STATISTICS IN TRANSITION-new series, October 2009
Vol. 10, No. 2, pp. 317—335

SEARCHING FOR PERIODICITIES IN DATA SERIES

Nicolas Farmakis[1]

ABSTRACT

A new idea for searching and mining latent periodicities from labeled data (time series, aminoacid or DNA sequences, etc.) is proposed. This new method is based on systematic sampling procedures. A materializing algorithm is proposed and some supporting theoretical results are given. This method is very useful for biologists, environmental researchers, financial researchers and managers, etc., etc. Also some illustrating examples are given.

Key words: Systematic Sampling, Periodicity, Data Series, DNA sequence.

1. Introduction

Suppose that we have a series of labeled data of size N:

$$X_1, X_2, X_3,\ldots,X_N. \qquad (1.1)$$

The size of N runs from some decades to some millions or billions (e.g. banking or stock exchange data series or meteorological data series, etc.).

The values in (1.1) are arithmetic ones but they may represent elements of a symbolic sequence, e.g. aminoacid or DNA sequences, Pasquier, et al. (1998), Korotkov and Korotkova (1995).

The values of the series in (1.1) are taken to be random, i.e. we have a random variable X (rv X) with its N values in implementation. The implemented values of labeled data in (1.1) may represent the temperature values at 12h 00m (noon) of every day for 50 years, i.e. N=18250 more or less. Also (1.1) can represent the humidity in a meteorological station of a city or the noise level in a place in the city center. Obviously, there are many cases of labeled data series like the price of gross oil per barrel, in USA dollars, during the last 40 years, etc., etc.

[1] Aristotle University of Thessaloniki, Department of Mathematics, GR-54124 – Thessaloniki GREECE, farmakis@math.auth.gr.

Having to study on such a labeled series as the one in (1.1), we deal (among many other things) with periodicities of the data. The next definition is useful for the purposes of the present paper.

Definition1.1: A series of labeled data, like the one in (1.1), is called *periodic*, if there is an integer $T>1$ for which we have the equality

$$X_{i+T}=X_i,\ \forall\ i=1,2,3,\ldots,N-T \tag{1.2}$$

or more generally

$$X_{i+nT}=X_i,\ \forall\ i, n \text{ such that } i, i+nT\in\{1,2,3,\ldots,N\} \tag{1.3}$$

The integer T is the ***period*** of series (1.1). ∫

Definition (1.1) says that "in a periodic labeled data series we have only T different values for its element, even if the number N is in some cases a very big number, e.g. $N=10$ millions". Moreover, if we write the values of the series (1.1) in a table with T columns, we face the same value as we run on the j^{th} column, $j=1,2,3,\ldots,T$ of this table with the T columns. If the 1^{st} element of the j^{th} column is the X_j, then all the elements in this column are $X_{j+(m-1)\cdot T}=X_j,\ m=1,2,3,\ldots,[\frac{N}{T}]+b$, where b is 1 for the first T-1 (at most) columns and 0 for all the next columns. Also $[x]$ stands for the integer part of the real number x. Thus, the mean value of Xs in the j^{th} column drawn as a systematic sample (see Definition 1.2 below) is $\bar{x}_j = X_j$ and the sample variance is $\sigma_j^2 = 0,$ for all the values $j=1,2,3,\ldots,T$.

Remark 1.1: Due to many sources of errors and due to several kinds of data noise the relations $X_{j+mT}=X_j,\ m=1,2,3,\ldots,[\frac{N}{T}]+b$, $\bar{x}_j = X_j$ and $\sigma_j^2 = 0$, seem to be more or less ideal and it is better to substitute them by the more realistic ones in the every day applications of a researcher: $X_{j+mT}\approx X_j,\ m=1,2,3,\ldots,[\frac{N}{T}]+b$, $\bar{x}_j \approx X_j$ and $\sigma_j^2 \approx 0$.

Finally, if the row length of the table, i.e. the number of columns-samples, is $k \neq T$, then in every column we face at least two different values of data and so the variance grows up quickly. In any case is $\sigma_j^2 > 0$. **o**

®

We need the next definition for the systematic sample:

Definition1.2: Suppose we have a series of labeled data $X_1, X_2, X_3,\ldots,X_N$. We write these data with the hierarchy of labels in a table of k columns and n rows. If k is a divisor of N, then $n=\left[\frac{N}{k}\right]$ exactly. In other cases $n=\left[\frac{N}{k}\right]+1$ and the last row is not full completed. The last row contains $u=N-k\cdot n>0$ elements and its last $k-u$ places are empty. Every column of this table is called "*a systematic sample*" with step k. Obviously $k>1$. ∫

Corollary1.1: The element staying on the *(i,j)* place of the table of definition 1.2 is the $\{(i-1)\cdot k+j\}^{th}$ element of the data series in (1.1).

Proof: Obviously, before the i^{th} row of the table there are $(i-1)\cdot k$ elements belonging to the previous $i-1$ rows. Up to the element situated in the *(i,j)* place of the table we have j more elements. If we represent this element of the table by $Y_{i,j}$, we have $Y_{i,j}=X_{(i-1)\cdot k+j}$, and the proposition is proved. ◊

◊

2. Dealing with systematic samples of labeled data

In what follows we call the table in definition 1.2 a ***systematic table***. It is also more convenient to imagine that k is a divisor of N. All the results remain the same for the other case and we will make some specializations if there is any need. We suppose that we work on a labeled data series like the (1.1) with period $T>1$. The calculations will take place in the ideal (theoretic) field. In the sense of remark 1.1, no statistical error neither data noise is supposed.

