Infectious Disease Modelling 7 (2022) 419–429

Contents lists available at ScienceDirect

Infectious Disease Modelling

journal homepage: www.keaipublishing.com/idm

Reaction to the COVID-19 pandemic in Seoul with biostatistics

Seungpil Jung [a], Seung-Sik Hwang [a], Kyoung-Nam Kim [b], Woojoo Lee [a, *]

[a] Department of Public Health Sciences, Graduate School of Public Health, Seoul National University, 08826, Republic of Korea
[b] Department of Preventive Medicine and Public Health, Ajou University School of Medicine, Suwon, 16499, Republic of Korea

ARTICLE INFO

Article history:
Received 26 January 2022
Received in revised form 15 June 2022
Accepted 24 June 2022
Available online 8 July 2022
Handling Editor: Dr HE DAIHAI HE

Keywords:
Count time series model
COVID-19
Endemic-epidemic model

ABSTRACT

This paper discusses our collaboration work with government officers in the health department of Seoul during the COVID-19 pandemic. First, we focus on short-term forecasting for the number of new confirmed cases and severe cases. Second, we focus on understanding how much of the current infections has been affected by external influx from neighborhood areas or internal transmission within the area. This understanding may be important because it is linked to the government policy determining non-pharmaceutical interventions. To obtain the decomposition of the effect, districts of Seoul should be considered simultaneously, and multivariate time series models are used. Third, we focus on predicting the number of new weekly confirmed cases for each district in Seoul. This detailed prediction may be important to the government policy on resource allocation. We consider an ensemble method to overcome poor prediction performance of simple models. This paper presents the methodological details and analysis results of the study.

1. Introduction

The COVID-19 pandemic has a huge impact on the daily lives of humans worldwide (Park et al., 2021; World Health Organization, 2020). In daily life, people on a commuter train wear mask and use hand sanitizer almost every time they move. The tourism industry (or the travel industry) has collapsed in 2020 and 2021 due to the sharp decline in the number of travelers. Telecommuting has become commonplace in many companies and university classes have been replaced by online education. In South Korea, as of August 6, 2021, there are 207,406 confirmed cases and 11,951,652 tests have been performed. The positivity rate, which is defined as positive tests divided by total number of tests $\times$ 100, is approximately 1.8%. These numbers are updated every day at http://ncov.mohw.go.kr/en/.

When COVID-19 begins to spread, it is very effective to do contact tracing and isolate people in close contact with an infected person (Lee et al., 2020). It is possible to secure time to understand the characteristics of COVID-19 by preventing the rapid spread of the infectious disease. This test–trace–isolate strategy has received attention from several researchers (Dighe et al., 2020). With the development of effective vaccines, it is important to consider non-pharmaceutical interventions (NPIs) as a device to mitigate the pandemic situation (Bo et al., 2021). Various measures such as social distancing and school closing were considered to reduce social contact. To understand the effect of the NPIs, mechanism-based epidemic models have been

* Corresponding author.
E-mail address: lwj221@gmail.com (W. Lee).
Peer review under responsibility of KeAi Communications Co., Ltd.

https://doi.org/10.1016/j.idm.2022.06.009

widely used in many researches. The Susceptible-Exposed-Infectious-Recovered (SEIR) model and its modified versions have been used for the purpose (Chang et al., 2021). Once the vaccine is developed, it is important to determine who will receive the vaccine first while considering the social contact structure and mortality. Under the various circumstances of this pandemic, it is important to keep a quick communication channel between policy makers in government and infectious disease experts including health professionals, medical doctors, and biostatisticians (Cramer et al., 2021).

From the perspective of the COVID-19 quarantine policy, the government needs to ensure adequate number of beds in hospitals to avoid medical collapse, because severe cases require oxygen treatment. In this study, we convey what we have done in collaboration with the government officers in the health department of Seoul during the COVID-19 pandemic. Among a wide range of cooperative activities, we first focus on short-term forecasting for the number of new confirmed cases and severe cases. In particular, we are concerned in the one week ahead prediction for the number of new confirmed cases and severe cases using univariate time series models. Second, we focus on understanding how much of the current number of infections has been affected by external influx from neighborhood areas or internal transmission within the area. To obtain the decomposition of the effect, we should simultaneously consider the districts of Seoul so that multivariate time series models are naturally employed. This understanding may be important because it is linked to the government policy determining NPIs. Third, we focus on predicting the number of new weekly confirmed cases for each district. This detailed prediction may be important to government policy on resource allocation. We consider an ensemble method to overcome the poor prediction performance of simple models. Based on the results of these efforts, we aim to create a reference material to assist policy making in Seoul.

2. Data

Seoul is the capital city of South Korea and is a megacity with approximately 10 million people. Seoul is made up of 25 districts, called gu in Korean, and its administrative divisions are marked by a gray solid line on the map in Fig. 1. There are several wards, called dong in Korean, within each gu, and there are 426 dong marked by a gray dotted line on the map. Seoul publicly reports the COVID-19 patients in severe or critical condition (severe cases) every day and available at https://www.seoul.go.kr/coronaV/coronaStatus.do?menu_code=07. Further, Seoul Open Data Plaza provides the COVID-19 confirmed cases by gu, which is available at https://data.seoul.go.kr/dataList/OA-20470/S/1/datasetView.do. A person is defined as a new confirmed case if they are confirmed to be infected with COVID-19 according to the COVID-19 gene (PCR) test for diagnosis, regardless of the clinical aspect. A severe case is defined as when a COVID-19 patient requires inpatient treatment.

Fig. 2 shows the three count time series of Seoul starting from September 2020 to March 2021 corresponding to the third wave of the COVID-19 pandemic. The black and gray solid lines represent the raw counts of the daily number of severe cases and new confirmed cases, respectively. Because the daily number of new confirmed cases is quite volatile, its 7-day moving average is provided as a dashed line to clearly show its trend. The lag time between the two peaks seems to be between seven and fourteen days. Similarly, Zhou et al. (2020) reported 12 days (95% CI, 8–15 days) as the time from illness onset to intensive care unit admission in China. This lag is expected because some of the new confirmed cases worsen to a critical condition.

