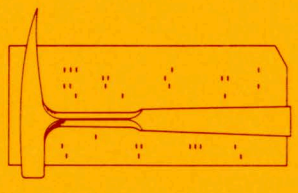
COMPUTER APPLICATIONS IN THE EARTH SCIENCES:

COLLOQUIUM ON TIME-SERIES ANALYSIS

Edited by

DANIEL F. MERRIAM



COMPUTER CONTRIBUTION 18
State Geological Survey
The University of Kansas, Lawrence
1967

Editor's Remarks

One of the great problems facing earth scientists today is that of becoming technically obsolete. I. Stambler and D.M. Graham reported in an article on Outracing Technical Erosion (Industrial Research, August, 1967) that 90 percent of all world scientists and innovators are alive today and are contributing to the "scientific explosion." Workers are becoming increasingly aware of the obsolesence problem, but find it ever more difficult to keep up with the vast amount of literature, even in specialized fields. Stambler and Graham estimate that about 50 percent of a scientist's knowledge will be "useless" in ten years after his graduation.

Many techniques of data storage and retrieval are being tested in an attempt to make pertinent information available quickly according to H.A. Simon (Information Can Be Merged, Think, v. 33, no. 3, 1967). He suggests it is necessary to distinguish between fundamental and transient knowledge, and "...There is little or no time available to stow away heaps of facts and particulars, specific narrow techniques that 'may be useful sometime'." He advocates as effective retrieval system based on a series of "indexes." Many data systems, especially those developed for the petroleum industry are now becoming available to earth scientists. Hopefully, these automated filing systems will increase our efficiency and give earth scientists the data necessary for proper decision-making.

Many organizations are now encouraging their employees to take refresher and continued education courses. In addition, they engage visiting lecturers and specialized consultants to aid in up-dating their personnel. Symposia, colloquia and seminars are held on special topics at all technical levels to keep workers abreast of latest developments. At present, meetings in computer technology are being sponsored by the University of Michigan, Oklahoma Research Institute, and the Kansas Geological Survey.

Special topics in information storage, retrieval, and analysis are being discussed in conjunction with some regional, national and international meetings, for example, meetings were held in conjunction with the recent International Sedimentological Congress and the 6th International Congress on Carboniferous Stratigraphy and Geology, both in Great Britain. These sessions promoted a free interchange of ideas among specialists from all over the world.

To help disseminate the latest information on computer applications in the earth sciences and keep researchers up-to-date, the Kansas Geological Survey publishes the COMPUTER CONTRIBUTION series. This issue, reporting the proceedings of the Colloquium on time-series analysis, brings together people from different disciplines with varied backgrounds to discuss research progress in this extremely interesting area. The sponsors of the Colloquium hope that participants will derive much benefit from their involvement in this meeting. The Colloquium is designed to allow maximum interchange of ideas and information in a limited amount of time.

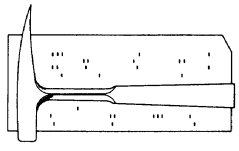
(continued on inside back cover)

COMPUTER APPLICATIONS IN THE EARTH SCIENCES:

COLLOQUIUM ON TIME-SERIES ANALYSIS

Edited by

DANIEL F. MERRIAM



1967

CONTENTS

A COMPARISON OF COHERENCE AND CORRELATION AS MEASURES OF ASSOCIATION FOR TIME OR SPACIALLY INDEXED DATA

by

L. H. Koopmans[1/]

University of New Mexico

INTRODUCTION

Time series or numerical sequences indexed by a space parameter are of considerable interest in geology. The thickness of varves as a function of time measured (yearly) from a convenient origin is an important example of the former, and terrain heights measured at equally spaced stations considered as a function of distance from the "first" station is an example of the latter. Two or more measurements may be made at each time or space index, thus, for varves, thickness, calcium carbonate content and several other measurements are made usually for each lamina in a varve sequence. Along with the various single-series parameters to be computed, it is of great interest to compute measures of association between the various series (see paper by R. Y. Anderson in this volume). For simplicity, I will restrict attention to time series taken two at a time. The corresponding information for spacially indexed series will be obtained easily by converting the terminology from time and frequency to space and wave number. A consideration of parameters for more than two series at a time is not in keeping with the subject matter of this paper.

The statistician's (and geologist's) old standby for measuring the linear association between two sets of numerical quantities $x_1, x_2, \ldots, x_n$ and $y_1, y_2, \ldots, y_n$ is the (sample) correlation coefficient. The formula for its calculation is well known and I will not repeat it here. If the subscript on the x's and y's is a time index; that is, if x_j denotes the x measurement corresponding to j equally spaced time units from a given origin and y_j is the y measurement taken for the same time index, j, then the x_j's and y_j's constitute a pair of time series and a frequency dependent measure of association, called coherence, can be computed. To save space, I will also avoid the details of the formulas and methods for calculating the (sample) coherence (see Jenkins, 1961; Amos and Koopmans, 1963).

As I will indicate in the next section, the sample correlation coefficient and sample coherences can be viewed as "estimates" of corresponding "overall" parameters for a mathematical idealization of the time series. If the number of time units (n) at which measurements are taken is large, there is little need for distinguishing between the sample quantities and the corresponding parameters as the sample quantities tend, in the limit as n goes to infinity, to these parameters. The purpose of this paper is to point out a simple mathematical relationship between the "overall" correlation coefficient and coherence, and on the basis of this relationship, indicate reasons for preferring coherence to correlation as a measure of linear dependence between pairs of time series. I will also describe an actual situation involving a particular varve sequence in which the correlation coefficient provided misleading information about the linear dependence between a particular pair of measurements which was later clarified by computing coherences. The reason for this phenomenon will be explained on the basis of the above mentioned relationship between correlation and coherence.

RELATIONSHIP BETWEEN CORRELATION AND COHERENCE

Imagine that both the x and y series are extended to the infinite past and into the infinite future so that our two sets of numerical quantities constitute n observations from the idealized doubly infinite series $\ldots, x_{-1}, x_0, x_1, \ldots$; $\ldots, y_{-1}, y_0, y_1, \ldots$. By adding constants to the x and y series we can guarantee, thus assume, that

$$\lim_{N\to\infty} \frac{1}{2N+1} \sum_{j=-N}^{N} x_j = \lim_{N\to\infty} \frac{1}{2N+1} \sum_{j=-N}^{N} y_j = 0.$$

(This is essentially the process of removing the D.C. component from each series.)

Now, let

$$\sigma_{xx} = \lim_{N\to\infty} \frac{1}{2N+1} \sum_{j=-N}^{N} x_j^2$$

and

$$\sigma_{xy} = \lim_{N\to\infty} \frac{1}{2N+1} \sum_{j=-N}^{N} x_j y_j \ .$$

[1/] Research supported by the National Science Foundation, grant GP 5217.

Similarly, σ_{yy} is defined as is σ_{xx} with y_i replacing x_i. The quantities σ_{xx} and σ_{yy} are the total mean square amplitudes of their corresponding series--also called power in the terminology of mathematical time-series analysis. If the time series has a stochastic origin and the x and y series are realizations of a stationary time series, then σ_{xx} and σ_{yy} are also the variances of the x and y measurements, respectively. In this situation, σ_{xy} is the covariance between the x and y measurements. Consequently, the "overall" or population correlation coefficient is

$$\rho = \sigma_{xy} / \sqrt{\sigma_{xx}\sigma_{yy}}.$$

In addition, under reasonable conditions, the quantities σ_{xx}, σ_{xy} and σ_{yy} have spectral representations (see Jenkins, 1961) of the form,

$$\sigma_{xx} = \int_{-\infty}^{\infty} f_{xx}(\lambda)d\lambda, \quad \sigma_{xy} = \int_{-\infty}^{\infty} f_{xy}(\lambda)d\lambda, \qquad (1)$$

$$\sigma_{yy} = \int_{-\infty}^{\infty} f_{yy}(\lambda)d\lambda .$$

The functions $f_{xx}(\lambda)$ and $f_{yy}(\lambda)$ are called the spectral densities or, simply, the spectra of the x and y series and $f_{xy}(\lambda)$ is called the cross-spectrum. Thus, $f_{xx}(\lambda)$ and $f_{yy}(\lambda)$ measure the intensity of the power in the x and y series at frequency λ and their integrals over all frequencies "explain" all of the power in these series by virtue of equation (1). The cross-spectral density contains information about the linear relationship between the two series as a function of frequency. This information is obtained more intuitively from two auxiliary functions, the coherence and phase angle:

$$\gamma_{xy}(\lambda) = \frac{|f_{xy}(\lambda)|}{\sqrt{f_{xx}(\lambda)f_{yy}(\lambda)}}$$

$$\theta_{xy}(\lambda) = \arctan \frac{\text{Im } f_{xy}(\lambda)}{\text{Re } f_{xy}(\lambda)} .$$

The functions $f_{xx}(\lambda)$ and $f_{yy}(\lambda)$ are real-valued whereas $f_{xy}(\lambda)$ is generally complex-valued. Then Im $f_{xy}(\lambda)$ and Re $f_{xy}(\lambda)$ stand for the imaginary and real parts of $f_{xy}(\lambda)$ and $|f_{xy}(\lambda)|$ for its absolute value.

The most useful interpretation of the correlation coefficient, ρ, is that ρ^2 is the proportion of the variance (σ_{yy}) of the y measurements which can be attributed to the linear dependence or regression of y on the x measurements. This important interpretation carries over almost directly to the coherence, which provides the reason for the importance of this parameter; $\gamma^2_{xy}(\lambda)$ is the proportion of the power (or more precisely power intensity) of the y series at frequency λ, which is attributable to the linear dependence of the y series on the x series. More intuitively, $\gamma_{xy}(\lambda)$ measures the extent (at frequency λ) to which a linear filter could be designed which would transform the x series into the y series.

The phase angle $\theta_{xy}(\lambda)$ is the average phase lead of the "harmonic component" of the x series at frequency γ over the corresponding component of the y series. It plays an important role in the relationship between correlation and coherence.

Without going into the algebraic details (which follow from (1) and are actually extremely simple), the correlation coefficient can be expressed in terms of the spectral parameters as follows:

$$\rho = \int_{-\infty}^{\infty} \gamma_{xy}(\lambda) \cos \theta_{xy}(\lambda) h_{xy}(\lambda)\, d\lambda . \qquad (2)$$

The weight function $h_{xy}(\lambda)$ is given by

$$h_{xy}(\lambda) = \left[\frac{f_{xx}(\lambda)f_{yy}(\lambda)}{\int_{-\infty}^{\infty} f_{xx}(\lambda)d\lambda \cdot \int_{-\infty}^{\infty} f_{yy}(\lambda)d\lambda} \right]^{1/2} .$$

It can be shown that $0 \le \int_{-\infty}^{\infty} h_{xy}(\lambda)d\lambda \le 1$,

where the 0 value is assumed only if $f_{xx}(\lambda)\ f_{yy}(\lambda) = 0$ for all λ and the value 1 is attained only if $f_{xx}(\lambda) = \text{const.} \cdot f_{yy}(\lambda)$ for all λ. Consequently, $h_{xy}(\lambda)$ measures, in a sense, the degree of similarity between the power spectra of the x and y series. For cyclic data such as varve measurements the spectra consist principally of isolated peaks (Anderson and Koopmans, 1963). Thus, this function will be appreciably different from zero only when peaks in the x and y spectra coincide. This means, returning to equation (2), that contributions to the correlation coefficient for varve series are made by the coherence only at frequencies corresponding to [strong] cycles in both the x and y series. It is possible for the coherence to be near its maximum value of one when $h_{xy}(\lambda)$ is near zero at a given frequency and no essential contribution to ρ will occur.

Note that $h_{xy}(\lambda)$ does not depend upon the

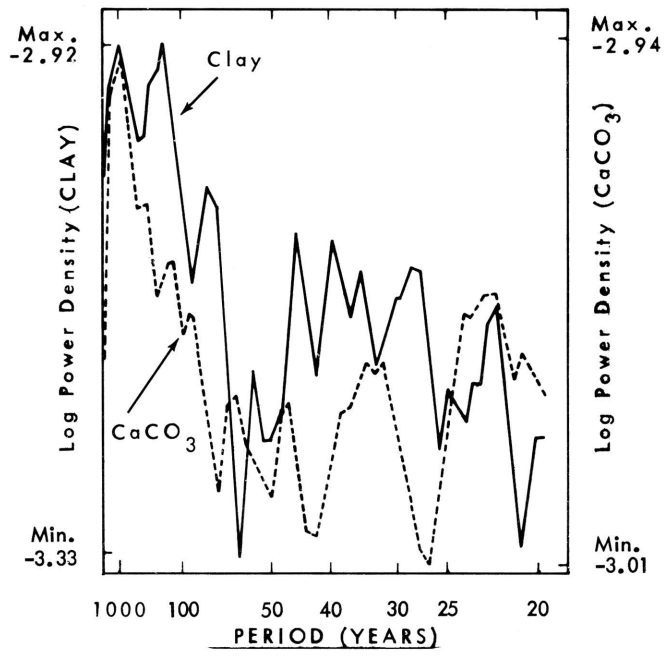


Figure 1.- Log-power spectra of clay and $CaCO_3$.

cross-spectrum, $f_{xy}(\lambda)$. The two terms in the integrand of (2) which do, provide the interesting possibility of high coherence at all frequencies for which $h_{xy}(\lambda)$ is large, yet a negligible correlation! This is brought about by the following possibilities, or a combination of them. If the series are $\pi/2$ or $3\pi/2$ radians out of phase, $\cos \theta_{xy}(\lambda) \cong 0$ and no contribution to ρ will be made at such frequencies. It is possible, however, for a direct relationship between the series to exist at some frequencies ($\theta_{xy}(\lambda) \cong 0$) and an inverse relation to exist at other frequencies ($\theta_{xy}(\lambda) \cong \pi$) which will lead to a cancellation of of the coherences from these two sets of frequencies when the integration is performed in (2). (Note that $\gamma_{xy}(\lambda)$ and $h_{xy}(\lambda)$ are both non-negative so this cancellation is due only to changes in the sign of $\cos \theta_{xy}(\lambda)$.) Thus, the harmonic components at certain frequencies can have "correlations" near +1 at certain frequencies and near -1 at others--indicating strong linear dependence between the series--but the correlation coefficient will be deceptively small. It is such contingencies which lead me to recommend coherence as a replacement or, at least, a supplement for correlation for measuring dependence between time or spacially indexed series. The following extract from a varve study will illustrate that such phenomena actually occur in practice.

COMPARISON OF COHERENCE AND CORRELATION IN VARVE STUDY

A variety of harmonic analyses were carried out on several kinds of measurements made on a 1400-year varve sequence from the Rita Blanca Lake deposits in Texas (Anderson and Koopmans, in press). One particular analysis--the comparison of clay content with calcium carbonate--provided some surprises. An anticipated large negative correlation between these two series failed to materialize. The computed correlation was only -0.12. On the other hand, the computed coherences over spectral regions of high common power for these two measurements were high. A study of Figures 1 and 2 shows that high coherence and large common power occur over the periods 22 years, 30-40 years, 70, 100 and 300-1000 years. By far the highest power is concentrated in the region of 70-1000 year periods (Fig. 1). The high coherence and low correlation seems contradictory until Figure 3 is studied. In the light of expression 2 and the above discussion, the contradiction is easily explained. Over the 70-1000 year periods the phase angle is almost uniformly near -180° which yields $\cos \theta_{xy}(\lambda) \cong -1$. On the other hand, over the two other regions of large common power and high coherence the phase angle is either near 0° or its image on the circle, -360°. For these regions, $\cos \theta_{xy}(\lambda) \cong 1$. Thus, the cancellation of coherence as discussed in the last section takes place here and yields the small correlation. The negative sign of the correlation is attributable to the comparatively large value of $h_{xy}(\lambda)$ over the 70-1000 year periods.

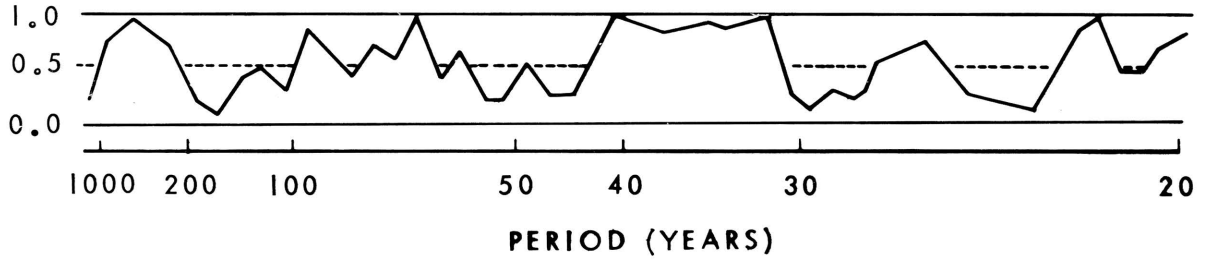


Figure 2.- Coherence between clay and $CaCO_3$.

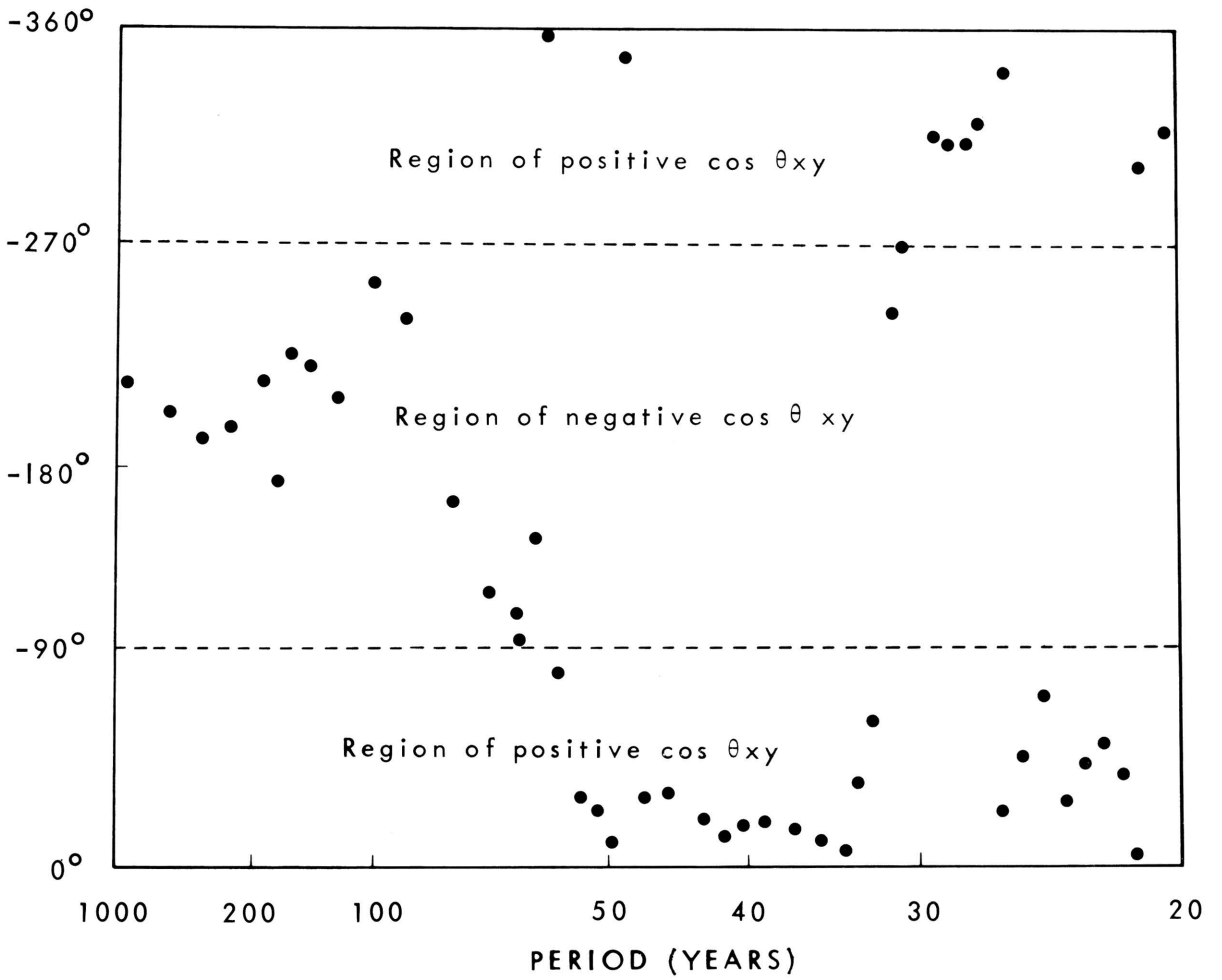


Figure 3.- Phase angle for clay and $CaCO_3$.

REFERENCES

Amos, D. E., and Koopmans, L. H., 1963, Tables of the distribution of the coefficient of coherence for stationary bivariate Gaussian processes: Sandia Corporation Monograph SCR-483 (available from the Office of Technical Services, Department of Commerce, Washington 25, D. C.).

Anderson, R. Y., and Koopmans, L. H., 1963, Harmonic analysis of varve time series: Jour. Geophys. Res., v. 68, no. 3, p. 877-893.

Anderson, R. Y., and Koopmans, L. H., in press, Statistical analysis of the Rita Blanca varve time series, in Anderson, R. Y., and Kirkland, D. W., eds., Paleoecology of an early Pleistocene lake on the High Plains of Texas: Geol. Soc. America Mem.

Jenkins, G. M., 1961, General considerations in the analysis of spectra: Technometrics, v. 3, p. 133-166.

SOME EXPERIMENTS TO SIMULATE THE PENNSYLVANIAN ROCK SEQUENCE OF KANSAS

by

W. Schwarzacher

Queen's University, Belfast, Northern Ireland

ABSTRACT

Power spectra were calculated of a simplified lithological section compiled from R. C. Moore's data for Kansas. The power spectra show that quasiperiodicity is a characteristic feature of cyclothem deposition. Such oscillating series can be generated only by a second- or higher order process. Either autoregression coefficients or transition probabilities of the series may be estimated. Both may be used as a model, and they are compared with reality by calculating the power spectra from simulated runs. The simplest model that gives reasonable agreement is a two-stage Markov chain. To go from one stage into the following in succession, transition probabilities are used that depend on lithologies found at a level of 155 feet below this stage in the section. The time lag of 155 feet is the dominant wavelength in the spectrum of the Kansas stratigraphic section.

INTRODUCTION

Time-series analysis applied to geological data is a relatively new development. The object is to obtain more information from the stratigraphic record than is apparent from casual inspection. Time-series analysis may be carried out for practical reasons, such as interpolation of incomplete sections or for stratigraphic correlation. The main problem, however, will always be the search for a mechanism which explains in physical terms the causes of sediment variation in geological sections. The development of time-series analysis in other subjects, such as economics, meteorology and geophysics, warns us not to expect too much from purely analytical procedures in our search for models. The generating processes or mathematical models responsible for observed time series are more complex than realized at first.

An important complication which is peculiar to geological data is the uncertainty of the time scale. Naturally we can substitute the vertical scale for time in our analysis of sections, but knowledge of the time history is essential to interpret lithologic variation in terms of physical processes. Because of this, it will always be necessary to treat geological data in two stages being the search for (1) a stratigraphic response model with a vertical scale response, and (2) a time process model. If we adopt the analytical approach, this will be the logical order of investigation. Unfortunately, it seems unlikely that this will lead to success in many instances. After a preliminary analysis has been made, it will be necessary to formulate a hypothesis in terms of geological processes. This geological model then may be expressed in mathematical terms and a response model obtained by a simulation experiment. The method not only is hit-and-miss but often it will be difficult to decide if a hit has been made. Nevertheless, it is felt that at present this is the only possible approach to geological time-series analysis. We can only hope that eventually geologically useful information will emerge.

This paper is intended to demonstrate such an approach by using R. C. Moore's rock-column data of Kansas (Moore and others, 1951). The author is unfamiliar with details of the section, but it is perhaps the "classic" example of cyclic sedimentation. The large amount of literature on cyclothems compensates only partially for this lack of geological familiarity, and the author realizes the limitations of such a "theoretical" study.

PRIMARY DATA

The ideal cyclothem (Moore, 1936) is defined in terms of lithologic phases which follow each other in a regular sequence. Moore originally differentiated ten phases, some of which are lithologically identical but can be defined by their position in the cyclothem. Pearn (1964) simplified Moore's classification to five lithologic types which can be used directly for coding lithologic variation. Further simplification leads to a three component system of sandstone, shale and limestone that has been essentially adopted for this study. However, shale containing coal was distinguished from shale without coal, thus, 1 = sandstone, 2 = shale with coal, 3 = shale without coal, 4 = limestone. The question of coding lithologies is obviously important and a problem which must be solved by geological argument rather than mathematical analysis.

The next decision to be made was the choice of the vertical measuring interval. Many authors avoid this problem by using beds of uniform lithologic composition as units; the thickness of the bed is deliberately ignored. This method may be useful, providing that the geological situation really indicates that each bed is a unit in itself and represents

a "phase" in the vertical development of the facies. This implies that a large amount of geological interpretation has to go into the designation of each unit. It is necessary, for instance, to decide on geological grounds whether two or more beds of identical lithology have followed each other in succession. The resulting series of alternating stages can be brought into a time relationship only if something is known about the duration of each stage. The available data certainly does not justify the bed-equal-unit approach. A rough analysis of thickness frequency distributions of sand, shale and limestone in the Kansas rock column indicates a large variability with bimodal or polymodal distribution. For example, shale has a mode both in the 2 to 4 foot thickness class and in the 12 to 14 foot class. If the bed-equal-unit approach is chosen, two shales which differ according to their behavior in thickness would be classified together. In fact, it seems doubtful if any statistical examination of vertical sections can afford to ignore the additional information contained in the thickness data.

In this investigation, the following procedure was adopted. Coded lithology and thickness to the nearest 0.5 of a foot were taken from the graphic columns in Moore's paper (Moore, 1936). A section starting with the Brownville Limestone at the Pennsylvanian-Permian boundary and extending downward to the Hepler Sandstone at the base of the Pleasanton Group was compiled. Measurements from Moore's section could not be made with great accuracy, but a check was provided by comparing the total thickness of the transferred values with Moore's total thickness. The values agree to within 1.5 percent of the total length of the 2000-foot section. A computer sorting program was used to determine percentage values of coded lithologies for specified increments on the vertical scale, together with lithologies which made up the highest percentage in each interval. The latter will be called dominant lithologies. For reasons to be discussed, an interval of 5 feet was found suitable for most problems, thus giving 400 values for the investigated time series.

PRELIMINARY ANALYSIS

An important feature of any time series is its stationarity. This property is difficult to examine, particularly if we are dealing with a single record of limited length. Stationarity in the broad sense is present if the series possesses a time-invariant mean and autocorrelation function. The determination of a mean or autocorrelation function, therefore, should not in practice depend on the position of the sample series within the section under investigation, and the presence of a definite trend would reduce such stationarity. We investigate stationarity by partitioning the entire section into subsamples and comparing the mean values of the variables. The mean percentage of sandstone for 50-foot intervals was used for a "runs" test (Bendat and Piersol, 1966), and results suggest that one may accept the hypothesis that subsequent 50-foot intervals are independent and there is no underlying trend. The graph shows, however, that the fluctuations are not without geological meaning, as they reflect the stratigraphic stages which have been used to subdivide the Kansas column. It is believed that this indicates significant deviation from stationarity but for the present this will be ignored.

To obtain a general description of the series, estimates of the power spectra were calculated. The method suggests itself because of the so-called "cyclic" nature of the sediments in this section. Estimates of power-spectral densities (Bendat and Piersol, 1966) were obtained by "Hanning" the periodogram of observed autocovariancies. The standard sample interval of 5 feet was determined by requirements of the spectral analysis. It seemed desirable to search for oscillations in the wavelength range of 500 to 10 feet. The highest frequency equals the number of lags used in the estimation of the autocovariance function. This number was kept at 50 in most examples.

The wavelength in feet therefore can be found from the frequency ν as shown on the diagrams by the relation: $\lambda = \frac{50}{0.1\nu}$. Furthermore, the wave length in feet also is indicated at the important peaks of the spectra.

Spectra can be calculated either from the coded data of dominant lithologies or from the percentage data; both results are comparable. In all instances individual spectra must be calculated for each lithology. Dominant lithologies are coded in a sequence of +1 (present) and -1 (absent) which is treated in exactly the same way as a series of percentage values.

Typical Pennsylvanian spectra are shown in Figure 1. Both shale and sandstone show a pronounced peak in the 166-foot wavelength. The limestone peak is shifted into the 125-foot wavelength, having a minimum coinciding with shale-sandstone peak. Subsidiary peaks occur at 50 and 35.7 feet and these are close to the 3rd and 4th harmonic of the main peak. The sandstone and shale maxima are without doubt due to cyclothemic sedimentation, as the wavelength of approximately 150 feet coincides well with the average thickness of cyclothems, particularly in the lower and middle part of the section. The limestone spectrum could be caused by an effect of erosion which frequently removed limestone as indicated where followed by sandstone. But there is evidence also of genuinely shorter limestone "cycles" and these may be individual members of megacyclothems. The interpretation, however, must be tentative at this stage for a number of reasons. The present system of coding, for instance, may not be detailed enough to enable one to recognize megacyclothems. Changes in cyclothem thicknesses may occur also and complicate the analysis.

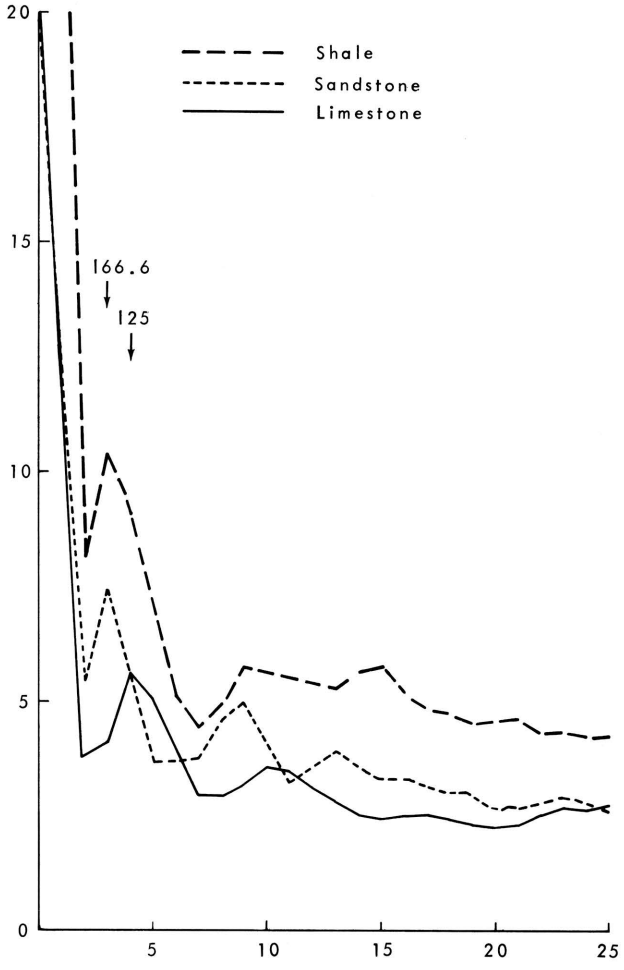


Figure 1.- Spectra of composite Kansas stratigraphic section. Dominant lithology at 5-foot intervals.

All spectra indicate that, in mathematical terms, the series can be described as being generated by a quasiperiodic random process. It is suggested therefore that the definition of cyclothem deposition logically should be the following "...the arrangement of one or more lithologies in a vertical section in such a way that there is a preferred period of occurrence when a complete record of the section is examined; more simply, when the power spectrum of the investigated record shows a significant peak." Unfortunately, geological definitions of cyclothems and "cycles" in general are directed usually more towards the order in which various rock types occur. This is of geological significance, but a different aspect from the problem considered here. It is obvious, for instance, that a completely ordered sequence of lithologies can give a white noise spectrum if the thickness of the lithologic units vary at random. If quasiperiodicity in the described sense is accepted as the characteristic feature of cyclothems, then the reality of "cyclic" sedimentation can be tested by investigating the significance of the peaks in the spectrum. This can be done in two ways either by the argument that (1) spurious peaks in the spectrum are not likely to have peaks at their harmonics as the Kansas spectra, or alternatively, (2) confidence limits to the power spectra can be constructed (Granger and Hatanka, 1964). There is no doubt about the reality of the frequency peaks in the Kansas spectra (Fig. 2).

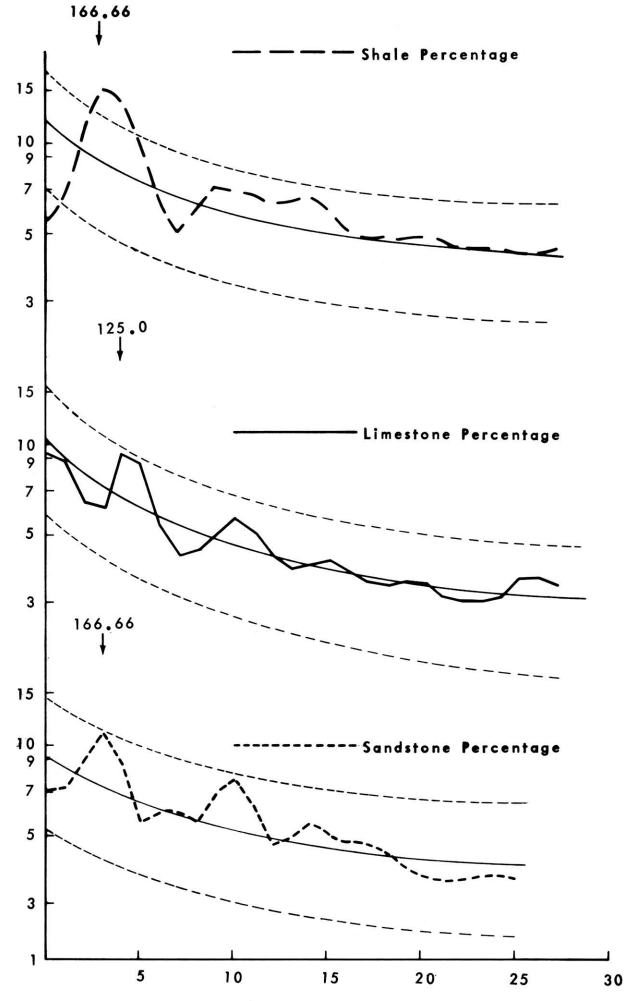


Figure 2.- Spectra of Kansas section showing percent of lithology with 10 percent confidence limits.

Further descriptions of cyclothems may be obtained by investigating the cospectral properties of individual lithologies. We consider, for instance, the records of the sandstone and shale sequence. The cospectrum then is calculated from the cross correlogram of the two records. The cospectral density function is a complex quantity and can be represented by two components, the coherence and phase. Coherence is a measure of amplitude and will be high if both series have maxima or minima in the

identical frequency band. The phase diagram shows how far the first series lags behind the second series. From the amount of lag, it is possible to calculate the average thickness of lithology for each frequency. For the 166-foot wavelength of the section, the following lags were found, sandstone 161 degrees behind shale, shale 166 degrees behind limestone. These can be translated into thickness, and we obtain for this frequency a cyclothem consisting of 13 feet sandstone, 137 feet shale and 17 feet limestone. This result should not be regarded simply as an average cyclothem that could have been obtained by averaging thickness measurements together with a so-called "modal cycle" (Duff and Walton, 1962), as the latter does not carry frequency information. At the same time, it must be realized that the cyclothem description derived from spectral analysis is not accurate, as for instance in this example where the frequency bands adjacent to the 166-foot band have a wavelength of 250 feet and 125 feet, respectively. This relative coarseness of estimation is not a fault of the method but of data, which are such that no more accurate and at the same time meaningful information could be obtained from them.

Cospectral analysis can be generalized to any number of variables by investigating the lag of cross-correlation matrices. Coded lithologies, which we may call different states, use a matrix of cross-association coefficients recording the frequency with which state i follows j at any specified lag. In this situation it is useful to transform the correlation matrices into stochastic matrices; this is simply achieved by making the rows add up to unity. Such a matrix gives the transition probabilities between states, i.e. lithologies, and can be useful for description and sometimes for the interpretation of lithologic variation. The transition probability matrix contains the same fundamental information as the correlation matrix; the selection of the method which one uses depends on the nature of the problem.

