

How To Improve Google Flu Trends.

Jurgen A. Doornik, University of Oxford, UK

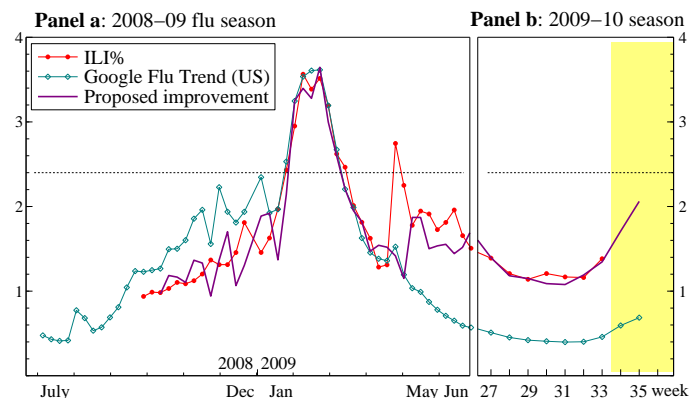
September 8, 2009

Google Flu Trends¹ reports the current flu activity in the US based on search activity indicators. A recent letter to Nature² reported how the model was obtained from historical search records. These results and the corresponding web site have received fairly widespread publicity.

Using internet search activity to improve short-term forecasts is an exciting new development, and the letter to Nature shows that it may contain useful information. Recently, however, there has been a dramatic increase in flu activity in the US — an episode that was missed entirely by the Google Flu Trends model. In this comment I propose how the Google Flu Trend model can be improved, and how the forecasts can be robustified. The objective of the latter is to limit the duration of a forecast failure. The former shows how the forecast errors can be reduced significantly.

The US Centers for Disease Control and Prevention (CDC) publishes flu activity with a two week delay, and the objective of Google Flu Trends is to fill that two week gap. In other words, to ‘predict the present’. The variable that is modeled by the Google researchers is the *percentage of visits for influenza-like illness, ILI%*.³ In week 17 of 2009 (starting April 26), the CDC reports that: ‘*The proportion of outpatient visits for influenza-like illness (ILI) was 2.6% which is above the national baseline.*’ That ILI is above the baseline in the

spring is quite exceptional, and, of course, associated with the current global swine-flu epidemic. However, Google Flu Trends does not report increased activity for May to July 2009, and thus misses this important event entirely. This can be seen in Panel a of the graph below. Panel b shows this more clearly: without the proposed improvement there is a danger that this epidemic could be missed entirely.



It should be emphasized that a Google Flu Trend estimate only has a lifespan of two weeks: after that we have the actual CDC information (possibly subject to minor revisions). The forecasts can be greatly improved by adopting a robustified⁴ version: compute the change for the current and previous period from the Google Flu Trend estimates, then apply this change to the actual outcome from two weeks ago.⁵ The figure shows how much better the robustified two-step ahead forecasts are than the original ones, particular in Panel b.

There are three reasons why there is such scope for improvement. The first is that previous modeling focused on simple correlations between potential individual candidate variables and the variable to be explained. This is a simplistic method relative to an efficient model selection procedure. The second is that the model was restricted to be static, while the dynamic properties are important for building a better model. Finally, forecast performance (and fit) is evaluated by the correlation between the outcomes and the predictions. However, the two can be far apart, and still highly correlated (the correlation between x_t and y_t is the same as that between $100 + x_t$ and y_t , say). Forecasters focus on the forecast errors, and compute summary statistics such as root mean squared error (RMSE) and mean absolute percentage error (MAPE).

A simple autoregressive model for the ILI% uses only past information on the dependent variable. Such a model, extended with calendar effects,⁶ has RMSE and MAPE that are similar to robustified Google Flu Trends.⁷ The autoregressive model explains why a sine wave will not be a good alternative: the change in a sine is a sine again, but the percentage changes in ILI% do not behave like that. The location and extent of flu activity changes from flu season to season, a feature that the dynamic model can handle better. This is also why the robustified forecasts give such an improvement.

It is important for health-care planning to know the current state of flu activity, and possibly to have forecasts of the near future. Google Flu Trends can assist, but only in its improved form. It indicates at the time of writing that ILI already exceeds the national baseline of 2.4%.

The current flu pandemic shows a brief peak before the summer, followed by a rapid pickup in the second half of August, which is revealed by the improved forecasts. This is similar to the pattern of the 1957-58 and 1968-69 pan-

demics.⁸

Notes

¹www.google.org/flutrends

²Ginsberg, 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**, 1012–1015.

³The % Weighted ILI in the weekly flu activity & surveillance report.

⁴Hendry, D.F. (2006). 'Robustifying forecasts from equilibrium-correction models', *J. of Econometrics*, **135**, 399–426.

⁵It is preferable to do this for the logit transform, undoing the transformation at the end. The bias correction is omitted throughout because it is very small for the estimated models.

⁶The model was selected from 73 weekly and holiday indicators. The final model for the logit of ILI% contains lags 1, 2 and 6, an intercept, and four composite calendar variables, as well as seven indicator variables to remove large outliers. This is documented in In supplementary research, 'Improving the Timeliness of Data on Influenza-like Illnesses using Google Search Data', mimeo, University of Oxford.

⁷ Pooling these two models may also be a good strategy: it improves the 2007-08 forecasts in particular.

⁸As described by Germann, T.C., Kadau, K., Longini Jr., I.M., Macken, C.A. (2006), *Proc. Natl. Acad. Sci. USA*, **103**, 5935–5940.

Helpful comments from Kate Doornik, Marius Ooms, David Hendry and Vivien Hendry are gratefully acknowledged.