World Journal of Information Technology
ISSN: 2959-9911
DOI: https://doi.org/10.61784/wjit3033

NASDAQ INDEX PREDICTION BASED ON ARIMA-GARCH MODEL AND DYNAMIC REGRESSION

YiLin Peng
South China Normal University, School of Mathematical Sciences, Guangzhou 510631, Guangdong, China.
Corresponding Email: 20212831009@m.scnu.edu.cn

Abstract: As the global market becomes more and more open and volatile, it is of great significance to grasp the financial temporal volatility and correlation and accurately predict stock price behavior. This paper takes the Nasdaq Composite Index in 2020-2023 as the research object, constructs the ARIMA(1,1,(1,5)) model with GARCH(1,1) conforming to t distribution disturbance term to fit the trend of the index in 2020-2022, and predicts the trend in the first half of 2023. The results show that the model with GARCH effect gives a wider forecast confidence interval and can indicate the potential risk, but can not accurately reflect the real trend of the index. To improve the prediction accuracy, this paper takes Nasdaq index as the response variable, introduces S&P 500 index as the input variable, constructs an effective dynamic regression model through Granger causality test and EG cointegration test, and improves the model through cross-correlation function analysis. The results show that the forecast trend of the model is closer to the actual series of fluctuations, indicating that the S&P 500 index plays a promoting role in predicting the Nasdaq index, which provides a more reliable reference for investors when weighing the benefits and risks.
Keywords: Volatility and correlation; Financial timing; ARIMA; GARCH; Cointegration; Dynamic regression

1 INTRODUCTION

Since 2020, the U.S. stock market has experienced sharp fluctuations under the impact of the COVID-19 pandemic. In response, the Federal Reserve implemented a zero interest rate and quantitative easing (QE) policies, alongside fiscal stimulus measures, leading to a rapid market rebound. Indices such as the Nasdaq and the S&P 500 recovered strongly. As a representative of technology stocks and innovative enterprises, the Nasdaq Index has significant spillover effects on global markets. Therefore, accurately forecasting its volatility not only helps investors grasp market trends but also enhances global risk warning capabilities and improves asset allocation efficiency.

Volatility and correlation are two core features of financial time series. How to effectively capture these characteristics, accurately describe stock price behaviors, and predict market movements has long been a central concern for both investors and the academic community.

In terms of volatility, market fluctuations are driven by multiple factors, including investor behavior, macroeconomic policies, and market structure, often exhibiting volatility clustering [1]. Engle and Bollerslev respectively proposed the ARCH and GARCH models [2], followed by variants such as EARCH, IGARCH, and GARCH-M [3], which enhance the models' ability to capture asymmetries and risk premiums. Marisetty N found that the GARCH(1,1) model performed well in balancing forecasting ability and simplicity across five major international indices [4]. Raza S, focusing on the Indian Green Finance Index, found that the APARCH(1,1) model best characterized the volatility of firms associated with carbon performance [5]. Roszyk N combined GARCH with deep learning models like LSTM and incorporated VIX information to construct a hybrid model [6], significantly improving the forecasting accuracy for the S&P 500.

In terms of correlation, after Granger introduced the concept of causality in time series [7], models such as ARIMAX and cointegration models have been widely used to quantify dynamic relationships among time series. Wang P C et al. combined XGBoost and ARIMAX to predict closing prices in the Vietnamese stock market [8], outperforming LSTM and other models. Akusta A introduced CEEMDAN decomposition in combination with ARIMAX, effectively improving the prediction accuracy for the Dow Jones Index [9].

In summary, although existing studies have made substantial progress in modeling volatility and correlation, most focus on forecasting within a single stock market, lacking dynamic characterization of inter-market linkages. Therefore, this paper takes the Nasdaq Index as the primary research object, incorporates related variables such as the S&P 500, and employs the ARIMA-GARCH model along with the ARIMAX dynamic regression model to capture volatility information in time series and quantify these complex interdependencies, while also comparing the predictive performance of different models.

2 RESEARCH METHODS AND THEORETICAL ANALYSIS

This section mainly introduces several commonly used models in financial time series analysis: ARIMA, GARCH, Granger causality test, cointegration modeling, and dynamic regression.

2.1 ARIMA-GARCH Model

2.1.1 ARIMA(p, d, q) model

The ARIMA model is a commonly used approach for fitting non-stationary time series with trends. The main steps include:

(1) Removing trend components through differencing

The differencing method extracts deterministic components of a time series, such as trends and cycles. For a series with a significant linear trend, first-order differencing can often achieve stationarity; for curved trends, second- or third-order differencing may be required. Although multiple rounds of differencing can help extract deterministic information from a non-stationary series, over-differencing can lead to the loss of valuable information, increased variance, and reduced fitting accuracy, and should therefore be avoided whenever possible.

