MEDINFO 2023 — The Future Is Accessible
J. Bichel-Findlay et al. (Eds.)

doi:10.3233/SHTI231086

Assessing Internet Search Models in Predicting Daily New COVID-19 Cases and Deaths in South Korea

Atina HUSNAYAIN[a,1] and Emily Chia-Yu SU[b]
[a] *Public Health Department, Monash University Indonesia, Banten, Indonesia*
[b] *Graduate Institute of Biomedical Informatics, Taipei Medical University, Taipei, Taiwan*
ORCiD ID: Atina Husnayain https://orcid.org/0000-0003-3002-8728, Emily Chia-Yu Su https://orcid.org/0000-0003-4801-5159

Abstract. Search data were found to be useful variables for COVID-19 trend prediction. In this study, we aimed to investigate the performance of online search models in state space models (SSMs), linear regression (LR) models, and generalized linear models (GLMs) for South Korean data from January 20, 2020, to July 31, 2021. Principal component analysis (PCA) was run to construct the composite features which were later used in model development. Values of root mean squared error (RMSE), peak day error (PDE), and peak magnitude error (PME) were defined as loss functions. Results showed that integrating search data in the models for short- and long-term prediction resulted in a low level of RMSE values, particularly for SSMs. Findings indicated that type of model used highly impacts the performance of prediction and interpretability of the model. Furthermore, PDE and PME could be beneficial to be included in the evaluation of peaks.

Keywords. Prediction, time series, internet search, COVID-19, digital epidemiology

1. Introduction

South Korea is known as one of the most aggressive countries in tackling the spread of the COVID-19 pandemic. The first case is reported on January 20, 2020, and since then multiple control measures were implemented including an integrated digital data platform that runs multiple artificial intelligence (AI) systems in producing information and responses [1]. Their AIs include warning systems for possible contacts and health providers, notification for hospitals, ambulances, mobile test laboratories, and nearest drive-through laboratory, possible cluster detection, as well as mask supply and distribution. With those approaches, South Korea has managed to bring the situation under control.

However, by the end of December 2021, there were an increased number of critically ill patients [2] and the spread of highly contagious Omicron variant [3]. This situation

[1] Corresponding Author: Atina Husnayain, Public Health Department, Monash University Indonesia, BSD Green Office Park 9, 6F, Sampora, Cisauk, Tangerang Selatan, Banten 15345, Indonesia, email: atina.husnayain@monash.edu.

showed that there was longer disease transmission in the population which may require trend predictions in preparing for upcoming future waves [4], in terms of human resources and medical equipment deployment [5]. A study by Rabiolo et al [6] found that models incorporating search data performed better in the first month of outbreak prediction. Similar results were also reported in two previous studies from Iran [7] and India, the United States, and the United Kingdom [8].

Utilizing search data for trend prediction was aimed to capture the patterns of online health-seeking behavior which may provide a real-time indication of symptoms in a population [9]. Therefore, new waves or peaks could possibly be detected at the earlier stage of the outbreak [10]. Our previous study [11] in South Korea demonstrated that search volumes were useful variables in predicting daily new COVID-19 cases and deaths in the first 6 months of the outbreak with higher feature effects. Although, it remains unclear whether the type of model used will affect the performance of online search models, particularly for longer prediction periods. In this study, we assessed the performance of internet search models in state space models (SSMs), linear regression (LR) models, and generalized linear models (GLMs) for COVID-19 cases and deaths prediction in South Korea.

2. Methods

2.1. Datasets

Country-level COVID-19 cases and deaths were downloaded from the Center for Systems Science and Engineering at Johns Hopkins University [12]. Besides, mobility data were collected from Google's Community Mobility Reports [13] along with Apple's Mobility Trends Reports [14]. In addition, NAVER search volumes were retrieved from NAVER's website [15] using terms in Korean language for coronavirus, coronavirus test, Middle East respiratory syndrome, face mask, social distancing, Shinchoenji, kf94 mask, disposable mask, thermometer, hand sanitizer, mask strap, and kf80 mask. Quotation marks were used for terms with more than two words and search data were retrieved for all types of searches, genders, and age groups. Case-related data were collected from January 20, 2020, as the first COVID-19 case was reported in South Korea to July 31, 2021, while mobility and search data were queried with a lag of 3 days. Data were then compiled into four subsets including 3, 6, 12, and 18 months after the first case was reported. Furthermore, missing values were filled using a mean of the subset.

