IOSR Journal of Mathematics (IOSR-JM)
e-ISSN: 2278-5728, p-ISSN: 2319-765X. Volume 14, Issue 4 Ver. III (Jul - Aug 2018), PP 25-30
www.iosrjournals.org

Periodicities in Discrete Time Series: The Case of Four Values

Farmakis Nikolaos[1], Makris C. Georgios[2]
[1]*Department of Mathematics, Aristotle University of Thessaloniki, Greece,*
[2]*Department of Mathematics, Aristotle University of Thessaloniki, Greece*
Corresponding Author:Makris C. Georgios

Abstract: *Suppose that we have a time series of data like the one with the form $X_1, X_2, X_3, \dots, X_N$. In the present paper, an effort takes place for searching on periodicities in the above series. More precisely this random variable X takes four distinct values: a, b, c and d. These values can be distributed in such a way that the periodicities can be seen even with the naked eye or, more likely, they may be latent ones.A special technique, proposed initially in 2009 and based on Systematic Sampling (SyS) is used to disclosure any periodicity latent or obvious. This technique helps to reveal periodicities and ultimately determines the period value T. The distribution of the values of the set {a, b, c, d} is not necessarily (discrete) uniform, although in the most popular applications it is the case, at least approximately.Several illustrative examples are given, in order to make clear the process of revealing the periodicity in the series with four discrete values.These can be exploited for the discovery of periodicities in chain of DNA. They can become a key parameter in a reading and identification effort for the chain of DNA.*

Keywords -*Frequency, Time series, Periodicity, DNA*

Date of Submission: 27-07-2018 Date of acceptance: 11-08-2018

I. Introduction

A key property of a time series [1],[2],[3],[4],[5]$X_1, X_2, X_3, \dots, X_N$. is it presents or not some kind of periodicity. There are time series with periodicity revealed with a naked eye. But there are some series where the periodicity (and the period of course) should be investigated by a systematic way which undoubtedly highlights the existence of any periodicity [6],[7],[8],[9].

One way for the enhancement of any periodicity in time series and determining the period T, is based on the use of Systematic Sampling (SyS) [10] [5]. The Target Population for this is the above time seriesn [11]. It has turned out that the use of SyS can reveal the existence or no of some periodicity in time series data[11]. This method calculates the period T, automatically. The data of the time series can be real numbers, generally. It is sometimes also of special interest, to reflect on a time series with respect to the periodicity, when a random variable X representing the time series takes its values from a set with a small number of items, e.g. A = {1,2,3,4}. This may be useful in the study of the DNA sequence, etc [12],[13],[14],[15],[16].

The case to have time series with discrete values (and specially) the above 4 is addressed specifically in this paper in theory and examples. In Farmakis work [11] the time series with two discrete values was fully studied.

II. TIME SERIES 1-2-3-4

We consider the time series $X_1, X_2, X_3, \dots, X_N$, with length value N large enough, e.g. $N > 500$ and where the values of $X_m, m = 1,2,3, \dots, N$ are random variables (rv) taking the values 1,2,3 and 4 only. Also, we are interested to see if the values of the time series have a periodicity with a period T, i.e. we like to see if applicable $X_{r+T} = X_r, r = 1,2,3, \dots, T$. To achieve this finding may help a simple view on the data. This view sometimes is not sufficient.

See the following example:

Example 1:

We have the following single-digit time series data is written into two tables-delta with different length marking line. In Table 1 we have k = 15 digits-data time series per line. In Table 2 we have n = 42 digits per line.

Table 1: *Line Length k=15*

123342312334231
233423123342312
334231233423123
342312334231233
423123342312334
231233423123342
312334231233423
123342312334231
233423123342312
334231233423123
342312334231233
423

Table 2:*Line Length k=42*

1233423123342312334231233423123342312334231233423
1233423123342312334231233423123342312334231233423
1233423123342312334231233423123342312334231233423
1233423123342312334231233423123342312334231233423

The fact that we are writing the data of the time series in terms of length k means that we try to have systematic samples of size n taken from N = 168 elements of the time series (part). Some samples have of size $n = \left\lfloor \frac{N}{k} \right\rfloor$ and for some of them have $n = \left\lfloor \frac{N}{k} \right\rfloor + 1$ where the symbol $\lfloor x \rfloor$ is the integer part of the real number x, Farmakis (2009a), chap. 4th pp 116-117 [10].

