German Council for Social
and Economic Data (RatSWD)

www.ratswd.de

RatSWD
Research Notes

Research Note No. 41

Google Econometrics and Unemployment Forecasting

Nikos Askitas, Klaus F. Zimmermann

June 2009

Federal Ministry
of Education
and Research

Research Notes of the Council for Social and Economic Data (RatSWD)

The *RatSWD Research Notes* series publishes empirical research findings based on data accessible through the data infrastructure recommended by the RatSWD. The pre-print series was launched at the end of 2007 under the title *RatSWD Working Papers*.

The series publishes studies from all disciplines of the social and economic sciences. The *RatSWD Research Notes* provide insights into the diverse scientific applications of empirical data and statistics, and are thus aimed at interested empirical researchers as well as representatives of official data collection agencies and research infrastructure organizations.

The *RatSWD Research Notes* provide a central, internationally visible platform for publishing findings based on empirical data as well as conceptual ideas for survey design. The *RatSWD Research Notes* are non-exclusive, which means that there is nothing to prevent you from publishing your work in another venue as well: all papers can and should also appear in professionally, institutionally, and locally specialized journals. The *RatSWD Research Notes* are not available in bookstores but can be ordered online through the RatSWD.

In order to make the series more accessible to readers not fluent in German, the English section of the *RatSWD Research Notes* website presents only those papers published in English, while the the German section lists the complete contents of all issues in the series in chronological order.

The views expressed in the *RatSWD Research Notes* are exclusively the opinions of their authors and not those of the RatSWD.

Google Econometrics and Unemployment Forecasting

Nikos Askitas
IZA

Klaus F. Zimmermann

Bonn University, IZA, and DIW Berlin: Zimmermann[at]iza.org

Abstract

The current economic crisis requires fast information to predict economic behavior early, which is difficult at times of structural changes. This paper suggests an innovative new method of using data on internet activity for that purpose. It demonstrates strong correlations between keyword searches and unemployment rates using monthly German data and exhibits a strong potential for the method used.

Keywords: Google, internet, keyword search, search engine, unemployment, predictions, time-series analysis
JEL classification: C22, C82, E17, E24, E37

1. Introduction

The internet contains an enormous amount of information which, to our knowledge, classical econometrics has yet to appropriately tap into. Such information comes timely on a continual basis. It is particularly welcome at times of an economic crisis where the traditional flow of information is too slow to provide a proper basis for sound economic decisions.[1] Not only has traditional (and typically official) statistical data a slow publication scheme, these data also do not reflect well the structural changes in the economy. While investigating many different kinds of internet activity, we focus here on Google search data to establish strong correlations between search activities for certain keywords or keyword groups and the unemployment rates in Germany. We call the relationship a Google predictor. Such an application is timely, since we have just experienced a turning-point in the fall of the unemployment rates after a longer decline caused by labor market reforms and the past economic boom. It is a particular challenge for the new proposed method to capture that turning-point properly.

Previous applications of Google search engine query data include Constant and Zimmermann (2008) measuring economic and political activities, and Ginsberg, Mohebbi, Patel, Brammer, Smolinski and Brilliant (2009) for studying influenza epidemics. While the former study purely documents the evolution of particular keyword searches before the US presidential elections, the latter investigates an epidemic process using more complex computational methods. The novel feature here in this paper is to demonstrate that the data can be used to predict economic behavior measured by traditional statistical sources.

The study is structured as follows. In Section 2, we explain how we use Google Insights and how we choose our indicator variables from the keyword searches. In Section 3, we provide the empirical results. Section 4 contains our conclusions and future plans.

2. Google Econometrics: Unemployment Rates and Choice of Indicators

In the summer of 2008, Google released a beta version of Google Insights (http://www.google.com/insights/search/). Using the service, search queries can be compared for keywords across countries and in some cases their regions, in narrow or wide time frames from 2004 onwards. A Google Insights query may have regional, temporal or keyword specific focus, i.e. you choose the region of interest, the time frame of interest and the

1 See Zimmermann (2008, 2009) for an analysis of the current challenges for economic forecasting.

keywords of interest (up to 5). The results are then delivered scaled and normalized within the query (for the region, the time frame and the selection of keywords)[2]. This presents some interesting but not insurmountable challenges in accessing the data. Google Insights has also been modifying the service since it was started, which caused changes in the way we were able to access the data ourselves. The data access is limited and restricted in many ways.

Ginsberg, Mohebbi, Patel, Brammer, Smolinski and Brilliant (2009) in their study of influenza epidemics obviously had better access to more data and consequently were able to apply more complex computational methods. They demonstrate how flu epidemics can be predicted using Google Insights as its data source. When we started to work on the idea to investigate human behavior measured by traditional statistical sources using internet queries and to apply it to correlate keyword searches and unemployment rates, among other things, we were not aware of their study. Knowledge of this work, however, encouraged us to proceed with our paper. Given our restricted access to the data, we decided to attempt a minimalist approach: theorize the choice of keywords, reduce our investigations to the parsimonious basics and demonstrate the power of the method.

In order to motivate our investigation as well as our use of the data, we need to set the stage by explaining the challenge we posed to ourselves. In Germany, the unemployment rates are announced monthly at a press conference by the Federal Employment Agency. The announcement dates are provided in advance for the next two years and are almost always at the end of the month, but sometimes early in the first week of the following month. This means that at the end of a given month M the unemployment rate "for the month" is made known. We will denote this by U_M. The data used to compute U_M is based on administrative data of the unemployment office between the middle of month M-1 and the middle of month M.

