Forecasting Stock Market Trends of S&P 500 based on Fractal Theory and Polynomial Regression Model

Baozhen Che*

School of Civil Aviation, Nanjing University of Aeronautics and Astronautics, Nanjing, 210016, China

* Corresponding Author Email: wmcvey74529@student.napavalley.edu

Abstract. The stock market is a key component of the national economy in the financial sector. Stock market analysis and prediction have consistently remained popular research directions, offering vital practical guidance for financial investments. Consequently, scholars from both domestic and international spheres have employed various methodologies to analyze and forecast stock market trends. This paper focuses on the Standard & Poor's 500 Index (S&P 500) as the primary research subject. By computing its box dimension, the paper provides insights into the future trends and complexities of the stock market using fractal theory and a polynomial regression model. Additionally, the paper presents trend-tracking strategies accordingly, offering valuable guidance to investors. The primary dataset used for this study is sourced from the Kaggle database, specifically comprising the closure prices for the S&P 500 Index ranging from September 3, 2013, to August 31, 2023. The results demonstrate the efficacy and validity of the proposed third-order polynomial regression model, with a root mean square error (RMSE) of 0.08. Moreover, a trading strategy that suggests appropriate buy or sell operations when the price trend coincides with the downward pattern of the box dimension is introduced. These findings provide theoretical and empirical references for the field of financial investment. However, it is essential to note that in order to enhance the predictive performance, future initiatives should incorporate a more extensive sample size and complement the polynomial regression model with combinations of other models.

Keywords: S&P 500, fractal theory, box dimension, polynomial regression model, forecast.

1. Introduction

The stock market holds a prominent position within the financial and securities industry, attracting considerable attention from investors. Accurate stock prediction is of utmost importance in the field of financial investment, bringing about potentially high market returns for investors and facilitating effective market regulation by governmental entities. Consequently, the analysis and prediction of stock prices carry significant practical and theoretical value. However, stock prices are subject to uncertainties due to the influence of diverse complex factors, including policies, the economy, and investor psychology. These uncertainties present significant challenges to stock prediction.

Consequently, the stock market can be viewed as a complex nonlinear system with multiple objectives, variables, and levels. Its dynamics exhibit irreversibility in time and multiple causal feedback, rendering it inherently uncertain. Therefore, traditional economic tools such as linear analysis and approximation analysis prove inadequate in accurately capturing the complexity of the stock market [1]. Hence, the adoption of nonlinear theories and methodologies becomes more appropriate for unraveling the fundamental characteristics of the stock market and introducing novel perspectives for its study. Fractal theory, a subordinate theory of nonlinear theory, offers a valuable tool for studying complex geometric shapes and patterns that display self-similarity and scale invariance. Given the non-linear and intricate nature of financial markets, fractal theory has emerged as an effective approach for exploring such phenomena.

Fractal dimension stands as a crucial concept within fractal theory, serving to measure the geometric dimension of fractal objects. Since the 1980s, the concept of fractal dimensions has found application in describing the self-similarity and multi-scale structure of financial markets, yielding promising outcomes in predicting volatility, risk management, asset pricing, and hedging strategies [2]. Peters utilized the R/S analysis method to conduct a test on the normality of asset price changes

in the capital market and discovered that asset prices or asset return sequences exhibit patterns of either fractal Brownian motion or biased random walk [3]. Nwarocki analyzed the S&P300 stock index and reported that this index experiences an average cycle of five years [4]. Pasquini & Serva analyzed the scaling behavior of generalized cumulative absolute returns by using daily returns as a probability measure of prices. The results showed that the volatility exhibits a power-law relationship, and the scaling exponent is not unique, providing evidence for the existence of multifractal phenomena. [5]. Traina proposed a fractal dimension-based attribute selection method in 2000, which has been widely used in various applications, such as the traveling salesman problem and the multi-objective assignment problem [6]. Tokinaga utilized wavelet analysis to measure the fractal dimension of time series and based on self-similarity, successfully predicted the time series with promising results [7]. Alvarez et al. discussed the multifractal features of international crude oil prices and identified two characteristic time scales connected to weekdays and quarters [8]. Wang et al. examined the fractal characteristics of the Chinese financial market by calculating the Lyapunov exponent of the Shanghai Stock Exchange Composite Index and the Shenzhen Component Index, and conducting correlation dimension analysis [9]. Lee et al. studied the multifractal characteristics of Korean Stock Price Index (KOSPI) by using the multiple-affine method [10]. Various studies reveal that fractal dimensions provide valuable insight into critical characteristics of these markets.

