An ensemble approach improves the prediction of the COVID-19 epidemic

Kyulhee Han[1], Catherine Apio[1], Liu Zhe[1], Taewan Goo[1], **Taesung Park**[2*]

[1]Interdisciplinary program of bioinformatics, Seoul National University, Seoul 08826, Korea. [2]Department of Statistics, Seoul National University, Seoul 08826, Korea.
[*]Corresponding author; Tel: 82-2-880-8924, Fax: 82-2-883-6144, E-mail: tspark@stats.snu.ac.kr

Abstract

- Various models can be used to forecast COVID-19 cases. However, no single model consistently performs best in all situations.
- To improve prediction accuracy, we developed ensemble models that combine the predictions from multiple models. We used both raw and smoothed COVID-19 data from Korea. Mathematical, statistical, and machine learning models models were used to get the prediction results, and these predictions were combined by ensemble models. Best models for each response variables were selected using five different error measures.
- Among the 15 top-performing models for the raw data (daily confirmed cases, daily confirmed deaths, and ICU patients), 10 were ensemble models. Notably, the stacking ensemble model with support vector regression (SVR) exhibited the lowest test Weighted Mean Absolute Percentage Error (WMAPE), indicating the best test performance.

Keywords: *COVID-19, forecasting, ensemble*

Materials and Methods

Data collection and preprocessing

- Daily confirmed COVID-19 cases, daily confirmed deaths, and ICU patients were considered as the response variables, sourced from the Our World in Data (OWID) database. The Stringency Index (SI), booster shot rate (BSR), and the proportion of the BA.5 variant sequences in Korea (BA.5 rate) were used as covariates to improve the prediction accuracy of individual prediction models. The analysis period was January 1, 2022, to September 21, 2022 (test periods - September 15, 2022, to September 21, 2022).

Error measures

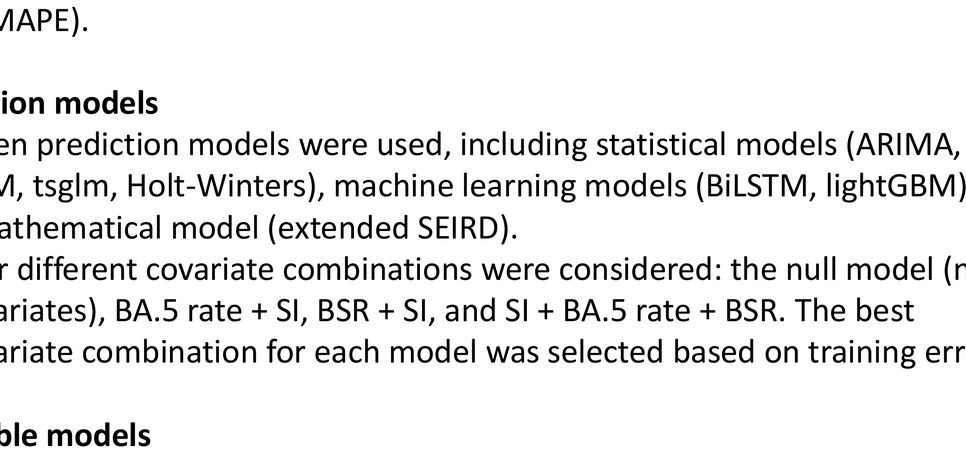
- To compare prediction accuracy, five error measures were considered: Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Mean Absolute Percentage Error (MAPE), r^2, and Weighted Mean Absolute Percentage Error (WMAPE).

Prediction models

- Seven prediction models were used, including statistical models (ARIMA, GAM, tsglm, Holt-Winters), machine learning models (BiLSTM, lightGBM), and a mathematical model (extended SEIRD).
- Four different covariate combinations were considered: the null model (no covariates), BA.5 rate + SI, BSR + SI, and SI + BA.5 rate + BSR. The best covariate combination for each model was selected based on training errors.

Ensemble models

- Four ensemble models (simple average, weighted average, stacking with linear regression, and stacking with support vector regression) were used to integrate the predictions from individual models.
- The average ensemble calculates the final prediction by taking the average of the predictions from individual models at each time point. The weighted average ensemble performs similarly, but it assigns different weights to each model. In this analysis, the weight for each model was determined by the normalized inverse of its training error, giving more weight to models with lower error.
- The stacking ensemble combines the predictions from individual models as input for another model (meta-model). The predictions from the training period were used to train the meta-model. In this study, multiple linear regression (LR) and support vector regression (SVR) were employed as meta-models.

