Predicting COVID-19 Spread using Simple Time-Series Statistical Models

Badriya Khayyat, Fouzi Harrou, Ying Sun
King Abdullah University of Science and Technology (KAUST)
Computer, Electrical and Mathematical Sciences and Engineering (CEMSE) Division
Thuwal, 23955-6900, Saudi Arabia
Email: badriya.Khayyat@kaust.edu.sa, fouzi.harrou@kaust.edu.sa

Abstract—Accurate and timely forecasts of new COVID-19 cases and recoveries would assist in the management of medical resources and bolster public policy formulation during the current pandemic. This study aims to forecast records of confirmed time-series data using simple time series models. Importantly, to predict COVID-19 data of limited size, the performance of statistical time series models, including Linear Regression (LR) and Exponential Smoothing (ES), was investigated. The daily records of confirmed and recovered cases from Saudi Arabia, India, and France were adopted to train and test the investigated models. The forecasting accuracy has been assessed based on three commonly used statistical indicators. Results reveal that the LR model did not forecast COVID-19 time-series data successfully. On the other hand, the ES model showed a promising forecasting performance for both recovered and confirmed timeseries data. Furthermore, results showed that ES outperformed the Decision Tree regression and support vector regression with linear kernel.

Index Terms—COVID-19, time series forecasting, data-driven model, exponential smoothing, SVR, Decision Tree.

I. INTRODUCTION

At the end of 2019, the first case of a novel coronavirus disease was registered in Wuhan, China, in December 31, 2019 [1]. COVID-19 is a contagious disease caused by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) [2]. Infected patients experienced symptoms similar to pneumonia. This virus spreads by human-to-human transmission, and it was international travelers who were responsible for its worldwide proliferation [3]. On January 5, 2019, the World Health Organization (WHO) published the first Disease Outbreak News on the new virus [1]. However, as the outbreak escalated around the world, it became more challenging to address. Strict standards had to be applied by countries to limit the spread of the disease. By the end of March 2020, more than 100 nations implemented a partial or complete lockdown, affecting the normal life behavior of billions of people. By wearing a mask, using a sterilizer, and practicing social distancing, individuals can help prevent COVID-19 from spreading.

Artificial intelligence (AI) can be defined as the simulation of human intelligence in machines programmed to react and work by mimicking the human brain. AI has confirmed its effectiveness in several areas, including healthcare, where it has generally been referred to as a powerful tool to help recognize diseases and make medical diagnoses [4]–[6]. In this pandemic, AI could help predict outbreaks and help assemble quickly evolving data to support general health specialists in complex decision-making [7].

Recently, many applications of AI have been developed in the healthcare field [8], [9]. For example, Health Map an automated electronic information system designed by a team of high-level educated people at Boston Children’s Hospital [10]. Health Map relies on unofficial online sources for disease outbreak observing and real-time surveillance of emerging public health threats. Blue Dot, a Canadian-based AI company that specializes in infectious disease epidemiology, is another example of AI in the healthcare field9. Both HealthMap and Blue Dot showed successful predictions for Zika and Coronavirus. On the other hand, there have been failures such as Google Flu Trends (GFT), which used search engines to enhance flu epidemic tracking. The model underperformed significantly, most likely due to the poor design of the variables used for predicting. To sum up, AI is an essential tool to forecast pandemics and take early steps to minimize the resulting negative impact.

Since Covid-19 appeared and affected the whole world, the research priority was not just finding a vaccine but predicting the number of cases to decrease its global impact. In order to understand and manage this epidemic, we turn to various works in the literature related to forecasting of COVID-19 spread. In [11], Namasudra et al. applied a novel Nonlinear Autoregressive (NAR) Neural Network-based approach to predict the transmission of COVID-19. Three algorithms have been implemented in training this model, and results reveal that the model trained with the Levenberg Marquardt procedure achieved better performance than other models. In [12], an approach based on Singular Spectrum Analysis (SSA) has been introduced to forecast daily COVID-19 cases till October 29, 2020. The SSA approach has been optimized, and its forecasting performance has been compared to other commonly used time series forecasting models, such as ARIMA and ES and Neural Network Autoregression. This study showed that optimized SSA outperformed the competing approaches. The authors in rostami2021forecasting proposed a Multiple Linear Regression model for forecasting new confirmed COVID-19 cases by utilizing calls recorded in a call center. Essentially,
this model incorporated information about the correlation linking the confirmed case and the NHS 111 calls and other features, including trends, the impact of weekends, and lags of confirmed cases. They show the superior performance of this approach compared to traditional models without call data.