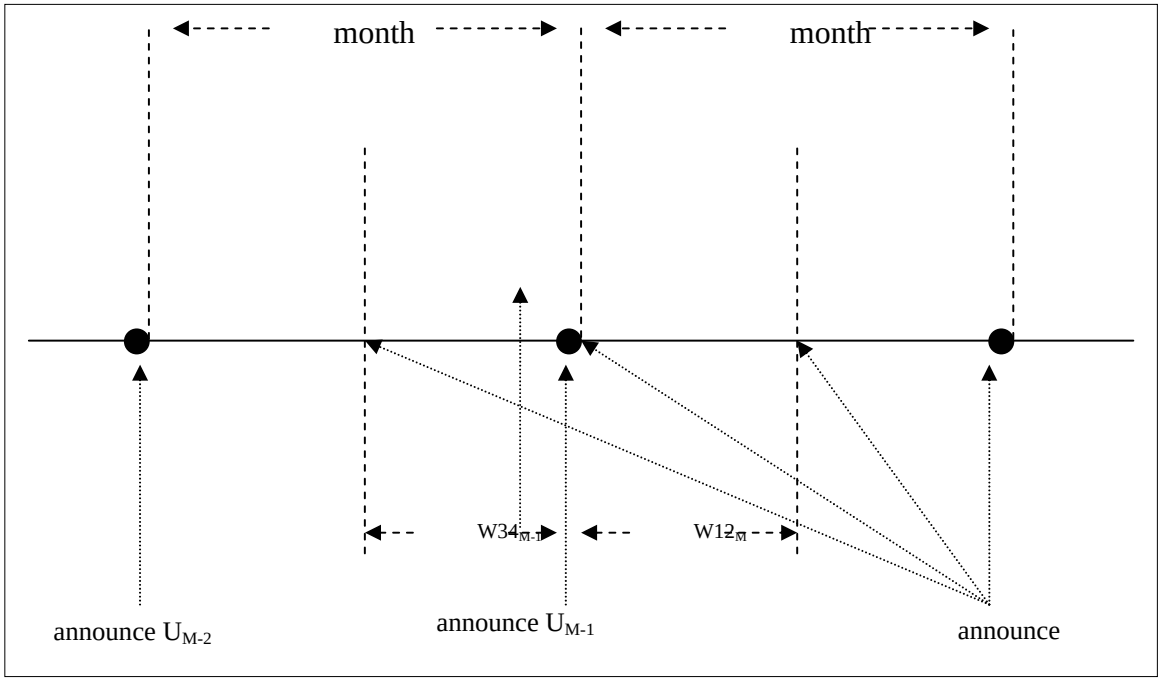
This means that the announced unemployment rates for month M, which are issued by the end of the month, are based on real unemployment processes occurring in the union of two time intervals:

- The first interval denoted by $W34_{M-1}$ is roughly speaking the 3rd and 4th week of month M-1.
- The second interval denoted by $W12_M$ is then the 1st and 2nd week of month M.

2 We decided to query Google Insights for keywords one at a time. This way we lost the information of the relative weight of keyword activity but freed ourselves of the problem of having a large volume variable trivialize a low volume one. The idea is that a smaller group of people may cause a low volume of keyword activity which contains as much or more information than a keyword with large volume.

We should point our that practically in the middle of the two time intervals (i.e. around the end of month M-1), we have the release of the unemployment rates for month M-1, which is based on unemployment occurring in the intervals $W34_{M-2}$ and $W12_{M-1}$. Figure 1 captures all the relevant information to set the stage for the real monthly unemployment rate: how it is measured and when is it made known.

Figure 1. The time structure of announcing unemployment by the German Federal Labor Office.



Note: U is announced unemployment, M month, and Wxy refers to weeks x and y in a particular month. Hence, U_M is unemployment U in month M and $W12_M$ refers to both weeks 1 and 2 in month M. 15 refers to the 15th day in the particular month.

Google collects, normalizes and scales the number of searches for all kinds of keywords, provided there is a "sufficient amount" of searches for this keyword. The exact threshold is not known to us. We will say that the data Google Insights returns for a certain keyword k is the "Google activity along the keyword k" and denote this by $g_{k,12,M}$ or $g_{k,34,M}$ in weeks 1 and 2, or 3 and 4 of month M.[3] As unemployment occurs, people are also using Google for all kinds of keyword searches. If we had access to the entire recorded Google activity along all keywords, we could attempt a more comprehensive approach, but even so we can ask whether we can figure out a core set of keywords whose Google activity would have predictive power

3 In our presentation and the resulting tables and graphs the variable convention we use is as follows: the variable containing the values gn,34,M is called w34kn and the one containing the values gn,12,M is called w12kn. Here w and k stand for week segment and keyword.

for the monthly unemployment rates. Google returns the data in weekly values, and the week boundaries are known to us. They do not contain the boundaries of our time intervals above, so we needed to re-split the activity proportionally to overcome this issue. We are aware that this introduces a certain amount of noise, and in fact this is the reason why we decided to use biweekly rather than weekly time intervals to minimize the noise we introduce.

Our aim is to investigate the extent to which we can locate keywords whose activity $g_{k,12,M}$ and $g_{k,34,M-1}$ may be used to predict U_M. We expect activities in the intervals $W34_{M-1}$ to have better predictive power than those in the interval $W12_M$, although the latter period is closer to the new announcement than the former. The reason for this is that the rate U_{M-1} is announced in between the two intervals and influences the activity $g_{k,12,M}$, i.e. people react to the announcement. A similar impact to $g_{k,34,M-1}$ may only come from $U_{M-2,}$ which was announced two weeks prior and is therefore less likely to be remembered.

