Failure Analysis of Google Flu Trends

BAX-401-002

Ashi Malik
Jason Stergiou
Shushmita MirazkarSubash
Thea Castañeda
Vinit Hasija

Table of Contents

Executive Summary

Google Flu Trends ("GFT") was a model that sought to detect influenza cases by utilizing user search activity but failed within a few years of launch. We examined GFT's development, implementation, failure, and update response to identify potential biases which led to failure: measurement and omitted variable bias in development, failure to recognize potential user social bias and user interaction bias in implementation, and potential emergent and temporal bias. We recommend management choose appropriate statistical methods for development; incorporation of the interaction between data, algorithm, and users; recalibration of the model; utilization of external resources; and constant re-evaluation and adjustment throughout implementation to adjust for potential unanticipated changes.

Introduction

GFT was a forecasting tool which used flu-related search activity to generate "nowcasts" of influenza-like illness ("ILI") cases to predict seasonal and pandemic influenza. At its launch in 2008, GFT accurately forecasted ILI cases from outpatient visits reported by the Centers for Disease Control and Prevention ("CDC") with a lead time of one to two weeks (Brammer et al., 2009). However, GFT has since failed significantly: it was unable to predict the 2009 A-H1N1 influenza epidemic (Walsh, 2014) and it overestimated cases beginning in the 2011-2012 flu season (Lazer et al., 2014), particularly in the 2012-2013 flu season (Butler, 2013).

Problem Formulation

We summarized updates to the GFT after each identified failure in Appendix A. These updates were not able to fully correct the model. We examine the ways in which GFT's algorithm failed in 2009 and in 2012-2013 seasons, as well as the implications of the respective responses to each failure. We focus on identifying underlying biases that precluded the model from working effectively and discuss other potential vulnerabilities of the GFT.

Model Development, Estimation, and Results

The initial model was based on a five-year aggregate of Google web searches which resulted in a set of 45 best scoring queries used as a single variable predictor for ILI cases (Ginsberg et al., 2009). This algorithm matched flu propensity by analyzing changes in relative search behavior, user utilization, to produce the data. While Google tested search terms for correlation, the methodology did not attempt to find causation between variables and ILI cases, showing measurement bias in GFT's development. GFT's inability to capture the 2009 outbreak and the subsequent GFT update provide further evidence of measurement bias in its original design. The failure revealed the search term variable's insufficiency as a measurement of flu-related search activity as it did not include terms necessary to forecast an outbreak outside the norm, showing bias in Google's judgement of what would be appropriate predictors for flu. An update of additional terms published in 2009, as seen in Appendix A, indicated recognition of this fault.

The 2009 episode showed omitted variable bias as GFT failed to account for a variable that would influence the volume of searches made by users and anticipation of cases. The new A-H1N1 strain caused changes in user behavior and relevant flu-associated language (Althouse et al., 2014), which the static predictive model was not prepared for. The failure to account for a new strain which displayed different seasonal behavior also revealed further weakness: the search terms' correlation may have been statistically significant in development, but not in practice. The 2009 outbreak began in summer (Cook et al., 2011), unusual compared to previous flu seasons normally occurring in colder months. Lazer et al. (2014) referred to the initial GFT as "part flu detector, part winter detector" suggesting that GFT's search terms, while correlated with flu, were simply related to the winter season which coincides with flu season.

GFT severely overestimated the number of flu cases in 2012-2013, around double the number reported by the CDC (Diakopoulos, 2013). Google's GFT team published an update discussing what they identified as the weakness which led to the mistake: the algorithm was particularly vulnerable to bias in short periods where search queries were unusually high (Copeland et al., 2013). GFT's failure here shows omitted variable bias and failure to recognize users' social bias. GFT was unable to account for effects of heightened media coverage or for social changes from signals such as New York's declaration of a public health emergency (Butler, 2013), which may have caused visible panic-like reactions among users affecting their judgement.

We also note that GFT was overestimating since 2011-2012, even prior to the 2012-2013 spike. A more constant explanation of these issues stemmed from Google's search algorithm and its effects on user utilization, particularly "blue team dynamics" which adjusts the search algorithm according to a business model (Lazer et al., 2014). Google's algorithm-generated additional and suggested search terms during a user's search may have interfered and affected the frequency of different queries made by users, resulting in user interaction bias.

We identified further two potential biases which may preclude GFT from working effectively, particularly long-term: emergent bias and temporal bias (Mehrabi et al., 2021). Had GFT continued, it likely would have faced emergent bias from the changing values and knowledge of the users, and temporal bias from their corresponding change in behavioral patterns online. Over a decade has passed since GFT's original launch and there has since been a boom in users and their knowledge of the internet, as well as in online medical services. Users may now be directly approaching online medical professionals for diagnostics instead of simply self-searching symptoms on Google, as they may have developed higher value for websites specifically for medical evaluation. These users may have developed a learned habit of going to these websites, skipping the search engine over time. Varian (2016) raised a relevant point for this scenario that a simple predictor of past behavior should not be the goal in forecasting. GFT would have to adapt and overcome the biases to remain an effective predictive model of ILI.

Recommendations and Managerial Implications

Google's responses to GFT's failures have been reactionary, though examination of biases present show the importance of constant adaptation over reactive response. We recommend regular re-evaluation and adjustment of the model using real time data to achieve better results and maintain its effectiveness at forecasting throughout change.