2.2. Statistical Analysis

Analyses and visualizations in this study were executed in SAS Studio (SAS Institute, Cary, NC, USA). Each subset was subjected to principal component analysis (PCA) using proc hpprincomp to eliminate the effect of multicollinearity. Composite variables that account for around 90% of the variance were then used in the model development. Proc glmselect with lasso (LR1), adaptive lasso (LR2), and elastic net regularization (LR3), steps of 25, and the lowest Akaike information criterion (AIC) in picking model were used to create the LR model. While GLM was created in proc hpgenselect using three distinct distributions including normal (GLM1), Poisson (GLM2), and negative binomial (GLM3), utilizing stepwise selection and an alpha of 0.05 for the selection technique. Besides, SSM with proc ssm was performed with three types of trends:

random walk (SSM1), local linear (SSM2), and damped local linear (SSM3), as well as white noise to represent irregular trends. The model was trained using all of the data from each subgroup.

Root mean squared error (RMSE) values were defined as loss functions as well as peak day error (PDE) and peak magnitude error (PME) to evaluate the performance of the model in predicting peaks of the pandemic. The formulas of root mean squared error (RMSE), peak day error (PDE), and peak magnitude error (PME) are written in Eqs.1-3:

$$RMSE = \frac{\sqrt{\sum_{i=1}^{n} (P_i - O_i)^2}}{n} \quad (1)$$

$$PDE = p` - p \quad (2)$$

$$PME = h` - h \quad (3)$$

In RMSE, Pi is defined as predicted and Oi as observed values of daily new COVID-19 cases and deaths in the time series. Whereas PDE was the difference between predicted and observed peak days and PME was calculated as the difference between the observed and predicted value in the peak of daily new COVID-19 cases and death. Where p and p` denote the observed and predicted peak day, while h and h` denote the maximum values reached by the actual and predicted target, respectively. The formula of PDE and PME was adopted from a previous study [16] with several changes.

3. Results

Prediction results (Table 1) revealed that the state space model with local linear trend type (SSM2) performed better in predicting cases and the state space model with damped local linear trend type (SSM3) for predicting deaths, as measured by root mean square error (RMSE) values. Despite the RMSE values being extremely low, when models were evaluated by peak day error (PDE) and peak magnitude error (PME), different findings were obtained. Figure 1 showed that in most cases, the timing of peaks was accurately predicted, with the exception of the second peak of the third subset (+1) and the last peak in the fourth subset (-3). Even while the model properly predicted the number of cases as indicated by the RMSE values, the magnitude of errors measured by PME ranged from -189.91 to 264.67. In addition, peaks for death prediction were also accurately predicted except for the last peak in the fourth subset (+2), with peak magnitude errors varied from -10.80 to -1.01.

Table 1. Performance of search models, assessed by root mean squared error (RMSE) values.

Model	Prediction of daily new COVID-19 cases				Prediction of daily new COVID-19 deaths			
	Set 1	Set 2	Set 3	Set 4	Set 1	Set 2	Set 3	Set 4
GLM1	6.17	3.88	0.26	3.43	0.04	0.01	0.03	0.09
GLM2	0.41	0.38	1.29	3.40	0.01	0.01	0.01	0.01
GLM3	0.12	0.18	0.29	0.70	0.02	0.00	0.01	0.01
LR1	4.14	3.38	0.66	2.33	0.08	0.06	0.03	0.11
LR2	1.50	3.93	1.12	1.57	0.02	0.05	0.07	0.05
LR3	0.77	4.81	2.33	0.19	0.07	0.02	0.02	0.21
SSM1	0.09	0.30	0.19	0.10	0.02	0.00	0.00	0.00
SSM2	0.07	0.06	0.08	0.03	0.00	0.00	0.00	0.00
SSM3	0.30	0.08	0.05	0.04	0.00	0.00	0.00	0.00

