Int. j. inf. tecnol.
https://doi.org/10.1007/s41870-021-00741-8

ORIGINAL RESEARCH

Multi-step time series analysis and forecasting strategy using ARIMA and evolutionary algorithms

Raghavendra Kumar[1,2] · **Pardeep Kumar**[1] · **Yugal Kumar**[1]

Received: 26 October 2020 / Accepted: 15 July 2021

Abstract Time series forecasting is a widely applied approach in sequential data series including the stock market. Time series forecasting can be examined through single step ahead as well as multi-step ahead forecasting despite its proven complex analysis and trends preserving limitations. Auto-regressive integrated moving average (ARIMA) is a widely accepted model for time series prediction. In this paper, we proposed a hyperparameter selection strategy for the ARIMA model using fusion of differential evolution (DE) and artificial bee colony (ABC) algorithm. Modified algorithms retain the exploration and exploitation strategies with the union of evolutionary algorithms with stock market time series data. The modified ABC using DE Optimization induces better generalization and efficient performance as compared to existing ARIMA models. In this paper, experiments are performed over 10 years of the dataset of Oil Drilling and Exploration and Refineries sector of National Stock Exchange and Bombay Stock Exchange from September 1, 2010 to August 31, 2020. Obtained result demonstrates that the proposed strategy using modified ABC-ARIMA hybrid model has superior performance than its counterparts. Proposed strategy achieves improved performance in forecasting accuracy along with preserving data trends in multi-step time series forecasting.

Keywords Time series · ARIMA · ABC and DE

✉ Raghavendra Kumar
raghavendra.dwivedi@gmail.com

[1] Department of Computer Science and Engineering, Jaypee University of Information Technology, Waknaghat, HP, India

[2] Department of Information Technology, KIET Group of Institutions, Delhi NCR, Ghaziabad, UP, India

1 Introduction

Stock market data as a time series identified as more complicated than other statistical data. Forecasting accuracy of a model is a most widely accepted criteria in time series analysis. Therefore, design and investigation of such models is a bigger challenge. ARIMA, is a widely accepted linear model to analyze time series data introduced by Box and Jenkins [1]. ARIMA is constituted using auto-regressive (AR) which processes the previous information to model and moving average (MA) that holds the control on noisy data information of previous instances. ARIMA checks the stationary process using augmented Dickey Fuller (ADF) test and Philippe-Perron (PP) test that are known as unit root tests. ADF test is to ensure that mean and variance will be constant over a given time frame. However, once the linearity of data gets lost ARIMA model shows a high error rate. ARIMA is limited to consider linear form of data and not suitable for a complex non-linear model [1]. In AR, if autoregressive coefficient ($\emptyset$) and number of previous instances are p then, lags function is obtained using Eq. (1):

$$X_t = \emptyset_1 X_{t-1} + \emptyset_2 X_{t-2} - - - - + \emptyset_p X_{t-p} = \sum_{j=1}^{p} \emptyset_j X_{t-j} \tag{1}$$

In MA, if MA coefficient (θ_j), q is the order of MA term and previous innovation process (ε_{t-j}) then MA equation will be

$$X_t = \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} - - - - + \theta_q \varepsilon_{t-q} = \sum_{j=1}^{q} \theta_j \varepsilon_{t-j} \tag{2}$$

The ARIMA parameters (p, d, q) hold the order of AR terms (p), order of MA terms (q) and significant difference

Published online: 02 August 2021

(d) to make stationary time series data. In most of the cases d is considered as 1. If parameter d is assigned as 0 then the model is known as auto-regressive moving average (ARMA). Finally, model parameters are estimated through auto-correlated function (ACF) and partial auto-correlated function (PACF). The values of p and q are traced from ACF and PACF plots. The behavior of both the plots set the ARIMA parameters trends. The generalized ARIMA is known as SARIMA where S indicates the *seasonal state*. In SARIMA terms P, D and Q indicate the seasonality of AR(p), MA(q) and differencing (d) where m is a periodicity (i.e. monthly, daily, etc.):

$$X_t = \text{SARIMA}(P, D, Q)(p, d, q)_m \quad (3)$$

Based on ACF and PACF plots of selected stocks, values of AR and MA are examined. Different combinations of ARIMA (p, d, q) and obtained values of AIC evaluated to finalize the best ARIMA model will be discussed in later sections. The ACF plot illustrates the stationary trends in time set evaluated through the ADF test successfully. ADF test confirms that default hypothesis to be accepted for stationary unit root test [1]. Existing models based on literature surveys such as ARIMA, SARIMA, ARCH, GARCH, etc. independently failed to achieve best performance in forecasting accuracy along with preserving data trends. Hybrid models are always found superior to get best results in both the segments [2]. After the study, it is observed that most of the models were applied on one step ahead prediction that is a major limitation of the existing works. Therefore, in this paper a modified ABC using DE is examined to optimize the parameters of ARIMA for multi-step forecasting of time series data. The proposed ABC-ARIMA obtains significant better results as compared to existing core and hybrid models.

Further, as the original contribution, following steps are streamlined to build the proposed hybrid model.

1. As a novel approach Initially, historical dataset of 10 years span is considered of Oil Drilling and Exploration and Refineries sectors of National Stock Exchange (NSE) and Bombay Stock Exchange (BSE) from Sept 1, 2010, to August 31, 2020. Detailed description for selected stocks is discussed in Table 1, where each historical dataset contains 2460 instances.
2. Secondary, effectively build ARIMA model for selected stocks and perform data preprocessing and stationary tests using unit root test. After that ACF and PACF plots are examined to identify AR, MA and differing values for p, d and q parameters for the ARIMA model.
3. Finally, proposed a novel strategy of hyperparameter selection of ABC-ARIMA model using modified artificial bee colony algorithm (ABC) and differential evolution (DE) algorithm to balance the exploitation issue in the onlooker bee phase of ABC. The ABC algorithm is also modified in this phase. However, complexity of the proposed algorithm is found on the higher side due to hybrid design.

Organization of the paper proceeds as follows. Section 2 involves the history and existing work of core and hybrid models. Section 3 demonstrates the methodology used in this paper in the form of a modified ABC-ARIMA model using DE. Section 4 simulates the result analysis and discussion regarding obtained accuracy from the proposed model. Finally, the conclusion of this overall investigation and future prospect are given in Sect. 5.

2 Related Works

Pai et al. proposed a hybrid model based on ARIMA and SVM models for stock forecasting. The proposed model obtains the error measures in the form of MAE, MSE, RMSE and MAPE for the 10 stocks. ARIMA (0,1,0) gets significant accuracy in stock forecasting for German Deutschmark exchange and Shanghai Share [2]. Babu et al. works on preserving data trends on forecasting time span and prediction accuracy using multi step ahead prediction. The experiment is performed on a linear hybrid model based on ARIMA and GARCH to preserve data trends and maintain prediction accuracy. Proposed model incorporates the partitioning and interpolation techniques on the NSE dataset [3]. Domingos et al. examines the linear and non-linear combination of machine learning models for time series forecasting. In the hybrid model ARIMA is identified as linear and MLP-SVR is identified to observe non-linear pattern trends. The proposed model is examined over six different applications including new and existing datasets like star brightness, airline passengers dataset, stock market, Colorado river flow, etc. [4]. Khashei et al. demonstrates a hybrid model of linear (ARIMA) and nonlinear (ANN) models to improve forecasting accuracy. Proposed model improves accuracy in terms of MAE and MSE against independent ARIMA, ANN and benchmark models for various datasets [5, 6]. Zhou evaluated the ARIMA model with attention mechanism-based LSTM to improve web traffic time series forecasting. Attention model between two LSTM layers improves the hyper parameters. ARIMA (1,0,0) and (1,1,2) selected for subsequent evaluation [7]. Vantuch et al. proposed an evolutionary based ARIMA model to optimize ARIMA parameters using genetic algorithm (GA) and particle swarm optimization (PSO). Proposed model is evaluated on a stationary dataset of the stock market. The parameters p, d, q (12,2,8) are obtained using the GA-ARIMA model along with lowest

Table 1 Dataset description based on closing value [24]

