Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID-19. The COVID-19 resource centre is hosted on Elsevier Connect, the company's public news and information website.

Elsevier hereby grants permission to make all its COVID-19-related research that is available on the COVID-19 resource centre - including this research content - immediately available in PubMed Central and other publicly funded repositories, such as the WHO COVID database with rights for unrestricted research re-use and analyses in any form or by any means with acknowledgement of the original source. These permissions are granted for free by Elsevier for as long as the COVID-19 resource centre remains active.
Modeling and forecasting the COVID-19 pandemic with heterogeneous autoregression approaches: South Korea

Eunjoo Hwang*, SeongMin Yu
Department of Applied Statistics, Gachon University, South Korea

Abstract

This paper deals with time series analysis for COVID-19 in South Korea. We adopt heterogeneous autoregressive (HAR) time series models and discuss the statistical inference for various COVID-19 data. Seven data sets such as cumulative confirmed (CC) case, cumulative recovered (CR) case and cumulative death (CD) case as well as recovery rate, fatality rate and infection rates for 14 and 21 days are handled for the statistical analysis. In the HAR models, model selections of orders are conducted by evaluating root mean square error (RMSE) and mean absolute error (MAE) as well as $R^2$, AIC, and BIC. As a result of estimation, we provide coefficients estimates, standard errors and 95% confidence intervals in the HAR models. Our results report that fitted values via the HAR models are not only well-matched with the real cumulative cases but also differenced values from the fitted HAR models are well-matched with real daily cases. Additionally, because the CC and the CD cases are strongly correlated, we use a bivariate HAR model for the two data sets. Out-of-sample forecasting is carried out with the COVID-19 data sets to obtain multi-step ahead predicted values and 95% prediction intervals. As for the forecasting performances, four accuracy measures such as RMSE, MAE, mean absolute percentage error (MAPE) and root relative square error (RRSE) are evaluated. Contributions of this work are three folds: First, it is shown that the HAR models fit well to cumulative numbers of the COVID-19 data along with good criterion results. Second, a variety of analysis are studied for the COVID-19 series: confirmed, recovered, death cases, as well as the related rates. Third, forecast accuracy measures are evaluated as small values of errors, and thus it is concluded that the HAR model provides a good prediction model for the COVID-19.

Introduction

The coronavirus 2019 (COVID-19) emerged in December 2019 has been seriously threatening the life of humans. The World Health Organization (WHO) declared the outbreak of COVID-19 as a pandemic in March 2020. Not only have governments around world tried to control the spread of the virus but also medical institutions are committed to finding vaccines and treatments. Many researchers in all fields are trying to support the health system to prevent the disaster of infection as well. Nevertheless, as of May 1, 2021, the COVID-19 pandemic has infected more than 151 million of the humans all over the world and caused 3 million deaths.

As one of the academia attempts to overcome the crisis, mathematical modeling and forecasting on the COVID-19 cases have been extensively carried out by statisticians and health scientists. The earliest work on the time series model for the COVID-19 analysis has been done in February 2020 by [1], who performed autoregressive integrated moving average (ARIMA) model prediction for the prevalence and incidence of COVID-19. The ARIMA model is well-known as a very powerful and effective model for epidemic diseases, for which we refer to Table 1 of [2]. The author of [2] also used the ARIMA time series models for estimation of COVID-19 in Italy, Spain and France. Many authors, for instance, [3–8] have adopted the ARIMA models for the COVID-19 analysis. [3] studied time series analysis and forecasting with the ARIMA models and other neural network techniques for main eight European countries, whereas [4] explored and compared using the ARIMA and various models of machine learning regression for short-term forecasting cumulative confirmed cases in Brazil. Also, [5] adopted the ARIMA and Prophet time series forecasting models and evaluated prediction performances for 10 mostly affected countries, where the Prophet model was proposed by [9] and used several non-linear and linear methods as components with time. [6] identified the best ARIMA model and then applied it to forecast future incidence of COVID-19 cases in India. Moreover, [7] used Box–Jenkins method to find the best ARIMA model for predicting the numbers of people infected with COVID-19 disease in Iraq. [8] presented numerous statistical analyses to forecast the future number of daily deaths and

* Corresponding author.

E-mail address: ehwang@gachon.ac.kr (E. Hwang).
infections in European and African countries, including Box-Jenkins ARIMA and fractal interpolations, fractal dimension.

As a more general or more refined probability model, \[10\] adopted an AR model with two-piece scale-mixed normal distributed errors (TP-SMN-AR) while \[11\] a TP-SMN-ARMA process in order to possess asymmetry of the distribution of the COVID-19 data. \[12\] introduced the autoregressive distributed lag (ADL) model to describe the development of the disease at the exponential phase. The ADL model allows describing non-monotonic changes in relative infection over the time, and predicting the outcomes of their decisions on public health. \[13\] proposed a mathematical model to characterize aspects of the COVID-19 pandemic in South Korea, Italy and Brazil, and compared main features of the three countries as examples of very different scenarios of the COVID-19 pandemic. \[14\] examined the pandemic for India via sensitivity analysis whereas \[15\] for China via an optimization method. \[16\] dealt with an evolution of the COVID-19 infection data with power-law growth and saturation by noticing that the total numbers of infected cases exhibits exponential growth and then power-law growth before the flattening of the curve. \[17\] proposed a new gray prediction model based on traditional Richards model and modified gray action quantity for the COVID-19 in China, Italy, Britain and Russia. \[18\] used an efficient prediction model based on Verhulst equation to analyze the COVID-19 in China, Italy and Spain.

Some efforts in the area of artificial intelligence have been conducted to analyze the COVID-19 epidemic. For instance, \[19\] predicted the number of COVID-19 daily cases in Turkey through a fuzzy rule basing system; \[20\] exploited the Gaussian process regression as a recently developed machine learning technique to predict the COVID-19 death cases and compared with artificial neural network supervised based methods in terms of their prediction abilities. \[21\] implemented several forecasting techniques such as naïve method, moving average, exponential smoothing, Holt–Winters method, ARIMA, etc. for comparison of prediction of COVID-19 worldwide cases. \[22\] proposed a new genetic programming based model for confirmed cases and death cases for India. \[23\] proposed a novel technique based on meta-heuristic GWO (gray wolf optimizer) algorithm to optimize hyper parameters for LSTM (long short term memory) network and compared with the baseline models including ARIMA model. \[24\] proposed a time series prediction model using machine learning to obtain the curve of disease and forecast the epidemic tendency. Linear regression, multilayer perception, random forest and support vector machines (SVM) machine learning methods were used. \[25\] examined the advantages of Singular Spectrum Analysis (SSA) for forecasting the number of daily confirmed cases, deaths, and recoveries caused by COVID-19. The results of V-SSA and R-SSA were compared to those from ARIMA, ARFIMA, Exponential Smoothing, TBATS, and NNAR. \[26\] studied parameter estimation and prediction of the COVID-19 epidemic using SIR/SQAIR models. \[27\] also presented a COVID-19 prediction study of artificial neural networks for the daily numbers of cases and deaths per million people, which are different from other works mentioned above. \[28--32\] investigated forecasting for the COVID-19 as well.

As for the infection fatality rates of the COVID-19, \[33--39\] discussed with the empirical results; \[33\] calculated an estimate of the infection fatality rate by dividing the number of deaths by the total infected cases from seroprevalence data. \[34\] presented empirical results on infection and fatality rates from cross-country regression and found a significant positive impact of local air pollution on the rates. \[39\] forecasted COVID-19 growth rates with statistical, epidemiological, machine-learning and deep-learning models, and a new hybrid forecasting method based on nearest neighbors and clustering. Meanwhile, \[40,41\] analyzed the COVID-19 data curve and \[42--44\] dealt with the finance and economics related to the COVID-19.