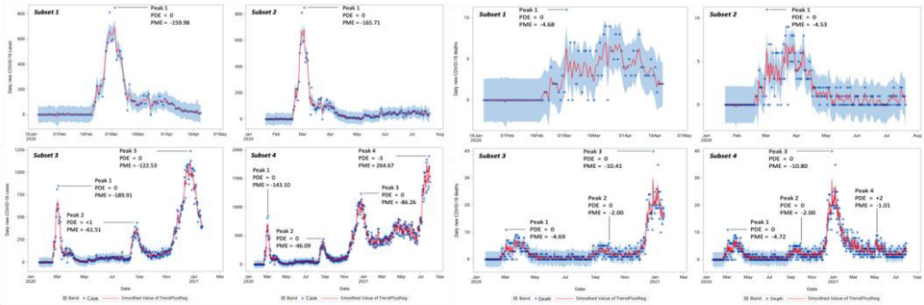

Figure 1. Time series of daily new COVID-19 cases and deaths in South Korea from January 20, 2020, to July 31, 2021, and predicted values in the state space models (SSMs). PDE: peak day error; PME: peak magnitude error.

Furthermore, developed models (Table 2) for case prediction indicated that the first composite variable only shared higher parameter estimates in the first and second subsets. Higher parameter estimates were observed in the model for trend components including level, slope, and noise variance. Similar results were also found in death prediction. This finding demonstrated that trend components had a greater influence than composite variables in the model.

Table 2. Parameter estimates of models.

Component	Prediction of daily new COVID-19 cases				Prediction of daily new COVID-19 deaths			
	Set 1	Set 2	Set 3	Set 4	Set 1	Set 2	Set 3	Set 4
Comp 1	36.17	19.73	11.30	6.69	0.17	0.10	0.06	0.02
Comp 2	-3.54	-11.38	-11.35	-3.22	-0.19	-0.18	-0.32	-0.13
Comp 3	-2.54	8.40	17.13	6.32	-0.13	-0.01	0.03	0.09
Comp 4	18.68	12.91	-0.55	-5.17	-0.22	-0.02	0.37	0.13
Comp 5	-4.63	4.92	1.42	5.76	-0.12	-0.11	0.05	0.35
Comp 6	1.70	10.44	8.02	-5.02	-0.08	-0.18	-0.14	0.02
Comp 7	14.18	-1.74	12.99	4.37	0.08	-0.10	-0.05	-0.13
Comp 8	24.85	0.53	4.91	6.73	-0.06	0.14	0.11	0.01
Comp 9	—	—	1.78	-1.98	—	—	—	-0.28
Level variance	2.47E+3	1.32E+3	1.50E+3	3.94E+3	1.05E-8	0.16	1.05E-8	1.05E-8
Slope variance	1.05E-8	1.05E-8	1.05E-8	1.05E-8	3.30	2.20	4.15	3.69
Phi	—	—	—	—	1.00E-5	0.00	1.00E-5	1.00E-5
Noise variance	2.16E+3	1.08E+3	1.70E+3	2.05E+3	2.71	1.74	3.17	2.83

4. Discussion

In this study, the use of search data in the models for short- and long-term prediction resulted in low RMSE values. State space models (SSMs) outperformed our prior models [11] in linear regression (LR) models and generalized linear models (GLMs). It might imply that the type of model utilized has a significant influence on prediction performance. If the target variable has a greater magnitude of trend component, a time series-based model that includes trend type in the model may perform better in prediction.

Besides, findings revealed that the type of model utilized also had a significant influence on model interpretability. In our previous study [11], search data had a stronger influence on the model. In this analysis, however, greater parameter estimates (Table 2) in the model were obtained in trend components including level, slope, and noise

variance when compared to composite variables constructed using search data, mobility reports, and COVID-19 metrics.