Sr. no	Stock name	Ticker value	Count	Mean	Std	Min	25%	50%	75%	Max
1	Bharat Petroleum Corporation Limited	BPCL	2460	212.91391	141.29129	39.906464	65.261169	202.25322	344.11088	516.3291
2	GAIL (India) Limited	GAIL	2460	123.12976	28.673799	69.400002	100.50175	118.568	139.7215	196.89999
3	Gujarat State Petronet Limited	GSPL	2460	132.23215	53.106373	49.049999	84.225	124.975	180.275	262.5
4	Hindustan Petroleum Corporation Limited	HINDPETRO	2460	182.15559	114.03113	36.2444	75.241699	169.739	268.31249	488.35001
5	Indraprastha Gas Limited	IGL	2460	164.16574	124.80265	38.669998	64.352503	93.84	270.775	522.29999
6	Indian Oil Corporation Limited	IOC	2460	110.57096	45.485315	48.224998	76.384377	94.549999	144.76875	227.35001
7	Mangalore Refinery and Petrochemicals Limited	MRPL	2460	69.224573	25.262513	22.299999	53.537499	65.050003	77.75	142.35001
8	Oil India Limited	OIL	2460	220.27907	49.747255	70.349998	187.45625	228.05	253.52	330.60001
9	Oil and Natural Gas Corporation	ONGC	2460	178.67292	40.75766	60	157.54575	180.60001	198.13775	310.43301
10	Reliance Industries Limited	RELIANCE	2460	700.97274	390.22876	338.04999	434.15001	501.425	929.96248	2177.7

AIC and BIC values [8]. In another approach of evolutionary hybrid model Musdholifah et al. uses firefly algorithm (FA) to extract lowest values of Akaike Information Center (AIC) among all the combinations tried on experiments [9]. In another hybrid approach, Kumar et al. proposed the hybrid model ABC-LSTM for time series forecasting using ABC and long short-term memory (LSTM). The proposed model obtains the best forecasting accuracy over its counterpart models [10]. However, core nonlinear models always produce satisfactory accuracy for stock market price and trends prediction as Kumar et al. proposed the modified stacked LSTM for BSE30 datasets and obtained superior performance than other benchmark models considered in this study [11]. Ballini et al. compares the ARIMA model with classical artificial neural network (ANN) including back propagation neural network (BPNN) and multilayer perceptron (MLP) and a neuro fuzzy based network. This study was examined on Brazilian Stock Exchange [12]. Wang builds an ARIMA model to forecast mid-term price trends of Taiwan Stock Exchange. ACF and PACF plots are examined on the ARIMA (1,2,1) model. Proposed model is based on RNN that features were extracted from ARIMA (1,2,1) model [13]. Existing study reflects the diverse approach of time series analysis. Noticeably, to achieve the objective of forecasting accuracy along with preserving data trends, we proposed a hybrid model using DE optimized ABC-ARIMA.

3 Research methodology

Karaboga et al. proposed a most effective typical swarm algorithm in 2005 known as ABC algorithm [14]. ABC is inspired from foraging behavior of swarms of bees. Swarm bees maintain the social harmony in populous colonies. ABC is considered as an effective solution to numerical optimization problems like the stock market. ABC algorithm provides multidimensional solutions to time series domain problems like feature selection and hyperparameters optimization [15, 16].

Algorithm1: DE Optimized Artificial Bee Colony (ABC) for ARIMA for Stock Price

Input: Initial Population x_i
Dataset (D) and Training & Testing dataset (X, X`)
No of optimization parameters D
Possible solutions (SN)
Output: Optimization parameters for LSTM
BEGIN
Step-1 Load training dataset (X)
Step-2 Generate the initial population x_i where i(1, 2,…, SN).
Step-3 Evaluate the fitness (fit_i) of population where I (1, 2,…,SN).
Step-4 set iteration *i* to 1
Step-5 Repeat
Step-6 for each employed-bee
{
Produce a new solution v_{ij} using equation-5
Compute the fitness (fit_i) for v_{ij} using equation-7
Using greedy approach
}
Step-7 Compute probability (p_i) for x_i using equation-6
Step-8 for each onlooker-bee
{
Apply selection on solution x_i based on p_i
Produce a new solution v_{ij}using equation-5
Compute the fitness (fit_i) for v_{ij} using equation-7
Generate new candidate solution v_{ij} as DE-mutation using equation- 13
Using greedy approach
}

Step-9 Replace SN with new solution produced by scout-bee using equation-4
Step-10 Repeat step-9 until (x_{ij}> limit) and gets optimized value using equation-8
Step-11 Iteration i=i+1
Step-12 Until i=max (i)

END

ABC algorithm creates food sources with SN random solutions. It also represents the combined population of employed and onlooker-bee. Optimization parameters are notified as D for the considered problem. Here, Lower and Upper boundary parameters for bee's movement are considered as x_{min} and x_{max}, respectively. Initialization counter is set as C for and x_j is a solution where j is from 1, 2,…, D with the range of Scaling factor rand [0,1]. In this problem x_i is carried in the range of 1, 2… SN and x_{ij} can be computed using Eq. (4). In the process of generating candidate solutions employed-bee finds x_{ij} using Eq. (5). As initial values, Where j in range (1, 2,… D), k in range (1, 2,… SN) and j ≠ SN. Here, Onlooker-bees forage the solutions based on nectar probability p_i using Eq. (6). At this stage onlooker-bee assesses the source and Eq. (7) helps to formulate the objective function f_i where fit_i is the fitness value of solution i. Now, scout-bees forage optimized food in the workspace and find a global optimized solution. Here, the number of iterations is considered as a limit to be computed with Eq. (8). In this equation D is optimization parameters, n_e is the number of employed-bees and c is considered as a coefficient factor.

$$x_{ij} = x_{min,j} + rand[0,1]\left(x_{max,j} - x_{min,j}\right) \tag{4}$$

$$v_{ij} = x_{ij} + rand[-1,1]\left(X_{max,j} - X_{min,j}\right) \tag{5}$$

$$p_i = \frac{fit_i}{\sum_1^{SN} fit_n} \tag{6}$$

$$fit_i = \begin{cases} \dfrac{1}{1+f_i} & fi \geq 0 \\ 1 + |f_i| & fi < 0 \end{cases} \tag{7}$$

$$limit = D * n_e * c \tag{8}$$

3.1 Differential evolution (DE)

Differential evolution is the most effective algorithm proposed by Storn et al. [17]. DE assists to solve global optimization problems of time series models such as stock market prediction. Standard DE algorithm consists of four operations that include several steps. Initialization of the length of chromosome (D), range of gene values for operations (U_{min}, U_{max}), considered crossover rate (CR), identified mutation factor (F) and selected population size (N) where, i = 1, 2… N, j = 1, 2… D and rand is uniform probability distribution [18, 19].

$$X_{ij} = U_{min} + rand * (U_{max} - U_{min}) \tag{9}$$

Mutation is an individual change in gene which we observe with the help of the below equation. Where r1, r2, r3 are different random values.

$$V_i^{G+1} = X_{r1}^G + F * \left(X_{r2}^G - X_{r3}^G\right) \tag{10}$$

Crossover is used to exchange the elements between previous (parent) and current (child) genes. Where j is an individual gene and random value of j, rand (j) is distributed in the range of 0 to 1 in crossover operation.

$$u_{ij}^{G+1} = \begin{cases} v_{ij}^{G+1}, & if\ rand(j) \le cror\, j = rand\, n(t) \\ x_{ij}^G, & otherwise \end{cases}$$

Selection is known as comparison between offspring and parents where better gets priority over others based on the fitness function.

$$x_i^{G+1} = \begin{cases} u_i^{G+1}, & if \quad f\left(u_i^{G+1}\right) < f\left(x_i^G\right) \\ x_i^G, & otherwise \end{cases} \tag{12}$$

3.2 Modified ABC using DE

ABC is known for handling global exploration issues in multi objective problems. However, it is mostly reported in local search efficiency due to local exploitation. DE modifies the onlooker bee phase of ABC algorithm to improve local search mechanism using Eq. (13). The solution equation obtains improved search efficiency using hybrid search strategies. Given equation is inspired from the current-to-best/1 mode of DE [20]. DE-current-to-best solution improves convergence as well as solve multi objective real time optimization problems using numerical functions. In addition, to validate the hybrid algorithm's complexity, several existing studies are referred to. Xiang, W. et al. proposed a novel hABCDE hybrid algorithm for solving numerical optimization problems and validate it with twenty benchmark functions ($f_1 - f_{20}$). Attainment of outcome assured the performance of hABCDE algorithm over other existing algorithms [21]. In another study, UCI machine learning repository is explored by Zorarpacı et al. to evaluate the performance of the proposed feature selection hybrid algorithm. Experiments results outperformed the standard ABC and standard DE algorithm in terms of F-measure values [22]. Similarly, Jadon et al. tested the hybrid algorithm HABCDE over 20 benchmark functions ($f_1 - f_{20}$). Proposed work also conducted on four real world optimization problems considering ($f_{21} - f_{24}$). Obtained result validates that the hybrid algorithm HABCDE yields leading accuracy, convergence speed, stability and robustness from its individual base algorithms ABC and DE [23]. Stock market time series data is one of the most dynamic data with linear and nonlinear patterns. Achieving accuracy in forecasting is a major challenge for any model for such data patterns due to characteristics of hyperparameters. The fusion of ABC-DE creates a higher probability of right hyperparameter tuning for the ARIMA model. Since, there must be a balance between exploitation and exploration of variable selection while computing hyperparameters values. Selection of appropriate values for AR and MA for statistical models in stock market movement has not been exploited earlier. Therefore, based on the DE-ABC performance of existing work on various studies fusion of DE-ABC is opted for hyper parameter selection of proposed model modified ABC-ARIMA.

