Journal of the Korean Data & Information Science Society
2016, **27**(4), 1075–1081

http://dx.doi.org/10.7465/jkdi.2016.27.4.1075
한국데이터정보과학회지

Comparison study of SARIMA and ARGO models for influenza epidemics prediction[†]

Jihoon Jung[1] · Sangyeol Lee[2]

[12]Department of Statistics, Seoul National University

Received 24 June 2016, revised 11 July 2016, accepted 13 July 2016

Abstract

The big data analysis has received much attention from the researchers working in various fields because the big data has a great potential in detecting or predicting future events such as epidemic outbreaks and changes in stock prices. Reflecting the current popularity of big data analysis, many authors have proposed methods tracking influenza epidemics based on internet-based information. The recently proposed 'autoregressive model using Google (ARGO) model' (Yang *et al.*, 2015) is one of those influenza tracking models that harness search queries from Google as well as the reports from the Centers for Disease Control (CDC), and appears to outperform the existing method such as 'Google Flu Trends (GFT)'. Although the ARGO predicts well the outbreaks of influenza, this study demonstrates that a classical seasonal autoregressive integrated moving average (SARIMA) model can outperform the ARGO. The SARIMA model incorporates more accurate seasonality of the past influenza activities and takes less input variables into account. Our findings show that the SARIMA model is a functional tool for monitoring influenza epidemics.

Keywords: ARGO model, big data, disease detection, Google flu trends, influenza epidemics, influenza-like illnesses activity estimation, SARIMA model.

1. Introduction

As the usage of internet grew rapidly coupled with the advent of smart phones, the amount of information about their online activities has significantly increased during the past decades. This huge amount of information collected from the web is called "big data". The issue of big data has received notable attention from practitioners in a variety of research and industrial fields (Preis *et al.*, 2013)) because the big data is considered to have a considerable potential in predicting or estimating future events (Labrinidis and Jagadish, 2012)). In recent years, a number of studies have proposed models that track epidemic outbreaks, such as influenza, Ebola (Wesolowski *et al.*, 2014), and dengue (Chan *et al.*, 2011),

[†] This research is supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Science, ICT and future Planning (No. 2015R1A2A2A010003894).

[1] Master student, Department of Statistics, Seoul National University, Seoul 08826, Korea.
[2] Corresponding author: Professor, Department of Statistics, Seoul National University, Seoul 08826, Korea. E-mail: sylee@stats.snu.ac.kr

by taking advantage of publicly available online information. Among such diseases, influenza has gotten special attention with regards to disease detection. Although majority of people consider influenza a relatively common disease that rarely causes fatal effects on them, influenza outbreaks actually bring about up to 500,000 deaths worldwide every year. From 1976 to 2007, the annual influenza-related deaths ranged from 3000 to 50,000 in the United States only, approximately with an average of about 24,000 annual deaths. Therefore, various methods based on information from the internet, such as Google, Yahoo, and Twitter, have been proposed recently to estimate influenza-like illnesses (ILI) (ILI is defined as a symptom of fever, greater than temperature of 37.8°C, and a cough or a sore throat that seems to be most likely caused by influenza) activities (ILI activity level refers to the percentage of patients diagnosed with ILI compared with the total number of patients who visited the hospital for any reason within a week) (Ginsberg *et al.*, 2009; Polgreen *et al.*, 2008; Santillana *et al.*, 2015; Bollen *et al.*, 2011). As a relevant reference, see Hwang and Oh (2016).

One of those methods based on a big data approach is the AutoRegression with GOogle (ARGO), proposed by Yang *et al.* (2015). The ARGO simultaneously uses the Google search data and ILI reports from the Centers for Disease Control (CDC) to track influenza outbreaks. Although the ARGO appears to improve existing methods such as Google Flu Trends (Cook *et al.*, 2011; Santillana *et al.*, 2014), it has two major shortcomings. The first is that it does not fully reflect the time series properties such as the seasonality in influenza epidemics, failing to accurately estimate the ILI activity. The second is to take too many input variables into account, so that the algorithm for implementation becomes quite complicated and also needs a long time till drawing a conclusion. To overcome these defects, we instead propose to employ a classical time series model, i.e. the seasonal autoregressive integrated moving average (SARIMA) model which only uses the CDC's ILI reports. In performing this procedure, we first remove a seasonal effect and stochastic trend by differencing the time series, and then apply model selection criteria such as Akaike's information criterion (AIC) to find an optimal SARIMA model. We then compare its performance with that of the ARGO using the accuracy metrics in Yang *et al.* (2015).