We also collect public mobility data from the Transport Operation & Information Service (TOPIS), which refers to the comprehensive traffic control center that operates and manages all traffic in Seoul. TOPIS provides origin-destination (OD) matrices representing the number of people moving from *j* dong to *i* dong through public transport on a daily basis. These OD matrices are available at https://topis.seoul.go.kr/refRoom/openRefRoom_3_4.do. Fig. 3 shows overall mobility, which corresponds to the average of all elements in the OD matrix over the past seven days, and the history of NPIs in Seoul. We can capture the sharp decline in population mobility when the South Korean government imposed social distancing (November

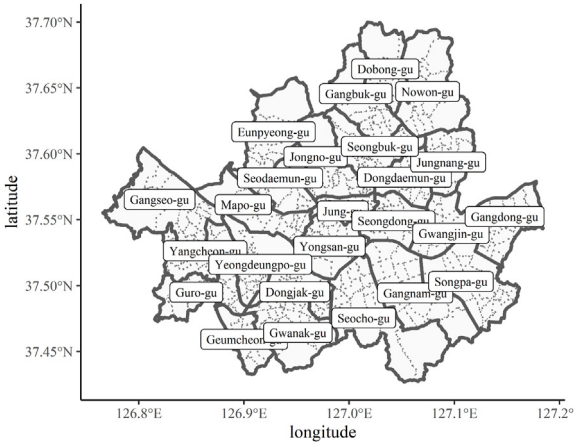

Fig. 1. Administrative map of Seoul.

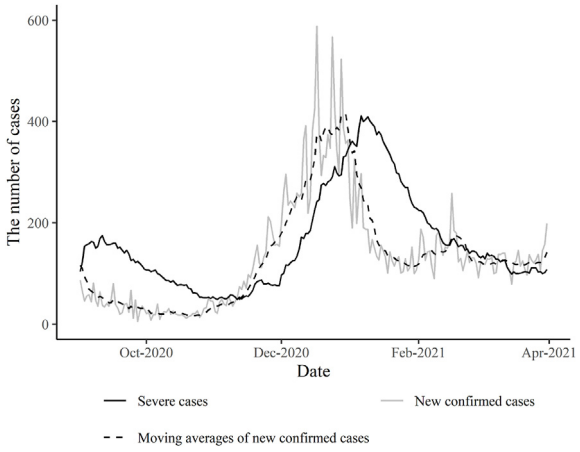

Fig. 2. Trends of new confirmed cases and severe cases.

19, November 24, and December 8, 2020) and ban on gatherings (December 23, 2020). This phenomenon reflects the effect of interventions in which various facilities with a high probability of transmission are suspended and complex amenities such as gyms and saunas are closed during the period.

3. Methods

The first aim of our analysis is to predict the daily number of new confirmed cases and severe cases for next seven days. We take univariate count time series approach to make these predictions. The second aim is to decompose the within effect (effect of internal transmission within each gu) from the between effect (effect of external influx from neighborhood areas) using multivariate time series models. The third aim is to predict the number of new confirmed cases in each gu for the next seven days. To enhance the model predictability, we apply an ensemble method.

3.1. Univariate count time series models

To provide more clarity and details about the count time series models, some notations are introduced. We denote a count time series and a time-varying r-dimensional covariate vector as $\{Y_t : t \in \mathbb{N}\}$ and $\boldsymbol{X}_t = (X_{t,1}, \ldots, X_{t,r})^T$, respectively. For example, a linear time trend and weekend effect can be included in the covariate vector $\boldsymbol{X_t}$.

Consider

$$Y_t|\mathcal{F}_{t-1} \sim \text{NB}(\mu_t, \varphi) \tag{1}$$

where NB is the negative binomial distribution, $\mathcal{F}_{t-1}$ denotes the history of the joint process $\{Y_{t-1}, \mu_{t-1}, \boldsymbol{X}_t : t \in \mathbb{N}\}$ up to $t-1$, and μ_t denotes the conditional mean of Y_t given $\mathcal{F}_{t-1}$, and φ is the dispersion parameter. Ferland et al. (2006) proposed the integer-valued GARCH (INGARCH) model of order p and q as

$$g(\mu_t) = \beta_0 + \sum_{k=1}^{p} \beta_k g(y_{t-k}) + \sum_{l=1}^{q} \alpha_l g(\mu_{t-l}) + \boldsymbol{\eta}^T \boldsymbol{X}_t \tag{2}$$

where g is a link function, and p and q are the maximum lag order of observations y_t and conditional means μ_t, respectively. The R-package "tscount" fits this model (Liboschik et al., 2017).

It is difficult to assume that the coefficients β_k and α_l are constant during the entire COVID-19 pandemic period because human behavior, society and the environment are interacting with each other continuously. As shown in Fig. 3, the government tries to respond to the pandemic situation with a timely policy such as social distancing and it has a large impact on people's mobility. This intervention changes some characteristics of the time series so that we should consider a proper length of time window to estimate parameters in univariate time series models. Therefore, we use only observations for a short period of time to estimate the coefficients in the time series models. The length of observations for estimating the coefficients is regarded as a tuning parameter and determined by cross-validation.

We also investigate how to predict the number of daily severe cases for the next seven days. A similar INGARCH model is applied to this problem, but there is a key difference between the two INGARCH models. The number of new confirmed cases has different characteristics from that of severe cases because some of the new infections develop into severe cases, but not the other way around. In addition, as shown in Fig. 2, the change in the new confirmed cases leads to a change in the severe

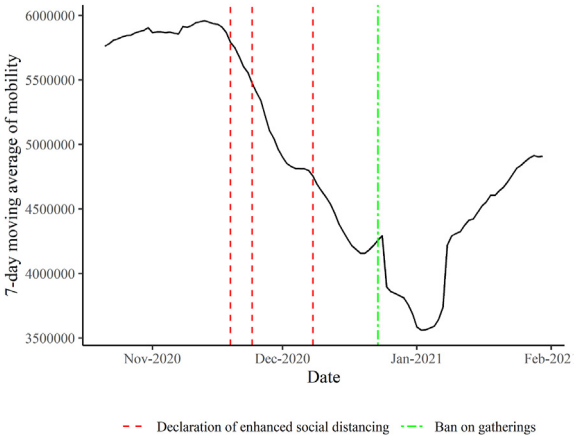

Fig. 3. Mobility trends in Seoul.

cases with a time lag. This means that the number of new confirmed cases is informative for predicting the number of severe cases, such that the former can be used as a predictor for the latter.

