Frontiers in Big Data

TYPE Original Research
PUBLISHED 14 June 2023
DOI 10.3389/fdata.2023.1081639

Nonstationary time series forecasting using optimized-EVDHM-ARIMA for COVID-19

Suraj Singh Nagvanshi[1]*, Inderjeet Kaur[1†], Charu Agarwal[1†] and Ashish Sharma[2†]

[1]Department of Computer Science & Engineering, Ajay Kumar Garg Engineering College, Ghaziabad, Uttar Pradesh, India, [2]Department of Computer Science & Engineering, Maharaja Agrasen Institute of Technology, New Delhi, India

OPEN ACCESS

EDITED BY
Shaolong Sun,
Xi'an Jiaotong University, China

REVIEWED BY
Isha Gupta,
Dr. A.P.J. Abdul Kalam Technical University, India
Basheer Al-Maqaleh,
Thamar University, Yemen
Mingchen Li,
Chinese Academy of Sciences (CAS), China

*CORRESPONDENCE
Suraj Singh Nagvanshi
✉ suraj2010012m@akgec.ac.in

†These authors have contributed equally to this work and share last authorship

RECEIVED 27 October 2022
ACCEPTED 18 April 2023
PUBLISHED 14 June 2023

CITATION
Nagvanshi SS, Kaur I, Agarwal C and Sharma A (2023) Nonstationary time series forecasting using optimized-EVDHM-ARIMA for COVID-19. *Front. Big Data* 6:1081639.
doi: 10.3389/fdata.2023.1081639

The Coronavirus (COVID-19) outbreak swept the world, infected millions of people, and caused many deaths. Multiple COVID-19 variations have been discovered since the initial case in December 2019, indicating that COVID-19 is highly mutable. COVID-19 variation "XE" is the most current of all COVID-19 variants found in January 2022. It is vital to detect the virus transmission rate and forecast instances of infection to be prepared for all scenarios, prepare healthcare services, and avoid deaths. Time-series forecasting helps predict future infected cases and determine the virus transmission rate to make timely decisions. A forecasting model for nonstationary time series has been created in this paper. The model comprises an optimized EigenValue Decomposition of Hankel Matrix (EVDHM) and an optimized AutoRegressive Integrated Moving Average (ARIMA). The Phillips Perron Test (PPT) has been used to determine whether a time series is nonstationary. A time series has been decomposed into components using EVDHM, and each component has been forecasted using ARIMA. The final forecasts have been formed by combining the predicted values of each component. A Genetic Algorithm (GA) to select ARIMA parameters resulting in the lowest Akaike Information Criterion (AIC) values has been used to discover the best ARIMA parameters. Another genetic algorithm has been used to optimize the decomposition results of EVDHM that ensures the minimum nonstationarity and maximal utilization of eigenvalues for each decomposed component.

KEYWORDS

time-series, forecasting, COVID-19, optimized-ARIMA, optimized-EVDHM, PPT

1. Introduction

Time series forecasting is a crucial area in machine learning, as it involves predicting future values of a variable or characteristic that is dependent on time and recorded at regular intervals. The ability to accurately forecast time series is essential in many fields, such as economics, finance, healthcare, transportation, and energy, to name a few. Time series forecasting helps organizations in projecting product demand, allocating resources efficiently, predicting maintenance schedules, and many other applications. Over the years, several approaches to time series forecasting have been proposed in the literature.

For instance, time series forecasting has been employed in various sectors, such as electricity prices (Contreras et al., 2003), wind power generation (Wan et al., 2014), electricity demand (Taylor and McSharry, 2007), traffic flow (Lippi et al., 2013), and taxi-passenger demand (Moreira-Matias et al., 2013). Moreover, time series forecasting has also been useful in predicting the spread of epidemic diseases such as Dengue (Appice et al., 2020) and Influenza (Saberian et al., 2014). These studies have demonstrated the effectiveness of time series forecasting in diverse fields and have inspired further research in this area.

Coronavirus Disease (COVID-19) is caused by the SARS-CoV-2 virus [Coronavirus disease (COVID-19) pandemic, 2021a]. COVID-19 was proclaimed a global pandemic on March 11th, 2020, after discovering the first known case in Wuhan, China [Coronavirus disease (COVID-19) pandemic, 2021b]. The illness has since spread around the globe, culminating in a pandemic. Various researchers have expressed interest in developing Artificial Intelligence based solutions to assist governments and enterprises in making decisions. The past COVID-19 forecasting investigations are summarized in Table 1.

ARIMA is a time-series forecasting model that is one of the most often utilized approaches. ARIMA-based models are commonly employed in stationary time-series analysis (Wilson, 2016), although the standard ARIMA-based approach is inefficient in real-world settings for time series with nonstationary properties (Li and Chiang, 2013; Yang and Lin, 2016). EVDHM is a modern approach for nonstationary time series forecasting that may be combined with ARIMA (Sharma et al., 2021).

