Developing new mathematical method for search of the time series periodicity with deletions and insertions

E V Korotkov[1,2], M A Korotkova[1]

[1] National Research Nuclear University MEPhI (Moscow Engineering Physics Institute), Kashirskoe highway, 31, 115409, Moscow, Russia
[2] Institute of Bioengineering, Research Center of Biotechnology of the Russian Academy of Sciences, Leninsky Ave. 33, bld. 2, 119071, Moscow, Russia

E-mail: genekorotkov@gmail.com

Abstract. The purpose of this study was to detect latent periodicity in the presence of deletions or insertions in the analyzed data, when the points of deletions or insertions are unknown. A mathematical method was developed to search for periodicity in the numerical series, using dynamic programming and random matrices. The developed method was applied to search for periodicity in the Euro/Dollar (Eu/$) exchange rate, since 2001. The presence of periodicity within the period length equal to 24 h in the analyzed financial series was shown. Periodicity can be detected only with insertions and deletions. The results of this study show that periodicity phase shifts, depend on the observation time. The reasons for the existence of the periodicity in the financial ranks are discussed.

1. Introduction.

Identification of the cyclic patterns in numeric time series and symbolical sequences may shed light on the processes occurring in systems of a different nature and give information about the structure of different time series. To identify periodicities in the time series and symbolical sequences, the methods mainly used are based on the Fourier transform, wavelet transform and dynamic programming, as well as some other method [1–10]. Previously, we proposed the method of informational decomposition, which allows the detection of periodicity in both symbolic and numeric sequences; however, the detection was severely impeded by the above approaches [11]. These difficulties emanate from the fact that methods based on Fourier and Wavelet transform, decompose the statistical significance of long periods (larger than the size of the analyzed sequence alphabet) for smaller periods with a multiple of the length [11]. Also, these methods are very sensitive to the insertion and deletion of symbols. This leads to the fact that spectral methods cannot detect periodicity at a statistically significant level, even in the presence of a few deletions or insertions. Dynamic programming, which allows the detection of deletions and insertions in the periods, cannot detect relatively "fuzzy" periodicity. This is due to the fact that this method is based on finding the similarity between pairs of periods in the studied sequence [4]. However, if a statistically significant similarity is absent between two separate periods, dynamic programming will fail to detect periodicity in the analyzed sequence. The lack of similarity between two separate periods can be observed for latent periodicity, where the periodicity occurs on the background of random noise [11]. To find such periodicity, the method of informational decomposition was used [11]. This method enabled the discovery of latent periodicity in the DNA sequences of many genes

[12,13] and revealed the latency of amino acid specific for the protein families [14,15]. These results suggest that latent periodicity with insertions and deletions can be detected in sequences of a different nature and in numerical sequences also.

Today, it is known that people, animals and plants have biological rhythms. The manifestation of these rhythms can be observed at all levels of biological organization. The interaction can be observed between the rhythms that affect the internal state of the person and on various social processes. A. Chizhevsky first drew attention to the influence of natural factors on social processes [16] The biological rhythms can also influence currency rates. If the periods in the exchange rates exist, then the various events in public life can affect them. Such events can lead to changes in the data and could be identified as a phase shift of the period. The phase shifts in the sequence could have resulted from deletion or insertion of values with respect to the existing period. Therefore, in order to detect this periodicity, it was necessary to develop a mathematical method for detecting the periodicity of the time series, taking into account the unknown location and unknown number of insertions or deletions, in the presence of large noise. It cannot be done with the help of all known mathematical methods. Previous studies have searched for insertions or deletions [17], but failed to find a periodicity with large noise [18]. It is right mainly for dynamic programming. Either of these methods works well in the presence of large noise, but fails to find periodicity in the presence of even small amounts of insertions and deletions [11].

