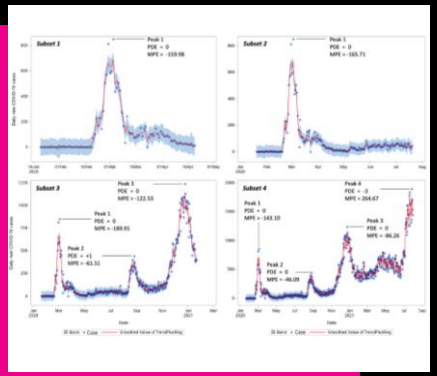
MEDINFO23
8 – 12 JULY 2023 | SYDNEY, AUSTRALIA



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

Atina Husnayain

Assistant Professor in Public Health

Monash University Indonesia

atina.husnayain@monash.edu

Infodemiology study

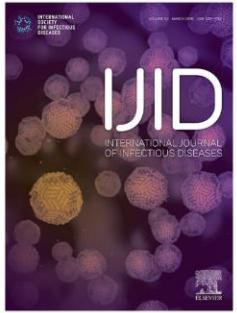
A promising approach in the field of epidemiology

1 Taiwan
2 Philippines
3 South Korea

Objective:
To **examine the potential use of search engine query data (1–3) for prediction purposes** and to **understand community online search behavior**.

4 USA
5 South Korea
6 South Korea

Objective:
To **evaluate the performance of online search models** in several aspects including **(4) locations (clustered and non-clustered areas), (5) times (pandemic stages), and (6) types of models**.



Journal of Medical Internet Research

JMIR Publicatons

PLOS ONE

IJID	PLoS ONE	JMIR
IF: 12.073	IF: 3.752	IF: 7.08

Publications

Predicting New Daily COVID-19 Cases and Deaths Using Search Engine Query Data in South Korea from 2020 to 2021: Infodemiology Study **(JMIR)**

High variability in model performance of Google relative search volume in spatially clustered COVID-19 areas of the US **(IJID)**

Understanding the Community Risk Perceptions of the COVID-19 Outbreak in South Korea: Infodemiology Study **(JMIR)**

Exploring online search behavior for COVID-19 preventive measures: The Philippine case **(PLoS ONE)**

Applications of Google Search Trends for risk communication in infectious disease management: A case study of the COVID-19 outbreak in Taiwan **(IJID)**

MAIN STUDIES

Type of models

South Korea

ABSTRACT

Background: It remains **unclear** whether the **type of model** used highly **impacts** the **performance** of **model incorporating online search volumes both at the national and regional level,** particularly **for longer prediction periods.**

Dataset: We used country-level **case-related data**, NAVER **search** volumes, and **mobility** data obtained from Google and Apple. Data were aggregated into four subsets (3, 6, 12, and 18 months).

Time frame: **January 20, 2020 to July 31, 2021**.

Objective: This study aimed to **examine** the **performance of online searches models in different types of models** for short- and long-term prediction of new COVID-19 cases and deaths in state space models (**SSMs**). We **compared** the results of analysis **with our previous model** in linear regression (**LR**) models and generalized linear models (**GLMs**).