The COVID-19 pandemic has brought about an urgent need for accurate forecasting of the spread of the disease, to inform decision-making and public health interventions. As shown in Table 1, multiple studies have explored the use of machine learning (ML) and time series forecasting models in predicting the spread of COVID-19. These studies have highlighted the potential of ML and time series forecasting models in guiding decision-making and administrations. For instance, Rustam et al. (2020) employed a supervised ML model to predict the number of COVID-19 cases, deaths, and recoveries and discovered that exponential smoothing outperformed other models. Similarly, Kumar and Susan (2020) used ARIMA and Prophet time series forecasting models and found that the ARIMA model was more effective in forecasting COVID-19 prevalence. Andreas et al. (2020) proposed an improved mathematical forecasting framework based on ML and cloud computing that uses real-time data to accurately predict the progress of the curve. Satu et al. (2021) developed a web portal that provides real-time information on COVID-19 cases in Bangladesh and worldwide, including an ML-based short-term forecasting tool. Other researchers, such as Darapaneni et al. (2020) and Kurniawan and Kurniawan (2021), have also presented models for forecasting COVID-19 prevalence in Indonesia and India, respectively. Finally, Sharma et al. (2021) proposed a new method for time-series forecasting of nonstationary data using a combination of EVDHM and ARIMA models.

Although these studies have shown promising results, the models used by the researchers are not completely automated and require manual analysis of the data to provide inputs to the models. This makes it difficult for non-technical individuals to use these models, and analysis of the data can be time-consuming and prone to human error. To address this limitation, this paper proposes an optimized EVDHM approach combined with the ARIMA model for automated nonstationary time series forecasting.

The proposed model is fully automated and requires no manual input, which makes it easy, efficient, and time-saving for forecasting COVID-19 cases. This model can be used by non-technical individuals and is less prone to human error. The optimized EVDHM approach is used to capture the nonstationarity of the COVID-19 time series data, while the ARIMA model is used to capture the autocorrelation in the data. The proposed model is optimized using a grid search algorithm to select the best hyperparameters. The performance of the proposed model is evaluated using RMSE and compared with the performance of the traditional ARIMA model.

In conclusion, the proposed automated nonstationary time series forecasting model has the potential to provide accurate and timely predictions of the spread of COVID-19. The model is easy to use, efficient, and less prone to human error, and can be used by non-technical individuals.

The remainder of this article is structured as follows. Section 2 covers the datasets used in this study. Section 3 introduces EVDHM and ARIMA, Section 4 provides the proposed model and the stage outcomes, Section 5 compares EVDHM and Optimized EVDHM, and Section 6 summarizes the findings. Finally, section 7 brings the article to an end.

2. Dataset used

The data set for COVID-19 new cases in India has been utilized in this research. Data from January 22nd to May 10th, 2020, has been used for training, and data from the 11th to the 30th of May 2020 has been used for testing to analyze the proposed model. The data set has been given by Johns Hopkins University's Center for Systems Science and Engineering (CSSE) and is accessible online (CSSE, 2021).

3. Methods used

3.1. EVDHM

In linear algebra, the Hankel matrix is a square matrix with skew-diagonals that are constants. A Hankel matrix is represented as follows (Sharma and Pachori, 2017):

$$A = \begin{bmatrix} A_1 & A_2 & A_3 & . & . & . & A_N \\ A_2 & A_3 & & & & & . \\ A_3 & & & & & & . \\ . & & & & & & . \\ . & & & & & & A_{2N-3} \\ . & & & & & A_{2N-3} & A_{2N-2} \\ A_N & . & . & . & A_{2N-3} & A_{2N-2} & A_{2N-1} \end{bmatrix} \tag{1}$$

Eigenvalue Decomposition may be used to decompose a square matrix into its eigenvalues and eigenvectors. In terms of eigenvalues

TABLE 1 A summary of predicting models for Covid-19 forecasts.

