Predictability of Stock Returns, Using Genetic Programming and Chaos Theory

Andrzej Małolepszy

Technical University of Lodz, Institute of Computer Science, Lodz, Poland
e-mail: malolep@ics.p.lodz.pl

Abstract The paper investigates possibilities of prediction of the Warsaw stock indexes returns, based on the chaos theory. Behavior of the time series, which represent the financial data, is very chaotic. In this paper, the Hurst exponent (H) was used to classify series of financial data, representing different indexes from the Warsaw Stock Exchange. The Hurst exponent may be treated as a measure of predictability. The series with H value greater than 0.5, can be predicted more accurately than series with H value close to 0.5. Then, for the selected time series it were reconstructed a higher-dimensional attractor. The values of the delay and the embedding dimension, which are needed to construct the attractor were obtained from auto-mutual information and from the false near neighbors' method. The time series which contain daily return of indexes were transformed according to a higher-dimensional attractor and prepared as data for a problem of symbolic regression. Genetic programming (GP) was used to generate the functions which match a target curve.

1 Introduction

A time series is a series of observations taken at regular time intervals. For the financial markets the values of the indexes or the prices of shares are the base for the time series. Observations of these data can be made with the different frequency: every hour, every day, every week or every month, yielding different time series for the same dynamic system. A question, which appears first, is whether a given time series is predictable. If the time series is fully random, there is no method to predict the future.

The Hurst exponent, proposed in [3], is used in fractal analysis and becomes very popular in determining the character of economic and financial time series. The Hurst exponent provides information about long-term memory and fractal nature of time series [6,7]. The values of H below 0.5 indicate an anti-persistent series in which the values are mean reverting, H = 0.5 indicates a random series, H above 0.5 indicates a persistent series, the trend being reinforced. The Hurst exponents were calculated with the use of the rescaled range analysis (R/S analysis) for selected time series from the Polish financial market. That allowed choosing a few time series with high H ($0.5<H<=1$) values and time series with H close to 0.5, for comparison of obtained results.

Another problem is the extent to which given scalar time series $x_1, x_2, ..., x_k$ reconstruct the original multidimensional attractor. The simplest method is to consider the vector $\boldsymbol{X}(t)=(x(t), x(t+\tau), ..., x(t+(n-1)\tau))$, which corresponds to the n-dimensional coordinate. It is important to select proper values for n and τ variables. If the delay τ is too small $\boldsymbol{X}(t)$, is very similar to $\boldsymbol{X}(t+\tau)$, when τ is too long, vital information can be lost. The value of τ can be calculated from auto-mutual information, i.e., the method proposed in [2]. The embedding dimension n can be found with the use of the false near neighbors' method [4].

As shown in [1], the time series with chaotic behavior can not be estimated by a linear model like AR, ARMA. None of the linear models has been able to learn the dynamics of this time series. On the other hand, satisfying results of the short-term prediction were obtained with help of the partial recurrent artificial neural network.

Genetic programming (GP) is a search technique [5], introduced by Koza. For forecasters the possibility of symbolic regression provided by GP is especially useful in predicting future values of time series. This is a numerical optimization tool to select model which best matches the time series. In this study GP was used to find the best models for the chosen (H>0.5) financial time series and for the series with H close to 0.5.

This paper is organized as follows: In the next sections below, there is a review of the R/S analysis and reconstructing attractors. It was then applied to measure the predictability of chosen financial time series and to find embedding space, where it can construct an attractor from scalar data that preserves the invariant characteristics of an originally unknown attractor. A conclusion is in the final section.

The auto-mutual information was calculated with use TSTOOL version 1.11. Genetic programming was performed with support lil-gp version 1.01.

2 Reconstruction of State Space

Hurst exponent – R/S analysis

The existence of long-term memory in a time series can be examined with the use of the analysis of rescaled range (R/S analysis). It allows identifying non-random behavior in these series. Based on [7], the Hurst exponent for time series x_t, $t = 1, 2, ...,n$ can be calculated by the R/S analysis as follows:

Share the series on d subseries and for every subseries $m = 1 ... d$ executes:

1. Calculate the mean value $M^{(m)}$ and the standard deviation $S^{(m)}$.
2. Construct the series

$$X_k^{(m)} = \sum_{t=1}^{k} \left(x_t - M^{(m)} \right) \tag{1}$$

for $k = 1, 2, ..., m$.

