Using internet search data as economic indicators

By Nick McLaren of the Bank's Conjunctural Assessment and Projections Division and Rachana Shanbhogue of the Bank's Structural Economic Analysis Division.[1]

Data on the volume of online searches can be used as indicators of economic activity. This article examines the use of these data for labour and housing markets in the United Kingdom. These data provide some additional information relative to existing surveys. And with further development, internet search data could become an important tool for economic analysis.

The increasingly widespread use of the internet by both businesses and consumers has led to the creation of a potentially useful data source: information on internet search behaviour. Search engine providers keep a record of the searches entered on their website. Some of this information has been made publicly available, enabling users to track the popularity of an extensive range of search terms. This vast database could be used to analyse various issues. For example data on searches for 'flatscreen televisions' and 'fridges' could help to analyse how demand for durable goods has changed over time.

Internet search data have the potential to be useful for economic policy making. Monitoring current economic activity closely is an important aspect of policymaking, but official economic statistics are generally published with a lag. Consumer and business surveys, which are published more quickly than official counterparts, have typically been used to monitor current activity. This type of analysis is often called 'nowcasting', since it tries to explain current, rather than forecast future, activity.

Numerous articles have already been published exploring the use of internet search data as economic indicators. This was initiated by Choi and Varian (2009a) who illustrated its use for predicting US retail sales, automotive sales, home sales and trends in travel destinations. In a preliminary study for the United Kingdom, Chamberlain (2010) finds that search terms are well correlated with disaggregated retail sales data. While the literature suggests that internet search data may be useful for nowcasting, comparisons against traditional survey indicators have yet to be made.

This article explores how internet search data can be used, now and in the future, to enhance understanding of the economy. It builds on the previous literature by evaluating the features of the data and by considering whether internet search data contain information over and above existing survey indicators for the UK housing and labour markets. The first section outlines the potential benefits, and some of the problems, of internet search data. The second section briefly describes the available internet search data. The third section applies the search data to analysis of the labour and housing markets, comparing their performance to existing survey indicators. The final section considers the potential of these data.

The potential benefits and problems of internet search data

Internet search data have a number of appealing properties as economic indicators. They are extremely timely and cover a potentially vast sample of respondents (approximately 60% of the adult population in the United Kingdom now use the internet every day).[2] In contrast to most traditional survey methods, they are collected as a by-product of normal activity, rather than requiring individuals or firms to respond to survey questions after the event. This can avoid problems associated with non-response or inaccurate responses. And it also means that information is continually collected on a wider range of issues, rather than just on a few pre-determined questions. As a result, search data can help analyse issues that arise unexpectedly.

In spite of these benefits, there remain difficulties with using these data. Widespread internet use is a relatively new phenomenon, so the data have a short backrun compared to other economic indicators. Internet use remains highly correlated with factors such as age and income, so the sample may not be representative. There are also issues surrounding the way search engines are used. Different users interested in the same topic could enter entirely different search queries.

(1) The authors would like to thank Hal Varian for his advice on using Google search data, and Madeleine Warwick for her help in producing this article.
(2) In 2010, 30.1 million adults used the internet every day or nearly every day (Office for National Statistics (ONS) (2010)).

Using the Google Insights for Search data set

Google data on search volumes are freely available from www.google.com/insights/search. This application allows the user to compare the popularity of search terms of their choice. The ability to track the popularity of such a wide range of search terms makes this the most suitable data source for this type of study. The comparison can be narrowed according to the country or region from which the internet search was made, and to a specific period in time. The data are extremely timely. Search data are available back to 2004.

The popularity of each search is reported as a weekly index.(1) This index is calculated by dividing the number of searches that include the query term by the total number of online search queries submitted during the week (since search volumes have risen over time, this controls for the upward trend). This fraction is then normalised so that its maximum value over the period is set to equal 100, and the rest of the series is scaled appropriately. There is no information on the actual number of searches, so there is a limit to how these data can be used.

