Scholar Publishing Group
International Journal of Big Data Intelligent Technology
https://doi.org/10.38007/IJBDIT.2026.070107
ISSN 2790-0932 Vol. 7, Issue 1: 52-59

Research on Financial Time Series Prediction and Multiscale Correlation Based on the Fusion of Network Big Data and Deep Learning

Jingzhi Yin

Department of Mathematics, Columbia University, New York 10017, New York, United States

Keywords: Network Big Data, Deep Learning, Financial Time Series Prediction, Multi-Scale Correlation, Complexity Analysis

Abstract: Due to the nonlinearity, high volatility, and multi factor driving characteristics of financial time series, the fitting ability of traditional prediction models is limited. This study focuses on financial time series prediction and multi-scale correlation analysis by integrating network big data and deep learning. The MF-DCCA method is used to verify the correlation between the RMB exchange rate and Baidu Index, WTI oil price and Google Trends. A WOA-STL-LSTM optimization model is constructed by integrating network search data as external features, significantly improving the accuracy of exchange rate and oil price prediction. Study the use of Hurst exponent, fractal dimension, Lyapunov exponent, and transfer entropy to quantify data complexity, revealing the marginal effect and keyword difference mechanism of Google Trends on WTI prediction. To address the issue of small sample size, fractal interpolation and linear interpolation are used to enhance data granularity. The adaptation strategy is validated by combining LSTM, GRU, and Bi LSTM models. It is found that fractal interpolation is superior to linear interpolation in simulating complex fluctuations, and the model selection needs to match data characteristics - Bi LSTM/LSTM is better in small sample scenarios, and the combination of GRU and fractal interpolation, Bi LSTM and linear interpolation after data augmentation yields the best results. Research provides an interpretable framework for financial time series forecasting, promoting the development towards precision and transparency.

1. Introduction

Time series [1], as a collection of statistical indicators arranged in chronological order, widely exists in fields such as finance, meteorology, population, and network traffic. Its core value lies in mining potential patterns from massive data to assist decision-making. However, financial time series face significant challenges due to their complex characteristics such as nonlinearity, time-varying nature, and high noise. Traditional linear analysis methods such as Pearson correlation coefficient [2] and Granger causality test are difficult to characterize the dynamic interaction patterns of non-stationary sequences; Although classical prediction models such as ARIMA and GARCH perform well in predicting stationary sequences, their adaptability is limited when facing nonlinear, multivariate, and complex scenarios; Although machine learning methods can capture nonlinear features, they often rely on artificial feature engineering, which limits their efficiency and generalization ability. With the development of the Internet and artificial intelligence technology, online big data and in-depth learning provide a new paradigm for financial time series research - online search behavior data [3] (such as Google Trends) reflects the trend information of real activities by quantifying users' attention to specific keywords, which has been proved to assist in influenza prediction, stock return analysis and industry sales prediction; Deep learning automatically extracts deep features from data through a multi-layer network structure, demonstrating stronger nonlinear fitting and long-term dependency modeling capabilities in time series prediction than traditional machine learning. Especially, recurrent neural networks represented by LSTM and GRU, as well as their improved models, effectively alleviate the problem of gradient vanishing and become the preferred tool for financial prediction. This study focuses on the research of financial time series prediction and multi-scale correlation based on the fusion of network big data and deep learning, aiming to address the following key issues: exploring the dynamic correlation mechanism between network big data (such as search engine trends) and financial time series, and quantifying cross effects through multi-scale correlation analysis (such as MF-DCCA); Integrating deep learning models with data augmentation techniques (such as interpolation algorithms) to improve the accuracy and interpretability of financial time series forecasting; Combining complexity analysis theories such as Hurst exponent, fractal dimension, and transfer entropy, explain the intrinsic driving logic of prediction results from dimensions such as long-term memory, chaotic characteristics, and information flow direction. Its innovative contributions are reflected in three aspects: in terms of method application, MF-DCCA is used as a prior tool to screen the influencing factors of network big data, and a hybrid prediction model (WOA-STL-LSTM) is constructed by combining whale optimization algorithm [4], STL decomposition, and LSTM to achieve end-to-end optimized prediction; In terms of research approach, the complexity theory is used to link the predicted results with dynamic causal mechanisms, enhancing the interpretability of the results; In terms of technical means, the system explores the matching mechanism between data augmentation methods such as linear interpolation and fractal interpolation and deep learning models to improve the model's generalization ability.