3. Calculate range

$$R^{(m)} = \max_k \left(X_k^{(m)} \right) - \min_k \left(X_k^{(m)} \right) \tag{2}$$

4. Calculate rescaled range series

$$\text{def. } R/S_m = R^{(m)}/S^{(m)} \tag{3}$$

Repeat the calculations for subsequent possible shares of the series x_t. The series of the values $(R/S)_n$ obtained in this way, is the base for calculation of the Hurst exponent, using the following expression:

$$\ln(R/S)_n = h(N)\ln n + \ln a \tag{4}$$

where *h(N)* is the Hurst exponent, *a* – any constant.

Estimated values of H can be obtained from the plot (R/S) versus t in log-log axes. The slope of the regression line approximates H (see Figure 1).

For Gaussian random series, H.E. Hurst has given the following formula to calculate the expected $(R/S)_n$ value as:

$$E(R/S)_n = \sqrt{n\frac{\pi}{2}} \tag{5}$$

The formula has been developed and the final form presented by Peters [6].

$$E(R/S)_n = \frac{n-0.5}{n}\sqrt{\frac{2}{n\pi}}\sum_{i=1}^{n-1}\sqrt{\frac{n-i}{i}} \tag{6}$$

So, in order to ensure that a series is chaotic, one should compare the empirical H with the anticipated value of H, calculated with the use of the above formula. The effect of the long-term memory occurs when the difference between empirical and theoretical values of H is greater than $\sqrt{1/N}$.

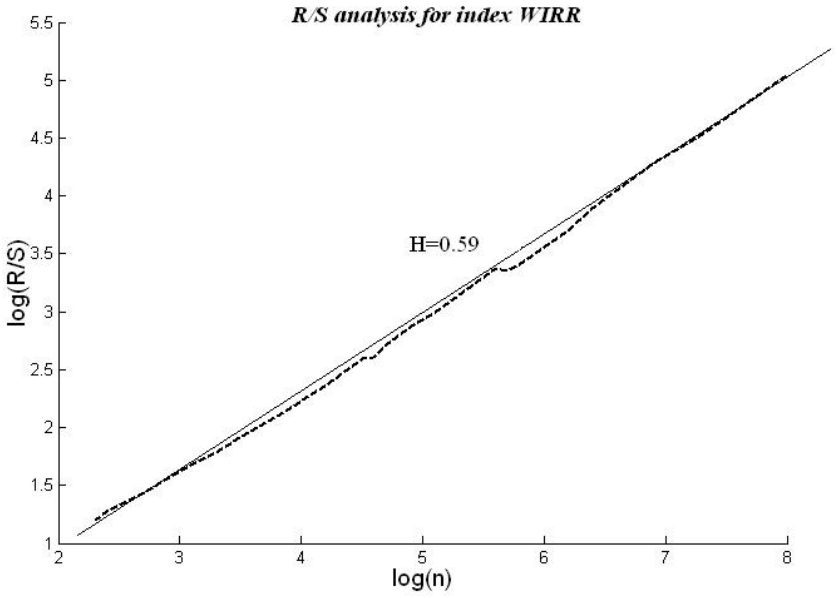

Figure 1. An example of the R/S analysis.

Auto-Mutual information

When we are sure that our financial time series is not fully random, we have to reconstruct an attractor from the scalar data. We seek an embedding space that preserves the original unknown attractor. The attractor can be constructed by expanding a scalar time series *x(t)* into a vector time series ***X(t)*** of delay coordinates ***X(t)=(x(t), x(t+τ), …, x(t+(n-1)τ))***, τ is an integer, called time delay. The introduction of τ allows skipping any samples during the reconstruction. For a too

small τ, the coordinates of x in the state space are fairly similar. Whether τ is too large, it causes loss of information, contained in the data.

One of the methods of choice for τ is the use of the autocorrelation function. This method of the τ calculation has been suggested [3], where was described criterion called auto-mutual information. Independence is measured as the information in bits, gained about $x(t+\tau)$ given the measurement of $x(t)$. This is the mutual information I and its first minimum is considered as a good estimator of τ (see Figure 2), where τ can be assumed as 4.

