Using R/S analysis for forecasting stock quotes with ARMA and ARIMA methods

Alena Stupina and *Anna* Zinenko*

Siberian Federal University, 79, Svobodny Prospect, Krasnoyarsk, 660049, Russian Federation

Abstract. The article observes the methods of forecasting time series Autoregressive Moving Average Model (ARMA) and Integrated Autoregressive Moving Average Model (ARIMA). The ARIMA model differs from the ARMA model only in that forecasting is performed not on absolute values of series levels, but on differences of order *d*, which makes it possible to apply to non-stationary time series. Financial time series are traditionally considered non-stationary. However, the Hurst exponent less than or equal to 0.5 indicates a random or anti-persistent nature of time series. The paper assumes that for random or anti-persistent time series according to Hurst, there is no need to take differences and it is sufficient to apply the ARMA model for forecasting. To test the hypothesis, we carried out forecasts of leading world indices stocks and currency pairs with the Hurst exponent less than or equal to 0.5 for 10 years using the ARMA and ARIMA methods and compared the results using the MAPE metric. According to the ARMA method forecasts, in most cases the error was smaller, that confirmed the initial hypothesis.

1 Introduction

R/S analysis is an invention of hydrologist Harold Edwin Hurst in the early 20th century for natural phenomena researches. Later, French mathematician Benoit Mandelbrot [1] discovered that market quotes also obey the patterns discovered by Hurst. Currently, R/S analysis for determining the persistence of time series and for determining their fractal dimension occasionally appears in scientific articles. The most complete guide to R/S analysis and calculating the Hurst exponent can be found in the books of trader and scientist Edgar Peters [2].

The essence of R/S analysis is to construct a linear regression equation in which the dependent feature is the logarithm of the range normalized by the standard deviation, and the factor is the logarithm of the number of elements in a subsample from the original sample. The slope coefficient of the regression line calls the Hurst exponent and its value indicates the persistent, random, or antipersistent nature of the time series.

Persistence in most sources is defined as dependence on past values. Thus, persistent time series have some trend. Now we recall another concept - stationarity of time series. Time series consider stationary in the narrow sense (strictly stationary) if a time shift does not

* Corresponding author: anna-z@mail.ru

change the distribution density function. This definition implies constant mathematical expectation and variance. Time series consider stationary in the broad sense (weakly stationary) if their mathematical expectation and variance exist and do not depend on time, and the autocorrelation function depends only on the difference between adjacent levels of the series [3]. Non-stationary time series are characterized by the presence of a trend or seasonality. Based on these definitions, we cannot identify the concepts of "non-stationarity" and "persistence", but we can say with confidence that persistent time series are non-stationary.

As mentioned above, the persistence of time series is determined by the value of the Hurst exponent. It fluctuates between zero and one. If the Hurst exponent is greater than 0.5, the time series is persistent, if it is strictly equal to 0.5, it is random, if it is less than 0.5, it is antipersistent (changes in the values of the time series occur faster than in a random time series). The mathematical basis for R/S analysis is the pattern discovered by A. Einstein: for large volumes of random sampling, its range is equal to the number of elements in the sample raised to the power of 0.5. Thus, the Hurst exponent for random time series is 0.5.

Now we turn to the autoregressive moving average (ARMA) model and the integrated autoregressive moving average (ARIMA) model, developed in the 1970s by George Box and Gwilym Jenkins [4]. The ARMA model shows the dependence of the current level of the series on past levels, lagging by a time lag (order p of the model) and on the error, lagging by a time lag (order q of the model). In our works, we examined in detail how to calculate the error in the second component of the model [5]. However, the ARMA model is applicable only to stationary time series.

The ARIMA model complements the ARMA model by taking differences of order d to reduce a non-stationary time series to a stationary one. As a rule, first-order differences are sufficient to achieve stationarity. The ARIMA model is successfully used in forecasting financial time series. While forecasting using the ARIMA method, the forecast should be cumulative, since it initially forecasts are not the levels of the series, but the differences between them. However, if the financial time series is random, it is questionable if it makes sense to take differences and complicate the model. This paper is devoted to this issue. We examined more than a thousand stocks and currencies included in large indices, selected those that are random according to Hurst, and compared forecasts using the ARMA and ARIMA methods using the MAPE error metric (mean absolute percentage error).