2. Correlation theory

2.1 Time series decomposition and multi-scale correlation analysis methods

The STL algorithm is based on the LOESS smoother to decompose the time series into trend component (Tt), seasonal component (St), and residual component (Rt). It is implemented through an iterative process of inner and outer loops: in the inner loop, the original sequence is first de trended, and then the seasonal component is updated through periodic subsequence smoothing, low-pass filtering, and seasonal component de trending steps; The outer loop uses the results of the inner loop to calculate the residual component, and reduces the influence of outliers by adjusting the weight of outliers, ultimately satisfying Xt=St+Tt+Rt. This method has strong robustness to outliers and is easy to operate, making it suitable for most time series decomposition scenarios. MF-DCCA[5] was proposed by Zhou in 2014 for analyzing the high-dimensional multifractal behavior of two time series. The process consists of five steps: centralizing the mean of the original sequence and reconstructing it into integral form; Divide into non overlapping subsequences of equal length (including repeated use of end data); Fit the local trend function using the least squares method and calculate the deviation covariance; Calculate the q-order mean ($q \neq 0$) or logarithmic mean (q=0) of the local covariance of 2N_2 windows to obtain the q-order wave function Fq (s); By analyzing the power-law characteristics through the linear relationship between log F_q (s) and log (s), the slope is the generalized Hurst exponent H_xy (q). When H_xy (q) varies with q, it indicates that the cross-correlation between the two sequences has multifractal characteristics; A value greater than 0.5 indicates positive long-range correlation (same increase or decrease), a value less than 0.5 indicates negative long-range correlation (reverse), and a value equal to 0.5 indicates no long-range correlation. This method can effectively reveal multifractal behavior and dynamic interaction patterns at different time scales.

2.2 Bi LSTM mechanism and evolution

LSTM was proposed by Schmidhuber in 1997 and later improved and extended by Alex Graves. It is a classic model for time series prediction. Its core structure includes a forget gate, input gate, output gate, and memory unit: the forget gate determines the information to be discarded in the memory unit at the previous moment through the sigmoid function; The input gate combines sigmoid and tanh layers to determine the information that needs to be updated at the current time and generate candidate states; The output gate controls the output of the memory unit state, and ultimately achieves selective information transmission and long-term dependency modeling through a gating mechanism. Bi LSTM was proposed by Schuster and Paliwal, which utilizes both past and future information through a bidirectional LSTM structure - forward LSTM processes the information from front to back of the sequence, and backward LSTM processes the information from back to front. The two outputs together form the final prediction result, enhancing the ability to understand the global features of the sequence. GRU simplifies the structure based on LSTM, integrates the forget gate and input gate into an update gate, and introduces a reset gate to control the degree of preservation of historical information: the update gate determines the inheritance ratio of the current state to the historical state, and the reset gate adjusts the influence of the historical state on the current candidate state, achieving faster convergence speed while maintaining comparable performance to LSTM, suitable for scenarios that require computational efficiency. All three effectively alleviate the problem of gradient vanishing through gating mechanisms, becoming core tools in the field of time series prediction.

