NASA CONTRACTOR REPORT

NASA CR-465

NASA CR-46

NON-LINEAR REGRESSION ANALYSIS AND ANALYSIS OF VARIANCE OF PERIODS DEFINED BY IRREGULAR OBSERVATIONS

by I. Jurkevich

Prepared under Contract No. NASw-880 *by*
GENERAL ELECTRIC COMPANY
Philadelphia, Pa.
for

NATIONAL AERONAUTICS AND SPACE ADMINISTRATION - WASHINGTON, D. C. - MAY 1966

NASA CR-465

NON-LINEAR REGRESSION ANALYSIS AND ANALYSIS OF VARIANCE OF PERIODS DEFINED BY IRREGULAR OBSERVATIONS

By I. Jurkevich

Prepared under Contract No. NASw-880 by
GENERAL ELECTRIC COMPANY
Philadelphia, Pa.

for

NATIONAL AERONAUTICS AND SPACE ADMINISTRATION

For sale by the Clearinghouse for Federal Scientific and Technical Information
Springfield, Virginia 22151 – Price $2.00

TABLE OF CONTENTS

SUMMARY

The purpose of the present study is to develop methods for the determination of accurate periods of periodic events observed at irregular intervals of time. It is pointed out that the powerful theoretical apparatus of time series analysis developed for equally spaced data becomes too restrictive when such spacing cannot be produced, and hence the need for methods applicable directly to unequally spaced ordinates.

Among various approaches to the above problem, two were selected for detailed study as the most promising ones. The first of these is based on the iterative harmonic analysis of observations. The technique employed is primarily concerned with the non-linear regression analysis of data. It is assumed that the residual sum of squares is a function of period and that the "best" estimate of the latter corresponds to the minimum value of the residual sum of squares. An iterative procedure is then employed to search for the above condition.

The second method is based on the analysis of variance technique for a single variable. This method requires that the data be grouped in such a manner as to permit the partitioning of the total variance into two components each of which is a function of period. The "best" estimate of the period is then assumed to correspond to the minimum value of squared deviations measuring the variation within groups. As in the first method an iterative procedure is required to search for the appropriate minimum.

Both methods were tested numerically against short runs of real data, and found to work in principle. From the computational point of view the "analysis of variance" approach is preferable because it involves only the very basic arithmetic operations resulting in significant economy and speed. However, to yield significant results the method requires that the available data record extend over several periods. On the other hand, the "harmonic analysis" approach yields useful results with data covering intervals not much longer than a full period. This gain is achieved at the expense of computational simplicity.

Numerical testing confirmed most features claimed for these methods in the main body of the report. In addition, it was found that in practice the methods suffer from a number of difficulties such as, the less than adequate rate of convergence, occasional appearance of spurious periods, the problem of finding an optimum number of data groups, etc. These problems have not been considered in detail.

1. INTRODUCTION

The present work is concerned with the problem of finding the period of a cyclic phenomenon which has been observed at irregularly spaced intervals of time. This problem is a special case of a broader class of problems dealing with the spectral analysis of an irregularly observed time series. The subject in question is not one on which there is an extensive literature. The existing works on the time series analysis are concerned exclusively with techniques which are applicable to observations equally spaced in the independent variable. The approach based on this fact permits many simplifications in the analysis, leading to theoretical elegance as well as to convenient computational schemes.

There exist, however, physical situations in which it is either impossible or impractical to obtain observations at fixed intervals. Difficulties of such a nature arise whenever experimental conditions are at least partially beyond the observer's control. In such cases, even if it is intended to produce equally spaced observations, the observational technique itself may cause the interval to depart by a large amount from its desired value. Note that this condition is quite different from that which occurs in sampling of data sources in communication and automatic control systems in which the sampling mechanism introduces a small timing error known as "time jitter" (Balakrishnan 1961 - 62, Brown 1963). The latter condition is of no concern in this work.

Consider now the problem of analyzing a time series which has been observed at unequally spaced intervals.

First, it is natural to inquire into the possibility of utilizing the powerful theoretical apparatus of time series analysis developed for equally spaced data.

In order to do this, one would have to replace actual observations by a new set generated from the original one by interpolation. The subsequent choice of one of the numerous existing techniques would depend primarily on the parameters to be estimated as well as on the computational means available.

For the specific case under consideration, namely the period search, many such techniques were described by Stumpff (1937). Among other approaches, not treated in the literature, the most natural one appears to be based on the representation of data by a trigonometric polynomial and subsequent utilization of non-linear regression analysis to determine the period. Equal spacing of observations is of crucial importance in such an approach because under this condition trigonometric polynomials become orthogonal and as a result one achieves significant computational economy and speed.

However, one finds that the results obtained by the above procedure depend significantly on the type of interpolation formula employed in generating equally spaced data points as well as on the number of "manufactured" points used. The degree of this sensitivity may not be serious in some problems, but in others such as those arising in astronomy it is highly disturbing. This situation is fully recognized (for the latest example see Wehlau, Leung, 1964) and yet it is accepted because it seems to be the best thing that one can do. For the reasons stated above as well as the fact that automatic computers make the use of equidistant ordinates less important, it seems desirable in certain cases to abandon the existing methods of analysis, despite their convenience, and attempt to develop methods applicable directly to irregularly spaced observations. Unfortunately, as soon as the condition of equal spacing is disallowed, one loses all advantages which normally accrue from the fact such as, the orthogonality of trigonometric functions. One immediately encounters not only theoretical difficulties, but numerical ones as well, particularly in cases employing non-linear regression analysis. Consequently, any work in this area must be concerned with the development of practical computational procedures in order to yield numerical estimates of the desired parameters.

The initial survey of possible ways of analyzing an irregularly observed time series for the presence of certain periods revealed two promising methods. The first of these is based on the "iterative harmonic

analysis" of the data and the second one on the "iterative analysis of variance" approach. In subsequent discussions these two methods will be referred to as the "Iterative Harmonic Analysis Method" and the "Iterative Analysis of Variance Method".

The remainder of this report is concerned with the details of these two methods.

2. DETERMINATION OF PERIODS - ITERATIVE HARMONIC ANALYSIS METHOD

2-1. General Consideration

Quantities obtained by observation of periodic phenomena are the observed response (mechanical displacements, light intensity, number of sun spots, rain fall, etc.) and the corresponding instants of time. It is pertinent to state that generally the precision with which these two quantities can be observed are vastly different. The precision with which time can be measured is orders of magnitude higher than that of the other quantities mentioned above. This observation has an important implication in the case of the Iterative Harmonic Analysis Method, presently under consideration.

As will be seen later, this method relies heavily on the method of least squares for its operation. For this reason the observational data must satisfy the Gauss-Markoff theorem on least squares. Briefly this theorem consists in the following. Recall that the least squares adjustment of data can be used to estimate the best numerical value of a quantity even though errors are not necessarily the observational ones. Generally, the Gauss-Markoff theorem is concerned with linear estimation of parameters appearing in linear equations. Assumptions which must be satisfied are:

(a) Estimators of parameters of interest are unbiased linear combinations of the observed values of the drawn sample.

(b) The "best" unbiased linear estimator is that one which minimizes the variance of the statistical variables (usually the observed quantity). The distribution law of the observed quantity is not restricted to any particular form.

If the expected value of a variable is

$$E(y_i) = b_o + b_1 x_{1i} + \ldots + b_k x_{ki} + \ldots + b_n x_{ni} \qquad (2\text{-}1)$$

then the sum of the squares of deviations s is given by

$$s = \sum_{i=1}^{N} (y_i - b_o - b_1 x_{1i} - \ldots - b_k x_{ki} - \ldots - b_n x_{ni})^2 W_i \qquad (2\text{-}2)$$

where W_i are known as weights. Minimization of the variance s results in a familiar set of equations known as the normal equations.

A very important fact to note is that for the Gauss-Markoff theorem to be applicable the coefficients x_{ki} must have known numerical values. In fact, they must be error free. In practice these coefficients are known from observations and thus contain the errors of observation. The Gauss-Markoff theorem is not applicable to this situation. It must be pointed out that the range of its application can be stretched if the coefficients are known with such accuracy that they affect the expectation of the variable y to a smaller degree than the standard deviation of any individually measured parameter. This condition will be satisfied as long as the measurement error of the independent variable is much smaller than that of the dependent one.

Although, strictly speaking, the Gauss-Markoff theorem applies to linear parameter estimation, it continues to be valid in non-linear cases provided the problem can be appropriately linearized. The subsequent discussion assumes that such linearization can indeed be carried out. Before proceeding with the main discussion it will be useful to recall that the normal equations resulting from the minimization of s have the following form

$$b_o \sum_i x_{ki} + b_1 \sum_i x_{1i} x_{ki} + \ldots + b_k \sum_i x_{ki}^2 + \ldots + b_n \sum x_{ni} x_{ki} =$$

$$= \sum_i y_i x_{ki} = C_k \qquad (2\text{-}3)$$

where $k = 0, 1, 2, \ldots n$; $x_{oi} \equiv 1$, and the weights W_i were set equal to unity.

