SIMPLIFIED METHODS FOR PERIOD ANALYSIS*

by

J. REIMANN

Department of Civil Engineering Mathematics, Technical University, Budapest

(Received February 8, 1972)

Presented by Prof. P. RÓZSA

1. Introduction

In connection with the statistical analysis of time series, several methods have been developed for the *analysis of periods* which were generally based on different *heuristic* considerations. Most of the methods developed are mathematically interesting, but their application in practice is rather difficult and laborious, and commonly requires a very long series of data. Most of the statistical methods evolved are appropriate for hypothesis testing, that is, if one has some guess on the length p of the period, then these methods may help to check whether p can really be the length of the period. The guess concerning the length of the period might result from considerations relating to the nature or from certain statistical properties of the process producing the data series (e.g., most frequent time intervals between outstanding values). One may not have any serious guess on the period length or there may be various periods in the data series, therefore it is desirable to develop a method for the period analysis to directly furnish the *length of the period*, so to say lending itself to calculate the period or, in other words, which procedure, in contradistinction to the hypothesis testing methods is rather of "estimation" nature.

True, alternatives of the "harmonic analysis" are suitable for the explicit determination of the period length, but only after a series of transformations by lengthy trial-and-error calculations.

This all made imperative to develop a simple statistical method for the relatively quick determination of the "main" period length convenient for computer use. To the author's knowledge, no comparable methods have been published in the special literature to now.

Prior to the description of the direct method for finding the length of the period, a hypothesis testing method reported of by WHITTAKER and ROBINSON [1] should briefly be outlined. It started from heuristic considerations which gave the author an inspiration to develop the direct method. Essentials of the former method are as follows.

* Based on research done at the Institute of Water Management and Hydraulic Engineering.

Be the data series observed:

$$u_1, u_2, \ldots, u_N.$$

In the considered case, the u_i values are the monthly normal water levels of the Lake Balaton (for example, u_1 is the mean level of November 1921 and u_n that of October 1958). Be the hypothesis that p is the real period length. Dividing the data series to subsets of length p:

$$\begin{array}{rllll} & u_1 & u_2 \ldots\ldots & u_p \\ & u_{p+1} & u_{p+2} \ldots\ldots & u_{2p} \\ \hline \Sigma u_i: & U_1 & U_2 & U_p \end{array}$$

If, in fact, p is the real period, then every row of the above table is by and large of the same course which means that the values $u_{ip+1}, \ldots, u_{(i+1)p}$ fit a kind of wave line.

The elements of every row fitting about the same curve, column sums $U_1, U_2, \ldots, U_p$ describe this curve with an m-fold amplitude. If, in turn, the data are divided to other than p lengths, then the data of each row cannot be said to approach the same curve; on the contrary, the rows so to say compensate each other. Hence, expressing the difference between the largest and the smallest column sum among $U_1, U_2, \ldots, U_p$:

$$\Delta U_{(p)} = \max_i U_i - \min_i U_i$$

the function $U(p)$ will be the maximum if p is the real period.

If for a different p' value $\Delta U(p') > \Delta U(p)$ then the hypothesis that p is the real period should be rejected.

This method has been applied for trials on the data series of the water levels of the Lake Balaton which will be reported below.

The data collected by SZESZTAY [2] were grouped for primary investigation according to periods 9, 10, 11, 12, 13, 14 and 15. The following $\Delta U(p) = \max_i U_i - \min_i U_i$ values were obtained for each of the assumed periods:

p:	9	10	11	12	13	14	15
$\Delta U(p)$:	109	141	325	*1193*	161	210	401

The result markedly shows the 12-month periodicity.

The calculations were done with a desk-top calculator by making use of a data series of 38 years.

Unfortunately, the distribution of $\Delta U(p)$ is not known exactly, thus no lower limit number for $\Delta U(p)$ may be established above that p may be considered as the real period. Anyhow, a principle can be laid down: the higher $\Delta U(p)$, the more p may be considered as a period.

2. Direct method for finding the period

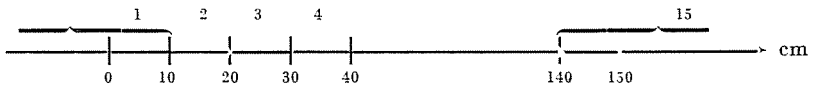
The basic idea that dividing a periodic series of data into subsets of lengths corresponding to the period, these lengths will be roughly of the same course (or, at least, for a certain criterion these subsets are more similar than if it would be divided into subsets of other lengths), will now be utilized in another way.