3. Research method

3.1 Time series prediction methods and evaluation framework

In terms of interpolation algorithms, linear interpolation determines the interpolation point through a linear relationship between two points, which is simple to calculate and highly similar to the original data distribution. It is suitable for filling missing values or increasing data intervals; Fractal interpolation was proposed by Barnsley and is based on an iterative function system to capture local wave characteristics, resulting in higher accuracy in complex scenes. The optimization algorithm adopts WOA (Whale Optimization Algorithm), proposed by Mirjalili in 2016, which simulates the hunting behavior of humpback whales in bubble nets. It achieves global and local search through contraction, spiral ascent, and random learning strategies, and is suitable for hyperparameter optimization. It has the characteristics of simple operation and avoiding local optima. The model evaluation uses MAE (Mean Absolute Error), RMSE (Root Mean Square Error), MAPE (Mean Absolute Percentage Error), and R^2 (Coefficient of Determination) to measure absolute error, root mean square error, relative deviation, and goodness of fit, respectively. The smaller or closer the value is to 1, the better the performance. The experimental environment configuration includes Windows 10 operating system, Intel Core CPU16GB memory, development tool VS Code/Jupyter notebook, programming language Python 3.7.8, deep learning framework using TensorFlow 2.1 and Keras 2.4.3, data processing relying on Pandas 1.0.1, NumPy 1.18.1, and complexity analysis toolkits nolds 0.5.2 and hfda 0.1.1.

3.2 Correlation analysis and prediction framework between online search data and financial time series

This paragraph focuses on the impact mechanism and predictive value of online search data as an external factor on financial time series. Firstly, the MF-DCCA method is used to quantify the cross correlation between the RMB exchange rate and online search data (comprehensive keyword search rate) as a prior indicator for feature selection. The experimental dataset covers daily frequency data from October 2015 to October 2020, with a total of 1311 samples. The network search data was obtained through a custom crawler program and transformed into search rates through logarithmic processing

$$index = \frac{index_1 + index_2 + index_3 + index_4}{4} \tag{1}$$

$$r_t = \log(Close_{t+1}) - \log(Close_t) + 1 \tag{2}$$

$$r_v^t = \log(index_{t+1}) - \log(index_t) + 1 \tag{3}$$

Descriptive statistics show that the standard deviation of exchange rate returns is small (0.22892) and the volatility is stable; The standard deviation of search rate is relatively large (6944.6340) and fluctuates violently. The Jarque Bela test rejects the assumption of normal distribution at the 1% significance level, confirming the non normal nature of the data. Further construction of WOA-STL-LSTM model - optimizing LSTM hyperparameters through Whale Optimization Algorithm (WOA), combining STL decomposition to split the sequence into trend, season, and residual components, and finally integrating prediction. This framework validates the effectiveness of network search data through cross correlation testing and compares its performance with other models such as ARIMA and ANN. Empirical evidence shows that network search data can significantly improve the accuracy of financial time series forecasting.

3.3 Correlation analysis and prediction between network big data and financial time series

This paragraph focuses on the cross correlation analysis and prediction framework construction between online search data and financial time series. The MF-DCCA method was used to quantify the cross correlation between the Chinese yuan exchange rate and the comprehensive Baidu index (obtained by averaging the search rates of the four keywords "Chinese yuan exchange rate", "Chinese yuan", "foreign exchange", and "exchange rate") in daily frequency data from October 2015 to October 2020. The generalized Hurst index H_xy (q) was used to verify the multifractal characteristics - short-term (s<137 days) H_xy (2)=0.2497, long-term (s>137 days) H_xy (2)=0.3944, both less than 0.5, indicating significant anti sustained cross correlation, and short-term multifractality (△ H=0.2592) is stronger than long-term (△ H=0.1263). Experimental construction of WOA-STL-LSTM decomposition ensemble model, combined with Whale Optimization Algorithm (WOA) to optimize LSTM hyperparameters. The sequence is decomposed into trend, season, and residual components through STL decomposition, and then predicted and summed separately. Compared with ARIMA (1,1,0), ANN (3 layers, 60-90-30 neurons), LSTM (2 layers, 64-32 neurons, time step 16) and other models, WOA-STL-LSTM (BI) performs the best in MAE (0.012937), RMSE (0.01772), MAPE (0.001858), and R^2 (0.974617) indicators, with an improvement of about 0.4% compared to LSTM (BI) and about 2% compared to ARIMA. DM test shows that at a 95% confidence level, WOA-STL-LSTM (BI) is significantly better than other models, and Baidu Index as an external feature can improve prediction accuracy (LSTM (BI) improves prediction accuracy by about 1% compared to LSTM). The yield prediction experiment further verified that the MAE (0.001825) and RMSE (0.002444) of WOA-STL-LSTM (BI) are still optimal, but the role of Baidu Index in yield prediction is weaker than trend prediction, mainly due to the random walk characteristics of yield and insufficient external data. The overall results indicate that network search data can serve as effective features for financial time series prediction, MF-DCCA can assist in feature screening, and the WOA-STL-LSTM model has significant advantages in nonlinear time series prediction.

