Hindawi Publishing Corporation
Discrete Dynamics in Nature and Society
Volume 2008, Article ID 526734, 8 pages
doi:10.1155/2008/526734

Research Article

Multivariate Nonlinear Analysis and Prediction of Shanghai Stock Market

Junhai Ma and Lixia Liu

School of Management, Tianjin University, Tianjin 300072, China

Correspondence should be addressed to Junhai Ma, mjhtju@yahoo.com.cn

Received 8 August 2007; Accepted 15 April 2008

Recommended by Masahiro Yabuta

This study attempts to characterize and predict stock returns series in Shanghai stock exchange using the concepts of nonlinear dynamical theory. Surrogate data method of multivariate time series shows that all the stock returns time series exhibit nonlinearity. Multivariate nonlinear prediction methods and univariate nonlinear prediction method, all of which use the concept of phase space reconstruction, are considered. The results indicate that multivariate nonlinear prediction model outperforms univariate nonlinear prediction model, local linear prediction method of multivariate time series outperforms local polynomial prediction method, and BP neural network method. Multivariate nonlinear prediction model is a useful tool for stock price prediction in emerging markets.

1. Introduction

Researchers in economics and finance have been interested in predicting stock price behavior for many years. A variety of forecasting methods have been proposed and implemented. Among them, nonlinear prediction method is a new method developed in the last decades. It is suitable for stock price short-term prediction for that stock market is seen as a nonlinear dynamical system.

An important aspect of nonlinear prediction is to detect nonlinear structures in time series. One of the most commonly applied tests for nonlinearity is surrogate data method of Theiler et al. [1]. In 1994, the method is extended to the multivariate case by Prichard and Theiler [2]. Now the method has been used in many fields such as electrocardiograms (EEG) [3] and finance [4]. Most prediction methods can be grouped into global and local methods. The class of local nonlinear prediction methods is based on next neighbor searches and is introduced by Lorenz [5]. Many introductions to next neighbor techniques have been published. A very simple next neighbor prediction method is proposed by Farmer and Sidorowich [6] in 1987. With the development of multidimensional phase space reconstruction,

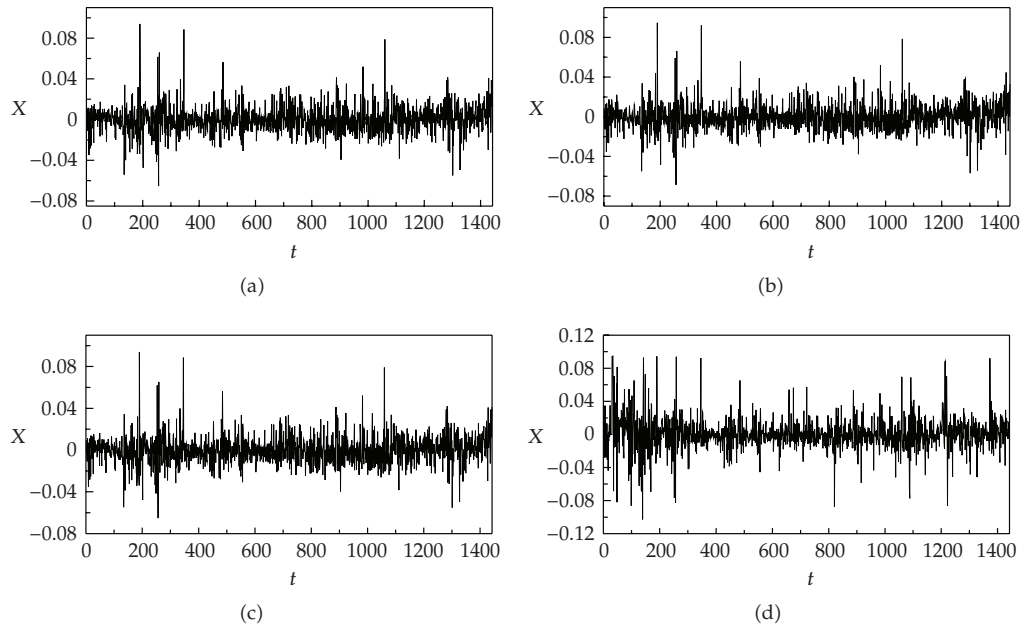

Figure 1: Time series plots of the daily close price time series of Shanghai stock exchange for (a) SSE Composite Index; (b) SSE 180 Index; (c) SSE A Share Index; (d) SSE B Share Index.

the method has been extended for the multivariate time series case and achieved satisfactory results [7, 8].

