British Journal of Multidisciplinary and Advanced Studies:
Health and Medical Sciences 4 (5),127-135, 2023
Print ISSN: 2517-276X
Online ISSN: 2517-2778
Website: https://bjmas.org/index.php/bjmas/index
Published by European Centre for Research Training and Development UK

On Predicting Growth Factor Data of Covid-19 Epidemic Using Hybrid Arima-Ann Model

Samir K. Safi
United Arab Emirates University
Email: ssafi@uaeu.ac.ae

doi: https://doi.org/10.37745/bjmas.2022.0335 Published October 22 2023

Citation: Samir K. Safi (2023) On Predicting Growth Factor Data of Covid-19 Epidemic Using Hybrid Arima-Ann Model, *British Journal of Multidisciplinary and Advanced Studies*: Health and Medical Sciences 4 (5),127-135

ABSTRACT: *The Autoregressive Integrated Moving Average (ARIMA) model cannot capture the nonlinear patterns exhibited by the 2019 coronavirus (COVID-19) in terms of daily growth factor. As a result, Artificial Neural Networks (ANNs) and Hybrid ARIMA-ANN models have been successfully applied to resolve problems with nonlinear estimation. We compare the forecasting performance of these models using real, worldwide, daily COVID-19 data. The best forecasting model selected was compared using the forecasting assessment criterion known as mean absolute error. The main finding results show that the ANN model is more efficient than the ARIMA and Hybrid ARIMA-ANN models. The main finding from the ANN model analysis indicates that the magnitude of the increase in growth factor over time is rising in general while the percentage change in the growth factor is declining. This may be the result of the social distancing, safety, and cautionary measures mandated by governments worldwide.*

KEYWORDS: forecasting, Covid-19, hybrid model, Ann, Arima.

LITERATURE REVIEW

In this paper, we aim to study the growth factor of COVID-19 using different models including the autoregressive integrated moving average (ARIMA), the artificial neural networks (ANNs), and the Hybrid ARIMA-ANN models to forecast the spread of COVID-19 around the world for the next 17 days using currently available data. The forecast of growth factor is presented to discover what should be expected in the coming days as well as to determine the best forecasting method.

Evidence from previous studies suggests that the hybrid model performs better than the linear or nonlinear models. For example, Wang. et. al. (2013) suggested that the linear ARIMA model and the nonlinear ANNs model were employed jointly, with the aim of capturing the different patterns in the time series data. They showed the effectiveness of the hybrid model (the multiplicative model) of ARIMA and ANNs models in obtaining more accurate forecasting as compared to ARIMA and ANNs models (Benvenuto et. al., 2020) indicated the effectiveness of the ARIMA model in predicting the epidemiological trend of the prevalence and incidence of COVID-2019. They mentioned that ARIMA (1,0,4) was chosen as the best ARIMA model for predicting the spread of COVID-19, while ARIMA (1,0,3) was selected as the best ARIMA model for determining the incidence of COVID-19.

(Ceylan, 2020) showed that ARIMA models are suitable for predicting the prevalence of COVID-19 in Italy, Spain, and France. The study formulated different ARIMA models with different ARIMA parameters and selected the best models based on the lowest MAPE values. (Perone, 2020) showed that the ARIMA models are trustworthy enough for forecasting COVID-19 incidence in Italy, Russia, and the USA when new daily cases begin to stabilize. (Saba and Elsheikh, 2020) showed that nonlinear autoregressive artificial neural networks (NARANN) have a better performance compared with ARIMA based on different statistical criteria such as mean absolute error (MAE), root mean square error (RMSE), mean absolute percentage error (MAPE), determination coefficient, deviation ratio (RD), and coefficient of residual mass (CRM). In this paper, an ARIMA, ANNs, and a combination of ARIMA and ANNs (hybrid model) were proposed to make forecasts of the growth factor of COVID-19 time series data.