4. Results and discussion

4.1 Research on the intrinsic mechanism of network search big data and crude oil price prediction

This paragraph focuses on the relationship and predictive value between internet search big data (Google Trends) and WTI crude oil futures prices. The research framework is divided into two parts: firstly, the predictive ability of Google Trends on WTI is verified through LSTM neural network. Secondly, complexity analysis (Hurst exponent, fractal dimension, Lyapunov exponent) and transmission entropy method are used to analyze its internal mechanism from the perspective of time series complexity and information transmission. Keyword filtering is based on mutual information measurement, selecting the top 4 crude oil related keywords with the highest mutual information from 23: "oil supply", "oil usage", "oil consumption", and "oil refining". Descriptive statistics show that both WTI and Google Trends for various keywords exhibit large standard deviations and non normal distribution characteristics (Jarque Bela test p<0.01). WTI prices fluctuate violently, while Google Trends has a higher frequency of fluctuations but a relatively stable trend (as shown in Table 1).

Table 1 Descriptive Statistics of WTI and Google Trends

Data	Mean	Maximum	Minimum	Standard Deviation	Skewness	Kurtosis	Jarque-Bera
WTI	68.3081	145.2900	16.9400	22.9588	0.4929	2.5857	42.9229***
Oil consumption	25.2712	100.0000	0.0000	10.8024	2.2963	11.8526	3733.9029***
Oil refining	13.5695	100.0000	0.0000	8.3791	3.2744	23.6245	17579.1872***
Oil usage	12.2499	100.0000	0.0000	10.1145	3.2627	20.2388	12754.9796***
Oil supply	28.6670	100.0000	0.0000	12.0564	2.2368	10.3905	2801.8609***

The study confirmed through the LSTM model that Google Trends can improve the accuracy of WTI prediction, and combined with the transfer entropy to quantify the direction of information flow, the complexity index reveals the nonlinear characteristics and interaction laws of the sequence, providing theoretical support for understanding the mechanism of improving crude oil price prediction through network search data.

4.2 Model experiment

This chapter focuses on the impact and internal mechanism analysis of online search big data (Google Trends) on WTI crude oil futures price prediction. The study uses an LSTM neural network model to incorporate Google Trends as an external feature into WTI historical data to enhance price prediction capabilities. The dataset consists of 3 hidden layers (with neuron configuration of 128-64-32), 300 iterations of training, batch size of 32, learning rate of 0.001, and Adam optimizer. It is divided into training and testing sets in an 8:2 ratio. Through 100 repeated prediction experiments (as shown in Table 2),

Table 2 Average Predictive Performance Across 100 Trials

Data Source	MAE	RMSE	MAPE	R²
WTI Only	4.0125	5.1102	0.0832	0.8215
Google Trends + WTI	3.8944	4.9805	0.0814	0.8304
Oil Consumption	4.0174	5.1143	0.0833	0.8212
Oil Refining	4.0022	5.1014	0.0830	0.8221
Oil Usage	3.9606	5.0531	0.0823	0.8255
Oil Supply	3.9406	5.0380	0.0820	0.8265

Complexity analysis reveals that both WTI and Google Trends exhibit nonlinear, chaotic, and long memory characteristics. The Hurst exponent of WTI (0.9132) is significantly higher than that of Google Trends keywords (0.5500-0.6944), while the fractal dimension (1.4473) is lower than that of Google Trends (>1.9), indicating that WTI has stronger long-term memory but more regular local fluctuations. The Lyapunov exponent shows that WTI has two positive values (0.0883, 0.0120), reflecting stronger chaotic behavior, which is consistent with the phenomenon that WTI prediction difficulty is higher than Google Trends.The transmission entropy analysis (shown in Table 3) shows a strong correlation between the direction of information flow and keywords. For example, "oil supply" presents bidirectional transmission (WTI → GT=0.0165, GT → WTI=0.1787), while "oil usage" is mainly unidirectional transmission (GT → WTI=0.1494).