To assess the prediction performance for the count time series models, Czado et al. (2009) proposed the ranked probability score (RPS), which is defined as

$$\mathrm{RPS}_t(P,x) = \sum_{y=0}^{\infty} [P_t(y) - \mathbf{1}(y_t \le y)]^2 \tag{3}$$

where $P_t(y) = P(Y_t \le y|\mathcal{F}_{t-1})$ denotes the predictive distribution function. This score is averaged over the time of making the predictions. This measure is recommended because it is less sensitive to low probability events.

3.2. Endemic-epidemic model for multivariate count time series

Held et al. (2005) developed the endemic-epidemic multivariate time series models for infectious diseases. Meyer et al. (2017) implemented endemic-epidemic modeling of areal count time series in the R package "surveillance". They called the model class hhh4.

Let $Y_{i,t}$ be the number of the new confirmed cases for the ith gu and time t. Assume that given the process history up to $t-1$, $Y_{i,t}$ follow a negative binomial distribution with mean $\mu_{i,t}$ and overdispersion parameter ψ where $\mathrm{Var}(Y_{i,t}|\mathcal{F}_{t-1}) = \mu_{i,t}(1+\psi\mu_{i,t})$ and

$$\mu_{i,t} = E(Y_{i,t}|\mathcal{F}_{t-1}) = e_{i,t}\nu_{i,t} + \lambda_i y_{i,t-1} + \varphi_i \sum_{j \ne i} w_{j,i} y_{j,t-1} \tag{4}$$

where $e_{i,t}$ is the population fraction in area i at time t (offset), and $\nu_{i,t}$ can vary over time t. For example, $\nu_{i,t}$ may include first- or second-order time trends (t, t^2). λ_i and φ_i are modeled as random effects introduced to explain the heterogeneity across the areal units. $e_{i,t}\nu_{i,t}$ denotes the endemic component. Held et al. (2005) called the last two terms in (4) "epidemic component." This component is designed to capture occasional outbreaks in area i and consists of between and within components. The within component $\lambda_i y_{i,t-1}$ denotes the autoregressive effect within the area, which describes the region-specific effect of how the $t-1$th outcome affects the tth outcome. The between component $\varphi_i \sum_{j \ne i} w_{j,i} y_{j,t-1}$ explains how neighbors at the $t-1$th time affect the tth time of the region of interest through the weight matrix $w_{j,i}$. Several options are possible for $w_{j,i}$. In our application, three weight matrices are considered. First, $w_{j,i} = 1$ only when the j region shares an administrative boundary with the i region; otherwise, $w_{j,i} = 0$. Second, $w_{j,i}$ reflects the physical distance between two regions, that is, $w_{j,i}$ is proportional to the inverse of the distance between the centers of the two regions. These weights represent only the geographical information and, consequently, are time-invariant. In contrast, Paul et al. (2008) considered the inclusion of travel information to model the spatio-temporal spread of influenza or SARS. Reflecting this perspective, we use public mobility information of Seoul from TOPIS and construct a time-varying weight matrix $w_{j,i} = w_{j,i}^{(t-1)}$. The weight $w_{j,i}^{(t-1)}$ represents the normalized number of people moving from region j to region i at time $t-1$, satisfying $\sum_i w_{j,i}^{(t-1)} = 1$.

Various components of the multivariate times series models for decomposing within effect from between effect are summarized in Table 1. The best model is found to minimize the proper scoring rules using the grid search algorithm. The R-package "surveillance" implements this class of models.

Table 1
List of model for the epidemic-endemic model: **Baseline** model inputs only intercept in all components. **Population** model is an extension of the **Baseline** model, and inputs the logarithm of population in the between-epidemic component. **Trend** model is an extension of the **Baseline** model, and inputs the second-order polynomial trend in the endemic component. In addition, we consider two other components: weight, which is in the between-epidemic component, and type of intercept. To define the neighborhood, we can use spatial adjacency and traffic counts. The power-law adjacency is defined as $w_{i,j} = o_{i,j}^{-d}/\sum_k o_{i,k}^{-d}$ where $o_{i,j}$ is the order of adjacency.

Intercept	Model name	Weight1	Weight2	Weight3
Fixed	Baseline	First-order neighborhood	Power-law adjacency	Traffic counts
	Population			
	Trend			
Random	Baseline			
	Population			
	Trend			

3.3. Ensemble model for regional prediction

Our goal is to predict the number of new confirmed cases in the next week for each district. First, consider simple Poisson generalized linear models (GLMs). Assume that $Y_{i,t}|\mathcal{G}_{i,t-1} \sim \text{Poisson}(\mu_{i,t})$, where $\mathcal{G}_{i,t-1}$ is the set of predictors $\{X_i^{pop}, y_{i,1}, \ldots, y_{i,t-1}\}$. Note that the unit of t is a week in this section. Here, X_i^{pop} describes the population of the i district and is assumed to be time-invariant during the pandemic period. $Y_{i,t}$ is the total number of confirmed cases of the i district during week t. Then, the Poisson GLM is described as

$$\log(\mu_{i,t}) = \beta_0 + \beta_1 X_i^{pop} + \beta_2 \log(Z_{i,t-1} + 1) \quad i = 1, \ldots, 25 \tag{5}$$

where $Z_{i,t-1}$ describes a function of the number of new confirmed cases of the i gu during week $t-1$. Various Poisson GLMs are possible depending on the shape of $Z_{i,t-1}$. Let $z_{i,t-1}^{(k)}$ represent the number of confirmed cases on the kth day during week $t-1$. Eight models are considered in our application: $Z_{i,t-1} = z_{i,t-1}^{(k)}$ for $k = 1, \ldots, 7$ and $Z_{i,t-1} = \sum_{k=1}^{7} z_{i,t-1}^{(k)}$. In addition, to improve the prediction performance of these Poisson GLMs, we consider an ensemble approach. To combine $M(=8)$ different models, the weight vector used in Altieri et al. (2021) is considered:

$$w_{i,t+1}^m \propto \exp\left(-c(1-\mu) \sum_{k=t_0}^{t} \mu^{t-k} \ell(\hat{y}_{i,k}^m, y_{i,k}) \right), \quad \sum_{m=1}^{M} w_{i,t+1}^m = 1 \tag{6}$$

where the hyper-parameter $c \geq 0$, $t_0 \in \{1, 2, \ldots, t\}$, $\mu \in (0, 1)$, and the loss function ℓ are tuned for our data using cross-validation. The superscript m denotes that the prediction is made from the mth model, and M is the number of Poisson GLMs considered in the ensemble method. To assess the prediction performance of these parameters, we use the RPS defined above. The detailed values in the weight vector are provided in the Result section. Using this weight vector, the final ensemble prediction value for week $t+1$ is computed as

$$\hat{y}_{i,t+1}^F = \sum_{m=1}^{M} w_{i,t+1}^m \hat{\mu}_{i,t+1}^m. \tag{7}$$

4. Results

We show how to select the best model and explain its prediction performance during the considered pandemic period in Seoul.

4.1. Univariate time series results for Seoul

Many combinations of model components in Table 2 are considered to find the best prediction model for the new confirmed cases: the type of link function $g(\cdot)$, use of the week indicator, length of samples to estimate the coefficients in the INGARCH model, and various p (how many lagged outcomes should be used) and q (how many lagged conditional means should be used). The details of the INGARCH settings are provided in Table 2. The weekend indicator is 1 for Sunday and Monday, and otherwise = 0, because the number of confirmed cases is reported one day later. For each combination of these settings, we evaluate its prediction performance using the proper scoring rule and select the best model minimizing the scoring rule. The ranked probability score values are provided in Table S2 and S3 in the supplementary material.

The best model for predicting the number of new confirmed cases is INGARCH(2,0) with the identity link function using 7 observations:

Table 2
Model settings for predicting the number of new confirmed cases (upper panel) and severe cases (lower panel). The linear time trend term ($\eta_1 t$) is included in all the model settings.

Target	Components	Parameter	Settings	Selected
New confirmed case	INGARCH options	Lagged observations	0, 1, 2	2
		Lagged conditional means	0, 1, 2	0
		Link function	Identity, Log	Identity
		Sample size	7, 14, 21	7
	Covariate	Weekend indicator	No, Yes	No
Severe cases	INGARCH options	Lagged observations	0, 1, 2	1
		Lagged conditional means	0, 1, 2	0
		Link function	Identity, Log	Log
		Sample size	7, 14, 21	21
	Covariate	Weekend indicator	No, Yes	Yes
		Moving average of confirmed cases	0 (No), 7, 14	14
		Lagged confirmed cases	0 (No), 7, 14	7

$$\mu_t = \beta_0 + \beta_1 y_{t-1} + \beta_2 y_{t-2} + \eta_1 t \tag{8}$$

The regression coefficients are not given as specific numbers because they vary over time. Figs. 4 and 5 show the prediction result of new confirmed cases and severe cases, respectively. In both figures, the black and the blue dots represent the observed and predicted values, respectively, and the blue vertical line represents the 95% prediction interval (PI). Looking at Fig. 4, the prediction performance is quite remarkable and its coverage rate of 95% bootstrap PI and prediction root-mean-square-error (RMSE) are 95.92% and 57.86, respectively.

Similar combinations of model components in Table 2 are considered to find the best prediction model for the severe cases. A notable difference is that the moving average of the new confirmed cases in the past days is added as a covariate. The best model for predicting the number of severe cases is INGARCH(1,0) with the log link described as

$$\log(\mu_t) = \beta_0 + \beta_1 \log(y_{t-1} + 1) + \eta_1 t + \eta_2 W_t + \eta_3 \tilde{C}_{t-7} \tag{9}$$

where W_t is the indicator of weekend, $\tilde{C}_t = \frac{1}{14}\sum_{i=0}^{13} \log(C_{t-i})$ denotes the 14-days moving average of the number of new confirmed cases on a logarithm scale, and our model uses 21 observations to make the predictions. Looking at Fig. 5 with gray vertical dotted lines to separate weeks, except for the last week of December 2020 and the second week of January 2021, the prediction performance is acceptable. The coverage rate of 95% bootstrap PI is 83.7% and the prediction RMSE is 31.16. The inconsistencies between the predictions and the actual observations during the last week of December 2020 and the second week of January 2021 are partially explained using the results of NPIs in Seoul. Fig. 5 shows the timings of social distancing (December 8, 2020) and ban on gatherings (December 23, 2020) by red and green lines, respectively. As shown in Fig. 2, the effects of NPIs on the number of severe cases are visible after two or three weeks. Our model's prediction may show some

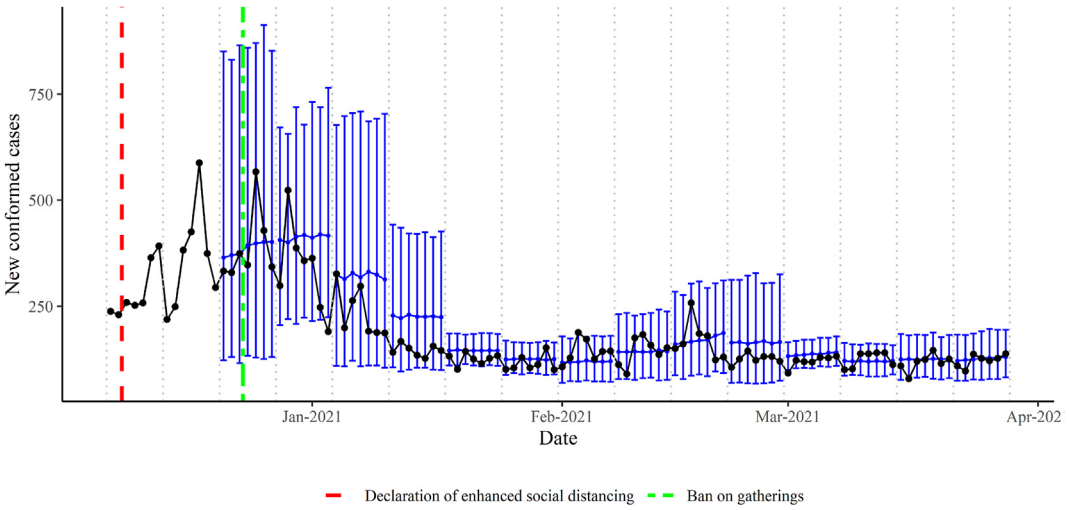

Fig. 4. Prediction of new confirmed cases.

