Prod. Mining	Business & Industrial	Agriculture & Forestry, Metals & Mining
Ind. Prod.	Autos & Vehicles, Business & Industrial	Classic Vehicles, Automotive Industry, Chemicals Industry, Industrial Materials & Equipment, Manufacturing
Energy Prod.	Business & Industrial, Home & Garden	Energy & Utilities, HVAC & Climate Control
Production Constr.	Business & Industrial, Home & Garden, Real Estate	Construction & Maintenance, Gardening & Landscaping, Property Development
Retail Sales	Autos & Vehicles, Sensitive Subjects, Shopping	Classic Vehicles, Vehicle Shopping, War & Conflict, Shopping Portals & Search Engines
Toll	Autos & Vehicles, Business & Industrial	Commercial Vehicles, Trucks & SUVs, Automotive Industry, Transportation & Logistics
Hotel Ind.	Food & Drink, Sensitive Subjects, Travel	Restaurants, War & Conflict, Hotels & Accommodations
VAT	Finance, Law & Government, News	Financial Planning & Management, Legal, Business News
Agric. & Fores.	Business & Industrial	Agriculture & Forestry
Info. & Comm.	Computer & Electronics, Internet & Telecom	Consumer Electronics, Communications Equipment, Email & Messaging
Housing	Home & Garden, Real Estate	Home Furnishing, Real Estate Agencies
Financial Services	Finance	Credit & Lending, Financial Planning & Management
Corporate Services	Business & Industrial, Finance, News	Business Services, Banking, Business News
Public Services, Health & Educ.	Finance, Health, Jobs & Education	Grants, Scholarships & Financial Aid, Medical Facilities & Services, Social Services, Education
Other Services	Business & Industrial, Internet & Telecom, Travel	Spas & Beauty Services, Domestic Services, Service Providers, Car Rental & Taxi Services, Travel Agencies & Services

Note: The first eight rows correspond to monthly indicators (x in Table 1), the remaining seven capture those GDP components that do not get augmented with an x-indicator such that the Google series enter directly in a time-aggregated fashion. All Google indicators enter with one (monthly) lag in case of a monthly indicator equation and with no (quarterly) lags in case of a quarterly GDP component equation.

Table 6: Google Variable Selection: Google Correlate

Macro Indicator / GDP component	Subcategories
Prod. Mining	Business Operations (+1 lag), Industrial Materials & Equipment
Ind. Prod.	Metals & Mining (+1 lag), Apartments & Residential Rentals (+1 lag)
Energy Prod.	Spas & Beauty Services (+2 lags), HVAC & Climate Control, Winter Sports (+2 lags)
Production Constr.	Construction & Maintenance, Manufacturing, Home Improvement
Retail Sales	Toys
Toll	Construction & Maintenance, Transportation & Logistics, Public Safety
Hotel Ind.	Gifts & Special Events Items
VAT	Restaurants, Car Rental & Taxi Services (+2 lags)
Agric. & Fores.	Chemicals Industry
Info. & Comm.	Software, Web Apps & Online Tools
Housing	Real Estate Listings
Financial Services	Business News
Corporate Services	Financial Planning & Management, Service Providers
Public Services, Health & Educ.	Web Services
Other Services	Education, Travel Agencies & Services

Note: Lags are indicated in brackets whenever present. For the rest see Table 5.

Table 7: Augmented BEM I, Survey & Google Data, 2013:M7-2016:M12, GDP Growth

Method		Subj.-Cat.	Subj.-Subcat.	Google Corr.	PCA-Cat	PCA	PLS-Cat	PLS	LASSO	AdaLASSO	Boosting
Forecast Horizon	17	0.98	0.95	1.00	1.02	0.98	1.00	0.97	0.99	0.99	1.00
	15	0.99	0.96	1.00	1.03	0.98	1.01	0.97	1.01	1.01	1.00
	13	0.96	0.93	1.00	1.05	0.98	1.02	0.99	1.03	1.03	1.02
	11	0.97	0.94	0.99	1.06	0.99	1.03	1.04	1.05	1.05	1.03
	9	0.99	0.98	0.98	1.09	1.02	0.98	0.99	1.09	1.09	1.06
	7	0.98	0.97	0.98	1.09	1.03	0.91	0.95	1.11	1.11	1.07
	5	1.09	0.99	0.99	1.02	1.00	1.04	1.03	1.06	1.06	1.01
	3	0.95	1.11	1.03	1.19	0.94	1.11	1.07	1.15	1.15	1.01
	1	0.94	1.08	1.02	1.18	0.93	1.14	1.11	1.16	1.16	0.99
	-1	1.09	1.10	0.98	1.21	0.98	1.25	1.24	1.09	1.09	1.00
	-3	1.09	1.10	0.98	1.19	1.00	1.25	1.24	1.09	1.09	1.00
	-5	1.20	1.12	0.97	1.23	1.06	1.25	1.37	1.03	1.03	0.89

Note: The figures represent RMSFE's of the Google-variable-augmented BEM in (1a)-(3b) relative to the benchmark BEM in (1)-(3). The various Google variable selection methods underlying the augmentations are described in Section 4. The forecast horizons -5, -3 and -1 correspond to backcasts, 1, 3 and 5 to nowcasts and 7 to 17 to forecasts.

Table 8: Augmented BEM I, Survey & Google Data, 2013:M7-2016:M12, GDP Components

		PCA							
		Mining	Manufact.	Energy & Water	Constr.	Trade	Traffic	Hotel Ind.	Net Taxes
Forecast Horizon	17	1.00	1.00	1.00	1.04	1.00	1.01	1.03	0.92
	13	1.00	1.01	1.00	1.03	0.96	1.00	1.01	0.93
	9	0.96	1.01	1.01	1.02	1.02	0.99	1.01	0.92
	5	0.96	1.00	1.03	1.01	1.00	1.00	0.96	0.91
	1	1.07	0.94	1.01	0.97	0.96	1.00	0.98	0.92
	-3	1.04	1.00	0.99	1.02	0.95	1.00	1.01	0.96
		PCA							
		Agr. & For.	Info. & Comm.	Housing	Fin. Services	Corp. Services	Publ. Services	Other Services	
Forecast Horizon	17	1.03	0.92	1.28	0.91	0.90	0.97	0.97	
	13	1.02	0.91	1.28	0.90	0.96	0.98	1.01	
	9	1.13	0.89	1.37	0.90	1.07	0.99	0.98	
	5	1.38	1.00	1.22	0.97	1.50	1.07	1.00	
	1	1.38	1.00	1.22	0.97	1.52	1.06	1.02	
	-3	1.39	1.00	1.38	1.03	1.39	1.06	1.01	

Note: Only a subset of the forecast horizons is shown to save on space. For the rest see Table 7.

