Stock market change prediction based on multiple linear regression and CNN-LSTM

*Zihao Yan, Huishan Zhang**

Anhui University, Anhui, China

ABSTRACT

The impact of climate change on the economy and society has become increasingly evident with the arrival of the 21st century. This impact has extended beyond the realm of economic development and has also affected the financial sector. In consideration of this, the article develops a set of models following a thorough analysis of data released by the National Bureau of Statistics of China, in order to investigate the relationship between climate change and stock indices. The article employs the Spearman correlation coefficient to examine the intricate connection between climate change and stock indices. This article aims to comprehensively analyze the potential impact of climate change on stock index trajectories through the establishment of multiple linear regression, ridge regression, and LASSO regression models. The objective is to facilitate the formulation of well-founded predictions. Finally, the article anticipates the future prices of oil and gas stocks for the next 100 days through the implementation of a CNN-LSTM model. The integration of Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM) architecture is anticipated to enhance the precision and adaptability of stock price trend forecasts within the context of dynamically evolving environmental conditions.

Keywords: Spearman Correlation; Thermodynamic Profile; Kendall Test; Multiple Linear Regression; LASSO Regression; Ridge Regression; CNN-LSTM

1 INTRODUCTION

In light of the ongoing escalation of global climate modification, escalating emphasis is being placed on the impact of environmental facets on the economy and financial sectors. This trajectory is not limited to the ecological sector, but progressively infiltrates the examination of economics and finance [1]. Severe weather conditions, such as hurricanes, floods and droughts, as well as alterations in air caliber, not only directly influence societal quality of life, but also yield profound economic and financial market implications on a global scale [2].

It is evident that first, severe weather phenomena may precipitate a decline in crop yield, subsequently impacting food prices and shareholder value of associated industries. Investors are required to take into account the potential risks posed by climate change for these sectors and subsequently reconfigure their portfolios [3]. Secondly, deteriorated air quality could engender elevated public health expenditures, which in turn would modify demand within the healthcare sphere and related insurance offerings. Simultaneously, heightened demand could also reverberate on related insurance products, necessitating investors and insurers to evaluate these potential risks and their repercussions on the healthcare and insurance sectors.

In summary, the influence of international climate change on the economy and financial circuits extends beyond environmental considerations to emerge as a significant subject demanding global

scrutiny and exploration. It is exclusively through interdisciplinary research and comprehensive comprehension that finance professionals will be able to understand and assimilate this intricate challenge more acutely.

Extensive foreign literature focuses on the effect of meteorological elements on the stock market, which is mature and covers an extensive time frame. The selected pertinent indicators of chosen weather factors primarily comprise sunlight, temperature, humidity, precipitation, wind velocity, air pressure, cloud cover and severe weather conditions, etc [4].

Saunders (1993) identified a negative correlation between New York weather, specifically cloud cover within Manhattan, and stock returns on the NYSE. A subsequent investigation by Hirshleifer and Shumway (2003) corroborated this negative correlation across 26 national (regional) stock exchanges from 1982-1997, although its implications are not universally transposable across all markets. Keef P S et al. (2002) established no substantial impact of cloud cover in Wellington, New Zealand, on the market capitalization-weighted returns of the New Zealand Stock Exchange stock index. Cao and Wei investigated the impact of temperature on stock markets across eight nations or regions from 1962 to 2001, revealing a significant negative influence on stock index returns associated with emotions triggered by temperature. Low temperatures tend to incite aggression, augmenting risk appetite and stock returns [5]. Conversely, Chang et al. (2008) underscored the significant impact of weather on stock market returns and investors' trading behavior, citing cloud cover's adverse effects on stock price volatility and market depth.

2 MATERIALS AND METHOLDS

2.1 Correlation analysis between environmental factors and stock market performance

2.1.1 Preprocessing of data

Our team meticulously pre-processed the data presented. Initially, we recognized the existence of missing values, subsequently, the outliers were isolated. After a thorough examination of vast literature and data, we discovered that the AQI signifies the air quality index, typically encompassing the quantification and evaluation of the concentration of pollutants in the atmosphere. The value of the AQI is never zero. Within the appendix, we observed that the value of the AQI on 20190122, 20190830, 20210524, and 20210718 was zero, potentially indicating outliers stemming from measurement inconsistencies or personnel statistical errors. Consequently, our team initially disregarded these outliers and then resorted to the nearest interpolation method to substitute the missing values [6]. The nearest interpolation serves to approximate the unknown points in the discrete data. The fundamental principle of this interpolation method is: for any point being estimated, identify the closest known data point, and subsequently assign the value of this point to the closest point.