Fortunately, vaccines of the COVID-19 have been recently developed and distributed worldwide. Even though the vaccines of the COVID-19 are now available almost all over the world, numbers of the confirmed cases are not decreasing and the virus still threatens the life of the humans and the health system on earth. At the beginning, South Korea was reported to be able to control the disease very well, as \[13\] mentioned. However, as the winter came with low temperature South Korea has been facing the crisis of the COVID-19 pandemic like other countries, and even as spring comes, the number of the COVID-19 confirmed cases still remains unchanged. The ceaseless spread of the COVID-19 leads us to study the accurate prediction of the time series on the COVID-19 cases data in order to give useful information and help to determine the government policy. Accuracy of the forecasting can provide optimized decisions on the policies such as social distancing stages and preparation of health systems, ready to accommodate the expecting numbers of patients.

In this work we explore Korean COVID-19 pandemic time series such as confirmed cases, released cases, death cases, as well as their related rates, and examine the estimation and forecast by means of a popular time series model. The model we adopt is a heterogeneous autoregressive (HAR) model, which is known to be powerful for long-memory financial volatility. The COVID-19 time series data reveal the long-memory feature as seen in Fig. 1. For this reason we employ the HAR models and investigate statistical analysis on the COVID-19. Some criteria such as \[R^2\], AIC and BIC are computed in the HAR models of order \(p = 2, 3, 4\). Another merit of the HAR model on the COVID-19 is that its coefficient of determination \(R^2\) is near one for the cumulative series data. \(R^2\) is a measure of the global fit of the model and \(R^2 = 1\) implies that the relative square error is zero. In conclusion, for these two reasons: long-memory of the COVID-19 data and the \(R^2\)-value of the HAR model with the COVID-19 data, by means of the HAR models, instead of other models, we study estimation as well as forecasting on the COVID-19 cases.

We handle seven data sets such as cumulative confirmed (CC) case, cumulative recovered (CR) case, cumulative death (CD) case as well as recovery rate (RR), fatality rate (FR), infection rates (IRs) for 14 and 21 days. The rates are considered as time-varying series and defined as follows: The RR on a day is computed as the ratio of cumulative recovered case to cumulative confirmed case, and the FR is the ratio of cumulative death case to cumulative confirmed case. The FR on a day during the period of \(r\) days is computed as the proportion of confirmed case on the day to the partial sum of confirmed cases for the last consecutive \(r\) days, and we set \(r = 14\) and 21 because the incubation period of the COVID-19 is between 14 and 21 according to \[45\]. Our results report that the \(R^2\)-values of HAR models fitted with the CC, CR, CD and RR are equal to one, and the value of the FR is 0.999, whereas the \(R^2\)-values of IRs are not one, but greater than 0.8, and yet their HAR models have good performance.

Model selections of orders in the HAR models are conducted by evaluating root mean square error (RMSE) and mean absolute error (MAE) as well as \(R^2\), AIC, and BIC. As for the estimation method we use the ordinary least square estimates (OLSE) in the HAR models. Asymptotic normality of the OLSE in the HAR model was established together with a Monte-Carlo simulation study by \[46\]. Thus, we defer to \[46\] for the simulation study and omit it in this work. Estimates of the parameters in the HAR models are computed along with standard errors and 95% confidence intervals, based on the theory of \[46\]. Our results address that fitted values via the HAR models with the OLSEs are not only well-matched to the real cumulative cases data, but also differentiated values from the fitted HAR models are well-matched to real daily cases data. Indeed, the proposed model makes small values of RMSE and MAE on the COVID-19 data, and therefore it is an efficient and reliable tool to predict the COVID-19.

Additionally, because the CC and the CD cases are strongly correlated, we consider a bivariate HAR time series model to see the co-movements of a pair of the confirmed and death cases. By comparing the bivariate HAR model with the univariate HAR models, applied to the CC and the CD, we determine better prediction models for the two cases. It is demonstrated that the bivariate HAR model has smaller errors than univariate HAR models.
Out-of-sample forecasts are conducted with the seven COVID-19 data sets to compute multi-step ahead predicted values and 95% prediction intervals. Four forecasting performance measures are evaluated: RMSE, MAE, mean absolute percentage error (MAPE) and root relative square error (RRSE). Most prediction intervals include the real values except for the FR that becomes suddenly flattened after the time epoch of the out-of-sample. However, predicted values and prediction interval of the FR have reasonable pattern, though its actual values in the out-of-sample period deviate from the interval. The CC and the CD cases, applied by two models: the univariate and the bivariate, have different prediction intervals in the two models. It is because the confirmed cases have regressors of the death cases in the bivariate HAR model, and after the time epoch of the out-of-sample, the number of death cases increases very slowly, while in the univariate HAR models the two cases are modeled independently. It is also addressed that in the bivariate prediction the errors of the forecast accuracy for the CC are similar to or smaller, and all the errors for the CD are smaller than those in the univariate prediction. Consequently, as we expected, the bivariate HAR model yields better fitting-performance than the univariate models.

The main findings of this work are three folds: First, HAR models are well-fitted to cumulative numbers of COVID-19 data with a good criterion result like coefficient of determination $R^2$ being near one, which means that the relative square error is near zero. Second, this study analyzes a variety of the COVID-19 time series data such as confirmed cases as well as recovery cases, death cases, and the related rates: recovery rate, fatality rate, infection rates for 14, 21 days, for confirmed cases as well as recovery cases, death cases, and the related rates: recovery rate, fatality rate, infection rates for 14, 21 days, for which all the HAR model can be adopted with good performance. Third, forecast accuracy measures of RMSE, MAE, MAPE and RRSE are evaluated as small errors to conclude that the HAR model provides a good prediction model for the COVID-19 pandemic.

The rest of the paper is organized as follows: In Section “Heterogeneous autoregressive (HAR) models” the HAR model and its background are briefly described. In Section “Estimation” data description and estimation are explored while in Section “Forecasting” forecasting is investigated. Conclusion and discussion are given in Section “Conclusion and discussion”.

**Heterogeneous autoregressive (HAR) models**

This section introduces briefly the heterogeneous autoregressive (HAR) model, which we adopt for analysis of Korean COVID-19 series. To do this, some background is stated here.

One of the theories that explain the financial market is a heterogeneous market hypothesis. The theory assumes that investors who accept investment information have different reactions. The heterogeneous market hypothesis in stock markets is the hypothesis proposed by [47]. Dealers and traders observe the situation in stock markets that are traded in a short period of time for trading. Investors holding stocks in the form of assets over the long term are sensitive to government policy changes and are not very sensitive to the frequency of stock trading. Market participants note long-term events along with share price volatility. This is called a heterogeneous market, which is composed by the entire economy on the premise that there is a temporal heterogeneity in making investment decisions.

[47] argued that stock market volatility should be analyzed using heterogeneous market hypotheses to identify dynamic changes in the market. Studies of heterogeneous market hypotheses include the HARCH (heterogeneous autoregressive conditional heteroscedastic) model of [47] and the HAR-RV (heterogeneous autoregressive-realized volatility) model of [48]. Inspired by the heterogeneous market hypothesis and the HARCH model, [48] proposed the HAR-RV model, which is a linear autoregressive model with daily, weekly, and monthly moving averages, theoretically equivalent to an autoregressive model of order 22. The HAR-RV model uses not only the heterogeneous market hypothesis of [47] but also the realized volatility of [49]. It represents heterogeneity over daily, weekly and monthly periods. A main feature of the HAR model is long-memory. In particular, it well represents the characteristics of variability, such as long-term memory, and is highly regarded for its good predictive performance on realized volatility.