2 Methods

2.1 ARMA and ARIMA Models

The ARMA (p, q) model is represented by the following equation

$$y_t = \sum_{i=1}^{p} \alpha_i y_{t-i} + \varepsilon_t + \sum_{i=1}^{p} \beta_i \varepsilon_{t-i}\,, \qquad (1)$$

where y_t – current level, y_{t-i} – level, laggard by lag i, ε_t – moving average forecast error, ε_{t-i} – moving average forecast error laggard by lag i. The first sum of the model is the autoregressive component, the second is the moving average component. For the ARIMA model, formula (1) is transformed as follows

$$\Delta_t^d = \sum_{i=1}^{p} \alpha_i \Delta_{t-i}^d + \varepsilon_t + \sum_{i=1}^{p} \beta_i \varepsilon_{t-i}\,, \qquad (2)$$

where Δ_t^d is the difference of d-order required to achieve stationarity [6].

The order of taking differences is determined by the Dickey-Fuller test. As for the parameters p and q, the order of autoregression is determined by the partial autocorrelation plot of the series levels, in which the time lags are indicated along the X axis, and the values of the correlation coefficient between the levels corresponding to the lag are along the Y axis [6]. The order of autoregression is chosen equal to the time lag at which the correlation coefficient takes the last maximum value different from zero. The order of the moving average is chosen in a similar way, only instead of the partial autocorrelation coefficients, the autocorrelation coefficients are calculated. Partial autocorrelation differs from autocorrelation in that it does not take into account the influence of levels located between the current level and the level lagging by a time lag. Obviously, with a single lag, autocorrelation and partial autocorrelation coincide [7].

3 R/S analysis

The purpose of R/S analysis is to find the Hurst exponent (in our case, to determine the persistence of a time series). The algorithm for implementing R/S analysis is as follows [5]. We select a time series of a sufficiently large size. An additional condition is a large number of proper divisors, since the number of subsamples and, accordingly, the volume of initial data for constructing the regression equation depend on this. We transform the sample to logarithms and obtain a time series by length *n* with level values $og\frac{y_i}{y_{i-1}}$.

The next step is to find the smallest proper divisor of *n* that is not less than 10. Denote it as *m*. Divide *n* into $k=\frac{n}{m}$groups and denote the elements in each group as t_i. Obviously, the number of elements in each group is *m*. Then we find the average values in each group:

$$\bar{t}_k = \frac{1}{m}\sum_{i=1}^{m} t_i,\ \frac{1}{m}\sum_{i=m+1}^{2m} t_i \ldots\ldots \frac{1}{m}\sum_{i=(k-1)m+1}^{n} t_i \quad (3)$$

and accumulated deviations from the mean X_i:

$$X_1 = t_1 - \frac{1}{m}\sum_{i=1}^{m} t_i,\ X_2 = \left(t_2 - \frac{1}{m}\sum_{i=1}^{m} t_i\right) + X_1 \ldots$$
$$X_m = \left(t_m - \frac{1}{m}\sum_{i=1}^{m} t_i\right) + X_{m-1}. \quad (4)$$

The normalized range for each group will be $R_k = \max(X_i) - \min(X_i)$. Also, for each group, we calculate the standard deviation $S_k$using the formula

$$S_k = \sqrt{\frac{1}{m}\sum_{i=1}^{m}(t_i - \bar{t}_k)^2}. \quad (5)$$

We calculate R/S indicator for each group as R_k/S_k. Then we find the average range of variation

$$\overline{R/S_j} = \frac{1}{k}\sum_{i=1}^{k} R/S_i. \quad (6)$$

The *j* index means that we have obtained the average variation range at the *j* -th step, which corresponds to the *j* -th proper divisor. We repeat the above procedure, using all possible proper divisors as *m*. At the last step, *m* will be equal to $\frac{n}{2}$.