The starting data series is again the series of the mean water levels of the Lake Balaton (456 data from 1921 to 1958 by SZESZTAY [2]). Be the data series

$$u_1, u_2, \ldots, u_{456}.$$

The monthly mean water levels are continuous random variables. Thus, in fact, this is a set of continuous random variables $\xi_1, \xi_2, \ldots, \xi_{456}$. The problem is much simplified without loss of efficiency by replacing the set of continuous variables by a set of discrete variables using the following simple transformation (rounding-off customary in statistics).

The possible water level values are grouped as:



The range of values is divided into intervals of 10 cm, the values lower than 10 cm are coded by 1, those between 10 and 20 cm by 2, and so on, and the values higher than 140 cm by 15, that is:

$$v_i = [u_i] + 1$$

where [ ] denotes the integer part, 10 cm being the unit.

Thus a coded table is obtained for the monthly mean water level values: $v_1, v_2, \ldots, v_{456}$ the actual values of which are listed in Table 1.

Suppose now that the same coded series of data $v_1, v_2, \ldots, v_{456}$ is written on two tapes (but the first tape contains it twice consecutively) which are then superimposed and shifted relatively to each other; count how many times two identical values are in coincidence at every shifting. With actual data, the following situation will be seen:

$$\begin{array}{l} 5\ 5\ 6\ 6\ 6\ 6\ \{6\ \{6\ 4\ 3\ \{6\ 1\ 1\ 1\ 1\ \{1\ 7\ 2\ 3\ 3\ \ldots \\ \qquad 5\ 5\ \{6\ \{6\ 6\ 6\ \{6\ 6\ 4\ 3\ 6\ \{1\ 1\ 1\ 1\ 1\ 7\ \ldots \end{array}$$

Table 1

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
	0	10	20	30	40	50	60	70	80	90	100	110	120	130	140	150
	Years	XI	XII	I	II	III	IV	V	VI	VII	VIII	IX	X			
1	1921	5	5	6	6	6	6	6	6	4	3	6	1			
2	1922	1	1	1	1	7	2	3	3	1	1	1	1			
3	1923	3	3	5	5	6	7	8	8	7	6	5	4			
4	1924	5	6	7	7	7	10	11	11	11	10	9	8			
5	1925	8	8	7	7	8	7	8	9	7	6	6	6			
6	1926	7	8	9	9	10	10	9	9	9	9	9	8			
7	1927	9	10	11	10	10	10	9	9	8	8	7	7			
8	1928	7	7	7	8	9	9	9	9	8	6	6	6			
9	1929	6	7	8	8	8	10	11	10	9	8	6	6			
10	1930	7	8	7	8	8	10	10	9	7	6	6	6			
11	1931	8	9	11	11	13	14	14	13	11	9	8	8			
12	1932	8	7	7	8	8	10	10	9	8	7	7	7			
13	1933	7	7	8	8	9	9	10	9	9	8	7	7			
14	1934	8	10	10	10	10	10	9	8	8	8	7	7			
15	1935	7	8	9	9	10	10	10	9	8	7	6	6			
16	1936	6	7	9	10	11	11	11	11	10	8	7	8			
17	1937	8	8	8	9	10	13	12	11	10	9	8	7			
18	1938	9	11	11	11	10	9	9	9	8	7	6	7			
19	1939	6	6	7	8	8	8	8	9	8	7	6	7			
20	1940	6	7	8	8	9	11	11	11	9	9	9	6			
21	1941	11	11	10	10	11	11	11	10	8	7	6	7			
22	1942	7	7	8	8	10	12	12	11	9	8	6	8			
23	1943	5	6	6	8	9	8	8	9	9	8	7	6			
24	1944	5	7	8	8	10	10	10	10	10	8	6	6			
25	1945	7	8	9	10	11	11	9	8	7	6	5	11			
26	1946	5	6	7	8	9	9	8	7	7	6	6	6			
27	1947	5	7	7	8	12	15	15	15	15	13	12	5			
28	1948	8	8	8	9	9	10	10	9	10	10	9	6			
29	1949	7	7	7	7	6	7	7	6	6	5	4	6			
30	1950	4	5	5	7	8	8	9	7	6	5	4	5			
31	1951	6	6	10	11	12	11	11	13	13	11	9	8			
32	1952	7	7	8	9	11	10	9	9	8	6	5	6			
33	1953	6	7	9	9	9	9	9	9	9	8	7	6			
34	1954	6	6	6	6	7	8	10	10	10	9	8	8			
35	1955	8	7	8	9	11	11	11	10	9	9	9	8			
36	1956	9	9	9	9	10	12	12	11	11	9	8	6			
37	1957	6	7	7	8	10	10	9	9	8	8	7	7			
38	1958	7	6	6	7	8	8	7	7	7	6	6	5			

