Improving the Timeliness of Data on Influenza-like Illnesses using Google Search Data

Jurgen A. Doornik[*]
University of Oxford, UK

September 8, 2009

Abstract

Google Flu Trends reports the current flu activity in the US based on search activity indicators. A recent letter to Nature reported how the model was obtained from historical search records.

Using internet search activity to improve short-term forecasts is an exciting new development. The letter to Nature shows that it does indeed 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. It is shown how the Google Flu Trend forecasts can be robustified, and how the model can be improved. The objective of the former is to limit the duration of a forecast failure. The latter shows how the forecasts error can be reduced significantly.

It turns out that a dynamic model with calendar effects has similar forecast performance as the robustified Google Trends model. Therefore, two further models are built that use Google Trends data. These improve on the model with calendar effects. The pooled model is better still, so for the periods considered, search data can indeed help with nowcasting.

keywords Autometrics, Autoregression, Google Flu Trends, Indicator saturation, Influenza-like illness, Nowcasting, Robustified forecasts, Web search data

1 Introduction

Recently, Google Inc. has started publishing aggregated data for the volume of search on two web sites: Google Insights for Search (*google.com/insights/search*) and Google Trends (*google.com/trends*). Weekly information on the search activity is available from 2004 onwards; some additional information on the data is in Appendices A and B. Because the search data is produced almost instantaneously, various researchers have started addressing the question

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

whether the new search data can be useful to shed light on the current state of the economy or public health.

Most economic time-series are published with a delay, and may still be subject to revision for quite a while afterwards. An example is the production of GDP data, which is important for economic policy, but only available with a substantial lag. In the U.S., e.g., the Bureau of Economic Analysis published the final GDP figure for the 1st quarter of 2009 on 25 June 2009. An initial estimate was produced a month earlier. There is a growing literature on estimating current economic activity, i.e. predicting the present, or 'nowcasting' (for a recent contribution see Castle, Fawcett and Hendry, 2009). Prediction markets, surveys, and regional or disaggregate data could be sources of contemporaneous information that may improve the quality of nowcasts. In addition, it is possible that search engine data related to economic activity may be an additional source of information. Choi and Varian (2009) consider some economic examples.

A second area where nowcasting may be useful is in epidemiology. In June 2009 the World Health Organization raised the pandemic alert level on a new influenza virus (swine flu) to phase 6, indicating that a global pandemic is underway. For disease prevention and health-care planning it is important to know what the actual incidence of flu-related illnesses is. There are several official institutions at national and regional levels that collect the necessary statistics. But again, this is published with some delay, although short compared to economic data: usually only a few weeks. Because of the possibility of rapid spread of a new strain, it could be useful to have more timely information available. Again, search data may help to fill the gap. Precisely this question was addressed by Ginsberg, Mohebbi, Patel, Brammer, Smolinski and Brilliant (2009), and the accompanying web site Google Flu Trends (*google.com/FluTrends*).

The question whether search activity can provide more accurate information on current activity is very interesting. Unfortunately, most analyses so far have been let down by the crude statistical methods that were employed. Google Flu Trends, e.g., has shown a massive forecast failure in spring/summer 2009. The models presented below only show failure for a short period of time, after which they get back on track. As a consequence, they have much better nowcast performance than Google Flu Trends.

Section 2 briefly discusses the flu data, followed by an overview of the Google Flu Trends model and web site (Sections 3 and 4). In Section 6 it is shown how their nowcasts can be improved using robustified forecasts.

The first model presented in Section 5 is purely autoregressive. It does not do well in capturing the annual cycle, but it shows how easy it is to outperform Google Flu Trends. The subsequent model (labelled $M2$) includes calendar effects, which are derived from weekly and holiday indicators. Section 7 explores two lines of research. The first is whether web search data provides information beyond that encapsulated in model $M2$. The second is whether search data alone is useful to capture the full annual dynamics. Pooled forecasts from these last two models are also considered.

2 Flu Surveillance

Weekly flu activity and surveillance reports for the US are produced by the Centers for Disease Control and Prevention (CDC, www.cdc.gov). Consulting the web site on 2009-07-16, the most up-to-date report on the CDC web site (see cdc.gov/flu/weekly/) relates to the week starting Sunday 2009-06-28.[1] The CDC reports that 'During week 26 (June 28-July 4, 2009), influenza activity decreased in the United States, however, there were still higher levels of influenza-like illness than is normal for this time of year.'

The variable of interest is the *percentage of visits for influenza-like illness* (ILI%). Using the data provided by the CDC in the report for week 33, together with data from earlier reports, we can produce a plot of the ILI percentage for the U.S. (as a weighted percentage of regional counts), see Figure 1. This only differs from the CDC plots in that calendar dates are used instead of week numbers: some years have 53 weeks, and the CDC handles years of 52 weeks by introducing an artificial week 53 as the average of week 52 and 1.

Historically, flu has a low incidence during the summer, and monitoring used to cease for a period during the summer (weeks 21 to 39). The current swine flu epidemic shows a need for continuous monitoring.

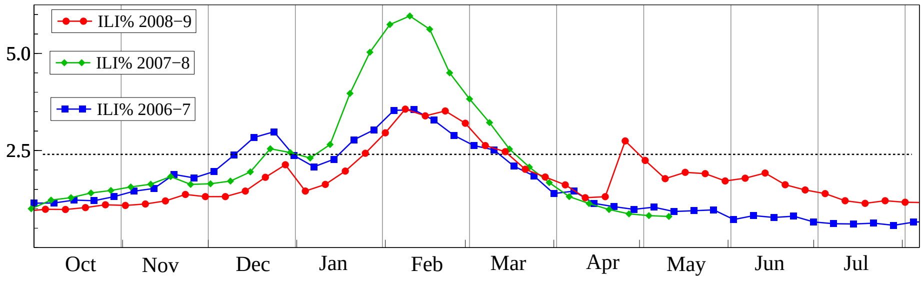


Figure 1: Percentage of visits for influenza-like illness (ILI%) reported by US Outpatient ILI surveillance network, national summary 2008-09 and previous two seasons. The national baseline is 2.4%.

3 The Google Flu Trends model

Ginsberg *et al.* (2009) have the full Google search data base at their disposal, using about 50 million of the most common search queries. After taking the logit transformation of all variables, they compute, for nine regions of the US, the correlations of each variable with the dependent variable. Then, for each variable, an average is computed over the nine regions based on the

[1] All dates are expressed in ISO year-month-day format. When referring to CDC weeks, we write year-Wweek, e.g.: 2009-W26. CDC weeks start on Sunday, while Google's weekly data starts on Monday. These are treated as being the same.

Z-transformed correlation. Next, the variables are ordered according to their correlation scores, and the best fitting set in terms of out-of-sample fit is selected.

The best fitting set consists of 45 queries, and the search index for these 45 (a single variable now) is used as the explanatory variable for all nine regions together. So the final model has just one right-hand side variable.

