2. Statistical Modelling (1)

Statistical Modelling & Machine Learning

Jaejik Kim

Department of Statistics
Sungkyunkwan University

STA3036

Statistical Modelling

- For accurate prediction, define your research question explicitly.
- Fancy models such as random forests, SVM, and neural networks, etc. do not guarantee better prediction.
- Regardless of algorithmic and data models, use a right model or method for your research question and data.
- An example of statistical modelling: ARGO (Auto Regression with GOole search data) proposed by Yang et al. (2015).

ARGO

Accurate estimation of influenza epidemics using Google search data via ARGO

Shihao Yang[a], Mauricio Santillana[b,c,1], and S. C. Kou[a,1]

[a]Department of Statistics, Harvard University, Cambridge, MA 02138; [b]School of Engineering and Applied Sciences, Harvard University, Cambridge, MA 02138; and [c]Computational Health Informatics Program, Boston Children's Hospital, Boston, MA 02115

Edited by Wing Hung Wong, Stanford University, Stanford, CA, and approved September 30, 2015 (received for review August 6, 2015)

Accurate real-time tracking of influenza outbreaks helps public health officials make timely and meaningful decisions that could save lives. We propose an influenza tracking model, ARGO (AutoRegression with GOogle search data), that uses publicly available online search data. In addition to having a rigorous statistical foundation, ARGO outperforms all previously available Google-search–based tracking models, including the latest version of Google Flu Trends, even though it uses only low-quality search data as input from publicly available Google Trends and Google Correlate websites. ARGO not only incorporates the seasonality in influenza epidemics but also captures changes in people's online search behavior over time. ARGO is also flexible, self-correcting, robust, and scalable, making it a potentially powerful tool that can be used for real-time tracking of other social events at multiple temporal and spatial resolutions.

digital disease detection | seasonal influenza | big data | influenza-like illnesses activity real-time estimation | autoregressive exogenous model

CDC's ILI reports have a delay of 1–3wk due to the time for processing and aggregating clinical information. This time lag is far from optimal for decision-making purposes. To alleviate this information gap, multiple methods combining climate, demographic, and epidemiological data with mathematical models have been proposed for real-time estimation of flu activity (18, 21–25). In recent years, methods that harness Internet-based information have also been proposed, such as Google (1), Yahoo (2), and Baidu (3) Internet searches, Twitter posts (4), Wikipedia article views (5), clinicians' queries (6), and crowdsourced self-reporting mobile apps such as Influenzanet (Europe) (26), Flutracking (Australia) (27), and Flu Near You (United States) (28). Among them, GFT has received the most attention and has inspired subsequent digital disease detection systems (3, 8, 29–32). Interestingly, Google has never made their raw data public, thus making it impossible to reproduce the exact results of GFT

APPLIED MATHEMATICS

Motivation

- Google search data (Big Data) $\Rightarrow$ Detecting epidemic outbreaks.
- Background: Accurate real-time tracking of influenza outbreaks helps public health officials make timely and meaningful decisions that could save lives.
- Google Flu Trends (GFT): A digital disease detection system that uses the volume of selected Google search terms to estimate current influenza-like illnesses (ILI) activity (The service was terminated in 2015).
- Problem: Significant discrepancy between GFT's estimates and measurements from CDC (Center for Disease Control).
- Goal: Accurate prediction of ILI activity level of this week in US using Google search data.

Data Collection

- Output: ILI activity level (provided by CDC in every week; 0$\sim$1 scale).
- Inputs: Google search queries (volumes of words related to flu for each week; 0$\sim$100 scale).
- Tools: Google trends and Google correlates.
  - Google trends (`https://trends.google.com`): Trends of volumes of google search queries by time (currently available).
  - Google correlates: A list of search queries that have a similar trend to your specific search word (currently not serviced).

Limitations of GFT

- GFT algorithm uses a static approach which does not fit for dynamic behavior of flu. So, it cannot use information in CDC's ILI weekly report.
- The input variables of GFT are generated by aggregating the multiple search words into a single variable (i.e., a function of volumes of fixed search words). It does not allow for changes in people's search term patterns over time.
- GFT ignores the intrinsic time series properties such as seasonality of ILI activity.

Properties of ARGO

- It can consider dynamically incorporating new information such as CDC report by an auto regressive model.
- It automatically selects the most useful Google search queries for estimation using Google correlates and lasso penalty.
- It considers the long term cyclic information (seasonality) using 2-year window for model training (training data: 104 week (2-year) data).
- Statistical model $\Rightarrow$ Statistical inference.

$\Rightarrow$ Even though inputs of ARGO are low-quality data from Google trends and Google correlates, it has significant improvement over the GFT.

