Kholodilin, Konstantin A.; Podstawski, Maximilian; Siliverstovs, Boriss

Working Paper

Do Google searches help in nowcasting private consumption? A real-time evidence for the US

KOF Working Papers, No. 256

Provided in Cooperation with:
KOF Swiss Economic Institute, ETH Zurich

Suggested Citation: Kholodilin, Konstantin A.; Podstawski, Maximilian; Siliverstovs, Boriss (2010) : Do Google searches help in nowcasting private consumption? A real-time evidence for the US, KOF Working Papers, No. 256, ETH Zurich, KOF Swiss Economic Institute, Zurich, https://doi.org/10.3929/ethz-a-006070977

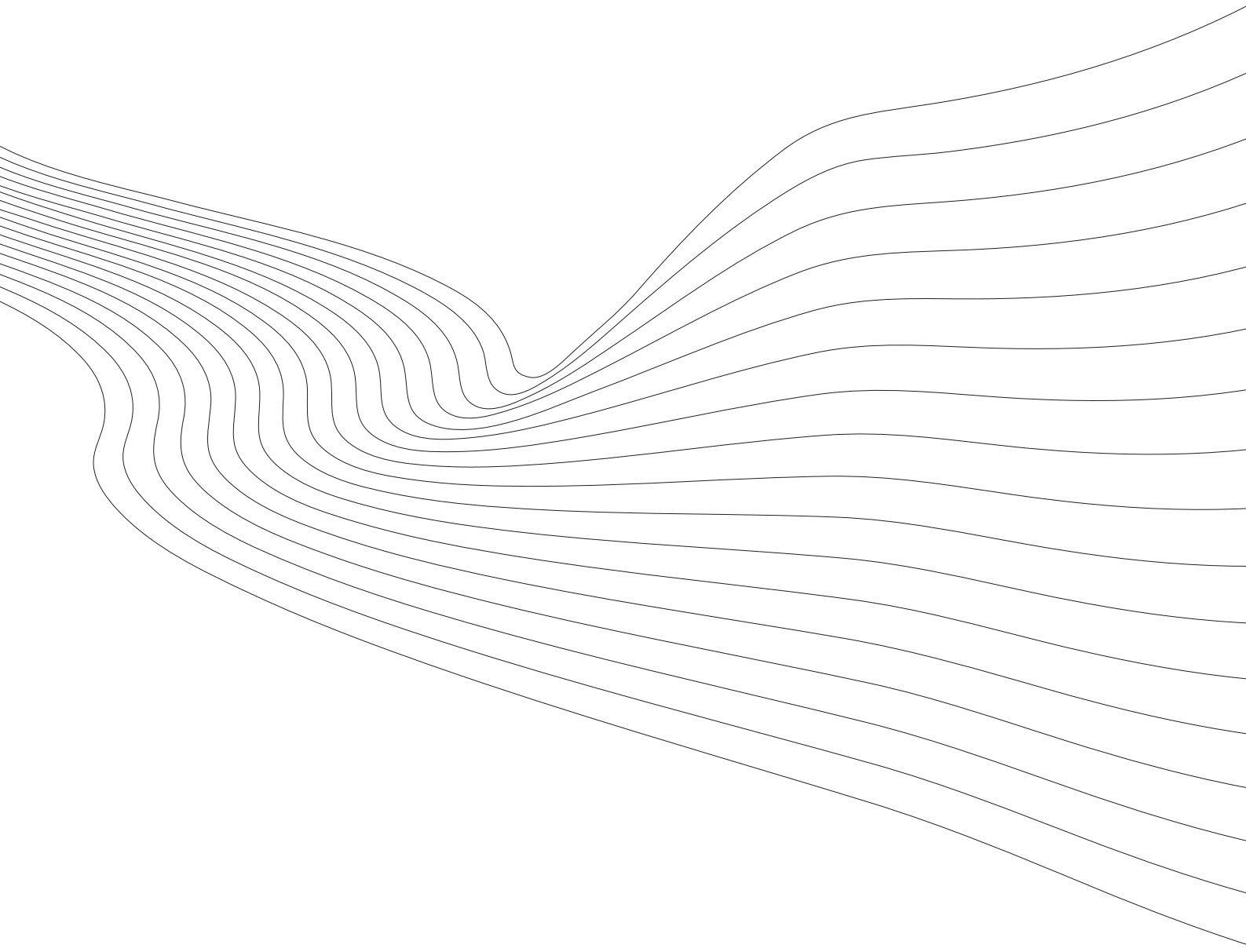
This Version is available at:
https://hdl.handle.net/10419/50431

KOF

KOF Working Papers

Do Google Searches Help in Nowcasting Private Consumption?
A Real-Time Evidence for the US

Konstantin A.Kholodilin, Maximilian Podstawski and Boriss Siliverstovs



No. 256
April 2010

ETH
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

KOF

ETH Zurich
KOF Swiss Economic Institute
WEH D 4
Weinbergstrasse 35
8092 Zurich
Switzerland

Phone +41 44 632 42 39
Fax +41 44 632 12 18
www.kof.ethz.ch
kof@kof.ethz.ch

Do Google Searches Help in Nowcasting Private Consumption?

A Real-Time Evidence for the US[‖]

Konstantin A. Kholodilin[*] Maximilian Podstawski[**] Boriss Siliverstovs[§]

Abstract

In this paper, we investigate whether the Google search activity can help in nowcasting the year-on-year growth rates of monthly US private consumption using a real-time data set. The Google-based forecasts are compared to those based on a benchmark AR(1) model and the models including the consumer surveys and financial indicators. According to the Diebold-Mariano test of equal predictive ability, the null hypothesis can be rejected suggesting that Google-based forecasts are significantly more accurate than those of the benchmark model. At the same time, the corresponding null hypothesis cannot be rejected for models with consumer surveys and financial variables. Moreover, when we apply the test of superior predictive ability (Hansen, 2005) that controls for possible data-snooping biases, we are able to reject the null hypothesis that the benchmark model is not inferior to any alternative model forecasts. Furthermore, the results of the model confidence set (MCS) procedure (Hansen et al., 2005) suggest that the autoregressive benchmark is not selected into a set of the best forecasting models. Apart from several Google-based models, the MCS contains also some models including survey-based indicators and financial variables. We conclude that Google searches do help improving the nowcasts of the private consumption in US.

Keywords: Google indicators; real-time nowcasting; principal components; US private consumption.

JEL classification: C22, C53, C82.

[‖]The earlier versions of this paper had been presented at a workshop "Neue Verfahren der Kurzfristprognose?", which took place in the German Ministry for Economy on July 30, 2009, at the Macroeconometric workshop at the DIW Berlin on December 18, 2009, as well as at the KOF Brown Bag Seminar at the ETH Zurich on March 29, 2010. We are grateful to the participants of the workshops for their insightful comments. Special thanks go to Vladimir Kuzin.

[*]DIW Berlin, Mohrenstrasse 58, 10117 Berlin, Germany, e-mail: kkholodilin@diw.de

[**]Universität Potsdam, Wirtschafts- und Sozialwissenschaftliche Fakultät, Potsdam, Germany, e-mail: mpod@gmx.net

[§]ETH Zurich, KOF Swiss Economic Institute, Weinbergstrasse 35, 8092 Zurich, Switzerland, e-mail: boriss.siliverstovs@kof.ethz.ch

1 Introduction

The pioneering study Ginsberg, Mohebbi, Patel, Brammer, Smolinski, and Brilliant (2009)—that appeared online in November 2008—showed how one can use disaggregated web searches filed by millions of users each day in order to study the intensity of influenza activity in the USA[1]. For this purpose Ginsberg et al. (2009) used web queries stored by Google Inc. that enabled them practically real-time influenza surveillance in areas with the large density of Internet connections. In a subsequent study that came out in April 2009, Choi and Varian (2009b) argued that web searches may not only be useful as a reliable indicator of the health-seeking behavior when facing the influenza pandemic but also they may contain a useful information for predicting the present stance of economic activity some time ahead of the official release of relevant data. Choi and Varian (2009b) provide examples of using the Google searches for in- as well as out-of-sample prediction of retail sales, automotive sales, and home sales for the US and of visitors arrivals in Hong Kong.

