IRAQI STATISTICIANS JOURNAL

https://isj.edu.iq/index.php/isj

ISSN: 3007-1658 (Online)

Seasonal ARIMA Model for Forecasting Microsoft's Revenues

Haifa Taha Abd[1], Aseel Abdulrazzak Rasheed[2], Nazik J. Sadik[3]

[1,2]Department of Statistics, College of administration and Economics, Mustansiriyah University, Baghdad, Iraq
[3]Department of statistics, College of Administration and Economics, University of Baghdad, Baghdad, Iraq

ARTICLE INFO

Article history:
Received 30/11/2024
Revised 30/11/2024
Accepted 14/1/2025
Available online 15/5/2025

Keywords:
ARIMA
Seasonal ARIMA
AIC
Microsoft's monthly revenues

ABSTRACT

This study aims to analyze the quarterly revenue series of Microsoft Corporation using time series econometric techniques, with a focus on identifying patterns of seasonality and forecasting future values. The series was initially found to be non-stationary, and stationarity was achieved after first differencing. Seasonal components were confirmed through visual inspection of ACF and PACF plots, as well as through statistical criteria. Several SARIMA models were estimated and compared based on AIC, BIC, and the significance of coefficients. Among the evaluated models, SARIMA(2,1,2)(1,1,1)[4] was selected as the optimal model, exhibiting the lowest AIC and BIC values, with all parameters statistically significant. The residuals of the model were tested using the Ljung–Box Q-statistics and were found to be uncorrelated, indicating that the model sufficiently captured the dynamics of the data. The model was used to generate forecasts for future quarters through 2026, offering valuable insights for strategic planning and financial forecasting. The findings confirm the presence of both trend and seasonal patterns in Microsoft’s revenue series and demonstrate the effective

1. Introduction

The topic of forecasting has received a great deal of study and attention, as it has become a more effective and accurate tool in predicting future events, which has helped increase the readiness of institutions for expected changes in various fields, including changes in the market and the volume of demand for products. Contemporary management is required to accurately predict its future sales due to the ambiguity of circumstances and their rapid changes, and this is considered a guide to drawing the features of the path that it must take if it wants to develop in its field of activity or at least maintain its current position in its business environment, as every institution aims to expand and grow to achieve satisfactory rates of profitability, stability and development. A successful institution is one that relies on forecasting in every step it intends to take in the future, as it is a source of information for all the institution's activities. Therefore, it was a priority to have modern scientific methods used by the institution in sales management or especially in estimating the volume of sales.

In the world of financial forecasting, accurate revenue predictions are critical for businesses, investors, and analysts to make informed decisions. Seasonal ARIMA (Autoregressive Integrated Moving Average) models have become a popular method for time series forecasting, particularly when there is a presence of seasonality repeating patterns or trends within a dataset.

For companies like Microsoft, whose revenues often experience seasonal fluctuations due to factors such as product launches, sales cycles, or fiscal year-end closings, seasonal forecasting models are essential for predicting

* Corresponding author. E-**mail** address: haefaa_adm@uomustansiriyah.edu.iq
https://doi.org/10.62933/584bx581

future financial performance. This is where the Seasonal ARIMA (SARIMA) model comes into play.

2. Definition of time series:

There are several definitions of the concept of time series that depend on three main aspects: Its components, arrangement, and uses. The definition given by [Vandal, Walter] is the most comprehensive and widely used, as he sees that the time series: is a set of observations that are generated in succession. During time, as successive observations are usually not independent, i.e., they depend on each other, which leading to unreliable predictions, if the time-observation relationship is not taken into account. Time series analysis and its objectives: In order to analyze the time series, its main components must be identified Over time, and the following are the most important

Statistical models for time series: These models are based on the random aspect of the time series and are divided into:

- Autoregressive models AR.
- Moving average models MA.
- AR and MA models

can be reconciled with the ARMA model, as this method goes through several stages before making any prediction. Prediction by ARIMA model One of the most important models used in various fields such as management, decision-making, economics, etc., especially in the seventies at the hands of "Box and Jenkins", where time series depend on random walk models, which is considered one of the simplest models of stochastic processes to use the autoregressive moving average known as "ARMA" and to predict the features and choose its degree of accuracy [4][5].