(2) Stationarity Test and White Noise Test

- ADF Test and PP Test: The null hypothesis of the Augmented Dickey-Fuller (ADF) test is that the time series is non-stationary. If the test statistic $\tau \leqslant \tau_\alpha$ (with α being the significance level), the null hypothesis is rejected, indicating that the series is stationary [10]. When heteroscedasticity is present in the series, the Phillips-Perron (PP) test is used instead, as its adjusted test statistic provides a more accurate assessment of stationarity.
- White Noise Test (Ljung–Box Q Test)
  Under the assumption that the series is stationary, the Ljung–Box Q test evaluates whether the autocorrelation function is significantly different from zero in order to determine whether the series is white noise. The null hypothesis of this test is that the series is white noise. If the test statistic $Q \geqslant \mathcal{X}_{1-\alpha}{}^2(m)$, the null hypothesis is rejected, indicating that the series is not white noise [11].
  If the differenced series is stationary and not white noise, it is necessary to proceed with fitting an ARMA(p, q) model.

(3) Model Order Selection

The key to ARMA(p, q) modeling lies in determining the appropriate values for parameters p and q, which can be judged based on the patterns of the autocorrelation function (ACF) and partial autocorrelation function (PACF):

- AR(p) model: PACF cuts off at lag p, ACF tails off.
- MA(q) model: ACF cuts off at lag q, PACF tails off.
- ARMA(p, q) model: Both ACF and PACF tail off.

(4) Parameter Estimation and Model Diagnostics

In this study, the conditional least squares method is used to estimate the parameters of the ARIMA model. This method has the advantage of not requiring a prior assumption about the distribution of the series, and it makes full use of sample information, resulting in high estimation accuracy.

Model diagnostics primarily include the significance test of the parameters (t-test) and the white noise test of the residuals (Ljung–Box Q test). If the residuals are not white noise, it indicates that the model has not fully captured the time series information and needs to be revised. If the residuals are white noise, it suggests that the model fits the time series well.

2.1.2 ARIMA(p, d, q) model with GARCH(p, q) disturbance terms

(1) Conditional Heteroscedasticity Test

The parameter estimation and testing of the ARIMA model require the assumption of homoscedasticity in the disturbance terms ε_t ; otherwise, it will affect the accuracy of the model's estimates. Therefore, after testing for zero mean and pure randomness of ε_t, it is also necessary to check whether ε_t has homogeneity of variance. Economists believe that this heteroscedasticity is caused by some autocorrelation relationship, which is usually modeled by an autoregressive model of the squared residual series $\{\varepsilon_t^2\}$. Common methods for testing heteroscedasticity include the following two:

- Graphical Test: If the time series plot shows a volatility clustering effect (alternating small and large fluctuations), conditional heteroscedasticity may be present.
- ARCH Test: The heteroscedasticity is examined by checking the autocorrelation of $\{\varepsilon_t^2\}$, mainly through the Ljung–Box Q test and LM test. The Q test examines the autocorrelation of $\{\varepsilon_t^2\}$ through the autocorrelation function coefficients, while the LM test establishes an autoregressive model for $\{\varepsilon_t^2\}$ to assess its autocorrelation. The null hypothesis for both tests is that the residual squared series has no autocorrelation, which is equivalent to stating that the residual series has no heteroscedasticity.

If the conditional variance of ε_t is not homogeneous, it is referred to as having ARCH effects, and further modeling of ε_t using ARCH or GARCH should be performed.

(2) GARCH(p, q) Model

A commonly used model for fitting time series with conditional heteroscedasticity is the GARCH family of models. The ARCH(q) model is suitable for short-term autocorrelation in $\{\varepsilon_t^2\}$, while the more broadly applicable GARCH(p, q) model is suitable for long-term autocorrelation in $\{\varepsilon_t^2\}$. When extracting higher-order autocorrelation information from $\{\varepsilon_t^2\}$, the GARCH(p, q) model, which has a relatively low order, is more effective in capturing the information than the

ARCH(q) model with a very high order.