Following the suggestion of Choi and Varian (2009b), in the same year of 2009 several studies investigated the usefulness of Google searches for forecasting unemployment developments in various countries such as Askitas and Zimmermann (2009, appeared in June) for Germany, Suhoy (2009, appeared in July) for Israel, D'Amuri (2009, appeared in October) for Italy, as well as Choi and Varian (2009a) and D'Amuri and Marcucci (2009) for United States published in July and October, respectively.

Our paper intends further to contribute to the rapidly developing area of using Internet data for prediction of macroeconomic variables by investigating whether Google searches can help to forecast the private consumption in the USA. An additional interesting question that we would like to address is whether Google searches have any predictive content beyond that of the sentiment indicators and financial variables that typically are used for monitoring developments in private consumption (e.g., see Croushore, 2005, for a relevant discussion and a brief historical review).

Our further contribution to the literature constitutes the use of the real-time data set, i.e., for every point of time, we utilized data vintages of private consumption that correspond to the available information. The importance of using real-time data instead of latest-available data has been emphasized in numerous studies

[1] In September 2009, Doornik (2009) demonstrated that the application of more sophisticated econometric models than the "crude statistical methods" employed in Ginsberg et al. (2009) to the Google search activity can dramatically improve the forecasts of the flu activity.

as it was shown, for example, by Diebold and Rudebusch (1991) and, more recently, by Croushore (2005) that the favorable conclusions on forecasting properties of leading indicators obtained using latest-available data may be substantially weakened or even reversed when forecasting exercise is replicated using real-time data sets. In this respect, our study distinguishes itself from the aforementioned studies on using Google searches for unemployment forecasting, where it is not clearly articulated how data revisions are treated.

Last but not least, in comparing predictive accuracy of the models with Google indicators against that of the competing models we go beyond the pairwise model comparison and employ both the tests for superior predictive ability and the tests based on the model confidence set introduced in Hansen (2005) and Hansen et al. (2005), respectively. These two approaches address complications arising when comparing multiple models related to difficulties of controlling the size of tests, which typically leads to spurious results.

Based on the pairwise model comparison we find a statistically significant evidence that models with Google indicators do offer an improvement in forecast accuracy with respect to the benchmark model. At the same time, according to the Diebold-Mariano test, for the models including sentiment indicators and financial variables we cannot reject the null hypothesis of equal predictive ability against the benchmark model. Being aware of the possible erroneous conclusions that can be reached when comparing multiple models, we cross-checked these encouraging results with the tests specifically developed to deal with possible data-snooping biases. According to the Superior Predictive Ability (SPA) test of Hansen (2005), we are able to reject at 10% significance level the null hypothesis that the benchmark model is not inferior to any alternative forecasts based on the leading indicators. This finding is confirmed by the Model Confidence Set (MCS) test of Hansen et al. (2005), which suggests that the benchmark AR model is not selected into a set of the best forecasting models. Moreover, apart from the Google-based models, the MCS contains also some models including the survey-based indicators and financial variables. This implies that the information content of the Google indicators as well as survey-based indicators and certain financial variables can help improving the nowcasts.

The paper is structured as follows. Section 2 describes the data used. Section 3 discusses the construction of our Google indicators. Sections 4 and 5 describe the forecasting models as well as forecast accuracy evaluation methods. Section 6 discusses the empirical results. The final section concludes.

2 Data

The variable to be nowcast in this study is the year-on-year growth rate of monthly US real private consumption. The source for monthly US private consumption real-time data is the ALFRED® database of the Federal Reserve Bank of St. Louis (http://alfred.stlouisfed.org/). ALFRED® provides vintages of economic data that were available on specific dates in history, which enables us to undertake real-time forecasts based on this historical data.

Three groups of leading indicators are used to forecast the US private consumption: 1) the conventional leading indicators based on consumer surveys; 2) several financial variables, that are typically used in the consumption forecasts; and 3) the data on Google searches.

Two conventional leading survey-based indicators available at monthly frequency are: 1) Consumer Sentiment Indicator produced by the University of Michigan and 2) Consumer Confidence Index constructed by the US Conference Board. The financial variables include: 1) 3-month US Treasury constant maturity rate, TBILL3M; 2) 10-year US Treasury constant maturity rate, USBOND10Y; 3) the yield spread, US10Ym3M (=USBOND10Y-TBILL3M); and 3) Standard and Poor's index of 500 large firms, S&P500. The data are taken from the Datastream.

Time series data on Google searches are available at weekly frequency from January 2004 onwards (http://www.google.com/insights/search/). Figure 1 depicts the interface of Google Insights. The Google data are not subject to any revisions. Google normalizes each time series by dividing the count for each query by the total number of online search queries submitted during the week, which results in a query fraction. A query fraction for the search query q is equivalent to the probability that a random search query submitted from a particular region at a particular time is exactly q (see Ginsberg et al., 2009).

Unlike Ginsberg et al. (2009) who had hundreds of computers at their disposal, our computational resources are quite limited. Therefore, we had to take a relatively small selection of the billions of Google searches. The data set used to predict US private consumption was built in two steps. In a first step, a pool of search items provided by the top ten searches of each of the 27 main Google search categories and respective subcategories was collected. Then, we intend to further eliminate queries that are not related to private consumption before fitting any models. Therefore, in the second step, we select only the search items from the entire pool of all

top ten queries in the categories provided by Google Insights that have enough variability and are economically relevant from the viewpoint of private consumption. Following this algorithm we were able to construct a data set comprising of 220 consumption-relevant Google searches.

3 Construction of Google indicators

Since Google Insights provides data at weekly frequency, while US private consumption data are published at monthly frequency, the Google searches time series have to be aggregated. Due to the overlaps of weeks and months the data are interpolated to daily frequency by applying a spline methodology as a first step and subsequently aggregated to monthly frequency. Many Google search time series show a distinct seasonal pattern and thus make a seasonal adjustment necessary. Therefore, we transform each of the 220 selected time series into monthly year-on-year growth rates, thus sacrifying the first 12 observations of the already short time series. This means that the effective sample that is available to us covers the period from January 2005 until December 2009.

Given the large amount of the Google searches data we have collected, we need to reduce their dimensionality to a small and yet relevant number of regressors to be used in the forecasting the growth rate of US private consumption. This is achieved by applying the factor model of Stock and Watson (1999) and Stock and Watson (2002), which is based on the method of principal components (PC) and which allows extracting a reduced number of common factors from the selected 220 time series. Since we conduct our exercise in a real-time framework, the principal components have been extracted recursively from the sample available at the time a nowcast is made such that they can be matched with the corresponding real-time vintages of private consumption.

In assessing the forecasting ability of the models with Google indicators we had to make a choice of how many principal components need to be included. Based on the information from the whole available sample (2005M1—2009M12) the selection criterion of Bai and Ng (2002) retains only three first principal components. The corresponding contributions of the first 25 out of 220 principal components to the total variance are shown in Figure 2. It can be seen that the first three principal components account for roughly half of the variance. After the fifth principal component the variance contributions start to decline relatively gradually. After the

first ten principal components the contributions of the remaining principal components turn so small, that they can be safely ignored. However, instead of relying on the in-sample selection rule in determining the number of relevant principal components, we chose to work with the first ten principal components due to the following reasons. First, the whole sample information was clearly not available to a forecaster at some earlier point of time. Secondly, by focusing on the first ten principal components we work with a substantially large model space, which enables us to identify those principal components that help in predicting private consumption in a genuine out-of-sample nowcasting exercise.

Figure 3 shows the first ten principal components extracted from the monthly year-on-year growth rates of Google searches. It can be seen that the first principal component is clearly upward trending. The second and fourth principal components seem to capture the current recession, whereas no distinct pattern can be observed in the case of the remaining components.

