THE USE OF LINEAR FUNCTIONS TO DETECT HIDDEN PERIODS IN DATA SEPARATED INTO SMALL SETS

By

EDWARD L. DODD

I.—INTRODUCTION

Readers who have access to the Handbook of Mathematical Statistics[1] will find in chapter XI a synopsis of a periodogram analysis by W. L. Crum, with references to some of the important papers on period testing.

My own interest in this subject was aroused several years ago by Dr. J. A. Udden,[2] Director of the Bureau of Economic Geology at the University of Texas, who had in his possession measurements of the thicknesses of successive layers of anhydrite ($CaSO_4$) taken from a Texas oil well. The material, Dr. Udden noted, was "suggestive of cycles" (p. 351); but one difficulty was mentioned: "Probably 2 per cent of the layers are indistinct." It was not always possible to tell whether the number recorded as the thickness of a layer represents a single deposit or two or more deposits insufficiently separated by the usual bituminous demarcation. The analogous difficulty in distinguishing consecutive rings of big trees[3] was met by comparison of the rings of trees from the same forest. But such companion records were not available for the rock lamina.

A little reflection will show that the usual method of testing

1 Rietz, Houghton Mifflin Co., 1924.

2 "Laminated Anhydrite in Texas." *Bulletin of the Geological Society of America*, Vol. 35 (1924), pp. 347-354.

3 A. E. Douglass, "Climatic Cycles and Tree Growth," Carnegie Institution of Washington, Publication No. 289 (1919).

for cycles, from data arranged in columns becomes vitiated if in several instances merging of layers has taken place—not so much because of the exaggerated size of certain items, but because the items get into the wrong columns. When a step is lost, all subsequent items are misplaced.

My purpose is then to explain how tests for periods can be made by first using the data in small sets—thus minimizing the vicious effects of a merger—and then by suitably combining the results obtained from these small sets.

We might as well admit at the start that a demonstration of a periodicity is in general impossible. Perhaps the revolution of the earth on its axis represents a demonstrated periodicity. But for the most part, announced periodicities are merely improbable or probable. There is no absolute proof that they exist. We know that what we call "pure chance," typified by the throws of a coin, will sometimes yield irregularities of oscillation between two states, the minimum and the maximum, which to all appearances is a "periodicity." The question arises: About how often will pure chance thus deceive us? All we can do is to compute certain relative frequencies or probabilities. If the probability found is very, very small, that the apparent periodicity had its origin in pure chance, we assert with some assurance that a real periodicity exists. In this mode of approach, this paper will follow rather closely Arthur Schuster,[1] whose work is fundamental.

Although Schuster's main interest was in the quadratic function, "intensity"—at first, in the square root of intensity, Terrestrial Magnetism, loc. cit.—he pointed out (p. 27) how certain con-

1 "On the investigation of hidden periodicities with application to a supposed 26-day period of meteorological phenomena." *Terrestrial Magnetism,* Vol. 3 (1898), pp. 13-41. In my paper, "The probability law for the intensity of a trial period, with data subject to the Gaussian law," *Bulletin of the American Mathematical Society,* Vol. 33 (1927), pp. 681-684, I referred to Schuster's paper in the *Proceedings of the Royal Society of London.* Reference should have been made also to the above paper in *Terrestrial Magnetism,* where the probability law is given for the square root of intensity (p. 21), which can easily be thrown into the form given in my paper. Schuster, however, postulated (p. 20) that "$2\pi p$ is a submultiple of a right angle"—a condition which would not always be satisfied—also (p. 21) that the vectors be distributed according to the law of errors centered at the *origin,* an inconvenient restriction, and his method did not bring out the different law of distribution needed for the case when the period is equal to two.

clusions could be reached through integrals—substantially linear functions, if integration is regarded as summation. It is this approach to period testing through linear functions that I am setting forth in his paper. Some special attention must be given to phase in the application of this method.

Most of the methods for detecting periodicities make use of the trigonometric functions, with their well known properties, in particular, use is made of the Sines and cosines of an angle and its multiples, as in harmonic analysis and Fourier series. With the aid of these harmonic multipliers, linear functions are first formed; and from these, by squaring and adding, a quadratic function, which plays the central role, as "intensity." In the method set forth in this paper, however, the linear functions themselves are the most important, not merely for graphical representation, but for determining probabilities.