The HAR-RV model of [48] is described as follows:

$$RV_t = a_0 + a_d RV_{t-1} + a_n RV_{t-n+1} + m RV_{t-22; t-1} + \epsilon_t$$

where $RV_{t-5; t-1}$ and $RV_{t-22; t-1}$ are weekly and monthly average realized volatility, and the parameters $a_d, a_n$ and $m_a$ are coefficients for each day, week, and month, respectively.

In this work, we adopt the HAR model to investigate estimation and forecasting of Korean COVID-19 cumulative cases series and related data. We consider two types of HAR models: univariate HAR and bivariate HAR models. Cumulative cases data have long-memory features as seen in Fig. 1 and thus the HAR models are appropriately fitted to the cumulative cases. Its theoretical analysis as well as simulation study can be seen in recent works, for example, in [46]. We describe a univariate HAR model and a bivariate HAR model in the following:

A (univariate) HAR model $\{X_t, t \in \mathbb{Z}\}$ of order $p$ is given by

$$X_t = \phi_1 X_{t-1}^{(1)} + \phi_2 X_{t-1}^{(2)} + \cdots + \phi_p X_{t-1}^{(p)} + \epsilon_t$$

where $X_t^{(p)} = \frac{1}{p} (X_{t-1} + \cdots + X_{t-p})$, $i = 1, 2, \ldots, p$, with positive integers $h_1, \ldots, p$, satisfying $1 = h_1 < h_2 < \cdots < h_p$, and $\{\epsilon_t\}$ is a sequence of random variables with mean zero and variance $\sigma^2$. $\phi_1, \ldots, \phi_p$ are coefficients of the model that is estimated and it is assumed that $\sum_{i=1}^p \phi_i < 1$.

A bivariate HAR model $\{(X_t, Y_t), t \in \mathbb{Z}\}$ of orders $(p, q)$ is defined as

$$X_t = a_{11} X_{t-1}^{(1)} + \cdots + a_{1q} X_{t-1}^{(q)} + \beta_{11} Y_{t-1}^{(1)} + \cdots + \beta_{1q} Y_{t-1}^{(q)} + \epsilon_{1t}$$
$$Y_t = a_{21} X_{t-1}^{(1)} + \cdots + a_{2q} X_{t-1}^{(q)} + \beta_{21} Y_{t-1}^{(1)} + \cdots + \beta_{2q} Y_{t-1}^{(q)} + \epsilon_{2t}$$

where $Y_t^{(q)}$ is given in the same way with $X_t^{(p)}$, $i = 1, 2, \ldots, q$: $\{\epsilon_{1t}\}$ and $\{\epsilon_{2t}\}$ are independent noise processes with mean zeros. Coefficients $a_{ij}, \beta_{ij}$ are assumed to be $\sum_{i=1}^p a_{ij} + \sum_{i=1}^q \beta_{ij} < 1$ for each $j = 1, 2$. The asymptotic property of the least squares estimators in the bivariate HAR model as well as the simulation study were established by [46] along with an application to financial data. In this work we adopt the HAR models in order to analysis Korean COVID-19 pandemic time series data such as the confirmed, recovered, death cases and their related rates.

**Estimation**

**Data**

In this work, some important COVID-19 cases data are considered for analysis of estimation and forecasting. The COVID-19 data used in this paper are available in <https://coronaboard.kr/> and <http://www.ecdc.europa.eu/en/covid-19-pandemic>. For all the computations, we use Python3.8 numpy, scipy and statsmodels.tsa. Each of the data is applied to the univariate HAR time series models, respectively, and furthermore a pair of the data is applied to the bivariate HAR model. Cumulative confirmed (CC), cumulative recovered (CR) and cumulative deaths (CD) cases time series data are basically used. As the related rates, recovery rate (RR) and fatality rate (FR) as well as infection rates (IR) for $r$ days ($r = 14, 21$) are analyzed. The RR on day $r$ is given as the ratio of the CR to the CC on the day and the FR on day $r$ as the ratio of the CD to the CC.

$$RR_r = \frac{CC_r}{CC_t} \times 100 \quad FR_r = \frac{CD_r}{CC_t} \times 100$$

The IR on day $r$ is defined as the proportion of (daily) confirmed (C) case to the sum of confirmed cases for $r$ days, where $r = 14, 21$ are the incubation period chosen according to [45]:

$$IR_r(t) = \frac{C_t}{\sum_{i=t-r+1}^{t} C_t} \times 100 \quad t = 14, 21$$

The data statistics of the seven data sets $CC_t, CR_t, CD_t, RR_r, FR_r, IR_r$, and $IR_r^{(a)}, t = 14, 21$ are described in Table 1, and plots of the data and their autocorrelation functions (ACFs) are depicted in Fig. 1.
Fig. 1. Korean COVID-19 time series data and their ACFs; Cumulative Confirmed, Cumulative Recovered and Cumulative Deaths cases (first row); Recovery Rate and Fatality Rate (second row); Infection Rates for 14 days and for 21 days (third row).

Fig. 2. Cumulative Confirmed, Cumulative Recovered and Cumulative Deaths cases data and their univariate HAR fitting (first column); Daily data and differences from the HAR fittings (second column). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)
as well as simulation study was conducted in the multivariate cases by for the estimation method we use the OLSE in the HAR model (1).

$$h$$

structure

Estimation in univariate HAR models

In the HAR($p$) models, orders $p = 2, 3, 4$ are set along with lag structure $h = (h_1, \ldots, h_p) = (1, 7, 14, 1, 7, 14, 21)$, respectively. As for the estimation method we use the OLSE in the HAR model (1). Theoretic asymptotic normality of the HAR model was developed as well as simulation study was conducted in the multivariate cases by [46]. Based on the work of [46], we perform the analysis of the OLSE for the HAR model and illustrate the estimation result with standard error (se) and 95% confidence interval (CI). Suppose we observed $\{X_{t-h_1}, X_{t-h_2}, \ldots, X_{t-p}, X_{t-1}, \ldots, X_{t}\}$ where $n$ is regarded as the sample size to compute the OLSE as follows:

Model (1) is rewritten as $X_t = \phi^T X_{t-1} + \epsilon_t$ where $\phi = (\phi_1, \ldots, \phi_p)^T$ and $X_{t-1} = (X_{t-1}^{(1)}, \ldots, X_{t-1}^{(p)})^T$. The OLSE is given by

$$\hat{\phi} = (\phi_1, \ldots, \phi_p)^T = \left( \sum_{t=1}^{n} X_{t-1} X_{t-1}^T \right)^{-1} \left( \sum_{t=1}^{n} X_{t-1} X_t \right).$$