4 Nowcasting models

Given a rather limited length of the time series of Google indicators covering five full years for the nowcasting purposes, we employ a parsimonious model in the following form:

$$y_t = \alpha + \beta y_{t-1} + \gamma x_t + \varepsilon_t, \tag{1}$$

where y_t is the year-on-year growth rate of the monthly real private consumption; x_t is an exogenous variable representing a leading indicator or a financial variable; and ε_t is the error term.

For the conventional sentiment indicator models, x_t is either the levels of the University of Michigan's Consumer Sentiment Index and the Conference Board's Consumer Confidence Index or the corresponding annual differences of these indicators. Thus, there are four different models based on the sentiment indicators.

For the financial models, x_t is either the levels of the short- and long-term interest rates, their yield spread, and Standard and Poor's 500 index, or the annual differences of these indicators. Notice that to the Standard and Poor's 500 index the logarithm was applied before differencing. In addition, a forecast combination of the three financial models including the 12th order differences of short-term interest rate, D12TBILL3M, long-term

interest rate, D12USBOND10Y, and Standard and Poor's 500 index, D12LSP500, was constructed as a simple average. Thus, the total number of the financial models is eight.

For Google searches, x_t is either one of the ten first principle components (10 models) or one of all possible combinations of pairs of the first ten principle components (45 models). Furthermore, we also report the results of a nowcasting exercise for averages of nowcasts based on the single-indicator models involving 2, 3,..., and up to 10 first principal components (9 models). We repeat the model averaging exercise also for models with pairs of principal components; we compute averages of nowcasts across all the pair combinations based on three up to ten principal components (8 models). Thus, the total number of models with Google-based indicators is 72. We have chosen to report the nowcasts based on model averaging, since the simple averaging of point forecasts is repeatedly found in the forecasting literature to outperform more elaborate forecast combination schemes; a finding that prompted Watson and Stock (2004, p. 428) to refer to it as a "forecast combination puzzle". Moreover, since the conclusions based on a superior performance of a single model may be sample dependent, by considering averages of model nowcasts we robustify our findings on the usefulness of Google indicators for predicting the economic variable of interest.

The benchmark model without any of the indicators is the first order autoregressive model:

$$y_t = \alpha + \beta y_{t-1} + \varepsilon_t. \tag{2}$$

We estimate the parameters of models given in equations (1) and (2) using a 24-month rolling window rather than an expanding estimation window for the following reasons. Firstly, we found that the forecast accuracy of the models estimated using an expanding rather than rolling window is uniformly inferior. This well corroborates an argument of Giacomini and White (2006, p. 1547) that models with limited memory can better track a time series of interest in the presence of (unmodeled) structural changes. In such situations the observations from a distant past are likely to lose their predictive relevance. Secondly, observe that the models given in equations (1) and (2) are nested. Hence, the application of the Diebold-Mariano test for equal predictive ability for comparison of nested models, whose parameters are estimated using the expanding window, is problematic, as it has been pointed out in Clark and McCracken (2001). On the contrary, Giacomini and White (2006) argue that as long as one keeps the size of the estimation window fixed, the Diebold-Mariano test

can be straightforwardly conducted. In addition, the tests for the superior forecasting ability as well as the test based on the model confidence set suggested in Hansen (2005) and Hansen et al. (2005), respectively, are based on the rolling estimation window.

The models given in equations (1) and (2) are used to make real-time out-of-sample nowcasts, i.e., forecasts for the current month using only information available up to that month. In other words, at every nowcast round we use the real-time vintages of private consumption as well as the corresponding recursively extracted principal components. After data transformation the whole available period is from 2005M1 until 2009M12. We have chosen the size of rolling window of 24 months as a compromise between choosing a shorter estimation window (e.g., 12 months) that would imply an estimation of a model with up to four parameters using only 12 observations and a longer estimation window (e.g., 36 months) that would imply a shortening of the forecast sample size to 24 observations. The whole available period is split into an initial estimation subperiod (2005M1-2006M12) and forecasting subperiod (2007M1-2009M12). After the nowcast for 2007M1 is made, the estimation period is moved forward by one observation— 2005M2-2007M1 —and the next nowcast is made for 2007M2, etc.

5 Nowcast evaluation

The nowcast accuracy of different models is evaluated using the root mean squared forecast error (RMSFE) measure:

$$RMSFE_l = \sqrt{\frac{1}{T}\sum_{t=1}^{T}(e_t^l)^2}, \tag{3}$$

where l is the l-th vintage of the data. Usually, the forecasts are compared to the so-called final, or last-revision, data, that is, $l = L > 1$. In some cases, L can be 24 months. In contrast, we assess the accuracy of the nowcasts with respect to the flash estimate, i.e., the first vintage, or release, $l = 1$, of the private consumption data. The use of the flash estimates as a reference series in our case can be justified as follows: (1) in the real-world setting the first release of data is the most relevant one for forecasters and policy-makers and (2) based on the analysis of the monthly time series of the US private consumption we could determine that the data are effectively revised until up to 23 months after the first publication, which would further shrink our already small sample.

In addition, the notion of the "final" revision, which used to be applied to the revision occurring three months after the current quarter's end, proved to be so misleading that it had been abandoned in 2009 by the Bureau of Economic Analysis, as communicated to us by an BEA economist Lisa Mataloni: "BEA recently stopped using the term "FINAL" to describe this estimate to avoid the implication that the estimates would not be subject to further revision". For more details on data revisions see chapter 1 of Bureau of Economic Analysis (2009).

We use the following approaches in order to compare the forecasting accuracy of the models with Google indicators, sentiment indices, financial variables, and of the univariate benchmark model. In the first place, we use the popular tests for equal predictive ability suggested in Diebold and Mariano (1995) based on the pairwise comparison of nowcast accuracy of competing models. We also compute the tests for nowcast encompassing in the form suggested by Harvey et al. (1998), which are also based on pairwise comparison of multiple models. In a second step, we employ the tests for superior predictive ability (SPA) as well as the tests based on the model confidence set (MCS) introduced in Hansen (2005) and Hansen et al. (2005), respectively, which allow us to control for possible data-mining biases. In the situations when multiple models are compared, it is very likely that a particular model will appear to be better than other models and the more models are being compared the more likely that some models will display a better predictive performance by chance. The tests suggested in Hansen (2005) and Hansen et al. (2005) allow us to control for that in a statistical framework.

6 Results

Table 1 compares the nowcast accuracy of the models examined in this paper. The first column contains the RMSFEs of the respective nowcast models, whereas the second column represents the RMSFE of the benchmark model (AR(1) process), to which all other models are compared. Column (3) reports the relative RMSFEs, which are obtained by dividing the respective model based on leading indicators, RMSFE(1), over that of the benchmark model, RMSFE(2). Column (4) contains the "p-values" of the Equal Predictive Ability (EPA) test[2], which is a Diebold-Mariano test based on bootstrap with a number of resamples, $B = 10000$. In fact, these "p-values" are "naive" p-values that compare the benchmark model to the best alternative model, ignoring the fact that the latter was chosen from a large set of models. Columns (5) and (6) report the p-values

[2]The test was conducted using an Ox package MulCom 1.0 for Ox written by P. R. Hansen and A. Lunde.

of the encompassing tests: in column (5) the null hypothesis of the corresponding model based on leading indicators encompassing the benchmark model is tested, whereas in column (6) the null of the benchmark model encompassing the respective model based on leading indicators is tested. Finally, column (7) compares the RMSFE of the expanding-window models to that of the rolling-window models.