In this paper, surrogate data method of multivariate time series and multivariate nonlinear prediction method are applied to Shanghai stock market in China. Local nonlinear prediction methods based on next neighbor techniques including linear, polynomial, and BP neural network are proposed and applied to predict stock price behavior. We also compare the accuracy of different prediction methods mentioned in this paper.

The remainder of this paper is organized as follows. Section 2 describes the data examined in this study. We test for nonlinearity in the data by surrogate data method as given in Section 3. Multivariate nonlinear prediction methods based on multidimensional phase space reconstruction are proposed in Section 4. The application of the prediction methods to Lorenz system and Shanghai stock market and the results are also given. Section 5 draws conclusion.

2. Data

Daily stock market index data are sourced from Shanghai stock exchange (SSE) for the time period of 1 January 2001 through 31 December 2006, given a total of 1443 observations. The following stock price indexes are studied: SSE Composite Index (SHCI), SSE Constituent Index (SSE 180 Index) (SHCI1), SSE A Share Index (SHAI), and SSE B Share Index (SHBI). The price series, which is not stationary and contains trends, seasonality, and cycles, is converted into continuously compounded returns series to obtain an accepted stationary series: $x_i = \ln(y_i/y_{i-1})$. The time series of daily close price are given in Figure 1.

3. Testing for nonlinearity using Surrogate data method

The wide-spread and powerful approach to test nonlinearity in observed time series is based on surrogate data method [1]. The idea is to generate many surrogate data sequences from the original data record, preserve its linear properties but destroy some nonlinear structures that possibly exist. In 1996, Paluš [9] proposed an extension to multivariate time series of the nonlinearity test, which combined the redundancy and line redundancy approach with the surrogate data technique. Suppose we have m measured variables $x_1(t), x_2(t), \ldots, x_m(t)$ with zero mean, unit variance, and correlation matrix C. The line redundancy of $x_1(t), x_2(t), \ldots, x_m(t)$ can be defined as

$$L\left(x_1(t);\ldots;x_m(t)\right) = -\frac{1}{2}\sum_{i=1}^{m}\log\left(\sigma_i\right), \tag{3.1}$$

where σ_i are the eigenvalues of the $m \times m$ correlation matrix C. The general redundancy of $x_1(t), x_2(t), \ldots, x_m(t)$ can be defined as

$$\begin{aligned} R_q\left(x_1(t);\ldots;x_m(t)\right) &= H\left(x_1(t)\right) + \cdots + H\left(x_m(t)\right) - H\left(x_1(t);\ldots;x_m(t)\right) \\ &= \frac{1}{2}\sum_{i=1}^{m}\log\left(c_{ii}\right) - \frac{1}{2}\sum_{i=1}^{m}\log\left(\sigma_i\right), \\ H(x,r) &= \frac{1}{q-1}\log_2\left(\frac{1}{N}\sum_{i=1}^{N}\left[\frac{1}{N-1}\sum_{j=1,\,j\neq i}^{N}\theta\left(r-\left\|x_i-x_j\right\|\right)\right]^{q-1}\right), \end{aligned} \tag{3.2}$$

where c_{ii} are the diagonal elements (variances) of the $m \times m$ correlation matrix C, and $\theta(x)$ is the entropy of discrete random variable x. If $x_1(t), x_2(t), \ldots, x_m(t)$ have an m-dimensional Gaussian distribution, $L(x_1(t);\ldots;x_m(t))$ and $R_q(x_1(t);\ldots;x_m(t))$ are theoretically equivalent. The general redundancies detect all dependences in data under study, while the line redundancies are sensitive only to linear structures. Due to stationary, the redundancies do not depend on time t and are the function of the time delays $\tau_1, \ldots, \tau_{m-1}$.