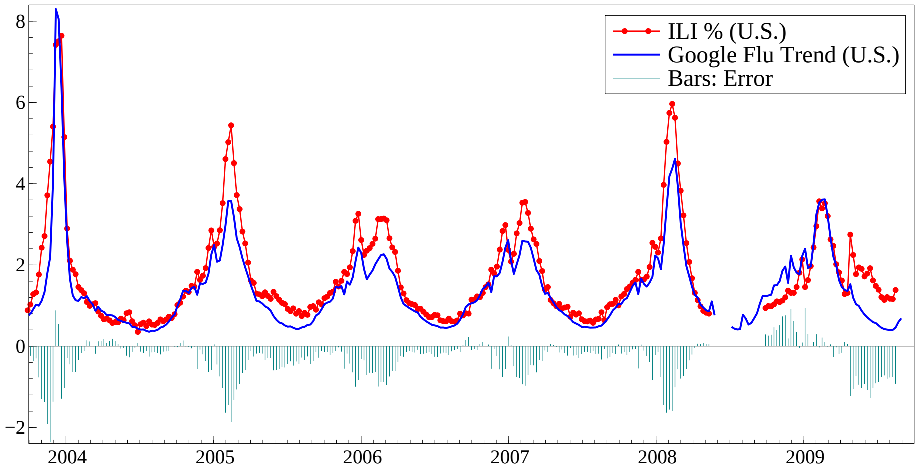


Figure 2: Comparison of ILI% as reported by the CDC and the Google Flu Trend indicator. The bars show the difference between the CDC data and Google Flu Trends.

Figure 2 compares the target variable (ILI%) to the Google Flu Trend indicator for the whole US.[2] The correlation is 0.92 for the first part of the sample (2003-10-05 to 2007-07-22), for the next 42 observations (2007-07-29 to 2008-05-11) it is 0.98. Both are in line with the numbers reported in the Nature paper. However, there are long periods with systematic errors: most peaks are underestimated, and the model seems to break down completely in 2008–2009 with over-estimation for November/December and under-estimation from April 2009. An immediate issue is that correlation is an inappropriate way to measure forecast performance: two variables can be far apart, but still highly correlated. The customary approach, used below, is to assess the forecast error using mean squared forecast error (MSE), or mean absolute percentage error (MAPE).

4 Google Flu Trends web site

Google Flu Trends presents a real-time version of the model developed by Ginsberg *et al.* (2009): based on aggregated search data an indicator of flu activity is computed, and plotted as in Figure 3.

[2] The data are available from *google.com/flutrends*.

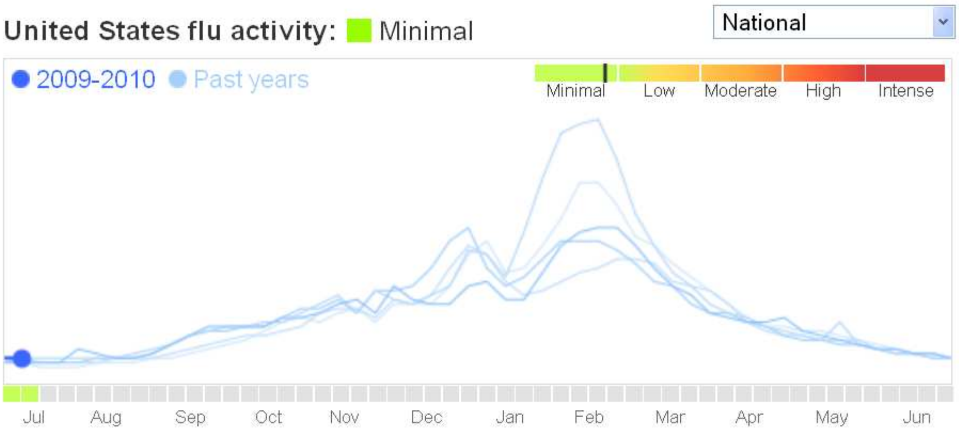


Figure 3: U.S. flu activity, as reported by google.com/flutrends/intl/en_us/ on 2009-07-16.

5 Modelling Percentage of Visits for Influenza-like Illness

We start by building a simple dynamic model for the logit of ILI%, the percentage of visits for influenza-like illness. This can provide a benchmark for comparison with extended models and Google Flu Trends. Appendix C discusses how the data was collected, and interpolated where necessary.[3]

The logit transformation is used to create an unbounded range (in principle) and to stabilize the variance:

$$\text{logit(ILI\%)} = \log\left(\frac{\text{ILI\%}}{100 - \text{ILI\%}}\right). \tag{1}$$

The variable with the missing values filled in is called lgILI*.

The initial model for the logit of ILI% is purely autoregressive: the only explanatory variables are the lagged dependent variable up to the 53rd lag. In the absence of another way to pick up the annual cycle, it is necessary to allow for such long lags. Analogous to Ginsberg *et al.* (2009), we first estimate up to the end of the 2006-7 season. Estimation is up to 2007-06-24 (week 26) using *Autometrics*[4] with reduction at 5%, outlier detection based on large residuals and without lag presearch.

Autometrics adds dummy variables for 7 potential outliers. The retained lags in the selected model are $1, 6, 26, 52, 53$. The coefficient on lags 26 is only moderately significant compared to the others, with a small coefficient. It may just have picked up some spurious effect, and the selection is run again using just lags 1, 6, 52, 53 and large outlier detection with reduction at 1%.

[3] All computations are done with OxMetrics 6 and PcGive 13, Hendry and Doornik (2009).

[4] Autometrics, Doornik (2009), implements the general-to-specific model selection approach developed by David Hendry, see Hendry (1995) for the foundations. All estimated models are linear regression models. Autometrics is particularly useful when variables are correlated, in which case stepwise regression works very badly. The procedure usually finds multiple candidate models, from which the final model is chosen by an information criterion.

The selected model is labelled $M1$ (standard errors are in parentheses):[5]

$$\begin{aligned} \textbf{M1}: \widehat{\text{lgILI*}}_t \;=\;& \underset{(0.020)}{1.018}\ \text{lgILI*}_{t-1} - \underset{(0.016)}{0.118}\ \text{lgILI*}_{t-6} + \underset{(0.041)}{0.286}\ \text{lgILI*}_{t-52} \\ & - \underset{(0.042)}{0.224}\ \text{lgILI*}_{t-53} - \underset{(0.053)}{0.162} + 6 \text{ dummies}, \end{aligned} \qquad (2)$$

$$\widehat{\sigma} = 0.124, \bar{R}^2 = 0.960, T = 350(\text{2000-10-08 – 2007-06-24}), k = 11.$$

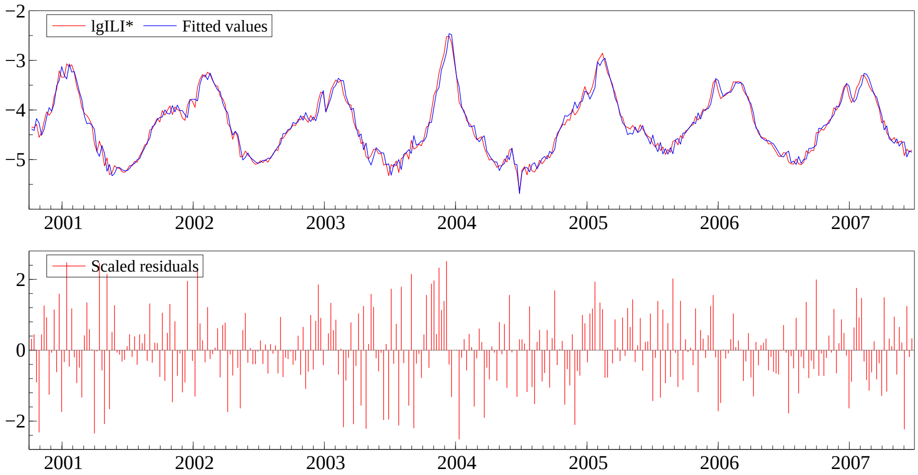


Figure 4: Model $M1$ for the logit of ILI: actual and fitted (top) and residuals (bottom).

