creative commons
COMMONS DEED

저작자표시-비영리-변경금지 2.0 대한민국

이용자는 아래의 조건을 따르는 경우에 한하여 자유롭게

- 이 저작물을 복제, 배포, 전송, 전시, 공연 및 방송할 수 있습니다.

다음과 같은 조건을 따라야 합니다:

저작자표시. 귀하는 원저작자를 표시하여야 합니다.

비영리. 귀하는 이 저작물을 영리 목적으로 이용할 수 없습니다.

변경금지. 귀하는 이 저작물을 개작, 변형 또는 가공할 수 없습니다.

- 귀하는, 이 저작물의 재이용이나 배포의 경우, 이 저작물에 적용된 이용허락조건을 명확하게 나타내어야 합니다.
- 저작권자로부터 별도의 허가를 받으면 이러한 조건들은 적용되지 않습니다.

저작권법에 따른 이용자의 권리는 위의 내용에 의하여 영향을 받지 않습니다.

이것은 이용허락규약(Legal Code)을 이해하기 쉽게 요약한 것입니다.

Disclaimer

dCollection

이 학 석 사 학 위 논 문

Efficient estimation of influenza epidemics based on seasonal ARIMA models

계절형 ARIMA모형을 기반으로 한 효율적인 유행성 독감 추정

2016년 2월

서울대학교 대학원

통계학과

정 지 훈

Efficient estimation of influenza epidemics based on seasonal ARIMA models

지도교수 이 상 열

이 논문을 이학석사 학위논문으로 제출함
2015년 10월

서울대학교 자연과학대학원
통계학과
정 지 훈

정지훈의 이학석사 학위논문을 인준함
2015년 12월

위 원 장	김 용 대	(인)
부위원장	이 상 열	(인)
위 원	박 병 욱	(인)

Efficient estimation of influenza epidemics based on seasonal ARIMA models

by

Jung Jihoon

A Thesis
submitted in fulfillment of the requirement
for the degree of
Master of Science
in
Statistics

The Department of Statistics
College of Natural Sciences
Seoul National University
February, 2016

Abstract

The issue of big data has received a lot of attention from researchers working in various fields as big data sets have been considered to have great potential in detecting or predicting future events such as epidemic outbreaks and changes in stock prices. Reflecting the current popularity of big data, a number of studies have been published to propose models that track influenza epidemics by using internet-based information. The most recent version of influenza tracking model was ARGO, which harnesses search queries from Google and reports from the Centers for Disease Control (CDC). Although the ARGO appears to predict the outbreaks of influenza accurately and outperform other existing methods such as Google Flu Trends (GFT), in this study, a classical seasonal autoregressive and integrated moving average (SARIMA) model is demonstrated to show that it outperforms the ARGO. Compared to the ARGO, the time series model incorporates more accurate seasonality of the past influenza activities, and takes less input variables into account since it utilizes only the reports provided by the CDC. The findings in this study show that the SARIMA model is a more efficient tool for estimating influenza epidemics.

Keywords: SARIMA model, ARGO, Seasonal effect, Influenza epidemics, Disease detection, Influenza-like illnesses activity estimation, Big data

Student Number: 2014-20288

Contents

1 Introduction **1**

2 Data description and seasonal time series model **4**

3 Results **12**

4 Discussion **16**

List of Tables

2.1 AIC values for candidate models 6

3.1 Comparison of SMA(52) and ARGO for the estimation of influenza epidemics 14

List of Figures

2.1 Time series plot of the logit-transformed CDC's weighted ILI activity level 7
2.2 Time series plots 8
2.3 ACF and PACF of the first-order difference of the seasonal differenced time series 9
2.4 ACF and PACF of the second-order difference of the seasonal differenced time series 10
2.5 ACF and PACF of the second-order difference of the seasonal differenced time series 11

3.1 Estimation result. The true ILI activity level is plotted with solid line whereas the estimated level is plotted with dotted line 15

Chapter 1

Introduction

With an increasing usage of the internet by people, coupled with the advent of smartphones, 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 now called "big data". The issue of big data has received notable attention from practitioners in a variety of research and industrial fields (see, for instance, Preis, Moat, and Stanley (2013)) since big data sets have been considered to have considerable potential in predicting or estimating future events (see Labrinidis and Jagadish (2012)). For example, 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)), by taking advantage of publicly available online information.