This paper contributed to filling the gap in mathematical methods for periodicity search. A mathematical method was developed to find periodicity in the symbolical sequences, in the presence of insertions and deletions and the big noise. The method was developed using the random matrices of periodicity and the method of dynamic programming. The developed method was applied to search for periodicity in the exchange rate of the Euro to the US Dollar. To search for periodicity, the numerical sequence was converted to a symbolical sequence. The calculations show that there is a periodicity equal to 24 h in the Euro-Dollar exchange rate. This periodicity of the exchange rate contains a lot of insertions and deletions and could not be detected by previously developed mathematical methods.

2. Methods and Algorithms

For study, we took the numerical sequence of the Euro/US$ rate. The candle opening and closing was separated by an interval of 1 and 4 h. Let $x_1(i)$ - be the rate at the time of opening of the candle, and $x_2(i)$- the rate at the close of the candle. At the first numerical sequence $A_{1,}$ we calculated the difference $a(i)=x_2(i)-x_1(i)$, where $x_1(i)$ and $x_2(i)$ are separated by a 1 h. The beginning of the candle falls on the start of each hour and the end of the candle at finish of each hour. The sequence A_1 was obtained from 14.08.2014 to 07.04.2015 and contained 3930 numbers ($a(i)$, i=1,...,3930). We used all days of the week except Saturday and Sunday. These days the trade is not made or the volume of trade is very low. Then, this sequence was translated into a symbolic form, as described below, and the sequence S_1 was obtained. The encoding is shown in Table 1.

At the sequence $A_{2,}$ the difference $a(i)=x_2(i)-x_1(i)$, was calculated, where $x_1(i)$ and $x_2(i)$ are separated by 4 hours. This means that the beginning of the candle occurs at the beginning of each 4 hours, and the end of the candle is the end of each 4 hours. The sequence A_2 was obtained for data from 05.12.2014 to 07.04.2015 and it contained 4027 numbers. After translation of this sequence in the symbolic sequence, the sequence S_2 was obtained. We used all days of the week except Saturday and Sunday also. Numerical data were taken from the Meta Trader 4 Alfa-Forex client terminal, site http://www.alfa-forex.ru. These data can be received from http://www.finam.ru also.

The numeric sequences A_1 and A_2 of length N_1 and N_2 were transformed in the symbolic sequence S_1 and S_2 with alphabet of 20 letters (Table 1). To convert the numerical sequence, the minimum and maximum elements of the sequences A_1 and A_2 were determined. Then, this interval was divided into 20 intervals, the number of elements of the numeric sequence in each interval was approximately equal to $N_1/20$ and $N_2/20$, respectively. Each interval received the letter of the Latin alphabet. If a numeric sequence contained many of the equal values than the boundaries of the intervals were varied in such a way that the same numbers were encoded by the same symbol. The coding of the sequence A_1 is shown in Table 1.

Table 1. The coding of sequence A_1 to obtain the sequence S_1 is shown.

K	N	I	M	T	R	S	L	Y	F
-.01042 -0.00215	-0.00215 -0.00149	-0.00149 -0.00108	-0.00108 -0.00082	-0.00082 -0.00062	-0.00062 -0.00047	-0.00047 -0.00034	-0.00034 -0.00021	-0.00021 0.00011	-0.00011 0.00002
C	W	P	H	Q	V	A	D	E	G
0.00002 0.00008	0.00008 0.00017	0.00017 0.00029	0.00029 0.00040	0.00040 0.00052	0.00052 0.00071	0.00071 0.00097	0.00097 0.00130	0.00130 0.00202	0.00202 1.00000