The first lag is highly significant with a coefficient almost equal to one; the sum of the autoregressive coeffients is 0.96. The long lags capture the annual cycle. The residual diagnostics are good with some mild serial correlation (at 2.4% significance). The estimated model has $\widehat{\sigma} = 12\%$, which is a substantial improvement over the original $\widehat{\sigma}_y = 65\%$. The graphical analysis from the model is in Figure 4. Model $M1$ will only serve as a baseline for comparison.

The data show signs of calendar effects, in particular a drop in ILI% after Christmas. The next step is to extend the model, allowing for 73 additional variables as listed in Table 1. The Labor to Easter variables are included in the initial model up to the second lag, to allow for a delayed impact. Many variables are linear combinations of others, but that does not affect the model selection procedure (but we switch lag-presearch off). Lags 1 to 9 and 50 to 52 of the dependent variable are used in the initial model, which therefore has 85 variables to select from. Several terminal candidate models are found, which differ in the chosen lag lenghts and the calendar effects. Using the one with only lags 1, 2, and 6 of the dependent variable, and

[5] The standard error of the residual is denoted $\widehat{\sigma}$, $\bar{R}^2$ is the adjusted R-squared, T the sample size, and k the number of regressors. The standard error of the dependent variable is denoted $\widehat{\sigma}_y$; this is the same as $\widehat{\sigma}$ when regressing the dependent variable on an intercept only.

Name	*Description*	*Count*
Labor	first week of September	3
Autumn	first week of October	3
ThanksGiving	week of Thanksgiving Day	3
Christmas	week with December 25th	3
Washington	week of Washington's Birthday	3
Spring	first week of April	3
Easter	week of Easter	3
week1 – week52	indicators for each week	52
	Total added	73

Table 1: Calendar effects considered in the model

reselecting from the 73 calendar effects suggests the following simplification:

$$\begin{aligned}
\text{Holidays1}_t &= \text{ThanksGiving}_t - \text{Christmas}_{t-2},\\
\text{Holidays2}_t &= -\text{Washington}_{t-1} + \text{Easter}_t - \text{Easter}_{t-1} - \text{Spring}_t - \text{Spring}_{t-1},\\
\text{Winter}_t &= \text{week50}_t + \text{week51}_t + \text{week52}_t + \text{week3}_t + \ldots + \text{week6}_t + \tfrac{1}{2}\text{week7}_t,\\
\text{Summer}_t &= \text{week23}_t + \text{week25}_t + \text{week27}_t + \text{week29}_t + \text{week31}_t.
\end{aligned}$$

There is an increase in ILI% when school starts (usually the week of Labor day), but this is offset the week before and afterwards. There is a similar up effect in the first week of October ('Autumn'), largely offsetting a downward effect the week before. This is reasonably significant, but omitted from the model. Next, there are similar increases in the week of Thanksgiving Day and Easter. The Easter effect is offset the next period. In addition, there is a drop after the week of Washington's birthday, which usually coincides with the mid-winter recess, and Spring (or perhaps lagged from the week before, with the reduction happening afterwards — so all negative effects happen immediately after school holidays). Spring and Easter are usually associated with the spring break. Similarly, there is an increase in ILI in the winter, partially reduced by the delayed effect of the Christmas holidays. The summer effect is quite peculiar, but could be caused by the interpolation for some years, or the method of data collection.

Rerunning the reduced model with these additions at 5%, adding dummies for large residuals produces model $M2$:

$$\begin{aligned}
\mathbf{M2}: \widehat{\text{lgILI*}}_t \;=\;& \underset{(0.012)}{0.864}\ \text{lgILI*}_{t-1} + \underset{(0.016)}{0.141}\ \left[\text{lgILI*}_{t-2} - \text{lgILI*}_{t-6}\right]\\
&+ \underset{(0.028)}{0.144}\ \text{Holidays1}_t + \underset{(0.018)}{0.203}\ \text{Winter}_t + \underset{(0.018)}{0.090}\ \text{Holidays2}_t \qquad (3)\\
&- \underset{(0.021)}{0.130}\ \text{Summer}_t - \underset{(0.053)}{0.597} + 7\ \text{dummies},\\
\widehat{\sigma} = 0.103,\ &\bar{R}^2 = 0.972, T = 350(\text{2000-10-08 – 2007-06-24}), k = 14.
\end{aligned}$$

One of the dummies is for November 2003. All the diagnostic tests for this model are fine. The autoregressive parameters in (3) add to unity, and Model $M2$ (like $M1$ before) is really a

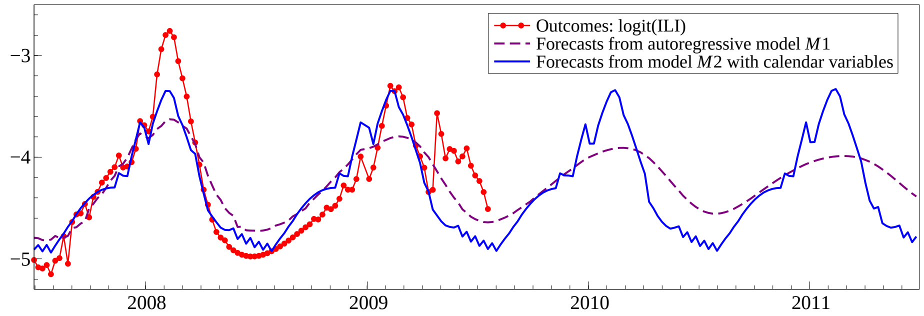


Figure 5: Dynamic forecasts for the logit of ILI%: 4-year ahead *ex ante* forecasts from 2007-W27 onwards.

model for the differences:

$$\begin{aligned}\Delta\widehat{\text{lgILI*}}_t &= \underset{(0.048)}{-0.604} - \underset{(0.011)}{0.138}\left[\text{lgILI*}_{t-1} - \text{lgILI*}_{t-2} + \text{lgILI*}_{t-6}\right] \\ &\quad +4 \text{ calendar effects} + 7 \text{ dummies}, \qquad (4) \\ &\widehat{\sigma} = 0.103, \bar{R}^2 = 0.575, T = 350(\text{2000-10-08 – 2007-06-24}), k = 13.\end{aligned}$$

In addition, using logarithms instead of logits would be virtually the same, so the models are effectively for the percentage change in ILI%. The MAPE's in particular would be very much bigger if expressed relative to the first differences. Correlations, as used by Ginsberg *et al.* (2009), will be considerably lower (as signalled by the change in $\bar{R}^2$).

The four-year ahead forecasts of models (2) and (3) are in Figure 5. The dynamic forecasts show the long-term cycle. The forecasts of the model with the calendar variables captures seasonal effects, as well as some asymmetry in the cycle. The sudden and unprecedented increase in 2009-04-26 (week 17) is not anticipated by any model, as should be expected. In practice, slightly different variations of the weekly variables could be found, but without much difference in fit or forecast performance.

