Communication

Short-Term Prediction Methodology of COVID-19 Infection in South Korea

Grace S. Ko *† and Taeseon Yoon †

Hankuk Academy of Foreign Studies, Yongin 17035, Gyeonggi, Korea; tsyoon@hafs.hs.kr
* Correspondence: seoyoungko@hafs.kr
† These authors contributed equally to this work.

Abstract: The purpose of this study is to predict the short-term trend of the COVID-19 pandemic and give insights into effective response strategies. Based on the basic SIR model, a compartment method for modeling the course of an epidemic, the short-term infection change ratio \( m_d \), is derived. The number of infected people can be predicted using this ratio. We calculated different \( m_d \) values on a weekly basis. As we tested different combinations of \( m_d \), the prediction from the combination of \( m_d \) based on a week and \( m_d \) based on 4 weeks was found to be statistically reliable. According to our regression analysis, our approach has an explanatory power of 96%. However, this method could only predict 1 week ahead of current data. Thus, we use LSTM, a deep learning method applied for time series data, to forecast the trend 4 weeks ahead. The forecasted trends show that the number of infected people in South Korea will reach its peak a week after the writing of this work and start to gradually decline after that.

Keywords: COVID-19; SIR model; regression; LSTM

1. Introduction

In December 2019, an unknown novel coronavirus, SARS-CoV-2 (COVID-19), was first reported in Wuhan, China. As globalization has increased the frequency of personal and business travel to other countries, COVID-19 has spread rapidly around the world. On 11 March 2020, the World Health Organization (WHO) declared a pandemic [1]. As of 21 August 2021, a total of 211,503,434 confirmed cases of COVID-19 were reported, and the death toll reached 4,426,543 [2].

As the pandemic continues, available information about COVID-19 is being collected, especially regarding its nature and characteristics. In particular, it is known that the virus tends to change its nature, evolving new variants based on genetic mutations. Therefore, thorough research is urgent in order to find the most effective measurements that can help end the COVID-19 pandemic [3].

Regarding the daily reported cases in South Korea, as presented in Figure 1, the number of cases first peaked and declined around March 2020. This decline was analyzed to be the result of measures such as school closures, social distancing, and the cancellation of all public gatherings. However, the number increased again in September 2020, which was repeated in January 2021, caused by various factors such as the increase and decrease of daily inspections, the significant relaxation and reinforcement of preventive measures, and the perception of members of society. Recently, the number of cases started to increase again in June 2021, and the numbers are currently still not decreasing. Therefore, predictions about the evolution of COVID-19 are crucial to adopting or readjusting preventive measures. Short-term predictions as well as long-term predictions about the end of the COVID-19 pandemic are very important.

There are three basic approaches to the prediction of the dynamics of an epidemic such as COVID-19: compartmental models, statistical methods, and ML-based methods [4,5]. Compartment models divide a population into mutually exclusive compartments using
a set of dynamic equations that describe the transition between compartments [6]. The Susceptible–Infected–Removed (SIR) model [7] is the most common model for modeling epidemics. Statistical methods involve the extraction of general statistics from data and fitting them to a mathematical model that describes the evolution of an epidemic [8,9]. Finally, ML-based methods are techniques that use machine learning algorithms (artificial intelligence) to analyze historical data and find patterns that accurately predict the number of new infections [4,10,11].

In this study, based on classical SIR models, we want to estimate the parameters of SIR models using statistical techniques from existing data and predict short-term infections through these estimated parameters in South Korea. We also want to apply the Recurrent Neural Network (RNN) technique for time-series prediction as a parametric prediction methodology for predicting extended periods.

2. Materials and Methods

2.1. Data Source

In this study, data from 21 January 2020 to 10 August 2021 are used; these data are taken from the data provided by the Korea Centers for Disease Control and Prevention (KCDC) [12]. In this dataset, there are four data fields: date, confirmed, death, and released. The number of current infections is calculated and added to the dataset, which is the number of confirmed cases subtracted by the number of deaths and released people. The trend of the number of infections in South Korea is shown in Figure 1. There have been four big waves of the pandemic in Korea. Although the fourth wave remains in progress at the time of the last data, the overall number of the infected is managed at a low level compared to the total population of Korea (51,829,136 people) [13]. However, considering the recent epidemic situation, there is still a possibility of an explosion, so continuous and thorough management is required.

Figure 1. The number of people infected with COVID-19 in South Korea.