In order to compare estimations in the HAR models of orders $p = 2, 3, 4$, we use some criteria such as $R^2$, AIC (Akaike’s Information Criterion)

\begin{table}[H]
\centering
\begin{tabular}{llllll}
\hline
\textbf{Data} & \textbf{Statistics} & \textbf{Size$^a$} & \textbf{Mean} & \textbf{SD} & \textbf{Min} & \textbf{Median} & \textbf{Max} \\
\hline
Cumulative confirmed cases & $CC_t$ & 330 & 15,351.89 & 9,723.06 & 1 & 12,935.5 & 43,484  \\
Cumulative recovered cases & $CR_t$ & 330 & 12,722.35 & 8,970.52 & 0 & 11,721.5 & 32,102  \\
Cumulative death cases & $CD_t$ & 330 & 276.91 & 156.49 & 0 & 262 & 587  \\
Recovery rate & $RR_t$ & 330 & 70.24 & 31.59 & 0 & 87.227 & 93.694  \\
Fatality rate & $FR_t$ & 330 & 1.643 & 0.691 & 0 & 2.396 & 2.966  \\
\hline
\end{tabular}
\caption{Korean COVID-19 Data during in-sample period from Jan.20.2020 to Dec.14.2020; (Values in Size$^a$ indicate $n + h_p$ in the HAR($p$) models, where $n$ is the sample size, $h_p = 7, 14, 21$ for $p = 2, 3, 4$, resp.).}
\end{table}

\begin{table}[H]
\centering
\begin{tabular}{llllll}
\hline
\textbf{Data} & \textbf{RMSE} & \textbf{MAE} & $R^2$ & \textbf{AIC} & \textbf{BIC} \\
\hline
Cumulative confirmed cases & 67.310 (0.00692) & 36.658 (0.00577) & 1.00 & 3640 & 3647  \\
Cumulative recovered cases & 63.136 (0.00649) & 34.647 (0.00535) & 1.00 & 3523 & 3534  \\
Cumulative death cases & 63.016 (0.00648) & 34.747 (0.00537) & 1.00 & 3446 & 3460  \\
Recovery rate & 54.936 (0.00625) & 36.857 (0.00419) & 1.00 & 3509 & 3516  \\
Fatality rate & 55.052 (0.00626) & 37.565 (0.00428) & 1.00 & 3436 & 3448  \\
\hline
\end{tabular}
\caption{Criteria of the OLSEs in the univariate HAR($p$) model on cumulative confirmed $CC_t$, cumulative recovered $CR_t$ and cumulative deaths $CD_t$ cases.}
\end{table}

\begin{table}[H]
\centering
\begin{tabular}{llllll}
\hline
\textbf{Data} & \textbf{RMSE} & \textbf{MAE} & $R^2$ & \textbf{AIC} & \textbf{BIC} \\
\hline
Cumulative confirmed cases & 67.310 (0.00692) & 36.658 (0.00577) & 1.00 & 3640 & 3647  \\
Cumulative recovered cases & 63.136 (0.00649) & 34.647 (0.00535) & 1.00 & 3523 & 3534  \\
Cumulative death cases & 63.016 (0.00648) & 34.747 (0.00537) & 1.00 & 3446 & 3460  \\
Recovery rate & 54.936 (0.00625) & 36.857 (0.00419) & 1.00 & 3509 & 3516  \\
Fatality rate & 55.052 (0.00626) & 37.565 (0.00428) & 1.00 & 3436 & 3448  \\
\hline
\end{tabular}
\caption{Criteria of the OLSEs in the univariate HAR($p$) model on recovery rate $RR_t$ and fatality rate $FR_t$.}
\end{table}
Table 4: Estimation in bivariate HAR models

| Model      | RMSE | MAE | $R^2$ | AIC | BIC |
|------------|------|-----|-------|-----|-----|
| $I(4)$     |      |     |       |     |     |
| $p = 2$    | 4.512 (0.5953) | 1.992 (0.2651) | 0.859 | 1621 | 1629 |
| $p = 4$    | 3.136 (0.4043) | 1.485 (0.2200) | 0.880 | 1283 | 1298 |

Values in the parentheses indicate the errors divided by standard deviation. *Denotes the best.

and BIC (Bayesian Information Criterion):

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (X_i - \hat{X}_i)^2}{\sum_{i=1}^{n} (X_i - \bar{X})^2}$$

where $\hat{X}_i = \hat{\phi}_1 X_{i-1}^{(1)} + \ldots + \hat{\phi}_p X_{i-p}^{(p)}$, and $\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$.

$AIC = -2LL + 2k, \quad BIC = -2LL + k \log(n)$

where $LL$ is the log-likelihood function and $k$ is the number of parameters.

We also compute some performance measures: root mean square error (RMSE) and mean absolute error (MAE):

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^{n} e_i^2}, \quad MAE = \frac{1}{n} \sum_{i=1}^{n} |e_i|$$

where $e_i$ is the difference between the observed and estimated values: $e_i = X_i - \hat{X}_i = X_i - \hat{\phi}_1 X_{i-1}^{(1)} - \ldots - \hat{\phi}_p X_{i-p}^{(p)}$.

One of the reasons why we adopt the HAR models is that the Korean COVID-19 data fit very well in the HAR models in the sense of the coefficient of determination $R^2$ being near one, as seen in Tables 2–3.

Comparisons in terms of $R^2$ values are commonly used in regression models, for instance, [50] recently computed $R^2$ values of univariate and multivariate HAR models for several financial data sets.

Tables 2–4 report the results of performance errors as well as the criterion rules for the seven data sets: CC, CR, CD, RR, FR and $I(4)$, $\tau = 14.21$. In the tables, values in the parentheses of RMSE and MAE are the RMSE and MAE divided by standard deviation, respectively, which are equivalent to RMSE and MAE of the HAR model fitted by the normalized data, that is, the data subtracted by mean and then divided by standard deviation. The reason why we compute the RMSE and MAE of the normalized data is because data considered have all different scales. To compare the different data adopted in the HAR models, we use the standardized data that have mean zero and variance one. In most of the cases of Tables 2–4, order $p$ has the best.

In Table 2, which has the criterion results of CC, CR, CD with all $R^2$ equal to one, all errors of RMSE and MAE in the parentheses are less than 0.01. Tables 5–7, which correspond to Tables 2–4, respectively, report coefficient estimates, standard errors as well as the confidence intervals of the parameters in the HAR models for Korean COVID-19 data. Estimation results of the parameters in the HAR(4) model for the CC are, letting $X_i = CC$, written by

$$X_i = 1.4920(0.0025) X_{i-1}^{(1)} - 0.705(0.008) X_{i-1}^{(2)} + 0.284(0.050) X_{i-1}^{(2)} - 0.07(0.025) X_{i-1}^{(4)} + e_i$$

where the estimates of coefficients with standard errors in the parentheses are cited from Table 5. The 95% confidence intervals of the coefficients are given as (1.447, 1.563), (−0.816, −0.594), (0.170, 0.399) and (−0.119, −0.022) for $\hat{\phi}_1, \hat{\phi}_2, \hat{\phi}_3$, and $\hat{\phi}_4$, respectively, and all the confidence intervals are significant.

Fig. 2 depicts the HAR fitted model with order chosen in Table 2 and with estimated coefficients in Table 5 for CC, CR, CD. Moreover, their daily fitted data, which are the differences from the HAR models, are plotted together with real daily data and residuals on the right column of Fig. 2. Also, Fig. 3 shows the HAR fitted model for RR, FR, $I(4)$, $\tau = 14.21$ with order chosen in Tables 3–4 and with estimated coefficients in Tables 6–7, respectively. As shown in Figs. 2–3, where real values, fits and residuals appear in blue, red and gray, respectively, HAR models have good performances with small residuals even for the cases of RR, FR and $I(4)$, $\tau = 14.21$.