Both models can be used to forecast one year ahead, then re-estimated with a new year of data (each time using the original dynamics), together with selection at 2.5% to add dummies for outliers if necessary. These one year ahead 'real-time' forecasts can be transformed to undo the logit transformation:[6]

$$\text{ILI\%} = 100\left[1 + \exp\left(-\text{logit(ILI\%)}\right)\right]^{-1}. \qquad (5)$$

When comparing the current results to Google Flu Trends, it must be remembered that Google Flu Trends is only designed to be two weeks ahead of the CDC data, it cannot be used to forecast further ahead. Therefore, the right metric for comparison are the two and one-step ahead forecasts. This is presented in Figure 6, where we compare the Google Flu Trend results with the two-step ahead forecasts of $M2$ and the actual outcomes. There is barely a need for

[6] We omit the bias correction for this transformation. Wallis (1987) gives an approximation to this, which, for the range of data considered here, barely exceeds 0.01.

summary statistics, because the difference is dramatic: the simple $M2$ model is very much better than Google Flu Trends. When there is a large unanticipated change, such as at the end of April 2009, it takes $M2$ two periods to correct (one period for one-step ahead forecasts), while Google Flu Trends never recovers. The reason is that the latter is a static model, while the former corrects because it has access to the actual past outcomes.

Table 2 presents the root mean squared forecast errors (RMSE) and mean absolute percentage error (MAPE) for Google Flu Trends and models $M1$ and $M2$.[7] This confirms that the simple autoregressive model is better, while model $M2$ provides further improvement. The dynamic forecasts (i.e. up to one year ahead) of $M2$ are given for comparison, and actually manage to be reasonably competitive with Google Flu Trends in 2008.

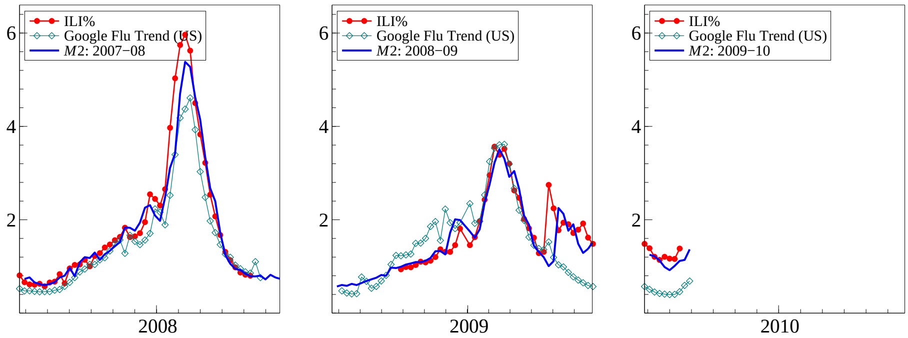


Figure 6: Two week ahead forecasts from model $M2$ (with School and Winter), Google Flu Trend estimates, and actual ILI. $M2$ models estimated up to 2007-W26 (left), 2008-W26 (middle) and 2009-W26.

6 Robustifying Google Flu Trends

The Google Flu Trend nowcasts are only useful for two weeks. After two weeks the actual ILI percentages are known (perhaps subject to some minor revisions). Because the Google Flu Trend model uses actual search volumes, it cannot produce *ex ante* forecasts, unless the search index is predicted. One insight from the estimated dynamic models is that the logit of ILI% is close to a random walk (i.e. the changes are very much closer to white noise). This explains why it is difficult to forecast the winter peaks (a clear failure of Google Flu Trends), and why a sudden shock persists, as seen at the end of April 2009 (causing a long period of forecast failure for Google Flu Trends). The two-step ahead forecasts from the dynamic model are effectively insured: for two or more periods ago it has access to the actual data, allowing it to 'self-correct'.

Hendry (2006) shows how forecasts in a non-stationary world with breaks can be robustified: use Google Flu Trends to estimate the change for the current and previous period, then apply

[7] The number of forecasts used for each year is reduced from 52 because of missing Flu Trend forecasts or ILI% outcomes. All forecast statistics are given for 52-week years, see Appendix C.

	Google Flu Trend	$M1$		$M2$		
		1-step	2-step	1-step	2-step	dynamic
	2007 Week 27 – 2008 Week 20 (46 forecasts)					
RMSE	0.58	0.30	0.54	0.22	0.36	0.82
MAPE	18	9	12	7	10	16
	2008 Week 42 – 2009 Week 26 (37 forecasts)					
RMSE	0.63	0.34	0.46	0.32	0.41	0.65
MAPE	30	10	14	10	12	30
	2009 Week 27 – 2009 Week 33 (7 forecasts)					
RMSE	0.80	0.13	0.17	0.14	0.18	0.30
MAPE	65	8	11	10	12	20

Table 2: Forecast statistics for ILI of Google Flu Trends and models $M1$ and $M2$.

this to the actual outcomes. Let F_t denote the Google Flu Trend nowcasts and ILI_t the actual ILI percentages (only known up to two periods ago), then:

$$\begin{aligned} \widetilde{\text{ILI}}_{t-1} &= \text{ILI}_{t-2} + (F_{t-1} - F_{t-2}), \\ \widetilde{\text{ILI}}_t &= \text{ILI}_{t-2} + (F_t - F_{t-2}). \end{aligned} \tag{6}$$

It seems preferable to apply this approach to the logit transformation, after which the anti-logit (5) can be taken:

$$\begin{aligned} \widetilde{\text{logit ILI}}_{t-1} &= \text{logit ILI}_{t-2} + (\text{logit}F_{t-1} - \text{logit}F_{t-2}), \\ \widetilde{\text{logit ILI}}_t &= \text{logit ILI}_{t-2} + (\text{logit}F_t - \text{logit}F_{t-2}). \end{aligned} \tag{7}$$

Table 3 shows how much (6) and (7) improve on the original Google Flu Trends nowcasts. Model $M2$ is now just beaten in 2008 on RMSE, but not on MAPE. Indeed, a visual inspection of the two-step forecasts from (7), see Figure 7, shows that the new approach corrects after two periods, unlike the original Flu Trends, which can go wrong for a long time. However, comparing Figure 7 to Figure 6 confirms a preference for model $M2$, as borne out by the MAPEs.

Table 3 also reports the pooled forecasts of $M2$ and robustified Google Flu Trends. This is based on the average of the logit forecasts, after which the logit transformation is undone.