Suppose, then, that a set of numbers is furnished us—perhaps from an unknown source—for example a set of ten numbers consisting of 5's and 1's alternating:

$$5,\ 1,\ 5,\ 1,\ 5,\ 1,\ 5,\ 1,\ 5,\ 1.$$

Has this set of numbers the period *two*? If this question means: Is there a function of period *two* which takes on these ten values, the answer is: Yes, namely—

$$3 + 2\cos \pi r \qquad\qquad r = 0,\ 1,\ 2,\ \ldots .\ 9$$

Here, as usual, π means 180°, obtained from a complete revolution of 360° by dividing by *two*. If, in place of an integer r, we take a continuous variable x, and plot

$$y = 3 + 2\cos \pi x$$

from $x = 0$ to $x = 10$, a wave curve is formed with each upper crest at 5, and each depression at 1.

But usually in period testing, something is desired beyond the mere possibility of making a mathematical curve fit the data. Perhaps a farmer on each 10 acres of his farm has raised 5 bales of cotton, 1 bale of cotton, 5 bales of cotton, etc., alternately for 10 years, under apparently the same conditions as to labor, fer-

tilizer, etc. He would like to know whether this is due to mere chance or to some recurrence at two-year intervals of droughts, pests, or adverse conditions. Stranger events do, indeed, occur by pure chance than the foregoing hypothetical yield of cotton. But the regularity postulated above would strike almost anyone as exceptional, and it would be prudent for our farmer to believe that there was some non-fortuitous cause of the regularity, and to try to discover it.

Let us, indeed, set up a chance situation to correspond to the foregoing yield of cotton. If the two faces of a coin are marked 5 and 1, and are recorded as such, the probability for ten throws starting with 5 and alternating between 1 and 5 is only 1/1024. A bet of $1,023 against $1 would measure the unusualness of the specified succession of 5's and 1's.

That this occurrence is unusual may be signalized by another test and method of approach. Let X_r denote the result of the r th trial of an independent chance variable, which with equal likelihood ($p_1 = 1/2 = p_2$) takes on the values 5 or 1, and can take on no other value. The "mean value" of X_r is then, by definition,

$$p_1(5) + p_2(1) = \tfrac{1}{2}(5) + \tfrac{1}{2}(1) = 3$$

This would, indeed, be also the average value of the five 5's and five 1's in the illustration. The "mean error" ε of X_r would be found from

$$\varepsilon^2 = \frac{1}{2}(5-3)^2 + \frac{1}{2}(1-3)^2 = 4\,, \qquad \varepsilon = 2$$

This would be also the standard deviation σ of the numbers in the illustration—that is

$$\sigma^2 = \frac{1}{10}\left[(5-3)^2 + (1-3)^2 + (5-3)^2 + \cdots + (1-3)^2\right] = 4\,, \qquad \sigma = 2$$

Now let

$$X = X_1 - X_2 + X_3 - \cdots - X_{10}$$

Since the signs alternate, the mean value of X is zero; since

there are ten terms, the mean error of X is $\varepsilon\sqrt{10} = 2(3.16) = 6.32$. If now X_r should take on alternately the values of 5 and 1, then X would become 20. It would thus exceed its mean value zero by more than three times its mean error, or more than four and one-half times its "probable error." This is commonly regarded as "significant."

To see a little more clearly into the mechanism of the above result, let us pass from the numbers X_r to their deviations from their mean value 3.

Let

$$x_1 = X_1 - 3, \quad x_2 = X_2 - 3, \quad \cdots, \quad x_r = X_r - 3, \quad \cdots$$

Then the mean value of x_r is zero, and its mean error is 2. Now let

$$x = x_1 - x_2 + x_3 - \cdots - x_{10}$$

Then the mean value of x is zero and its mean error is $\varepsilon\sqrt{10}$; in both respects it resembles X. Furthermore it takes on the same value 20 that X takes on when the 5 and 1 alternate; since $x_1 - x_2 = (X_1 - 3) - (X_2 - 3) = X_1 - X_2$, etc. And here again 20 is a remarkable value for x since it represents an excess of more than three times its mean error. But let us now find x directly from the values taken on by x_r, when X_r alternates between 5 and 1.

$$x = 1\ (2) - 1\ (-2) + 1\ (2) - \ldots - 1\ (-2) = 20$$

The feature to be noted is that the successive values of x_r and of $\cos \pi (r-1)$ match in sign, for $r = 1, 2, 3 \ldots 10$.

$$x_r = 2, -2, 2, -2, \ldots, -2$$

$$\cos \pi (r-1) = 1, -1, 1, -1, \ldots, -1$$

Each product $x_r \cos \pi (r-1)$ is then positive; and this accounts for the large value of x. This matching in sign of the deviations of the data with the successive terms of a test function $\cos 2\pi (r-1)/k$ or perhaps $\cos 2\pi r/k$ when k is given

a particular value—here $k = 2$—is, indeed, fundamental. Also the similarity between the properties of X and x will be found to be maintained in more general cases.