Larger levels of error were discovered in the measurement of PDE and PME when compared to RMSE. PDE and PME (adopted from a previous study [16] with several changes) were defined in this study as the difference between expected and observed peak days, as well as the difference between the observed and predicted value in the peak of daily new COVID-19 cases and deaths. This peak assessment which includes an evaluation of peak timing and amplitude may help public health practitioners and policymakers to anticipate future waves.

5. Conclusions

Utilizing search data for trend prediction was aimed to capture the patterns of online health-seeking behavior which may provide a real-time indication of symptoms in a population. In addition, types of models have a significant influence on the model's prediction performance and interpretability.

References

[1] M. Islam, South Korea winning the fight against coronavirus using big-data and AI, in, The Daily Star, Seoul, 2020.
[2] J. McCurry and S. Lock, Covid cases rise across Asia as South Korea sees record numbers of seriously ill, Thailand restarts quarantine, in: *The Guardian*, The Guardian, 2021.
[3] H. Shin, South Korea to extend curbs amid Omicron surge, serious COVID-19 cases, in: *Reuters*, 2021.
[4] B.A. Panuganti, A. Jafari, B. MacDonald, and A.S. DeConde, Predicting COVID-19 Incidence Using Anosmia and Other COVID-19 Symptomatology: Preliminary Analysis Using Google and Twitter, *Otolaryngol Head Neck Surg* **163** (2020), 491-497.
[5] I. Ahmad, R. Flanagan, and K. Staller, Increased Internet Search Interest for GI Symptoms May Predict COVID-19 Cases in US Hotspots, *Clin Gastroenterol Hepatol* **18** (2020), 2833-2834.e2833.
[6] A. Rabiolo, E. Alladio, E. Morales, A.I. McNaught, F. Bandello, A.A. Afifi, and A. Marchese, Forecasting the COVID-19 Epidemic by Integrating Symptom Search Behavior Into Predictive Models: Infoveillance Study, *J Med Internet Res* **23** (2021), e28876.
[7] S.M. Ayyoubzadeh, S.M. Ayyoubzadeh, H. Zahedi, M. Ahmadi, and R.N.K. S, Predicting COVID-19 Incidence Through Analysis of Google Trends Data in Iran: Data Mining and Deep Learning Pilot Study, *JMIR Public Health Surveill* **6** (2020), e18828.
[8] S. Prasanth, U. Singh, A. Kumar, V.A. Tikkiwal, and P.H.J. Chong, Forecasting spread of COVID-19 using google trends: A hybrid GWO-deep learning approach, *Chaos Solitons Fractals* **142** (2021), 110336.
[9] Y. Ortiz-Martínez, J.E. Garcia-Robledo, D.L. Vásquez-Castañeda, D.K. Bonilla-Aldana, and A.J. Rodriguez-Morales, Can Google® trends predict COVID-19 incidence and help preparedness? The situation in Colombia, *Travel Med Infect Dis* **37** (2020), 101703.
[10] Y.H. Lin, C.H. Liu, and Y.C. Chiu, Google searches for the keywords of "wash hands" predict the speed of national spread of COVID-19 outbreak among 21 countries, *Brain Behav Immun* **87** (2020), 30-32.
[11] A. Husnayain, E. Shim, A. Fuad, and E.C. Su, Predicting New Daily COVID-19 Cases and Deaths Using Search Engine Query Data in South Korea From 2020 to 2021: Infodemiology Study, *J Med Internet Res* **23** (2021), e34178.
[12] E. Dong, H. Du, and L. Gardner, An interactive web-based dashboard to track COVID-19 in real time, *Lancet Infect Dis* **20** (2020), 533-534.
[13] Google, COVID-19 Community Mobility Reports, in, 2021.
[14] Apple, Apple Mobility Report, in, 2021.
[15] NAVER, NAVER search volumes, in, 2021.
[16] Z. Ertem, D. Raymond, and L.A. Meyers, Optimal multi-source forecasting of seasonal influenza, *PLoS Comput Biol* **14** (2018), e1006236.