	Google	Robustified Google Flu Trends					
	Flu Trends	*levels*		*logits*		*pooled with* $M2$	
		1-step	2-step	1-step	2-step	1-step	2-step
	2007 Week 27 – 2008 Week 20 (46 forecasts)						
RMSE	0.58	0.26	0.37	0.29	0.38	0.19	0.27
MAPE	18	10	12	11	13	7	8
	2008 Week 42 – 2009 Week 26 (37 forecasts)						
RMSE	0.63	0.34	0.40	0.32	0.38	0.30	0.37
MAPE	30	14	16	12	15	10	11
	2009 Week 27 – 2009 Week 33 (7 forecasts)						
RMSE	0.80	0.09	0.12	0.05	0.06	0.09	0.11
MAPE	65	5	9	3	4	7	7

Table 3: Forecast statistics for ILI% of Google Flu Trends and robustified nowcasts; levels corresponds to (6), logits to (7). The pooled model takes the average of the logit forecasts

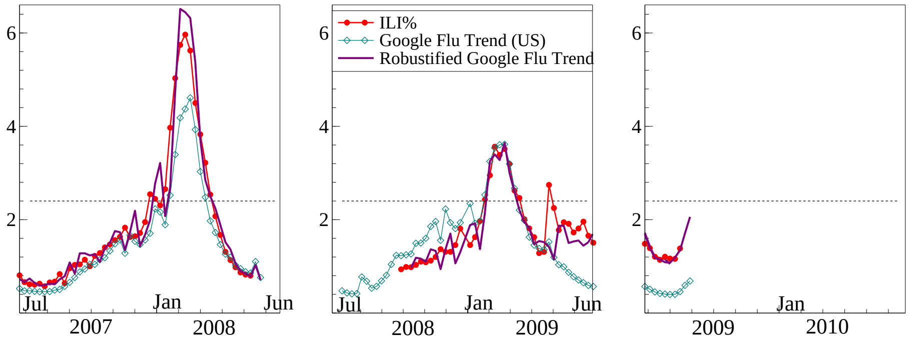


Figure 7: Google Flu Trend estimates, the robustified nowcasts (based on logits, see (7)), and actual ILI%.

7 Modelling ILI% with Google Trends

The CDC web site lists the following flu symptoms:

- Fever (usually high)
- Headache
- Tiredness (can be extreme)
- Cough
- Sore throat
- Runny or stuffy nose
- Body aches
- Diarrhea and vomiting (more common among children than adults)

We now investigate whether there is some benefit for one and two-step forecasting from adding Google Trend variables to the model. For this purpose we used the flu symptoms as search terms. There is not enough search volume for the joint query on diarrhea and vomiting, so these were done as separate terms. An additional four flu related and five holiday related queries were added to the set of potential variables. The holiday variables could be important, given the influence of holiday effects found earlier. The variables are listed in Table 4, and plotted in Figures 8 and 9. Tiredness peaks in the summer, unlike most others. Vomiting has a strong peak in the week of Christmas, next to an upward trend, both of which seem unrelated to flu prevalence (c.f. Figs. 4 and 5). Similar patterns in the explanatory variables might induce extra calendar effects or trend terms in the regression model. In our analysis extra effects in relation with calendar effects and trends in the extra explanatory variables were not significant. Flu symptoms is mostly flat, except for a pronounced peak in 2009W18 and W19 (last week of April, first week of May). If selected, this could have a large impact on the forecasts.

Flu symptoms	*Other flu*	*Holiday terms*
body aches	cold remedy	child care
cough	flu remedy	homework
diarrhea	flu symptoms	kids camp
headache	flu vaccine	school holidays
high fever		Walt Disney
runny nose		
sore throat		
tiredness		
vomiting		

Table 4: Search terms used in the extended model

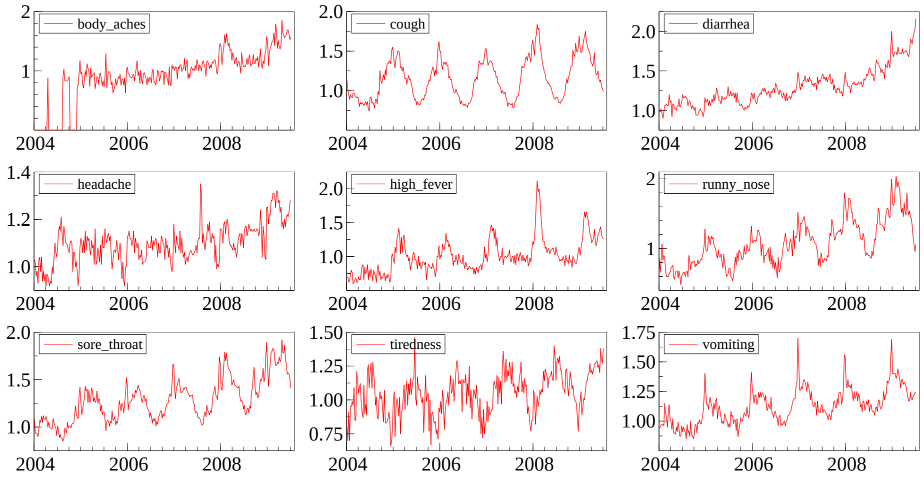


Figure 8: SVI of flu-symptom search terms.

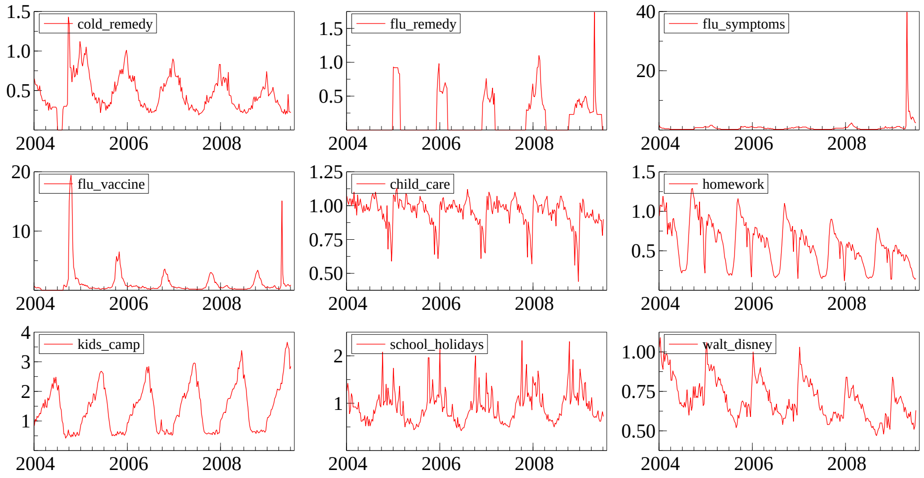


Figure 9: SVI of other flu-related and holiday search terms.

The additional explanatory variables are added in logs (they can be larger than 100 in principle, although most are between zero and two).[8]

The starting point is model $M2$, equation (3), without the dummy variables, but augmented with the 18 search variables up to the first lag, so an additional 36 regressors. The intercept

[8] The body aches, cold remedy, and flu vaccine variables have some zeros in 2004. We replaced the zeros by the 2004 average for body aches, and the minimum for the other two (0.89, 0.25, 0.3 respectively) before taking logarithms. Flu remedy, on the other hand, has genuine zeros throughout the summer period; we added 0.01 to the variable before taking the log.