The remainder of this article is organized as follows. Section 2 provides the Description of the dataset. The Forecasting models are presented in section 3. The Empirical Study for the training and test datasets using three different time series models is discussed in section 4, and the final section concludes by summarizing the key results.

DESCRIPTION OF THE DATASET

There were 123,450,040 confirmed cases and 2,722,302 deaths (2.21%) from the coronavirus COVID-19 outbreak as of March 21, 2021, 07:16 GMT. There were 21295138 (17.25%) cases currently infected and 102,154,902 (82.75%) cases with outcomes. Among the infected cases, there are 21,205,071 (99.58%) cases that are in mild condition, and 90,067 (0.42%) cases are in serious or critical condition. In addition, among cases with the outcome, there are 99,432,600 (97.34%) cases recovered/discharged, and 2,722,302 (2.66%) deaths (Worldometer, (2021).

We consider the growth factor from January 24, 2020, to March 20, 2021. The Growth factor is the factor by which a quantity multiplies itself over time. The formula used for the growth factor is given by

$$Growth\,Factor = \frac{Every\ day's\ new\ cases}{New\ cases\ on\ the\ previous\ day} \qquad (1)$$

A ratio of the growth factor greater than one indicates exponential growth and one which remains smaller than 1 is a signal of decreasing.

The growth factor ranged between 0.2840 and 6.9171 with a mean of 1.0395, a median of 1.0190, and a standard deviation of 0.3276 (with an interquartile range of 0.1725). These summaries indicate that the spread of growth factor varies among the various countries. Figure 1 shows the trend of the growth factor of COVID-19. From this plot, we can see that the growth factor is not linear over time and shows large fluctuations. We must be cautious using ARIMA models as they may not provide accurate forecasts in this case.

British Journal of Multidisciplinary and Advanced Studies:
Health and Medical Sciences 4 (5),127-135, 2023
Print ISSN: 2517-276X
Online ISSN: 2517-2778
Website: https://bjmas.org/index.php/bjmas/index
Published by European Centre for Research Training and Development UK

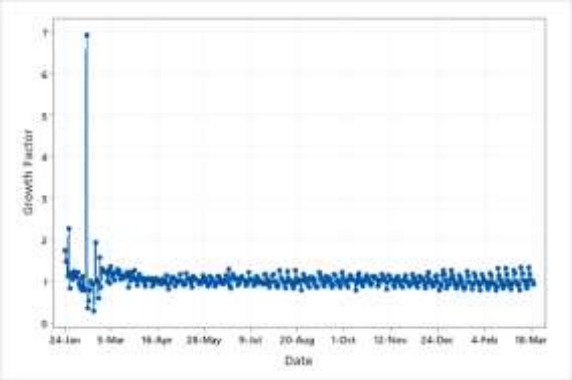

Figure 1: Growth Factor of COVID-19

In this study, 10% of the sample size is used as the testing sample. A training sample is used for the model building, and the testing sample is used for the model validation at the end of the analysis. We considered the first 380 observations as training sample over the period, of Jan 24, 2020 – Feb. 6, 2021, and 42 observations as testing sample over the period, of Feb. 7, 2021 – March 20, 2021.

The accuracy of forecasts can be determined by considering how well a model performs on new data that were not used when fitting the model. Accuracy is an important issue in forecasting; therefore, researchers tend to add more and more variables to their proposed model. Safi. and White. (2017) considered the issue of whether a complex model actually does a better job than a simple one.

Several measures of forecasting accuracy have been developed and discussed, the fundamental usage of these measurements compared the accuracy of forecasting methods with univariate time series data (Cryer and Chan, 2008; Hyndman. and Athanasopoulos, 2018; and Wei, 2006). The best forecasting models selected will be compared using one of the three different forecasting accuracy measuring criteria: MAE, RMSE, and MAPE. RMSE is more sensitive to outliers than the MAE, which is preferable in cases of the existence of outliers. Using the MAE or RMSE is recommended when comparing forecasting methods on a single data set. This means the MAE and RMSE should be used if all forecasts are measured on the same scale. The MAPE is used when comparing the accuracy of the same or different methods on different time series data with different scales unless the data contain zeros or small values (Hyndman. and Koehler, 2006). The evaluation criterion for these measures of forecasting accuracy is that the smaller the value obtained, the better the model's forecasting ability (McKenzie, 2011).

