Contents lists available at Science-Gate

International Journal of Advanced and Applied Sciences

Journal homepage: http://www.science-gate.com/IJAAS.html

The effect of an accumulation algorithm on the predictive accuracy of ARIMA models

Mubarak H. Elhafian *, Hamid H. Hussien, Abdelmgid O. M. Sidahmed, Muhammed Aljifri

Department of Mathematics, College of Science and Arts, King Abdulaziz University, Jeddah, Saudi Arabia

ARTICLE INFO

Article history:
Received 25 December 2024
Received in revised form
2 May 2025
Accepted 3 October 2025

Keywords:
Forecasting accuracy
ARIMA model
Accumulated data
Time series
Model evaluation

ABSTRACT

This study explores the use of the autoregressive integrated moving average (ARIMA) data-driven modeling approach for forecasting peanut yields in Sudan. Two tests were conducted: one using the original dataset and another using accumulated data. The main objective was to improve forecasting accuracy by applying a method that incorporates accumulated data for future predictions. The results, based on a comparison of the two tests, indicate that the proposed approach enhances prediction clarity. Model identification showed an increase in the coefficient of determination, a decrease in the Bayesian information criterion (BIC), and a reduction in the mean absolute error. These outcomes suggest that the proposed method may provide more accurate forecasts and could be useful for forecasting in various fields.

© 2025 The Authors. Published by IASE. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

1. Introduction

Solving forecasting problems is essential for successful agricultural planning and development. Therefore, it has become important to develop methods that enable decision-makers to understand underlying phenomena in the future (Pankratz, 2009). Many types of time series analyses are available, for example, the classical method, the Box-Jenkins method, and artificial neural networks. Over the years, several research studies examining predictions about economic phenomena have been conducted with a variety of solutions proposed (Shumway and Stoffer, 2017). Statistical models have been used to make predictions based on historical data. The time series model is the most important model for forecasting the future of economic phenomena (Shumway and Stoffer, 2017). They are popular in practice because their computation is simple. In addition, a variety of other intelligent methods have been used for forecasting time series data, namely artificial neural networks (ANN) (Wang et al., 2018), as well as the fuzzy logic method, the hybrid method, the support vector machine, and others. Because of their superior forecasting performance and ability to detect and extract non-linear relationships from a given set of information, a wide variety of ANNs are commonly used to model time series in various applications.

The Sudanese economy's dependence on the expansion of its resources for the growth of agricultural production makes it vulnerable to the fluctuations that characterize agricultural production due to its dependence on rainfall. In its effect on the Sudanese economy's rate of growth, it is clear that the most important constraint and limitation of production in the agricultural sector is the absence of a clear agricultural policy regulating this sector, establishing sustainable structuring, or increasing productivity. Oilseed crops are important plant-based foods in most developing countries because they are cheap sources of protein and alternatives to costly animal proteins. One of the most important oilseed plants cultivated in Sudan is the peanut, which is one of the country's principal cash crops (Elshafie et al., 2011). In the early 1970s, the peanut ranked second after cotton among Sudan's top exports, but it dropped to fifth place after cotton, sorghum, sesame, and gum Arabic in the early 1980s, when the quantity exported fluctuated relative to fluctuations in production, productivity, and cultivated area (Zhuo et al., 2016). With the secession of South Sudan in 2011, the country retreated from the leading African countries in terms of area cultivated and peanut production; consequently, Sudan ranked second after Nigeria and seventh globally. However, over the last 30 years, Sudan's production of peanut seeds has substantially decreased for two reasons. The first is that the government neglected agriculture to focus more on increasing oil production, which greatly decreased

* Corresponding Author.
Email Address: hafian10@yahoo.com (M. H. Elhafian)
https://doi.org/10.21833/ijaas.2025.10.021
Corresponding author's ORCID profile:
https://orcid.org/0000-0002-3619-1617
2313-626X/© 2025 The Authors. Published by IASE.
This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/)

with the secession of South Sudan in 2011. The second reason is that the country has been caught up in a vicious cycle of insecurity and conflict in the area of agriculture.

The peanut is one of the most important oilseed crops since it is a main source of concentrated calories necessary for human and animal nutrition, while oilseed, cake, and haulms are traditionally used as animal feed. There are two types of this sandy summer crop grown in Sudan. In the western part of the county, in Darfur, it comprises approximately 60% to 70% of the total production and is known to be of better quality, while in Gazeria, the central part of Sudan, and in East Sudan, the other type is grown. The crop yields a quick cash return for farms because the lands are characterized by double or triple agricultural cycles. This results in better economic yields than other summer crops. Furthermore, the country consumes about 65% to 70% of the product locally and exports about 30% to 35% to European and Arab countries, although the quantity exported has declined in recent years due to reductions in Sudanese production.

Forecasting plays a pivotal role in decision-making. It is an important and necessary tool for governments to develop future production policies, identify problems, and devise solutions. Good crop yield estimates affect seasonal crop management decisions. Many studies in Sudan have sought to find models that might accurately predict the future value of the peanut crop due to its importance as a cooking oil and its significant contribution to the gross national product (GNP). An important aspect of the modeling process is finding the best model to describe a phenomenon. In this work, we aspire to use a new method of modeling, using data transformation to determine a better model to estimate the production function for peanuts in Sudan. The rest of this paper is organized as follows. Section 2 summarizes the literature review of the study, and Section 3 describes the forecasting techniques, including the autoregressive integrated moving average (ARIMA) model. Section 4 describes the procedure for using ARIMA to construct the proposed method. Section 5 provides the results and discussion, while Section 6 serves as the conclusion.

2. literature review

Improving the accuracy of time-series forecasting models is a matter of constant attention from researchers. A large body of literature has demonstrated that basic strategies for combining frequently yield significantly higher accuracy than more complicated and sophisticated procedures (Jose and Winkler, 2008). The constrained ordinary least squares approach was an early tool for combining linear forecasts. It computes the combining weights by solving a quadratic programming problem that minimizes the sum of squared errors between the original and forecasted datasets, with the condition that the weights be nonnegative and unbiased, and hence predict dependent variables more accurately (Wang et al., 2023; Granger and Newbold, 2014). Several research studies in the literature have revealed that the naive simple average produced substantially better forecasting outcomes on several occasions than various additional complicated combination techniques (Jose and Winkler, 2008). In a recent comprehensive study, Jose and Winkler (2008) and Granger and Newbold (2014) found that trimmed means are marginally more accurate than simple averages and lower the probability of high errors. Moreover, a study suggests using the median as a remedial measure. It is far less sensitive to extreme values than the simple average. But there are varied results regarding the superiority of the simple average and median. The former produced better accuracy in Stock and Watson's work (Stock and Watson, 2004). Wu and Huang came up with the ensemble empirical mode decomposition, which fixes the problem of mode mixing in the original empirical mode decomposition (Huang and Wu, 2008). They made a noise-assisted data analysis method. A lot of different areas, like the solar cycle, seismic waves, figuring out the price of crude oil, and speaker identification systems, have used empirical mode decomposition to get signals out of noisy data that comes from nonlinear and stationary processes. Empirical mode decomposition is better than other methods because it can change on its own and is very local in both physical and frequency space (Wu and Tsai, 2011).