Table 3 Transfer Entropy Between WTI and Google Trends Keywords

Keyword	WTI → GT	GT → WTI
Oil Supply	0.0165	0.1787
Oil Usage	-0.0899	0.1494
Oil Consumption	-0.0065	0.0850
Oil Refining	0.0373	-0.0514

The research conclusion indicates that Google Trends can marginally improve WTI prediction, but the effectiveness is limited by the inherent complexity characteristics of both; Keyword selection directly affects the direction and intensity of information transmission; Google Trends is more valuable in short-term forecasting, while long-term forecasting requires a combination of chaos theory and dynamic analysis methods such as sliding windows. This study provides a mechanistic explanation and practical guidance for the application of online search data in financial time series prediction.

4.3 Effect analysis

This paragraph focuses on the application of interpolation algorithms [6]and deep learning in financial time series prediction, improving prediction performance through data augmentation, and exploring the matching mechanism between the model and interpolation algorithms. The study focuses on five types of financial time-series data (RMB exchange rate, WTI crude oil futures, Shanghai Composite Index, Bitcoin, Twitter Happiness Index) and applies linear interpolation and fractal interpolation algorithms to increase data granularity. LSTM, GRU, Bi LSTM deep learning models are combined for prediction, and fractal dimension is used to quantify data complexity to explain the prediction results. The data covers daily frequency data from June 3, 2019 to June 3, 2021. Missing values are filled in linearly, and the model parameters are uniformly set to 2 hidden layers (64, 32 neurons), 200 iterations, batch size 11, learning rate 0.001, Adam optimizer, time step 7, and training test ratio 8:2.Experiments have shown that fractal interpolation is closer to the complex changes in financial time series by increasing small fluctuations, while linear interpolation smooths fluctuations and reduces nonlinearity. The fractal dimension analysis shows that the original data has a fractal dimension close to 1.5 (random walk state), and the Twitter happiness index reaches 1.7391, indicating higher volatility complexity; After interpolation, the fractal dimension is significantly reduced, reducing short-term fluctuations and potentially improving short-term prediction performance. The prediction results reveal that the prediction accuracy of all five types of data has been improved after data augmentation, and fractal interpolation has better performance than linear interpolation. In terms of model adaptability, Bi LSTM performed the best in three datasets, LSTM in two, and GRU performed the worst due to its structure merging forget gate and input gate, small sample size parameters, and weak nonlinear fitting ability. After interpolation, Bi LSTM performed the best among the four datasets in linear interpolation data (thanks to the bidirectional structure capturing sequence information); GRU leads in 3 datasets and Bi LSTM leads in 2 datasets in fractal interpolation data. The research conclusion emphasizes that interpolation algorithms effectively improve the prediction accuracy of deep learning models by increasing data granularity, and fractal interpolation has significant advantages in simulating complex fluctuations; The matching between models and interpolation algorithms needs to consider data characteristics such as fluctuation level and randomness. It is recommended to use a combination of multiple models to improve prediction robustness. Overall, interpolation algorithms have good applicability for financial time series forecasting, but the selection needs to be based on data characteristics and model advantages to cope with the complexity of financial time series forecasting.