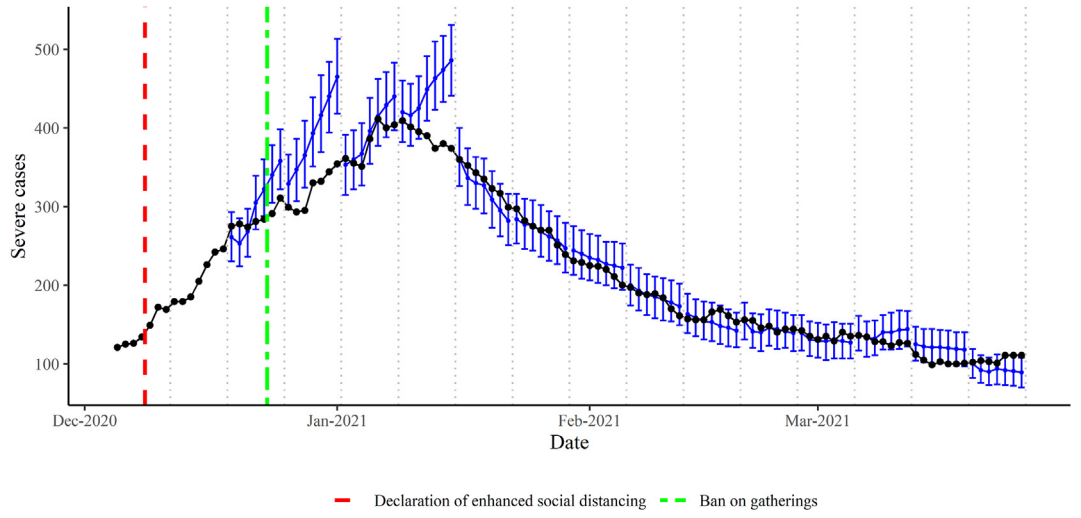

Fig. 5. Prediction of severe cases.

mismatches between actual and predicted values if only observations before the effect of NPIs were used, because it uses 21 days as its time window.

4.2. Within vs between epidemic model results

For areal units $i = 1, \ldots, I$ and time: $t = 1, \ldots, T$, the best model is described as

$$\mu_{i,t} = e_i \nu_{i,t} + \lambda_i y_{i,t-1} + \varphi_i \sum_{j \neq i} w_{j,i}^{(t-1)} y_{j,t-1}, \quad \nu_{i,t}, \lambda_i, \varphi_i > 0 \tag{10}$$

where $\log(\nu_{i,t}) = \alpha_i^{(\nu)} + \gamma_1^{(\nu)} t + \gamma_2^{(\nu)} t^2$, $\log(\lambda_i) = \alpha_i^{(\lambda)}$, and $\log(\varphi_i) = \alpha_i^{(\varphi)}$. The weight $w_{j,i}^{(t-1)}$ is the normalized traffic count of public transportation, including subway and bus from j gu to i gu at time $t-1$, and it satisfies $\sum_i w_{j,i}^{(t-1)} = 1$. Here, $\alpha_i^{(\nu)}, \alpha_i^{(\lambda)}$, and $\alpha_i^{(\varphi)}$ are unit-specific random intercepts.

The endemic-epidemic multivariate time series model (4) provides the within and between effects for every gu. Fig. 6 shows the contributions of the between and within effects of the best model, and they are defined as the differences between the within and between effect:

$$D_i = \sum_{k=1}^{d} \lambda_i y_{i,k} - \sum_{k=1}^{d} \left\{ \varphi_i \sum_{j \neq i} w_{j,i}^{(k)} y_{j,k} \right\} \tag{11}$$

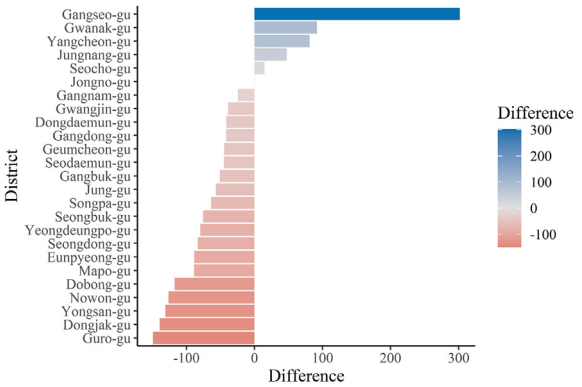

Fig. 6. Difference between effect and within effect for each gu.

where d represents the last day of the data set. The three regions (gu) with the smallest D_i are Guro-gu, Dongjak-gu, and Yongsan-gu. In other words, these regions have the largest between effects compared to their corresponding within effects. The three regions are considered to have such a characteristic because they have major transport centers with subway transfer stations. In contrast, Gangseo-gu, Gwanak-gu, and Yangcheon-gu show the largest D_i. This observation can be partially explained because they are known as typical residential areas. Fig. 7 shows the relative contributions of the between effect $\varphi_i \sum_{j \neq i} w_{j,i}^{(t-1)} y_{j,t-1}$, the within effect $\lambda_i y_{i,t-1}$, and the endemic component $e_i \nu_{i,t}$ for each gu, which are marked in orange, blue and gray colors, respectively. For example, the within effect is overwhelming in Gangseo-gu and this observation is consistent with the largest D_i. Considering that the two blue spikes are associated with the two reported mass infection cases of Gangseo-gu, they seem to be affected by internal transmission within the area rather than external influx from