South Korea

STUDY 6

Principal component analysis of cases & deaths

The first component

Variable[a]		Prediction of daily new COVID-19 cases				Prediction of daily new COVID-19 deaths			
		Subset 1[b]	Subset 2[b]	Subset 3[b]	Subset 4[b]	Subset 1[b]	Subset 2[b]	Subset 3[b]	Subset 4[b]
Case-based variables	1	0.27	0.25	0.08	-0.05	0.27	0.25	0.08	-0.05
	2	0.08	0.11	0.03	0.00	0.08	0.11	0.03	0.00
Google mobility	3	-0.27	-0.25	-0.16	-0.18	-0.27	-0.25	-0.16	-0.18
	4	-0.19	-0.22	-0.19	-0.20	-0.19	-0.22	-0.19	-0.20
	5	-0.08	-0.16	-0.17	-0.15	-0.08	-0.16	-0.17	-0.15
	6	-0.26	-0.24	-0.16	-0.16	-0.26	-0.24	-0.16	-0.16
	7	-0.14	-0.05	-0.00	-0.02	-0.14	-0.05	-0.00	-0.02
	8	0.24	0.20	0.13	0.12	0.24	0.20	0.13	0.12
Apple mobility	9	-0.21	-0.15	-0.07	-0.04	-0.21	-0.15	-0.07	-0.04
	10	-0.20	-0.06	-0.01	-0.02	-0.20	-0.06	-0.01	-0.02
NAVER search volumes	11	0.16	0.24	0.30	0.31	0.16	0.24	0.30	0.31
	12	0.26	0.29	0.32	0.32	0.26	0.29	0.32	0.32
	13	0.07	0.15	0.22	0.23	0.07	0.15	0.22	0.23
	14	0.32	0.32	0.34	0.34	0.32	0.32	0.34	0.34
	15	-0.02	-0.01	-0.01	0.02	-0.02	-0.01	-0.01	0.02
	16	0.06	0.09	0.12	0.13	0.06	0.09	0.12	0.13
	17	0.30	0.31	0.34	0.34	0.29	0.31	0.34	0.34
	18	0.30	0.31	0.34	0.33	0.30	0.31	0.34	0.33
	19	0.30	0.29	0.30	0.30	0.30	0.29	0.30	0.29
	20	0.11	0.18	0.25	0.26	0.11	0.18	0.25	0.26
	21	0.04	-0.09	-0.08	-0.04	0.04	-0.09	-0.08	-0.04
	22	0.30	0.27	0.30	0.31	0.30	0.27	0.30	0.31
Proportion of variance		0.39	0.41	0.34	0.35	0.39	0.41	0.34	0.35

Positive eigenvectors are mostly **found** in **case-based variables** and **search data**. Yet, **negative** eigenvectors are frequently **found** in **mobility data**, both from Google and Apple mobility reports.

Higher eigenvalues were **reported** for **search data** (**coronavirus, coronavirus test, face mask, and thermometer**) in all subsets for cases and deaths, compared to case-based variables and mobility variables.

South Korea

STUDY 6

Model performance

Better performance of model **found** in state space model with **local linear** trend type (**SSM2**) for **predicting cases** and state space model with **damped local linear** trend type (**SSM3**) for **predicting deaths**, **assessed by** root mean square error (**RMSE**) values.

RMSE values showed extremely low values.

State space model (**SSM**) **outperformed LR and GLM models** in all subset for predicting COVID-19 cases and deaths.