And using the interpolated data we plot the AQI time series for the last 5 years from 1 January 2019 - 31 December 2023, as shown in Figure 1.

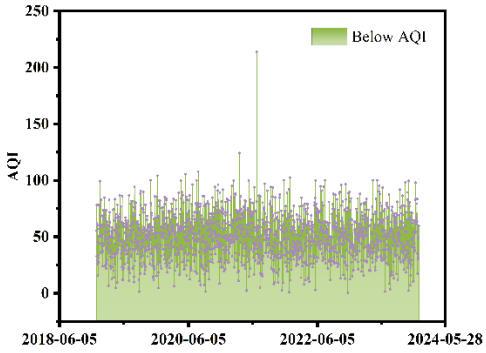


Figure 1: AQI time series for the last 5 years

From the analysis of the above figure, it is found that most of the values of AQI fluctuate between [0,150] and some of the values appear, using the data after interpolation thus making the AQI changes more complete and increasing the ease and accuracy of modelling.

2.1.2 Model building and solving

2.1.2.1 Introduction to the principles of the Spearman correlation coefficient method

The spearman correlation coefficient is a non-parametric measure of correlation that grades the correlation between variables and is used to analyze the correlation between two continuous variables. Since the change in the stock index is a continuous variable over time and since the data does not conform to a normal distribution, Spearman's correlation coefficient is used in this modeling to analyze the correlation between environmental factors and overall stock market performance [7]. X, Y are two independent and identically distributed sets of data and their sample size is N. X, Y are two independent and identically distributed sets of data and their sample size is N. X, Y are two independent and identically distributed sets of data. The number of samples is N. Xi, Yi denotes the ith value of the two sets of random variables, where $i = 1,2,N$ respectively.

Firstly, the X and Y sets are sorted simultaneously in descending or ascending order to obtain two element-ranked sets x, y, where the elements x_i, y_i are the ordering of X and Y in the respective sets. Set d-set as the difference between the ordering of the same bit elements in the X and Y sets, and the computational formula for each element of d-set is as follows.

$$d_i = x_i - y_i \tag{1}$$

Spearman correlation coefficient r_s is calculated as follows:

$$r_s = 1 - 6\frac{\sum_{i=1}^{N} d_i^2}{N(N^2-1)} \tag{2}$$

2.1.2.2 Analysis of correlation thermodynamic diagrams

A heat map is a chart that represents the values in a matrix by color coding, while a correlation row heat map is a special type of heat map that is used to visualize the correlation between the variables, the following figure shows the correlation heat map for the above results as shown in Figure 2, Also since extreme weather events are discrete data, we removed this factor from our processing [8].

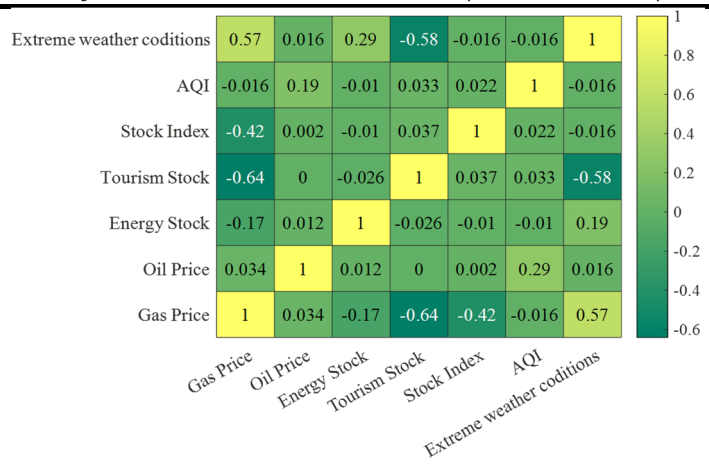


Figure 2: Heat map of correlation coefficients

As discernible from the aforementioned figure:

1) Only extreme weather anomalies and other variables exhibit a correlation coefficient greater than 0.3, indicating a positive association, suggesting that there exists some positive correlation between them, albeit of non-robust magnitude.
2) Among these, the correlation coefficient between extreme weather anomalies and extreme weather phenomena is 0.7, signifying a robust correlation per se.
3) Policy shift displays a correlation coefficient of approximately 0.3 with all other variables except itself, indicating a favorable correlation.
4) The extreme weather conditions are negatively correlated with the stock index, tourism index, energy stocks, and natural gas prices all, positively correlated with oil prices, and most negatively correlated with tourism stocks, which is consistent with the actual situation, extreme weather factors affect people's traveling, which has the greatest impact on the tourism industry, and the oil and gas storage is difficult to transport due to the impact of the extreme weather, leading to price increases.