$$v_{ij} = x_{ij} + rand[-1, 1]\left(x_{best\, i} - x_{ij} + x_{r1,j} - x_{r2,j}\right) \tag{13}$$

3.3 Proposed modified ABC-ARIMA model

Step-1 As the preprocessing step, historical datasets of Oil Drilling and Exploration and Refineries of oil and gas sector of NSE and BSE from Sept 1, 2010 to August 31, 2020 are examined for stationary test. ADF test is a well-known unit root test performed over a selected dataset.

Step-2 In step-2 of the algorithm process, parameters of AR and MA are initialized with default values for the ARIMA model. The parameters values of p, d and q are to be assigned with (0,1,0).

Step-3 Based on the assigned values ACF and PACF plots are generated and examined for least residual. Parameter estimation for ARIMA models holds the coefficient factors and residual of variance. If the residual has lower values, then the ARIMA model is set for the forecasting with obtained parameters values (p, d, q).

Step-4 If the parameters values are not performed well for residuals, the proposed hybrid algorithm optimizes the hyper parameters p, d and q. Upgraded hyper parameters values regenerate the ACF and PACF plots and find the residual. The minimum of root mean square error (RMSE) and AIC bring the best model for time series forecasting.

RMSE is considered as fitness function for the modified ABC algorithm.

The weights of the ABC based model are trained on the validation set. The population size SN of the ABC strategy is set to 20 with dimension size 7. The maximum cycle number is set to 10 and the limit is set to 7. The flow chart of DE-optimized ABC algorithm is illustrated in Fig. 1.

4 Result analysis and experiments setup

For the proposed algorithm discussed in Sect. 3.3, experiment is conducted on Intel(R) Core (TM) i5-8265UC @ 1.8 GHz, secondary storage (2 TB) and RAM (16 GB) being considered as commodity hardware. Python3.0 and tensor flow environments are selected to examine the proposed hybrid model. However, Analysis and modeling of time series data and multi-step ahead forecasting is performed on SAS Studio.

4.1 Dataset description

For the experiment, 10 years of dataset is considered Oil Drilling and Exploration and Refineries of the oil and gas sector of NSE and BSE from Sept 1, 2010 to August 31, 2020 [24]. Historical data set contains 2460 instances for each stock of the oil and gas sector as described in Table 1. Stock dataset with 2460 instances is further divided into subsets training 1968 instances and testing 492 instances to bring generalization ability to the proposed hybrid model. To make the training dataset free from overfitting and underfitting constraints, training set is subsets into 25% that is 492 instances as validation set. Finally, the z-score normalization process is used to normalize the data values in the range of 0–1 using Eq. (14) since the data range has a wide range of stock prices. In the given equation X' is a scaled values of raw data X and stock price range is considered between X_{max} and X_{min}.

$$X' = X - X_{min}/X_{max} - X_{min} \tag{14}$$

4.2 Time series analysis

In the descriptive analysis phase of financial time series data, white noise test is performed in the experiment. White noise test is one of statistical tests to validate the hypothesis. This test yields the best result for random values generated for continuous time series data [25]. Figure 2a–j illustrates the autocorrelation of selected stocks of the oil and gas sector. Obtained autocorrelations are different in some of the cases from 0 for the given To-Lag range [6, 12,

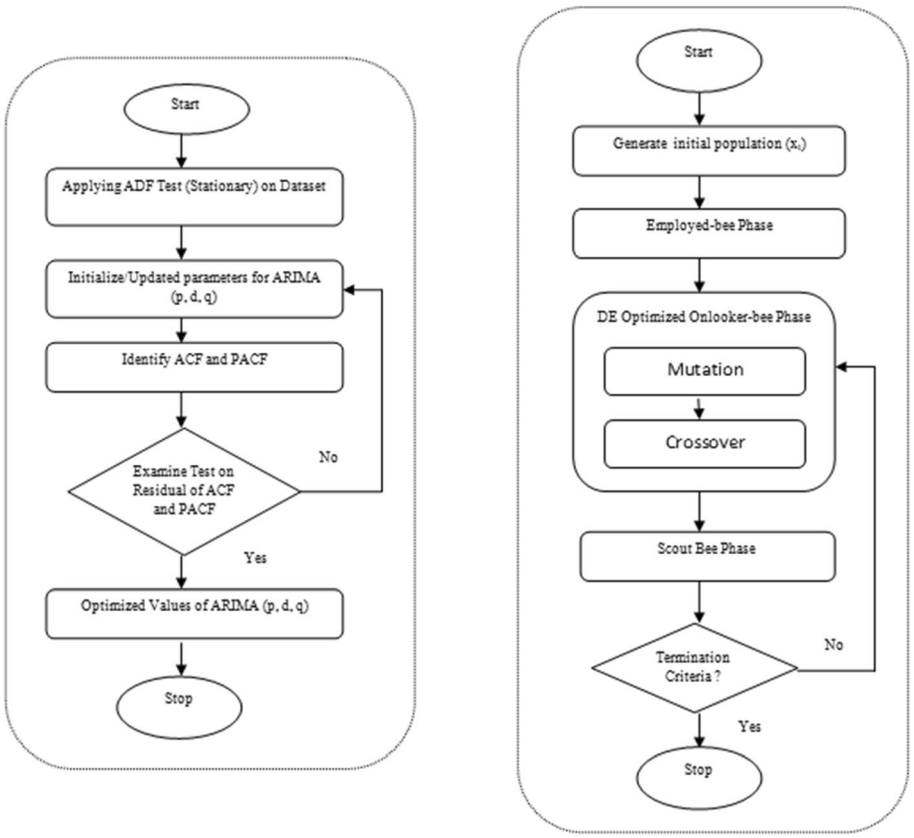


Fig. 1 Flow chart of proposed modified ABC-ARIMA (left) and modified algorithm ABC using DE (right)

Table 2 Performance analysis of proposed model

Variable	ARIMA			ABC-ARIMA			Proposed model (modified ABC-ARIMA)		
	RMSE	AIC	BIC	RMSE	AIC	BIC	RMSE	AIC	BIC
BPCL	0.01685	− 9535.673	− 10,302.865	0.01531	− 9647.235	− 9812.745	0.01436	− 9947.857	− 9936.243
GAIL	0.08892	− 10,200.465	− 10,990.251	0.08764	− 10,335.148	− 10,690.251	0.08629	− 10,663.208	− 10,761.594
GSPL	0.03574	− 10,001.886	− 10,856.144	0.03496	− 10,201.012	− 10,156.144	0.03008	− 10,410.250	− 10,398.635
HINDPETRO	0.01455	− 9286.231	− 9188.579	0.01399	− 9344.815	− 9176.124	0.01351	− 9362.110	− 9350.496
IGL	0.02754	− 10,430.648	− 10,653.444	0.02738	− 10,580.438	− 10,660.234	0.02629	− 10,773.208	− 10,761.594
IOC	0.02602	− 10,045.663	− 10,377.549	0.02588	− 10,245.543	− 10,437.449	0.02509	− 10,492.343	− 10,480.729
MRPL	0.01623	− 9428.112	− 9198.845	0.01603	− 9430.112	− 9288.338	0.01584	− 9443.802	− 9432.188
OIL	0.01020	− 10,481.538	− 10,388.125	0.00993	− 10,480.024	− 10,511.005	0.00908	− 10,735.044	− 10,723.430
ONGC	0.02971	− 10,034.546	− 10,090.782	0.02932	− 10,188.006	− 10,114.052	0.02807	− 10,268.356	− 10,256.742
RELIANCE	0.02270	− 10,712.761	− 10,777.054	0.02253	− 10,722.112	− 10,807.503	0.02200	− 10,988.622	− 10,977.008