VERTICAL SCALE RESPONSE MODEL

It is now our aim to formulate a mathematical model in such a way that it can be used to simulate the stratigraphic record. At the same time, we will try to keep the geological implications of any such model in mind.

In the previous section we have proposed that quasiperiodicity should be regarded as the typical feature of "cyclic" sedimentation, and we can therefore specify that the theoretical model must be a process of second or higher order. A first-order Markov process, for instance, can only be regarded as a limiting case, because its power spectrum has a maximum only at zero frequency (Bartlett, 1962). Nevertheless, the first-order transition probability matrix gives a considerably better description than can be obtained by assuming random changes in lithology. The estimated transition probabilities for 5-foot intervals in the Kansas section are given in Table 1.

Table 1.- Estimated transition probabilities for 5-foot intervals in Kansas section.

0.6456	0.1266	0.1519	0.0759
0.0625	0.3125	0.3125	0.3125
0.1272	0.0060	0.7212	0.1456
0.0476	0.0119	0.3452	0.5953

Testing the hypothesis (Anderson and Goodman, 1957) that the events in the sequence are independent against the alternative that they are controlled by a first-order Markov process gives a $-2ln\lambda$ value of 176.01, which indicates that the hypothesis of independent states can be rejected with great confidence.

In order to see the type of section which would result from a first-order Markov model, a simulated section has been calculated together with the power spectra (Fig. 3).

Figure 3.- First-order Markov model.

As predicted, there is no significant peak in this spectrum apart from the zero frequency maximum. The simulated section (Fig. 4) indicates, as one would expect, the most likely sequence from sandstone to shale to limestone, but it is difficult to pick out successive cyclothems. Geologically, the first-order Markov process can be interpreted as a facies continuity, in the sense that if changes in the environment take place they have to follow a definite pattern which is logically fixed. Thus, it is not possible (as a rule) to change from shallow into deep water without passing through a state of intermediate water depth.

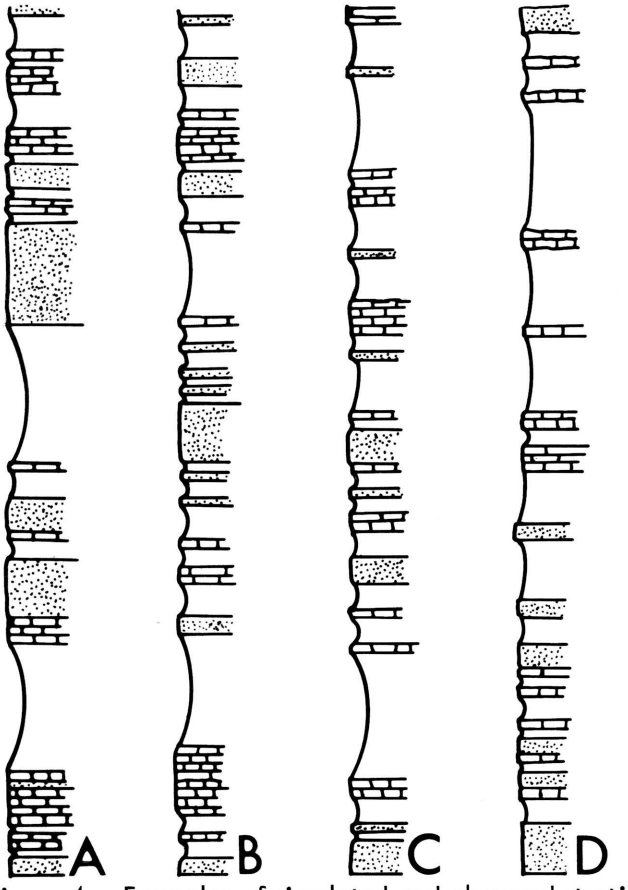


Figure 4.- Examples of simulated and observed stratigraphic sections; A, Kansas; B, second-order Markov; C, autoregression; D, first-order Markov.

Before discussing higher order processes, it seems profitable to review briefly how quasiperiodicity can be introduced by means of geological processes. We consider two crude models, both based on the structural behaviour of a piece of crust.

Model 1

Consider a piece of crust which we assume to be floating and perhaps supported by a central pivot (Fig. 5A). If this is brought out of equilibrium, it will oscillate around its pivotal point with a characteristic frequency. The oscillations may fall off rapidly if damping is present. If next we assume that the oscillations are excited by random impulses, then we have an example of Yule's famous random disturbed pendulum which is the classic model for an autoregressive process (Yule, 1927). The movement of the pendulum is described by the stochastic differential equation

$$\frac{d^2x_t}{dt^2} + b_1\frac{dx_t}{dt} + b_2x_t = \epsilon_t$$

Model 2

Consider an area of uplift separated by rigid crust from an area of subsidence. Assume downwarp is linked casually with uplift, for instance, due to increased sedimentation when uplift occurs. In turn, uplift is caused by subsidence after a more-or-less constant time lag, which could be due to a slow-moving undercurrent (Fig. 5B). In this model, the movement of the basin area depends not only on the immediate past but also is strongly determined by the state of the basin at a time before the present. This is precisely the model which was chosen by Whittle (1954) to represent Alfven's theory of sunspot formation. It can be summarized by the stochastic difference equation

$$\frac{dx_t}{dt} + C_1X_t + C_2X_{t-} = \epsilon_t$$

The two models do not only apply to tectonic control as used in this illustration but could be developed for a variety of environmental factors such as climatic or biological control. For the present it is only

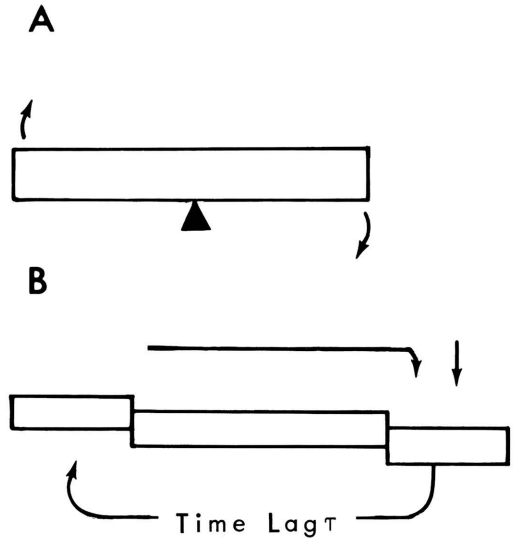


Figure 5.- Models: A, disturbed pendulum; B, time-lag.

necessary to investigate if either of the models can be used as a stratigraphic response model. At the same time, it must be realized that much more complex processes can be constructed which may be mixed differential difference equations of any order and which also will lead to quasiperiodicity.

Estimation of an Autoregressive Scheme

Whittle's (1954) method of examining the two types of models is to fit autoregression coefficients to a time series and to compare the residual variance v_1 of a scheme in which α coefficients have been fitted against the residual variance v_2 of a scheme in which only $\alpha-\beta$ coefficients were fitted. The statistic

$$\psi^2 = (n-\alpha)\ln\left(\frac{v_1}{v_2}\right)$$

is approximately χ^2 distributed with β degrees of freedom. Thus, if ψ^2 is significantly large, the β excluded coefficients are essential in representing the scheme. The full procedure can be found in Whittle's paper.

We have examined the autocorrelations of the sandstone, shale and limestone percentages. In each instance, the following model has been fitted

$$X_{(t)} = A_1X_{t-1} + A_2X_{t-z} + A_{30}X_{t-30} + \epsilon_{(t)}$$

Where x stands for lithological percentage at time t and the time lag 30 (i.e. 150 feet) represents the dominant frequency of the cyclothems. The results are interesting. For sandstone, the scheme is

$$X_{(t)} = .826X_{t-1} - .301X_{t-z} + .072X_{t-30} + \epsilon_t \qquad (1)$$

Although the coefficients for lag 30 gave a small value, Whittle's test shows that it contributes significantly. Comparing the variance of the scheme derived from coefficients A_1, A_2 with the variance of a scheme incorporating coefficients A_1, A_2, A_{30} we find

$$\psi^2 = 397\ln\left(\frac{.4585}{.3917}\right) = 28.31$$

The probability of obtaining ψ^2 as large by chance is <0.001. The shale and limestone models differ considerably from the sandstone model. For shale we find

$$X_t = .662X_{t-1}\ .005X_{t-2} - .001X_{t-30} \qquad (2)$$

with $\psi^2 = .75$, $p \approx 0.4$;

and for limestone

$$X_t = .564X_{t-1} + .013X_{t-2} + .041X_{t-30} \qquad (3)$$

with $\psi^2 = 1.01$, $p \approx 0.2$

In both the limestone and shale regressions, the lag 30 coefficients do not contribute significantly to the scheme. In fact, the lag 2 coefficients are so small that we will not be too wrong if we assume first-order Markov properties for shale and limestone.

The three series have been considered here as independent for simplicity. Further, it is realized that the estimation of the autoregressive coefficient is inaccurate at the relatively small sample size (n = 400) used. We use this method principally to aid our intuition in constructing the first simple model to simulate the series.

We assume that it is only the sandstone which determines the quasiperiodicity of the sequence. Sandstone sedimentation is caused by a random process incorporating a memory which extends at least over the length of the cyclothem. The simulation is carried out in two steps: first, a sandstone series is generated by making use of equation (1) and a random number generator. The series is standardized to unit mean. Thus, if the mean sandstone series is multiplied with the first row of the first-order transition probability matrix, the latter stays unchanged. If the series oscillates, the first row of the transition probabilities will be biased correspondingly. In stage two, a series of transition probability matrices is calculated and from these the section is simulated. Simulation runs for about 2000 feet give the power spectra which are shown in Figure 6 and the section given in Figure 4C. The results are an improvement on those obtained with the first-order Markov model. The most prominent peak in the spectrum is at approximately 50 feet and coincides reasonably well with the theoretical oscillation period of a two-term autoregressive scheme. Making use of the coefficients a_1 and a_2, we obtain 43.75 feet for this period (Kendall, 1945). Peaks around the 150-foot period are indicated but obviously the 30-lag term has been underestimated. Nevertheless, if one studies the simulated section one can make out the larger periods (megacyclothems?). The model is primitive in the sense that it regards sandstone sedimentation up to the first-order Markov relationship as independent from the other variables and this is not a realistic assumption. If we investigated the limestone sedimentation as an independent variable we find that the 30-lag term is insignificant but the correlogram indicates that a 20-lag term (wavelength = 100 feet) may be significant. Fitting the three autoregressive coefficients we find for limestone

$$X_t = .534X_{t-1} + .002X_{t-2} + .066X_{t-20} \qquad (4)$$

and

$$\psi^2 = 397 \ln \frac{.6230}{.4866} = 98.00$$

which is highly significant. Similar estimates for the sandstone and shale sequence show that the 20-lag term does not significantly contribute to the scheme.

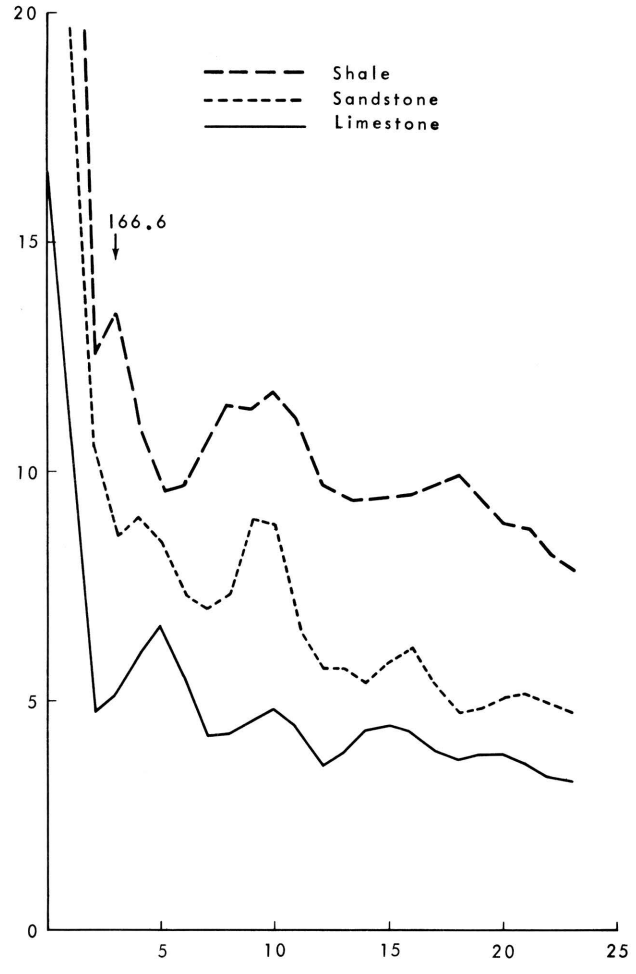


Figure 6.- Spectral analysis of lagged autocorrelation run.

This may indicate a separate mechanism at work that determines the limestone sedimentation, and more sophisticated models may have to incorporate terms for both limestone and sandstone sedimentation. Obviously, two regression equations could no longer be regarded as independent and more has to be learned about the phase relationship of sandstone and limestone. This may be attempted at a later stage with more suitable data.

Estimation of Transition Probabilities

A more direct approach to the simulation problem is the estimation of transition probabilities of lithologic states in the sequence. Although this estimation procedure is straightforward from a mathematical point of view, the resulting matrices of transition probabilities are not as easily interpreted by the geologist as the autoregressive model. We stress once more, however, that fundamentally the two methods are identical.

We concern ourselves again with a second-order model. Thus, if at time $t - \tau$ the state i occurred, we find a matrix of probabilities such that at time $t + 1$ the state k occurs provided that state j occurred at time t. In the instant of the four lithology system, we will therefore be concerned with four $[4 \times 4]$ matrices for each time lag τ. If τ becomes zero, we have a simple Markov matrix. In order to establish the most significant value of τ we use the likelihood ratio criterion

$$\lambda = \prod_{i,j,k}^{4} (\hat{P}_{jk} / \hat{P}_{ijk})^{n_{i,j,k}}$$

where $\hat{P}_{jk}$ and $\hat{P}_{ijk}$ are estimated transition probabilities and n_{ijk} is the number of individuals in state i at $t - \tau$, state j at t and state k at $t + 1$. The statistic $-2\ln\lambda$ has an assymptotic χ^2 distribution under the null hypothesis that the chain is first order against the alternative that it is second order (Anderson and Goodman, 1957). The test criterion $-2\ln\lambda$ has been calculated for $\tau = 0$ to $\tau = 45$ and is shown in Figure 7 together with the 0.001 confidence limits of the distribution. At zero lag, the null hypothesis tested is the hypothesis discussed earlier, that the series consists of independent random variables against the alternative of a first-order Markov chain. The test statistic gave the highly significant value of 176.01. Taking into account the possibility of a second-order Markov chain we obtain a highly significant value of 103.13 at $\tau = 31$. Thus the hypothesis of the first-order Markov chain is abandoned in favor of the second-order chain and we formulate our model. Provided the lithology at $t - 31$ has been 1, 2, 3 or 4, then the transition probability from the state at time t to the state at time $t + 1$ is given by four $[4 \times 4]$ matrices (Table 2). The transition probabilities of Table 2 have been used for simulation experiments to obtain spectra and sections (Fig. 8, and 4B) which show considerable agreement with reality. Curiously enough, the best agreement now is in the limestone spectra which show similar peaks in the 100-foot wavelength band, closely followed by the shale spectrum. The sandstone maximum that should occur at the 166-foot band is not well developed. This again may suggest that a different mechanism is needed for the control of sandstone sedimentation. The comparison of the simulated with the observed spectra generally indicates that the actual sedimentation processes are more regular than a second-order model.

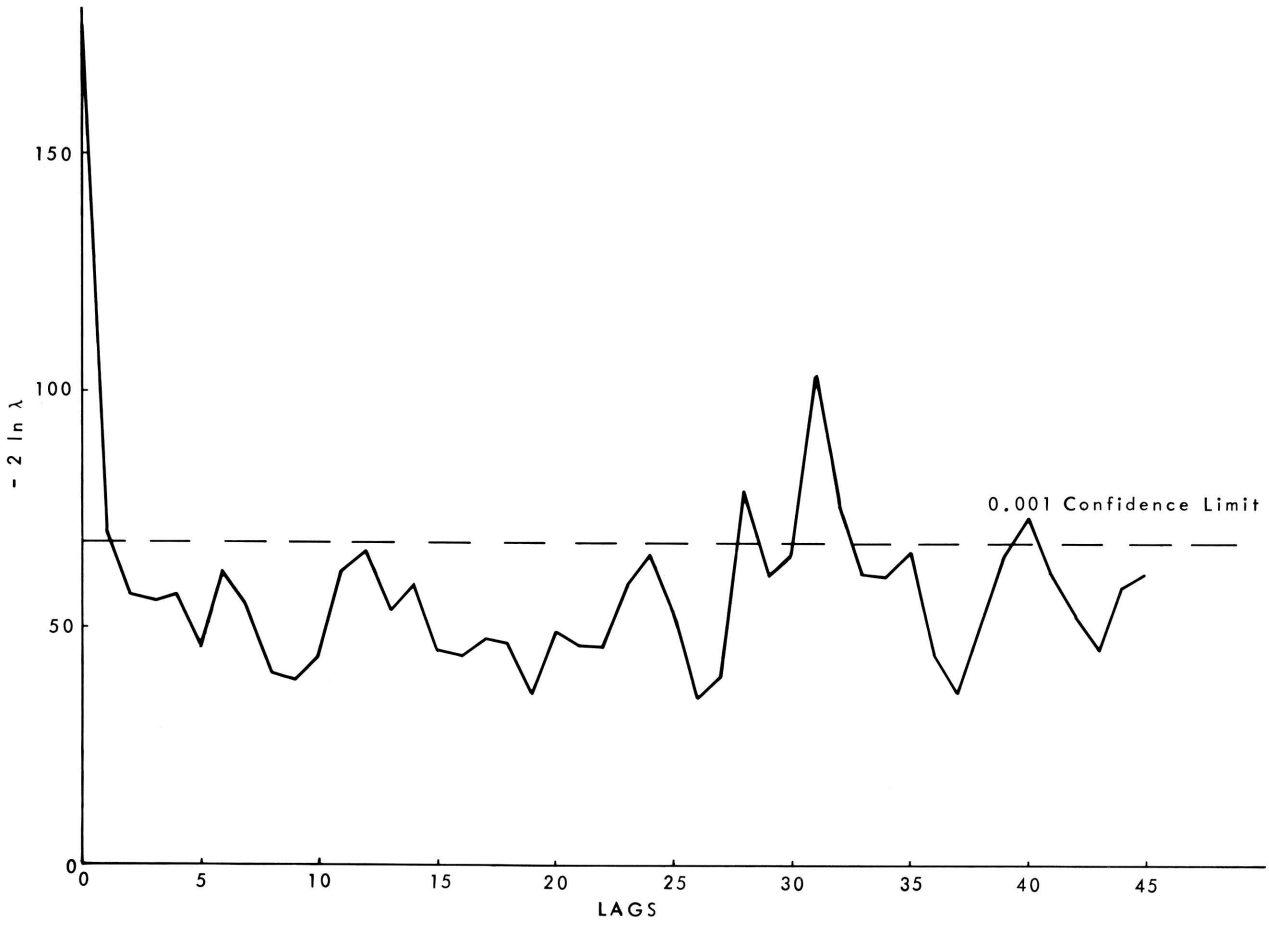


Figure 7.- $-2\ln\lambda$ values for lags 1 to 45.

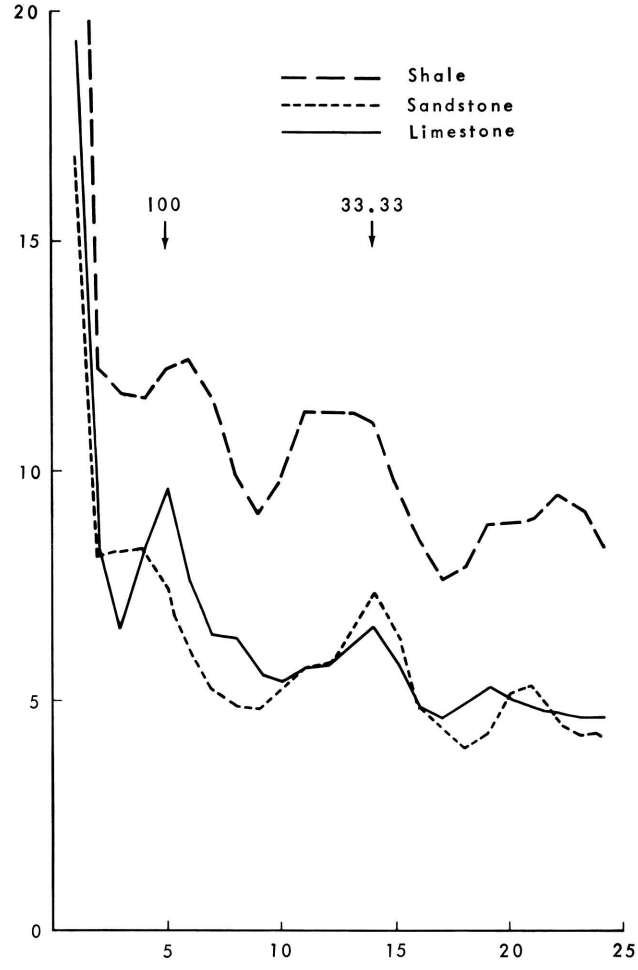
Table 2.- Probabilities from the state at time t to the state at time t + 1.

T - 31 = 1				T - 31 = 3			
0.7143	0.0952	0.1429	0.0476	0.2222	0.3333	0.3333	0.1112
0.0000	0.0000	0.0000	1.0000	0.10000	0.3000	0.50000	0.1000
0.0455	0.0000	0.5454	0.4091	0.1270	0.0000	0.7301	0.1429
0.0000	0.0000	1.0000	0.0000	0.0000	0.0285	0.2858	0.6957
T - 31 = 2				T - 31 = 4			
1.0000	0.0000	0.0000	0.0000	0.8823	0.0588	0.0000	0.0589
0.2500	0.2500	0.2500	0.2500	0.0000	0.5000	0.0000	0.5000
0.3750	0.1250	0.3750	0.1250	0.1316	0.0000	0.8158	0.0526
0.0000	0.0000	1.0000	0.0000	0.1875	0.0000	0.2500	0.5625

CONCLUSIONS

The geological interpretation of the experiments must be approached with caution. Assuming that the stratigraphic record is an unbiased estimate of the time history of sedimentation, then the hypothesis of a time-lag mechanism seems to provide an adequate model. It is not intended here to speculate about the nature of such a mechanism beyond indicating that it is most likely to involve some form of lateral transport that must have proceeded over a constant distance at a constant rate.

Assuming that the stratigraphic record contains systematically disturbed gaps or a systematic variation in sedimentation rates, our conclusions are more uncertain. It is believed that the experiments definitely indicate that the simpler random models such as a first-order Markov process or a second-order autoregressive process can be eliminated. This is more positive if we allow for randomly fluctuating sedimentation rates, as such fluctuations would blur the spectra rather than accentuate them. The time quasi-periodicity must have been more pronounced than is found in the stratigraphic record. Whether the model should be improved by making use of higher order terms or possibly by introducing a deterministic periodic component depends ultimately on the geological theories which are available to justify such procedures. In the meantime, simulation techniques can be used to study the effect of such improvements.



REFERENCES

Anderson, T. W., and Goodman, L. A., 1957, Statistical inference about Markov chains: Am. Math. Stat., v. 28, p. 89-110.

Bartlett, M. S., 1962, Stochastic processes: Cambridge University Press, Cambridge, 312 p.

Bendat, J. S., and Piersol, A. G., 1966, Measurement and analysis of random data, John Wiley and Sons, New York, 390 p.

Duff, P. M. D., and Walton, E. K., 1962, Statistical basis for cyclothems: A quantitative study of sedimentary successions in the east Pennine coalfield: Sedimentology, v. 1, p. 235-255.

Granger, C. W. J., and Hatanka, M., 1964, Spectral analysis of economic time series: Princeton University Press, Princeton, 299 p.

Kendall, M. G., 1945, On the analysis of oscillatory time-series: Jour. Roy. Stat. Soc., v. 108, p. 93-129.

Moore, R. C., 1936, Stratigraphic classification of the Pennsylvanian rocks of Kansas: Kansas Geol. Survey Bull. 22, 256 p.

Moore, R. C., and others, 1951, The Kansas rock column: Kansas Geol. Survey Bull. 89, 132 p.

Pearn, W. C., 1964, Finding the ideal cyclothem: Kansas Geol. Survey Bull. 169, p. 399-413.

Whittle, P., 1954, The statistical analysis of a sunspot series. Astrophys. Jour., v. 119, p. 251-260.

Yule, G. U., 1927, On a method of investigating periodicities in disturbed series, with special reference to Wolfer's sunspot numbers: Phil. Trans., Sec. A, v. 226, p. 267.

FREQUENCY ANALYSIS FOR SPARSE AND BADLY SAMPLED DATA IN THE EARTH SCIENCES

by

Norman S. Neidell[1/]

Gulf Research and Development Company

INTRODUCTION

Cyclicities, periodicities, and oscillatory transients of long-time duration are ubiquitous characteristics of many earth-science studies. Second-order differential equations govern motions of heavenly bodies, wave phenomena, and heat transport processes. These represent some of the most basic physical systems which admit solutions of the nature described, and their solutions become imprinted in the geological record to varying degrees. Hence, frequency analysis can assume a role of central importance in deciphering and understanding these aspects of geology.

The current literature abounds with examples where frequency analysis might be employed to advantage. Xanthakis (1967) is concerned about the rise time initiating sunspot cycles, whereas Donn and Shaw (1967) debate with C. Emiliani on paleotemperatures. A pulsating "constant" of gravitation is related to the major tectonic episodes by Machado (1967). All too often, however, the material for study barely qualifies as data. It can be sparse, limited in range, badly sampled, and highly inaccurate. Additional information may be extremely difficult to obtain.

Conventional methods of frequency analysis date from Fourier in 1822 (Lanczos, 1956) and were designed to give a maximum resolution in frequency approximately proportional to the inverse of the number of data samples multiplied with some mean sampling interval. The power-spectral methods currently in vogue can improve the reliability of calculated amplitudes but sacrifice frequency resolution by averaging over adjacent bandwidths (Blackman and Tukey, 1958; Jenkins, 1961; Parzen, 1961; Goodman, 1961). Alternative methods of frequency analysis exist, although they have not been commonly used in the earth sciences. Two of them, nonlinear estimation and differential modeling, will be mentioned in this paper. Appropriately used, both of these methods attain a degree of amplitude and frequency resolution which cannot be matched by the conventional analysis.

[1/] I would like to thank officers of the Gulf Research and Development Company for their permission to publish this paper.

NONLINEAR ESTIMATION

Frequency analysis by nonlinear estimation might equivalently be called "doing the obvious." Take the data and fit it in the least-squares sense with a trigonometric series of a fixed finite number of terms. Both the amplitudes and frequencies are considered as unknowns thus making for a nonlinear problem. There is, of course, no requirement that the trigonometric functions be harmonics so the frequency resolution is governed only by the accuracy of the data and not by the number of observations as in conventional analysis. Considerations of aliasing and such phenomena exist but do not cause undue difficulty for the experienced analyst.

Obviously, much technology has been developed to allow the doing of the obvious. In particular, the pioneering work of Kuhn and Tukey discussed by Hadley (1964) on the algorithmics and the advances in digital computing must be cited. For our studies, the algorithm of Shanno (1966) has been used.

DIFFERENTIAL MODELING

Before discussing frequency analysis via differential modeling, it would be helpful to note how modern digital computing has changed our perspective on what constitutes a solution. This is an observation made by the mathematician Richard Bellman. Suppose a differential equation is to be solved as an initial value problem and also the analytic function which satisfies it. A difference approximation can be made for the equation and using the FORTRAN language this leads to a single DO loop. The analytic function can be expanded as a series or its values sought in a book of tables. Now, if it is desired to know a single value of the function then we might seek it in a book of tables or compute it using the series. Suppose, however, that we wish to have available for further calculations a range of values of the function. These might be most efficiently and accurately calculated using the DO loop on a computer. In effect this is solving the differential equation. Is it fair to consider the function rather than the differential equation a solution?

Table 1 indicates how an advance in technology has interchanged the classical or 17th century view of problem and solution. Philosophically it is a basis for techniques of differential modeling.

Table 1.- Problem or solution.

Differential Equation:	Function:
$\frac{du}{dx} = -cu \quad u(0) = a$	$u = ae^{-cx}$
Difference Approximation:	**Series Expansion:**
$u_{i+1} = (1-c\Delta x)u_i$	$u(x) = ae^{-cx} =$
$u_0 = u(0) = a$	$a(1-cx+\frac{(cx)^2}{2!}-\frac{(cx)^3}{3!}$
$[\Delta x = \frac{b}{c},\ b < 1]$	$+ \ldots)$

"DO" Loop:

```
      UZERO = A
      DO 76 I = 1, N
      U(I) = (1.0 D0-B)*UZERO
76    UZERO = U(I)
```

Table of Values:

Fitting of data using differential equations owes much to Richard Bellman. The generality of these models in relation to the number of parameters to be estimated is unmatched while their applicability to problems in the physical universe is most dramatic. For instance, reflect upon how much more often theoretical studies lead to expressions of differential calculus rather than algebra. Further, differential models are intrinsically recursive or self-adaptive in nature and embody the essence of "learning" processes.

For frequency analysis, we need only fit the data with a differential system whose solutions are sine and cosine functions. Again, the analysis is not restricted to include only harmonics of some fundamental frequency. In terms of numerical stability these approaches to frequency analysis have been demonstrated to be superior. Many alternative methods for differential modeling exist especially in the literature of filter theory, control theory, and system identification. In our studies we use the quasilinearization method developed by Bellman and Kalaba (1965).

AN EXAMPLE - TWO WAYS

Consider the example of "periodogram analysis" given by Bellman and Kalaba (1965). Seven nonuniformly spaced data points were generated from the mathematical expression

$$U(t) = \sum_{i=1}^{3} \alpha_i \cos \omega_i t$$

using known values of α_i and ω_i. Next, the values of α_i and ω_i were treated as unknowns to be found by fitting the data using the differential system

$$u_i'' + \omega_i^2 u_i = 0 \quad u_i(0) = \alpha_i \quad u_i'(0) = 0 \quad i = 1, 2, 3.$$

In essence, we are asking to find six unknown values from seven data points. The procedure used first was quasilinearization and required initial guesses for the α_i and ω_i. Data and the results of four iterations are shown in Table 2.

Table 2.- Results of frequency analysis

Data:		Results:			
t	U(t)	Method	i	α_i	ω_i
0.00	1.600000	Initial	1	0.900000	1.00000
0.83	0.452413	Guesses	2	0.600000	2.00000
1.67	-0.679691		3	0.200000	3.00000
2.50	-0.820313	True	1	1.000000	1.11000
3.33	-0.367504	Values	2	0.500000	2.03000
4.17	-0.381874		3	0.100000	3.42000
5.00	0.351082	Quasi-	1	1.000000	1.11000
		linear-	2	0.500000	2.03000
		ization	3	0.100000	3.42000
		Nonlinear	1	0.999976	1.11012
		Estimation	2	0.499886	2.02995
			3	0.100073	3.41921

Table 2 also shows the results of fitting the same data using the identical initial guesses with a sum of three cosine functions of unknown parameters. This nonlinear programming problem was accomplished using the algorithm of Shanno (1966). These results are somewhat inferior because they were computed using less digits of accuracy than the quasilinearization solution. The frequencies in this example were severely truncated by the range of sampling, and there were few samples. Both quasilinearization and nonlinear estimation give results which cannot be matched by conventional analysis yet neither technique has been used much as indicated in earth-science literature.

CLOSING REMARKS

In brief, the results obtained for the synthetic example indicate that these alternative methods of frequency analysis merit further study and attention, especially when treating sparse and badly sampled data, a common situation in the earth sciences.

REFERENCES

Bellman, R. E., and Kalaba, R. E., 1965, Quasilinearization and nonlinear boundary value problems: Elsevier, New York, 206 p.

Blackman, R. B., and Tukey, J. W., 1958, The measurement of power spectra: Dover, New York, 190 p.

Donn, W. L., and Shaw, D. M., 1967, Isotopic paleotemperatures: discussion: Science, v. 157, no. 3789, p. 722-723.

Goodman, N. R., 1961, Some Comments on spectral analysis of time series: Technometrics, v. 3, no. 1, p. 221-228.

Hadley, G., 1964, Nonlinear and dynamic programming: Addison-Wesley, Reading, Massachusetts, 484 p.

Jenkins, G. M., 1961, General considerations in the analysis of spectra: Technometrics, v. 3, no. 1, p. 133-166.

Lanczos, C., 1956, Applied analysis: Prentice Hall, New Jersey, p. 207 and 305.

Machado, F., 1967, Geological evidence for a pulsating gravitation: Nature, v. 214, no. 5095, p. 1317-1318.

Parzen, E., 1961, Mathematical considerations in the estimation of spectra: Technometrics, v. 3, no. 1, p. 167-190.

Shanno, D. F., 1966, Nonlinear estimation programs REEP and CREEP: Gulf Research & Development Company, C. & E. Div., Mem. 133M6020, Re. 1335CN01, 88 p.

Xanthakis, J., 1967, Probable values of the time of rise of the forthcoming sunspot cycles: Nature, v. 215, no. 5105, p. 1046-1048.

SIMULATION MODELS OF TIME-TREND CURVES FOR PALEOECOLOGIC INTERPRETATION

by

William T. Fox

Williams College

INTRODUCTION

Time-trend curves can be used to plot the distribution of organisms or rock types as a function of time. A series of time-trend curves was computed and plotted to show the relationship between fossils and the enclosing strata (Fox and Brown, 1965). Individual curves have been plotted for brachiopod genera in the upper part of the Richmond Group (Upper Ordovician) of southeastern Indiana (Fox, in press). Based on these curves, it is possible to make a generalized interpretation of the paleoenvironment.

To explain the faunal distribution in terms of actual changes in the physical environment, a mathematical simulation model was constructed to reproduce the faunal distribution pattern shown by the time-trend curves. In the simulation model, organisms are assigned hypothetical tolerances and respond to changes in the temperature, salinity and depth. By adjusting the tolerances in the model, the distribution patterns are brought into line with the observed time-trend curves.

TEMPERATURE, SALINITY AND DEPTH

To test hypotheses involving changes in temperature, salinity and depth as a function of time, a method was devised for computing and plotting simulated time-trend curves. For each simulation model, it is possible to change the shape of the curve and the maximum and minimum values. To make the curves closely approximate the observed data, a random component can be superposed on the simulated curve using a pseudorandom number generator. The smooth curve represents the general trend of the environmental factor through time with the random components mimicking local fluctuations.

Several parameters are used to control the shape of the curves and their limits. For temperature, salinity and depth, each point on the simulated curve is considered the average value during a short interval of geologic time, represented by a small thickness of rock strata. In simulating the fossil distribution of an actual stratigraphic section, it is convenient to select the number of units so that the simulated time-trend curve will be plotted on the same scale as the time-trend curve for the observed section.

In computing and plotting the curves for each variable, the vertical section is divided into four equal segments. It is possible to treat the full section as a single unit, to split the section into two segments and treat the upper half and the lower half independently, or to treat the four quarters separately, making certain that the ends of successive segments join. The attitude of each segment of the curve and the maximum and minimum values are controlled by the function control card which is read in with the data. The nine different options available for each segment include: (1) remain constant at the minimum value, (2) remain constant at the mid-point, (3) remain constant at the maximum, (4) increase from the minimum to the maximum, (5) increase from the minimum to the mid-point, (6) increase from the mid-point to the maximum, (7) decrease from the maximum to the minimum, (8) decrease from the maximum to the mid-point, and (9) decrease from the mid-point to the minimum. By using different combinations of the above options with different segments of the curve, several different hypotheses about how the variables change with time can be tested. Each segment of the curve can be plotted as a linear function or as a portion of a cosine curve. In plotting a linear function which increases from the minimum to the maximum, the segment is plotted as a straight line with the minimum value at the lower end of the segment and the maximum value at the upper end of the segment.

