RUHR
ECONOMIC PAPERS

Torsten Schmidt
Simeon Vosen

Using Internet Data to Account for Special Events in Economic Forecasting

rwi

#382

Imprint

Ruhr Economic Papers

Published by

Ruhr-Universität Bochum (RUB), Department of Economics
Universitätsstr. 150, 44801 Bochum, Germany

Technische Universität Dortmund, Department of Economic and Social Sciences
Vogelpothsweg 87, 44227 Dortmund, Germany

Universität Duisburg-Essen, Department of Economics
Universitätsstr. 12, 45117 Essen, Germany

Rheinisch-Westfälisches Institut für Wirtschaftsforschung (RWI)
Hohenzollernstr. 1-3, 45128 Essen, Germany

Editors

Editorial Office

Ruhr Economic Papers #382

Responsible Editor: Christoph M. Schmidt

ISSN 1864-4872 (online) – ISBN 978-3-86788-437-2

Ruhr Economic Papers #382

Torsten Schmidt and Simeon Vosen

Using Internet Data to Account for Special Events in Economic Forecasting

Bibliografische Informationen der Deutschen Nationalbibliothek

Die Deutsche Bibliothek verzeichnet diese Publikation in der deutschen Nationalbibliografie; detaillierte bibliografische Daten sind im Internet über: *http://dnb.d-nb.de* abrufbar.

http://dx.doi.org/10.4419/86788437
ISSN 1864-4872 (online)
ISBN 978-3-86788-437-2

Torsten Schmidt and Simeon Vosen[1]

Using Internet Data to Account for Special Events in Economic Forecasting

Abstract

Information about special events can improve economic forecasts substantially. However, due to the lack of timely quantitative data about these events, it has been difficult for professional forecasters to utilise such information in their forecasts. This paper investigates whether Internet search data can improve economic predictions in times of special events. An analysis of "cash for clunkers" programs in four selected countries exemplifies that including search query data into statistical forecasting models improves the forecasting performance in almost all cases. However, the challenge to identify irregular events and to find the appropriate time series from Google Insights for search remains.

JEL Classification: C53, E21, E27

Keywords: Forecast adjustment; Google Trends; private consumption

November 2012

1 Both RWI. – We thank Roland Döhrn for helpful comments. – All correspondence to Torsten Schmidt, RWI, Hohenzollernstr. 1-3, 45128 Essen, Germany, E-Mail: torsten.schmidt@rwi-essen.

1. Introduction

In recent years, the Internet has emerged as a promising new source of data for many fields of economic activity. For example, people use the Internet to look for information on goods and services, to compare prices and to buy or sell products. For each of these activities they use specialized platforms. The providers of these platforms collect, store and – at least in some cases – offer data related to these activities without delay. This data can thus fruitfully be employed for economic forecasting.

In most cases, economic forecasts are based on quantitative methods (Fildes, Steckler 2002: 456). An appropriate econometric model captures regularities of a time series and relations between time series in the past. In addition, it has to be assumed that these regularities hold over the forecast horizon. It is also shown in the literature that in practice the outcome of formal forecasting models can be improved by including additional information available for the forecaster but not incorporated in the data of the model (Goodwin, Fildes 1999). This is particularly the case in presence of unusual or irregular events. In some of these cases the forecaster has information about the events in advance. For example, a forecaster of a firm's sales should have knowledge about a promotion campaign. Likewise, a forecaster of private consumption should know about anticipation effects of a value added tax increase or government programs that seek to boost private consumption during recessions such as consumption checks or the "cash for clunkers" programs.

Due to the lack of timely data related to these unusual events, this information often cannot be implemented directly in the econometric model (Goodwin, Fildes 1999). To deal with this problem, the literature provides two strategies: (Goodwin 2000, 2005). Firstly, the forecaster can adjust the model forecast based on expert judgement. Secondly, the forecaster can produce a forecast based on a statistical model and a judgmental forecast and combine these forecasts in a mechanical way

(Lawrence et al. 2006, Marmier et al. 2009). However, informal adjustment methods may lead to substantial biases of the forecasts.

