Macroeconomic Nowcasting Using Google Probabilities

Gary Koop and Luca Onorante

University of Strathclyde and Central Bank of Ireland

April 7^{th} 2014

Google search data

1. Macroeconomic data are typically published with a time lag. This has led to a growing body of research on nowcasting.
2. Internet search data provide a new resource for researchers interested in nowcasts or short-term forecasts.
3. Google search data, available since January 2004, is a particularly popular source.
   - Choi and Varian (2009, 2011) have led to an explosion of nowcasting work using Google data including, among many others, Artola and Galan (2012), Askitas and Zimmermann (2009), Carriere-Swallow and Labbe (2011), Chamberlin (2010), D'Amuri and Marcucci (2009), Hellerstein and Middeldorp (2012), Kholodilin, Podstawski and Siliverstovs (2010), McLaren and Shanbhoge (2011), Scott and Varian (2012), Schmidt and Vosen (2009), Suhoy (2009) and Wu and Brynjolfsson (2010).
4. Google data is also being used for nowcasting in other fields such as finance and epidemiology.

Google search data

Few general themes emerge.

- First, Google data is potentially useful in nowcasting or short-term forecasting, but there is little evidence that it can be successfully used for long-term forecasting.
- Second, Google data is only rarely found to be useful for broad macroeconomic variables (e.g. inflation, industrial production, etc.)[1] and is more commonly used to nowcast specific variables relating to consumption, housing or labor markets. For instance, Choi and Varian (2011) successfully nowcast the variables motor vehicles and car parts, initial claims for unemployment benefits and tourist arrivals in Hong Kong.
- Third, most existing literature uses linear regression methods.

The present paper deals with the second and third of these points.

[1] A notable exception is the nowcasting of U.S. unemployment in D'Amuri and Marcucci (2009).

Our contribution

- We add to the literature by nowcasting using dynamic model averaging (DMA/DMS) methods which allow for model switching between time-varying parameter regression models: Raftery et al (2010).
  - This is potentially useful in an environment of coefficient instability and over-parameterization which can arise when forecasting with Google variables.
- We extend the methodology by allowing for the model switching to be controlled by the Google variables through what we call Google probabilities.
  - Instead of using Google variables as regressors, we allow them to determine which nowcasting model should be used at each point in time.
- In an empirical exercise involving nine major monthly US macroeconomic variables, we find DMS methods to provide large improvements in nowcasting.
- Google model probabilities within DMS often performs better than conventional DMS.

Our contribution

We nowcast nine US monthly macroeconomic variables and see if Google variables provide additional nowcasting power
We use Google variables in different ways:

1. DMA and DMS using Google variables as additional predictors in TVP regressions.
   - Extension over existing nowcasting methods, such as Choi and Varian (2009, 2011), who use linear regression methods with constant coefficients
2. Inclusion probability of each macro explanatory variable depending on the Google data: "Google probabilities".
   - Literature (e.g. Choi and Varian, 2011) suggests that Google variables might not be good linear predictors. However, they collect "collective wisdom" about which macro variables are important in the model at different points in time, either directly or by influencing the outcomes through expectations. (example: oil prices)

Dependent and Explanatory Variables

- Macro variables: Inflation, Wage inflation, Unemployment rate, Term spread, Financial Conditions Index, Comm. price inflation, Industrial production, Oil price inflation, Money supply growth
- Corresponding to each, the average Google search variable reflecting internet search activity relating to the underlying macroeconomic concept.

How to make a Google variable

In the literature: Scott and Varian (2012) use 151 search categories (top-level categories are e.g. 'Food and beverages' or 'News and current events')

In this paper we use a standardized, non-judgement procedure

1. Search for the name of the macro variable of interest
2. Collect the Google search volume related terms.
   - These are the most popular terms related to the search: Google chooses them by examining searches conducted by users immediately before and after.
3. Fetch related searches, and repeat the procedure for each of them, finding new terms.
4. Final Google database: 259 search results. Convert weekly series by taking the last observation available for every month.
5. Average all the Google search variables to produce a single "Google variable" corresponding to each macroeconomic variable.

Works well, more sophisticated methods (e.g. using principal components methods) would be possible.