Among such diseases, influenza has gotten particular attention with regards to disease detection. For majority of people, influenza has been considered to be relatively common disease that does not cause any fatal effect on them. Yet, influenza outbreaks, in fact, bring about up to 500,000 deaths worldwide every year. In the case of the United States only, from 1976 to 2007, the annual influenza-related deaths ranged from 3,000 to 50,000 ap-

proximately, with an average of about 24,000 annual deaths. Therefore, in recent years, various methods that utilize information from the internet, such as Google and Twitter, have been proposed to estimate influenza-like illnesses (ILI) activities: see Ginsberg et al. (2009), Polgreen et al. (2008), Santillana et al. (2015), and Bollen, Mao, and Zeng (2011).

One of those methods based on a big data approach is the AutoRegression with GOogle data (ARGO), proposed by Yang, Santillana, and Kou (2015). The ARGO harnesses both the Google search data and ILI reports from the Centers for Disease Control (CDC) to track influenza outbreaks.

Although the ARGO seems to have better performance than existing methods such as Google Flu Trends (see Cook et al. (2011) and Santillana et al. (2014)) in terms of the estimation of influenza outbreaks, it has two major shortcomings. First of all, it does not fully reflect the time series properties, especially the seasonality in influenza epidemics, thus failing to produce accurate estimates of the ILI activities. Second, it takes too many input variables into account, causing its algorithm to be so much complicated and to take a quite long time to draw conclusion.

To resolve such defects that the ARGO has, this study proposes to employ a classical time series model, a seasonal autoregressive integrated moving average (SARIMA) model which only uses the CDC's ILI reports. In implementation, an optimal SARIMA model is selected based on the value of AIC. In this process of finding the optimal time series model, seasonal effect and stochastic trend are removed by differencing the time series. After the model is selected, its performance is compared with that of the ARGO by using the accuracy metrics adopted in Yang et al. (2015).

The findings in this study 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, as the model uses

the data provided by the CDC only, it has an advantage over the ARGO only to take one variable into consideration, which enables it to save the computing time significantly. In terms of the accuracy of estimation, the SARIMA model seems to significantly outperform the ARGO even though it does not involve any publicly available online search data. This result can be explained by the fact that all important information is already locked in the time series itself and the SARIMA model can successfully extract the information without harnessing any exogenous variable.

This study is organized as follows. Chapter 2 introduces the ARGO and SARIMA model and shows the process of selecting an optimal SARIMA model. In Chapter 3 the SARIMA model is compared with the ARGO based on several accuracy metrics.

Chapter 2

Data description and seasonal time series model

The ARGO model introduced by Yang et al. (2015) is given by

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

where $y_t = logit(p_t)$ refers to the logit-transformed ILI activity level at time t and $X_{i,t}$ is the Google search frequency of term i among 100 Google search terms at time t. They claimed that the ARGO incorporates the seasonality in the CDC-reported ILI activity level dataset. However, a close inspection of the formulation of the model led to an idea that the ARGO does not fully reflect a seasonal effect. Motivated by the idea, this study considers using a seasonal time series model such as a seasonal ARIMA (SARIMA) model.

As in Yang et al. (2015), 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) is used in this study. The CDC's ILI activity level refers to the percentage of patients diagnosed positive with regards to the ILI among total patients who visited the hospital with ILI

symptoms over a week. Also, the logit-transformation is performed on the CDC's weighted ILI activity level, that is, $x_t = logit(w_t)$, where w_t refers to the weighted ILI activity level at time t.

For the first step, x_t is plotted from March 29, 2009 to July 11, 2015. Figure 2.1 clearly shows that the time series has a seasonal effect as well as nonconstant variance. The seasonal pattern seems to occur every year or 52 weeks. Thus, a seasonal differencing is applied to the data to resolve the issue of the seasonality. Figure 2.2 shows the plots of seasonal differenced, the first and second-order difference of the seasonal differenced time series. The plot in (c) appears to best incorporate the stationarity. The Dicky-Fuller test actually shows that the seasonal differenced time series has a unit root and its first-order differenced time series does not have any unit root. Thus, in this study, both the first and second-order difference of the seasonal differenced time series are taken into consideration.

