



Strathprints Institutional Repository

Ross, Andrew (2013) *Nowcasting with Google Trends : a keyword selection method*. Fraser of Allander Economic Commentary, 37 (2). pp. 54-64. ISSN 2046-5378

Strathprints is designed to allow users to access the research output of the University of Strathclyde. Copyright © and Moral Rights for the papers on this site are retained by the individual authors and/or other copyright owners. You may not engage in further distribution of the material for any profitmaking activities or any commercial gain. You may freely distribute both the url (<http://strathprints.strath.ac.uk/>) and the content of this paper for research or study, educational, or not-for-profit purposes without prior permission or charge.

Any correspondence concerning this service should be sent to Strathprints administrator: <mailto:strathprints@strath.ac.uk>

Nowcasting with Google Trends: a keyword selection method

Andrew Ross ⁱ, Fraser of Allander Institute, University of Strathclyde

Abstract

Search engines, such as Google, keep a log of searches entered into their websites. Google makes this data publicly available with Google Trends in the form of aggregate weekly search term volume. Aggregate search volume has been shown to be able to nowcast (i.e. compute real-time assessment of current activity) a variety of variables such as influenza outbreaks, financial market fluctuations, unemployment and retail sales. Although identifying appropriate keywords in Google Trends is an essential element of using search data, the recurring difficulty identified in the literature is the lack of a technique to do so. Given this, the main goal of this paper is to put forward a method (the “backward induction method”) of identifying and extracting keywords from Google Trends relevant to economic variables.

Introduction

The growing use of the internet has made available a number of new data sources. For example, the increasing use of the internet as an information finding tool has led to the creation of new data sources to measure consumer sentiment and behaviour. Search engine providers, such as Google, keep a record of searches entered in their websites (McLaren & Shanbhogue, 2011). Google has made some of this data available by publicising aggregate search volumes for specific search terms.

Data on search term volume can be used to analyse a variety of issues and variables. For example, the query volume for ‘dishwashers’, ‘fridges’, or ‘flat screen televisions’ can be used to explain demand for durable goods. The major catalyst in this area of research, however, was the research conducted by Ginsberg et al. (2009), who used query volume of influenza and flu related search terms (e.g. flu symptoms) to monitor flu outbreaks in real time.

Given the availability of aggregate internet search data, there is now the possibility to add a further method of economic analysis, which attempts to explain current, rather than future activity. Internet search data is therefore mainly used to provide real time assessment of current activity i.e. to nowcast rather than forecast (Aruoba & Diebold, 2010).

Policy making relies upon the availability of accurate and timely micro, sub-macro, and macro-level data. Yet, the majority of official data is published with a reporting lag of several weeks, and may subsequently even be revised (Choi & Varian, 2009a). Even though there are numerous methods and econometric models employed to provide for timely economic analysis, the lag in official data may delay and distort rational policy making. Real-time policy making is particularly significant in times of structural change or economic uncertainty where the predictive power of models can break down (Castle et al., 2009).

At such times, it is necessary to obtain timely high-frequency data which remains robust during structural changes. Thus, the main advantage of using internet search data is that it is made available without lags (maximum of one week) whilst covering a representative sample of the population. Internet search data is therefore mainly used to provide real time assessment of current activity.

Using internet search terms to nowcast economic variables requires the selection of explanatory keywords. Although economic intuition can be used to identify keywords explaining sales of flatscreen TVs, for example, when it comes to nowcasting complex economic variables such as intuition, however, may not be sufficient. Thus, the main difficulty when using search term data is the selection of individual search terms in Google Trends (GT) ⁱⁱ that are significant in explaining the economic variable investigated.

Google Trends

In 2008 Google made aggregate query data freely available through GT. Aggregate search logs (query volumes) are accessible and downloadable from 2004:M1 onwards. Queries are 'broad matched' so, for example, queries such as 'Strathclyde University' are accounted for in the query index for 'Strathclyde'. Thus, it must be stressed that relying on single keywords may not yield robust findings as the data used may be contaminated (or contain significant amount of noise) by queries unrelated to the topic investigated.

GT data is available as a query share index. That is, the total popularity of a search term is determined by the volume of a search term in a geographic region, divided by the total number of queries within the same location and time parameters (op.cit.).