When the conditional volatility of ε_t is not homogeneous, heteroscedasticity is introduced with a sequence h_t, and the expression for fitting ε_t using the GARCH(p, q) model is as follows:

$$\varepsilon_t = \sqrt{h_t} e_t \ , e_t \sim WN(0, \sigma^2) \tag{1}$$

$$h_t = \lambda_0 + \sum_{j=1}^{p} \eta_j h_{t-j} + \sum_{i=1}^{q} \lambda_i \varepsilon_{t-i}^2 \tag{2}$$

$$\lambda_i, \eta_j \in [0,1) \ , \ \sum_{j=1}^{p} \eta_j + \sum_{i=1}^{q} \lambda_i \in [0,1) \tag{3}$$

It can be proven that $Var(\varepsilon_t|\varepsilon_{t-1},...) = h_t$, meaning that h_t is the conditional heteroscedasticity of ε_t. Therefore, it can be seen that the GARCH(p, q) model essentially fits the ARMA(p, q) model to the conditional heteroscedasticity h_t of ε_t.

For the order selection of GARCH, similar to ARMA, the order p and q of the GARCH model can be determined by examining the ACF and PACF plots of $\{\varepsilon_t^2\}$.

(3) Standardized Residual Test and Final Model

After fitting ε_t using the GARCH(p, q) model, the focus should be on e_t in Equation (1), which represents the standardized residuals of the GARCH(p, q) model. This paper will sequentially test the conditional heteroscedasticity of e_t and the distribution assumption, in order to determine whether the volatility information of ε_t has been fully extracted and whether the coefficient significance tests are valid.

First, to assess the conditional heteroscedasticity of e_t, the autocorrelation of $\left(\varepsilon_t/\sqrt{h_t}\right)^2$ can be tested using the Ljung-Box Q test or LM test. Second, the GARCH(p, q) model generally assumes $e_t \sim N(0,1)$ because parameter estimation and tests are conducted under the normality assumption. However, considering that financial time series often exhibit leptokurtosis (fat tails), it is frequently assumed that $e_t \sim t(m)$. This paper uses the Jarque-Bera (JB) test(H_0:$\varepsilon_t/\sqrt{h_t} \sim N(0,1)$, H_1:$\varepsilon_t/\sqrt{h_t} \overset{not}{\sim} N(0,1)$) and the Kolmogorov-Smirnov (K-S) test(H_0:$\varepsilon_t/\sqrt{h_t}$~$t(m)$, H_1:$\varepsilon_t/\sqrt{h_t} \overset{not}{\sim} t(m)$) to determine the distribution of e_t.

When both the conditional homoscedasticity and distribution tests of e_t are passed, it is considered that the volatility information of ε_t has been fully extracted. The final expression for the ARIMA(p, d, q) model with GARCH(p', q') disturbance terms is:

$$\nabla^d x_t = \phi_0 + \frac{1-\theta_1 B - ... - \theta_q B^q}{1-\phi_1 B - ... - \phi_p B^p} \varepsilon_t \tag{4}$$

$$\varepsilon_t = \sqrt{h_t} e_t \ , e_t \sim N(0,1) \ or \ e_t \sim t(m) \tag{5}$$

$$h_t = \lambda_0 + \sum_{j=1}^{p'} \eta_j h_{t-j} + \sum_{i=1}^{q'} \lambda_i \varepsilon_{t-i}^2 \tag{6}$$

where x_t is the time series to be fitted, B is the lag operator, ε_t is the error term, h_t is the conditional heteroscedasticity, e_t is the standardized error term, θ_i, ϕ_i are the ARIMA model parameters, and η_i, λ_i are the GARCH model parameters.

2.1.3 ARIMA+GARCH forecasting

The introduction of the GARCH model is aimed at better extracting volatility information from financial time series in order to more accurately assess future risks. Therefore, the introduction of the GARCH disturbance term will alter the standard error of the forecasted value $\hat{X}_{t+k}$, thus changing the width of the confidence interval for the forecasted value. Let the variance of $\hat{X}_{t+k}$ be $Var\left(\hat{X}_{t+k}\right)$, then:

Under the assumption of homoscedasticity (using only ARIMA for forecasting):

$$Var\left(\hat{X}_{t+k}\right) = (1 + G_1^2 + ... + G_{K-1}^2)\sigma_\varepsilon^2 \tag{7}$$

Under the assumption of heteroscedasticity (using ARIMA+GARCH disturbance term for forecasting):

$$Var\left(\hat{X}_{t+k}\right) = \hat{h}_{t+k} + G_1^2 \hat{h}_{t+k-1} + ... + G_{k-1}^2 \hat{h}_{t+1} \tag{8}$$

where G_n is the Green's function, and under the normality assumption, the 95% confidence interval for the forecast is

$$\hat{X}_{t+k} \pm 1.96\sqrt{Var\left(\hat{X}_{t+k}\right)}.$$

A large number of empirical studies have shown that when the volatility of a series is high (low), the confidence interval provided by the GARCH model is also wider (narrower). Therefore, using the ARIMA model with GARCH disturbance terms to predict stock index trends will yield results that are closer to reality.