If the variables are measured from their respective means, equation (2-3) transforms into a well-known form given by

$$b_1 \sum_i (x_{1i} - \bar{x}_1)(x_{ki} - \bar{x}_k) + b_2 \sum_i (x_{2i} - \bar{x}_2)(x_{ki} - \bar{x}_k) + \ldots$$

$$+ b_k \sum_i (x_{ki} - \bar{x}_k)^2 + \ldots + b_n \sum_i (x_{ni} - \bar{x}_n)(x_{ki} - \bar{x}_k) =$$

$$= \sum_i (y_i - \bar{y})(x_{ki} - \bar{x}_k) = C_k$$

$$b_o = \bar{y} - b_1 \bar{x}_1 - \ldots - b_n \bar{x}_n . \qquad (2\text{-}4)$$

For our purposes, it is immaterial which of these two forms is used for computation and, therefore, we shall limit ourselves to equation (2-3).

In the above equations quantities x_k can represent either explicit independent variables or functions defined in terms of independent variables.

2-2. Formulation of the Problem

The periodic processes are most conveniently represented by trigonometric sequences. If the period of the phenomenon is known, its analytical description in terms of an approximating trigonometric sequence is obtained by deriving the appropriate Fourier coefficients of the sequence. Techniques to treat such problems are well known. The converse problem, that of extracting as accurate a value of the period as the unequally spaced observations will allow is much more difficult and has received little attention.

The numerical difficulties arising from unequal spacing are compounded by theoretical problems associated with the fact the desired parameter - period - enters the regression equation non-linearly.

Let it be assumed that the run of observations y_i' as a function of time t_i can be represented by equation (2-5)

$$y_i(t_i, T, t_o, n) = \frac{1}{2} A_o + \sum_{p=1}^{n} \left\{ A_p \cos \left[\frac{2\pi p}{T} (t_i - t_o) \right] \right.$$

$$\left. + B_p \sin \left[\frac{2\pi p}{T} (t_i - t_o) \right] \right\} . \qquad (2\text{-}5)$$

It is evident that y_i, within our scheme, is not only a function of time, but also of three additional variables. It should be clear that the representation of the observations in terms of our model should improve with the number of terms retained in the sequence.

If we had the correct values of T and t_o the problem would be reduced to finding coefficients A_o, A_p, B_p by the straightforward application of the method of least squares. Of course, a problem which normally remains unanswered is the value of n. In dealing with normal Fourier series where the function y(t) is given by an analytical expression, n can be taken arbitrarily large. In our case, y is defined observationally at discrete points; therefore, one would expect that beyond a certain value of n, coefficients A_n and B_n are unreliable due to a strong effect of observational error. This indicates that the sequence (2-5) must be terminated at some n based on a suitable statistical test. Such a test will be discussed later.

In our case, a further complication is the fact that T and t_o are known but approximately. The quantity t_o represents the origin on the time axis. For the purpose of determining the period, t_o is an arbitrary parameter which can be considered free of observational error and therefore left out of further consideration. As far as T is concerned an approximate value can always be obtained from a plot of the observations. Consider now the question of determining the value of the period such that it is "best" in some agreed sense.

In the first approximation the above problem can be approached as follows. Assume that the observed data can be represented by a truncated Fourier Series. The amplitudes of the various harmonics can be derived from observations by means of any suitable method such as the method of least squares. A complication in this approach is the fact that the period enters the regression expression in a non-linear manner and therefore the straightforward application of the least squares procedure will not directly yield the best value of the period. However, if one makes an assumption that in the neighborhood of the true value of the period the sum of the squares of deviations of the observed and expected values of the function in question

should reach a minimum and furthermore that it behaves according to some reasonable power law, one has here the beginnings of an iterative scheme which should yield improved values of the period.

Consequently to meet our objective it is necessary to implement an iterative computational scheme based on the regression analysis in which one of the parameters to be estimated occurs non-linearly.

The desired procedure would involve the following steps:

1. The regression expression assumed to represent the expected value of the observed variable is a truncated Fourier Series. We start with a constant term plus the first harmonic. The initial value of the period can be estimated from observations. Note that at this point there are three coefficients to be estimated.
2. Using the estimated value of the period, the coefficients are determined by the straightforward application of the method of least squares.
3. The initial value of the period is improved by repeating the entire computation for suitably small increments in the period and searching for that value of the period which results in a minimum value of the residual sum of the squares.
4. Following this the second harmonic is added to the regression expression. At this point the expected value of the observed variable is described by five unknown coefficients and the period which is to be further improved. Taking the value of the period obtained in 3 we again use the least squares method to obtain the coefficients in question.
5. The iterative improvement of the period proceeds now according to the prescription given in 3. At the end of this computation one has at his disposal two values of the period; one resulting from the regression on the first harmonic, the second from the regression on the first and second harmonics. It should be

apparent that the representation of the data by the regression expression is not only a function of the period, but also a function of the number of harmonics included in the regression. However, since we are dealing with the observational data it is not possible to include arbitrarily high harmonics in this analysis, because at some point the amplitudes of higher harmonics will be submerged in observational "noise".

6. The highest usable harmonic can be ascertained by means of the standard statistical technique of testing for the significance of added terms in the regression expression. The most common test utilizes the well known F ratio. This test is then used to establish whether the addition of the second harmonic produced a significant improvement in the data representation. If it is found that the improvement is not significant, the computation is discontinued and the last value of the period is taken as the best value in the sense of the least squares.

7. On the other hand should the improvement be significant, the third harmonic is added and the computations similar to those under 4 and 5 are carried out. The F test is now applied to ascertain whether the addition of the third harmonic is significant. Depending on the result of this test one either terminates the computation or proceeds to add the fourth harmonic, etc.

2-3. Formal Solution

Let T_m denote an approximate value of the true period T. The index m will denote the sequential order of trial values of T. The starting value of T, T_1, can be obtained most readily from a plot of the observations. Furthermore, select a suitable value of t_o. This value can be selected in a number of ways. It can be taken arbitrarily, or it may correspond to a specific state of the physical system under study. In the latter case t_o may be the result of a separate computation. If now n is set equal to unity, equation (2-5) can be written as:

$$y_i(t_i, T, 1) = \frac{1}{2} A_o + A_1 \cos\left[2\pi(t_i - t_o)/T_1\right] + B_1 \sin\left[2\pi(t_i - t_o)/T_1\right] \qquad (2\text{-}6)$$

Coefficients A_o, A_1, and B_1 are obtained by fitting expression (2-6) to the data by the method of least squares. Thus, if equation (2-6) is taken as the equation of condition, the normal equations can be obtained by the use of equation (2-3), in which we identify the following quantities:

$$b_o = A_o$$

$$b_1 = A_1$$

$$b_2 = B_1$$

$$x_1 = \cos\left[2\pi\left(\frac{t_i - t_o}{T_i}\right)\right]$$

$$x_2 = \sin\left[2\pi\left(\frac{t_i - t_o}{T_i}\right)\right]$$

Employing these in (2-3) and writing out the result in matrix form, we have

$$\mathbf{CA = C_y} \qquad \text{or}$$

$$\begin{pmatrix} C_{11} & C_{12} & C_{13} \\ C_{21} & C_{22} & C_{23} \\ C_{31} & C_{32} & C_{33} \end{pmatrix} \begin{pmatrix} A_o \\ A_1 \\ B_1 \end{pmatrix} = \begin{pmatrix} C_{y1} \\ C_{y2} \\ C_{y3} \end{pmatrix} \qquad (2\text{-}7)$$

where

$C_{11} = \left(\frac{1}{2}\right)^2 N$, with N representing the total number of observations

$$C_{12} = C_{21} = \frac{1}{2} \sum_i \cos \left[2\pi \left(\frac{t_i - t_o}{T_1} \right) \right]$$

$$C_{13} = C_{31} = \frac{1}{2} \sum_i \sin \left[2\pi \left(\frac{t_i - t_o}{T_1} \right) \right]$$

$$C_{22} = \sum_i \cos^2 \left[2\pi \left(\frac{t_i - t_o}{T_1} \right) \right]$$

$$C_{23} = C_{32} = \sum_i \cos \left[2\pi \left(\frac{t_i - t_o}{T_1} \right) \right] \sin \left[2\pi \left(\frac{t_i - t_o}{T_1} \right) \right]$$

$$C_{33} = \sum_i \sin^2 \left[2\pi \left(\frac{t_i - t_o}{T_1} \right) \right]$$

$$C_{y1} = \frac{1}{2} \sum_i y_i'$$

$$C_{y2} = \sum_i y_i' \cos \left[2\pi \left(\frac{t_i - t_o}{T_1} \right) \right]$$

$$C_{y3} = \sum_i y_i' \sin \left[2\pi \left(\frac{t_i - t_o}{T_1} \right) \right]$$

$$C_{yy} = \sum_i y_i'^2$$

The coefficients A_o, A_1, B_1 are obtained by matrix inversion

$$A = C^{-1} C_y$$

or

$$\begin{pmatrix} A_o \\ A_1 \\ B_1 \end{pmatrix} = \begin{pmatrix} C^{11} & C^{12} & C^{13} \\ C^{21} & C^{22} & C^{23} \\ C^{31} & C^{32} & C^{33} \end{pmatrix} \begin{pmatrix} C_{y1} \\ C_{y2} \\ C_{y3} \end{pmatrix} \qquad (2\text{-}8)$$