The foregoing illustrates the method of period testing to be set forth in this paper. A general assumption is at first made, that the data contain no periodic constituent, but on the contrary represent mere chance fluctuations. Certain linear functions of the data are found with coefficients which are the cosines or sines of multiples of the angle associated with a given period. For these functions, the fluctuations usually to be expected are to be computed—assuming that the measurements represent chance data. If the actual values which these functions take on are greatly in excess of what is expected of them, the initial assumption that the data are due to chance is called into question. It may be more reasonable to suppose that to some extent the data conform to the period associated with the cosine multipliers involved in the test. These "harmonic" multipliers, indeed, pass through a succession of positive and negative values in a regular way. If the positive and negative fluctuations of the measurements from their average value are well "timed" with those of the harmonic multipliers, we get a sum of products nearly all positive, thus a much larger result than if positive numbers were not matched with positive, negative with negative numbers.

As preliminary to all tests, the data may be divided into fairly large groups of consecutive measurements—say with 120 measurements in a group; for 120 is a multiple of 2, 3, 4, 5, 6, 8, 10, 12, 15, 20, 24, 40, 60, numbers quite suitable for trial periods. The arithmetic mean and standard deviation of each such group may be computed. These may usually be accepted as close approximations to the mean value and mean error of the measurements of the group.

To illustrate further the nature of the tests to be applied, let us imagine that the 120 measurements of a group are recorded on slips of papers, these slips put into a bag, drawn at random, and recorded as drawn. This set of numbers would have the same arithmetic mean and standard deviation, noted above, no matter in what order they are drawn and recorded. But periodicities depend upon the order of the measurements. A chance order of measurements x_r, such as established by drawing from a bag, would very seldom match sufficiently well a periodic function like

cos $2\pi r/2$ with period 2, or cos $2\pi r/3$ with period 3, etc., to make a test function $C(k)=\Sigma X_r \cos 2\pi r/k$ noticeably large. Thus, if for some particuler k, the function $C(k)$, computed from the data in their actual given order, turns out to be significantly large, the indication is that the data contain a constituent with period k.

We mean here that each measurement of the set may be thought of as the sum of certain constituents, one of which is periodic with period k. Another constituent may perhaps have a different period k'. Still another constituent may be a chance variable with no regularity which can properly be called periodic.

II. Trigonometric Formulas.

Of considerable use are the simple formulas:

$$(1) \qquad \sin a \sin b = \tfrac{1}{2}\cos(a-b) - \tfrac{1}{2}\cos(a+b)$$

$$(2) \qquad \cos a \cos b = \tfrac{1}{2}\cos(a-b) + \tfrac{1}{2}\cos(a+b)$$

Indeed, by using (1) in summing the product $\sin(r\theta+\alpha)\sin\theta/2$ from $r=0$ to $(n-1)$ there is obtained,[1] in case θ is not a multiple of 360°,

$$(3) \qquad \sum_{r=0}^{n-1} \sin(r\theta+\alpha) = \sin\left(\frac{(n-1)\theta}{2}+\alpha\right)\frac{\sin n\theta/2}{\sin\theta/2}$$

Likewise, for $\theta \not\equiv 0 \pmod{360^\circ}$; i e., θ not a multiple of 360°,

$$(4) \qquad \sum_{r=0}^{n-1} \cos(r\theta+\alpha) = \cos\left(\frac{(n-1)\theta}{2}+\alpha\right)\frac{\sin n\theta/2}{\sin\theta/2}$$

As important special cases, we have when $n\theta$ is a multiple of 360°,

1 For formulas suitable for period testing and for a historical review of this subject with references, the reader is referred to the article of H. Burkhardt in *Encyklopädie der Mathematischen Wissenschaften*, II A 9a, pp. 642-694.

$$(5)\qquad \sum_{r=0}^{n-1} \sin(r\theta+\alpha) = 0 = \sum_{r=0}^{n-1} \cos(r\theta+\alpha), \quad \begin{matrix} \theta \not\equiv 0 \pmod{360^\circ} \\ n\theta \equiv 0 \pmod{360^\circ} \end{matrix}$$

As an application of (5), let $X_1, X_2, \cdots X_r \cdots X_n$ be any set of n numbers, let C be any constant. Then

$$(6)\qquad \sum_{r=0}^{n-1} (X_r - C)\cos(r\theta+\alpha) = \sum_{r=0}^{n-1} X_r \cos(r\theta+\alpha), \quad \begin{matrix} \theta \not\equiv 0 \pmod{360^\circ} \\ n\theta \equiv 0 \pmod{360^\circ} \end{matrix}$$

Likewise for $\sin(r\theta+\alpha)$.