Table 9: Augmented BEM II, Survey & Google Data, 2013:M7-2016:M12, GDP Growth

Method		Subj.-Cat.	Subj.-Subcat.	Google Corr.	PCA-Cat	PCA	PLS-Cat	PLS	LASSO	AdaLASSO	Boosting
Forecast Horizon	17	1.00	1.01	1.01	0.99	0.99	0.99	0.92	0.97	0.97	0.96
	15	1.00	1.01	1.01	1.00	0.99	0.99	0.92	0.98	0.98	0.97
	13	0.98	0.99	1.01	1.01	1.00	1.00	0.95	0.98	0.98	0.98
	11	0.98	0.99	1.00	1.02	1.02	1.01	0.98	1.01	1.01	1.00
	9	0.99	1.00	0.99	1.03	1.03	0.94	0.92	1.03	1.03	1.01
	7	1.00	0.96	1.00	1.00	1.03	0.86	0.86	1.02	1.02	1.00
	5	1.00	1.02	1.00	1.03	1.02	0.94	0.91	1.06	1.06	1.04
	3	0.92	1.15	1.04	1.07	0.95	1.03	1.05	1.07	1.07	1.07
	1	0.92	1.10	1.03	1.04	0.94	1.04	1.09	1.07	1.07	1.05
	-1	1.03	1.09	0.99	1.01	1.00	1.03	0.96	1.05	1.05	1.03
	-3	1.02	1.10	1.00	0.98	0.98	1.02	0.95	1.06	1.06	1.04
	-5	1.01	0.97	1.00	1.00	1.00	1.01	1.00	0.99	0.99	0.96

Note: The figures represent RMSFE's of the Google-variable-augmented BEM in (1),(2a),(3a) and (3b) relative to the benchmark BEM in (1)-(3). For the rest see Table 7.

Table 10: Augmented BEM II, Survey & Google Data, 2013:M7-2016:M12, GDP Components & Monthly Indicators

		Mining		Manufact.		Energy & Water		Constr.		Trade		Traffic		Hotel Ind.		Net Taxes	
		S-Cat	PLS	S-Cat	PLS	S-Cat	PLS	S-Cat	PLS	S-Cat	PLS	S-Cat	PLS	S-Cat	PLS	S-Cat	PLS
Forecast Horizon	17	0.99	0.94	0.98	0.94	1.00	1.00	1.05	1.09	1.04	0.98	0.99	1.05	1.04	0.99	1.03	1.00
	13	0.99	0.97	0.98	0.96	1.00	0.98	1.03	1.08	1.00	0.99	0.98	1.01	1.01	0.96	1.04	1.01
	9	0.99	1.02	0.99	0.88	1.00	1.01	1.04	0.98	1.02	1.17	1.01	1.02	1.02	0.99	1.04	1.01
	5	0.99	1.05	1.02	0.87	1.00	1.08	1.07	1.11	0.98	1.16	1.01	1.00	1.01	1.00	1.07	1.02
	1	1.03	1.10	0.98	0.97	1.01	1.02	0.99	1.12	1.04	1.10	1.00	0.99	1.02	1.09	1.07	1.06
	-3	1.03	1.06	0.99	1.00	1.00	0.99	1.01	1.07	1.06	1.11	1.00	1.00	1.03	1.09	1.04	1.03

		Prod. Mining		Ind. Prod.		Energy Prod.		Prod. Constr.		Retail Sales		Toll		Sales Hotel Ind.		VAT	
		S-Cat	PLS	S-Cat	PLS	S-Cat	PLS	S-Cat	PLS	S-Cat	PLS	S-Cat	PLS	S-Cat	PLS	S-Cat	PLS
Forecast Horizon	17	1.03	1.03	0.99	0.87	1.01	0.94	1.10	1.21	0.97	0.99	1.04	0.95	0.99	1.00	0.94	0.96
	15	1.00	0.96	0.98	0.88	1.01	0.96	1.09	1.26	0.97	0.98	1.05	0.96	1.00	0.98	0.98	0.95
	13	1.00	0.97	0.99	0.86	1.00	1.00	1.05	1.14	0.96	0.99	1.03	0.94	0.99	0.96	0.97	0.92
	11	1.01	0.98	0.98	0.90	1.01	1.02	1.00	1.02	0.97	0.98	1.03	0.92	0.99	0.97	1.01	0.95
	9	1.01	1.00	0.99	0.91	1.01	1.04	0.99	1.00	0.96	0.99	1.04	0.98	0.98	1.00	1.01	0.97
	7	1.03	0.98	0.99	0.88	1.01	1.01	1.02	1.07	0.97	0.98	1.05	0.98	0.98	0.99	1.00	0.98
	5	1.01	0.99	1.00	0.86	1.04	1.07	1.02	1.05	0.96	1.00	1.08	1.00	0.99	0.92	1.03	1.02
	3	0.99	0.98	0.99	0.93	1.03	1.07	1.01	1.08	0.94	0.93	1.09	1.01	0.99	0.89	1.04	1.04
	1	1.00	1.09	1.00	0.99	1.05	1.13	1.02	1.02	0.96	1.01	1.16	1.25	0.99	0.89	1.10	1.21
	-1	1.03	1.13	0.99	1.01	1.09	1.23	0.98	1.04	0.96	1.03	1.17	1.25	0.99	0.89	1.09	1.18
	-3	1.03	1.13	0.99	1.01	1.09	1.23	0.98	1.04	0.96	1.03			1.03	0.98	1.09	1.18
	-5									1.00	1.10			1.02	0.98	1.05	1.11
	-7									1.00	1.10					1.05	1.11

Note: For the monthly indicators the forecast horizons -7 to -1 correspond to backcasts, 1 and 3 to nowcasts and 5 to 17 to forecasts. S-Cat – Subj.-Cat. For the rest see Table 9.