To search the periodicity with insertions and deletions, the next algorithm was used. Firstly, a set of random matrices with size 20xn was generated, where n is the length of the period, and 20 is the alphabet size of the studied sequence. Then, a local alignment of the studied sequence (S_1 *or* S_2) was built relative to each of the generated random matrices. Dynamic programming was used to build the local alignment and in determining the similarity function *F*. Then, the matrices were transformed because the distribution of the maximum of the similarity function *F* for each of the matrices for random sequences (set *Q*, volume of the set is 1000 sequences) should be similar. The transformed matrix having the highest value of the similarity function *F,* with the studied sequence *S,* was chosen. Then, this matrix was optimized to achieve the highest value of the similarity function *F* (*maxF*) with the studied sequence *S* and the transformed matrix was called *T*. For this purpose we used the genetic algorithm [19,20]. Then, the value of *maxF* for each random sequence from the set *Q* and for matrix *T* was calculated. It allowed the mean value and variance for *maxF* to be determined. This algorithm was applied for periods of different lengths and for each length of the period *n,* the corresponding value of $Z = (maxF - \overline{maxF}) / D(maxF)^{0.5}$ was calculated. As a result of the algorithm, the dependence of *Z* on *n* was obtained and denoted as *Z*(*n*). It should be noted that in this study, dynamic programming was used to find a local alignment. This means that the boundaries of the regions with *maxF,* may differ from the beginning and end of the studied sequence. It means also that the values of *Z*(*n*) for different *n* can be obtained for different fragments of the studied sequence. The boundaries of the fragments, obtained for relevant values of *Z*(*n*) are shown. More details of the applied mathematical method are shown in [19,20].

3. Results

Figure 1 shows the application of the developed approach to the random sequence and the sequence S_3 obtained from the sequence S_4=(EFKLWNMSTWRYLQKLWQSMETMQ)$_{16}$. Randomly, 75% substitutions were done in the sequence S_4, after which our developed approach was applied. The results of the analysis of random sequences show that *Z*(*n*) fluctuated around 4.0. Noticeable, a small trend toward increasing values of *Z* for the random sequence is only for lengths of periods of more than 70. From these results, it can be concluded that the periods with Z-values>7.0 were found to be interesting. The results of the study of sequence S_4 show that the developed approach was able to identify fuzzy frequency and determine the optimal weighting matrix [20] for the detection of periodicity with length equal to 24 symbols. It is possible to see periodicity equal to 48 and 72 symbols. However, large values of *Z* also received periods with lengths close to 48 and 72 symbols, because there is the possibility of insertions or deletions. It is always possible to do insertions, or deletions to align the sequence S_4 against the matrices and to receive periodicity near to 48 to 72 symbols with high *Z*. However, this leads to a slight drop in the values of *Z,* which is related to the penalties for insertions or deletions. So the graph around 48 and 72 symbols, looks like a gently sloping mountain.

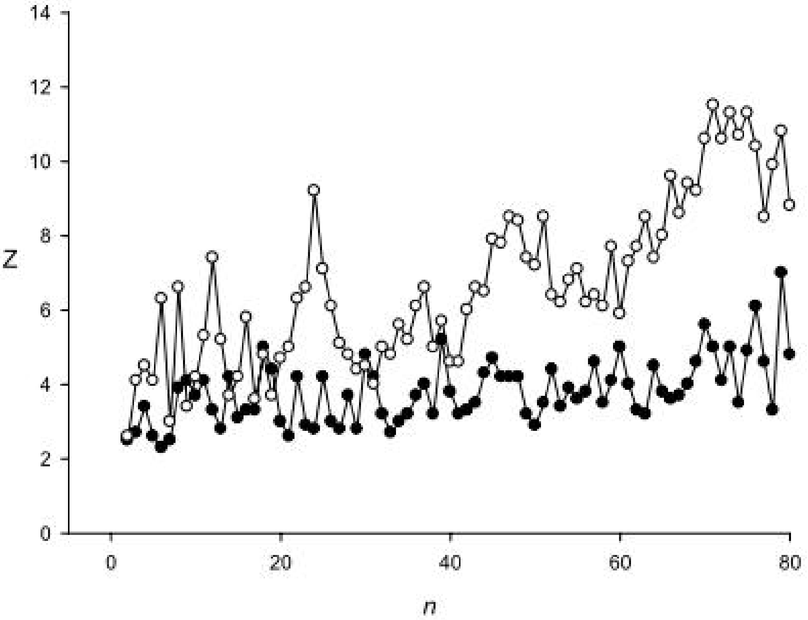

Figure 1. Black circles show the *Z*(*n*) for the random sequence. White circles show the *Z*(*n*) for the artificial sequence with a period of 24 symbols with total length of 384 symbols and 75% random substitutions (S_4). It is evident that despite the large number of random substitutions, the developed mathematical method could detect the periodicity with *n*=24.