S.no	Year	Author	Investigated region(s)	Method(s) used	Accuracy measures
1	2020	Rustam et al. (2020)	Australia, Canada, Algeria, and Afghanistan	Support Vector Regression (SVR), Linear regression (LR), Exponential Smoothing (ES), Least Absolute Shrinkage, Selection operator (LASSO), and Linear regression (LR)	Mean Square Error (MSE), Root Mean Square Error (RMSE), Mean Absolute error (MAE), *R*-square, Adjusted *R*-square
2	2020	Kumar and Susan (2020)	Global cases, Spain, Italy, India, France, Russia, Iran, UK, US, Turkey, Germany	ARIMA, PROPHET	MAE, RMSE, Root Relative Squared Error (RRSE), Mean Absolute Percentage Error (MAPE)
3	2020	Andreas et al. (2020)	Global cases	Curve fitting	*R*-square
4	2021	Dash et al. (2021)	Brazil, France, India, Russia, United Kingdom, US	ARIMA	RMSE, MAE, MAPE
5	2021	Satu et al. (2021)	Bangladesh and Global cases	PROPHET, LR, Polynomial-regression (PR), SVR, Multilayer Perceptron (MLP), and Polynomial-MLP (poly-MLP)	RMSE, R-squared
6	2021	Kurniawan and Kurniawan (2021)	Indonesia	Curve fitting	MSE
7	2020	Darapaneni et al. (2020)	India	ARIMA	R-square, Bayesian Information Criteria (BIC), Akaike's Information Criteria (AIC), and MSE
8	2021	Sharma et al. (2021)	India, USA, Brazil	ARIMA	RMSE
9	2020	Mustafa and Fareed (2020)	Iraq	ARIMA	MSE and MAE
10	2021	Kumar and Kaur (2021)	Delhi (India)	ARIMA, Gaussian Process Regression (GPR), LR, M5 Rule MLP, Support Vector Regression, Multi-Criteria Decision Making (MCDM), and Self-organized maps and fuzzy time series (SOMFTS)	Normalized-RMSE, Mean Magnitude of Forecasting Error (MMFE) Square root of the variance of the magnitude of residual errors (SdARE), The proportion of anticipated instances with a relative error magnitude of <0.20
11	2021	Iqbal et al. (2021)	Pakistan	Long short-term memory (LSTM)	MAPE
12	2021	Zhan et al. (2021)	24 North American countries, 11 South American countries, 45 Asian countries, 46 European countries, 54 African countries, and 4 Oceanian countries	Gated Recurrent Unit (GRU), LSTM, Artificial Neural Network (ANN), Particle swarm optimization broad learning system (PSO- BLS)	R-square, MAE, RMSE

and eigenvectors, matrix A may be written as follows (Sharma and Pachori, 2017):

$$A = V\lambda V^{-1} \tag{2}$$

V, λ, and V^{-1} are the eigenvector matrix, eigenvalue matrix, and inverse eigenvalue matrix, respectively. A stationary time series with statistical features such as mean and variance is either time-invariant or has a reasonable variation. Because of the significant variance, forecasting a nonstationary time series is often more difficult or complex than forecasting a stationary time series. Therefore, it is usually processed to reduce the time series' nonstationarity when working with nonstationary time series. EVDHM is an algorithm that can be utilized for analyzing nonstationary time series. It decomposes a time series into components with changing trends, noise, and oscillating patterns (Sharma and Pachori, 2017). Other applications of EVDHM include cardiovascular signal analysis (Sharma and Pachori, 2018b; Sharma et al., 2019), muscle signal analysis (Sharma et al., 2019), and complex data processing (Sharma and Pachori, 2018a).

A time-series $S_t = t_1, t_2, t_3, \ldots t_{2N-1}$ can be written as a Hankel matrix H of size $N \times N$ as follows (Sharma and Pachori, 2017):

$$H = \begin{bmatrix} t_1 & t_2 & t_3 & . & . & . & t_N \\ t_2 & t_3 & & & & & . \\ t_3 & & & & & & . \\ . & & & & & & . \\ . & & & & & & t_{2N-3} \\ . & & & & & t_{2N-3} & t_{2N-2} \\ t_N & . & . & . & t_{2N-3} & t_{2N-2} & t_{2N-1} \end{bmatrix} \tag{3}$$

Then H can be expressed as follows (Sharma and Pachori, 2017):

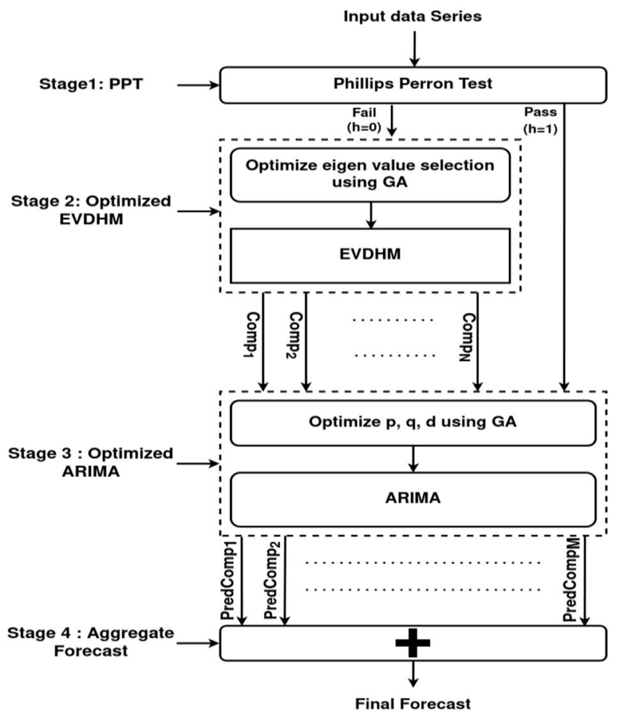

FIGURE 1
Model's block diagram based on O-EVDHM and O-ARIMA.

$$H = V_s \lambda V_s^{-1} \tag{4}$$

where eigenvalue matrix λ can be expressed as follows (Sharma and Pachori, 2017).

$$\lambda = \begin{bmatrix} \lambda_1 & 0 & 0 & . & . & . & 0 \\ 0 & \lambda_2 & & & & & . \\ 0 & & & & & & . \\ . & & & & & & . \\ . & & & & & & 0 \\ . & & & & & \lambda_{N-1} & 0 \\ 0 & . & . & . & 0 & 0 & \lambda_N \end{bmatrix} \tag{5}$$

