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.
ARIMA and NAR based prediction model for time series analysis of COVID-19 cases in India

Farhan Mohammad Khan a,∗, Rajiv Gupta b

a Research Scholar, Department of Civil Engineering, BITS Pilani, Pilani Campus, India
b Senior Professor, Department of Civil Engineering, BITS Pilani, Pilani Campus, India

A R T I C L E   I N F O

Keywords:
Time series
Novel coronavirus
SARS-CoV-2
Forecasting
ARIMA
NAR
India

A B S T R A C T

In this paper, we have applied the univariate time series model to predict the number of COVID-19 infected cases that can be expected in upcoming days in India. We adopted an Auto-Regressive Integrated Moving Average (ARIMA) model on the data collected from 31st January 2020 to 25th March 2020 and verified it using the data collected from 26th March 2020 to 04th April 2020. A nonlinear autoregressive (NAR) neural network was developed to compare the accuracy of predicted models. The model has been used for daily prediction of COVID-19 cases for next 50 days without any additional intervention. Statistics from various sources, including the Ministry of Health and Family Welfare (MoHFW) and http://covid19india.org/ are used for the study. The results showed an increasing trend in the actual and forecasted numbers of COVID-19 cases with approximately 1500 cases per day, based on available data as on 04th April 2020. The appropriate ARIMA (1,1,0) model was selected based on the Bayesian Information Criteria (BIC) values and the overall highest R2 values of 0.95. The NAR model architecture constitutes ten neurons, which was optimized using the Levenberg-Marquardt optimization training algorithm (LM) with the overall highest R2 values of 0.97.

Introduction

COVID-19

The outbreak of transmissible person-to-person pneumonia caused by the extreme acute respiratory coronavirus 2 syndrome (SARS-COV-2, also known as COVID-19), has sparked a global warning. The COVID-19 virus spreads mainly by droplets of saliva or nose discharge while an infected person is coughing or sneezing. Milan Batista [1] used the logistic model in prediction of the total number of cases and peak time of the coronavirus epidemic in China, South Korea, and the rest of the World, which gave a reasonable description of the epidemic in those places. A time series is a sequence where a pattern is recorded over regular time intervals. In this paper, we have applied the univariate time series model to predict the number of COVID-19 infected cases that can be expected in upcoming days in India.

ARIMA model

The ARIMA method is also known as the Box-Jenkins method [2]. The Box-Jenkins method relates to the fitting of a mixed ARIMA model to a given data set. ARIMA, short for ‘Auto-Regressive Integrated Moving Average’ is a class of models explaining a given time series based on its past values, i.e., its own lags and the lagged forecast errors, so that equation can be used to forecast future values. An ARIMA model is characterized by 3 terms: p, d, q.

Where,

\[ p \] is the order of the AR term.
\[ q \] is the order of the MA term.
\[ d \] is the number of differencing required to make the time series stationary.

The most common approach to make a series stationary is to subtract the previous value from the current value. Sometimes more than one differencing may be required, depending on the complexity of the series. Therefore, the value of d is the minimum amount of differentiation needed to render the sequence stationary, and if the time series is stationary already, then \( d = 0 \). The ‘p’ is the order of ‘Auto-Regressive’ (AR) term; it refers to the number of Y lags which should be used as predictors. The ‘q’ is the order of the ‘Moving Average’ (MA) term; it refers to the number of lagged errors in the forecast that should go into the ARIMA model. The principal objective of the fitting ARIMA model is to correctly recognize the stochastic mechanism of the time series and forecast future values. Such approaches have also proven useful in other types of scenarios in which models for discrete-time series and dynamic systems are created. This method is, however, not appropriate for lead times or seasonal series with a broad random variable [3,4]. This univariate time series model has been used in the present study to forecast the number of COVID-19 cases in India. The steps of the ARIMA model building methodology are presented in a flowchart below in Fig. 1.

https://doi.org/10.1016/j.jnlssr.2020.06.007

Received 18 April 2020; Received in revised form 27 May 2020; Accepted 5 June 2020

© 2020 China Science Publishing & Media Ltd. Publishing Services by Elsevier B.V. on behalf of KeAi Communications Co. Ltd. This is an open access article under the CC BY-NC-ND license. (http://creativecommons.org/licenses/by-nc-nd/4.0/)
Artificial neural network based nonlinear forecasting

Nonlinear autoregressive (NAR) neural networks were designed to forecast a time series from past values [5]. In this study, the NAR based interface was developed using the app designer programming environment of MATLAB software [6] using the standard commands. NAR based networks can be built trained and predict future values [7]. Our objective is for users who do not have any programming experience to predict future values over a time series. The steps of NAR model building methodology are presented in a flowchart below in Fig. 2.