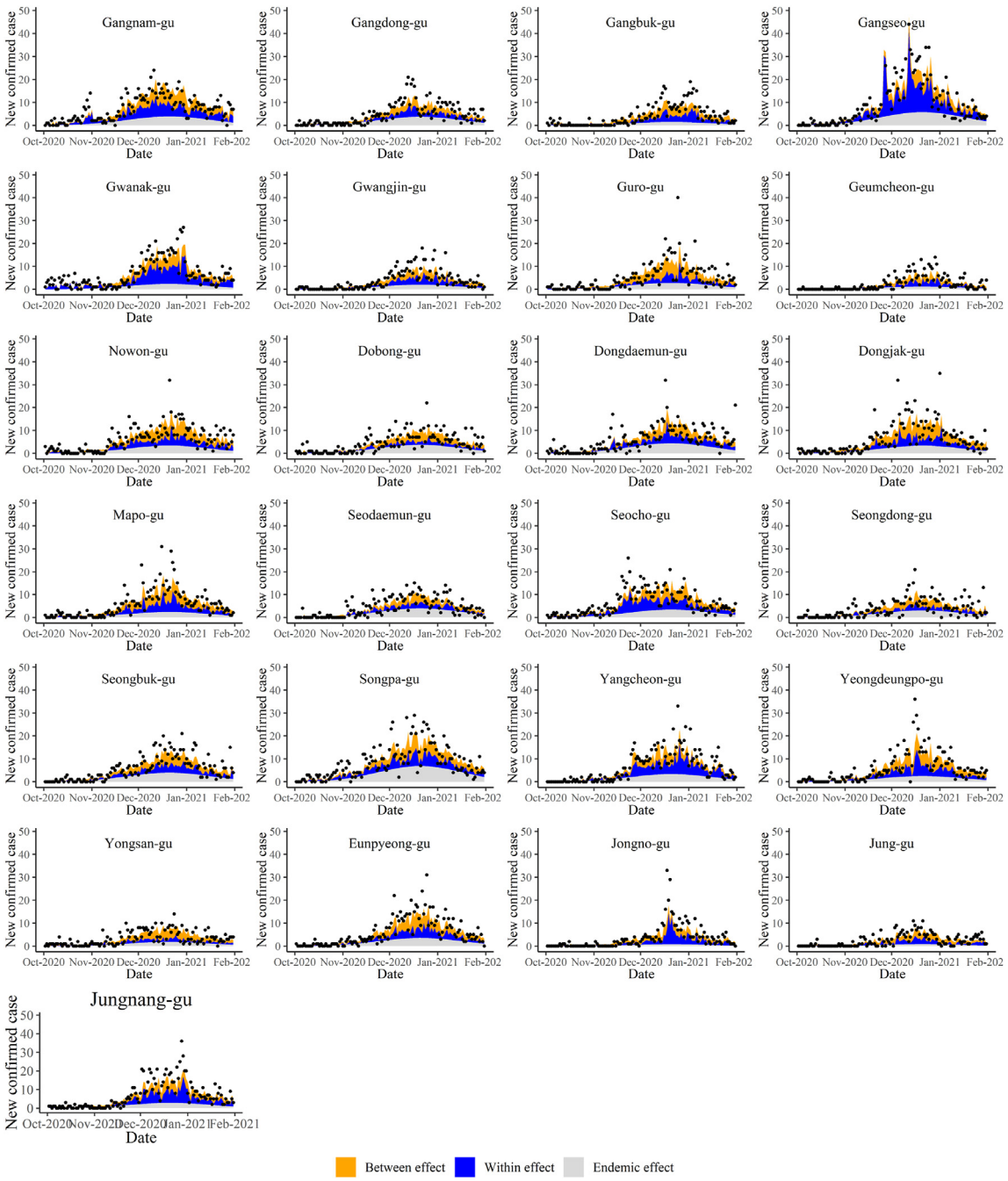

Fig. 7. Contributions of between(orange), within(blue), and endemic(gray) components for each gu.

neighborhood areas. In contrast, the between effect is overwhelming in Guro-gu and this observation is consistent with the smallest D_i. The reported mass infection case of Guro-gu seems to be affected by external influx from neighborhood areas, but a rigorous epidemiological investigation is needed to draw confirmatory conclusions.

4.3. Spatial prediction results for each gu

The numerical details of the weight vector used in the ensemble model are summarized in Table 3. Following Altieri et al. (2021), the parameter c is fixed at 1.

To combine the eight models described in the Method section while minimizing the proper scoring rule, the following weight

$$w_{i,t+1}^{m} = \exp\left(-0.6 \sum_{k=t-2}^{t} (0.4)^{t-k} \left| \hat{y}_{i,k}^{m} - y_{i,k} \right| \right) \tag{12}$$

is used, where $\hat{y}_{i,k}^{m}$ is the prediction of the weekly new confirmed cases of i gu which starts at the time k from model m. This form reflects that a good prediction model has a larger weight.

Prediction results for each gu is shown in Fig. 8. The x-axis shows the predicted values obtained from the past observations and the y-axis shows the actual values. Fig. 8 show high correlations, but the slopes are often greater or less than one. The 25th, 50th and 75th percentiles of the slopes are 0.547, 0.751 and 1.150, respectively. The maximum slope is 3.161 in mid-February, which represents an occasional outbreak as shown in Fig. 2. Prediction results for each gu are further examined when the selected model is fitted to the updated data from November to December 2021. There is no change in the qualitative conclusion. See Fig. S6 in the supplementary material.

5. Discussion

We have considered the following four tasks to create a reference material to assist policy making in Seoul: (1) predicting the daily number of new confirmed cases in the next seven days, (2) predicting the daily number of severe cases in the next seven days, (3) assessing between and within effects using mobility information, and (4) predicting the number of new confirmed cases for the next week by gu.