We use general redundancy and line redundancy as the test statistics to compare the original data and the surrogates. The method of generating surrogate data based on phase-randomized Fourier transform algorithm proposed by Prichard and Theiler [2] is used in our paper. Typically, the confidence of rejection is given in terms of significance

$$Z = \frac{\left|D_{\text{orig}} - \left\langle D_{\text{surr}}\right\rangle\right|}{\sigma_{\text{surr}}}, \tag{3.3}$$

where D_{orig} is the test statistic of the original data, $\langle D_{\text{surr}}\rangle$ is the average, and σ_{surr} is the standard deviation of the test statistics of $D_1, D_2, \ldots, D_n$. A $Z > 1.96$ indicates that the null hypothesis is rejected at a level of 0.05, and the time series is nonlinear with probability 95%. If the null hypothesis is not rejected, it means that the time series can be described by a linear Gaussian process with probability 95%.

The surrogate data method is first used to detect nonlinearity of Lorenz system. The Lorenz system is given by a set of nonlinear equations: $\dot{x} = \sigma(y - x)$, $\dot{y} = x(R - z) - y$, $\dot{z} = xy - bz$, where the parameters are chosen as $\sigma = 16$, $R = 45.92$, $b = 4.0$. We create time series of the x, y, and z components by Runge-Kutta algorithm with a time step of 0.001 with

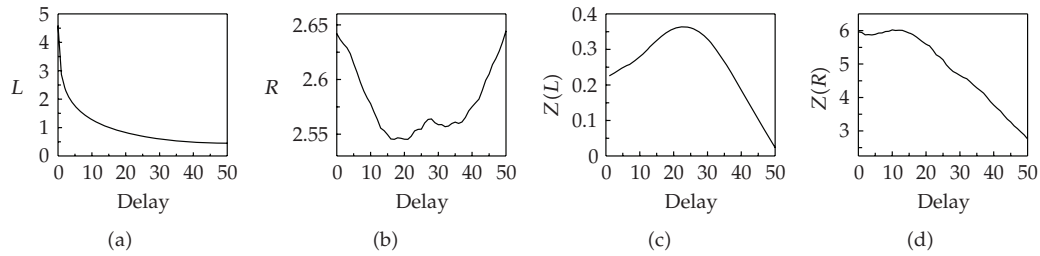

Figure 2: The results of surrogate data for Lorenz system: (a) linear redundancy L; (b) general redundancy R; (c) the test statistic of linear redundancy; (d) the test statistic of general redundancy.

the assumed initial values of $x_0 = 0.1$, $y_0 = 0.2$, and $z_0 = 0.2$. We use 4500 samples of the three variables time series and generate 39 multivariate surrogate data sets. Line redundancy $L(x_1(t), x_2(t+\tau), x_3(t))$ and general redundancy $R(x_1(t), x_2(t+\tau), x_3(t))$ as functions against delay time are presented in Figures 2(a) and 2(b), respectively. Obviously, linear redundancy is different from general redundancy. The test statistic of the line redundancy and the test statistic of the general redundancy as functions against delay time are also presented in Figure 2. While the nonlinear statistic presented in the figures are larger than 1.95 when $r = 0.2$, the null hypothesis is rejected. The nonlinearity is reliably detected; it is also consistent with the fact.

We analyze the nonlinearity of stock returns time series in Shanghai stock market. 39 surrogates are generated using phase randomization algorithm. Line redundancy $L(x_1(t), x_2(t+\tau), x_3(t))$, general redundancy $R(x_1(t), x_2(t+\tau), x_3(t))$, and its test statistics as functions against delay time are presented in Figure 3. Obviously, linear redundancy is different from general redundancy for all 4 stock returns time series. In other words, it shows that the surrogate data is technically good and should not be a source of spurious results in the test. While the nonlinear statistic presented in the figures are larger than 1.95 when $r = 0.1$, the null hypothesis is rejected. The nonlinearity of all 4 stock returns time series is reliably detected.