To test the accuracy of the developed model, Fattah et al. (2018) conducted comparative studies between experimental sales and simulations that were adopted in the Same period. It was reached that the selected model has high accuracy and the ability to simulate the dynamic behavior of sales (Fattah et al., 2018).

Adopted ANN Application Studies in pattern recognition, classification, and prediction, more than one algorithm has been used to obtain the results and then compare these algorithms and determine the best method through predictive accuracy (Allende et al., 2002).

3. Forecasting techniques

Solving forecasting problems is an essential task for successful planning and development. Many statistical techniques for time series forecasting have been developed. Of the conventional statistical methods, the ARIMA model is widely used in developing forecasting models (Štulajter, 2002).

3.1. Stationary time series

A series is said to be stationary when there is no systematic change in the mean (i.e., no trend), there is no systematic change in variance, and strictly periodic variations have been removed. In other words, one part of the data has properties exactly like those of every other part (Chatfield and Xing,

2019). Suppose that X_t represents the value of phenomena at time t. Stationarity can be defined with respect to the moments of the stochastic process $\{X_t\}$ (Kirchgässner et al., 2012).

Mean stationarity: A process is mean stationary if $E[X_t] = \mu_t = \mu$ for all t. Variance stationarity: A process is variance stationary if $V[X_t] = E[(X_t - \mu_t)^2] = \sigma_x^2 = \gamma(0)$ for all t. Covariance stationarity: A process is covariance stationary if $Cov[X_t, X_s] = E[(X_t - \mu_t)(X_s - \mu_s)] = \gamma(|t - s|)$. Weak stationarity: A stochastic process is said to be weakly stationary when it is both mean and covariance stationary. The correlation coefficient between the values of r_t and $\gamma_{t-\ell}$ is called the lag-ℓ autocorrelation of r_t and is commonly denoted by ρ_h, which, under the weak stationarity assumption, is a function of h only. Specifically, we define (Tsay, 2005).

$$\boldsymbol{\rho_\ell} = \frac{\boldsymbol{Cov(\gamma_t, \gamma_{t-\ell})}}{\sqrt{\boldsymbol{Var(\gamma_t)}\,\mathrm{Var}(\boldsymbol{\gamma_{t-\ell}})}} = \frac{\boldsymbol{Cov(\gamma_t, \gamma_{t-\ell})}}{\boldsymbol{Var(\gamma_t)}} = \frac{\boldsymbol{\gamma(\ell)}}{\boldsymbol{\gamma(0)}} \quad (1)$$

Eq. 1 is achieved under the assumption $Var(\gamma_t) = \mathrm{Var}(\gamma_{t-\ell})$ this assumption is used for weak stationarity.

The value of the autocorrelation is $-1 \le \rho_\ell \le 1$, where ρ_ℓ is the value of the autocorrelation.

3.2. Stationarity tests

To test the stationarity of a time series, Box and Pierce (1970) proposed the Portmanteau statistics:

$$Q^*(m) = T \sum_{\ell=1}^{m} \hat{\rho}_\ell^2 \quad (2)$$

under the following hypotheses:

$H_0 : \rho_1 = \rho_2 = \cdots = \rho_m = 0$
$H_a : \rho_i \neq 0$ for some i $\in [1, \cdots, m]$

where, Q_m^* is a chi-squared random variable with m degrees of freedom.

Ljung and Box (1978) modified the Q_m^* statistic, as shown below, to increase the power of the test in finite samples.

$$Q(m) = T(T+2) \sum_{\ell=1}^{m} \frac{\hat{\rho}_\ell^2}{T-\ell} \quad (3)$$

Simulated studies suggest that $m = ln(T)$, where T is the total number of observations.

Most of the time series probability theory concerns stationary time series, which is why time series analysis frequently enables one to transform a non-stationary series into a stationary one to use this theory (Wollstadt et al., 2014).

Although the value of the stationarity test is easy to calculate, most statistical programs depend on finding a range for autocorrelations (Mizon, 1995).

3.3. Autoregressive integrated moving average (ARIMA) model

Autoregressive Moving Average (ARMA) models, also called Box–Jenkins models, were introduced by Box et al. (2015) and have since become widely used for time series forecasting. These models are applied in many prediction tasks because of their effectiveness in modeling time-dependent data. The ARIMA model consists of three components: (1) the autoregressive (AR) process, (2) the moving average (MA) process, and (3) differencing to achieve stationarity, which leads to the ARIMA form (ARMA with integration). The general non-seasonal ARIMA model is expressed as ARIMA(p, d, q), where p represents the order of the autoregressive part, d indicates the number of differences required to make the series stationary, and q represents the order of the moving average part.

Model ARIMA(0, 0, 0) is classified as a white noise model in which there is no AR part, no MA part, and no difference involved. Similarly, model ARIMA(0, 1, 0) is known as a random walk model because it involves only one difference. The equation of the AR model is written as (Jonathan and Kung-Sik, 2008):

$$X_t = C + \phi_1 X_{t-1} + \phi_2 X_{t-2} + \cdots + \phi_p X_{t-p} \quad (4)$$

where, X_{t-1}, X_{t-2},...,$X_{t-\mathrm{p}}$ are the values for previous years, C is a constant, $\emptyset_1, \ldots, \emptyset_P$ are the AR parameters, p is the order of the AR, and e_t is the white noise series.

Likewise, the equation of the MA model can be written as:

$$X_t = \mu + e_t + \theta_1 e_{t-1} + \theta_2 e_{t-2} + \cdots + \theta_q e_{t-q} \quad (5)$$

where, μ is the expectation of X_t, θ_1,...,θ_q are the MA parameters, q is the order of MA, and $e_t, e_{t-1} + \cdots + e_{t-q}$ are the white noise error terms.

Integrating the two models yields the ARIMA model [ARIMA(p, q)] as follows:

$$X_t = C + \phi_1 x_{t-1} + \phi_2 x_{t-2} + \cdots + \phi_p x_{t-p} + e_t + \theta_1 e_{t-1} + \theta_2 e_{t-2} + \cdots + \theta_q e_{t-q} \quad (6)$$

where, p and q are the AR and MA terms, respectively.

The basic condition of the model is that the time series data properties are stationary, while statistical measures such as the mean, variance, and autocorrelation remain constant. However, if the time series data is non-stationary, the ARIMA model requires differenced data to be transformed to stationarity and is denoted as ARIMA(p, d, q) (Yuan et al., 2016).

The process of constructing ARIMA models requires the identification of the order of ARIMA(p, d, q), the estimation of model parameters, the checking of model validity, the selection of the best model, and, finally, forecasting. After determining the model and estimating its diagnostic parameters, we must check the form to determine if the model is suitable for the data. There are key criteria that should be used to verify this step so that any necessary revisions can be made to the form (Kumar and Anand, 2014).

3.4. Choosing the appropriate value of p, d, and q

Initially, possible temporary values are determined for p, d, and q (Table 1). The suitability of the proposed model is then verified (Fig. 1). If the proposed model is not suitable, the nature of inadequacy should be studied to arrive at another model (Cryer and Chan, 2008).