All of the values in the eigenvalue matrix are zero, except for the diagonal elements, which have eigenvalues ranging from λ_1 to λ_N. The decomposition of the time series relies heavily on this eigenvalue matrix. It may be written as a sum of matrices with one or more distinct eigenvalues.

$$\lambda = \lambda_1 + \lambda_2 + \ldots + \lambda_M \tag{6}$$

(6)

Now, H becomes

$$H = V_s \lambda_1 V_s^{-1} + V_s \lambda_2 V_s^{-1} + \ldots + V_s \lambda_M V_s^{-1} \tag{7}$$

Let,

$$H = H_1 + H_2 + \ldots + H_M \tag{8}$$

The first decomposed component S_t^1 is computed using the mean of the skew-diagonal elements of the matrix H_1. The remaining components are computed in the same way, using the H_i matrices $i \in \{2,3, \ldots, M\}$.

3.2. ARIMA

ARIMA comprises Autoregressive (AR), Moving Average (MA), and Integration (I) models combined. The Autoregressive (AR) model is a regression-based model whose current value is determined by previous values. The lagged forecast errors create the moving average (MA) model. It makes the next forecast based on prior errors. To model a nonstationary series, ARIMA employs differencing, which is represented by the letter I in ARIMA. An ARIMA model having parameters p, d, and q, if applied on time series $S_t = \{s_1, s_2, s_3, \ldots\}$ then, it will be expressed as follows (Wilson, 2016):

$$\left(1 - \sum_{n=1}^{p} \phi_n \beta^n\right) (1-\beta)^d s_t = \theta^0 + \left(1 + \sum_{n=1}^{q} \theta^n \beta^n\right) e_t \tag{9}$$

where p is the AR component's order and q is the MA component's order, p and q are positive integers. The backshift operator β is interpreted as $\beta^n\ s_t = s_{t-n}$, where d is the needed degree of differencing to keep the time series stationary. The deterministic trend term is denoted by the number θ^0. e_t is the white Gaussian noise with zero mean and σ_e^2 variance.

4. Optimized EVDHM and ARIMA based model

ARIMA does not work well with nonstationary time series. It has a parameter d that takes the lagged series to cope with nonstationarity in the time series. However, it cannot be estimated successfully for nonlinear time series (Sharma et al., 2021). Compared to the original series, the EVDHM decomposes the actual time series into numerous components with high stationarity. Because the components are stationary, this breakdown strategy helps decrease predicting errors since the components are stationary.

The EVDHM-ARIMA-based model has been used for time series forecasting (Sharma et al., 2021). In this work, an optimized EVDHM and ARIMA-based hybrid model has been used to create a time series forecasting model that is both efficient and automated. The Phillips Perron Test (PPT) test has been initially used to test for the stationarity of the time series. Then, EVDHM is applied to the time series in case

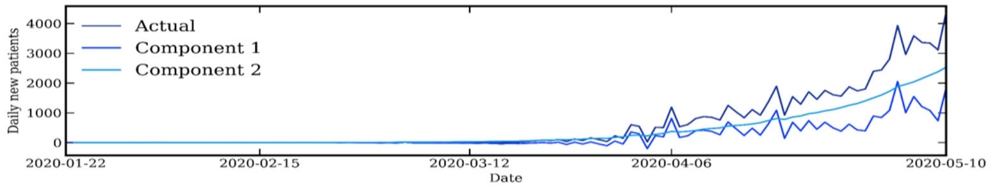

FIGURE 2
Trend (Component 1) and Variability (Component 2) of the actual series.

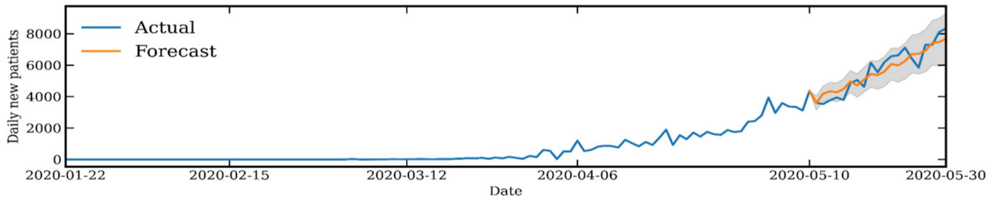

FIGURE 3
Final forecast from May 11, 2020 to May 30, 2020.

of failure of PPT that results in the decomposed components. EVDHM employs a genetic algorithm to select the eigenvalues to decompose the series. The genetic algorithm aims to increase the number of eigenvalues selected while minimizing the series' nonstationarity.