18, 24]. However, most cases are close to 0 that this financial time series dataset may yield best results with the ARIMA model. Each stock panel contains the time series plot of the selected historical dataset. Behavior of the autocorrelation function (ACF) reflects with respect to different ranges of lags. Inverse autocorrelation function (IACF) is estimated in proposed hybrid ABC-ARIMA using Yule-Walker equations. The high order autoregressive model is computed using MA. In other words, IACF is a pure MA model of ACF. The impact of IACF getting smaller while lag value is found greater than the p value. Partial autocorrelation function (PACF) function at lag k in the time series dataset is influenced by the autoregressive Gaussian process of order k − 1. ACF plots for selected indices reflect significant drops with respect to k lags that proves the stationary behavior of time series data. Diagnosed statistics in Table 2 and plots of trends and correlation analysis (Fig. 2a–j) helps to validate the proposed model fitness. Proposed algorithm of DE Optimized ABC for ARIMA discussed in Sect. 3 applied to test the model. If a model does not meet the objective function as a minimum of AIC then parameters selection and model verification process typically gets repeated until a satisfactory model is finalized with optimized parameters. Remarkably, all selected stocks of the oil drilling and exploration and refineries of the oil and gas sector produce similar patterns. ACF and PACF plots show the constant performance of stationarity of NSE and BSE, India for selected 10 years of datasets.

4.3 Forecasting analysis

Existing works [7, 8] suggested the range of p, d and q of ARIMA optimization model between (0,0,0) to (12,12,12) to compute the smallest values of AIC. Number of Employed bees is equal to the number of parameters to be optimized for the ARIMA model. Therefore, the length of the food source of ABC is equivalent to the number of ARIMA parameters (3). The number of iterations 100 are considered for the experiments of historical dataset. The selected stocks of NSE and BSE, India obtain RMSE for the proposed ABC-ARIMA model. Optimized values for AR(p), I(d) and MA(q) are obtained as p d q (2,1,3). Table 2 reflects low AIC and BIC values of selected experiments performed to yield best prediction accuracy. Finally, for all the selected historical stock values it is evaluated through core ARIMA, ABC-ARIMA and modified ABC-ARIMA (proposed model). RMSE is a scale dependent measure that measures the error between real value of the stock and estimated value of the stock by proposed model equation-15 [26]. Noticeably, the proposed model had sufficient low values for RMSE and AIC and BIC as likelihood penalization criteria. The proposed model improves the RMSE values for the selected stocks BPCL (14.77%), GAIL (2.95%), GSPL (24.33%), HINDPETRO (7.14%), IGL (4.53%), IOC (3.57%), MRPL (2.40%), OIL (10.98%), ONGC (5.52%) and RELIANCE (3.08%) from core ARIMA. Remarkably, the range of improvements of RMSE from 2.40% to 24.33% based on stock's values and volatile behavior.

$$\text{Root Mean Square Error}(RMSE) = \sqrt{\sum_{i=1}^{n}\left(Y_t - Y_t'\right)^2/N} \tag{15}$$

4.4 Multi-step ahead prediction

In this paper, multi-step time series forecasting is examined through selected stock's future prediction. In this approach

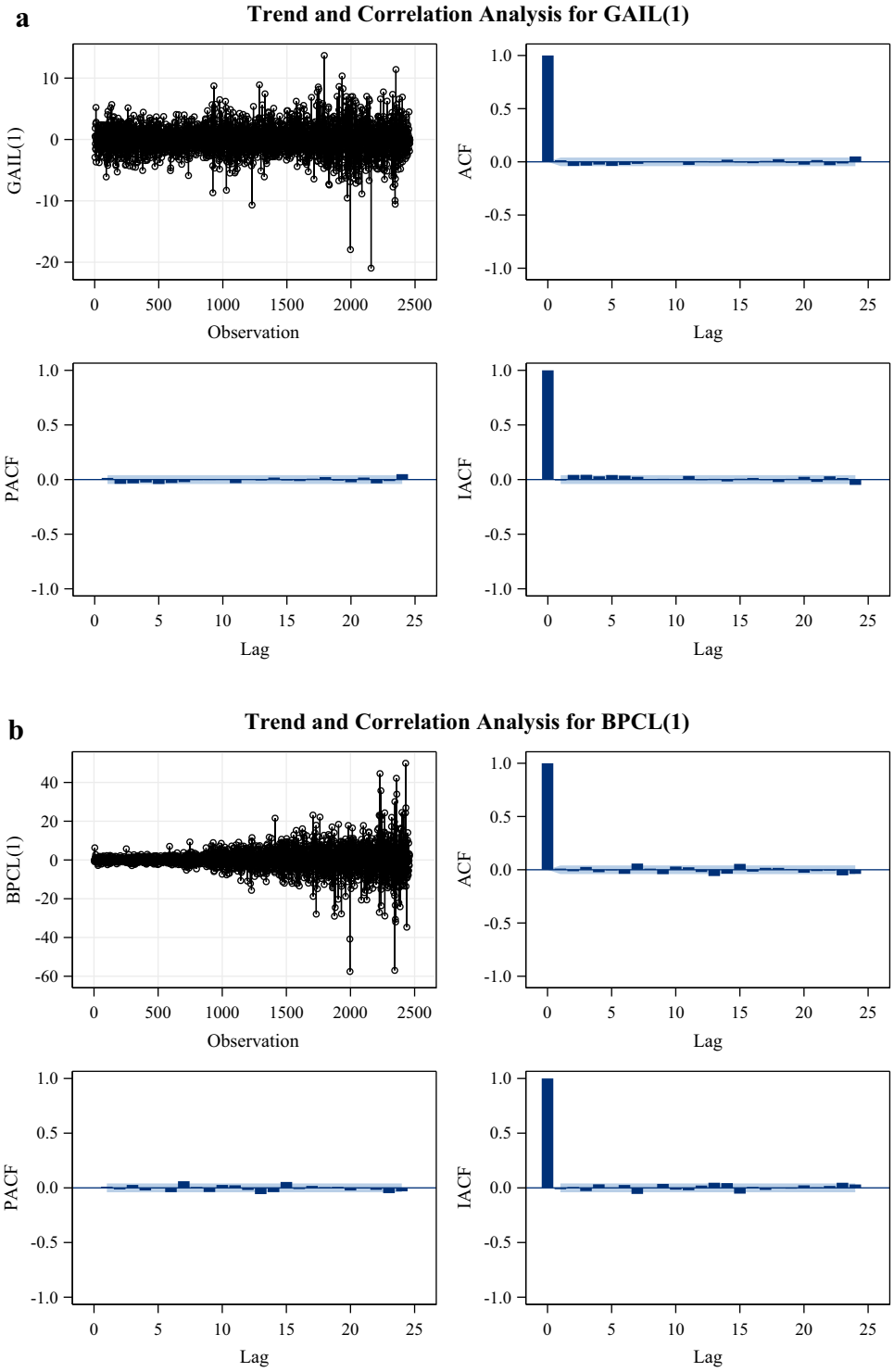


Fig. 2 **a** Trend and correlation analysis for GAIL. **b** Trend and correlation analysis for BPCL. **c** Trend and correlation analysis for GSPL. **d** Trend and correlation analysis for HINDPETRO. **e** Trend and correlation analysis for IGL. **f** Trend and correlation analysis for IOC. **g** Trend and correlation analysis for MRPL. **h** Trend and correlation analysis for OIL. **i** Trend and correlation analysis for GSPL. **j** Trend and correlation analysis for HINDPETRO

we investigate Oil Drilling and Exploration and Refineries of the oil and gas sector of NSE and BSE, India from Sept 1, 2010 to August 31, 2020. Historical datasets of four stock companies that contain 2460 instances for each stock of the oil and gas sector. As one-step time series prediction has its own limitation since it can reflect the market trends not the actual momentum of the real time scenario. On the other hand, multi-step time series prediction deals with market uncertainties and momentum which help in financial decision making like stock returns and portfolio optimization. Based on existing works, there are five methods in multi-step ahead forecasting proposed by Taieb et al.