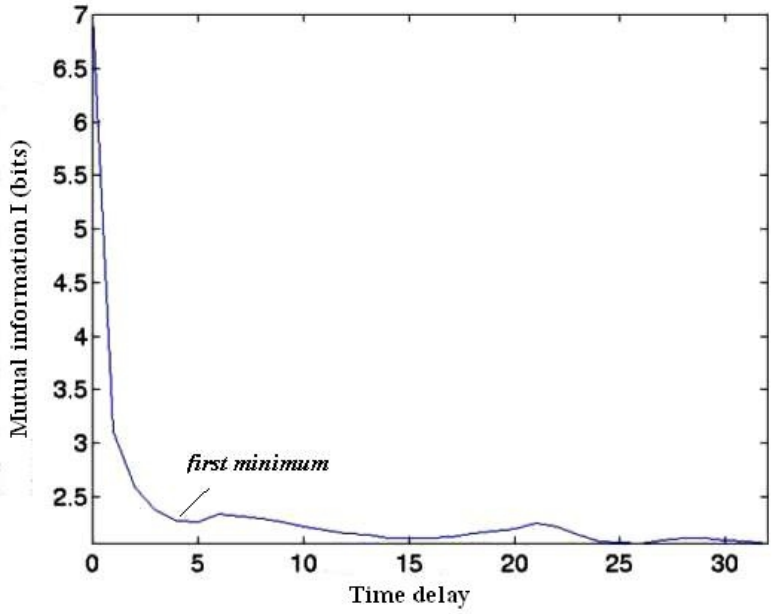

Figure 2. An example of mutual information, where the first minimum approximates 4.

False near neighbors'

A subsequent parameter, needed for reconstruction of the state space, is embedding of n dimension. The effective method of choice, using embedding dimension, is described in [4]. Calculations are based on the idea that the passage from dimension d to dimension $d+1$ allows to distinguishing between "true" and "false" neighbors. A false neighbor is a point in a data set that is a neighbor only because the assumed dimension d of the attractor is too small. So, the dimension should be increased till the first value of d, for which all the neighbors of each point in the state space will be true neighbors. A calculation example of false nearest neighbors is graphically shown on Figure 3.

3 Data Preparation and Pre-Processing

The Hurst exponent values were calculated for selected shares and indexes from the Warsaw Stock Exchange. For the majority of the selected time series, H values were close to 0.5, what suggests a fully random behavior of those series. The greatest H value among the tested series

was obtained for the WIRR index and amounted to 0.59. Finally, for further research, four indexes: WIG, MIDWIG, WIRR and WIG20 were chosen. The initial time series contained the daily close values of the indexes from a few last years: depending on index, it was between nine and fourteen years. Then, those series were transformed into series of the logarithmic daily rates of return, according to the expression below:

$$s_t = \ln\left(x_t / x_{(t-1)}\right) \tag{7}$$

where s_t - logarithmic rate of return.

Figure 1 shows an example of R/S analysis for the WIRR index. In this case, the regression line has a slope equal to 0.59 and it is also the value of H.

Then, the following values were successively calculated:

- estimated H value for series of *N* length,
- minimal distance between empirical and estimated H values to test the occurrence of long-term memory,
- time delay from mutual information,
- embedding dimension of the state space.

The results of these calculations are presented in Table 1.