The above signifies that if the X_r represent data to be subjected to tests with harmonic multipliers, where an integral number of complete cycles is taken, it is immaterial where the origin for the data is taken. In the theory, the C will be usually taken as the arithmetic mean of the data; in computation, the C may be some simple number which will reduce the number of significant figures in the data.

III. Chance Data in Distributions with close contact at extremities

Chance data distributed normally will be considered first. Given n numbers or variates $X_1, X_2, \ldots, X_n$, the arithmetic mean M and standard deviation σ are determined by

$$(7)\qquad M = \frac{1}{n}(X_1+X_2+\cdots+X_n)\ ;\ \sigma^2 = \frac{1}{n}\left[(X_1-M)^2+\cdots(X_n-M)^2\right].$$

Let

$$(8)\qquad \phi(t) = \frac{2}{\sqrt{2\pi}}\int_0^t e^{-\frac{t^2}{2}}\,dt$$

The data will be said to be normally distributed if the number of variates lying between $M+\lambda_1$ and $M+\lambda_2$ is approximately

$\frac{n}{2}\left[\phi(\lambda_2/\sigma)-\phi(\lambda_1/\sigma)\right]$ for all values of $\lambda_1 < \lambda_2$. Here n is supposed to be at least moderately large. To express this in the language of probability, suppose the n numbers $X_1, X_2, \cdots$ are recorded on slips and put into a bag, and suppose a slip is drawn out. Then, for the X_r thus drawn—

(9) Probability that $\lambda < X_r < \lambda + d\lambda$ is $\frac{d\lambda}{\sigma\sqrt{2\pi}} e^{-(\lambda-m)^2/2\sigma^2}$

where, if $d\lambda$ is taken rather small, the n is to be thought of as rather large.

The important theorem needed here—substantially explained, if not proven, in most books on probability—is that if sets of k of these variates are drawn at random, and linear functions with fixed coefficients, such as

$$F(k) = a_1 X_1 + a_2 X_2 + \cdots + a_k X_k \tag{10}$$

are formed, these functions $F(k)$ as determined in sets of drawings will be normally distributed with standard deviation σ_k, where

$$\sigma_k^2 = \sigma^2 (a_1^2 + a_2^2 + \cdots + a_k^2) . \tag{11}$$

If, in particular $a_r = \cos(r\theta + \alpha)$ making $2a_r^2 = 1 + \cos(2r\theta + 2\alpha)$ and if further $k\theta = 360°$ with $k > 2$, it follows from (5) that

$$\sigma_k^2 = k\sigma^2/2 \tag{12}$$

Let us now in (6) set $C = M$; or rather, what amounts to the same thing, change the origin for the data so as to make $M = 0$. Then the "mean value" or "expected value" of $F(k)$ in (10) is zero. Then, with the use of (8) and (12), it follows that

(13) Probability that $|F(k)| > 3\sigma\sqrt{k/2}$ is $1 - \phi(3)$ =0.0027.

This small probability by no means implies impossibility. However, if the computed $|F(k)|$ exceeds $3\sigma\sqrt{k/2}$, there

is some ground for doubting the original hypothesis that the data under consideration exhibit a chance arrangement. Sometimes such evidence gathered from different sections of the data can be made cumulative. A comparatively large value for $F(k)$ in (10) is likely to result when the signs of the X_r match the signs of the a_r, taken as in (11) and (12) as $\cos(r\theta+\alpha)$, giving a cycle or period of k items.

To what extent evidence is thus afforded for the specific period of k needs further consideration. But, until we have found an adequate number of instances in which some inequality like (13) is satisfied we have obtained little evidence of any periodicity at all.

Thus far we have considered normally distributed data, conforming to the well-known symmetric bell-shaped probability curve. But this is more restrictive than necessary. Results substantially the same can usually be obtained for distributions—even those not symmetric and not mesokurtic—which at both ends taper off in slender tails. Although the particular numerical value of the probability given in (13) is no longer applicable to these curves, the probability nevertheless is usually very small, as presented geometrically as slices of the two tails.

Moreover, the equations (11) and (12) arise from the general theory of expected values. Suppose that p_r is the probability that the chance variable X will take on the value ξ_i, where $p_1+p_2+\cdots+p_s=1$. Then the expected value of X is, by definition

$$(14)\qquad E(X)=p_1\xi_1+p_2\xi_2+\cdots+p_s\xi_s=E_1,$$

and its mean error $\varepsilon(X)$ is defined by

$$(15)\qquad \varepsilon^2(X)=p_1(\xi_1-E_1)^2+\cdots+p_s(\xi_s-E_1)^2=E(X-E_1)^2$$

It is common to identify expected value and mean error with arithmetic mean and standard deviation as approximations. In applying (6), the supposition was made that the origin be taken so as to make $M=0$. With this adjustment, we may take $E(X)=0=E$, in (14) and (15). As the X_j are regarded as independent, the theory of expected values applied to (10) leads first from $E(X)=0$ to $E[F(k)]=0$, as mentioned before;

and then to (11)—noting that when $i \neq j$, $E(X_i X_j) = 0$, in the expression for $E[F(k)-O]^2$.