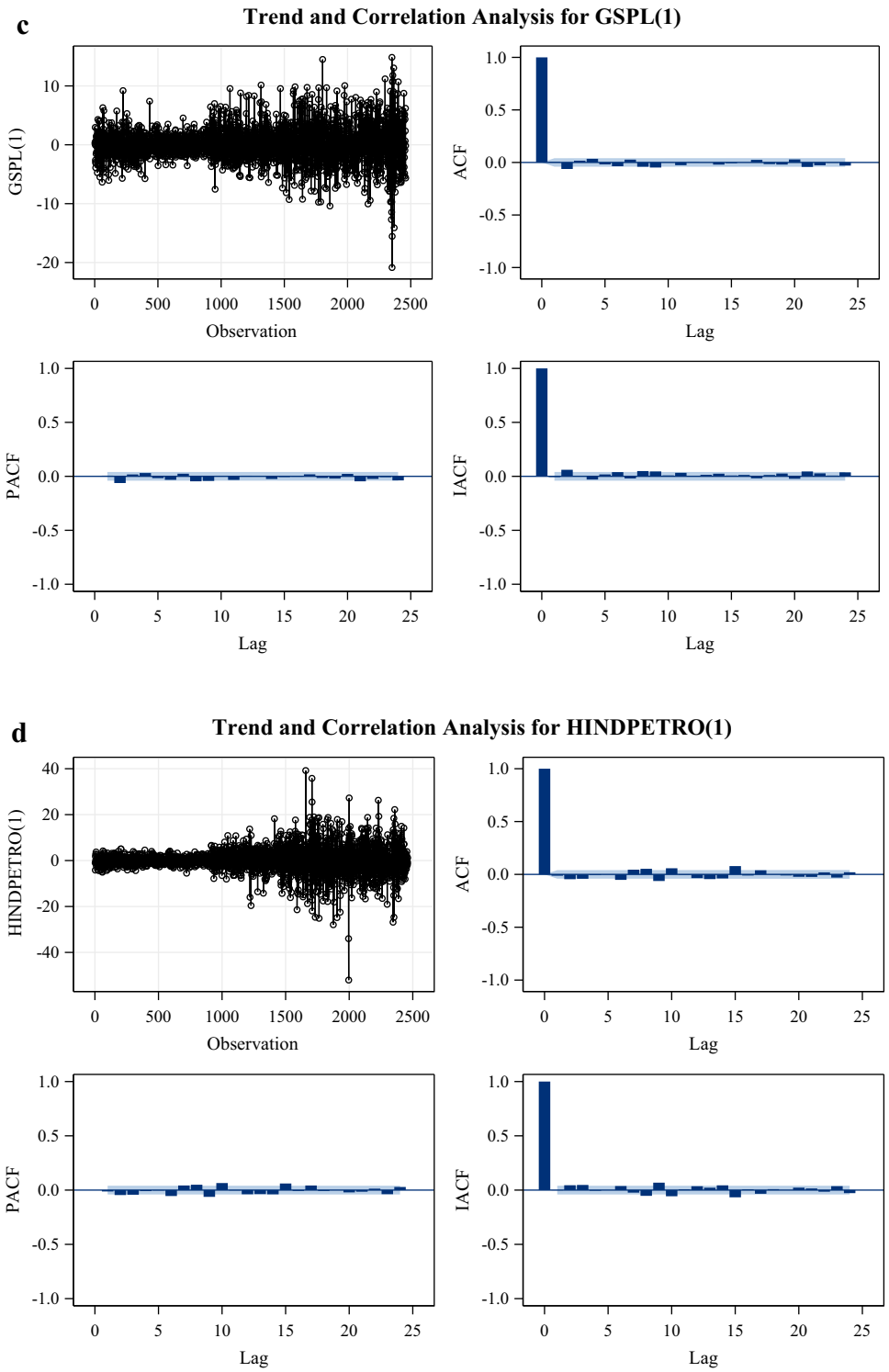


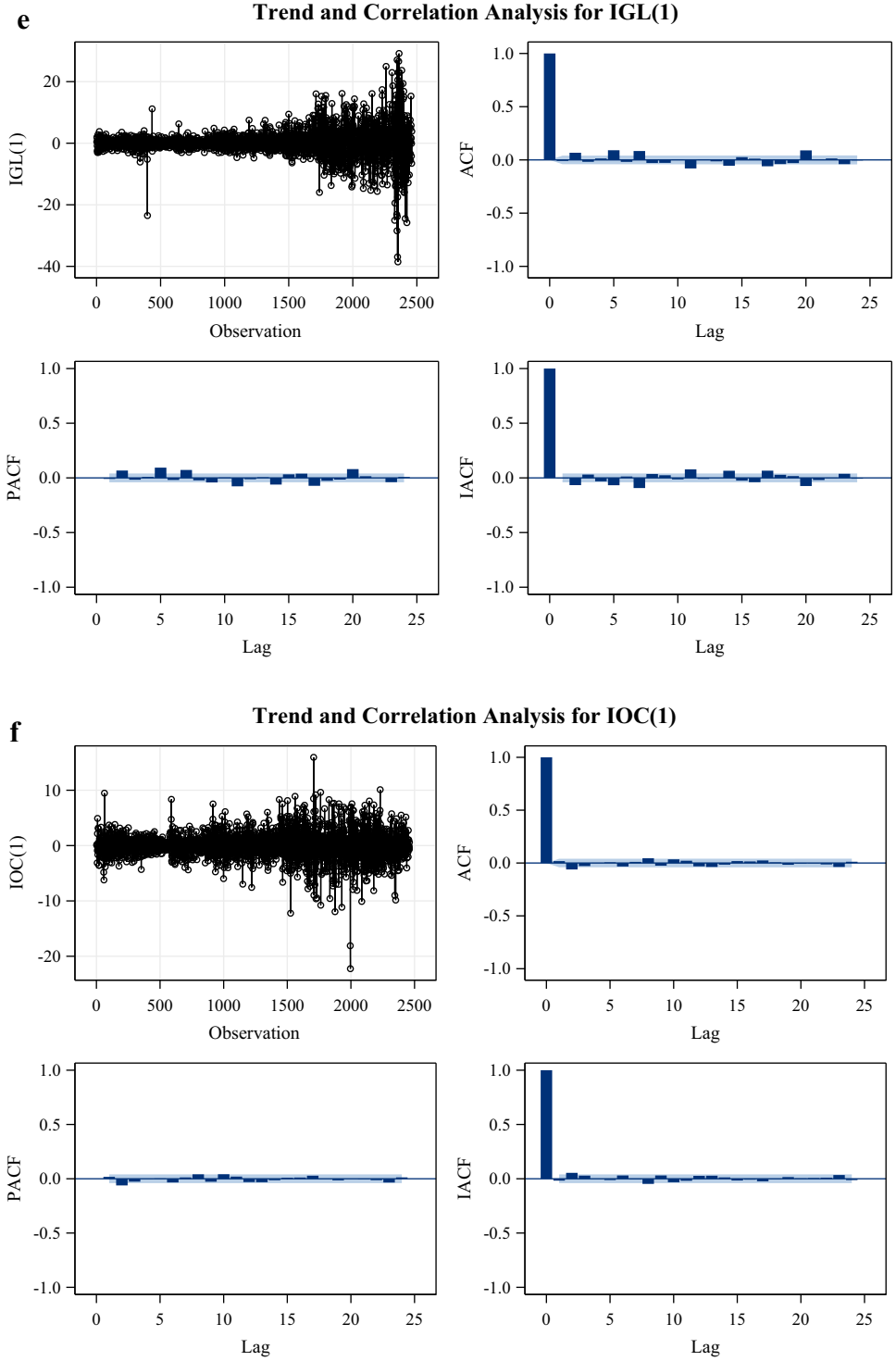
Fig. 2 continued

[27]. These methods are driven from direct, recursive and multiple input and multiple output (M IMO) strategies [28].

In this paper, direct recursive strategy is adopted for multi-step ahead forecasting of selected stocks that is the most intuitive method among all. direct recursive (DirREC) strategy works for model (H) from f_h time series (Y_1, Y_2, Y_3–Y_N) stock market data formulated in Eq. (16) [29].