2.2. SIR Model

In this study, the standard Susceptible–Infected–Recovered (SIR) model, proposed by Kermack–McKendrick [7], is considered. In this model, the total population (N) is composed of three categories: Susceptible (S), Infected (I), and Recovered (R). The Susceptible category is the part of the total population that is vulnerable and at risk of becoming infected. The Infected category is the part of the total population that is managed by the disease, and the Recovered category is the fraction of the total population that has recovered or has been removed from the data (by death or recovery). The basic structure of the SIR mathematical model is described in Figure 2.
The SIR model is expressed as an ordinary differential equation modeling the spread of epidemics in a fixed population assuming permanent immunity [14]. These equations are

\[
\begin{align*}
\frac{dS}{dt} &= -\frac{\beta SI}{N} \\
\frac{dI}{dt} &= \frac{\beta SI}{N} - \gamma I \\
\frac{dR}{dt} &= \gamma I
\end{align*}
\] (1)

where $\beta$ represents the effective transmission rate—i.e., the average number of contacts per person over time—multiplied by the probability of disease transmission in a contact between a susceptible and an infectious subject. $\gamma$ represents the recovery rate, which is defined as the inverse of the duration of recovery $r_d$ ($\gamma = 1/r_d$).

$N$, $S$, $I$, and $R$ are assumed to be functions of $t$, as they fluctuate with time $t$ as infection occurs. The Susceptible population changes with the number of infections and the vaccinated. If we let $\alpha(t)$ be the ratio of infected and vaccinated people at time $t$, then $S(t) = \alpha(t)N$. Consequently, Equation (1) can be rewritten as follows:

\[
\frac{dI}{dt} = (\beta\alpha(t) - \gamma)I(t) = m(t)I
\]

where $m(t) = \beta\alpha(t) - \gamma$, which indicates the change ratio of the number of infections at time $t$, since

\[
\frac{dI(t)/dt}{I(t)} = m(t)
\] (2)

Here, we assume that $m(t)$ is constant in some interval $(t - d, t + d)$ and $m(t) = m^d_t$. Integrating Equation (2), we can obtain the following result:

\[
\int_{t-d}^{t} \frac{dI(t)/dt}{I(t)} dt = \int_{t-d}^{t} m^d_t dt
\]

\[
ln(I(t)) = ln(I(t-d)) + m^d_t d
\] (3)

then

\[
m^d_t = \frac{ln(I(t)) - ln(I(t-d))}{d}
\] (4)

We can calculate the value of $m^d_t$ from the used data and Equation (3). The prediction formulation can be derived as follows:

\[
ln(I^p(t + d)) = ln(I(t)) + m^d_t d
\]

\[
I^p(t + d) = I(t)e^{m^d_t d}
\] (5)

where $I^p(t + d)$ is the predicted number of infected people at time $t + d$. 

Figure 2. Schematic representation of the SIR model.
3. Results

3.1. Calculation of \( m_d \)

By calculating \( m_d \), we can predict the rate of infection after \( d \) days. If \( d \) is small, we can satisfy the assumption that \( m_d \) is a constant. If \( d \) is large, the uncertainty in the change over the period increases, which can violate the assumption. Thus, the selection of an appropriate \( d \) is important. Furthermore, an important factor in the choice of \( d \) is the forecast period. We set \( d \) on a weekly basis, from 1 week (\( m_7 \)) to 4 weeks (\( m_28 \)), as shown in Figure 3.

![Figure 3. Trend of \( m_d \) values.](image)

In Figure 3, all \( m_d \)s show a similar pattern. This pattern is also shown well in the correlation coefficient table in Table 1. The volatility of \( m_7 \) and \( m_21 \) is relatively large, and \( m_14 \) and \( m_28 \) are relatively stable. In particular, \( m_28 \) indicates a stable state during the whole interval. The overall trend started in a very unstable state at the beginning, but it entered a stable state as time passed. It can be seen that the volatility expands in the large wave interval. Since the initial trend in \( m_d \) is very unstable and stabilizes after the first large wave, we analyze data after May 2020, excluding data from the initial period of instability, as these data may negatively affect the overall analysis results.

Table 1 shows that the difference between the value \( d \) and the correlation is inversely related. This means that it is useful to use variables with large differences in \( d \) to increase the effectiveness of predictions.

|       | \( m_7 \) | \( m_14 \) | \( m_21 \) | \( m_28 \) |
|-------|----------|----------|----------|----------|
| \( m_7 \) | 1        | 0.9248   | 0.7480   | 0.6854   |
| \( m_14 \) | 0.9248   | 1        | 0.8659   | 0.7080   |
| \( m_21 \) | 0.7480   | 0.8659   | 1        | 0.8307   |
| \( m_28 \) | 0.6854   | 0.7080   | 0.8307   | 1        |

3.2. Regression Analysis

Each \( m_d \) can be used to predict the number of infected people at time \( t \). However, predicting with a combination of these values can increase the predictive power. Therefore, this study presents an equation for prediction through regression analysis. The full regression model can be expressed as follows:

\[
\ln(I^p(t)) = \beta_0 + \beta_7 \ln(I^p_7(t)) + \beta_1 \ln(I^p_14(t)) + \beta_2 \ln(I^p_21(t)) + \beta_3 \ln(I^p_28(t))
\]
where
\[ \ln(I_p^d(t)) = \ln(I(t - d)) + m^d \times d \quad \text{when } d = 7, 14, 21, 28 \]

By applying the backward elimination method, we can obtain the final results in Table 2. The regression model is well fitted with adj-R^2 = 0.969.