The reported weekly index is based on a random sample of all searches conducted on Google. A new sample is drawn when the user enters a query. The query is stored for a day, so repeated queries for the same term on the same day will return the same results. But data for the same search term conducted on different days could differ. Since this introduces volatility, this article uses a more stable data set by taking the average of the data generated on seven consecutive days.

As the economic variables of interest are reported at a monthly frequency, the weekly index obtained from the database is not directly comparable. Comparison is complicated because some weeks overlap two months. To overcome this, the data used in this article are first transformed into an implied daily series by assigning each day of the week the same value, and then these daily values are aggregated into a monthly average. For certain search terms there are clear seasonal patterns. So the calculated monthly indices are seasonally adjusted using a standard census X-12 procedure.

(1) The data include searches for a particular term even if it is searched for as part of a longer string of words; for example, data for the term 'dishwasher' would include searches for 'energy efficient dishwasher'.

Equally, users with entirely different intentions could enter very similar search queries. For example, a lot of searches will be purely out of curiosity. So there is often significant noise in the search data. There are also many economic activities that still involve little use of the internet — for example, firms' investment in new production facilities — and so are unlikely to be related to internet search activity. Finally, there are also some limitations to the data as they are available now, which are related to how they are extracted from the search engine. This will be discussed further below.

The available internet search data

In line with previous studies on internet searches, this article uses data from the Google Insights for Search application: for more information on the data set and how the data are used in this article, see the box above. Although many search engines publish lists of the most popular search terms, the Insights application is currently the only one with a flexible interface that reports the popularity of a search term specified by the user. As Google currently has such a large proportion of the search engine market,(1) it is likely that its data cover the largest possible sample of internet users. Of course, these techniques could equally well be applied to data from other search engines if they were to make similar statistics available.

In their current form, there are some limitations to the available data. The popularity of each search is reported as an index rather than a volume of searches. So the data are not informative of the actual level of interest in the search term. Furthermore, because the reported index is based on a random sample of total searches, the backrun of data can change. This appears to be a particular issue for less popular search terms. Therefore users of these data must be careful that the results of their analysis are not specific to the index reported on any given day. To overcome this issue, this article averages the index reported on seven consecutive days, and uses more popular search terms which tend to be more stable.

Deciding which search queries to consider is a crucial element of using internet search data. To keep the analysis simple and transparent, only individual search terms are considered in this article. As discussed below, preferred queries are selected based on economic intuition. But further work into the selection of search terms could be helpful in fully exploiting the information in the search data.(2)

Analysing the labour and housing markets

This article evaluates the usefulness of the data for two specific markets: the labour and housing markets. These are two areas where the internet has become an increasingly important tool for companies and the public alike.

(1) Google's share of the UK internet search market, by search volume, was 85% for the four weeks ending 21 May 2011 (Experian Hitwise).
(2) A variety of approaches have been used in the literature. Perhaps most notably, in their study of influenza trends, Ginsberg *et al* (2009) select their preferred search terms using a purely statistical procedure involving running 450 million different models to choose between candidate queries. This exhaustive process is beyond the scope of this analysis.

For example, it is now likely that people who are unemployed, or fear they may soon lose their job, will search on the internet to find out about the benefits system and to search for new jobs. So internet search terms may be useful for monitoring the labour market. Labour market studies using internet search data have been carried out in a wide range of countries. For the United States, Choi and Varian (2009b) find that unemployment and welfare-related searches can improve predictions of initial jobless benefit claims. Askitas and Zimmerman (2009), D'Amuri (2009) and Suhoy (2009) find similar results for Germany, Italy and Israel respectively.

