Journal of Statistics Applications & Probability Letters
An International Journal

http://dx.doi.org/10.18576/jsapl/090101

A Comparative Simulation Study of ARIMA and Computational Intelligent Techniques for Forecasting Time Series Data

*Haitham Fawzy**, *EL Houssainy A. Rady and Amal Mohamed Abdel Fattah*

Department of Applied Statistics and Econometrics, Faculty of Graduate Studies for Statistical Research, Cairo University, Cairo, Egypt.

Received: 28 Mar. 2021, Revised: 12 Jun. 2021, Accepted: 23 Jul. 2021

Published online: 1 Jan. 2022

Abstract: This paper aims to use the computational intelligent techniques and hybrid models for forecasting time series data based on 100 generated data of the Autoregressive integrated moving average (ARIMA) models. There are three scenarios used in this study. Furthermore, the performances were evaluated based on three metrics mean square error (MSE), mean absolute error (MAE) and mean absolute percentage error (MAPE) to determine the more appropriate method and performance of model. The results of this study show that the hybrid ARIMA-ANN model was better than other models. The results also proved that applying hybrid models can improve the forecasting accuracy over the ARIMA and ANN models.

Keywords: Time Series, ARIMA, SVM, ANN, Hybrid Models, MSE, MAE, MAPE.

1 Introduction

Time series analysis and forecasting is a dynamic research area that plays important role in planning and decision making in several practical applications. Different kinds of forecasting models have been developed, and researchers have relied on statistical techniques to predict time series data. The idea of predicting future events of a time series based on its previous values received a strong impulse after Box and Jenkins (1970). They introduced a modeling cycle for the autoregressive (AR) model, which assumes that future events of a time series data can be expressed as a linear dependency to the past values. However, in real life, time series are mostly non-linear and nonstationary, such as financial and economic time series. Computational intelligence methods are computational methods that have been designed to handle the problems which traditional methods cannot solve effectively. Support vector machines (SVM), artificial neural networks (ANN), and fuzzy systems are the main categorical of computational intelligence methods [1]. These algorithms are considered as intelligent because of their ability to adapt to complex systems. SVM and ANN are very popular alternative algorithms to the ARIMA models for non-linear time series forecasting. However, time series data can contain both linear and nonlinear patterns. To address this, using a hybrid model can give better results in forecasting. Both linear and nonlinear patterns of the time series can be modeled by this method. While ARIMA model is used to capture the linear behavior of the time series data. ANN is used to model the nonlinearity in the series. In this paper we carried out a simulation study to evaluate and compare the performance of the ARIMA, computational intelligence methods, and hybrid method for time series forecasting under some time series models. One sample size, with size n=100 is used, and data are generated from a set of autoregressive integrated moving average model with three different combination of parameters. The following sections contains the simulation design and the simulation results that provide a comparison of the of the hybrid model, and the other models in terms of their accuracy.

* Corresponding author e-mail: haitham.fawzy@yahoo.com

2 Related Works

In the literature, there are a small number of studies on forecasting time series using machine learning algorithms, from researchers of diversified areas of economic, statistics, engineering, and science. [2] have revealed that the machine learning methods are the most popularly used in financial time series prediction. This method is popular because of its ability to identify the changes in price of financial markets with near accurate and the income will be generated just because of this accuracy in predictions. The researchers have tried to disprove the general theory of the financial economists that the predictability and profitable trading in financial market could not be made exactly by any technique rather than econometric methods. The authors have tried to prove that machine learning models are well be used in the accurate prediction of financial market than the best econometric methods. The authors have concluded that the financial forecasting is influenced by the factors such as market maturity, the methods used for the prediction, the scope for which it develops forecasting, the methodology used to access the model, and last but not the least the simulate model-based training. The investigational analysis shows that advanced prediction models are very much useful to predict the price changes in financial markets. [3] presented machine learning methods to statistical time series forecasting and compared the correctness of those methods with the correctness of conventional statistical methods and found that the machine learning methods are better and out top using both measures of accuracy. They provide the reason for the accuracy of machine learning methods is less that of statistical models and suggested some other achievable ways. [4] used two machine learning methods: Support Vector Machine (SVM) and K-Nearest Neighbor (KNN) for gold price forecasting. The sample data of the gold price were taken from November 1989 to December 2019. Data up till November 2016 were used to build the model while the remaining data were used to forecast the gold price and to check the accuracy of the models. The results indicated that the SVM method had a better forecast quality (in terms of MSE, MAE and MAPE) than KNN. [5]proposed a hybrid model of Autoregressive Integrated Moving Average (ARIMA) and Support Vector Machine (SVM) to predict the future value of Natural Rubber (NR) prices. The experimental results show that the proposed model performs the best whereby compared to single ARIMA and SVM models. The results also show hybrid model can be an effective tool in improving the forecasting accuracy by reducing the model forecast error. [6] used tree-based methods for time series forecasting and compared the correctness of those methods with the correctness of conventional statistical methods. The results indicated that the random forest method had a better forecast quality (in terms of MSE, MAE and MAPE) than other methods.