The first four rows of Table 1 include the models based on the conventional sentiment indicators: the levels and 12th order differences of the Consumer Sentiment Indicator of the University of Michigan and Consumer Confidence Index of the Conference Board. The next eight rows present the nowcasts based on the levels and 12th order differences of the short- and long-term interest rates, yield spread as well as monthly and annual log-difference of the Standard and Poor's 500 index, representing monthly and annual stock returns. Observe that the nowcasts obtained by model averaging of the models including 12th order difference of the short- and long-term interest rates as well as the annual stock returns is labeled as FINCOMB. The following ten rows report the results obtained for the models including one of the first 10 principal components extracted from the Google-search series. The models from c(PC1,PC2) through c(PC9,PC10) are the models, in which the respective pair of principal components are included as regressors. The rows from Single.PC2 to Single.PC10 represent combinations of nowcasts based on the models containing the single principal components. The combined nowcasts are constructed by simple averaging. Single.PC2, for example, is the average of the nowcasts based on the model including PC1 and model involving PC2, and hence Single PC.10 averages the nowcasts based on the models containing single principle components from the first up to the tenth component. The rest of the rows in the table ought to be understood accordingly, with the only difference being that here nowcasts obtained by models, which include two principal components simultaneously, are averaged.

Several observations can be made on the basis of the results summarized in the table. Firstly, as column (2) shows, the nowcasts based on the sentiment indicators produce no or only small nowcast accuracy gains compared to the benchmark model. Notice that the models based on the 12th order differences of the sentiment indicators perform somewhat better than the models containing their levels. The best sentiment-indicator-based nowcast, D12CBCCI, in terms of the RMSFE is just 9% more accurate than the benchmark nowcast. However, according to the Diebold-Mariano test for equal forecast accuracy we cannot reject the corresponding null hypothesis for any of the sentiment indicators or their transformation. The encompassing tests indicate

that only the D12UMCSI- and D12CBCCI-based models encompasse the benchmark model without being encompassed by it. Thus, based on the model pairwise comparison we conclude that there is at best only a weak evidence that the conventional sentiment indicators possess any predictive power for private consumption growth in the USA; a result that is consistent with the findings of Croushore (2005), who had shown that the levels of sentiment indicators are not able to add any additional information to the nowcast of the consumption aggregate.

Secondly, the nowcasts based on the financial variables as well as combination of such nowcasts produce even smaller nowcast accuracy gain than the nowcasts based on the sentiment indicators. The best nowcasts based on financial variables, FINCOMB and D12LSP500, are 8% and 7.5%, respectively, more accurate than the nowcast based on the simple AR(1) model. According to the Diebold-Mariano test reported in column (4), for none of the financial models the null hypothesis of equal forecast accuracy can be rejected. However, the encompassing tests show that all financial nowcasts, except for those based on DLSP500 and USBOND10Y, encompass the benchmark model without being encompassed by it.

Thirdly, the best performing Google-based models are those including either PC2 or PC5, the best nowcast being that based on the model including both of them simultaneously, c(PC2,PC5). This nowcast outperforms the benchmark AR(1) process by 17%. According to the pairwise test for equal forecast accuracy for this model the null hypothesis can be rejected at the 5% significance level. The nowcast of c(PC2,PC5) model is plotted together with the benchmark nowcast and the actual values in Figure 4. It can be seen that the noticeable improvement over the benchmark model is attained during the recession period.

Fourthly, a further encouraging observation supporting the ability of Google-based indicators to predict private consumption in the USA is that the null hypothesis of equal forecast accuracy can be rejected for several models with either a single or a pair of principal components such as PC2, PC5, c(PC2,PC5), c(PC2,PC9), c(PC3,PC5), c(PC4,PC5). Please note that all of these models include either PC2 or PC5. For most of the nowcasts based on model averaging one can reject the null hypothesis of equal predictive ability with the benchmark model.

Fourthly, the results of the Diebold-Mariano tests are supported by the forecast encompassing tests of Harvey et al. (1998). As seen, the corresponding null hypothesis that nowcasts from each of the models PC2,

PC5, c(PC2,PC5), c(PC2,PC9), c(PC3,PC5), c(PC4,PC5) encompass those of the benchmark model cannot be rejected at the usual significance level, whereas the opposite hypothesis that these nowcasts are encompassed by those of the univariate autoregressive model can be rejected. This also holds for all nowcasts based on model averaging.

All in all, our results of the pairwise model comparison so far suggest that the models with Google indicators have some predicive ability beyond not only a benchmark univariate model but also models with the conventional sentiment indicators. In particular, we find that PC2 and PC5 are of informational value for predicting US private consumption. We also conclude that extending the information set by including the principal components beyond the fifth one does not result in further noticeable improvement over that achieved with the first five principal components.

It might be useful to understand how the PCs that allowed significantly improving the nowcasts of US real private consumption—PC2 and PC5—can be interpreted. One way to do this is to analyze their factor loadings. However, the factor loading do not provide a clear picture allowing an unambiguous interpretation of the principal components. To gain further insight in possible ways of interpretation of the principal components Table 4 reports the correlations of the first ten principal components with year-on-year growth rates of the disaggregated consumer spending components obtained from the BEA. Several observations can be made. Firstly, the correlation with the components of consumer spending fades out as the order of the principal components rises. Secondly, the first principal component is highly correlated with 11 out of 15 components of consumer spending. Thirdly, the second principal component captures the variance of *motor vehicles and parts*, *gasoline and other energy goods*, *transportation services*, *financial services and insurance* and *other services* well, clearly pointing towards an interpretation of the component capturing consumption activity related to mobility. It should be noted that *motor vehicles and parts*, *gasoline and other energy goods* and *transportation services* are the only consumer spending components that have a negative average growth over the time period 2005M1 to 2009M12, for which mainly the recession period since early 2008 is responsible. As already stated above with regards to Figure 3, PC2 seems to capture mainly recession effects. Fourthly, while there is no clear pattern for PC3 and PC4, except for slightly higher correlations in the subcategory of *durable goods* for the latter, the PC5 captures some variance of *health care* that was not captured by the other principal components and of *other*

services.

Finally, it can be seen from column (7) of Table 1, that the expanding-window models are systematically worse than the rolling-window models, which supports our choice of the rolling-window models as a basis for the nowcasts. This might be due to the inability of the nowcasting method to account for structural breaks in the time series during a period of economic turbulence.

In addition to the results reported in Table 1, we conducted two tests allowing to determine whether our nowcasting results are robust to the data-mining biases, namely: the Superior Predictive Ability (SPA) test of Hansen (2005) and the Model Confidence Set (MCS) test of Hansen et al. (2005)[3].

The null hypothesis of the SPA test is that the benchmark model is not inferior to any alternative nowcasts based on the leading indicators. The test results are presented in Table 2. Column (1) contains the name of the model. Column (2) reports the sample loss, which is in this case the mean squared error (MSE). While columns (3) and (4) report the t-statistic and "p-value" corresponding to the EPA test, respectively. The best model, according to the sample loss, is c(PC2,PC5), whereas the most significant one is Single.PC5. Columns (3) and (4) basically summarize the results in Table 1. In the lower panel of Table 2, the lower, consistent, and upper p-values of the SPA test are presented. The lower bound is the p-value of a liberal test, whose null hypothesis assumes that the alternative models with worse performance than the benchmark are poor models. The consistent p-value is produced by the test for SPA of Hansen (2005), that determines, which models are worse than the benchmark. The upper bound is the p-value of a conservative test, which assumes that all the competing models are as accurate as the benchmark in terms of expected loss. Whereas the conservative test is sensitive to including poor and irrelevant models in the comparison, the consistent and liberal tests are not influenced by it. In our case, none of the p-values is lower than 0.1. Given that the consistent p-value is 0.08, there is a statistical evidence that the Google- or sentiment indicator-based nowcasts as well as their combinations are better than the benchmark model nowcast. This finding corroborates the optimistic conclusion reached in the earlier literature on the usefulness of Google searches for short-term prediction of financial variables.