Table 11: Augmented BEM III, Google Data, 2013:M7-2016:M12, GDP Growth

Method		Subj.-Cat.	Subj.-Subcat.	Google Corr.	PCA-Cat	PCA	PLS-Cat	PLS	LASSO	AdaLASSO	Boosting
Forecast Horizon	17	0.82	0.91	0.83	0.84	0.78	0.72	0.71	0.74	0.74	0.78
	15	0.83	0.91	0.84	0.86	0.81	0.73	0.74	0.74	0.74	0.77
	13	0.90	0.98	0.91	1.02	0.91	0.85	0.89	0.78	0.78	0.85
	11	0.85	0.93	0.86	0.96	0.88	0.84	0.88	0.81	0.81	0.85
	9	0.83	0.89	0.79	0.93	0.83	0.69	0.67	0.84	0.84	0.86
	7	0.73	0.78	0.71	0.80	0.73	0.57	0.54	0.83	0.83	0.80
	5	0.79	0.89	0.77	0.89	0.77	0.68	0.65	0.87	0.87	0.89
	3	0.99	1.24	1.09	1.18	1.01	1.12	1.13	1.02	1.02	1.12
	1	1.04	1.27	1.12	1.22	1.07	1.21	1.25	1.06	1.06	1.15
	-1	0.96	1.10	0.96	0.99	0.93	0.94	0.92	0.97	0.97	0.97
	-3	0.95	1.11	0.96	0.96	0.92	0.94	0.91	0.98	0.98	0.98
	-5	1.00	0.97	1.00	0.99	1.00	1.01	1.00	0.99	0.99	0.97

Note: The figures represent RMSFE's of the Google-variable-augmented BEM in (1),(2a*),(3a) and (3b) relative to the benchmark BEM in (1)-(3). For the rest see Table 7.

Table 12: Augmented BEM III, Google Data, 2013:M7-2016:M12, GDP Components & Monthly Indicators

		Mining		Manufact.		Energy & Water		Constr.		Trade		Traffic		Hotel Ind.		Net Taxes	
		PLS	Las.	PLS	Las.	PLS	Las.	PLS	Las.	PLS	Las.	PLS	Las.	PLS	Las.	PLS	Las.
Forecast Horizon	17	0.87	0.91	0.90	0.88	1.01	1.01	1.33	1.29	0.89	0.97	1.08	1.03	0.94	0.95	1.00	1.02
	13	0.92	0.93	1.14	1.00	0.99	1.00	1.29	1.17	0.93	1.00	1.02	0.97	0.90	0.93	1.02	1.02
	9	0.98	0.87	0.84	0.97	1.02	1.02	1.16	1.13	1.15	1.16	1.01	0.98	1.03	0.98	1.02	1.00
	5	1.04	1.01	0.68	1.01	1.07	1.01	1.26	1.21	1.15	1.10	1.01	1.02	1.01	0.94	1.05	1.05
	1	1.05	1.00	1.22	1.12	1.02	1.01	1.07	1.03	1.09	1.00	0.99	0.99	1.06	0.99	1.08	1.06
	-3	1.05	0.99	1.12	1.05	1.00	1.00	1.02	0.99	1.10	1.07	1.00	1.00	1.08	1.04	1.06	1.03

		Prod. Mining		Ind. Prod.		Energy Prod.		Prod. Constr.		Retail Sales		Toll		Sales Hotel Ind.		VAT	
		PLS	Las.	PLS	Las.	PLS	Las.	PLS	Las.	PLS	Las.	PLS	Las.	PLS	Las.	PLS	Las.
Forecast Horizon	17	0.95	0.90	0.81	0.80	0.91	0.91	1.40	1.30	0.97	0.99	0.91	1.00	1.02	0.98	0.91	0.91
	15	0.90	0.88	0.87	0.85	0.93	0.91	1.44	1.25	0.95	0.97	0.91	1.00	1.01	0.97	0.92	0.97
	13	0.91	0.88	0.80	0.81	0.96	0.92	1.32	1.19	0.98	0.99	0.90	0.95	0.99	0.95	0.87	0.93
	11	0.93	0.92	0.88	0.84	0.98	0.94	1.10	1.02	0.97	0.99	0.87	0.97	1.01	0.96	0.92	0.94
	9	0.95	0.92	0.80	0.82	1.00	0.95	1.09	1.00	0.99	1.00	0.94	0.95	0.98	0.98	0.92	0.95
	7	0.94	0.94	0.78	0.81	0.98	0.97	1.09	0.97	0.99	1.02	0.94	0.96	0.97	0.98	0.96	0.97
	5	0.95	0.92	0.72	0.81	1.02	0.96	1.11	0.95	1.01	1.01	0.97	1.00	0.83	0.95	0.99	0.98
	3	0.94	0.94	0.93	0.90	1.03	0.95	1.10	0.96	0.95	0.99	0.98	1.01	0.80	0.95	1.04	0.99
	1	1.06	0.97	0.94	0.89	1.11	0.98	1.04	1.00	1.04	1.02	1.15	1.12	0.85	0.87	1.22	1.01
	-1	1.07	1.01	0.99	0.89	1.18	1.03	1.05	1.01	1.07	1.04	1.16	1.14	0.86	0.87	1.19	0.98
	-3	1.06	1.01	0.99	0.89	1.18	1.03	1.05	1.01	1.06	1.04			1.01	0.87	1.19	0.98
	-5									1.12	1.10			1.01	0.88	1.12	1.00
	-7									1.12	1.10					1.12	1.00

Note: For the underlying augmented BEM see Table 11. Las. – LASSO. For the rest see Table 10.

Table 13: Augmented BEM II, Survey & Google Data, 2010:M1-2016:M12, GDP Growth

Method		Subj.-Cat.	Subj.-Subcat.	Google Corr.	PCA-Cat	PCA	PLS-Cat	PLS	LASSO	AdaLASSO	Boosting
Forecast Horizon	17	1.02	1.07	1.02	1.02	1.02	0.97	0.93	0.97	0.97	0.99
	15	0.99	1.06	1.02	1.01	1.01	0.89	0.89	0.95	0.95	0.98
	13	1.00	1.05	1.02	1.04	1.04	0.94	0.93	0.92	0.92	0.94
	11	0.98	1.03	1.00	1.03	1.00	0.89	1.01	0.99	0.99	0.98
	9	0.97	1.04	0.99	1.01	1.00	0.86	0.89	0.98	0.98	0.95
	7	0.98	1.09	0.99	1.01	1.02	0.85	0.87	1.00	1.00	0.95
	5	1.02	1.07	0.99	0.95	0.99	0.87	0.90	0.97	0.97	0.94
	3	0.93	1.13	1.00	0.93	0.96	0.83	0.85	0.95	0.95	0.99
	1	0.93	1.15	0.99	0.90	0.95	0.82	0.85	0.95	0.95	0.99
	-1	1.00	1.07	1.01	0.94	0.95	0.97	1.02	1.04	1.04	1.01
	-3	1.01	1.07	1.01	0.95	0.96	0.97	1.03	1.04	1.04	1.01
	-5	0.99	1.00	1.00	0.98	0.99	1.02	1.02	1.01	1.01	1.01

Note: The evaluation period is 2010:M1-2016:M12. For the rest see Table 9.