Methodology

ARIMA model

- Identification of the model: This includes identifying the most suitable lags for the components of the AR and MA, and deciding whether the variable needs first differencing to induce stationarity. The Auto Correlation (ACF) function and the Partial Auto Correlation (PACF) function are used to determine the best model.
- Estimation: This usually involves the use of a least-squares estimation process.
- Diagnostic testing: This usually is the test for autocorrelation. If this part fails, then the process returns to the identification section and begins again, usually by the addition of extra variables.
- Forecasting: The ARIMA models are particularly useful for forecasting due to the use of lagged variables.

Once the model was identified, and the model parameter can be estimated, then the model is determined with a different set of parameters. It is basically checked with the assumption that the model about the random error is satisfied using statistical diagnostic tests and residual plots that can be used to analyze the suitability of various models to historical data. The model selection can be made based on the values of specific criteria like Normalized Bayesian Information Criteria (BIC).

Nonlinear autoregressive neural networks (NAR)

The data containing confirmed COVID-19 cases is used for the time series analysis. The statistical calculations for raw data include data count, mean, standard deviation, maximum and minimum values, and mode. The dataset is divided into training (70%), testing (15%), and validation (15%). The NAR model optimization is performed using three training functions, TRAINBR, TRAINLM, and TRAINSCG. The performance of the model is measured using the value of the determination coefficient ($R^2$); it reflects the proportion of variance in the dependent parameter described by the independent parameter. The higher value of $R^2$ indicates that the model explains variation in the dependent parameter. Mean Squared Error (MSE) is one of several ways of quantifying the difference between the values predicted and the real values of the measured quantity.

Analysis, results, and discussions

Data regarding the number of cases reported in India till 04th April 2020, data were collected from the Ministry of Health and Family Welfare (MoHFW) and covid19india.org [8]. This data was plotted on a graph to see the trend, as shown in Fig. 3. The ACF stands for Autocorrelation function, and the PACF for Partial Autocorrelation function. Autocorrelation computes and plots the autocorrelations of a time series. Autocorrelation is the correlation between observations of a time series separated by k time units, while the ACF is the plot used to show the correlation between the points, up to and including the lag unit. In ACF, the coefficient of correlation is in the x-axis whereas the number of lags in the y-axis is shown. Similarly, partial autocorrelations measure the strength of the relationship with other terms that are accounted for. In this case, other terms are the lags in the model that are intervening.

A partial autocorrelation is a combination of the relationship between an observation in a time series with observations being excluded at the initiation phase with the relationships of intervening observations. Auto Correlation Function (ACF) is given in Fig. 4 shows that the series has positive autocorrelations to a large number of lags, i.e., 10, so a higher order of differentiation is required. Fig. 5 shows that the
Fig. 3. Time series plot of COVID-19 cases as on 04th April 2020.

Fig. 4. Autocorrelation factor.

Fig. 5. Partial autocorrelation factor.
autocorrelation of lag-1 is small and patternless, so the series does not need a higher order of differencing. If the autocorrelation of lag-1 is zero or more negative, then the series may be over differenced. The partial autocorrelation function (PACF) of the differenced series shows a sharp cutoff due to the positive lag-1 autocorrelation, and the series appears to be slightly under differenced, so one or more AR terms should be added to the model. The lag beyond which the PACF cutoff is the number of AR terms indicated. The ACF and PACF showed an irregular increasing pattern in the number of cases of COVID-19.

Henceforth, ARIMA models (p, d, q), apt for such a scenario was applied. In terms of choosing a Box-Jenkins model, the smaller the goodness-of-fit measures the better. The best suitable Box-Jenkins model was selected based on minimal Bayesian Information Criteria (BIC) value. In this study, the least BIC value is 354.4367 as given below in Table 1, and the corresponding model is ARIMA (1,1,0) with the overall highest R² values of 0.95, as shown in Fig. 6.

The verification of the model is done by checking the residuals of the model using autocorrelation and correlation of the residuals of different orders. For a time series model, the residuals are proportional to the difference between the measurements and the corresponding fit values. Residuals are used for testing the model’s suitability to capture the information in the data. The estimated autocorrelations and partial autocorrelations between the residuals at various lags are depicted in Fig. 7. It is evident from the figure that all the lags are well within the 95% confidence level. This implies that residuals are random, i.e., white noise, indicating that the model is a good fit. It is also observed that all autocorrelation coefficients are not statistically significant, implying that residuals are not autocorrelated with each other. A model with the lowest value of normalized BIC is found to be ARIMA (1,1,0), which can be considered as the best fit model and can be further used to generate the forecasts. Based on the ACF and PACF, the daily prediction of COVID-19 cases is calculated, as shown in Fig. 8.

Using the fitted model to generate MMSE forecasts corresponding mean square errors over a 50-day horizon, as shown in Fig. 9. The
predicted ARIMA model is compared with Monte Carlo Forecasts for validation. Results show that the forecast for MMSE and the mean for simulation are practically indistinguishable, as shown in Fig. 10. Slight differences occur between the theoretical 95% forecast intervals and the simulation-based 95% forecast intervals.