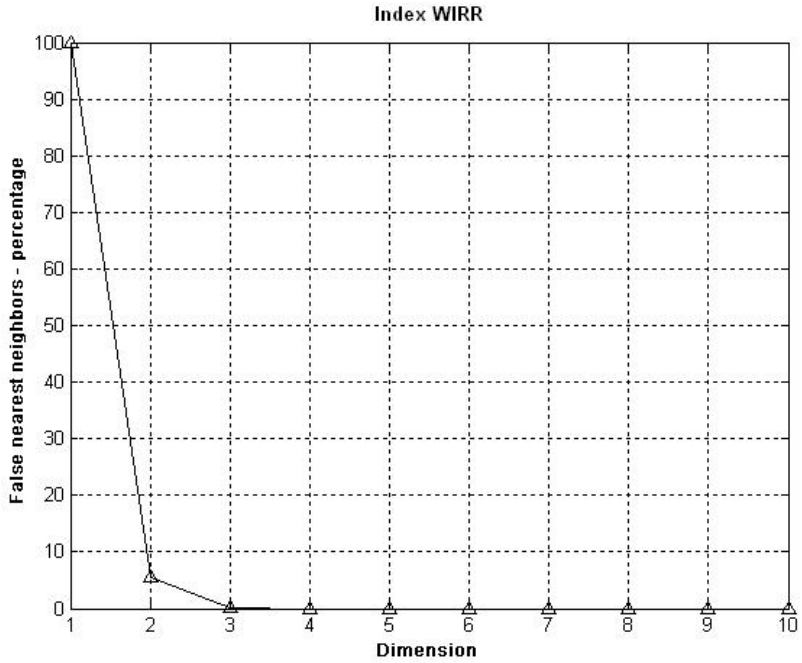

Figure 3. False nearest neighbors versus dimension

The WIG20 series is fully random and intended to compare the obtained results. Additionally, for the remaining three samples, the series were scrambled and then H values were calculated. If there is any structure in these series, scrambling should destroy it and the Hurst exponent should be close to the random series. H values calculated for the scrambled series were about 0.52-0.53, allowing for a conclusion that, some arrangement may exist in these series, differentiating them from random series.

Table 1. The results of calculations

Index	Length of the series N	Empirical Hurst exponent	Estimated Hurst exponent	Minimal distance $\sqrt{1/N}$	Time delay τ	Embedding dimension D
WIG20	3150	0.53	0.526	0.018	-	-
MIDWIG	2280	0.58	0.527	0.021	1	4
WIG	3480	0.57	0.523	0.017	1	4
WIRR	3150	0.59	0.526	0.018	1	4

The WIG20 series is fully random and intended to compare the obtained results. Additionally, for the remaining three samples, the series were scrambled and then H values were calculated. If there is any structure in these series, scrambling should destroy it and the Hurst exponent should be close to the random series. H values calculated for the scrambled series were about 0.52-0.53, allowing for a conclusion that, some arrangement may exist in these series, differentiating them from random series.

The above observation gives us some hope that it is possible to re-create hidden structures in the series and exploit predictability in the series.

Based on the performed calculations and analysis, the original multidimensional attractors were re-created. Researched time series have the same time delay $\tau=1$ and the same embedding dimension $n=4$, thus, for a given time series $x_1, x_2, \ldots, x_k$ time-delay embedding vectors $\boldsymbol{X_i}=(x_i, x_{i+1}, x_{i+2}, x_{i+3})$ may be constructed to predict x_{i+4}. For all the time series, new series were built, containing $\boldsymbol{X_i}$ and x_{i+4}, $i=1\ldots1024$ (data from the last four years), $\boldsymbol{X_i}$ as an input data and x_{i+4} as an expected output data.

GP is a universal tool that provides a possibility to discover functions and dependencies hidden in the data sets, so it is expected that using symbolic regression, we may obtain appropriate results.

4 Genetic Programming Construction

Four time series were used: MIDWIG, WIG and WIRR, with long-term memory, and random WIG20, containing 1024 cases, provided $\boldsymbol{X_i}$ as an input and x_{i+4} to calculate the output error.

A special form of GP, called symbolic regression, will be used in effort to discover hidden functions in the data. The symbolic regression is to discover a function that can fit a finite set of sample data. It is necessary to emphasize the fact that the object of search is a symbolic description of a model, not just a set of coefficients in a prespecified model. This is in sharp contrast with other methods of regression.

Below, there is a specification of GP configuration, as used in that research:

- The set of terminals $T=\{X1, X2, X3, X4, R\}$, R – random constant
- The set of functions $F=\{+, -, *, /, \sin, \cos, \exp, rlog, \text{if-else}\}$; $rlog(x) = \begin{cases} 0 & for\ x = 0 \\ \log(|x|) & for\ x \neq 0 \end{cases}$
- Population size 5000
- Max. generations 200
- Fitness function: raw fitness $r=\sum_{i=1}^{1024} \left| y_i - y_i^* \right|$ y_i – real output, y_i^* - expected output.

Raw fitness should tend towards zero.

GP "randomly" searches for the best-fit expression and computes raw fitness for a given series y_i. The best results for all the series are presented in Table 2.