3 Predictive Models and Data

In this Section, we present the predictive models and data used in this study. We present the forecasting models, as well as the models used for simulation. Additionally, we present the model evaluation criteria carried out in this study.

3.1 Predictive Models

3.1.1 ARIMA Model

Autoregressive integrated moving average (ARIMA) model forecasts variable based on linear dependency to the past values to it. The models defined as AR, MA, and ARMA are preferred for stationary time series analysis. However, in real life, time series are mostly non-stationary. To fit stationary models, it is essential to get rid of the variation of non-stationary sources in time series. One solution to this, Box and Jenkins introduced the ARIMA model which can effectively transform the non-stationary data into stationary by introducing a differencing process and overcome the limitation [7]. In ARIMA models, the initial step is to eliminate this non-stationarity using differencing. It is done by subtracting a current observation from observation at the previous time step. As an example, a first-order differencing can be done by replacing y_t by $y_t - y_{t-1}$. By this procedure, the stationary model that is fitted to the differenced data must be summed or integrated so that it can provide a model for original non-stationary data [8]. Therefore, the ARIMA model is called Integrated ARMA. The general form of the ARIMA(p, d, q) process is described as:

$$\Delta y_t = \Phi + \theta_1 \Delta y_{t-1} + \theta_2 \Delta y_{t-2} + ... + \theta_p \Delta y_{t-p} + \varepsilon_t - \lambda_1 \varepsilon_{t-1} - \lambda_2 \varepsilon_{t-2} - ... - \lambda_q \varepsilon_{t-q} \quad (1)$$

where Δy_t is the differenced new series (after the subtractions) The right-hand side ”predictors” include both lagged errors and lagged values of y_t, ε_t is $WN(0, \sigma^2)$, $\theta = (1, ..., p)$ and $\lambda = (1, ..., q)$.

3.1.2 Support Vector Machine (SVM)

SVM is a computational intelligence method that attempts to find a hyperplane in the original input space to separate a given training set correctly and leave as much distance as possible from the closest instances to the hyperplane on both sides. The idea of SVM is to separate the dataset into a high-dimensional feature space and find the hyperplane that maximizes the margin [9]. The objective of SVM is to find a decision rule with good generalization ability through selecting some particular subset of training data, called support vectors. Assume a non-linear function, given by $y(x)$:

$$y(x) = w^T \phi(x) + b \tag{2}$$

where, w is the weight vector, b is the bias or threshold, and (x) represents a high-dimensional feature space that is nonlinearly mapped from the input space x. The goal of SVM is to determine the values of w and b to orientate the hyperplane to be as far as possible from the closest samples. The coefficients w and b are estimated by minimizing the following function:

$$min(\frac{1}{2}w^T w) \tag{3}$$

subject to the following constraints:

$$\begin{cases} y_i - w^T \phi(x_i) - b \leq \varepsilon \\ w^T \phi(x_i) + b - y_i \leq \varepsilon \end{cases}$$