We see that in Table 2 each column-sample consists of equal elements and of course the mean value of the sample is equal to the unique element of the sample, i.e. $1 \leq \bar{x}_b \leq 4$. Also, in Table 1 and in each column-sample there are elements of several sizes from {1,2,3,4}. So the mean value of the sample is $1 < \bar{x}_a < 4$. Obviously, it follows that $Var\bar{\bar{x}}_a < Var\bar{\bar{x}}_b$. The difference between the two variances of the samples is very large. We face variance of two different class of magnitude. In the present sample of n = 168 elements are:

1st) In Table 1 mean value of 15 sample mean values is $\bar{\bar{x}}_a = \frac{424}{165} = 2.5697$ and the dispersion of the 15 sample values is $Var\bar{\bar{x}}_a = 0.0187$.

2) In Table 2 the mean value of the 42 means is $\bar{\bar{x}}_b = \frac{18}{7} = 2.5714$ and the dispersion of the 42 of sample means values is $Var\bar{\bar{x}}_b = 0.8163$. Obvious that $Var\bar{\bar{x}}_a < Var\bar{\bar{x}}_b$ and for the class of the magnitude we have: $\frac{Var\,\bar{\bar{x}}_b}{Var\,\bar{\bar{x}}_a} = \frac{0.8163}{0.0187} = 43.65$.

It is essential to note that from the Table 2 we have the obvious conclusion that the time series period is T = 7. In the contrary from Table 1 is not an easy thing a conclusion like this.

Example 2:

We have the following single-digit time series data, according to Example1 data written back into "pages" of different widths, the tables 3 and 4 Population size N = 168.

Table 3: *Line Length k=15*

123412341234123
412341234123412
341234123412341
234123412341234
123412341234123
412341234123412
341234123412341
234123412341234
123412341234123
412341234123412
341234123412341
234

Mean Value $\overline{x}_a = 2.5$ and VEariance $Var\bar{\bar{x}}_a = \frac{1}{11^2} = 0.008264$

Table 4: *Line Length k=40*

1234123412341234123412341234123412341234 1234123412341234123412341234123412341234 1234123412341234123412341234123412341234 1234123412341234123412341234123412341234 12341234
Mean Value $\overline{x}_b = 2.5$ Variance $Var\bar{\bar{x}}_b = \frac{5}{4} = 1.25$

In this case two things occur essentially:

1st) It is noted with naked eye a periodicity with period T = 4, based on the ("page") length k = 40 in table 4 lines, which is a multiple of four.

2) The dispersion of the sampling average multiples in Table 4 than it is in Table 3. In fact we have not just $Var\bar{\bar{x}}_a < Var\bar{\bar{x}}_b$ but $\frac{Var\,\bar{\bar{x}}_b}{Var\,\bar{\bar{x}}_a} = \frac{1.25}{0.008264} = 151.25$.

III. A More General View, Algorithm

The conclusions of the two examples of the previous paragraph bring to the surface a more general but important conclusion for the time series and, more generally, for numbered data, which show some periodicity. In general, this conclusion is summarized and proved in [11] as:

"*When the data length k is a multiple of the T period then the dispersion of the sample mean value of the SyS becomes very large. It becomes disproportionately greater than the dispersion we have for the sample mean value when the data length k is not a multiple of the T period*"

It is clear that the time series $X_1, X_2, X_3, \dots, X_N$ can be written on pages with an index line length equal to $k = 2,3,4, \dots, \left\lfloor \frac{N}{2} \right\rfloor$.

When k = T, or when k is a multiple of T, then there is a disproportionately large dispersion value of the sample mean value $Var\bar{\bar{x}}$.

The above finding leads to the computation of the following algorithm:

Step 1: For each value of the display line length k, we create the corresponding table of k-column samples.

Step 2: We find for each sample column of the length table k the mean value and we have k mean values.

Step 3: For each length table k we find the dispersion of the k values of step 2, $Var\bar{\bar{x}}(k)$.