The decomposition of a time series into components using EVDHM are completely reliant on the matrix λ and how its N eigenvalues are split among the matrices λ^1, λ^2, ... λ^M Eq. (7). The simplest approach is to divide N eigenvalues for N matrices to get N components, but in this case, all of the components will not be stationary, and the number of components will be large, causing the proposed model to take a long time to forecast because each component will be fitted to ARIMA for individual component forecasting. To deal with this, a genetic algorithm has been used. The goal of this GA is to distribute N eigenvalues among M matrices in such a way that the majority of the components are stationary and the number of components is as low as possible Eq. (6). The following parameters and their values have been used by this genetic algorithm number of iterations = 50, number of bits (n_bits) = M (number of Eigenvalues), population size = 100, crossover rate = 0.9, mutation rate = 1.0/(n_bits).

An M-digit binary number is used to represent individuals in this evolutionary method. For each binary digit, 1 and 0 denote the relevant eigenvalue selection or rejection. The goal of the genetic algorithm is to find the best possible combination of eigenvalues, including the maximum possible eigenvalues both at the same time that give a nonstationary component. The best possible combination of eigenvalues is picked after all iterations. With the remaining eigenvalues, this process is continued until no more stationary components are feasible. Because of the genetic algorithm, the generated components have a low count and are also stationary. Finally, the components are subjected to ARIMA, and ARIMA parameters are tuned using another genetic algorithm.

Figure 1 depicts the proposed model's block diagram and also the stages of the optimized EVDHM and ARIMA-based model. The performance of the proposed model has been compared with the EVDHM-ARIMA-based model (Sharma et al., 2021). As indicated in Section 2, the data set from January 22nd to May 10th, 2020, is used for training, and the data set from May 11th to May 30th, 2020 is used for testing. The proposed model mainly consists of 4 stages PPT, Optimized EVDHM, Optimized ARIMA, and Aggregate forecast. A detailed explanation of each stage is given in the upcoming subsections.

4.1. Phillips Perron Test

The study employs the Phillips Perron test, which is a unit root test utilized to determine the stationarity of a time series. This test is distinct from other unit root tests such as the Dickey-Fuller test (DFT) and Augmented Dickey-Fuller test (ADFT) since it addresses serial correlation and heteroskedasticity in the errors. The automated features of PPT make it a better option than DFT and ADFT for this study (Patterson, 2011). In this study, the Boolean decision vector for the PPT is denoted by h, and the p-score is a left-tailed probability ranging from 0 to 1. The null hypothesis ($h = 0$) assumes that the underlying time series is non-stationary, while the alternative hypothesis ($h = 1$) asserts that it is stationary. A p-score <0.05 leads to the rejection of the null hypothesis, indicating that the time series is nonstationary. Conversely, a p-score >0.05 leads to the acceptance of the null hypothesis and the rejection of the alternative hypothesis, indicating that the time series is stationary. The results of applying PPT to the training dataset in this study reveal that the dataset is nonstationary since the h and p scores are 0 and 1, respectively.

TABLE 2 *h*-score, *p*-score, and SS score of decomposed components.

Components	EVDHM (first iteration)										Optimized EVDHM	
	C_1	C_2	C_3	C_4	C_5	C_6	C_7	C_8	C_9	C_{10}	C_1	C_2
h score	0	0	1	1	1	1	1	1	1	1	1	0
p score	0.9990	0.0764	0.0010	0.0010	0.0010	0.0010	0.0010	0.0010	0.0010	0.0010	0.04114	1
SS score	1.46×10^8	$7.5\ 10^5$	7.64×10^5	4.6×10^5	1.09×10^5	8.03×10^4	1.07×10^5	6.38×10^4	2.5×10^4	9.6×10^3	2.40×10^7	5.91×10^7

4.2. Optimized-EVDHM

EVDHM transforms a data series into a Hankel matrix H and then decomposes the Hankel matrix into its eigenvalues and eigenvectors, as shown in Eq. (4). The eigenvalue matrix λ is decomposed into M matrices such that each matrix has one or more distinct eigenvalues of the matrix H. For an effective and efficient decomposition of a data series, the decomposed components should be less in count, and all the components should be stationary. The number of components should be less because ARIMA is applied to each component. Therefore, a high number of components will increase the running time of ARIMA. Moreover, the stationarity of the component will help the ARIMA for better predictions. A genetic algorithm has been used in the proposed model to decompose the eigenvalue matrix λ into λ_1, λ_2, . . ., λ_M matrices. The objective of this genetic algorithm is to select the eigenvalues of λ for λ_1 such that it will try to maximize the eigenvalues utilization and minimize the nonstationarity of the component formed with λ_1. Similarly, subsequent λ's are calculated. The resultant two components and the actual data series are shown in Figure 2.

4.3. Optimized ARIMA

ARIMA needs three parameters: *p, q*, and *d* for each decomposed component to fit the model. A genetic algorithm has been used in this research to automate parameter selection. The genetic algorithm (GA) selects the model's variables *p*, *q*, and *d*, having the lowest Akaike information criteria (AIC) value. The AIC is a mathematical tool for determining if a model sufficiently fits the data. Generally, the best fitting models are considered good. Therefore, the GA has been set to use 0–5 for all three parameters to find out the best set of parameters with the lowest AIC value. The derived parameters for each component have been fitted into the ARIMA model, and the fitted model was used to anticipate future values from May 10th through May 30th, 2020.