We use measurements of Google activity along the disjunction of four groups of keywords (Google Insights supports queries for disjunctions of keywords):

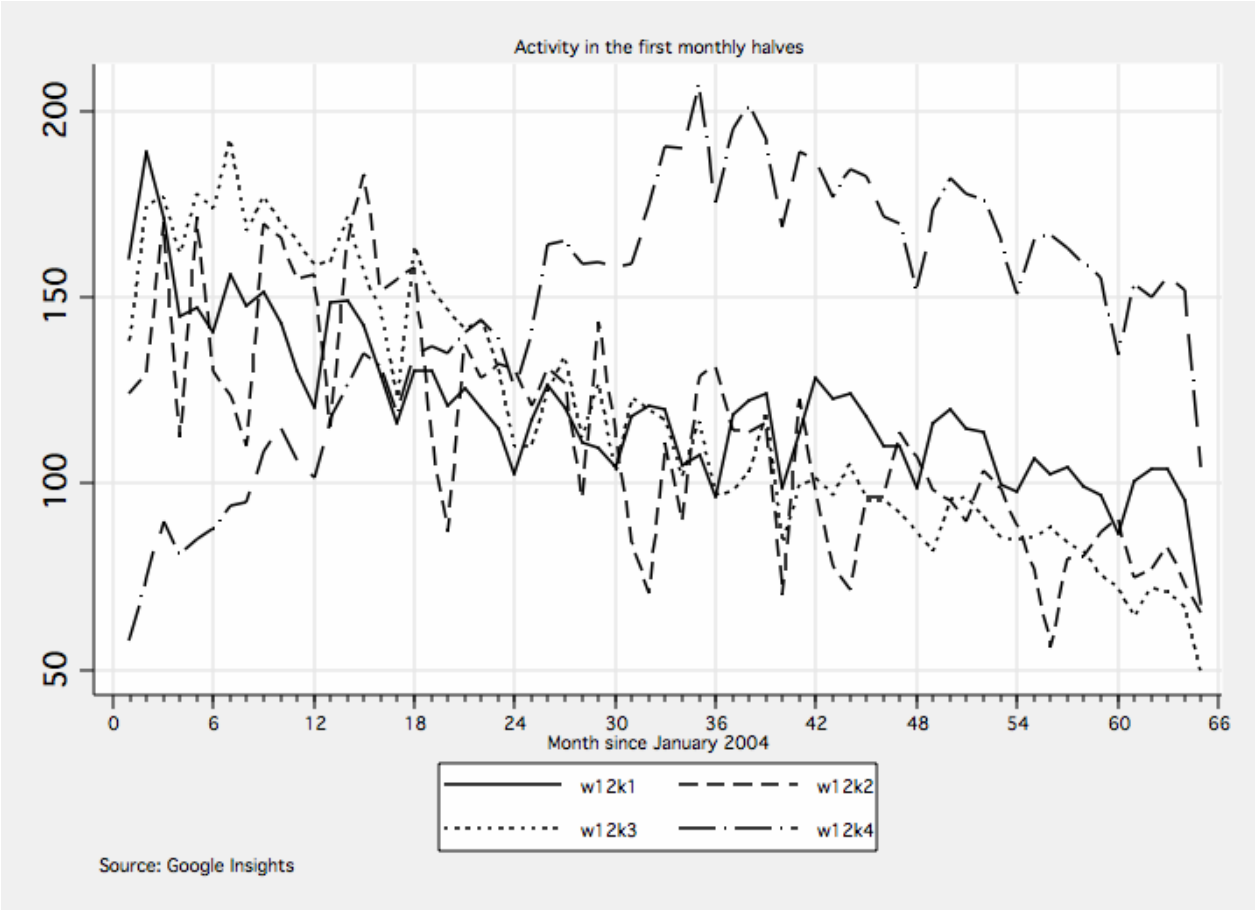
k1	Arbeitsamt OR Arbeitsagentur ("unemployment office or agency")
k2	Arbeitslosenquote ("unemployment rate")
k3	Personalberater OR Personalberatung ("Personnel Consultant")
k4	Stepstone OR Jobworld OR Jobscout OR Meinestadt OR meine Stadt OR Monster Jobs OR Monster de OR Jobboerse ("most popular job search engines in Germany")

We expect Google activity along k1 (Arbeitsamt or Arbeitsagentur) to be connected with people having contacted or being in the process of contacting the unemployment office. As such it should have something to do with the "flow into unemployment". The keyword k2 (Arbeitslosenquote) is just the easiest and most natural keyword to think of when dealing with unemployment. The activity around the disjunction k3 (Personalberater or Personalberatung) is expected to correlate with high-skilled workers reacting to fears of layoffs and companies preparing for layoffs or personnel restructuring. The keyword k4 (Stepstone or Jobworld or Jobscout or Meinestadt or meine Stadt or Monster Jobs or Monster de or Jobboerse) is expected to be related to job searching activities, and hence should be associated with the "flow out of unemployment."

Figure 2 shows plots of Google activity along the keyword sets above in the first monthly halves, while Figure 3 exhibits Google activity along the keyword in the second monthly halves. All indicators seem to follow a somewhat similar seasonal pattern, while activities