Our findings show that the SARIMA model improves the accuracy of estimation by incorporating the long-term cyclic information or seasonality of the past ILI activities. Furthermore, since our method only uses the data provided by the CDC, it has an advantage over the ARGO in that it only takes one variable into consideration, which leads to saving the computing time significantly. Our method appears to remarkably outperform the ARGO model even though it does not involve any publicly available online search data. It may be because all important information is already locked in the time series themselves, so that the SARIMA model without exogenous information can successfully extract out the information. This coincides with the spirit of 'autoregression' scheme (Lee *et al.*, 2013), broadly appreciated among time series analysts: see also Lee and Kim (2013). Lazer *et al.* (2014) mentions that Google constantly modifies its search algorithm and returns different recommended additional search terms over time to support its business model. Thus, an influenza-tracking model based on the Google searches would be affected by the change in the algorithm because the modification would eventually affect people's search behavior. In contrast, our SARIMA model based method is free from the algorithm underlying Google's method.

The organization of this study is as follows. Section 2 introduces the ARGO and SARIMA

models and presents the procedure to choose an optimal model. Section 3 compares the SARIMA model with the ARGO based on several metrics. Section 4 provides concluding remarks.

2. Data description and seasonal model

Yang *et al.* (2015) introduces the ARGO model:

$$y_t = \mu_y + \sum_{j=1}^{52} \alpha_j y_{t-j} + \sum_{i=1}^{100} \beta_i X_{i,t} + \epsilon_t, \qquad \epsilon_t \overset{iid}{\sim} N(0.\sigma^2), \tag{2.1}$$

as-where weakly-based observation $y_t = \text{logit}(p_t) = \log \frac{p_t}{1-p_t}$ refers to the logit-transformed ILI activity level p_t and $X_{i,t}$ denotes the Google search frequency of term i among 100 Google search terms at time t. Although the ARGO is designed to incorporate the seasonality in the CDC-reported ILI activity level data, a closer inspection of the model shows that the ARGO does not fully reflect seasonal effect. Motivated by this, we consider employing a classical seasonal time series model, such as the seasonal ARIMA (SARIMA) model. Furthermore, we discard the role of the $X_{i,t}$ in building up a SARIMA model.

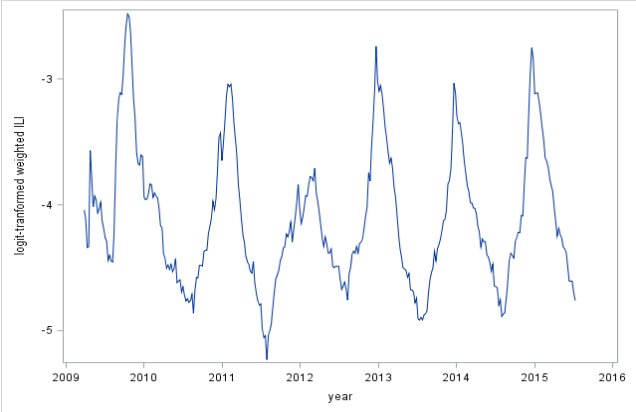
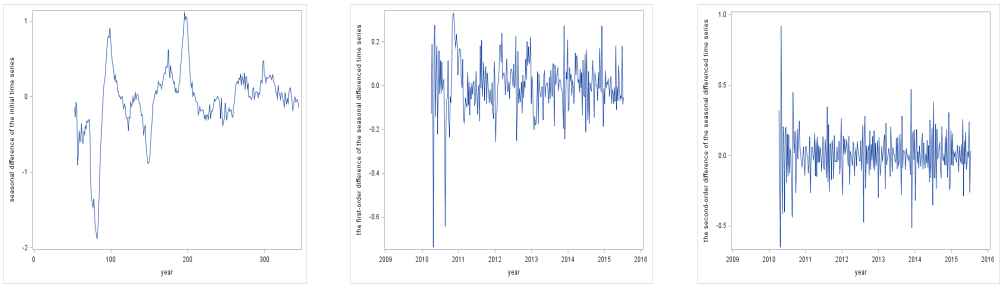


Figure 2.1 Time series plot of the logit-transformed CDC's weighted ILI activity level