Figure 1 shows examples of the linear and cosine functions plotted as several curves. Curve A is a linear function for the full section which increases from the minimum to the maximum. Curve B is a cosine function increasing from the minimum to the maximum. Curve C is a linear function in which the bottom segment increases from the minimum to the maximum, and the top segment decreases from the maximum to the minimum. In Curve D a cosine function is used which increases from the mid-point to the maximum, then decreases from the maximum to the minimum. Curve E represents a linear function in which the bottom quarter is constant, the second quarter increases from the minimum to the maximum, the third quarter decreases from the maximum to the minimum, and the top quarter remains constant at the minimum. Curve F is based on a cosine function with the lower quarter increasing from the minimum to the maximum, the second quarter decreasing to the mid-point, the third quarter increasing to the maximum and the top quarter decreasing to the minimum.

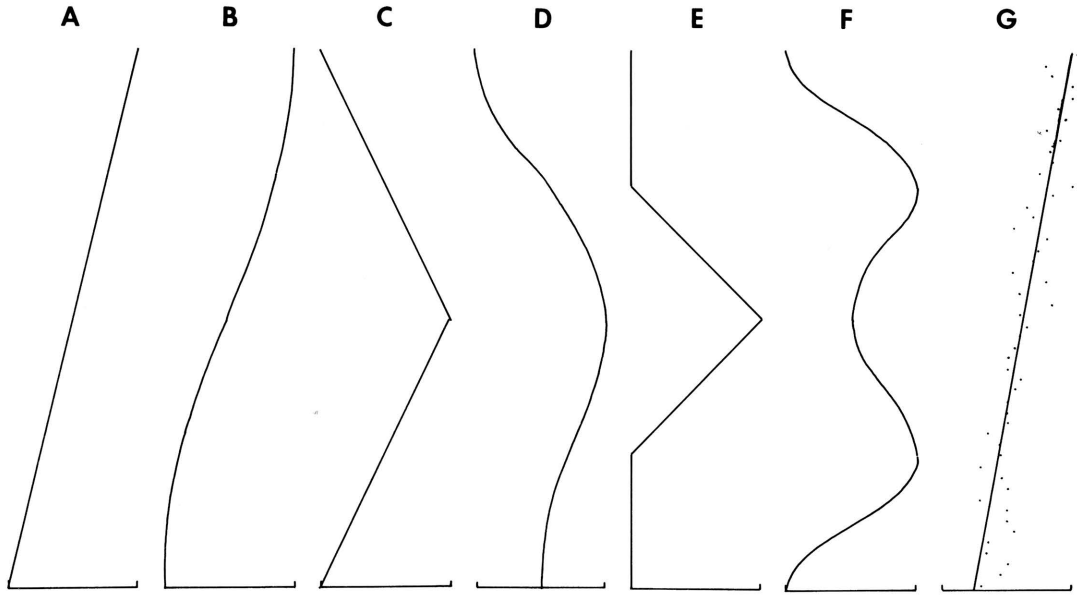


Figure 1.- Linear and cosine function curves used for plotting environmental parameters.

It is possible to superpose a random fluctuation on the curves, as an option to the simulation program. Using a pseudorandom number generator with a normal distribution, a standard deviation can be assigned to the environmental function. Curve G is a linear function which increases from a minimum to a maximum and has a standard deviation of 1.0. For curves A through F, the standard deviation is set at 0. It is possible to control the amount of fluctuation by increasing or decreasing the standard deviation.

LIGHT INTENSITY AND PHOTOSYNTHESIS

Light intensity is the most important limiting factor to photosynthesis within the oceans. Photosynthesis is limited to the euphotic zone, which extends to a maximum depth of 100 meters where the light intensity reaches 1 percent of the surface illumination. The depth of the euphotic zone is a function of surface radiation and the turbidity of the sea water. In the middle latitudes, the average surface illumination at noon on a clear day is about 10,000 foot-candles.

The turbidity is affected by both absorption by sea water, dissolved organic substances and colored particles, and scattering by the water molecules and particles in suspension.

Jerlov (1951) subdivided the oceanic waters into three categories and coastal waters into nine, based on the transmissibility of light. For each water type, Jerlov recorded the transmissibility at a depth of 1 meter for several wave lengths of light. The extinction coefficient can be computed directly as a function of light transmissibility. Light with a shorter wave length has a greater transmissibility in clear oceanic water, but in more turbid coastal water, the longer wave lengths have a greater transmissibility. Although the extinction coefficient increases with water type in moving from oceanic through coastal water, there is not a direct mathematical relationship between water types selected by Jerlov and their extinction coefficients. Equation (1) was derived empirically to relate the turbidity coefficient (Tb) used in the simulation program to the extinction coefficient (k)

$$k = 0.018 + 0.051\ Tb \qquad (1)$$

A plot of extinction coefficient versus turbidity coefficient is given in Figure 2 with the three oceanic water types and nine coastal water types from Jerlov (1951) included for reference. The oceanic types are numbered beneath the line with the coastal types above the line. In using the turbidity coefficient, 0 represents clear oceanic water with an extinction coefficient of 0.018, 2 is the boundary between oceanic and coastal water approximately at the outer margin of the continental shelf with an extinction coefficient of 0.120, and 10 is the most turbid coastal water with an extinction coefficient of 0.528.

The light intensity (L) at any depth is computed as a function of surface illumination (Ls), depth (h) and extinction coefficient (k) using equation (2).

$$L = Ls \cdot e^{-hk} \qquad (2)$$

Ryther (1956) made a study of the relationship between light intensity and photosynthesis of marine phytoplankton. In experiments made at the Woods Hole Oceanographic Institution, he exposed phytoplankton to a full range of light intensities and measured photosynthesis by the uptake of radioactive carbon, C^{14}. A series of additional experiments also was made by measuring photosynthesis at different depths in the Woods Hole Harbor with simultaneous measurements of incident radiation and the extinction coefficient of the water.

A generalized photosynthesis light-intensity curve was constructed by averaging the relative photosynthesis values at each intensity for three groups of marine phytoplankton. In order to use the generalized curve in the program for simulating time-trend curves, it was necessary to derive the mathematical equation for the curve. The curve plotted by Ryther (1956) has a marked asymmetry tailing off toward the higher values of light intensity which is best approximated by a gamma-density distribution. A computer program was written to compute relative photosynthesis as a function of light intensity using equation (3) for the gamma distribution (Krumbein and Graybill, 1965, p. 121).

$$G(X,r,\beta) = \frac{X^{r-1} e^{-X/\beta}}{\Gamma(r)\,\beta^{r}} \qquad (3)$$

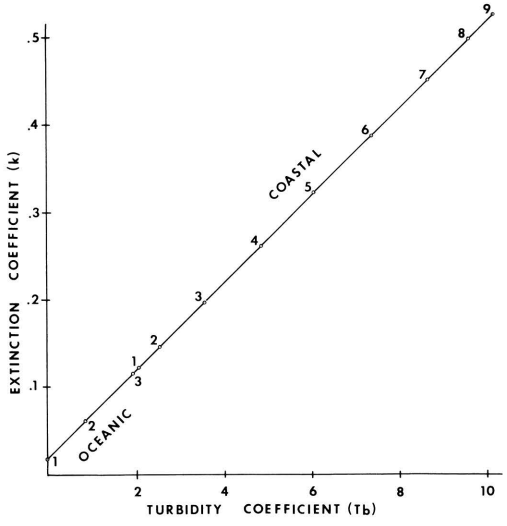


Figure 2.- Plot of extinction coefficient (k) versus turbidity coefficient (Tb) with coastal and oceanic water types.

The population parameters for the gamma-density distribution include a location parameter, r, and a scale factor, β, which can be found in terms of the mean X and the standard deviation S by using equations (4) and (5).

$$r = \left(\frac{\bar{X}}{S}\right)^{2} \qquad (4)$$

$$\beta = \frac{S^{2}}{\bar{X}} \qquad (5)$$

In using the gamma density distribution to compute relative photosynthesis, $\bar{X}$ is the light intensity measured in thousands of foot-candles. The value of light intensity for which the relative photosynthesis is at its peak is found by equation (6).

$$X \text{ peak} = \frac{S^{2}}{\bar{X}\left(\frac{\bar{X}}{S}\right)^{2} - 1} \qquad (6)$$

In computing the relative photosynthesis curve, the peak value is set equal to 1 and the remaining values are computed relative to the peak.

By computing and plotting the gamma distribution through several iterations changing the mean and standard deviation, a function was determined with a close fit to the observed curve from Ryther (1956). The gamma distribution with a mean of 4.075 and a standard deviation of 2.574 (Fig. 3) gives a good approximation of relative photosynthesis as a function of light intensity in thousands of foot-candles.

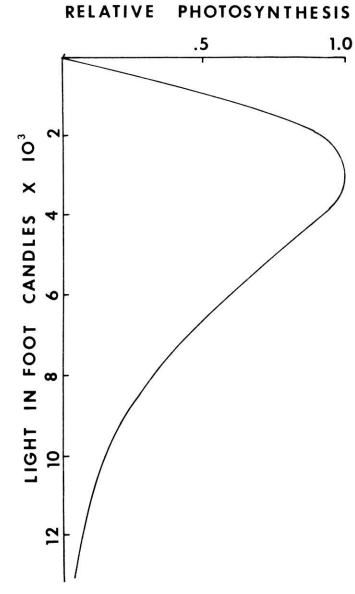


Figure 3.- Relative photosynthesis as a function of light intensity in thousands of foot-candles.

The program for computing light intensity was combined with the program for relative photosynthesis to produce a program which computes relative photosynthesis as a function of surface illumination, depth and turbidity. The program also computes the depth for the maximum phytoplankton production and depth for the base of the euphotic zone where light inten-

sity equals 1 percent of the surface-light intensity. Curves for light intensity and phytoplankton distribution versus depth with different turbidity values are given in Figure 4.

ORGANISM TOLERANCES

Organisms are restricted in their geographic and temporal range by tolerances to the physical and chemical factors of their environment. Although the environment consists of many factors involved in a complex interaction, only water temperature, salinity and light intensity are considered in time-trend simulation models. For every factor, each organism is assigned a tolerance mean which represents the optimum and a standard deviation which controls the range.

For many living species of plants and animals, the temperature range can be firmly established by climatic observations. Along with the lethal temperature limits, feeding, reproduction and general activity are modified by temperatures less extreme than those which actually cause death. An increase in temperature accelerates chemical reactions which regulate the body processes within organisms (Moore, 1958). Two types of reactions, one set building up and the other set breaking down, are taking place within organisms. A combination of the two sets of reactions gives a reaction curve which rises to a peak and then drops off with an increase in temperature. The peak in the curve is considered the optimum temperature for a particular organism and tails of the curve mark lethal limits.

In time-trend simulation, the normal curve is used as a model of the temperature reaction curve with each species assigned a temperature tolerance mean and standard deviation. With an increase or decrease in temperature away from the mean, there is a decrease in abundance (Fig. 5). Organisms with a wide temperature range are called eurythermal, and those with a narrow range are termed stenothermal. Figure 5 is an example of a simulation run with a constant increase in temperature, to demonstrate the difference between eurythermal and stenothermal organisms with different optimum temperatures. Organisms A through E are stenothermal with a temperature tolerance standard deviation of 1.0° C. The remaining organisms F through J are eurythermal with a temperature tolerance standard deviation of 2.0° C.

The distribution of organisms in the marine environment also is dependent on salinity. Those organisms with a narrow salinity tolerance are considered stenohaline and those with a wide salinity tolerance are called euryhaline. Normally those organisms which live in the open ocean with a

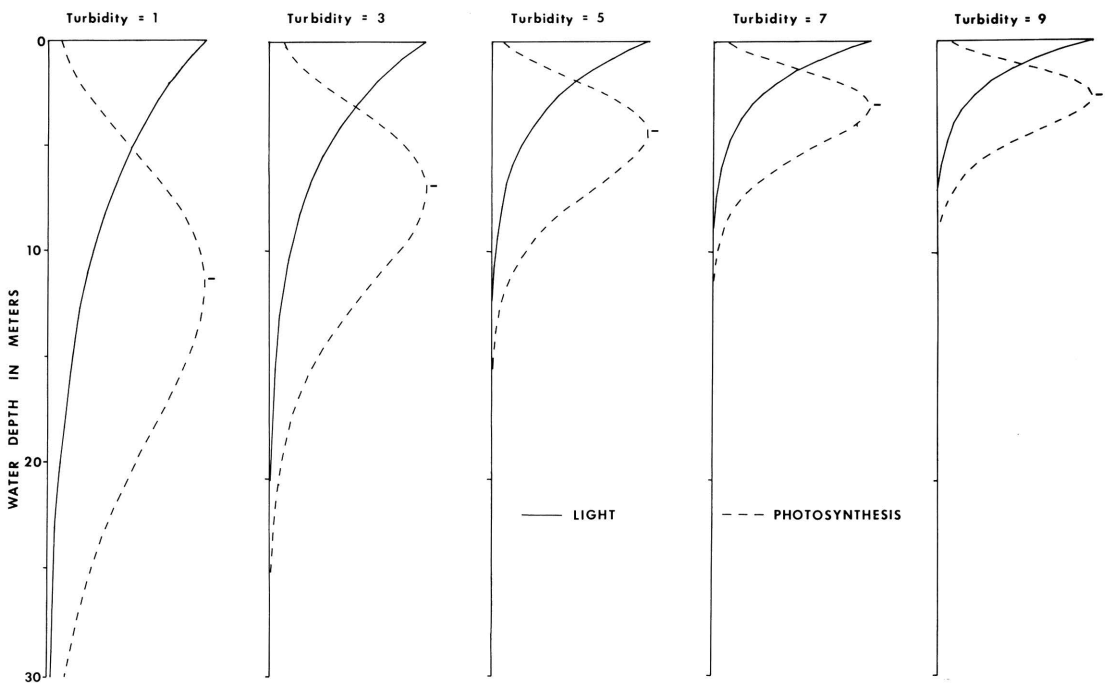


Figure 4.- Light intensity and phytoplankton distribution versus depth with different turbidity values.

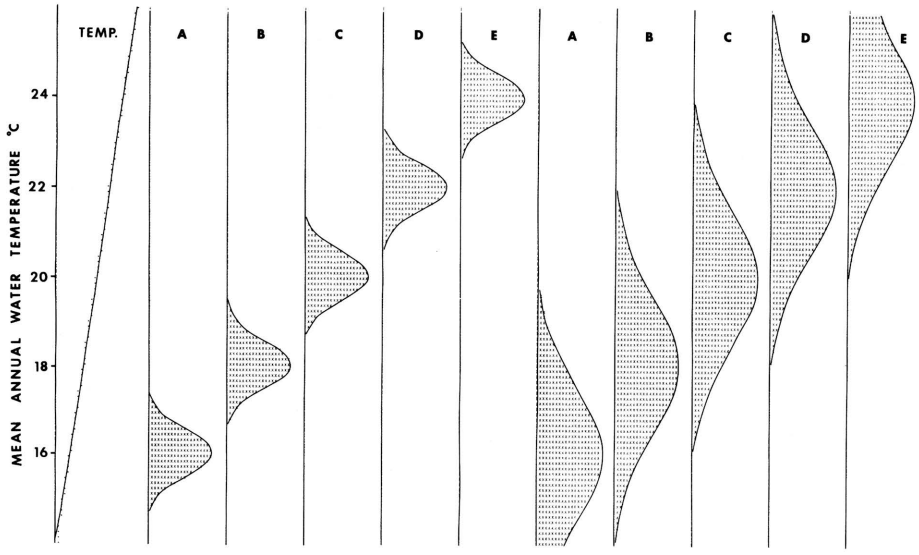


Figure 5.- Organism abundance versus temperature for organisms with different optimum temperatures and temperature ranges.

salinity near 35°/oo have a narrow salinity tolerance, whereas those in the intertidal zone, tidal flats or estuaries usually have a broader salinity tolerance. As with temperature, the organisms in the simulation models are assigned salinity tolerance means and standard deviations.

The benthonic species also have different tolerances to light intensity, which provides a vertical zonation with water depth. Based on the curve for light intensity versus photosynthesis of phytoplankton plotted in Figure 3, the gamma distribution is used as a model for light intensity versus population density in the marine environment. In Figure 6, three species (A, B and C) with different light intensity means and the same standard deviation are used to test the relationship between light and depth for different conditions. Temperature was held constant at 20°C, salinity at 35°/oo, and surface illumination at 10,000 foot-candles. In moving through time, depth was decreased linearly from 25 meters to sea level.

A turbidity coefficient of 3.0 was used for the first test, giving an extinction coefficient of 0.153 (Fig. 6). Species A is assigned a light intensity tolerance mean of 4.1 which is the value derived for marine phytoplankton. For species B, the light intensity mean has been increased to 4.6 and for species C, the mean is 5.1. The standard deviation derived for the phytoplankton, 2.6, is used for the three species. With an increase in light intensity tolerance, the peak abundance of species A to C occurs at a higher level and the range is more restricted. If the turbidity coefficient is increased to 5.0, the extinction coefficient goes up to 0.273. As a result of the decrease in light penetration, species A, B and C are found in higher levels of the sea. For the three species, the increased turbidity results in a higher peak and a decrease in vertical range. If the turbidity is raised to 7.0, the extinction coefficient is raised to 0.375 and less light reaches the bottom. With the increase in turbidity, the organisms are restricted to a depth of 10 meters or less.

FREQUENCY, ABUNDANCE AND AGGREGATION

The terms frequency, abundance and aggregation are used to describe the distribution of a group of organisms under optimum environmental conditions. For living communities, Fager (1963) defined indices to measure frequency, abundance and aggregation.

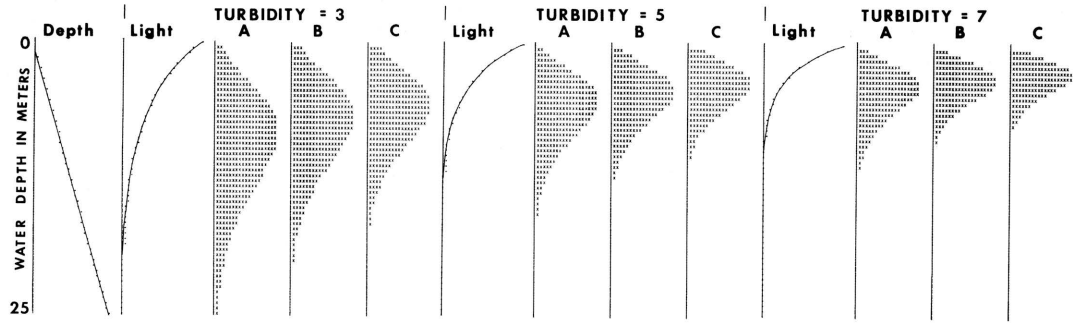


Figure 6.- Depth, light intensity and organism abundance for different turbidity values.

For the time-trend simulation models, each taxonomic group is given a frequency index, an abundance mean and standard deviation, and an aggregation index. These indices are used with the environmental tolerances to determine the population density in each increment of time or rock thickness.

According to Fager (1963), frequency is measured on the spread of a species throughout a community. For the simulation program, frequency is defined as the ratio of the number of horizons at which the species is present to the total number of horizons. If a species is found in 20 limestone layers in a sequence of 100, it has a frequency index of 0.2. A pseudorandom number generator is used in the simulation program to determine if a species is present or absent in a particular layer. The pseudorandom numbers have a rectangular distribution ranging between 0 and 1.0. If a species has a frequency index of 0.4, the species will be counted as present if the random number is less than or equal to 0.4, and will be considered absent if the random number is greater than 0.4. Because a new random number is selected for each stratigraphic interval, a species with a frequency index of 0.4 will be present in 40 percent of the layers and absent in 60 percent. As an example of how the frequency index is used, three species which are distributed normally with a linear increase in temperature are plotted in Figure 7. The frequency indices (0.2, 0.5 and 0.8) are given at the top of each column.

The mean abundance is defined as the mean number of individuals of a species per unit volume or surface area in the samples in which the species is present. Because mean abundance is computed using only those samples where the species is present, mean abundance and frequency are independent. Species with different abundance means but the same frequency are plotted in the second part of Figure 7. Natural fluctuations in population density are taken into account by the abundance standard deviation. For computing population density, a random number is selected from a normally distributed population with a mean of 0 and a standard deviation of 1. The effect of changing abundance standard deviation can be seen from the third part of Figure 7. Species with a stable population have a small abundance standard deviation, and those with a highly fluctuating population have a large standard deviation.

The aggregation index is used as a measure of the clustering or aggregation of a species within a stratigraphic sequence. The aggregation index ranges from 0 to 1.0 with 0 for negative clustering or dispersion and 1.0 for maximum clustering. The aggregation index acts as a weighting factor in a Markov process to adjust the frequency index. If the aggregation index is greater than 0.5, the frequency index is increased if the species was present in the previous layer, and decreased if the species was absent. If the aggregation index equals 0.5, the presence or absence of a species in a layer has no effect on its frequency in the following layer. The effect which the aggregation index has on a population is shown in the final group in Figure 7. The species which have a short larval stage and form colonies would have a high aggregation index, whereas species with a long larval stage which are dispersed widely by the currents would have an aggregation index close to 0.5.

SIMULATION MODEL 1

To test the simulation program with actual data, a model was made of the distribution of ten brachiopod genera in the upper part of the Richmond Group in southeastern Indiana (Fox, 1962). In Figure 8, time-trend curves of brachiopod distribution are plotted using a modified version of the FORTRAN program by Fox (1964). The data used in the time-trend curves are plotted also as bar graphs in Figure 9 for direct comparison with the simulated time-trend curves in Figures 10 through 13.

The simulation models are an attempt to reproduce the time-trend curves using environmental factors and organism tolerances to control the curves. The optimum temperature is the temperature tolerance

mean, and the temperature range is the standard deviation.

For Simulation Model 1 which is plotted in Figure 10, temperature increases as a linear function from 16 to 24°C. Water depth is held constant at 5 meters and the turbidity coefficient is set at 5. According to Jerlov (1951), this would correspond to the middle of the continental shelf. The surface illumination is set at 10,000 foot-candles and salinity is held constant at 35°/oo which is close to the average salinity for the ocean.

The distribution of Sowerbyella cannot be satisfactorily explained by a linear increase in temperature. Sowerbyella is found near the base of the section, is missing from the middle, and is fairly abundant at the top (Fig. 9). In order to account for its presence both near the base and at the top of the section, Sowerbyella is given an intermediate optimum temperature (21.0) and a high range (5.0). Because the plot for Sowerbyella in Figure 10 does not match its distribution in Figures 8 and 9, Sowerbyella is considered relatively

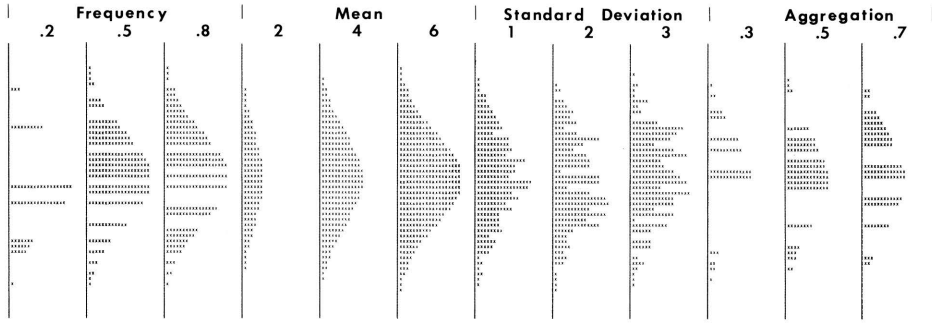


Figure 7.- Faunal abundance versus temperature with different values for frequency index, mean abundance, abundance standard deviation and aggregation index.

Because there is a linear increase in temperature with time, the organisms with a low optimum temperature mean will occur near the base of the section and those with a high optimum temperature will be concentrated near the top.

The results of the first simulation model using temperature as the only variable (Fig. 10) can be compared with the observed distribution in Figure 9. Resserella is limited to the lower part of the section and has been assigned a low optimum temperature (17.0) and a narrow range (1.0). Platystrophia is most abundant in the lower part of the section but has a wide range, so it is given an optimum temperature of 18.0 and a range of 2.5. Leptaena occurs in the middle of the section and is assigned an optimum of 20.0 and a range of 1.5. Rafinesquina and Zygospira are fairly evenly spread over the entire section, so they are given an intermediate optimum temperature (20.0) and a high range (5.0). Strophomena and Hebertella extend from near the base to the top of the section with their greatest abundance in the upper portion, therefore, they are given a high optimum (23.0) and a medium range (3.0). Rhynchotrema and Plaesiomys are found only in the upper few meters, so they are given a high temperature optimum (24.0). Rhynchotrema has a wider range and is given a larger deviation (1.5) than Plaesiomys (1.0).

independent of temperature and its distribution must be accounted for by other environmental factors.

SIMULATION MODEL 2

Depth is changed and the other environmental factors are held constant for Simulation Model 2 (Fig. 11). A cosine function is used to compute the depth values with a minimum of 2.0 meters and a maximum of 16 meters. Between the base and middle of the section, the depth is increased from the midpoint value (9 meters) to the maximum (16 meters). Between the middle and top of the section, depth is decreased from the maximum to the minimum (2). The original assumption for an increase in depth followed by a decrease is based on point counts of distribution of micrite and sparite in the Richmond Group. In the lower part of the section, there is a high sparite-to-micrite ratio which would indicate relatively shallow water. The middle part of the section has a lower sparite-to-micrite ratio which has been interpreted as the result of an increase in depth. The top part of the section has the highest sparite-to-micrite ratio and is overlain by a primary dolomite, the Saluda Formation, which contains mudcracks and other evidence of subaerial exposure.

The turbidity coefficient for Model 2 is set

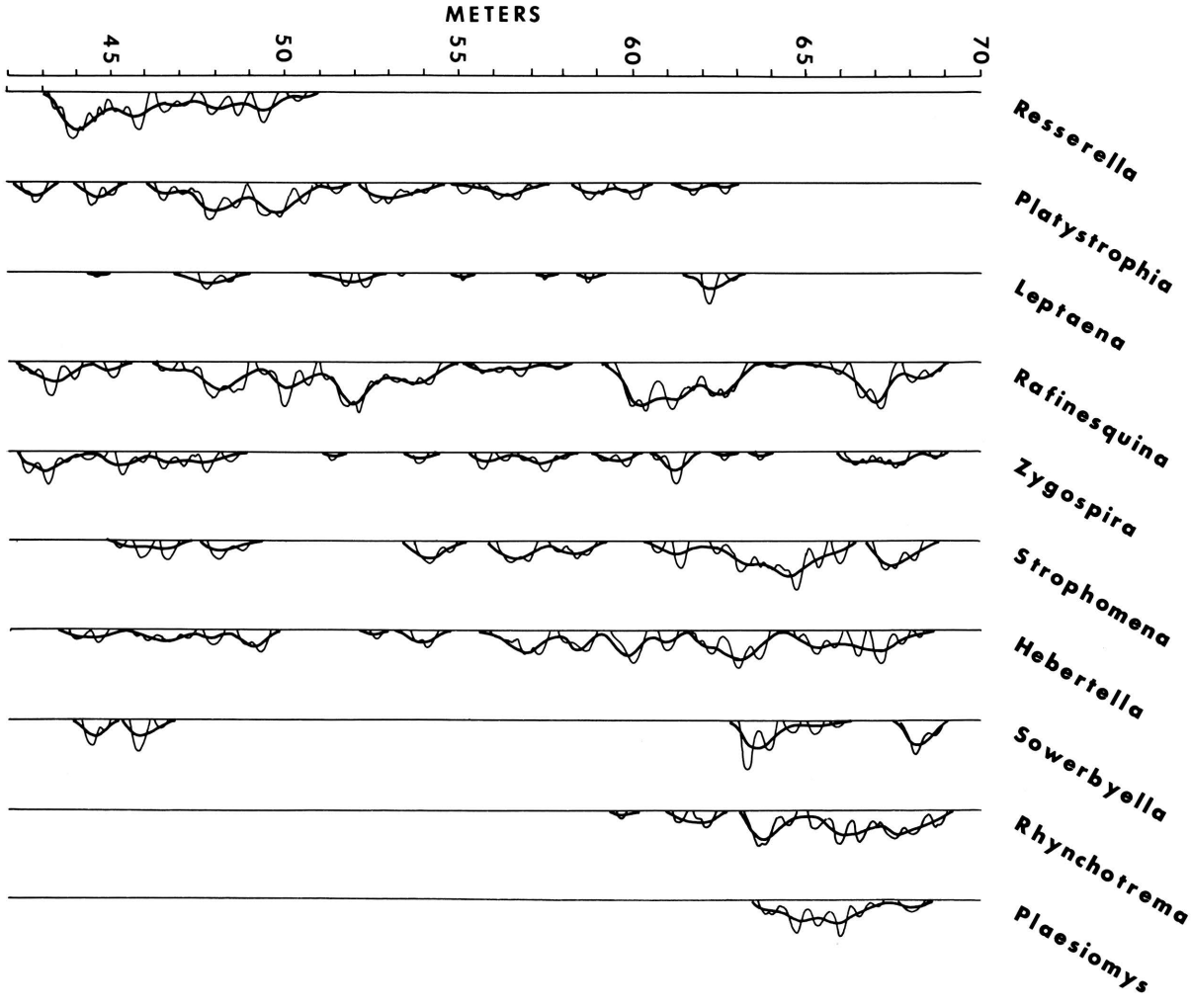


Figure 8.- Time-trend curves for brachiopod distribution in upper part of Richmond Group.

at 3.0, giving an extinction coefficient of 0.171. This corresponds to Jerlov's (1951) coastal type 2, found in clear water on the continental shelf. Because this is a carbonate shelf environment with a low influx of terrigenous clastics, a low turbidity value was selected for this model. The bottom illumination, which is a function of depth, turbidity and surface illumination, changes inversely with depth. The bottom illumination plotted in column 3 of Figure 11 decreases from 2.15×10^3 foot-candles at the base of the section to 0.65×10^3 at the midpoint. At the top of the section where the depth reaches 2.0 meters, the bottom illumination reaches a maximum of 7.10×10^3 foot-candles.

The abundance curves based on the last of a series of computer runs for Model 2 are plotted in Figure 11. The temperature and salinity tolerances

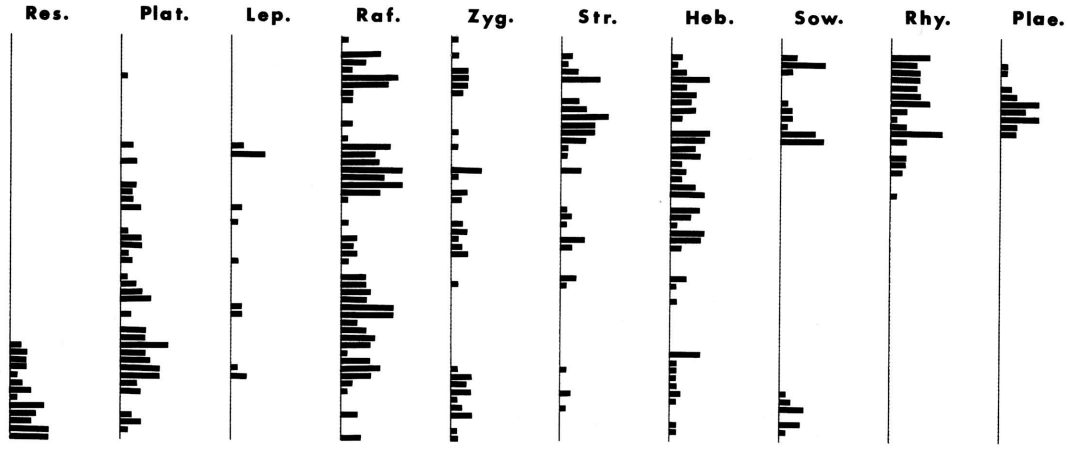


Figure 9.- Bar graph for brachiopod distribution in Richmond Group. Each bar represents average abundance for a 50-centimeter interval.

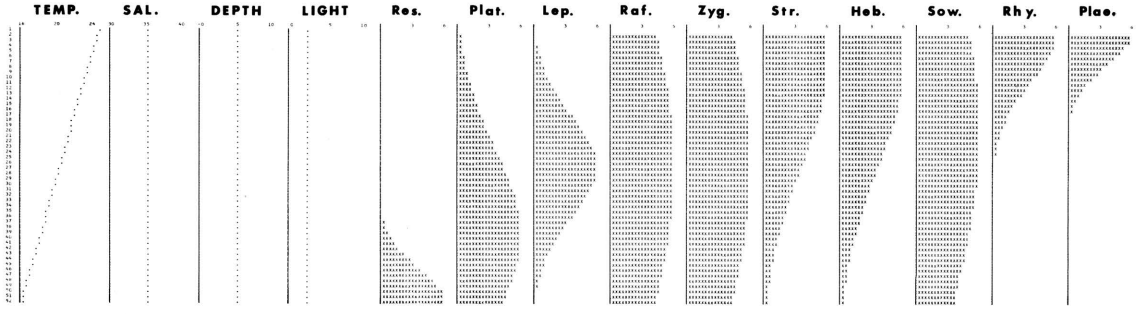


Figure 10.- Simulation Model 1 with temperature as only variable.

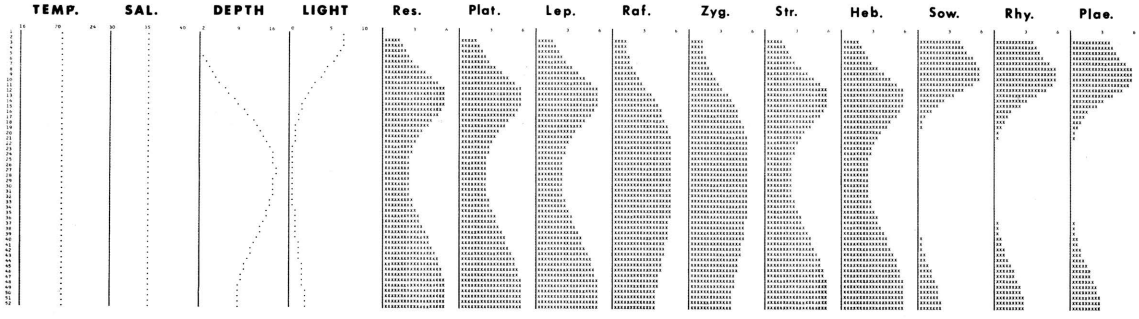


Figure 11.- Simulation Model 2 with depth as variable and light intensity as a function of depth.

are set at their optimum values, 20° C and 35°/oo respectively, and the light tolerances are different for the different genera. Although light tolerance seemingly is not an important factor for a majority of the organisms, it seems to be the critical factor controlling the distribution of Sowerbyella in Model 2. For the distribution of brachiopods which can be explained by temperature tolerances in Model 1, the light tolerances observed for phytoplankton are used. Sowerbyella, Rhynchotrema and Plaesiomys are given a higher light optimum so that their distribution is split into 2 modes. In their observed occurrence in Figure 9, Rhynchotrema and Plaesiomys are limited to the upper mode, but Sowerbyella is found in both the upper and lower part.

SIMULATION MODEL 3

The depth function from Model 2 and the temperature function from Model 1 are united in Model 3. The tolerance means and standard deviations are based on the temperature tolerances for Model 1 and the light tolerances for Model 2. The population density curves in Figure 12 portray the effect of changing both temperature and depth through time. At this point in the simulation process, it is possible only to line up the maxima and minima. Because the frequency and abundance means are set at their maximum values, the simulated abundances in Model 3 will be higher than the observed abundances. With the abundance standard deviation set at 0.0, the simulated model lacks the fluctuations in abundance which are seen in the observed data. The abundances in Model 3 form an envelope which would fit the time-trend curves in Figures 8 and 9 as closely as possible.

SIMULATION MODEL 4

Simulation Model 4 in Figure 13 is based on the temperature and depth functions and the environmental tolerances from Model 3. To make the model more realistic by adding a random component, hypothetical frequency, abundance and aggregation indices are established for each genus. With these final adjustments, the model is brought in tune with the time-trend curves in Figure 8. The tolerances are listed in Table 1.