Among others, from Table 7 we report the estimation results for the $I(4)$ in the HAR(4) model as follows: for $X_i = I_t^{(4)}$, $X_i = 0.599(0.049) X_{i-1}^{(1)} + 0.241(0.077) X_{i-1}^{(2)} - 0.093(0.118) X_{i-1}^{(3)} + 0.162(0.089) X_{i-1}^{(4)} + e_i$.

The 95% confidence intervals of the four coefficients are given as (0.502, 0.696), (0.089, 0.393), (−0.325, 0.139) and (−0.012, 0.337), and the first two confidence intervals are significant. The HAR(4) model of the $I(4)$ is given by

$$X_i = 0.612(0.061) X_{i-1}^{(1)} + 0.576(0.121) X_{i-1}^{(2)} - 0.498(0.119) X_{i-1}^{(3)} + 0.204(0.070) X_{i-1}^{(4)} + e_i$$

for $X_i = I_{t-1}^{(4)}$. The 95% confidence intervals are (0.492, 0.733), (0.337, 0.814), (−0.672, −0.225) and (0.066, 0.343), and all the confidence intervals have significant result.

Estimation in bivariate HAR models

Now we consider a bivariate HAR model for bivariate data of CC and CD. The matrix form of the model (2) is written as $Z_{t} = BW_{t-1} + \epsilon_t$ where

$$Z_t = (X_t, Y_t)^T, \quad W_{t-1} = (X_{t-1}^{(1)}, X_{t-1}^{(2)}, X_{t-1}^{(3)}, X_{t-1}^{(4)})^T, \quad \epsilon_t = (\epsilon_{1t}, \epsilon_{2t})^T$$

and

$$B = \begin{bmatrix} a_{11} & \ldots & a_{1p} \\ \vdots & \ddots & \vdots \\ a_{p1} & \ldots & a_{pp} \end{bmatrix}, \quad \beta_{1} = \begin{bmatrix} \beta_{11} \\ \vdots \\ \beta_{p1} \end{bmatrix}, \quad \beta_{2} = \begin{bmatrix} \beta_{12} \\ \vdots \\ \beta_{p2} \end{bmatrix}$$

The OLSE $\hat{B}$ of $B$ is given by $\hat{B} = \arg \min_{B} \sum_{t=1}^{n} \sum_{i=1}^{p} e_{ij}^2$ and it is obtained by

$$\hat{B} = \left( \sum_{t=1}^{n} W_{t-1} W_{t-1}^T \right)^{-1} \left( \sum_{t=1}^{n} W_{t-1} Z_t^T \right)$$

Its asymptotic normality theory was established by [46] together with a simulation study in the cases of i.i.d. and correlated errors. We refer to [46] for the simulation, and thus we omit the simulation experiment in this present work. Table 8 presents the results of the criteria for the bivariate HAR($p,q$) models with some pairs of orders for two data sets $CC_t$ and $CD_t$, where we use the normalized data because the two data sets have different scales. It is also intended to compare performance of the univariate HAR models of $CC_t$ and $CD_t$ with that of the bivariate HAR model.

In the bivariate HAR model, results of coefficient estimates, standard errors and 95% confidence intervals are given in Table 9 for the model of order $(p,q) = (4, 4)$, which has the smallest RMSE among the considered models. The bivariate HAR(4,4) model for $X_t = CC_t$ and $Y_t = CD_t$ in Table 9 is given by

$$X_t = 1.512(0.023) X_{t-1}^{(1)} - 0.751(0.059) X_{t-1}^{(2)} + 0.341(0.062) X_{t-1}^{(3)} + 0.103(0.026) X_{t-1}^{(4)} - 0.051(0.036) X_{t-1}^{(4)} + 0.085(0.048) Y_{t-1}^{(2)} - 0.040(0.101) Y_{t-1}^{(3)} + 0.006(0.046) Y_{t-1}^{(4)} + \epsilon_{1t}$$

$$Y_t = 0.143(0.032) X_{t-1}^{(1)} - 0.286(0.081) X_{t-1}^{(2)} + 0.268(0.087) X_{t-1}^{(3)} - 0.119(0.039) X_{t-1}^{(4)} + 0.973(0.051) Y_{t-1}^{(1)} + 0.197(0.117) Y_{t-1}^{(2)} - 0.128(0.142) Y_{t-1}^{(3)} - 0.047(0.065) Y_{t-1}^{(4)} + \epsilon_{2t}$$
From the results in Table 9, where the 95% confidence intervals have the star mark if they are significant, we can see that the CC affects the CD whereas the CD does not affect the CC so much. Notice that in the univariate HAR models of order \( p = 4 \) the RMSEs of \( CC \) and \( CD \) are 0.00648 and 0.00978, respectively, while in the bivariate case, those of \( CC \) and \( CD \) are 0.00647 and 0.00910. We see that the MAEs also have smaller values in the bivariate models. It implies that the bivariate HAR model fits better with the two data sets.

### Forecasting

Now in this section we investigate forecasts of the Korean COVID-19 data by means of the HAR models. For the forecasting, we use the data during the in-sample-period from Jan.20.2020 to Jan.20.2021, and compute the out-of-sample forecasts between Jan.21.2021 and Feb.18.2021.
E. Hwang and S. Yu

Table 9
Coefficient estimates (standard errors) and 95% CI in bivariate HAR(4,4) model for \((CC_C, CD_c)\).

| \(\ell = 1\) | \(\ell = 7\) | \(\ell = 14\) | \(\ell = 21\) | \(\ell = 1\) | \(\ell = 7\) | \(\ell = 14\) | \(\ell = 21\) |
|---|---|---|---|---|---|---|---|
| \(a_1\) | 1.5120(0.023) | 1.4671,1.558 | 0.143(0.032) | 0.079,0.206 |
| \(a_2\) | -0.751(0.058) | -0.864,-0.637 | -0.286(0.081) | -0.447,-0.126 |
| \(a_3\) | 0.341(0.062) | 0.220,0.462 | 0.268(0.087) | 0.098,0.431 |
| \(a_4\) | -0.103(0.028) | -0.157,-0.048 | -0.119(0.039) | -0.196,-0.043 |
| \(\beta_1\) | -0.051(0.036) | -0.121,0.020 | 0.973(0.051) | 0.873,1.072 |
| \(\beta_2\) | 0.085(0.084) | -0.079,0.249 | 0.197(0.117) | -0.034,0.429 |
| \(\gamma_1\) | -0.040(0.101) | -0.238,0.158 | -0.128(0.142) | -0.407,0.150 |
| \(\gamma_2\) | 0.006(0.046) | -0.085,0.096 | -0.047(0.065) | -0.175,0.079 |

*Indicates significant CI.