Regressions

Dependent variable, y_t, other macroeconomic variables as potential explanatory variables, X_t. The Google variables corresponding to X_t will be labelled Z_t.

$$y_t = X_{t-1}'\beta + \varepsilon_t. \tag{1}$$

$$y_t = X_{t-1}'\beta + Z_t'\gamma + \varepsilon_t. \tag{2}$$

Two potential problems:

1. coefficients are constant over time which, for many macroeconomic time series, is rejected by the data (see, among many other, Stock and Watson, 1996)
2. they may be over-parameterized

1 - Time variance: TVP regression model

$$\begin{aligned} y_t &= W_t'\theta_t + \varepsilon_t \\ \theta_{t+1} &= \theta_t + \eta_t, \end{aligned} \tag{3}$$

we consider both $W_t = X_{t-1}$ and $W_t = [X_{t-1}\prime, Z_t']'$. We also allow for time variation in the error variance: ε_t i.i.d. $N\left(0, \sigma_t^2\right)$, and σ_t^2 is an Exponentially Weighted Moving Average:

$$\widehat{\sigma}_t = \kappa\widehat{\sigma}_{t-1} + (1-\kappa)\,\widehat{\varepsilon}_t\widehat{\varepsilon}_t', \tag{4}$$

where $\widehat{\varepsilon}_t$ are the estimated regression errors, $\kappa = 0.96$ following suggestions in Riskmetrics (1996), η_t are independent $N(0, Q_t)$ random variables (also independent of ε_t).

$\oplus$ These are state space models and, thus, use Kalman filter.

$\ominus$ They can be over-parameterized, exacerbating the second problem noted above.

2 - Parsimony: model averaging

- The pioneering paper which developed methods for DMA and DMS was Raftery et al (2010).
- DMA or DMS used to ensure shrinkage in over-parameterized models (see, e.g., Koop and Korobilis, 2012). They allow for a different model to be selected at each point in time (with DMS) or different weighs used in model averaging at each point in time (with DMA).
  - For instance, in light of Choi and Varian (2011)'s finding that Google variables predict better at some points in time than others, one may wish to include the Google variables at some times but not others. DMS allows for this. It can switch between models which include Google variables and models which do not, as necessary.

2 - Parsimony: model averaging

We have $j = 1, .., J$ TVP regression models, each of the form:

$$\begin{aligned} y_t &= W_t^{(j)} \theta_t^{(j)} + \varepsilon_t^{(j)} \\ \theta_{t+1}^{(j)} &= \theta_t^{(j)} + \eta_t^{(j)}, \end{aligned} \tag{5}$$

where $\varepsilon_t^{(j)}$ is $N\left(0, \sigma_t^{2(j)}\right)$ and $\eta_t^{(j)}$ is $N\left(0, Q_t^{(j)}\right)$. there are $J = 2^S$ possible TVP regressions involving every possible combination of the S explanatory variables.
As discussed in Koop and Korobilis (2012), exact Bayesian estimation is computationally infeasible. Within a single TVP regression model we estimate $\sigma_t^{2(j)}$ using EWMA methods (as described above) and $Q_t^{(j)}$ using forgetting factor methods.

[Forgetting factors]

Forgetting factors have long been used in the state space literature to simplify estimation. Sources such as Raftery et al (2010) and West and Harrison (1997) describe forgetting factor estimation of state space models. Suffice it to note that:

- they involve choice of a scalar forgetting factor $\lambda \in [0, 1]$
- lead to estimates of $\theta_t^{(j)}$ where observations j periods in the past have weight λ^j.

An alternative way of interpreting λ is to note that it implies an effective window size of $\frac{1}{1-\lambda}$.

With EWMA and forgetting factor methods used to estimate $\sigma_t^{2(j)}$ and $Q_t^{(j)}$, use of the Kalman filter in order to provide estimates of the states and, crucially for our purposes, the predictive density, $p_j\left(y_t | W_{1:t}, y_{1:t-1}\right)$, where $W_{1:t} = (W_1, .., W_t)$ and $y_{1:t-1} = (y_1, .., y_{t-1})$.

Model updating equations

- DMA and DMS involve a recursive updating scheme using quantities which we label $q_{t|t,j}$ and $q_{t|t-1,j}$.
  - The latter is the probability that model j is the model used for nowcasting y_t, at time t, using data available at time $t-1$.
  - The former updates $q_{t|t-1,j}$ using data available at time t.
- DMS involves selecting the single model with the highest value for $q_{t|t-1,j}$ and using it for forecasting y_t. Note that DMS allows for model switching: at each point in time it is possible that a different model is used for forecasting.
- DMA uses forecasts which average over all $j = 1, .., J$ models using $q_{t|t-1,j}$ as weights.

Model updating equations

Raftery et al (2010) derive the following model updating equation:

$$q_{t|t,j} = \frac{q_{t|t-1,j} p_j \left(y_t | W_{1:t}, y_{1:t-1}\right)}{\sum_{l=1}^{J} q_{t|t-1,l} p_l \left(y_t | W_{1:t}, y_{1:t-1}\right)} \tag{6}$$

where $p_j \left(y_t | W_{1:t}, y_{1:t-1}\right)$ is the predictive likelihood (i.e. the predictive density for y_t produced by the Kalman filter run for model j evaluated at the realized value for y_t). The algorithm then uses a forgetting factor, α, set to 0.99 following Raftery et al (2010), to produce a model prediction equation:

$$q_{t|t-1,j} = \frac{q_{t-1|t-1,j}^{\alpha}}{\sum_{l=1}^{J} q_{t-1|t-1,l}^{\alpha}}. \tag{7}$$

Thus, starting with $q_{0|0,j}$ (for which we use the noninformative choice of $q_{0|0,j} = \frac{1}{J}$ for $j = 1, .., J$) we can recursively calculate the key elements of DMA: $q_{t|t,j}$ and $q_{t|t-1,j}$ for $j = 1, .., J$.