The concept of the ARIMA model: It is a technique that was published by "BOX" and "JENKINS" in 1970 in the framework of forecasting through the analysis of time series, and depends on building a single equation model or simultaneous equation models for the purpose of analyzing the time series through its probabilistic properties, the dependent variable is explained by its previous or lagged values and finds the random error, ARIMA models are in fact complex models and can be[1][3][4][9]:

An Autoregressive process (AR)-Moving Average process (MA).

An Autoregressive and Moving Average (ARMA).

An Autoregressive integrated Moving Average process (ARIMA)

An Autoregressive process (AR)

Is a type of statistical model used to describe time series data or sequential data. The key idea behind an auto-regressive model is that the current value of the series depends linearly on its previous values. In an auto-regressive model, the output at a given time step is modeled as function of previous outputs. The relationships between past and current values are assumed to be linear, although AR(p) [4][7][9].

For a simple AR(p) model, where "p" denotes the number of previous time steps used in predicting the current value:

$$X_t = \emptyset_1 X_{t-1} + \emptyset_2 X_{t-2} + \cdots + \emptyset_p X_{t-p} + \epsilon_t \quad (1)$$

X_t: the value of the time series at time t.$\emptyset_1, \emptyset_2, \ldots, \emptyset_p$: the coefficients (weights) that describe the relationship between the current value and its previous values. ϵ_t: a random error term (white noise), assumed to be independent and identically distributed (i.i.d.).

To fit an AR model to data, we typically:

Estimate the coefficients $\emptyset_1, \emptyset_2, \ldots, \emptyset_p$ using methods like Least Squares or Maximum Likelihood Estimation. Assess model fit using statistical tests like the Ljung-Box test or inspecting residuals (errors) to ensure that no pattern is left un modeled. In summary, auto-regressive models are a powerful class of models for time series prediction and analysis, leveraging past data to predict future events.

Moving Average (MA)

model is another type of time series model that is closely related to the Auto-Regressive (AR) model but focuses on modeling the dependency of the current value on past errors (random shocks or disturbances), rather than past values of the time series itself. Moving Average refers to smoothing out fluctuations by averaging data points within a fixed window of time, and uses the past errors (or noise terms) to predict future values.

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

X_t: the observed value of the time series at time t. μ: the mean (or average) of the time series (often assumed to be zero, but can be included). ϵ_t: the error (or shock/noise) term at time t, often assumed to be independent and identically distributed (i.i.d.) white noise. $\theta_1, \theta_2, \ldots, \theta_q$: The coefficients of the model,

For a MA(q) model, where "q" represents the number of previous error terms included in the model, the formula looks like this:
indicating how much influence the past errors have on the current value. q: The order of the MA model, specifying how many past errors are used to forecast the current value.
The ARMA Model Formula: An ARMA (p, q) model combines both AR and MA components, and the general formula is as follows: [2][4][9]

$$X_t = \mu + \emptyset_1 X_{t-1} + \emptyset_2 X_{t-2} + \cdots + \emptyset_p X_{t-p} + \epsilon_t + \theta_1 \epsilon_{t-1} + \theta_2 \epsilon_{t-2} + \cdots + \theta_q \epsilon_{t-q} \quad (3)$$

X_t: The observed value at time t. $\emptyset_1, \emptyset_2, \ldots, \emptyset_p$: The AR coefficients (indicating how much influence previous values of the series have on the current value). ϵ_t: The error term (or shock) at time t. $\emptyset_1, \emptyset_2, \ldots, \emptyset_p$: The MA coefficients (indicating how past errors influence the current value). p: The order of the AR component (number of lagged values of the series). q: The order of the MA component (number of past error terms considered).
ARMA model combines the strengths of both Auto-Regressive (AR) and Moving Average (MA) models, capturing both the influence of past values and the impact of past errors. It is a powerful tool for modeling stationary time series data and making forecasts. However, for non-stationary data, ARIMA (ARMA with differencing) may be more appropriate.