In this study, as in Yang *et al.* (2015), we use the weighted version of the CDC's ILI activity level (available at http::/gis.cdc.gov/grasp/fluview/fluportaldashboard.html; date of access: November 12, 2015). Also, we use the logit-transformed the CDC's weighted ILI activity level, that is, $x_t = \text{logit}(w_t)$, where w_t denotes the weighted ILI activity level at time t. Figure 2.1 presents the plot of x_t from March 29, 2009 to July 11, 2015, which clearly shows the existence of strong seasonal effect. To remove the seasonal effect, we take seasonal difference, say, $x_t - x_{t-52}$. Figure 2.2 shows the plots of seasonal differenced time series, and also, the first and second-order differenced time series after the seasonal differencing. The Dickey-Fuller test actually shows that the seasonal differenced time series has a unit root and its first-order differenced time series has no unit root. However, in this case, we use the second-order differenced time series because it better captures the characteristics of stationary time series. The autocorrelation function (ACF) and partial autocorrelation function (PACF) of the first and second differenced time series are plotted in Figures 2.3 and

2.4. Figure 2.3 shows that the ACF of the first-order differenced time series has significant peaks at too many lags compared with that of the second-order differenced time series. This suggests that the second-order differenced time series is more tractable for analyzing the ILI activity data set.

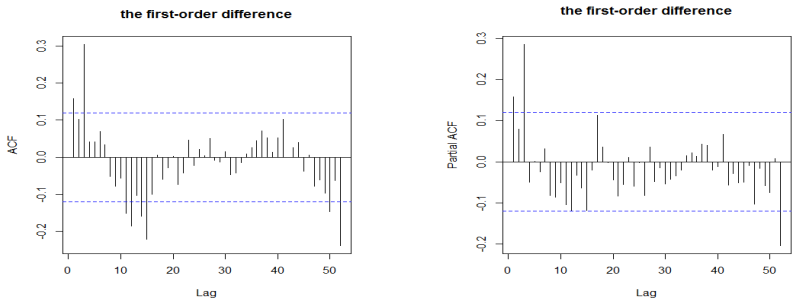


(a) Seasonal differenced time series

(b) The first-order difference of the seasonal differenced time series

(c) The second-order difference of the seasonal differenced time series

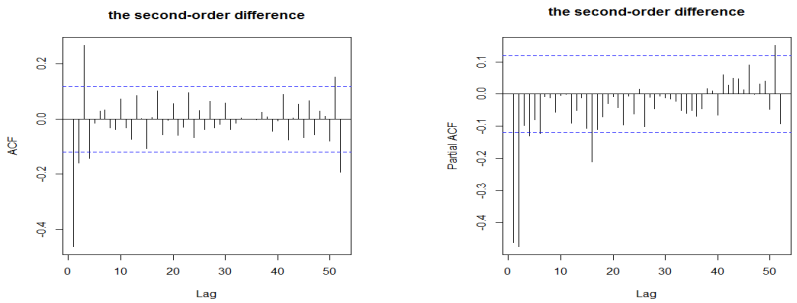
Figure 2.2 Time series plots



(a) ACF of the first-order difference of the seasonal differenced time series

(b) PACF of the first-order difference of the seasonal differenced time series

Figure 2.3 ACF and PACF of the first-order difference of the seasonal differenced time series



(a) ACF of the second-order difference of the seasonal differenced time series

(b) PACF of the second-order difference of the seasonal differenced time series

Figure 2.4 ACF and PACF of the second-order difference of the seasonal differenced time series

The SARIMA model, i.e. ARIMA$(p,d,q)\times(P,D,Q)_S$ under consideration is as follows:

$$\phi(B)\Phi(B^s)(1-B^s)^D(1-B)^d x_t=\theta(B)\Theta(B^s)\epsilon_t, \tag{2.2}$$

where $s=52$, $D=1$ and $d=2$ are preassigned as discussed earlier, B denotes the back-shift operator, and ϕ,Φ,θ,Θ are characteristic polynomials with orders p,P,q,Q, respectively, that should be determined from the data. Comparing the AIC values of candidate ARMA models, we conclude that ARIMA$(0,2,4)\times(0,1,1)_{52}$ is an optimal model. More precisely, the estimated SARIMA model is given by

$$(1-B)^2(1-B^{52})x_t=(1+0.9076B-0.1770B^3+0.2693B^4)(1+0.4148B^{52})\epsilon_t. \tag{2.3}$$

Since the above model does not contain autoregressive (AR) part, hereafter, it is named SIMA(52) for abbreviation.