2.2 Dynamic Regression Model

2.2.1 Granger Causality Test

The Granger Causality Test is commonly used to determine the causal relationship between time series and is the foundation of multivariate time series cointegration modeling. The idea is that, given two time series x_t, y_t, if x_{t-k} has a significant impact on y_t, then x_t is considered the cause and y_t is considered the effect, meaning the cause precedes the effect. The null hypothesis of the test is H_0: x *is not the Granger cause of* $y\,(x \nrightarrow y)$ (with the requirement that both x and y are stationary series). The problem is transformed into testing the significance of the linear model between y and x (see equation (9)), with the null hypothesis being equivalent to $H_0: \alpha_1 = \ldots = \alpha_q = 0$.

$$y_t = \beta_0 + \sum_{k=1}^{p} \beta_k y_{t-k} + \sum_{k=1}^{q} \alpha_k x_{t-k} + \varepsilon_t \tag{9}$$

2.2.2 Cointegration modeling (ARIMAX dynamic regression)

The concept of cointegration provides a theoretical basis for modeling multivariate non-stationary time series. Suppose the linear model of two time series x_t, y_t is $y_t = \beta_0 + \beta_1 x_t + \varepsilon_t$. If x_t, y_t are non-stationary, the stationarity of ε_t cannot be guaranteed, leading to issues like spurious regression, which makes parameter significance tests ineffective. Cointegration testing examines the stationarity of ε_t. If ε_t is stationary, then x_t, y_t are cointegrated, which also means that the linear model between x_t, y_t is valid. The hypothesis for the EG cointegration test is: H_0: x_t, y_t do not have a cointegration relationship, which is equivalent to ε_t being non-stationary. H_1: : x_t, y_t have a cointegration relationship, which is equivalent to ε_t being stationary.

After the cointegration test, a dynamic regression model between x_t, y_t can be established. The lag variables to be introduced are determined by examining the cross-correlation function between y and x, and the specific form of the dynamic regression model is then determined for forecasting purposes.

3 EMPIRICAL ANALYSIS AND RESULTS

3.1 Data Source

This study selects weekly data of the NASDAQ Composite Index from January 5, 2020, to June 25, 2023, obtained from the official Statista website. This period covers several major global events, including the outbreak of the COVID-19 pandemic, large-scale fiscal and monetary stimulus policies by various governments, and the onset of the Federal Reserve's interest rate hikes. These events caused significant fluctuations in the capital markets, making the data highly representative. As a key benchmark for the global technology stock market, the NASDAQ Composite Index includes nearly all common stocks listed on the NASDAQ exchange. Its performance reflects not only the development of U.S. technology firms but also global tech sector trends, which highlights the practical significance of modeling and forecasting this index.

3.2 ARIMA-GARCH Empirical Analysis and Results

To explore the dynamic characteristics of the index and evaluate its future volatility trend, the ARIMA-GARCH model is first used to fit the weekly data from 2020 to 2022 (a total of 156 periods), and predict the trend of the index for the first half of 2023. The model's effectiveness is tested by comparing the predicted values with the actual values.

3.2.1 ARIMA fitting of the nasdaq time series

Since the Nasdaq time series X_t from 2020 to 2022 exhibits a linear trend of first increasing and then decreasing (Figure 1), a first-order difference is applied. The result (Figure 2) shows that the differenced series ∇x_t has no obvious trend, but the volatility increases in the later periods, suggesting the possibility of conditional heteroscedasticity. The PP test ($p < 0.0001$) and the Ljung-Box Q test for lags 18-30 ($p < 0.05$) indicate that ∇x_t is stationary and not white noise, so the ARMA model fitting should continue.