4. Prediction method based on multivariate time series

In this section, multivariate prediction methods, which extend from univariate prediction [6] proposed by Farmer and Sidorowich, are introduced. The one-dimensional versions will be obtained as a special case. Suppose we have an M-dimensional time series $X_1, X_2, \ldots, X_M$, where $X_i = (x_{i,1}, x_{i,2}, \ldots, x_{i,N})$, $i = 1, 2, \ldots, M$. The embedding of multivariate time series [10] is given as

$$V_n = \left(x_{1,n}, x_{1,n-\tau_1}, \ldots, x_{1,n-(d_1-1)\tau_1}, x_{2,n}, x_{2,n-\tau_2}, \ldots, x_{2,n-(d_2-1)\tau_2}, \ldots, x_{M,n}, x_{M,n-\tau_M}, \ldots, x_{M,n-(d_M-1)\tau_M}\right), \tag{4.1}$$

where $n = \max_{1\le i\le M}(d_i - 1)\tau_i + 1, \ldots, N$, τ_i is the time delays and d_i is the embedding dimensions. In this paper, we use mutual information function [11] to choose the time delays τ_i separately for each scalar time series. The method to get the minimum embedding dimension is based on minimum forecasting error.

Following the delay embedding theorem, if d or d_i is sufficiently large, there exists a map $F : R^d \to R^d$ ($d = \sum_{i=1}^{M} d_i$) such that $V_{n+1} = F(V_n)$. In local prediction method, the change of V_n with time on the attractor is assumed to be the same as that of nearby points V_n^i, $i = 1, 2, \ldots, L$,