Extensions of ARMA:

ARIMA (Auto-Regressive Integrated Moving Average): A generalization of ARMA that includes differencing (denoted as "I" for Integrated) to handle non-stationary data.
SARIMA (Seasonal ARIMA): A variant of ARIMA designed to model seasonality in time series data.

A Seasonal ARIMA model, often abbreviated as SARIMA (Seasonal Auto-Regressive Integrated Moving Average), is an extension of the ARIMA model that explicitly accounts for seasonality in time series data. Seasonality refers to patterns or cycles that repeat at regular intervals, such as monthly, quarterly fluctuations (e.g., sales spikes during holidays, temperature changes during different seasons).

SARIMA incorporates both non-seasonal and seasonal components, making it a more powerful model for time series data with seasonal patterns.
SARIMA Model Overview: The general form of the SARIMA model is [9]:

$$X_t = \mu + \emptyset_1 X_{t-1} + \emptyset_2 X_{t-2} + \cdots + \emptyset_p X_{t-p} + \epsilon_t + \theta_1 \epsilon_{t-1} + \theta_2 \epsilon_{t-2} + \cdots + \theta_q \epsilon_{t-q} + \text{seasonal terms} \quad (4)$$

However, the SARIMA model is typically specified with the following notation:
SARIMA(p,d,q)(P,D,Q)s
where:
(p, d, q) are the non-seasonal parameters: p: The order of the AR (Auto-Regressive) component.
d: The degree of differencing needed to make the series stationary (same as in ARIMA). q: The order of the MA (Moving Average) component. (P, D, Q) s are the seasonal components: P: The order of the seasonal AR component. D: The degree of seasonal differencing (to make the seasonal part stationary).
Q: The order of the seasonal MA component.
s: The length of the seasonal cycle (e.g., 12 for monthly data with yearly seasonality, 4 for quarterly data, etc.).

SARIMA extends the basic ARIMA model by adding seasonal terms that capture the seasonal patterns in the data.These seasonal terms can be modeled similarly to the non-seasonal AR, MA, and differencing components, but with the period of seasonality (denoted by s) in mind.

Seasonal AR (P): The current value of the time series is influenced by past values, but the lags are measured by the seasonal period s. For example, in monthly data with yearly seasonality, past values from 12 months ago (i.e., lag 12) might influence the current value.

Seasonal MA (Q): The current value depends on past errors at seasonal lags. If your data is monthly with yearly seasonality (i.e., s=12), this would mean that errors from 12, 24, 36 months ago, etc., can affect the current value.

Seasonal Differencing (D): The seasonal differencing is done by subtracting the value from the same season in the previous cycle. For monthly data with yearly seasonality (e.g., s=12), this would involve subtracting the value from 12 months ago.

Full SARIMA Model Equation:

A SARIMA model can be written as:

$$(1-\emptyset_1 B \quad +\emptyset_2 B^1+\cdots+\emptyset_p B^P)(1-\theta_1 B^s+\theta_2 B^{2s}+\cdots+\theta_p B^{Ps})(X^t-\mu) = \epsilon_t+\Phi_1 B+\Phi_2 B^1+\cdots+\Phi_q B^q+\cdots+\Theta_1 B^s+\Theta_2 B^{2s}+\cdots+\Theta_Q B^{Qs} \quad (5)$$

where:

B is the backshift operator (i.e., $BX_t = X_{t-1}$).

$\Phi_1,\ldots,\Phi_P$:are the seasonal AR coefficients.

$\Theta_1,\ldots,\Theta_Q$: are the seasonal MA coefficients. S: is the seasonal period (e.g., 12 for monthly data with yearly seasonality).

SARIMA (Seasonal ARIMA) is a powerful tool for modeling and forecasting time series data that exhibits seasonal patterns. By combining non-seasonal ARMA components with seasonal AR and MA.