That way we get a sample of $\overline{R/S_j}$, $j = \overline{1, \frac{n}{2}}$, the number of elements in the sample corresponds to the number of proper divisors of *n*. Now we can construct a linear regression equation in which the dependent variable is the logarithm of the *R/S* indicator, and the factor feature is the logarithm of the number of elements in the *j* -th group *k*:

$$\log R/S = log\ c + H * \log k. \quad (7)$$

The authors have developed and patented an algorithm in Python that allows calculating the Hurst exponent for various financial instruments and plotting a regression line. A change in the slope of the regression line means that as the number of elements in the subsample increases, the range of the subsample decreases. According to Peters, that means that when the subsample reaches such a number of elements, a cycle occurs. During the analysis, we found very few instruments (about 0.3%) that exhibit such a phenomenon. From this, we can conclude that the Hurst exponent is permanent for each instrument over a long period. Using this statement, we created a database of the Hurst exponents of the instruments we analyzed in order to reduce the execution time of the algorithm described in the next paragraph.

4 Results and discussion

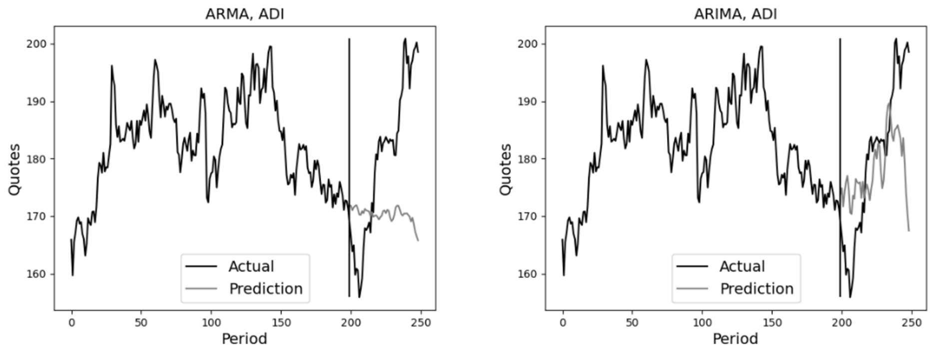
For the analysis, we took the AMEX, NASDAQ, NYSE indices of American stock exchanges and the FOREX currency market index. Since we took data for a long period of 10 years, data on some instruments were missed and had to be cut off. Also, the calculation of the Hurst exponent required deeper data cleaning. For example, to calculate the R/S ratio, it is necessary to place the standard deviation in the denominator. Thus, stocks and currencies whose quotes do not change over long periods and whose standard deviation is zero are cut off. As a result, 131 instruments were taken for the AMEX index, 162 for NASDAQ, 401 for NYSE, and 338 instruments for FOREX. From these stocks and currencies, we selected those with a Hurst exponent less than or equal to 0.5, i.e. random or anti-persistent stocks.

As we assume, it makes no sense to take differences (use the ARIMA model, not ARMA) for random and anti-persistent stocks. In order to confirm our assumption, we made a forecast on annual time intervals using the ARIMA and ARMA methods and compared them using the MAPE metric (mean absolute error in percent). The total period was divided into training and forecast in a ratio of 80:20. Thus, quotes were analyzed for 10 years - from 2015 to 2024. The results are shown in Table 1.

Table 1. Averaged by indices MAPE (in %) using the ARMA and ARIMA methods.