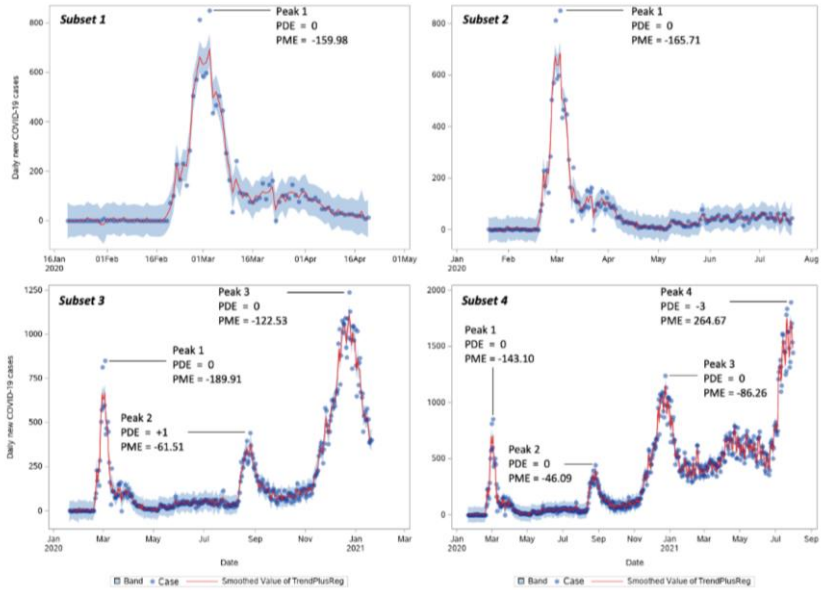
Model	Subset 1[a]	Subset 2[a]	Subset 3[a]	Subset 4[a]
Prediction of daily new COVID-19 cases				
GLM1[b]	6.173009	3.884628	0.262750	3.425700
GLM2[c]	0.407860	0.376861	1.293891	3.404890
GLM3[d]	0.120144	0.177100	0.285655	0.704238
LR1[e]	4.139906	3.375861	0.664009	2.325877
LR2[f]	1.495288	3.925300	1.123901	1.572179
LR3[g]	0.770830	4.814684	2.326752	0.187781
SSM1[h]	0.090701	0.302443	0.188837	0.100363
SSM2[i]	0.066934[k]	0.062925[k]	0.082488	0.031008[k]
SSM3[j]	0.295103	0.077743	0.054705[k]	0.036752
Prediction of daily new COVID-19 deaths				
GLM1	0.035888	0.013321	0.028742	0.090924
GLM2	0.012831	0.006588	0.013428	0.011810
GLM3	0.021062	0.003945	0.011148	0.010633
LR1	0.084272	0.057706	0.032435	0.108230
LR2	0.021080	0.047259	0.068419	0.045460
LR3	0.068250	0.022627	0.021701	0.209836
SSM1	0.018501	0.004651	0.004530	0.001543
SSM2	0.000921	0.003830	0.001998	0.002185
SSM3	0.003252[k]	0.001177[k]	0.000023[k]	0.000001[k]

GLM1: generalized linear model with a normal distribution.
GLM2: generalized linear model with a Poisson distribution.
GLM3: generalized linear model with a negative binomial distribution.
LR1: linear regression model with lasso regularization.
LR2: linear regression model with adaptive lasso regularization.
LR3: linear regression model with elastic net regularization.
SSM1: state space model with random walk trend type.
SSM2: state space model with local linear trend type.
SSM3: state space model with damped local linear trend type.

South Korea

STUDY 6

Model performance



Peak Day Error (PDE)

$$PDE = p` - p$$

p and p` denote the observed and predicted peak day.

Peak Magnitude Error (PME)

$$PME = h` - h$$

h and h` denote the maximum values reached by the actual and predicted peak, respectively.

In **predicting cases**, time of **peaks was correctly predicted, except** for the second peak of the third subset (**+1**) and the last peak in the fourth subset (**-3**).

Even model was accurately predicted the number of cases shown by RMSE values, the **magnitude of errors** measured by PME were **ranged from -189.91 to 264.67**.