The objective of this research is to explore the potential application of the fractal dimension method in predicting the American stock index. By computing the fractal dimension of the price series linked to the S&P 500 index and examining the changing trend of this fractal dimension, the study specifically tries to forecast the future price trend of the stock index. Widely recognized as a bellwether of the US stock market, the S&P 500 index exhibits substantial price fluctuations, thereby possessing considerable potential for analysis and prediction.

2. Methods

2.1. Data sources

The S&P 500 Index was established by the American company Standard & Poor's, recording the stock index of 500 publicly listed companies. Standard & Poor's was founded in 1860 and is a global information services provider, offering services such as investment consulting, research analysis, and credit ratings. Additionally, the business offers credit ratings for more than 220,000 securities and funds globally, establishing itself as a premier source of information and a trusted authority in the field of analysis. In 1957, the first S&P 500 index was created. The earliest parts were made up of 60 utility companies, 15 railroad stocks, and 425 industrial firms. The components were amended to include 400 industrial firms as of 1976. The base index for the index, which is based on the years 1941 to 1943, is set at 10. It is derived using a weighted average technique, with the weight being determined by the base period and the stock listing volume. The calculating equation are as follows:

$$S\&P\ 500 = \frac{\text{Base Period Market Value} * \text{Current Total Market Value}}{\text{Base Period Total Market Value}} \quad (1)$$

The S&P 500 index exhibits the traits of extensive sampling coverage, great representativeness, high accuracy, and good continuity when compared to the Dow Jones Industrial Average. It is often regarded as the perfect underlying asset for futures contracts on stock indices. Therefore, in the present paper, this index is selected as a proxy variable to represent the fluctuation of the U.S. stock market.

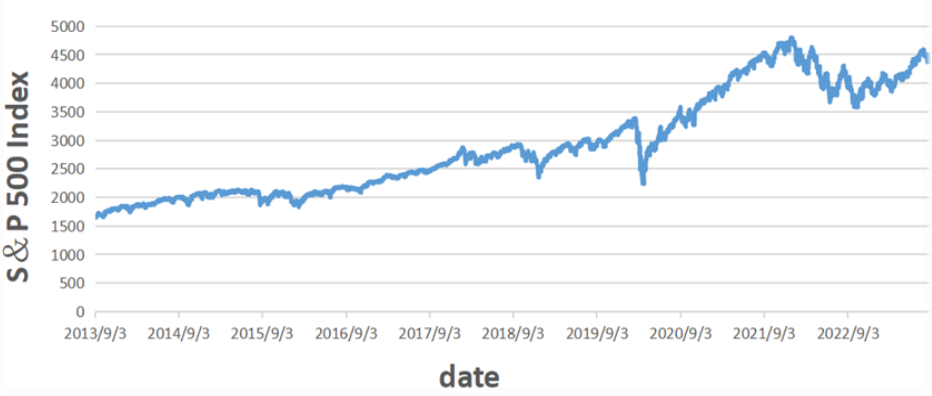

Figure 1. S&P 500 Index.

In this study, a sample of 2517 daily trading data sets for the S&P 500 index was obtained via Kaggle Data Online, covering the period from September 3, 2013 to August 31, 2023. The Figure 1 shows original data.

2.2. Variable description

Fractal theory is applied to analyze the future trends and complexities of the stock market by examining the irregularity and self-similarity of stock index time series data. The fractal dimension provides a measure of the degree of self-similarity in a time series and describes the ability to fill the space of the time series [11]. In the context of stock index time series, fractal dimension characterizes the irregularity of the time series, describes the ability to fill the space of the time series, and quantifies the degree of "bias" in the observed values. This allows for the identification of factors causing the "bias" in the time series. Currently, there are several types of fractal dimensions, such as box dimension, Hausdorff dimension, similarity dimension, capacity dimension, and grid dimension. Within these, the box dimension is widely used due to its relatively easier mathematical approximation and empirical assessment. The computation of the box dimension entails partitioning the time series dataset into non-overlapping boxes of specified size, and then determining the minimum number of boxes required to cover the entire time series. The box dimension is ascertained by computing the slope of the linear regression model over the log-log plot of the number of boxes against the corresponding box size. Functioning as a measure of the data regularity, the box dimension quantifies the level of bias evidenced in the observed data values. The formula for calculating box dimension is shown as follows:

$$D_c = \lim_{\delta \to 0} \frac{\ln N(A,\delta)}{\ln(1/\delta)} \tag{2}$$

Where D_c is the box dimension, A indicates an arbitrary non-empty bounded subset of R^n space ($\delta>0$), and N (A, δ) denote the minimum number of closed balls with radius δ required to cover A.

2.3. Mathematical statistics method

Polynomial regression model is a regression model that fits data by introducing nonlinear terms through polynomials. It can be used to explore the nonlinear relationship between independent and dependent variables. When forecasting the stock market index, there are several benefits to utilizing a polynomial regression model. This model is a form of regression analysis that introduces nonlinear terms through polynomial functions to better fit data and explore potential nonlinear relationships between independent and dependent variables. In the context of stock market forecasting, where stock prices often behave in a nonlinear manner, a polynomial regression model can more accurately capture and model this behavior than a linear regression model [12]. Furthermore, a polynomial regression model offers flexibility and adaptability when dealing with different types of data distributions. This adaptability enables the model to account for unique features in the stock market index, ultimately improving the accuracy of forecasts and enhancing the model's applicability across various datasets. Additionally, polynomial regression models can identify complex patterns in the

data more effectively, making it a favorable approach for analyzing intricate stock market data. Another benefit of a polynomial regression model is the increased reliability and accuracy of forecasts, even when dealing with noisy or sparse data. As a result of its ability to capture potential nonlinearities in data, the model can more effectively capture underlying trends and fluctuations in the stock market index, ultimately leading to more precise predictions. The basic form of the polynomial regression model is represented as:

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \cdots + \beta_k x^k + \varepsilon \quad (3)$$

In this equation, the dependent variable is represented by y, and the independent variable is represented by x. $\beta_0 \sim \beta_k$ are the regression coefficients, x is raised to the power of 1 to k, and ε represents the error term. The value of k represents the order of the polynomial, which controls the complexity and degree of nonlinearity of the model. For instance, when k = 1, it represents a linear polynomial regression model; when k = 2, it represents a quadratic polynomial regression model; when k = 3, it represents a cubic polynomial regression model, and so forth.

When using polynomial regression to fit and predict stock market index, the order of the polynomial needs to be carefully chosen. Generally, 2-3 order polynomial regression models are preferred over 1st order or higher than 4th order models due to several reasons. Firstly, a 1st order polynomial regression model can only describe a linear relationship between the predictor variable and response variable, which might not be sufficient to capture complex trends and patterns in the stock market index [13]. Secondly, given the dataset size and the significance of the trend, using a polynomial model with a much higher order could lead to over-parameterization, which might lead to over-fitting [14]. A quadratic curve is often a prevalent pattern in stock market index, and a second-order polynomial can capture such changes with accuracy. Moreover, a third-order polynomial has the advantages of fitting the local features of the data and enhancing the model's fitting ability, while maintaining reasonable generalization capability and model simplicity compared to higher-order polynomial models. Therefore, a selection of 2-3 order polynomial regression models facilitates a balance between the accuracy and complexity of stock market index prediction [15]. The final order of the polynomial regression model needs to be determined based on the specific evaluation metrics of the generated model (as shown in section 3).

3. Results And Discussion

First, the parameters for calculating the box dimension have to be determined. It is significant to select an appropriate box size in order to obtain more accurate results. The S&P 500 index was chosen as the research subject, which includes market capitalizations from multiple industries and companies, and is considered to be complex and highly volatile. To determine the most suitable box size, the study chose two sizes for comparison: 5 and 7. These two sizes reflect different data granularity and can affect the calculation results of the box dimension. Figure 2 illustrates the calculation results of the box dimension for the S&P 500 index using box sizes of 5 and 7.