GT data is normalised i.e. no absolute data is given, so that regions generating the highest search volumes do not always rank on top. Google divides the sets of data by a common variable to cancel out the effect of the variable on the data. Following the normalisation, each data point on the graph is divided by the highest value and then multiplied by one hundred. GT therefore provides aggregate data that is normalised and made available on a scale from zero to one hundred. If required, GT data can be further filtered by time period, geographically, and/or by category (Google, 2011).

GT data 'indicates the likelihood of a random user to search for a particular search term from a certain location at a certain time' (op. cit.). With GT data being computed on a daily basis by a sampling method, results vary from day to day by a few per cent (op.cit.). This adds additional noise to the data due to the sampling errors. GT data is made available on a weekly basis (a week being from one Sunday to the next) giving a maximum lag of one week.

Even though GT data is widely considered to be unbiased (i.e. unknowingly provided by Google users) it is prone to manipulation. This weakness has not been identified in the current literature. It is extremely difficult to identify 'unnatural' changes in search volumes as they can be due to public campaigns, changes in trends, or due to automated queries i.e. manipulation by automated queries submitted by "robots". Thus, Goodhart's lawⁱⁱⁱ has to be taken into account when using GT to underpin policy decisions.

With widespread use of the internet being a rather new phenomenon, data obtained from search engines has a short back-run compared to other economic indicators. Also, internet usage tends to remain highly correlated with factors such as age and income, leaving the sample not representative of the population. Even though Google holds a large proportion of the search engine market, the sample could still be skewed by the fact that different users prefer different search engines. Also, different users interested in the same topic use different search queries (and vice versa), and queries can be made with entirely different intentions. Therefore significant noise in the search data can be present through e.g. queries made out of curiosity rather than the intent to take the according action.

It must also be kept in mind throughout that whilst query volumes are good indicators of future consumer activities, such as attending movies or purchasing video games, there is wide variability in the predictive power of query data (Goel et al., 2010). Even though several possible reasons for this have been identified, such as size of relevant population and making searches to inform rather than take action, this area of research still remains largely unexplored and unexplained.

As identified by Askitas and Zimmermann (2009), and others, keywords used underlay significant dynamicity, i.e. keywords and websites searched for may come into and go out of existence. Moreover, search behaviour is a constantly evolving process where for example, search patterns can be predominantly 'one word' queries today, and 'multiple word' queries tomorrow. Keyword dynamicity is impacted by, among other things, generational patterns, linguistic developments, and social and economic levels.

Given the limitations and caveats identified, GT datasets used in research would either have to be constantly evolving where search queries are constantly added/removed, or more realistically, a core set of keywords must be identified which have the power to predict/nowcast selected economic variables. This again emphasises the need for a reliable process in identifying keywords.

What can query data "predict"

Research in the field of epidemiology, where researchers were able to link query data with influenza outbreaks (e.g. Ginsberg et al., op.cit.; Chan et al., 2011), provided the foundations. Their findings

suggest that the relative frequency of certain queries (e.g. related to influenza symptoms) correlates with the percentage of physician visits in the US. Also, they found that it is possible to accurately estimate the current level of influenza activity, with a reporting lag of one day, when using GT data as an independent variable. This research was used to make available estimates of flu^{iv} and dengue^v trends around the world.

These findings and the newly available data through GT stimulated research in the area of Economics. Choi and Varian (2009b) provided significant findings when using GT data to “predict” economic unemployment, retail sales, automotive sales, home sales and travel plans. Their research found that GT data does not necessarily predict the future, but it does help to predict the present i.e. to nowcast. Their nowcasting model and the findings from Ginsberg et al. (op. cit.) provided a catalyst for further research.

Building upon these findings, a large body of research has evolved using GT data to nowcast a vast array of diverse variables. Amongst others, GT data was used to nowcast trading volatility (Vlastakis & Markellos, 2012), consumer sentiment indices (Penna & Huang, 2009), private consumption (Schmidt & Vosen, 2012), and inflation expectations (Guzman, 2011).

Importantly for this research paper, Askitas and Zimmermann (op. cit.) examined the correlations between keyword searches and unemployment rates. The research found a strong correlation, and suggested that GT data is particularly useful in times of economic crisis where decision makers require faster flows of information. Similar findings were made by McLaren and Shanbhogue (op. cit.) and Baker and Fradkin (2011), who also found that employment related queries contain relevant information for explaining changes in the labour market.