Table 14: Augmented BEM II, Survey & Google Data, 2010:M1-2016:M12, GDP Components & Monthly Indicators

		Mining		Manufact.		Energy & Water		Constr.		Trade		Traffic		Hotel Ind.		Net Taxes	
		PLS-Cat	PLS	PLS-Cat	PLS	PLS-Cat	PLS	PLS-Cat	PLS	PLS-Cat	PLS	PLS-Cat	PLS	PLS-Cat	PLS	PLS-Cat	PLS
Forecast Horizon	17	0.94	0.94	0.90	0.82	0.99	1.00	1.03	1.03	0.78	0.77	1.01	1.03	1.04	1.04	0.99	0.99
	13	1.01	1.01	0.84	0.82	0.98	0.98	1.02	1.03	0.83	0.83	1.03	1.04	1.05	1.09	0.99	0.99
	9	1.12	1.12	0.85	0.91	1.00	1.01	1.01	1.03	1.02	1.08	1.02	1.02	1.04	1.04	1.00	1.00
	5	1.02	1.02	0.91	0.97	1.03	1.04	0.95	1.03	0.99	1.04	1.01	1.00	1.06	1.03	1.04	1.08
	1	1.04	1.02	0.86	0.91	0.99	0.99	1.00	1.00	1.06	1.15	1.00	1.00	0.96	0.96	1.00	1.02
	-3	1.05	1.05	0.98	0.99	0.99	0.99	1.01	0.99	0.92	0.88	1.00	1.00	1.00	1.00	1.02	1.01

		Prod. Mining		Ind. Prod.		Energy Prod.		Prod. Constr.		Retail Sales		Toll		Sales Hotel Ind.		VAT	
		PLS-Cat	PLS	PLS-Cat	PLS	PLS-Cat	PLS	PLS-Cat	PLS	PLS-Cat	PLS	PLS-Cat	PLS	PLS-Cat	PLS	PLS-Cat	PLS
Forecast Horizon	17	0.94	0.96	0.89	0.90	1.00	1.02	1.07	1.15	1.12	1.07	0.93	0.92	0.96	0.96	0.97	1.04
	15	0.90	0.91	0.90	0.92	1.02	1.03	1.05	1.12	1.10	1.04	0.94	0.93	0.96	0.97	0.97	1.07
	13	0.91	0.93	0.88	0.89	1.04	1.06	1.03	1.11	1.17	1.12	0.90	0.90	0.96	0.94	0.95	1.03
	11	0.84	0.85	0.89	0.92	1.03	1.01	1.01	1.06	1.17	1.11	0.89	0.89	0.96	0.96	1.01	1.07
	9	0.87	0.90	0.89	0.91	1.03	1.05	1.02	1.09	1.22	1.19	0.92	0.92	0.97	0.99	0.99	1.01
	7	0.89	0.90	0.90	0.90	1.05	1.08	1.03	1.09	1.22	1.17	0.92	0.91	0.97	0.99	0.99	1.06
	5	0.96	1.00	0.93	0.93	1.09	1.10	1.01	1.08	1.21	1.21	0.87	0.87	1.04	1.05	0.97	1.05
	3	0.97	0.98	0.94	0.98	1.09	1.07	1.02	1.05	1.15	1.11	0.86	0.86	1.05	1.05	0.99	1.04
	1	1.06	1.11	1.05	1.06	1.08	1.11	1.04	1.07	1.20	1.22	0.99	1.03	0.98	1.00	1.04	1.10
	-1	1.08	1.09	1.14	1.11	1.09	1.10	1.07	1.07	1.15	1.19	0.96	1.01	0.97	1.00	1.09	1.19
	-3	1.08	1.09	1.14	1.11	1.09	1.10	1.07	1.07	1.15	1.18			0.99	0.98	1.10	1.19
	-5									1.08	1.14			1.00	1.00	1.08	1.15
	-7									1.08	1.14					1.08	1.15

Note: The evaluation period is 2010:M1-2016:M12. For the rest see Table 10.

Table 15: Augmented BEM III, Google Data, 2010:M1-2016:M12, GDP Growth

Method		Subj.-Cat.	Subj.-Subcat.	Google Corr.	PCA-Cat	PCA	PLS-Cat	PLS	LASSO	AdaLASSO	Boosting
Forecast Horizon	17	0.82	0.96	0.92	0.93	0.91	0.86	0.88	0.80	0.80	0.89
	15	0.74	0.94	0.80	0.80	0.79	0.80	0.83	0.75	0.75	0.78
	13	0.81	1.21	0.90	0.92	0.86	0.89	0.95	0.73	0.73	0.83
	11	1.02	1.16	1.06	1.05	1.03	1.01	1.04	0.98	0.98	1.02
	9	0.88	0.99	0.96	0.96	0.91	0.83	0.78	0.97	0.97	0.90
	7	0.75	0.99	0.82	0.82	0.77	0.72	0.72	0.93	0.93	0.83
	5	0.85	1.03	0.93	0.89	0.86	0.75	0.68	1.01	1.01	0.94
	3	0.98	1.27	1.14	1.05	1.08	0.94	0.88	1.06	1.06	1.12
	1	1.01	1.35	1.17	1.05	1.08	0.91	0.87	1.07	1.07	1.12
	-1	1.08	1.17	1.06	1.00	1.02	1.00	0.96	1.12	1.12	1.04
	-3	1.08	1.16	1.06	1.02	1.03	1.00	0.96	1.12	1.12	1.03
	-5	1.00	1.01	1.01	0.98	0.99	1.02	1.03	1.02	1.02	1.01

Note: The evaluation period is 2010:M1-2016:M12. For the rest see Table 11.