**NAR model**

The trial and error method is implemented to achieve the optimal network structure in which a rigorous analysis with one hidden layer is carried out. Fig. 11 shows the architecture of the NAR neural network. The back-propagation of error is carried out by the Bayesian regularization (BR), which is a standard second-order nonlinear least-squares technique using the back-propagation process to increase the speed and efficiency of the training. Fig. 12 shows the performance graph obtained for
Fig. 13. Close Loop Non-Linear Auto-regressive Neural Network.

Fig. 14. Open Loop Non-Linear Auto-regressive Neural Network.

Fig. 15. Error Histogram of trained NAR neural network.

Fig. 16. Time-Series Response of NAR neural network.

the model; the best validation performance of the model was 168.8105 obtained at epoch 34. The closed-loop training form of the created neural network is shown in Fig. 13. After the closed-loop training period, one step ahead form is created using an open loop, as shown in Fig. 14. Error histogram is plotted in Fig. 15, error histogram is the histogram of the errors between target values and predicted values after a neural network has been trained. These error values reflect how the expected values vary from the target values, and they may also be negative. The results from the error histogram show that the maximum correlations at zero lag, the time-series response of the NAR model, is shown in Fig. 16.

Comparison of ARIMA and NAR model

The data related to is collected from 30th Jan. Subsequently, the probable variation for the next 50 days is forecasted using the two different techniques, as discussed above. In this study machine, learning-based algorithms are proposed to observe the transmission pattern of
COVID-19. Performance evaluation of the forecast is estimated in terms of Root Mean Square Error (RMSE) and coefficient of determination, and the statistics are found to be within satisfactory limits. The accuracy of the ARIMA and NAR model developed is compared for the daily predicted confirmed COVID-19 cases in India, as shown in Fig. 17. It can be seen clearly from the curve that at the end of June, daily new confirmed cases will be almost 1500 cases per day if the conditions remain the same in the country.

As early mentioned above, the ARIMA model consists of four steps. The first step was the identification of the model. The model identification is made by ACF and PACF (Fig. 4 and Fig. 5). The time-series response of the NAR neural network model was observed using an open loop and close loop network. Both the models revealed an increasing pattern in the number of cases of COVID-19. Model parameters were estimated using MATLAB2019b. In the forthcoming days, the model predicted a steady increase. The first case of COVID-19 is reported in India on 30th January 2020, is expected to increase approximately 1500 cases per day in the next 50 days. This implies that the existing interventions in the country will not be adequate to control the on-going epidemic. Based on the forecasting model and assuming the current conditions of the COVID-19 outbreak, it is expected a quick and exponential increase in the number of cases.

**Conclusion**

This study considered data related to daily confirmed cases of COVID-19 in India. Based on available data as on 04th April 2020, the daily predicted number of COVID-19 cases in India using the ARIMA model and NAR neural network is approximately 1500 cases per day for the next 50 days. This prediction is made with existing conditions. However, it can be improved by taking a few preventive steps. We intend to further improve our model by collecting more data from the upcoming days. For future modifications, the new methodology will be used for forecasting the scenario of confirmed, deceased, and recovered cases of COVID-19.

**Declarations of Competing Interest**

None.

**References**

[1] Batista, M. (2020). Estimation of the final size of the COVID-19 epidemic. Preprint.J medRxiv.
[2] S. Bhatnagar, V. Lal, S.D. Gupta, O.P. Gupta, Forecasting incidence of dengue in Rajasthan, using time series analyses, Indian J. Public Health 56 (4) (2012) 281.
[3] M. Widenström, M. Omsberg, M. Fern, A.K. Pettersson, M.R. Eriksson, I. Eckerdal, J. Wiström, Autoregressive integrated moving average (ARIMA) modeling of time series of local telephone triage data for syndromic surveillance, Online J. Public Health Inform. 6 (1) (2014).
[4] M. Manikandan, A. Velavan, Z. Singh, Forecasting the trend in cases of Ebola virus disease in west African countries using auto regressive integrated moving average models, Int. J. Commun. Med. Public Health 3 (2016) 615–618.
[5] D. Sayeedkhale, Nonlinear autoregressive neural network with exogenous inputs based solution for local minimum problem of agent tracking using quadrotor, Инженерный вестник Дона 29 (2) (2014).
[6] MathWorks: Neural Network Toolbox Release 2018b– MATLAB & Simulink – MathWorks India, available at, last access: 05th April 2020.
[7] I. Kirbas, NAR based forecasting interface for time series analytic t-seer, IV Int. Conf. Eng. Nat. Sci. (ICENS) (2018) 144–149.
[8] Ministry of Health and Family Welfare, https://www.mohfw.gov.in/ and http://Covid19india.org/ portal, last access: 05th April 2020.