Table 10
Out-of-sample \(\ell\)-step forecast performance of the univariate HAR(4) model on cumulative confirmed \(CC_C\), cumulative recovered \(CR\), and cumulative deaths \(CD\) cases. (Values in parentheses indicate the errors divided by standard deviation.)

| \(\ell = 1\) | \(\ell = 7\) | \(\ell = 14\) | \(\ell = 21\) | \(\ell = 1\) | \(\ell = 7\) | \(\ell = 14\) | \(\ell = 21\) |
|---|---|---|---|---|---|---|---|
| \(RMSE_{CC}\) | 142.76 | 904.05 | 2519.56 | 3231.81 | 338.53 | 1049.06 | 3154.64 | 6074.15 |
| \(MAE_{CC}\) | 114.35 | 772.86 | 1844.05 | 3001.86 | 206.82 | 805.05 | 2644.25 | 5988.07 |
| \(MAPE_{CC}\) | 0.196 | 1.258 | 2.896 | 4.710 | 0.489 | 1.911 | 5.822 | 12.147 |
| \(RRSE_{CC}\) | 0.0187 | 0.1571 | 0.7069 | 1.9811 | 0.0587 | 0.2129 | 0.8555 | 2.7305 |

Table 11
Out-of-sample \(\ell\)-step forecast performance of the univariate HAR(4) model on recovery rate \(RR\), and fatality rate \(FR\). (Values in parentheses indicate the errors divided by standard deviation.)

| \(\ell = 1\) | \(\ell = 7\) | \(\ell = 14\) | \(\ell = 21\) | \(\ell = 1\) | \(\ell = 7\) | \(\ell = 14\) | \(\ell = 21\) |
|---|---|---|---|---|---|---|---|
| \(RMSE_{RR}\) | 0.5931 | 1.8122 | 4.7101 | 4.6843 | 0.0143 | 0.0869 | 0.1998 | 0.3322 |
| \(MAE_{RR}\) | 0.4038 | 1.2792 | 1.2792 | 3.5534 | 0.0097 | 0.0845 | 0.1990 | 0.3316 |
| \(MAPE_{RR}\) | 0.5656 | 1.8034 | 1.8034 | 4.8277 | 0.6669 | 5.7143 | 12.959 | 20.864 |
| \(RRSE_{RR}\) | 0.3443 | 1.0360 | 1.0360 | 2.8141 | 0.1538 | 1.0592 | 3.0011 | 7.9863 |

Table 12
Out-of-sample \(\ell\)-step forecast performance of the univariate HAR(4) model on individual infection rate \(IK_C\), \(\ell = 14,21\). (Values in parentheses indicate the errors divided by standard deviation.)

| \(\ell = 1\) | \(\ell = 7\) | \(\ell = 14\) | \(\ell = 21\) | \(\ell = 1\) | \(\ell = 7\) | \(\ell = 14\) | \(\ell = 21\) |
|---|---|---|---|---|---|---|---|
| \(RMSE_{IK_C}\) | 1.0921 | 1.082 | 1.123 | 1.393 | 0.7733 | 0.9360 | 1.0026 | 0.9103 |
| \(MAE_{IK_C}\) | 0.8923 | 0.9043 | 0.9611 | 0.9998 | 0.6507 | 0.7604 | 0.8440 | 0.6973 |
| \(MAPE_{IK_C}\) | 13.601 | 15.342 | 16.135 | 16.395 | 14.313 | 20.441 | 24.325 | 24.325 |
| \(RRSE_{IK_C}\) | 0.6855 | 0.9108 | 1.1141 | 1.408 | 0.5398 | 0.8646 | 1.2576 | 1.298 |

We examine forecast performance by evaluating four measures: root mean square error (RMSE), mean absolute error (MAE), mean absolute percentage error (MAPE), root relative square error (RRSE) for \(\ell\)-step ahead forecast, \(\ell = 1,2,\ldots\). Suppose that \(\{z_{1},z_{2},\ldots,z_{n}\}\) are observed data during the in-sample period and \(\{z_{\ell+n} : \ell = 1,2,\ldots\}\) are data used to compute out-of-sample forecasts. Let \(\hat{z}_{\ell+n}\) be the \(\ell\)-step ahead forecast. The four forecast performance measures of the \(\ell\)-step ahead forecasts are given as follows: \(N\) indicates a general time epoch of out-of-sample to obtain the \(\ell\)-step ahead forecast so that \(N-n+\ell,\ldots,N-1+\ell\) are time indices used to compute the difference between the real value and the forecast, as well as \(z_{N-n+\ell}\) is the last observation in the total sample for each \(\ell\), where \(n = m_{\ell}\) is the size for the forecasts, which depends on \(\ell\).

Root Mean Square Error (RMSE):

\[
RMSE_{\ell} = \sqrt{\frac{1}{m} \sum_{n=N-m}^{N-1} (z_{\ell+n} - \hat{z}_{\ell+n})^2}.
\]

Mean Absolute Error (MAE):

\[
MAE_{\ell} = \frac{1}{m} \sum_{n=N-m}^{N-1} |z_{\ell+n} - \hat{z}_{\ell+n}|.
\]

Mean Absolute Percentage Error (MAPE):

\[
MAPE_{\ell} = \frac{100}{m} \sum_{n=N-m}^{N-1} \left| \frac{z_{\ell+n} - \hat{z}_{\ell+n}}{z_{\ell+n}} \right|.
\]
Root Relative Square Error (RRSE): \[
RRSE_\ell = \sqrt{\frac{\sum_{m=N-\ell}^{N-1}(\hat{z}_m - z_m)^2}{\sum_{m=1}^{N-\ell}(\hat{z}_m - z_m)^2}}
\]
where \( \hat{z}_m = \frac{1}{m} \sum_{n=1}^{m} z_{N-n+\ell} \).

We evaluate the performance results of the first, 7th, 14th, 21th step ahead forecasts (i.e. \( \ell = 1, 7, 14, 21 \)) for the Korean COVID-19 time series data in Tables 10–13. In Table 10, forecasting performances of CC, CR, CD adopted in univariate HAR models are described whereas bivariate HAR forecasting performances of (CC, CD) are given in Table 13. Tables 11–12 report the multi-step ahead forecasts of RR, FR and IR. In Tables 10–13, values in parentheses of RMSE and MAE indicate the errors of forecasts by means of the normalized data with mean zero and variance one, as in Tables 2–4.

Moreover, the forecasts plots can be seen in Figs. 4–6, where predicted values as well as 95% prediction intervals are depicted between Jan.21.2021 and Feb.18.2021. Most of the cases show that forecasts are close to real data and the 95% prediction intervals include the real values except for the fatality rate. In the fatality rate case, predicted values and prediction interval have reasonably increasing pattern, although the actual values are not contained in the prediction interval. It is because the real fatality rate becomes suddenly flattened after Jan.21.2021.

We see, in the first plots of Figs. 4 and 6, that somewhat different prediction intervals for the CC are obtained in univariate and bivariate HAR models. This happens because the univariate model of the CC in Fig. 4 is independent of the CD, but the bivariate model of CC depends on the CD. As seen in the plots of real data of the CC in Figs. 4 and 6, the numbers of death cases are increasing very slowly during the period from Jan.21.2021 to Feb.18.2021, and this fact affects the slow increasing predicted values of the CC in the bivariate model in Fig. 6. Also, we note in Tables 10 and 13 that in the bivariate prediction model the errors for the CC are similar to or smaller, and all the errors for the CD are smaller than those in the univariate prediction model. For example, in the univariate CD has RMSEs (7.66, 37.92, 94.59, 172.85) whereas in the bivariate (7.087, 27.48, 62.07, 104.08) for \( \ell = 1, 7, 14, 21 \). Consequently, we conclude that the bivariate HAR model yields better fitting performance than the univariate model as we expected.

Conclusion and discussion

This present work concerns with the COVID-19 time series analysis in South Korea. The COVID-19 pandemic lasts longer than one year, and humans are still threatened and are not free from the virus. We deal with a variety of Korean COVID-19 time series data sets such as confirmed case, recovery case, death case as well as recovery rate, fatality rate and infection rates. An interesting and robust time series model for long-memory structured data is a HAR model that has recently attracted much attention from econometrician and statistician in the field of time series analysis.