3. Results and discussion

In this section, we conduct a comparison study of the performance of the SIMA(52) and ARGO in the same settings as that of Yang *et al.* (2015). To this end, we examine the root mean squared error (RMSE), mean absolute error (MAE), mean absolute percentage error (MAPE), correlation with the observed values or the CDC-reported weighted ILI activity level, and correlation of increment (CI) with the observed values as follows:

$$\begin{aligned}
&\mathrm{RMSE}(\hat{w}_t,w_t)=\Big[\frac{1}{n}\sum_{t=1}^{n}(\hat{w}_t-w_t)^2\Big]^{\frac{1}{2}};\\
&\mathrm{MAE}(\hat{w}_t,w_t)=\frac{1}{n}\sum_{t=1}^{n}|\hat{w}_t-w_t|;\\
&\mathrm{MAPE}(\hat{w}_t,w_t)=\frac{1}{n}\sum_{t=1}^{n}\frac{|\hat{w}_t-w_t|}{w_t};\\
&\mathrm{CI}(\hat{w}_t,w_t)=\mathrm{Corr}(\hat{w}_t-\hat{w}_{t-1},w_t-w_{t-1}).
\end{aligned}$$

The correlation of fitted value $\hat{w}$ to the true value of ILI activity level w refers to their sample correlation coefficient.

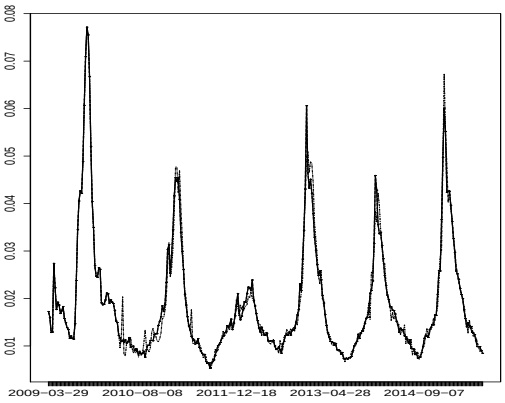


Figure 3.1 The plots of the true ILI activity levels (solid line) and the estimated ones (dotted line)

Table 3.1 Comparison of SIMA(52) and ARGO for the estimation of influenza epidemics

Metrics	Model	Whole period	Regular flu seasons (week 40 to week 20 next year)				
		2009-2015	2010-2011	2011-2012	2012-2013	2013-2014	2014-2015
RMSE	SIMA(52)	**0.002**	**0.003**	**0.001**	**0.003**	**0.003**	**0.002**
	ARGO	0.611	0.599	0.801	0,687	0.306	0.418
MAE	SIMA(52)	**0.001**	**0.002**	**0.001**	**0.002**	**0.002**	**0.001**
	ARGO	0.649	0.560	0.738	0.655	0.392	0.473
MAPE	SIMA(52)	**0.065**	**0.101**	**0.062**	**0.071**	**0.072**	**0.053**
	ARGO	0.789	0.648	0.763	0.727	0.458	0.532
Correlation	SIMA(52)	**0.987**	0.974	0.918	**0.973**	0.951	0.986
	ARGO	0.986	**0.988**	**0.931**	0.968	**0.993**	**0.994**
Corr. of increment	SIMA(52)	0.737	0.728	0.263	**0.680**	0.392	0.763
	ARGO	**0.748**	**0.796**	**0.280**	0.526	**0.945**	**0.911**

Table 3.1 shows the summary of the metrics for the SIMA(52) and ARGO for different time periods. It presents the values of the performance metrics for the whole period (2009-2015) and the regular flu seasons (winter time) in each year from 2010 to 2015. The table shows that for the whole time period, the SIMA(52) outperforms the ARGO in all metrics except for the correlation of increment. More precisely in terms of the RMSE, MAE, and MAPE, the SIMA(52) appears to produce the values considerably lower than those of the ARGO. With regard to the correlation, the SIMA(52) ($r = 98.7\%$) slightly outperforms the ARGO ($r = 98.6\%$) for the whole time period, whereas during the regular flu seasons, the former has a higher correlation only for 2012-2013 season and a slightly lower correlation in the remaining flu seasons. In short, the SIMA(52) has a performance similar to the ARGO in terms of correlation. A similar conclusion can be made for the correlation of increments: Table 3.1 shows that the SIMA(52) has a performance similar to the ARGO except during the 2013-2014 flu season. Figure 3.1 shows the fitted values of the SIMA(52) (dotted line) against the observed ILI activity level reported by the CDC.

Overall, our findings show that the SARIMA model outperforms the ARGO in accuracy. What is more, considerably less time is taken to fit the data when using the former (about 27 seconds) than when utilizing the latter (about 8 minutes). Because the ARGO outperforms influenza tracking models based on Google searches, it is believed that the SIMA(52) is more suitable in tracking influenza than flu tracking methodologies using publicly available Google search data.