This gives:

$$y_i(w^T \phi(x_i) + b) - 1 \leq 0 \tag{4}$$

The goal of the objective function in Equation (3) is to make the function as ”flat” as possible; that is, to make as ”small” as possible while satisfying the constraints. In order to solve Equation (3), slack variables are introduced to cope with possible infeasible optimization problems. One silent assumption here is that $f(x)$ exists; in other words, the convex optimization problem is feasible. However, this is not always the case; therefore, one might want to trade off errors by flatness of the estimate. This idea leads to the following primal formulations as stated in Vapnik (1995):

$$minimize\ \frac{1}{2}w^T w + C\sum_{i=1}^{m}(\zeta_i^+ + \zeta_i^-) \tag{5}$$

subject to:

$$\begin{cases} y_i - w^T \phi(x_i) - b \leq \varepsilon + \zeta_i^+ \\ w^T \phi(x_i) + b - y_i \leq \varepsilon + \zeta_i^- \\ \zeta_i^+, \zeta_i^- \geq 0 \end{cases}$$

Where $\phi(x)$ is a kernel function and and C ($>$ 0) is a pre-specified regularization constant and represents the weight of the loss function.

3.1.3 Artificial Neural Networks (ANN)

ANN are a part of Artificial Intelligence that have improved the mechanism of human thinking. ANN can be considered as a computer model or mathematical algorithms based on biological Neural Networks (brain). The idea of ANN revolves around how to simulate the brain through computers. The main feature of ANN is its capability to learn. ANN have aroused considerable interest in such diverse fields as medicine, biology, psychology, computer science, mathematics, economics, and statistics. The main reason behind this interest lies in the fact that ANNs are a general, flexible, nonlinear tool adept of approximating any kind of arbitrary function [10]. For the problems of time series forecasting, it is suitable to use the dynamic neural networks (DNN), where the network output depends on the present and previous values. NAR network makes the future forecasting of the data by using that data previous values. NAR network structure can be written as:

$$y(t) = f(y(t-1), y(t-2), ..., y(t-n)) + e(t) \tag{6}$$

Where $y(t)$ is a time series which values network is trying to forecast using series own n number of lags, $e(t)$ is the error term that occurs as result of difference between forecasted and actual values, while $f(.)$ is the transfer function of network which is often log-sigmoid (as pre-determined by MATLAB), however it could be re-specified if desired.

$y(t-1), y(t-2), ..., y(t-n)$ are called feedback delays, and can be thought of as an input layers in the system, while $y(t)$ is the output of the network.

3.1.4 Hybrid ARIMA-ANN Models

In recent years, the researchers have tried to use hybrid models instead of a single model for predicting time series data, where one of them offset the shortage of other. Combination between time series models and neural networks will give the most accurate results than if we use all of them alone. A given time series data may have both linear and nonlinear characteristics. So, a suitable combination of both linear and nonlinear models, yields a more accurate prediction model than individual models for forecasting time series data of different origin [11]. The hybrid model consists of a linear model ARIMA and a nonlinear model ANN. According to Khashei and Bijari [12], it is reasonable to consider a time series can be considered as a function of a linear and a nonlinear component, as shown in the Equation (7):

$$y_t = f(L_t + N_t) \tag{7}$$

Where L_t denotes the linear component and N_t denotes the nonlinear component. These two components are estimated from data using the following three steps. First, ARIMA model is used to model linear components in the data. Second, calculate the residual from the linear. Let r_t denote the residual which is represented by

$$r_t = f(y_t - \hat{L}_t) \tag{8}$$

The residual represents nonlinear components that cannot be modeled by ARIMA model. Finally, ANN is used to implement functional relationship between the past observed data $y_{t-1}, y_{t-2}, ..., y_{t-n}$, present forecast of linear component $\hat{L}_t$, and data , present forecast of linear component , and past error data $r_{t-1}, r_{t-2}, ..., r_{t-n}$ as indicated in Equation (9):

$$y_t = f(y_{t-1}, y_{t-2}, ..., y_{t-n}, r_{t-1}, \hat{L}_t, r_{t-2}, ..., r_{t-n}) \tag{9}$$

Fig.1 shows the flowchart for the hybrid model, ARIMA -ANN.