Table 1: The behavior of the ACF and PACF for ARMA models

	AR(p)	MA(q)	ARMA(p, q)
ACF	Tails off	Cuts off after lag q	Tails of
PACF	Cuts off after lag p	Tails off	Tails of

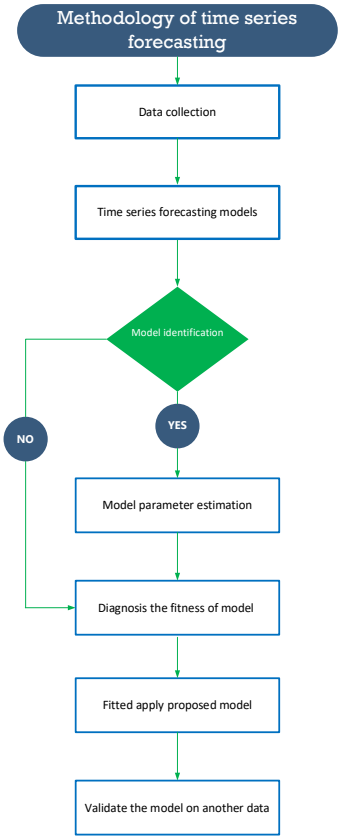

Fig. 1: Time series methodology framework

When we need to verify the independence of random error, a suitable model must be studied for autocorrelation patterns in the data that were not captured in the form during the estimation phase. Then, the remaining ACF should be examined to ensure that it was an insignificant statistical self-correlation coefficient. This step is very useful in proving that the model cannot be further improved (Mateos et al., 2022).

4. The proposed method

The aim of this paper is to improve the accuracy of the Box-Jenkins model using a transformation of the data to achieve this aim. We will use the following transformation:

$$\text{Let } X_t \equiv \textit{the series values no } t \quad (7)$$

$$\text{then} \quad Y_t = \sum_{t=1}^{t} X_t \quad (8)$$

$$\text{and} \quad z_t = \log Y_t \quad (9)$$

We propose applying the transformation of Eq. 9 to the formula of the Box-Jenkins model (Eq. 6).

$$Z_t = c + \phi_1 Z_{t-1} + \phi_2 Z_{t-2} + \cdots + \phi_p Z_p + \theta_1 e_{t-1} + \theta_2 e_{t-2} + \cdots + \theta_q e_{t-q} \quad (10)$$

The basic idea in using this transformation with agricultural data is that the production of the crops in any specific year generally depends on the value from the previous year. We are using aggregation to assist in controlling for the effects of the random error limit and logarithms to help make the values closely related.

By applying Eq. 10, we take into consideration the prediction values of the Z variable. To arrive at the real predictions, the concept of the inverse function must be used, that is;

$$Y_t = 10^{Z_t} \quad (11)$$

$$X_t = Y_{t+1} - Y_t \quad (12)$$

5. Results and discussion

5.1. The dataset

In this study, the data consist of time series data on peanut production in Sudan during the period between 1966 and 2018, which were obtained from the Central Bureau of Statistics. The Central Bureau of Statistics collects data annually from all agricultural projects (irrigated and rainy). The percentage of cultivated area of peanuts in the rainy sector is about 91% of the total area planted in Sudan, which produces about 67% of the total production in Sudan.

Fig. 2 presents annual peanut crop production in Sudan during the period between 1966 and 2018. The trend demonstrates that the quantity produced fluctuated between increasing and decreasing during the study period. Generally, the quantity of peanuts produced during recent years has significantly decreased, which confirms the need for conducting such studies to identify problems and determine solutions.

Table 2 presents the descriptive statistics regarding peanut crop production. The annual average production is 105,031.06 tons with a standard deviation of 80959.74 tons. The maximum value of production is 345546 tons, and the minimum production is 1835 tons in 1996. The range of production was 343711.

5.2. Results

In this study, we used the EViews statistical package to perform all ARIMA modeling. To determine if the proposed method increased the accuracy of prediction, we estimated the model first using the original data and then using the transformed data. We then checked the accuracy of the two methods. The results are provided in Tables 2, 3, 4, and 5. To obtain the best model, we used four evaluation measures: namely, the number of

significant coefficients in the model, the estimator of the error variance of the residual (labeled as SIGMASQ), the coefficient of determination (R-squared), Akaike's Information Criterion (AIC), and Schwarz's Information Criterion (SIC). The comparison between accuracy metrics is represented in Table 6. The model with a smaller value for AIC or SIC suggests a better fit. The following sections present the results of the two methods.

Method 1: ARIMA model using original data: In this method, the peanut data without transformation were used, and followed the steps mentioned in the theoretical part of this paper. First, we checked the stationarity of the series by plotting the autocorrelation function (ACF) and the partial autocorrelation function (PACF).

Using Eq. 2 to calculate the value of Q by substituting the values of ACF from Table 3, the value of $Q = 97.5$ and the critical value of ${\chi^2}_{24,0.05} = 36.415$ that means H0 is not accepted, and the series is not stationary. Fig. 3 shows the results of the ACF and PACF of the original series. It is evident that the series is not stationary, as seven ACF values lie outside the acceptable range. To achieve stationarity, a log transformation and first differencing were applied.

Table 2: Descriptive statistics on the quantity of the peanut crop, 1966–2016

Parameter	Value
Mean	105031.06
Std. error of mean	11120.653
Std. deviation	80959.574
Skewness	1.099
Std. error of skewness	0.327
Kurtosis	0.919
Std. error of kurtosis	0.644
Range	343711
Minimum	1835
Maximum	345546

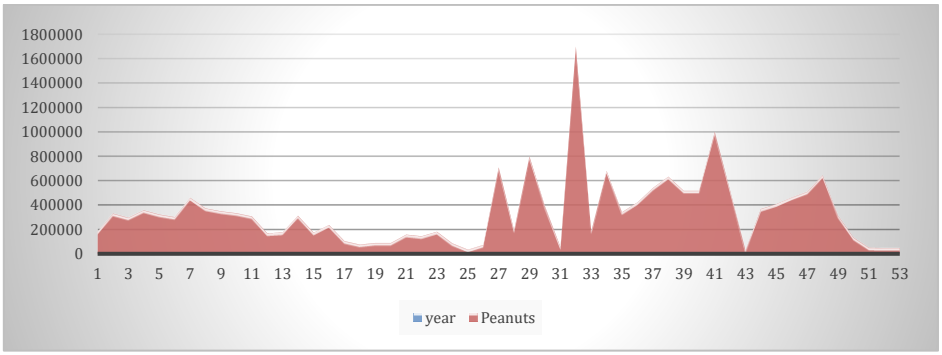

Fig. 2: Time series plot of peanut production data

Table 3: The ACF and the PACF of the original data

	AC	PAC	Q-Stat	Prob.
1	0.627	0.627	22.062	0.000
2	0.662	0.443	47.109	0.000
3	0.454	-011	59.108	0.000
4	0.379	-0.10	67.655	0.000
5	0.318	0.089	73.803	0.000
6	0.303	0.133	79.503	0.000
7	0.312	0.098	85.682	0.000
8	0.174	-0.27	87.634	0.000
9	0.290	0.212	93.223	0.000
10	0.145	0.003	94.656	0.000
11	0.211	-0.01	97.733	0.000
12	0.112	-0.11	98.618	0.000
13	0.105	-0.04	99.414	0.000
14	0.071	0.103	99.795	0.000
15	0.047	-0.00	99.967	0.000
16	0.130	0.090	101.30	0.000
17	0.034	-0.09	101.40	0.000
18	0.064	-0.14	101.73	0.000
19	0.003	0.106	101.73	0.000
20	-0.06	-0.17	102.15	0.000
21	-0.17	-0.26	105.01	0.000
22	-0.18	0.019	108.35	0.000
23	-0.22	0.094	113.37	0.000
24	-0.24	0.032	119.47	0.000