Quantities C^{ij} are elements of the inverted matrix of the coefficients of the normal equations. The inverted matrix has a well-known property that its diagonal terms are related to the standard errors of the estimated regression coefficients. The off-diagonal elements are related to covariances of the estimates. In the above case, these relations are as follows:

$$\begin{aligned} SE(A_o) &= \sigma\sqrt{C^{11}} \\ SE(A_1) &= \sigma\sqrt{C^{22}} \\ SE(B_1) &= \sigma\sqrt{C^{33}} \\ \text{cov}(A_o A_1) &= \sigma^2 C^{12} \\ \text{cov}(A_o B_1) &= \sigma^2 C^{13} \\ \text{cov}(A_1 B_1) &= \sigma^2 C^{23} \end{aligned} \qquad (2\text{-}9)$$

It is useful to keep in mind that, since the matrix of coefficients of the normal equations is symmetric, the inverted matrix must also be symmetric. Thus,

$$C_{ij} = C_{ji} \text{ and } C^{ij} = C^{ji}.$$

In equations (2-9) σ^2 is the error variance of the observed y′ values. The variance is generally not known a priori. It can be estimated, however,

from the residual variation of the observed response y′ about the values predicted by the regression relation. The residual sum of the squares expressed in the notation of equation (2-3) is given by

$$\sum_i (y_{iobs} - y_{icom})^2 = \sum_i y_{iobs}^2 - b_o \sum_i y_{iobs} - b_1 \sum_i x_{1i}\, y_{iobs} - \cdots$$

$$- b_n \sum_i x_{ni}\, y_{iobs} = v^2.$$

where y_{iobs} = observed quantities

y_{icom} = quantities computed from the regression equation

As applied to equation (2-5), this becomes:

$$\sum_i (y_{iobs} - y_{icom})^2 = C_{yy} - A_o C_{y1} - A_1 C_{y2} - B_1 C_{y3}$$

$$- \cdots = v^2.$$

For subsequent analysis, it may be useful to recall that $\sum_i (y_{iobs} - y_{icom})^2$ is known as the residual sum of squares and $(b_o \sum_i y_i - b_1 \sum_i x_{1i}\, y_i - \cdots - b_n \sum_i x_{ni}\, y_i)$ is known as the sum of squares due to regression. In terms of these two quantities, the estimate s^2 of error variance σ^2 is given by:

$$s^2 = \frac{1}{(N-p-1)} (C_{yy} - b_o \sum_i y_i - b_1 \sum_i x_{1i}\, y_i - \cdots - b_n \sum_i x_{ni}\, y_i) \qquad (2\text{-}10)$$

where N is the number of independent observations and p is the number of parameters estimated by least squares.

If T and t_o used in (2-6) were known precisely, computation would essentially be completed by carrying out operations indicated in equations (2-7), (2-8), (2-9), and (2-10).

It should be clear from equation (2-5) that, for a fixed value of n, $v^2 = \sum_i (y_{iobs} - y_{icom})^2$ is a function of both T and t_o and, therefore, one would expect that for proper values of these parameters, the quantity v^2 should reach a minimum. As pointed out earlier, it is only the quantity T which can be employed to minimize v^2. The reason for this is that t_o serves merely as a reference point and, as such, remains in the present scheme as an arbitrary parameter which is not subject to error. Let us now consider ways of obtaining a minimum of the function $v^2(T)$.

The most direct approach to this problem is to assume that, in the neighborhood of the true value of T, T_o, the quantity $v^2(T)$ can be represented by a quadratic function

$$v^2(T) = A + BT + CT^2$$

This assumption is based on the following simple argument. If y_i is, for the moment, taken to be a function of T alone, then in the neighborhood of T_o we can write

$$\Delta y_i = y_{iobs} - y_{icom} = \left(\frac{\partial y_i}{\partial T}\right)(T - T_o) .$$

From this expression we have

$$v^2(T) = \sum_i \Delta y_i^2 = \sum_i \left(\frac{\partial y_i}{\partial T}\right)^2 (T - T_o)^2 = \sum_i \left(\frac{\partial y_i}{\partial T}\right)^2 T^2$$
$$- 2T_o \sum_i \left(\frac{\partial y_i}{\partial T}\right)^2 T + T_o^2 \sum_i \left(\frac{\partial y_i}{\partial T}\right)^2$$

Thus, $v^2(T)$ has the postulated form, provided we set

$$C = \sum_i \left(\frac{\partial y_i}{\partial T}\right)^2$$
$$B = -2T_o \sum_i \left(\frac{\partial y_i}{\partial T}\right)^2 \qquad (2\text{-}11)$$
$$A = T_o^2 \sum_i \left(\frac{\partial y_i}{\partial T}\right)^2$$

Having obtained, in some manner, quantities A, B, and C, the point at which the minimum of $v^2(T)$ occurs is given by:

$$T = -\frac{1}{2}\frac{B}{C} \quad . \tag{2-12}$$

Approximate values of the constants A, B, C can be obtained either from their defining expressions (2-11) or, more directly, as follows:

For some starting values T_1 and t_o the trigonometric sequence approximation to y_i' is computed. This computation also yields the quantity $v^2(T_1)$. The above procedure is repeated twice for different values T_2 and T_3, yielding the corresponding values of $v^2(T_2)$ and $v^2(T_3)$. Consequently, we have produced coefficients for a system of three equations in three unknown constants. This system is given by:

$$\mathbf{V^2 = TA}$$

where:

$$\mathbf{V^2} = \begin{pmatrix} v_1^2 \\ v_2^2 \\ v_3^2 \end{pmatrix}$$

$$\mathbf{T} = \begin{pmatrix} 1 & T_1 & T_1^2 \\ 1 & T_2 & T_2^2 \\ 1 & T_3 & T_3^2 \end{pmatrix}$$

$$\mathbf{A} = \begin{pmatrix} A \\ B \\ C \end{pmatrix}$$

The solution of these equations is given by:

$$\mathbf{A = T^{-1}V^2}$$

Explicitly, the inverse matrix $\mathbf{T}^{-1}$ is given by

$$\begin{pmatrix} \frac{T_2 T_3}{(T_3 - T_1)(T_2 - T_1)} & - \frac{T_1 T_3}{(T_3 - T_2)(T_2 - T_1)} & \frac{T_1 T_2}{(T_3 - T_1)(T_3 - T_2)} \\ - \frac{T_3 + T_2}{(T_3 - T_1)(T_2 - T_1)} & \frac{T_1 + T_3}{(T_3 - T_2)(T_2 - T_1)} & - \frac{T_1 + T_2}{(T_3 - T_1)(T_3 - T_2)} \\ \frac{1}{(T_3 - T_1)(T_2 - T_1)} & - \frac{1}{(T_3 - T_2)(T_2 - T_1)} & \frac{1}{(T_3 - T_1)(T_3 - T_2)} \end{pmatrix}$$

Substituting the constants B and C in equation (2-12) will produce an estimate of T at which the quadratic approximation to $v^2(T)$ has a minimum. However, this is not the best minimum which can be obtained under the circumstances. To continue the process, the estimate of T just obtained is used to recompute the trigonometric approximation to y_i' and hence a new value of $v^2(T)$ is obtained. Hopefully, the fourth value of $v^2(T)$ is smaller than any of the three previously obtained values. Therefore, the largest $v^2(T)$ in the set of four available is discarded and the remaining three with the corresponding periods are used to recompute A, B, C. Employing these in equation (2-12), a new and presumably improved value of the period T is obtained. In addition, of course, we produce an estimate of the corresponding $v^2(T)$. This procedure can, in principle, be continued indefinitely. In practice, several reasons conspire to force suspension of the iteration procedure. First, in computers handling a fixed number of digits, the progressive loss of significant figures in v^2 eventually makes v^2 insensitive to small changes in T. Second, as the order of approximation to y_i' increases, the resulting improvement in representation ceases to be statistically significant. Consequently, it is necessary to employ some rules for discontinuing the computation. In the following paragraphs, one possible version of such a procedure is described.