Table 16: Augmented BEM III, Google Data, 2010:M1-2016:M12, GDP Components & Monthly Indicators

		Mining		Manufact.		Energy & Water		Constr.		Trade		Traffic		Hotel Ind.		Net Taxes	
		PLS-Cat	PLS	PLS-Cat	PLS	PLS-Cat	PLS	PLS-Cat	PLS	PLS-Cat	PLS	PLS-Cat	PLS	PLS-Cat	PLS	PLS-Cat	PLS
Forecast Horizon	17	0.96	0.96	0.77	0.79	1.02	1.02	1.15	1.15	0.88	0.85	1.02	1.02	1.07	1.07	0.99	0.98
	13	1.05	1.03	0.87	0.90	1.01	1.00	1.15	1.17	0.90	0.86	1.01	1.02	1.08	1.11	0.99	0.99
	9	1.10	1.10	0.87	0.78	1.01	1.01	1.14	1.14	0.92	0.97	1.05	1.06	1.07	1.08	0.99	1.00
	5	1.02	1.04	0.83	0.73	1.04	1.04	1.00	1.05	0.92	0.98	1.03	1.04	1.11	1.11	1.05	1.06
	1	1.03	1.00	0.89	0.89	1.00	1.00	1.07	1.06	1.00	1.04	1.01	1.00	1.00	1.00	1.02	1.01
	-3	1.03	1.03	1.01	0.98	1.00	1.00	1.08	1.07	0.91	0.88	0.99	0.99	1.01	0.99	1.03	1.00

		Prod. Mining		Ind. Prod.		Energy Prod.		Prod. Constr.		Retail Sales		Toll		Sales Hotel Ind.		VAT	
		PLS-Cat	PLS	PLS-Cat	PLS	PLS-Cat	PLS	PLS-Cat	PLS	PLS-Cat	PLS	PLS-Cat	PLS	PLS-Cat	PLS	PLS-Cat	PLS
Forecast Horizon	17	0.97	0.96	1.01	0.99	0.96	1.00	1.28	1.29	0.99	1.06	0.98	0.99	0.97	0.95	0.88	0.95
	15	0.96	0.94	1.01	1.06	0.98	1.01	1.26	1.27	0.98	1.06	0.98	1.00	1.00	0.97	0.94	1.01
	13	0.93	0.92	0.98	0.98	0.98	1.04	1.28	1.29	1.02	1.10	0.92	0.94	0.99	0.97	0.91	0.96
	11	0.90	0.89	1.00	1.03	0.95	0.95	1.21	1.22	1.03	1.12	0.91	0.92	1.01	0.98	0.99	1.01
	9	0.91	0.90	0.98	0.95	0.94	0.97	1.21	1.23	1.04	1.16	0.95	0.97	0.99	0.99	0.94	0.94
	7	0.96	0.94	0.96	0.97	0.99	1.01	1.20	1.20	1.07	1.16	0.95	0.97	1.00	0.99	0.97	1.04
	5	0.98	0.96	0.99	0.94	0.95	1.00	1.19	1.20	1.04	1.13	0.91	0.91	0.91	0.97	0.92	1.00
	3	1.00	0.97	1.05	1.04	1.00	1.03	1.18	1.16	1.04	1.08	0.91	0.91	0.91	0.94	0.97	1.01
	1	1.05	1.04	1.12	1.04	0.94	1.03	1.15	1.09	1.06	1.13	0.98	1.00	0.96	0.98	1.00	1.07
	-1	1.06	1.05	1.12	1.08	1.04	1.05	1.05	1.01	1.10	1.17	0.98	1.00	1.00	1.00	1.05	1.14
	-3	1.05	1.04	1.12	1.08	1.03	1.05	1.04	1.01	1.10	1.15			1.01	1.04	1.05	1.14
	-5									1.09	1.14			1.02	1.05	1.04	1.12
	-7									1.09	1.13					1.04	1.12

Note: The evaluation period is 2010:M1-2016:M12. For the rest see Table 12.

Table 17: Augmented BEM II, Survey & Google Data, 2013:M7-2016:M12, GDP Growth, Rolling window

Method		Subj.-Cat.	Subj.-Subcat.	Google Corr.	PCA-Cat	PCA	PLS-Cat	PLS	LASSO	AdaLASSO	Boosting
Forecast Horizon	17	1.09	1.15	1.14	1.10	1.08	1.09	0.97	0.98	0.98	1.10
	15	1.10	1.17	1.16	1.11	1.11	1.11	0.99	1.06	1.06	1.12
	13	1.05	1.11	1.12	1.10	1.08	1.05	0.98	1.01	1.01	1.09
	11	1.02	1.13	1.12	1.08	1.08	1.03	1.00	1.01	1.01	1.12
	9	1.10	1.18	1.15	1.13	1.13	1.00	0.83	1.11	1.11	1.11
	7	1.06	1.05	1.11	1.11	1.13	0.93	0.81	1.09	1.09	1.03
	5	0.97	1.05	1.00	1.09	1.08	0.86	0.76	1.07	1.08	0.97
	3	1.00	1.26	1.07	1.15	1.06	1.09	1.04	1.04	1.04	1.02
	1	1.00	1.27	1.08	1.16	1.07	1.10	1.12	1.00	1.00	1.01
	-1	1.12	1.15	1.08	1.06	1.05	1.06	1.06	0.95	0.94	1.00
	-3	1.12	1.15	1.08	1.07	1.03	1.05	1.05	0.95	0.94	1.01
	-5	1.03	1.00	1.00	1.02	1.02	1.05	0.98	1.01	1.00	1.01

Note: Estimation is undertaken using a rolling window of six years. For the rest see Table 9.

Table 18: Augmented BEM II, Survey & Google Data, 2013:M7-2016:M12, GDP Components & Monthly Indicators, Rolling window

		Mining		Manufact.		Energy & Water		Constr.		Trade		Traffic		Hotel Ind.		Net Taxes	
		PLS	Las.	PLS	Las.	PLS	Las.	PLS	Las.	PLS	Las.	PLS	Las.	PLS	Las.	PLS	Las.
Forecast Horizon	17	0.88	0.97	0.96	0.97	1.05	1.07	1.09	1.01	1.14	0.92	1.07	1.06	1.03	0.98	1.02	1.01
	13	0.91	1.09	0.95	0.95	0.99	1.09	1.08	1.00	1.23	1.06	1.03	1.00	0.95	0.99	0.94	1.04
	9	0.91	1.05	0.75	1.06	1.09	1.10	1.05	1.01	1.34	1.29	1.03	1.02	1.06	1.01	0.90	1.00
	5	0.98	1.30	0.67	1.12	1.10	1.05	1.08	1.06	1.17	1.06	1.01	1.01	0.90	0.97	0.82	0.95
	1	1.10	1.20	0.87	1.02	1.06	1.03	1.21	1.06	1.29	1.01	1.00	1.00	1.00	1.02	0.97	0.91
	-3	1.00	1.01	1.00	0.92	1.03	1.01	1.04	0.99	1.43	1.34	1.01	1.01	1.02	1.05	0.92	0.92