Table 4 presents the ACF and PACF of the transformed series. After applying the log transformation and first differencing, the Ljung–Box Q statistic was calculated using Eq. 2. Substituting the ACF values from Table 4 gives $Q = 17.96$. Since the critical value is ${\chi^2}_{24,0.05} = 36.415$, the null hypothesis H0 is accepted. This result indicates that the residual series is stationary. Fig. 4 presents the results of the ACF and the PACF of the original series after using a log transformation. It reveals that the time series of the data becomes stationary because all the ACFs are within the range.

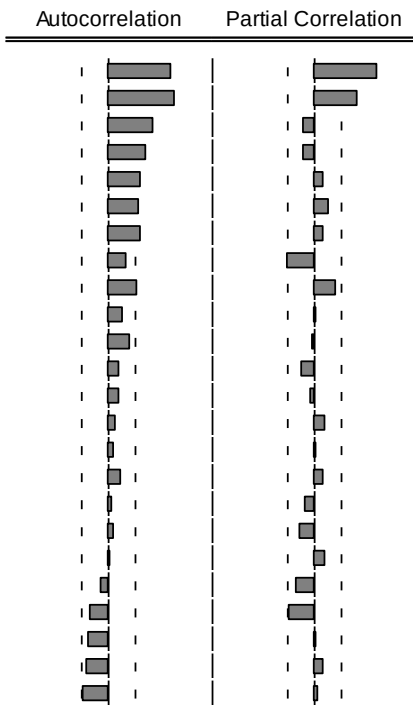

Fig. 3: The ACF and the PACF of the original data

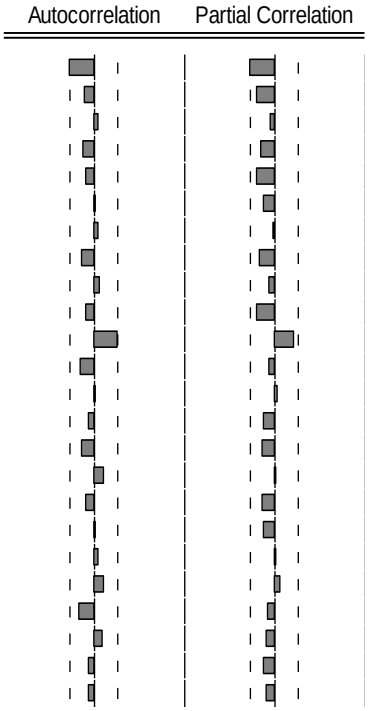

Fig. 4: The ACF and the PACF of the original series after using a log transformation

Table 4: The ACF and the PACF of the original series after using log transformation and the first difference

	AC	PAC	Q-stat	Prob.
1	-0.278	-0.278	4.2568	0.039
2	-0.105	-0.198	4.8765	0.087
3	0.051	-0.044	5.0266	0.170
4	-0.110	-0.143	5.7330	0.220
5	-0.089	-0.190	6.2074	0.287
6	0.030	-0.119	6.2619	0.394
7	0.064	-0.015	6.5179	0.481
8	-0.127	-0.169	7.5548	0.478
9	0.082	-0.056	7.9933	0.535
10	-0.092	-0.191	8.5621	0.574
11	0.273	0.226	13.666	0.252
12	-0.142	0.056	15.072	0.238
13	0.025	0.048	15.119	0.300
14	-0.054	-0.113	15.336	0.356
15	-0.135	-0.137	16.718	0.336
16	0.115	0.018	17.749	0.339
17	-0.086	-0.126	18.345	0.367
18	0.028	-0.117	18.704	0.429
19	0.059	0.011	18.703	0.476
20	0.109	0.076	19.738	0.474
21	-0.170	-0.072	22.348	0.380
22	0.095	-0.088	23.198	0.391
23	-0.065	-0.111	23.605	0.426
24	-0.060	-0.087	23.971	0.463

To choose a suitable model, we estimated the parameters of the model and compared them. The correlogram clearly illustrates that the proposed model of log transformation for the data is ARIMA(1, 0, 0). For comparison, the estimated number of models is presented in Table 4 to determine the best one according to the specified criteria. To estimate the ARIMA model of the time series unit using the EVIEWS program, the model rank (p, d, q) should be determined.

The model parameters are calculated by using the maximum likelihood estimation algorithm.

Table5: Comparison metrics of different ARIMA models for peanut data

Metrics	ARIMA(p, d, q)				
	(1, 0, 0)	(0, 0, 1)	(1, 0,1)	(0, 0, 2)	(2, 0, 0)
Significant coefficient	1	1	1	1	1
SIGMASQ	0.576	0.689	0.568	0.641	0.569
R-squared	0.402	0.285	0.399	0.320	0.397
Akaike info criterion	2.411	2.585	2.434	2.552	2.585
Schwarz criterion	2.523	2.697	2.584	2.701	2.494

Table 5 compares the metrics of various ARIMA models for peanut data. The comparison shows that the ARIMA(1, 0, 0) model is the best fit. This is because it has the most significant coefficients (1), the highest coefficient of determination R-squared (0.402), the lowest values for AIC (2.411), and the Schwarz criterion (2.523). ARIMA(1,0,0) will be estimated.

Table 6 shows the coefficients of the best model. Thus, the best equation in estimating the future production of peanuts becomes as follows:

$$X_t = 11.15033 + 0.663762\, X_{t-1}$$

To ensure the suitability of the model, the residuals are tested (whether the residuals distribution follows the normal distribution or not). From Table 6, it is noted that the residuals are relatively normally distributed around zero because all the probability values are greater than 0.05, which means no one is significant.

Table 7 presents the ACF and PACF of the residuals from the AR(1,0,0) model, along with the Ljung-Box test results. The residuals show no strong autocorrelation, indicating that the model effectively captures the data structure. The Ljung-Box test confirms the independence of residuals, with p-values above 0.05 for most lags, supporting the model's adequacy. Using Eq. 2 to calculate the value of Q by substituting the values of ACF from Table 7 (the ACF of the residuals), the value of $Q = 11.168$ and the critical value of $\chi^2_{24,0.05} = 36.415$ that means H0 is accepted and the residuals is stationary.