The recurring difficulty identified in the literature is the lack of a scientific technique to identify appropriate keywords in GT. For example, McLaren and Shanbhogue (op. cit.) point out that deciding which queries to consider is a crucial element of using search data.

The nowcasting methodology

Bank of England researchers suggest using GT data for the keyword ‘jsa’ (a short form of the UK labour market assistance programme ‘Jobseekers Allowance’) to nowcast unemployment (McLaren and Shanbhogue, op. cit.). The keyword ‘jsa’ is subsequently used as a benchmark when assessing the power of the “backward induction method” in selecting statistically significant keywords from GT.

Monthly UK unemployment data $\{U\}$ was sourced from the ONS (2012) with the time-frame analysed being between 2004:M1 and 2012:M4. Weekly GT query volume for the keyword ‘jsa’ was downloaded, for the same time period, restricted to UK data only. GT data was aggregated from weekly to monthly observations. GT weekly frequencies are in sets of Sunday to Sunday, so that some weeks overlap two months and findings may therefore contain some additional noise.

The baseline model (1) is set up as a simple autoregressive model, where only changes in unemployment in previous months $\{U_{t-1}\}$ and $\{U_{t-2}\}$ are used as explanatory variables. Monthly GT data $\{x_t\}$ for the keyword ‘jsa’ is then added (2), and compared to the baseline model.

$$\log(U_t) = \alpha + \beta_1 \log(U_{t-1}) + \beta_2 \log(U_{t-2}) + \varepsilon_t \quad (1)$$

$$\log(U_t) = \alpha + \beta_1 \log(U_{t-1}) + \beta_2 \log(U_{t-2}) + \phi x_t + \varepsilon_t \quad (2)$$

To measure the fit of the model, in-sample criteria are used (adjusted R^2 and AIC). The model providing the better fit has the higher adjusted R^2 and the lower AIC. Out-of-sample observations of forecasts and forecast errors are used to determine whether GT data helps to predict the variable investigated. To determine this, a series of one-month ahead predictions are made and the prediction errors are computed. From this, the RMSE and the MAE are computed. The preferred model is the one with the smallest out-of-sample RMSE or MAE.

The unemployment regression results for model (1) and (2) are summarised in **Table A**. GT data for the keyword ‘jsa’ has not been found to be statistically significant in explaining changes in unemployment.

Also, improvements in error reduction and adjusted R^2 are only marginal compared to the baseline model. It must be noted, however, that sample errors in GT and the aggregation of the data may have caused these results to be statistically insignificant. This is further explored in the following sections

where different unemployment indicators and additional keywords (selected by the “backward induction” method) are tested.

Table A: Unemployment regression results

| Independent variables | Baseline | ‘jsa’ |
|----------------------------------|---------------------|---------------------|
| α | 0.02791 (0.514) | 0.03024 (0.622) |
| $\log(U_{t-1})$ | 1.44649 (0.000) | 1.44629 (0.000) |
| $\log(U_{t-2})$ | -0.44976 (0.000) | -0.44989 (0.000) |
| x_t | - - | 0.00001 (0.960) |
| Adjusted R^2 | 0.99643 | 0.99639 |
| AIC | -557.53450 | -555.53610 |
| MAE | 0.01020 | 0.01020 |
| RMSE | 0.01365 | 0.01365 |

P-values for heteroskedasticity robust standard errors (HC3) are shown in parentheses.

Keyword selection process

GT categories provide a strong starting point when selecting keywords to nowcast economic variables. These categories are classifications of industries or markets, and are commonly referred to as verticals (Google, 2012a). The category: All Categories > Jobs and Education > Jobs, for example, includes keywords such as ‘jobs’, ‘resume’ and ‘careers’.

Additionally, Google Correlate^{vi} can be used (essentially GT in reverse). In GT a specific query is typed in to obtain a time-series dataset of query activity. In contrast, in Google Correlate a data series can be entered (the target) to obtain a list of queries whose data series follows a similar pattern, i.e. correlates (Google, 2012b).

There are, however, situations where both GT categories and Google Correlate fail to suggest relevant keywords. In this case, a third method, the “backward induction method” is suggested. It must, however, be emphasised that this method should not be considered as the panacea to the keyword selection problem as it should be used in addition to the methods previously outlined. More specifically, the appropriate method depends solely on the needs of the researcher and the variable investigated.