In the housing market, people interested in both buying and selling properties make use of the internet to monitor market conditions and advertise their properties. Therefore internet searches may also be related to conditions in the housing market. Most previous studies for the housing market focus on the United States. Choi and Varian (2009a) find that real estate related searches can improve on standard nowcasts for house sales. A similar study by Wu and Brynjolfsson (2009) finds that this applies at a state, as well as national level. They also find evidence that search data can be informative for future housing transactions and prices. Webb (2009) finds evidence that searches for 'foreclosure' are highly correlated with actual US home foreclosures, and so suggests search trends could be used as an early warning system of troubles in the US housing market.

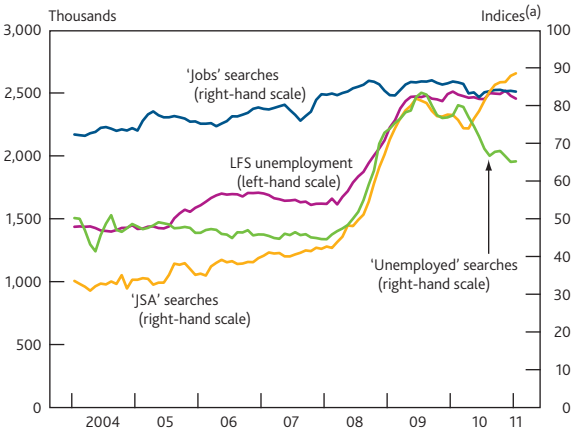
Both the survey data and the internet search data are timelier than official statistics, and consequently can help to 'nowcast', or enhance understanding of the current state of the economy. To assess the value of search data in the United Kingdom, this article compares simple regression models for unemployment and house prices, to those augmented with internet search variables. Existing indicators are also considered to see if internet search data can better explain the official data. The performance of each model in nowcasting the official data is then compared. The simple models used are a benchmark against which to compare our results, and are not intended to illustrate the Bank's approach to modelling these markets.

Labour market

A range of labour market related searches which could be used to nowcast unemployment (such as 'jobs', 'Jobseeker's Allowance', 'JSA', 'unemployment benefit', 'unemployed', 'unemployment') were considered. **Chart 1** illustrates that, over the available sample, some of these have behaved similarly to actual unemployment as measured by the Labour Force Survey (LFS) published by the ONS.

It is notable that 'jobs', which is likely to have been searched for by both those in and out of employment, did not increase much during the recession. Searches for 'unemployed' rose markedly during the recession. The term 'JSA' (acronym for Jobseeker's Allowance) was chosen because its movements best correlated with those in the official data. It is also a term likely to be used by those who think they may soon become unemployed and so search for more information on unemployment benefit.

Chart 1 LFS unemployment and unemployment-related searches



Sources: Google, ONS and Bank calculations.

(a) The weekly search data are calculated as an index, where the highest point in the series is rescaled to 100. The index here is a monthly average of that weekly data.

When trying to investigate the econometric relationship between the official unemployment data and the search term data, it is important to note that both have trended upwards over this period. Therefore, to avoid the results being dominated by the correlation between the trends, the change in unemployment on the previous three months (ΔU) is modelled. The baseline model is a simple autoregressive model. This includes only changes in unemployment in previous months as explanatory variables. Different unemployment indicators (X) are then added, and compared to the baseline model. The unemployment equation takes the form:

$$\Delta U_t = \alpha + \beta_1 \Delta U_{t-1} + \beta_2 \Delta U_{t-2} + \phi \Delta X_t$$

First, internet search data are included and its performance compared with the baseline. These data are available from 2004 so the estimation is from June 2004 to January 2011. Second, the performance of a model with internet search data is compared to models that use alternative indicators of unemployment, such as the claimant count, and the GfK consumer confidence question on changes in expected unemployment for the next year.(1) Since some indicators are timelier than LFS unemployment, indicators for the current period are used to 'nowcast' current unemployment.

(1) The GfK survey question asks respondents: 'How do you expect the number of people unemployed in this country will change over the next twelve months?'. The GfK data lagged by four months are used, since these best correlate with the dependent variable.