4.4. Aggregate forecast

The ARIMA's separate component predictions have been combined to get the final projection. Figure 3 depicts the final forecast.

4.5. Performance of O-EVDHM

In this study, the performance of the Optimized EVDHM approach has been evaluated by calculating the sum of squared (SS) scores for all the decomposed components. The SS score is an important metric that measures the strength of the decomposition of the time series. The formula for calculating the SS score has been given in Eq. (10).

$$SS = \sum_{i=1}^{n} (t_i)^2 \qquad (10)$$

where t_i is an n-sample data series with i being a positive integer less than or equal to n.

The SS scores for the decomposed components have been presented in Table 2. From the table, it can be observed that the SS score for the main signal is 1.47×10^8, while the SS values for C_1 and C_2 are 2.40×10^7 and 5.91×10^7, respectively. The stationary component C_1 captures the variability of the actual series, whereas the trend component C_2 is nonstationary. Therefore, no further decomposition is required beyond C_2.

The performance of the proposed approach has been further assessed by comparing it with the EVDHM-based approach. As mentioned earlier, the EVDHM-based approach decomposes the time series into 10 components. The SS score of the trend component C_1 calculated using the EVDHM technique is 1.46×10^8, which is much higher than the trend component C_2 calculated using the Optimized EVDHM method, 5.91×10^7. This indicates that the Optimized EVDHM approach is more successful in decomposing the actual nonstationary series.

To evaluate the forecasting accuracy of the proposed approach, the RMSE has been calculated for both the EVDHM-based ARIMA model and the Optimized EVDHM-based ARIMA model. The RMSE for the EVDHM-based ARIMA model is 702.6, whereas the RMSE for the Optimized EVDHM-based ARIMA model is 538, which is significantly lower.

5. Comparison of EVDHM and optimized EVDHM

The Optimized EVDHM-based decomposition approach was developed to address the limitations of the EVDHM method used in (Sharma et al., 2021). For measuring the performance of the Optimized EVDHM method, it was compared to the EVDHM method in terms of h-score, p-value, and SS score for the decomposed components of the actual series.

Table 2 shows the comparison of the results. After the first iteration, the EVDHM decomposes the actual series into ten components C_1-C_{10}, as shown in Table 2. Component C_1 has a trend, whereas the remaining components, C_2-C_{10}, have variability. The SS score of the trend component C_1 calculated using the EVDHM technique is 1.46×10^8, which is much higher than the trend component C_2 calculated using the Optimized EVDHM method, 5.91×10^7. This indicates that the Optimized EVDHM is more successful in decomposing actual nonstationary series.

Moreover, after the first iteration, the EVDHM method resulted in a total of 10 components, while the Optimized EVDHM method yielded only two. This reduction in the number of components implies that the Optimized EVDHM method can effectively and efficiently decompose nonstationary time series. In addition, the decomposition using the Optimized EVDHM does not require several iterations, making it a more suitable approach for nonstationary time series.

Overall, the results demonstrate that the Optimized EVDHM-based decomposition approach outperforms the EVDHM method in efficiency and effectiveness in decomposing nonstationary time series. This improvement is attributed to the use of a genetic algorithm in the process of selecting the eigenvalues best suited for each decomposition, which optimizes the decomposition process and yields better results. As such, the Optimized EVDHM method holds great promise for future applications in various fields, including finance, healthcare, and environmental studies, where nonstationary time series are ubiquitous.

6. Results and discussions

This paper presents an Optimized EVDHM and ARIMA-based time series forecasting model used to anticipate the COVID-19 Indian cases. As mentioned in section 2, 109 days of data were utilized for training the proposed model from January 22nd to May 10th, 2020. Forecasting has been done for the following 20 days till May 30th, as indicated in Figure 3. The blue line in Figure 3 depicts the actual data series, while the red line depicts predictions from May 11th to May 30th, 2020. The gray region represents the 95% percent confidence interval. The root means squared error (RMSE) for the anticipated values has been determined to assess the suggested model's performance.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(P_i - A_i)} \tag{11}$$

P and A are the series' predicted and observed values, respectively. The RMSE of the suggested model is 538 for the 20 days of projected values, but the RMSE of the EVDHM-ARIMA-based model is 702.6, which is much higher than the RMSE of the proposed model.

7. Conclusions

In conclusion, the proposed Optimized EVDHM and ARIMA-based approach for time series forecasting is demonstrated to be effective in predicting new cases of COVID-19. The numerical comparison shows that the Optimized EVDHM-based ARIMA model outperforms the EVDHM-based ARIMA model with an RMSE of 538, indicating the practical significance of this study. The approach utilizes a genetic algorithm-based approach for decomposing nonstationary time series into its constituent components, followed by the application of ARIMA for forecasting. The proposed technique can be applied to various signals in the future, such as power load, sales forecasts, and inventory research, among others, making it a versatile tool for time series forecasting.

The points of innovation in this study include the use of genetic algorithms for optimizing the EVDHM decomposition method, which leads to more efficient and effective decomposition of nonstationary time series. The proposed approach also utilizes ARIMA for forecasting, which is a widely used and reliable method for time series forecasting.