Note that for any order of approximation n, the computation can be discontinued when either the period or the variance change by less than a preset amount ϵ. Furthermore, in going from the approximation level n to the $(n+1)^{st}$ level, two terms are added, namely those describing the contribution of the next higher harmonic. The question which one would like to answer is whether the addition of these two terms contributes significantly to the description of the periodic process of interest. Such a decision can be made on the basis of the analysis of variance as follows: To carry out the analysis of variance, it is necessary that the results at two levels of approximation be available. Thus, one has at one's disposal the sums of squares due to regression on the first (p-2) variables and due to regression on the p^{th} and $(p-1)^{st}$ variables. In addition, the residual sum of squares is available. Under these conditions the analysis proceeds according to the following table:

Source	Sum of Squares	Degrees of Freedom (D.O.F.)	Mean Square
Regression on p^{th} and $(p-1)^{st}$ variables	ξ	p-(p-2)	$\xi/2$
Regression on the first (p-2) variables	η	p-2	$\frac{\eta}{p-2}$
Residual	ζ	N-p-1	$\frac{\zeta}{N-p-1}$
Total	C_{yy}	N-1	

As an example, note that at the second approximation level, the quantities listed in the table are given by:

$$\xi = \frac{1}{2} A_o^{(2)} \Sigma y'_i + A_1^{(2)} \sum_i y'_i \cos \theta_i + B_1^{(2)} \sum_i y'_i \sin \theta_i$$

$$+ A_2^{(2)} \sum_i y'_i \cos 2\theta_i + B_2^{(2)} \sum_i y'_i \sin 2\theta_i - \eta$$

$$\eta = \frac{1}{2} A_o^{(1)} \sum_i y'_i + A_1^{(1)} \sum_i y'_i \cos \theta_i + B_1^{(1)} \sum_i y'_i \sin \theta_i$$

The quantities ξ and η have two degrees of freedom. Furthermore,

$$\zeta = \sum_i y'^2_i - \frac{1}{2} A_o^{(2)} - A_1^{(2)} \sum_i y'_i \cos \theta_i - B_1^{(2)} \sum_i y'_i \sin \theta_i$$

$$- A_2^{(2)} \sum_i y'_i \cos 2\theta_i - B_2^{(2)} \sum_i y'_i \sin 2\theta_i$$

$$C_{yy} = \sum_i y'^2_i$$

In the above, y'_i are the observed quantities, p represents the number of constants determined in the regression analysis, N is the number of observations, $\theta_i = \frac{2\pi}{T}(t_i - t_o)$, and superscripts indicate the order of approximation.

Returning now to the table, we note that the F-ratio for the effect of added terms is given by:

$$F = \frac{\xi/2}{\zeta/(N-p-1)}$$

Now we make recourse to the table of critical F values computed for some desired probability level. Such tables can be found in most standard texts on mathematical statistics. A hypothesis is made that there is no significant difference between the two variances; that is, that the added terms contribute nothing to the description of the data. If, for the assigned probability level, the computed F ratio is smaller than the critical value, the hypothesis is accepted and the computation is discontinued. Should the

computed value of F be larger than the critical value, n is stepped by 1 and the computation is repeated in its entirety.

When the computation is discontinued, the corresponding value of T is taken as the final value of the period. An estimate of the standard deviation of the final value of the period is given by

$$\sigma_{T_o} \approx \sqrt{\frac{\Sigma(y_{iobs} - y_{icom})^2}{C\ (N-2)}}\ .$$

The above expression is based on the following argument. The procedure for finding the minimum value of T, T_o, can be interpreted as being equivalent to a least squares solution with one unknown T_o. Here the equation

$$\Delta y_i = y_{iobs} - y_{icom} = \left(\frac{\partial y_i}{\partial T}\right)(T - T_o) \qquad (2\text{-}13)$$

serves as the equation of condition which in turn yields

$$v^2(T) = A + BT + CT^2 \qquad (2\text{-}14)$$

as the normal equation. The quantity $v^2(T)$ reaches a minimum for the value of T given by

$$CT = -\frac{B}{2}$$

Here $\frac{1}{C}$ can be interpreted as the inverse matrix $\mathbf{C}^{-1}$ of equation (2-8). In this case this matrix contains a single element. Consequently the standard error of T_o can be written as

$$SE\,(T_o) = \sigma\sqrt{C^{11}} = \frac{\sigma}{\sqrt{C}}$$

Now σ can be estimated from the residual variance by

$$s^2 = \frac{\Sigma(y_{iobs} - y_{icom})^2}{N-p-1}$$

where in the present case p will be equal to 1. Combining these two expressions we obtain σ_{T_o} given earlier.

The following remarks are now in order with regard to the proposed method.

The fact that the observations are not equally spaced results in the loss of advantages which normally accrue from the orthogonality of the sine and cosine functions. Consequently the sums of cross products no longer vanish. Furthermore, even if the period were fixed, the addition of higher harmonics would require recomputation of all coefficients of the normal equations. A similar effect is produced by the fact that the period continuously changes.

These factors conspire to produce an enormous increase in the numerical work required because they lead to nondiagonal matrices of increasingly larger size and to solutions by iterative procedures.

2-4. An Alternate Approach - Differential Corrections Scheme

The method of searching for the minimum values of the sum of squares of deviations is quite awkward in the scheme just described. Furthermore the precision of the final value of the period is obtained from a questionable interpretation of equations (2-13) and (2-14).

The computation and interpretation can be made more direct by linearizing the problem in such a manner that a differential correction to the preliminary value of the period is explicitly included in the appropriate regression expression. For this purpose it will be convenient to re-write equation (2-5) in the following form

$$y_i = A_o + \sum_{p=1}^{n} \left\{ A_p \cos\left[mp\,(t - t_o)\right] + B_p \sin\left[mp\,(t - t_o)\right] \right\} \qquad (2\text{-}15)$$

where the factor 1/2 of equation (2-5) has been absorbed in A_o and $m = 2\pi/T$. Assume now that an approximate value of T is available. This in turn implies that for any value of n approximate values of A_o, A_1, B_1, ..., A_n, B_n are available since they can be obtained by the straight forward application of the method of least squares as outlined in section 2-3. At this point one can ignore the fact that for a fixed n there is only one independent variable, namely T, and consider y_i to be a function of A's, B's, and of m. Assuming furthermore that the first estimate of T is sufficiently close to its true value so that the second and higher powers of the required corrections are negligible, the Taylor's expansion of $y_i(A_o + \Delta A_o, \ldots A_n + \Delta A_n, B_n + \Delta B_n, \ldots m + \Delta m)$ will yield

$$\Delta y_i = \frac{\partial y_i}{\partial A_o} \Delta A_o + \ldots + \frac{\partial y_i}{\partial A_n} \Delta A_n + \frac{\partial y_i}{\partial B_n} \Delta B_n + \frac{\partial y_i}{\partial m} \Delta m$$

or explicitly in terms of equation (2-15)

$$\Delta y_i = \Delta A_o + \sum_{p=1}^{n} \left\{ \Delta A_p \cos\left[mp(t - t_o)\right] + \Delta B_p \sin\left[mp(t - t_o)\right] \right\} +$$

$$+ \sum_{p=1}^{n} \left\{ - A_p \sin\left[mp(t - t_o)\right] + B_p \cos\left[mp(t - t_o)\right] \right\} (t - t_o) \Delta m \qquad (2\text{-}16)$$

The quantity Δy_i is taken to represent the difference between the observed value of y and the value obtained from the first approximation.

Equation (2-16) is now taken as the new equation of condition in which the appropriate parameters $\Delta A_o, \ldots \Delta m$ will be estimated by the method of least squares. The corrections obtained from these linear regression estimates are then used to produce improved values of the starting quantities A's and B's as well as of the non-linear parameter T. The estimate of precision of the resulting value of T is obtained directly from the solution of normal equations. Note that

$$\Delta m = \frac{\partial m}{\partial T} \Delta T = - \frac{2\pi}{T^2} \Delta T = - m \frac{\Delta T}{T}$$

If now Δm as computed from normal equations is given by

$$|\Delta m| = |a \pm SE_{\Delta m}| = m \frac{\Delta T}{T}$$

then

$$\Delta T = |a \pm SE_{\Delta m}| \frac{T^2}{2\pi}$$

Therefore

$$SE_{\Delta T} = \left(\frac{T^2}{2\pi}\right) SE_{\Delta m}$$

This completes the discussion of the operations necessary to implement the iterative harmonic analysis method.

3. DETERMINATION OF PERIOD - ITERATIVE ANALYSIS OF VARIANCE METHOD

3-1. General Remarks

The fact that the methods of sections 2-3 and 2-4 are based on regression analysis leads to a large number of computations which rapidly increase with both the order of approximation n and the number of data points N. The difficulty is further compunded by the necessity of computing a large number of trigonometric functions. As a result the required computations become combersome and slow.