3 - Google Model Probabilities

Search volume might show the relevance of a certain variable for nowcasting at one point in time rather than a precise and signed cause-effect relationship.
We propose to modify the conventional DMA/DMS methodology as follows.
Let $Z_t = (Z_{1t}, .., Z_{kt})'$ be the vector of Google variables. Re-sized, this number can be interpreted as a probability.
Consider the same model space as before, defined in (5), with $W_t = X_{t-1}$.
For each of these models and for each time t we define $p_{t,j}$, which we call a Google probability:

$$p_{t,j} = \prod_{s \in I^j} Z_{st} \prod_{s \in I^{\sim j}} (1 - Z_{st}) \tag{8}$$

where I^j indicates which variables are in model j and $I^{\sim j}$ which are excluded.

3 - Google Model Probabilities

It can be seen that $\sum_{j=1}^{J} p_{t,j} = 1$ and that each Google model probability reflects increases or decreases in internet searches. [2] Time varying model probabilities reflect the Google model probabilities as:

$$q_{t|t-1,j} = \omega \frac{q_{t-1|t-1,j}^{\alpha}}{\sum_{l=1}^{J} q_{t-1|t-1,l}^{\alpha}} + (1-\omega)\, p_{t,j} \tag{9}$$

where $0 \leq \omega \leq 1$. If $\omega = 1$ we are back in conventional DMA or DMS as done by Raftery et al (2010), if $\omega = 0$ then $p_{t,s}$ replaces $q_{t|t-1,s}$ in the algorithm (and, hence, the Google model probabilities are driving model switching). Intermediate values of ω will combine the information.

[2] For example, when $I^j = \{3, 7\}$, $p_{t,j}$ will be large in times when internet searches on terms relating to the third and seventh explanatory variables are unusually high and it will be low when such searches are unusually low.

Results

Table 2a: Nowcast Performance (post-2004 data)				
	LogPL		MSFE	
	DMA	DMS	DMA	DMS
	Inflation			
	Google Variables Not Used			
$\omega = 1$	-236.73	-235.38	24.95	23.28
Rec. OLS	-	-	30.75	-
Rec. AR(2)	-	-	24.22	-
No change	-	-	31.20	-
	Google Variables Used in Model Probs			
$\omega = 0.5$	-239.41	-232.29	24.69	19.35
$\omega = 0$	-239.48	-232.36	24.75	19.13
	Google Variables Used as Regressors			
$\omega = 1$	-237.64	-233.23	26.28	21.08
Rec. OLS	-	-	37.23	-
	Industrial Production			
	Google Variables Not Used			
$\omega = 1$	-289.10	-287.78	107.04	104.46
Rec. OLS	-	-	165.51	-
Rec. AR(2)	-	-	114.13	-
No change	-	-	113.83	-
	Google Variables Used in Model Probs			
$\omega = 0.5$	-291.74	-286.46	116.96	110.12
$\omega = 0$	-291.94	-284.42	117.49	109.74
	Google Variables Used as Regressors			
$\omega = 1$	-288.28	-284.98	102.88	95.90
Rec. OLS	-	-	158.12	-

Results

Table 2b: Nowcast Performance (post-2004 data)				
	LogPL		MSFE	
	DMA	DMS	DMA	DMS
	Unemployment			
	Google Variables Not Used			
$\omega = 1$	-124.10	-123.04	0.033	0.033
Rec. OLS	-	-	0.036	-
Rec. AR(2)	-	-	0.038	-
No change	-	-	5.44	-
	Google Variables Used in Model Probs			
$\omega = 0.5$	-133.75	-127.30	0.032	0.033
$\omega = 0$	-134.41	-128.38	0.032	0.035
	Google Variables Used as Regressors			
$\omega = 1$	-127.91	-123.04	0.034	0.033
Rec. OLS	-	-	0.047	-
	Wage Inflation			
	Google Variables Not Used			
$\omega = 1$	-192.95	-190.10	6.52	5.77
Rec. OLS	-	-	9.78	-
Rec. AR(2)	-	-	6.83	-
No change	-	-	6.15	-
	Google Variables Used in Model Probs			
$\omega = 0.5$	-197.80	-194.11	7.16	5.71
$\omega = 0$	-198.02	-194.49	7.17	5.89
	Google Variables Used as Regressors			
$\omega = 1$	-195.25	-189.50	6.72	5.53
Rec. OLS	-	-	11.48	-