In [13], authors used a partial derivative regression and nonlinear machine learning (PDR-NML) method to execute a prediction process on the COVID-19 epidemic. First, a Progressive Partial Derivative Linear Regression model was used to normalize the used features for predictions. Second, a Nonlinear Global Pandemic Machine Learning model was applied for accurate predictions. Results prove that the ML method outperforms LR and AI-based prediction models. In [14], Linear Regression (LR), Least Absolute Shrinkage and Selection Operator (LASSO), Support Vector Machine (SVM), and Exponential Smoothing (ES) all these models were used for forecasting the numbers of three types of cases: newly infected, deaths, and recoveries for the following ten days after that, they were compared with each other. This study concludes that ES performed the best, LR came in second, and LASSO was third, while SVM performed poorly in all three types of cases. In [15], the performances of four models, including auto-regressive integrated moving average (ARIMA), nonlinear auto-regression neural network, and long-short term memory (LSTM), have been compared to forecast confirmed COVID-19 cases. Results show that the LSTM model achieved superior performance. In [16], Behnooda et al. investigated the impacts of weather conditions and the density of the population on the infection rate of the COVID-19. To this end, they used a coupled adaptive network-based fuzzy inference system and virus optimization algorithms. Results reveal that the population density presents the most crucial effect on the performance of the proposed models, which highlights the importance of social distancing to slow down the spread of COVID-19. In [17], Torrealba-Rodriguez et al. conducted a prediction of COVID-19 in Mexico based on daily confirmed cases until May 8th. They applied Gompertz, logistic, and artificial neural networks to model the number of cases of COVID-19 infection and showed that these models could describe well the daily COVID-19 confirmed cases. The method in [18] used reduced-space Gaussian process regression to forecast and evaluate COVID-19 spreading in the United States from January 21th to April 12th, 2020. In [19], Maleki et al. considered an AR time series model based on a two-piece scale mixture normal distributions to forecast recovered and confirmed cases of COVID-19. In [20], Khan et al. applied a vector autoregressive (VAR) model to forecast new confirmed cases, deaths, and recovered cases of COVID-19 in Pakistan. They used data from March 8th to June 27th, 2020, and demonstrated that the VAR model is effective in forecasting COVID-19 spread. However, the major shortcoming of this study is the availability of a small amount of data, which makes the model identification a difficult task.

This study aims to develop some time series models to forecast recovered and confirmed COVID-19 time-series data. Specifically, an Exponential Smoothing (ES) and a Linear Regression (LR) model are presented to forecast the number of recovered and confirmed cases of COVID-19. The spread of this contagious disease is directly linked to existing confirmed cases, while its recovery is restricted by existing treatment capacity, which allows for forecasting without exogenous information. Specifically, we would like to investigate the efficiency of simple models such as ES and LR in forecasting confirmed and recovered COVID-19 cases. Datasets from three countries (Saudi Arabia, India, and France) are used to evaluate and compare the two methods’ forecasting accuracy. First, we build these models (i.e., ES and LR) using COVID-19 time-series datasets from January 22, 2020, to May 27, 2021, and apply them to forecast the COVID-19 cases from May 28, 2021, to June 6, 2021. The forecast data are compared to the actual records of COVID-19 cases. The ES and LR models are assessed using three statistical metrics. Results reveal that the ES can capture the future trend of COVID-19 spread and provides superior forecasting performance compared to the LR model. Decision tree regression model and support vector regression (SVR) with linear kernel. Of course, in this study, we discovered that even using a simple model (i.e., ES), we could forecast the COVID-19 trends successfully.

The remaining of this study is structured as follows. Section II describes the used COVID-19 time-series datasets and briefly overviews the applied forecasting methods (i.e., ES and LR). Section III presents the forecasting results, and Section IV concludes this paper.

II. METHODS AND DATASETS:

A. Datasets:

In the current study, we used the confirmed and recovered cases time-series data from Saudi Arabia, India, and France, recorded from January 22, 2020, to June 6, 2021. These data are available in (https://github.com/CSSEGISandData/COVID-19), and automatically updated for delayed data in the website with no missing value. Figure 1(a-c) displays the confirmed and recovered COVID-19 cases time-series data used in this study. It can be seen that India has the largest number of confirmed cases. Regarding the population in the three countries, India is getting the most significant impact from COVID-19.