In this paper, we test whether Internet resources provide quantitative information that can be used to timely account for unusual events in statistical forecasting models. To be more precise, we use time series data from the Google Trends application *Insights for search*. Google provides time series information on the search intensities for specific search queries on a weekly basis. This may provide information related to irregular events that can be used in the forecasting process. We illustrate our approach using search query data on cash for clunkers programs during the financial crisis for four countries. We use this data to improve the fit of statistical forecasting equations for private consumption and test the forecasting performance of these adjusted equations.

The paper is structured as follows: In the next section we provide a brief overview of the literature on using search query date as a forecasting tool. In section three we describe the data and present our forecasting approach. The results are reported in section four and section five concludes.

2. Using search query data for forecasting

Ettredge et al. (2005) were the first to examine the potential of using Internet data for economic forecasting. They found search engine keyword usage data extracted from WordTracker's Top 500 Keyword Report to be useful in predicting the number of unemployed workers in the U.S. The literature on Google Trends data as a tool to predict economic statistics goes back to Choi and Varian (2009a), who used selected search categories to predict the dynamics in retail sales, automotive sales, home sales and travel.

Subsequently, search query data were used for predictions of many other variables of economic activity. A large number of papers investigated the use of Google search data in predicting labour market outcomes. Choi and Varian

(2009b) identified a relationship of what search query data to initial jobless claims in the US. D'Amuri and Marcucci (2009), Suhoy (2009), and Askitas and Zimmermann (2009) found similar relations for several other countries. Baker and Fradkin (2011) investigated how job search activity on the Internet responds to extensions of unemployment benefits.

The predictive power of search data has also been explored for other economic variables. Vosen and Schmidt (2011, 2012) demonstrated how Google Trends data can be used to improve forecasts of aggregate consumption in the US and in Germany. Guzman (2011) used Google data to predict inflation. Chamberlin (2010) used examined retail sales, car sales, and home sales. Wu and Brynjolfson (2010) and McLaren and Shanbhogue (2011) found that Google trends data to be a powerful tool to predict housing sales and prices. Finally, a number studies also found web search data to be a good predictor of financial indicators (Andrade et. al. 2012, Preis et al. 2010, Vlastakis and Markellos 2010, Da et al. 2010, 2011).

3. Data and Methodology

In response to the financial crisis and the worldwide recession, governments of many countries took stimulus measures to mitigate the economic downturn. Many of the measures, such as tax cuts or public investment programs, are regularly used to stabilize economic activity. Historical data on these measures are thus available and can be incorporated into the models of economic forecasters.

However, during the Great Recession, a relatively new measure, publicly known as the "scrappage subsidy" or "cash for clunkers", was introduced in several countries. It is essentially a targeted subsidy to the car industry, incentifying car owners to purchase a new car while scrapping their old ones. Knowledge about such programs is important for forecasting private consumption because cars form a substantial portion of private consumption expenditures. In contrast to other parts of the public expenditure programs, however, historical data on the timing of such new measures is rare. One exception is France where a car subsidy was introduced in 1992 (Adda, Cooper 2000) For this reason, it is particularly interesting to test the usefulness of search query data related to the "cash for clunkers" programs for forecasts of consumer spending.

We use data from four countries (France, Germany, Italy and the United States) that introduced a "cash for clunkers" program during the recent economic crisis. Among the first countries that took this measure was France. The program was introduced in December 2008 and terminated at the end of 2010. The total volume was roughly 1 billion euro. In January 2009, the German government started a similar program but the total volume was much larger (5 billion euro). The program ended in September 2009. The first scrappage scheme in Italy ran from January 2007 to December 2008. One month later the government launched a second program to promote car sales that ended in December 2009. Compared to other countries the U.S. program with a total volume of 3 billion dollar was introduced quite late.

To get information about the impact of these programs on private consumption, it seems reasonable to assume that people start their decision making process whether to buy a new car or not by collecting information about the program. Due to the fact that using web search engines have emerged as an increasingly important way to collect information, we employ data about the search intensity provided by Google Trends. The *Insights for Search* application provides indices of volumes of search queries sorted by countries and search categories. The indices are calculated from the raw data in two steps. In a first step the number of search queries for a phrase of interest at a special point in time in a given area is divided by the total number of search queries. In the second step these query shares are normalized to zero in the first week of January 2004 (Choi, Varian 2009).