4. Concluding remarks

In this study, we employed a SARIMA model, named SIMA(52), to predict the influenza epidemics and compared its performance with the ARGO method based on Google searches, and demonstrated that our method outperforms the ARGO. Reflecting one year's seasonal effect of the historical ILI activity plays a key role in our analysis: the twice differencing also helps stabilize the time-varying variances. Another aspect is that our approach does not harness the Google searches unlike the ARGO but only uses officially-reported data provided by the national institution, which enables us to avoid the high sensitivity to public's overreaction to the disease on internet. Our approach also shares the same spirit as in Lazer *et al.* (2014) who pointed out the limitation of influenza-tracking models based on the Google searches. This also coincides with the spirit of 'autoregression' such that all the information is already locked in time series themselves other than exogenous variables.

Although the SIMA(52) demonstrates an outstanding performance somewhat superior to that of the ARGO, it still has the same defect as the ARGO has because the CDC reports ILI activity level one∼two weeks after the target date, and henceforth, make an information gap, which seems inevitable as far as only a weekly data is available for prediction. This study reminds the practitioners of the importance of classical methods in advance of adopting a trendy one. Both methods have their own merit, and therefore, a great care is necessary when implementing them for actual usage.

References

Bollen, J., Mao, H. and Zeng, X. (2011). Twitter mood predicts the stock market. *Journal of Computational Science*, **2**, 1-8.

Chan, E. H., Sahai, V., Conrad, C. and Brownstein, J. S. (2011). Using web search query data to monitor dengue epidemics: A new model for neglected tropical disease surveillance. *PLoS Neglected Tropical Diseases*, **5**, e1206.

Cook, S., Conrad C., Fowlkes, A. L. and Mohebbi, M. H. (2011). Assessing Google flu trends performance in the United States during the 2009 influenza vrius A (H1N1) pandemic. *PLoS One*, **6**, e23610.

Ginsberg, J. Mohebbi1, M. H., Patel1, R. S., Brammer, L., Smolinski1, M. S. and Brilliant, L. (2009). Detecting influenza epidemics using search engine query data. *Nature*, **457**, 1012-1014.

Hwang, S. and Oh, C. (2016). Estimation of the case fatality ratio of MERS epidemics using information on patients' severity condition. *Journal of the Korean Data & Information Science Society*, **27**, 599-607.

Labrinidis, A. and Jagadish, H. V. (2012). Challenges and opportunities with big data. *Proceedings of the VLDB Endowment*, **5**, 2032-2033.

Lazer, D., Kennedy, R., King, G. and Vespignani, A. (2014). The parable of Google Flu: Traps in big data analysis. *Science*, **343**, 1203-1205.

Lee, S. and Kim, B. (2013). Dependence structure analysis of KOSPI and NYSE based on time-varying copula models. *Journal of the Korea & Information Science Society*, **24**, 1477-1488.

Lee, S., Lee, J. and Noh, J. (2013). Maximum entropy test for infinite order autoregressive models. *Journal of the Korean Data & Information Science Society*, **24**, 637-642.

Overview of Influenza Surveillance in the United States. (2016). Retrived from `http://www.cdc.gov/flu/weekly/overview.htm`.

Polgreen, P. M., Chen, Y., Pennock, D. M. and Nelson, F. D. (2008). Using internet searches for influenza surveillance. *Clinical Infectious Diseases*, **47**, 1443-1448.

Preis, T., Moat, H. S. and Stanley H. E. (2013). Quantifying trading behavior in financial markets using Google trends. *Scientific Reports*, **3**, 1684.

Santillana, M., Nguyen, A. T., Dredze, M., Paul,M. J., Nsoesie, E. O. and Brownstein, J. S. (2015). Combining search, social media, and traditional data sources to improve influenza surveillance. *PLoS Computational Biology*, **11**, e1004513.

Santillana, M., Zhang, D. W., Althouse, B. M. and Ayers, J. W. (2014). What can digital disease detection learn from (an external revision to) Google Flu Trends? *American journal of preventive medicine*, **47**, 314-347.

Wesolowski, A., Buckee, C. O., Bengtsson, L., Wetter, E., Lu, X. and Tatem, A. J. (2014). Commentary: Containing the Ebola outbreak–the potential and challenge of mobile network data. *PLOS Currents Outbreaks*, `10.1371/currents.outbreaks.0177e7fcf52217b8b634376e2f3efc5e`.

Yang, S., Santillana, M. and Kou, S. C. (2015). *ARGO: A model for accurate estimation of influenza epidemics using Google search data*, arXiv preprint arXiv:1505.00864.