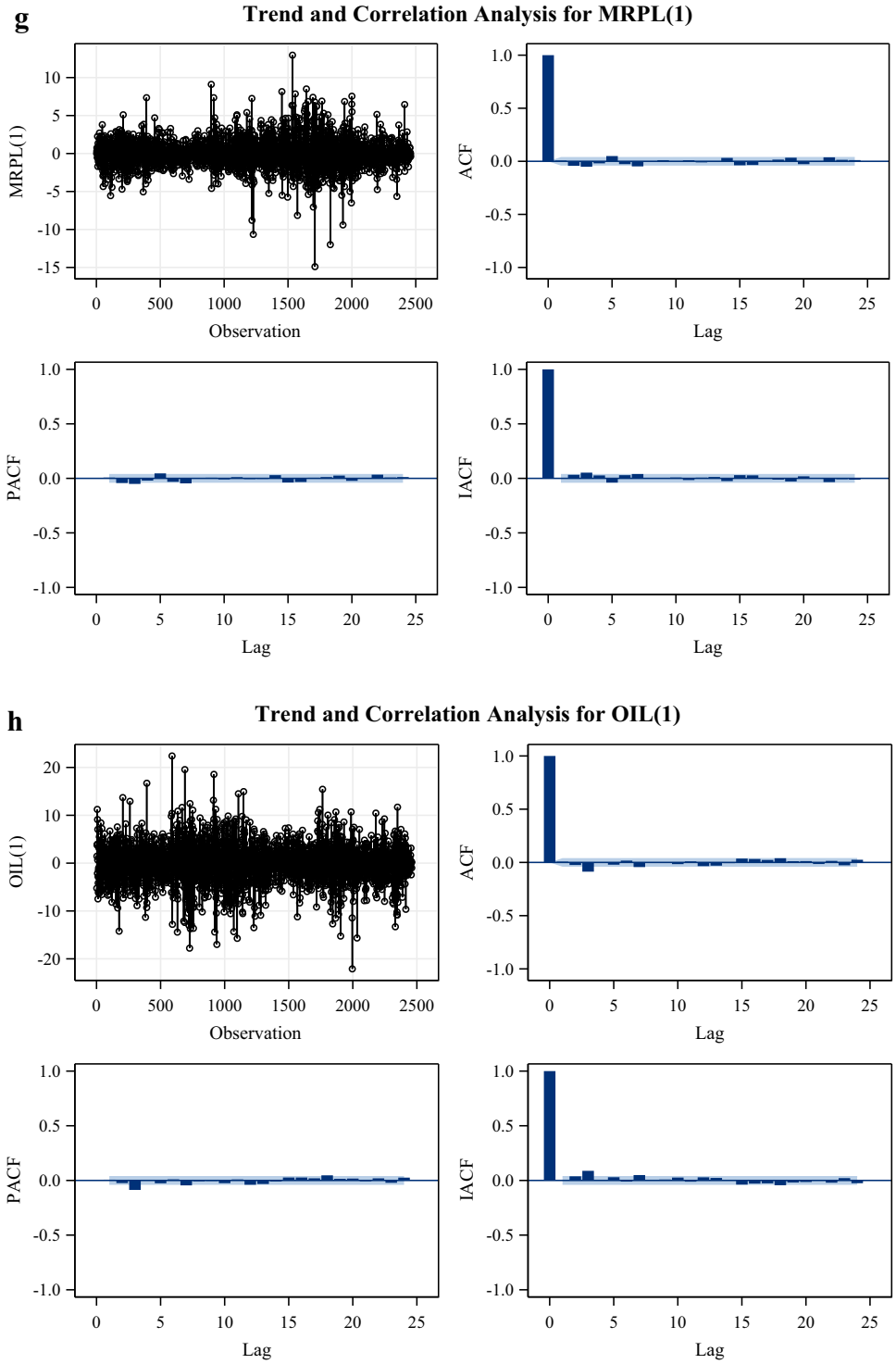
$$Y_{t+h} = f_h(Y_{t+h-1}, \ldots Y_{t-d+1}) \tag{16}$$

Fig. 2 continued



where, Y_{t+h} is the prediction of closing values of selected stocks of NSE and BSE illustrated on (Table 1) from Sept 1, 2010 to August 31,2020. Estimated values of the proposed model, modified ABC-ARIMA for test datasets of selected stocks are plotted using SAS studio in Figs. 3, 4, 5, 6, 7, 8, 9, 10, 11 and 12, respectively. Proposed model validates the hypothesis and outperforms other core ARIMA and ABC-ARIMA. The obtained accuracy as RMSE is observed least for the selected stocks. The 12 months future prediction of mentioned stocks (Table 1) are predicted with 95% confidence limit. Initially, ARIMA model has set parameters range from (0,0,0–12,12,12) thereafter, proposed model optimizes parameters of p, d

Fig. 2 continued

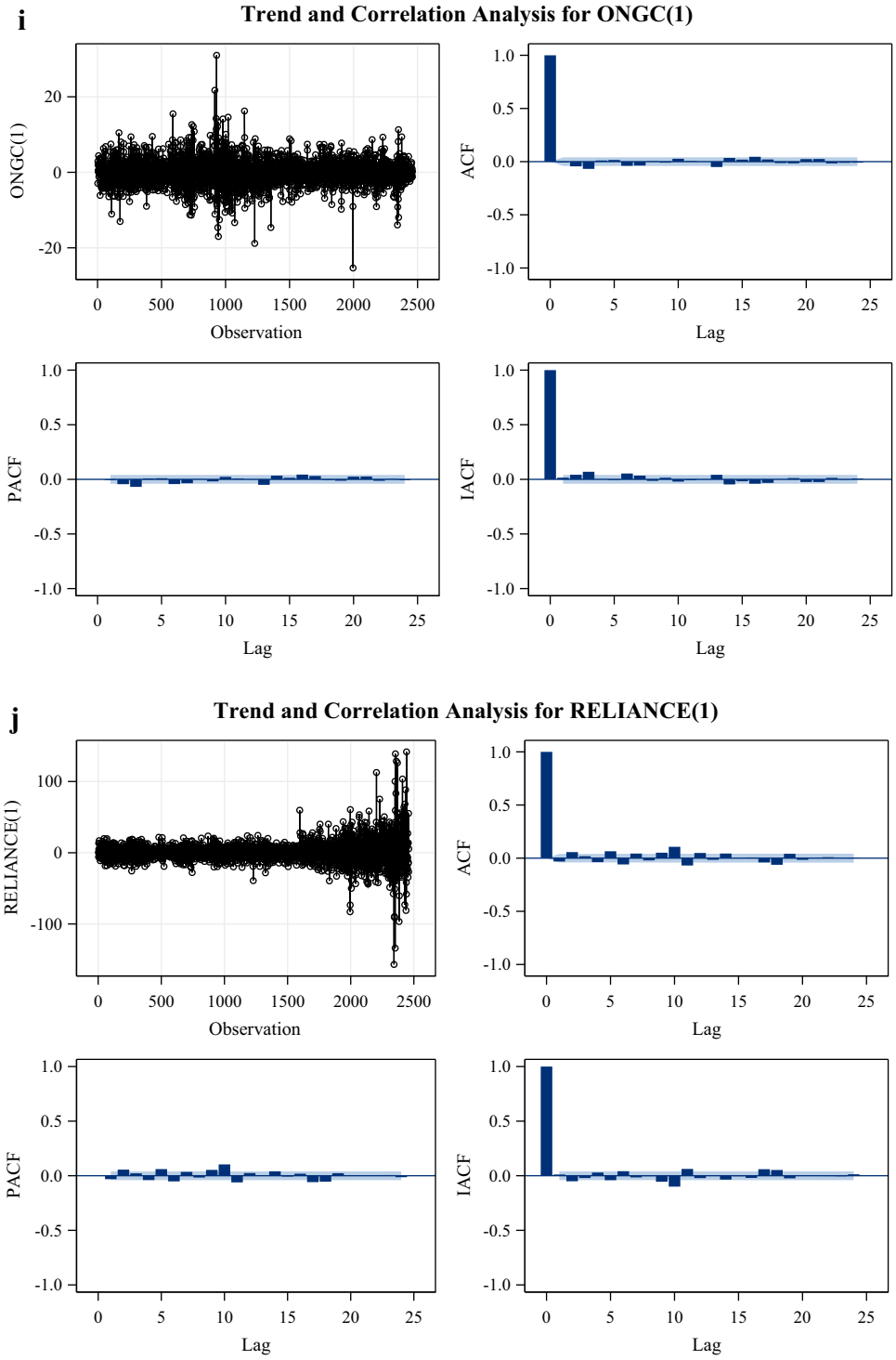


and q and obtains (2,1,3) as best fit combination. Multi-step ahead forecasting plots (Figs. 3, 4, 5, 6, 7, 8, 9, 10, 11, 12) illustrate that predicted values are closer to the actual values of selected stocks. Noticeably, this paper establishes and sets the constant trend for 12 months as multi-step ahead forecasting for all the selected stocks.