However, there are some current shortcomings in this study. The proposed approach has only been tested on COVID-19 data, and its performance on other datasets needs to be evaluated. Additionally, the proposed approach can only be applied to univariate time series, and its extension to multivariate time series remains an area for future research. Nonetheless, the proposed Optimized EVDHM and ARIMA-based approach demonstrates promising results and can be considered a valuable addition to the existing literature on time series analysis and forecasting.

Data availability statement

Publicly available datasets were analyzed in this study. This data can be found at: https://github.com/CSSEGISandData/COVID-19.

Author contributions

All authors listed have made a substantial, direct, and intellectual contribution to the work and approved it for publication.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Andreas, A., Mavromoustakis, C. X., Mastorakis, G., Mumtaz, S., Batalla, J. M., Pallis, E., et al. (2020). "Modified machine learning technique for curve fitting on regression models for COVID-19 projections," in *2020 IEEE 25th International Workshop on Computer Aided Modeling and Design of Communication Links and Networks (CAMAD)* (Pisa: IEEE), 1–6. doi: 10.1109/CAMAD50429.2020.9209264

Appice, A., Gel, Y. R., Iliev, I., Lyubchich, V., and Malerba, D. A. (2020). Multi-stage machine learning approach to predict dengue incidence: a case study in Mexico. *IEEE Access* 8, 52713–52725. doi: 10.1109/ACCESS.2020.2980634

Contreras, J., Espinola, R., Nogales, F. J., and Conejo, A. J. (2003). ARIMA models to predict next-day electricity prices. *IEEE Trans. Power Syst.* 18, 1014–1020. doi: 10.1109/TPWRS.2002.804943

Coronavirus disease (COVID-19) pandemic (2021b). Available online at: https://www.who.int/emergencies/diseases/novel-coronavirus-2019 (accessed December 07, 2021).

Coronavirus disease (COVID-19) pandemic (2021a). Available online at: https://www.who.int/health-topics/coronavirus (accessed December 07, 2021).

CSSE (2021). *Center for Systems Science and Engineering (CSSE) at Johns Hopkins University, COVID-19 Data*. Available online at: https://github.com/CSSEGISandData/COVID-19 (accessed December 07, 2021).

Darapaneni, N., Reddy, D., Paduri, A. R., Acharya, P., and Nithin, H. S. (2020). "Forecasting of COVID-19 in India using ARIMA model," in *2020 11th IEEE Annual Ubiquitous Computing, Electronics Mobile Communication Conference (UEMCON)* (New York, NY: IEEE), 0894–0899. doi: 10.1109/UEMCON51285.2020.9298045

Dash, S., Chakraborty, C., Giri, S. K., Pani, S. K., and Frnda, J. (2021). BIFM: big-data driven intelligent forecasting model for COVID-19. *IEEE Access* 9, 97505–97517. doi: 10.1109/ACCESS.2021.3094658

Iqbal, M., Al-Obeidat, F., Maqbool, F., Razzaq, S., Anwar, S., Tubaishat, A., et al. (2021). COVID-19 patient count prediction using LSTM. *IEEE Trans. Comput. Soc. Syst.* 8, 974–981. doi: 10.1109/TCSS.2021.3056769

Kumar, A., and Kaur, K. (2021). "A hybrid SOM-Fuzzy time series (SOMFTS) technique for future forecasting of COVID-19 cases and MCDM based evaluation of COVID-19 forecasting models," in *2021 International Conference on Computing, Communication, and Intelligent Systems (ICCCIS)* (Greater Noida, IEEE), 612–617. doi: 10.1109/ICCCIS51004.2021.9397216

Kumar, N., and Susan, S. (2020). "COVID-19 pandemic prediction using time series forecasting models," in *2020 11th International Conference on Computing, Communication and Networking Technologies (ICCCNT)* (Kharagpur: IEEE), 1–7. doi: 10.1109/ICCCNT49239.2020.9225319

Kurniawan, A., and Kurniawan, F. (2021). "Time series forecasting for the spread of COVID-19 in Indonesia using curve fitting," in *2021 3rd East Indonesia Conference on Computer and Information Technology (EIConCIT)* (Surabaya: IEEE), 45–48. doi: 10.1109/EIConCIT50028.2021.9431936

Li, C., and Chiang, T. W. (2013). Complex neurofuzzy ARIMA forecasting—a new approach using complex fuzzy sets. *IEEE Trans. Fuzzy Syst.* 21, 567–584. doi: 10.1109/TFUZZ.2012.2226890

Lippi, M., Bertini, M., and Frasconi, P. (2013). Short-term traffic flow forecasting: an experimental comparison of time-series analysis and supervised learning. *IEEE Trans. Intell. Transp. Syst.* 14, 871–882. doi: 10.1109/TITS.2013.2247040

Moreira-Matias, L., Gama, J., Ferreira, M., Mendes-Moreira, J., and Damas, L. (2013). Predicting taxi–passenger demand using streaming data. *IEEE Trans. Intell. Transp. Syst.* 14, 1393–1402. doi: 10.1109/TITS.2013.2262376