The difficulties can largely be eliminated by the method based on the analysis of variance technique (e.g. Hoel, 1947, Brooks and Caruthers, 1953 etc.)

To apply this method to our case we attempt to group the available data in such a manner that the total sum of squared deviations can be analyzed into two components such that one of these will measure the variation of appropriate quantities between the groups whereas the other measures the variation within the groups. This, of course, represents the basic idea behind the fundamental form of the analysis-of-variance technique. Let it be assumed that such a grouping is possible, and further, that the component variances are the statistical parameters which permit one to make a decision concerning the periodicity of the phenomenon. This decision will be based on the fact that if the data contain a given period, the variance should exhibit a relative minimum for this period. The practical implementation of this idea takes the following form:

1. The computation starts by selecting an arbitrary trial frequency or period.
2. All observations are then reduced to a single cycle by the use of the assumed period. Effectively, this procedure transforms the original independent variable into a new one - the phase.
3. The entire phase range is now arbitrarily divided into a certain number of intervals, each interval containing a certain number of observations.

4. We now compute the means and variances of each group. These will allow us to partition the overall variance into several components each of which can be utilized to obtain an independent estimate of the population variance of individual observations.
5. We can now employ the F-test to ascertain whether the estimated variances are significantly different. This is equivalent to searching for systematic differences between the groups. For a given grouping one expects significant differences between the groups whenever the value of the period under test is too far from its true value. Also, as shown later, the sum of group variances is expected to be smaller than the overall variance. The difference between these can be systematic or it can arise by chance. As before, the F-test can be used to test for its significance.
6. The entire procedure outlined above is repeated to cover the range of periods which are of interest.
7. This procedure will indicate whether the data contains a true period in the neighborhood of the trail period. To establish an improved value a search is conducted over the region in question until the variance exhibits a minimum.

3-2. Formal Description of the Method

Consider for the moment the problem of data grouping. In particular consider a true sine wave defined by Figure 1. Note that for illustrative purposes the values of the period, the spacing interval between observations, and the number of groups are of no consequence.

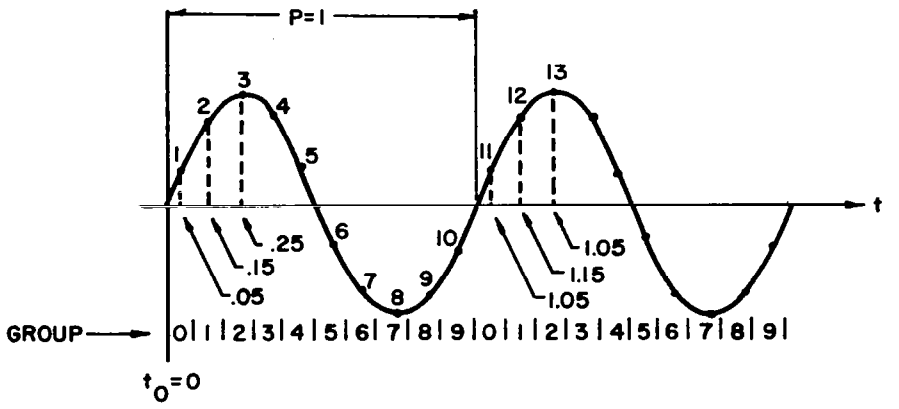

FIG. I GROUPING OF DATA

Using the quantity $q = (t - t_o)/P$ to express the group limits as fractions of the period we can write down the following table

Δq	$10\ \Delta q$	Group Index (INDGR)
.0 - .1	0 - 1	0
.1 - .2	1 - 2	1
.2 - .3	2 - 3	2
...	...	...
...	...	...
1.0 - 1.1	(10—11) - 10 x P	0
1.1 - 1.2	(11—12) - 10 x P	1
1.2 - 1.3	(12—13) - 10 x P	2
...	...	...
...	...	...

This table suggests the following rule for assigning a given observation to a particular class. To obtain INDGR, express time associated with the observation as a fraction of the period and discard the characteristic, if any, of the resulting number. This converts the original variable into a new one - the phase. Multiply the phase obtained in this manner by m, the number of groups into which the observations are being assigned, and

discard the mantissa if any. The resulting integer is then used to denote the group index of the observation in question.

It should be quite clear that the content of a particular group changes with the changing value of the period. To clarify consider 1st, 2nd, 11th, and 12th points in Figure 1. Following the prescription outlined above one can produce the following schematic table summarizing the membership of the points in question in the appropriate group.

TRIAL PERIOD / POINT	.9	1.0	1.1
1	0	0	0
2	1	1	1
11	1	0	9
12	2	1	0

The varying content of groups, will of course, produce different group means and variances which in turn affect the partitioning of the overall variance. It should also be noted that if the spacing of the original observations is random the assignment of a particular point will also be random. For the time being we shall ignore the fact that in practice the observations may not be strictly random and that they may be influenced by systematic effects.

Let us assume now that we have effected classification of data according to the above rule. This will result in a table such as the following one.

GROUP					GROUP MEANS
1	x_{11}	x_{12}	...	x_{1n_1}	$\bar{x}_1$
2	x_{21}	x_{22}	...	x_{2n_2}	$\bar{x}_2$
3	x_{31}	x_{32}	...	x_{3n_3}	$\bar{x}_3$
...	...	...	...	...	...
...	...	...	...	...	...
m	x_{m1}	x_{m2}	...	x_{mn_m}	$\bar{x}_m$
Grand Mean					$\bar{x}$

The sum of the squared deviations from the grand mean $\bar{x}$ for any given row is given by $\sum_{j=1}^{n_i} (x_{ij} - \bar{x})^2$. This quantity can also be expressed as

$$\sum_{j=1}^{n_i} (x_{ij} - \bar{x})^2 = \sum_{j=1}^{n_i} \left[(x_{ij} - \bar{x}_i) + (\bar{x}_i - \bar{x})\right]^2 \qquad (3\text{-}1)$$

$$= \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_i)^2 + \sum_{j=1}^{n_i} (\bar{x}_i - \bar{x})^2 + 2 \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_i)(\bar{x}_i - \bar{x})$$

In the last term the factor $(\bar{x}_i - \bar{x})$ does not depend on j. Consequently this term can be written as

$$2 \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_i)(\bar{x}_i - \bar{x}) = 2(\bar{x}_i - \bar{x}) \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_i)$$

However,

$$\sum_{j=1}^{n_i} (x_{ij} - \bar{x}_i) = \sum_{j=1}^{n_i} x_{ij} - \sum_{j=1}^{n_i} \bar{x}_i = n_i \bar{x}_i - n_i \bar{x}_i = 0$$

Consequently the third term in (3-1) vanishes and we have

$$\sum_{j=1}^{n_i} (x_{ij} - \bar{x})^2 = \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_i)^2 + n_i(\bar{x}_i - \bar{x})^2 \qquad (3\text{-}2)$$

For each group there is an equation of this form. If all these equations are added together one obtains

$$\sum_{i=1}^{m} \sum_{j=1}^{n_i} (x_{ij} - \bar{x})^2 = \sum_{i=1}^{m} n_i(\bar{x}_i - \bar{x})^2 + \sum_{i=1}^{m} \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_i)^2 \qquad (3\text{-}3)$$

The left hand side of this equation represents the overall sum of squared deviations from the grand mean. Since the unbiased estimate of population variance σ^2 is given by the quantity

$$S_T^2 = \frac{\sum_{i=1}^{m} \sum_{j=1}^{n_i} (x_{ij} - \bar{x})^2}{N - 1}$$

It follows that the left hand side of (3-3) is an estimate of $(N - 1)\sigma^2$.