Table 2. The ANOVA table of the reduced model.

| df  | SS      | MS       | F         | Significance F |
|-----|---------|----------|-----------|----------------|
| Regression | 2       | 495.664  | 247.832   | 7248.166       | 0               |
| Residual   | 464     | 15.865   | 0.034     |                |                 |
| Total      | 466     | 511.529  |           |                |                 |

| df | SS      | MS       | F         | Significance F |
|----|---------|----------|-----------|----------------|
| Intercept | 0.501 | 0.066    | 7.628     | 0.000          |
| ln(I_p^7(t)) | 0.889 | 0.011    | 83.283    | 0.000          |
| ln(I_p^28(t)) | 0.052 | 0.008    | 6.225     | 0.000          |

From the ANOVA table above, Equation (6) can be reduced as follows:
\[ \ln(I_p^7(t)) = 0.501 + 0.889\ln(I_p^7(t)) + 0.052\ln(I_p^{28}(t)) \quad (7) \]

Using data from 1 May 2020 to 10 August 2021, Figure 4 shows the actual number of infected people, the predicted value \( I_p \) of infected people obtained from the Equation (7), and the predicted value \( I_p^7 \) of infected people using \( m^7 \). Figure 4 shows that \( I_p \) is closer to the true value while reducing the volatility compared to \( I_p^7 \). It can also be seen that it shows the same pattern as the number of infected people fluctuates. The prediction of these movements shows that it is possible to predict the change in the number of infected people due to the variants.

![Figure 4. Comparison between the actual number of infected people and the predicted value.](image-url)

It can be seen that the proposed regression analysis model fits the actual values well. However, due to \( I_p^7 \), the prediction interval of this equation based on the current information is as short as 1 week. To establish a strategy to deal with the current COVID-19 situation, a forecast of at least 4 weeks is required. To achieve this, a forecast of the next 3 weeks for \( m^7 \) must be included. To forecast the value of \( m^7 \), we adopt the LSTM method—the most popular deep learning model in time series prediction.

LSTM consists of three types of layers: an input layer, output layer, and hidden layer. The input layer and output layer are composed of one layer each, while the number of hidden layers can be set. In this study, the LSTM is run with three hidden layers, using...
**Mean Squared Error** as the LSTM loss function and Adam as the optimization algorithm. A batch size of 64 is used for training. The number of epochs is fixed at 200. With this LSTM, we obtain $m^7$ values for the next 21 days. Using these values, we can obtain the predicted values of infected people from the 8th day after the last data point to the 28th day by Equations (5) and (7). The prediction values for first 7 days after the last data point are based on current data. These values are shown in Figure 5.

![Figure 5. Four-week forecast for the number of infected people.](image-url)

In Figure 5, we can see that the wave is heading towards a peak and will decline over the following week. A gradual decrease is expected rather than a sharp decrease. This indicates that an active response strategy is needed to reduce the number of infected people. It also shows that the cycle is occurring on a weekly basis. This phenomenon occurs continuously under the influence of the number of inspections, and this is reflected in the forecast data.

**4. Discussion**

This study focuses on short-term prediction as a strategic response to COVID-19 rather than predicting the end of the COVID-19 epidemic. Under the assumption that the short-term infection change ratio ($m^d$) is constant, this study was conducted based on the differential equation of the basic SIR model.

The results of the first study derived a short-term (1 to 4 week) prediction formula (3) for the number of infected people ($I$) as a function of the change ratio ($m^d$) and itself. The change ratio of the number of infected people ($m^d$) induced by the SIR model can be simply calculated from the current data.

For more accurate prediction, a combination of short-term predictions was constructed through regression analysis, and a combination of $m^d$ values of 1 week and 4 weeks was proposed. This formula (Equation (7) had an explanatory power (adj-$R^2$) of 96%, indicating that it reflects the actual data well.

Finally, in order to predict the trend of the pandemic a month ahead using this formula, which can predict a week based on current data, the change ratio for the number of infected people ($m^d$) must be predicted. For this prediction, we applied LSTM—the most representative time series analysis method. The results of South Korea’s forecast for infections, for which we applied the predicted infected change ratio ($m^d$), showed that the current wave is expected to peak soon and show a gradual decline, but the sharp decline shows that the situation will remain difficult.