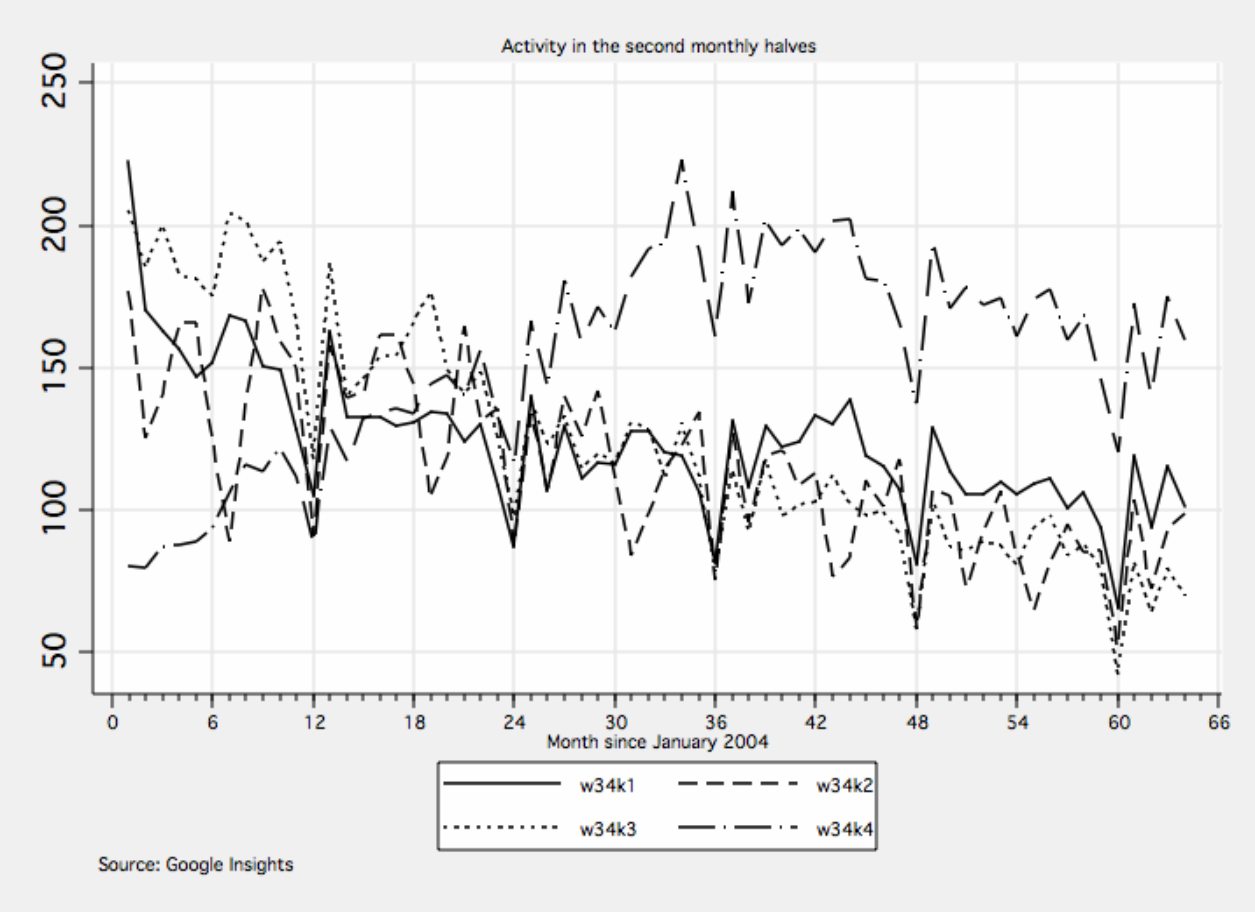
measured through variables k1 - k3 move down and activity k4 moves up. Data collected in weeks 1 and 2 (Figure 2) provide somewhat different signals than those collected in weeks 3 and 4 (Figure 3). As discussed above, the special sequence of announcements makes it likely that the more recent information before an announcement is clouded by the previous announcement. Furthermore, in terms of using the data for predictions, it is useful to rely on earlier information because this may allow the analyst to obtain forecasts much faster. Below we investigate further whether the older search activity data predicts unemployment rates better than the more recent data.

Figure 2



Note: w12kx is variable k1 collected in weeks 1 and 2; the variables are k1: Arbeitsamt or Arbeitsagentur, k2: Arbeitslosenquote, k3: Personalberater or Personalberatung, and k4: Stepstone or Jobworld or Jobscout or Meinestadt or meine Stadt or Monster Jobs or Monster de or Jobboerse.

Figure 3



Note: w34kx is variable k1 collected in weeks 1 and 2; the variables are k1: Arbeitsamt or Arbeitsagentur, k2: Arbeitslosenquote, k3: Personalberater or Personalberatung, and k4: Stepstone or Jobworld or Jobscout or Meinestadt or meine Stadt or Monster Jobs or Monster de or Jobboerse.

We close this section with some comments on the Google access data and its usefulness. Expecting that search engine keyword searching contains information which correlates with people's lives is a natural and, we believe, commonly accepted expectation. In fact, provided we are able to weed out the noisy activity and get to the signal in any kind of effective way, this approach may be thought of as an indirect form of anonymous interviewing resulting in a noisy aggregate time-series data set. It is not surprising that the study of Google search activity contains a large portion of the general search engine activity. As of December 2008, Google's share of the search engine market was close to 63%, with Yahoo being a distant second at about 17%, MSN third at about 10%, followed by AOL search at 4% and ask.com at 2%.[4]

Most people use Google not just as a search engine but also as a directory of their sites of

4 See for example the Nielsen Online December 2008 Search Engine Share Rankings, http://searchenginewatch.com/3632382.

interest. It is quite common for someone to first google a familiar website and click on the appropriate URL, rather than enter the required site in the address bar. Not only does search activity contain residual information on the Google user but it also contains information on the sites the Google user intends to visit.

Lastly, we need to discuss issues of keyword choice. In constructing the “search for a job” keyword set k4, for example, we had to ascertain what kind of online job directory services there were. The keyword set which defines k4 is not constant over time: sites may come in and out of existence etc. The concept under study captured by the choice of keywords may depend on linguistic developments, generational parameters, social and economic levels and a host of other factors. It is therefore important to use keywords which remain constant during the period observed. We tried a wide range of other keyword families capturing such concepts as consumption, retail activity and online dating, but we restricted ourselves to k1, k2, k3 and k4, as they seem to be sufficient in order to model the process of unemployment we aim to investigate.

3. Empirical Results

To investigate the usefulness of the Google search activity data for predicting real economic behavior, we employ a time-series causality approach using the well-known error-correction model specification (Engle and Granger, 1987; Greene, 2008). This approach implies that the change of the variable of interest is regressed on its past level, the change of the explanatory variables of interest, and their past levels. The real data variable to explain used here is the seasonally unadjusted monthly unemployment rate of Germany[5] from January 2004 to April 2009. This particular time-frame has been enforced by the availability of the Google query data and the latest available data point at the time when this investigation had to be carried out. To calculate the change of the variables used, a 12 month lag operator is used; consequently, the past stock variables are of lag 12. This has the advantage that we do not need to model seasonality explicitly.