IV. Data with Periodic Constituents.

We now consider data of the form

$$W_r = X_r + Y_r + Z_r \quad (16)$$

where X_r is, as before, a chance variable; but

$$Y_r = b\cos(r\theta+\beta) \; ; \quad Z_r = c\cos(r\theta'+\gamma) \quad (17)$$

$$k\theta - 2\pi = 360^\circ = k'\theta' \quad (18)$$

Here Y_r and Z_r are periodic with periods k and k', not necessarily integral, amplitudes b an c, phases β and γ, respectively. Dealing first with Y_r, let m and n be whole numbers such that $n = mk$. Then, in analogy with (10), but applied to n of the Y's take

$$F(n) = \sum_{r=0}^{n-1} Y_r \cos(r\theta+\alpha) = \frac{nb}{2}\cos(\alpha-\beta) \; ; \quad k>2 \quad (19)$$

as may be shown from (2) and (5). The magnitude of $F(n)$ depends materially upon the phase difference $(\alpha-\beta)$ But

$$\left|\cos(\alpha-\beta)\right| > 0.92 \; , \quad \text{if} \quad \left|\alpha-\beta\right| \leq 22\tfrac{1}{2}^\circ \quad (20)$$

Thus if the phase α of the test function $\cos(r\theta+\alpha)$ differs from the phase β of the data, taken now as Y_r in (17), by not more than $22\frac{1}{2}^\circ$, the absolue value of $F(n)$ in (19) will fall below its maximum, $nb/2$, by less than 8%. The phase β of the Y constituent of data would in general be unknown; but if for α we take *eight* consecutive multiples of 45°, one of these would fall within $22\frac{1}{2}^\circ$ of any designated angle β, (mod 360°). Moreover, if in (19), α is increased by 180°, $F(n)$ merely changes its sign, and thus gives no essentially new information. Hence, instead of *eight* multiples of

45°, the four multiples—90°, 0°, −45°, 45°—will be adequate. These, taken in the above order, give

$$(21) \qquad S = \sum_{r=0}^{n-1} Y_r \sin r\theta \quad ; \quad C = \sum_{r=0}^{n-1} Y_r \cos r\theta$$

$$(22) \qquad S' = \sum_{r=0}^{n-1} Y_r \sin (r\theta + 45°) \; ; \quad C' = \sum_{r=0}^{n-1} Y_r \cos (r\theta + 45°) .$$

Furthermore, it is not necessary to compute S' and C' in (22) directly from the data, since

$$(23) \qquad S' \; \frac{\sqrt{2}}{2} (C+S) \quad ; \qquad C' \; \frac{\sqrt{2}}{2} (C-S) \quad ;$$

but a direct computation of S' or C' would serve well as a check upon (21). Thus, if in (19) we assign to α the four values mentioned above, we get S, C, S', C', in (21), (22) such that for one of these quantities (20) is satisfied, which makes $F(n)$ in (19) take a value almost equal to $nb/2$. This increases as n itself—not merely as the square root of n, an increase typical for $\sum X_r \cos(r\theta+\alpha)$, see (12), (16), with k replaced by n.

Let us now consider the function:

$$(24) \qquad G(n) = \sum_{r=0}^{n-1} Z_r \cos (r\theta + \alpha) = C \sum_{r=0}^{n-1} \cos (r\theta + \alpha) \cdot \cos (r\theta' + \gamma).$$

By (2), the terms above have the form

$$(25) \qquad \frac{C}{2} \cos \left[r(\theta+\theta') + \alpha + \gamma \right] + \frac{C}{2} \cos \left[r(\theta-\theta') + \alpha - \gamma \right] .$$

In order to use (5), we postulate that neither $\theta+\theta'$ nor $\theta-\theta'$ is zero or any other multiple of 360°, in particular $\theta' \neq \theta$. With $n = mk$, as before, (18) gives $n\theta = mk\theta = m(2\pi)$. Hence, it follows that

$$\sin n(\theta+\theta')/2 = \pm \sin n\theta'/2 = \pm \sin mk\pi/k' .$$

Likewise, $\sin n(\theta-\theta')/2 = \mp \sin mk\pi/k'$.