In this work, instead of the traditional ARIMA model, the HAR models are applied to study the estimation and forecasting for the Korean COVID-19 time series data. The reason for choosing the HAR models in the present paper is that the Korean COVID-19 data have long-memory features, and moreover some criteria such as \( R^2 \), AIC and BIC have good performances. Indeed, the \( R^2 \)-values of the four data sets: cumulative confirmed, cumulative recovery, cumulative death cases and recovery rate, are one, and also the value of fatality rate is 0.999. Those of infection rates for 14 and 21 days are larger than 0.84.

For the model selection, we compute RMSE, MAE as well as \( R^2 \), AIC, BIC to decide the order of the HAR models, and evaluate coefficients estimates by means of the least square method for each model. Moreover, 95% confidence intervals along with standard errors are constructed. Adopting the HAR models, we see that the actual values of COVID-19 cases and the fits by the HAR models have very small errors in residuals, with well-matched fitting plots as shown in Figs. 2 and 3. Also, real daily data of confirmed, recovered and death cases are well-matched with the differences from our proposed HAR models. Furthermore, noticing that confirmed and death cases are strongly correlated with each other, a bivariate HAR model is applied to the two data sets. Better fitting is demonstrated with smaller RMSE and MAE in the bivariate HAR model than in the univariate HAR models.

To investigate the out-of-sample forecasting, HAR models with the orders, chosen optimally according to the criteria rules, are used to compute multi-step ahead forecasts. As forecasting performance measures, RMSE, MAE, MAPE and RRSE are evaluated at the 1, 7, 14 and 21-step ahead forecasts. Predicted values as well as 95% prediction intervals are illustrated from the out-of-sample forecasting. Most of the cases show that forecasts are close to real values and the 95% prediction intervals contain the real values except for the fatality rate. In the fatality rate case, predicted values and prediction interval have reasonably increasing pattern, although the actual values of the real data deviate from the prediction interval. The reason is that the fatality rates become suddenly flattened after the time epoch of the out-of-sample.

For the two data sets of cumulative confirmed and cumulative death cases, as mentioned above, the univariate and bivariate models are considered for forecasting. We see that somewhat different prediction intervals for the cumulative confirmed cases are found. It is because the univariate model for the confirmed cases is independent of death cases, whereas the bivariate model depends on the death cases as regressors. The numbers of death cases are increasing very slowly during the out-of-sample period, and for this reason, slowly increasing predicted values of the confirmed cases seem to appear in the bivariate model. Also, the forecast results report that in the bivariate prediction model the forecast accuracy errors for the cumulative confirmed case are similar to or smaller, and all the errors for the cumulative death case are smaller than those in the univariate prediction model. In conclusion, the bivariate HAR model improves the fitting-performance of the univariate models.

The novelty of this paper is as follows: This work is the first attempt of the HAR models for the COVID-19 analysis in literature, and moreover we validate that the HAR model provides a good prediction model for the COVID-19 time series data. Various data sets from the Korean COVID-19, that is, time series data such as confirmed, recovered, death cases as well as recovery, fatality, and infection rates, are applied to the HAR models. Parameter estimation and forecasting are illustrated with good performances of fitting and accuracy. In the pair of the confirmed and death cases, the bivariate HAR model has better fitting-performance than the univariate HAR models. Therefore, the proposed model can be an efficient tool for the COVID-19 analysis, which will help governments and health systems to manage social policies of the COVID-19 disease prevention during the pandemic.

We have explored the estimation and prediction using the HAR models for the COVID-19 time series data in South Korea. However, Korean COVID-19 cumulative confirmed or death cases are not much similar to those of worldwide or US COVID-19 data, which have some pattern of oscillation with 7 days periodicity. By imposing the periodicity to the HAR models, we can extend the discussion of this work to the COVID-19 in other countries. For example, recently [51] proposed an integer-valued AR model with oscillating weighted cosine geometric innovations for modeling the COVID-19 series in some small island developing states. The weighted cosine geometric process accounts for oscillating patterns and, according to [52], it outperforms well-known competing discrete models. The HAR model with the weighted cosine geometric innovation terms will be an interesting model to fit the COVID-19 series worldwide or in other countries with oscillation feature. This extension will be studied in a future research. Another interesting time series model we suggest is a partially periodic HAR model, which has some truncation in the oscillation. In general, as seen
Fig. 4. Forecasting and 95% prediction intervals of Cumulative Confirmed, Cumulative Recovery, Cumulative Death cases using the univariate HAR models.

Fig. 5. Forecasting and 95% prediction intervals of Recovery Rate, Fatality Rate (first row); Infection Rates for 14 and 21 days (second row).

Fig. 6. Forecasting and 95% prediction intervals of Cumulative Confirmed, Cumulative Death cases using the bivariate HAR models.

Table 13
Out-of-sample $\ell$-step forecast performance of the bivariate HAR(4, 4) model on joint cumulative confirmed $CC$, and cumulative deaths $CD$, cases. (Values in parentheses indicate the errors divided by standard deviation.)

| Bivariate       | $CC$          | $CD$          |
|-----------------|---------------|---------------|
|                 | $(j = 1)$     | $(j = 2)$     |
|                 | $\ell = 1$    | $\ell = 7$    | $\ell = 14$ | $\ell = 21$ |
| RMSE$\ell$      | 146.12        | 999.56        | 2984.10     | 2283.31      | 7.087        | 27.482       | 62.07       | 104.08      |
|                 | (0.0126)      | (0.0904)      | (0.2750)    | (0.2215)     | (0.0417)     | (0.1718)     | (0.3897)    | (0.6633)    |
| MAE$\ell$       | 117.72        | 768.20        | 2270.15     | 1952.30      | 5.218        | 22.56        | 54.52       | 86.61       |
|                 | (0.0101)      | (0.0691)      | (0.2106)    | (0.1912)     | (0.0305)     | (0.1392)     | (0.3405)    | (0.5509)    |
| MAPE$\ell$      | 0.2025        | 1.2614        | 3.4759      | 2.8907       | 0.6363       | 2.7278       | 5.7639      | 8.4292      |
|                 | (0.0192)      | (0.1737)      | (0.8372)    | (1.3996)     | (0.0429)     | (0.2031)     | (0.6403)    | (1.9449)    |
| RRSE$\ell$      | 0.0192        | 0.1737        | 0.8372      | 1.3996       | 0.0429       | 0.2031       | 0.6403      | 1.9449      |

the numbers of the confirmed cases in US., Germany, Brazil, India, etc., the oscillation seems to appear as the magnitude of the COVID-19 series is large, and furthermore, the larger magnitude the stronger oscillation. In order to represent this phenomenon, partial oscillating which has some truncation as well as is proportional to the magnitude of oscillation might be desirable rather than pure oscillating. This topic
on the partially periodic HAR model will be also very interesting and promising.

**CRediT authorship contribution statement**

**Eunjoo Hwang:** Conceptualization, Methodology, Software, Formal analysis, Writing – original draft, Project administration, Supervision, Funding acquisition. **SeongMin Yu:** Conceptualization, Methodology, Software, Formal analysis, Writing – original draft, Visualization.

**Declaration of competing interest**

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

**Acknowledgment**

This work was supported by National Research Foundation of Korea (NRF-2018R1D1A1B07048745).

**References**

[1] Benvenuto D, Giovannetti M, Vassallo L, Angelissi S, Giccozzi M. Application of the ARIMA model on the COVID-19 epidemic dataset. Date Brief 2020:105340.

[2] Croux Z. Estimation of COVID-19 prevalence in Italy, Spain, and France. Sci Total Environ 2020;729:138817.