		Prod. Mining		Ind. Prod.		Energy Prod.		Prod. Constr.		Retail Sales		Toll		Sales Hotel Ind.		VAT	
		PLS	Las.	PLS	Las.	PLS	Las.	PLS	Las.	PLS	Las.	PLS	Las.	PLS	Las.	PLS	Las.
Forecast Horizon	17	0.92	0.98	0.81	1.00	1.02	1.07	1.34	1.01	0.97	1.07	0.95	1.07	1.06	0.90	1.14	0.89
	15	0.90	0.95	0.84	0.93	1.01	1.07	1.30	1.05	0.98	1.11	0.95	1.09	1.03	0.86	0.95	0.84
	13	0.86	1.04	0.86	0.98	1.04	1.09	1.26	0.98	0.99	1.08	0.93	1.05	1.04	0.95	0.87	0.83
	11	0.94	1.05	0.89	0.93	1.05	1.06	1.12	1.00	1.01	1.15	0.95	1.08	1.03	0.94	0.85	0.89
	9	0.93	1.16	0.89	0.99	1.10	1.13	1.15	1.01	0.99	1.10	0.93	1.08	1.00	0.98	0.91	0.88
	7	0.99	1.15	0.90	0.93	1.07	1.06	1.13	1.07	1.00	1.16	0.93	1.08	0.99	0.96	0.87	0.92
	5	0.92	1.13	0.86	0.99	1.12	1.13	1.13	1.04	1.01	1.12	0.90	1.07	0.90	0.94	0.96	0.92
	3	0.97	1.11	0.86	0.94	1.11	1.08	1.06	1.06	0.97	1.13	0.91	1.06	0.88	0.93	0.99	0.98
	1	0.96	1.14	0.93	1.00	1.27	1.22	1.26	1.07	0.95	1.07	1.10	1.25	0.87	0.92	1.35	1.06
	-1	1.15	1.26	0.92	0.96	1.26	1.29	1.25	1.14	0.90	1.08	1.09	1.23	0.87	0.90	1.41	1.05
	-3	1.17	1.30	0.92	0.99	1.20	1.23	1.28	1.11	0.91	1.00			0.90	0.98	1.34	1.02
	-5									0.92	0.99			0.90	1.00	1.28	1.08
	-7									0.91	1.03					1.30	1.12

Note: Estimation is undertaken using a rolling window of six years. Las. – LASSO. For the rest see Table 10.

Table 19: Augmented BEM III, Google Data, 2013:M7-2016:M12, GDP Growth, Rolling window

Method		Subj.-Cat.	Subj.-Subcat.	Google Corr.	PCA-Cat	PCA	PLS-Cat	PLS	LASSO	AdaLASSO	Boosting
Forecast Horizon	17	1.04	1.16	1.06	1.09	0.96	1.04	0.63	0.79	0.79	0.88
	15	1.05	1.13	1.10	1.12	1.01	1.06	0.62	0.81	0.81	0.83
	13	1.14	1.21	1.16	1.29	1.13	1.15	0.74	0.93	0.93	0.89
	11	1.09	1.32	1.18	1.20	1.12	0.95	0.84	0.97	0.97	1.09
	9	1.11	1.30	1.11	1.15	1.09	0.89	0.56	0.91	0.91	1.02
	7	0.76	0.80	0.78	0.84	0.84	0.70	0.58	0.68	0.68	0.69
	5	0.77	0.81	0.68	0.80	0.74	0.60	0.45	0.60	0.61	0.65
	3	0.95	1.16	1.00	1.08	0.92	1.07	0.94	0.88	0.88	0.92
	1	0.99	1.23	1.03	1.11	0.98	1.15	1.07	0.85	0.86	0.92
	-1	1.13	1.18	1.08	1.06	1.01	1.04	1.02	1.02	1.03	1.03
	-3	1.13	1.19	1.08	1.07	0.99	1.04	1.01	1.03	1.03	1.03
	-5	1.01	0.99	0.99	1.00	1.02	1.05	0.98	1.03	1.03	1.01

Note: Estimation is undertaken using a rolling window of six years. For the rest see Table 11.

Table 20: Augmented BEM III, Google Data, 2013:M7-2016:M12, GDP Components & Monthly Indicators, Rolling window

		Mining		Manufact.		Energy & Water		Constr.		Trade		Traffic		Hotel Ind.		Net Taxes	
		PLS	Las.	PLS	Las.	PLS	Las.	PLS	Las.	PLS	Las.	PLS	Las.	PLS	Las.	PLS	Las.
Forecast Horizon	17	0.76	0.77	0.82	0.98	1.06	1.07	1.42	1.39	1.00	1.09	1.08	1.07	0.98	1.00	1.09	1.03
	13	0.86	0.86	0.90	1.07	1.07	1.14	1.40	1.41	1.14	1.04	1.02	1.00	0.94	1.00	1.00	1.01
	9	0.77	0.80	0.73	0.96	1.17	1.13	1.30	1.44	1.33	1.11	1.02	1.01	1.09	1.01	0.89	1.02
	5	0.88	1.19	0.66	0.80	1.13	1.04	1.32	1.39	1.02	0.97	1.01	1.01	0.94	0.94	0.83	1.00
	1	0.98	1.09	1.02	1.13	1.05	1.04	1.13	1.06	1.17	1.10	1.00	1.00	1.01	1.00	0.99	0.95
	-3	0.98	0.97	1.04	0.98	1.03	1.03	0.99	0.96	1.48	1.58	1.01	1.01	1.02	1.05	0.97	1.04