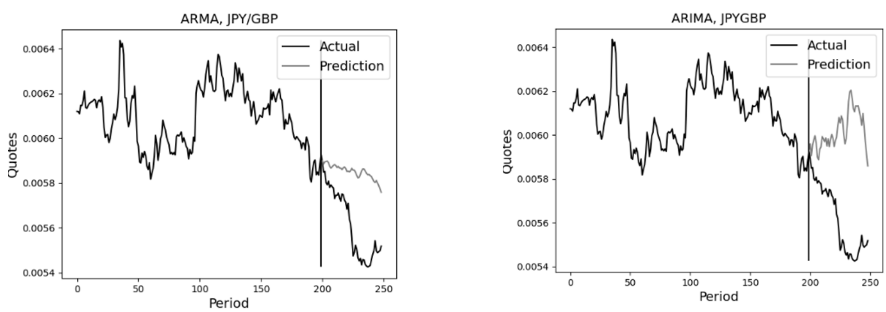
Period	AMEX		NASDAQ		NYSE		FOREX	
	ARMA	ARIMA	ARMA	ARIMA	ARMA	ARIMA	ARMA	ARIMA
2015	9.7	20	6.4	6	6	7.3	44.6	20.4
2016	15.2	26.5	5.3	5.3	7.7	5.9	42.7	41.5
2017	9.8	17.3	6.2	5.7	6.3	6.3	42	17.5
2018	25	19	6.9	8.2	9	7.1	5.1	4.2
2019	8	16.1	5.5	5	5.8	5.4	6.1	20.6
2020	18.3	21.2	5.9	6.4	7.3	6	9	21
2021	10.8	16.7	10.5	6.7	9.8	6.7	2.4	3.1
2022	13	27.3	5.5	5.6	5.5	5.9	2.7	3.3
2023	15	28	6	7	5	5	2.8	3.2
2024	8.7	12.2	5.8	6.3	5.7	5	2.2	2.8

Table 1 shows that in most cases the ARMA model shows a smaller average error. Only in 2018 ARMA showed a larger average error for three of the two indices, with a significant difference only for the AMEX index. Therefore, when forecasting using the ARIMA method, it makes sense to separate stocks by the Hurst exponent and not forecast by absolute values of the series levels, but by first-order differences, random and anti-persistent time series. Figures 1–2 visualize the forecasts of some instruments using the ARMA and ARIMA methods.



a) b)

Fig 1. Forecast for ADI stock with methods a) ARMA b) ARIMA.



a) b)

Fig 2. Forecast for currency pair Japanese Yen/British Pound with methods a) ARMA b) ARIMA.

5 Conclusion

The paper examined forecasting financial instruments using the ARIMA method, which shows some of the best results in financial time series forecasting [8]. We also considered the Hurst exponent, which allows us to draw conclusions about the persistent, random or antipersistent nature of time series. We hypothesized that random and antipersistent series do not need to be subtracted when forecasting using the ARIMA method, and thus the ARIMA model degenerates into the ARMA model. Using the algorithm we developed in the Python programming language; we calculated the Hurst exponents for about a thousand stocks and currency pairs included in American stock indices and FOREX market instruments. Then we selected those with a Hurst exponent less than or equal to 0.5 and made forecasts for annual samples for 10 periods from 2015 to 2024 using the ARMA and ARIMA methods. For each method and for each period, the averaged over the index MAPE metric was calculated. In most cases, the error by the ARMA method was smaller (in some cases significantly smaller) than by the ARIMA method, from which we can conclude that it makes sense to use the ARMA method instead of ARIMA for random and Hurst antipersistent instruments.

This work was supported by the Ministry of Science and Higher Education of the Russian Federation (Grant No.075-15-2022-1121).

References

1. B. Mandelbrot, *Fractals, case and finance* (Research Center "Regular and Chaotic Dynamics", 2004)
2. E. Peters, *Fractal analysis of financial markets. Application of chaos in investment and economics* (Internet trading, 2004)
3. G. G. Kantorovich, HSE Econ., **1**, 85 (2002)
4. J. Box, G. Jenkins, *Time Series Analysis: Forecast and Control* (Mir, 1974)
5. A. V. Zinenko, Bus. Inform., **3 (21)**, 21 (2012)
6. J. Zhang, H. Liu, W. Bai, X. Li, North Am. J. Econ. Finance. **69**, 102022 (2024)
7. E. Nielsen, *Practical analysis of time series: forecasting with statistics and machine learning* (Dialectics LLC., 2021)
8. V. P. Nosko, *Econometrics. Introduction to regression analysis of time series* (National Foundation for Personnel Training, 2002)