[3] Kirbas I, Sozen A, Tuncer AD, Kazancioglu FS. Comparative analysis and forecasting of COVID-19 cases in various European countries with ARIMA, NARNN and LSTM approaches. Chaos Solitons Fractals 2020;138:110015.

[4] Ribeiro MBMD, Silva RDG, Mariani VC, Coelho LJS. Short-term forecasting COVID-19 cumulative confirmed cases: Perspectives for Brazil. Chaos Solitons Fractals 2020;135:109853.

[5] Kumar N, Susan S. COVID-19 pandemic prediction using time series forecasting models. In: 11th ICCI 2020 Conference. 2020.

[6] Tandion H, Ranjan P, Shakhrobaty T, Suhag V. Coronavirus (COVID-19): ARIMA based time-series analysis model to forecast near future. 2020. arXiv:2004.07859.

[7] Mustafa HI, Faried NY. COVID-19 cases in Iraq: forecasting incidents using box–jenkins ARIMA model. In: 2nd Al-Noor International Conference for Science and Technology (ICSTC). 2020. p. 22–6.

[8] Atanaga A, Arar IS. Modeling and forecasting the spread of COVID-19 with stochastic and deterministic approaches: Africa and europe. Adv Difference Equ 2021;57.

[9] Taylor SJ, Letham B. Forecasting at scale. Amer Statist 2018;72:37–45.

[10] Maleki M, Mahmoudi M, Wraith D, Pho K. Time series modelling to forecast the confirmed and recovered cases of COVID-19. Travel Med Infect Dis 2020;37:101742.

[11] Maleki M, Mahmoudi MR, Heydari MH. Modeling and forecasting the spread and death rate of coronavirus (COVID-19) in the world using time series models. Chaos Solitons Fractals 2020;140:110151.

[12] Soukhovolsky V, Kovalov A, Pitt A, Kessel B. A new modelling of the COVID 19 pandemic. Chaos Solitons Fractals 2020;139:110039.

[13] Reis RF, Quintela BDM, Campos JDO, Gomes JM, Rocha BM, Lobosco M. Characterization of the COVID-19 pandemic and the impact of uncertainties, mitigation strategies, and underreporting of cases in South Korea, Italy, and Brazil. Chaos Solitons Fractals 2020;136:109888.

[14] Sarkar K, Khajanchi S, Nieto JJ. Modeling and forecasting the COVID-19 pandemic in India. Chaos Solitons Fractals 2020;139:110049.

[15] Mohammed AA, Ewees Ahmed A, Fan Hong. Abd El Aziz Mohamed. Optimization method for forecasting confirmed cases of COVID-19 in China. J Clin Med 2020;9:674.

[16] Chatterjee S, Asad A, Shayan B, Bhattacharya S, Alam S, Verma MK. A systematic review and meta-analysis of published research data on COVID-19 infection fatality rates. Int J Infect Dis 2020;91:1318–48.

[17] Lee S, Liao Y, Mei SH, Yin Y. Sparse HP filter: Finding kinks in the COVID-19 contact rate. J Econometrics 2021;220:158–80.

[18] Nikolopoulos K, Punia S, Schäfers A, Tsinopoulos C, Verma MK. Evolution and death rate of coronavirus (COVID-19) in the world using time series models. Chaos Solitons Fractals 2020;140:110196.

[19] Chaurasia V, Pal S. Application of machine learning time series analysis for prediction COVID-19 pandemic. Res Biomed Eng 2020.

[20] Salgotta R, Gandomi M, Gandomi AH. Time series analysis and forecast of the COVID-19 pandemic in India using genetic programming. Chaos Solitons Fractals 2020;138:109945.

[21] Prasanth S, Singh U, Kumar A, Tikkaliya VA, Chong PIU. Forecasting spread of COVID-19 using google trends: A hybrid GWO-deep learning approach. Chaos Solitons Fractals 2021;142:110336.

[22] Balli S. Data analysis of Covid-19 pandemic and short-term cumulative case forecasting using machine learning time series methods. Chaos Solitons Fractals 2021;142:110512.

[23] Kalantari M. Forecasting COVID-19 pandemic using optimal singular spectrum analysis. Chaos Solitons Fractals 2021;142:110547.

[24] Mehra AHA, Shafeeirad M, Abbasi Z, Zamani I. Parameter estimation and prediction of COVID-19 epidemic teruning point and ending time of a case study on SIR/SQAR epidemic models. Comput Math Methods Model 2020;465923.

[25] Carvalho K, Vicente JP, Jakovljevic M, Teixeira JP. Forecasted incidence, intensive care unit admissions and projected mortality attributable to Covid-19 in Portugal, UK, Germany, Italy and France - 4 weeks ahead. 2021, 2021050116. http://dx.doi.org/10.20994/preprints202105.0116.v1, Preprints.

[26] Feroze N. Forecasting the patterns of COVID-19 and causal impacts of lockdown in top five affected countries using bayesian structural time series models. Chaos Solitons Fractals 2020;140:110196.

[27] Karats C. A time series-based statistical approach for outbreak spread forecasting: Application of COVID-19 in Greece. Expert Syst Appl 2021;166:114077.

[28] Marmarelis VZ. Predictive modeling of COVID-19 data in the US: Adaptive phase-space approach. J Eng Med Biol 2020;30:08313.

[29] Liu L, Moon HR, Schoferde F. Panel forecasts of country-level COVID-19 infections. J Econometrics 2021;220:22.

[30] Sun J, Chen X, Zhang Z, Lai S, Zhao B, Liu H, Wang S, Huan W, Zhao R, Ng MTA, Zheng Y. Forecasting the long-term trend of COVID-10 epidemic using a dynamic model. Sci Rep 2020;10:21122.

[31] Ioannidis JPA. Infection fatality rate of COVID-19 inferred from seroprevalence data. Bull World Health Organ 2020;0304–4076.

[32] James Nick, Menzies Max. Association between COVID-19 cases and international equity indices. Physica D 2021;417:132809.

[33] Mehlhennrich F, Molinari E. Estimating the COVID-19 infection rate: Anatomy of an inference problem. J Econometrics 2020;60268–2690.

[34] Feng J, Zhao Z, Shao X. Time series analysis of COVID-19 infection curve: A change-point perspective. J Econometrics 2020;220:139–57.

[35] Keane M, Neal T. Consumer panic in the COVID-19 pandemic. J Econometrics 2020;204:4076.

[36] Hanke M, Kostolopova M, Weissensteiner A. COVID-19 and market expectations: Evidence from option-implied densities. Econ Lett 2020;195:109441.

[37] James Nick, Menzies Max. Association between COVID-19 cases and international equity indices. Physica D 2021;417:132809.

[38] Lee S, Liao Y, Mei SH, Yin Y. Sparse HP filter: Finding kinks in the COVID-19 contact rate. J Econometrics 2021;220:158–80.

[39] Nikolopoulos K, Punia S, Schäfers A, Tsinopoulos C, Verma MK. Forecasting and planning during a pandemic: COVID-19 growth rates, supply chain disruptions, and governmental decisions. European J Oper Res 2021;290:99–115.

[40] Lee KB, Han S, Jeong Y. COVID-19, flattening the curve, and Benford’s law. Physica A 2020;559:125090.

[41] Li S, Linton O. When will the COVID-19 pandemic peak? J Econometrics 2020;220:139–57.

[42] Keane M, Neal T. Consumer panic in the COVID-19 pandemic. J Econometrics 2020;204:4076.