Results

The baseline model can account for a large proportion of the variation in unemployment. The second column in **Table A** shows that when the 'JSA' internet search term is added to the model, it has the expected positive coefficient, and is significant at the 1% level. The last two rows of **Table A** show that the 'JSA' model also improves the fit according to in-sample goodness of fit measures: it has a higher adjusted R-squared and a lower Akaike information criterion than the baseline.(1) This provides clear evidence that search terms do contain relevant information for explaining changes in unemployment. The 'JSA' model is outperformed by the claimant count model, which (as shown in the third column of the table) has the lowest Akaike information criterion. But both the search data and the claimant count are significant at the 5% level when all indicators are simultaneously included in the equation. So the results suggest that search data contain useful information in addition to existing surveys.

Table A Unemployment regression results

Independent variables	Baseline	'JSA'	Claimant count	GfK	All
	(1)	(2)	(3)	(4)	(5)
α	5.36 (0.16)	1.25 (0.73)	11.16 (0.01)	8.35 (0.04)	9.02 (0.03)
ΔU_{t-1}	1.08 (0.00)	0.85 (0.00)	0.76 (0.00)	0.91 (0.00)	0.69 (0.00)
ΔU_{t-2}	-0.20 (0.03)	-0.13 (0.17)	-0.28 (0.00)	-0.15 (0.08)	-0.21 (0.02)
Δ'JSA'$_t$	–	5.02 (0.00)	–	–	2.37 (0.04)
ΔCC_t	–	–	0.44 (0.00)	–	0.32 (0.00)
ΔGfK_{t-4}	–	–	–	2.22 (0.00)	0.57 (0.38)
Adjusted R-squared	0.81	0.85	0.86	0.83	**0.87**
Akaike information criterion	9.85	9.65	**9.52**	9.75	**9.52**

Dependent variable: Change in LFS unemployment, latest three months on previous three months. Delta denotes change on previous period.
Sample: 2004 M6 to 2011 M1.
P-values for heteroskedasticity robust standard errors are shown in parentheses.

An out-of-sample test is also conducted. This is used to compare how well each model nowcasts current unemployment data. For the test, the model is first estimated up to June 2008, and a nowcast produced for July 2008. The difference between the nowcast and the unemployment data for that particular month is then recorded — this is referred to as the one month ahead nowcast error. The exercise is then repeated, with the model estimated up until July 2008, and with a nowcast for August 2008 being compared with the data. This is continued up to the end of the sample. The one month ahead nowcast errors are then compared across models. **Table B** shows that the claimant count model produces the smallest errors. In line with the in-sample results above, the out-of-sample test suggests that the 'JSA' is outperformed by the claimant count model but improves upon the GfK model.

Table B Unemployment equations out-of-sample forecast test

	Baseline	'JSA'	Claimant count	GfK	All
	(1)	(2)	(3)	(4)	(5)
RMSE	40.4	35.3	**33.8**	37.1	37.1

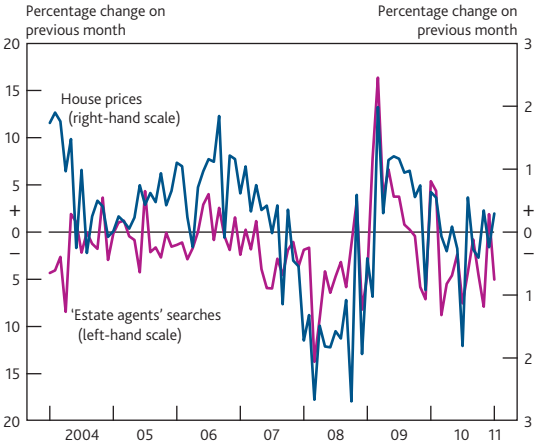
Each model is first estimated for the period up to 2008 M6. The square root of the mean-squared forecast error (RMSE) for one month ahead forecasts is then compared. These results are robust to different starting periods for the out-of-sample testing.