Table 2. Hurst exponents and the best fitness

Index	Hurst exponent	Raw fitness
WIG20	0.53	10.07
MIDWIG	0.58	6.71
WIG	0.57	8.35
WIRR	0.59	8.02

Comparison of the Hurst exponent and the best raw fitness for the time series shows that the series were higher H values, i.e., containing long-term memory, are more predictable and generate lower raw-fitness values, as well as lower errors in the output.

It should be checked whether the above results can be applied in real transactions on financial markets. In Figure 4, GP output is shown for the best-fit program, together with expected values for the last 100 quotations of WIRR.

The ranges of changes of the quotations are very different for both graphs so, this information is not useful from the point of view of market transactions. Instead, it seems that it is possible to draw some conclusions from this chart about trends of future movements, what is also a very valuable information in planning market transactions. This last conclusion needs further investigations.

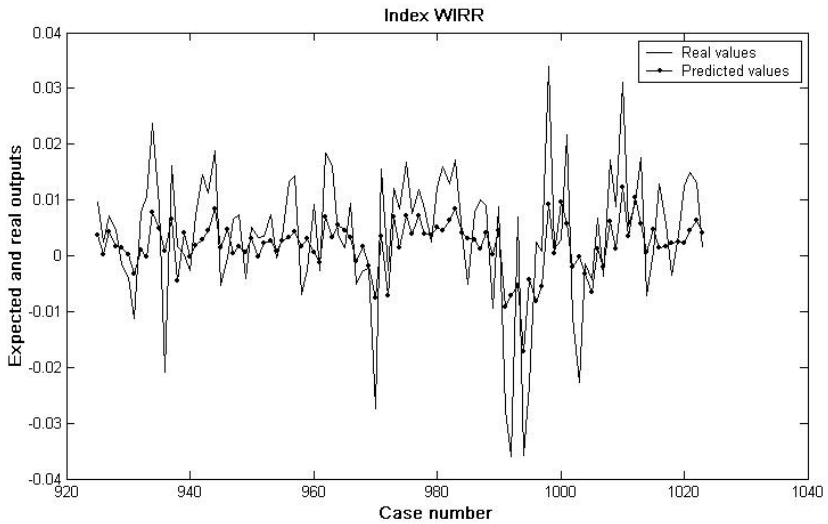

Figure 4. Real and expected output for the last 100 quotations of WIRR

5 Conclusions

In this paper, the Hurst exponent was analyzed as a measure of predictability of financial time series. It has been shown that the time series with higher H are more predictable, contain

long-term memory and differ from series with H close to 0.5 that are fully random. Series with high H can be transformed from their original values into series that reconstruct the original multidimensional attractor. The presented methods, i.e., the auto-mutual information and the false near neighbors', allowed calculating time-delays and embedding dimension for linear time series and constructing embedding vectors. GP has been used to search the expression that, in the best way approximates the time series. The use of GP has confirmed that series with higher H generate lower errors in the output. Although the ranges of changes are considerably different, they seem to be able to predict future changes and trends. So, the connection chaos theory and GP can become a powerful tool for investors.

Bibliography

[1] Dudul, S. V. Prediction of a Lorenz chaotic attractor using two-layer perceptron neural network. *Applied Soft Computing* 5:333-355, 2005.

[2] Fraser, A. M., and Swinney, H. L. Independent coordinates for strange attractors from mutual information, *Physical Review A* 33:1134-1140, 1986.

[3] Hurst, H. E. Long-term storage of reservoirs: an experimental study. *Transaction of the American Society of Civil Engineers* 116:770-799, 1951.

[4] Kennel, M. B., Brown, R., and Abarbanel, H. D. I. Determining embedding dimension for phase-space reconstruction using a geometrical construction. *Physical Review A* 45:3403-3411, 1992.

[5] Koza, J. *Genetic Programming III.* San Francisco, CA: Morgan Kaufmann Publishers, Inc, 1999.

[6] Peters, E. E. *Fractal Market Analysis. Applying Chaos Theory to Investment and Economics.* New York: John Wiley & Sons Inc, 1994.

[7] Peters, E. E. *Chaos and Order in the Capital Markets. A New View of Cycle, Prices and Market Volatility*. New York: John Wiley & Sons Inc, 1996.