5 Conclusion

Financial time series analysis and prediction is one of the most challenging tasks. On one hand, one-step ahead prediction strategy works well to achieve higher accuracy in

Fig. 2 continued



the stock market domain. On the other hand, multi-step ahead prediction is yet to be established as a common tool and method among market traders and investors as it is a relatively complex task. This article proposed a hybrid model modified ABC-ARIMA inspired by optimization algorithms DE. Proposed model not only addresses the exploration and exploitation issues in time series prediction but also improves the forecasting accuracy and preserving data trends in multi-step time series stock forecasting. In this paper, experiments are performed over 10 years of dataset of Oil Drilling and Exploration and Refineries of oil and gas sector of NSE and BSE, India from Sept 1, 2010, to

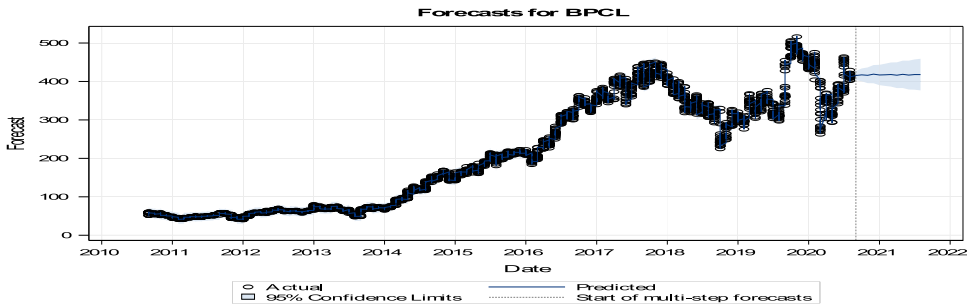


Fig. 3 Multi-step ahead forecasting for BPCL

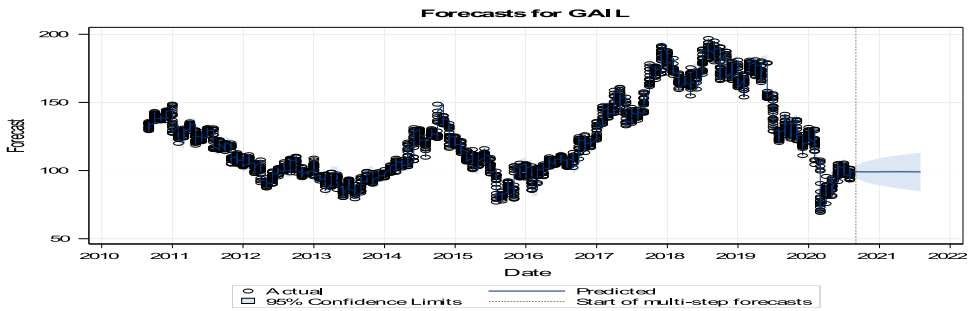


Fig. 4 Multi-step ahead forecasting for GAIL India Limited

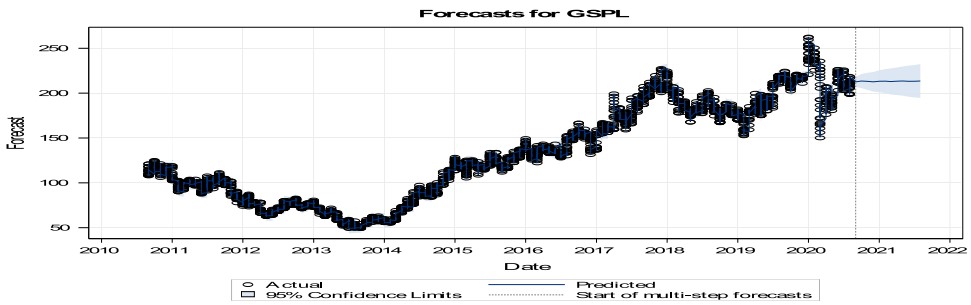


Fig. 5 Multi-step ahead forecasting for GSPL

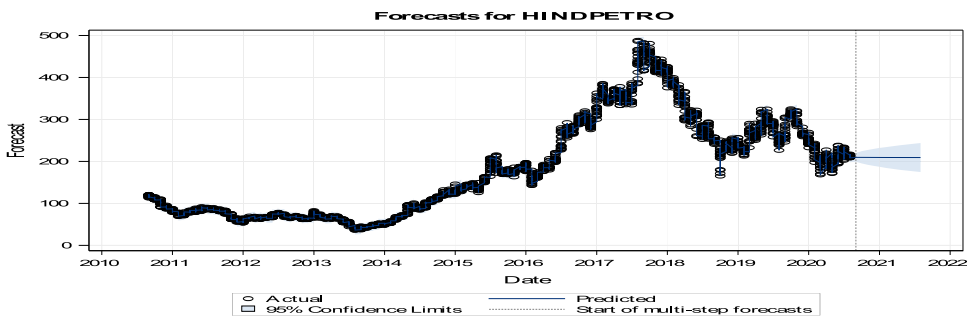


Fig. 6 Multi-step ahead forecasting for HINDPETRO

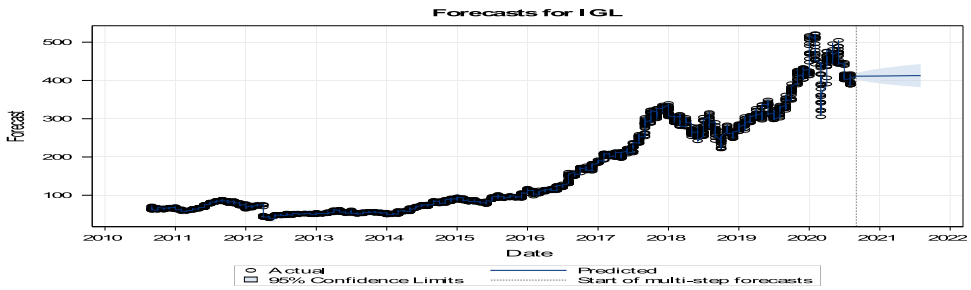


Fig. 7 Multi-step ahead forecasting for IGL

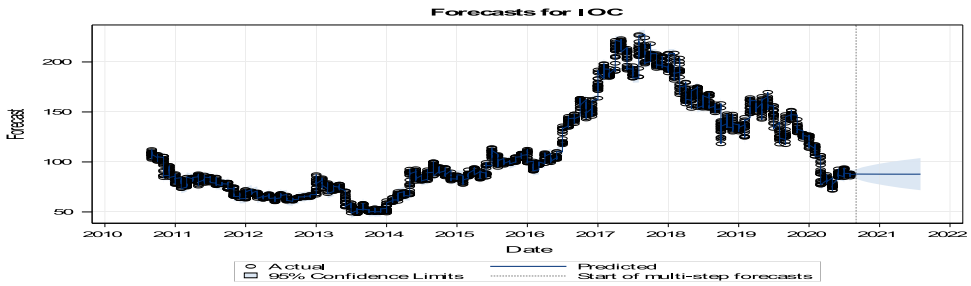


Fig. 8 Multi-step ahead forecasting for IOC

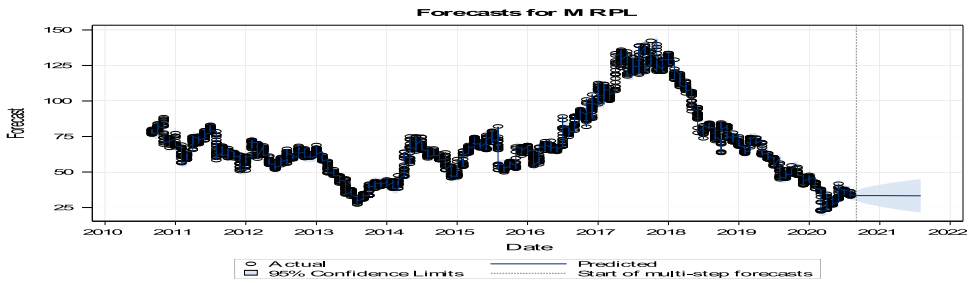


Fig. 9 Multi-step ahead forecasting for MRPL

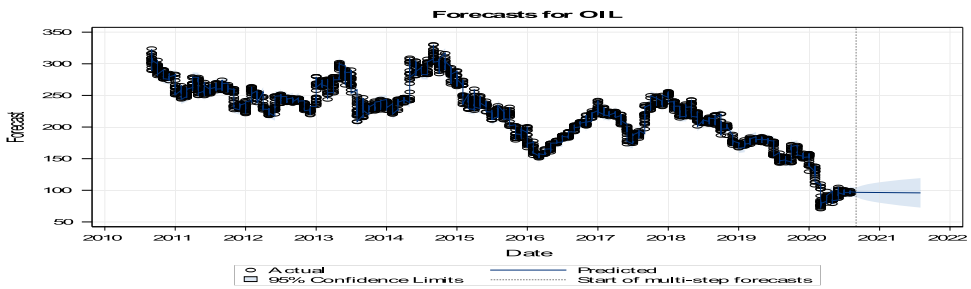


Fig. 10 Multi-step ahead forecasting for OIL

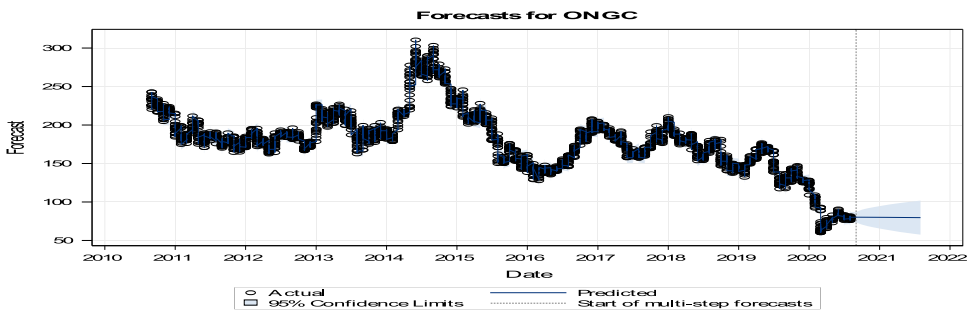


Fig. 11 Multi-step ahead forecasting for ONGC

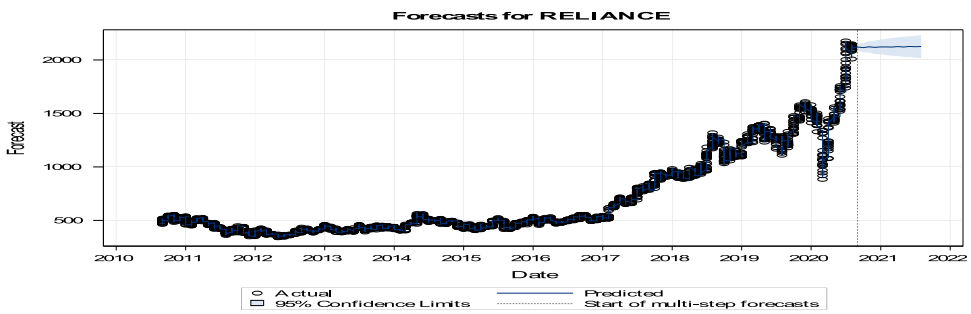


Fig. 12 Multi-step ahead forecasting for Reliance

August 31,2020. The selected stocks obtain remarkably less RMSE for the proposed modified ABC-ARIMA model in comparison with core ARIMA and ABC-ARIMA models. Optimized values for AR(p), I(d) and MA(q) are obtained as p d q (2,1,3). Remarkably, the range of improvements of RMSE from 2.40 to 24.33% for MRPL and GSPL respectively, based on stock's values and volatile behavior. Obtained result demonstrates that the proposed modified ABC-ARIMA hybrid model has superior performance than its counterparts in multi-step time series forecasting. As future scope, examining nonlinear patterns are key issues in time series domain specially, stock market prediction. Recurrent neural network (RNN) and LSTM) are well known nonlinear models preferred by research communities. Hybridization of such models can be effective to address linear as well as nonlinear patterns of dataset. Therefore, this issue can be considered as a future prospect of this study and to be explored further.

References

1. Box GE, Jenkins GM, Reinsel GC, Ljung GM (2015) Time series analysis: forecasting and control. Wiley, New York
2. Pai PF, Lin CS (2005) A hybrid ARIMA and support vector machines model in stock price forecasting. Omega 33(6):497–505
3. Babu CN, Reddy BE (2015) Prediction of selected Indian stock using a partitioning—interpolation based ARIMA–GARCH model. Appl Comput Inform 11(2):130–143
4. Domingos SDO, de Oliveira JF, de Mattos Neto PS (2019) An intelligent hybridization of ARIMA with machine learning models for time series forecasting. Knowl Based Syst 175:72–86
5. Khashei M, Bijari M (2010) An artificial neural network (p, d, q) model for time series forecasting. Expert Syst Appl 37(1):479–489
6. Khashei M, Bijari M (2011) A novel hybridization of artificial neural networks and ARIMA models for time series forecasting. Appl Soft Comput 11(2):2664–2675
7. Zhou K, Wang WY, Hu T, Wu CH (2020) Comparison of time series forecasting based on statistical ARIMA model and LSTM with attention mechanism. J Phys Conf Ser 1631(1):012141