In Models 1 through 3, the frequency was held at 1.0 to study the response of the organisms to changes in the environmental parameters. The frequency index is independent of temperature or depth and hypothetically is a measure of dispersal rate. A high frequency would correspond to a high dispersal rate and a low frequency to a low rate of dispersal. The frequency values in Table 1 range from 0.95 for Rhynchotrema to 0.40 for Leptaena and Zygospira. This would indicate that Rhynchotrema occurs in almost all horizons with the proper environmental conditions. Leptaena and Zygospira, on the other hand, have a spotty occurrence although the environmental conditions are suitable for their existence. The frequency index controls the presence or absence of a genus at a particular horizon, whereas the abundance mean controls its relative abundance. The abundance means and standard deviations are used to fit the abundance curves in Figure 13 more closely to the trend curves in Figure 8. In the first three models, the abundance mean was set at the maximum, 6.0, and the standard deviation was held at 0.0. The abundance mean ranges from a high of 3.0 for Strophomena, Rhynchotrema and Plaesiomys to a low of 1.5 for Zygospira. The adjustments in abundance mean are based on the curves plotted in Figure 12 for Model 3.

The abundance standard deviation is used to control the amount of fluctuation about the abundance mean. In Models 1 through 3, the standard deviation was set at 0.0 to eliminate random fluctuations in the curves. For Model 4, the random fluctuations are introduced into the model to simulate the fluctuations in the observed data. Rhynchotrema is given the highest standard deviation (3.0) while Zygospira, Strophomena and Hebertella are given low deviations (1.0). Most of the organisms are given in standard deviation of 1.5.

The aggregation index is used to control the clustering or aggregation within a genus. An

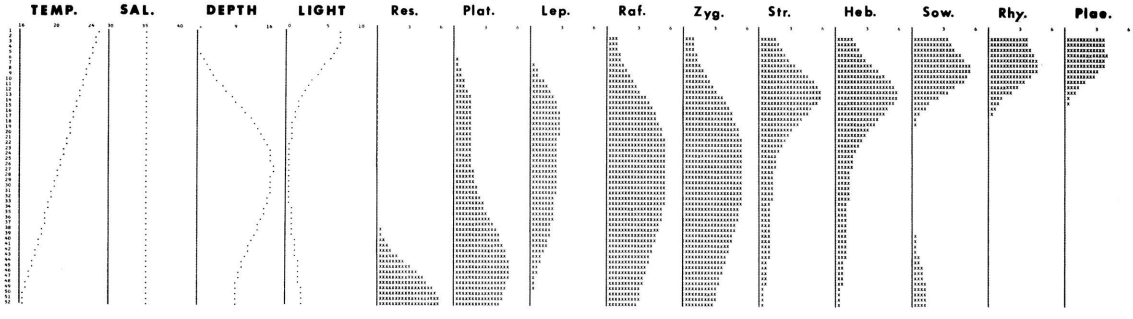


Figure 12.- Simulation Model 3 with both temperature and depth as variables.

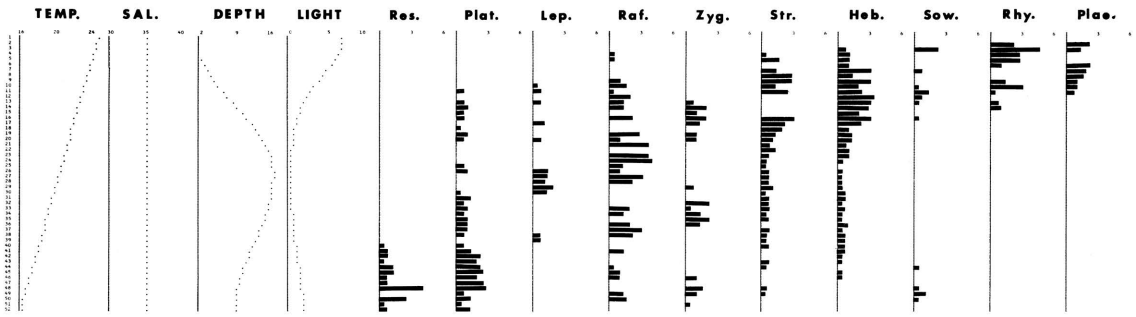


Figure 13.- Simulation Model 4 with temperature and depth as variables, and frequency, abundance mean and standard deviation and aggregation index used for organisms.

aggregation index of 0.5 would mean that the frequency has a purely random distribution. With an aggregation index greater than 0.5, the frequency index is raised if the organism was present in the preceding layer and lowered if it was absent. With a high aggregation index, the organism has a higher probability of occurring if it was present in the preceding layer. For Models 1 through 3, the aggregation index was set at the neutral position, 0.5. For Model 4, Leptaena and Zygospira are given the highest aggregation values (0.7) and their clustered occurrence, especially for Zygospira, can be seen from Figure 12. Rhynchotrema and Resserella, which have the highest frequencies, are given a normal aggregation index (0.5). The remaining genera are given an aggregation index of 0.6 which would cause them to have a slight tendency toward clustering.

CONCLUSIONS

By plotting temperature, salinity and depth as a function of time and assigning tolerances to the organisms, it is possible to reproduce the time-trend curves based on observed data. The simulated time-trend curves provide one possible explanation for brachiopod distribution in the Richmond Group. It would be possible also to reproduce the faunal distribution with other combinations of environmental factors and tolerances. Thus, the time-trend simulation model provides a means of testing whether a hypothesis involving environmental changes is possible. For example, it was not possible to account for the entire brachiopod distribution in the Richmond Group by either temperature or depth in Models 1 and 2. It was necessary to change both temperature and depth in Models 3 and 4 to reproduce the original pattern.

Table 1.- Tolerance values used for simulation Model 4.

	OPTIMUM ABUNDANCE FACTORS				TEMPERATURE		SALINITY		LIGHT	
	FREQ.	MEAN	DEV.	AGGR.	MEAN	DEV.	MEAN	DEV.	MEAN	DEV.
RESSERELLA	0.99	2.00	1.50	0.50	16.0	1.0	35.0	1.0	4.1	2.6
PLATYSTROPHIA	0.80	2.00	1.50	0.60	18.0	2.5	35.0	1.0	4.1	2.6
LEPTAENA	0.40	2.50	1.50	0.70	20.0	1.5	35.0	1.0	4.1	2.6
RAFINESQUINA	0.70	2.50	1.50	0.60	20.0	5.0	35.0	1.0	4.1	4.0
ZYGOSPIRA	0.40	1.50	1.00	0.70	20.0	5.0	35.0	1.0	4.1	4.0
STROPHOMENA	0.65	3.00	1.00	0.60	23.0	3.0	35.0	1.0	4.1	2.6
HEBERTELLA	0.80	2.50	1.00	0.60	23.0	3.0	35.0	1.0	4.1	2.6
SOWERBYELLA	0.70	2.00	1.50	0.50	21.0	5.0	35.0	1.0	5.9	2.6
RHYNCHOTREMA	0.95	3.00	2.50	0.50	24.0	1.5	35.0	1.0	5.6	2.6
PLAESIOMYS	0.80	3.00	1.00	0.60	24.0	1.0	35.0	1.0	5.6	2.6

REFERENCES

Fager, E. W., 1963, Communities of organisms: The sea, v. 2, p. 415-437.

Fox, W. T., 1962, Stratigraphy and paleoecology of the Richmond Group in southeastern Indiana: Geol. Soc. America Bull., v. 73, no. 5, p. 621-642.

Fox, W. T., 1964, FORTRAN and FAP program for calculating and plotting time-trend curves using an IBM 7090 or 7094/1401 computer system: Kansas Geol. Survey, Sp. Dist. Pub. 12, p. 1-24.

Fox, W. T., in press, The use of the computer for quantitative analysis of fossil distribution: Vistelius Volume, Leningrad.

Fox, W. T., and Brown, J. A., 1965, The use of time-trend analysis for environmental interpretation of limestone: Jour. Geology, v. 73, no. 3, p. 510-518.

Jerlov, N. G., 1951, Optical studies of ocean waters, in Reports of the Swedish Deep Sea Expedition: v. 3, no. 1.

Krumbein, W. C., and Graybill, F. A., 1965, An introduction to statistical models in geology: McGraw-Hill, New York, 475 p.

Moore, H. B., 1958, Marine ecology: John Wiley and Sons, London, 493 p.

Ryther, J. H., 1956, Photosynthesis in the ocean as a function of light intensity: Limnology and Oceanography, v. 1, p. 61-70.

PREDICTION OF MULTIPLE TIME SERIES GENERATED BY STATIONARY RANDOM PROCESS

by

Gangu G. Hingorani and Louis F. Marczynski
Northern Natural Gas Company

INTRODUCTION

A time series can be any collection of data where each point is associated with a moment in time, i.e. a time series can be defined as a set of ordered pairs (t_i, x_i) for $i = 0, 1, 2, \ldots, n$.

The time series generated by a random phenomenon has the property that each observation is unique. A given observation will represent only one of many possible values that might be generated at that particular point in time. Because of this property, a random process cannot be described by an explicit mathematical relationship, but must be looked at in terms of its statistical properties.

The random processes can be classified as stationary and nonstationary processes. The underlying mechanism that generates a random process can be described in physical or mathematical terms.

The underlying mechanism that generates ocean waves is essentially wind force in conjunction with the earth's gravity. The outcome of dice throwing is determined by probability of 1/6 for each face of each die independent of all others.

If the generating mechanism does not change with time, any measured average property of the random process is independent of the time of measurement aside from some statistical fluctuations, and the random process is called stationary. For instance, ocean wave height in a given sea state, a telegraphic signal of a certain language are examples of stationary random processes. If the generating mechanism does change, the random process is called nonstationary, so that the generating mechanism may change in a predetermined fashion or at random.

An example of a stationary random process is shown in Figure 1. The time series represents the annual mean air temperature in London from 1763 to 1900. An example of a nonstationary random process is shown in Figure 2. The time series represents the average daily temperature in Minneapolis from January 1, 1966 to December 31, 1966.

The time series representing a stationary random process can be generated by the expression

$$x(t) = \sum_{k=1}^{N} (A_k \sin w_k t + B_k \cos w_k t)$$

where:

$$w_k = \frac{\pi}{2k}$$

A_k and B_k are uncorrelated random variables with a zero mean and a standard deviation of σ.

MATHEMATICAL DESCRIPTION

Definition of the Problem

Consider the set of time series $x_i(t)$ for $i = 1, 2, \ldots, m$ and $t = 0, 1, 2, \ldots, N$ which are generated by stationary random process.

$$x_1(t) : x_1(0), x_1(1), x_1(2), \ldots, x_1(N)$$

$$x_2(t) : x_2(0), x_2(1), x_2(2), \ldots, x_2(N)$$

$$\vdots$$

$$x_m(t) : x_m(0), x_m(1), x_m(2), \ldots, x_m(N)$$

The problem is to predict $x_1(N+\alpha), x_2(N+\alpha), \ldots, x_m(N+\alpha)$ for $\alpha = 1, 2, \ldots$ where α is the prediction distance.

The Correlation Function

The correlation function between x(t) and y(t) is defined as

$$\phi_{xy}(\tau) = \lim_{T \to \infty} \frac{1}{T} \int_0^T x(t)\, y(t-\tau)\, dt.$$

The correlation function between two input time series $x_1(t)$ and $x_2(t)$ (or the same time series) is computed exactly the same way as it is defined, except that the sample time series has finite length T_1

$$\phi_{12}(\tau) = \overline{x_1(t)\, x_2(t-\tau)} = \frac{1}{T_1 - \tau} \int_0^{T_1-\tau} x_1(t)\, x_2(t-\tau)\, dt$$

where the bar refers to the lagged mean crossproduct.

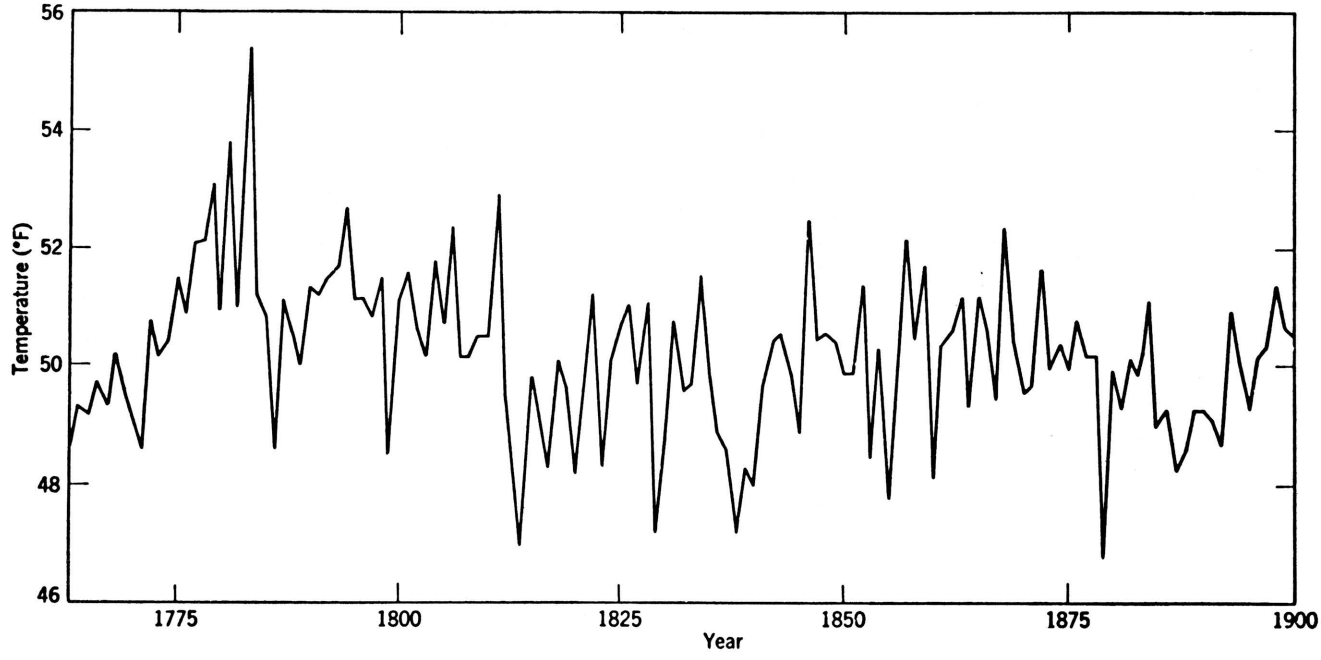


Figure 1.- Annual mean air temperature in London from 1763-1900.

The equations indicate three essential steps

(i) Time shift: Either advancing x_1 or delaying x_2 by an interval τ

(ii) Multiplication

(iii) Averaging

Because only a finite number of data points can enter into the computations, we select an interval Δt, $\Delta t = \frac{T_1}{N}$ and take readings of $x_1(t)$ and $x_2(t)$ at $t = 0, \Delta t, 2\Delta t, \ldots, N\Delta t$. Now we have two time series $x_1(t)$ and $x_2(t)$ for $t = 0, \Delta t, \ldots, N\Delta t$. The correlation function can be written as

$$\phi_{12}(\tau) = \frac{1}{N-\tau+1} \sum_{t=0}^{N-\tau} x_1(t)\, x_2(t-\tau)$$

Prediction Equation

The basic equation for predicting multiple time series is defined as

$$x(t+\alpha) = a_0\, x(t) + a_1\, x(t-1) + a_2\, x(t-2) + \ldots + a_\tau\, x(t-\tau) \qquad (1)$$

where

$$x(t) = \begin{bmatrix} x_1(t) \\ x_2(t) \\ \cdot \\ \cdot \\ \cdot \\ x_m(t) \end{bmatrix} = \text{a mx1 vector containing the values of each time series for } t = 0, 1, 2, \ldots, N.$$

$a_i = (a_0, a_1, \ldots, a_\tau)$ = a set of mxm coefficient matrices for $i = 0, 1, \ldots, \tau$

τ = time lag

α = prediction distance

Minimization of Mean Square Error

In designing a "digital filter" or a "set of weights" to predict multiple time series we use the mean square error criterion.

The input to our digital system is $x(t)$. The desired output is defined as $x(t+\alpha)$ and the actual output is $y(t)$ where

$$y(t) = a_0\, x(t) + a_1\, x(t-1) + a_2\, x(t-2) + \ldots + a_\tau\, x(t-\tau) \qquad (1a)$$

To obtain the "best" prediction equation, we minimize

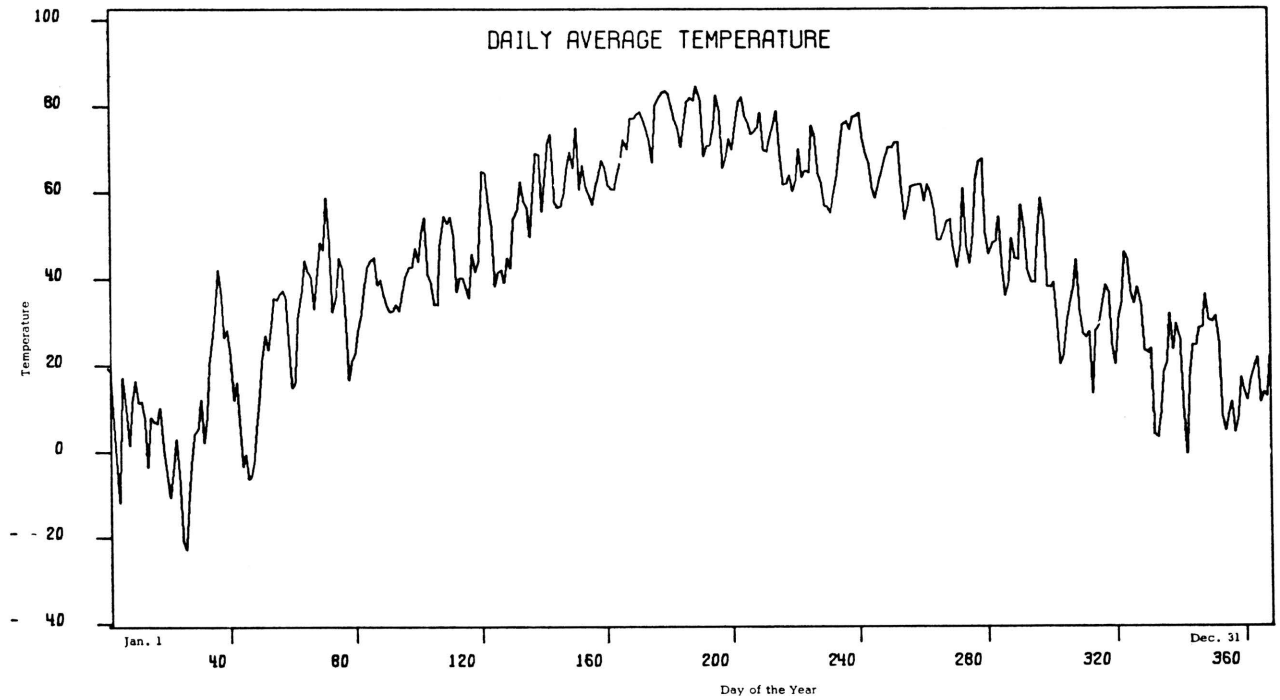


Figure 2.- Daily average temperature.

the mean squared error between the desired output $x(t+\alpha)$ and the actual output $y(t)$. The error is defined as

$$\sum_{t=0}^{N} e^2(t) = \sum_{t=0}^{N} [x(t+\alpha)-y(t)]^2 \quad (2)$$

where

$$e^2(t) = \begin{bmatrix} e_1^2(t) \\ e_2^2(t) \\ \cdot \\ \cdot \\ \cdot \\ e_m^2(t) \end{bmatrix}$$

Taking the partial derivatives of $\sum_{t=0}^{N} e^2(t)$ with respect to each coefficient matrix $[a_i]$ and setting the resulting equations to zero, we obtain the following matrix equation.

$$\begin{bmatrix} [r_{11}] & [r_{12}] & \cdots & [r_{1k}] \\ [r_{12}]^T & [r_{22}] & \cdots & [r_{k-1,k}] \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ [r_{1k}]^T & [r_{k-1,k}]^T & \cdots & [r_{k,k}] \end{bmatrix} \begin{bmatrix} [a_0] \\ [a_1] \\ \cdot \\ \cdot \\ \cdot \\ [a_\tau] \end{bmatrix} = \begin{bmatrix} [r_{1,\alpha+1}]^T \\ [r_{1,\alpha+2}]^T \\ \cdot \\ \cdot \\ \cdot \\ [r_{1,\alpha+k}]^T \end{bmatrix} \quad (3)$$

Each matrix $[r_{ij}]$ is a mxm square matrix for m time-series problem $[r_{ij}]$ contains all terms of the $\pm\,|(j-1)|^{th}$ lag of the autocorrelations and cross-correlations of the input time series.

For stationary random time series, the correlation matrix $[R]$ in equation (3) can be simplified. We can show

$$[r_{11}] = [r_{22}] = \ldots = [r_{k,k}].$$

Let us define the matrices $[r_{11}]$, $[r_{22}]$, ..., $[r_{k,k}]$.

$$[r_{11}] = \sum_{t=0}^{N} x(t)\, x(t)^T$$

$$[r_{22}] = \sum_{t=0}^{N} x(t-1)\, x(t-1)^T$$

$$\vdots \qquad\qquad \vdots$$

$$[r_{k,k}] = \sum_{t=0}^{N} x(t-\tau)\, x(t-\tau)^T$$

where $\tau = k-1$

To show that $[r_{11}] = [r_{22}]$ the following equation must hold.

$$\sum_{t=0}^{N} x(t)\, x(t)^T \qquad \sum_{t=0}^{N} x(t-1)\, x(t-1)^T$$

$$\boxed{x(-1)\, x(-1)^T} \text{ END EFFECT}$$

$$x(0)\, x(0)^T \qquad x(0)\, x(0)^T$$
$$x(1)\, x(1)^T \qquad x(1)\, x(1)^T$$
$$\vdots \qquad\qquad \vdots$$
$$x(N-1)\, x(N-1)^T \qquad x(N-1)\, x(N-1)^T$$

END EFFECT $\boxed{x(N)\, x(N)^T}$

The N terms are the same and the (N + 1) terms are different. If we assume the time series to be random and stationary and N is large, then $x(-1)\, x(-1)^T$ is nearly equal to $x(N)\, x(N)^T$. Under the same assumptions

$$[r_{11}] = [r_{22}] = \ldots = [r_{k,k}]$$

Similarly, we can show that all matrices on each subdiagonal and superdiagonal are equal.

Final Equations for Digital Filters

The equations which have to be solved for the set of coefficient matrices $[a_i]$ are given by:

$$\begin{bmatrix} [r_{11}] & [r_{12}] & \ldots & [r_{1,k}] \\ [r_{12}]^T & [r_{11}] & \ldots & [r_{1,k-1}] \\ \vdots & \vdots & & \vdots \\ [r_{1,k}]^T & [r_{1,k-1}]^T & \ldots & [r_{11}] \end{bmatrix} \begin{bmatrix} [a_0] \\ [a_1] \\ \vdots \\ [a_\tau] \end{bmatrix} = \begin{bmatrix} [r_{1,\alpha+1}]^T \\ [r_{1,\alpha+2}]^T \\ \vdots \\ [r_{1,\alpha+k}]^T \end{bmatrix} \qquad (4)$$

All the elements along each diagonal of the correlation matrix [R] are the same and $[r_{ii}] = [r_{ii}]^T$.

Note that if we know the first row

$$[r_{11}], [r_{12}] \ldots, [r_{1k}]$$

where

$$[r_{11}] = \sum_{t=0}^{N} x(t)\, x(t)^T = \begin{bmatrix} \phi_{11}(0) & \phi_{12}(0) & \ldots & \phi_{1m}(0) \\ \phi_{21}(0) & \phi_{22}(0) & & \phi_{m-1,m}(0) \\ \vdots & \vdots & & \vdots \\ \phi_{m1}(0) & \phi_{m-1,m}(0) & \ldots & \phi_{mm}(0) \end{bmatrix}$$

$$[r_{12}] = \sum_{t=0}^{N} x(t)\, x(t-1)^T = \begin{bmatrix} \phi_{11}(1) & \phi_{12}(1) & \ldots & \phi_{1m}(1) \\ \phi_{21}(1) & \phi_{22}(1) & & \phi_{m-1,m}(1) \\ \vdots & & & \vdots \\ \phi_{m1}(1) & \phi_{m-1,m}(1) & \ldots & \phi_{mm}(1) \end{bmatrix}$$

$$\vdots \qquad\qquad \vdots$$

$$[r_{1,k}] = \sum x(t)\, x(t-\tau)^T = \begin{bmatrix} \phi_{11}(\tau) & \phi_{12}(\tau) & \ldots & \phi_{1m}(\tau) \\ \phi_{21}(\tau) & \phi_{22}(\tau) & & \phi_{m-1,m}(\tau) \\ \vdots & & & \vdots \\ \phi_{m1}(\tau) & \phi_{m-1,m}(1) & \ldots & \phi_{mm}(\tau) \end{bmatrix}$$

then all the elements of the matrix [R] are shown.

For a different value of α, the matrix on the left-hand side of equation (4) is the same and it is only necessary to change the right-hand side and solve for the set of coefficients.

The predicted values $x_1(N+\alpha)$, $x_2(N+\alpha)$, ..., $x_m(N+\alpha)$ are obtained by solving the system of equations (4) and applying the solution to the original set of time series $x_i(t)$.

NUMERICAL EXAMPLE

To illustrate the method of predicting "multiple time series," let us consider two time series representing economic data. The data were obtained from "Business Statistics 1965" published by the United States Department of Commerce. One time series represents an index of the value of total industrial production in the United States by month for the period of January 1947 to June 1960 (Fig. 3). The other time series represents the total production of crude petroleum in the United States by month for the same period (Fig. 4).

Results were obtained in predicting each time series for five consecutive periods (July 1960 through November 1960). Three different filter lengths were tried giving the following average percent error for both time series.

Filter Length	Average percent Error
8	1.86
10	1.83
12	2.13

The best filter length was 10 with an average error of 1.83 percent. The results using a filter of length 10 for each time series are shown in Tables 1 and 2.

Table 1.- Time-series results of industrial production index using a filter length of 10.

Date	Actual Values	Predicted Values	Absolute percent Error
July 60	103.9	105.1	1.15
Aug 60	107.6	106.8	0.74
Sept 60	108.9	105.5	3.12
Oct 60	110.3	103.9	5.80
Nov 60	106.5	101.9	4.32
Average error			3.03

FUTURE WORK

Nonstationary data represent any class of data whose statistical properties change with time. Consequently, the vast majority of physical data actually falls in this area. Data are arbitrarily assumed to be stationary for reasons of approximation and simplicity. Also if the data are slowly

Figure 3.- Industrial production index.

changing with time, we consider it to be stationary.

Table 2.- Time-series results of crude petroleum production using a filter length of 10.

Date	Actual Values	Predicted Values	Absolute percent Error
July 60	212.6	212.0	0.28
Aug 60	215.1	215.4	0.14
Sept 60	209.1	210.8	0.81
Oct 60	215.7	215.0	0.32
Nov 60	214.0	210.5	1.64
Average error			0.64

standard derivation, and (3) a combination of the above. For joint statistical properties between values of a single nonstationary process at different times, the process can be described further in terms of its nonstationary autocorrelation function.

The selective partitioning approach groups the given nonstationary time series into sequence of stationary time series. The finite state machine is set up which consists of linear predictor operators for each partition of the time series and the director which tells the machine when to change from one state to the next state. That is, at each state a linear filter is given but these filters change in time and space according to the action of the finite state machine.

The basic idea of this approach is to construct a class of generalized linear predictors which fit a set of data for different intervals of time. If we are interested in predicting a particular interval of time in the future, we choose the predictor which will

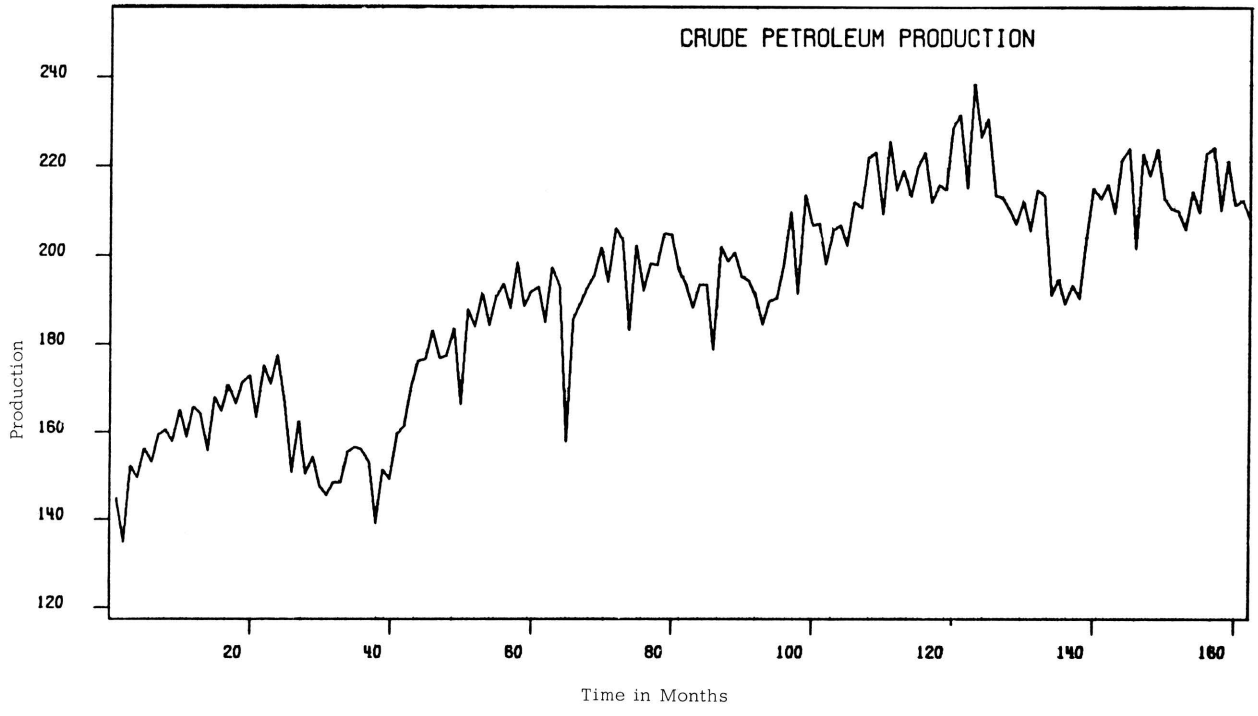


Figure 4.- Crude petroleum production.

A totally adequate methodology does not exist as yet for the analysis of all types of nonstationary data. This is partly because a nonstationary conclusion is generally a negative statement specifying the lack of stationary properties, rather than defining the precise nature of the nonstationarity.

There are various types of nonstationary data. Three basic and important types which can represent certain physically occurring nonstationary data are (1) a time-varying mean value, (2) time-varying

operate best on those data. The approach is based on the assumption that not all past data are significant to the prediction of the future values. The selective partitioning approach may be considered as combined detection-estimation procedure.

The seismic correlation problem involves seismic data from which we would like to infer the geologic structure. To infer the geologic structure may be described roughly as follows. Look at the collections of all inferences over past and present

which have characteristics in agreement with the given past data and which can be performed by present machines. Then choose the simplest member of this class of machines.

For example, suppose we are given some sequences of seismic data (the inputs) and the resulting geologic inferences (the outputs) as well as the true geologic structures (the desired outputs). Then we would consider the collection of all finite state machines which agree with the given data. From this collection we could choose the one with the smallest complexity. This machine would be the one into which we would feed fresh seismic data in order to make the geologic inferences.

Consider the problem of estimating the gas load for a particular day when we are given the data representing the temperature, wind velocity and gas load for each day for the last three years. The underlying temperature and wind velocity which affect the gas load can be divided in three periods (1) winter, (2) summer, and (3) marginal, i.e. periods during fall and spring which cannot be considered as part of winter or summer. For this problem the finite state machine will consist of three states, one state for each period. For each state the optimum linear filter is computed using the minimum mean square error criteria.

The finite state machine acts as a predictor-director. The director tells the machine when to change from one state to the next state. Using this approach, one can forecast the gas load when estimated value of temperature and wind velocity are given for any day of the year.

This problem of designing near optimum nonlinear filters has been reduced to two separate subproblems, namely the design of linear filters on local criteria and the design of finite state machines to control these linear filters on global criteria. The key word for the first subproblem is "linear"; for the second, "finite." There are methods for handling each of these subproblems.

REFERENCES

Bendat, J. S., and Piersal, A. G., 1966, Measurement and analysis of random data: John Wiley and Sons, New York, 390 p.

Cramer, H., and Leadbetter, M. R., 1967, Stationary and related stochastic processes: John Wiley and Sons, New York, 348 p.

Hingorani, G. G., 1966, Forecasting economic time series generated by random processes, Bus. and Econ. Stat. Sec., Proc. Am. Stat. Assoc., p. 409-422.

Hingorani, G. G., and Marczynski, L. F., 1967, Report on the Minneapolis Gas Company load-weather forecasting study: Gas Supply Res. Dept., Northern Natural Gas Co., internal rept., 34 p.

Lee, Y. W., 1960, Statistical theory of communication: John Wiley and Sons, New York, 509 p.

Robinson, E. A., 1953, Predictive decomposition of time series with applications to seismic exploration: Ph. D. Thesis, M. I. T., Geophysical Analysis Group, Rept. no. 7, 66 p.

Robinson, E. A., in press, Multichannel time series analysis with digital computer programs: Holdenday, Inc., San Francisco, 325 p.

Robinson, E. A., and Treitel, S., 1964, Principles of digital filtering: Geophysics, v. 19, no. 3, p. 395-404.

Wiener, N., 1949, The extrapolation, interpolation and smoothing of stationary time series: John Wiley and Sons, New York, 163 p.

Wozencraft, J. M., and Jacobs, I. M., 1965, Principles of communication engineering: John Wiley and Sons, New York, 720 p.

SOME DISTRIBUTION PROBLEMS IN TIME-SERIES SIMULATION[1/]

by

N. C. Matalas

U. S. Geological Survey

ABSTRACT

To generate synthetic sequences by means of a first-order Markov process, the probability distribution of the random component, ϵ, in the process must be considered in terms of the form of resemblance that is desired between the historical and synthetic sequences. For resemblance in terms of moments of order $m \leq 3$, no special difficulties are encountered. If the historical events are assumed to follow a particular probability distribution, then the probability distribution of ϵ is uniquely defined and this may give rise to difficulties in generating synthetic sequences.

INTRODUCTION

The responses of complex systems to time dependent inputs are difficult to determine analytically. Simulation offers a means, and in some instances the only means, of evaluating input-output relations. If a historical sequence of inputs is "routed" through a system, the sequence will yield a single response of the system. The historical sequence is unlikely to be repeated in the future, so that the single response is not representative of future responses of the system. With several simulated input sequences, each "routed" through the system, a set of responses may be obtained from which properties of the probability distribution of the responses may be assessed.

For the statistical properties of the simulated responses to have practical utility, the simulated sequences, referred to hereafter as synthetic sequences (Thomas and Fiering, 1962), must bear some resemblance to the historical sequence. Resemblance may refer to the moments that characterize the historical sequence or to the assumed underlying probability distribution of the historical events. If a historical sequence is characterized by moments of order 1 through m, then the moments of order 1 through m for a synthetic sequence must converge to the corresponding historical moments as the length of the synthetic sequence tends to infinity. If the historical events are assumed to follow a specific probability distribution, then the synthetic events for a sequence of finite length must represent a random sample from a population that has the assumed underlying probability distribution.

Various time-series models may be used to generate synthetic sequences, however, only one model, the first-order Markov process, will be considered. The following paragraphs discuss some difficulties associated with the use of this model in attempting to achieve resemblance either in terms of moments or with respect to an assumed underlying probability distribution.

1/ Publication authorized by Director, U. S. Geological Survey.

FIRST-ORDER MARKOV PROCESS

The first-order Markov process is defined as

$$(x_{i+1} - \mu_x) = \rho_x(1)\,(x_i - \mu_x) + \left[1 - \rho_x^2(1)\right]^{1/2} \sigma_x\, \epsilon_{i+1} \qquad (1)$$

where x_i and x_{i+1} denote the events at the time points i and i+1, respectively, μ_x and σ_x are the mean and standard deviation of x, respectively, $\rho_x(1)$ is the lag-one autocorrelation coefficient for x, and ϵ_{i+1} is a random component that is independent of x_i. For an historical sequence, the values of μ_x, σ_x, and $\rho_x(1)$ are unknown, but their estimates $\hat{\mu}_x$, $\hat{\sigma}_x$, and $\hat{\rho}_x(1)$ may be obtained.