$\log$ of	Dynamic model $M4$			Static model $M5$		
body aches		$./-$				$./-$
cough	$+/.$	$+/.$	$+/.$	$./+$	$+/.$	$+/.$
high fever	$+/.$	$+/.$	$+/.$	$+/+$	$+/+$	$+/+$
runny nose				$./-$	$./-$	$./-$
sore throat				$./-$		
tiredness	$./-$	$./-$	$./-$	$-/.$	$-/.$	$-/.$
vomiting				$+/-$		
cold remedy	$+/.$		$+/.$	$+/+$	$+/+$	$+/.$
flu symptoms			$+/.$	$+/.$	$+/.$	$+/+$
flu vaccine				$-/.$	$./-$	$./-$
child care			$./+$			
homework	$./+$			$+/.$	$./+$	$./+$
kids camp	$./+$		$./+$	$./+$	$./+$	
school holidays				$./-$		$./-$
Walt Disney		$+/.$				$+/.$
Dummies	9	15	9	31	43	50
$\widehat{\sigma}$	0.0731	0.0698	0.0755	0.0805	0.0851	0.0820
Sample ends week 26 of	2007	2008	2009	2007	2008	2009

Table 5: $M4$: Sign of selected flu and holiday related search variables from Google Trends. Samples starting 2004-01-18 and ending in week 26 of 2007, 2008 and 2009 respectively. Notation is sign at t/sign at $t-1$, with a dot indicating absence.

and the four calendar variables are always forced into the model, but the two terms involving lags of the dependent variable are allowed to be deselected. Now *Autometrics* reduction is run at 1% (2.5% for estimation up to 2007-W26), with indicator saturation, so there are $T + 38$ variables to select from.[9] The estimated models are given in Table 5, with the report limited to the signs of the selected search variables. For example, an entry of +/. for cough means that $\log(\text{cough})_t$ has a positive sign, and that $\log(\text{cough})_{t-1}$ is not in the model. There is a difference in the variables that are selected, although the signs are remarkably consistent between samples and models. The residual diagnostics for all models support the assumption of independent and normally distributed errors.

Most selected variables relate to flu symptoms, with cough, high fever and tiredness selected throughout. Tiredness has the opposite cycle to most others, and enters with a negative sign. Vomiting and sore throat are quite marginal where selected. From the remaining search terms, cold remedy, child care and kids camp are the most important ones in model $M4$. Essentially the same forecast performance as $M4$ is achieved if only $\log(\text{cough})_t$, $\log(\text{tiredness})_{t-1}$, $\log(\text{child care})_{t-1}$ are used.

[9] Indicator saturation is a method of robust estimation which allows for simultaneous model selection and robustness. See Hendry, Johansen and Santos (2008) and Johansen and Nielsen (2009); Doornik (2008) discusses the method implemented in Autometrics.

	$M2$		$M4$		$M5$		pooled $M2$ and $M5$	
	1-step	2-step	1-step	2-step	1-step	2-step	1-step	2-step
	2007 Week 27 – 2008 Week 20 (46 forecasts)							
RMSE	0.22	0.36	0.18	0.25	0.24	0.33	0.18	0.27
MAPE	7	10	7	10	10	13	8	9
	2008 Week 42 – 2009 Week 26 (37 forecasts)							
RMSE	0.32	0.41	0.31	0.39	0.23	0.29	0.22	0.26
MAPE	10	12	10	13	9	12	7	9

Table 6: Forecast statistics for ILI% of models $M2$ (autoregressive with calendar effects), $M4$ (autoregressive, calendar effects and Google Trends data), $M5$ (with Google Trends data only), and $M6$ (pooled $M4$ and $M5$).

The improvement in forecasting over model $M2$ is not so easy to see in a graph, but the summary statistics in Table 6 indicate that $M4$ is an improvement over $M2$. However, it is not better than the pooled forecasts of $M2$ and robustified Google Flu Trends, cf. Table 3.

As a closer analogue to Google Flu Trends, I estimate a model that only relies on the search index variables: the candidate variables consist of the intercept and the 18 search together with their first lag.[10] Robust estimation using indicator saturation at 2.5% is used. More variables are selected into model $M5$, consistently across samples, although to a lesser extent for the holiday related search variables. Robustified forecasts are used for $M5$, because it has no dynamics. Finally, the pooled forecasts of $M2$ and $M5$ are reported, using equal weights. In this case there was a small benefit from combining in levels. These pooled forecasts are easily the best of those considered here for the 2008-09 season, with two-step RMSE and MAPE at least twice as good as Google Flu Trends. For the 2007-08 season there is not much between $M4$ and the pooled $M2 + M5$, but in all cases the use of Google Trends data has improved on the forecasts from the dynamic model $M2$ which uses calendar effects only.

Table 7 shows what happens when Google Flu Trend estimates are added as an additional regressor to models, $M2$, $M4$ and $M5$. It is never significant in $M4$. In the others it has t-values ranging from 3.8 to 2.1. Adding it to $M2$ makes the forecasts worse, showing the difference from pooling with Google Flu Trends, which always gave an improvement.

8 Conclusions

The Google Flu Trends model is designed to fill the two week gap between the release of CDC's flu report and the present. The objective is to show that search activity data can be used to estimate current levels of activity. The second objective is to provide an early warning system that can aide with planning and improve the state of public health. Unfortunately, as was shown, the estimates (i.e. forecasts of the two most recent weeks) failed to detect a recent large decrease

[10] Models with only the contemporaneous search indices were also tried, but considerably inferior to those with one lag.

Estimation up to	Model $M2$		Model $M4$		Model $M5$	
2007-W26	0.127	(0.037)	0.017	(0.047)	0.220	(0.066)
2008-W26	0.132	(0.034)	0.030	(0.039)	0.190	(0.062)
2009-W26	0.054	(0.024)	-0.022	(0.029)	0.084	(0.040)

Table 7: Coefficients and (standard errors) on logit of Google Flu Trends when added as an additional variable to models $M2$, $M4$ and $M5$.

in flu activity. Even a simple autoregressive model was shown to have better two-step ahead forecasts than the Google Flu Trend estimates (as measured by RMSE and MAPE). A third objective might be to provide forecasts for each state, so at a lower level of aggregation than the CDC provides. However, this requires a good model at the more aggregate level before trying to apply it at a disaggregate level.

Robustified forecasting, as detailed in Hendry (2006) turned out to be a very useful procedure in this case. For Google Flu Trends it almost halves the RMSE as well as the MAPE. I also used it for the static model with search data, although the effect is not quite so dramatic.

The primary purpuse of the purely autoregressive model was to serve as a baseline. The next stage was to build a serious model with calendar effects. The weekly terms could be condensed into four variables: Winter (weeks 50 to 7, except for 1 and 2), two holiday variables and a summer effect. This model has forecast performance that is comparable to Google Flu Trends.

Two additional models are formulated to investigate whether search engine data can help. The search index for 18 terms and their first lag provided the extended data base. Autometrics was used again to select models from this large candidate set. A dynamic and a static model was developed, in both cases providing an improvement for the 2007 and 2008 nowcasts. Pooling these two models did provide substantially better nowcasts, so search data can indeed be useful. It was found that search activity for ‘cough’, ‘high fever’ and ‘child care’/‘homework’ has a positive impact on the percentage of visits for influenza-like illness, while ‘school holidays’ and ‘tiredness’ have a negative impact.

References

Castle, J. L., Fawcett, N. W. P., and Hendry, D. F. (2009). Nowcasting is not just contemporaneous forecasting. *National Institute Economic Review*, **Forthcoming**.

Castle, J. L., and Shephard, N. (eds.)(2009). *The Methodology and Practice of Econometrics: Festschrift in Honour of David F. Hendry*. Forthcoming, Oxford: Oxford University Press.

Choi, H., and Varian, H. (2009). Predicting the present with google trends. mimeo, Google Inc.

Doornik, J. A. (2008). General-to-specific model selection with more variables than observations. Mimeo, Department of Economics, University of Oxford.

Doornik, J. A. (2009). Autometrics. in Castle, and Shephard (2009).

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*, **457**, 1012–1015. Letters, doi:10.138/nature07634.