B. Methods:

In this study, two improved time series methods are applied to predict the confirmed and recovered cases of COVID-19: ES and LR.

1) Exponential smoothing (ES): Exponential Smoothing is a statistical approach and a successful method used to predict data, especially univariate data, and has shown successful forecasting results. This method depends on the past and current data to predict the future. ES uses the following recessive expiration to perform forecasting:

$$F_t = \alpha x_{t-1} + (1 - \alpha) F_{t-1},$$

(1)
Where $x$ is our COVID-19 time series data, $F_{t+1}$ is the predicted value of the prior data, $\alpha$ is a forgetting or smoothing parameter, and the $\alpha$ value is within $[0, 1]$, and $F_1 = x_1$

2) Linear Regression:: LR is a statistical technique used for determining the relationship between the dependent and independent variables and is commonly used in predicting. The LR model depends on two values: the independent and the dependent variables. By LR, a line is determined between these variables. The equation below presents the LR method used for making predictions:

$$y_t = \beta_0 + \beta_1 x_t + \epsilon,$$  

(2)

here $y_t$ is the forecasted time series data, $\beta_0$ is the intercept, $\beta_1$ is the slope, $x$ is the previous confirmed and recovered cases, and $\epsilon$ is the error term [21].

3) forecasting performance evaluation:: In this paper, MAE, RMSE, and MAPE are employed to assess the forecasting models’ performance.

$$RMSE = \sqrt{\frac{1}{n} \sum_{t=1}^{n} (y_t - \hat{y}_t)^2},$$  

(3)

$$MAE = \frac{1}{n} \sum_{t=1}^{n} |y_t - \hat{y}_t|,$$  

(4)

$$MAPE = 100 \frac{1}{n} \sum_{t=1}^{n} \left| \frac{y_t - \hat{y}_t}{y_t} \right| \%,$$  

(5)

where $y_t$ is the number of COVID cases, $\hat{y}_t$ is its corresponding forecasted values, and $n$ is the number of data points. Lower RMSE and MAE values would imply better precision and forecasting quality.

C. The used forecasting strategy

The general strategy implemented in this work forecasting COVID-19 cases is depicted in Figure 2. At first, the daily recovered and confirmed time-series data are divided into training and test subsets. We first train the models using the training set and evaluate them using the testing set. The performance of these models (i.e., ES, LR, DT, and SVR) are compared by four statistical criteria, namely RMSE, MAE, and MAPE.

III. RESULTS AND DISCUSSION

The ES, LR, DT, and SVR models built based on train COVID-19 data are employed to forecast COVID-19 confirmed and recovered cases for 7 days from May 31, 2021. The recorded test set together with model forecasts of confirmed and recovered cases in the three considered countries are displayed in Figure 3 and Figure 4, respectively.

To quantitatively show the performance of both models, we compute the RMSE, MAE, and MAPE for COVID-19 data recorded in Saudi Arabia, India, and France. Tables I and I show the performances of two models when applied to confirmed and recovered COVID-19 cases, respectively.

In terms of all metrics calculated, the ES showed the best performance. It could be attributed to its flexibility in taking past and actual data in forecasting COVID-19 data.

The overall response is that ES can successfully forecast the future trend of recovered and confirmed COVID-19 cases.
in the three considered countries. On the other hand, LR performs well in both KSA cases and in India’s recovered cases; however, it performs poorly for the rest of the cases. Consequently, ES-based forecasting is better compared to LR for the recovered and confirmed cases. This is mainly due to the capacity of ES to take into consideration the past and actual data in forecasting. Also, we observe that ES outperforms machine learning models (i.e., DT and SVR with linear kernel). This maybe that the machine learning models are not suitable in the presence of small sized dataset.
In this study, to forecast the COVID-19 data, we investigated the performances of simple statistical time series models, including ES and LR. The daily records of confirmed and recovered cases from three countries, namely Saudi Arabia, India, and France, were adopted to test the two models. Four statistical indicators assessed the forecasting accuracy: MAE, RMSE, and MAPE. Metrics showed that the ES displayed superior performance compared to the LR, DT and SVR model. It could be attributed to its capacity to merge information from past and actual data. This study showed that even by using simple statistical models, we could successfully forecast the future trends of COVID-19 spread.