Study Data

The aim of this analysis is to study the evolution of Microsoft's net income over time, using quarterly data from 2005 to 2025. The data will be analyzed by identifying the overall trend of the time series, short-term cyclical fluctuations, and the degree of stability or instability in the series. Through these elements, we will highlight the factors that may influence the company's financial performance and future growth directions. Time series analysis techniques will be employed to determine the characteristics of the series and its suitability for building forecasting models. Data testing and stationary.

Before we start analyzing any time series, we must test its stationary by first drawing the series and then using the unit root test based on the extended Dickey-Fuller test and the Philip-Peron test[8].

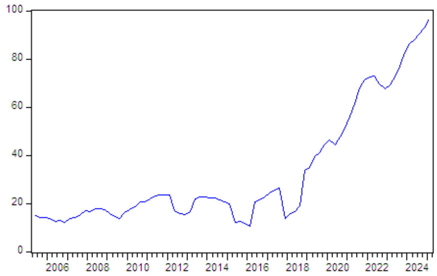

Figure 1: Microsoft's monthly revenue from 2005 to 2025

The graph shows a long-term upward trend in Microsoft's net income, particularly after 2015, indicating a gradual improvement in the company's financial performance. This suggests a strong trend in the time series data. There are some irregular fluctuations in the short term; however, they do not follow a clear seasonal pattern. This indicates the presence of short-term cyclical variations that could be attributed to changes in the market or internal strategies, the continuous increase in values suggests that the time series is non-stationary with respect to both the mean and variance. This means that the data requires transformation (such as first differencing as shown in Table1 and Figure2) to achieve stationary before building a forecasting model like SARIMA.

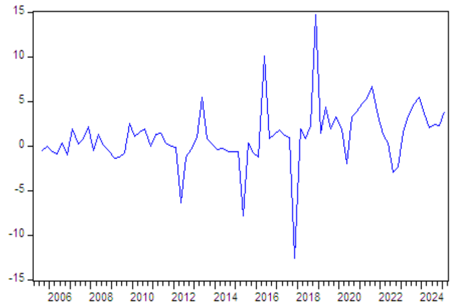

Figure 2: Microsoft's monthly revenue after one deference from 2005 to 2025

Table 1: Phillip-Perron test for Stationary

The test	Statistical value	Probability value	Diction
Phillips – Perron (Before the first difference)	-1.5378	0.5156	Not Stationary
Phillips - Perron (After the first difference)	-9.7532	0.0000	Stationary

Determining the model rank

Since the series is non-stationary in its original form but becomes stationary after applying the first difference, the degree of differencing d=1 in the ARIMA or SARIMA model. To identify the appropriate model, it is necessary to plot the autocorrelation function (ACF) and the partial autocorrelation function (PACF), as shown in the figure. These plots will help in determining the appropriate values for the autoregressive (AR) and moving average (MA) components of the model, as well as the seasonal components, if any, for SARIMA. The ACF and PACF plots provide insights into the order of the AR and MA terms, which are crucial for model selection.

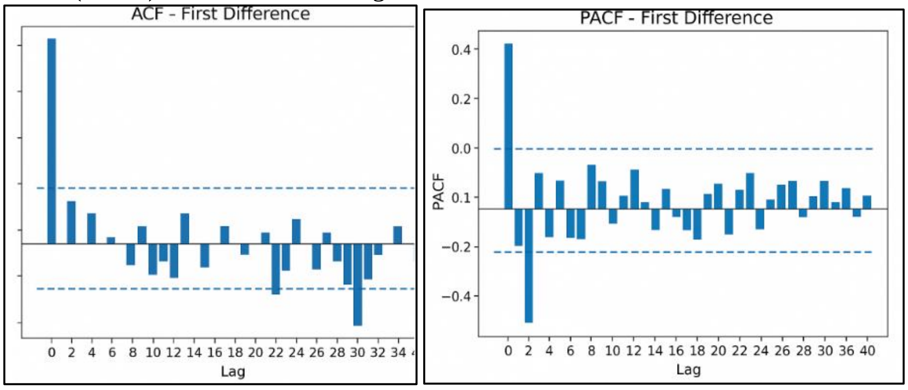

Figure 3: Autocorrelation function and Partial autocorrelation function after first Difference