In the case we have $k=T$ systematic samples of size n, i.e. $N=n\cdot k= n\cdot T$ is the size of the series in (1.1) and the number of elements of the systematic table. This also means that we have the next relations for every column $j=1,2,3,...,k$ of the table:

$$X_{j+m\cdot T}=X_j,\ m=1,2,3,\ldots,n-1,\quad \overline{x}_j(k)=X_j \quad \text{and} \quad \sigma_j^2(k)=0\,. \tag{2.1}$$

It is quite easy to note that the mean values of the samples are different and thus their variance is not zero, in general. It is given by

$$Var\overline{x}(T)=\sigma_{\overline{x}}^2(T)=E(\overline{x}(T)-\overline{X}(T))^2>0\,,\quad \overline{X}(T)=T^{-1}\cdot\sum_{j=1}^{T}\overline{x}_j(T) \tag{2.2}$$

Obviously, we have the same results with $k=\lambda\cdot T$ in (2.2). We are going to prove:

Theorem 2.1: For k=T we have

$$Var\overline{x}(T)=\sigma_{\overline{x}}^2(T)>Var(\overline{x}'(T))\geq 0 \tag{2.3}$$

$\overline{x}$ is the sample mean value before any permutation of elements between any two columns and $\overline{x}'$ is the sample mean after some permutations of elements between at least two columns of the systematic table.

Consequently

$$Var\overline{x}(T)=\sigma_{\overline{x}}^2(T)>Var(\overline{x}(k))=\sigma_{\overline{x}}^2(k)\geq 0\,,\ k\neq T\,,$$

i.e. finally it is $\max_k Var(\overline{x}(k))=Var(\overline{x}(T))$.

Proof: For $k=T$ and since all the elements of the j^{th} column are all equal we have

$$\bar{x}_j(T) = x_j \text{ , } j=1,2,3,...,T,$$

for every column-sample of the systematic table. This equality is destroyed for every other value of k.

Similarly, the equality is destroyed for every permutation of two at least elements belonging to different columns-samples.

Suppose now that, for any reason, the q^{th} sample exchanges an element with the r^{th} sample of the systematic table. This simply means that the mean values of the new systematic samples will be all the same with the previous except for the two above cases of order q and r. If we define the new sample means as $\bar{x}'_j(T)$, we have of course $\bar{x}'_j(T) = \bar{x}_j(T)$, $j \neq q, r$, and $\bar{x}'_t(T) \neq \bar{x}_t(T),\ t = q, r$. Thus, we arrive to the next two quantities for the variances of the sample means:

$$Var\bar{x}(T) = M + \frac{1}{k}\left\{\left(\bar{x}_q(T) - \bar{X}\right)^2 + \left(\bar{x}_r(T) - \bar{X}\right)^2\right\},\ \ \bar{X} = N^{-1} \cdot \sum_{i=1}^{N} X_i \qquad (2.4)$$

and

$$Var\bar{x}'(T) = M + \frac{1}{k}\left\{\left(\bar{x}'_q(T) - \bar{X}'\right)^2 + \left(\bar{x}'_r(T) - \bar{X}'\right)^2\right\},\ \ \bar{X}' = N^{-1} \cdot \sum_{i=1}^{N} X'_i \,. \qquad (2.5)$$

Obviously, in (2.4) , (2.5) it is $\bar{X}' = \bar{X}$ because the summation is over the same set of values via a different order only. Thus, we need to compare only the two next quantities:

$$k \cdot Q_1 = \left(\bar{x}_q(T) - \bar{X}\right)^2 + \left(\bar{x}_r(T) - \bar{X}\right)^2$$

and

$$k \cdot Q_2 = \left(\bar{x}'_q(T) - \bar{X}\right)^2 + \left(\bar{x}'_r(T) - \bar{X}\right)^2 .$$

If we prove that $Q_1 > Q_2$, then the theorem is also proved.

First of all, we note that

$$\bar{x}'_q = \frac{n-1}{n} \cdot \bar{x}_q + \frac{x_r}{n} = \frac{n-1}{n} \cdot \bar{x}_q + \frac{\bar{x}_r}{n}$$

$$\text{and } \bar{x}'_r = \frac{n-1}{n} \cdot \bar{x}_r + \frac{x_q}{n} = \frac{n-1}{n} \cdot \bar{x}_r + \frac{\bar{x}_q}{n} \qquad (2.6)$$

and thus it is

$$Q_2 = \left(\frac{n-1}{n} \cdot \bar{x}_q + \frac{\bar{x}_r}{n} - \bar{X} \right)^2 + \left(\frac{n-1}{n} \cdot \bar{x}_r + \frac{\bar{x}_q}{n} - \bar{X} \right)^2 = \ldots =$$

$$Q_1 - \frac{2 \cdot (n-1)}{n^2} \cdot \left(\bar{x}_q - \bar{x}_r \right)^2 < Q_1 \text{ q.d.e.}$$

Obviously, for *k=2,3,4,...,T-1* we have $\bar{x}_j(k) \neq x_j$ for some values *j=1,2,3,...k,* due to permutations. After this remark it is obvious that the next relation is valid

$$\max_k Var(\bar{x}(k)) = Var(\bar{x}(T)) . \quad \Diamond \qquad \Diamond$$

Now, a sharper version of the above theorem 2.1 is when we have exchanges of elements among more than two columns-samples of a systematic table with *k* columns and *n* rows. The biggest difference between Q_1 and Q_2 takes place when (evidently) we have exchanges among all the *k* columns. An element from, e.g. the m^{th} column, goes to the r^{th} one and one from the r^{th} goes to another, say to q^{th} one, and so on until all the columns give an element to another and every column receives an element from a column of the systematic table. The next theorem 2.2 is related to the range of the difference between Q_1 and Q_2:

Theorem 2.2: For every k=2,3,4,…,T-1 we have

$$Var\bar{x}(T) = \sigma^2_{\bar{x}}(T) > Var(\bar{x}'(T)) = \sigma^2_{\bar{x}'}(T) \geq 0 \tag{2.7}$$

Consequently

$$Var\bar{x}(T) = \sigma^2_{\bar{x}}(T) > Var(\bar{x}(k)) = \sigma^2_{\bar{x}}(k) \geq 0 \tag{2.8}$$

$$\text{i.e. } \max_k Var(\bar{x}(k)) = Var(\bar{x}(T)) .$$

Proof: For *k=T* and for every column-sample of the systematic table we have the $\bar{x}_j(T) = x_j$, *j=1,2,3,...,T,* because all the elements of the j^{th} column are all equal. For this case with the *k=T* movements we have the equalities:

$Var\bar{x}(T) = \frac{1}{T}\sum_{q=1}^{T}\left(\bar{x}_q - \bar{X}\right)^2$ and $Var\bar{x}'(T) = \frac{1}{T}\sum_{q=1}^{T}\left(\bar{x}'_q - \bar{X}\right)^2$ where it is

$$\bar{x}'_q = \frac{n-1}{n} \cdot \bar{x}_q + \frac{x_r}{n} = \frac{n-1}{n} \cdot \bar{x}_q + \frac{\bar{x}_r}{n}, \text{ with } q \neq r, \quad q, r = 1, 2, 3, \ldots, T \tag{2.9}$$

and the one

$$\bar{x}'_q - \bar{X} = \frac{(n-1) \cdot x_q + x_r - n \cdot \bar{X}}{n}, \text{ with } q \neq r, \quad q, r = 1, 2, 3, \ldots, T \tag{2.10}$$

From (2.9) and (2.10) we obtain

$$Q_2 = T \cdot Var(\overline{x}'(T)) = \sum_{q=1}^{T}\left(\overline{x}' - \overline{X}\right)^2 = \frac{1}{n^2}\cdot\sum_{\substack{q=1\\ q\neq j}}^{T}\left\{(n-1)\cdot x_q + x_j - n\cdot\overline{X}\right\}^2 = \ldots =$$

$$= \frac{1}{n^2}\cdot\left\{\left(n^2-2n+2\right)\cdot\sum_{q=1}^{T}x_q^2 + T\cdot n^2\cdot\overline{X}^2 + 2\cdot(n-1)\cdot\sum_{\substack{q=1\\ j\neq q}}^{T}x_j\cdot x_q - 2\cdot n^2\cdot\overline{X}\cdot\sum_{q=1}^{T}x_q\right\} = \ldots =$$

$$= Q_1 - \frac{n-1}{n^2}\cdot\sum_{\substack{q=1\\ j\neq q}}^{T}\left(x_q - x_j\right)^2 = T\cdot Var(\overline{x}(T)) - \frac{n-1}{n^2}\cdot\sum_{\substack{q=1\\ j\neq q}}^{T}\left(x_q - x_j\right)^2 \text{ and (2.7) is proved.}$$

Note that for a $k \neq T$ some exchanges of elements take place among the columns which now are a little different than the above assumed with $k=T$. The exchanges are more complicated than the described for the relation (2.7) but more chaotic than the above mentioned and they provide mean values with less variation than for those in (2.7). So we expect to arrive to (2.8) with freedom to say that the next equality holds:

$$\max_k Var(\overline{x}(k)) = Var(\overline{x}(T)). \qquad \Diamond \quad \Diamond$$

All that was mentioned in this paragraph could be more clear by some examples like that in the next paragraph.

3. Examples on systematic samples of labeled data

Some examples on systematic samples of size n from a Population **P** of labeled data are written in rows of size k=2,3,..., $\left[\frac{N}{2}\right]$. For reasons of space, and for simplicity reasons too, we use population with relatively small size in the present paper.

Example 3.1: We have some (say 50) values of a random variable X over the time (time series), the x_m, m=1,2,3,…,50. We can of course write them in a table with rows of size k=2,3,4,…,25. We will try first to write them in a table with rows of size:

(a) k=11, in Table 3.1

Table 3.1. In the last row we see the variances of the (every) column-sample data and in the previous one we see the mean values of the respective column-sample data

1st	2nd	3rd	4th	5th	6th	7th	8th	9th	10th	11th
2	100	40	30	50	1	200	40	40	40	0
180	40	50	40	2	300	40	50	50	2	300
30	30	40	1	250	40	50	40	0	200	40
45	45	3	400	80	40	40	2	320	40	50
40	1	250	40	40	50					
59.4	**43.2**	**76.6**	**102.2**	**84.4**	**86.2**	**82.5**	**33.0**	**102.5**	**70.5**	**97.5**
4822	**1299**	**9718**	**27969**	**9347**	**14637**	**6158**	**449**	**21492**	**7774**	**18692**

and afterwards with rows of size

(b) k=10, in Table 3.2

Table 3.2. In the last row we see the variances of the (every) column-sample data and in the previous one we see the mean values of the respective column-sample data