Results

Table 2c: Nowcast Performance (post-2004 data)				
	LogPL		MSFE	
	DMA	DMS	DMA	DMS
	Money			
	Google Variables Not Used			
$\omega = 1$	-245.69	-244.53	30.02	29.71
Rec. OLS	-	-	33.50	-
Rec. AR(2)	-	-	28.99	-
No change	-	-	28.69	-
	Google Variables Used in Model Probs			
$\omega = 0.5$	-249.97	-242.72	29.28	27.34
$\omega = 0$	-250.81	-243.97	29.75	26.07
	Google Variables Used as Regressors			
$\omega = 1$	-247.07	-242.90	31.20	28.12
Rec. OLS	-	-	42.77	-
	FCI			
	Google Variables Not Used			
$\omega = 1$	-53.22	-53.64	0.29	0.29
Rec. OLS	-	-	0.30	-
Rec. AR(2)	-	-	0.32	-
No change	-	-	0.45	-
	Google Variables Used in Model Probs			
$\omega = 0.5$	-58.92	-53.29	0.28	0.21
$\omega = 0$	-59.56	-54.56	0.28	0.21
	Google Variables Used as Regressors			
$\omega = 1$	-55.51	-51.56	0.32	0.26
Rec. OLS	-	-	0.48	-

Results

Table 2d: Nowcast Performance (post-2004 data)				
	LogPL		MSFE	
	DMA	DMS	DMA	DMS
	Oil Price Inflation			
	Google Variables Not Used			
$\omega = 1$	-484.51	-479.54	13,219	10,407
Rec. OLS	-	-	17,465	-
Rec. AR(2)	-	-	11,253	-
No change	-	-	12,185	-
	Google Variables Used in Model Probs			
$\omega = 0.5$	-481.39	-475.00	11,678	8,961
$\omega = 0$	-481.60	-474.71	11,857	8,555
	Google Variables Used as Regressors			
$\omega = 1$	-484.63	-479.72	13,241	10,415
Rec. OLS	-	-	29,333	-
	Commodity Price Inflation			
	Google Variables Not Used			
$\omega = 1$	-429.24	-425.50	3,115	2,706
Rec. OLS	-	-	3,925	-
Rec. AR(2)	-	-	2,950	-
No change	-	-	3,254	-
	Google Variables Used in Model Probs			
$\omega = 0.5$	-429.74	-427.97	3,169	2,964
$\omega = 0$	-429.85	-428.49	3,168	2,986
	Google Variables Used as Regressors			
$\omega = 1$	-429.23	-424.71	3,120	2,635
Rec. OLS	-	-	5,193	-

Results

Table 2e: Nowcast Performance (post-2004 data)				
	LogPL		MSFE	
	DMA	DMS	DMA	DMS
	Term Spread			
	Google Variables Not Used			
$\omega = 1$	-87.68	-86.67	0.072	0.072
Rec. OLS	-	-	0.092	-
Rec. AR(2)	-	-	0.068	-
No change	-	-	1.476	-
	Google Variables Used in Model Probs			
$\omega = 0.5$	-99.42	-91.28	0.069	0.081
$\omega = 0$	-100.53	-93.32	0.069	0.091
	Google Variables Used as Regressors			
$\omega = 1$	-91.44	-86.67	0.068	0.072
Rec. OLS	-	-	0.103	-

Discussion and tentative conclusions

The following main conclusions emerge:

- The inclusion of Google data leads to sizeable improvements in nowcast performance. This result complements the existing literature by showing that Google search variables can be used to improve nowcasting of broad macroeconomic aggregates.
- Second, and despite the crude procedure we adopted to create the Google variables, it is often (albeit not invariably) the case that the information in the Google variables is best included in the form of model probabilities. Google search volumes provide the econometrician with useful information about which variable is important at each point in time. More extensive use of this vast database.
- Finally, Google probabilities make sense if the economy is not constant. DMS proved to be a particularly good method. It often nowcasts best, when it does not it does not go too far wrong. Simple benchmarks such as OLS methods occasionally produce very bad nowcasts.

Way ahead

- This is a first and so far successful attempt to use Google variables to improve macroeconomic nowcasting.
- We proposed two different uses of these variables, one of which, to our knowledge, completely new and close to the spirit ("what are people concerned about?") in which these variables are collected.
- Additional research will be needed to make these results more robust. Our construction of the Google variables, in particular, is extremely simple, and it is not unlikely that a more accurate choice in the searches or a different method of averaging may lead to further improvements in their use.

Thank you for your attention!