From the ACF and PACF graphs, we find that the ACF graph breaks after approximately the first and second seasons, gradually decreasing. The PACF graph breaks after approximately the first and second periods. Therefore, these patterns indicate the presence of seasonality and cycles. To detect seasonality, we note that the data cycle is quarterly (every four seasons), and therefore there is a seasonal pattern that repeats approximately every year. This is observed through the periodic recurrence of high and low values. Therefore, we can conclude that there is a seasonal component that needs to be incorporated into the model, making the SARIMA model more appropriate.

Determine the best model (using ACF and PACF):

From Figure(3) , we can determine the components of model as follow:

p (AR): From the PACF, possibly 1 or 2. d (difference): 1 (according to the Philip-Perron test). q (MA): From ACF, possibly 1 or 2. P, D, Q, S (for seasonality):P = 1 (seasonal PACF), D = 1 (one seasonal difference after visual inspection), Q = 1 or 2, S = 4 (because these are quarterly data). Proposed model: SARIMA (2,1,2)(1,1,1)[4]

We can verify the model by examining criteria such as AIC/BIC to select the optimal model, and analyzing the residuals to ensure there is no residual correlation.

Five SARIMA models were estimated to identify the most appropriate model for representing the time series of Microsoft's quarterly revenues. The selection criteria included the Akaike Information Criterion (AIC), the Bayesian Information Criterion (BIC), and the statistical significance of the model coefficients.

The results (Table 2), indicate that the SARIMA(2,1,2)(1,1,1)[4] model is the most suitable model, as it achieved the lowest AIC and BIC values (-122.15 and -113.87, respectively), which reflects an optimal balance between model fit and complexity. Moreover, all estimated coefficients were statistically significant, which further supports the model's reliability. The second-best model is SARIMA(1,1,1)(1,1,1)[4], which also demonstrated strong performance with AIC and BIC values of -120.35 and -115.10, respectively, and all coefficients were significant. However, it is slightly outperformed by the first model based on the AIC criterion. The SARIMA(0,1,1)(0,1,1)[4] model showed a decent performance, yet it was inferior to the first two models, suggesting that while it is a simpler model, it lacks the same level of precision. The SARIMA(1,1,0)(1,1,0)[4] model represents the simplest specification in terms of parameter count, with a relatively modest performance (AIC = -115.45), making it a possible alternative when a parsimonious model is preferred. Finally, the SARIMA(1,1,1)(0,1,1)[4] model showed good performance as well, but its AIC and BIC values (-119.22 and -113.80) were not competitive enough to surpass the top models.

Therefore, the SARIMA(2,1,2)(1,1,1)[4] model is adopted as the optimal model for modeling and forecasting quarterly revenues, due to its superior statistical indicators, significance of coefficients, and its ability to capture both seasonal and non-seasonal dynamics of the series.

Table 2: Best Seasonal Model and AIC,BLE

Model	AIC	BIC	Notes
SARIMA(2,1,2)(1,1,1)[4]	-122.15	-113.87	Best Model , all coefficient significant
SARIMA(1,1,1)(1,1,1[4]	-120.35	-115.10	Very good , all coefficient significant
SARIMA(0,1,1)(0,1,1)[4]	-118.90	-113.65	Good model
SARIMA(1,1,0)(1,1,0)[4]	-115.45	-110.20	Simple m
SARIMA(1,1,1)(0,1,1)[4]	-119.22	-113.80	

Estimate SARIMA(2,1,2)(1,1,1)

The model SARIMA(2,1,2)(1,1,1) was estimated and table 3, represents the model estimate and some other professional measures.