Given the severe economic crisis and the sudden strong decline in economic activity, the unemployment variable is currently of particular interest to the general public and for

5 Collecting the information from the Federal Employment Agency on the internet was a bit cumbersome. The current monthly report is posted in PDF format under: http://www.pub.arbeitsagentur.de/hst/services/statistik/000000/html/start/monat/aktuell.pdf. In order to collect the monthly data we needed to download and parse all pdf documents. The authors believe that the data posted by the Federal Employment Agency would be more complete if it included "machine actionable" data streams (in SDMX standard for example) in addition to PDF reports for historical data. The work done at the European Central Bank in that direction (http://www.ecb.eu/stats/services/sdmx/html/index.en.html) is a good example of such a service.

scientists. A surprisingly long continual decline in unemployment rates in the first quarters of the German recession until December 2004 were observed, which was mainly driven by a long period of economic boom in connection with the significant and effective labor market reforms undertaken in the previous years. The economic decline, however, became suddenly very pronounced in the fourth quarter of 2008, and in specific economic sectors: namely the export oriented high-quality investment goods industries. It resulted in a labor policy measure which sought to encourage government supported short-time working and was accompanied by a strong PR campaign by the Federal Employment Agency. The period of short-time working was increased from previously 6 months to first 18 and finally 24 months. The short-time working allowance increased: Employers do only have to pay half of the normal social security contributions for short-time workers, and even nothing if short-time workers engage in further education. Also, access to short-time working has been improved. This all resulted in strong incentives to retain staff, encouraged further education, and lead to a reduction of a possible loss of income by employees. Companies adopted the policy at unprecedented levels, contributing to the only moderate increase in unemployment in early 2009. In this environment, unemployment predictions are very difficult even in the short-term, and a soft approach using internet activity data might be even more warranted. We want to evaluate its potential here.

Tables 1 to 4 contain the estimated error correction models for two and more regressors capturing the effects of weeks 1 and 2 of the current month (see Tables 1 and 2) and weeks 3 and 4 of the previous month (see Tables 3 and 4). The regressors used are the four variables k1 (Arbeitsamt or Arbeitsagentur), k2 (Arbeitslosenquote), k3 (Personalberater or Personalberatung), and k4 (Stepstone or Jobworld or Jobscout or Meinestadt or meine Stadt or Monster Jobs or Monster de or Jobboerse). The estimates are created in a systematic way and presented in the tables together with coefficients, their t-ratios, and information criteria (R^2, log-likelihood values, AIC, and BIC). We will base our judgement of the statistical performance of a model on the BIC; the other measures are for comparison only.

The correct choice of model has to be seen in the context of parsimony, prediction success, usefulness, and sound economic basis. The economic variables included should have short- and long-run effects in line with economic intuition, and there should be a long-run stationary solution of the model. The statistical model should be parsimonious, and therefore we want to use as few explanatory variables as possible. This is typically investigated with an information criteria like the BIC or the AIC. The approach is useful if it employs regressors that are available early, and hence enables early forecasts. Finally, prediction success can only

be judged in practice after the model has been used a number of times ex ante.

Our findings based on the BIC suggest that using the earlier data of weeks 3 and 4 of the previous month is statistically acceptable. This makes the Google activity data even more useful, since one gains in practice two weeks for prediction purposes due to their earlier availability. We also find that a more parsimonious specification is justified, since using the BIC the models including k1 and k4 only are doing best in comparison to other or more complex specifications; the BIC also chooses the model using data from weeks 3 and 4 of the previous month against the data from weeks 1 and 2 of the current month. Therefore, the model of the third column of Table 3 is the best, based on statistical grounds. The lagged level variable of unemployment has a negative sign and is significant, and hence there is a stable long-run solution. k1 measuring the process of contacting the unemployment office have a positive and statistically significant impact on unemployment in the short- and long-run. Jobsearch activities measured by k4 predict a strong and significant decline in unemployment in the short-term, but somewhat less strong and significant in the long-run.

Forecasts and realizations of the unemployment rate are shown in Figure 4, and move together quite well. In a few events the forecasts indicate much earlier that there is a change in trend; for instance, the predictions for October to December 2008 were conservative, and they anticipated the turning point to the rise in unemployment early on. However, after a perfect fit in January, the two curves split increasingly in the sequel. Our understanding is that this is a result of a change in labor policy which was announced only during December 2008 and came into effect in January 2009 concerning the role of government supported short-time working already discussed above. The increased interest in short-time work unmeasured in our regression models have likely contributed to the predicted decline in unemployment. To examine this hypothesis in an informal way, we have replaced k1 in our final regression model by the search activity on "Kurzarbeit" (short-time work) and obtained Figure 5 for evaluation. This graph demonstrates that through this variable most of the differences between forecasts and realizations disappear. However, the actual prediction for a decline in future unemployment remains. Please also note that the policy change has been quite recent, and in May, the German labor minister announced an even larger increase in the duration of short-time work.[6] Hence, it is more difficult to adjust the modeling to a realistic approach at this time; we would like to wait for more data points to make a realistic effort to do so. What remains important for the purpose of this paper is that we can demonstrate that the internet

6 Originally, workers were only able to receive the program for 6 months. The increase in the duration of the program in January 2009 was for 18 months, and a further increase to 24 months was decided at the end of April and put into practice on May 1, 2009.

activity data is useful to help predict under complex and fast changing conditions.