Another possible safeguard is to ensure that appropriate statistical methods are used, and that the team considers measurement, validity, and reliability of the data in context before performing any statistical analysis or implementing a predictive model based on the data. The model's performance illustrated problems with the GFT's algorithm which used correlations. GFT targeted simplicity with correlation but for its purposes, GFT's execution has shown the shortcomings of the data and methods utilized for the model.

As the GFT used the CDC data simultaneously, recalibration of the search data with lagged CDC data would help adapt the model to improve accuracy and performance. The data design currently uses a methodology to match propensity instead of prediction. Considering prediction in the data generating process would further improve the model to provide accuracy in flu monitoring.

The algorithm mechanics also need to consider the data, algorithm, and user interaction model, and take into account the interaction between the three as unanticipated change in user behavior led to more than double the ILI cases recorded. User behavior can be addressed by explanation through use of relative research to feed into the data that produces the algorithm.

External resources could also be utilized as part of preventative measures. Google could offer the data to individuals or groups that are able to perform their own replication of the data. This would be especially useful as GFT was planned to be rolled out to other countries and for use with other diseases. Allowing others to perform replications would help identify potential weaknesses not initially identified and would help in development and customization specific to their target areas in a rollout.

Conclusion

GFT was an attempt to create a simple predictive model that generated ILI case estimates based on the search activity of its users. However, GFT was found to be ineffective at certain key points (2009, 2012-2013) and over the long term (overestimation beginning in 2011-2012). Our examination of Google's development, implementation, ineffectiveness, and failure responses identified multiple biases that led to its failure: measurement bias and omitted variable bias in the model's initial development and failure to recognize potential user social bias and user interaction bias over the course of GFT implementation. We also identified two further potential biases which may have led to failure, emergent and temporal bias, had GFT been continued. To negate these biases, we recommend the following: more appropriate statistical methods beyond simple correlation for the model's development; acknowledgement and incorporation of the effects of interaction between data, algorithm, and user into the model; recalibration of the model using prior information; utilization of external resources; and constant re-evaluation and adjustment using real-time data throughout implementation for continued effectivity.

References

Althouse, B.M., Ayers, J.W., Santillana, M., and Zhang, D.W. (2014). What Can Digital Disease Detection Learn from (an External Revision to) Google Flu Trends? *American Journal of Preventive Medicine*, *47*(3), (341-347). **https://doi.org/10.1016/j.amepre.2014.05.020**

Ginsberg, J., Mohebbi, M., Patel, R. et al. (2009). Detecting influenza epidemics using search engine query data. *Nature*, 457, (1012–1014). **https://doi.org/10.1038/nature07634**

Cook, S., Conrad, C., Fowlkes, A.L., and Mohebbi, M.H. (2011, August 19). Assessing Google Flu Trends Performance in the United States during the 2009 Influenza Virus A (H1N1) Pandemic. *PLOS ONE* 6(8): e23610. **https://doi.org/10.1371/journal.pone.0023610**

Copeland, P., Romano, R., Hecht, G., Stefansen, C., Zhang, T., and Zigmond, D. (2013). Google disease trends: An update. *International Society of Neglected Tropical Diseases 2013, International Society of Neglected Tropical Diseases*. **https://research.google/pubs/pub41763.pdf**

Diakopoulos, N. (2013, July 5). *How Google Flu Trends Is Getting to the Bottom of Messy Data*. Harvard Business Review. https://hbr.org/2013/07/how-google-flu-trends-is-getting-to-the-bottom

Galstyan, A., Lerman, K., Mehrabi, N., Morstatter, F., and Saxena, N., (2021, July). A Survey on Bias and Fairness in Machine Learning. *ACM Computing Surveys, 54*(6), Article 115 (1-35). **https://doi.org/10.1145/3457607**

Goldberg, C. (2013, January 13). *Is 'Google Flu Trends' Prescient Or Wrong?* WBUR. https://www.wbur.org/news/2013/01/13/google-flu-trends-cdc

Lazer, D., Kennedy, R., King, G., and Vespignani, A. (2014, March 14). The parable of Google flu: traps in big data analysis. *Science Magazine,* 343, 6176, 1203-1205.

Varian, H.R. (2016, July). Causal inference in economics and marketing. *Proceedings of the National Academy of Sciences, 113*(27), (7310-7315). **DOI: 10.1073/pnas.1510479113**

Walsh, B. (2014, March 3). *Google's Flu Project Shows the Failings of Big Data*. Time. https://time.com/23782/google-flu-trends-big-data-problems/

Appendix

APPENDIX A - Failure Responses

Time Period	Failure	Response
2009	GFT was unable to predict the 2009 A-H1N1 flu outbreak.	An updated GFT model was made in which updated search terms utilized from 45 to 160 search terms, which were overall more directly related to the influenza infection instead of simply complications or terms associated with influenza infection or season (Cook et al., 2011).
2011-2012	GFT overestimated flu cases for 100 out of 108 weeks beginning in August 2011 (Lazer et al., 2014).	No immediate response was identified.
2012-2013	GFT severely overestimated the number of flu cases in the 2012-2013 flu season, around double the number reported by the CDC.	Google attempted to address effects of prolonged media exposure in two ways: (1) by “dampening” anomalous media spikes and (2) using ElasticNet to adjust the regression after the 2012-2013 overestimation (Copeland et. al., 2013).