Table 6: Output of the plausible model AR(1, 0, 0)

Variable	Coefficient	Std. error	t-statistic	Prob.
C	11.15033	0.385690	28.91013	0.0000
AR(1)	0.663762	0.092659	7.163523	0.0000
SIGMASQ	0.576391	0.088327	6.525633	0.0000
R-squared	0.424584	Mean dependent var		11.19131
Adjusted R-squared	0.401568	S.D. dependent var		1.010425
S.E. of regression	0.781649	Akaike info criterion		2.411076
Sum squared resid	30.54874	Schwarz criterion		2.522602
Log likelihood	-60.89351	Hannan-Quinn criterion		2.453963
F-statistic	18.44684	Durbin-Watson stat		2.103321
Prob(F-statistic)	0.000001			
Inverted AR roots	.66			

Table 7: The ACF and the PACF of residuals of the model AR(1)

	AC	PAC	Q-stat	Prob.
1	-0.06	-0.06	0.2223	
2	0.022	0.018	0.2491	0.618
3	0.129	0.132	1.2263	0.542
4	-0.05	-0.04	1.4090	0.703
5	-0.02	-0.03	1.4487	0.836
6	0.054	0.037	1.6299	0.898
7	0.096	0.119	2.2161	0.899
8	-0.06	-0.05	2.5158	0.926
9	0.115	0.089	3.3973	0.907
10	-0.02	-0.03	3.4297	0.945
11	0.268	0.304	8.4015	0.590
12	-0.07	-0.08	8.7613	0.644
13	0.018	0.014	8.7860	0.721
14	-0.02	-0.12	8.8503	0.784
15	-0.11	-0.05	9.8613	0.772
16	0.100	0.078	10.648	0.777
17	-0.06	-0.05	11.003	0.809
18	0.033	-0.02	11.094	0.852
19	0.077	0.114	11.601	0.867
20	0.101	0.082	12.510	0.863
21	-0.11	-0.07	13.638	0.848
22	0.111	-0.01	14.786	0.834
23	-0.07	-0.06	15.304	0.849
24	-0.11	-0.04	16.564	0.830

Fig. 5 presents a residual diagnostics correlogram for residuals of Model AR(1). Based on the correlogram, which is flat, we conclude that there is no serial correlation and that all information has been captured. Therefore, the predictions will be based on the Model ARIMA(1, 0, 0) for the log transformation of the peanut data.

Method 2: ARIMA model using transformed data: In the previous analysis, the main objective was to perform the time series analysis in the usual manner to compare the results with the method proposed by the study. The accumulated method is used for the peanut series by using Eq. 8 to calculate the accumulated series. Firstly, the stationarity of the new series was tested.

Using Eq. 2 to calculate the value of Q by substituting the values of ACF from Table 8, the value of $Q = 257.0885$ and the critical value of $\chi^2_{24,0.05} = 36.415$ that means H0 is accepted, and the series is not stationary.

Fig. 6 illustrates the ACF and PACF of the accumulated data. It is clear that the series is not stationary because there are 12 ACF out of range.

To achieve stationary, the method of log transformation and differences is used. After using the log transformation and the first difference and calculating the value of Q using Eq. 2 by substituting the values of ACF from Table 9, the value of $Q = 26.68$ and the critical value of ${\chi^2}_{24,0.05} = 36.415$ that means H0 is accepted and the series is stationary.

Fig. 7 illustrates the ACF and the PACF of the accumulated data after using a log transformation. It reveals that the time series of data becomes stationary because only one ACF is out of range. Through Fig. 7, the best model for the accumulated data can be determined as ARIMA(2,1,2). Some of the models are estimated for comparison. Based on model accuracy, as shown in Table 10, the best-fitting model is ARIMA(2,1,2) because it satisfied the standard requirements, included three significant parameters, achieved the highest adjusted R-squared (0.796), and recorded the lowest values of the Akaike information criterion (-2.095) and Schwarz criterion (-1.982). Table 11 illustrates the parameters and model diagnoses of ARIMA(2, 1, 2). From Table 11, the best model can be written as follows:

$$Z_t = 0.280703 + 0.074490Z_{t-1} + 0.885160Z_{t-2} + 0.797396\,\theta_{t-1} + 0.535891\theta_{t-2}$$

Based on Fig. 8, the residuals are flat, which indicates that all relevant information has been captured. Therefore, the prediction will be based on the Model ARIMA(2, 1, 2).

Table 8: The ACF and the PACF of the accumulated data

	AC	PAC	Q-stat	Prob.
1	0.939	0.939	49.478	0.000
2	0.878	-0.03	93.546	0.000
3	0.815	-0.04	132.27	0.000
4	0.752	-0.03	165.94	0.000
5	0.686	-0.06	194.55	0.000
6	0.620	-0.04	218.41	0.000
7	0.553	-0.04	237.82	0.000
8	0.489	-0.02	253.32	0.000
9	0.427	-0.02	265.39	0.000
10	0.370	0.004	274.69	0.000
11	0.318	-0.00	281.70	0.000
12	0.268	-0.02	286.81	0.000
13	0.225	0.012	290.49	0.000
14	0.181	-0.04	292.95	0.000
15	0.141	-0.01	294.48	0.000
16	0.104	-0.02	295.33	0.000
17	0.069	-0.01	295.71	0.000
18	0.035	-0.02	295.81	0.000
19	0.002	-0.02	295.81	0.000
20	-0.03	-0.03	295.89	0.000
21	-0.06	-0.04	296.26	0.000
22	-0.09	-0.02	297.12	0.000
23	-0.12	-0.03	298.69	0.000
24	-0.15	-0.03	301.22	0.000

Table 9: The ACF and the PACF of the accumulated data after log transformation and the first difference

	AC	PAC	Q-Stat	Prob.
1	0.527	0.527	15.284	0.000
2	0.452	0.241	26.735	0.000
3	0.220	-0.12	29.509	0.000
4	0.300	0.211	34.789	0.000
5	0.236	0.056	38.119	0.000
6	0.283	0.066	42.996	0.000
7	0.230	0.050	46.296	0.000
8	0.224	0.013	49.506	0.000
9	0.168	-0.00	51.343	0.000
10	0.104	-0.08	52.063	0.000
11	0.111	0.047	52.913	0.000
12	0.022	-0.11	52.946	0.000
13	0.042	-0.00	53.073	0.000
14	0.011	0.017	53.082	0.000
15	0.003	-0.07	53.082	0.000
16	-0.02	0.005	53.124	0.000
17	-0.03	-0.02	53.216	0.000
18	-0.04	-0.01	53.408	0.000
19	-0.04	0.002	53.598	0.000
20	-0.04	0.001	53.806	0.000
21	-0.06	-0.01	54.199	0.000
22	-0.06.	-0.02	54.632	0.000
23	-0.07	0.017	55.145	0.000
24	-0.06	-0.01	55.602	0.000