Our finding has epidemiological implications. First, an approximately 10-days lag exists between the new confirmed cases and the severe cases. Government should be careful to ensure that an adequate time has been secured to manage the increasing number of severe cases, which requires beds in hospitals. Therefore, considering this information on the time lag, the government can plan when there is a rapid increase in the number of new confirmed cases. Second, the current univariate time series models show an accurate prediction performance. This finding holds even when the selected models are fitted to the updated data until December 2021 (new confirmed cases) or April 15, 2021 (severe cases). See Fig. S1-S3 in the supplementary material. Therefore, under the premise of a recent major social intervention, the big discrepancy between prediction and actual values may be interpreted as the effect of the corresponding non-pharmaceutical strategy. Using accurate prediction models, we enhance our understanding about whether NPIs are effective by comparing predictions and actual values. However, it is difficult to isolate the effect of NPIs because the effect of a policy appears with a time lag and the effect of a subsequent policy overlaps. Third, looking at the decomposed between and within effects, we gain some epidemiological insights to reduce the rapid spread of the infectious disease in each region. For example, if several districts (gu) showing high between effects share a common commercial area as a source of external influx, it seems desirable to have a policy that limits the operating hours and size of meetings in the relevant commercial area. However, the relative contributions of the decomposed between and within effects seem to rely on the research period so that such epidemiological insights should be interpreted carefully. See Fig. S4 and S5 in the supplementary material.

Our study also has some limitations. First, the new confirmed cases are officially registered at the place mentioned in the resident register. Obviously, the risk of infection at the current address will not match the risk at the location where the actual

Table 3
List of parameters for the weight vector in our ensemble model.

Parameter	Settings	Selected
μ	0.1, 0.2, …, 0.9	0.4
t_0	t, $t-1$, $t-2$	$t-2$
ℓ	$\left\|\hat{y}_{i,k}^{m} - y_{i,k}\right\|$	$\left\|\hat{y}_{i,k}^{m} - y_{i,k}\right\|$
	$\left\|\sqrt{\hat{y}_{i,k}^{m}} - \sqrt{y_{i,k}}\right\|$	
	$\left\|\log(\hat{y}_{i,k}^{m}+1) - \log(y_{i,k}+1)\right\|$	

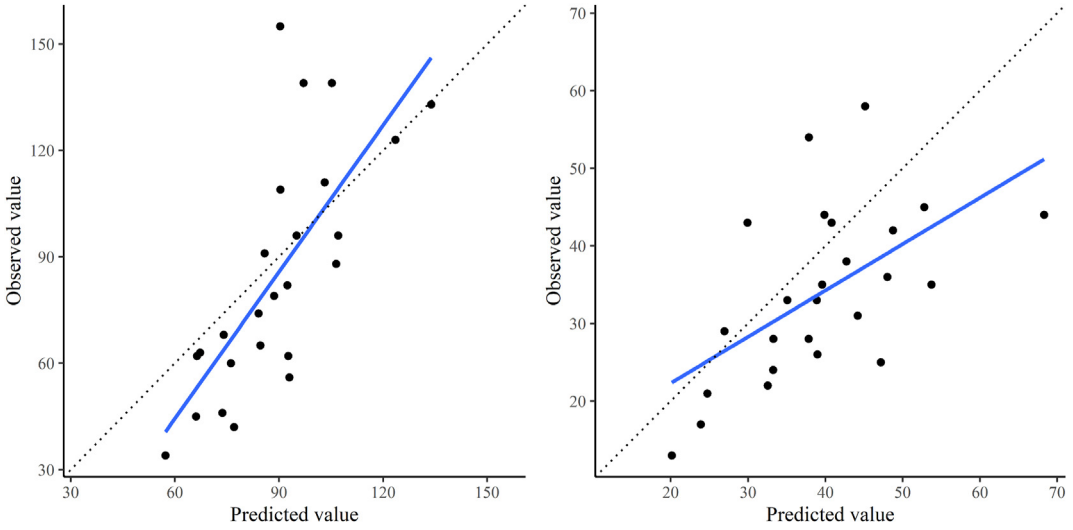

Fig. 8. Prediction of confirmed cases in each gu.

infection occurred. Therefore, especially for predictions by district, we think that infection risk assessment based on address will have a different actual risk. Second, in this collaboration work, we used empirical time-series models and ensemble models, which do not have a mechanistic background. In contrast, as a mechanism-based compartment model using differential equations, the SEIR model has been widely used in COVID-19 research. This approach has a strong advantage in studying the hypothetical situation under various intervention scenarios. For example, Chang et al. (2021) predicts the number of confirmed cases when population mobility has significantly decreased through NPIs such as social distancing. Foy et al. (2021) predicts how this pandemic situation will progress if vaccines are given to people over the age of 60 or middle-aged people in their 40s–60s. Considering this advantage of the SEIR models, our empirical modeling has difficulty in studying such hypothetical scenarios because parameters in empirical models often do not have a mechanistic background; therefore, it is difficult to link them with NPIs in a quantitative way. Third, a further calibration study is necessary to predict the new confirmed cases in each gu. By plotting the predicted value on the *x*-axis and the actual value on *y*-axis, the fitted line deviated from the identity line. It is an interesting research topic to identify regional characteristics that influence the deviation. In addition, a recent study by Oh et al. (2021) pointed out that mobility restrictions were associated with reductions in COVID-19 incidence early in the pandemic. This means that the NPIs can have different effect sizes depending on how tired people are with the policy. How to incorporate this observation in our empirical modeling requires further study.

Declaration of competing interest

The authors declare no competing interests.

Acknowledgements

This study was exempted from review by the Institutional Review Board (IRB) of Seoul National University (SNU IRB no.21-08-109) because the data were aggregated and anonymized. This work was supported by the National Research Foundation of Korea (BK21 Center for Integrative Response to Health Disasters, Graduate School of Public Health, Seoul National University) (NO.419 999 0514025). Woojoo Lee was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (no. 2021R1A2C1014409).

Appendix A. Supplementary data

Supplementary data related to this article can be found at https://doi.org/10.1016/j.idm.2022.06.009.