Backward induction (generally used within Game Theory) is the process of reasoning backwards, starting from the end of a problem or situation. This backward reasoning can be applied to the keyword selection process, where the approach is taken that relevant keywords have already been selected. That is, keywords have been selected by people using search engines, and these simply need to be identified and extracted.

People searching for websites, for example, to find employment related websites, will search for ‘jobs’ or ‘career’ for example, to then be presented with a website (e.g. website *A*) offering the requested products or services. Reasoning backwards, top referring keywords (top keywords used by people to find a specific website) from website *A* can therefore be extracted and used to obtain variable relevant GT data.

Instead of selecting keywords by economic intuition, this approach extracts top keywords employed by search engine users in trying to find specific goods/services/information. This ensures that these keywords are actually being used, and secondly ensures that these are relevant to the economic variable investigated. This is, however, best outlined by means of an example. In the following, an example is given, assuming the need to identify keywords relevant in explaining the job search market.

The first step would be to select a representative number of dominant websites within the area investigated. This can be done by using directory services such as Open Directory Project, Yahoo

Directory or Alexa categories. Alternatively a search for ‘jobs’, for example, in Google will present a number of relevant websites.

Following the identification of dominant job search related websites, top keywords employed by users to find these websites can be extracted. Keyword extraction can be done by means of several online services (some of which require subscription). The ones tested within this research were Alexa, Semrush, Sistrix, and the AdWords Keyword Tool. Keywords extracted from Sistrix (2012) seemed to be most promising as they did not contain a large amount of noise created by non-relevant, or domain related keywords. Also, in contrast to Alexa and Semrush, Sistrix provides for both subpage and subdomain keyword information.

Being able to identify keywords used for subpages and subdomains is a significant advantage and is indispensable when using backward induction and GT data to nowcast economic variables. Being able to extract these keywords allows identification of specific keywords used to find subpages (e.g. <http://direct.gov.uk/en/Employment/Jobseekers/>), instead of extracting keywords used to find the main webpage (e.g. <http://direct.gov.uk>). This allows for more topic-specific keyword extraction.

As such, keywords used to find the UK’s Job Seekers Money, tax and benefits website^{vii} can be obtained through Sistrix (2012). **Table B** summarises keywords which seemed promising in explaining changes in unemployment.

Table B: Unemployment related keywords

| | | | | |
|-------------------|-------------------------|------------------------|---------------|--------------|
| ‘made redundant’ | ‘job seekers allowance’ | ‘jobseekers allowance’ | ‘job centre’ | ‘jobcentre’ |
| ‘job centre plus’ | ‘unemployment benefits’ | ‘employment support’ | ‘job seekers’ | ‘jobseekers’ |

Nowcasting Unemployment

Keywords extracted using the backward induction method were added to the model derived in the previous section (see equation 1 and 2). To reduce data volatility, and thereby provide more stable and robust findings, the average of the data generated is taken from keyword data downloaded on seven consecutive days. Thus, GT data was obtained for the time period 2004:M1 to 2012:M4, restricted to UK data only, downloaded once a day for seven consecutive days starting on July 26, 2012. Importantly, the data for each keyword was downloaded individually instead of downloading the maximum of five keywords at a time, as the dominant keyword degrades the query volume of the less dominant keywords.

Regression results are summarised in **Appendix A**. Similarly to what was found in the previous section, the results are not very promising. The keywords ‘made redundant’, ‘job seekers allowance’, ‘jobseekers allowance’ attained the highest significance, within this data set, of only 5 per cent. The keyword ‘job centre’ attained a significance of 10 per cent, whilst the remaining keywords are statistically insignificant in explaining unemployment growth (including the hurdle keyword ‘jsa’).

Most noteworthy, however, is that the majority of keywords identified using backward induction outperform the baseline model, and the second hurdle set by the keyword ‘jsa’ in terms of significance and out-of-sample nowcasting ability. This makes a strong case for the backward induction method in its ability to identify and extract a set of relevant keywords.