Preprocessing of Data

- Time period of data: Apr. 4, 2009 $\sim$ May 16, 2015.
- Google search words as Inputs:
  - Search queries highly correlation with CDC's ILI activity level for given periods (Google correlates).
  - Two different time period (Pre and post H1N1 pandemic).
- Output variable (y_t): The logit-transformed CDC ILI activity level at time (week) t.
  - Logit-transformation: $[0, 1] \rightarrow \mathbb{R}$.
- Input variable ($\boldsymbol{X}_t$): The vector of log-transformed normalized volume of Google search queries at time t.
  - Log-transformation: Google search frequencies usually have an exponential growth rate near peaks and are artificially scaled to [0,100].
  - Log-transformation: $[0, 100] \rightarrow \mathbb{R}$

Search Queries for Pre H1N1 Pandemic Period

influenza.type.a	painful.cough	treatment.for.the.flu	weather.march
flu.incubation	fever.flu	basketball.standing	fevers
bronchitis	over.the.counter.flu	flu.test	duration.of.flu
influenza.contagious	pneumonia	tussionex	flu.contagious.period
flu.fever	how.long.is.the.flu	reduce.a.fever	cold.vs.flu
influenza.a	flu.how.long	how.long.is.the.flu.contagious	cure.the.flu
influenza.incubation	treatment.for.flu	treat.flu	walking.pneumonia
flu.contagious	fever.cough	spring.break.family	flu.vs..cold
treating.the.flu	flu.medicine	las.vegas.shows.march	length.of.flu
type.a.influenza	dangerous.fever	how.to.reduce.a.fever	influenza.a.and.b
symptoms.of.the.flu	high.fever	flu.or.cold	flu.and.pregnancy
influenza.symptoms	is.flu.contagious	incubation.period.for.the.flu	sinus.infections
flu.duration	normal.body	harlem.globe	influenza.treatment
flu.report	normal.body.temperature	tussin	jiminy.peak.ski
symptoms.of.flu	how.long.does.the.flu.last.	basketball.standings	baseball.preseason
influenza.incubation.period	symptoms.of.pneumonia	sinus	spring.break.date
how.to.treat.the.flu	signs.of.the.flu	upper.respiratory	indoor.driving
treat.the.flu	flu.vs.cold	get.over.the.flu	z.pack
symptoms.of.bronchitis	low.body	acute.bronchitis	college.spring.break.dates
flu.treatment	cough.fever	body.temperature	aloha.ski
symptoms.of.influenza	vegas.shows.march	college.basketball.standings	concerts.in.march
treating.flu	is.the.flu.contagious	strep	break.a.fever
flu.in.children	type.a.flu	march.weather	influenza.duration
fever.reducer	flu.treatments	getting.over.the.flu	robitussin
cold.or.flu	remedies.for.the.flu	march.vacation	virginia.wrestling

Search Queries for Post H1N1 Pandemic Period

influenza.type.a	get.over.the.flu	type.a.influenza	flu.care
symptoms.of.flu	treating.flu	i.have.the.flu	how.long.contagious
flu.duration	flu.vs..cold	taking.temperature	fight.the.flu
flu.contagious	having.the.flu	flu.versus.cold	reduce.a.fever
flu.fever	treatment.for.flu	bronchitis	cure.the.flu
treat.the.flu	human.temperature	how.long.flu	medicine.for.flu
how.to.treat.the.flu	dangerous.fever	flu.germs	flu.length
signs.of.the.flu	the.flu	cold.vs..flu	cure.flu
over.the.counter.flu	remedies.for.flu	flu.and.cold	exposed.to.flu
how.long.is.the.flu	influenza.a.and.b	thermoscan	low.body
symptoms.of.the.flu	contagious.flu	flu.complications	early.flu.symptoms
flu.recovery	how.long.does.the.flu.last	high.fever	remedies.for.the.flu
cold.or.flu	fever.flu	flu.children	flu.report
flu.medicine	oscillococcinum	the.flu.virus	incubation.period.for.flu
flu.or.cold	flu.remedies	how.to.treat.flu	break.a.fever
normal.body	how.long.is.flu.contagious	pneumonia	flu.contagious.period
is.flu.contagious	flu.treatments	flu.headache	influenza.incubation.period
treat.flu	influenza.symptoms	flu.cough	cold.versus.flu
body.temperature	cold.vs.flu	ear.thermometer	flu.in.children
is.the.flu.contagious	braun.thermoscan	how.to.get.rid.of.the.flu	what.to.do.if.you.have.the.flu
reduce.fever	fever.cough	flu.how.long	medicine.for.the.flu
flu.treatment	signs.of.flu	symptoms.of.bronchitis	flu.and.fever
flu.vs.cold	how.long.does.flu.last	cold.and.flu	flu.lasts
how.long.is.the.flu.contagious	normal.body.temperature	over.the.counter.flu.medicine	incubation.period.for.the.flu
fever.reducer	get.rid.of.the.flu	treating.the.flu	do.i.have.the.flu

