GOOGLE DISEASE TRENDS: AN UPDATE

by Patrick Copeland, Raquel Romano, Tom Zhang, Greg Hecht, Dan Zigmond, Christian Stefansen

ABSTRACT: The purpose of Google Flu Trends (GFT) is to use search keyword trends from Google.com to produce a daily estimate, or nowcast, of the occurrence of flu two weeks in advance of publication of official surveillance data. While not covered in detail in this paper, Google Dengue Trends, launched in June 2011, is a service that uses similar techniques to track Dengue fever. During the 2012 flu season we observed our algorithm overestimating influenza-like illness (ILI). We have concluded that our algorithm for Flu and Dengue were susceptible to heightened media coverage and have since developed several improvements.

Background

Our goal for the project is to produce a system that is accurate, fine-grained, and timely. The primary measures of forecast accuracy are for season start and peak; the magnitude of cases is generally a secondarily concern.

In addition to the general public, an important target audience for GFT has been public health officials, who can benefit from reliable daily estimates and often make far-reaching decisions based on predicted flu incidence (such as how to stock and distribute vaccine, and the content of public health messaging). During the development of GFT we met regularly with a variety of health officials, and we convened with more than a dozen leaders from around the world in 2010.

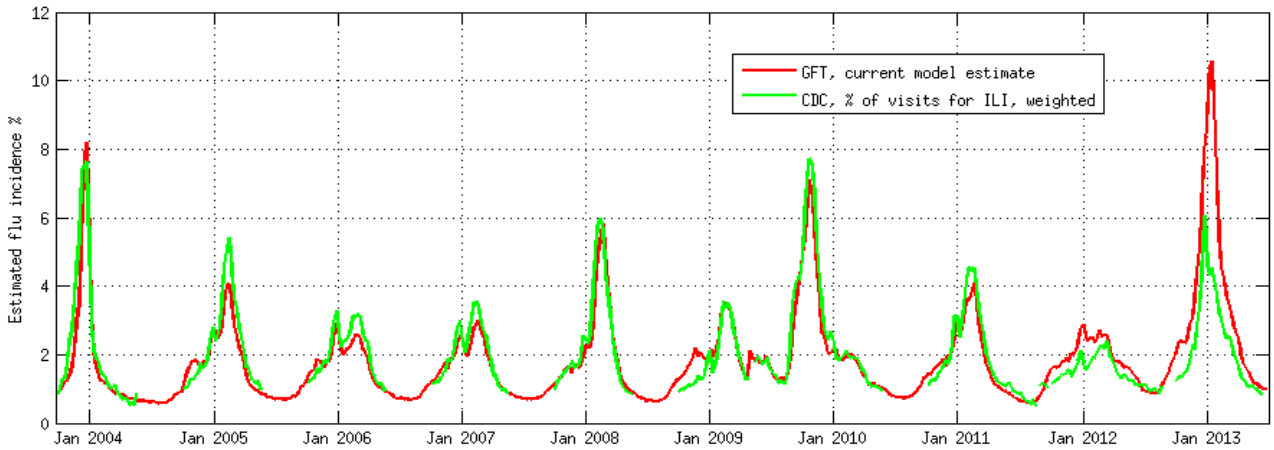
The original GFT model was created in 2008 and released in multiple countries. The country selection was limited by availability of "ground truth" data in the form of incidence reports of ILI, typically provided by a national or international public health agency. The flu surveillance data itself was publicly available or acquired via a partnership license. Since the initial model's release, there has been one update in response to slightly underestimating 2009 H1N1 swine flu (PloS 2011). From the launch in 2008 until the 2012-13 season, the highest estimation error for national flu incidence was 1.13 percentage points (week starting Jan. 1, 2012: CDC data 1.74%, GFT estimate 2.86%), and the mean absolute error during this period across all weekly estimates was 0.30 percentage points. However, in the 2012-13 season, the overestimation peaked at 6.04 percentage points, an estimate more than twice the CDC-reported incidence (week starting Jan. 13: CDC data 4.52%, GFT estimate 10.56%). (Also see Nature 2/13/13, When Google got Flu Wrong for an external report.)