The quantity $\sum_{j=1}^{n_i} (x_{ij} - \bar{x}_i)^2$ represents the sum of n_i squared deviations of observations of a given row about its own mean. As long as the assignment of observations into rows is random each of the above sums represents an estimate of the quantity $(n_i - 1)\sigma^2$. Consequently the quantity $\sum_{i=1}^{m} \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_i)^2$ is an estimate of

$$\sum_{i=1}^{m} (n_i - 1)\sigma^2 = (N - m)\sigma^2 \qquad (3\text{-}4)$$

Let M now represent the population mean. Consider the first term on the right hand side of (3-3). This quantity can be re-written as follows

$$\sum_{i=1}^{m} n_i (\bar{x}_i - \bar{x})^2 = \sum_{i=1}^{m} n_i \left[(\bar{x}_i - M) - (\bar{x} - M)\right]^2$$

$$= \sum_{i=1}^{m} n_i (\bar{x}_i - M)^2 - 2 \sum_{i=1}^{m} n_i (\bar{x}_i - M)(\bar{x} - M) + \sum_{i=1}^{m} n_i (\bar{x} - M)^2$$

Since $(\bar{x} - M)$ is independent with respect to index i, the second term can be written as

$$-2 \sum_{i=1}^{m} n_i (\bar{x}_i - M)(\bar{x} - M) = -2(\bar{x} - M) \sum_{i=1}^{m} n_i (\bar{x}_i - M) = 2(\bar{x} - M)\left\{\sum_{i=1}^{m} n_i \bar{x}_i - M \sum_{i=1}^{m} n_i\right\} \quad (3\text{-}5)$$

Note that since the sample mean $\bar{x}$ is defined as

$$\bar{x} = \frac{\sum_{i=1}^{m} n_i \bar{x}_i}{N}$$

(3-5) becomes

$$-2 \sum_{i=1}^{m} n_i (\bar{x}_i - M)(\bar{x} - M) = -2(\bar{x} - M)\left[N\bar{x} - NM\right] = -2N(\bar{x} - M)^2$$

Consequently we have

$$\sum_{i=1}^{m} n_i (\bar{x}_i - \bar{x})^2 = \sum_{i=1}^{m} n_i (\bar{x}_i - M)^2 - 2N(\bar{x} - M) + \sum_{i=1}^{m} n_i (\bar{x} - M)^2$$

Consider now the expected value of $\sum_{i=1}^{m} n_i (\bar{x}_i - \bar{x})^2$. We have

$$E\left[\sum_{i=1}^{m} n_i (\bar{x}_i - \bar{x})^2\right] = E\left[\sum_{i=1}^{m} n_i (\bar{x}_i - M)^2\right] - E\left[2N(\bar{x} - M)^2\right] + E\left[\sum_{i=1}^{m} n_i (\bar{x} - M)^2\right]$$

$$= \sum_{i=1}^{m} n_i E\left[(\bar{x}_i - M)^2\right] - 2NE\left[(\bar{x} - M)^2\right] + \sum_{i=1}^{m} n_i E\left[(\bar{x} - M)^2\right] \quad (3\text{-}6)$$

At this point let us invoke the following theorem (e.g. Hoel, 1947, p. 65)

Theorem: If x is normally distributed with mean M and variance σ^2 and random samples of size n are drawn, then the sample means $\bar{x}$ will be normally distributed with mean M and variance σ^2/n.

Note that $\bar{x}_i$ is based on n_i observations and $\bar{x}$ on N observations. The application of the above theorem to equation (3-6) yields the following expression

$$E\left[\sum_{i=1}^{m} n_i(\bar{x}_i - \bar{x})^2\right] = \sum_{i=1}^{m} n_i \frac{\sigma^2}{N} - 2N \frac{\sigma^2}{N} + \sum_{i=1}^{m} n_i \frac{\sigma^2}{N}$$

$$= m\sigma^2 - 2\sigma^2 + \sigma^2 = (m-1)\sigma^2 \qquad (3\text{-}7)$$

Consequently the two terms on the right hand side of (3-3) yield additional estimates of population variance, namely,

$$S^2_{BG} = \frac{\sum_{i=1}^{m} n_i (\bar{x}_i - \bar{x})^2}{m-1}$$

$$S^2_{WG} = \frac{\sum_{i=1}^{m} \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_i)^2}{N-m}$$

It must be noted (e.g. Caruthers and Brooks, p. 140; Kendall, 1950, p. 507) that the estimates S^2_{BG} and S^2_{WG} are independent of each other although not of the estimate S^2_T since the latter incorporates both S^2_{BG} and S^2_{WG}. This fact is significant for the subsequent use of the F-test in ascertaining the existence of any systematic differences between the groups.

The quantities S^2_{BG} and S^2_{WG} measuring the variation between groups and within groups provide all the information necessary to ascertain the presence of a given period in the data under consideration.

The convenient standard form of the analysis of variance for the case of a single independent variable is summarized in the following table:

Sources of the Observed Sums of Squared Deviations and the Corresponding Expressions		Degrees of Freedom	Estimates of Variance σ^2
Variation Between Groups	$\sum_{i=1}^{m} n_i(\bar{x}_i - \bar{\bar{x}})^2 = D_{BG}$	m-1	$S^2_{BG} = \frac{D_{BG}}{m-1}$
Variation Within Groups	$\sum_{i=1}^{m}\sum_{j=1}^{n_i} (x_{ij} - \bar{x}_i)^2 = D_{WG}$	N-m	$S^2_{WG} = \frac{D_{WG}}{N-m}$
Total Sum of Squared Deviations	$\sum_{i=1}^{m}\sum_{j=1}^{n_i} (x_{ij} - \bar{\bar{x}})^2 = D_T$	N - 1	$S^2_T = \frac{D_T}{N-1}$

It is of interest to note certain relations between the component variances as well as to examine their behavior with changing period. These relations will be found useful in practical computational schemes based on the present method.

3-3. Auxiliary Relations

A. Relation Between D_T, D_{BG}, and D_{WG}

It is immediately evident from equation (3-3) that the total sum of squared deviations D_T equals in value or exceeds the sum of squared deviations in m groups D_{BG} or the residual sum of squares D_{WG}.

Thus, in numerical work D_T provides an upper limit which cannot be exceeded by either D_{BG} or D_{WG}. Note furthermore that in the problem as formulated here the total sum of squared deviations D_T is not a function of period. The changes in the latter affect only D_{BG} and D_{WG}. This fact can be utilized to search for the presence of a given period by examining the behavior of either D_{BC} or D_{WG} as a function of period and establishing the minima of these functions. In practical work one would favor the use of D_{WG} because of the larger number of degrees of freedom associated with this quantity.

It has already been mentioned that only S^2_{BG} and S^2_{WG} can be used to test the significance of differences between the appropriate variance estimates. This is unfortunate since S^2_{BG} is associated with a rather small number of degrees of freedom unless the number of groups into which

the data is classified is taken deliberately large. The two variances S_T^2 and S_{WG}^2 possessing the largest numbers of degrees of freedom cannot be tested against each other because they do not represent independent estimates of the population variance.

Generally the variance between the groups is larger than the variance within the groups with the exception of those cases for which there is negative intraclass correlation (Kendall 1947, pp. 512-513). Such cases are rarely met in practical work. Thus, with rare exceptions, we can compute the sample value of F, F_S, from the ratio

$$F_S = \frac{S_{BG}^2}{S_{WG}^2} = \frac{(N - m)}{(m - 1)} \cdot \frac{D_{BG}}{D_{WG}} \qquad (3\text{-}8)$$

If now $F_S > F_C$ the difference between S_{BG}^2 and S_{WG}^2 is judged significantly different. If on the other hand $F_S < F_C$ there is no reason to assume that the groups differ systematically from each other. Here F_C denotes the critical value of F taken from the appropriate tables.

The condition for acceptance or rejection of the hypothesis of no difference between groups can be put in the following convenient form

$$\frac{(N - m)}{(m - 1)} \cdot \frac{D_{BG}}{D_{WG}} \gtrless F_C$$

or

$$D_{WG} \gtrless \frac{(N - m)}{(m - 1)} \cdot \frac{D_{BG}}{F_C} \qquad (3\text{-}9)$$

In (3-9) the upper inequality sign applies to the case when the reduction in D_{WG} is not significant whereas the lower sign indicates a significant difference between S_{BG}^2 and S_{WG}^2.

B. Limiting Values of the Period

From the earlier discussion concerning the determination of group indices it is evident that the phase expressed as a fraction of period and restricted to the range $0 \leq \varphi \leq 1$ can be written as

$$\varphi = \frac{t - t_o}{P} - \text{Char}\left(\frac{t - t_o}{P}\right) \qquad (3\text{-}10)$$

In this expression t represents the value of the independent variable (e.g. time) associated with the given observation, t_o is the arbitrary reference time, P is the period, and Char indicates that only the characteristic of the number enclosed in the parenthesis is to be retained.

The phase determined by equation (3-10) will vanish whenever $\frac{t - t_o}{P}$ is exactly equal to its own characteristic that is whenever the mantissa of the resulting number is zero. This will occur for $P = 10^{-k}$ where k is the number of decimal digits in $(t - t_o)$. For example, let $(t - t_o)$ be expressed as X.XXX. If now $P = 10^{-3}$ we have

$$\varphi = \text{XXXX} - \text{Char (XXXX)} = 0$$

This situation will persist for any value of P smaller than 10^{-k} provided one strictly adheres to the rules of retaining the proper number of significant figures in arithmetic operations. It is evident that if these rules are violated there will be values of P between 10^{-k} and $10^{-(k+1)}$ for which equation (3-10) yields phases having non-zero mantissas. In automatic computation the problem of handling significant digits properly is a very inconvenient one and therefore it is generally ignored. Consequently phase computed by this means for periods less than 10^{-k} will exhibit fluctuations causing the value of φ to depart from zero. However, let us ignore this practical difficulty and assume that the arithmetic is carried out properly. In this case every $P \leq 10^{-k}$ will yield zero value for φ and consequently our rule for assigning data into classes will force every point into one class whose index number is

zero. Therefore, all class variances vanish except one, and the value of the remaining variance must be equal to the total variance. The latter represents the upper limit of possible variances. Evidently our procedure will fail to produce any variation in D_{BG} or D_{WG} in response to periods smaller than 10^{-k}. Such periods, even should they exist, could not be detected by the analysis of variance method.