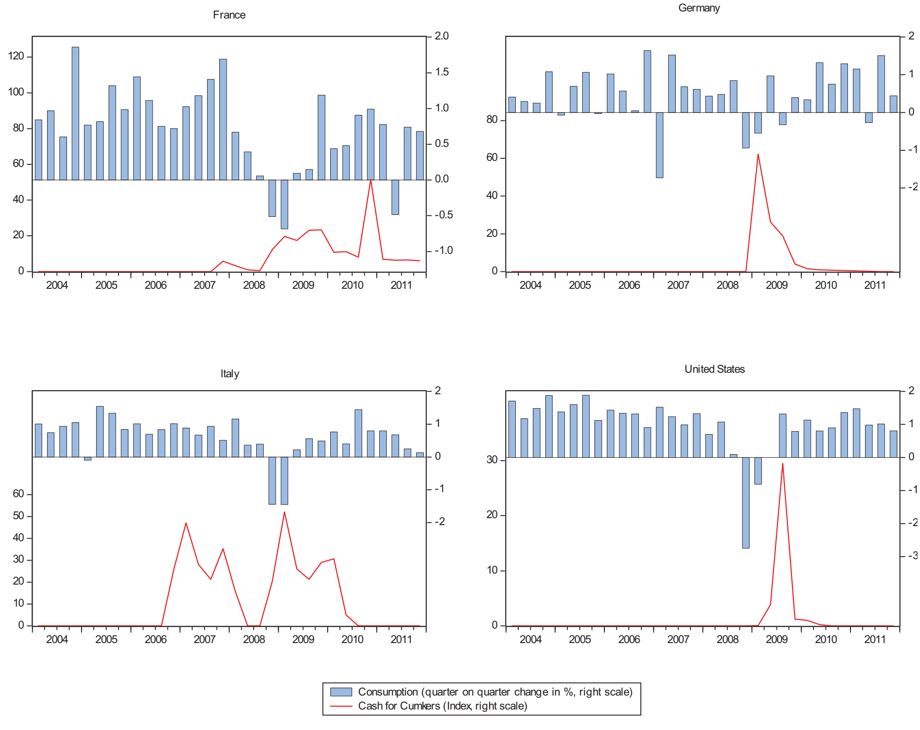
For the US we use the *Insights for Search* time series for the expression "cash for clunkers". In France the program was called "prime à la casse" in Germany

“Abwrackprämie” and in Italy “incentive alla rottamazione” so we use the data for these expressions. In addition, we use seasonally adjusted quarterly nominal data for private consumption and disposable income from national statistical offices to specify the baseline forecasting model.[1] We obtain stationary series by using first differences of the logs of variables.

In Figure 1 we plot search query indices and quarterly growth rates of private consumption for the selected countries. The search query data offer some interesting insights. In all four countries people started to search for the “cash for clunkers” programs before the official implementation of these programs. In France there was already substantial interest at the end of 2007, one year before the program was started. In Germany and the U.S. people started to search for “cash for clunkers” one quarter in advance of the official start. In Italy the lead of the “cash for clunkers” index before the launch of the programs was two quarters. However, it is not obvious from this picture whether these indices provide useful information for forecasting private consumption.

[1] For Italy we use nominal GDP instead of disposable income due to an insufficient number of observations for disposable income.

Figure 1: Google search indices for cash for clunkers and quarterly change of private consumption in selected countries (2004, 1 to 2011, 4)



Source: BEA, Destatis, INSEE, ONS, Google.