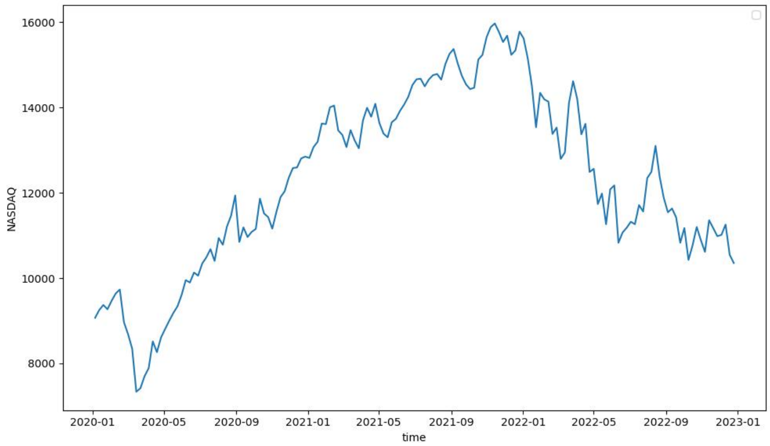

Figure 1 Time Series Plot of the NASDAQ Index

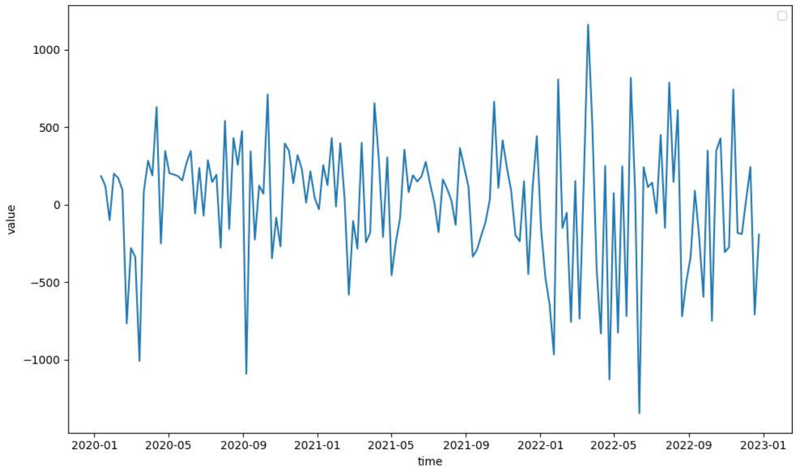

Figure 2 Differenced NASDAQ Time Series Plot

Since the ACF and PACF of ∇x_t (Figure 3) decay to zero after the 5th lag without clear cutoff, an ARMA(5,5) model is considered to fit ∇x_t. After removing the insignificant coefficients, the sparse coefficient model ARMA(1, (1,5)) is obtained. Model parameters are significant, and the disturbance terms can be considered as white noise (Tables 1 and 2), indicating that the model has fully extracted the deterministic components of X_t.

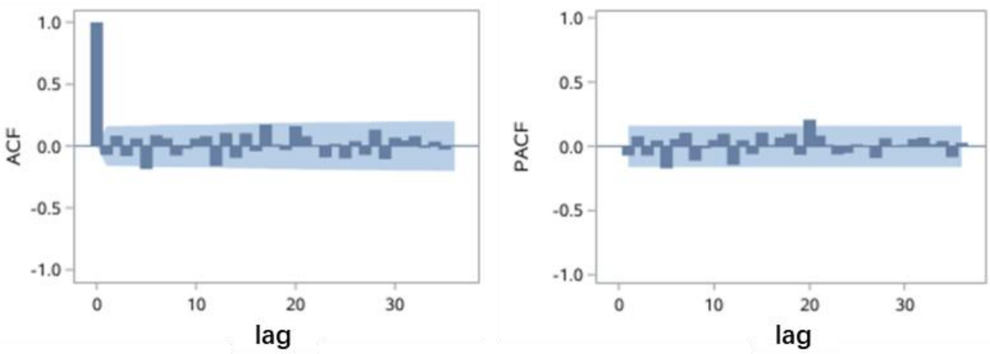

Figure 3 The ACF and PACF of ∇x_t

Table 1 Parameter Estimation of the ARIMA(1,1,(1,5)) Model($\sigma = 408.858$, SBC=2316.116)

parameter	estimate	std	t	p
MA1,1	-0.53864	0.23327	-2.31	0.0223
MA1,5	0.18164	0.0804	2.26	0.0253
AR1,1	-0.61029	0.22813	-2.68	0.0083

Table 2 White Noise Test of ARIMA Residuals

lag	Q	df	p
6	0.61	3	0.8939
12	8.23	9	0.5115
18	17.06	15	0.3152

3.2.2 GARCH fitting of nasdaq index residuals

Since the Nasdaq index still exhibits volatility clustering effects after differencing (Figure 2), and the Q test and LM test of the squared residual sequence ε_t^2 from the ARIMA model (Table 3) indicate long-term autocorrelation, considering that a low-order GARCH model can effectively capture this long-term autocorrelation, we directly use the GARCH(1,1) model to fit ε_t. The model expression is as follows:

$$\varepsilon_t = \sqrt{h_t} e_t \ , e_t \sim WN(0, \sigma^2) \tag{10}$$

$$h_t = \lambda_0 + \eta_1 h_{t-1} + \lambda_1 \varepsilon_{t-1}^2 \tag{11}$$

Table 3 Conditional Heteroskedasticity Test of ε_t

ARCH lag	Q	Pr>Q	LM	Pr>LM
6	8.6607	0.1936	8.4581	0.2064
12	44.7741	<. 0001	28.7565	0.0043

Since the JB test indicates that the standardized residuals e_t significantly deviate from a normal distribution (JB = 9.2894, p = 0.0096), the GARCH(1,1) model with a t-distribution assumption is used to fit ε_t. The results show that e_t passed the KS test (KS Statistic = 0.1025, p = 0.07168). Therefore, parameter estimation and model testing can be effectively conducted (Table 4). The Q test and LM test(Table 5) both show that e_t^2 has no autocorrelation, suggesting that the conditional volatility information of ε_t has been adequately extracted.