The AQI and the stocks as a whole all show a positive correlation, and AQI represents how good or bad the air quality is.Therefore, our team concludes that environmental factors have a strong correlation with overall stock market performance.

2.2 The impact of extreme weather on specific stock markets

2.2.1 Establishment and Resolution of Multiple Linear Regression Model

The multiple linear regression analysis is a statistical methodology for investigating the relationship between random variables, by examining the empirical observations of the variables, calculating the establishment of a quantitative relationship between one variable and another, i.e., the regression equation. After the statistical examination that ascertain the significance of the regression effect, it can be utilized to ascertain the degree of influence [9].

Let the random variable y be associated with the variable $x_1, x_2 \cdots x_m$, then its m-element linear regression model is:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_m x_m + \varepsilon \tag{3}$$

Where ε denotes the random error exhibiting the normal distribution N(0, σ^2), $\beta_0, \beta_1, \ldots \beta_m$ signifies the regression coefficient. We designate y_{SI} as the stock index, y_{TS} the tourism stock, y_{ES} the energy stock, and x_1, x_2, x_3 represents the extreme weather, the AQI measure, and the quantity of

climate change news, respectively, to formulate the multi-factor linear regression equation, and utilize the least squares methodology to resolve the system of multi-factor linear regression equations, as per the following procedure:

$$x = \begin{bmatrix} 1 & x_{11} & \cdots & x_{1m} \\ \vdots & x_{21} & \cdots & x_{2m} \\ \vdots & \vdots & \cdots & \vdots \\ 1 & x_{n1} & \cdots & x_{nm} \end{bmatrix} \tag{4}$$

The above equation can be expressed as:

$$\begin{cases} Y = X\beta + \varepsilon \\ \varepsilon \sim N(0, \sigma^2 E_n) \end{cases} \tag{5}$$

$$Q = \sum_{i=1}^{n} \varepsilon^2 \tag{6}$$

Where E_n denotes the unit matrix of order n:

$$Q = \sum_{i=1}^{N} \varepsilon^2 = \sum_{i=1}^{N} (y_2 - -\beta_0 - \beta_1 x_{21} - \cdots - \beta_m x_{2m}) = 0 \tag{7}$$

To do this, let $\frac{\partial Q}{\partial \beta_0} = 0, J = 1,2.\ldots.m.$

$$\begin{cases} \dfrac{\partial Q}{\partial \beta_0} = -2\sum_{i=1}^{n} (y_2 - \beta_0 - \beta_1 x_{21} - \cdots - \beta_m x_{2m}) \\ \dfrac{\partial Q}{\partial \beta_0} = -2\sum_{i=1}^{n} (y_2 - \beta_0 - \beta_1 x_{21} - \cdots - \beta_m x_{2m}) x_{ij} = 0 \end{cases} \tag{8}$$

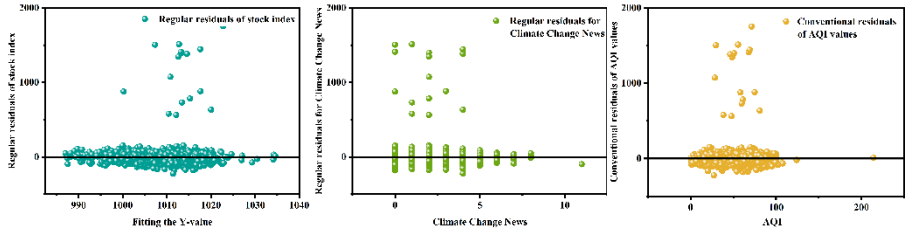
Thus after collation, the matrix form of the regular system of equations is $X^T X\beta = X^T Y$.When the matrix x columns are full rank, X^T is an invertible square matrix, giving $\hat{\beta} = (X^T X)^T X^T Y$.Substituting $\hat{y} = [b_0, \quad b_1, \cdots b_m]$ into Eq. y is obtained as an estimate of y:

$$\hat{y} = b_0 + b_1 x_1 + \cdots + b_m x_m \tag{9}$$

Where the fitting error $e = y - y$ is called the residual and is used as an estimate of the random error, so our team used Origin software to solve the above linear regression equation:

$$\begin{aligned} y_{SI} &= 102.8 - 117x_1 + 0.13x_2 + 1.19x_3, R^2 = 0.00219 \\ y_{TS} &= 99.61 - 0.56x_1 + 0.028x_2 - 0.094x_3, R^2 = 0.00265 \\ y_{ES} &= 99.02 + 0.17x_1 + 0.0121x_2 - 0.0035x_3, R^2 = 0.00041 \end{aligned} \tag{10}$$