Nowcasting the Claimant count

With unemployment results lacking robustness, the same keywords were applied to nowcast an alternative unemployment measure. That is, the Claimant count, which measures the number of people claiming unemployment-related benefits, is analysed. This data set was obtained from the ONS (2012) for the time period 2004:M1 to 2012:M4. The model outlined above (see equation 1 and 2) is applied, where only changes in the Claimant count $\{CC\}$ in previous months $\{CC_{t-1}\}$ and $\{CC_{t-2}\}$ are used as explanatory variables. GT data for each of the keywords $\{x_t\}$ is then added separately.

Regression results are summarised in **Appendix B**. All keywords (except ‘jsa’ and ‘job centre’) attained a significance of at least 10 per cent. The keywords ‘made redundant’, ‘job centre plus’ and ‘employment support’ are significant at a 1 per cent level, thus providing robust results. Moreover, the out-of-sample results also show that the majority of selected keywords were able to produce smaller errors than the

baseline model. Therefore, the selected unemployment-related keywords are able to explain a significant amount of changes in the Claimant count.

With the exception of 'job centre' all keywords identified using backward induction outperformed the baseline model and also the additional hurdle set by the keyword 'jsa' in terms of significance and also, partially, in the out-of-sample testing. This again underpins the ability of the backward induction method to provide robust and significant keywords.

The tests are repeated using GT data for the first week of each month to assess whether it is possible to forecast the monthly Claimant count using GT data for only the first week of the month. The Claimant count is the t -th month, denoted as $\{y_t: t = 1, 2, \dots, T\}$ and the GT data is the k -th week of the month, denoted as $\{x_t^{(k)}: t = 1, 2, \dots, T; k = 1, \dots, 4\}$.

Regression results are summarised in **Appendix C**. The majority of keywords attained the minimum significance of 10 per cent. The keywords 'jobseekers allowance', 'jobcentre' and 'job centre plus' attained a significance level of 1 per cent. Keywords 'made redundant' and 'employment support' attained a 5 per cent significance, 'job seekers' and 'jobseekers' attained a 10 per cent significance, whilst the remaining keywords continued to be statistically insignificant in explaining Claimant count growth.

The significant keywords showed strong results within the in-sample and out-of-sample tests of predictability. This indicates that GT data for the selected keywords for only the first week of a month contain a significant amount of information to enable forecasts of the monthly Claimant count.

Summary

With growing use of the internet as an information finding tool, new data sources such as GT have become very appealing for policy makers, for example, as a proxy to monitor economic activity and sentiment. Even though there is no agreement in the literature on the ability of GT data to "predict the future", there is unanimous agreement that GT data is highly useful in nowcasting economic variables.

It was found, however, that the significance of GT data may be limited due to the short back-run, and the amount of noise the data contains due to the sampling method employed by Google. Within the results of this research it was, however, found that large amounts of data volatility, due to sampling errors, can be reduced by downloading GT data for each keyword individually, and over several consecutive days.

The major recurring difficulty identified in the literature is the lack of a technique to identify appropriate keywords in GT. The selection of keywords is, however, a crucial element of using search data. Currently, keywords are mostly selected in GT on the basis of economic intuition, rather than by following a set of strategies or guidelines.

Thus, the core of this paper describes and tests the backward induction method which identifies relevant keywords by extracting these from variable relevant websites. To evaluate and examine this method, this research tested the keywords identified using the backward induction method against keywords identified in the literature review (the benchmark).

This backward induction method was applied to nowcast UK unemployment growth using a small set of keywords. The majority of keywords identified using the backward induction method outperformed the baseline model and the benchmark in terms of in-sample and out-of-sample tests of predictability indicating that the backward induction method is effective in identifying relevant keywords.

When nowcasting unemployment growth, it was found that several keywords (including the benchmark keyword) lacked robustness in terms of statistical significance. To provide further evidence that the backward induction method is applicable, the same set of keywords was successfully tested to nowcast growth in the monthly UK Claimant count. Notably, the initial research was also able to successfully nowcast house price inflation and individual insolvencies using the backward induction method. This has shown that, even though relevant keywords can now be extracted using the backward induction method, the issue still remains that the right questions have to be asked, using the right model, and the right data.

"Prediction is very difficult, especially if it's about the future" - Niels Bohr (1885-1962)

Acknowledgements

The initial research was supervised by Professor G. Koop^{viii} and was submitted in 2012 in partial fulfilment of the requirements for the degree of MSc in Economic Management and Policy in the Department of Economics at the University of Strathclyde.