Table 4 Parameter Estimation of GARCH(1,1) Model

Parameter	coef	std err	t	P>|t|	95.0% Conf. Int.
λ_0	3285.2352	2994.1250	1.0970	0.2730	[-2.583e+03,9.154e+03]
λ_1	0.1051	0.0623	1.6860	0.0919	[-1.710e-02, 0.227]
η_1	0.8812	0.0444	19.8310	0.0000	[0.794, 0.968]

Table 5 Conditional Heteroskedasticity Test of e_t

Lag	LM Statistic	p-value (LM)	Q Statistic	p-value (Q)
6	3.9167596	0.6879401	4.573374	0.599572
12	9.8372461	0.6302365	12.25077	0.425756

3.2.3 ARIMA-GARCH forecasting performance

In summary, this paper first uses ARIMA(1,1,(1,5)) to extract the deterministic information from the time series, and then employs GARCH(1,1) to capture the conditional volatility information of the residuals. The final model expression for fitting X_t is:

$$(1-B)X_t = \frac{1+0.53864B-0.18164B^5}{1+0.61029B}\varepsilon_t \tag{12}$$

$$\varepsilon_t = \sqrt{h_t} e_t \ , e_t \sim t(17.1962) \tag{13}$$

$$h_t = 3285.2352 + 0.8812 h_{t-1} + 0.1051 \varepsilon_{t-1}^2 \tag{14}$$

The model prediction results (Figure 4) show that the ARIMA model with GARCH effects has a wider confidence interval for the predicted values, indicating that the GARCH model is better at capturing the volatility in financial time series. The wider confidence interval suggests that investors need to be more cautious when weighing the returns and risks of Nasdaq stocks, especially in the context of the Federal Reserve's frequent interest rate hikes to curb high inflation. This policy shift has directly impacted the liquidity and valuation in the capital markets.

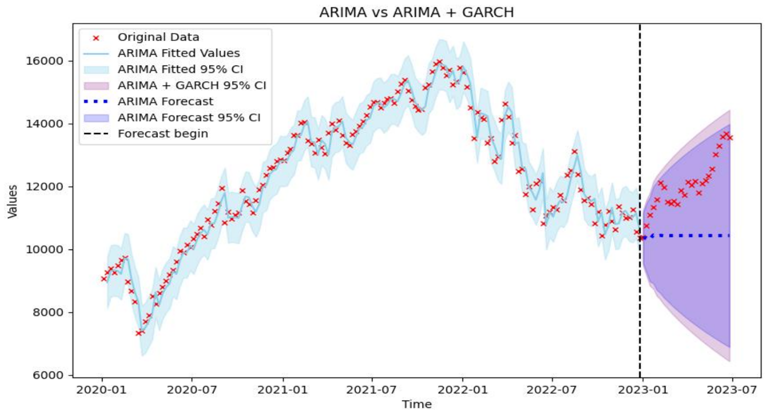

Figure 4 Comparison Chart of Forecasting Performance with and without GARCH Effects

3.3 Empirical Analysis and Results of the Dynamic Regression Model

The main limitation of the ARIMA model in forecasting the Nasdaq index time series is that the forecast period is relatively short, and the predicted values tend to stabilize, making it difficult to reflect the real volatility increase trend (Figure 4). To improve the forecasting accuracy and explore the interconnectedness of the U.S. stock market, this study introduces input variables and constructs a dynamic regression model to reforecast the Nasdaq index trends.

3.3.1 Selection of input variables and data preprocessing

In this study, the S&P 500 index $z_t^{(1)}$, which is also highly watched, is selected as a potential input variable alongside the NASDAQ Composite Index X_t. To avoid heteroscedasticity, both stock time series are first transformed using logarithms. After transformation (Figure 5), the trends of both indices are largely consistent. Since Granger causality tests require stationary series, a first-order difference is applied to the logarithmic series of both indices. The ADF test ($p_x, p_z < 0.001$) shows that both $\nabla \ln x_t$ and $\nabla \ln z_t^{(1)}$ are stationary. The results of the Granger causality test and cointegration test (Table 6) indicate that it is appropriate to fit a dynamic regression model for $\nabla \ln x_t$ and $\nabla \ln z_t^{(1)}$.