To achieve some specified degree of resemblance between the historical and synthetic sequences, an appropriate choice of the probability distribution of ϵ must be made. The choice of this distribution is discussed below. Once the probability distribution of ϵ is specified, a synthetic sequence may be generated by equation (1), with μ_x, σ_x, and $\rho_x(1)$ replaced by their respective historical estimates, in the following manner. The most recent historical event is represented by x_i and a value for ϵ_{i+1} is randomly selected from a population that has a specified underlying probability distribution. With these two values, equation (1) yields a value for x_{i+1}, which is the first synthetic event. The value for x_{i+1} now assumes the role of x_i and with a new random selection of a value for ϵ_{i+1}, equation (1) yields a new value for x_{i+1}, which is the second synthetic event. This procedure is repeated N times to obtain a sequence of N synthetic events.

As N tends to infinity, certain properties, depending upon the desired degree of resemblance, of the synthetic sequence will converge to the corresponding properties of the historical sequence. Essentially, the values of these properties for the historical sequence act as population values relative to the corresponding values of the properties for the synthetic sequence. In practice N will be finite, however large, so that resemblance between the historical and synthetic sequences is said to exist if the values of the synthetic properties do not depart from the corresponding values of the historical properties by more than is expected by chance.

MOMENT RESEMBLANCE

The following discussions are limited to statistical parameters that are functions of moments of order m = 1, ..., 4. More specifically, m = 1 and m = 2 refer to the mean and standard deviation, respectively, and m = 3 and m = 4 refer to the coefficients of skewness and kurtosis, respectively. For a specified value of m, resemblance is with respect to all parameters that are functions of moments of order equal to and less than m. The lag-one serial correlation coefficient is defined in terms of second order moments and thus resemblance of degree $m \geq 2$ requires this coefficient be preserved as well as the standard deviation.

For any value of $m \geq 1$, ϵ must follow a probability distribution that has zero mean, and for $m \geq 2$, the probability distribution of ϵ must have zero mean and unit variance. The skewness, $\beta_1(\epsilon)$, and kurtosis, $\beta_2(\epsilon)$, of ϵ will depend upon the sample values of skewness, $\hat{\beta}_1(x)$, and kurtosis, $\hat{\beta}_2(x)$, of x. The relations between the coefficients of skewness and coefficients of kurtosis of x and ϵ are

$$\beta_1(\epsilon) = \frac{[1 - \hat{\rho}_x^3(1)]^2}{[1 - \hat{\rho}_x^2(1)]^3} \hat{\beta}_1(x) \qquad (2)$$

$$\beta_2(\epsilon) = \frac{[1 + \hat{\rho}_x^2(1)]}{[1 - \hat{\rho}_x^2(x)]} \beta_2(x) - \frac{6\hat{\rho}_x^2(1)}{[1 - \hat{\rho}_x^2(1)]} \qquad (3)$$

For $\hat{\rho}_x^2(1) \geq 0$, $\beta_1(\epsilon) \geq \hat{\beta}_1(x)$, where equality holds if $\hat{\rho}_x(1) = 0$ or if $\hat{\beta}_1(x) = 0$. Note that kurtosis is defined as the fourth central moment divided by the square of the variance, and therefore kurtosis is a positive value. If $\hat{\beta}_2(x) = 3$, then $\beta_2(\epsilon) = 3$ for all values of $\hat{\rho}_x^2(1) \geq 0$. For $\hat{\rho}_x^2(1) > 0$, $\beta_2(\epsilon) > \hat{\beta}_2(x)$ if $\hat{\beta}_2(x) > 3$ and $\beta_2(\epsilon) < \hat{\beta}_2(x)$ if $\hat{\beta}_2(x) < 3$, and for $\hat{\rho}_x(1) = 0$, $\beta_2(\epsilon) = \hat{\beta}_2(x)$. However, if

$$\hat{\beta}_2(x) \leq \frac{6\hat{\rho}_x^2(1)}{[1 + \hat{\rho}_x^2(1)]} \qquad (4)$$

then $\beta_2(\epsilon) \leq 0$, which is not an admissible range of values for kurtosis, in which instance, a first-order Markov process cannot be used to generate synthetic sequences and achieve resemblance of degree m = 4 for such a case.

For resemblance of degree m = 2, there is considerable latitude in the choice of the probability distribution of ϵ. All that matters is that the probability distribution have zero mean and unit variance. The choice of distribution of ϵ is narrowed if resemblance is extended to m = 3. Here, ϵ must be distributed with zero mean, unit variance, and skewness as specified by equation (2). If resemblance is extended to m = 4, the probability distribution of ϵ, the choice of which is limited, must have zero mean, unit variance, and skewness and kurtosis as specified by equations (2) and (3), respectively. The choice of the probability distribution of ϵ is said to be wide or narrow in the sense that the choice is likely to be made among the relatively few distributions commonly used in statistical practice where generally there is a unique relation between the coefficients of skewness and kurtosis.

As long as resemblance is desired in terms of moments, the probability distribution that underlies the historical sequence need not be considered. Therefore, the choice of probability distribution of ϵ is an operational one, that is, any choice is allowable that permits the desired degree of moment resemblance to be achieved. More than one probability distribution of ϵ may be used, in which case, the choice of one of them may be guided by the simplicity with which the random sequence for ϵ may be generated.

DISTRIBUTION RESEMBLANCE

If x is assumed to follow a specific distribution and if resemblance of degree m is desired, then the distribution of ϵ must be uniquely defined. The following discussions of resemblance in terms of the probability distribution and certain moments of x are based on the assumption of strict stationarity (Papoulis, 1964). With this assumption, the probability distributions of x_i and x_{i+1} are identical and independent of i. Because for the first-order Markov process x_i and ϵ_{i+1} are independent for all values of i,

$$\Phi[\theta : x] = \Phi[\theta : z]\, \Phi[\theta : \eta] \qquad (5)$$

where $\Phi[\theta : x]$, $\Phi[\theta : z]$, and $\Phi[\theta : \eta]$ denote the characteristic functions of x, z, and η, respectively. In this notation, $z = \hat{\rho}_x(1)x$ and $\eta = [1 - \hat{\rho}_x^2(1)]^{1/2} \hat{\sigma}_x \epsilon$.

If the characteristic function for x is known, the characteristic function for z is easily obtained, whereby, the ratio of $\Phi[\theta:x]$ to $\Phi[\theta:z]$ gives the characteristic function for η. However, the derivation of $f(\eta)$, the probability density function of η, may be a difficult task.

For a given probability density function f(x), the probability density function $f(\eta)$ that satisfies equation (5) may present operational difficulties. That is, it may not be possible to generate in a convenient manner random numbers that follow $f(\eta)$. If, however, $f(\eta)$ is integrable in closed form over the range $(-\infty, \eta)$, random numbers that follow $f(\eta)$ can be generated easily.

If x is assumed to be normally distributed, then ϵ is normally distributed. Thus in special instances where x can be transformed to a normal variate, there is no need to derive $f(\eta)$ from $\Phi[\theta:\eta]$. For example, suppose x follows a log-normal distribution, so that y = log x, where logarithm is to the base e, is normally distributed. For x, the mean, $\hat{\mu}_x$, and variance, $\hat{\sigma}_x^2$, are related to the mean, μ_y, and variance, σ_y^2, for y by

$$\hat{\mu} = \exp\left[\frac{1}{2}\sigma_y^2 + \mu_y\right] \tag{6}$$

$$\hat{\sigma}_x^2 = \exp[2(\sigma_y^2 + \mu_y)] - \exp[\sigma_y^2 + 2\mu_y] \tag{7}$$

(Aitchison and Brown, 1957). If y instead of x is assumed to be generated by a first-order Markov process, then in terms of x, the generating process is

$$x_{i+1} = \{\exp[\mu_y(1-\rho)]\} x_i^{\rho} \delta_{i+1} \tag{8}$$

where $\rho = \rho_y(1)$, $\delta_{i+1} = \exp\{[1-\rho^2]^{1/2}\sigma_y \epsilon_{i+1}\}$, and ϵ_{i+1} is normally distributed with zero mean and unit variance. $\hat{\rho}_x(1)$ is related to ρ by

$$\hat{\rho}_x(1) = \{\exp[\sigma_y^2 \rho] - 1\}/\{\exp[\sigma_y^2] - 1\} \tag{9}$$

From the observed values of $\hat{\mu}_x$, $\hat{\sigma}_x$, and $\hat{\rho}_x(1)$, the values of μ_y, σ_y, and ρ may be obtained from the above relations, whereupon, equation (8) may be used to generate synthetic events that are log-normally distributed and where moment resemblance is of degree m = 2. The log-normal case with m = 3 and the case where x follows a gamma distribution with m = 3 has been discussed elsewhere (Matalas, in press).

If moment resemblance of degree m = 4 is desired, then the assumed probability density function, f(x), must be considered in terms of the relation between $\beta_1(x)$ and $\beta_2(x)$. If f(x) implies a unique relation between $\beta_1(x)$ and $\beta_2(x)$, the historical values $\hat{\beta}_1(x)$ and $\hat{\beta}_2(x)$ will have to satisfy this relation, which is unlikely. Consequently to generate synthetic sequences some degree of resemblance, either in terms of f(x) or in terms of statistical parameters, will need to be sacrificed. In practice the number of historical events is seldom large enough to permit a strong discrimination on the basis of goodness of fit among various assumed probability density functions. Moreover, the statistical parameters estimated from the historical events are subject to standard errors that tend to increase rapidly as m increases. Therefore, resemblance of degree $m \geq 3$ may be misleading, in the sense that the high order moments estimated from the historical sequence may be poorly representative of their respective population values.

SUMMARY AND CONCLUSIONS

The time dependent structure of an historical sequence of events may be approximated by first-order Markov process, whereby, synthetic sequences may be generated. The generation and application of the synthetic sequences requires that some degree of resemblance between the historical and synthetic sequences must be specified. Two forms of resemblance have been considered. The first pertains to parameters defined in terms of moments, and the second, to the probability distribution that is assumed to underly the historical events. For either form, the probability distribution of the random component, ϵ, in the first-order Markov process must be considered. For moment resemblance, however, the choice of the probability distribution of ϵ is an operational one, and in general more than one distribution of ϵ may be used. For distribution resemblance, the choice of the probability distribution of ϵ is uniquely defined by the probability distribution that is assumed to underly the historical events.

No special difficulties are encountered in generating synthetic sequences by means of a first-order Markov process if resemblance is desired in terms of parameters defined by moments of order $m \leq 3$. Where $m > 3$ or where resemblance is desired in terms of an assumed underlying probability distribution of the historical events, it is not easily handled by means of a first-order Markov process. With respect to distribution resemblance, the difficulty in simulation arises from the relation that may exist between the coefficients of skewness and kurtosis for the probability density function, f(x), that is assumed for the historical events. The historical values of the coefficients of skewness and kurtosis are unlikely to satisfy the relation, so that the particular f(x) cannot be assumed.

In practical situations, the exact form of f(x) may not be important, and only moment resemblance need be considered in the generation of

synthetic sequences. To say that the form of f(x) is important implies that moments of all orders must be considered. To determine if the form of f(x) matters, moment resemblance may be limited to m = 3, and for different forms of f(x), synthetic sequences may be generated. These sequences may be used to assess the sensitivity of the system's responses to the various forms of f(x).

REFERENCES

Aitchison, J., and Brown, J. A. C., 1957, The log-normal distribution: Cambridge Univ. Press, London

Matalas, N. C., in press, Mathematical assessment of synthetic hydrology: Water Resources Res.

Papoulis, A., 1965, Probability, random variables, and stochastic processes: McGraw-Hill Book Co., New York, 583 p.

Thomas, H. A., and Fiering, M. B., 1962, Mathematical synthesis of streamflow sequences for the analysis of river basins by simulation, in Design of Water-Resource Systems: Harvard Univ. Press, Cambridge, Mass.

SPECTRAL-DENSITY ANALYSIS OF STRATIGRAPHIC DATA

by

C. John Mann

University of Illinois

ABSTRACT

Numerically coded data on lithology, color, bedding, and thickness for two stratigraphic sections of Missourian (Pennsylvanian) rocks were analyzed by spectral densities. Geological assumptions required that deposition was continuous and at a uniform rate for 4.14 million years throughout the sequence. Although noise level is high for the analysis of stratigraphic data, periodicities of 109,000; 30,800; and 24,250 years may be interpreted from the spectral density.

INTRODUCTION

Determination of the spectral density (power-spectral density, power spectrum, covariance spectrum, or second-moment spectrum) is a method which may be used to study cyclic components of a time series. The technique has been developed extensively and employed widely by engineers in studying communications, electrical systems, and data-processing systems. Less frequently, it has been utilized successfully by geophysicists, economists, astronomers, oceanographers and meteorologists. Thorough mathematical treatment of the measurement and calculation of power-spectral densities has been given by Blackman and Tukey (1959), Cox and Lewis (1966), Grenander and Rosenblatt (1957) and Rice (1944, 1945).

SPECTRAL-DENSITY ANALYSIS

Briefly and simply, the power-spectral density is a function of frequency and represents that contribution to the total variance of the time series from frequencies within a given interval of frequencies (see the Appendix). It provides a harmonic analysis of a series variate x_i as a function of time. Blackman and Tukey (1959) define the spectral density as the "...value of a function whose integral over any frequency interval represents the contribution to the variance from that frequency interval." The term, power, is derived from one of its first applications in electrical engineering. If we consider the voltage across, or the current through, a resistance of 1 ohm, the average power dissipated in the resistance theoretically will be proportional to the variance of the voltage, or of the current.

Mathematically, the spectral density in a continuous time series is simply the Fourier transform of the autocovariance function. The autocovariance is a function of the statistical correlation existing between values of a variate, $x_i(t)$, in a series and the value of that same variate at a constant interval or lag of time, $x_{i+p}(t)$. Thus the range of frequencies represented by the spectral density is determined by the arbitrarily chosen interval of lag used to determine the autocovariance.

Assumptions must be made commonly in applying spectral-density analysis to practical problems. Although these assumptions may not in fact be true and may result in a theoretically incorrect spectral density, the answer may be a sufficiently good estimate of the actual spectrum so that the analysis is beneficial for the investigator (Blackman and Tukey, 1959). Spectral-density measurement in theory requires that the data are stationary Gaussian random processes with zero means. Exact determination of the spectrum would require an infinitely long piece of a continuous random function which has been measured perfectly. Inaccuracies and uncertainties necessarily are introduced into the analysis if the sequence is not time invariant, the data sequence is finite in length, the data lack normality of distribution, or the data are discrete observations of a continuous series.

The spectral densities for this investigation were calculated following a procedure proposed for digital computation by Southworth (1960). The hamming-type spectral window for smoothing values of the raw spectral-density estimates was employed. Southworth's method assumes that the discrete values are taken at equal intervals of time from the stationary series. Digital computation was performed on the University of Illinois' IBM 7094 Computer. Basic mathematical relations are provided in the Appendix.

STRATIGRAPHY

Two stratigraphic sequences of Missourian age (Pennsylvanian) were analyzed for their spectral densities. The sections, one from Superior, Arizona, and the other from Honaker Trail, Utah, (Fig. 1), were selected for the seemingly continuous deposition they represent and because they were

known to be cyclic in lithology and equivalent to better known cyclothemic sequences of the Midcontinent region. The sections are portions of a thicker sequence of Atokan, Desmoinesian, Missourian, and Virgilian strata approximately 426 km apart. The Superior section is mainly carbonate strata, whereas the Honaker Trail rocks are primarily shale and fine clastics (Fig. 1). Both sections were measured and described in detail by Professor Harold R. Wanless.

Several assumptions concerning the geological data were made for the purposes of this study. These strata are assumed to represent uninterrupted deposition during Missourian time, an interval taken to be 4.14 million years (based on data from Francis and Woodland, 1964), and the sediments are assumed to have accumulated at a uniform rate irrespective of lithology. Under these assumptions, discrete stratigraphic thickness may be related to specific intervals of time depending solely upon the total thickness of the Missourian section and the average depositional rate for each stratigraphic section. Both sections were subsequently "sampled" from the field descriptions at 10,000-year intervals for digital computation of the spectral densities. Ten thousand years, under these assumptions, are represented in the Superior section by 15.91 cm (0.522 ft) and in the Honaker Trail section by 37.33 cm (1.225 ft).

Qualitative elements of the field descriptions including lithology, color, and bedding were quantified by establishing arbitrarily a coding scale which for most aspects was between 0 and 10.

RESULTS OF SPECTRAL-DENSITY ANALYSIS

Spectral densities were computed for thickness, lithology, color, and bedding for both stratigraphic sections. These rock properties are not equally good indicators of strata cyclicity as might be surmised a priori. Thickness and lithology are more consistent than bedding and better than color as an index to periodicities (Fig. 2). The noise level for the stratigraphic data is generally high; noise levels for the individual factors vary considerably. In spite of these differences, the periodicities which are revealed by analysis of the individual properties are strikingly similar (Fig. 2).

Under conditions imposed by the data and the parameters of computation, the theoretical frequency resolution resulting from this spectral-density analysis is 20,000 to 600,000 years per cycle. The high noise level, however, generally tends to obscure frequencies near the extreme values.

Three frequencies appear to stand out more than others above the general noise level in these two stratigraphic sections. These frequencies are 109,000; 30,800; and 24,250 years. Interestingly, the first periodicity agrees well with the perihelion cycle of 112,000 years (van den Heuvel, 1966) and the last may be associated with the precession cycle

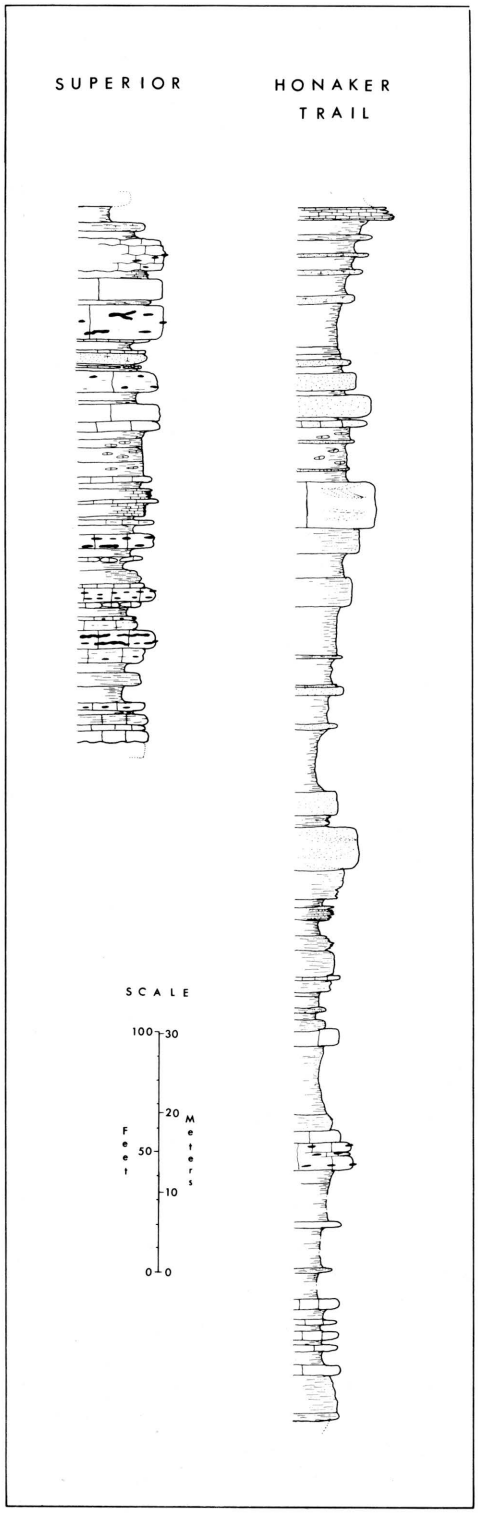


Figure 1.- Columnar sections of Missourian strata in Queen Creek Canyon (sec. 36, T. 1 S., R. 12 E.) 2 miles northeast of Superior, Arizona, and along Honaker Trail in San Juan River Canyon about 6 miles northwest of Mexican Hat Lodge, Utah.

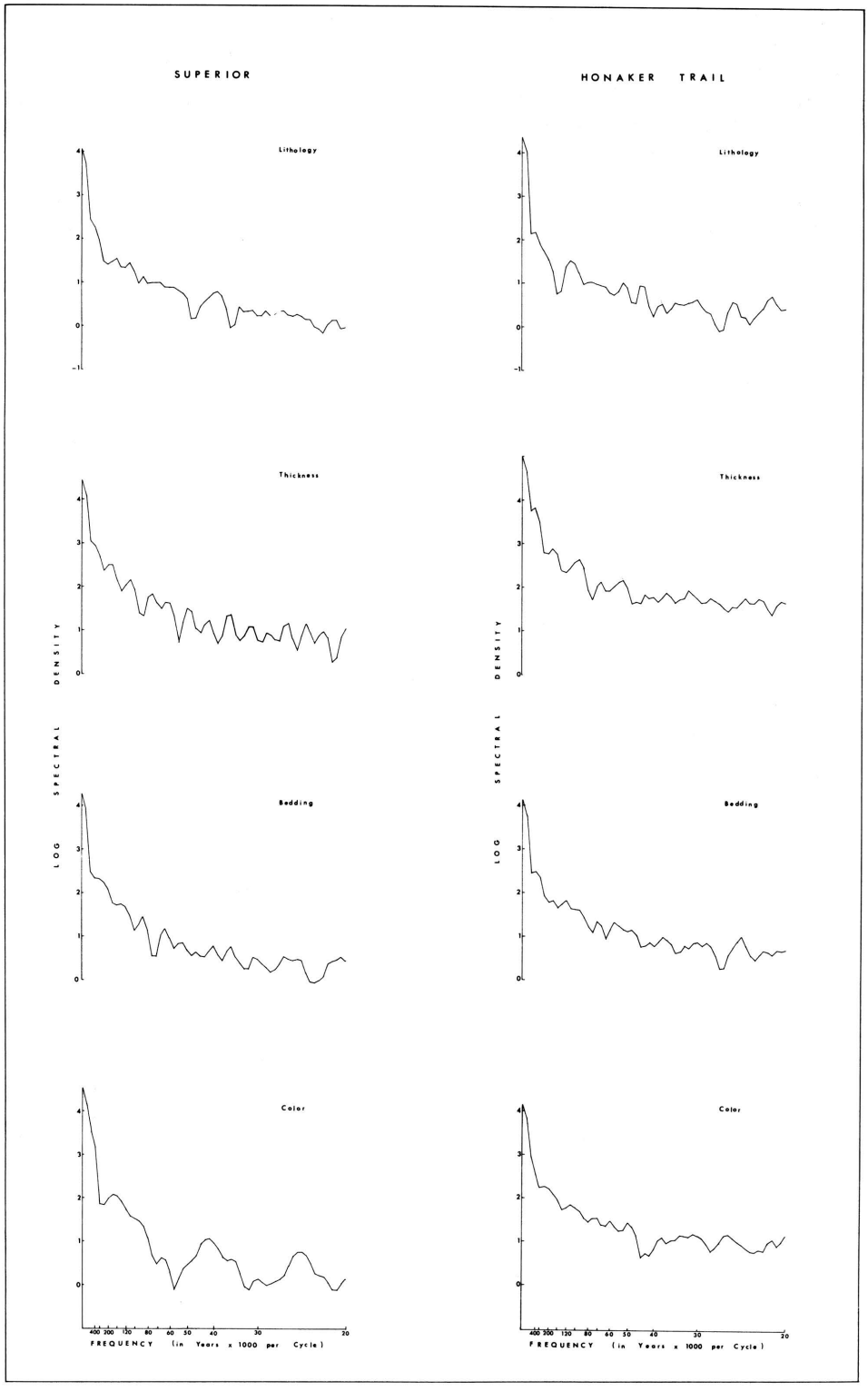


Figure 2.- Spectral densities calculated for lithology, thickness, bedding and color for Missourian strata in Superior and Honaker Trail sections. Noise level is approximately intermediate between minor peaks and troughs of curves.

of 25,789 years. The intermediate periodicity recognized is unassociated as yet with known solar cycles and may represent a harmonic of one of the primary frequencies. Obviously, additional investigation and study is necessary to substantiate these preliminary conclusions; but nonetheless, these suggested periodicities are stimulating and intriguing.

Although other frequencies are evident on the spectrograms (Fig. 2), they appear to be less consistent or to be simple multiples of other primary frequencies and hence are discounted.

CONCLUSIONS

Results obtained in this preliminary analysis of the cyclic nature of Missourian strata suggests that numerical analysis of time series by digital computation may reveal many aspects concerning the nature of strata and geological processes.

An admitted shortcoming in this investigation has been the geological assumptions necessary to fit the quantitative requirements of the technique employed. Further investigation into the power, accuracy, and limitations of these numerical analysis methods applicable to geological data and geological situations, however, should reveal their validity and provide insight for a re-evaluation of the geological assumptions and interpretations.

REFERENCES

Blackman, R. B., and Tukey, J. W., 1959, The measurement of power spectra: Dover Publ., Inc., New York, 190 p.

Cox, D. R., and Lewis, P. A. W., 1966, The statistical analysis of series of events: Methuen and Co., Ltd., London, 283 p.

Francis, E. H., and Woodland, A. W., 1964, The Carboniferous Period, in The Phanerozoic Time-Scale: Quart. Jour. Geol. Soc. London, v. 120S, p. 221-232.

Grenander, Ulf, and Rosenblatt, Murray, 1957, Statistical analysis of stationary time series: John Wiley and Sons, New York, 300 p.

Rice, S. O., 1944, 1945, Mathematical analysis of random noise: Bell System Tech. Jour., pt. I, v. 23, p. 283-332; and Pt. II, v. 24, p. 46-156.

Southworth, R. W., 1960, Autocorrelation and spectral analysis, in Mathematical methods for digital computers, John Wiley and Sons, New York, p. 213-220.

van den Heuvel, E. P. J., 1966, On the precession as a cause of Pleistocene variations of the Atlantic Ocean water temperatures: Geophys. Jour. Royal Astronomical Soc., v. 11, p. 323-336.

APPENDIX

The autocovariance function W(p) for a continuous variate x(t) with that same variate at a constant lag interval x(t+p) is

$$W(p) = \lim_{T\to\infty} 1/T \int_{-T/2}^{T/2} x(t)\, x(t+p)\, dt$$

where T is the total time over which observations of the variate are made and p is the lag interval. The autocovariance function may also be shown (Blackman and Tukey, 1959; Cox and Lewis, 1966) to be the Fourier transform of a distribution function P(f);

$$W(p) = \int_{-\infty}^{\infty} P(f)\, e^{i\, 2\pi fp} df$$

where

$$P(f) = \lim_{T\to\infty} 1/T \left[\int_{-T/2}^{T/2} x(t)\, e^{-i\, 2\pi ft} dt \right]^2$$

with f being the frequency. Conversely, P(f) is the Fourier transform of W(p);

$$P(f) = \int_{-\infty}^{\infty} W(p)\, e^{-i\, 2\pi fp} dp.$$

Because the function of P(f) may be shown to represent the contribution to the variance of x(t) from frequencies between f and f+df, it describes the spectrum of the process. Thus P(f) is called the spectral-density function for a stationary time series x(t).

For equally spaced discrete values $x_i(t)$ of a

stationary time series, a finite Fourier series transformation must be used instead of the infinite integral transformation. Raw estimates of the spectral density (Blackman and Tukey, 1959) may be calculated by

$$L(p) = W_o + 2 \sum_{q=1}^{M-1} W_q \cos qp\pi/M + W_M \cos p\pi$$

where the autocovariance is

$$W(p) = 1/(N-p) \sum_{i=1}^{N-p} x_i x_i + p$$

with M being the maximum lag and N the total number of observed discrete values of the series. The problem of smoothing raw estimates is complex and without definitive guidelines; it has been discussed and examined at great length by many workers in spectral-density analysis. The hamming-type of corrected estimates C(p) of the smoothed power density has been used in this study: here

$$C(p) = 0.23\, L_{(p-1)} + 0.54\, L_p + 0.23\, L_{(p+1)}$$

with $L_{-1} = L_1$ and $L_{(M+1)} = L_{(M-1)}$.

For a time series which has a mean value of zero, an assumption which is made for a stationary series, and discrete observations

$$W(p) = r_p S^2_x$$

where r_p is the autocorrelation coefficient for a lag p and S_x is the standard deviation of the observations, the autocorrelation coefficient r_p for a discrete variate x_i of a series with that same variate at a constant lag x_{i+p} is

$$r_p = \frac{(N-p) \Sigma x_i x_{i+p} - (\Sigma x_i)(\Sigma x_{i+p})}{[(N-p) \Sigma (x^2_i) - (\Sigma x_i)^2]^{1/2} [(N-p) \Sigma (x^2_{i+p}) - (\Sigma x_{i+p})^2]^{1/2}}$$

For economy of time in computation, the autocovariance was calculated with the autocorrelation coefficient.

Spectral-density estimates for finite series theoretically are reasonably accurate if the lag length does not exceed 5 to 10 percent of the total series observed. In this instance the resolution is equal to the maximum lag interval.

A WAVE STATISTICS MODEL FOR CLIMATIC TIME SERIES

by

Leslie Curry

University of Toronto

INTRODUCTION

Godske (1962) has stated "...the aim of statistical climatology is to arrive at mathematical models giving an adequate description of the statistical properties of the atmospheric variables." He then goes on to detail various stages such as distribution functions, time and space autocorrelations and cross-correlations between different elements. It is assumed that such models can be linked back genetically and forward functionally and thus provide a firm basis for an integrated climatology. Clearly it will be necessary to provide the models with a basis in hydrodynamics and thermodynamics, but at this stage the task is mathematical rather than physical.

In the progression of weather systems over a sampling point, we note the physical interrelations between our measured values and develop theory to account for them. If mathematical functions describe the relations between the instantaneous values of the weather elements and their rates of change, statistical functions should relate the aggregations of data of the elements which we know as climate. It is the purpose of this paper to propose a stochastic generating mechanism for these statistics. Initially, time variation of the elements is of concern, spatial variation being neglected; one-dimensional processes are, of course, easier to deal with but work in oceanography is available to aid future extensions into two dimensions. The strictly periodic modulation of weather processes by diurnal and annual cycles is excluded.

The purpose here is to outline a theory which can be checked almost immediately with available data. In this setting a good deal of heuristic argument is acceptable, the effort of formalizing being delayed until the physically more meaningful two-dimensional processes are tackled. Most of the statistical results we have, however, are for time variation, and many of these apparently disparate results seemingly are capable of being linked in a common schema.

A safer procedure may be to have a theoretical model to test on the data rather than examining the data alone. Not only can one make sense of the results by inferring processes but the model suggests ways of looking at the data not immediately apparent.

Previous Work.- With rare exceptions, the study of the form of climatic statistics has been pursued either in its own right or as a necessity to an understanding of the relations to other features of the landscape. This has led to the neglect of the relationships between the statistics of different climatic elements, both in terms of the general forms of density functions describing, say, cloud cover and rainfall, and the numerical values of the parameters of these functions. Also, there has been little attempt to describe possible probability generating mechanisms for these elements, particularly mechanisms which are consistent for the whole range of elements.

There have been several approaches to the probabilistic analysis of climate.

1. Probability density and distribution functions have been fitted to time collections of some elements.

2. Serial or autocorrelations have been calculated for time series and spectral densities worked out. Probability generating mechanisms, for example, without physical plausibility of the urn type, have been tested for rain days, no-rain days and temperatures.

3. One example exists of the use of one-dimensional wave statistics in studying the duration of temperature oscillations.

4. Le Cam (1961) provides the only example of an attempt to phrase physical processes in stochastic terms in his work on precipitation. This is a tentative, exploratory study which does not claim substantive results. Such a model also has been suggested for the analysis of cloud-seeding experiments.

5. The vexing problem of extreme values has been ascribed some probabilistic basis, but this is of a limited, ad hoc nature.

There has been a considerable amount of purely statistical analysis of such topics as singularities, synoptic-climatological forecasting, rainmaking and atmospheric budget-keeping.

WAVE BASIS OF THE MODEL

From observation and theory we know waves occur in the zonal winds. Recent work has shown many variables in time and space may be represented as additive harmonics. If we take the sine wave as the basic unit for our climatological models, we are within reach of physical theory. While not pursuing a rigorous interpretation, we may rely on qualitative

understanding. The statistical properties may be examined of a time sample taken at a point of characteristics of a succession of moving waves. The well-known resolution of summation of waves into standing waves and transient eddies is made. The latter are most tractable for the use of probability theory in our context because they exhibit stationarity in the time domain. It is important particularly to note that standing waves are not time dependent over periods on the order of a month or at least can be so regarded.

CLOUD COVER

The formation of clouds in level terrain is dependent on the degree of convergence experienced, moisture content of the air, and change in static stability as a result of heating or cooling at different levels. Passage of a single wave in the westerlies will produce periodic fluctuations in all quantities and a symmetric wave thus will give alternating periods of cloudy and clear weather. Note that in order to discuss cloud formation by waves, we need not refer to the amplitude of the waves but only to their phase.

Consider a collection of sine waves with a uniformly random phase distribution. Samples are taken at regularly spaced moments of time, i.e. once a day. Because cloud statistics are collected on this basis we can ignore any autocorrelation between consecutive readings. It may be shown (Bendat, 1957) that the probability density function of the amplitude (or similarly of the rate of change of amplitude) is

$$P(y) = (\pi\sqrt{A_0^2 - y^2})^{-1} \qquad -A_0 < y < A_0$$

where A_0 is the constant maximum amplitude of the waves. This is the well-known but misnamed arc-sine density. The corresponding distribution function is

$$P(y) = \frac{1}{\pi}\left(\frac{\pi}{2} + \sin^{-1}(y/A_0)\right).$$

It may be postulated that these functions describe the distribution of cloud cover in the absence of a standing wave. The latter now is added to the transient waves as a sinusoid of fixed phase for any sampling location. Where the rate of change of amplitude for the standing wave is zero (i.e. at the ridge and trough lines), the combined density function again will be arc-sine. Where the rate of change is positive, the number of occasions when negative values are recorded will be reduced and the frequency of positive values increased. The U-shaped function will become asymmetric, the asymmetry depending on the position of the sampling point within the standing wave. Such a distribution is the Beta-density function, a generalization of the arc-sine function (Feller, 1966):

$$\beta_{\mu,\nu}(x) = \left[\frac{\Gamma(\mu+\nu)}{\Gamma(\mu)\Gamma(\nu)}\right](1-x)^{\mu-1}x^{\nu-1}$$

for $0 < x < 1$

To maintain a U-shape, $\mu < 1$, $\nu < 1$, the term within the square brackets may be written

$$1/\int_{x=0}^{1} x^{\nu-1}(1-x)^{\mu-1}dx$$

for $\mu > 0$, $\nu > 0$, so that

$$\beta_{\mu,\nu}(x) = \frac{(1-x)^{\mu-1}x^{\nu-1}}{\int_{x=0}^{1}(1-x)^{\mu-1}x^{\nu-1}\,dx}$$

For $\mu = \frac{1}{2}$, $\nu = \frac{1}{2}$ this reduces to $(\pi/\sqrt{x(1-x)})^{-1}$, which is the arc-sine density. Note that the Beta integral produces a probability measure from the product of x and (1-x) with appropriate weights. It is likely that μ and ν are not independent for cloud statistics; this can be checked for a number of stations. Certainly they are available for physical interpretation. If $\mu + \nu = 1$, a possible result, the Beta densities are known as generalized arc-sine densities.

SUNSHINE DURATION

The next atmospheric variable to be considered is hours of sunshine. We make the assumption that sunshine is related inversely to percent cloud cover. Thus, the density function describing its probable duration is an integration over the day of instantaneous values of cloud cover subtracted from unity. To perform this integration, the form of the autocorrelation function for cloud cover should be known, but for a period as short as a day we can simply assume a negatively sloping ramp function without worrying about where and in what manner it reaches zero. If each period of, say, an hour was independent, we would be summing independent random variables and the central limit theorem would ensure normality. If each hour was the same as 9 am then, of course, a Beta distribution would result. Because the Beta-density converges to a normal one, it seems likely that sunshine duration can be specified by a Beta with the cloud parameters modified by autocorrelation. To the extent that intensity of sunshine is not a direct function of cloudiness, in particular the importance attached to height of cloud and low sun angles, there may be

need for some revision.