{ 78 coincidences

If it is true that subsets of actual period length are more similar than are those of other lengths, then the most of coincidences may be observed when the tapes are shifted by the true length of the period. This fact may also serve as the statistical definition of the period, if it is completed by certain numerical stipulations. *The series of data is a stochastic process*, hence, a random function, thus, the *number of coincidences is a random variable.* In the case of a strictly periodic, non-random function, if the shifting equals the period, then all of the data coincide, that is, if N is the length of the data series, then shifting by the period length would bring about N coincidences whereas shifting by a different length would result in less of coincidences.

Let us examine now that in the case of a stochastic data series how many coincidences may be expected in a certain position.

The number of coincidences depends partly on the statistical distribution of each value in the data series and partly on the relative position of the individual values, i.e. on the spacing of identical digits.

Be the statistical distribution of the data series:

II	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
	p_1	p_2	p_3	$\cdots$											p_{15}

Confronting the data series with itself in a certain shifting is the same as taking two random permutations of the data

$$v_1, v_2, \ldots, v_{456}$$

and writing under each other. What is the probability of a coincidence at the ith place?

Denote by $A_1, A_2, \ldots, A_{15}$ the event that at the ith place the digit $1, 2, \ldots, 15$ is found twice under each other. Then

$$P(A_1) = p_1^2, \quad P(A_2) = p_2^2, \ldots, P(A_{15}) = p_{15}^2 .$$

Since $A_1, A_2, \ldots, A_{15}$ are mutually exclusive events,

$$P(A_1 + A_2 + \ldots . A_{15}) = \sum_{k=1}^{15} p_k^2 .$$

If ζ_i is a characteristic variable allotted to the ith place which assumes the value 1 if at the ith place there is a coincidence and the value 0 if not, then

$$P(\zeta_i = 1) = \sum_{k=1}^{15} p_k^2, \qquad P(\zeta_i = 0) = 1 - \sum_{k=1}^{15} p_k^2 .$$

If the data series is of length N, then the random variable $\xi_i = \sum_{i=1}^{N} \zeta_i$ (number of coincidences in a data series of length N) is a random variable of binomial distribution, with mathematical expectation and standard deviation

$$M(\xi) = N \sum_{k=1}^{15} p_k^2; \qquad D(\xi) = \sqrt{N \sum_{1}^{15} p_k^2 \left(1 - \sum_{1}^{15} p_k^2\right)}$$

respectively.

On the basis of the *Moivre—Laplace* limit theorem, ξ is of approximately normal distribution, thus, if the data series is of statistical distribution (II), then the relationship

$$P\left(\frac{\xi - N \sum_{1}^{15} p_k^2}{\sqrt{N \Sigma p_k^2 (1 - \Sigma p_k^2)}} > 1{,}28\right) \leq 0{,}1$$

holds for the coincidences in a data series shifted at random.

The statistical distribution for the data series of the water levels of the Lake Balaton is the following:

$p_1 = 0.02$	$p_1^2 = 0.0004$
$p_2 = 0.00\ldots$	$p_2^2 = 0.\ldots$
$p_3 = 0.01$	$p_3^2 = 0.0001$
$p_4 = 0.01$	$p_4^2 = 0.0001$
$p_5 = 0.04$	$p_5^2 = 0.0016$
$p_6 = 0.15$	$p_6^2 = 0.0225$
$p_7 = 0.17$	$p_7^2 = 0.0289$
$p_8 = 0.19$	$p_8^2 = 0.0361$
$p_9 = 0.17$	$p_9^2 = 0.0289$
$p_{10} = 0.11$	$p_{10}^2 = 0.0121$
$p_{11} = 0.09$	$p_{11}^2 = 0.0081$
$p_{12} = 0.02$	$p_{12}^2 = 0.0004$
$p_{13} = 0.01$	$p_{13}^2 = 0.0001$
$p_{14} = 0.00\ldots$	$p_{14}^2 = 0.\ldots$
$p_{15} = 0.01$	$p_{15}^2 = 0.0001$
	$\sum_{1}^{15} p_k^2 = 0.1394$