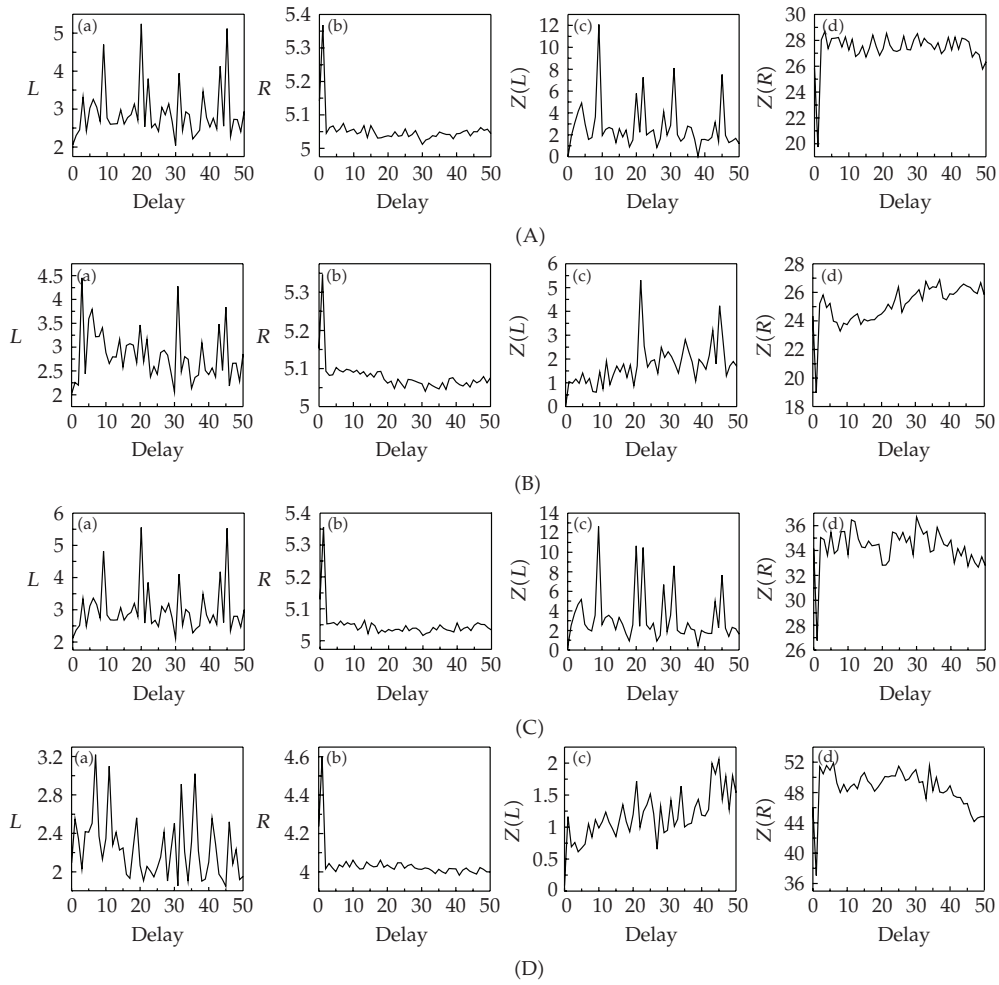

Figure 3: The results of surrogate data of Shanghai stock market for (A) SSE Composite Index; (B) SSE 180 Index; (C) SSE A Share Index; (D) SSE B Share Index: (a) linear redundancy L, (b) general redundancy R, (c) the test statistic of linear redundancy, (d) the test statistic of general redundancy.

where L is the number of neighbor points. In general, for a time series prediction problem, a predictor fits a model to given data and finds an approximate mapping function between the input and the output values. To estimate the kth predictor of vector x_t, $t = 1, 2, \ldots, N$, we employ a set of L nearest neighbors, $x_t^1, x_t^2, \ldots, x_t^L$ of x_t,

$$y_{t+k} = F\left(x_t^1, x_t^2, \ldots, x_t^L\right), \quad k = 1, 2, \ldots, K. \tag{4.2}$$

The real value is defined as $x_{t+k} = y_{t+k} + e_{t+k}$, where e_{t+k} is the prediction error, y_{t+k} is the prediction value. Three prediction models used in this paper are defined as follows.

(1) Line regression prediction of M-dimension inputs (MLP) is defined as

$$y_{t+k} = a_0 + a_1 x_t^1 + \cdots + a_L x_t^L. \tag{4.3}$$

When $M = 1$, it is the line regression prediction of univariate input (ULP).

Table 1: Prediction results for Lorenz system.

Lorenz system	ULP	MLP	MPP	MBP
RMSE	0.0114	4.2046×10^{-6}	0.0142	0.0139
NMSE	1.178×10^{-4}	1.6018×10^{-11}	1.8316×10^{-4}	0.0130

(2) Polynomial prediction of M-dimension inputs (MPP) is defined as

$$y_t^k = a_0 + \sum_{i=1}^{M} b_i x_t^i + \sum_{i=1}^{M} c_i \left(x_t^i\right)^2. \tag{4.4}$$

(3) Back-propagation neural network of M-dimension inputs (MBP) is defined as

$$y_t^k = F\left(x_t^1, x_t^2, \ldots, x_t^L; w_k\right), \tag{4.5}$$

where w_k is a matrix of weights.

To evaluate forecasting performance, the prediction values are compared with actual values according to root mean squared error (RMSE) and normalized mean squared error (NMAE) criteria.

5. Numerical simulation

5.1. Nonlinear prediction of Lorenz system

The proposed method is used to predict Lorenz system mentioned above. The forecast variable here is x. The first 1400 data points are taken as the training data. Then the model that is used to predict the last 100 data points is predicted. The univariate time series prediction is performed at first. The delay time $\tau = 4$ and embedding dimension $m = 5$ are selected based on mutual information function [11] and false nearest neighbor method [7]. The results of one-step prediction with univariate time series (ULP) are given in Table 1.