References

Altieri, N., Barter, R. L., Duncan, J., Dwivedi, R., Kumbier, K., Li, X., Netzorg, R., Park, B., Singh, C., Tan, Y. S., Tang, T., Wang, Y., Zhang, C., & Yu, B. (2021). Curating a COVID-19 data repository and forecasting county-level death counts in the United States. *Harvard Data Science Review*. https://doi.org/10.1162/99608f92.1d4e0dae

Bo, Y., Guo, C., Lin, C., Zeng, Y., Li, H. B., Zhang, Y., Hossain, M. S., Chan, J. W. M., Yeung, D. W., Kwok, K. O., Wong, S. Y. S., Lau, A. K. H., & Lao, X. Q. (2021). Effectiveness of non-pharmaceutical interventions on COVID-19 transmission in 190 countries from 23 January to 13 April 2020. *International Journal of Infectious Diseases, 102*, 247–253. https://doi.org/10.1016/j.ijid.2020.10.066

Chang, S., Pierson, E., Koh, P. W., Gerardin, J., Redbird, B., Grusky, D., & Leskovec, J. (2021). Mobility network models of COVID-19 explain inequities and inform reopening. *Nature, 589*(7840), 82–87. https://doi.org/10.1038/s41586-020-2923-3

Cramer, E. Y., Ray, E. L., Lopez, V. K., Bracher, J., Brennen, A., Castro Rivadeneira, A. J., Gerding, A., Gneiting, T., House, K. H., Huang, Y., Jayawardena, D., Kanji, A. H., Khandelwal, A., Le, K., Mühlemann, A., Niemi, J., Shah, A., Stark, A., Wang, Y., … Reich, N. G. (2021). *Evaluation of individual and ensemble probabilistic forecasts of COVID-19 mortality in the US*. medRxiv. https://doi.org/10.1101/2021.02.03.21250974, 2002.2003.21250974, 2021.

Czado, C., Gneiting, T., & Held, L. (2009). Predictive model assessment for count data. *Biometrics, 65*(4), 1254–1261. https://doi.org/10.1111/j.1541-0420.2009.01191.x

Dighe, A., Cattarino, L., Cuomo-Dannenburg, G., Skarp, J., Imai, N., Bhatia, S., Gaythorpe, K. A. M., Ainslie, K. E. C., Baguelin, M., Bhatt, S., Boonyasiri, A., Brazeau, N. F., Cooper, L. V., Coupland, H., Cucunuba, Z., Dorigatti, I., Eales, O. D., van Elsland, S. L., FitzJohn, R. G., … Riley, S. (2020). Response to COVID-19 in South Korea and implications for lifting stringent interventions. *BMC Medicine, 18*(1), 321. https://doi.org/10.1186/s12916-020-01791-8

Ferland, R., Latour, A., & Oraichi, D. (2006). Integer-valued GARCH process. *Journal of Time Series Analysis, 27*(6), 923–942. https://doi.org/10.1111/j.1467-9892.2006.00496.x

Foy, B. H., Wahl, B., Mehta, K., Shet, A., Menon, G. I., & Britto, C. (2021). Comparing COVID-19 vaccine allocation strategies in India: A mathematical modelling study. *International Journal of Infectious Diseases, 103*, 431–438. https://doi.org/10.1016/j.ijid.2020.12.075

Held, L., Hőhle, M., & Hofmann, M. (2005). A statistical framework for the analysis of multivariate infectious disease surveillance counts. *Statistical Modelling, 5*(3), 187–199. https://doi.org/10.1191/1471082X05st098oa

Lee, S. W., Yuh, W. T., Yang, J. M., Cho, Y. S., Yoo, I. K., Koh, H. Y., Marshall, D., Oh, D., Ha, E. K., Han, M. Y., & Yon, D. K. (2020). Nationwide results of COVID-19 contact tracing in South Korea: Individual participant data from an epidemiological survey. *Jmir Medical Informatics, 8*(8). https://doi.org/10.2196/20992

Liboschik, T., Fokianos, K., & Fried, R. (2017). tscount: An R package for analysis of count time series following generalized linear models. *Journal of Statistical Software, 82*(5). https://doi.org/10.18637/jss.v082.i05

Meyer, S., Held, L., & Hőhle, M. (2017). Spatio-temporal analysis of epidemic phenomena using the R package surveillance. *Journal of Statistical Software, 77*(11). https://doi.org/10.18637/jss.v077.i11

Oh, J., Lee, H. Y., Khuong, Q. L., Markuns, J. F., Bullen, C., Barrios, O. E. A., Hwang, S. S., Suh, Y. S., McCool, J., Kachur, S. P., Chan, C. C., Kwon, S., Kondo, N., Hoang, V., Moon, J. R., Rostila, M., Norheim, O. F., You, M., Withers, M., … Gostin, L. O. (2021). Mobility restrictions were associated with reductions in COVID-19 incidence early in the pandemic: Evidence from a real-time evaluation in 34 countries. *Scientific Reports, 11*(1). https://doi.org/10.1038/s41598-021-92766-z

Park, K. H., Kim, A. R., Yang, M. A., Lim, S. J., & Park, J. H. (2021). Impact of the COVID-19 pandemic on the lifestyle, mental health, and quality of life of adults in South Korea. *PLoS One, 16*(2). https://doi.org/10.1371/journal.pone.0247970

Paul, M., Held, L., & Toschke, A. M. (2008). Multivariate modelling of infectious disease surveillance data. *Statistics in Medicine, 27*(29), 6250–6267. https://doi.org/10.1002/sim.3440

World Health Organization. (2020). WHO Director-General's opening remarks at the media briefing on COVID-19 - 11 March 2020. https://www.who.int/director-general/speeches/detail/who-director-general-s-opening-remarks-at-the-media-briefing-on-covid-19-11-march-2020.

Zhou, F., Yu, T., Du, R., Fan, G., Liu, Y., Liu, Z., Xiang, J., Wang, Y., Song, B., Gu, X., Guan, L., Wei, Y., Li, H., Wu, X., Xu, J., Tu, S., Zhang, Y., Chen, H., & Cao, B. (2020). Clinical course and risk factors for mortality of adult inpatients with COVID-19 in Wuhan, China: A retrospective cohort study. *Lancet, 395*(10229), 1054–1062. https://doi.org/10.1016/S0140-6736(20)30566-3