The graphs of the autocorrelation function (ACF) and partial autocorrelation function (PACF) of the first-order and second-order difference of the seasonal differenced time series are shown in Figures 2.3 and 2.4. Figure 2.3 shows that the ACF of the first-order difference has statistically significant peaks at too many lags. This suggests that the second-order differenced time series is more adequate than the first-order differenced one for fitting the ILI activity dataset. The second-order difference of the seasonal differenced time series can be written as

$$\nabla_{52}\nabla^2 x_t = (1 - B^{52})(1 - B)^2 x_t.$$

Based on the ACF and PACF of the second-order difference within the seasonal lag, (0,2,4) appears to be the optimal order selection for the dataset. Next, another graph of the ACF and PACF of the second-order difference is produced to decide appropriate order of P and Q in ARIMA$(0, 2, 4) \times (P, 1, Q)_{52}$. Figure 2.5 shows that the ACF and PACF have strong peaks at

lag 52, namely, the seasonal lag. In particular, the ACF shows that the peaks at lag 51 and 53 tend to be symmetrical around the seasonal lag. Also, the second-order difference is estimated to have mean 0 based on the plot (c) in Figure 2.2. Thus, ARIMA$(0,2,4)(0,1,1)_{52}$ without constant term is a reasonable candidate model for analyzing the dataset. Since the PACF shows statistically significant peak at the seasonal lag, ARIMA$(0,2,4)(1,1,0)_{52}$ is also considered as a candidate model.

Table 2.1: AIC values for candidate models

AIC values	
$(0,2,4) \times (0,1,1)_{52}$	-401.14
$(0,2,4) \times (1,1,0)_{52}$	-399.07

The value of the Akaike's Information Criterion (AIC) goodness of fit measure is used as criterion to determine the optimal model. That is, the model with smaller AIC value would be the better fit model than another candidate. As shown in Table 2.1, the combination of $(0,2,4) \times (0,1,1)_{52}$ has lower AIC value. As a result, ARIMA$(0,2,4) \times (0,1,1)_{52}$ is chosen as the optimal fit model.

In general, the SARIMA model, i.e. ARIMA$(p,d,q) \times (P,D,Q)_S$, is given by

$$\phi(B)\Phi(B^s)(1-B^s)^D(1-b)^d x_t = \delta + \theta(B)\Theta(B^s)\epsilon_t.$$

According to the analysis, the estimated SARIMA model for x_t, i.e. ARIMA$(0,2,4) \times (0,1,1)_{52}$, is given by