PRECIPITATION

A physical approach to a statistical theory of rain amounts is virtually impossible at the present time. In cloud physics, a quantitative theory of shower rains is nonexistent and qualitative theory controversial. At the microscale, we need some form of branching process from random initiations by precipitation nuclei, in turn affected by updraught conditions, windshear, temperatures, etc., and the whole affected by the distribution of storms. Of course, some of these conditions will correspond to the general process we are describing: lower level convergence, thermal stability and moisture content. We shall make the sweeping assumption that the microprocesses are functions of the macroconditions and these are described by our model.

A number of urn models have been devised for studying rainfall persistence. Gold (1929) obtained the probable number of runs of length r out of m events, assuming rainy and fine days to be equally probable. Cochran (1938) allowed for unequal probabilities but used independent successive trials. Gabriel and Neumann (1962) assumed a first-order Markov chain so that transition probabilities are constant and sequences are independent with lengths being geometrically distributed. In 1964, Feyerherm and Bark used this model for daily rain amounts after deterministic seasonal variations were removed by harmonic analysis. Wiser (1965) sought a more general solution with a variety of contagious models, i.e. transition probabilities were allowed to vary with run length. It is interesting to note that the schemes which worked best were those which caused persistence to increase with run length and then to decline beyond a certain point. Thus a Polya urn which increases chances of success with run length needed to be limited by a Bernoulli urn beyond a certain point or be generalized by a Friedman urn which essentially does the same in a more continuous fashion. Seemingly, we need a model which generates an autocorrelation function which declines slowly at first and then rapidly to zero. So far as is known no models have attempted to take both runs of rainy and fine periods into account, yet casual observation suggests that the autocorrelation function may well pass through zero into negative values and then return into positive values. This corresponds to our experience of rainy weather followed by fine, followed by rain again, and also to a succession of cyclones and anticyclones passing overhead whose phase becomes increasingly uncorrelated with time. We might thus expect an autocorrelation function of the form

$$R(\tau) = Ae^{-k|T|}\cos c\tau$$

where $A \geq 0$, $k > 0$, $c \geq 0$, and a spectral density of

$$G(\omega) = \left[\frac{2Ak}{\pi}\right]\left\{\frac{[\omega^2 + (k^2 + c^2)]}{[\omega^4 + 2(k^2 + c^2)\omega^2 + (k^2+c^2)^2]}\right\}.$$

Depending on the relative values of k and c, this exhibits either a monotonic decrease with ω or a single maximum occurs, a situation typical of macroclimatological variables. Bendat points out that the more complex actual realizations may be approximated by composites of two functions.

We shall now generalize our model to generate such a series. Given that waves of different wavelength and frequency are arriving at a station in a manner so that the integral effect is a surface of varying amplitude and wavelength, we may assume reasonably that the phases within any narrow bandwidth of frequencies are unrelated. Assuming a uniformly random distribution of phase and invoking the central limit theorem, it may be shown that the amplitudes of the surface will be distributed normally. Consider measurements of the height of an isobaric surface x(t) taken either at random times or at equal intervals of time and assume no standing waves are present. Subtract the mean height and assume the phases ϕ_n of the series are distributed rectangularly

$$x(t) = \Sigma a_n \cos(T_n t - \phi_n)$$

Not only will x(t) have a normal density function with zero mean, but time derivatives x't, x"t, etc., and any groupings of such quantities will also be normal. Thus they can be expressed entirely in terms of the variances and covariances between these quantities.

In terms of the wave model, precipitation amounts should depend on (a) maximum values of the wave form, and (b) time it takes to reach (or descend from) the maximum from (or to) some arbitrary level. This should be a measure of the period and intensity of lower-level convergence and includes as well some notion of thermal stability. The distribution of maxima is somewhat confusing. Rice (1954) and Longuet-Higgins (1957, 1958a, 1958b) both assume a narrow-band spectrum (i.e. sharply peaked) and show that if amplitudes are defined as the difference in height between a crest and the preceding trough, maxima have a Rayleigh density

$$P(a) = \left(\frac{G}{m_0}\right) e^{-0.5a^2/m_0}.$$

In fact, the Rayleigh distribution describes the maxima for any spectrum, provided we refer now to the height of the maxima formed on the envelope of fluctuations, i.e. the curve joining peaks of the

summed sinusoidal series (Sveshnikov, 1966). It is only with a narrow-band series in which the waves assume the shape of a single sinusoid with slowly varying amplitude that the envelope maxima and series maxima become the same to within a constant term. Cartwright and Longuet-Higgins (1958) give the equation for maxima for any normal-series spectrum (meaning heights of maxima above the mean level be made zero), but this seemingly is less useful for our purpose. If we use the Rayleigh density, it is difficult to know exactly what assumptions we are making. No natural series fulfills the strict requirements of the theoretical narrow-band process. On the other hand, the maxima of the envelope may be sufficient for our heuristic reasoning. Consequently, we are not sure if we are assuming a narrow-band spectrum or not. It is of interest that Roden's (1964) application of wave statistics to screen temperatures on the Pacific coast used Rice as a source so oscillations of the westerlies may well be a narrow band.

The Rayleigh density function is of interest because squaring such variables produces a Chi-square function that can be related readily to Gamma-distributed rainfall per seven day period. Summing over a five to seven day period seems to take care of any important autocorrelation; thus such periods can be regarded as independently distributed. Rice shows that the energy content (i.e. the square of deviations from the mean) within a fixed period beginning from an arbitrary time approximates a Gamma distribution.

SUMMARY

We presumably want to "explain" climatic statistics in rigorous terms. How do we do this? Current circulation models of the Smagorinsky type, if they could generate the statistics exactly, do not seem to be geared to answering questions about the statistics. I rather doubt, for example, that a statistician charged with designing a set of rain-making experiments for one or two sites would find a hemispheric general-circulation model which begins with the primitive equations of much value. Nevertheless, it will be useful if the differential equations of physics can be represented in terms of stochastic wave equations and in particular in terms of the transfer functions of linear and nonlinear systems. Bendat (1958) discusses this approach in general terms. Sveshnikov (1966) gives a relevant example where an exponential cosine plus sine autocorrelation function is obtained from white noise input, Y(t), through a system having a "restoring" force $k^2Y(t)$ and a damping force $2hd\ Y(t)/dt$.

A reasonable physical model must consider a two-dimentional surface and, because the earth is small, spherical coordinates are necessary. Its construction will not be easy but should be rewarding. How do standing and moving waves interact? Could the families of storms we recognize be waves beating in phase and cancelling out of phase within a narrow spectrum? The prospect is exciting but our present task is to see how these ideas work out in contact with data.

REFERENCES

Bendat, J. S., 1957, Mathematical analysis and analog simulation of atmospheric turbulence gust velocities: Jour. Aeronautical Sciences, v. 24, p. 69-70.

Bendat, J. S., 1958, Principles and applications of random noise theory: John Wiley and Sons, New York, 431 p.

Cartwright, D. E., and Longuet-Higgins, M. S., 1958, The statistical distribution of the maxima of a random function: Royal Soc. London Proc., v. A247, p. 22-48.

Cochran, W. G., 1938, An extension of Gold's method of examining the apparent persistence of one type of weather: Royal Meteorological Soc. Quart. Jour., v. 64, no. 277, p. 631-634.

Feller, W., 1966, An introduction to probability theory and its applications, v. 2: John Wiley and Sons, New York, 626 p.

Feyerherm, A. M., and Bark, L. D., 1964, Statistical methods for persistent precipitation patterns: Jour. Appl. Meteorology, v. 4, no. 3, p. 320-328.

Gabriel, K. R., and Neumann, J., 1962, A Markov chain model for daily rainfall occurrence at Tel Aviv: Royal Meteorological Soc. Quart. Jour., v. 88, p. 90-95.

Godske, C. L., 1962, Contribution to statistical meteorology: Geofysiske Publikasjoner, Geophysica Norvegica, Oslo, v. 24, no. 5, p. 161-210.

Gold, E., 1929, Note on the frequency of occurrence of sequences in a series of events of two types: Royal Meteorological Soc. Quart. Jour., v. 55, p. 307-309.

Le Cam, L., 1961, A stochastic description of precipitation, in Fourth Berkeley Symp. Math. Statistics and Probability: Univ. of California Press, v. 3, p. 165-186

Longuet-Higgins, M. S., 1957, The statistical analysis of a random, moving surface: Royal Soc. London Phil. Trans., v. A249, no. 966, p. 321-387.

Longuet-Higgins, M. S., 1958a, The statistical distribution of the curvature of a random Gaussian surface: Cambridge Phil. Soc. Proc., v. 54, p. 439-453.

Longuet-Higgins, M. S., 1958b, On the intervals between successive zeros of a random function: Royal Soc. London Proc., v. A246, p. 99-118.

Rice, S. O., 1954, Mathematical analysis of random noise: Bell System Tech. Jour., v. 23 and 24; reprinted in Selected papers on noise and stochastic processes, N. Wax, ed., Dover Publ. Inc., New York, 337 p.

Roden, G. I., 1964, On the duration of nonseasonal temperature oscillations: Jour. Atmospheric Sci., v. 21, no. 5, p. 520-528.

Sveshnikov, A. A., 1964, Applied methods of the theory of random functions (English translation): Pergammon Press, New York, 321 p.

Wiser, E. H., 1965, Modified Markov probability models of sequences of precipitation events: Monthly Weather Rev., v. 93, no. 8, p. 511-516.

ADDITIONAL REFERENCES

Cartwright, D. E., 1962, Analysis and statistics, in Physical oceanography of the sea: Intersci. Publ., New York, p. 567-589.

Caskey, J. B., 1963, A Markov chain model for the probability of precipitation occurrence in intervals of various lengths: Monthly Weather Rev., v. 91, p. 298-300.

Green, J. R., 1965, Two probability models for sequences of wet or dry days: Monthly Weather Rev., v. 93, no. 3, p. 155-156.

Roden, G. I., 1966, A modern statistical analysis and documentation of historical temperature records in California, Oregon and Washington, 1821-1964: Jour. Appl. Meteorology, v. 5, no. 1, p. 3-24.

Thom, H. C. S., 1958, A note on the Gamma distribution: Monthly Weather Rev., v. 86, no. 4, p. 117-121.

Weiss, L. L., 1964, Sequences of wet or dry days described by a Markov chain probability model: Monthly Weather Rev., v. 92, p. 169-176.

IN SEARCH OF GEOLOGICAL CYCLES USING A TECHNIQUE FROM COMMUNICATIONS THEORY

by

Brian W. Carss

University of Illinois

INTRODUCTION

Research Report No. 51 of the Radio Research Laboratory of Harvard University entitled "The spectrum of clipped noise" is an important milestone in a number of disciplines including radio communications, radar, radio astronomy and geophysics. The latest discipline to be added to the list is geology.

J. H. Van Vleck (see Van Vleck and Middleton, 1966), author of Research Report No. 51, was concerned particularly in this project with the calculation of the power-density spectrum of a Gaussian signal (a time series) after it had been clipped. The process of clipping involves chopping off the extreme amplitude values of the time series. Figure 1 is a sequence of drawings showing what happens to a time series if it is clipped. Figure 1A shows the original series; Figure 1B shows the same series clipped at an arbitrary amplitude of +25, -25; and Figure 1C shows extreme clipping where practically all the amplitude information has been destroyed, and only the position of the zero crossings remains. For extreme clipping the resulting wave is rectangular as only two values for the amplitude remain, either +1 or -1.

The real contribution made by Van Vleck (see Van Vleck and Middleton, 1966) in this study was to show how the power-density spectrum of the clipped time series is related to the power-density spectrum of the unclipped series. He showed that the effect of extreme clipping is to make the clipped autocorrleation function $2/\pi$ times the arc sin of the unclipped autocorrelation function

$$R(t_i) = 2/\pi * \text{arc sin } r(t_i) \qquad (1)$$

where $R(t_i)$ is the i th term of the clipped autocorrelation function, and
$r(t_i)$ is the i th term of the unclipped autocorrelation function.

It should be obvious, therefore, that it is possible to calculate the power spectrum of the continuous data from a clipped time series by rearrangement of the equation so that

$$r(t_i) = \sin (\pi/2 * R(t_i)), \qquad (2)$$

and to obtain a normalized and corrected autocorrelation

$$r(t_i) = \sin \left(\pi/2 * \frac{R(t_i)}{R(t_1)}\right). \qquad (3)$$

This means that any Gaussian random time series can be specified by numbers of finite accuracy and a normalized autocorrelation function calculated from these numbers. It is known that less accurate numbers require a greater length of series (more data) to estimate the autocorrelation function with the same accuracy. If this reasoning is carried to the limit, the zero-line crossings of the time series contain enough information to calculate an autocorrelation. Hence, frequency information then can be derived from the power-density spectrum. It is this frequency information that is of particular interest to geologists.

DESCRIPTION OF TECHNIQUE

Until recently geological data have been predominantly qualitative or descriptive. Many times, however, it is possible to make a generalized interpretation, say of the environment of deposition of a rock. It is possible also to quantize generalized interpretation by some ordinal scale. For example, an alternating sequence of limestone, shale, sandstone and coal was deposited during Pennsylvanian time in many parts of the world. It is assumed that some limestones were deposited under marine conditions and the coal under quasimarine (terrestrial?) conditions. It is therefore relatively simple to quantize a description of a section of Pennsylvanian rocks. The rock is either terrestrial (coal) and may be coded +1, or marine (limestone, shale or sandstone) and may be coded -1.

A portion of the lithological log (Lumsden, 1961) of the Archerbeck Borehole, Canonbie, Dumfriesshire, Scotland was processed after being quantized by means of the technique of polarity coincidence correlation as just described. Rocks of Carboniferous age from the base of the Catsbit Limestone to top of the Glencartholm Volcanic Beds in the Archerbeck Borehole (about 3,450 feet) consist of a sequence of alternating marine limestone, shale, sandstone and coal. This sequence (Fig. 2) represents a continuous distribution of marine to nonmarine environments. The environments, defined by coal, sandstone, shale, etc., range arbitrarily within this continuous distribution. Consequently it is difficult to place a boundary precisely between two adjacent environments. On the other hand, it is possible to define broader classes such as marine and nonmarine. In doing this a continuous distribution has been reduced to a two-state discrete form.

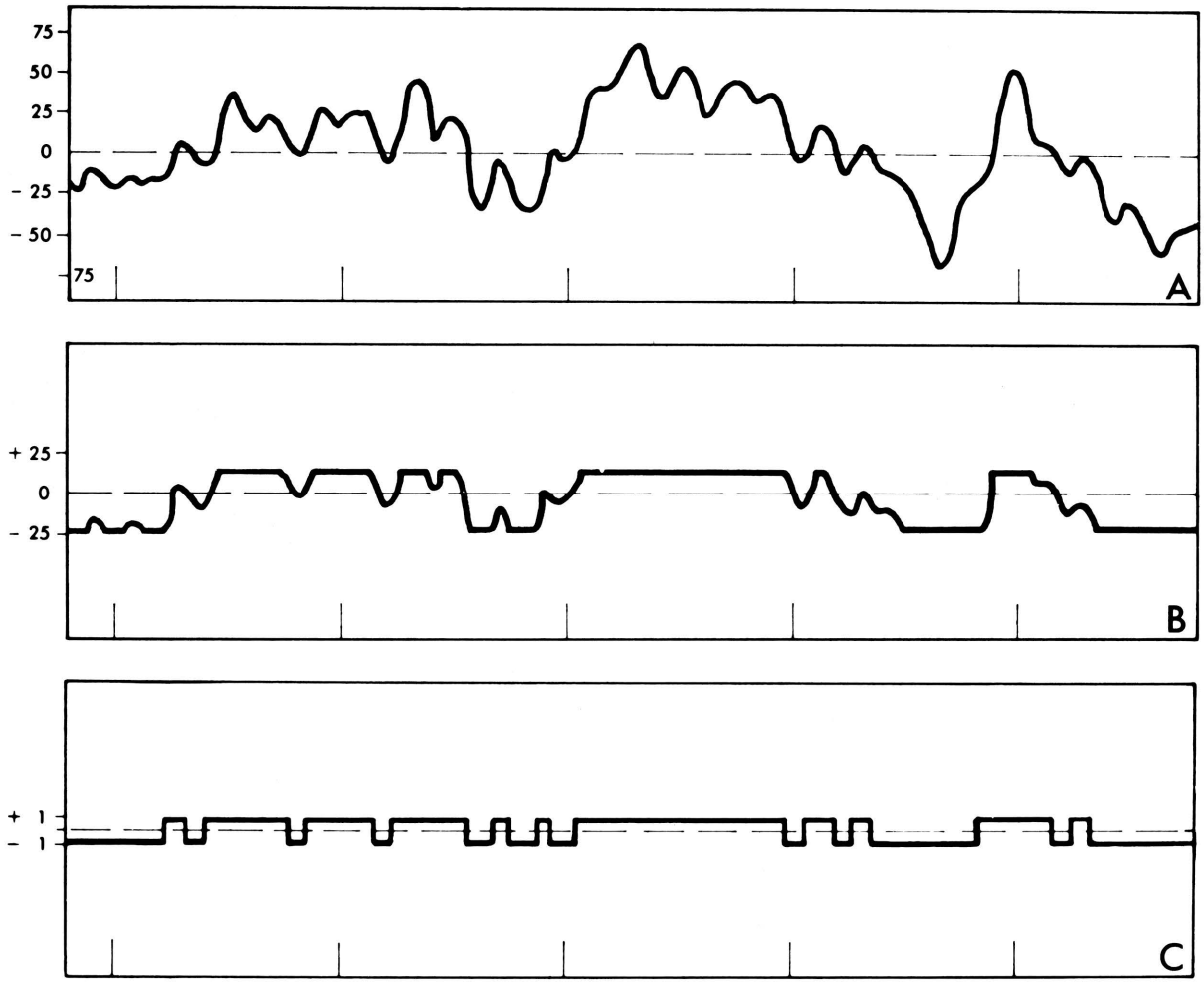


Figure 1.- A, unclipped time series; B, time series clipped at arbitrary level; C, time series with extreme clipping, all amplitude values reduced to +1 or -1.

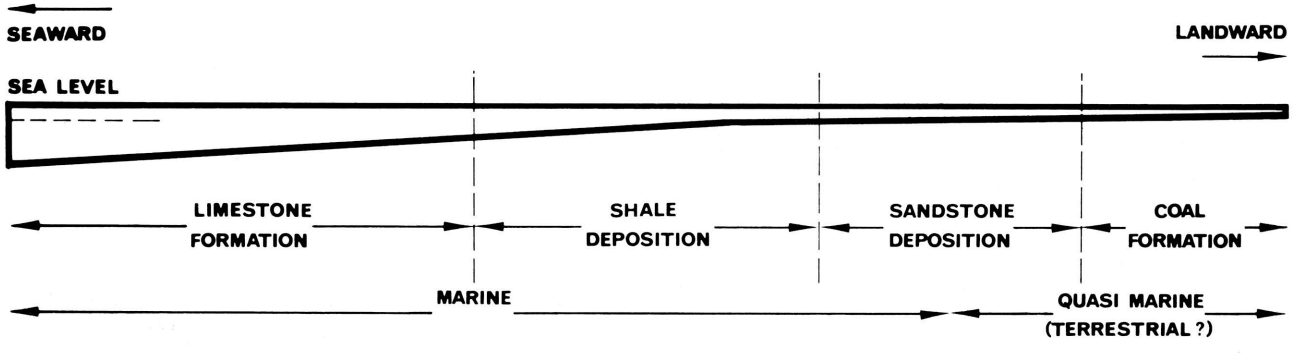


Figure 2.- Diagrammatic horizontal succession of environments for Carboniferous sedimentation.

The borehole data was considered as a time series ordered on depth. Each 6-inch interval was assigned a value of +1 or -1 according to whether the environment was judged to be nonmarine or marine. In all, there were 6,900 data points. Criteria used to define nonmarine or quasimarine conditions were the presence of coal, plant material in situ, seatearth or roots in place.

Implicit in applying this technique of analysis to the data are two assumptions; first, that the statistics of the continuous distribution of environment occurrence do not depart too far from those of a Gaussian distribution; and second, that the borderline between marine and the quasimarine conditions is near the mean value of the continuous distribution. These conditions seemed not to be unreasonable on the basis of a priori knowledge.

An autocorrelation function was calculated from the two-state (+1, -1) data and corrected to the normalized autocorrelation for continuous data using equation (3). The normalized autocorrelation is shown in Figure 3. The first zero crossing in Figure 3 gives an approximate indication of the quarter wavelength of the dominant frequency. In this instance the wavelength of the fundamental (dominant) frequency is about 145 feet. The quarter wavelength criterion would be exact if the original data had been a sine wave.

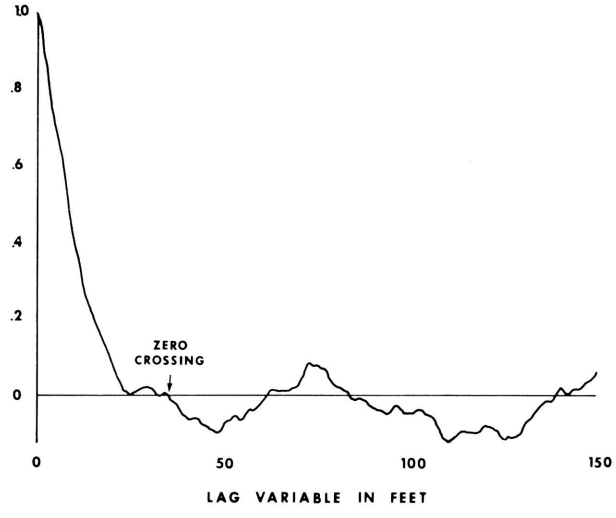


Figure 3.- Normalized autocorrelation function corrected for 1-bit quantization for Carboniferous lithologic data, Archerbeck Borehole.

The normalized autocorrelation function was smoothed using a cosine or hanning window (Blackman and Tukey, 1958), and then a Fourier transform was performed on this smoothed, corrected, normalized autocorrelation using the following relationship to compute the power-density spectrum:

$$S(t) = \sum_n a_n \cos \frac{2\pi nt}{T_o} + \sum_n b_n \sin \frac{2\pi nt}{T_o}$$

SUMMARY OF CALCULATION PROCEDURE

Step 1. Remove any trend from the time series by subtracting the mean value from each sample point, or by fitting a best line in the least-squares sense.

Step 2. Calculate the clipped autocorrelation function.

Step 3. Correct and normalize the clipped autocorrelation function.

Step 4. Generate a hanning or cosine window.

Step 5. Smooth the normalized autocorrelation by multiplying it point for point with the cosine window.

Step 6. Carry out Fourier transform.

Step 7. Calculate logarithms of the independent estimates.

Step 8. Plot the resulting power-density spectrum.

INTERPRETATION OF THE POWER-DENSITY SPECTRUM

The power-density spectrum, as shown in Figure 4, has a background shape such that the Gaussian assumption is reasonable. Recurrent peaks (Neidell, 1966) are present at about equal frequency intervals. Nine of these peaks have been labelled on Figure 4 by letters of the alphabet (Table 1).

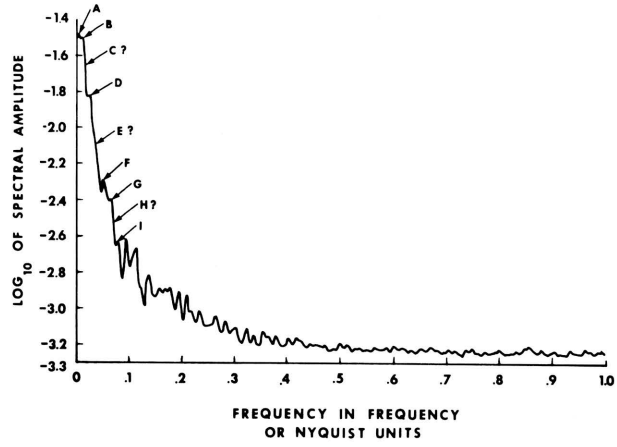


Figure 4.- Power spectrum estimated from Carboniferous lithologic data, Archerbeck Borehole.

The harmonics of a 145 feet fundamental are:

Fundamental	145.0	6	20.7
1	72.5	7	18.2
2	48.3	8	16.2
3	36.2	9	14.5
4	29.0	10	13.2
5	24.2	11	12.1

Table 1.- Nine peaks with associated wavelengths compared with harmonics of a fundamental wavelength of 145 feet.

Peak	Approximate wavelength, feet	Harmonic indicated
A	200	Fundamental
B	71.3	1
C	(?)	2
D	37.5	3
E	25	4, 5, & 6
F	19.3	7
G	15.4	8 & 9
H	(?)	10
I	12.8	11

(?) denotes position uncertain, but presence indicated for comparison.

There is no reason to expect all harmonics to be present, or in a particular proportion. It is possible for some averaging to take place for example at E. The presence of many harmonics can be taken as a good indication of cyclicity rather than periodicity.

GEOLOGICAL INTERPRETATION OF CYCLICITY

Both Johnson (1962) and Westoll (1962) state that the thickness of cyclothems in northern England ranges from 90-150 feet. Westoll also records an average thickness of 100 feet for a complete cycle. The fundamental cycle detected in the Archerbeck Borehole agrees well with the thickness recorded for this type of sedimentation. Using rates of sedimentation from 1,000-3,000 years per foot as Westoll did, the Archerbeck Borehole cycle would have a fundamental time span from 145,000-435,000 years. Average rates of sedimentation are not too meaningful because the rate of deposition of any sedimentary sequence is not linear with time. It seems, therefore, that an interpretation which is not linked to time would offer a more satisfactory explanation.

Dunham (1950) suggested that small isostatic readjustments take place in the earth's crust as sediments accumulate. The thickness of the fundamental would represent then the maximum thickness of sediment that the crust can support by its own strength. Once this thickness has been exceeded, the strength of the material is exceeded and isostatic readjustment takes place.

SEARCH FOR CYCLICITY IN LIMESTONES

A section of 1,846 feet of limestone belonging to the Bird Spring Group (Pennsylvanian-Permian) in the Arrow Canyon Range, Nevada, was studied by Heath (1965) and Lumsden (1965). They were interested particularly in the petrography of the microfacies found in the Bird Spring limestones.

Their interpretation of environmental conditions again showed a continuous sequence of sediment type ranging from a calcisiltite to an oolitic calcarenite (Fig. 5). The calcisiltite was interpreted as being deposited under quiet marine conditions, with poor circulation, and the oolitic calcarenite deposited under agitated conditions. The boundary between the nonagitated and the agitated sediments coincides approximately with wave base.

Both workers presented curves showing an interpretation of how the environmental energy varied with time. These curves were quantized. In quantizing these curves, microfacies 3, 4, and 5 were coded +1; and 0, 1, and 2 were coded as -1 at a sampling interval of 6 inches. The same calculating procedure was used as before and the power-density spectrum plotted. Analysis of the power-density spectrum points to the existence of cyclicity. The sequence of peaks (Fig. 6) labelled A through K are given also in Table 2. It became necessary to compute broadband spectral estimates as well as narrower ones in order to detect the fundamental. From the several autocorrelation functions calculated, the fundamental was estimated to be approximately 400 feet.

Table 2.- Sequence of peaks with corresponding wavelengths and indicated harmonics.

Peak	Approximate wavelength, feet	Harmonic indicated
A	80	5
B	64 (?)	6
C	50	8
D	33.3	12
E	25.0	16
F	18.2	22
G	16.6	25
H	12.1	33
I	10	40
J	9	44
K	8	50

INTERPRETATION OF LIMESTONE CYCLICITY

Schwarzacher (1964) has detected a cyclical phenomenon in a group of Lower Carboniferous limestones and shales in Ireland. The wavelength of the cycle that he detected is 9.85 feet (300 cms.) which agrees with the 40th harmonic in the present study. In another study on the Lower Carboniferous in the north of England, Schwarzacher (1958) gives a cycle wavelength of approximately 30 feet. This would

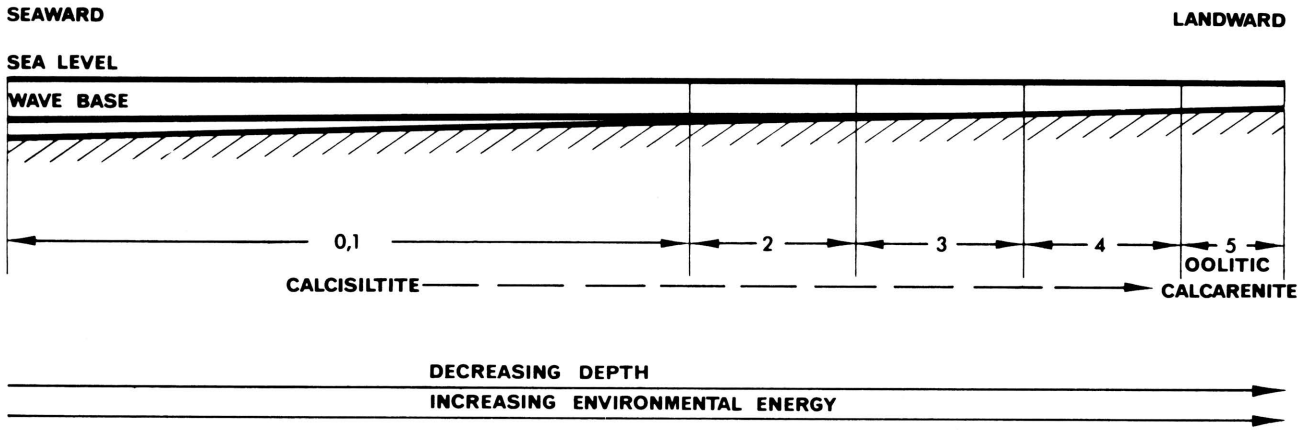


Figure 5.- Horizontal interpretation of microfacies relationships.

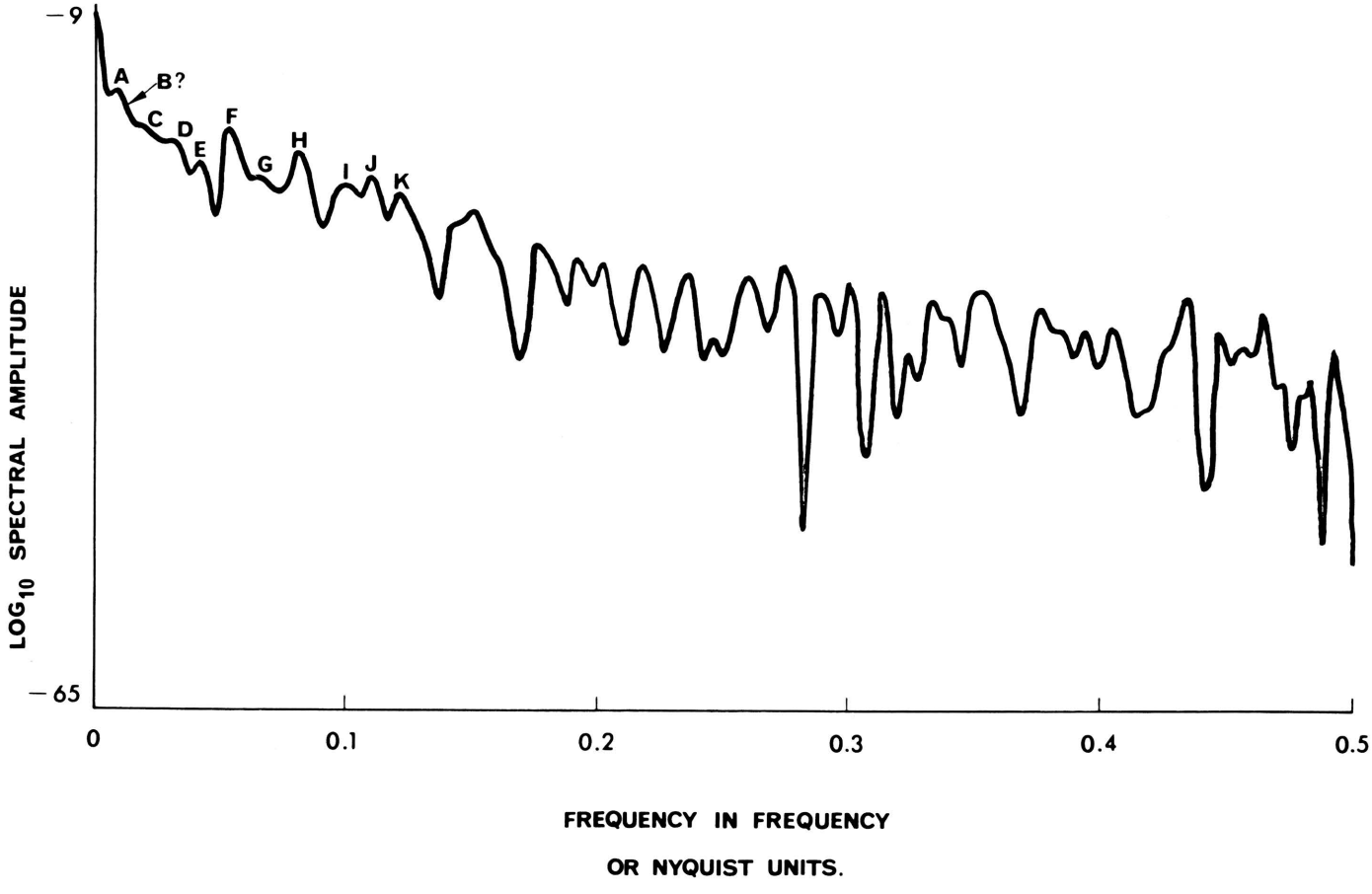


Figure 6.- Power spectrum estimated from Bird Spring Group data.

correspond to the 12th harmonic of the present study.

It is not possible to suggest any definite mechanism for cyclical sedimentation phenomena, but it is possible to make comparisons between the number of cyclothems in the midwestern United States for a comparable interval. There are 37 complete cycles in the Arrow Canyon section and 36 cyclothems in the midwestern United States. A coincidence perhaps?

CONCLUSIONS

There are too many recorded occurrences of cyclical phenomena to be any doubt about the existence of cycles, but very little, if anything, is known about the mechanism that causes them. With the development of techniques such as the one described here, however, more data will be accumulated and from this hopefully an understanding will result.

REFERENCES

Blackman, R. B., and Tukey, J. W., 1958, The measurement of power spectra: Dover Publ., New York, 190 p.

Dunham, K. C., 1950, Lower Carboniferous sedimentation in the Northern Pennines (England): 18th Internat. Geol. Congress Rept., pt. 4, p. 46-62.

Heath, C. R. P., 1965, Microfacies of the lower Bird Spring Group (Penn-Permian), Arrow Canyon Range, Clark County, Nevada: Unpubl. doctoral dissertation, Univ. Illinois.

Johnson, G. A. L., 1962, Lateral variation of marine and deltaic sediments in cyclothemic deposits with particular reference to the Visean and Namurian of northern England: 4th Internat. Congress Strat. and Carboniferous Geol., Heerlen 1958, Tome 2, p. 323-330.

Lumsden, D. N., 1965, Microfacies of the middle Bird Spring Group (Penn-Permian) Arrow Canyon Range, Clark County, Nevada: Unpubl. doctoral dissertation, Univ. Illinois.

Lumsden, G. I., 1961, Appendix A, in The stratigraphy of the Archerbeck Borehole, Canonbie Dumfriesshire: Bull. Geol. Surv. Gt. Brit., no. 18, p. 17-46.

Neidell, N. S., 1966, Spectral studies of marine geophysical profiles: Geophysics, v. 31, no. 1, p. 122-134.

Schwarzacher, W., 1958, The stratification of the Great Scar Limestone in the Settle District of Yorkshire: Liverpool and Manchester Geol. Jour., v. 2, p. 124-142.

Schwarzacher, W., 1964, An application of statistical time-series analysis of a limestone-shale sequence: Jour. Geology, v. 72, no. 2, p. 195-213.

Van Vleck, J. H., and Middleton, D., 1966, The spectrum of clipped noise: Proc. Inst. Electrical and Electronics Engineers, v. 54, no. 1, p. 2-19.

Westoll, T. S., 1962, The standard model cyclothem of the Visean and Namurian sequence in northern England: 4th Internat. Congress Strat. and Carboniferous Geol., Heerlen 1958, Tome 3, p. 767-773.