Hence, from (4), (25) it follows that $G(n)$ in (24) contains the factor $\sin mk\pi/k$. Thus $G(n)=0$, if

$$(26)\quad k' = mk,\ \frac{mk}{2},\ \frac{mk}{3},\ \ldots\ \frac{mk}{m-1},\ \frac{mk}{m+1},\ \frac{mk}{m+2},\ \ldots$$

This may also be written

$$(27)\quad qk' = mk,\quad q = \text{any whole number} \neq m.$$

Thus, if m cycles of a period k are used as multipliers in the form (24) upon a set of mk numbers Z_r with period $k' = mk/q$, where q is any whole number except m, the result is zero. It should be noted that in order to apply (4) to (24) (25), to get (26), it was necessary to require that $k' \neq k$, which would make $q \neq m$ in (27). To illustrate: 3 cycles, each with period $k=4$, will "annihilate" a set of 12 numbers if these are the successive terms of $C\cos(\gamma + 2\pi r/k')$ with period k' equal to 12, or 12/2, or 12/4, or 12/5, etc., but not 12/3.

Indeed, $G(n)$, instead of vanishing when k' is set equal to k in (24), making $\theta' = \theta$, takes on just about its maximum value $nc/2$ in this case when the phase α is properly chosen—see (19), (20). Inasmuch as $G(n)$ in (24) is a continuous function of θ', it follows that if k' is taken very close to k, $G(n)$ would be almost as large as for $k' = k$. But from (26) we learn that $G(n)$ goes down to zero if k' is allowed to be as small as $mk/(m+1)$ or as large as $mk/(m-1)$.

Thus, if significantly large results are obtained when using the test function $\cos(r\theta + \alpha)$ with period $k = 2\pi/\theta$, the individual period k itself is not necessarily indicated. But rather, the test furnishes evidence that *some period close to k is present in the data*, this proximity being expressed by the inequality (see 26)

$$(28)\quad \frac{m}{m+1}k < k' < \frac{m}{m-1}k.$$

The relations involved here can perhaps be set forth in greatest simplicity by using integration to effect summations—cf. (34). In the test function cos ($r\theta + \alpha$), set the phase $\alpha = 0$, and take $x = r\theta$, where $\theta = 2\pi/k$. Suppose k is rational, and take an even integer m such that $n = mk$ is an even integer. Consider the test as covering the data, from $x = -m\pi$ to $x = m\pi$. Also, in (24), take $\gamma = 0$, $\theta' = t\theta$, $= 1$. This leads naturally to

$$g(t, m) = \frac{1}{m\pi} \int_{-m\pi}^{m\pi} \cos x \cdot \cos tx \, dx \tag{29}$$

where the coefficient $1/m\pi$ is chosen to make $g(1, m) = 1$ With the aid of (2), it is easily seen that

$$g(t, m) = \frac{2t \sin m\pi t}{m\pi(t^2 - 1)}, \qquad t \neq 1 \tag{30}$$

for a given m, the plot of $g(t, m)$ as a function of t consists of a crest above the interval from $t = 1 - 1/m$ to $t = 1 + 1/m$, flanked on each side by depressions only about one-fourth or one-fifth as great in size or amplitude followed by waves of still smaller size—a "vibration" *strongly* "damped" on each side of $t = 1$. It has essentially the same characteristics as curves frequently occurring in periodogram analysis.[1] Only the interval from $t = 1 - 1/m$ to $1 + 1/m$ has in general much significance. Sometimes the two adjacent waves[2] need a little attention. But as $\theta' = t\theta$, the above interval is described by

$$1 - \frac{1}{m} < \frac{k}{k'} < 1 + \frac{1}{m}, \tag{31}$$

which is another way of writing (28).

As an illustration, suppose that 4 cycles of 12 terms each of cos ($r\,30^\circ + \alpha$) are used in a test with a significantly large result. Here $k = 12$, $m = 4$. Then (28) would recommend to our consideration periods between 9.6 and 16. Perhaps only those between 10 and 15 would deserve serious attention. Since at points $t = 1 \pm \frac{1}{4}m$, the curve (30) is less than half as high as at $t = 1$. Another interesting form[3] of (28) is

---

1 Rietz-Handbook Loc. cit. p. 172, Figure 17.

2 Schuster, *Terrestrial Magnetism*, Vol. 3 (1898), p. 30.

3 Cf. the Schuster criterion, Rietz Loc. cit. p. 173; Schuster, Loc. cit., p. 30.

$$(32) \qquad |k - k| < \frac{k'}{m}$$

Before leaving (30), it may be well to note that $g(t, m)$ does *not* take its maximum exactly at $t = 1$; but at