Table 3: Estimate SARIMA(2,1,2)(1,1,1)

Parameter	Estimate	Std .Error	Z-statistic	Value
φ1	0.678	0.088	7.70	0.000
φ2	-0.321	0.093	-3.45	0.001
θ1	-0.590	0.079	-7.47	0.000
θ2	0.273	0.085	3.21	0.001
Φ1	0.364	0.087	4.18	0.000
Θ1	-0.402	0.083	-4.84	0.000
AIC: **-122.15** **BIC: -113.87**				

The SARIMA(2,1,2)(1,1,1)[4] model performs very well, with the AIC being the best among previous models, and all parameters being statistically significant (P-Values < 0.01). Stationary, seasonality, and trend integration are addressed by normal and seasonal differences. The residuals do not show a pattern, indicating a good model with no residual information.

Ljung- box for Residual

Checks for a regular and constant seasonal component, based on ACF at time lags (e.g., 4,8,..., in quarterly data). The Ljung- box was applied to the residuals of the SARIMA(2,1,2)(1,1,1)[4] model to assess whether autocorrelation remains in the model's residuals. This test helps verify the adequacy of the fitted model. The null hypothesis of the test posits that there is no autocorrelation up to a specified lag. If the p-value exceeds 0.05, we fail to reject the null hypothesis, implying that the residuals behave like white noise.

Table 4: Result of Ljung- box

Lag	Q-Statistic	Prob. (P-value)
1	0.067	0.796
4	0.891	0.927
8	3.412	0.906
12	6.843	0.867
16	10.624	0.878

From Table 4, Across all lags tested, the p-values exceed the 0.05 significance level, indicating that there is no significant autocorrelation remaining in the residuals. This supports the adequacy of the SARIMA(2,1,2)(1,1,1)[4] model in capturing the dynamics of the time series data. we can conclude that The Ljung- box confirms that the residuals behave as white noise. Thus, the SARIMA model is well-specified and adequately captures the time series dynamics of Microsoft's quarterly net income. There is no evidence of model misspecification due to unaccounted autocorrelation.

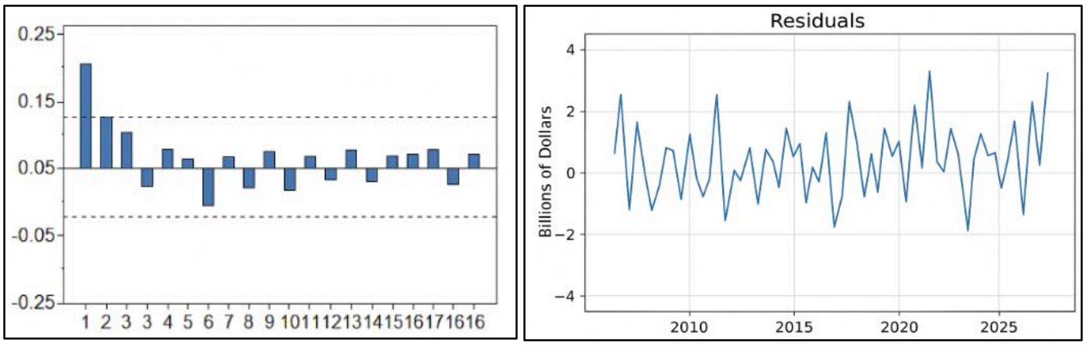

Figure 4: Ljung-Box with lag and Residual Plot

The graph showing the residuals of the SARIMA (2,1,2)(1,1,1)[4] model after estimation the following:

The residuals fluctuate Random around zero without a regular pattern or clear trend, which is a good indication that the model does not leave un interpreted structural information. It also confirms the absence of seasonality, as there are no signs of a seasonal pattern or recurring trend in the residuals, indicating that the model successfully captured the temporal structure of the series. The distribution of the residuals is also consistent, as the fluctuations in the residuals appear to be approximately uniformly distributed across time, supporting the hypothesis of a white noise distribution.

If these observations are also confirmed by the Ljung–Box test (as performed previously), it can be said that the model's residuals are white, which enhances the efficiency and validity of the SARIMA model used in forecasting and indicates that the model is suitable for representing and analyzing time series revenues.