We combine x, y, and z time series into one multivariate time series to predict the evolvement of variable x. Time delays are found to be 4 for each variable with mutual information functions. We select embedding dimensions $m = 2$ for each variable based on false nearest neighbor method. Three-layer BP neural networks is constructed with 6 input neurons, 15 hidden neurons, and 1 output neurons for Lorenz system. We use RMSE and NMSE to monitor prediction performance of the methods. Table 1 gives the performance measure of three methods (MLP, MPP, and MBP). As it can be seen from Table 1, the application of the MLP took on the smallest RMSE and NMSE, ULP ranked second, followed by MBP and MPP. The results of MLP with more hits of minimal errors seem to give the best performance among the four methods. MLP is more suitable and can be applied to the prediction of Lorenz system.

5.2. Nonlinear prediction for Shanghai stock market

In the following, we apply our prediction algorithms to stock returns series. The forecast variable here is next day's close price. The total number of data points measured in this period

Table 2: Prediction results for stock returns series.

Stock name	ULP		MLP		MPP		MBP	
	RMSE	NMSE	RMSE	NMSE	RMSE	NMSE	RMSE	NMSE
SHCI	0.0139	1.0979	0.0139	1.0952	0.0165	1.5440	0.017	1.6302
SHCI1	0.0150	1.0991	0.0138	0.9375	0.0166	1.3524	0.0175	1.5037
SHAI	0.0140	1.1065	0.0138	1.0744	0.0180	1.8301	0.0147	1.2144
SHBI	0.0170	1.1143	0.0170	1.1109	0.0180	1.2414	0.0192	1.4121

Table 3: Prediction results for the natural logarithms series of stock price.

Stock name	ULP		MLP		MPP		MBP	
	RMSE	NMSE	RMSE	NMSE	RMSE	NMSE	RMSE	NMSE
SHCI	0.0141	0.0134	0.0134	0.0120	0.0138	0.0123	0.0181	0.0219
SHCI1	0.0160	0.0151	0.0154	0.0139	0.0149	0.0130	0.0199	0.0232
SHAI	0.0139	0.0130	0.0138	0.0128	0.0146	0.0143	0.0153	0.0156
SHBI	0.019	0.0251	0.0178	0.0222	0.0169	0.0199	0.0187	0.0243

is 1442. The first 1319 data points are taken as the training data. Then the model is used to predict the last 123 data points.

Firstly, univariate prediction of stock price is performed. We choose the delay time $\tau = 1$ for all four daily close price series based on mutual information function. Simultaneously, we set embedding dimension $m = 6, 9, 6, 9$ for the daily close price time series of SHCI, SHCI1, SHAI, and SHBI, respectively, which is based on false nearest neighbor method. The results of one-step prediction with univariate time series (ULP) are given in Table 2.

We combine close price, open price, high price, and low price time series into one multivariate time series to predict the evolvement of the close price. Time delays are found to be 1 for each variable with mutual information functions. The embedding dimensions are selected as follows: $m = 2, 2, 2, 2$ for the close price, the open price time series, the high price time series, and the low price time series of SHCI; $m = 3, 2, 4, 2$ for SHCI1, $m = 2, 2, 2, 3$ for SHAI, and $m = 2, 2, 2, 4$ for SHBI, respectively. After the reconstruction of phase space, one-step prediction based on multivariate time series is applied to the close price. In the ANN section of model, we also choose three-layer BP neural networks. The BP was constructed with input neurons $a = 8, 11, 9, 10$, hidden neurons $b = 20, 20, 20, 20$, and output neurons $c = 1, 1, 1, 1$ for SHCI, SHCI1, SHAI, and SHBI. Table 2 gives the performance measure of three methods (MLP, MPP, and MBP). As it can be seen from Table 2, the application of the MLP takes on the smallest RMSE and NMSE.

From Tables 1 and 2, we can see that the errors between the observation and the prediction are all very small, which suggest that prediction obtained from multivariate time series is very effective. Thus, the practical application results show the effectiveness of the proposed approaches. Comparison of the errors among models shows that the multivariate model is superior to the univariate model, and local linear prediction method of multivariate time series outperform local polynomial prediction method and BP neural network method.