Step 4: By studying the values $Var\bar{\bar{x}}(k)$, we try to see for which k values we have disproportionately large values of the dispersion. If there is a periodicity of the time series values, say period T, then there will be such extreme large values of the dispersion of the sampling mean value for k=T length of display. The fraction of the extreme value of the dispersion to any other dispersion (with length k not a multiple of T) is usually a few tens and depends quite significantly on the range of $R = X_{max} - X_{min}$ of the time series values. □

In all the above it is generally assumed that in the time series $X_1, X_2, X_3, \dots, X_N$ there are no deviations from the definition rule of $X_\lambda = X_{\lambda+\rho T}, \lambda = 1,2,3, \dots, T \; and \; \rho = 1,2,3, \dots$

Here comes the question: What will happen if we have deviations from the rule $X_\lambda = X_{\lambda+\rho T}$? The next paragraph tries to give some answers, mainly through examples. There are some outlets and the general problem has not yet found the complete solution.

IV. Existence Of Derogations In Frequency

When the time series is periodic, its definition equation applies:

$$X_{\lambda+\rho T} = X_\lambda, \lambda = 1,2,3, \dots, T \; and \; \rho = 1,2,3, \dots \quad (4.1)$$

This is the case without derogations. If any derogations exist, then the relation (4.1) takes the form

$$X_{\lambda+\rho T} = X_\lambda + e_{\rho,\lambda}, \lambda = 1,2,3, \dots, T \; and \; \rho = 1,2,3, \dots$$

If it is $e_{\rho,\lambda} = 0$, then they apply almost as they are. Simply the fraction of dispersions of the sample mean values $\frac{Var\,\bar{\bar{x}}_b}{Var\,\bar{\bar{x}}_a}$ in paragraph 2 is a little less than the ones in paragraph II.

If $e_{\rho,\lambda} \neq 0$, then it is interesting to see the quantity $Ee^2_{\rho,\lambda}$ (in principal) in relation to the mean value EX of the time series. Sometimes a deviation of squares of a size like $\frac{Ee^2_{\rho,\lambda}}{EX} = 0{,}30$ leaves indifferent the periodicity as it is found through the algorithm.

Some more examples:

Example 3:

We have a series of numbered elements with values of the set A = {1,2,3,4} of length N = 1000 and its data follows totally and iteratively the rate 12341234 ... This is the time series X-1

From the above series another length is produced, again 1000 whose elements are the elements of X-1 except that in 21 positions randomly selected we have moved the elements of these positions to one of their neighbors. The new series is the X-2 time series. The sum of the squares of the differences in the two time series components is 122. Obviously the two rows are identified at 958 points. What is running on if we apply the algorithm in paragraph 3 to the two series?

We note that it is $Ee^2_{\rho,\lambda} = 0.122$, and $\frac{Ee^2_{\rho,\lambda}}{EX} = \frac{0.122}{2.5} = 0.0488 < 0.05$, that is a very small relative deviation of X-2. Such a deviation leaves almost intact the X-2 periodicity character that behaves almost the same as X-1, as shown by the two figures at the end of this example. Note that from X-1 we can take with other interventions and other series like X-2. Indeed, with similar interventions, we took four other time series, X-3, X-4, X-5 and X-6. The following table 5 gives the sums of squares of differences (deviations) of 6 rows and for each of them:

Table 5:*Sums of squares of differences (deviations)for X-1,X-2,..., X-6*

Time Series	***X*-1**	***X*-2**	***X*-3**	***X*-4**	***X*-5**	***X*-6**
$\sum e^2$	0	122	149	440	109	100
$Var_{max}\,\bar{\bar{x}}$	1.6667	1.5083	1.4867	1.1316	1.5299	1.5632

A simple observation shows that the quantities $\sum e^2$ and $Var_{max}\,\bar{\bar{x}}$are connected with a decreasing function. Calculations have shown that the relationship is linear with a linear correlation coefficient r = -0.9987, i.e. that 97.74% of its variations $Var_{max}\,\bar{\bar{x}}$ are due to the variations of the derogations and $Var_{max}\,\bar{\bar{x}}$ is descending against derogations. The relative regression equation is:

$$Var_{max}\,\bar{\bar{x}} = 1.6682 - 0.00122(4.2)$$

We also list 2 of the 6 figures illustrating the periodic nature of the time series. They correspond to X-1 and X-2:

Figure 1:*The periodic nature of the time seriescorrespond to X-1*

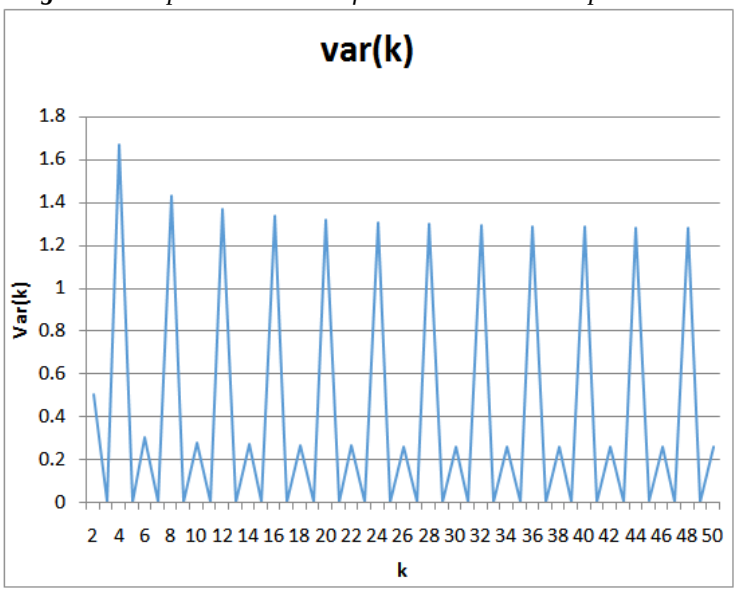
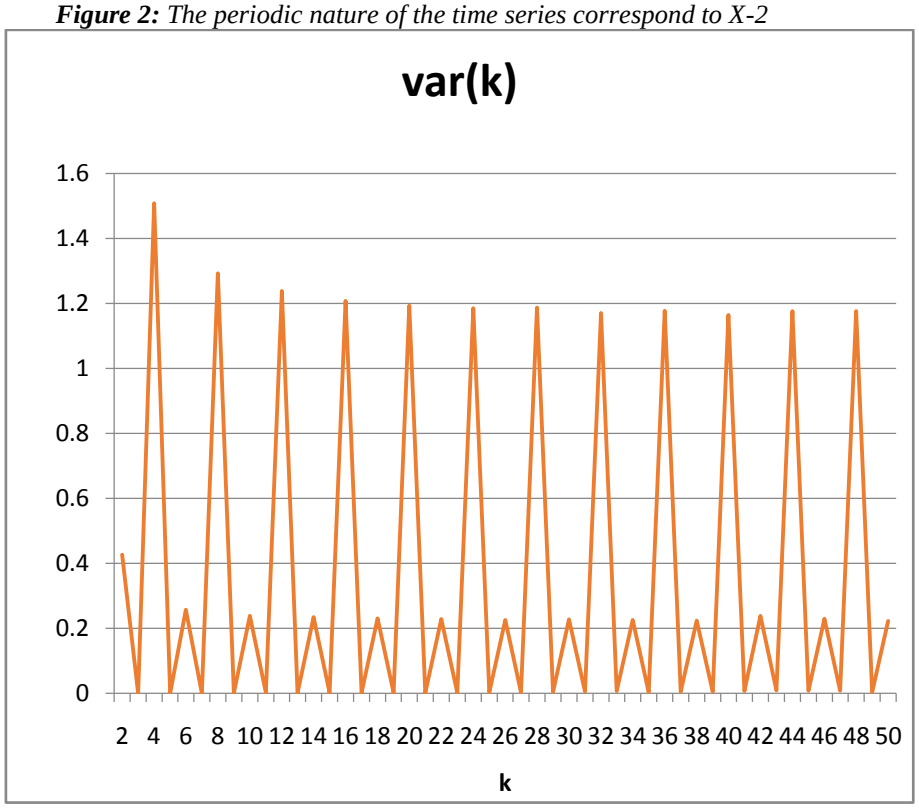