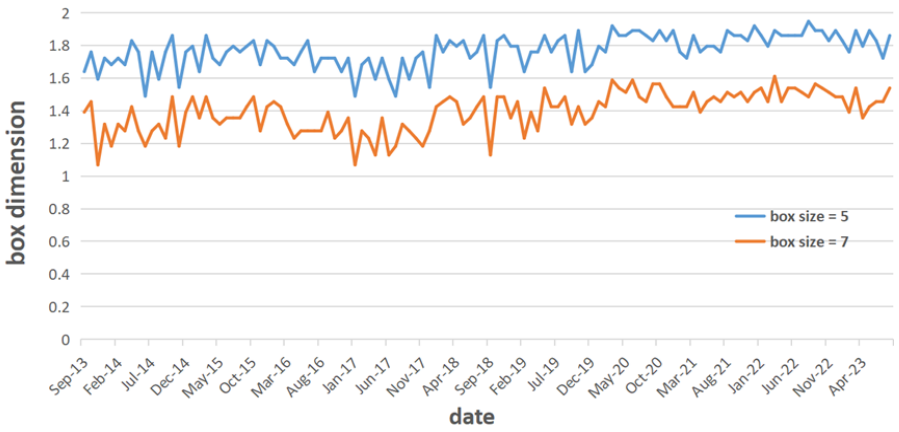

Figure 2. Box dimension with different box sizes

By observing the graph, the following conclusions can be drawn: For a box size of 5, the calculated box dimension ranges from approximately 1.5 to 2. For a box size of 7, the calculated box dimension ranges from approximately 1 to 1.5. Since the S&P 500 index is more complex and volatile as compared to individual stock from a single industry or company as said above, it is inferred that the box dimension of the S&P 500 index should generally range from 1.5 to 2.

Based on the above inference, the research chose a box size of 5 as the parameter for calculating the box dimension of the S&P 500 index. This choice shall provide more accurate results and better reflect the complexity and volatility of the index.

From Figure 2, it can be clearly noticed that the box dimension of the S&P 500 exhibited non-linear distribution patterns and high volatility over time. To capture this non-linear behavior effectively, the study aimed to model and analyze the box dimension using polynomial regression. To determine the optimal degree of the polynomial regression model, the root mean. Generally, higher degree means better fitting accuracy. However, caution should be exercised to strike a balance and avoid overfitting.

In this paper, multiple polynomial regression models with degrees ranging from 1 to 3 were established. To validate the accuracy of the model, the squared error value was calculated. Squared error (RMSE) was employed as an evaluation metric that measures the discrepancy between the observed data and the predicted values from the polynomial fit. Typically, an RMSE value ranging from 0.2 to 0.5 is considered indicative of an effective predictive model. The results indicated that the model achieved a relatively lower RMSE value of 0.08 at the 3th degree, indicating a valid model and a better fit to the original data while minimizing the risk of overfitting.

Consequently, the 3th degree polynomial regression was selected as the optimal fitting model. The equation for this model is:

$$Dimension(x) = 1.734 - 0.003 \times x + 8.130e - 05 \times x^2 - 3.981e - 07 \times x^3 \quad (4)$$

Here, the variable 'x' represents the numerical identifiers assigned to the monthly time-based data in the dataset. Specifically, this paper assigned the number '1' to represent the starting time of the data (September 2013), '2' to represent October 2013, and so forth. Figure 3 illustrates the fitted curve and the observed data.

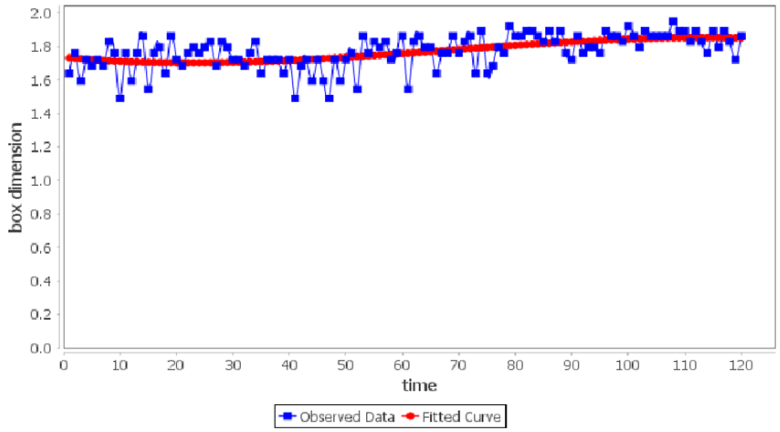

Figure 3. Fitted graph for the observed data

The S&P 500 index monthly box dimension of 2024 was forecasted using the aforementioned predictive model. The results are displayed in Table 1.