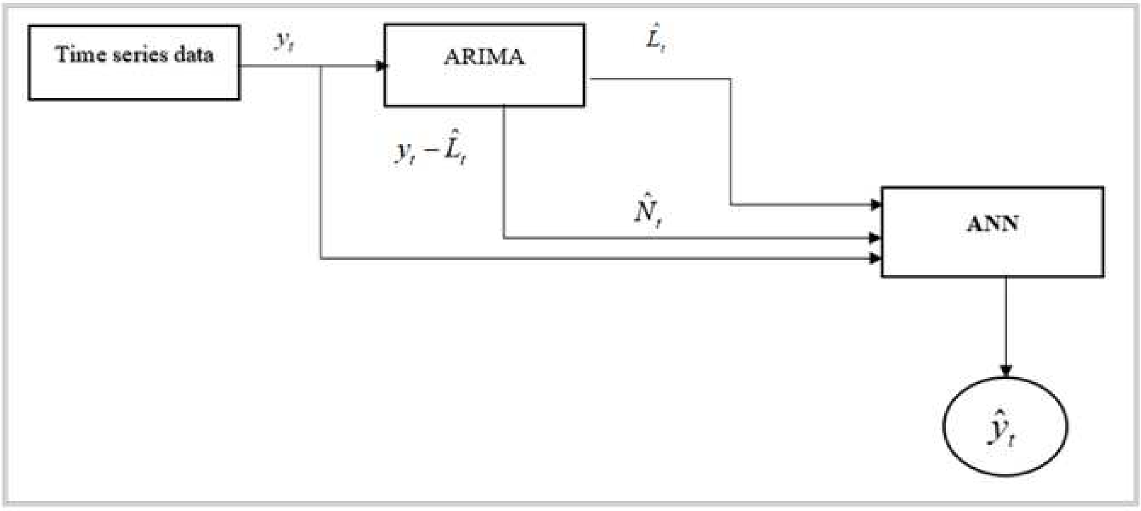

Fig. 1: Flowchart of the hybrid ARIMA -ANN model

3.2 Data

Data were simulated from a set of ARIMA models ((ARIMA (0,1,1), (ARIMA (1,1,0) and (ARIMA (1,1,1)) models with size n=100 using the arima.sim function in R. Each model is replicated 100 times using different initial random seeds for the error term. The series are described in Table 1 and depicted in Fig. 2.

Table 1: Simulation experiments, their respective models, and their parameters

Experiment	Model	parameters	Size (total, testing)
1s	ARIMA (0,1,1)	$\theta_1 = 0.8$	(100,10)
2	ARIMA (11,0)	$\phi_1 = 0.6$	(100,10)
3	ARIMA (1,1,1)	$\theta_1 = 0.4, \phi_1 = 0.5$	(100,10)

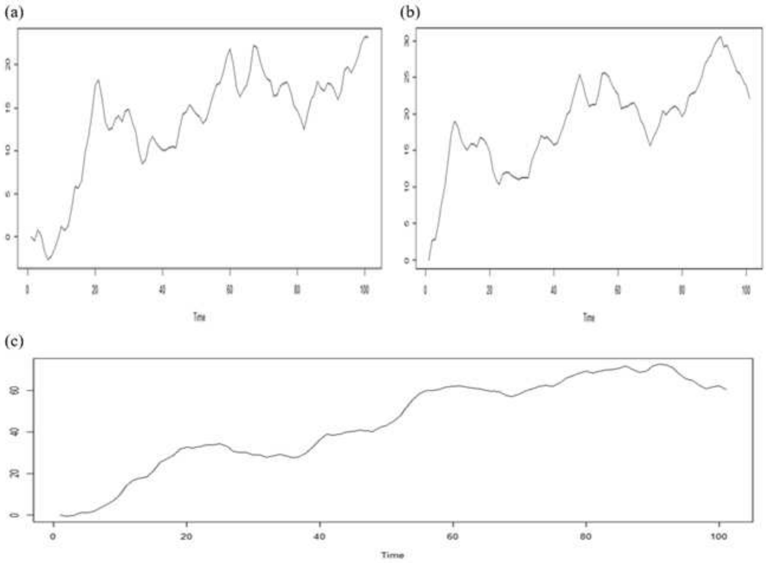

Fig. 2: The time plots of the series: (a) Experiment one, (b) Experiment two, (c) Experiment three

3.3 Model Evaluation

Model evaluation process is as important as model development process. According to accuracy performance results, the process of model development including selection of a proper method, hyperparameter optimization, etc. could be reevaluated until obtaining the most appropriate model. The accuracy performance of predictions can only be determined by considering how accurate a model performs on a new dataset which is not used while developing the model. It is important to find the best prediction method that produces the most accurate results in the evaluation process. Various error metrics have been developed in the literature to measure ”the most accurate” one. The error metrics define ”error” e_t as the difference between actual observed value Y and its prediction $\hat{Y}$ at time t. This difference refers to the unpredictable part of the corresponding observation. In this study we used three well-known error metrics to evaluate models performances.