5. Conclusion

Summarize the research work of the entire article and look forward to future research directions. As a complex system, the time series data of financial markets are limited in their predictive ability by traditional models due to their nonlinearity, high volatility, and multi factor driving characteristics. This study systematically explores the improvement path of financial time series prediction by integrating network big data, data augmentation technology, and deep learning models, and deepens the explanation of the prediction mechanism with the help of complexity analysis theory. The core contribution of the research lies in three aspects: firstly, verifying the predictive value of network big data. The MF-DCCA method was used to confirm the correlation between the RMB exchange rate and Baidu Index, WTI crude oil prices and Google Trends. The WOA-STL-LSTM optimization model constructed effectively integrates network search data as external features, significantly improving the accuracy of exchange rate and oil price prediction. This indicates that network big data can be used as an effective feature dimension to supplement traditional time-series data. Secondly, reveal the impact mechanism of data complexity on prediction. Using metrics such as Hurst exponent, fractal dimension, and Lyapunov exponent [7] to quantify data self similarity, chaotic characteristics, and information transmission direction, explain the marginal effects of Google Trends on WTI prediction and the differences in the effects of different keywords, and provide theoretical support for the limitations of prediction. Thirdly, explore strategies for data augmentation and model adaptation. To address the issue of small sample size, fractal interpolation and linear interpolation were used to increase data granularity. Prediction experiments were conducted using LSTM, GRU, and Bi LSTM models[8], and it was found that fractal interpolation was superior to linear interpolation in simulating complex financial time series fluctuations. The model selection needed to match data characteristics - for example, Bi LSTM/LSTM performed better in small sample scenarios, and the combination of GRU and fractal interpolation, Bi LSTM and linear interpolation had the best effect after data augmentation.

Although the research has achieved phased results, there is still room for expansion: in the future, multimodal network data such as social media public opinion and news texts can be included to broaden the sources of features; Attempt to use temporal decomposition methods such as empirical mode decomposition and wavelet transform [9], combined with swarm intelligence optimization algorithms to achieve adaptive hyperparameter selection; Optimize keyword filtering strategies, such as introducing topic models or reinforcement learning for dynamic keyword mining; Deepen the interpretability research of deep learning models, combine attention mechanisms, SHAP values and other methods to reveal feature contribution, and solve the "black box" dilemma [10]. These directions will promote the development of financial time series forecasting towards more accurate and interpretable directions, providing scientific support for risk management, investment decision-making, and other practices.

References

[1] Yuan J, Li J, Hao J. A reliable ensemble forecasting modeling approach for complex time series with distributionally robust optimization. Computers and Operations Research, 2025, 173.

[2] Okwonu F Z, Chiyeaka O M, Ahad N A, et al. ROBUST PEARSON CORRELATION COEFFICIENT FOR IMBALANCED SAMPLE SIZE AND HIGH DIMENSIONAL DATA SET. Science World Journal, 2025, 20(1).

[3] Zheng L, Chai H, Chen X, et al. Search-based Time-aware Graph-enhanced Recommendation with Sequential Behavior Data. ACM Transactions on Recommender Systems, 2024, 2(4).

[4] Gharehbaghi A, Ghasemlounia R, Ahmadi F, et al. Developing a novel hybrid model based on GRU deep neural network and Whale optimization algorithm for precise forecasting of river's streamflow. Scientific Reports, 2025, 15(1).

[5] Xindi Wei. Optimization of Machine Learning Models and Application Supported by Data Engineering. Machine Learning Theory and Practice (2025), Vol. 5, Issue 1: 117-124

[6] Yiting Gu. The Strategic Application of Front-End Technology in The Process of Digital Transformation. Machine Learning Theory and Practice (2025), Vol. 5, Issue 1: 125-132.

[7] Huijie Pan. Design of Data-Driven Social Network Platforms and Optimization of Big Data Analysis. Machine Learning Theory and Practice (2025), Vol. 5, Issue 1: 133-140.

[8] Yixian Jiang. Research on Integration and Optimization Strategies of Cross-platform Machine Learning Services. Machine Learning Theory and Practice (2025), Vol. 5, Issue 1: 141-148.

[9] Shuang Yuan. Integration and Optimization of Network Security Protection Strategies and Vulnerability Detection Technologies. International Journal of Neural Network (2025), Vol. 4, Issue 1: 32-39.

[10] Jiangnan Huang. Application of AI-driven Personalized Recommendation Technology in E-commerce. International Journal of Neural Network (2025), Vol. 4, Issue 1: 40-47.