The efficiency of the proposed forecasting method relative to that of the benchmark method in terms of the RMSE is defined by

$$\Omega = \frac{RMSE_p}{RMSE_b}, \qquad (2)$$

where, $RMSE_p$ and $RMSE_b$ represent the RMSE from the proposed and the benchmark methods, respectively. Usually, the benchmark method is the most naïve method (Hyndman and Koehler, 2006). A ratio of less than one indicates that the forecasting performance of the proposed method is more efficient than the benchmark method and if this ratio is close to one, then the proposed forecasting method is nearly as efficient as the benchmark forecasting method. Otherwise, the proposed method performed poorly (White and Safi, 2016; and Safi, 2013).

British Journal of Multidisciplinary and Advanced Studies:
Health and Medical Sciences 4 (5),127-135, 2023
Print ISSN: 2517-276X
Online ISSN: 2517-2778
Website: https://bjmas.org/index.php/bjmas/index
Published by European Centre for Research Training and Development UK

FORECASTING MODELS

The ARIMA Model

The general ARIMA (p,d,q) model is given (by Box et. al. (2015)):

$$\phi(B)\nabla^d Y_i = \theta(B)\varepsilon_i, \tag{3}$$

where, d is the degree of differencing, $\nabla = 1 - B$ is the differencing operator, and the *lag* operator B is defined as $BY_t = Y_{t-1}$, the operator that provides the previous value of the series. $\phi(B)$ and $\theta(B)$ are polynomials of degrees p and q in B,

$$\phi(B) = 1 - \phi_1 B - \phi_2 B^2 - \cdots - \phi_p B^p, \tag{4}$$

and

$$\theta(B) = 1 - \theta_1 B - \theta_2 B^2 - \cdots - \theta_q B^q. \tag{5}$$

The best ARIMA model is chosen according to its Akaike information criterion (AIC), AICc, or BIC value.

The ANN Model

The *nnetar* function is used to fit neural networks. This function is described as a feed-forward neural network with a single hidden layer and lagged inputs for forecasting univariate time series. The *nnetar* function fits a neural network autoregressive $\mathrm{NNAR}(p,P,k)$ model. For a non-seasonal time series, the default is the optimal number of lags, according to the AIC value, for a linear autoregressive $\mathrm{AR}(p)$ model. For a seasonal time series, the default values is $P = 1$ where p is chosen from the optimal linear model fitted to the seasonally adjusted data and $k = \frac{1}{2}(p + P + 1)$ (rounded to the nearest integer). By default, 25 networks with random starting values are trained and their predictions are averaged (by Hyndman, 2004).

The Hybrid Model

The hybrid model fits multiple individual model specifications to enable the easy creation of ensemble forecasts. The hybrid model consists of a combination of three models: the ARIMA, the exponential smoothing, and the ANN models.

Looking at a time series composed of autocorrelated linear and nonlinear components, we have:

$$y_t = L_t + N_t \tag{6}$$

Fitting $\widehat{L_t}$ using the ARIMA model with e_t as the residual yields:

$$e_t = y_t - \widehat{L_t} \tag{7}$$

The error term consists of nonlinear relationships with previous errors. The nonlinear relationships can be modeled from the past residuals as follows:

$$e_t = f(e_{t-1}, e_{t-2}, \ \ldots, e_{t-n}) + \varepsilon_t \tag{8}$$

Then, using an ANN model to predict e_t as an estimate for N_t, we can calculate the forecast:

British Journal of Multidisciplinary and Advanced Studies:
Health and Medical Sciences 4 (5),127-135, 2023
Print ISSN: 2517-276X
Online ISSN: 2517-2778
Website: https://bjmas.org/index.php/bjmas/index
Published by European Centre for Research Training and Development UK

$$\widehat{y_t} = \widehat{L_t} + \widehat{N_t} \tag{9}$$

These models were built using *hybridModel* command in the *forecastHybrid* package. This package fits multiple models and combines them using either equal weights or weights based on in-sample errors. There are six models: ARIMA, exponential smoothing, theta, neural network autoregression (NNAR), seasonal and trend decomposition, and the trigonometric seasonal + exponential smoothing method + Box-Cox Transformation (TBATS) model for heterogeneity and the autoregressive moving average model for residuals + trends + seasonal (including multiples and non-integer periods).

EMPIRICAL STUDY

This section presents the empirical results for the training and test datasets of the models used to forecast the growth factor using three different time series models; the ARIMA, ANN models, and the hybrid combination of the two. The forecasting results are presented in the following sub-sections.

We carried out the Anderson-Darling normality test to determine if the data followed a normal distribution (Thode, (2002. The Anderson-Darling (AD) test is an empirical distribution function omnibus test for the composite hypothesis of normality. The test statistic is:

$$\mathrm{AD} = -n - n^{-1}\sum_{i=1}^{n}(2i-1)\left[\ln p_i + \ln\left(1 - p_{n-i+1}\right)\right], \tag{10}$$

where, $p_i = \Phi\left(S^{-1}(x_i - \bar{x})\right)$. Here, Φ is the cumulative distribution function of the standard normal distribution and $\bar{x}$ and S are the mean and the standard deviation, respectively, of the data values.

For using the training dataset, the normality test for the growth factor residuals' of COVID-19 data yielded the AD values of 0.80794, 1.0139, and 0.35662, with corresponding p-values of 0.03356, 0.0102, and 0.4405 for ARIMA, ANN, and Hybrid models, respectively.

Therefore, this result indicates that the normality assumption was satisfied at a 0.01 level of significance for all residuals of the three selected models.

Since the data are normally distributed, to compare the performance of the models for the given dataset, we used the forecasting accuracy measure, the RMSE, over the forecasting period for each model. The smaller values of RMSE indicate higher forecasting accuracy. Therefore, the ratios of the RMSE of the ANN to those of the ARIMA, and hybrid models were calculated for analysis.

Table 1 lists the empirical results for the RMSE, MAE, and MAPE and the ratios of the ANN model's RMSE, MAE, and MAPE to those of the ARIMA and hybrid models for growth factor. The ratios of the RMSE of the ANN model to those of the ARIMA and hybrid models were calculated for analysis.

Applying the ANN model for with an average of 1,000 networks, each of which is a 22-25-1 network, with 601 weights and an estimated noise variance of 0.000655. This indicates that 25 networks were trained and that their predictions were averaged. For the ARIMA model, the result shows that the best-fit model was the ARIMA (1,1,2) with drift $\mu = -0.0008$ and the equation is given by

$$w_t = \theta_0 + 0.70756 w_{t-1} + e_t + 1.9530 e_{t-1} - 0.9624 e_{t-1} \tag{14}$$

British Journal of Multidisciplinary and Advanced Studies:
Health and Medical Sciences 4 (5),127-135, 2023
Print ISSN: 2517-276X
Online ISSN: 2517-2778
Website: https://bjmas.org/index.php/bjmas/index
Published by European Centre for Research Training and Development UK

where $w_t = y_t - y_{t-1}$, $\theta_0 = \mu(1-\phi_1) = -0.000234$, with estimated noise variance of 0.1033, AIC=221.84, AICc=222.0, and BIC=241.51.

Table 1: Forecasting Criteria and Ratios of the RMSE, MAE, and MAPE for the ANN model to the ARIMA and hybrid models.