1st	2nd	3rd	4th	5th	6th	7th	8th	9th	10th
2	100	40	30	50	1	200	40	40	40
0	180	40	50	40	2	300	40	50	50
2	300	30	30	40	1	250	40	50	40
0	200	40	45	45	3	400	80	40	40
2	320	40	50	40	1	250	40	40	50
1.2	**220.0**	**38.0**	**41.0**	**43.0**	**1.6**	**280.0**	**48.0**	**44.0**	**44.0**
1.2	**8200**	**20**	**105**	**20**	**0.8**	**5750**	**320**	**30**	**30**

Its obvious that we have two (among the 24) ways of presentation of the same data x_m, m=1,2,3,…,50. The second way enables us to know a periodicity of the values of data, with a period T=5. Every graph of the five values $x_{5(t-1)+r}$, r=1,2,3,4,5 is giving the same (about) scheme for all t=1,2,3,…,10. The same icon is coming out from the mean values of the columns of data in pre-last row of Table 3.2, i.e. period T=5. Every column of this table is a systematic sample from the population of the 50 data, taken as 1 from 10; see in Cochran (1977), Farmakis (2002), ch. 4th, Thompson (1992) and Thompson (1997). The variance of the sample mean in Table 3.2 is $var\bar{x}_2 = 8007.23$ and the mean value of the sample variances is $E(s^2) = 1448$ with one unit accuracy.

From the other hand the presentation of Table 3.1 encrypts the above mentioned periodicity of the data series. There is not any discretional power even from the row of means in Table 3.1. Thus, the variance of the sample mean in

Table 3.1 is very low i.e. $var\overline{x}_1 = 481.68$ and a mean value of the sample variances $E(s^2) = 11123$, with one unit accuracy.

It becomes more obvious now that a presentation of the above data in a table with rows of size k=5 or k=15 or in general of $k=5t$, t=1,2,3,4,5 gives always the opportunity to "see" the period $T=5$ coming out. In the opposite, a size of rows $k=5t+p \geq 2$, p=1,2,3,4, t=0,1,2,3,4 keeps out of our optical field, the truth of the existence of a period T=5. For instance, in Table 3.3 with row size k=8. As in Table 3.1, we cannot see the period as we are checking the 50 values of data in Table 3.3. We cannot also see this period in the pre-last row of mean values of the columns (systematic samples). It seems to the observer now that the period T=5 is "near" (let us say so) the row size k=10=5·2 in the presentation of Table 3.2 and the same period is "far" from the row size k=11 of Table 3.1, even if the Euclidean distance between 10 and 11 is less than between 10 and 5.

From the pre-last row of Table 3.3, with the mean values of the suitable columns, we get the variance of the sample mean as $var\overline{x}_3 = 193.51$. It is very small in comparison with the $var\overline{x}_2 = 8007.23$ got from the last row of Table 3.2. We also note that the mean value of the sample variances is $E(s^2) = 10657$ (accuracy=1).

Table 3.3. In the last row we see the variances of the (every) column-sample data and in the previous one we see the mean values of the respective column-sample data

1st	2nd	3rd	4th	5th	6th	7th	8th
2	100	40	30	50	1	200	40
40	40	0	180	40	50	40	2
300	40	50	50	2	300	30	30
40	1	250	40	50	40	0	200
40	45	45	3	400	80	40	40
2	320	40	50	40	1	250	40
40	50						
66.29	**85.14**	**70.83**	**58.83**	**97.00**	**78.67**	**93.33**	**58.67**
10942	**11563**	**8024**	**3828**	**22350**	**12674**	**10867**	**5011**

Example 3.2: We have the same values of a random variable X over the time, the x_m, m=1,2,3,…,50, as in the previous example. We try to write them in a table with rows of size k=2,3,4,…,25 (in general it must be k=2,3,4,...,$\left[\frac{N}{2}\right]$, here it is N=50) and we found the results of the next Table 3.4, dealing with the same data:

Table 3.4.

k	2	3	4	5	6	7	8	9	10	11
$Var\bar{x}_k$	27	10	98	**7822**	151	290	194	130	**8007**	482
$E(s^2)$	9520	9718	9715	**1493**	10254	10349	10657	11095	**1448**	11123

Table 3.4. (continued)

k	12	13	14	15	16	17	18	19	20
$Var\bar{x}_k$	645	948	1395	**8386**	1403	2471	2333	2558	**8749**
$E(s^2)$	11032	10709	11700	**1534**	11580	10997	10780	12331	**1101**

Table 3.4. (continued)

k	21	22	23	24	25
$Var\bar{x}_k$	2849	3958	3839	3283	**8401**
$E(s^2)$	12306	10904	10547	12274	**1529**

From the point of view of $Var\bar{x}_k$, the values of row size $k=5t$, t=1,2,3,4,5 seem to be isolated among all the values of row size from 2 to 25. This result becomes sharper if we take as critical index the ratio $\frac{Var\bar{x}_k}{E(s^2)}$. This is obvious because the relatively small values of $Var\bar{x}_k$ are to be divided by the very big values of $E(s^2)$ and the (much) bigger values of $Var\bar{x}_k$ have to be divided by the relatively little values of $E(s^2)$.