QUALITY AND QUANTITY OF AVAILABLE GEOLOGIC INFORMATION FOR STUDIES IN TIME

by

P. H. A. Sneath

University of Leicester (UK)

Studies that involve time are found often to make exacting demands on the quality and quantity of data that is needed. This discussion will not be restricted to time series in the narrow sense, but will consider also other problems in which there is a time element.

Experiences in numerical taxonomy and medical diagnosis have shown that usually a large quantity of data is required to be reasonably sure of getting a satisfactory result. This stems largely from the difficulty of selecting the best data for answering a given question. There are two approaches: (1) one may choose the minimum amount of data of the "right kind," based on some simple hypothesis or model, or (2) one uses much data, in the hope that the "right kind" of data can be sifted effectively from the mass. It is particularly difficult in complex biological fields to formulate the hypotheses required for the first approach. We may defeat therefore our own ends if we try to restrict the data to a minimum. Also, up to a point one may compensate for poor quality if the quantity of available data is increased, and this is important if the cost of improving the quality is prohibitive. These considerations apply to many geological problems.

My colleagues D. F. Merriam and M. J. Sackin and I are interested in studies on sequences of sedimentary rocks, which of course involves the time element. We have been developing a method of cross-association, which is useful in fields as diverse as molecular biology of proteins and sedimentation (Sackin, Sneath, and Merriam, 1965). The method is analagous to cross-correlation, or rather to cross-multiple correlation because it considers many variables simultaneously. Essentially it consists of sliding one sequence past another, and counting the matches between the two sequences at every position of trial. The matches may arise from comparing many variables of strata that are opposite each other. We are experiencing difficulties, as one might expect, especially with different thicknesses of strata in different sections, and thus have used only sequences of rock types in rank order, ignoring thickness, and taking an arbitrary criterion of what is a single rock type. The method is capable in theory of finding the geological correlation between rock sections, of detecting insertions and deletions in the sections and of revealing periodic or cyclic phenomena such as cyclothems. In preliminary work (Sackin, Sneath, and Merriam, 1965; Merriam and Sneath, 1967) we used only a few descriptors of rock types, and though encouraging, the results suggested that more information is required. Ideally we would like to put into the computer the same information the geologist uses, and in some respects we may be able to approach this ideal.

We have been looking, therefore, into the quantity and quality of information in coded sections. Most of this has been done with the idealized case, viz. the two sections are from the same spot and are identical, but further studies of sections from different localities are in progress. The example shown here is taken from the Kansas City Group, Middle Pennsylvanian (Jewett, Hornbaker, and Press, 1967). Thirty rock units were recorded in average or better than average detail, and 36 descriptors were found to be reasonably accurate and complete (Tables 1 and 2). A number of subsections of different size were chosen to investigate how the results depend on the number of units in an unknown section that is being compared with a standard section. Cross-association studies were run on these. From the results we can answer a number of questions, and these answers may well hold true for much material of this type. The number of descriptors available for a typical section as ordinarily recorded by geologists is evidently large. Here there were 36 relevant descriptors, but the number levels off as the number of strata is increased (thus there are about 10 descriptors for any one stratum taken at random, and five strata require nearly 30 descriptors). This increase in descriptors depends primarily on new features of the rocks that differentiate them. Thus, a new fossil type introduces a new descriptor capable of differentiating some rocks, and as more strata are added, more fossil types in general will be added. We would expect that for long sections there would be about twice as many relevant descriptors as there were strata, but this figure no doubt depends mostly on the diligence of the investigator.

In the section studied about 3/4 of the descriptors were recorded and applicable to any one stratum. The remaining 1/4 is due to lack of precise recording, and also to certain combinations of features that are logically incompatible (though careful coding of features should keep this to a minimum; the principles of coding of data of this type are given in Sokal and Sneath, 1963, p. 74-79). In this section,

Table 1.- Strata used in study.

Stratum No. in Section	Rock Type	Member	Formation
29	Limestone	Westerville	Cherryvale Shale
28	Limestone	Westerville	Cherryvale Shale
27	Limestone and Shale	Westerville	Cherryvale Shale
26	Limestone	Westerville	Cherryvale Shale
25	Shale	Westerville	Cherryvale Shale
24	Limestone	Westerville	Cherryvale Shale
23	Shale	Wea	Cherryvale Shale
22	Limestone	Wea	Cherryvale Shale
21	Shale	Wea	Cherryvale Shale
20	Limestone	Wea	Cherryvale Shale
19	Shale	Wea	Cherryvale Shale
18	Limestone	? Block	Cherryvale Shale
17	Shale	Fontana	Cherryvale Shale
16	Limestone and Chert	Winterset	Dennis Limestone
15	Limestone	Winterset	Dennis Limestone
14	Shale	Winterset	Dennis Limestone
13	Limestone	Winterset	Dennis Limestone
12	Shale	Winterset	Dennis Limestone
11b	Limestone	Winterset	Dennis Limestone
11a	Limestone	Winterset	Dennis Limestone
10	Shale	Winterset	Dennis Limestone
9	Limestone	Winterset	Dennis Limestone
8	Shale	Winterset	Dennis Limestone
7	Limestone	Winterset	Dennis Limestone
6	Shale	Stark	Dennis Limestone
5	Shale	Stark	Dennis Limestone
4	Shale	Stark	Dennis Limestone
3	Shale	Galesburg	Galesburg Shale
2	Shale	Galesburg	Galesburg Shale
1	Limestone	Bethany Falls	Swope Limestone

also, the probability of a match between any two strata on a given descriptor is 72.9 percent. That is, observed matching greater than this percentage indicates better-than-expected agreement, and a lower match indicates poorer matching than expected. The degree of better (or worse) matching is given by our computer program in standard deviations (SD) above (or below) the expected match.

As an experiment, a number of short subsections were run against the complete (reference) section. The number of strata in the subsections, m_1, varied from 1 to 30 (the latter was the complete section itself). There was, for each subsection, one position where it gave a perfect match against the reference section. The proportion of matches was then, of course, 1·0, and the number of descriptors that matched, summed over all the relevant strata, increased linearly with m_1, as one would expect. Thus for $m_1 = 1$ there were about 28 matches, for $m_1 = 5$ about 135 matches, while for $m_1 = 30$ there were 832 matches. The SD's do not increase linearly, of course; thus for $m_1 = 1$, matching 28/28 is about 6·07 SD above the expected proportion of about 73 percent, whereas 832/832 corresponds to 31·60 SD above the expected. The statistical significance increases much faster than linearly, but we do not wish to put a literal interpretation on the astronomic probabilities that correspond to figures like 30 SD's.

Of more consequence is the distribution of the SD values that correspond to positions that are not the position of perfect match. The most interesting are positive SD values, which might lead to misleading conclusions on the correct matching positions between different sections or subsections, because these might be mistaken for near-perfect match positions. In our example there was little likelihood of such confusion (Table 3). The next highest SD's were always well below the perfect-match SD, even for subsections that consisted of only three adjacent strata. Indeed, the results suggest that it may be possible to match with some confidence a single stratum correctly in a reference column, when 30 or more descriptors are available. The remarkable facility of expert geologists to identify a single stratum at an outcrop without knowing the exact location or the beds above and below is no less remarkable for knowing that one stratum does, very likely, contain enough information for this feat to be possible. Yet the results suggest that we may be able to increase the geologist's powers by supplementary computer methods.

We do not have enough data to study in detail the distribution of the SD's of the "next best" matches. It is clear that this is not simple mathematically, and at present we are simply studying it empirically, by tabulating observed SD values and by simulation studies. Present results suggest that the "next best" SD's seldom exceed 4·0, whereas the perfect match SD is usually above 9 even for a subsection of only three strata. Of course in a more realistic study, where the sections being compared are not identical, the perfect match SD would not be so high, but there is evidently a considerable margin. Strong cyclic or periodic structure would raise also the "next best" SD's. The negative SD's indicate worse than expected matching, for example if many shales are positioned opposite limestones. As yet we do not understand the significance of these, but it may be noted that negative values of over -4·0 are not uncommon (Table 3).

Confirmation of the general conclusions noted has come from simulation studies, where the order of strata in the subsequences has been randomized. We do not now expect any position of perfect match. The highest SD values are mostly below 4·0 (Table 4). In passing, one may note that as m_1 is increased the chance of high "false match" SD's will become less. The longer sections will generate more match positions, however, and more SD values will be tabulated, so there will be a greater chance of an extreme value arising "by chance." In the present study these opposing effects almost cancel out; whether they will do so in general is not known. Reversing the order of strata entirely ("reverse matching") behaved like a randomized order.

It may be mentioned that the main type of

Table 2.- Descriptors used for section.

No.	Descriptor	Coding (0=unrecorded or not applicable)
1	Major rock type	1=limestone, 2=limestone and shale, 3=shale
2	Fresh surface, color depth	1=light, 2=medium, 3=dark, 4=black
3	Fresh surface, color shade	1=white, 2=gray, 3=pink, 4=buff, 5=yellow-brown
4	Weathered surface, color depth	1=light, 2=medium, 3=dark, 4=black
5	Weathered surface, color shade	1=white, 2=gray, 3=pink, 4=buff, 5=yellow-brown
6	Secondary rock	1=No shale, 2=Shale present
7	2° rock fresh, color depth	1=light, 2=medium, 3=dark
8	2° rock fresh, color shade	1=white, 2=gray, 3=pink, 4=buff, 5=yellow-brown
9	2° rock weathered, color depth	1=light, 2=medium, 3=dark
10	2° rock weathered, color shade	1=white, 2=gray, 3=pink, 4=buff, 5=yellow-brown
11	Other components: argillaceous	1=none, 2=argillaceous material present
12	calcareous bands	1=none, 2=present
13	silty bands	1=none, 2=present
14	chert present	1=none, 2=present
15	calcite present	1=none, 2=sparry, 3=crystalline in fossils
16	phosphate nodules	1=none, 2=present
17	Texture	1=dense, 2=poorly consolidated
18	Fracture	1=conchoidal or speudoconchoidal
19	Fissile, for shales	1=nonfissile, 2=fissile or splintery
20	Bedding, spacing	1=thin, 2=medium, 3=blocky, 4=massive
21	Bedding, regularity	1=regular, 2=irregular
22	Bedding, wavyness	1=flat, 2=wavy, 3=cross-bedded
23	Vertical joints, spacing	1=close, 2=distant
24	Undulated upper surface	1=flat, 2=undulated, 3=pitted
25	Prominent ledge formed	1=no, 2=yes
26	Nodular or lenticular bed	1=no, 2=yes
27	Dark band in bed	1=no, 2=yes
28	Limestone type	1=amorphous, 2=oolitic, 3=crinoid, 4=algal 5=fine crystalline
29	Fossil abundance	1=nonfossiliferous, 2=sparse, 3=abundant, 4=fossil "hash"
30	Specific fossils recorded: crinoids	1=no, 2=yes
31	bryozoans	1=no, 2=yes
32	snails	1=no, 2=yes
33	clams	1=no, 2=yes
34	brachiopods	1=no, 2=yes
35	Composita	1=no, 2=yes
36	echinoderms	1=no, 2=yes

rock in the subsections (shale or limestone) had little influence in any of the above effects. The section showed no marked evidence of cyclic or periodic structure, which can be detected in principle by cross-association, but it contained only one cyclothem and a small part of another (Merriam, 1963) so this is not unexpected. In this connection the study of Schwarzacher (1964) is interesting. We have not been able to make much use of the sum of Chi-square statistic referred to in an early study (Merriam and Sneath, 1967).

Clearly we need more work on this problem, but its implication is obvious. A standard computer coded geological section for areas could be made based on careful recording (and no doubt re-recording) in standard format (such as printed record-sheets). Cross-correlation methods are capable of great sensitivity, and one can illustrate this by the remarkable results of tree-ring dating, e.g. Fritts (1963) where correct dating was obtained that could not be detected by eye, and where distances up to 1,000 miles may nevertheless give significant correlations. Similar results are being obtained with varves (Anderson and Kirkland, 1966).

One approach that might be tried would be to regard the rock at unit distances up the section as a series of vectors, each vector being given by the descriptors of the rock at that point. In principle,

Table 3.- Highest positive and highest negative match standard deviations from cross-associations of different subsections compared with the complete section of Table 1 ("forward matches" only); m_1 indicates number of strata in subsection.

Subsection (strata numbers)	m_1	Perfect match SD	Next three highest positive SD's			Three highest negative SD's			Main rock type
24	1	6·07	1·74	1·33	0·97	-2·78	-1·52	-1·11	Limestone
5	1	5·39	2·43	2·43	2·43	-4·08	-2·06	-1·24	Shale
23-25	3	9·55	2·31	2·07	1·45	-3·03	-2·45	-1·53	Shale
4- 6	3	9·50	1·68	1·29	1·16	-4·07	-3·54	-2·60	Shale
22-26	5	12·39	2·87	2·65	2·18	-2·98	-2·93	-2·14	Limestone
3- 7	5	12·57	2·12	1·76	1·44	-3·12	-2·58	-2·33	Shale
19-28	10	18·39	3·09	1·82	1·06	-3·24	-3·18	-3·15	Limestone
2-11a	10	17·96	1·61	1·47	1·16	-3·26	-3·18	-2·63	Shale
1-14	15	22·04	1·27	1·18	1·09	-3·42	-3·35	-3·25	About equal
1-29 (complete section)	30	31·60	1·34	1·34	0·99	-3·39	-3·31	-3·21	About equal

Table 4.- Highest positive and highest negative match standard deviations for subsections with order of strata randomized, if compared by cross-association against complete section in its original order of strata.

Randomized subsection (strata numbers before randomizing)	m_1	Three highest positive SD's			Three highest negative SD's			Main rock type
22-26	5	4·28	2·67	2·14	-2·91	-2·44	-2·27	Limestone
3- 7	5	6·07	2·28	2·15	-3·91	-3·12	-2·88	Shale
19-28	10	2·36	1·75	1·71	-3·39	-3·00	-1·99	Limestone
2-11a	10	3·24	2·10	2·08	-2·70	-2·66	-2·57	Shale
1-29 (complete section)	30	2·80	2·61	2·38	-3·79	-3·73	-3·57	About equal
1-29 (complete section)	30	2·46	2·20	2·04	-3·08	-2·52	-2·09	About equal

therefore, one would have a continuous series of vectors with respect to height (and by implication also with respect to time). One might then look for specified types of movement of the vector tips, from given hypercubes to other hypercubes, or from hyperspheres, or for specified angular movements of the vectors. This might be more useful in some data than transition matrices. Any vector model, however, poses acute problems on the proper weights to be given to each of its dimensions. Comparison between cross-association and vector models should prove interesting.

If one turns from detailed multivariate data of this sort to the equally detailed but predominently univariate (or oligovariate) data of things like electrical logs the nonspecialist like myself is reminded of the puzzles that are set by electro-encephalogram records of brain activity. One does not doubt that there is much information in them, but it is not clear how it can be extracted. It may be that we need new kinds of information about rocks--and the rocks must surely be full of information as yet unguessed. Anything that could help toward determining synchronicity in rocks would be of the greatest theoretical and practical interest. With exception of varves and volcanic ash bands, we have little here as yet. It is the privilege of the amateur to make uninformed, even outrageous, suggestions in the hope that one or other may point a new pathway. Clearly we await new technical advances in petrology. Nevertheless there are possibilities, if remote, that may be considered. Recent volcanic eruptions seemingly are detectable by delicate ash bands in deep sea sediments (e.g. Kuenen, 1950; Nayudi, 1964); would sensitive analytic methods detect them in chalks, limestones, and shales?

The secular and long term vectors of remnant magnetism perhaps could be subjected to cross-correlation (or cross-vector) studies if paleomagnetic techniques could be improved so that they could be performed quickly on small rock samples. Oxygen isotope studies of ancient temperatures perhaps could be exploited further. Sea-level changes might be dissociated from local land mass movements if we had sufficient comparative data on a world-wide scale.

Lastly, it might be instructive to apply to geological studies the multidimensional scaling methods of Sheperd and Kruskal. An example in an archaeological context is given by Hodson, Sneath, and Doran (1966) where good agreement was found between time and complex patterns. This technique makes few assumptions about the relation between variables (whether linear, curvilinear, etc.) and has shown itself to be flexible and reasonably sensitive.

REFERENCES

Anderson, R. Y., and Kirkland, D. W., 1966, Interbasin varve correlation: Geol. Soc. America Bull., v. 77, no. 3, p. 241-256.

Fritts, H. C., 1963, Computer programs for tree-ring research: Tree-Ring Bulletin, v. 25, no's. 3-4, p. 2-7.

Hodson, F. R., Sneath, P. H. A., and Doran, J. E., 1966, Some experiments in the numerical analysis of archaeological data: Biometrika, v. 53, no's. 3-4, p. 311-324.

Jewett, J. M., Hornbaker, A. L., and Press, J. E., 1967, Inland underground facilities division of Beatrice Foods Co.: Guidebook, 3rd national forum on the geology of industrial minerals, Kansas Geol. Survey, 13 p.

Kuenan, P. H., 1950, Marine geology: John Wiley and Sons, New York, 551 p.

Merriam, D. F., 1963, The geologic history of Kansas: Kansas Geol. Survey Bull. 162, 317 p.

Merriam, D. F., and Sneath, P. H. A., 1967, Comparison of cyclic rock sequences using cross-association, in Essays in paleontology and stratigraphy, R. C. Moore Comm. vol.; Teichert, C., and Yochelson, E. L., eds.: Univ. Kansas Dept. Geol. Spec. Publ. 2, p. 523-538.

Nayudi, Y. R., 1964, Volcanic ash deposits in the Gulf of Alaska and problems of correlation of deep-sea ash deposits: Marine Geol., v. 1, no. 3, p. 194-212.

Sackin, M. J., Sneath, P. H. A., and Merriam, D. F., 1965, ALGOL program for cross-association of nonnumeric sequences using a medium-size computer: Kansas Geol. Survey Sp. Dist. Publ. 23, 36 p.

Schwarzacher, W., 1964, An application of statistical time-serial analysis of a limestone-shale sequence: Jour. Geol., v. 72, no. 2, p. 195-213.

Sokal, R. R., and Sneath, P. H. A., 1963, Principles of numerical taxonomy: W. H. Freeman and Co., San Francisco and London, 359 p.

COMPARISON OF SUBSET TREND SURFACES BY UTILIZATION OF INFORMATION THEORY

by

S. V. L. N. Rao

and

G. S. Srivastava

Indian Institute of Technology and National Mineral Development Corporation

ABSTRACT

A detailed discussion on different methods employed for the comparison of polynomial trend surfaces is outlined. The trend and residual components are treated by modern analytical methods that are used currently in communication theory, as the signal and noise content of spatial-spectral series.

INTRODUCTION

In the computation of trend-surfaces, it is necessary to know the degree needed and to evaluate the goodness of fit of the determined polynomial surface. At present evaluation and interpretation of trend maps is subjective rather than statistical. An element of empiricism points to the need for a collaborative attempt by geoscientists and statisticians to apply statistical theories more rigorously in evaluation of trends (Krumbein and Graybill, 1965).

Krumbein and Graybill advocated use of analyses of variance and confidence intervals around the fitted surface. Wilks (1963) indicated the need for a "two sample validity cross check" procedure that aids in evaluation of the regression techniques. Agterberg (1964) applied this concept of using two subsets and by subtracting one set of trend contours from the other, he showed that it is possible to obtain satisfactory maps that portray the actual regional trends in this manner. He showed, however, only the "difference trend maps" and did not indicate procedure for computation. Because evaluation of trends involve comparison of polynomial surfaces, emphasis in this study is placed on critical evaluation of present methods for the comparison of the trend surfaces.

The technique of trend-surface analysis can be viewed either on the basis of statistical theories or on the basis of information theory. It can be thought of as a statistical procedure (a method of variance analyses), where the observed parameters can be decomposed into trend and residual components. If the trend and residual components are treated as the signal and the noise contents of a spatial-spectral series (Tobler, 1966), it is possible to apply certain modern analytical methods that are used currently in communication theory.

Rao and Rao (in press) used trend-surface analysis as a technique for filtering the local fluctuations (noise) in grain size distributions. In this way the relation between the mean and the degree of sorting of bed-load sediment in a curved channel becomes more meaningful. Rao (in press) approached the problem of grading in a magnesite deposit on the basis of information theory; trend-surface analysis is followed by use of grade specifications as "numeric filters" in the grading of a homogeneous mineral deposit.

In 1966 Tobler indicated the possibility of applying time-series concepts to spatial data directly. In a personal communication regarding the comparison of trend-surface maps of subsets, W. R. Tobler mentioned cross-correlation analysis and distance function techniques. In an important contribution, Merriam and Sneath (1966) detailed a procedure for quantitative comparison of contour maps. They adapted two techniques (correlation analysis and taxonomic distance) for preparing dendrograms. This was followed by comments from Mandelbaum (1966) and a reply by the authors. One of the earliest contributions on comparisons of trend surfaces is by Miller (1964). He used the surfaces themselves or the residuals for comparison. Miller (1964) says: "The problem of devising quantitative standards for comparing contour maps is very complicated. Difficulties appear at the very onset, in asking just what is to be compared."

Tobler (personal communication, 1967) essentially agrees with the stated view and says "...the question of comparison of trend maps is very difficult. The question is clearly of considerable importance, especially for the case, in which one wishes to compare a theoretical surface with an empirical surface." In the present study, the authors computed two sets of trend maps from the same region to estimate parameters by the following methods:

1. Use of variance analyses techniques.
2. Comparison of error measures as obtained from the Davis and Sampson program (1966).
3. Computation of root mean square error measure.

4. Computation of distance function using a modified Mandelbaum (1966) procedure.

5. Use of serial and cross-correlation functions.

Methods (3) and (5) are justified by viewing the data as part of a spatial-spectral series (an analogue of time series).

DETAILS OF COMPUTATION

The test data were taken from a region near Salem, Madras State. The region represented in Figure 1 is dunite intrusive for which a series of specific-gravity determinations were made by D. K. Shrivastava. The determinations were by the pycnometer method and Metlar's monopan balance. It is estimated that the values are accurate to ±0.005. The samples were divided into two sets as indicated (Fig. 1).

Trend surfaces up to and including the fourth degree were fitted by the Davis and Sampson (1966) program. This program operates in single precision, so that the coefficients of determination of higher degree trends may give a smaller index than higher degree trends. In one instance the authors corrected ill-conditioned matrices by either increasing the mantissa length or by removal or addition of some samples. By and large, however, this effect was ignored.

Krumbein and Graybill (1965) showed that variance analysis can be used to evaluate the "strength" of the trend. The results of variance analysis are given in Table 1. The analysis indicates that the F value is significant up to the third degree in both subsets. This finding corroborates conclusions derived from the serial and cross-correlation coefficients.

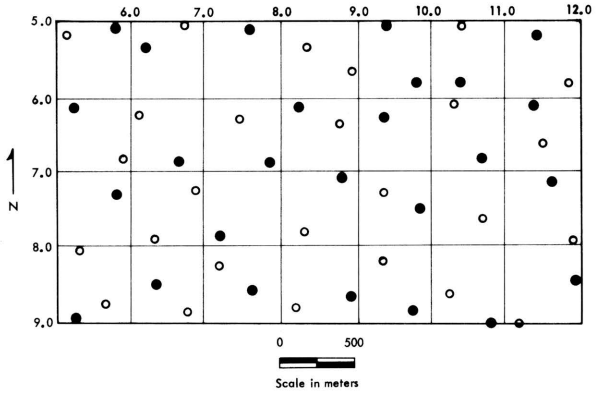


Figure 1.- Sampling pattern, ● Subset A and ○ Subset B, in dunite outcrop of Chalk Hills, Salem area, Madras State.

Table 2 shows the error measures obtained by processing the data with the Davis and Sampson program. With one exception, the error measures for both subsets are in substantial agreement.

Table 1.- Variance analysis and results.

Source	Sum of squares	Degrees of freedom	Mean square	F
Subset A				
Due to linear	45.09	2	22.04	3.15
Dev. from linear	167.66	24	6.98	
Due to quad.	55.01	3	18.34	3.45
Dev. from quad.	112.65	21	5.32	
Due to cubic	32.33	4	8.08	1.71
Dev. from cubic	80.32	17	4.72	
Due to quartic	9.53	5	1.90	.32
Dev. from quartic	70.78	12	5.90	
Total	212.75	26		
Subset B				
Due to linear	58.03	2	29.01	3.61
Dev. from linear	168.68	24	7.02	
Due to quad.	39.40	3	13.13	2.28
Dev. from quad.	129.28	21	6.15	
Due to cubic	67.81	4	16.95	4.69
Dev. from cubic	61.47	17	3.61	
Due to quartic	- 13.19	5	- 2.64	.43
Dev. from quartic	74.66	12	6.22	
Total	226.70	26		

Another method for comparing the subset trend maps is to compare the dependent variable as computed for fixed grid-point locations. Because the equations for the four surfaces are known, values for the dependent variable can be generated. In this study the origin was taken at the northwest corner of the map. For a fixed value of the vertical coordinate, 71 values of Y were computed along the east-west line at grid intervals of 0.1. These values were generated for each surface, then one set was subtracted from the other. To annul the effect of algebraic sign, the difference is squared and then summed along the grid line. The mean value is obtained by division by 71, and the square root of the summed value is computed. Next, the value of the vertical coordinate is incremented by 0.1 and a similar set of values is obtained for the second row. Thus summation was determined for 41 rows to cover the entire mapped area. The root mean square values for all rows are again summed and divided by the number of rows to obtain the total root mean square error for the entire map. This is equivalent to the generation of two signals along particular directions, and computing the root mean

square error, as is usually done in time-series analysis. Some representative data computed for the entire map are given in Table 3. The error measure spread

Table 2.- Comparison of error measures.

		Subset A	Subset B
Standard deviation:	first degree	0.2364	0.2499
	second degree	0.1938	0.2189
	third degree	0.1636	0.1509
	fourth degree	0.1536	0.1663
Variation explained by surface:	first degree	0.4509	0.5803
	second degree	1.0010	0.9743
	third degree	1.3244	1.6524
	fourth degree	1.4197	1.5205
Variation not explained by surface:	first degree	1.6766	1.6868
	second degree	1.1264	1.2928
	third degree	0.8032	0.6146
	fourth degree	0.7077	0.7465
Total variation:		2.1275	2.2671
Coefficient of determination:	first degree	0.2119	0.2559
	second degree	0.4705	0.4297
	third degree	0.6225	0.7289*
	fourth degree	0.6673	0.6707

*Anomalous value discussed in text.

at these different intersections are graphically represented in Figure 2. This diagram indicates that in selected regions the RMS error for the third-degree surface is lower than the linear and quadratic surfaces (about a value of 6.00). The quadratic surface has in general a lower value than the linear within region 5.5 to 7.5. The quartic surface indicates a higher RMS value over the entire region. This row-by-row scanning procedure discloses certain trends that are not obvious from a study of the total RMS value, which generally shows a progressive increase as surfaces with higher terms are computed.

As a corollary to the computation, one can obtain the difference value (including the sign) at these 2,911 fixed grid points. These differences can be used as a dependent variable to determine a polynomial function of appropriate degree, and for each surface, one can obtain the "error polynomial surface." If the surface is flat and devoid of large-scale undulations, the flatness of the surface can be taken as a comparative measure. This phase of computation was not done here.

Table 3.- Determination of root mean square error measures.

Line no.*	Linear surface	Quadratic surface	Cubic surface	Quartic surface
5.00	0.0176	0.0247	0.0342	0.0338
5.50	0.0155	0.0163	0.0128	0.0147
6.00	0.0135	0.0161	0.0099	0.0140
6.50	0.0119	0.0088	0.0124	0.0165
7.00	0.0107	0.0099	0.0134	0.0149
7.50	0.0103	0.0117	0.0148	0.0118
8.00	0.0106	0.0129	0.0162	0.0116
8.50	0.0116	0.0139	0.0156	0.0146
9.00	0.0131	0.0151	0.0157	0.0286

Total root mean square value using 41 x 71 = 2,911 grid locations

Linear	0.0124
Quadratic	0.0130
Cubic	0.0149
Quartic	0.0160

*Scaled value along western (vertical) margin.

The authors also have used computation procedures outlined by Merriam and Sneath (1966) for the determination of the distance function between the two subsets. As Mandelbaum (1966) mentions, it is not possible to adopt the Merriam and Sneath procedure for comparison of two trend surfaces as they will have no standard deviation. No weighting has been assigned for two reasons. Because both samples are drawn from the same population and equal numbers of samples are used, the given weightings cancel each other. The authors feel that Mandelbaum's (1966) suggestion of a method of assigning weights will make the calculation procedure biased.

Procedures adopted for distance-function calculation have yielded the following generalized distances:

Between linear surfaces of Subset A and Subset B	1.035
Between quadratic surfaces of Subset A and Subset B	0.596
Between cubic surfaces of Subset A and Subset B	24.903
Between quartic surfaces of Subset A and Subset B	6.846

These results do not correspond to conclusions derived from other functions. An inspection of the trend surfaces indicate that the high value of the distance function can be accounted for by the constant term in the equations for the third degree which differ greatly. This is also the case with the fourth-

degree distance function (polynomial surface equations are given in the Appendix). In these polynomial equations the constant term usually will be large, followed by lesser and lesser values for the higher degree and cross-product terms. In usual calculations of distance functions, this situation does not arise. Parks (1966) adopted a normalized data matrix in similar situations so that equal weighting was given to all variables. Following Mandelbaum's comment, Merriam and Sneath did not recalculate the distance functions, so the effect of adding the constant term to the computations could be evaluated.

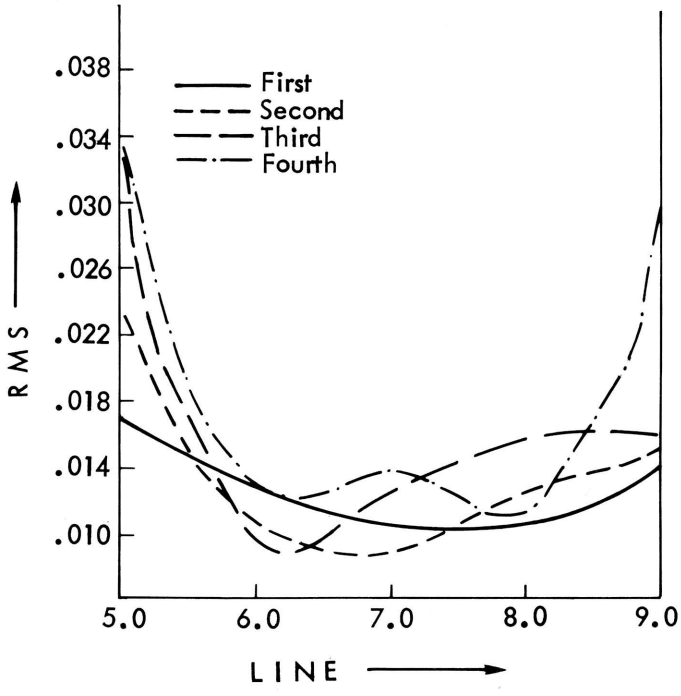


Figure 2.- Error analysis.

The authors are of the opinion that some specific investigations are needed to evaluate the efficiency of distance functions in comparing contour maps. Based on Tobler's (1967) suggestion, the authors have computed a series of cross- and serial-correlation coefficients at specified locations. The computation of serial- and cross-correlation coefficients requires the determination of residuals at every grid location. Because the data are irregularly spaced, it is possible to obtain the residuals. The interpolated value at any specified location may be obtained by triangulation or by the polygonal method. Trend values can be computed and by subtraction, the residuals can be found at all grid locations needed for computation. It is possible also to interpolate observed values in any specific direction along a line which cuts across some sampling sites and then use Lagrangian or other functions to obtain the interpolated values along the cross section. For this cross section, the residuals can be obtained which are used for the computation of serial- and cross-correlation coefficients. The authors, however, followed a different method. They subtracted one surface from the other and used these values as residuals. For the four polynomial surfaces of each set, the following six sets of residuals were obtained.

Set 1 residuals by subtracting 1st from 2nd degree
Set 2 residuals by subtracting 1st from 3rd degree
Set 3 residuals by subtracting 1st from 4th degree
Set 4 residuals by subtracting 2nd from 3rd degree
Set 5 residuals by subtracting 2nd from 4th degree
Set 6 residuals by subtracting 3rd from 4th degree

Thus for both sets, we get 12 sets of residuals. The serial-correlation graphs and cross-correlation graphs are machine plotted along the east-west direction at the 7.5 vertical coordinate. As generally noted the serial- or cross-correlation coefficient initially will have a maximum value which falls rapidly as lag increases. It eventually becomes equal to 0 and then attains a negative value. The lag values at which the serial- or cross-correlation coefficients become 0 are given in Table 4. In the serial-correlation, residuals of sets 4 and 10 have minimum lag. These are obtained by the subtraction of the second degree from the third degree surface. The cross-correlation coefficients show a minimum lag for sets 4 to 10 and 6 to 12. These represent residuals obtained by subtraction of the second-degree surface from the third degree and the third degree from the fourth-degree surface.

These results corroborate the results obtained by variance analyses, that the third-degree surface

Table 4.- Serial- and cross-correlation coefficients.

Serial-Correlation

Subset A		Subset B	
Set No.	Lag value at which S.C. is zero	Set No.	Lag value at which S.C. is zero
1.	16.	7.	16.
2.	12.	8.	10.
3.	16.	9.	11.
4.	6.	10.	6.
5.	9.	11.	7.
6.	17.	12.	12.

Cross-Correlation

Set No.	Lag value at which C.C. is zero
1-7	17.
2-8	14.
3-9	13.
4-10	5.
5-11	8.
6-12	4.

almost fully represents trends in this mapped region. It may be mentioned that the structural correlation coefficient of Mirchink and Bukhartsev (1959) represents a cross-correlation coefficient for a curviplanar surface.

CONCLUSIONS

The authors have discussed the various methods available for the comparison of polynomial surfaces from the viewpoint of statistics and communication theory. This paper is confined to polynomial surfaces, which are widely used. Trigonometric functions, such as are used by Rayner (1967) and others, are not explicitly included.

ACKNOWLEDGMENTS

The authors thank Dr. T. C. Bagchi, Head of Geology and Geophysics Department, I. I. T., Kharagpur, for his continued encouragement. They also wish to express their gratitude to Dr. D. F. Merriam, Dr. John C. Davis, and Mr. Robert Sampson of the Kansas Geological Survey; Dr. W. R. Tobler of the University of Michigan; and Mr. A. Roy and Mr. A. Maity.

REFERENCES

Agterberg, F. P., 1964, Methods of trend surface analysis: Colorado Sch. Mines Quart., v. 59, no. 4, pt. A, p. 111-130.

Davis, J. C., and Sampson, R. J., 1966, FORTRAN II trend surface fitting program with unrestricted input for the IBM 1620 computer: Kansas Geol. Survey Sp. Dist. Publ. 26, 12 p.

Krumbein, W. C., and Graybill, F. A., 1965, An introduction to statistical models in geology: McGraw-Hill Book Co., New York, 475 p.

Mandelbaum, H., 1966, Comments on a paper by Daniel F. Merriam and Peter H. A. Sneath, 'Quantitative comparison of contour maps': Jour. Geophysical Res., v. 71, no. 18, p. 4431-4432.

Merriam, D. F., and Sneath, P. H. A., 1966, Quantitative comparison of contour maps: Jour. Geophysical Res., v. 71, no. 4, 1105-1115.

Miller, R. L., 1964, Comparison-analysis of trend maps: Stanford Univ. Publ., Geol. Sci., v. 9, no. 2, p. 669-685.

Miller, R. L., and Kahn, J. S., 1962, Statistical analysis in the geological sciences: John Wiley and Sons, New York, 483 p.

Mirchink, M. F., and Bukhartsev, V. P., 1959, The possibility of a statistical study of structural correlations: Doklady Acad. Nauk SSSR (English transl.), v. 126, no. 5, p. 1062-1065.

Parks, J. M., 1966, Cluster analysis applied to multivariate geologic problems: Jour. Geology, v. 74, no. 5, pt. 2, p. 703-715.

Rao, S. V. L. N., in press, Theory of grading a homogeneous economic mineral deposit: Jour. Geol. Soc. India.

Rao, S. V. L. N., and Rao, C. N., in press, Study of the grain size distribution in a curved channel: Utilization of trend surfaces as a cleaning device: Sedimentology.