Next, the developed approach was applied to the periodicity search for the symbolic sequence S_1. The spectrum of *Z*(*n*) obtained for the sequence S_1 is shown in figure 2A. From this figure, it can be seen that the sequence S_1 has two periods. The first period equals 24 hours and has the largest value of *Z* from 341 to 3533, the hours of the sequence S_1. 39 insertions and deletions were made to find this period. The second period has a length equal to 23 hours and has the largest value of *Z* from 1044 to 3478 hours. The period equal to 24 h is the most statistically expressed ($Z \approx 10.3$) compared to the period having length of 23 h ($Z \approx 8.1$). It can be assumed that building a significant alignment for two periods is associated with the ability to create insertions or deletions of the symbols in the sequence S_1.

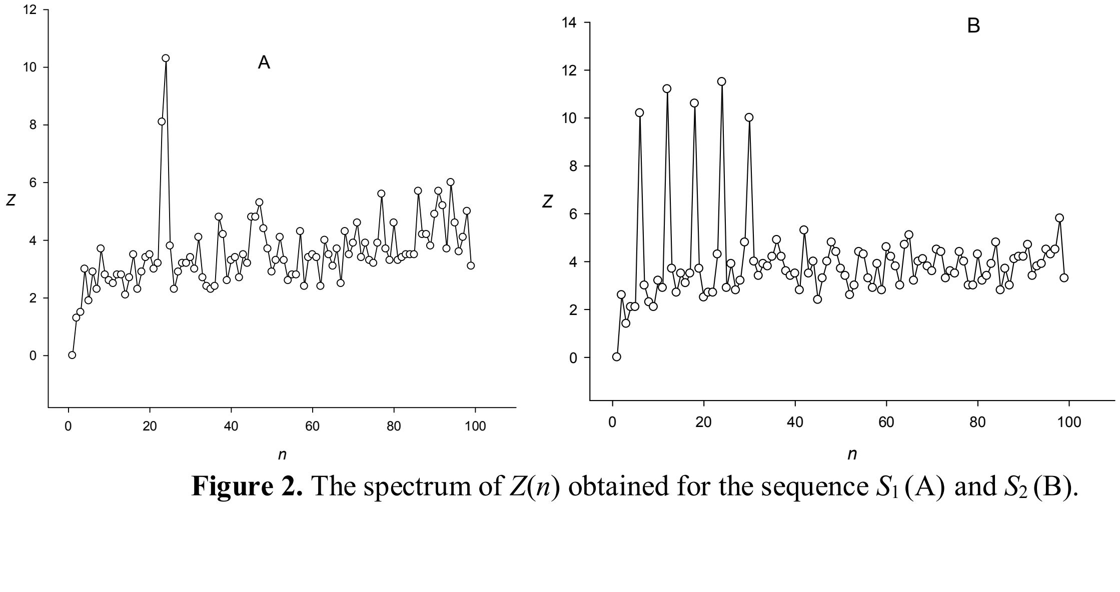

Figure 2. The spectrum of *Z*(*n*) obtained for the sequence S_1 (A) and S_2 (B).

The figure 2B shows $Z(n)$ for the sequence S_2. For the sequence S_2, periodicity was observed for n=48 which corresponds to 24 h also. This period was observed for the 970 to 3893 numbers of the sequence S_2. 44 insertions and deletions were made to find this period. The periodicity of sequence S_3 is shown in figure 5. It is possible to see the periods for n=6, 12, 18, 24 and 30. It corresponds to 24, 48, 72, 96 and 120 hours. The period equal to 96 h is the most statistically expressed ($Z \approx 11.5$). These results demonstrate that the periodicity equal to 24 hours can be seen for different candles (half hour, one hour and four hours). It is also seen that there is a little period equal to 4 days apart from a period of 24 hours since Z(24)> Z(6).