At the same time we plot the residuals of the independent variables for the three equations and the overall factor, as shown in Figure 3 below:



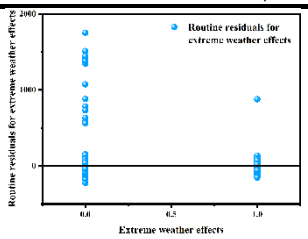


Figure 3: Dependent product of residuals of the independent variables

Upon scrutiny of the aforementioned figure, it becomes discernable that the amalgamation of the fit R^2 and the independent variable Residual plot ascertained that the application of multiple linear regression yields a superior fitting effect, albeit substantial deviation exists in the actual degree of influence, thus, our team elected to enhance the conventional linear regression equations, employing LASSO regression to fit the equation [10].

2.3 Industry-specific stock price forecast movements

In this segment, we employ a CNN-LSTM deep learning architecture to prognosticate prices for oil and gas. The inputs to the system encompass various potential influencers such as AQI, severe weather, etc., along with historical information on oil and gas itself. The layout of the paradigm is illustrated in Figure 11. CNN is employed to capture the interrelation between features, LSTM is utilized to capture the features on the time series, and the output is a 2D vector of the estimated duration, which finally transits through a multi-layered linear layer, which is implemented to enhance the complexity of the model and facilitate a superior fit.

2.3.1 Establishment of CNN-LSTM model

The CNN-LSTM time series prediction model represents an advanced model integrating Convolutional Neural Network (CNN) and Long Short-Term Memory (LSTM). This model is primarily utilized to manage time series data and possesses the capability to capture spatio-temporal elements within the sequence for prediction.

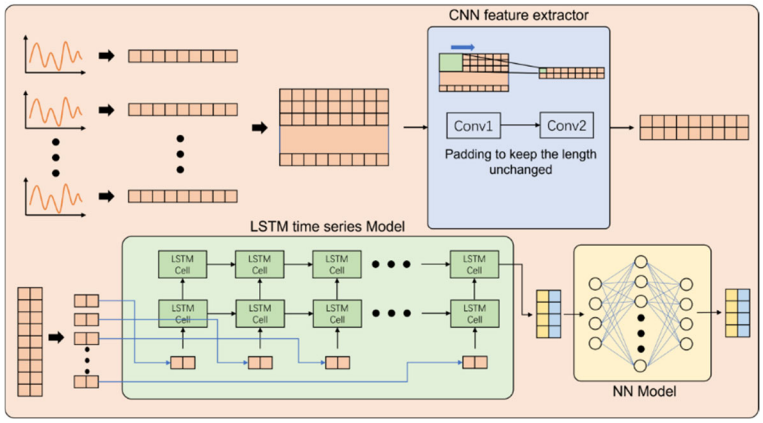


Figure 4: CNN-LSTM network

2.3.2 Training of the model

The training parameters are set: the number of training rounds is 3000, the Adam optimizer is used, the initial learning rate is 0.001, and the loss function is the average MSE. and we normalise the data before training. The training process is visualized as shown in Figure 5 and finally trained to convergence. The performance on the test set is shown in Table 1.

Table 1: Test set training results

RMSE	0.2116
MSE	0.04475
MAE	0.1511
R^2	0.9989

The results show that the validation error of the model is small and the accuracy of the prediction is high enough to be used as a predictive model for oil and gas.