Appendix A
Unemployment regression results

| Independent variables | Baseline | made redundant | job seekers allowance | jobseekers allowance | job centre | Jobcentre | job centre plus | unemployment benefits | employment support | job seekers | jobseekers | isa |
|-------------------------------------|---------------------|---------------------|-----------------------|----------------------|---------------------|---------------------|---------------------|-----------------------|---------------------|---------------------|---------------------|---------------------|
| α | 0.02791 (0.514) | 0.22886 (0.015) | 0.15850 (0.029) | 0.15189 (0.020) | 0.02813 (0.507) | 0.03649 (0.475) | 0.03085 (0.564) | 0.04872 (0.403) | 0.01867 (0.876) | 0.08329 (0.144) | 0.08190 (0.113) | 0.03024 (0.622) |
| $\log(U_{t-1})$ | 1.44649 (0.000) | 1.32114 (0.000) | 1.38966 (0.000) | 1.37823 (0.000) | 1.39349 (0.000) | 1.44726 (0.000) | 1.44696 (0.000) | 1.42505 (0.000) | 1.44682 (0.000) | 1.43116 (0.000) | 1.42762 (0.000) | 1.44629 (0.000) |
| $\log(U_{t-2})$ | -0.44976 (0.000) | -0.35207 (0.003) | -0.41147 (0.000) | -0.39940 (0.000) | -0.39785 (0.001) | -0.45182 (0.000) | -0.45066 (0.000) | -0.43145 (0.000) | -0.44882 (0.000) | -0.44234 (0.000) | -0.43864 (0.000) | -0.44989 (0.000) |
| x_t | - (0.000) | 0.00034 (0.014) | 0.00022 (0.044) | 0.00028 (0.015) | 0.00013 (0.090) | 0.00003 (0.803) | 0.00001 (0.938) | 0.00010 (0.411) | -0.00001 (0.932) | 0.00010 (0.250) | 0.00012 (0.147) | 0.00001 (0.960) |
| Adjusted R^2 | 0.99643 | 0.99666 | 0.99650 | 0.99656 | 0.99650 | 0.99639 | 0.99639 | 0.99642 | 0.99639 | 0.99642 | 0.99643 | 0.99639 |
| Akaike information criterion | -557.53450 | -563.24250 | -558.57850 | -560.41960 | -558.71310 | -555.60110 | -555.54090 | -556.47780 | -555.53890 | -556.29570 | -556.79530 | -555.53610 |
| MAE | 0.01020 | 0.00994 | 0.01012 | 0.01005 | 0.01035 | 0.01018 | 0.01019 | 0.01006 | 0.01020 | 0.01016 | 0.01011 | 0.01020 |
| RMSE | 0.01365 | 0.01312 | 0.01344 | 0.01331 | 0.01343 | 0.01364 | 0.01365 | 0.01358 | 0.01365 | 0.01360 | 0.01356 | 0.01365 |

P-values for heteroskedasticity robust standard errors (HC3) are shown in parentheses.