• Mean Squared Error (MSE) is defined as follows:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{10}$$

• Mean Absolute Error (MAE) is defined as follows:

$$MAE = \sum_{i=1}^{n}\frac{|y_i - \hat{y}_i|}{n} \tag{11}$$

• Mean Absolute Percentage Error (MAPE) is defined as follows:

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\frac{|y_i - \hat{y}_i|}{y_i} \times 100 \qquad , y_i \neq 0 \tag{12}$$

4 Results

To evaluate the prediction capability of the predictive models, the predictive models are applied to the three simulated series. The prediction performance measures involved in this paper consist of three measures: mean square error (MSE), mean absolute error (MAE) and mean absolute error (MAPE). Table 2 presents the obtained prediction performance results through ARIMA, SVR, ANN, and the hybrid model in terms of MSE, MAP, and MAPE. From Table 2, it can be clearly seen that hybrid model have achieved lower errors than other models. This may suggest that neither ARIMA nor ANN model captures all of patterns in the data. For example, model one, in terms of MAPE, the hybrid ARIMA-ANN model can improve 81.03% over than ARIMA model in the test data. In addition, the hybrid ARIMA-ANN can improve 53.24% over than ANN model in the test data.

Table 2: The obtained prediction performance results

Model	Experiment one			Experiment two			Experiment three		
	MSE	MAE	MAPE	MSE	MAE	MAPE	MSE	MAE	MAPE
ARIMA-ANN	0.35	0.44	2.31	0.65	0.55	2.02	0.63	0.65	1.01
ANN	1.15	0.98	4.94	0.88	0.71	2.61	1.34	0.96	1.51
SVR	1.16	0.99	5.01	1.01	0.83	3.08	1.93	1.19	1.82
ARIMA	9.06	2.59	12.18	29.19	4.45	18.02	161.42	11.14	17.8

Fig. 3 shows the comparison of actual values and forecast values for all three time series. The ANN, SVR, and hybrid model were found to predict closer to the actual value and have a similar pattern with the actual data. Meanwhile, ARIMA and other model show unsatisfactory forecasting performance with the actual data.