From Table 3.4 we conclude that the values of row size equal to $k=5t$, $t=1,2,3,4,5$ correspond with very big values of $Var\bar{x}_k$. They are big values in comparison with the values of $Var\bar{x}_k$ for $k=5t+s$, $s=1,2,3,4$, so that they seem to be isolated among the 24 values from 2 to 25. This result means that we take that the period is:

$$T=5=min\ \{k:\ \text{with comparably big value of}\ Var\bar{x}_k\ \}.$$

Resuming the results of all the tables in example 3.1 and 3.2 we have some interesting conclusions:

1st) In order to affirm the existence of a periodicity among the data series (expressed by the random variable X) an observer must see the $Var\bar{x}_k$ for all the values of $k=2,3,4,\ldots, \left[\frac{N}{2}\right]$, where k is the step of the Systematic Sampling Procedure (SySP), i.e. the row size of the systematic table. The observer checks

the values of k with extremely big values of $Var\bar{x}_k$. This is because these sizes of $Var\bar{x}_k$ are the results of a (kind of) timbre between the values T of the period and k of the row size of the systematic table, when it is $k=\lambda \cdot T$, λ is an integer like 1, 2, 3,…

2nd) The quality of estimation for the mean value of data is the worst one when we choose in SySP such a sample size n as the period T is a divisor of $k=\left[\frac{N}{n}\right]$. Thus, a profit of the knowledge of the value of period T is to avoid confusion via those suitable values of n. The best would be to choose $k < \frac{T}{2}$ or at least $k<T$.

4. An algorithm mining periodicities

The previous paragraphs and their examples and conclusions lead directly to a four steps algorithm. This algorithm is based on the results until now, especially the remark that there is a timbre between T and k.

With the next algorithm we ask the computer to run for all values of k from 2 to $\left[\frac{N}{2}\right]$, in order to compare the values of variance when we have timbre with the others when we have not any timbre:

Algorithm 3.1:

We imagine all data of a series put in a table with row size k. Every column of the table is a systematic sample of size $n=\left[\frac{N}{\kappa}\right]$ or, for at most $k-1$ samples of size $n+1$. The random variable A represents the data with values a_j, $j=1,...,N$.

Step 1: For every $k=2,3,4,\ldots, \left[\frac{N}{2}\right]$ find the mean values $\bar{a}_j(k)$, $j=1,2,3,\ldots,k$ for all the suitable samples.

Step 2: For every $k=2,3,4,\ldots, \left[\frac{N}{2}\right]$ calculate the variance of the mean value estimator of Step 1, i.e. $Var\bar{a}(k)$.

Step 3: Check the values of k with extremely big values of the corresponding variance $Var\bar{a}(k)$, (write a list, e.g. as the two first rows of table 3.4).

Step 4: Find the value of $k_0=min\ \{k$: with a big suitable value of $Var\bar{a}(k)\}$.

Stop

End

If there is any (integer) period T for the checked data then it is $T= k_0$, in 100% of the cases. Some problem arises when T is not an integer. If (e.g.) it is $T=5.5$, then the algorithm will give the value $T=11$ as a period. In this case we get (probably) for $k=5$ and $k=6$ relatively big values of $Var\bar{a}(k)$. More complicated cases are faced as the quantity $T-[T]$ is different from 0 or 0.5, given that the symbol $[x]$ is the bigger integer which does not exceed the real x.

5. Sharper processes for mining periodicities

We are going to give a more sharp process for mining periodicities. This process will be based on some observations on tables 3.2, 3.4 and on 5.1 and 5.2. This leads us to some constitutive and very useful remarks afterwards. The basic remarks are:

Remark 5.1: Looking on tables 3.2 and 5.1 and more over on 3.4 we can state:

"If $k=\lambda \cdot T$, $\lambda=1,2,3,\ldots$ then we have the maximum of the variance of the mean values of the samples (i.e. max $Var\bar{a}(k)$) and also we have the minimum of the mean value of the sample variances (i.e. min $E\left(s^2(k)\right)$)". ®

Remark 5.2: Also looking on tables 3.2 and 5.1 and more over on 5.2 we can state:

"If $k=\lambda \cdot T$, $\lambda=1,2,3,\ldots$ then we have the maximum of the variance of the mean values of the samples (i.e. max $Var\bar{x}_k$) and also we have the minimum of the mean value of the sample coefficient of variation (i.e. min $E\left(CV(A(k))\right)$)". ®

The above remarks lead to two new indices for mining periodicities. These indices could be from 3.4 the $\frac{Var\bar{x}_k}{E\left(s^2\right)}$ and from 5.2 the $\frac{Var\bar{x}_k}{E(CVy_k)}$. We are going to prove the sharpness of them which is equivalent with the outgrowth of the gap between the relatively very big values of $Var\bar{x}_k$ (when T is an divisor of k) and the other values of $Var\bar{x}_k$ (when T is not an divisor of k) which are relatively very small.

Table 5.1.

1st	2nd	3rd	4th	5th
2	100	40	30	50
1	200	40	40	40
0	180	40	50	40
2	300	40	50	50
2	300	30	30	40
1	250	40	50	40
0	200	40	45	45
3	400	80	40	40
2	320	40	50	40
1	250	40	40	50
1.4	**250.0**	**43.0**	**42.5**	**43.5**
0.9	**7200**	**179**	**63**	**23**

In the last row we see the variances of the (every) column-sample data, accuracy=1 and in the previous one we see the mean values of the respective column-sample data with accuracy=0.1.

Table 5.2.

k	2	3	4	5	6	7	8	9	10	11
$Var\bar{x}_k$	27	10	98	**7822**	151	290	194	130	**8007**	482
$E(CVx_k)$	1.27	1.29	1.26	**0.33**	1.29	1.32	1.31	1.31	**0.35**	1.20

Table 5.2. (continued)

k	12	13	14	15	16	17	18	19	20
$Var\bar{x}_k$	645	948	1395	**8386**	1403	2471	2333	2558	**8749**
$E(CVx_k)$	1.14	1.10	1.08	**0.28**	0.93	1.03	1.02	1.04	**0.16**

Table 5.2. (continued)

k	21	22	23	24	25
$Var\bar{x}_k$	2849	3958	3839	3283	**8401**
$E(CVx_k)$	0.81	0.95	0.97	0.79	**0.32**

From the point of view of $Var\bar{x}_k$, the values of row size $k=5t$, $t=1,2,3,4,5$ seem to be isolated among all the values of row size from 2 to 25. This result becomes more sharp if we take as critical index the ratio $\frac{Var\bar{x}_k}{E(CVx_k)}$. This is obvious because the small values of $Var\bar{x}_k$ are divided by the very big values of $E(s^2)$ and the (much) bigger values of $E(CVx_k)$ have to be divided by the relatively small values of $E(CVx_k)$.