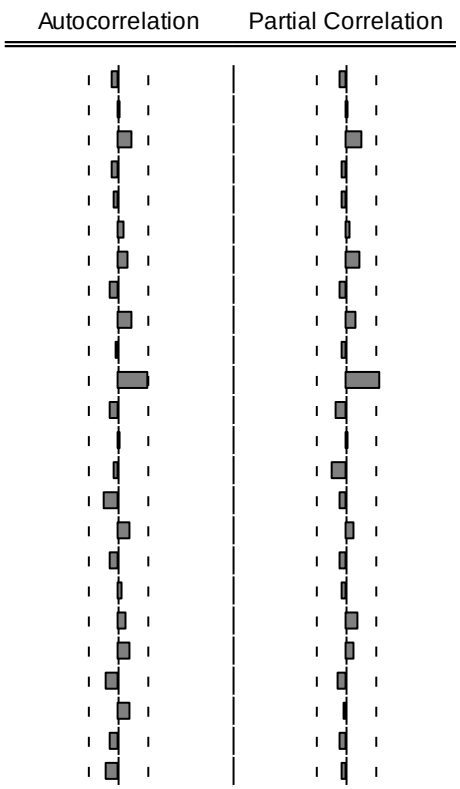

Fig. 5: The ACF and the PACF of residuals of the model AR(1)

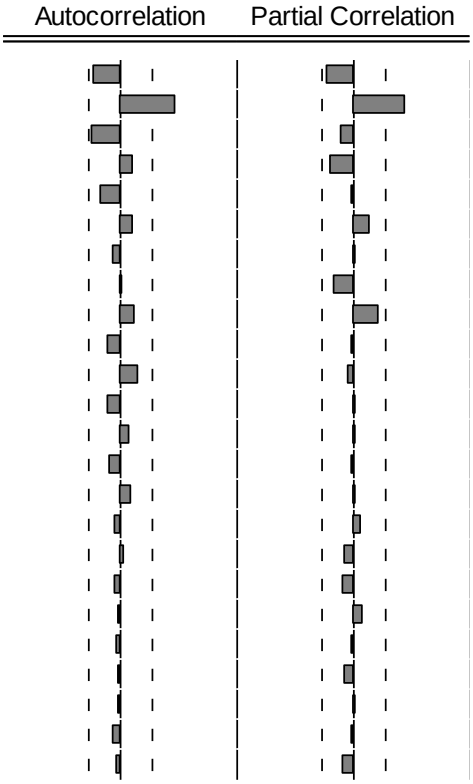

Fig. 7: The ACF and the PACF of the accumulated data after log transformation and the first difference

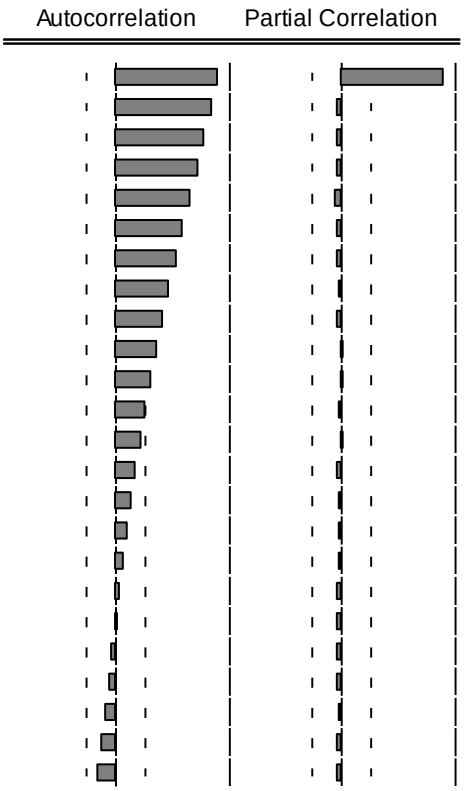

Fig. 6: The ACF and the PACF of the accumulated data

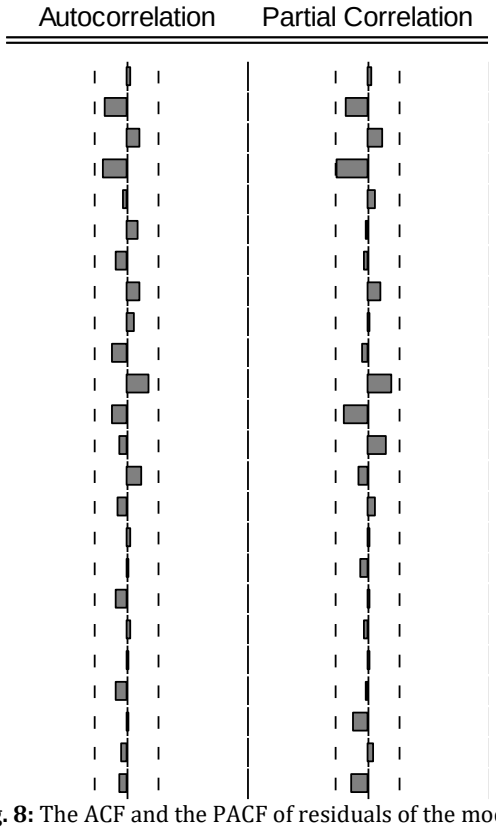

Fig. 8: The ACF and the PACF of residuals of the model AR(2, 1, 2)

Table 10: The best ARIMA models for transformed data

ARIMA model	ARIMA(p, d, q)					
	(1, 1, 0)	(2, 1, 0)	(0, 1, 1)	(0, 1, 2)	(1, 1, 1)	(2, 1, 2)
Significant coefficient	1	2	1	2	1	3
SIGMASQ	0.006	0.006	0.013	0.006	0.006	0.003
Adjusted R-squared	0.634	0.648	0.232	0.625	0.633	0.796
Akaike info criterion	-2.095	-2.115	-1.402	-2.050	-2.076	-2.574
Schwarz criterion	-1.982	-1.964	-1.289	-1.810	-1.925	-2.349

Table 11: The Output of the plausible model AR(2, 1, 2), coefficient covariance computed using outer product of gradients

Variable	Coefficient	Std. error	t-statistic	Prob.
C	0.280703	0.391285	0.717386	0.4768
AR(1)	0.074490	0.051522	1.445781	0.1550
AR(2)	0.885160	0.089126	9.931501	0.0000
MA(1)	0.797396	0.128571	6.202006	0.0000
MA(2)	0.535891	0.169563	3.160428	0.0028
SIGMASQ	0.003182	0.000568	5.606080	0.0000
R-squared	0.816161	Mean dependent var		0.070255
Adjusted R-squared	0.796178	S.D. dependent var		0.132840
S.E. of regression	0.059973	Akaike info criterion		-2.574128
Sum squared resid	0.165449	Schwarz criterion		-2.348985
Log likelihood	72.92733	Hannan-Quinn criterion		-2.487814
F-statistic	40.84371	Durbin-Watson stat		1.737754
Prob(F-statistic)	0.000000			
Inverted AR roots	.98	-.90		
Inverted MA roots	-.40-.61i	-.40+.61i		

Table 12 shows the ACF and PACF of the residuals from the AR(2, 1, 2) model, together with the Ljung–Box test results. The residuals do not show strong autocorrelation, which means the model explains the main structure of the data. The Ljung–Box test also supports this, because most p-values are greater than 0.05, so we fail to reject the null hypothesis of no residual autocorrelation. Small autocorrelations appear at lags 2 and 4, but the Q-statistics remain within acceptable limits. Using Eq. 2 and the residual ACF values from Table 6, the Q value is 11.168. This is less than the critical value $\chi^2_{24,0.05} = 36.415$, so H0 is accepted and the residuals can be treated as uncorrelated (white noise). Overall, the AR(2, 1, 2) model appears well specified, with no major unexplained patterns, although additional validation could still be considered.