As a first test of the usefulness of search query data to account for unusual events in econometric models, we estimate equations for private consumption over a sample that includes the whole phase of the cash for clunkers program and test for statistical significance of the Google search query index for "cash for clunkers" in the four selected countries. As a second test, we perform one period ahead forecasts with and without the "cash for clunkers" data. To get a baseline specification of the forecast equation we use a sample that ends before the launch of the programs. For each country we included a constant and four lags of the change in private consumption (PC) and disposable income (DI). Insignificant lags of the variables were removed. We estimate our equations using OLS with

Newey-West standard errors. For the duration of the "cash for clunkers" programs which was known in advance, we re-estimate the equations each quarter and perform one step ahead forecasts of private consumption with and without the inclusion of the search query data for each quarter during the program. To keep the exercise simple, we use the mean of the weekly data to get quarterly observations. In each step we test whether the current value, a lagged value or the first difference of the cash for clunkers variable (CFC) is significant.

We let the "cash for clunkers" variable enter into the equations with current values and perform our forecast at the end of the quarter. This exercise is typically called "nowcasting". Since the statistical offices publish the official figures of the National Accounts with a substantial time lag, such predictions of the present still provide valuable information.

4. Empirical results

The estimation results for the period that includes the "cash for clunkers" programs are presented in Table 1. In all four equations, the "cash for clunkers" index enters significantly into the equations, although the lags differ between countries. For France and the U.S. the index enters without a lag while for Germany and Italy the index is included with one lag.

Table 1: Estimation results

France

$$\Delta \log PC_t = \underset{(9.557)}{0.008} + \underset{(3.019)}{0.272}\,\Delta \log DI_{t-1} - \underset{(9.467)}{0.009}\,D_{084} - \underset{(2.084)}{0.0002}\,CFC_t$$

Sample: 2002:1 – 2009:4, S.E. = 0.004, Adj. $R^2 = 0.58$

Germany

$$\Delta \log PC_t = \underset{(2.246)}{0.003} - \underset{(3.299)}{0.519}\,\Delta \log PC_{t-1} + \underset{(2.601)}{0.523}\,\Delta \log DI_{t-1} + \underset{(11.037)}{0.001}\,D_{064} + \Delta \underset{(4.014)}{(0.0002)}\,CFC_{t-1}$$

Sample: 2002:1 – 2009:4, S.E. = 0.005, Adj. $R^2 = 0.47$

Italy

$$\Delta \log PC_t = \underset{(7.563)}{0.009} + \underset{(3.318)}{0.251}\,\Delta \log PC_{t-2} - \underset{(3.044)}{0.113}\,\Delta \log DI_{t-1} - \underset{(22.593)}{0.025}\,D_{084} + \underset{(2.407)}{0.0001}\,CFC_{t-1}$$

Sample: 1995:1 – 2009:4, S.E. = 0.005, Adj. $R^2 = 0.50$

United States

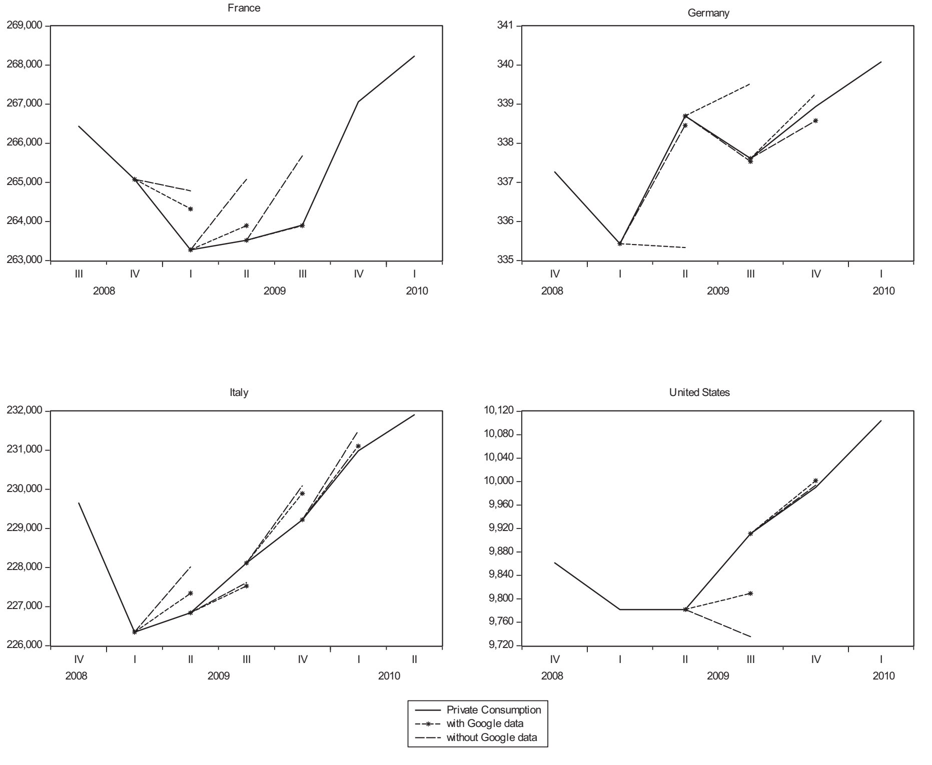
$$\Delta \log PC_t = \underset{(3.864)}{0.007} + \underset{(3.564)}{0.396}\,\Delta \log PC_{t-2} + \underset{(1.593)}{0.261}\,\Delta \log PC_{t-3} - \underset{(1.770)}{0.232}\,\Delta \log DI_{t-2}$$
$$- \underset{(12.230)}{0.027}\,D_{084} + \underset{(2.100)}{0.0004}\,CFC_t$$