Housing market

For the housing market a similar approach is followed to that taken above for the labour market. A significant proportion of housing-related searches are for specific companies' websites. However, these searches vary over time depending on the popularity of each website. So a wide range of more generic search terms are considered (including 'house prices', 'buy house', 'sell house', 'mortgage' and 'estate agents'). The search terms 'buy house' and 'sell house' were initially considered, since they would capture the demand for and supply of houses. But the data for these search terms vary significantly when downloaded on different days, perhaps because of low search volumes. This volatility affected the robustness of the results. Instead, the search term 'estate agents' was chosen as it is much more stable when downloaded on different days. The term is correlated with both house prices and housing transactions, but appears to move more closely with house prices over our sample period. As a result, this article considers a model of house prices.

The term 'estate agents' may capture both demand and supply-related searches. But it appears that demand searches dominate so that there is a positive relationship between prices and searches (**Chart 2**).

Chart 2 House price inflation and 'estate agents' searches



Sources: Google, Halifax, Nationwide and Bank calculations.

(1) Both are measures of the goodness of fit of a model. The higher the R-squared, the greater the variation in the data that can be explained by the regression model. The Akaike information criterion measures the goodness of fit that can be achieved using the smallest number of explanatory variables: the lower the number, the better the fit.

The dependent variable in the model is monthly house price growth ($\dot{HP}$). In the previous section, contemporaneous searches and indicators were used to nowcast current unemployment. But since the house price data are timelier than the equivalent labour market data, the previous month's searches and indicators must be used to produce nowcasts. So these terms enter the equation with a lag. The models suggest that lagged variables tend to correlate more strongly in any case. This may be because there is more of a lag in the housing market between internet search activity and actual market activity, due to the time taken for negotiation, and administrative and legal processes. There are several alternative indicators of house prices. The house price growth balances from the Home Builders Federation (HBF) and the Royal Institution of Chartered Surveyors (RICS) are both used here (X). The house price growth equation takes the form:

$$\dot{HP}_t = \alpha + \beta_1 \dot{HP}_{t-1} + \beta_2 \dot{HP}_{t-2} + \phi X_{t-1}$$

Results

The results for the house price equation seem to be even more encouraging than the unemployment equation. Column 1 in **Table C** shows that, as with the unemployment equation, the baseline model is able to explain a significant proportion of the growth in house prices and each of the variables have the appropriate sign. Column 2 shows that when the search variable is included, it enters with a positive coefficient and is significant at the 1% level. Both of the alternative surveys are significant when they are included individually at the 10% level. But the search term model performs better according to in-sample criteria such as the adjusted R-squared and the Akaike information criterion. And when all the indicators are included simultaneously, the search variable remains significant at the 1% level, while the other surveys are insignificant.

These results are supported by the out-of-sample test, conducted in the same manner as for unemployment (**Table D**). When added individually, the model with the internet search term variable has a lower root mean square error for one month ahead nowcasts compared to models with other survey indicators. So there is evidence that the search data can improve understanding of the current state of the housing market.

Table C House price regression results

Independent variables	Baseline	'Estate agents'	RICS	HBF	All
	(1)	(2)	(3)	(4)	(5)
α	0.00	0.00	0.00	0.00	0.00
	(0.97)	(0.04)	(0.10)	(0.25)	(0.00)
$\dot{HP}_{t-1}$	0.20	0.12	0.05	0.05	-0.02
	(0.06)	(0.13)	(0.68)	(0.73)	(0.86)
$\dot{HP}_{t-2}$	0.58	0.58	0.41	0.44	0.41
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
$\dot{'Estate\ agents'}_{t-1}$	–	0.09	–	–	0.09
		(0.00)			(0.00)
$RICS_{t-1}$	–	–	0.00	–	0.00
			(0.01)		(0.09)
HBF_{t-1}	–	–	–	0.00	0.00
				(0.07)	(0.84)
Adjusted R-squared	0.53	0.68	0.56	0.55	**0.70**
Akaike information criterion	-7.05	-7.42	-7.09	-7.07	**-7.47**

Dependent variable: Change in house prices on previous month.
Sample: 2004 M5 to 2011 M3.
P-values for heteroskedasticity robust standard errors are shown in parentheses.
The dot above the variables denotes growth rate.