Table 13 compares two best-fit models from different approaches: ARIMA(1, 0, 0) fitted to the original data and ARIMA(2, 1, 2) fitted to the cumulative data. According to the reported criteria, the proposed approach (ARIMA(2, 1, 2) on the cumulative data) provides better performance than ARIMA(1, 0, 0) on the original data, as indicated by improved goodness-of-fit measures (e.g., lower error metrics and information criteria). This comparison shows that modeling the cumulative series yields a more accurate and reliable fit.

Table 12: The ACF and the PACF of residuals for the AR(2, 1, 2) model

	AC	PAC	Q-stat	Prob.
1	0.044	0.044	0.1083	
2	-0.182	-0.184	1.9618	
3	0.109	0.131	2.6421	
4	-0.196	-0.258	4.8795	
5	-0.024	0.071	4.9137	0.027
6	0.103	-0.014	5.5653	0.062
7	-0.083	-0.029	5.9922	0.112
8	0.114	0.109	6.8236	0.146
9	0.068	0.006	7.1237	0.212
10	-0.120	-0.040	8.0897	0.232
11	0.204	0.208	8.4015	0.141
12	-0.115	-0.201	10.393	0.158
13	-0.050	0.164	11.860	0.211
14	0.130	-0.079	12.037	0.208
15	-0.075	0.078	13.291	0.249
16	0.041	0.003	13.717	0.310
17	0.018	-0.058	13.877	0.383
18	-0.092	-0.022	14.575	0.408
19	0.048	-0.024	14.767	0.468
20	0.029	0.011	14.843	0.536
21	-0.087	-0.012	15.532	0.557
22	0.019	-0.114	15.567	0.623
23	-0.047	-0.057	15.781	0.672
24	-0.058	-0.132	16.115	0.709

Table 13: The compression of the best models using the two methods

Metrics	The original data ARIMA(1, 0, 0)	The cumulative data ARIMA(2, 1, 2)
Significant coefficient	1	3
SIGMASQ	0.576	0.003
Adjusted R-squared	0.402	0.796
Akaike info criterion	2.411	-2.57
Schwarz criterion	2.523	-2.349

The comparison between ARIMA(1, 0, 0) and ARIMA(2, 1, 2) models highlights significant improvements in model performance. The ARIMA(2, 1, 2) model, with three significant coefficients

compared to just one in ARIMA(1, 0, 0), demonstrates greater flexibility and a better fit to the data. Additionally, the variance of residuals (SIGMASQ) is drastically reduced from 0.576 to 0.003, indicating that ARIMA(2, 1, 2) more effectively captures the underlying data structure and minimizes prediction errors. The adjusted R-squared value, a key measure of model fit, increases from 0.402 to 0.796, reinforcing the stronger explanatory power of ARIMA(2, 1, 2). Furthermore, the Akaike Information Criterion (AIC) and Schwarz Criterion (SC) show substantial reductions, decreasing from 2.411 to -2.57 and 2.523 to -2.349, respectively. These lower values suggest that ARIMA(2, 1, 2) provides a superior balance between model complexity and goodness of fit. Overall, these improvements confirm that ARIMA(2, 1, 2) is a more effective model for forecasting, offering greater accuracy and efficiency.

The results indicate that the ARIMA(2, 1, 2) model applied to cumulative data outperforms the ARIMA(1, 0, 0) model based on multiple evaluation metrics. The increased adjusted R-squared value demonstrates improved explanatory power, while the reduced AIC and SC confirm a more efficient model fit. Additionally, the drastic reduction in SIGMASQ suggests that the new approach effectively minimizes prediction errors. These findings support the conclusion that the proposed method significantly enhances predictive accuracy, making it a more suitable choice for forecasting in this context.

5.3. Discussion

Accurate forecasting of the productivity of any agricultural crop is critical for economic growth, food security, and poverty reduction to avoid the risks associated with poor diets leading to disease and health crises. It is important to develop methods that assist decision-makers in understanding underlying future phenomena. There are many types of time series forecasting methods, for example, the classical method, the Box-Jenkins method, and artificial neural networks. Accuracy needs to be the key factor considered whenever one is deciding between various methods of forecasting. Over the past few decades, several techniques for increasing the accuracy of forecasting models have been created (Cerqueira et al., 2020; Bergmeir et al., 2018). However, there isn't a method that everyone agrees on. In this paper, we contribute to the existing body of research by carrying out an empirical study that compares two time series methods to determine which one will increase the model's accuracy. In the first method, we used original time series data and performed our analysis to find the best-fit mode. This method showed that the data were nonstationary since autocorrelation values were out of the stationary range (Table 2). In the second method, we used accumulated data from the same time series and conducted the analysis using a logarithmic transformation. According to the illustration in Table 8, the time series data is stationary and contributes significantly to the autocorrelations being confined within the range.

Model comparisons (Table 12) in AIC and parameter estimates showed that Model ARIMA(2, 1, 2) of the second method (the cumulative method) contributed significantly. According to the metric comparison, the proposed model (ARIMA(2, 1, 2)) performed significantly and managed to produce better forecasts than the first methods with an adjusted R-squared of 0.796 versus 0.40, an AIC of -2.57 versus 2.41, and a SIC of -2.35 versus 2.52. Based on this result, Model ARIMA(2, 1, 2) of the second method (i.e., the cumulative data method) is superior. Therefore, one can say that the proposed technique leads to more clarity in the identification of the model, increases the value of the coefficients of determination, decreases the value of BIC, and produces more accurate forecasting.

Through Table 12, it is clear that the proposed method showed an improvement in predictive accuracy by 39%. comparing our study with the study of Wang et al. (2015), which used the method of ensemble empirical mode decomposition and showed a development in predictive accuracy by 12%, our study has proved better results.

6. Conclusion

In this study, the time series method was used to analyze peanut yields in Sudan using conventional time series data. ARIMA model with the original data of crops over successive years, as well as the proposed ARIMA model with the original cumulative data after transformation. The results of the two methods were compared. The precision of the proposed method, which makes use of the transformation data, led to its selection as the best option for forecasting. The method uses the logarithm of cumulative values of peanut yield, resulting in significantly better results than the traditional method. This is especially important for studying how crops have been affected by production values from previous years. Therefore, the proposed methods should be used when studying the crop time series in order to obtain more accurate results, leading to more accurate planning to achieve more productivity and a greater abundance of these crops. This study added a new method to increase the predictive accuracy of these models.

This paper makes a valuable contribution on both the scientific and practical sides. The study demonstrates that the new method, the cumulative method, enhances the predictive power of the Box-Jenkins models from a scientific perspective. On the practical side, we find that this paper made a great contribution to Sudan's government agencies and decision-makers. The model's significant predictive power enables the creation of highly accurate future forecasts for the peanut crop, a crucial national product that boosts the state's income by exporting its raw form or processing it into oil or peanut butter.