We also see that the errors NMSE of all stock returns series for all the methods are relatively big, possibly because the values of stock returns are very small. After all the prices were transformed into natural logarithms, we use our methods to predict logarithms series of stock close price. Table 2 gives the performance measure of all the methods (ULP, MLP, MPP,

and MBP). From Table 2, we can see that no significant differences among the errors RMSE of logarithms series of stock close price and those of stock returns, but the errors NMSE of logarithms series of stock close price are far smaller than those of stock returns, which suggest that the forecast results of stock price itself or its natural logarithm sequence are more ideal.

6. Conclusion

In this work, multivariate time series of stock returns in Shanghai stock exchange has been analyzed in order to discover whether a nonlinear dynamical approach can provide better predictions than others. Several kinds of analysis have been conducted on the data. Surrogate data method of multivariate time series provides evidence of the presence of a nonlinear deterministic component in the dynamics considered. Predictions are obtained approximating the nonlinear dynamics by linear autoregressive, polynomials autoregressive, and BP neural network, in the context of the nearest neighbor method. One-step prediction is performed. The predictive results are, on the whole, satisfactory, regarding the comparison between the observed and predicted time series. Comparison of the errors among models shows that the prediction quality of multivariate time series approximating the nonlinear dynamics by linear autoregressive outperforms that of univariate time series, and local linear prediction method of multivariate time series outperforms local polynomial prediction method and BP neural network method. The forecast results of stock price itself or its natural logarithm sequence are more ideal. It is also conjectured that stock price time series could be modeled and predicted better by the dynamical systems approach.

References

[1] J. Theiler, S. Eubank, A. Longtin, B. Galdrikian, and J. D. Farmer, "Testing for nonlinearity in time series: the method of surrogate data," *Physica D*, vol. 58, no. 1–4, pp. 77–94, 1992.
[2] D. Prichard and J. Theiler, "Generating surrogate data for time series with several simultaneously measured variables," *Physical Review Letters*, vol. 73, no. 7, pp. 951–954, 1994.
[3] S. A. R. B. Rombouts, R. W. M. Keunen, and C. J. Stam, "Investigation of nonlinear structure in multichannel EEG," *Physics Letters A*, vol. 202, no. 5-6, pp. 352–358, 1995.
[4] H. Wang and L. Tang, "Testing for nonlinearity in Shanghai stock market," *International Journal of Modern Physics B*, vol. 18, no. 17–19, pp. 2720–2724, 2004.
[5] E. Lorenz, "Atmospheric predictability as revealed by naturally occurring analogues," *Journal of the Atmospheric Sciences*, vol. 26, no. 4, pp. 636–641, 1969.
[6] J. D. Farmer and J. J. Sidorowich, "Predicting chaotic time series," *Physical Review Letters*, vol. 59, no. 8, pp. 845–848, 1987.
[7] A. Porporato and L. Ridolfi, "Multivariate nonlinear prediction of river flows," *Journal of Hydrology*, vol. 248, no. 1–4, pp. 109–122, 2001.
[8] K. Koçak, L. Şaylan, and J. Eitzinger, "Nonlinear prediction of near-surface temperature via univariate and multivariate time series embedding," *Ecological Modeling*, vol. 173, no. 1, pp. 1–7, 2004.
[9] M. Paluš, "Detecting nonlinearity in multivariate time series," *Physics Letters A*, vol. 213, no. 3-4, pp. 138–147, 1996.
[10] C. Liangyue, M. Alistair, and J. Kevin, "Dynamics from multivariate time series," *Physica D*, vol. 121, no. 1-2, pp. 75–88, 1998.
[11] M. Kennel, R. Brown, and H. Abarbanel, "Determining embedding dimension for phase-space reconstruction using a geometrical construction," *Physical Review A*, vol. 45, no. 6, pp. 3403–3411, 1992.