Sample: 2002:1 – 2009:4, S.E. = 0.004, Adj. $R^2 = 0.84$

HAC standard errors. t-values in parenthesis.

The results of the forecasting exercise are presented in Figure 2. For each country we present the forecasts of the equations with and without the "cash for clunkers" index if this variable is significant for this specific sample. Following this procedure for all countries we get for France predictions for 2009:Q1 – 2009:Q3. In all cases the prediction errors are substantially smaller if the search query data were included. For Germany we get predictions for 2009:Q2 – 2009:Q4. Again, in the second and third quarter the forecast error were reduced substantially by including the "cash for clunkers" index while for the fourth quarter the magnitude of the error is roughly the same. The results for the U.S. are similar. We get forecasts for 2009:Q3 – 2009:Q4. While the forecast error of the extended equation is substantially smaller for Q3, the error is slightly larger. For Italy we get predictions for 2009:Q2 – 2010:Q1. In most cases the errors of the extended equation are slightly smaller. The only exception is 2009:Q3, were the error of this equation is slightly larger.

Figure 2: Predictions of private consumption with and without Google data during the cash for clunkers program in selected countries



The overall picture of this exercise demonstrates that search query data has the potential to provide information on unusual events that can be used to increase the accuracy of statistical forecasting models. This reduces the need for judgmental adjustment of model based forecasts. However, these results are not a formal test of the superiority of this adjustment compared to other approaches. Moreover, this approach still depends on certain judgments, e.g. regarding the appropriate search query.

5. Conclusions

Incorporating information on special events into forecasting models can improve economic forecasts substantially. Several approaches have been proposed to improve model based forecasts in presence of a lack of timely quantitative data on these events. Most of these seek to adjust the model based forecast by judgement or combine a statistical and a judgemental forecast. They are able significantly reduce the forecast error of the baseline forecast, but it seems plausible that they can be further improved by adding quantitative data about these events.

This paper demonstrates that Google's *Insights for Search* data can in many occasions reduce this lack of information about unusual events. Due to the enormous importance of the Internet in providing and distributing information, web search engines generate data about almost all major current events. The case of the "cash for clunkers" programs, explored in this paper, illustrates that the inclusion of search query data into statistical forecasting models improves the forecasting performance in almost all cases. However, the challenge of identifying irregular events and the corresponding Google time series remains.

References

Adda, J. and Cooper, R. (2000), "Balladurette and Juppette: A Discrete Analysis of Scraping Subsidies", *Journal of Political Economy*, Vol. 108 No. 4, pp. 778-806.

Andrade, S.C., Bian, J. and Burch, T.R. (2012), "Analyst Coverage, Information and Bubbles", Working Paper, University of Miami, Miami, May.

Askitas, N. and Zimmermann, K.F. (2009), "Google Econometrics and Unemployment Forecasting", *Applied Economics Quarterly*, Vol. 55 No. 2, pp. 107-120.