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

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

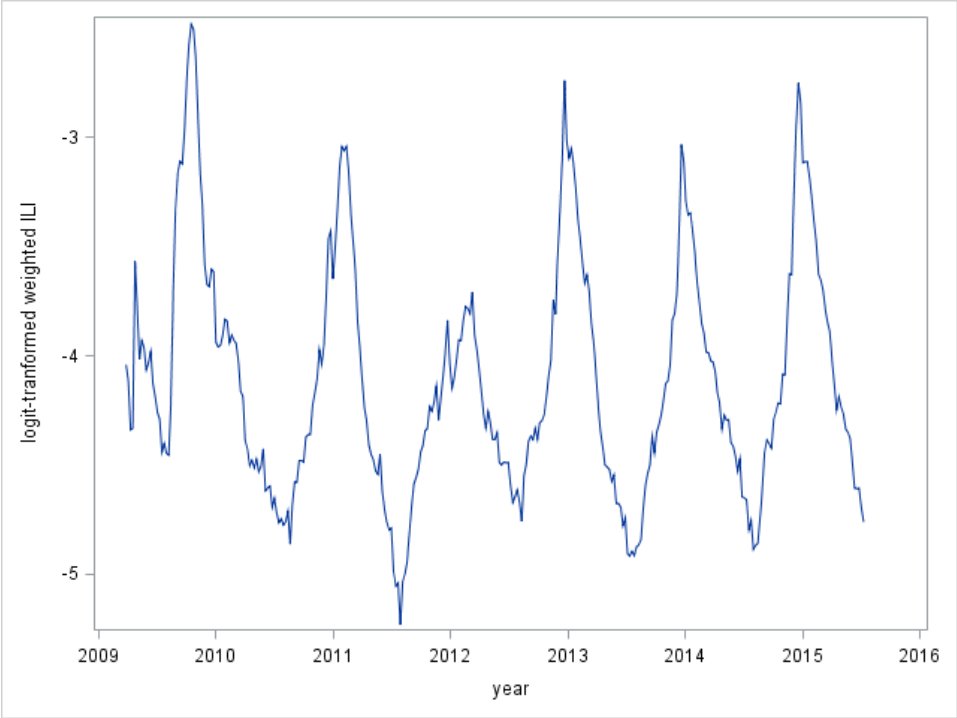
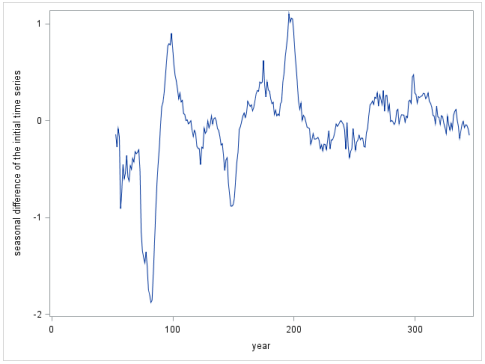
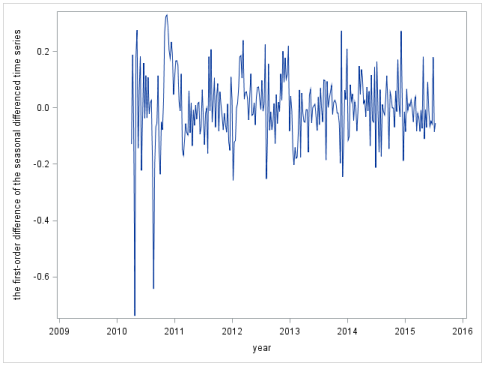


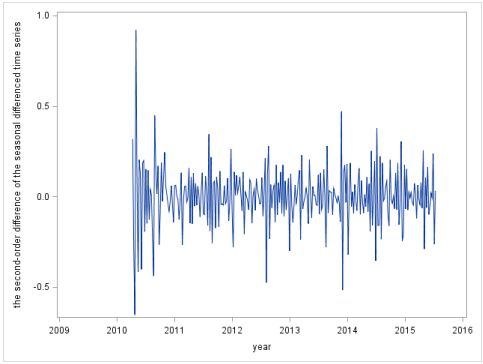
Figure 2.1: Time series plot of the logit-transformed CDC's weighted ILI activity level