Appendix B
Claimant count regression results

| Independent variables | Baseline | made redundant | job seekers allowance | jobseekers allowance | job centre | Jobcentre | job centre plus | unemployment benefits | employment support | job seekers | jobseekers | jsa |
|-------------------------------|------------|----------------|-----------------------|----------------------|------------|------------|-----------------|-----------------------|--------------------|-------------|------------|------------|
| α | 0.04317 | 0.17382 | 0.14809 | 0.12314 | 0.04117 | 0.07450 | 0.07809 | 0.06343 | 0.19516 | 0.11253 | 0.08826 | 0.08122 |
| | (0.047) | (0.017) | (0.021) | (0.015) | (0.061) | (0.007) | (0.005) | (0.015) | (0.005) | (0.026) | (0.014) | (0.050) |
| $\log(CC_{t-1})$ | 1.85895 | 1.72286 | 1.80949 | 1.80078 | 1.84959 | 1.85564 | 1.85854 | 1.80773 | 1.83853 | 1.83124 | 1.83859 | 1.85135 |
| | (0.000) | (0.000) | (0.000) | (0.000) | (0.000) | (0.000) | (0.000) | (0.000) | (0.000) | (0.000) | (0.000) | (0.000) |
| $\log(CC_{t-2})$ | -0.86500 | -0.74863 | -0.83195 | -0.81967 | -0.85568 | -0.86707 | -0.87042 | -0.81744 | -0.86729 | -0.84825 | -0.85182 | -0.86336 |
| | (0.000) | (0.000) | (0.000) | (0.000) | (0.000) | (0.000) | (0.000) | (0.000) | (0.000) | (0.000) | (0.000) | (0.000) |
| x_t | - | 0.00033 | 0.00024 | 0.00025 | 0.00004 | 0.00016 | 0.00016 | 0.00019 | 0.00027 | 0.00017 | 0.00014 | 0.08122 |
| | - | (0.043) | (0.049) | (0.040) | (0.395) | (0.007) | (0.003) | (0.054) | (0.007) | (0.063) | (0.034) | (0.050) |
| Adjusted R² | 0.99831 | 0.99848 | 0.99837 | 0.99838 | 0.99830 | 0.99835 | 0.99835 | 0.99838 | 0.99839 | 0.99835 | 0.99833 | 0.99831 |
| AIC | -594.56720 | -604.50110 | -597.59240 | -598.20480 | -592.95350 | -596.23760 | -596.13170 | -597.74390 | -598.31490 | -596.06220 | -595.27210 | -593.81200 |
| MAE | 0.00676 | 0.00637 | 0.00650 | 0.00643 | 0.00676 | 0.00634 | 0.00635 | 0.00687 | 0.00638 | 0.00655 | 0.00654 | 0.00665 |
| RMSE | 0.01130 | 0.01063 | 0.01101 | 0.01098 | 0.01128 | 0.01109 | 0.01110 | 0.01100 | 0.01097 | 0.01110 | 0.01114 | 0.01123 |

P-values for heteroskedasticity robust standard errors (HC3) are shown in parentheses.

Appendix C

Claimant count regression results (using $x_t^{(1)}$ only)

| Independent variables | Baseline | made redundant | job seekers allowance | jobseekers allowance | job centre | Jobcentre | job centre plus | unemployment benefits | employment support | job seekers | jobseekers | jsa |
|-------------------------------|---------------------|---------------------|-----------------------|----------------------|---------------------|---------------------|---------------------|-----------------------|---------------------|---------------------|---------------------|---------------------|
| α | 0.04317 (0.047) | 0.12873 (0.013) | 0.13600 (0.044) | 0.15137 (0.002) | 0.04311 (0.049) | 0.07700 (0.005) | 0.07869 (0.005) | 0.05411 (0.023) | 0.22828 (0.011) | 0.11494 (0.042) | 0.08971 (0.021) | 0.08181 (0.088) |
| $\log(CC_{t-1})$ | 1.85895 (0.000) | 1.76183 (0.000) | 1.82345 (0.000) | 1.79064 (0.000) | 1.84814 (0.000) | 1.85732 (0.000) | 1.85975 (0.000) | 1.83246 (0.000) | 1.83845 (0.000) | 1.83675 (0.000) | 1.84075 (0.000) | 1.85467 (0.000) |
| $\log(CC_{t-2})$ | -0.86500 (0.000) | -0.78078 (0.000) | -0.84397 (0.000) | -0.81392 (0.000) | -0.85461 (0.000) | -0.86909 (0.000) | -0.87167 (0.000) | -0.84051 (0.000) | -0.87216 (0.000) | -0.85408 (0.000) | -0.85416 (0.000) | -0.86673 (0.000) |
| $x_t^{(1)}$ | - (0.040) | 0.00021 (0.040) | 0.00020 (0.101) | 0.00031 (0.003) | 0.00005 (0.454) | 0.00016 (0.006) | 0.00015 (0.006) | 0.00010 (0.127) | 0.00032 (0.019) | 0.00017 (0.094) | 0.00013 (0.055) | 0.00011 (0.256) |
| <i>Adjusted R²</i> | 0.99831 | 0.99837 | 0.99833 | 0.99841 | 0.99829 | 0.99834 | 0.99834 | 0.99833 | 0.99842 | 0.99833 | 0.99833 | 0.99830 |
| <i>AIC</i> | -594.56720 | -597.64120 | -595.23420 | -599.77380 | -592.93640 | -595.73600 | -595.51550 | -594.72290 | -600.22300 | -594.97540 | -594.85670 | -593.42810 |
| <i>MAE</i> | 0.00676 | 0.00643 | 0.00660 | 0.00622 | 0.00673 | 0.00632 | 0.00633 | 0.00673 | 0.00624 | 0.00660 | 0.00647 | 0.00662 |
| <i>RMSE</i> | 0.01130 | 0.01101 | 0.01115 | 0.01089 | 0.01128 | 0.01112 | 0.01113 | 0.01118 | 0.01087 | 0.01116 | 0.01117 | 0.01125 |