$$(33) \qquad t = 1 + \frac{3}{3 + 2m^2\pi^2}$$

as may be seen by setting $t = 1 + \tau$ in (30), expanding $\sin m\pi t = \sin m\pi\tau$ in powers of τ, and setting the first derivative equal to zero. When $m = 1$, $t = 1.13$; when $m = 2$, $t = 1.04$; when m is moderately large, t is very close to 1. In all cases, however, the test function which yields the largest result, when applied to a cosine function with period k' is not that one which exactly fits, but one with period $k = k't$, where in the ideal case represented by (29) this value of t is given by (33). Inasmuch as $t > 1$, there is some *danger*, then, of *overestimating the size of the unknown period* k', if the attempt is made to get a close approximation to k' by using several test periods k in the immediate vicinity of k', and selecting the k giving the maximum result. This is not due to the fact that $G(n)$ in (24) is a *linear* function of the Z_r's. For, if in (29), we should change cos x to sin x, to get the mate of $g(t, m)$, this mate would be zero. Thus, the usual *quadratic* function would reduce to the square of $g(t, m)$, and would have its maximum at the same place given by (33). If the main purpose of an investigation is merely to locate with fair precision those periods whose existence have high probabilities, it may not be necessary to refer to (33).

Going back to the constituents of W_r in (16) we see that if n terms of $\Sigma W_r \cos(r\theta + \alpha)$ are taken, the Y contribution to this sum increases directly as n; the X contribution, being of chance origin, increases usually about as the square root of n; while the Z contribution oscillates about zero.

V. Convenient Forms for Test Functions

The main points of the theory needed for testing data for periods with the aid of linear functions have now been set forth. In the first place, it appears impossible to demonstrate a periodicity. At best, we

can merely make certain suppositions appear more or less probable or improbable. The method outlined here starts with the assumption that the *order* of the sizes exhibited in the data is a *chance* arrangement. Certain functions are to be computed which under chance conditions would ordinarily keep generally within a certain range. If these functions show no marked tendency to jump the bounds, then the tests yield no positive evidence of periodicity. On the other hand, if these functions take on extremely high values, it appears reasonable to relinquish the supposition that the order of sizes is a chance arrangement and to suppose, rather, that such a periodicity exists as would naturally make the function large. If the data have as a constituent a cosine fluctuation and this is matched by a test cosine curve of the same period and phase, it is easy to see that the sum of products all positive obtained from similarly placed ordinates may be abnormally large. Any k which gives these large results is to be regarded as approximating a probable period.

That the tests may all be conducted in a systematic and uniform manner, some further properties and details may well be noted.

In the first place, only values of $k \geqq 2$ need be considered if the data are regarded as representing a sequence of discrete values, corresponding to values of the time (or other argument) spaced at unit intervals. For suppose that p/q, the period of $\cos\left[2\pi rq/p\right]$ is less than 2. Then for each integer r, $\cos\left[2\pi rq/p\right] = \cos\left[2\pi r(p-q)/p\right]$, the latter with period $p/(p-q) > 2$. This applies, indeed, to the case where the discrete values are integrated values. In fact, since

$$\int_r^{r+1} \cos\left(\frac{2\pi t}{k}+\beta\right)dt = A\cos\left(\frac{2\pi r}{k}+\beta'\right), \tag{34}$$

where $A = (k/\pi)\sin \pi/k$, $\beta' = (\pi/k)+\beta$, it follows that if there is growth or deposit of $K+\cos\left[2\pi t/k+\beta\right]dt$ in time dt —thus, with period k —then the total deposits in time intervals 0 to 1, 1 to 2, 2 to 3, etc., form a sequence with the same period k.

In the second place, it should be noticed that the case of $k=2$ is peculiar. In place of (12), we have

$$\sigma_2^2 = 2\sigma^2\cos^2\alpha \tag{35}$$

as follows directly from the fact that when $k=2$, $\theta=180°$, and $\cos(180°+\alpha)=-\cos\alpha$. With the phase α small, we have approximately $\sigma_2=\sigma\sqrt{2}$.

Let us now suppose that the data in given order are divided into sets of convenient size—say sets of 120 measurements. Let the arithmetic mean and standard deviation of each set be found. If these quantities—in particular, the standard deviation—show violent fluctuations as we pass from one set to the next set, it may be necessary to handle the material in different sets. But suppose these fluctuations appear to keep within reasonable bounds.