Baker, S. and Fradkin, A. (2011), "What Drives Job Search? Evidence from Google Search Data", SIEPR Discussion Paper No. 10-020, Stanford Institute for Economic Policy Research. 30 March.

Chamberlin, G. (2010), "Googling the Present", *Economic and Labour Market Review*, Vol. 4 No. 12, pp. 59–95.

Choi, H. and Varian, H. (2009a), "Predicting the Present with Google Trends", Google, 10 April.

Choi, H. and Varian, H. (2009b), "Predicting Initial Claims for Unemployment Benefits", Google, 5 July.

Da, Z., Engelberg, J. and Gao, P. (2011), "In Search of Attention", *Journal of Finance*, Vol. 6 No. 5, pp. 1461–1499.

Da, Z., Engelberg, J. and Gao, P. (2010), "In Search of Earnings Predictability", Working Paper, University of Notre Dame and University of North Carolina at Chapel Hill, 9 April.

D'Amuri, F. and Marcucci, J. (2009), "Google it! Forecasting th US unemployment rate with a Google job search index", ISER Working Paper Series, No. 2009-32, University of Essex, November.

Ettredge, M., Gerdes, J. and Karuga, G. (2005), "Using Web-based Search Data to Predict Macroeconomic Statistics", *Communications of the Association for Computing Machinery*, Vol. 48 No. 11, pp. 87-92.

Fildes, R. and Stekler, H. (2002), "The State of Macroeconomic Forecasting", *Journal of Macroeconomics*, Vol. 24, pp. 435-468.

Goodwin, P. (2000), "Correct or Combine? Mechanically Integrating Judgemental Forecasts with Statistical Methods", *International Journal of Forecasting*, Vol. 16, pp. 261-275.

Goodwin, P. (2005), "How to Integrate Management Judgement with Statistical Forecasts", *Foresight*, Vol. 1 No. 1, pp. 8-12.

Goodwin, P. and Fildes, R. (1999), "Judgmental Forecasts of Time Series Affected by Special Events: Does Providing a Statistical Forecast Improve Accuracy?", *Journal of Behavioral Decision Making*, Vol. 12, pp. 37-53.

Guzman, G.C. (2010), "Internet Search Behavior as an Economic Forecasting Tool: The Case of Inflation Expectations", *Journal of Economic and Social Measurement*, Vol. 36 No. 3, pp. 119-167.

Lawrence, M., Goodwin, P., O'Connor, M. and Önkal, D. (2006), "Judgmental Forecasting: A Review of Progress over the last 25 Years", *International Journal of Forecasting*, Vol. 22, pp. 493-518.

Marmier, F., Gonzales-Blanch, M. and Cheikhrouhou, N. (2009), "A New Structured Adjustment Approach for Demand Forecasting", *International Conference on Computing and Industrial Engineering*, pp. 773-778.

McLaren, N. and Shanbhogue, R. (2011), "Using Internet Search Data as Economic Indicators", Quarterly Bulletin No. 2011, Bank of England, London.

Preis, T., Reith, D. and Stanley, H.E. (2010), "Complex Dynamics of our Economic Life on Different Scales: Insights from Search Engine Query Data", *Philosophical Transactions of the Royal Society A* 368, pp. 5707-5719.

Suhoy T. (2009), "Query Indices and a 2008 Downturn: Israeli Data", Discussion Paper, 2009/06, Bank of Israel, Jerusalem, July.

Vlastakis, N. and Markellos, R. (2012), "Information Demand and Stock Market Volatility", *Journal of Banking and Finance* (In Press).

Vosen, S. and Schmidt, T. (2011), "Forecasting Private Consumption: Survey-based Indicators vs. Google-Trends", *Journal of Forecasting*, Vol. 30 No. 6, pp. 565-578.

Vosen, S. and Schmidt T. (2012), "A Monthly Consumption Indicator for Germany Based on Internet Search Query Data", *Applied Economics Letters*, Vol. 19 no. 7, pp. 683-687.

Wu, L. and Brynjolfsson, E. (2009), "The Future of Prediction: How Google Searches Foreshadow Housing Prices and Quantities", *ICIS 2009 Proceedings* Paper 147.