Figure 4

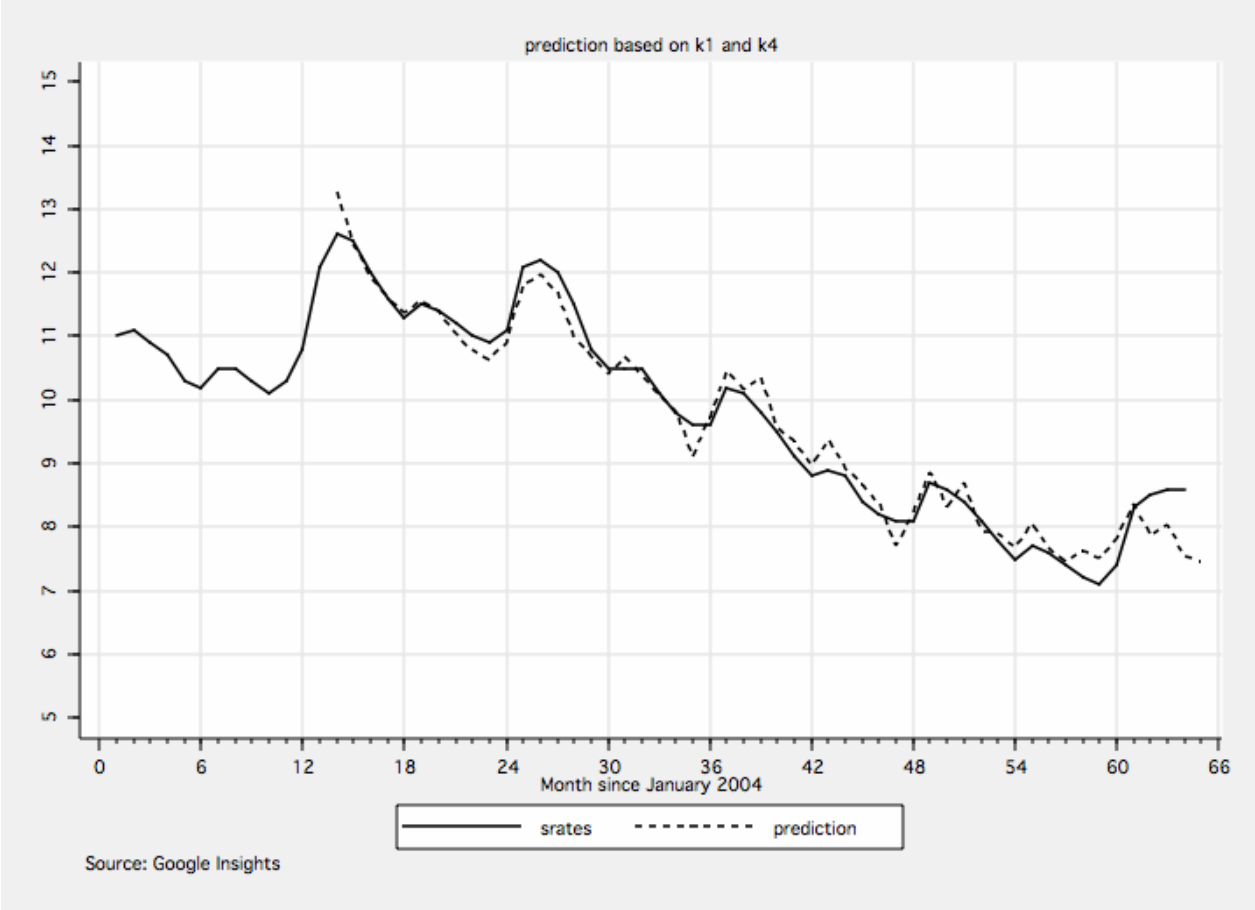
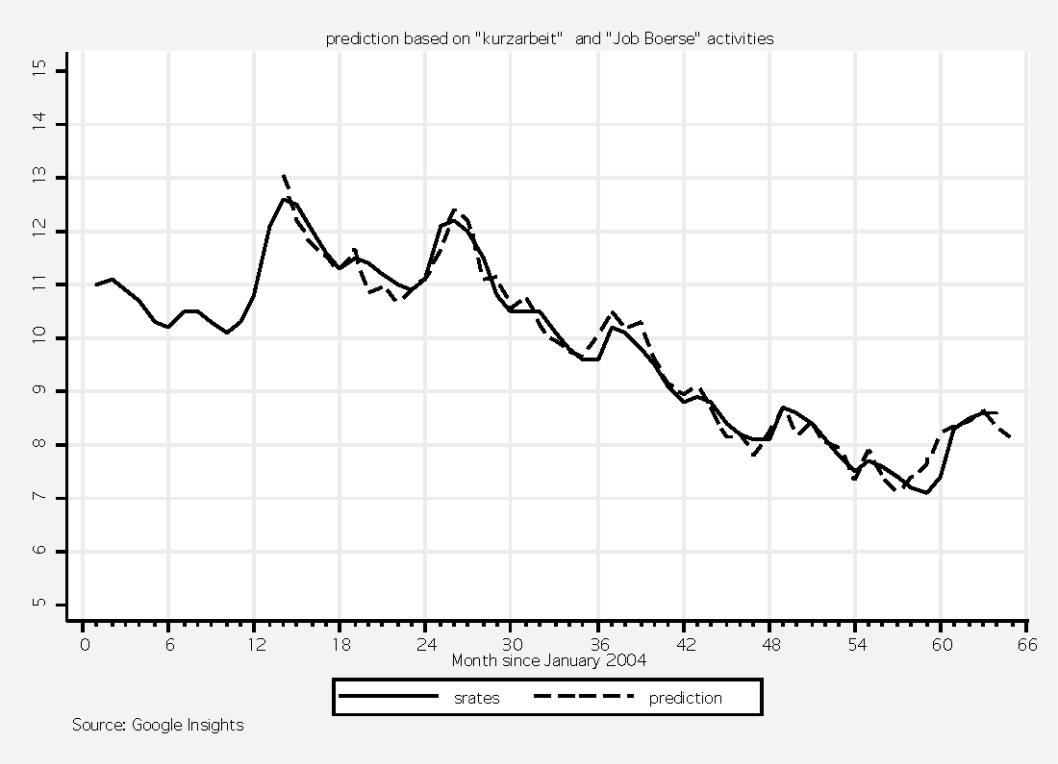