		Prod. Mining		Ind. Prod.		Energy Prod.		Prod. Constr.		Retail Sales		Toll		Sales Hotel Ind.		VAT	
		PLS	Las.	PLS	Las.	PLS	Las.	PLS	Las.	PLS	Las.	PLS	Las.	PLS	Las.	PLS	Las.
Forecast Horizon	17	0.85	0.91	0.69	0.90	0.91	0.92	1.57	1.52	0.92	1.00	0.88	0.99	1.03	0.98	1.13	0.96
	15	0.82	0.86	0.75	0.90	0.92	0.95	1.66	1.50	0.90	0.96	0.88	1.01	1.00	0.95	0.97	0.92
	13	0.84	0.95	0.72	0.90	0.96	0.95	1.53	1.42	0.93	0.98	0.88	0.99	1.00	1.02	0.87	0.86
	11	0.87	0.91	0.81	0.92	0.98	0.96	1.33	1.27	0.92	0.97	0.89	1.02	0.99	1.01	0.86	0.88
	9	0.89	1.02	0.76	0.90	1.05	0.99	1.30	1.25	0.93	0.99	0.86	0.98	0.94	1.00	0.89	0.87
	7	0.90	0.95	0.72	0.85	1.00	0.97	1.26	1.18	0.92	0.97	0.85	0.99	0.93	0.98	0.85	0.90
	5	0.86	1.00	0.69	0.89	1.05	0.97	1.27	1.18	0.96	0.99	0.90	1.03	0.84	0.95	0.91	0.89
	3	0.87	0.97	0.76	0.88	1.05	0.98	1.14	1.10	0.89	0.95	0.90	1.03	0.82	0.94	0.97	1.00
	1	0.87	1.06	0.83	0.90	1.21	1.05	1.24	1.33	0.92	0.95	1.07	1.21	0.82	0.88	1.32	1.05
	-1	0.96	1.13	0.93	0.90	1.23	1.21	1.21	1.25	0.91	0.87	1.07	1.24	0.81	0.86	1.39	1.05
	-3	1.01	1.20	0.93	0.91	1.15	1.14	1.21	1.33	0.90	0.89			0.88	0.88	1.34	1.05
	-5									0.92	0.90			0.87	0.88	1.25	1.10
	-7									0.93	0.95					1.24	1.10

Note: Estimation is undertaken using a rolling window of six years. Las. – LASSO. For the rest see Table 12.

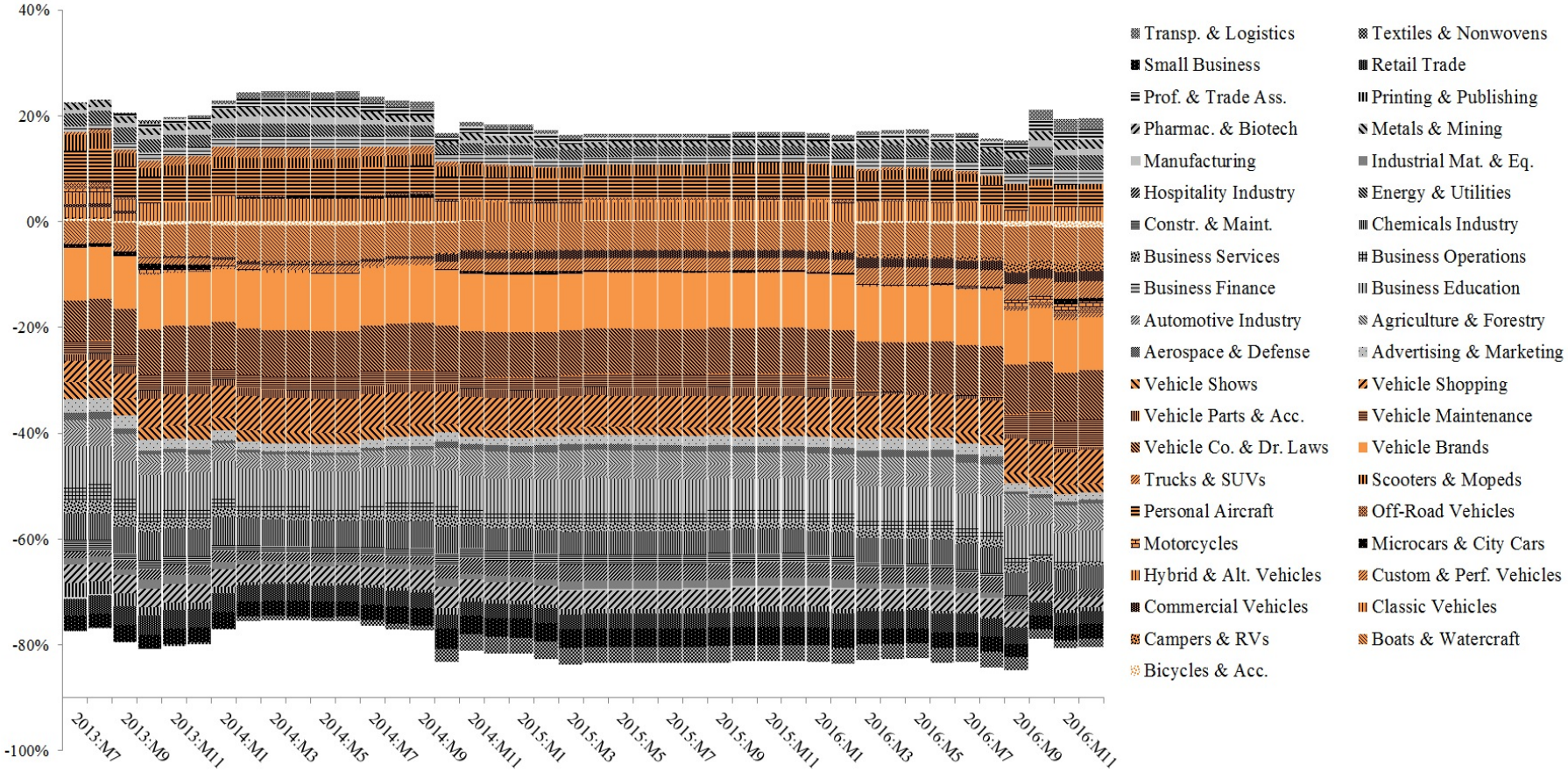
Table 21: Augmented BEM II*, Survey & Google Data, 2013:M7-2016:M12, GDP Growth & GDP Components & Monthly Indicators

			PLS							
		GDP Growth	Mining	Manufact.	Energy & Water	Constr.	Trade	Traffic	Hotel Ind.	Net Taxes
Forecast Horizon	17	0.66	0.90	0.75	1.02	1.11	0.84	1.06	0.94	1.02
	13	0.83	0.90	0.93	1.00	1.12	0.97	1.00	0.89	0.99
	9	0.73	0.89	0.74	1.06	1.04	1.26	0.99	0.96	1.00
	5	0.66	0.95	0.71	1.10	1.15	1.34	1.01	1.04	0.96
	1	1.12	1.02	1.10	1.02	1.12	1.24	0.99	1.07	1.05
	-3	0.87	1.00	1.01	0.99	1.10	1.06	1.00	1.12	0.99