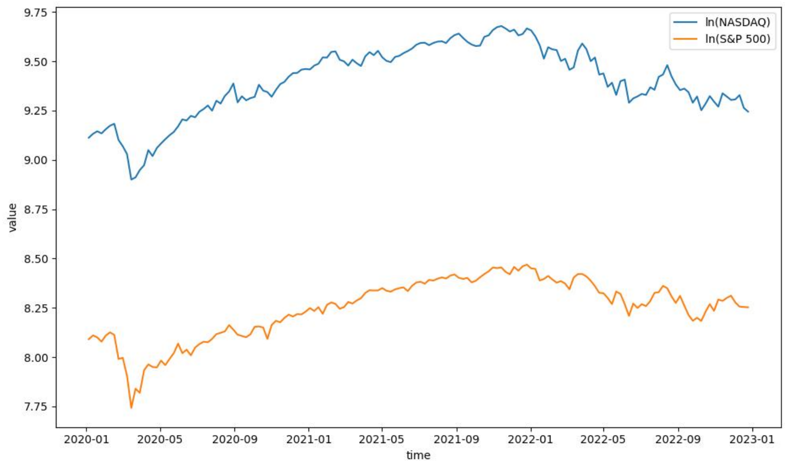

Figure 5 Log-Transformed S&P 500 and Nasdaq Index Trends from 2020 to 2023

Table 6 Granger Causality Diagnosis and Cointegration Test for S&P 500 and Nasdaq

null hypothesis	Statistics	p
The S&P 500 ($\nabla \ln z_t^{(1)}$) is not the Granger cause of the Nasdaq($\nabla \ln x_t$).	F=72.3113	0.0000
no cointegration relationship between $\nabla \ln z_t^{(1)}$ and $\nabla \ln x_t$.	τ =-6.7179	0.0000

3.3.2 Fitting the dynamic regression model to the NASDAQ time series

To determine the specific form of the dynamic regression model, it is necessary to identify several input variables that are strongly correlated with the response series $\nabla \ln x_t$. By examining the cross-correlation function between $\nabla \ln x_t$ and $\nabla \ln z_t^{(1)}$ (Figure 6), significant cross-correlation coefficients at lags 0 and 1 were found. Therefore, $\nabla \ln z_t^{(1)}$ and $\nabla \ln z_{t-1}^{(1)}$ were selected as input variables to be included in the regression equation. The dynamic regression model is constructed as follows:

$$\nabla \ln x_t = \beta_0 + \beta_1 \nabla \ln z_t^{(1)} + \beta_2 \nabla \ln z_{t-1}^{(1)} + \varepsilon_t \tag{15}$$

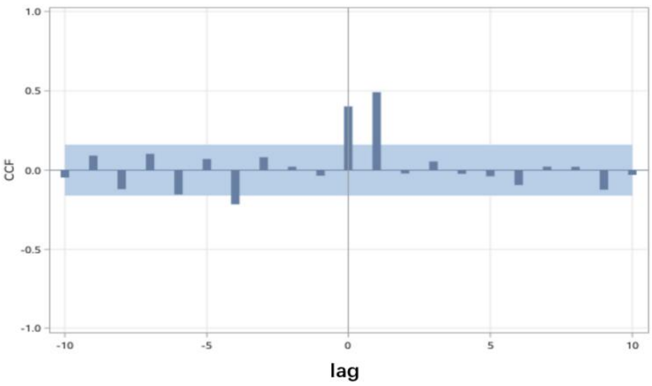

Figure 6 The Cross-Correlation Function Plot between $\nabla \ln x_t$ and $\nabla \ln z_t^{(1)}$

The test reveals that ε_t is stationary and non-white noise, with an ACF tailing off and a PACF exhibiting a first-order cutoff. Therefore, an AR(1) model is fitted to ε_t. After removing the insignificant coefficients, the final expression of the model is as follows:

$$\nabla \ln x_t = 0.4676 \nabla \ln z_t^{(1)} + 0.5583 \nabla \ln z_{t-1}^{(1)} + \varepsilon_t \tag{16}$$

$$\varepsilon_t = \frac{1}{1 + 0.46214B} a_t \text{ , } a_t \sim WN(0, 0.000569) \tag{17}$$

3.3.3 Dynamic Regression Forecasting Performance

The forecasting results of the Nasdaq for the first half of 2023 based on the dynamic regression model (Figure 7) show that although there is still some deviation between the predicted values and the actual values, the predicted trend aligns with the real trend, both showing a fluctuating upward trend. This indicates that the inclusion of the S&P 500 index has improved the forecasting accuracy of the Nasdaq index.