ARGO Model

- ARGO model:

$$y_t = \mu_y + \sum_{j=1}^{52} \alpha_j y_{t-j} + \sum_{i=1}^{100} \beta_i X_{it} + \epsilon_t, \qquad (1)$$

- Model assumptions:
  - $\epsilon_t \sim N(0, \sigma^2)$.
  - $\boldsymbol{X}_t$ depends only on the ILI activity at the same time point t (i.e., $\boldsymbol{X}_t | y_t$ is independent of y_l and $\boldsymbol{X}_l$, $t \neq l$).
  - y_t has auto regressive terms with 52 weeks (1-year) for seasonality.
  - $\boldsymbol{X}_t | y_t \sim MVN(\boldsymbol{\mu}_x + y_t \boldsymbol{\theta},\ \boldsymbol{Q})$.

Estimation of ARGO

- Estimation: Penalized method

$$\underset{\mu_y, \boldsymbol{\alpha}, \boldsymbol{\beta}}{\arg\min} \sum_t \left(y_t - \mu_y - \sum_{j=1}^{52} \alpha_j y_{t-j} - \sum_{i=1}^{100} \beta_i X_{i,t} \right)^2$$
$$+ \lambda_\alpha \|\boldsymbol{\alpha}\|_1 + \eta_\alpha \|\boldsymbol{\alpha}\|_2^2 + \lambda_\beta \|\boldsymbol{\beta}\|_1 + \eta_\beta \|\boldsymbol{\beta}\|_2^2.$$

  - Penalty term: Lasso, ridge, or Elastic net is possible.
  - This paper set $\eta_\alpha = \eta_\beta = 0$ (i.e., lasso penalty was used).

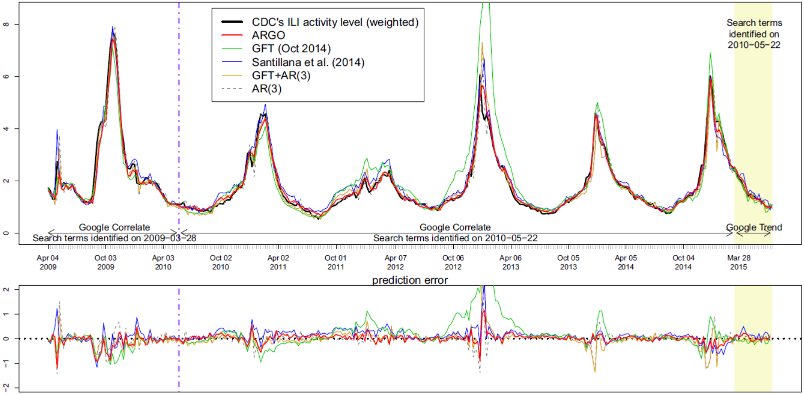
Result (1)

Table 1. Comparison of different models for the estimation of influenza epidemics