			PLS							
			Prod. Mining	Ind. Prod.	Energy Prod.	Prod. Constr.	Retail Sales	Toll	Sales Hotel Ind.	VAT
Forecast Horizon	17		0.91	0.75	0.93	1.23	1.21	0.87	0.98	1.60
	15		0.87	0.76	0.94	1.34	1.20	0.88	0.96	1.48
	13		0.86	0.73	1.00	1.19	1.24	0.88	0.91	1.49
	11		0.90	0.79	1.01	1.07	1.23	0.88	0.92	1.34
	9		0.89	0.75	1.04	1.08	1.27	0.90	0.91	1.47
	7		0.92	0.71	0.99	1.14	1.27	0.91	0.91	1.38
	5		0.88	0.75	1.05	1.13	1.30	0.95	0.80	1.47
	3		0.91	0.89	1.11	1.13	1.18	0.95	0.76	1.37
	1		0.93	0.92	1.22	1.02	1.28	1.14	0.83	1.57
	-1		0.96	0.88	1.25	1.02	1.22	1.15	0.83	1.46
	-3		0.96	0.89	1.25	1.02	1.22		0.94	1.46
	-5						1.17		0.94	1.22
	-7						1.18			1.23

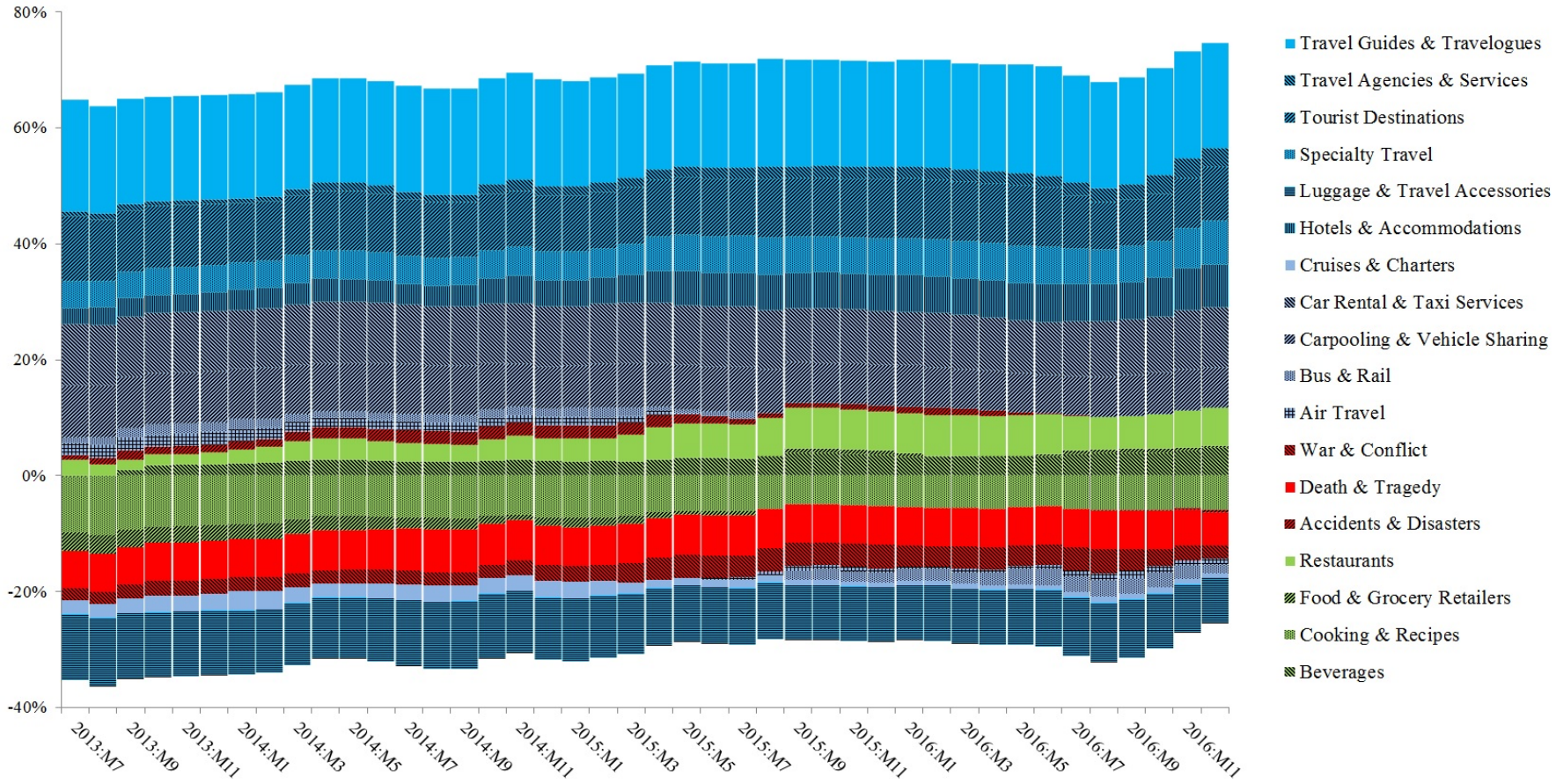
Note: For each monthly indicator in question, PLS factors are computed on the eligible subcategories and the survey indicator simultaneously. For the rest see Tables 9 and 10.

Figure 3: First PLS-factor Loadings, Industrial Production, Augmented BEM III, 2013:M7-2016:M12



Note: The graph is based on the the Google-variable-augmented BEM in (1),(2a*),(3a) and (3b). The loadings are scaled such that they add up to 100%. Orange-shaded loadings correspond to the subcategories of Autos & Vehicles, grey-shaded ones to those of Business & Industrial.

Figure 4: First PLS-factor Loadings, Sales Hotel Industry, Augmented BEM III, 2013:M7-2016:M12



Note: Green-shaded loadings correspond to the subcategories of Food & Drink, red-shaded ones to those of Sensitive Subjects and blue-shaded ones to those of Travel. For the rest see Figure 3.

Figure 5: Google Variable Selection by LASSO, Augmented BEM III, 2013:M7-2016:M12