Figure 7 Dynamic Regression Model Prediction Results Plot

4 CONCLUSIONS AND OUTLOOKS

This study focuses on the weekly data of the Nasdaq Composite Index (NASDAQ_Index) from January 5, 2020, to June 25, 2023. In order to better capture both the deterministic and conditional volatility information of this financial time series, an ARIMA(1,1,(1,5)) model with GARCH(1,1) disturbances following a t-distribution was initially established to fit the index time series from 2020 to 2022. Subsequently, the index trends for the first half of 2023 were predicted using both models with and without GARCH effects. The results show that the model with GARCH effects produced a wider prediction confidence interval, indicating that this model is more capable of forecasting potential risks in future financial environments. This also suggests that investors need to be more cautious when balancing the returns and risks of Nasdaq stocks.

However, the forecasts generated by the ARIMA-GARCH model became more stable as the forecast period increased, failing to reflect the true rising volatility trend of the Nasdaq index effectively. To improve prediction accuracy, this study then applied the concept of cointegration by introducing the S&P 500 index as an input variable. Through Granger causality and EG cointegration tests, an effective dynamic regression model was initially built. By fully exploiting the information of the input variable using the cross-correlation function between the Nasdaq index and the S&P 500 index, the model's specific form was determined and refined. The results show that the forecast trend of this model largely follows the true sequence's rising volatility, indicating that the S&P 500 index plays a facilitative role in predicting the Nasdaq index.

Although the dynamic regression model constructed in this study has significantly improved the forecasting accuracy of the NASDAQ Index under the condition of known input variables, it still has certain limitations. Firstly, the model's effectiveness depends on the availability of future values of input variables (such as the S&P 500 Index), which are often unknown in real-world forecasting scenarios. Secondly, even if these input variables are forecasted separately, the associated prediction errors may propagate and negatively impact the accuracy of the response variable forecast. Future research could consider integrating dynamic regression with multivariate time series modeling or machine learning techniques to enhance the model's robustness and adaptability, thereby better capturing the complex volatility patterns in financial markets.

COMPETING INTERESTS

The authors have no relevant financial or non-financial interests to disclose.

REFERENCES

[1] Akin I, Akin M. Behavioral finance impacts on US stock market volatility: An analysis of market anomalies. Behavioural Public Policy, 2024, 1-25.

[2] Muliabanta N H. Peramalan Volatilitas Return IHSG Menggunakan GARCH-Genetic Algorithm-Support Vector Regression. Institut Teknologi Sepuluh Nopember, 2025.

[3] Nugroho D B, Kurniawati D, Panjaitan L P, et al. Empirical performance of GARCH, GARCH-M, GJR-GARCH and log-GARCH models for returns volatility[C]//Journal of Physics: Conference Series. IOP Publishing, 2019, 1307(1): 012003.

[4] Marisetty N. Prediction of Popular Global Stock Indexes Volatility by Using ARCH/GARCH Models. GARCH Models, 2024.

[5] Raza S, Shreevastava A, Meher B K, et al. Predictive Analysis of Volatility for BSE S&P GREENEX index using GARCH Family Models: A case study for Indian stock market. Revista de Științe Politice. Revue des Sciences Politiques, 2024, 84, 54-67.

[6] Roszyk N, Ślepaczuk R. The Hybrid Forecast of S&P 500 Volatility ensembled from VIX, GARCH and LSTM models. arXiv preprint arXiv:2407.16780, 2024.

[7] Shojaie A, Fox E B. Granger causality: A review and recent advances. Annual Review of Statistics and Its Application, 2022, 9(1): 289-319.

[8] Wang P C, Vo T T H. Stock price prediction based on dual important indicators using ARIMAX: A case study in Vietnam. Journal of Intelligent Systems, 2025, 34(1): 20240101.

[9] Akusta A. Exploring ceemdan decomposition for improved financial market forecasting: A case study on dow Jones index. Pamukkale Üniversitesi Sosyal Bilimler Enstitüsü Dergisi, 2024, (62): 19-35.

[10] Aylar E, Smeekes S, Westerlund J. Lag truncation and the local asymptotic distribution of the ADF test for a unit root. Statistical Papers, 2019, 60: 2109-2118.

[11] Yasmin S, Moniruzzaman M. Forecasting of area, production, and yield of jute in Bangladesh using Box-Jenkins ARIMA model. Journal of Agriculture and Food Research, 2024, 16: 101203.