Figure 2: *The periodic nature of the time series correspond to X-2*



V. Implementation Algorithm By Matlab

To implement the algorithm a MatLab program was developed which accepts the time series x and returns a table (pin_var) with fluctuations for each $k = 2,3,4, \dots, \left\lfloor \frac{N}{2} \right\rfloor$. The graph of this table (pin_var) gives us figures 1 and 2. From where one can very easily discover the periodicity of the time series.

```
[N M]=size(x);
pin_var=[];
for k=2:N/2
i=N/k;
   p=zeros(i,k);
ti=1;
for i1=1:i
for k1=1:k
      p(i1,k1)=x(ti,1);
ti=ti+1;
end
end
   x1=[];
for k1=1:k;
tempx=mean(p(:,k1));
      x1=[x1;tempx];
end
pin_var=[pin_var;kvar(x1)];
end
```

VI. Conclusion

In this paper we presented a relatively simple algorithm for calculating time series periodicity with discrete values.

In case of non-numeric data (discrete data), then we transform the data into numeric by replacing each symbol with a number of 1,2,3, ...

The T-calculation algorithm is based on the use of Systemic Sampling (SyS). The use of the SyS may reveal the occurrence or non-periodicity of the elements of a time series and the method also automatically calculates the value of the T period.

The study was done on the elements of a time series with respect to its periodicity with elements of a particular set with a small number of elements: A = {1,2,3,4}. These elements could be a sequence of DNA.

References

[1] Hamming, Richard. Numerical methods for scientists and engineers. Courier Corporation, 2012.
[2] Kantz, Holger; Thomas, Schreiber (2004). Nonlinear Time Series Analysis. London: Cambridge University Press. ISBN 978-0521529020.
[3] Abarbanel, Henry (Nov 25, 1997). Analysis of Observed Chaotic Data. New York: Springer. ISBN 978-0387983721.
[4] Boashash, B. (ed.), (2003) Time-Frequency Signal Analysis and Processing: A Comprehensive Reference, Elsevier Science, Oxford, 2003 ISBN 0-08-044335-4
[5] Cochran W., (1977) "*Sampling Techniques*", John Wiley & Sons, New York.
[6] Artis M, Hoffmann M, Nachane D, Toro J. The detection of hidden periodicities: A comparison of alternative methods. 2004. p. 10. EUI Working Paper ECO.
[7] Benedetto JJ, Pfander GE. Periodic wavelet transforms and periodicity detection. SIAM Journal of Applied Mathematics. 2002;62(4):1329–68.
[8] Okamura H, Semba Y. A novel statistical method for validating the periodicity of vertebral growth band formation in elasmobranch fishes. Canadian Journal of Fisheries and Aquatic Sciences. 2009;66(5):771–80.
[9] Hogg RV, McKean JW, Craig AT. Introduction to Mathematical Statistics. 6th ed. Peason Prentice Hall; 2005
[10] Farmakis N., (2009a) "*Introduction to Sampling*", A & P Christodoulidi, Thessaloniki, (in Greek)
[11] Farmakis N., (2009b) "Searching for Periodicities in Data Series", *Statistics in Transition, new series*, Vol. 10, No 2, pp 317-335.
[12] Korotkov E. V., Korotkova M. A. (1995) "Latent Periodicity of DNA Sequences from Some Human Gene Regions", *DNA Sequence-The Journal of Sequencing & Mapping*, Vol. 5, pp 353-358.
[13] M. Buchner and S. Janjarasjitt, "Detection and visualization of tandem repeats in DNA sequences", IEEE Trans. SP, Vol. 51, No. 9, pp. 2280-2287, September 2003.
[14] J. Butler, "Forensic DNA typing: biology and technology behind STR markers", Academic Press, 2003.
[15] D. Holste and I. Grosse, "Repeats and correlations in human DNA sequences", Physical Review E, 67, 2003.
[16] J. K. Perkus, "Mathematics of genome analysis", Cambridge University Press, 2002.

Makris C. Georgios."Periodicities in Discrete Time Series: The Case of Four Values. IOSR Journal of Mathematics (IOSR-JM) 14.4 (2018) PP: 25-30.