List of abbreviations

AC	Autocorrelation
ACF	Autocorrelation function
AIC	Akaike's information criterion
ANN	Artificial neural networks
AR	Autoregressive process
ARIMA	Autoregressive integrated moving average
ARMA	Autoregressive moving average
BIC	Bayesian information criterion
GNP	Gross national product
MA	Moving average process
PAC	Partial autocorrelation
PACF	Partial autocorrelation function
Prob.	Probability
Q-Stat	Q-Statistics
S.E.	Standard error
SC	Schwarz's criterion
SIC	Schwarz's information criterion
SIGMASQ	The estimator of the error variance of the residual
Std.	Standard
Var	Variable

Acknowledgment

This work was funded by the Deanship of Scientific Research (DSR) at King Abdulaziz University, Jeddah, under grant no. G: 331-662-1436. The authors, therefore, acknowledge with thanks to DSR technical and financial support.

Compliance with ethical standards

Conflict of interest

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

References

Allende H, Moraga C, and Salas R (2002). Artificial neural networks in time series forecasting: A comparative analysis. Kybernetika, 38(6): 685-707.

Bergmeir C, Hyndman RJ, and Koo B (2018). A note on the validity of cross-validation for evaluating autoregressive time series prediction. Computational Statistics & Data Analysis, 120: 70-83. https://doi.org/10.1016/j.csda.2017.11.003

Box GE and Pierce DA (1970). Distribution of residual autocorrelations in autoregressive-integrated moving average time series models. Journal of the American Statistical Association, 65(332): 1509-1526. https://doi.org/10.1080/01621459.1970.10481180

Box GE, Jenkins GM, Reinsel GC, and Ljung GM (2015). Time series analysis: Forecasting and control. 5th Edition, John Wiley & Sons, Hoboken, USA.

Cerqueira V, Torgo L, and Mozetič I (2020). Evaluating time series forecasting models: An empirical study on performance estimation methods. Machine Learning, 109: 1997-2028. https://doi.org/10.1007/s10994-020-05910-7

Chatfield C and Xing H (2019). The analysis of time series: An introduction with R. Chapman and Hall/CRC, London, UK. https://doi.org/10.1201/9781351259446

Cryer JD and Chan KS (2008). Time series regression models. In: Cryer JD and Chan KS (Eds.), Time series analysis: With applications in R: 249-276. Springer, New York, USA. https://doi.org/10.1007/978-0-387-75959-3_11

Elshafie SZ, ElMubarak A, El-Nagerabi SA, and Elshafie AE (2011). Aflatoxin B_1 contamination of traditionally processed peanuts butter for human consumption in Sudan. Mycopathologia, 171: 435-439. https://doi.org/10.1007/s11046-010-9378-2 **PMid:21104323**

Fattah J, Ezzine L, Aman Z, El Moussami H, and Lachhab A (2018). Forecasting of demand using ARIMA model. International Journal of Engineering Business Management, 10: 1-9. https://doi.org/10.1177/1847979018808673

Granger CWJ and Newbold P (2014). Forecasting economic time series. Academic Press, Cambridge, USA.

Huang NE and Wu Z (2008). A review on Hilbert-Huang transform: Method and its applications to geophysical studies. Reviews of Geophysics, 46: RG2006. https://doi.org/10.1029/2007RG000228

Jonathan DC and Kung-Sik C (2008). Time series analysis with applications in R. Springer, Berlin, Germany.

Jose VRR and Winkler RL (2008). Simple robust averages of forecasts: Some empirical results. International Journal of Forecasting, 24(1): 163-169. https://doi.org/10.1016/j.ijforecast.2007.06.001

Kirchgässner G, Wolters J, and Hassler U (2012). Introduction to modern time series analysis. Springer Science and Business Media, Berlin, Germany. https://doi.org/10.1007/978-3-642-33436-8 **PMid:24144469**

Kumar M and Anand M (2014). An application of time series ARIMA forecasting model for predicting sugarcane production in India. Studies in Business and Economics, 9(1): 81-94.

Ljung GM and Box GE (1978). On a measure of lack of fit in time series models. Biometrika, 65(2): 297-303. https://doi.org/10.1093/biomet/65.2.297

Mateos H, Gentile L, Murgia S, Colafemmina G, Collu M, Smets J, and Palazzo G (2022). Understanding the self-assembly of the polymeric drug solubilizer Soluplus®. Journal of Colloid and Interface Science, 611: 224-234. https://doi.org/10.1016/j.jcis.2021.12.016 **PMid:34952275**

Mizon GE (1995). A simple message for autocorrelation correctors: Don't. Journal of Econometrics, 69(1): 267-288. https://doi.org/10.1016/0304-4076(94)01671-L

Pankratz A (2009). Forecasting with univariate Box-Jenkins models: Concepts and cases. John Wiley & Sons, Hoboken, USA.

Shumway RH and Stoffer DS (2017). Time series analysis and its applications: With R examples. 4th Edition, Springer, Cham, Switzerland. https://doi.org/10.1007/978-3-319-52452-8

Stock JH and Watson MW (2004). Combination forecasts of output growth in a seven-country data set. Journal of Forecasting, 23(6): 405-430. https://doi.org/10.1002/for.928

Štulajter F (2002). Predictions in time series using regression models. Springer, New York, USA. https://doi.org/10.1007/978-1-4757-3629-8

Tsay RS (2005). Analysis of financial time series. John Wiley & Sons, Hoboken, USA. https://doi.org/10.1002/0471746193

Wang L, Wang Z, Qu H, and Liu S (2018). Optimal forecast combination based on neural networks for time series forecasting. Applied Soft Computing, 66: 1-17. https://doi.org/10.1016/j.asoc.2018.02.004

Wang WC, Chau KW, Xu DM, and Chen XY (2015). Improving forecasting accuracy of annual runoff time series using ARIMA based on EEMD decomposition. Water Resources Management, 29: 2655-2675. https://doi.org/10.1007/s11269-015-0962-6

Wang X, Hyndman RJ, Li F, and Kang Y (2023). Forecast combinations: An over 50-year review. International Journal of Forecasting, 39(4): 1518-1547. https://doi.org/10.1016/j.ijforecast.2022.11.005

Wollstadt P, Martínez-Zarzuela M, Vicente R, Díaz-Pernas FJ, and Wibral M (2014). Efficient transfer entropy analysis of non-stationary neural time series. PLOS ONE, 9(7): e102833. https://doi.org/10.1371/journal.pone.0102833 **PMid:25068489 PMCid:PMC4113280**

Wu JD and Tsai YJ (2011). Speaker identification system using empirical mode decomposition and an artificial neural network. Expert Systems with Applications, 38(5): 6112-6117. https://doi.org/10.1016/j.eswa.2010.11.013

Yuan C, Liu S, and Fang Z (2016). Comparison of China's primary energy consumption forecasting by using ARIMA (the autoregressive integrated moving average) model and GM(1,1) model. Energy, 100: 384-390. https://doi.org/10.1016/j.energy.2016.02.001

Zhuo L, Mekonnen MM, and Hoekstra AY (2016). The effect of inter-annual variability of consumption, production, trade and climate on crop-related green and blue water footprints and inter-regional virtual water trade: A study for China (1978–2008). Water Research, 94: 73-85. https://doi.org/10.1016/j.watres.2016.02.037 **PMid:26938494**