Figure 5



4. Conclusions

The internet contains an enormous amount of information which, to our knowledge, classical econometrics has yet to appropriately tap into. Such information comes timely on a continual basis. It is particularly welcome at times of an economic crisis where the traditional flow of information is too slow to provide a proper basis for sound economic decisions. To examine this potential, this paper has examined the use of internet activity data to predict economic behavior under complex and changing circumstances. Of much interest is when and how, and at what magnitude unemployment is affected after a long period of strong recovery. Therefore, we have suggested an innovative new method of using data on internet activity for that purpose and have demonstrated strong correlations between keyword searches and unemployment rates using monthly German data on a simple and parsimonious level. This suggests that there is a strong potential for the method used, which needs to be further explored.

References:

Constant, A. / Zimmermann, K. F. (2008): Im Angesicht der Krise: US-Präsidentschaftswahlen in transnationaler Sicht, DIW Wochenbericht 44, 688 - 701.

Engle, R. F. / Granger, C. W. J. (1987): Co-Integration and Error Correction: Representation, Estimation, and Testing, Econometrica 55, 251-276.

Ginsberg, J. / Mohebbi, M. H. / Patel, R. S. / Brammer, L. / Smolinski, M. S. / Brilliant, L. (2009): Detecting Influenza Epidemics using Search Engine Query Data, Nature 457, 1012 – 1014.

Greene, W. H. (2008): Econometric Analysis, 6th Edition, Upper Saddle River: Wharton School Publishing.

Zimmermann, K. F. (2008): Schadensbegrenzung oder Kapriolen wie im Finanzsektor?, Wirtschaftsdienst 12, 18 - 20.

Zimmermann, K. F. (2009): Prognosekrise: Warum weniger manchmal mehr ist, Wirtschaftsdienst 2, 86 - 90.

Table 1. Models with two variables involving activity in weeks 1, 2

	w12k1_2 b/t	w12k1_3 b/t	w12k1_4 b/t	w12k2_3 b/t	w12k2_4 b/t	w12k3_4 b/t
L12.srates	-0.387*** (-4.41)	-0.475*** (-5.68)	-0.310*** (-4.60)	-0.611*** (-7.30)	-0.256** (-3.00)	-0.350*** (-4.51)
L12.w12k1	0.037*** (4.50)	0.011 (1.09)	0.021*** (4.25)			
S12.w12k1	-0.010 (-0.92)	-0.019 (-1.65)	-0.000 (-0.06)			
L12.w12k2	0.014*** (3.58)			0.007 (1.62)	0.006 (1.79)	
S12.w12k2	0.011* (2.07)			0.007 (1.24)	0.003 (0.65)	
L12.w12k3		0.028*** (4.86)		0.036*** (7.71)		0.016*** (3.82)
S12.w12k3		0.016* (2.12)		0.013 (1.76)		0.004 (0.86)
L12.w12k4			-0.028*** (-8.09)		-0.036*** (-10.92)	-0.029*** (-7.91)
S12.w12k4			-0.013** (-2.84)		-0.020*** (-3.91)	-0.018*** (-3.91)
_cons	-2.968*** (-3.69)	-0.686 (-0.74)	4.121*** (4.57)	0.522 (0.77)	6.942*** (8.00)	5.435*** (6.17)
N	52	52	52	52	52	52
AIC	94.191	84.961	45.895	95.708	66.811	53.245
BIC	105.899	96.669	57.603	107.415	78.518	64.952
Log Lik.	-41.096	-36.481	-16.948	-41.854	-27.405	-20.622
R^2	0.756	0.795	0.904	0.749	0.856	0.889

Ln and Sn are the nth monthly lag and difference operators respectively. . The variable naming convention is as follows: w12=first monthly half, w34=second monthly half; k1, k2, k3, k4 are the

keywords defined in Section 2. A model which is denoted by eg w12ki_j is one involving the two activity variables in the first monthly halves i and j whereas w34ki_j_l is a model with 3 keywords i, j and l in the second monthly halves. The variable srates is the seasonal unemployment rates. Finally the significance stars mean: * p<0.05; ** p<0.01; *** p<0.001

Table 2. Models with more than two variables involving activity in weeks 1, 2

	w12k1_2_3 b/t	w12k1_2_4 b/t	w12k2_3_4 b/t	w12k1_3_4 b/t	w12k1_2_3_4 b/t
L12.srates	-0.490*** (-5.72)	-0.371*** (-5.17)	-0.359*** (-4.46)	-0.380*** (-5.38)	-0.411*** (-5.64)
L12.w12k1	0.014 (1.32)	0.023*** (4.45)		0.023** (3.19)	0.026** (3.51)
S12.w12k1	-0.016 (-1.38)	0.004 (0.54)		0.008 (0.95)	0.011 (1.28)
L12.w12k2	0.007 (1.82)	0.005 (1.87)	0.003 (0.93)		0.005 (1.70)
S12.w12k2	0.007 (1.45)	-0.000 (-0.15)	0.002 (0.57)		0.001 (0.32)
L12.w12k3	0.022** (3.42)		0.015** (3.36)	0.003 (0.59)	-0.000 (-0.01)
S12.w12k3	0.011 (1.29)		0.003 (0.62)	-0.007 (-1.13)	-0.009 (-1.38)
L12.w12k4		-0.026*** (-7.52)	-0.028*** (-7.50)	-0.027*** (-7.33)	-0.026*** (-7.13)
S12.w12k4		-0.012** (-2.75)	-0.018*** (-3.80)	-0.013** (-2.88)	-0.012** (-2.74)
_cons	-0.964 (-1.03)	3.723*** (4.09)	5.251*** (5.72)	4.068*** (4.64)	3.750*** (4.21)
N	52	52	52	52	52
AIC	84.662	44.280	56.208	43.190	43.395
BIC	100.272	59.890	71.818	58.800	62.908