8. Vantuch T, Zelinka I (2015) Evolutionary based ARIMA models for stock price forecasting. In: ISCS 2014: interdisciplinary symposium on complex systems. Springer, Cham, pp 239–247
9. Musdholifah A, Sari AK (2019) Optimization of ARIMA forecasting model using firefly algorithm. Indones J Comput Cybern Syst 13(2):127–136
10. Kumar R, Kumar P, Kumar Y (2021) Integrating big data driven sentiments polarity and ABC-optimized LSTM for time series forecasting. Multimed Tools Appl. https://doi.org/10.1007/s11042-021-11029-1
11. Kumar R, Kumar P, Kumar Y (2021) Analysis of financial time series forecasting using deep learning model. In: 2021 11th international conference on cloud computing, data science & engineering (confluence). IEEE, pp 877–881
12. Ballini R, Luna I, Lima LD, da Silveira RLF (1995) A comparative analysis of neurofuzzy, ANN and ARIMA models for Brazilian stock index forecasting. SCE-Computing in Economics and Finance
13. Wang JH, Leu JY (1996) Stock market trend prediction using ARIMA-based neural networks. In: Proceedings of international conference on neural networks (ICNN'96), vol 4. IEEE, pp 2160–2165
14. Karaboga D (2005) An idea based on honey bee swarm for numerical optimization, vol 200. Technical report-tr06, Erciyes University, Engineering Faculty, Computer Engineering Department, pp 1–10
15. Karaboga D, Gorkemli B, Ozturk C, Karaboga N (2014) A comprehensive survey: artificial bee colony (ABC) algorithm and applications. Artif Intell Rev 42(1):21–57
16. Mernik M, Liu SH, Karaboga D, Črepinšek M (2015) On clarifying misconceptions when comparing variants of the artificial bee colony algorithm by offering a new implementation. Inf Sci 291:115–127
17. Storn R, Price K (1997) Differential evolution—a simple and efficient heuristic for global optimization over continuous spaces. J Glob Optim 11(4):341–359
18. Das S, Mullick SS, Suganthan PN (2016) Recent advances in differential evolution—an updated survey. Swarm Evol Comput 27:1–30
19. Qin AK, Huang VL, Suganthan PN (2008) Differential evolution algorithm with strategy adaptation for global numerical optimization. IEEE Trans Evol Comput 13(2):398–417
20. Price KV (1996) Differential evolution: a fast and simple numerical optimizer. In: Proceedings of North American fuzzy information processing. IEEE, pp 524–527
21. Xiang W, Ma S, An M (2014) Habcde: a hybrid evolutionary algorithm based on artificial bee colony algorithm and differential evolution. Appl Math Comput 238:370–386
22. Zorarpacı E, Özel SA (2016) A hybrid approach of differential evolution and artificial bee colony for feature selection. Expert Syst Appl 62:91–103
23. Jadon SS, Tiwari R, Sharma H, Bansal JC (2017) Hybrid artificial bee colony algorithm with differential evolution. Appl Soft Comput 58:11–24
24. National Stock Exchange (NSE), Bombay Stock Exchange (BSE) (2020) Historical datasets. https://finance.yahoo.com/quote/. Accessed 10 Aug 2020
25. Mahan MY, Chorn CR, Georgopoulos AP (2015) White noise test: detecting autocorrelation and non-stationarities in long time series after ARIMA modeling. In: Proceedings 14th python in science conference (Scipy 2015), Austin, TX
26. Kumar R, Kumar P, Kumar Y (2020) Time series data prediction using IoT and machine learning technique. Procedia Comput Sci 167:373–381
27. Taieb SB, Sorjamaa A, Bontempi G (2010) Multiple-output modeling for multi-step-ahead time series forecasting. Neurocomputing 73(10–12):1950–1957
28. Taieb SB, Bontempi G, Atiya AF, Sorjamaa A (2012) A review and comparison of strategies for multi-step ahead time series forecasting based on the NN5 forecasting competition. Expert Syst Appl 39(8):7067–7083
29. An NH, Anh DT (2015) Comparison of strategies for multi-step-ahead prediction of time series using neural network. In: 2015 international conference on advanced computing and applications (ACOMP). IEEE, pp 142–149