Hendry, D. F. (1995). *Dynamic Econometrics*. Oxford: Oxford University Press.

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

Hendry, D. F., and Doornik, J. A. (2009). *Empirical Econometric Modelling using PcGive: Volume I* 6th edn. London: Timberlake Consultants Press.

Hendry, D. F., Johansen, S., and Santos, C. (2008). Automatic selection of indicators in a fully saturated regression. *Computational Statistics*, **33**, 317–335. Erratum, 337–339.

Johansen, S., and Nielsen, B. (2009). Saturation by indicators in regression models. in Castle, and Shephard (2009).

Wallis, K. F. (1987). Time series analysis of bounded economic variables. *Journal of Time Series Analysis*, **8**, 115–123.

A Google Trends

Figure 10 shows a typical graph that is produced by Google Trends. The top shows the Search Volume Index (SVI), based on a subset of Google's search database. The SVI is available world-wide, by country or for different regions, provided there was enough search volume. Currently this information is updated daily. The bottom graph displays the news reference volume: the number of times 'car insurance' appeared in Google News stories. After registering, the SVI data in the graph can be downloaded.[11]

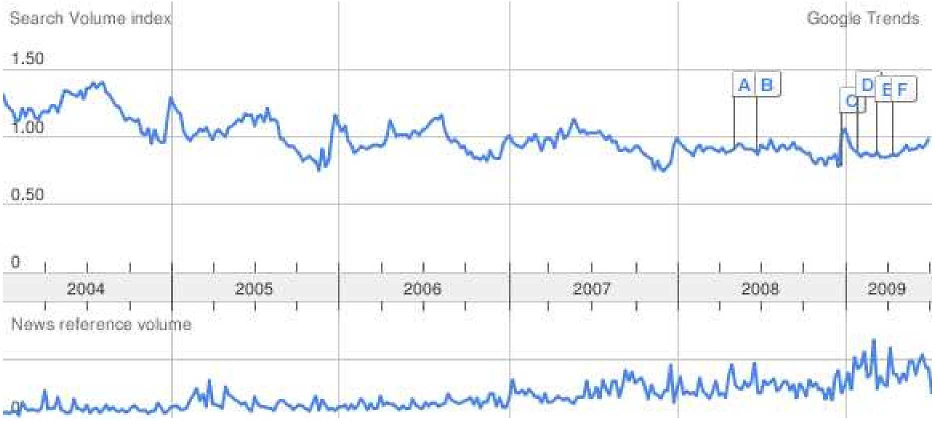


Figure 10: Search volume index for search term 'car insurance' in the U.S. produced by Google Trends on 2009-07-16.

The SVI for region r is constructed as follows.[12] First the percentage of the total search volume that

[11] The news data can be downloaded from Google Insights for Search, and this could perhaps be useful to control for wide fluctions in the SVI for some terms.

[12] All dates are expressed in ISO year-month-day format: YYYY-MM-DD.

relates to the term is computed for every day for the specified region. The search data in Figure 10 are divided by the full sample mean to give the variable a mean of one over the displayed period. Google calls this ‘relative scaling’ and reports the data with two decimals. The data that is used for modelling uses ‘fixed scaling’, i.e. scaled to the average for January 2004. Finally, weekly observations are computed as an average of the daily data.[13] Figure 11 shows the two versions of the variables.

Using subscript τ for daily data and t for weekly data, and $V_{\tau,r}$ for the search volume on term V in region r, with $T_{\tau,r}$ the total search volume:

search share	$S_{\tau,r} = \dfrac{V_{\tau,r}}{T_{\tau,r}}, \quad \tau = 1, ..., 7T,$	
fixed SVI	$s_{t,r} = \frac{1}{7}\frac{1}{\mu_r} \sum_{\tau=\text{Sunday}}^{\text{Saturday}} S_{\tau,r},$	$\mu_r = \frac{1}{31} \sum_{\tau=2004-01-01}^{2004-01-31} S_{\tau,r},$
relative SVI	$s^R_{t,r} = \frac{1}{7}\frac{1}{\mu^R_r} \sum_{\tau=\text{Sunday}}^{\text{Saturday}} S_{\tau,r}$	$\mu^R_r = \frac{1}{7T} \sum_{\tau=1}^{7T} S_{\tau,r}.$

The fixed SVI cannot be computed when $\mu_r = 0$; the relative SVI has a mean of unity over the selected sample. Both are positive, and, in principle, unbounded.

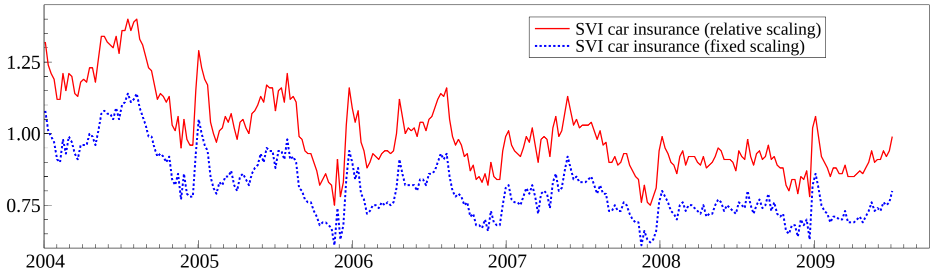


Figure 11: Search Volume Index using relative and fixed scaling

Unfortunately, the Google Trends data is subject to revisions. Downloading the SVI with fixed scaling for US car insurance on 2009-08-05 gives different observations for the entire historical period, as the first panel of Figure. 12 shows. The second panel shows the residuals from regressing the new generation of the variable on the old generation. The shaded area corresponds to ± 2 standard errors. The correlation between the two versions is 0.97625. Such revisions hamper the use of Google Trends for statistical modelling.

B Google Insights for Search

The search volume data are represented somewhat differently on Google Insights for Search, as illustrated in Figure 13. The plotted data is monthly, unlike Figure 10, however the downloadable data remains

[13] This is my hypothesis, based on the fact that standard errors are given for the weekly data.

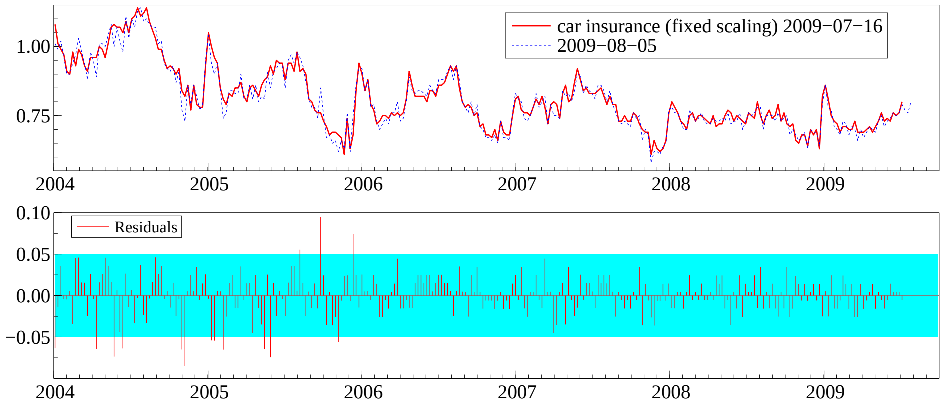
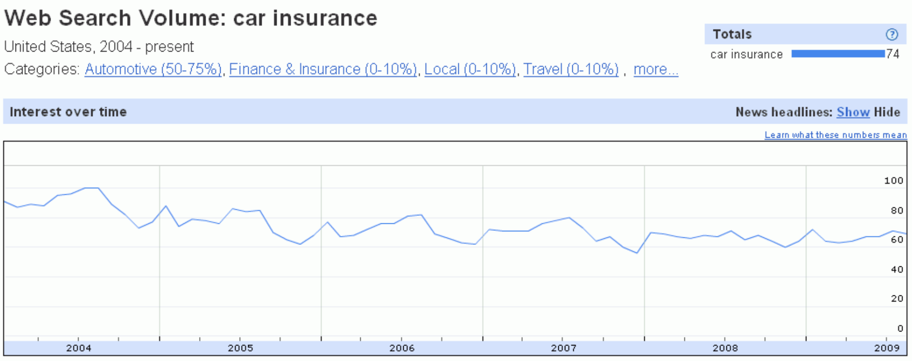