In (6), the data were represented by X_r. Later, in order to emphasize the possibility of different constituents, W_r was used in (16). But, for simplicity, let us now return to X_r as a symbol for the rth element of the data. In the first tests, let the period k be a whole number. Moreover, in place of the functions (21), (22) let us introduce the following, for $k>2$.

$$(36)\qquad u=u_j(k)=\frac{1}{\sigma\sqrt{k}}\sum_{jk-k}^{jk-1}X_r\sin r\theta,\quad j=1,2,3,\cdots$$

$$(37)\qquad v=v_j(k)=\frac{1}{\sigma\sqrt{k}}\sum X_r\cos r\theta,\quad k\theta=360°$$

$$(38)\qquad u'=u_j'(k)=\frac{1}{\sigma}\sqrt{\frac{2}{k}}\sum X_r\sin(r\theta+45°)$$

$$(39)\qquad v'=v_j'(k)=\frac{1}{\sigma}\sqrt{\frac{2}{k}}\sum X_r\cos(r\theta+45°)$$

In the case of $k=2$, replace the radical by $1/\sqrt{2}$. If tests for fractional k are desirable, replace k in (36) to (39) by n, where $n=mk$, m and n whole numbers as in (21).

Here for each individual set—say of 120 measurements—it is assumed that each measurement has the same "expected value" or "Probable value," approximated by the arithmetic mean, and the same "mean error," approximated by the standard deviation σ. In this case, u, v, u', and v' all have the same expected value, zero, by (5), noting that the distributive law holds for expected values. Moreover,

when $k > 2$—see (11), (12), (14), (15)—the mean error of u, v, u', and v' is in each case unity. This is also true when $k = 2$, if the phase has been properly matched—see (35).

To make the tests, then, the functions u, v, u', and v' are computed for certain values of k —perhaps for the sub-multiples 2, 3, 4, 5, 6, 8, . . . of 120. In this way, for $k = 6$, twenty values would be found for each of the four functions. The information thus found may not be very significant. But if not, we may combine results as follows. Let $s = q^2$, where q is a whole number. Let

$$(40) \qquad U_1 = \frac{1}{q}(u_1 + u_2 + \cdots + u_s) \; ; \quad U_2 = \frac{1}{q}(u_{s+1} + \cdots + u_{2s}),$$

etc., and form similar expressions for $V_1, V_2, \ldots U_1', U_2', \ldots V_1', V_2', \ldots$ Each of these functions has expected value *zero* and mean error *unity*. To illustrate—suppose that u_1 (6) = 1.8; $u_2(6) = 2.1$; $u_3(6) = 1.7$; $u_4(6) = 1.8$. These results taken individually would not furnish strong evidence for a period of 6. Some statisticians regard a variation equal to three times the "probable error" or two times the standard deviation as "significant"—in which case (6) 2.1 would be significant. But such evidence is not overwhelming. But, by (40), $U_1 = 3.7$. Here U_1, with mean value zero, has jumped up to an absolute value 3.7 times its standard deviation, unity. On a pure chance basis, in normal distributions, this would happen only about twice in 10,000 trials, on the average. Altho $U_1 = 3.7$ affords no demonstration of a period of 6, the result is at least highly significant. If such high values occur repeatedly in using $k = 6$, we would be justified in asserting that the data contain a constituent with period somewhere near 6.

Moreover, the process (40) is subject to iteration—as long as the data hold out. If $s = q'^2$, then $(U_1 + U_2 + \ldots + U_s)/q'$ is a function with mean value zero, and mean error unity.

When the change in standard deviation is fairly gradual from set to set, the values of $u_1, u_2 \ldots$ can be computed without interruption, using proper adjustments for those values of u_j whose terms arise partly from two sets, such as u_8 (16).

Such a result as $u_2(6) = 2.1$ would furnish evidence only for the six measurements from which it was computed; and in the light of (28), with $m = 1$, the implication at most would be for some period

greater than 3. But $U_i = 3.7$ would furnish strong evidence that in the 24 measurements covered there was a constituent with period between 4.8 and 8—taking $m = 4$ in (28).

The technique of computation would present a few problems. In some cases (6) would be utilized. Certain tables[1] of products with the harmonic factors as multiplicands may be of assistance. Or certain tables may be constructed for use with the aid of an adding machine—with complements listed to take the place of negative numbers. Only u and v in (36), (37) need be computed directly; for u' and v' may be found at once—see (23). But it would seem advisable to compute u' or v' as a check. Graphs showing the progress of the functions u, v, etc., may be constructed.

The interpretation of the results would often be difficult because different sections of the data would frequently give different indications. Again, if two layers of rock are counted as one, an error would be introduced. But this would affect the u_j, v_j, ... involved, not the preceding or following u_j, v_j. Indeed, if an actual period is present, as indicated by the u's, an error of merging may merely shift the "burden of proof" to one of the other functions v, u', or v'. Certain cyclic changes, bringing u, u', v, v' into prominence in rotation, may indicate that the test period k is close to an actual period but with a discrepancy large enough to produce a systematic advance of phase. Many similar principles commonly employed in period testing could be used to advantage in the method here outlined.

---

1 E. g., L. W. Pollak, "Rechentafeln zur Harmonischen Analyse."

Edward L. Dodd