The outcome of the test for superior predictive ability that the models with leading indicators are superior to the univariate model is confirmed by the result of the MCS test, which is reported in Table 3. It shows

[3]The SPA and MCS tests were carried out using the MulCom package for Ox written by P. R. Hansen and A. Lunde.

the model confidence set, which is a set of models that is constructed so that it should contain the best model with a given level of confidence. In this particular case, the confidence level is set at 10%. In addition, the block-length parameter, d, was set to 2 and the number of bootstrap resamples was 10,000. Notice that the benchmark model, AR(1), does not appear in the reported model confidence set. Out of the models based on the conventional sentiment indicators only those including the 12th order differences of these indicators are included in this set. From the set of the models based on the financial variables only two were included into the MCS, namely: D12LSP500 and D12USBOND10Y. The majority of the models contained in the MCS are Google-based ones, in particular, those including PC2 and PC5 and model combinations. Notice also that the model c(PC2,PC5) has the highest among other models p-value equal to 1. The higher the p-value of the MCS test the more likely the corresponding model to be one of the best models.

To summarize, we find evidence that models with Google indicators provide a statistically significant forecast accuracy gain with respect to the benchmark model. Moreover, according to the results of the Diebold-Mariano test, for the models with conventional sentiment indicators and financial variables the null of equal predictive ability against that of the benchmark model cannot be rejected.

According to the forecast encompassing tests, we conclude that the Google indicator models with one or two principal components yield mixed results. As a rule, for those models where the relative RMSFE with respect to the benchmark model is less than unity, the corresponding nowcasts tend to encompass those of the latter model, whereas in cases when the relative RMSFE is larger than one the opposite generally is observed. An encouraging fact is that for all nowcasts based on averaging of Google indicator models the null hypothesis that the respective nowcasts encompass those of the benchmark model cannot be rejected at the usual significance levels, whereas the null hypothesis that these nowcasts can be encompassed by those of the benchmark model can be rejected. At the same time, the models based on the 12th order difference of both sentiment and financial indicators as well as the model with yield spread and the model based on averaged nowcasts of financial variables encompass the benchmark model without being encompassed by it themselves.

Taking into account the possibility of reaching erroneous conclusions when comparing multiple models, we cross-checked our results of the Diebold-Mariano tests using the testing procedures, which are specifically designed to deal with biases caused by data snooping. The results of the SPA and MCS tests concerning the

predictive ability of Google-based models point to the similar direction. Firstly, according to the SPA test, we are able to reject the null hypothesis that the benchmark model is not inferior to any alternative nowcasts based on the leading indicators, suggesting that the models augmented with Google or sentiment indicator or a financial variable do provide a substantial improvement of forecast accuracy beyond that of the AR(1) model. Secondly, his is confirmed by the fact that the autoregressive benchmark has not been selected into the model confidence set, i.e., the forecasting performance of the benchmark AR(1) model is not as good as of the models selected into MCS.

7 Conclusion

In this paper, using a real-time setting, where the forecaster disposes only of the information available up to current period, we examined the forecast accuracy gains obtained thanks to the Google indicators compared to an autoregressive benchmark and models based on the conventional sentiment indicators and financial variables. Our objective was to see whether the employment of this new data source allows significantly improving the nowcasts of US monthly real private consumption.

The Google indicators were constructed as common factors extracted using the methodology of Stock and Watson (1999) and Stock and Watson (2002) from 220 Google searches. Then, they were included into nowcast regressions along with the first lag of the private consumption. For the sake of comparison, the models based on the various leading indicators —two conventional sentiment indicators and several financial variables— and the first lag of the private consumption were estimated. As a benchmark model, to which all other models are compared, a simple AR(1) specification was chosen.

Based on the pairwise model comparison we find statistically significant evidence that models with Google indicators do offer an improvement in nowcast accuracy over the benchmark model. At the same time, according to the Diebold-Mariano test, for the models with conventional sentiment indicators and financial variables we cannot reject the null hypothesis of equal predictive ability against the benchmark model. Being aware of the possible erroneous conclusions that can be reached when comparing multiple models, we cross-checked these encouraging results with the tests specifically developed to deal with possible data-snooping biases. The evidence on the predictive power of Google-based models that came out of application of the SPA and MCS

tests points out to the usefulness of the Google indicators. Firstly, according to the SPA test, we can reject the null hypothesis that the benchmark model is not inferior to any alternative nowcasts based on the leading indicators. Secondly, the MCS test suggests that the benchmark AR model is not selected into a set of the best nowcasting models. Moreover, apart from the Google-based models, the MCS contains also some models including the survey-based indicators and financial variables. This implies that the information content of the Google indicators, survey-based indicators, and certain financial variables do help improving the nowcasts of the private consumption in US.

References

Askitas, N. and K. F. Zimmermann (2009). Google econometrics and unemployment forecasting. IZA Discussion Papers 4201, Institute for the Study of Labor (IZA).

Bai, J. and S. Ng (2002). Determining the number of factors in approximate factor models. *Econometrica 70*(1), 191–221.

Bureau of Economic Analysis (2009). *Concepts and Methods of the U.S. National Income and Product Accounts*. Bureau of Economic Analysis.

Choi, H. and H. Varian (2009a). Predicting initial claims for unemployment benefits. Technical report, Google Inc.

Choi, H. and H. Varian (2009b). Predicting the present with Google trends. Technical report, Google Inc.

Clark, T. E. and M. W. McCracken (2001). Tests of equal forecast accuracy and encompassing for nested models. *Journal of Econometrics 105*(1), 85–110.

Croushore, D. (2005). Do consumer-confidence indexes help forecast consumer spending in real time? *The North American Journal of Economics and Finance 16*(3), 435–450.

D'Amuri, F. (2009). Predicting unemployment in short samples with internet job search query data. MPRA Paper 18403, University Library of Munich, Germany.

D'Amuri, F. and J. Marcucci (2009). "Google it!" Forecasting the US unemployment rate with a Google job search index. MPRA paper, University Library of Munich, Germany.

Diebold, F. X. and R. S. Mariano (1995). Comparing predictive accuracy. *Journal of Business & Economic Statistics 13*(3), 253–263.

Diebold, F. X. and G. D. Rudebusch (1991). Forecasting output with the composite leading index: A real-time analysis. *Journal of the American Statistical Association 86*, 603610.

Doornik, J. A. (2009). Improving the timeliness of data on influenza-like illnesses using Google search data. Technical report, University of Oxford.

Giacomini, R. and H. White (2006). Tests of conditional predictive ability. *Econometrica 74*(6), 1545–1578.

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

Hansen, P. R. (2005). A test for superior predictive ability. *Journal of Business and Economic Statistics 23*(4), 365–380.

Hansen, P. R., A. Lunde, and J. M. Nason (2005). Model confidence sets for forecasting models. Working Paper 2005-07, Federal Reserve Bank of Atlanta.

Harvey, D. I., S. J. Leybourne, and P. Newbold (1998). Tests for forecast encompassing. *Journal of Business & Economic Statistics 16*(2), 254–259.

Stock, J. and M. Watson (1999). Forecasting inflation. *Journal of Monetary Economcs 44*, 293–335.

Stock, J. H. and M. W. Watson (2002). Macroeconomic forecasting using diffusion indexes. *Journal of Business and Economic Statistics 20*(2), 147–162.

Suhoy, T. (2009). Query indices and a 2008 downturn: Israeli data. Technical report, Bank of Israel.

Watson, M. W. and J. H. Stock (2004). Combination forecasts of output growth in a seven-country data set. *Journal of Forecasting 23*(6), 405–430.