The conditions specified above for assigning all observations into a single group are not the only ones for which such a situation obtains.

Note that the group index number INDGR is given by

$$\text{INDGR} = \text{Char}(m\varphi) = \text{Char}\left\{\frac{m(t - t_o)}{P} - m\,\text{Char}\left(\frac{t - t_o}{P}\right)\right\} \qquad (3\text{-}11)$$

It is evident that $\text{Char}\ \frac{t - t_o}{P}$ will vanish whenever $P > \text{Max}\left|t - t_o\right|$ where $\text{Max}\left|t - t_o\right|$ denotes the largest of all the differences $(t - t_o)$ which can be produced in the given data set. Thus

$$\text{Char}\left(\frac{t - t_o}{\text{Max}\left|t - t_o\right| + \epsilon}\right) = 0$$

where ϵ is arbitrarily small. Evidently this will also be the case for any P given by $m(t - t_o)$ where m is any integer larger than unity. Under these conditions (3-11) becomes

$$\text{INDGR} = \text{Char}\left\{\frac{m(t - t_o)}{P}\right\}$$

This expression will consistently yield zero values for INDGR if the denominator P is equal to or exceeds the value given by

$$P = m\ \text{Max}\left|t - t_o\right|$$

This fact shows that there exists a maximum value of P beyond which all observations will be assigned to the 0^{th} group and again it will be impossible to detect any changes in the component variances. Consequently the region of possible trial periods is restricted at both ends.

C. Behavior of Component Variances S^2_{BG} and S^2_{WG}

The total variance S^2_T is not a function of period and therefore it will remain constant throughout the analysis. However, the values of S^2_{BG} and S^2_{WG} will depend on the specific composition of individual groups. A change in these quantities can be effected only when at least one of the x_{ij}'s switches to a different group in response to changes in the trial period.

Consider now the behavior of INDGR (P) as given by equation (3-11). The variation of INDGR for m = 2 is exhibited in the following table.

P	$2(t - t_o)/P$	$2\ \mathrm{Char}\left(\frac{t - t_o}{P}\right)$	INDGR
$(t - t_o)$	2	2	0
$.9(t - t_o)$	$2.\overline{2}$	2	0
$.8(t - t_o)$	2.5	2	0
$2/3(t - t_o)$	3	2	1
$.6(t - t_o)$	3.332	2	1
$.55(t - t_o)$	3.636	2	1
$.5(t - t_o)$	4	4	0
...	...	...	...

It is apparent that for an observation at time t the index number INDGR remains constant in the intervals

$$\cdots$$

$$2/3(t - t_o) < P < (t - t_o)$$

$$2/4(t - t_o) < P < 2/3\,(t - t_o)$$

$$2/5(t - t_o) < P < 2/4\,(t - t_o)$$

$$\cdots$$

Similar table for m = 3 shows that INDGR remains constant in the intervals

$$\cdots$$

$$3/4(t - t_o) < P < (t - t_o)$$

$$3/5(t - t_o) < P < 3/4(t - t_o)$$

$$\cdots$$

In general INDGR changes by 1 for P given by $\frac{m}{3}(t - t_o)$, $\frac{m}{4}(t - t_o)$, $\frac{m}{5}(t - t_o)$, etc. The above pattern indicates that these limits can be expressed as

$$P_{limit.} = \frac{m(t - t_o)}{\text{Char}\left\{\frac{m(t - t_o)}{P}\right\} + 1} \qquad (3\text{-}12)$$

The above quantity yields the lower limit of the range within which INDGR remains constant. Clearly the upper limit is given by $m(t - t_o)/\text{Char}\left\{m(t - t_o)/P\right\}$. Thus, for a given t, the corresponding INDGR remains constant within the range

$$P_A = \frac{m(t - t_o)}{\text{Char}\left\{\frac{m(t - t_o)}{P}\right\}} > P > \frac{m(t - t_o)}{\text{Char}\left\{\frac{m(t - t_o)}{P}\right\} + 1} = P_B \qquad (3\text{-}13)$$

Within a given data set one expects to find one or more points which yield the maximum value of P_A, $P_{A_{MAX}}$, and similarly points which produce

the minimum value of P_B, P_{BMIN}. Within the range $P_{BMIN} < P < P_{AMAX}$ not a single observation changes its group association and therefore S^2_{BG} and S^2_{WG} remain constant. This shows that the component variances exhibit stepwise variation with P.

D. Location of the Origin, t_o

Expression (3-13) indicates that for fixed values of m, t, and P the range over which the component variances remain constant becomes smaller with increasing t_o. This is clear from the expression for $\Delta P = P_A - P_B$, namely

$$\Delta P = P_A - P_B = m(t - t_o)\left[\frac{1}{\text{Char}\left\{\frac{m(t - t_o)}{P}\right\}} - \frac{1}{\text{Char}\left\{\frac{m(t - t_o)}{P}\right\} + 1}\right]$$

For large values of t_o, $1 << \text{Char}\left\{m(t - t_o)/P\right\}$ and furthermore numerically

$$\text{Char}\left\{\frac{m(t - t_o)}{P}\right\} \approx \frac{m(t - t_o)}{P}. \quad \text{Consequently } \Delta P \approx \frac{P^2}{m(t - t_o)}$$

The range in question can be made as small as desired by taking t_o sufficiently large.

Note now that if the range over which S^2_{BG} and S^2_{WG} are constant is wide the precision with which the period can be determined is low. It appears that the difficulty can be alleviated by choosing t_o sufficiently large thus narrowing the range of constancy of the variances in question. At present it is not clear whether the resulting improvement in precision of the computed period is real or illusory. Basically it is difficult to see why an arbitrary reference number t_o should have any effect on precision with which P can be determined since the choice of t_o does not influence the number of observations used, their inherent precision, the length of the time interval covered,

or the computational technique employed. These are the quantities which would be expected to determine the precision of the final result.

Finally, numerical tests had shown that the depths of the minima of the S^2_{WG} function which was chosen as the indicator of periodicity are functions of m, the number of classification groups. The smaller this number, the shallower is the given minimum.

4. TESTING OF METHODS

The techniques described in section 2 and 3 have been tested numerically by means of computer programs written for the IBM 1620 data processing system. It must be emphasized that the three programs used were intended only for testing the essential features of the proposed methods, and therefore cannot be considered a practical computing tool for large scale data processing by these methods. However, it is believed that the work done so far can serve as the basis for the preparation of practical programs.

Both versions of the iterative harmonic analysis method were tested by applying it to a light curve of the VW Cephei eclipsing system as observed by K. K. Kwee at the Leiden Observatory. These observations are shown in Figure 2. To avoid unnecessarily lengthy computations, only 54 points were selected for processing. These points are shown by open circles. It should be noted that if one hopes to determine the period of length P the data must cover an interval at least that long.

The reference time was arbitrarily taken as JD 2436232.2904 and the starting value of the period was estimated from the plot of the light curve as JD 0.276.

It was found that for n = 1 the first version of the iterative harmonic analysis method diverged, indicating that observation cannot be adequately represented by a simple sinusoid. This is not an unexpected result considering the fairly complex nature of the curve.

The iterative process for n = 2 was found to converge, although slowly.

Repetition of these computations for n = 3, 4, and 5 indicated that the last significant improvement occurs for n = 4.

The value of the period obtained at this stage is equal to .278387 days and is associated with a standard error of 1.261×10^{-3} days. The commonly claimed value determined from observations covering many periods is .27831 days (Kwee 1958, Kopal 1956) with the sixth figure being uncertain. This is a reasonable agreement despite the fact that the observational material employed was vastly different.