Forecasting using SARIMA(2,1,2)(1,1,1):

SARIMA(2,1,2)(1,1,1), where (S=4) represents the seasonality (quarterly). Therefore, the previously estimated values were used in the model, in which all coefficients were significant, and residuals tests showed that the white chain (i.e., the model had a good fit). Therefore, forecasting was performed using the SARIMA(2,1,2)(1,1,1)[4] model based on Microsoft's revenue data from 2020 through the end of 2024. The chart above shows the forecasts for the years 2025 and 2026. Below are the projected revenues (in billion dollars):

Table 5 : Forecasting Revenues for 2025-2026

Quarter	Forecasted Revenue
2025Q1	59.43
2025Q2	64.96
2025Q3	74.76
2025Q4	75.94
2026Q1	74.75
2026Q2	79.05
2026Q3	90.91
2026Q4	94.02

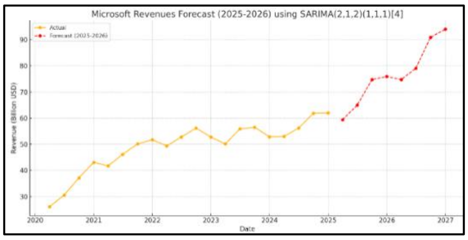

Figure 5: Microsoft Revenue Forecast (2025-2026)

Conclusion

This research aimed to analyze and model the quarterly revenue of Microsoft using advanced time series techniques to develop a robust forecasting model that can inform future strategic decisions. The Phillips-Perron unit root test results indicated that the revenue series was non-stationary in levels but achieved stationary after applying first-order and seasonal differencing. This confirmed the appropriateness of employing a Seasonal ARIMA (SARIMA) model that accounts for both trend and seasonal components. Autocorrelation and partial autocorrelation analyses revealed a clear seasonal structure with a periodicity of four quarters, which was further supported by formal seasonal tests. Several SARIMA models were estimated, and based on model selection criteria—specifically the Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC)—the SARIMA(2,1,2)(1,1,1)[4] model was identified as the optimal specification. This model exhibited the lowest AIC and BIC values, with all estimated coefficients statistically significant. Diagnostic checking using the Ljung–Box Q-test confirmed that the residuals were free from serial correlation, indicating that the model was well-specified. The forecasting performance of the selected model was validated through comparison with actual data from 2005Q2 to 2025Q1, showing a close alignment and affirming the model's predictive strength. Accordingly, the SARIMA(2,1,2)(1,1,1)[4] model is recommended for forecasting Microsoft's quarterly revenues, with the suggestion to update it periodically and consider incorporating exogenous variables for enhanced accuracy in future studies.

Reference

[1] Abd, H., Rasheed, A., & Al-Saffar, R. (2023). Studying Some Air Pollutants by Using the Nonlinear Autoregressive Distributed Lags (NARDL) Model. Journal of Al-Rafidain University College For Sciences (Print ISSN: 1681-6870, Online ISSN: 2790-2293), (1), 364-373.

[2] Al-Manaseer, W. S. I. (2023). Predict the number of Husayni processions for the Arbaeen visit using modified Box-Jenkins methodology. AL ARBA'IN, (ج 2)

[3] Box,G.E and Jenkins, G.C Reinsel(1994)"Time Series Analysis Forecasting and Control",3th ed.,Englewood Cliffs: prentice Hall.

[4] Damodar N . Gujarati ,2009, Domnc porter , "Basic Econometric " , 15th Edition , McGraw, Hill .

[5] Hyndman.R, 2014,"Forcasting Principle & Practice", University of Western Australia.

[6] Olivera.P.J.,Steffen.J.L and Cheung. P,2017,"Parameter Estimation of Seasonal Arima Model for Water Demand Forecasting Using the Harmony Search Algorithm", Procedia Engineering 186,177-185.

[7] Pank .R , 1983" Forecasting with Univariate Box – Jenkins models" , John Wiley & sons , 1983.

[8] Phillips- Perron,1986 , testing For à Unit roots in time series Regression, Biometrica, vol. 75..,

[9] Voind, H.D.(1999)"Time Series analysis" Economic Fordham University, Bronex, New York, USA