8 Appendix

Table 1: RMSFE, 2007M1 - 2009M12

	RMSFE(1) (1)	RMSFE(2) (2)	RMSFE(1/2) (3)	EPA (4)	(1)ENC(2) (5)	(2)ENC(1) (6)	Expanding window RMSFE/RMSFE(1) (7)
UMCSI	0.552	0.554	0.996	0.436	0.185	0.147	1.102
D12UMCSI	0.530	0.554	0.957	0.272	0.235	0.031	1.142
CBCCI	0.559	0.554	1.008	0.626	0.109	0.150	1.064
D12CBCCI	0.502	0.554	0.906	0.125	0.267	0.007	1.120
TBILL3M	0.543	0.554	0.980	0.313	0.206	0.099	1.116
D12TBILL3M	0.543	0.554	0.980	0.332	0.299	0.096	1.106
DLSP500	0.566	0.554	1.021	0.791	0.170	0.523	1.099
D12LSP500	0.513	0.554	0.925	0.127	0.169	0.014	1.070
USBOND10Y	0.570	0.554	1.028	0.908	0.052	0.590	1.096
D12USBOND10Y	0.524	0.554	0.946	0.229	0.225	0.109	1.189
US10Ym3M	0.534	0.554	0.964	0.217	0.287	0.077	1.138
AVE3	0.509	0.554	0.919	0.101	0.556	0.043	1.086
PC1	0.576	0.554	1.038	0.711	0.059	0.224	1.184
PC2	0.468	0.554	0.844	0.015	0.855	0.004	1.141
PC3	0.520	0.554	0.939	0.060	0.801	0.056	1.140
PC4	0.551	0.554	0.994	0.385	0.239	0.252	1.104
PC5	0.510	0.554	0.920	0.031	0.302	0.001	1.187
PC6	0.589	0.554	1.062	0.976	0.016	0.967	1.149
PC7	0.618	0.554	1.114	0.893	0.120	0.720	1.054
PC8	0.593	0.554	1.070	0.989	0.006	0.982	1.088
PC9	0.577	0.554	1.042	0.914	0.066	0.282	1.069
PC10	0.581	0.554	1.048	0.811	0.189	0.505	1.075
c(PC1,PC2)	0.527	0.554	0.950	0.278	0.267	0.040	1.175
c(PC1,PC3)	0.585	0.554	1.055	0.780	0.065	0.329	1.151
c(PC1,PC4)	0.600	0.554	1.083	0.864	0.014	0.171	1.116
c(PC1,PC5)	0.531	0.554	0.958	0.313	0.115	0.018	1.274
c(PC1,PC6)	0.604	0.554	1.090	0.928	0.015	0.327	1.197
c(PC1,PC7)	0.628	0.554	1.133	0.979	0.001	0.352	1.182
c(PC1,PC8)	0.591	0.554	1.067	0.808	0.046	0.452	1.165
c(PC1,PC9)	0.565	0.554	1.019	0.638	0.074	0.170	1.217
c(PC1,PC10)	0.642	0.554	1.158	0.812	0.125	0.327	1.095
c(PC2,PC3)	0.499	0.554	0.899	0.114	0.424	0.014	1.045
c(PC2,PC4)	0.495	0.554	0.894	0.117	0.238	0.008	1.050
c(PC2,PC5)	0.463	0.554	0.835	0.008	0.355	0.001	1.170
c(PC2,PC6)	0.513	0.554	0.925	0.162	0.226	0.013	1.162
c(PC2,PC7)	0.529	0.554	0.953	0.265	0.118	0.055	1.096
c(PC2,PC8)	0.490	0.554	0.885	0.081	0.581	0.012	1.159
c(PC2,PC9)	0.479	0.554	0.863	0.019	0.488	0.002	1.091
c(PC2,PC10)	0.543	0.554	0.979	0.437	0.243	0.044	1.001
c(PC3,PC4)	0.533	0.554	0.961	0.151	0.438	0.083	1.107
c(PC3,PC5)	0.491	0.554	0.887	0.050	0.489	0.005	1.200
c(PC3,PC6)	0.571	0.554	1.031	0.748	0.150	0.331	1.159
c(PC3,PC7)	0.543	0.554	0.980	0.327	0.071	0.138	1.165
c(PC3,PC8)	0.533	0.554	0.962	0.293	0.416	0.124	1.166
c(PC3,PC9)	0.542	0.554	0.977	0.308	0.163	0.045	1.089
c(PC3,PC10)	0.549	0.554	0.990	0.445	0.313	0.113	1.103
c(PC4,PC5)	0.488	0.554	0.881	0.048	0.585	0.004	1.213
c(PC4,PC6)	0.565	0.554	1.018	0.696	0.128	0.413	1.168
c(PC4,PC7)	0.582	0.554	1.049	0.823	0.030	0.383	1.135
c(PC4,PC8)	0.582	0.554	1.050	0.831	0.047	0.578	1.112
c(PC4,PC9)	0.577	0.554	1.041	0.799	0.075	0.168	1.059
c(PC4,PC10)	0.601	0.554	1.084	0.785	0.177	0.340	1.037
c(PC5,PC6)	0.534	0.554	0.963	0.242	0.178	0.007	1.260
c(PC5,PC7)	0.541	0.554	0.976	0.290	0.142	0.076	1.156
c(PC5,PC8)	0.533	0.554	0.961	0.246	0.231	0.018	1.212
c(PC5,PC9)	0.568	0.554	1.025	0.696	0.078	0.001	1.074
c(PC5,PC10)	0.547	0.554	0.987	0.408	0.127	0.029	1.129
c(PC6,PC7)	0.660	0.554	1.190	0.962	0.035	0.875	1.077
c(PC6,PC8)	0.628	0.554	1.134	1.000	0.000	0.999	1.139
c(PC6,PC9)	0.599	0.554	1.080	0.992	0.023	0.497	1.139
c(PC6,PC10)	0.629	0.554	1.135	0.981	0.050	0.797	1.094
c(PC7,PC8)	0.643	0.554	1.160	0.996	0.015	0.929	1.063
c(PC7,PC9)	0.654	0.554	1.180	0.969	0.039	0.654	1.002
c(PC7,PC10)	0.668	0.554	1.206	0.985	0.058	0.743	0.994
c(PC8,PC9)	0.620	0.554	1.119	0.994	0.002	0.927	1.058
c(PC8,PC10)	0.616	0.554	1.112	0.963	0.056	0.836	1.079
c(PC9,PC10)	0.592	0.554	1.068	0.837	0.091	0.282	1.053
Single.PC2	0.512	0.554	0.923	0.117	0.560	0.016	1.166
Single.PC3	0.508	0.554	0.917	0.065	0.780	0.016	1.167
Single.PC4	0.514	0.554	0.928	0.077	0.812	0.030	1.157
Single.PC5	0.503	0.554	0.907	0.025	0.931	0.009	1.180
Single.PC6	0.513	0.554	0.926	0.027	0.941	0.012	1.177
Single.PC7	0.522	0.554	0.942	0.026	0.954	0.009	1.162
Single.PC8	0.529	0.554	0.954	0.050	0.942	0.016	1.153
Single.PC9	0.530	0.554	0.956	0.026	0.965	0.012	1.149
Single.PC10	0.533	0.554	0.961	0.040	0.930	0.021	1.145
Pair.PC3	0.528	0.554	0.953	0.283	0.338	0.038	1.120
Pair.PC4	0.519	0.554	0.936	0.184	0.422	0.024	1.126
Pair.PC5	0.488	0.554	0.880	0.039	0.786	0.006	1.186
Pair.PC6	0.498	0.554	0.898	0.042	0.794	0.010	1.197
Pair.PC7	0.506	0.554	0.913	0.045	0.807	0.014	1.192
Pair.PC8	0.514	0.554	0.928	0.078	0.807	0.020	1.185
Pair.PC9	0.513	0.554	0.925	0.034	0.900	0.014	1.181
Pair.PC10	0.519	0.554	0.937	0.073	0.771	0.028	1.171

Notes:

1. Column (1) contains the RMSFEs of the respective nowcast models.
2. Column (2) reports the RMSFE of the benchmark AR(1) model.
3. Column (3) reports the relative RMSFEs, which are obtained by dividing the respective model in row, RMSFE(1), over that of the benchmark model, RMSFE(2).
4. Column (4) contains the "naive" p-values that compares the pairwise performance of indicator models with the benchmark model, ignoring possible data-snooping bias.
5. Columns (5) and (6) report the p-values of the forecast encompassing tests: in column (5) the null hypothesis of the corresponding model based on leading indicators encompassing the benchmark model is tested, whereas in column (6) the null of the benchmark model encompassing the respective model based on leading indicators is tested.
6. Column (7) compares the RMSFE of the expanding-window models to that of the rolling-window models.