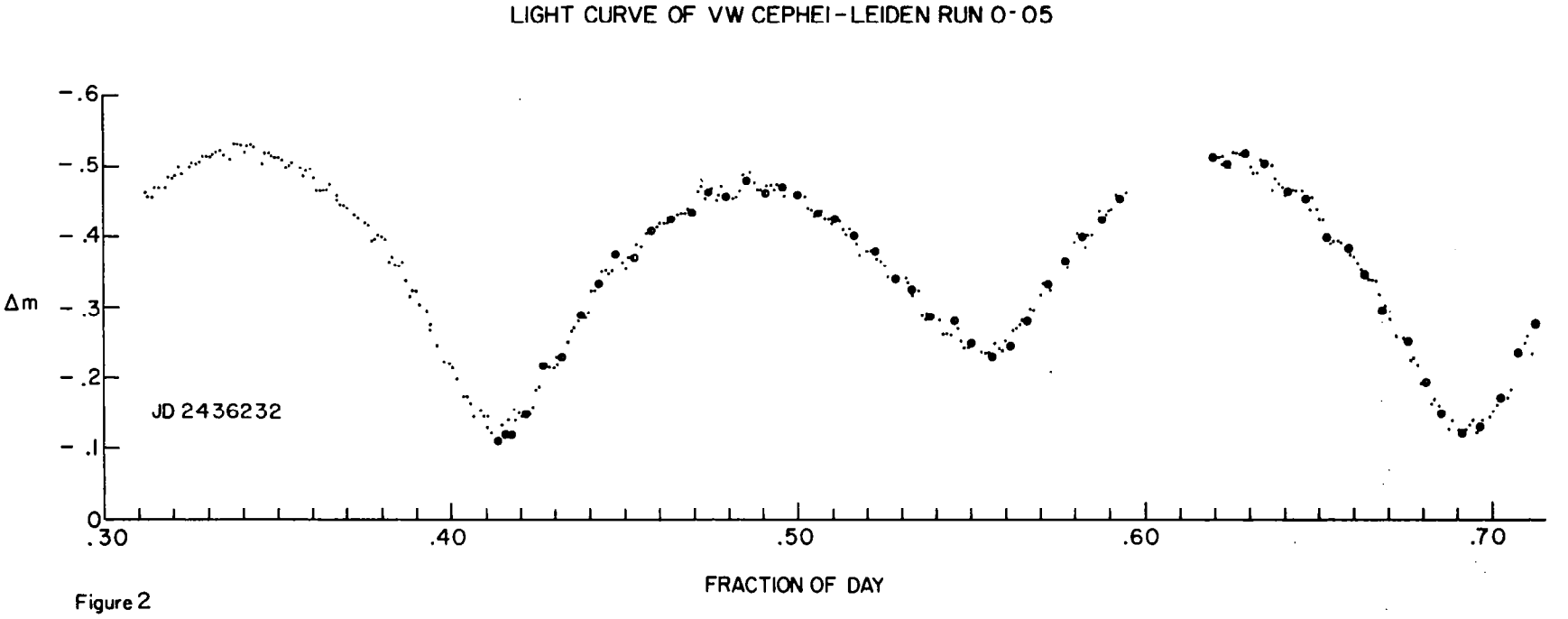
LIGHT CURVE OF VW CEPHEI - LEIDEN RUN 0-05
Δm
-.6
-.5
-.4
-.3
-.2
-.1
0
JD 2436232
.30
.40
.50
.60
.70
FRACTION OF DAY

Figure 2

The differential correction modification of the above method when applied to the same set of data, shows much the same behavior. Thus, the iteration with n = 1 was found to diverge. Computations with successive n values exhibit convergence in fewer iteration steps, but the overall computing time increases due to the larger number of coefficients involved in the process.

Clearly the two procedures must exhibit the same termination point. For n = 4, the last significant improvement in data representation occurs. The period corresponding to this approximation level is .278377 days with a standard error of 1.458×10^{-3}. Thus, the results produced by these two versions are practically the same. From the numerical point of view the differential correction procedure appears to offer little advantage over the unmodified version which requires fewer computational steps per iteration.

Numerical testing of the iterative analysis of variance method proved to be more difficult primarily because of different observational requirements. The iterative harmonic analysis method appears to be applicable to data covering intervals not much longer than a full period. However, the analysis of variance approach proved ineffective under these conditions. The performance of the latter method appears to improve steadily with the increasing length of the data record.

In the VW Cephei case, shown in Fig. 2, this method started yielding results comparable to those of the harmonic analysis method only when the interval covered by observations reached a value of 3 to 4 periods. Consequently the test problem as presently formulated required 150 to 200 points. Unfortunately these were not available in one unbroken sequence and for this reason had to be taken from different, rather widely separated cycles. The above circumstances indicate that the comparison of the two methods by the use of the same data set covering the same overall interval is not possible.

Numerical work with this method revealed all features claimed for it in section 3 particularly with regard to the behavior of the component variance S^2_{WG} in response to changes in P, m, and t_o. In addition a disturbing feature came to light, namely, the presence of spurious periods.

The true period is always found in association with the deepest and probably the absolute minimum of S^2_{WG}. However, the present test revealed that other minima exist in the neighborhood of the deepest one and that such minima are not associated with real periods. The depth, location, and presence or absence of the spurious minima appear to depend at least on m, the number of classification groups.

It must be emphasized that the testing to which this method has been subjected is insufficient to have revealed other hidden features which may be present.

5. CONCLUDING REMARKS

The present work resulted in two possible methods of determining unknown periods of cyclic phenomena which are observed at irregular time intervals. The computational feasibility of each has been demonstrated by numerical tests. The basic differences between the methods stem from the different ways in which the minimum variance is sought. In the harmonic analysis method the least squares technique is used to minimize the residual variance in response to the number of harmonics used and to changes in the period. Since the latter enters the analysis nonlinearly the whole problem becomes concerned with the non-linear regression analysis of irregularly spaced observations. For these reasons the computational aspects of the problem become very cumbersome. Furthermore the method is suitable for the determination of only the fundamental period and it is not capable of revealing periods which are not related harmonically. However, this method has the advantage of yielding useful results from very short runs of data.

The second method is based on the standard analysis of variance technique for a single variable. In this method one of the component variances is minimized as a function of the period. From the computational point of view this method is very desirable because it involves only the basic arithmetic operations in the sense that no special functions need be computed, no matrices formed and inverted, etc. However, theoretically this method contains a greater number of loose ends than the first method. Most prominent among these are the questions associated with the choice of m and t_o as well as the problem of what interactions give rise to spurious periods.

Finally in both methods one is faced with the troublesome questions of convergence of the iterative process and the nature of the solution obtained. These problems have been considered in the literature to a limited extent for the non-linear regression analyses (e.g. Hartley 1961). As far as it is known to this writer problems of the latter type have not been considered for the analysis of variance technique.

REFERENCES

1. Balakrishnan, A. V., 1961 - 1962, On the Problem of Time Jitter in Sampling, IRE Transactions on Information Theory, IT 7-8, 226-236.

2. Brooks, C.E. P., Carruthers, N., 1953, Handbook of Statistical Methods in Meteorology (Her Majesty's Stationery Office, 1953, London).

3. Brown, W. M., 1963, Sampling with Random Jitter, J. Soc. Indust. and Applied Math., 11, No. 2, 460-473.

4. Hartley, H. O., 1961, The Modified Gauss-Newton Method for Fitting of Non-Linear Regression Functions by Least Squares, Technometrics, 3, No. 2, 269-280.

5. Hoel, P. G., 1949, Introduction to Mathematical Statistics, (John Wiley & Sons, Inc., New York).

6. Kale, B. K., 1962, On the Solution of Likelihood Equations by Iteration Processes. The Multiparametric Case, Biometrika, 49, 3 and 4, 479-486.

7. Kendall, M. G., Udny Yule, G., 1950, An Introduction to the Theory of Statistics (Hafner Publishing Company, New York).

8. Kopal, Z., Shapley, M. B., 1956, Catalogue of the Element of Eclipsing Binary Systems, Astronomical Contributions from the University of Manchester (Jodrell Bank Annals), Series I, Vol. 1, No. 4, 99-221.

9. Kozik, S. M., 1964, Otyskanie Perioda po Neskol'kim Razroznennym Nobliudeniam Periodicheskogo Yavlenia (Gidrometeoizdat, Leingrad).

10. Kwee, K. K., 1958, Investigation of Variations in the Period of Sixteen Bright Short Period Eclipsing Binary Stars, Bull. Astron. Inst. Neth., 14, No. 485, 131-142.

11. Parzen, E., 1962, On Spectral Analysis with Missing Observations and Amplitude Modulation, Technical Report No. 46, Statistics Lab. Stanford Univ.

12. Stumpff, K., 1937, Grundlagen und Methoden der Periodenforschung (Julius Springer, Berlin, repreprinted by J. W. Edwards, Ann Arbor, Mich., 1947).

13. Turner, M. E., Monroe, R. J., Lucas Jr., H. L., 1961, Generalized Asymptomic Regression and Non-Linear Estimation Path Analysis, Biometrics, 17, 120-143.

REFERENCES (Cont'd)

14. Wehlau, W., Kam-Ching Leung, 1964, The Multiple Periodicity of Delta Delphini, Ap. J., 143, No. 3, 843-863.

NASA-Langley, 1966