Theorem 5.1: If we have periodic labeled data, then the gap between the extremely big values of $Var\bar{x}_k$ and its other common values is smaller or equal to the gap of the corresponding values of the parameter $\frac{Var\bar{x}_k}{E(s^2)}$.

Proof: Obvious, because we divide the maximum of $Var\bar{x}_k$ (nominator) with the minimum of $E(s^2)$ (denominator). ◊

Theorem 5.2: If we have periodic labeled data, then the gap between the extremely big values of $\boldsymbol{Var\bar{x}_k}$ and its other common values is smaller or equal to the gap of the corresponding values of the parameter $\frac{\boldsymbol{Var\bar{x}_k}}{\boldsymbol{E(CVx_k)}}$.

Proof: Obvious, because we divide the maximum of $\boldsymbol{Var\bar{x}_k}$ (nominator) with the minimum of $\boldsymbol{E(CVx_k)}$ (denominator). ◊

6. The trivial case *T=2*

Suppose we have labeled data with period T=2.

This is equal to the supposition that the series of these data has the form a, b, a, b, a, b,….a, b,… where a, b are real numbers or that we have

$$X_{2t-1}=a \text{ and } X_{2t}=b,\ t=1,2,3,\ldots$$

The population of all these values has N elements. N could be odd or even.

All that was mentioned above lead to the idea of a series with the least period T=2, i.e. we deal with the trivial case of the labeled data series.

We are going to study this data series and to prove again that the variance of the sample mean values takes its maximum value when T is a divisor of k.

We have obviously the two cases for the population size:

$$N=2\cdot\lambda+1 \text{ and } N=2\cdot\lambda,\ \lambda=0,1,2,3,\ldots.$$

In order to be more precisely to the environment of the present problem we adopt the next symbolism for the population size N:

$N=T\cdot k\cdot\lambda+2\cdot\theta+1$ and $N=T\cdot k\cdot\lambda+2\cdot\theta$, θ=0,1,2,3,..., λ=1,2,3,… and k=2,3,4,…

Thus, the minimum value for N seems to be the $2\cdot T$ (here it is 4) but a well working value is much bigger, e.g. N=50 for even small T.

We are going to prove the next two theorems:

Theorem 6.1: Suppose we have the above mentioned series of labeled data

a, b, a, b, a, b, …

with population size

$N=T\cdot k\cdot\lambda+2\cdot\theta+1=2\cdot k\cdot\lambda+2\cdot\theta+1$, $\theta=0,1,2,\ldots,k-1$ and $\lambda=1,2,3,\ldots$ Thus, we can note that

$$\lambda=\frac{N-2\cdot\theta-1}{2\cdot\kappa},\ 2\cdot\lambda+1=\frac{N+\kappa-2\cdot\theta-1}{\kappa}$$

$$\text{and } \lambda+1=\frac{N+2\cdot k-2\cdot\theta-1}{2\cdot k}.$$

Then

(a) for $k=2\cdot\mu$, $\mu=1,2,3,\ldots$ we have $Var\bar{x}_k = \frac{(b-a)^2}{4}$ and

(b) *for* $k=2\cdot\mu+1$, $\mu=1,2,3,\ldots$ we have $Var\bar{x}_k < \frac{(b-a)^2}{4}$, sometimes

$Var\bar{x}_k < \frac{(b-a)^2}{16}$.

Proof: (a) Since we have only two values for X, the a and b, in the case with $k=2\cdot\mu$, $\mu=1,2,3,\ldots$ the systematic sampling table has in all its odd columns the value a and in all its even columns the value b. Thus, every odd column-sample gives a sample mean value equal to a and every even column gives a sample mean value equal to b. Since the number of columns is even, we have μ times sample mean the $\bar{x}_k = a$ and μ times sample mean the $\bar{x}_k = b$. Thus, it is $E(\bar{x}_k) = \frac{a+b}{2}$ and $Var\bar{x}_k = \frac{(b-a)^2}{4}$, q.d.e.

(b) Now, it is $k=2\cdot\mu+1$, $\mu=1,2,3,\ldots$

(b1) For $1+2\cdot\theta<k$ in the systematic sampling table we have λ pairs of rows, like

$$a\ b\ a\ b\ldots a$$

$$b\ a\ b\ a\ldots b$$

and a truncated row with $1+2\cdot\theta$ elements like

$$a\ b\ a\ldots a$$

After this, there are

$\theta+1$ columns-samples with mean value $\bar{x} = \frac{(\lambda+1)\cdot a+\lambda\cdot b}{2\cdot\lambda+1}$

θ columns-samples with mean value $\bar{x} = \frac{\lambda\cdot a+(\lambda+1)\cdot b}{2\cdot\lambda+1}$ and

$k-2\cdot\theta-1$ columns-samples with mean value $\bar{x} = \frac{a+b}{2}$.