P-values for heteroskedasticity robust standard errors (HC3) are shown in parentheses.

References

- Aruoba, S., & Diebold, X. (2010). Real-time macroeconomic monitoring: Real activity, inflation, and interactions. *American Economic Review*, 100(2), 20–24.
- Askatas, N., & Zimmermann, F. (2009). Google econometrics and unemployment forecasting. *Applied Economics Quarterly*, 55(2), 107–120.
- Baker, S., & Fradkin, A. (2011). What drives job search? evidence from Google search data. *Stanford Institute for Economic Policy Research, Discussion Papers*, 10(020), 1–45.
- Castle, J., Fawcett, W., & Hendry, F. (2009). Nowcasting is not just contemporaneous forecasting. *National Institute Economic Review*, 210(1), 71–89.
- Choi, H., & Varian, H. (2009a). Predicting the present with Google Trends. *Google Inc.*, 1–20.
- Choi, H., & Varian, H. (2009b). Predicting initial claims for unemployment insurance using Google Trends. *Google Inc.*, 1–5.
- Ginsberg, J., Mohebbi, H., Patel, S., Brammer, L., Smolinski, S., & Brilliant, L. (2009). Detecting influenza epidemics using search engine query data. *Nature*, 457, 1012–1014.
- Goel, S., Hofman, M., Lahaie, S., Pennock, M., & Watts, J. (2010). Predicting consumer behaviour with web search. *PNAS*, 107(41), 17486–17490.
- Google. (2011). Trends Help [Online]. Retrieved 04/06/2012, from <https://support.google.com/trends/>
- Google (2012a). Categories [Online]. Retrieved 06/07/2012, from <https://support.google.com/insights/bin/topic.py?hl=en-GB&topic=19357>
- Google (2012b). Google correlate faq [Online]. Retrieved 16/07/2012, from <https://www.google.com/trends/correlate/faq>
- Guzman, G. (2011). Internet search behaviour as an economic forecasting tool: The case of inflation expectations. *The Journal of Economic and Social Measurement*, 36(3), 3–67
- McLaren, N., & Shanbhogue, R. (2011). Using internet search data as economic indicators. *Bank of England Quarterly Bulletin*, 2011(Q2), 134–140.
- ONS (2012). *Datasets and reference tables* [Online]. Retrieved 24/06/2012, from <http://www.ons.gov.uk/ons/datasets-and-tables/index.html>
- Penna, D., & Huang, H. (2009). Constructing consumer sentiment index for U.S. using internet search patterns. *University of Alberta, Department of Economics*, 2009(26), 1–22.
- Schmidt, T., & Vosen, S. (2012). A monthly consumption indicator for Germany based on internet search query data. *Applied Economics Letters*, 19, 683–687.
- Sistrix (2012a). Tools [Online]. Retrieved 18/07/2012, from <https://tools.sistrix.co.uk/>
- Vlastakis, N., & Markellos, N. (2012). Information demand and stock market volatility. *Journal of Banking and Finance*, 36(6), 1808–1821.

ⁱ Andrew Ross: <http://andrewross.de>

ⁱⁱ Google Trends: <https://www.google.com/trends/>

ⁱⁱⁱ When a measure becomes a target, it ceases to be a good measure (Goodhard, 1975).

^{iv} Google Flu Trends: <https://www.google.org/flutrends/>

^v Google Dengue Trends: <https://www.google.org/denguetrends/>

^{vi} Google Correlate: <https://www.google.com/trends/correlate/>

^{vii} Directgov Money, tax and benefits: <http://www.direct.gov.uk/en/moneytaxandbenefits/>

^{viii} Professor Gary Koop: <http://personal.strath.ac.uk/gary.koop/>