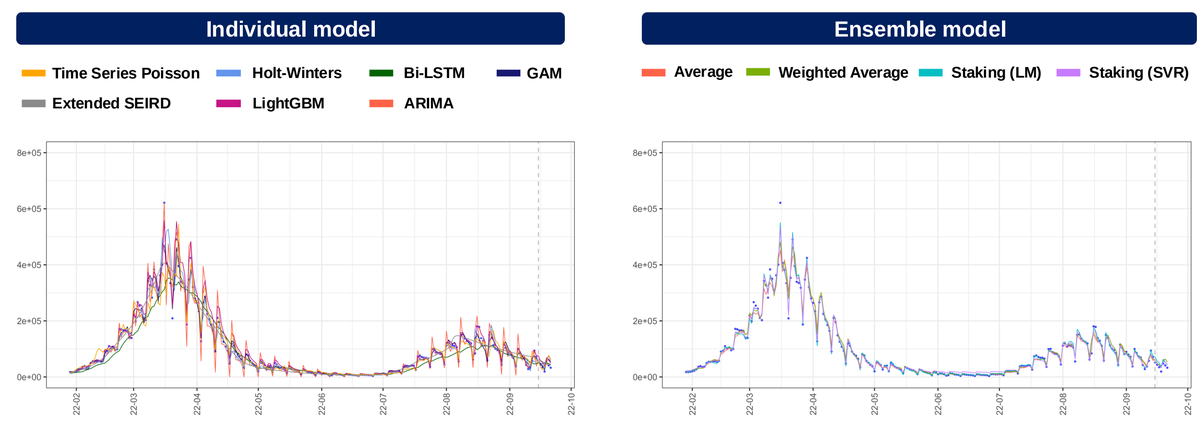


<Figure 1> Stacking ensemble
This figure shows the structure of stacking method. As a meta-model, LR and SVR were considered in this study.

Results

Forecasting results

- Six different response variables (raw / smoothed daily confirmed cases, daily confirmed deaths, and ICU patients) were forecasted using seven prediction models and four ensemble models.



<Figure 2> Forecasting results for raw daily confirmed cases
This figure shows the forecasting results (lines) and observed values (blue points) for the raw daily confirmed cases of seven individual models (left) and four ensemble models (right).

Comparison of forecasting results

- For the raw data, 10 out of the top 15 best performing models were Stacking Ensemble (SVM), and 23 out of the top 30 first- and second-best models were ensemble models.
- In contrast, for the smoothed data, less than half of the first- and second-best models were ensemble models (13 out of 30, data not shown). Simple time series models, such as Holt-Winters and ARIMA, also performed well.

Response variable	Error measure	First best model	Second best model
Raw Daily Confirmed Cases	WMAPE	Stacking Ensemble (SVM)	GAM
	MAPE	GAM	Stacking Ensemble (SVM)
	MSE	Stacking Ensemble (SVM)	GAM
	RMSE	Stacking Ensemble (SVM)	GAM
	r^2	Stacking Ensemble (SVM)	LightGBM
Raw Daily Confirmed Deaths	WMAPE	Stacking Ensemble (SVM)	Weighted Average Ensemble
	MAPE	Stacking Ensemble (SVM)	Average Ensemble
	MSE	Stacking Ensemble (SVM)	Average Ensemble
	RMSE	Stacking Ensemble (SVM)	Average Ensemble
	r^2	GAM	Time Series Poisson
Raw Daily ICU Patients	WMAPE	Stacking Ensemble (SVM)	Time Series Poisson
	MAPE	Stacking Ensemble (SVM)	Time Series Poisson
	MSE	Holt-Winters	Stacking Ensemble (SVM)
	RMSE	Holt-Winters	Stacking Ensemble (SVM)
	r^2	Holt-Winters	Average Ensemble

<Table 1> Best performing models for raw data
The table shows the first- and second-best performing models for predicting each response variable using five different error measures. Ensemble models are highlighted in blue

Conclusion

- We applied four ensemble models to predict the COVID-19 epidemic in Korea by combining the prediction results of seven individual models. Both raw and smoothed data were considered.
- While simple statistical models (e.g. Holts-Winters model, ARIMA model) performed as well as ensemble models on smoothed data, ensemble models outperformed individual models on raw data. This suggests that ensemble models are particularly effective for forecasting more complex data.

Reference

1. OWID COVID-19 Data. 2023 [cited 2022 December 21]; Available from: https://github.com/owid/covid-19-data/tree/master/public/data.
2. Korea COVID-19 Dashboard 2023 [cited 2022 December 21]; Available from: http://ncov.mohw.go.kr/en/.

Funding

This research was supported by the Bio and Medical Technology Development Program of the National Research Foundation (NRF) funded by the Korean government (MSIT) (No. 2021M3E5E3081425).