Statistics	ARIMA	ANN	Hybrid	ANN/ARIMA	ANN/hybrid
RMSE	0.3169	0.0255	0.0744	0.0805	0.3427
MAE	0.1312	0.0165	0.0561	0.1260	0.2947
MAPE	12.5308	1.6526	5.9294	0.1319	0.2787

The RMSE of the ARIMA, ANN, and hybrid models equal 0.3169, 0.0255 and 0.0744, respectively. This result indicates that the relative efficiencies of the ANN model to the ARIMA and hybrid models equaled Ω=0.0805, and 0.3427, respectively. This result indicates that the ANN model's RMSEs equaled 8.05% and 34.27% of that of the ARIMA and hybrid models, respectively. Therefore, the ANN model was more efficient than the ARIMA and hybrid models for the growth factor of COVID-19 data. Therefore, we could use the ANN model since it outperforms the ARIMA and the hybrid models. However, as a second choice in this case, we could use the Hybrid model since it is outperforming the ARIMA, keeping in mind that it is not a perfect substitute.

The forecast for growth factor using the ANN model is compared with the actual values as shown in Figure 2. From the plot we observe the closeness of forecast values using ANN model to the actual values, indicating that the selected model in forecasting the growth factor is relatively close to the actual values, hence, this can be reliable for policy implementation. This substantiates the valid use of the model. In Figure 3, we show the percentage changes in growth factor calculated using the actual testing data (Feb. 7 – March 20). Figures 2 and 3 show that the daily growth factor is fluctuating over time in general. This result indicates that the growth factor is not stable over time.

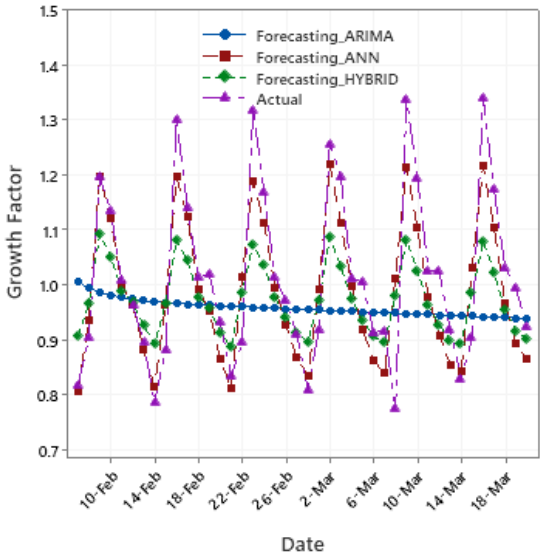

Figure 2: Forecasts and Actual values in Growth factor.

British Journal of Multidisciplinary and Advanced Studies:
Health and Medical Sciences 4 (5),127-135, 2023
Print ISSN: 2517-276X
Online ISSN: 2517-2778
Website: https://bjmas.org/index.php/bjmas/index
Published by European Centre for Research Training and Development UK