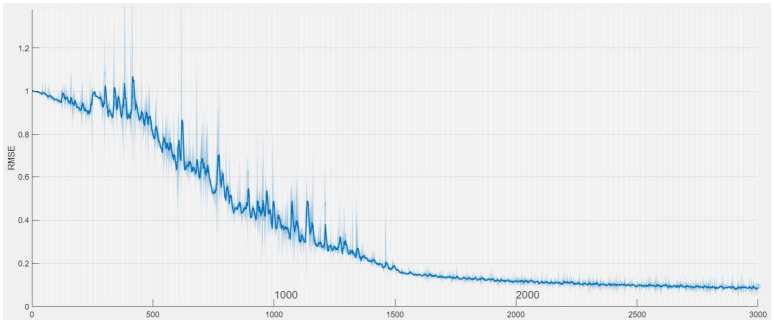


Figure 5: Diagram of the training process

2.5.3 Predictions for solving the model

The fitted curves of the model and the forecast results for the next 100 days (including oil and gas prices) are shown in Figure 6, from which it can be seen that oil and gas prices will fluctuate steadily over the next 100 days.

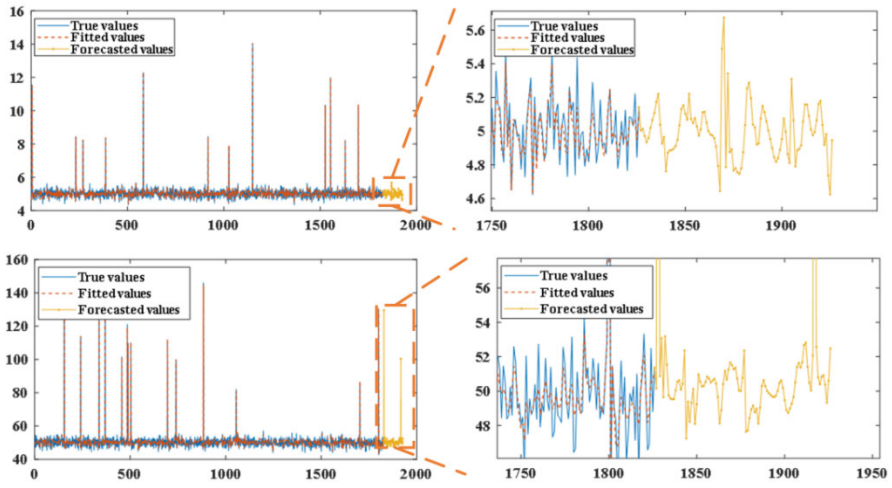


Figure 6: Price projections for oil and gas

3 CONCLUSION

This study elucidates the profound influence of climate change on financial markets and economic growth in the 21st century. Specifically, the effects of environmental elements on the stock market are scrutinized, their correlations examined, and thermodynamic correlation charts developed. Multiple linear regression and LASSO regression are utilized to examine the connection between stock indices, tourism stocks, energy stocks, weather events, AQI levels, and climate change news. Finally, the predicted trajectory of natural gas and oil stock indices over the next 100 days utilizing CNN-LSTM networks is illustrated.

REFERENCES

[1] Saunders, E. M. (1993). Stock prices and Wall Street weather. The American Econo mic Review, 83(5), 1337-1345.

[2] Hirshleifer, D., & Shumway, T. (2003). Good day sunshine: Stock returns and the weather. The journal of Finance, 58(3), 1009-1032.

[3] Keef, S. P., & Roush, M. L. (2002). The weather and stock returns in New Zealan d. Quarterly Journal of Business and Economics, 61-79.

[4] Cao, M., & Wei, J. (2005). Stock market returns: A note on temperature anomaly. Journal of Banking & Finance, 29(6), 1559-1573.

[5] Chang, S. C., Chen, S. S., Chou, R. K., & Lin, Y. H. (2008). Weather and intraday patterns in stock returns and trading activity. Journal of Banking & Finance, 32(9), 1754-1766.

[6] Li Hongmei. Image scaling technique based on interpolation algorithm[J]. Journal of Xinxiang College,2017,34(03):31-33.

[7] Xu, K., Yin, H., Chen, M., & Zhang, J. (2023). A robust permutation test for Kend all's tau. Journal of Statistical Computation and Simulation, 1-21.

[8] Li, Y., He, X., & Liu, X. (2023). Fuzzy multiple linear least squares regression anal ysis. Fuzzy Sets and Systems, 459, 118-143.

[9] Levy, J., Brunner, M., Keller, U., & Fischbach, A. (2023). How sensitive are the ev aluations of a school's effectiveness to the selection of covariates in the applied va lue-added model?. Educational Assessment, Evaluation and Accountability, 35(1), 1 29-164.

[10] Li, Q., Guan, X., & Liu, J. (2023). A CNN-LSTM framework for flight delay predi ction. Expert Systems with Applications, 227, 120287.