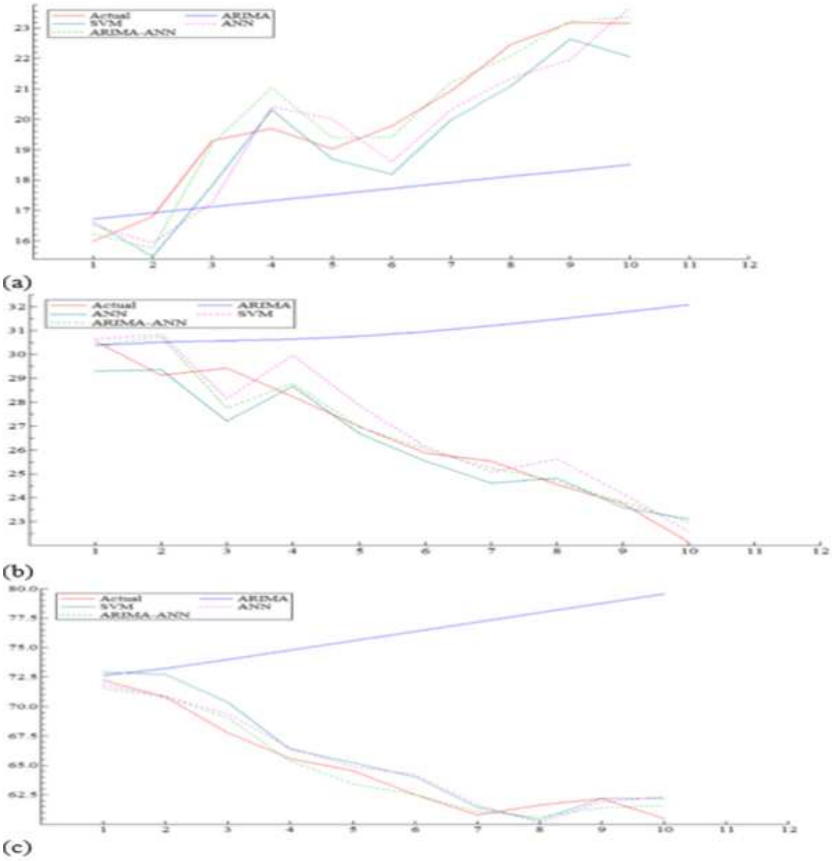

Fig. 3: Actual and Forecast Results for: (a) Experiment one, (b) Experiment two, (c) Experiment three

5 Conclusion

The paper aimed to find the best performing model for the time series data, to find out whether statistical, computational intelligent techniques or hybrid models are more accurate for forecasting time series data. The predictive models were applied to the simulated data generated from ARIMA models and comparing between models to see which one is better in forecasting the time series data. The results show that the hybrid model can give the best performance in the data set and measures (i.e., MSE, MAE and MAPE) and has an acceptable performance for modeling and forecasting of time series data in all the considered situations. The most important finding was that applying hybrid models can improve the forecasting accuracy over the ARIMA and ANN models.

Acknowledgement

The authors are grateful to the anonymous referee for a careful checking of the details and for helpful comments that improved this paper.

References

[1] Berthold, M., & Hand, D. J. (2003). Intelligent data analysis (Vol. 2). Berlin: Springer.

[2] Hsu, Ming-Wei, Stefan Lessmann, Ming-Chien Sung, Tiejun Ma, & Johnnie EV Johnson. ”Bridging the divide in financial market forecasting: machine learners vs. financial economists. Expert Systems with Applications., 61 (2016), 215-234.

[3] Makridakis, S., Spiliotis, E., & Assimakopoulos, V. (2018). The M4 Competition: Results, findings, conclusion and way forward. International Journal of Forecasting, 34(4), 802-808.

[4] Fawzy, H., Rady, E. H. A., & Abdel Fattah, A. M. (2020). Comparison between support vector machines and K-nearest neighbor for time series forecasting. J. Math. Comput. Sci., 10(6), 2342-2359.

[5] Jong, L. J., Ismail, S., Mustapha, A., Abd Wahab, M. H., & Idrus, S. Z. S. (2020). The Combination of Autoregressive Integrated Moving Average (ARIMA) and Support Vector Machines (SVM) for Daily Rubber Price Forecasting. In IOP Conference Series: Materials Science and Engineering (Vol. 917, No. 1, p. 012044). IOP Publishing

[6] El Houssainy A. Rady, Fawzy, H., & Abdel Fattah, A.M. (2021). Time Series Forecasting Using Tree Based Methods. Journal of Statistics Applications & Probability., 10(1), 229-244.

[7] Chatfield, C., & Xing, H. (2019). The analysis of time series: an introduction with R. CRC press.

[8] Hyndman, R. J., & Athanasopoulos, G. (2018). Forecasting: principles and practice. OTexts.

[9] Yaseen, Z. M., Sulaiman, S. O., Deo, R. C., & Chau, K. W. (2019). An enhanced extreme learning machine model for river flow forecasting: State-of-the-art, practical applications in water resource engineering area and future research direction. Journal of Hydrology, 569, 387-408.

[10] Ghiassi, M., Saidane, H., & Zimbra, D. K. (2005). A dynamic artificial neural network model for forecasting time series events. International Journal of Forecasting, 21(2), 341-362.

[11] Alwee, R., Shamsuddin, S. M., & Sallehuddin, R. (2013). Hybrid support vector regression and autoregressive integrated moving average models improved by particle swarm optimization for property crime rates forecasting with economic indicators. The Scientific World Journal, 2013.

[12] Khashei, M., & Bijari, M. (2010). An artificial neural network (p, d, q) model for time series forecasting. Expert Systems with applications, 37(1), 479-489.