Mustafa, H. I., and Fareed, N. Y. (2020). "COVID-19 cases in Iraq; Forecasting incidents using box - Jenkins ARIMA Model," in *2020 2nd Al-Noor International Conference for Science and Technology (NICST)* (Baku: IEEE), 22–26. doi: 10.1109/NICST50904.2020.9280304

Patterson, K. (2011). *Unit Root Tests in Time Series Volume 1: Key Concepts and Problems*. Belfast: Kerry Patterson. doi: 10.1057/9780230299306

Rustam, F., Reshi, A. A., Mehmood, A., Ullah, S., On, B. W., Aslam, W., et al. (2020). COVID-19 future forecasting using supervised machine learning models. *IEEE Access* 8, 101489–101499. doi: 10.1109/ACCESS.2020.2997311

Saberian, F., Zamani, A., Gooya, M. M., Hemmati, P., Shoorehdeli, M. A., Teshnehlab, M., et al. (2014). "Prediction of seasonal influenza epidemics in Tehran using artificial neural networks," in *2014 22nd Iranian Conference on Electrical Engineering (ICEE)* (Tehran: IEEE), 1921–1923. doi: 10.1109/IranianCEE.2014.6999855

Satu, M. S., Rahman, M. K., Rony, M. A., Shovon, A. R., Adnan, M. J. A., Howlader, K. C., et al. (2021). "COVID-19: update, forecast and assistant - an interactive web portal to provide real-time information and forecast COVID-19 cases in Bangladesh," in *2021 International Conference on Information and Communication Technology for Sustainable Development (ICICT4SD)* (Dhaka: IEEE), 456–460. doi: 10.1109/ICICT4SD50815.2021.9396786

Sharma, R., Chandra, P., and Pachori, R. (2019). "Electromyogram signal analysis using eigenvalue decomposition of the hankel matrix," in *Intelligence and Signal Analysis. Advances in Intelligent Systems and Computing, Vol 748*, eds M. Tanveer and R. Pachori (Singapore: Springer), 671–682. doi: 10.1007/978-981-13-0923-6_57

Sharma, R., and Pachori, R. (2017). Time-frequency representation using IEVDHM-HT with application to classification of epileptic EEG signals. *IET Sci. Meas. Technol.* 09, 12. doi: 10.1049/iet-smt.2017.0058

Sharma, R., and Pachori, R. (2018a). Eigenvalue decomposition of hankel matrix-based time-frequency representation for complex signals. *Circuits, Syst. Signal Process.* 37, 3313–3329. doi: 10.1007/s00034-018-0834-4

Sharma, R. R., Kumar, M., Maheshwari, S., and Ray, K. P. (2021). EVDHM-ARIMA-based time series forecasting model and its application for COVID-19 Cases. *IEEE Trans. Instrum. Meas.* 70, 1–10. doi: 10.1109/TIM.2020.3041833

Sharma, R. R., Kumar, M., and Pachori, R. B. (2019). Joint time-frequency domain-based CAD disease sensing system using ECG signals. *IEEE Sens. J.* 19, 3912–3920. doi: 10.1109/JSEN.2019.2894706

Sharma, R. R., and Pachori, R. B. (2018b). Baseline wander and power line interference removal from ECG signals using eigenvalue de- composition. *Biomed. Signal Process. Control.* 45, 33–49. doi: 10.1016/j.bspc.2018.05.002

Taylor, J. W., and McSharry, P. E. (2007). Short-term load forecasting methods: an evaluation based on european data. *IEEE Trans. Power Syst.* 22, 2213–2219. doi: 10.1109/TPWRS.2007.907583

Wan, C., Xu, Z., Pinson, P., Dong, Z. Y., and Wong, K. P. (2014). Probabilistic forecasting of wind power generation using extreme learning machine. *IEEE Trans. Power Syst*. 29, 1033–1044. doi: 10.1109/TPWRS.2013.2287871

Wilson, G. T. (2016). Time Series Analysis: Forecasting and Control, 5th Edition, by George E. P. Box, Gwilym M. Jenkins, Gregory C. Reinsel and Greta M. Ljung, 2015. Published by John Wiley and Sons Inc., Hoboken, New Jersey, pp. 712. ISBN: 978-1-118-67502-1. *J. Time Ser. Anal.* 37, 709–711. doi: 10.1111/jtsa.12194

Yang, H. L., and Lin, H. C. (2016). An integrated model combined ARIMA, EMD with SVR for stock indices forecasting. *Int. J. Arti. Intell. Tools*. 25, 1650005. doi: 10.1142/S0218213016500056

Zhan, C., Wu, Z., Wen, Q., Gao, Y., and Zhang, H. (2021). "Optimizing broad learning system hyper-parameters through particle swarm optimization for predicting COVID-19 in 184 countries," in *2020 IEEE International Conference on E-health Networking, Application Services (HEALTHCOM)* (Shenzhen: IEEE), 1–6. doi: 10.1109/HEALTHCOM49281.2021.9399020