This paper addresses several questions related to our model's recent performance: Why were this season's predictions so high? Is our model too simple? Were there unforeseen side effects from the 2009 update? Does this reveal a phenomenon not captured in incidence data provided by the US Centers for Disease Control and Prevention (CDC)?

Algorithm

The premise of our model is that certain search query terms on

Google Flu Trends predictions vs. CDC, 2004-2013



Google.com, such as "flu symptoms," have a high historical correlation with doctor visits for ILI and so may be useful predictors of such visits in the future.

The basis for our algorithm is the continually updated ILI target signal data for a particular region, such

as the percentage of physician visits in which people report symptoms of the flu. Usually these data are provided at the national level, but in some places (such as Utah) it is also offered at the state level.

The second key element in our algorithm is a set of approximately 50 million query terms run through Google's servers. A challenge with this approach is that volumes of a particular query are not constant and can vary over time, both short-term and long-term, and by location and language. For instance, during the holiday season, more people search for "gift" than at any other period. Similarly, overall usage of Google search varies throughout the year and is growing over time. We handle this by computing the **query fraction** of each query term: the total count of a query term in a given location is aggregated weekly and normalized by the total count of all queries issued in that week at that location.

The third step in our algorithm is to identify a small subset of the millions of query terms that provide the highest correlation with the CDC published target signal. The summed query fractions of this subset are used to obtain a fraction history of ILI-related queries. We then fit the query fraction and target signal curves to a univariate linear regression model (per country or region) that predicts the daily target signal from daily queries. For a more detailed discussion of the algorithm, see Nature 457, pp. 1012-1014.

What happened this year?

The current model, while a well performing predictor in previous years, did not do very well in the 2012-2013 flu season and significantly deviated from the source of truth, predicting substantially higher incidence of ILI than the CDC actually found in their surveys. It became clear that our algorithm was susceptible to bias in situations where searches for flu-related terms on Google.com were uncharacteristically high within a short time period. We hypothesized that concerned people were reacting to heightened media coverage, which in turn created unexpected spikes in the query volume. This assumption led to a deep investigation into the algorithm that looked for ways to insulate the model from this type of media influence.

The sensitivity of our algorithm to sudden changes in query volume, and thus the importance of keeping the list of queries confidential, has been known for some time. When we launched GFT in 2008 the New York Times published a story that included an example query that was actually used in the model. We immediately saw traffic increase on that query term. We expect that divulging the query list would result in skewing the model, negating the usefulness of the service.

To compensate, we have "spike detectors" in place to identify patterns of sharp increases in query traffic as "inorganic" and remove them from the model. The system receives time series data of the flu-related queries as input and validates whether the latest counts are within expectation, based on statistical variations from what we have seen in the past. As far back as 2008, we knew that most query spikes caused by news attention tend to last for 3 to 7 days. The problem is that our detector solved for short-term spikes, but didn't consider unusually high query volume that lasted for an entire season.

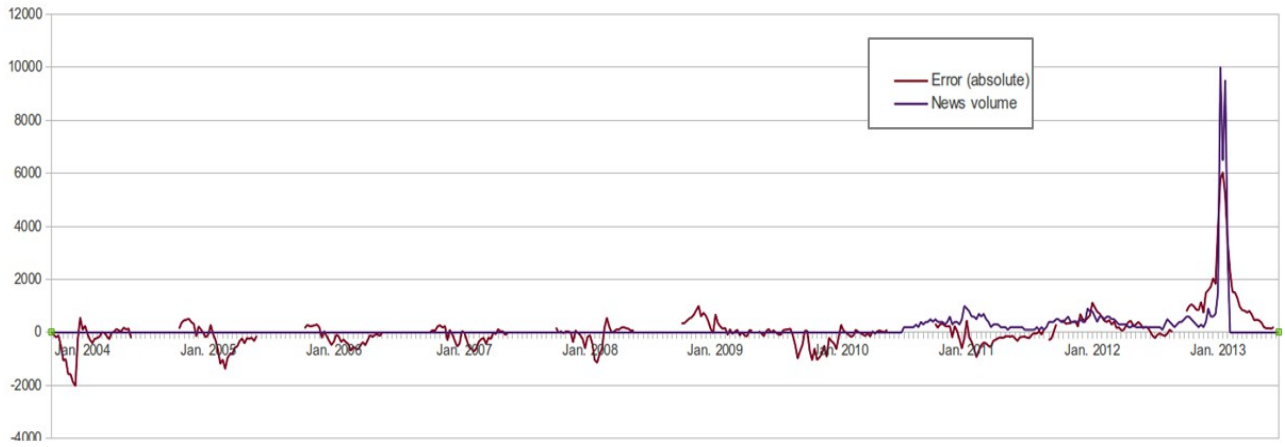
Finally, while we evaluate the model after every flu season, we have not updated the model annually because the model built in 2009 performed quite well on subsequent years. Updating the model after each flu season should improve its estimation accuracy by informing it with longitudinal data, although additional data alone do not address the open question of how to deal with truly anomalous years.