Currently, COVID-19 has presented several large waves in South Korea, but long-term prediction methods show difficulties in predicting these big waves. On the other hand, the model presented in this study focuses on forecasting for about 4 weeks considering various...
environmental changes. Thus, our approach will be useful for short-term COVID-19 response strategies.

**Author Contributions:** Conceptualization, G.S.K. and T.Y.; methodology, G.S.K. and T.Y.; software, G.S.K.; validation, T.Y.; formal analysis, G.S.K. and T.Y.; investigation, G.S.K.; resources, G.S.K.; data curation, T.Y.; writing—original draft preparation, G.S.K.; writing—review and editing, T.Y.; visualization, G.S.K.; supervision, T.Y.; project administration, T.Y. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research received no external funding.

**Institutional Review Board Statement:** Not applicable.

**Informed Consent Statement:** Not applicable.

**Conflicts of Interest:** The authors declare no conflict of interest.

**References**

1. World Health Organization (WHO) Director, General’s Opening Remarks at the Media Briefing on COVID-19. Available online: https://www.who.int/director-general/speeches/detail/who-director-general-s-opening-remarks-at-the-media-briefing-on-covid-19---11-march-2020 (accessed on 3 August 2021).
2. Worldometer. Available online: https://www.worldometers.info/coronavirus/ (accessed on 3 August 2021).
3. Yang, P.; Liu, P.; Li, D.; Zhao, D. Corona virus disease 2019, a growing threat to children? *J. Infect.* 2020, 80, 671. [CrossRef] [PubMed]
4. Watson, G.L.; Xiong, D.; Zhang, L.; Zoller, J.A.; Shamshoian, J.; Sundin, P.; Bufford, T.; Rimoin, A.W.; Suchard, M.A.; Ramirez, C.M. Pandemic velocity: Forecasting covid-19 in the us with a machine learning and bayesian time series compartmental model. *PLoS Comput. Biol.* 2021, 17, e1008837. [CrossRef] [PubMed]
5. Arik, S.; Li, C.L.; Yoon, J.; Sinha, R.; Epshteyn, A.; Le, L.; Menon, V.; Singh, S.; Zhang, L.; Nikoltchev, M.; et al. Interpretable Sequence Learning for Covid-19 Forecasting. *arXiv* 2020, arXiv:2008.00646.
6. Blackwood, J.C.; Childs, L.M. An introduction to compartmental modeling for the budding infectious disease modeler. *Lett. Biomath.* 2018, 5, 195–221. [CrossRef]
7. Kermack, W.O.; McKendrick A.G. Contributions to the mathematical theory of epidemics. *Proc. R. Soc. Lond. Ser. A Math. Phys. Eng. Sci.* 1927, 115, 700–721.
8. Liao, Z.; Lan, P.; Liao, Z.; Zhang, Y.; Liu, S. TW-SIR: Time-Window based sir for covid-19 forecasts. *Sci. Rep.* 2020, 10, 1–15. [CrossRef] [PubMed]
9. Zou, Y.; Pan, S.; Zhao, P.; Han, L.; Wang, X.; Hemerik, L.; Knops, J.; van der Werf, W. Outbreak analysis with a logistic growth model shows COVID-19 suppression dynamics in China. *PLoS ONE* 2020, 15, e0235247. [CrossRef] [PubMed]
10. Murphy, K.P. *Machine Learning: A Probabilistic Perspective*; MIT Press: London, UK, 2012.
11. Malik, Z.; Atlam, E.S.; Ewis, A.; Dagnew, G.; Ghoneim, O.A.; Mohamed, A.A.; Mohamed, M.A.; Gad. I. The COVID-19 pandemic: Prediction study based on machine learning models. *Environ. Sci. Pollut. Res.* 2021, 28, 1–11. [CrossRef] [PubMed]
12. Korea Centers for Disease Control and Prevention. Current Status of COVID-19 Outbreak in the Republic of Korea. Available online: http://ncov.mohw.go.kr/tcmBoardList.do?brdId=&brdGubun=&dataGubun=&ncvContSeq=&contSeq=&board_id= (accessed on 3 August 2021).
13. Korean Statistical Information Service. Population Census. Available online: http://kosis.kr/statisticsList/statisticsListIndex.do?menuid=M_01_01&viewcd=MT_ZTITLE&parmTabId=M_01_01 (accessed on 3 August 2021).
14. Hethcote, H.W. Qualitative analyses of communicable disease models. *Math Biosci.* 1976, 28, 335–356. [CrossRef]