Figure 12: SVI for U.S. car insurance using fixed scaling, data for 2009-07-16 and 2009-08-05 (top panel) and residuals of regressing the newer generation on the older variable.

weekly. Now the data, labelled Web Search Volume, are scaled by the maximum over the selected sample, then multiplied by 100 and reported without any decimals:

$$\text{Web Search Volume} \quad s^*_{t,r} = \frac{1}{7}\frac{100}{\max_t(s_{t,r})}\sum_{\tau=\text{Sunday}}^{\text{Saturday}} S_{\tau,r}.$$



Web Search Volume: car insurance

United States, 2004 - present

Categories: Automotive (50-75%), Finance & Insurance (0-10%), Local (0-10%), Travel (0-10%) , more...

Totals

car insurance 74

Interest over time

News headlines: Show Hide

Learn what these numbers mean

Figure 13: Search volume index for search term ‘car insurance’ in the U.S. produced by Google Insights for Search on 2009-08-05.

There is some discrepancy between the data reported by Google Trends and Google Insights, which exceeds the rounding errors, see Figure 14. Just like Google Trends, the data changes from day to day, as illustrated in Figure 15.

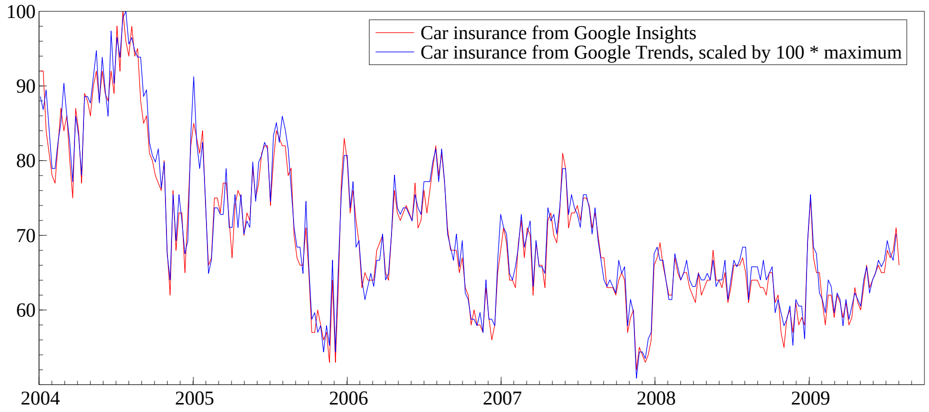


Figure 14: U.S. car insurance from Google Insights compared to Google Trends, with the latter scaled as in Insights.

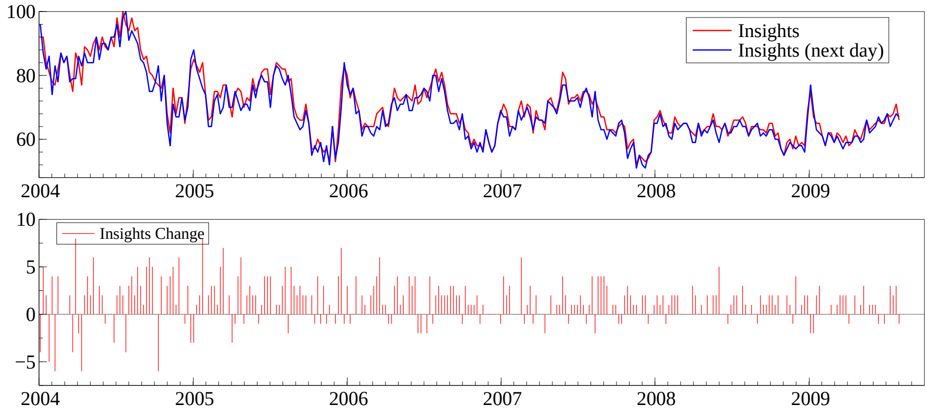


Figure 15: Car insurance data for 2009-08-05 and 2009-08-06 (top panel) and the difference between the two (bottom panel).

C Data issues

The variable of interest is the %Weighted ILI from Sentinel Providers, as reported on the CDC web site. The data for 2008-W40 to 2008-W29 are taken from the 2008-W29 report; for 2006-W40 to 2008-W20 from the final report for the 2007-08 season; for 2003-W40 to 2006-W39 from the final data tables; For 1999-W40 to 2003-W20 (with no data for W21 to W39) from the 2006-06 end report. The 2003-W21 to 2003-W39 data is taken as the reginal data, weighted by the 2002 population estimates from the Census Bureau; this is very accurate.

The ILI data used for modelling starts in 1999-10-03, but there are gaps, as can be seen in Figure 16. These gaps are during the summer (week 21 to 39) of 2000, 2001, 2003 and 2008, when the ILI% is low.

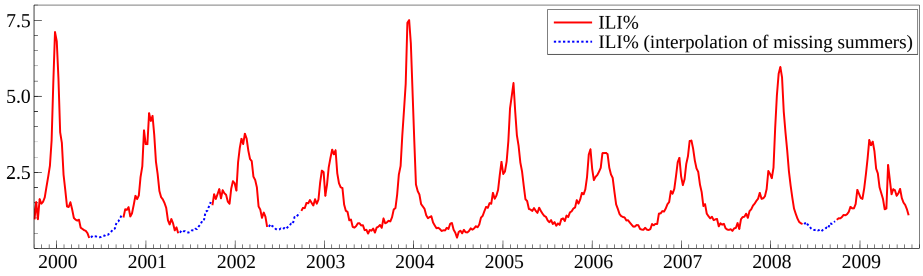


Figure 16: ILI%, together with interpolated missing values.

For modelling purposes it is useful to ‘fill the gaps’. We do this as follows. After taking the logit, compute the average change for weeks 21 to 40 from years 2003 to 2008. Apply this to each year with missing data, each time spreading the required total change (to make interpolated week 40 the same as actual week 40) evenly over the period. Finally, undo the logit transformation to obtain the interpolated ILI%. The created values are shown with a dotted line in Figure 16. This interpolation is somewhat ad hoc, of course, but the benefits of a larger sample are likely to outweigh the error that we make.

The data sample contains two years with 53 weeks (2003 and 2008), with week 53 starting on Sunday 28 December. Weeks 53 are removed from the sample by assigning 4/7th to week 52 and 3/7th to week one. This is the final data adjustment before modelling.