Table D House price equations out-of-sample forecast test

	Baseline	'Estate agents'	RICS	HBF	All
	(1)	(2)	(3)	(4)	(5)
RMSE	0.87	0.69	0.87	0.87	**0.67**

The potential of the data

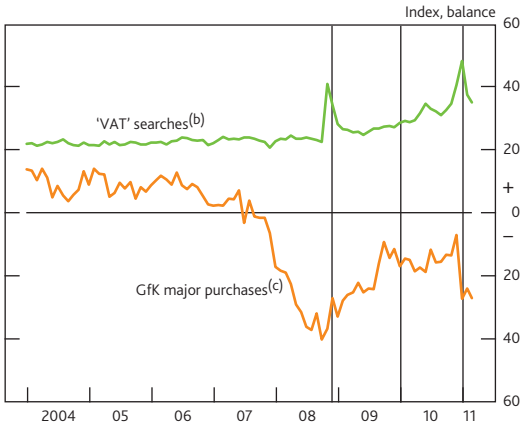
This analysis suggests that internet search data contain valuable information for analysis of unemployment and house prices. These applications treated the search data in a similar manner to existing surveys in conducting standard regression analysis. But internet search data also have the potential to answer different sorts of questions to existing indicators. They have the particular advantage that they can help analyse issues that arise unexpectedly. Whereas survey data must be consciously collected based on pre-determined questions, internet data are collected based on behaviour at the time, and a backrun will be available provided the term was searched widely on the internet.

An example of this type of issue is analysis of the public reaction to the recent changes in the rate of VAT. Data on internet searches including 'VAT' can provide an insight into the way consumer confidence survey balances moved in the months surrounding the VAT changes.

The orange line in **Chart 3** shows the GfK consumer confidence question asking about whether now is a good time to make a major purchase, with the timing of VAT changes shown by the vertical lines. As expected, there is a clear relationship between changes in VAT and the major purchases survey balance. However, the survey balance fell much more following the January 2011 increase to 20%, than following the increase to 17.5% in January 2010. This difference could be due to a more muted consumer response to the 2010 VAT increase. Or it might reflect other changes on the month offsetting the negative VAT impact.

The consumer confidence survey does not ask specifically about the VAT impact, so it is difficult to distinguish between these two explanations. However, data on internet searches

Chart 3 Internet search responses to VAT changes[a]



Sources: GfK/EC Consumer Confidence Barometer, Google Insights for Search and Bank calculations.

(a) Vertical lines indicate changes in standard rate of VAT.
(b) The weekly search data are calculated as an index, where the highest point in the series is rescaled to 100. The index here is a monthly average of that weekly data (see the box on page 135).
(c) The GfK survey question asks respondents: 'In view of the general economic situation, do you think now is the right time for people to make major purchases such as furniture or electrical goods?'.

including 'VAT' collected over this period can help provide some insight into the consumer reaction. The green line in **Chart 3** shows that searches including 'VAT' increased in the period surrounding VAT changes. And consistent with the consumer confidence balance, there was only a small increase in searches following the 2010 VAT increase. This appears to provide evidence that the impact on consumers of the VAT increase in 2010 was not as significant as the other two changes in VAT.

The internet data are helpful for determining how consumers responded. But in this case neither the consumer confidence survey nor the internet data are able to explain the observed differences in consumer behaviour. Given the similarities between the two VAT increases, the difference in responses for both data sources may be surprising. In both cases the changes had been pre-announced, and in both cases the changes were permanent. One possible explanation may be related to the previous movements in the rate of VAT. The increase to 17.5% was the reversal of a temporary reduction in the VAT rate. By contrast, the increase to 20% was a permanent increase to a higher rate of taxation.