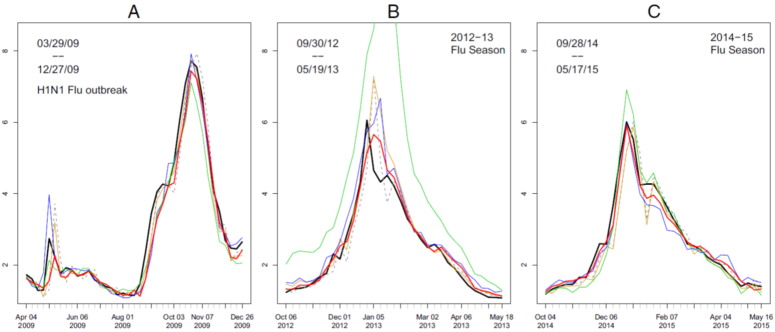
	Whole period (Mar 29, 2009 to Jul 11, 2015)	Off-season flu H1N1	Regular flu seasons (week 40 to week 20 next year) 2010–2011	2011–2012	2012–2013	2013–2014	2014–15
RMSE							
ARGO	**0.608**	0.640	**0.596**	**0.807**	**0.687**	**0.306**	**0.438**
GFT (Oct 2014)	2.216	0.773	1.110	3.023	4.451	0.986	0.700
Ref. 16	0.915	0.833	0.881	2.027	1.090	0.446	0.663
GFT+AR(3)	0.912	**0.580**	0.602	1.382	1.279	0.993	0.906
AR(3)	0.957	0.813	0.794	1.051	1.191	0.969	0.928
Naive	1 (0.348)	1 (0.600)	1 (0.339)	1 (0.163)	1 (0.499)	1 (0.350)	1 (0.465)
MAE							
ARGO	**0.649**	0.584	**0.574**	**0.748**	**0.650**	**0.391**	**0.530**
GFT (Oct 2014)	1.834	0.777	1.260	3.277	5.028	0.891	0.770
Ref. 16	1.052	0.719	1.010	2.211	1.029	0.610	0.820
GFT+AR(3)	0.888	**0.570**	0.613	1.308	1.016	1.034	0.839
AR(3)	0.925	0.777	0.787	0.951	0.988	0.917	0.934
Naive	1 (0.201)	1 (0.425)	1 (0.259)	1 (0.135)	1 (0.325)	1 (0.212)	1 (0.295)
MAPE							
ARGO	**0.787**	**0.620**	**0.663**	**0.770**	**0.719**	**0.453**	**0.620**
GFT (Oct 2014)	1.937	0.721	1.394	3.442	5.419	0.892	0.895
Ref. 16	1.381	0.765	1.380	2.306	1.251	0.754	0.958
GFT+AR(3)	1.037	0.683	0.698	1.407	0.986	1.062	0.828
AR(3)	1.003	0.894	0.814	0.947	0.939	0.891	0.916
Naive	1 (0.090)	1 (0.139)	1 (0.105)	1 (0.081)	1 (0.110)	1 (0.084)	1 (0.097)
Correlation							
ARGO	**0.986**	0.985	**0.989**	**0.928**	**0.968**	**0.993**	**0.993**
GFT (Oct 2014)	0.875	**0.989**	0.968	0.833	0.926	0.969	0.986
Ref. 16	0.971	0.967	0.983	0.927	0.956	0.985	0.984
GFT+AR(3)	0.967	0.986	0.985	0.879	0.929	0.945	0.957
AR(3)	0.964	0.968	0.971	0.877	0.903	0.927	0.945
Naive	0.961	0.951	0.954	0.887	0.924	0.923	0.937
Correlation of increment							
ARGO	**0.758**	0.806	**0.810**	0.286	**0.527**	**0.938**	0.912
GFT (Oct 2014)	0.706	**0.863**	0.702	0.484	0.502	0.847	**0.918**
Ref. 16	0.690	0.776	0.693	**0.510**	0.367	0.915	0.889
GFT+AR(3)	0.512	0.708	0.708	0.165	0.141	0.534	0.587
AR(3)	0.385	0.585	0.569	0.077	0.011	0.404	0.493
Naive	0.436	0.602	0.570	0.095	0.134	0.406	0.514

GFT+AR(3) stands for the model $p_t = \mu + \alpha_1 p_{t-1} + \alpha_2 p_{t-2} + \alpha_3 p_{t-3} + \beta \mathrm{GFT}(t)$, where the GFT estimate is treated as an exogenous variable. Boldface highlights the best performance for each metric in each study period. RMSE, MAE, and MAPE are relative to the error of naive method; that is, the number reported is the ratio of error of a given method to that of the naive method. The absolute error of the naive method is reported in parentheses. All comparisons are based on the original scale of ILI activity level.

Result (2)


CDC's ILI activity level (weighted)
ARGO
GFT (Oct 2014)
Santillana et al. (2014)
GFT+AR(3)
AR(3)
Search terms identified on 2010−05−22
Google Correlate
Search terms identified on 2009−03−28
Google Correlate
Search terms identified on 2010−05−22
Google Trend
Apr 04 2009
Oct 03 2009
Apr 03 2010
Oct 02 2010
Apr 02 2011
Oct 01 2011
Apr 07 2012
Oct 06 2012
Apr 06 2013
Oct 05 2013
Apr 05 2014
Oct 04 2014
Mar 28 2015
prediction error

Result (3)


A
03/29/09
--
12/27/09
H1N1 Flu outbreak
Apr 04 2009
Jun 06 2009
Aug 01 2009
Oct 03 2009
Nov 07 2009
Dec 26 2009
B
09/30/12
--
05/19/13
2012−13 Flu Season
Oct 06 2012
Dec 01 2012
Jan 05 2013
Mar 02 2013
Apr 06 2013
May 18 2013
C
09/28/14
--
05/17/15
2014−15 Flu Season
Oct 04 2014
Dec 06 2014
Feb 07 2015
Apr 04 2015
May 16 2015
2
4
6
8