The question arises about the nature of the periodicity discovered in this work. The different periodic processes influence currency exchange rates. Man himself has inherent rhythms of different frequencies; for example, Halberg [21] classified the biological rhythms of humans according to the period length. He identified several groups of rhythms: 1. Low frequency group has periods from 4 days to 12 months; 2. Mid-range group has period from 20 to 72 h; 3. High frequency area has period less than 20 h. It can therefore be assumed that the observed periodicity of the exchange rate at 24 h, reflects the influence of mid-frequency rhythms on the exchange rate.

An attempt was made to find the periodicity for the sequence $a(i)$=$x_2(i)$-$x_1(i)$,, where $x_1(i)$ и $x_2(i)$ are separated by minutes in the interval from 1 to 5 min. No statistically significant periodicity was found in a given interval, for a "minute" of the sequence. Sequences S_1 and S_2 were also built for other time intervals with the same length. In these cases, the $Z(n)$ and behavior of the phase shifts were similar.

It was also interesting to explain the presence of insertions or deletions of symbols (it was impossible to find frequency without it) and the behavior of the phase shift for the periodicity of 24 h. Probably, there is great instability in the behavior of large human masses and this instability may create phase shifts with insertions or deletions. Also, it can be assumed that some events of public life may be the cause of phase shifts, insertions, or deletions. For a more accurate consideration of this issue, a separate study is required. The correlations between the phase shifts, insertions and deletions in the sequences S_1 and S_2 should be searched and the events of public life and other factors, are of both social and physical nature.

References

[1] Stoica P, Moses R 2005 *Spectral Analysis of Signals* (New-York: Prentice Hall)
[2] Struzik Z R 2001 Phys. *A Stat. Mech. Its Appl.* **296** 307
[3] Hamilton J D 1994 *Time Series Analysis* (Princeton: Princeton University Press)
[4] Benson G 1999 *Nucleic Acids Res.* **27** 573
[5] Stankovic R S, Moraga C and Astola J 1987 *Fourier Analysis on Finite Groups with Applications in Signal Processing and System Design* (New-York: John Wiley & Sons)
[6] Marple S L 1987 *Digital Spectral Analysis: With Applications* (London:Prentice Hall)
[7] Dodin G, Vandergheynst P, Levoir P, Cordier C and Marcourt L 2000 *J. Theor. Biol.* **206** 323
[8] Jackson J H, George R and Herring P A 2000 *Biochem. Biophys. Res. Commun.* **268** 289
[9] Coward E and Drabløs F 1998 *Bioinformatics* **14** 498
[10] Chechetkin V R and Turygin AYu 1995 *J. Theor. Biol.* **175** 477
[11] Korotkov E V, Korotkova M A and Kudryashov N A 2003 *Phys. Lett. Sect. A* **312** 198
[12] Korotkov E V, Korotkova M A and Tulko J S 1997 *Comput. Appl. Biosci.* **13** 37
[13] Chaley M B, Korotkov E V and Skryabin K G 1999 *DNA Res.* **6** 153
[14] Turutina V P, Laskin A A, Kudryashov N A, Skryabin K G and Korotkov E V 2006 *J. Comput. Biol.* **13** 946
[15] Korotkov E V, Korotkova M A and Rudenko V M 1999 *J. Mol. Model.* **5** 103
[16] Chizhevsky A L 1976 *The Terrestrial Echo of Solar Storms* (Moscow: Misl')
[17] Rastogi S C,Mendiratta N and Rastogi P 2006 *Bioinformatics Methods and Applications: Genomics, Proteomics and Drug Discovery* (Delhi: PHI Learning)
[18] Suvorova Y M Korotkova M A and Korotkov E V 2014 *Comput. Biol. Chem.* **53** 43
[19] Pugacheva E, Korotkov E V and Korotkov A E 2016 in *Proc. 9th Int. Jt. Conf. Biomed. Eng. Syst. Technol. BIOSTEC* pp.117–127 (Setúbal: Scitepress)

[20] Pugacheva V M, Korotkov A E and Korotkov E V 2016 *Stat. Appl. Genet. Mol. Biol.* **15 (**ahead of print**).**

[21] Halberg F 1969 *Annu. Rev. Physiol.* **31** 675