Thus, the mean value of all the column mean values is given by the next quantity:

$$E(\bar{x}_k) = \frac{a\cdot(N+k-2\cdot\theta)+b\cdot(N+k-2\cdot\theta-2)}{2\cdot(N+k-2\cdot\theta-1)}.$$

The calculation of the variance of the ample mean values gives the

$$\boldsymbol{Var\bar{x}_k} = \frac{\{(2\cdot\theta+1)\cdot k-1\}\cdot(b-a)^2}{4\cdot(N+k-2\cdot\theta-1)^2} < \frac{(b-a)^2}{4}\cdot\frac{k^2}{N^2} < \frac{(b-a)^2}{16} \quad .$$

(b2) For $2\cdot\theta+1=k$ there are $\theta+1$ columns-samples with mean value $\bar{x} = \frac{(\lambda+1)\cdot a+\lambda\cdot b}{2\cdot\lambda+1}$ and θ columns-samples with mean value $\bar{x} = \frac{\lambda\cdot a+(\lambda+1)\cdot b}{2\cdot\lambda+1}$ and the mean value of these means is

$$E(\bar{x}_k) = \frac{(a+b)\cdot(N+k-2\cdot\theta-1)(\theta+1)}{2\cdot k\cdot(N+k-2\cdot\theta-1)}.$$

Thus, the variance of the sample mean values is

$$\boldsymbol{Var\bar{x}_k} = \frac{(b-a)^2\cdot(k+1)\cdot(k-1)}{4\cdot(N+k-2\cdot\theta-1)^2} < \frac{(b-a)^2\cdot k^2}{4\cdot N^2} < \frac{(b-a)^2}{16}.$$

(b3) For $k+1 \le 2\cdot\theta+1 \le 2\cdot k-1$ there are $2\cdot\theta+1-k$ columns-samples with mean value $\bar{x} = \frac{a+b}{2}$, $k-\theta$ with mean value $\bar{x} = \frac{(\lambda+1)\cdot a+\lambda\cdot b}{2\cdot\lambda+1}$ and $k-\theta-1$ with mean value $\bar{x} = \frac{\lambda\cdot a+(\lambda+1)\cdot b}{2\cdot\lambda+1}$.

Thus, the mean value of the sample means is

$$E(\bar{x}_k) = \frac{a\cdot(N-k-2\cdot\theta)+b\cdot(N+k-2\cdot\theta-2)}{2\cdot(N+k-2\cdot\theta-1)}$$

and also the $\boldsymbol{Var\bar{x}_k} = \frac{\{2\cdot k^2-(2\cdot\theta+1)\cdot k-1\}\cdot(b-a)^2\cdot}{4\cdot(N+k-2\cdot\theta-1)^2} < \frac{(b-a)^2}{4}\cdot\frac{k^2}{(N-k)^2} < \frac{(b-a)^2}{4}$.

Thus, the theorem is proved. ◊

Theorem 6.2: Suppose we have the above mentioned series of labeled data

$$a, b, a, b, a, b, \ldots$$

with population size $N=T\cdot k\cdot\lambda+2\cdot\theta=2\cdot k\cdot\lambda+2\cdot\theta$, $\theta=0,1,2,\ldots,2k-2$ and $\lambda=1,2,3,\ldots$

Then (a) for $k=2\cdot\mu$, $\mu=1,2,3,\ldots$ we have $\boldsymbol{Var\bar{x}_k} = \frac{(b-a)^2}{4}$

and (b) for $k=2\cdot\mu+1$, $\mu=1,2,3,\ldots$ we have $0 \le \boldsymbol{Var\bar{x}_k} < \frac{(b-a)^2}{4}$.

Proof: (a) Since we have only two values for X, the a and b, in the case with $k=2\cdot\mu$, $\mu=1,2,3,\ldots$ the systematic sampling table has in all its odd columns the

value *a* and in all its even columns the value *b*. Thus, every odd column (sample) gives as sample mean value the value *a* and every even column gives as sample mean value the *b*. Since the number of columns is even we have μ times sample mean the $\bar{x}_k = a$ and μ times sample mean the $\bar{x}_k = b$. Thus it is $E(\bar{x}_k) = \frac{a+b}{2}$ and (q.d.e.) $Var\bar{x}_k = \frac{(b-a)^2}{4}$, q.d.e.

(b) Now, it is k=2·μ+1, μ=1,2,3,…

(b1) For $\theta=0$ in the systematic sampling table we have λ pairs of rows, like

$$a\ b\ a\ b \ldots a$$

$$b\ a\ b\ a \ldots b$$

and for all the columns-samples the mean value is $\bar{x}_k = \frac{a+b}{2}$. Thus, the $Var\bar{x}_k = 0$.

(b2) For θ=1,2,3,…, $\frac{k-1}{2}$ there are θ columns-samples with mean value $\bar{x} = \frac{(\lambda+1)\cdot a + \lambda \cdot b}{2\cdot\lambda+1}$ and θ columns-samples with mean value $\bar{x} = \frac{\lambda \cdot a + (\lambda+1)\cdot b}{2\cdot\lambda+1}$ and $k-2\cdot\theta$ columns with mean value $\bar{x} = \frac{a+b}{2}$. Thus, the mean value of the sample means is $E(\bar{x}_k) = \frac{a+b}{2}$ and the

$$Var\bar{x}_k = \frac{\theta\cdot(b-a)^2\cdot k}{2\cdot(N+k-2\cdot\theta)^2} < \frac{(b-a)^2}{2}\cdot\frac{k^2}{2\cdot(N+k-2\cdot\theta)^2} < \frac{(b-a)^2}{16}.$$

Note that $N=2\cdot k\cdot\lambda+2\cdot\theta$ gives $\lambda = \frac{N-2\cdot\theta}{2\cdot k}$ and $2\cdot\lambda+1 = \frac{N+k-2\cdot\theta}{k}$. Also the inequality $2\cdot\theta \le k-1$ was adopted.