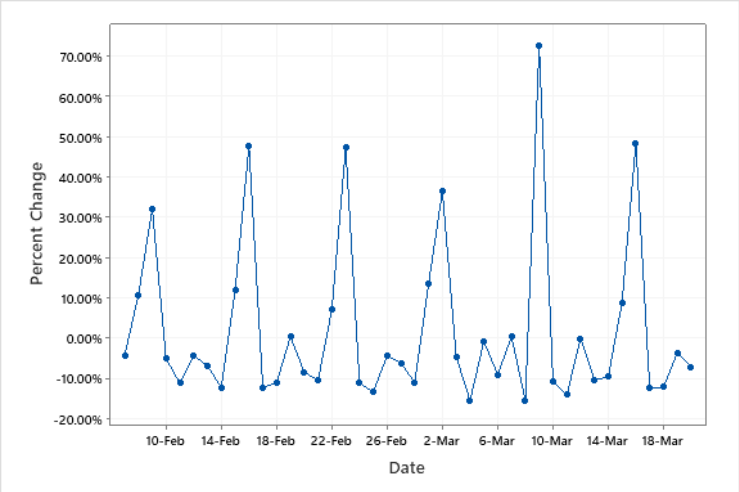

Figure 3: Actual Percentage changes for Growth factor

After feeding the model with the data from Feb. 7 – March 20, and repeating the procedure using ANN "The best-chosen model", we show the forecasts in the growth factor for the next 41 days (March 21 – April 30), this is shown in Table 3 and Figure 3. In the forecast for March 21st till April 30th, the daily growth factor is shown to be fluctuating over time in general. The growth factor exhibits a fluctuating trend which decreases our optimism as the fight to contain COVID-19 continues. The chosen forecasting model which is relatively higher prepares policymakers for the worst-case scenario as we go through the next wave of the pandemic.

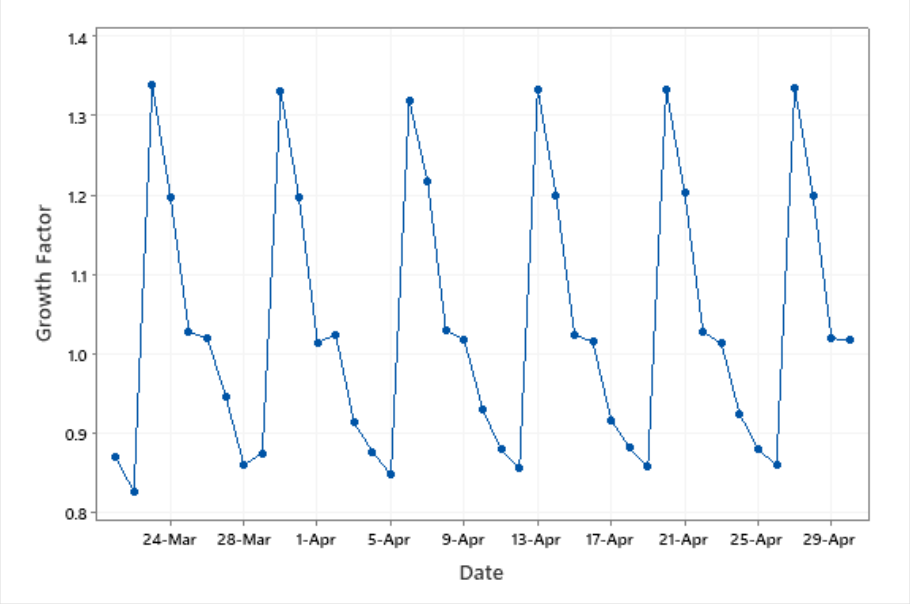

Figure 3: Forecasts in Growth Factor for the full data Using ANN model.

Table 3: Forecasts in Growth Factor for the full data Using ANN model.

Date	Forecast	Date	Forecast	Date	Forecast
21-Mar	0.8706	4-Apr	0.8773	18-Apr	0.8818
22-Mar	0.8266	5-Apr	0.8481	19-Apr	0.8587
23-Mar	1.3394	6-Apr	1.3194	20-Apr	1.3339
24-Mar	1.1974	7-Apr	1.2176	21-Apr	1.2033
25-Mar	1.0281	8-Apr	1.0298	22-Apr	1.0289
26-Mar	1.0204	9-Apr	1.0180	23-Apr	1.0134
27-Mar	0.9470	10-Apr	0.9303	24-Apr	0.9247
28-Mar	0.8603	11-Apr	0.8799	25-Apr	0.8803
29-Mar	0.8755	12-Apr	0.8557	26-Apr	0.8599
30-Mar	1.3313	13-Apr	1.3337	27-Apr	1.3355
31-Mar	1.1968	14-Apr	1.2003	28-Apr	1.1995
1-Apr	1.0152	15-Apr	1.0251	29-Apr	1.0197
2-Apr	1.0239	16-Apr	1.0152	30-Apr	1.0179
3-Apr	0.9137	17-Apr	0.9168		

CONCLUSION

We used the ANN to forecast the growth factor of COVID-19. Forecasting performance was compared for different models using real daily data for COVID-19 around the world in the upcoming days. We discussed various forecasting techniques for choosing the best forecasts for growth factors for COVID-19.