(a) Seasonal differenced time series

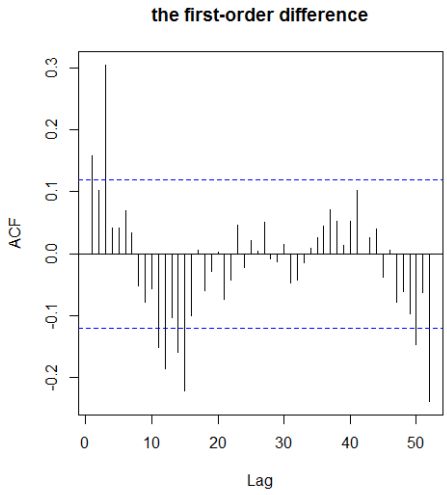


(b) The first-order difference of the seasonal differenced time series

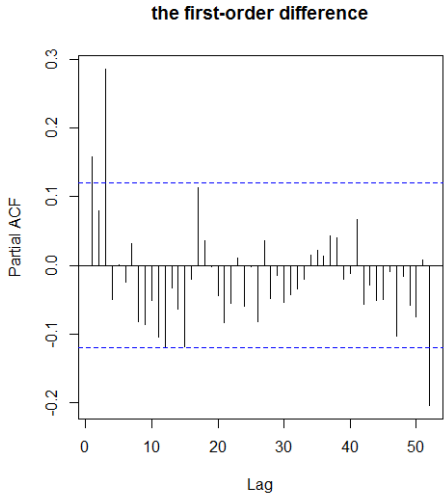


(c) The second-order difference of the seasonal differenced time series

Figure 2.2: Time series plots

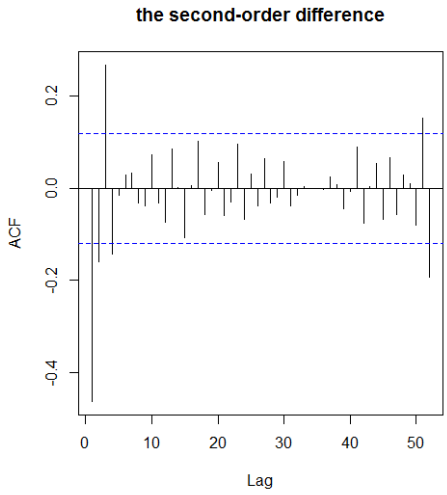


(a) ACF of the first-order difference of the seasonal differenced time series

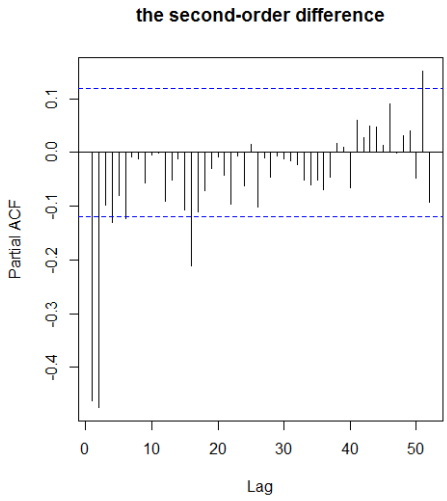


(b) PACF of the first-order difference of the seasonal differenced time series

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

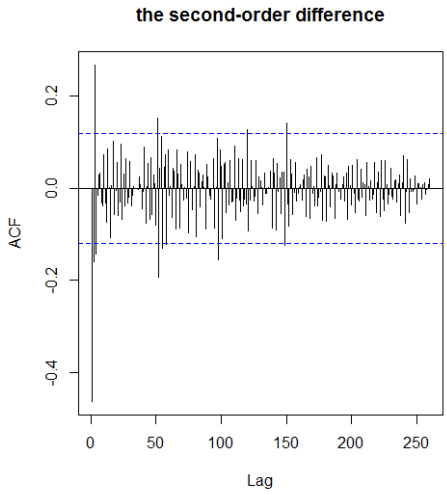


(a) ACF of the second-order difference of the seasonal differenced time series

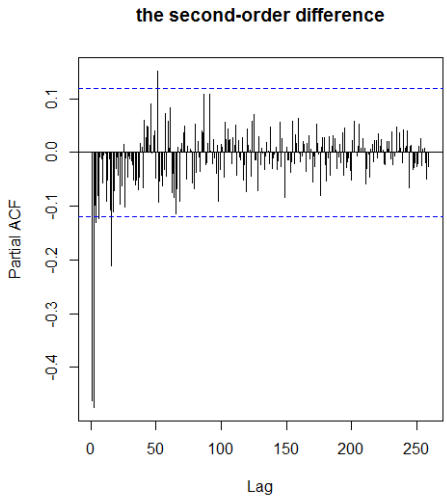


(b) PACF of the second-order difference of the seasonal differenced time series

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



(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 2.5: ACF and PACF of the second-order difference of the seasonal differenced time series

Chapter 3

Results

In this chapter, a comparison study of the performance of the SMA(52) and ARGO was conducted in the same setting as that of Yang et al. (2015). Therefore, the five accuracy metrics used by Yang et al. (2015) are calculated in this study. The metrics include 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 with the observed values, all of which are defined respectively as