Table 1. Predicted results

Date	Predict	Date	Predict
2024-01	1.851	2024-07	1.841
2024-02	1.850	2024-08	1.839
2024-03	1.849	2024-09	1.837
2024-04	1.847	2024-10	1.834
2024-05	1.845	2024-11	1.831
2024-06	1.843	2024-12	1.829

It is manifest from Table 1 that the model gets the conclusion that in the upcoming year, it is anticipated that the box dimension of the S&P 500 index will maintain its relatively high value, centered around 1.8. This indicates a persistent level of market volatility and uncertainty, as well as a complexity in trading behaviors among market participants, leading to the potential for substantial fluctuations and unexpected changes in market prices. However, the S&P 500 index's box-counting dimension appears to be gradually declining, and this suggests that the market's dynamic complexity is also declining. Leveraging this trend, a viable approach would involve implementing a trend-following strategy, where suitable buy or sell operations are considered when the price trend aligns with the descending pattern of the box dimension. By incorporating such an analytical framework, market participants may be equipped to make informed investment decisions in response to the evolving market dynamics.

4. Conclusion

This paper proposed a polynomial regression-based time series forecasting method for stock market prediction, utilizing data from September 2013 to August 2023. The prediction model employed a third-order polynomial, yielding a root RMSE value of 0.08, indicating its efficacy and practical applicability. However, the research process highlighted several limitations and challenges. Specifically, the box dimension of the index demonstrated high volatility and unpredictability, complicating the data prediction process. Improving data forecasting accuracy will require supplementing the proposed approach with additional data analysis models and incorporating comprehensive evaluation metrics beyond the RMSE value. Additionally, the study highlighted the need for an expanded sample size for the box dimension data to enhance the reliability of results. Finally, it is essential to recognize the diverse influences on market price volatility in different stages. Thus, it would be optimal to analyze each stage individually rather than generalizing across the entire dataset.

References

[1] Deng, J. X., "Stock price volatility analysis and forecasting using the multi-fractal theory," Jinan University, (2008).

[2] Wang, Y. J., "Stock forecasting based on fractal theory and machine learning," Henan Polytechnic University, (2018).

[3] Peters, E. E. "Chaos and Order in the Capital Markets," Beijing: Economics Science Press, (1999).

[4] Chen, Y. and Zhou, J., "Review of the theory of fractal research on stock market, " Economic Research Guide 33(251), 115-116 (2014).

[5] Pasquini, M. and Serva, M., "Multi-scaling and clustering of volatility," Physica A 269, 140-147 (1999).

[6] Traina, C., Agma, Jr. and Leejay, T., "Fast feature selection using the fractal dimension," National 62 Science Foundation, (2000).

[7] Tokinaga, S., et al. "Forecasting of time series with fractal geometry by using scale transformations and parameter estimations obtained by the wavelet transform," Electronics and Communications in Japan 80(8), 20-30 (1997).

[8] Alvarez-Ramirez, J., et al. "Time-varying Hurst exponent for US stock markets," Physica A 387(24), 6159-6169 (2008).

[9] Wang, H. C., et al. "Daily data series' complex dynamical patterns in the stock market," Physics Letters A 333, 246-255 (2004).

[10] Lee, J. W., Lee, K. E. and Rikvold, P. A., "Multifractal behavior of the Korean stock-market index KOSPI," Physica A 364, 355-361 (2006).

[11] Yang, L. J., "Fractal characteristics and quantitative strategies of China's stock market under fractal theory," Lanzhou University, (2023).

[12] Yang, G. Y., "Comparative analysis of linear regression methods for forecasting five stocks based on stock correlation," Modern Business 29, 42-45 (2022).

[13] Liu, X., "Research on poisoning attack and defense technology for polynomial regression models," National University of Defense Technology, (2022).

[14] António, A., "Polynomial regression with reduced over-fitting—The PALS technique, " Measurement. Volume, 515-521 (2018).

[15] Asoke, K. N., "Model order selection from noisy polynomial data without using any polynomial coefficients," IEEE Access, (2020).