Table 2: Superior predictive ability test, 2007M1 - 2009M12

	Model	RMSFE	t-statistic	"p-value"
	(1)	(2)	(3)	(4)
Benchmark	AR(1)	0.554	—	—
Most significant	c(PC2,PC5)	0.503	2.488	0.008
Best	c(PC2,PC5)	0.463	2.488	0.008
Model 25%	Single.PC4	0.514	1.471	0.077
Median	c(PC3,PC9)	0.542	0.453	0.308
Model 75%	c(PC4,PC7)	0.582	-0.917	0.823
Worst	c(PC7,PC10)	0.668	-2.315	0.985
	Lower	Consistent	Upper	
SPA p-values:	0.065	0.080	0.106	
Critical values:				
10%	2.264	2.377	2.513	
5%	2.600	2.699	2.811	
1%	3.267	3.345	3.443	

Notes:

1. t-statistic and "p-value" refer to the EPA test that compares the benchmark model to the best alternative model, ignoring the fact that the latter was chosen from a large set of models.
2. The p-value of a liberal test, whose null hypothesis assumes that the alternative models with worse performance than the benchmark are poor models in limit, is reported; it provides a lower bound for the p-values of the SPA test.
3. Consistent p-value are reported for the SPA test of Hansen (2005).
4. The p-value of a conservative test, which assumes that all the competing models are as accurate as the benchmark in terms of expected loss, is reported; it provides an upper bound for the p-values of the SPA test.

Table 3: Model confidence set obtained at 10% level, 2007M1 - 2009M12

Model name	RMSFE	MCS p-value	
D12CBCCI	0.502	0.537	*
D12LSP500	0.513	0.493	*
D12USBOND10Y	0.524	0.145	**
FINCOMB	0.509	0.537	*
PC2	0.468	0.874	*
PC3	0.520	0.145	**
PC5	0.510	0.476	*
c(PC1,PC2)	0.527	0.117	**
c(PC2,PC3)	0.499	0.537	*
c(PC2,PC4)	0.495	0.537	*
c(PC2,PC5)	0.463	1.000	*
c(PC2,PC6)	0.513	0.234	*
c(PC2,PC7)	0.529	0.117	**
c(PC2,PC8)	0.490	0.599	*
c(PC2,PC9)	0.479	0.874	*
c(PC2,PC10)	0.543	0.262	*
c(PC3,PC5)	0.491	0.637	*
c(PC4,PC5)	0.488	0.637	*
c(PC5,PC7)	0.541	0.145	**
Single.PC6	0.513	0.234	*
Single.PC5	0.503	0.537	*
Single.PC4	0.514	0.140	**
Single.PC3	0.508	0.195	**
Single.PC2	0.512	0.125	**
Pair.PC9	0.513	0.140	**
Pair.PC8	0.514	0.117	**
Pair.PC7	0.506	0.234	*
Pair.PC6	0.498	0.537	*
Pair.PC5	0.488	0.599	*

Notes:

1. '*' and '**' indicate that a model belongs to the model 80% and 90% confidence sets, respectively.

Table 4: Correlation of principal components with year-on-year growth rates of monthly consumer spending disaggregates, 2005M1 - 2009M12

	Average share of total, %	Average monthly y-o-y growth rate, %	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10
Durable goods												
Motor vehicles and parts	0.04	-4.618	0.345	**0.593**	0.043	0.409	0.009	0.139	0.171	0.073	0.305	0.014
Furnishings and durable household equipment	0.03	0.926	**0.858**	0.078	0.219	0.384	0.017	0.099	0.022	0.010	0.078	0.018
Recreational goods and vehicles	0.04	8.996	**0.775**	0.130	0.061	**0.504**	0.087	0.093	0.039	0.060	0.035	0.002
Other durable goods	0.01	0.535	**0.631**	0.237	0.393	0.422	0.162	0.241	0.015	0.132	0.006	0.082
Nondurable goods												
Food and beverages purchased for off-premises consumption	0.08	1.785	**0.732**	0.203	0.276	0.476	0.013	0.063	0.118	0.145	0.041	0.085
Clothing and footwear	0.04	2.215	**0.876**	0.069	0.010	0.298	0.073	0.148	0.058	0.002	0.049	0.086
Gasoline and other energy goods	0.03	-0.888	0.225	**0.766**	0.061	0.001	0.037	0.027	0.116	0.045	0.062	0.074
Other nondurable goods	0.08	1.710	**0.826**	0.145	0.011	0.208	0.254	0.025	0.139	0.085	0.106	0.128
Services												
Housing and utilities	0.18	1.849	**0.722**	0.022	0.349	0.186	0.382	0.016	0.001	0.096	0.112	0.132
Health care	0.15	2.665	0.335	0.097	0.177	0.175	**0.789**	0.041	0.170	0.111	0.052	0.169
Transportation services	0.03	-1.376	**0.638**	**0.591**	0.116	0.328	0.068	0.134	0.244	0.015	0.050	0.018
Recreation services	0.04	1.506	**0.690**	0.293	0.201	0.177	0.469	0.058	0.190	0.107	0.029	0.030
Food services and accommodations	0.06	0.846	**0.879**	0.069	0.122	0.204	0.008	0.222	0.008	0.056	0.010	0.129
Financial services and insurance	0.08	1.880	**0.570**	**0.585**	0.004	0.232	0.221	0.295	0.287	0.026	0.059	0.068
Other services	0.09	1.517	0.311	**0.584**	0.138	0.113	**0.633**	0.064	0.074	0.045	0.020	0.102

Notes:

1. Entries in bold font indicate correlations larger than 50%.

Figure 1: Google Insights Interface

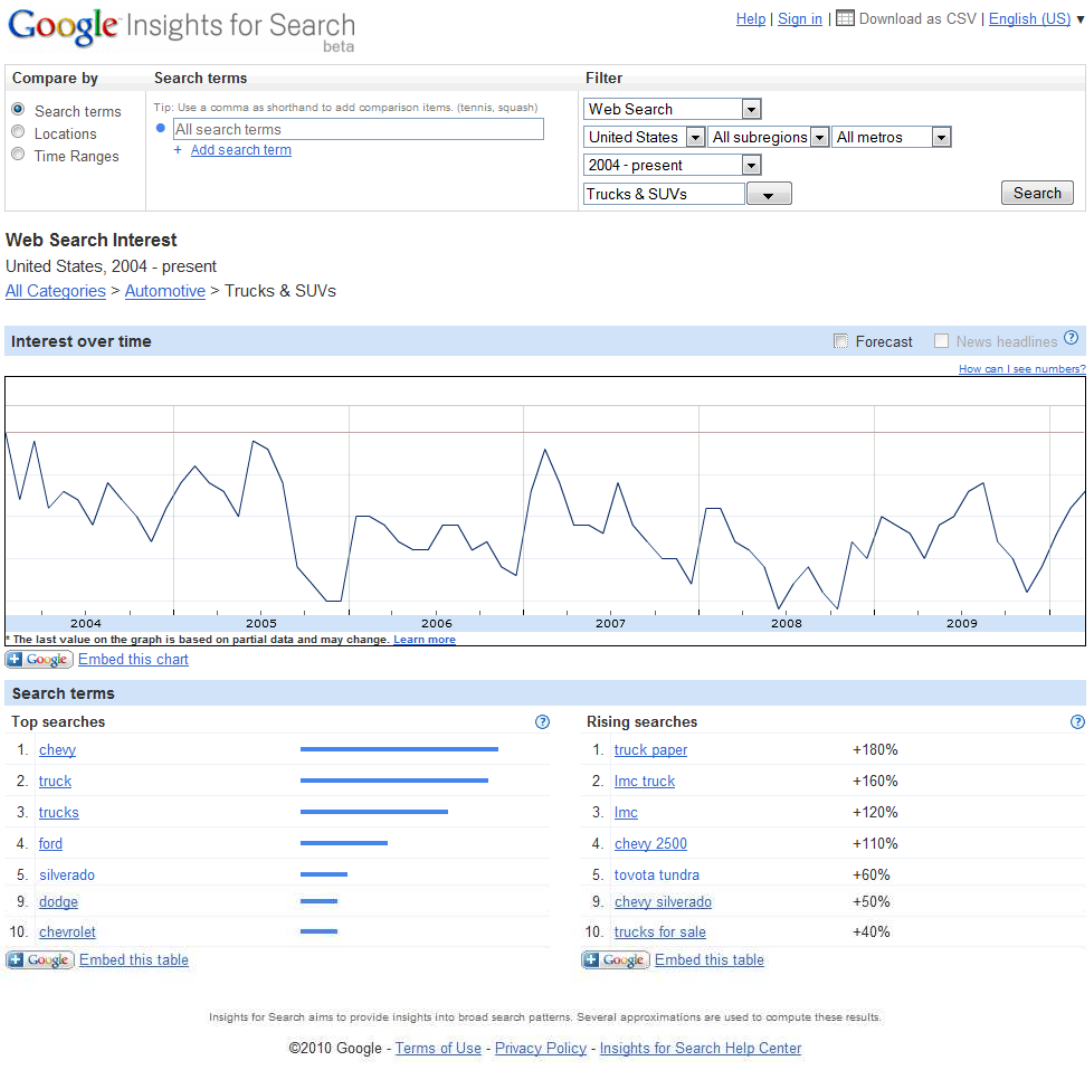


Figure 2: The contributions of the principal components to the total variance

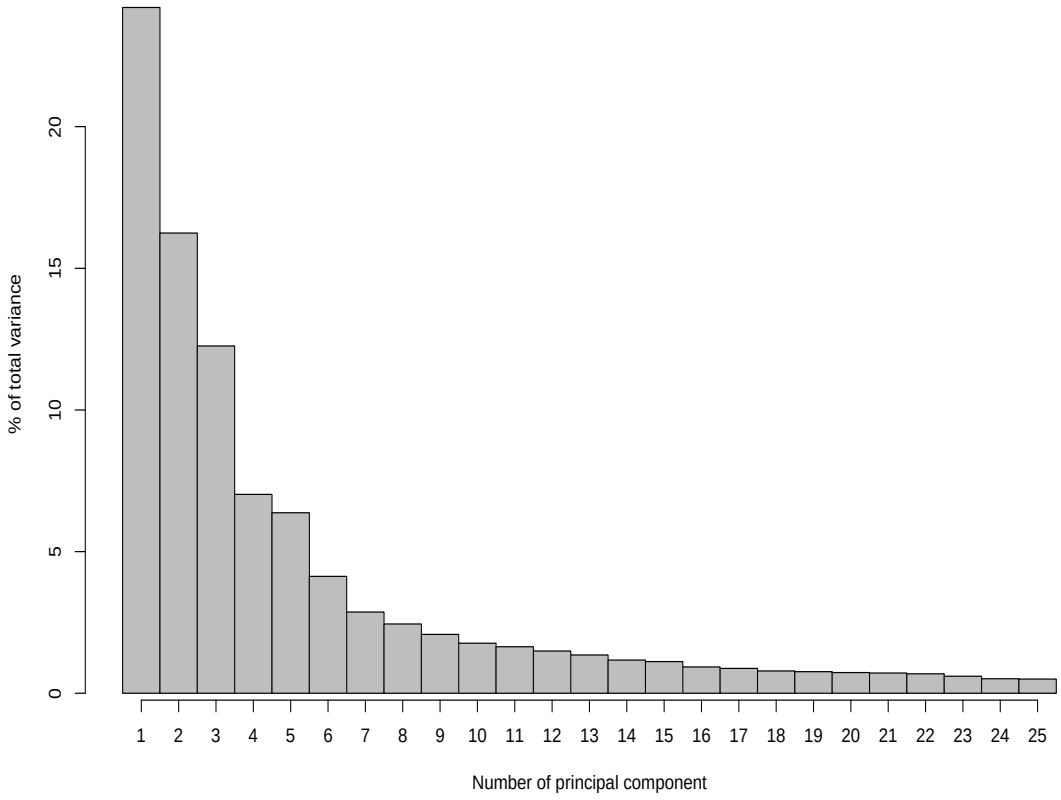


Figure 3: The first 10 prinicipal components

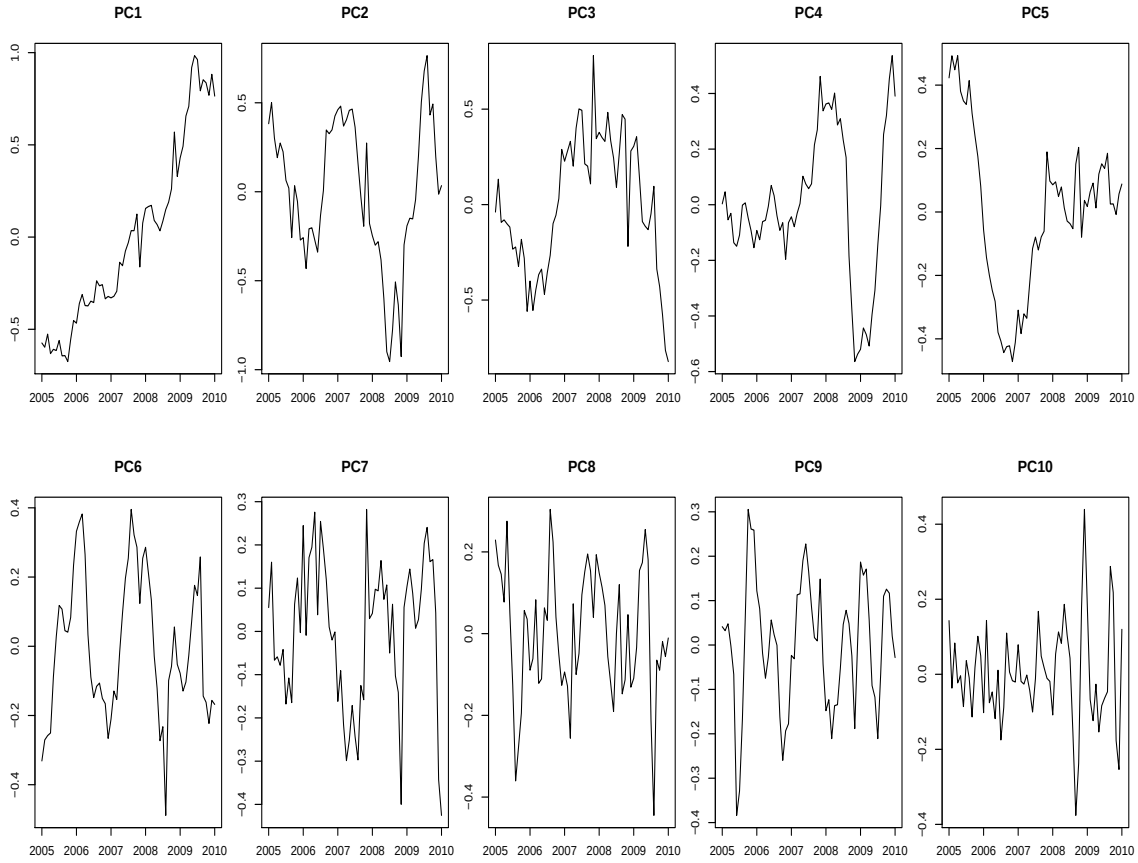


Figure 4: Actual growth rates of private consumption vs. benchmark nowcast, AR(1), and best nowcast, c(PC2,PC5)