Log Lik.	-34.331	-14.140	-20.104	-13.595	-11.698
R^2	0.812	0.913	0.891	0.915	0.921

Ln and Sn are the nth monthly lag and difference operators respectively. . The variable naming convention is as follows: w12=first monthly half, w34=second monthly half; k1, k2, k3, k4 are the keywords defined in Section 2. A model which is denoted by eg w12ki_j is one involving the two activity variables in the first monthly halves i and j whereas w34ki_j_l is a model with 3 keywords i, j and l in the second monthly halves. The variable srates is the seasonal unemployment rates. Finally the significance stars mean: * p<0.05; ** p<0.01; *** p<0.001

Table 3. Models with two variables involving activity in weeks 3,4

	w34k1_2 b/t	w34k1_3 b/t	w34k1_4 b/t	w34k2_3 b/t	w34k2_4 b/t	w34k3_4 b/t
L12.srates	-0.156 (-1.95)	-0.378*** (-4.01)	-0.406*** (-6.65)	-0.414*** (-4.51)	-0.429*** (-4.95)	-0.455*** (-6.81)
L13.w34k1	0.000 (0.04)	-0.021* (-2.11)	0.024*** (6.65)			
LS12.w34k1	-0.045*** (-4.80)	-0.039*** (-4.08)	0.016* (2.34)			
L13.w34k2	0.014** (2.95)			0.004 (0.75)	0.012*** (3.66)	
LS12.w34k2	0.032*** (4.71)			0.021* (2.48)	0.013* (2.45)	
L13.w34k3		0.031*** (4.32)		0.020*** (3.58)		0.017*** (6.41)
LS12.w34k3		0.017 (1.58)		-0.010 (-0.83)		0.009 (1.50)
L13.w34k4			-0.031*** (-11.41)		-0.028*** (-10.72)	-0.025*** (-11.79)
LS12.w34k4			-0.014** (-3.46)		-0.010* (-2.19)	-0.013** (-3.40)
_cons	-0.918 (-1.06)	1.600 (1.41)	5.503*** (6.98)	0.482 (0.58)	6.911*** (6.61)	5.977*** (7.41)

N	51	51	51	51	51	51
AIC	102.860	103.445	40.470	112.394	72.194	46.150
BIC	114.451	115.036	52.061	123.985	83.785	57.741
Log Lik.	-45.430	-45.722	-14.235	-50.197	-30.097	-17.075
R^2	0.692	0.688	0.909	0.628	0.831	0.899

Ln and Sn are the nth monthly lag and difference operators respectively. . The variable naming convention is as follows: w12=first monthly half, w34=second monthly half; k1, k2, k3, k4 are the keywords defined in Section 2. A model which is denoted by eg w12ki_j is one involving the two activity variables in the first monthly halves i and j whereas w34ki_j_l is a model with 3 keywords i, j and l in the second monthly halves. The variable srates is the seasonal unemployment rates. Finally the significance stars mean: * p<0.05; ** p<0.01; *** p<0.001

Table 4. Models with more than two variables involving activity in weeks 3, 4

	w34k1_2_3 b/t	w34k1_2_4 b/t	w34k2_3_4 b/t	w34k1_3_4 b/t	w34k1_2_3_4 b/t
L12.srates	-0.321*** (-3.81)	-0.391*** (-6.01)	-0.432*** (-6.39)	-0.432*** (-6.82)	-0.412*** (-6.34)
L13.w34k1	-0.017 (-1.91)	0.022*** (4.84)		0.018** (2.82)	0.016* (2.46)
LS12.w34k1	-0.043*** (-5.11)	0.013 (1.58)		0.014 (1.91)	0.009 (1.14)
L13.w34k2	0.008 (1.76)	0.002 (0.74)	0.003 (0.89)		0.002 (0.60)
LS12.w34k2	0.028*** (3.94)	0.004 (0.93)	0.008 (1.65)		0.007 (1.32)
L13.w34k3	0.018* (2.38)		0.015*** (4.77)	0.005 (0.92)	0.003 (0.66)
LS12.w34k3	-0.004 (-0.37)		0.004 (0.63)	-0.001 (-0.20)	-0.005 (-0.69)
L13.w34k4		-0.030*** (-9.54)	-0.024*** (-11.10)	-0.029*** (-10.08)	-0.027*** (-8.23)

LS12.w34k4		-0.014** (-3.40)	-0.014*** (-3.60)	-0.014** (-3.36)	-0.013** (-3.10)
_cons	1.185 (1.19)	5.199*** (6.04)	5.567*** (6.63)	5.589*** (7.10)	5.188*** (6.12)
N	51	51	51	51	51
AIC	91.379	43.401	46.997	41.494	43.288
BIC	106.834	58.855	62.452	56.949	62.606
Log Lik.	-37.690	-13.700	-15.499	-12.747	-11.644
R^2	0.772	0.911	0.905	0.914	0.918

Ln and Sn are the nth monthly lag and difference operators respectively. . The variable naming convention is as follows: w12=first monthly half, w34=second monthly half; k1, k2, k3, k4 are the keywords defined in Section 2. A model which is denoted by eg w12ki_j is one involving the two activity variables in the first monthly halves i and j whereas w34ki_j_l is a model with 3 keywords i, j and l in the second monthly halves. The variable srates is the seasonal unemployment rates. Finally the significance stars mean: * p<0.05; ** p<0.01; *** p<0.001