$$P(\zeta_i = 1) = 0.14\ ; \quad P(\zeta_i = 0) = 0.86$$

$$M(\xi) = 456 \cdot 0{,}14 \approx 64, \qquad D(\xi) = \sqrt{457 \cdot 0{,}14 \cdot 0{,}86} \approx 7.$$

From relationship (3):

$$P\left(\frac{\xi - 64}{7} \geq 1{,}28\right) = P(\xi \geq 73) < 0{,}1\,.$$

This means that the shift including 73 or more coincidences should be considered as a period. (Here, a 90% reliability is sufficient.)

Remember that from the data in Table 1 it can directly be read off that shifting by 12 yields 78 coincidences, therefore the 12-month periodicity of the data series is beyond doubt.

In order to see whether there is another longer period in the data series, the coincidences are to be tested by the outlined procedure. (It should be noted that the data series in Table 1 is rather short for this purpose.)

It would be rather impractical and hardly feasible to actually write down the data on tapes and count the number of coincidences in different shiftings. Therefore, development of an algorithm likely to be used for computer programming has been aimed at.

This algorithm delivers the exact number of coincidences to be expected from a given shift.

For an easier understanding of the algorithm let us again imagine that the data of Table 1 are written down on long paper tapes with centimetre graduation, consecutively twice on the upper tape.

Assume that, for example, in the 9th shifting k_1 coincidences have been observed between digits 1, further k_2 coincidences between digits 2, and . . . k_{15} coincidences between digits 15. Examining now in the data series (on the centimetre-tape) at which graduations the digit 1 occurs, recording them and finding their distances, among these distances that of 9 will be encountered just k_1 times, among the distances between location marks of digits 2 that of 9 occurs k_2 times, and so on.

For the sake of clearness let us see the location marks of the water level value 1. In Table 1 the digits 1 may be found at graduations 12, 13, 14, 15, 21, 22, 23, 24. Since there are 456 data, the coincidences will be observed in 455 shiftings, therefore this data series will be written consecutively twice on the upper tape. In the data series of double length, the graduations at digit 1 and their distances yield Table 2.

Table 2

Graduations at digit 1	12	13	14	15	21	22	23	24	468	469	470	471	477	478	479	480
12		1	2	3	9}	10	11	12								
13			1	2	8	9}	10	11	455							
14				1	7	8	9}	10	454	455						
15					6	7	8	9}	453	454	455					
21						1	2	3	447	448	449	450				
22							1	2	446	447	448	449	455			
23								1	445	446	447	448	454	455		
24									444	445	446	447	453	454	455	
468										1	2	3	...			
469																
470																
471																
477																
478																

The distances of every graduation to all higher graduations should be established up to 456. In Table 2, for shifting by e.g. 9, four digits of 1 are seen to coincide because among the distances that of 9 occurs four times.

This table of distances should be compiled for the graduations at digits 2, 3, 4, . . ., 15. Counting in each of the 15 tables of distances the occurrences of graduation 9 yields the number of coincidences for a shift by 9. And counting in each of the 15 tables the repetitions of every encountered distance (1, 2, . . ., 455) delivers the number of coincidences for every shifting. As concerns the data in Table 1, the frequency of the distance corresponding to the actual period must be above 72.

Summary

Probability criteria are involved to define the concept of periodicity, then the length of an eventual period is determined from the number of coincidences of random sets of data. A numerical example is presented to illustrate that the method suggested in the paper is, from computing aspects, more advantageous than the methods used so far for the analysis of periods.

References

1. WHITTAKER, E. T., ROBINSON, G.: Calculus of Observations. 3rd ed. Blackie and Sons, 1940.
2. SZESZTAY K.: Water Conservancy of the Lake Balaton. Studies and Research Results. No. 9, Bp. 1962, VITUKI (In Hungarian).

Ass. Prof. Dr. József REIMANN, 1111 Budapest, Műegyetem rkp. 3, Hungary