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

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

Table 3.1 shows the summary of the metrics for the SMA(52) and ARGO for different time periods. It presents the values of performance metrics for the whole period (2009-2015) and regular flu seasons (winter time) each year from 2010 to 2015. From the table, it can be seen that, for the whole time period, the SMA(52) outperforms the ARGO in all accuracy metrics except for the correlation of increment. To be specific, in terms of the RMSE, MAE, and MAPE, the SMA(52) has outstanding values that are considerably lower than those of the ARGO. With regard to the correlation, SMA(52) ($r = 98.7\%$) slightly outperforms the ARGO ($r = 98.6\%$) for the whole time period. However, during the regular flu seasons, the former has a higher correlation only for the 2012-2013 season and a slightly lower correlation in the rest of the flu seasons. Overall, in terms of the correlation, the SMA(52) has a similar performance to the ARGO. The correlation of increment between a model (the SMA(52) or ARGO) and the CDC-reported ILI activity level is estimated for the purpose of checking how well the model responds to the change in the ILI activity. According to Table 3.1, the SMA(52) has a performance similar to the ARGO except during the 2013-2014 flu season. Figure 3.1 shows the fitted values of the SMA(52) against the observed ILI activity level.

Table 3.1: Comparison of SMA(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	SMA(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	SMA(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	SMA(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	SMA(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	SMA(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**

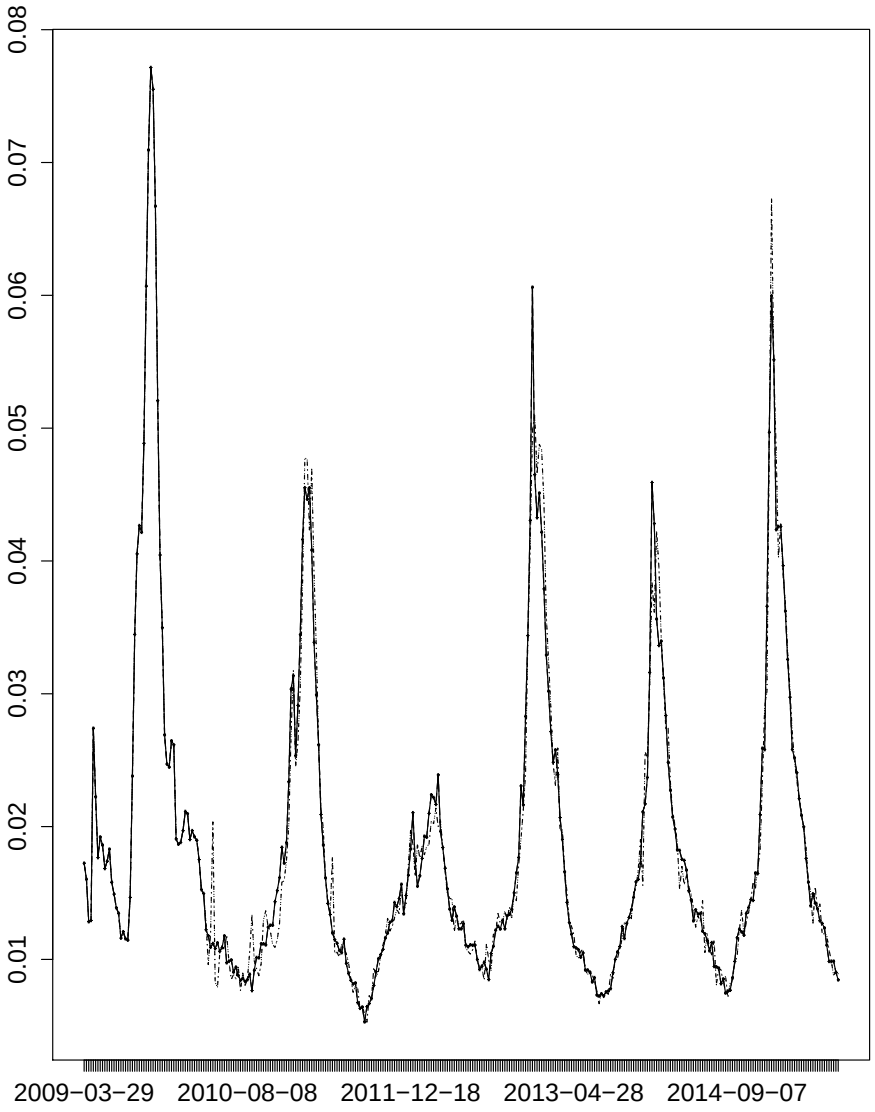


Figure 3.1: Estimation result. The true ILI activity level is plotted with solid line whereas the estimated level is plotted with dotted line

Chapter 4

Discussion

The findings show that the time series model, the SMA(52), is superior to the ARGO in terms of accuracy. Furthermore, since the ARGO outperforms all other influenza tracking models based on Google searches, the findings imply that the SMA(52) is more suitable in estimating influenza than any of existing flu tracking methodologies that utilize publicly available Google search data.