The results show that the ANN performed better than the ARIMA and ARIMA-ANN Hybrid models. This is not a surprising result because ANN is designed to capture the nonlinear trend of the data that were exhibited during the time period of the sample. These results add to the growing body of literature that seeks to accurately forecast the spread of COVID-19 by combining multiple models used by other researchers.

The results are useful because they provide an accurate forecast for growth factors for the COVID-19 pandemic. All Governments and institutions involved in public health can benefit from these results for forecasting purposes using more a reliable and accurate forecast model for the novel COVID-2019. The additional value of results is not encouraging as the world struggles to contain the spread of COVID-19.

The growth factor exhibits a fluctuating trend which decreases our optimism as the fight to contain COVID-19 continues. Further research could be carried out in this area by studying the impact of COVID-19 on economic variables using the most appropriate forecasting techniques. In addition, these results can be used to make a relationship between its forecasts and some economic variables.

REFERENCES

Benvenuto, D., Giovanetti, M., Vassallo, L., Angeletti, S., & Ciccozzi, M. (2020). Application of the ARIMA model on the COVID-2019 epidemic dataset. Data in brief, 105340.

Box, G. E., Jenkins, G. M., Reinsel, G. C., & Ljung, G. M. (2015). Time series analysis: forecasting and control. John

Wiley & Sons.
Ceylan, Z. (2020). Estimation of COVID-19 prevalence in Italy, Spain, and France. Science of The Total Environment, 138817.
Cryer, J. D., & Chan, K. S. (2008). Time series regression models. Time series analysis: with applications in R, 249-276.
Hyndman, R. J., & Athanasopoulos, G. (2018). Forecasting: principles and practice. OTexts.
Hyndman, R. J., & Koehler, A. B. (2006). Another look at measures of forecast accuracy. International journal of forecasting, 22(4), 679-688.
Hyndman, R.J. (2004). New in Forecast 4.0. http://robjhyndman.com/hyndsight/forecast4/
McKenzie, J. (2011). Mean absolute percentage error and bias in economic forecasting. Economics Letters, 113(3), 259-262.
Perone, G. (2020). ARIMA forecasting of COVID-19 incidence in Italy, Russia, and the USA. Russia, and the USA (May 26, 2020).
Saba, A. I., & Elsheikh, A. H. (2020). Forecasting the prevalence of COVID-19 outbreak in Egypt using nonlinear autoregressive artificial neural networks. Process Safety and Environmental Protection.
Safi, S. K. (2013). Artificial neural networks approach to time series forecasting for electricity consumption in gaza strip. Artificial Neural Networks Approach to Time Series Forecasting for Electricity Consumption in Gaza Strip, 21(2).
Safi, S. K., & White, A. K. (2017). Short and long-term forecasting using artificial neural networks for stock prices in Palestine: a comparative study. Electronic Journal of Applied Statistical Analysis, 10(1), 14-28.
Thode, H. C. (2002). Testing for normality (Vol. 164). CRC press.
Wang, L., Zou, H., Su, J., Li, L., & Chaudhry, S. (2013). An ARIMA-ANN hybrid model for time series forecasting. Systems Research and Behavioral Science, 30(3), 244-259.
Wei, W. W. (2006). Time series analysis. In The Oxford Handbook of Quantitative Methods in Psychology: Vol. 2.
White, A., & Safi, S. K. (2016). The efficiency of artificial neural networks for forecasting in the presence of autocorrelated disturbances. The efficiency of artificial neural networks for forecasting in the presence of autocorrelated disturbances, 5(2).
Worldometer, C. C. Worldometer.(2020) 1–22. Doi, 10(2020.01), 23-20018549.