### TABLE I
<table>
<thead>
<tr>
<th>Model</th>
<th>RMSE</th>
<th>MAE</th>
<th>MAPE</th>
</tr>
</thead>
<tbody>
<tr>
<td>ES -Saudi Arabia</td>
<td>256.576</td>
<td>233.517</td>
<td>0.051</td>
</tr>
<tr>
<td>LR -Saudi Arabia</td>
<td>481.413</td>
<td>471.357</td>
<td>0.104</td>
</tr>
<tr>
<td>SVR -Saudi Arabia</td>
<td>22343.127</td>
<td>22342.783</td>
<td>5.167</td>
</tr>
<tr>
<td>DT -Saudi Arabia</td>
<td>9110.550</td>
<td>8884.286</td>
<td>1.993</td>
</tr>
<tr>
<td>ES -India</td>
<td>12024.030</td>
<td>10940.735</td>
<td>0.038</td>
</tr>
<tr>
<td>LR -India</td>
<td>140830.854</td>
<td>140128.058</td>
<td>0.487</td>
</tr>
<tr>
<td>SVR -India</td>
<td>760662.991</td>
<td>760648.180</td>
<td>2.730</td>
</tr>
<tr>
<td>DT -India</td>
<td>1018627.964</td>
<td>997405.833</td>
<td>3.610</td>
</tr>
<tr>
<td>ES -France</td>
<td>808.601</td>
<td>611.512</td>
<td>0.011</td>
</tr>
<tr>
<td>LR -France</td>
<td>18264.854</td>
<td>17653.074</td>
<td>0.312</td>
</tr>
<tr>
<td>SVR -France</td>
<td>149980.040</td>
<td>149912.473</td>
<td>2.733</td>
</tr>
<tr>
<td>DT -France</td>
<td>33756.013</td>
<td>29323.967</td>
<td>0.520</td>
</tr>
</tbody>
</table>

### TABLE II
<table>
<thead>
<tr>
<th>Model</th>
<th>RMSE</th>
<th>MAE</th>
<th>MAPE</th>
</tr>
</thead>
<tbody>
<tr>
<td>ES -Saudi Arabia</td>
<td>272.035</td>
<td>246.032</td>
<td>0.056</td>
</tr>
<tr>
<td>LR -Saudi Arabia</td>
<td>471.915</td>
<td>457.749</td>
<td>0.104</td>
</tr>
<tr>
<td>SVR -Saudi Arabia</td>
<td>23763.726</td>
<td>23761.868</td>
<td>5.743</td>
</tr>
<tr>
<td>DT -Saudi Arabia</td>
<td>100942.461</td>
<td>9808.333</td>
<td>0.293</td>
</tr>
<tr>
<td>ES -India</td>
<td>42792.414</td>
<td>38985.991</td>
<td>0.146</td>
</tr>
<tr>
<td>LR -India</td>
<td>58289.970</td>
<td>54230.752</td>
<td>0.202</td>
</tr>
<tr>
<td>SVR -India</td>
<td>755010.759</td>
<td>755004.471</td>
<td>2.912</td>
</tr>
<tr>
<td>DT -India</td>
<td>1965186.319</td>
<td>1936257.083</td>
<td>7.824</td>
</tr>
<tr>
<td>ES -France</td>
<td>164.913</td>
<td>119.239</td>
<td>0.036</td>
</tr>
<tr>
<td>LR -France</td>
<td>873.247</td>
<td>767.689</td>
<td>0.228</td>
</tr>
<tr>
<td>SVR -France</td>
<td>7527.941</td>
<td>7512.021</td>
<td>2.293</td>
</tr>
<tr>
<td>DT -France</td>
<td>4954.377</td>
<td>4804.667</td>
<td>1.455</td>
</tr>
</tbody>
</table>

### IV. CONCLUSION

In this study, to forecast the COVID-19 data, we investigated the performances of simple statistical time series models, including ES and LR. The daily records of confirmed and recovered cases from three countries, namely Saudi Arabia, India, and France, were adopted to test the two models. Four statistical indicators assessed the forecasting accuracy: MAE, RMSE, and MAPE. Metrics showed that the ES displayed superior performance compared to the LR, DT and SVR model. It could be attributed to its capacity to merge information from past and actual data. This study showed that even by using simple statistical models, we could successfully forecast the future trends of COVID-19 spread.

### REFERENCES