The key factor for the SMA(52) to have better performance than the ARGO is the incorporation of the seasonality of the historical ILI activities, which is clearly shown in Figure 2.1. Furthermore, the second-order differencing helps stabilize the time-varying variances. Another advantage of the SMA(52) is that, unlike the ARGO, it does not use the Google searches, which are highly sensitive to the public's overreaction to the disease. As the model only harnesses officially-reported data provided by the national institution, its prediction does not get influenced by people's behavior reflected on the internet, which can mislead the conclusion. Moreover, Lazer, Kennedy, King, and Vespignani (2014) said 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. However, the SMA(52) is not under the influence of the algorithm underlying Google.

Although the SMA(52) demonstrates outstanding performance that is highly superior to that of the ARGO, it still has the same limitation as the ARGO has becasue the CDC reports ILI activity level from one to two weeks after the target date. Thus, the information gap is unavoidable and inevitable as long as only a weekly dataset is available for prediction.

Acknowledgements. A working paper entitled "Prediction of Influenza Epidemics based on Seasonal ARIMA Models" (with S. Lee) is now on going based on this thesis.

Bibliography

Bollen, J., Mao, H., and Zeng, X. (2011). Twitter mood predicts the stock market. *J. Comput. Sci.*, **2(1)**, 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 Negl. Trop. Dis.*, **5(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(8)**, e23610.

Ginsberg, J., et al. (2009). Detecting influenza epidemics using search engine query data. *Nature*, **457(7232)**, 1012–1014.

Jung, J. and Lee S. (2015). Prediction of Influenza Epidemics based on Seasonal ARIMA Models. Unpublished manuscript.

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.

Polgreen, P. M., Chen, Y., Pennock, D. M., and Nelson, F. D. (2008). Using Internet searches for influenza surveillance. *Clin. Infect. Dis.*, **47(11)**, 1443–1448.

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

Santillana, M., et al. (2015). Combining search, social media, and traditional data sources to improve influenza surveillance. *PLoS Comput. Biol.*, **11(10)**, 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?. *Am J Prev Med*, **47(3)**, 314–347.

Wesolowski, A., et al. (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. Unpublished manuscript.

국문초록

빅데이터가 전염병의 발생이나 주가 변화 등 미래에 일어날 사건들을 감지하거나 예측하는 데 있어 잠재력을 가지고 있다고 여겨지면서 다양한 분야의 전문가들로부터 많은 관심을 받고 있다. 이러한 현재의 빅데이터 인기를 반영하듯이 인터넷 기반의 정보를 이용하여 유행성 독감 발생을 예측할 수 있는 모형을 제시하는 논문들이 상당수 발표되었다. 가장 최신의 독감 추정 모형은 구글 기반의 검색 데이터와 미국 질병관리센터 (CDC)에서 제공하는 자료를 이용한 아르고(ARGO)이다. 아르고는 독감 발생을 상당히 잘 예측하고 GFT와 같은 현존하는 모형들을 독감 발생 추정능력에서 능가하는 것으로 보여진다. 그러나 본 논문에서는 전통 시계열모형인 계절형 자기회귀누적이동평균모형 (Seasonal AutoRegressive Integrated Moving Average, SARIMA)을 제시하고 이 모형이 독감 발생 추정능력에서 아르고를 능가한다는 것을 보여준다. 아르고와 비교하였을 때 이 모형은 과거의 독감 활동이 가지고 있는 계절성을 더 정확히 반영하고 있다. 또한 이 모형은 CDC에서 제공하는 자료만을 사용하기 때문에 더 적은 수의 변수를 고려한다. 본 논문의 결과는 SARIMA모형이 독감 발생을 더욱 효과적으로 추정한다는 것을 보여준다.

주 요 어 : 계절형 자기회귀누적이동평균모형 (SARIMA), 아르고 (ARGO), 계절효과, 유행성 독감, 질병 감지, 독감 유사 질병 (ILI) 추정, 빅데이터

학 번 : 2014-20288