(b3) For $\theta = \frac{k+1}{2}, \frac{k+3}{2}, \ldots, k-1$, or $k+1 \le 2\cdot\theta \le 2\cdot k-2$, there are $2\cdot\theta-k$ columns-samples with mean value $\bar{x} = \frac{a+b}{2}$, $k-\theta$ with mean value $\bar{x} = \frac{\lambda\cdot a + (\lambda+1)\cdot b}{2\cdot\lambda+1}$ and $k-\theta$ with mean value $\bar{x} = \frac{(\lambda+1)\cdot a + \lambda\cdot b}{2\cdot\lambda+1}$. Thus, the

mean value of the sample means is $E(\bar{x}_k) = \frac{a+b}{2}$ and the

$$Var\bar{x}_k = \frac{(k-\theta)\cdot(b-a)^2\cdot k}{2\cdot(N+k-2\cdot\theta)^2} < \frac{(b-a)^2}{2}\cdot\frac{k\cdot(k-1)}{2\cdot(N+k-2\cdot\theta)^2} < \frac{(b-a)^2}{4}.$$

Note that $N=2\cdot k\cdot\lambda+2\cdot\theta$ gives $\lambda = \frac{N-2\cdot\theta}{2\cdot k}$ and $2\cdot\lambda+1 = \frac{N+k-2\cdot\theta}{k}$.

Thus, the theorem is proved. ◊ ◊

From a special point of view the case T=2 becomes more trivial if we adopt $b=1$ and $a=0$. The general form is $N = 2\cdot k\cdot\lambda + 2\cdot\theta + \upsilon$, where $\upsilon = 0,1$ and $\theta = 0,1,2,...,k-1$. We face two different cases dependent on k, if it is even or odd. The results are given in the next Table 6.1

Table 6.1.

$k \rightarrow$ / $\downarrow \upsilon$	$2\cdot\mu$	$2\cdot\mu+1$
0	$Var\bar{x}_k = \frac{1}{4}$	$0 \le Var\bar{x}_k < \frac{1}{4}$
1	$Var\bar{x}_k = \frac{1}{4}$	$0 \le Var\bar{x}_k < \frac{1}{4}$

The variance in the very trivial case $T=2, a=0, b=1$

7. Results and discussion

Our basic goal was to propose and support by theoretic tools an algorithm dealing with periodicities of labeled data.

The coming out spirit of the present article is that we can take help from systematic sampling with a *step* k=2, 3, 4, 5,…,$\left[\frac{N}{2}\right]$. When $k=\xi\cdot T$, ξ=1,2,3,… and T is the period of the labeled data, we have k samples from the systematic table and the mean values coming out from those samples have the maximum variance. The related variance, when it is $k \neq \xi\cdot T$ is not the maximum and it is usually extremely lower than the maximum one. For instance, T=2 means that data series has the form *a, b, a, b,…,a, b,…* and when it is k=2, 4, 6, 8, …, $2\cdot\xi$ we have very big values for the variance of the sample mean value. On the other hand, the variance in the case $k=2\cdot\lambda+1$ tends to zero and sometimes is exactly zero when we face even values of N.

Sometimes the maximum variance is extremely bigger than the other values. It is even one hundred times or more bigger, see Table 3.4. These extremely big values of the variance are a chance to use the proposed algorithm and its results. The systematic sampling algorithm becomes a tool to copy the motive of data via the timbre of period T and the number of systematic samples k. Moreover, if we want to see more clearly the periodicity and the value of the period T we can use the ratio indices

$$\frac{Var\bar{x}_k}{E\left(s^2\right)} \text{ and } \frac{Var\bar{x}_k}{E(CVy_k)} .$$

For each of these indices the difference between its maximum value and its other common values becomes more and more bigger than the appropriate difference between maximum and common values of the simple $Var\bar{x}_k$.

It is obvious that all the indices presented in this paper are very useful and applicable for many sectors of production and administration. For instance, the administration of a hospital wants to plan the supply of an antivirus vaccine. Their first duty is to see the number of events per months in the past (say) 10 years. If from the 120 values a period T=6 comes out, they have to be careful every six months with the picks in demand for the suitable vaccine.

An implication of the presented indices is their use by the researchers in every case of analysing data. Among many goals, one could be to see if we face any cyclic behaviour of the data separated or in parallel with a linear trend, etc., etc.

Announcement

I thank very much the referees for their suggestions and corrections.

REFERENCES

COCHRAN, W. (1977) "*Sampling Techniques*", John Wiley & Sons, New York, Toronto.

FARMAKIS, N. (2002) "*Introduction to Sampling*", Editing A & P Cristodoulidi, Thessaloniki, (in Greek).

KOROTKOV, E.V., KOROTKOVA, M.A. (1995) "Latent Periodicity of DNA Sequences from Some Human Gene Regions", *DNA Sequence-The Journal of Sequencing and Mapping*, **Vol. 5**, pp 353–358.

PASQUIER, C., PROBONAS, V., VARVAYANNIS, N., HAMODRAKAS, S.(1998) "A Web Server to Locate Periodicities in a Sequence", *Bioinformatics App. Note,* **Vol. 14**, No 8, pp 749–750.

THOMSON, M. E. (1997) "*Theory of Sample Surveys*", Chapman & Hall, London, New York.

THOMSON, S. K. (1992) "*Sampling*", John Wiley & Sons, New York, Toronto.