Conclusion

We have concluded that our algorithm for Flu and Dengue were susceptible to heightened media coverage. While we haven't observed an effect on our predictions for Dengue from media coverage, we believe that like Flu Trends, it was vulnerable to similar spikes. We've addressed this with two areas of improvement: 1) dampening anomalous media spikes and 2) using ElasticNet.

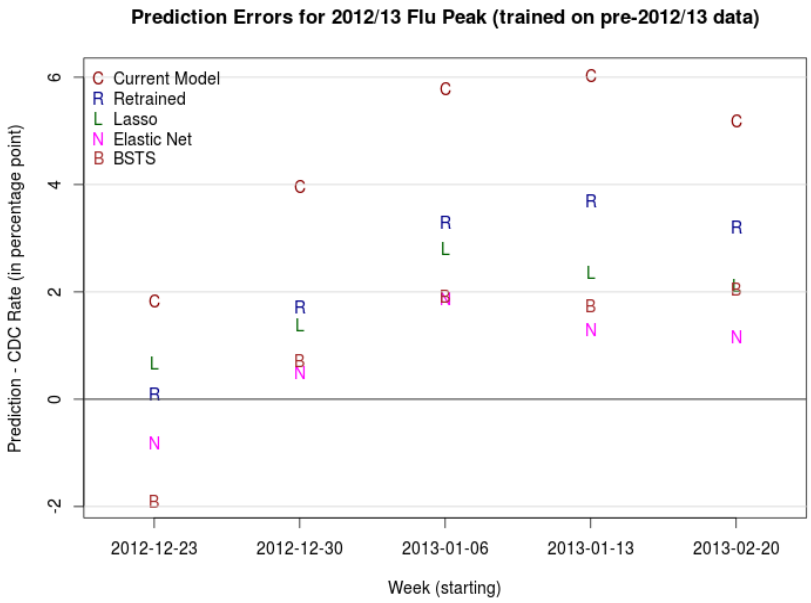
First, a given query term in the model has an influence proportional to its contribution to the total query fraction of flu-related terms. Hence, the model estimates are susceptible to significant changes in the seasonal

Media volume and Prediction Error Rate, 2004-2013



shape of even a single query term. Indeed, the 2012-13 season did experience a protracted surge in several flu-related queries that were not indicative of high flu incidence. These anomalous surges in query volume were due to flu-related media reports, and we can use an independent measure of flu in the news media to modulate the contribution of certain flu-related queries during estimation.

The second improvement addresses the absence of explicit coefficients for query terms in the model. We experimented with regularized regression models to the query data, e.g. Lasso [Tibshirani] and Elastic Net [Zou, et. al.] models, where we made improvement to the Least Angle Regression algorithm [Efron et. al.] to handle large number of query terms (in the order of millions). These regression models significantly improve over the incumbent, but still slightly overpredict the 2012-13 flu levels.



Week	CDC Sentry Data, % Weighted ILI [CDC]	Current model in production [GFT]	Current model retrained '03-'12	Lasso	ElasticNet	BSTS
2012-12-23	6.07%	7.90%	6.17%	6.74%	5.26%	4.18%
2013-12-30	4.65%	8.62%	6.38%	6.03%	5.15%	5.38%
2013-01-06	4.33%	10.11%	7.62%	7.14%	6.20%	6.25%
2013-01-13	4.52%	10.56%	8.21%	6.88%	5.82%	6.27%
2013-01-20	4.22%	9.41%	7.44%	6.35%	5.39%	6.27%

For more information, please contact:

google.org