This simple example illustrates the potential value internet search data have for providing added detail on the way consumers are behaving. The internet data are particularly useful in this type of situation because traditional survey indicators would not necessarily have been adapted to ask specifically about VAT changes.

This is just one area where internet search data have the potential to shape the information we have about economic behaviour. As the backrun of the data increases, and more activities become internet orientated, it is likely that the importance of this data source will increase further. Already the data can be informative if the appropriate search terms are used. And this could be the key to future development of this data source. As consumer search queries become more complex, it will be important to develop better ways to extract the economic content contained in these data. Determining which search terms to use, and how to distinguish noise from signal will be important future developments in this area. It is likely that these data can help answer important economic questions; it is a case of making sure the right questions are asked of the data.

Conclusion

This article has considered the potential usefulness of internet search data as economic indicators. There remain some limitations of these data: there is only a short backrun, there is no information on the actual volume of searches, and as the index is based on a subsample the backrun of data can change. However, even in their current form, initial results suggest these data can be useful. In line with studies for other countries, internet search data can help predict changes in unemployment in the United Kingdom. These appear to be as useful as existing indicators. For house prices, the results are somewhat stronger: search term variables can outperform some existing indicators over the period since 2004. There is also evidence that these data may be used to provide additional insight on a wider range of issues which traditional business surveys might not cover.

The Bank will continue to monitor these data as part of the range of different indicators it considers in forming its view about the outlook for the economy of the United Kingdom. As further developments are made in this area, and the backrun of the data increases, these data are likely to become an increasingly useful source of information about economic behaviour.

References

Askitas, N and Zimmerman, K F (2009), 'Google econometrics and unemployment forecasting', *Applied Economics Quarterly*, Vol. 55, No. 2, pages 107–20, available at www.atypon-link.com/DH/doi/abs/10.3790/aeq.55.2.107.

Chamberlain, G (2010), 'Googling the present', *Economic and Labour Market Review*, Office for National Statistics, Vol. 4, No. 12, available at www.statistics.gov.uk/cci/article.asp?ID=2621.

Choi, H and Varian, H (2009a), 'Predicting the present with Google Trends', Google Inc., available at http://google.com/googleblogs/pdfs/google_predicting_the_present.pdf.

Choi, H and Varian, H (2009b), 'Predicting initial claims for unemployment benefits', Google Inc., available at http://research.google.com/archive/papers/initialclaimsUS.pdf.

D'Amuri, F (2009), 'Predicting unemployment in short samples with internet job search query data', Bank of Italy Research Department, available at http://mpra.ub.uni-muenchen.de/18403/1/MPRA_paper_18403.pdf.

Ginsberg, J, Mohebbi, M H, Patel, R S, Brammer, L, Smolinski, M S and Brilliant, L (2009), 'Detecting influenza epidemics using search engine query data', *Nature*, Vol. 457, pages 1,012–14, available at www.nature.com/nature/journal/v457/n7232/full/nature07634.html.

Office for National Statistics (2010), 'Internet access 2010: households and individuals', available at www.statistics.gov.uk/pdfdir/iahi0810.pdf.

Suhoy, T (2009), 'Query indices and a 2008 downturn', *Bank of Israel Discussion Paper no. 2009.06*, available at www.bankisrael.gov.il/deptdata/mehkar/papers/dp0906e.pdf.

Webb, G K (2009), 'Internet search statistics as a source of business intelligence: searches on foreclosure as an estimate of actual home foreclosures', *Issues in Information Systems*, Vol. 10, No. 2, available at www.iacis.org/iis/2009_iis/pdf/P2009_1169.pdf.

Wu, L and Brynjolfsson, E (2009), 'The future of prediction: how Google searches foreshadow housing prices and sales', MIT Sloan School of Management, available at http://pages.stern.nyu.edu/~bakos/wise/papers/wise2009-3b3_paper.pdf.