South Korea

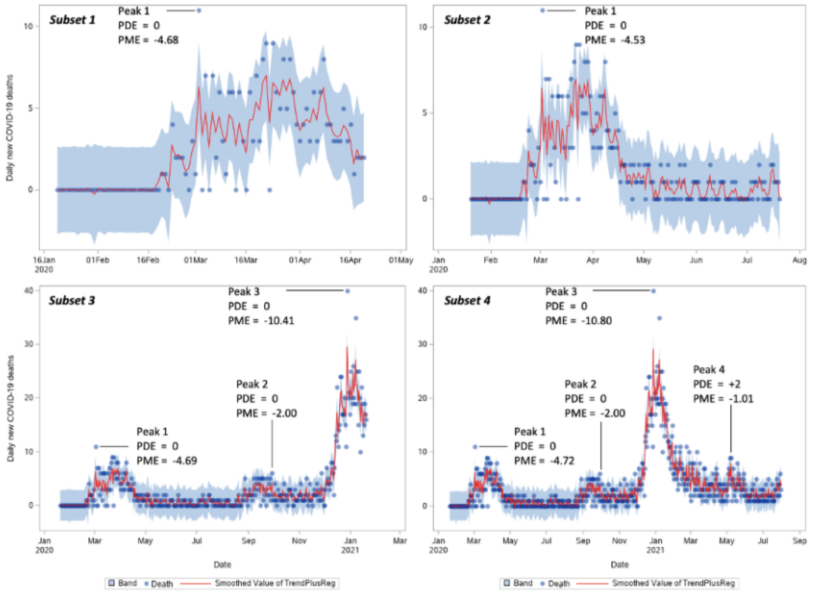
STUDY 6

Model performance

In **predicting deaths, RMSE values showed extremely low values**.

Peaks were correctly predicted, except for the last peak in the fourth subset (**+2**).

The **lower magnitude of peak errors** found in death prediction **ranged from −10.80 to −1.01**.



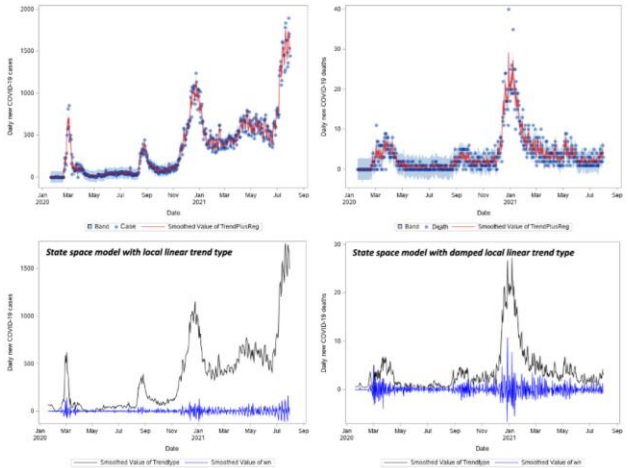
STUDY 6 — South Korea

Feature effects

Component	Prediction of daily new COVID-19 cases				Prediction of daily new COVID-19 deaths			
	Subset 1	Subset 2	Subset 3	Subset 4	Subset 1	Subset 2	Subset 3	Subset 4
Composite 1	36.17	19.73	11.30	6.69	0.17	0.10	0.06	0.02
Composite 2	-3.54	-11.38	-11.35	-3.22	-0.19	-0.18	-0.32	-0.13
Composite 3	-2.54	8.40	17.13	6.32	-0.13	-0.01	0.03	0.09
Composite 4	18.68	12.91	-0.55	-5.17	-0.22	-0.02	0.37	0.13
Composite 5	-4.63	4.92	1.42	5.76	-0.12	-0.11	0.05	0.35
Composite 6	1.70	10.44	8.02	-5.02	-0.08	-0.18	-0.14	0.02
Composite 7	14.18	-1.74	12.99	4.37	0.08	-0.10	-0.05	-0.13
Composite 8	24.85	0.53	4.91	6.73	-0.06	0.14	0.11	0.01
Composite 9	—[a]	—	1.78	-1.98	—	—	—	-0.28
Trend 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
Trend 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

The first composite variable only **shared higher parameter estimates** in the **first and second subsets** of **case** prediction.

Higher parameter estimates in the model were **found** in **trend components** including level and slope variance as well as **noise variance**. Patterns of **trend components were closely similar to cases and deaths** in the prediction model.



STUDY 6

Highlighted findings

South Korea

- **Search data** were **important variables** that **could be beneficial to be used in the prediction of cases and deaths**.
- **Type of model used highly impacts the performance of prediction**. If the higher magnitude of trend component was found in the target variable, time series-based model that includes trend type in the model may perform better in the prediction.
- **Type of model used highly impacts the interpretability of model**. Higher parameter estimates were found in trend components and noise variance.
- **Higher values of error** were **found in** the measurement of **PDE and PME compared to RMSE**. Assessment of peak may contribute to predicting future waves.