Rayner, J. N., 1967, Correlation between surfaces by spectral methods: Kansas Geol. Survey Computer Contr. 12, p. 31-37.

Tobler, W. R., 1966, Spectral analysis of spectral series: paper delivered at 4th Ann. Conf. on Urban Planning, Information System and Programs, Berkeley, California.

Wilks, S. S., 1963, Statistical inference in geology, in The earth sciences problems and progress in current research: Univ. of Chicago Press, p. 105-136.

APPENDIX

Equations for Subset A

1st Degree

$$Z = 1.8943793 + 0.0454694\,X + 0.0650713\,Y$$

2nd Degree

$$Z = 7.5151330 - 0.5193206\,X - 0.8897343\,Y + 0.0212037\,X2 + 0.0287213\,XY + 0.0501776\,Y2$$

3rd Degree

$$Z = -0.3216236 + 1.7521444\,X + 0.0768555\,Y - 0.3261589\,X2 + 0.2051868\,XY - 0.2147443\,Y2 + 0.0107994\,X3 + 0.0100178\,X2Y - 0.0241175\,XY2 + 0.0224141\,Y3$$

4th Degree

$$Z = 25.1280760 - 2.7416320\,X + 10.1476370\,Y + 0.3899428\,X2 + 0.4462897\,XY + 1.9949123\,Y2 - 0.0507990\,X3 + 0.0387204\,X2Y + 0.1033560\,XY2 - 0.1667794\,Y3 + 0.0004221\,X4 + 0.0063774\,X3Y - 0.0133861\,X2Y2 + 0.0143222\,XY3 + 0.0029754\,Y4$$

Equations for Subset B

1st Degree

$$Z = 1.8914432 + 0.0035416\,X + 0.1126443\,Y$$

2nd Degree

$$Z = 5.8800829 - 0.6126081\,X - 0.3174948\,Y + 0.0291434\,X2 + 0.0174408\,XY + 0.0195861\,Y2$$

3rd Degree

$$Z = -30.2970910 + 5.5425986\,X + 8.5133967\,Y - 0.6156923\,X2 - 0.1950124\,XY - 1.1544107\,Y2 + 0.0188511\,X3 + 0.0222879\,X2Y - 0.0114727\,XY2 + 0.0605195\,Y3$$

4th Degree

$$Z = 17.9726880 - 5.4009023\,X - 7.0069909\,Y + 1.5115974\,X2 - 0.5946450\,XY + 2.5174640\,Y2 - 0.1394446\,X3 - 0.0079996\,X2Y + 0.0860757\,XY2 - 0.3357144\,Y3 + 0.0039719\,X4 + 0.0030459\,X3Y - 0.0036486\,X2Y2 - 0.0017753\,XY3 + 0.0148537\,Y4$$

SEDIMENTARY LAMINATIONS IN TIME-SERIES STUDY[1/]

by

R. Y. Anderson

University of New Mexico

INTRODUCTION

Laminations are alternations of textures or components of sediments (i.e. clay, silt, sand, organic matter, carbonate, sulfate, halite, etc.). The simple repetitions represent the basic or shortest term aspects of sedimentary processes. Because the time involved in their formation is on the order of minutes or seasons, the processes are apt to persist for intervals of time that are sufficient to produce series of laminations that lend themselves to time-series methods. The persistent repetition of a few components is a distinctive characteristic of laminations; this is a more distinctive characteristic than the generally accepted arbitrary size limit of 1 cm (Kelley, 1956). Something can be learned of the shorter and longer term aspects of related processes through careful examination of the laminae and the repetitious patterns in which they are arranged. Laminations also form under a variety of conditions. In this report we are concerned only with those laminae that were deposited in quiet saline or unoxygenated basins where the organisms that normally mix and destroy laminations have been excluded. Under these conditions, the laminae tend to be well preserved in long, fairly continuous series.

SEDIMENT "RAIN"

In quiet basins, the sediments "rain" down on the floor of the basin from above. Chemical and organic constituents are derived generally from the water mass itself as precipitates or flocculates. The allocthonous fraction is mainly a pelagic clay that also can be considered to settle as a "rain" of particles. The continuity of sediment "rains" is an aspect of sedimentology that has been studied only slightly but the concept of a more or less geographically uniform "rain" of sedimentary particles has been established through stratigraphic correlations of individual laminae for many kilometers (Anderson and Kirkland, 1966).

ASSOCIATION WITH TIME

Two time-related concepts, first applied to ecologic changes (Clements and Shelford, 1939), are the basis for interpreting the significance of laminations that are formed as the result of sediment rains. The first, called aspection, involves changes from season to season. The production or deposition of each component usually has defineable limits in time. That is, a particular component (i.e. the carbonate layer in the Rita Blanca varves; Anderson, in press) is produced or deposited within a known season or some other period of time.

The other time concept (annuation, or changes from year to year) relates to the frequency of production or influx of a particular laminae-forming component. A range of frequencies and patterns from precisely annual and seasonal to random and non-seasonal is possible but some generalizations about different types of laminae can be made. In glacial varve series, for example, where the controlling mechanism is fairly well understood (Antevs, 1925, 1951), the deposition of laminae is seasonal (aspection) and at a frequency that is nearly annual (annuation). In many nonglacial varve series a clastic (clay) component may accumulate with only a slight seasonal variation in quantity. A second component such as a layer of sapropel or plankton bloom may settle in a strongly seasonal manner and "set off" the clay layer into annual laminae. In other instances, multiple blooms or precipitates may occur in a single season or entire years may be skipped.

The lamination process may be complicated further if clastic materials larger than clay size (sand, silt, tuff, pumice, woody fragments, etc.) are involved. The frequency of influx of these components tends to be more erratic or random than for the finer grained materials (Anderson, 1964), and they may be associated with other laminae with different time relationships. Hence, in the Oligocene Florissant laminated series, seasonally blooming diatomite was interlaminated with a more constantly accumulating sapropel and erratically deposited graded tuff and inversely graded pumice laminae (McLeroy and Anderson, 1966).

In all these examples, the aspection (changes from season to season) concept is used to help interpret the environmental significance of a laminae or component. The concept of frequency (annuation) is used to calibrate the rate of production or accumulation of the components.

[1/] This work was sponsored by the earth sciences program of the National Science Foundation grant AG 922, GP-4200 and earlier grants. I wish to thank Walter E. Dean, Jr. for his help in developing some of the ideas in this study.

CALIBRATED STUDIES

Calibration (the counting and measuring of laminae) makes it possible to sample and analyze stratigraphic samples on a time-series basis. This, in turn, permits the study of the "true" associations of major and minor constituents as well as the study of other phenomena that are unrelated to the environment of the laminated sequence. It leads also to more detailed petrographic interpretations based upon lamina by lamina correlation.

Component Association

Conventional stratigraphic sampling is based upon either standard thickness, standard volume, or changes in physical appearance. It is difficult to avoid interpretations that assume that the rate of deposition is relatively constant. Hence, a change from shale to limestone is assumed to be a time of increased carbonate deposition when, in fact, a reduction in both clay and carbonate deposition might be equally plausible. It is impossible to determine the true associations of several components if only changes in proportion are known.

In time-series sampling, however, the number of laminae is held constant and the thickness or volume of the sample interval varies. Percentage estimates of components based upon standard volume within the sampling intervals (relative values) are converted to a quantity-per-unit time basis (absolute values) by obtaining the product of the thickness or volume of the sampling interval and the percentage estimate. The result is an independent time series for each component, expressed on a quantity-per-unit time basis, that may be used to determine the time associations. The method is applicable whether the frequency of deposition is known or not because the laminations are assumed to be the result of a regular process. In the situation of known or proven varve series, this assumption is the most valid and interpretations the most reliable.

Two laminated series illustrate the method. The Permian Castile Anhydrite consists of organic-rich calcite laminae alternating with nearly pure anhydrite laminae (some carbonate grains are mixed in with the anhydrite). Plots of carbonate and sulfate on a relative (%), and "absolute" basis (Fig. 1) show the difference in association for each sampling technique. The relative (%) plots naturally show an inverse association of carbonate and sulfate because there are only two major components; note also that percent carbonate declines in the last 1,000 years.

The "absolute" plot, however, shows that the two components have a positive association ($r = +0.56$, $n = 60$, first 600 units) and the amount of carbonate actually deposited holds fairly steady during the last 1,000 years. This positive association is not constant throughout the series and there are times when the association is negative. There are also places in the Castile series where sulfate deposition stops completely with little or no change in carbonate deposition and the two parameters appear to behave independently. Hence, the derived time series is a more informative record than percentage estimates derived from ordinary stratigraphic sampling.

"Absolute" time series were constructed also for the clay and carbonate components in the Rita Blanca lake deposits of Texas (Pleistocene). Smoothed plots of the two independent 1,400-year series show a strong negative association of the two components (Fig. 2). The correlation coefficient, however, did not substantiate the interpretation of a strong negative association (-0.012; $n = 51$). As for the Permian Castile example, associations change with time, and in this instance it was suspected that the strong negative association was cancelled by positive associations elsewhere in the series. Bivariate spectral analysis was used to determine the association for different frequencies of change and a comparison of correlation and coherence methods, using the Rita Blanca series as an example is discussed by L. H. Koopmans in this Colloquium.

The phase angles derived from the bivariate analysis (see Koopmans, Fig. 3) revealed that the clay-carbonate association was negative for frequencies greater than 70 years and positive for the shorter frequencies. Environmentally, this was interpreted to mean that some of the carbonate was washed in with the clay from the caliche soil, probably as a result of runoff from local storms or melting snow. But another form of carbonate was precipitated in the basin in response to long-term changes in temperature, evaporation or precipitation. This information, if combined with biologic and petrographic data was the basis for interpreting air-mass movements that alternately brought in cool-moist or warm-dry conditions (Anderson and Koopmans, in press).

Stratigraphic Correlation and Association

Evaporitic, organic, and some pelagic clay laminations persist for great distances and lend themselves to stratigraphic correlation techniques. Laminae by laminae stratigraphic correlations can be located and demonstrated by sliding a short "slave" series along a "master" series until a significant correlation has been obtained (Anderson and Kirkland, 1966). In practice, however, several statistically significant correlations will be obtained, and although the true correlation generally has the highest value, correlation is determined more easily by direct visual comparison. Where a correlation is known to exist but cannot be located visually, the sliding method will extend the range of correlation, but this is based upon interpreting the same number of laminae in each series, which is not always possible.

Stratigraphic correlation has proved to be more useful as a method for determining the degree

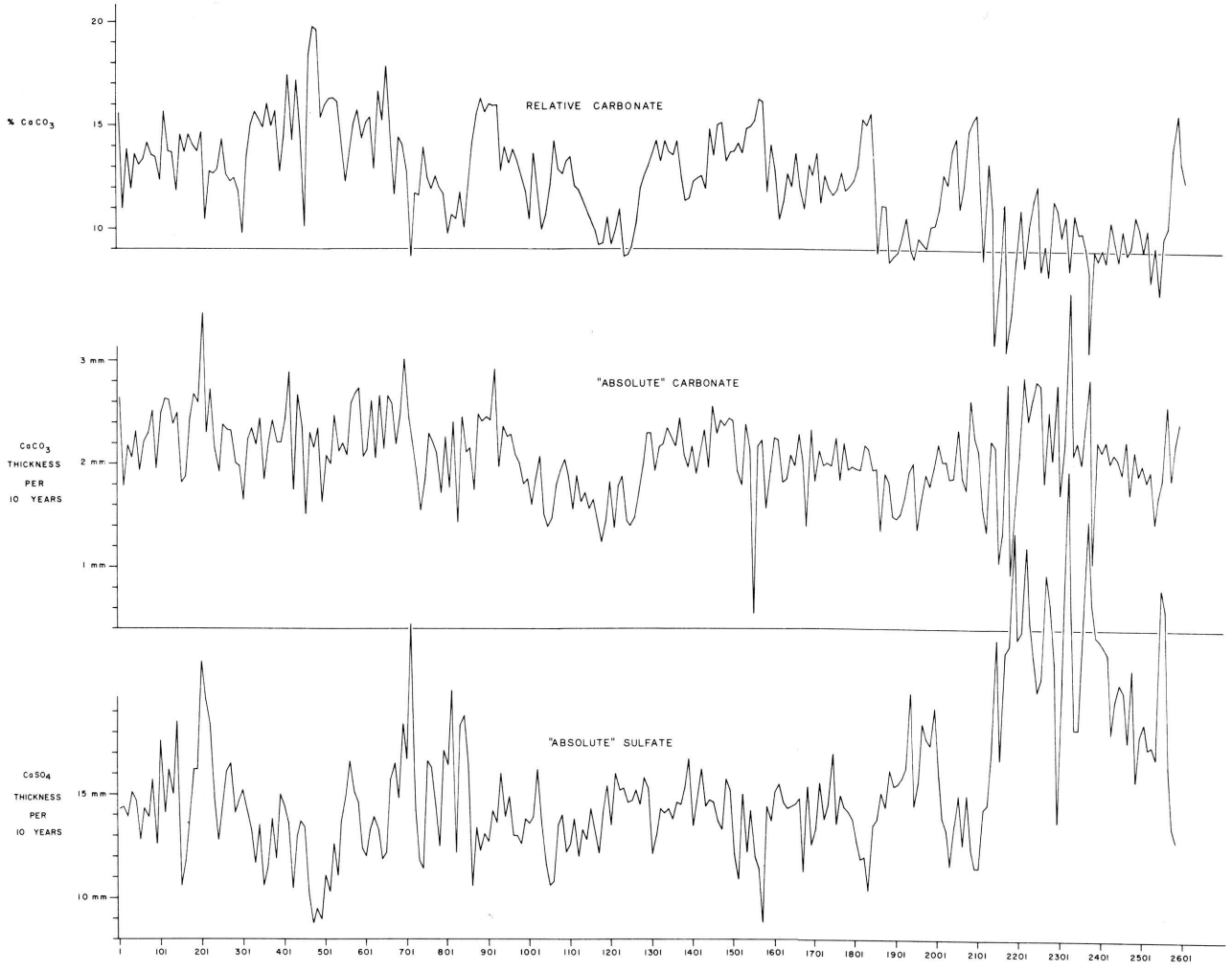


Figure 1.- "Relative" and "absolute" time series of $CaSO_4$ and $CaCO_3$ in Permian Castile Anhydrite. Samples were collected and analyzed on 10-unit basis.

of association of correlative series of laminae. In this way it was possible to infer that certain stratigraphic sections in the Permian Zechstein and Jurassic Todilto Formations were probably disturbed by currents throughout deposition and had lost some of their continuity as a record of precipitated material (Anderson and Kirkland, 1966).

A simple moving correlation technique can be applied to correlative stratigraphic sections to determine the degree of lateral continuity in different parts of the same series. After two correlative sections have been measured and interpretive problems about the number of laminae resolved (same n in each series), a moving correlation for some arbitrarily selected length of data can be computed. A plot of such a series will show the zones of greatest lateral continuity.

The moving correlation technique may be applied also to associated but different parameters in the same varve series. In the Permian Castile series, a moving correlation (110-year interval) revealed alternate times of positive and negative association of carbonate and sulfate on a quantity-per-unit-time (absolute) basis (Fig. 3). The same figure also shows the changing association of sulfate, carbonate, and organic matter with time (Dean, 1967). This technique is a valuable adjunct to ordinary correlation studies and to bivariate spectral analysis.

Lamina by lamina correlation has proved also to be a valuable petrographic tool. Comparison of two or more correlative thin sections may reveal stages in the processes of reorganization and recrystallization and allows for the recognition and separation of local and regional characteristics. Correlation also makes possible a comparison of normal and structurally deformed textures and probably has a good many other applications.

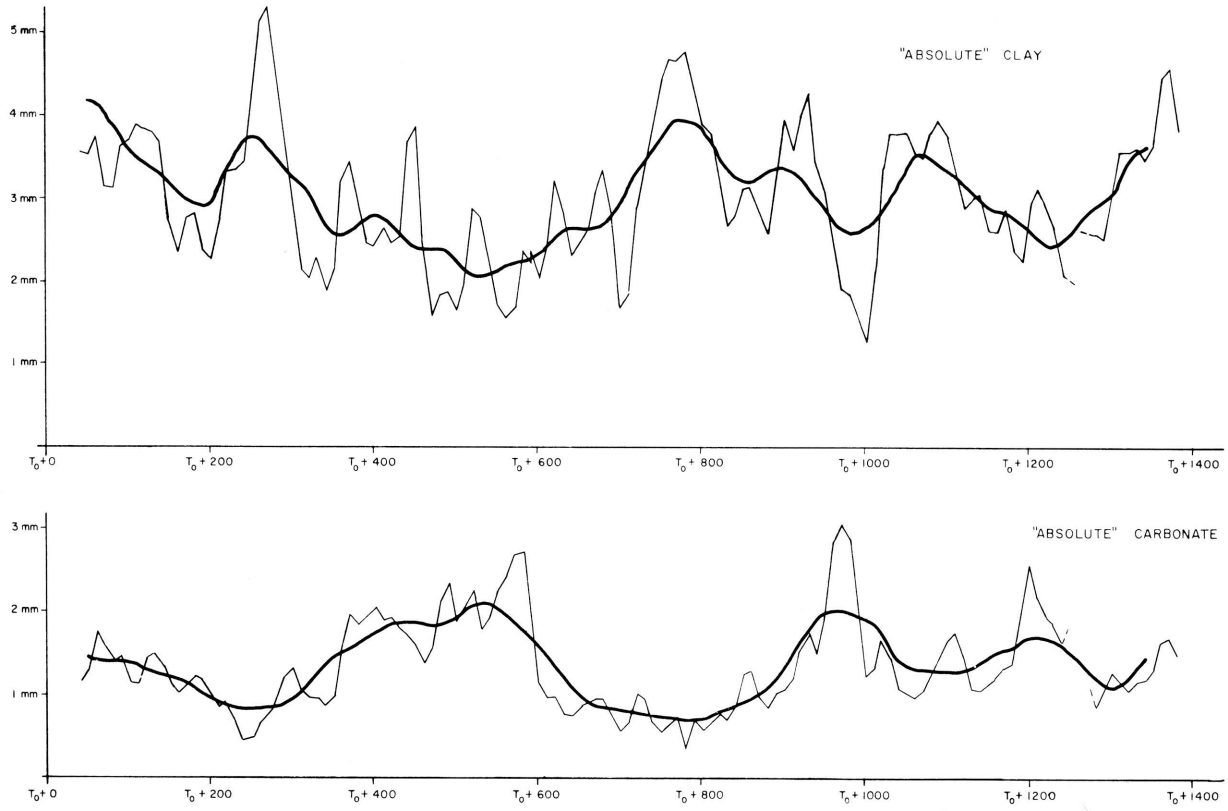


Figure 2.- Variation in quantity of clay and calcium carbonate deposited in Early Pleistocene Rita Blanca varve time series. Note strong negative association between clay and carbonate in smoothed curve. Thickness in mm per 5-year sample.

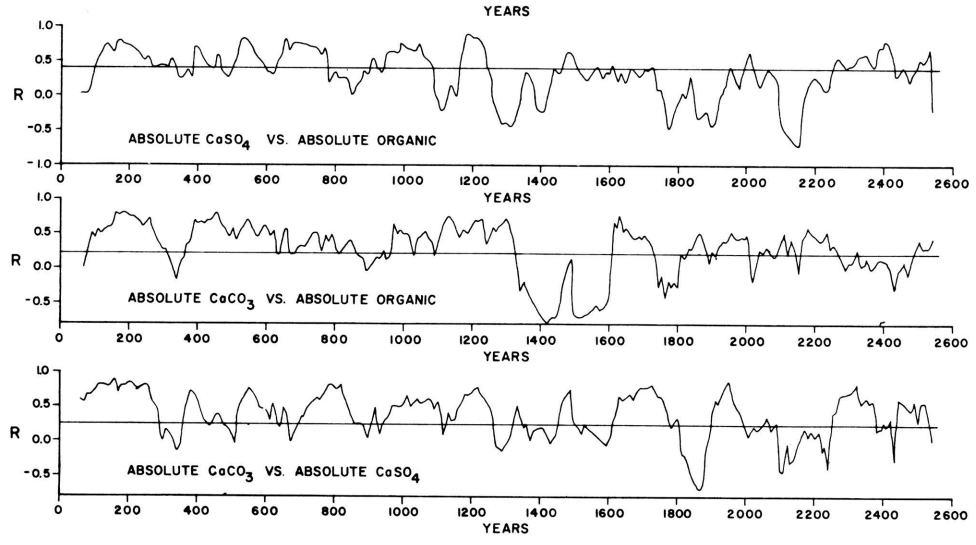


Figure 3.- Moving correlation coefficients for $CaCO_3$, $CaSO_4$, and organic matter determined on a quantity per unit-time (absolute) basis. Sample interval, 10 units. Length of moving correlation, 110 units.

Unrelated Phenomena

In the previous discussion, some simple time-series methods have been adapted to the study of laminations with a view toward interpreting the local environments and processes with which the laminae are associated. Some phenomena, such as climatic changes resulting from planetary perturbations or solar variations, are related only indirectly to the environment in which the laminations formed. For other phenomena, such as secular changes in the magnetic field or magnetic particle influx, the relationship is even more remote. In spite of the remoteness of these phenomena, it is already apparent that any attempt to attack these problems without first understanding most of the local environmental factors would be futile.

REFERENCES

Anderson, R. Y., 1964, Varve calibration of stratification, in Merriam, D. F., ed., Symposium on cyclic sedimentation: Kansas Geol. Survey Bull. 169, p. 1-20.

Anderson, R. Y., in press, Paleoecology of the Rita Blanca lake area, in Anderson, R. Y., and Kirkland, D. W., eds., Paleoecology of an early Pleistocene lake on the High Plains of Texas: Geol. Soc. America Mem.

Anderson, R. Y., and Kirkland, D. W., 1966, Intrabasin varve correlation: Geol. Soc. America Bull., v. 77, no. 3, p. 241-256.

Anderson, R. Y., and Koopmans, L. H., in press, Statistical analysis of the Rita Blanca varve time series, in Anderson, R. Y., and Kirkland, D. W., eds., Paleoecology of an early Pleistocene lake on the High Plains of Texas: Geol. Soc. America Mem.

Antevs, E., 1925, Retreat of the last ice sheet in eastern Canada: Geol. Survey Canada Mem. 146, 142 p.

Antevs, E., 1951, Glacial clays in Steep Rock lake, Ontario, Canada: Geol. Soc. America Bull., v. 62, no. 10, p. 1223-1262.

Clements, F. E., and Shelford, V. E., 1939, Bio-ecology: John Wiley and Sons, New York, 425 p.

Dean, W. E., 1967, Petrologic and geochemical variations in the Permian Castile varved anhydrite, Delaware basin, Texas and New Mexico: Unpub. doctoral dissertation, Univ. New Mexico, 326 p.

Kelley, V. C., 1956, Thickness of strata: Jour. Sed. Pet., v. 26, no. 4, p. 289-300.

McLeroy, C. A., and Anderson, R. Y., 1966, Laminations of the Oligocene Florissant lake deposits, Colorado: Geol. Soc. America Bull., v. 77, no. 6, p. 605-618.

ABSENCE OF DETECTABLE TRENDS IN THE RATE OF BENTONITE OCCURRENCES IN THE MOWRY SHALE (CRETACEOUS) OF WYOMING

by

John C. Davis

Kansas Geological Survey

INTRODUCTION

Geologists are performing time-series analyses on stratigraphic sequences in the hope of obtaining information on rates of occurrence of geologic events. In most instances, time scales are not available, and the stratigraphic succession is considered as a time equivalent. Information is needed about the expected consistency of depositional rates at a location within a sedimentary basin.

Information gathered during an investigation on the petrography of a Cretaceous black shale (Davis, 1967) provided an unusual opportunity to test the application of standard time-series procedures. The unit measured, the Mowry Shale, is a siliceous, marine black shale deposited far from its inferred shoreline. Only minor variations in grain size occur through the section; sedimentary structures indicate that deposition occurred in relatively undisturbed, stagnant water. The section used probably represents as uniform a depositional situation as can be found in marine environments. Distributed throughout the sequence are bentonite deposits that represent the only disturbance in an area of seemingly uniform, slow deposition.

R. A. Reyment (personal communication, 1967), using the same time-series analysis procedures, has found strong trends in the rates of eruptions of modern volcanoes. Detection of similar trends in bentonites in the Mowry Shale could be interpreted as indicating a uniform rate of sedimentation throughout the sequence. Absence of these trends, however, would have no particular significance.

FREQUENCY OF BENTONITE OCCURRENCE

Bentonites in the Mowry Shale are distinctive beds with sharp lower contacts in an otherwise uniform sequence of marine shales. Stratigraphic distribution of bentonites is the result of sudden, catastrophic volcanic eruptions which were separated by differing intervals of time. A random variable through time may be distributed according to the Poisson model, so randomness of bentonites may be tested by calculating goodness of fit to the Poisson distribution. The following assumptions are necessary for this model:

(1) Volcanic eruptions occurring in one time interval are independent of those in any other interval.

(2) The probability that an eruption occurs is proportional to the length of the time interval.

(3) The probability of two or more eruptions occurring in a very short time interval is so small that it can be neglected.

(4) The length of time intervals may be approximated by the thickness of shale that accumulates during that interval.

All assumptions of the model may be questioned. Volcanic eruptions may not be independent events if, for example, one eruption triggers subsequent eruptions in the area or changes the natural pattern of vulcanism in the region. For the same reasons, assumption (2) may be invalid. Graded bedding in bentonites, noted by Slaughter and Earley (1965) may indicate successive, closely spaced eruptions, a condition violating assumption (3). The tenuousness of assumption (4) is obvious.

A search was made for trends in bentonite beds in a section measured in sec. 15, T. 33 N., R. 94 W., Fremont County, near Sand Draw, Wyoming. This section was selected because it is exceptionally well exposed and bentonites are thicker, making them easier to find. From the nature of the exposure, it is unlikely that any bentonites were missed during field examination. This removes the confounding effect of possible incomplete selection from available bentonites, a hazard at less well-exposed sections.

Data, listed in Table 1, consist of distances between successive bentonite beds, measured from center of bed to center of bed. Twenty-eight bentonites are found in 289.5 feet of section. Thickness of individual bentonite beds are not considered, but range from 7 feet to less than 1 inch. Lithology of the enclosing rock consists of rather uniformly silty, siliceous shale. Some fluctuation in apparent grain size is noticeable.

Data were tested using a statistical analysis of series of events program by Lewis and Kelly (1966); procedures used in this program are discussed in Cox and Lewis (1966). The first computation is a test for trend in the rate of occurrence of bentonites, represented by a smooth change in frequency of occurrence through the section. It may be postulated that the distribution is not stationary Poisson, but time dependent Poisson ($\lambda(t) = e^{\alpha+\beta t}$). If the process is stationary, $\beta = 0$, so trend may be detected by testing this assumption. The appropriate testing statistic is

$$U = \frac{S - \frac{T}{2}}{\left(\frac{T}{\sqrt{12n}}\right)},$$

which is normally distributed except for cases of very small n. [T=total thickness, n=number of events, $S=\Sigma t_i/n$, where t_i is the thickness from the bottom of the section to each successive bed.] For the Mowry bentonites, U=0.253, which is not significant above the 60 percent level. It can be concluded, therefore, that no significant trend is apparent in rate of bentonite occurrences.

Table 1.- Feet between midpoints of successive bentonites in Mowry Shale in section 19. Intervals listed from bottom upward.

4	4	14
26	35	17
4	2	5
5	15	10
4	10	5
17	23	6
3	8	11
6	7	29
	47	

A series of values then are computed that can be used to graphically analyze the trend, if present, using standard regression methods. These values are based in part on an arbitrary constant, K, which specifies successive intervals (X_i) containing exactly K bentonites. In this run, K=4. Because any trend present may be exponential, it is also appropriate to work with logs of thicknesses. Graphs of thickness of intervals and log thickness vs. midpoints of intervals are shown in Figure 1. No trends are apparent, nor are fitted regressions significant.

Next, the data are ordered so the form of the distribution of intervals between beds may be displayed graphically as an empirical distribution function. Figure 2 is a cumulative plot of feet between bentonites vs. number of bentonites. Note that the curve appears exponential. Four distribution-free statistics were generated for testing the Poisson hypothesis against three broad general categories of alternatives: (1) trends, (2) renewal processes, and (3) serially correlated stationary series. The first of these alternatives has already been eliminated by testing for trends, leaving alternatives (2) and (3). The empirical distribution function may be compared to a Poisson distribution by Kolmogorov-Smirnov statistics. Neither one-sided nor two-sided tests reject the hypothesis of goodness of fit at the 95 percent level. The Anderson-Darling statistic, W_n^2, which is exceptionally sensitive to departures in the tails of distributions, is then computed. This statistic is not significant at the 95 percent level. Moran's statistic, $\hat{\kappa}_n$, is a test against renewal hypotheses, having a Chi-square distribution with n-1 degrees of freedom. Computed value of $\hat{\kappa}_n$ is 22.7, not significant at the 95 percent level. In conclusion, all four tests fail to reject the hypothesis that the distribution is Poisson.

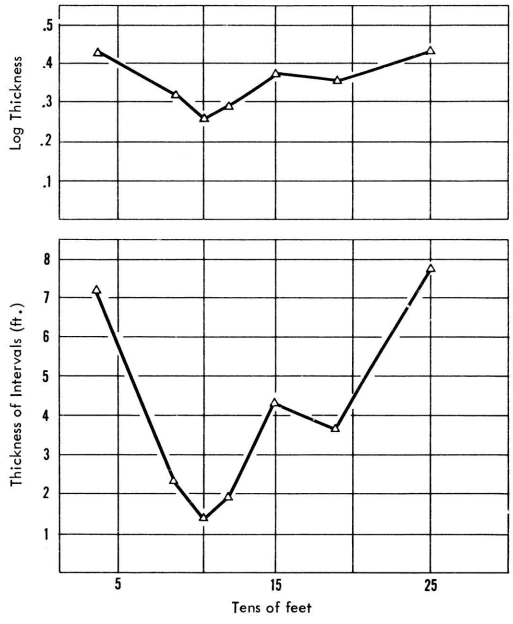


Figure 1.- Thicknesses of intervals containing four bentonites versus stratigraphic height of interval midpoints.

Serial-correlation coefficients may be computed for lags up to n/2 = 13. The distribution of serial correlation coefficients is an estimate of the spectral-density function and can be tested for trend using the U-statistic. Here, the computed value of U = .513 is not significant above the 70 percent level. Highest correlation found is 0.328 for a lag of 2. Distribution-free statistics can be calculated and applied to the estimated spectral-density function. Kolmogorov-Smirnov and Anderson-Darling statistics for testing the observed distribution against a Poisson distribution are not significant at the 95 percent level.

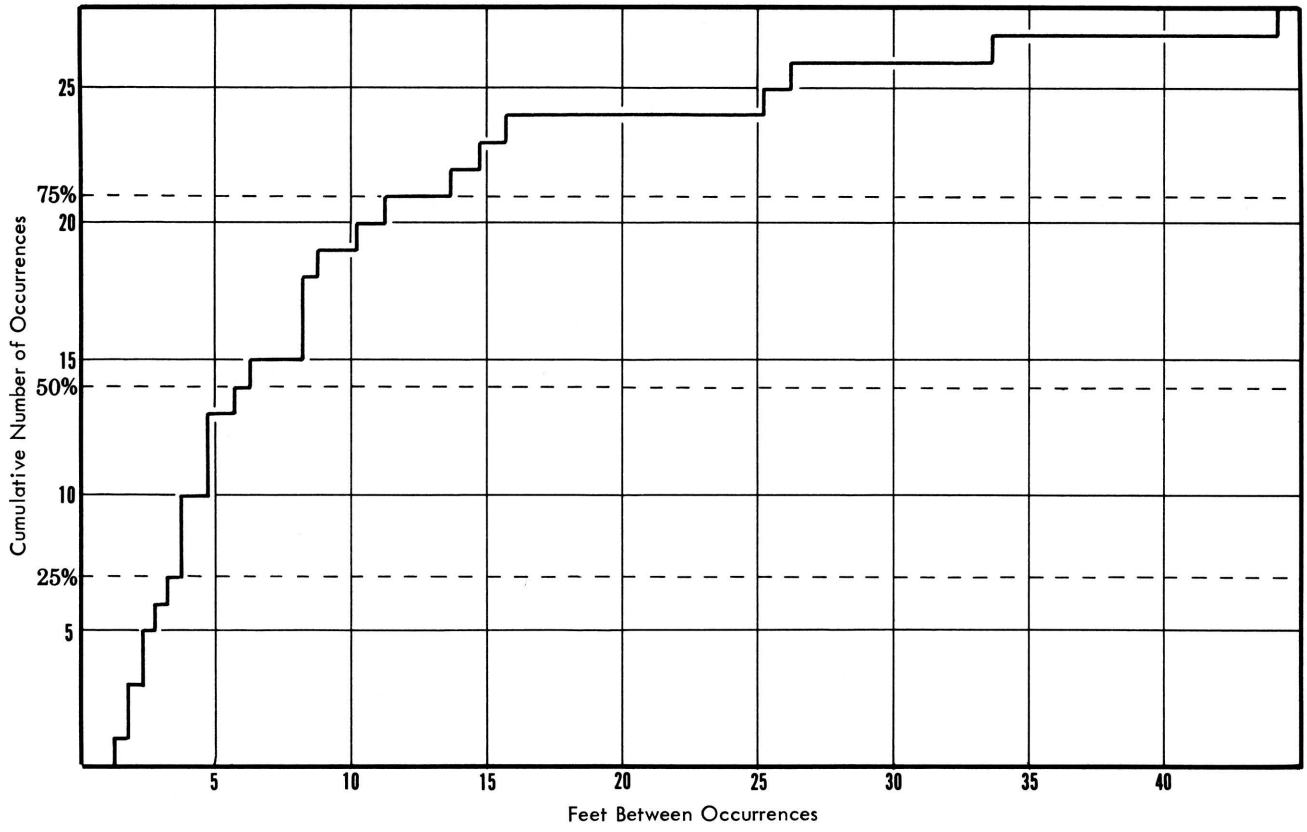


Figure 2.- Cumulative plot of number of bentonites versus feet between occurrences.

A series of additional tests were performed, including calculation of variance time curves, serial covariance curves, and the normalized spectral-density function. In all instances, statistics of these functions are not significantly different from those expected from a Poisson process.

These analyses were run in an attempt to discern a trend or pattern in the frequency of bentonites present in the measured section. Operating on the assumption that segments of black shale represent equal increments of time, bentonite occurrences fit into a stationary Poissonlike distribution. This may be interpreted in various ways: (1) the model assumptions are correct and bentonite-producing eruptions occurred at random time intervals; (2) the model assumptions are incorrect and the true trend is confounded with changes in sedimentation rates throughout the section; and (3) the model assumptions are partially correct and the effect of randomly spaced eruptions are confounded with random fluctuations in sedimentation. The best that can be done is to render the verdict, "not proved." Significant information on rates of volcanic eruptions cannot be extracted from thickness measurements of this section or, presumably, from any other section of the Mowry Shale.

REFERENCES

Cox, D. R., and Lewis, P. A. W., 1966, The statistical analysis of series of events: Methuen and Co., Ltd., London, 283 p.

Davis, J. C., 1967, Petrology of the Mowry Shale: Unpub. doctoral dissertation, Univ. Wyoming, 141 p.

Lewis, P. A. W., and Kelly, T. C., 1965, A computer program for the statistical analysis of series of events: IBM Research Rept. RJ-362, 73 p.

Slaughter, M., and Earley, J. W., 1965, Mineralogy and geological significance of the Mowry bentonites, Wyoming: Geol. Soc. America Sp. Paper 83, 116 p.

GEOPHYSICAL DIGITAL FILTERING

by

Sven Treitel
Pan American Petroleum Corporation

ABSTRACT

Statistical communication theory has contributed substantially to the art of signal extraction from geophysical recordings. Filter design techniques based on the principle of least squares have been extensively studied. Minimization of the error energy in the difference between a distorted and an undistorted signal leads to the Wiener filter. This operator is particularly useful in high resolution work, and can be implemented either in the single-channel or the multichannel mode. Simple models of the real layered earth aid in the formulation of filter design parameters.

AUTOCORRELATION, SPECTRAL ANALYSIS, AND MARKOV CHAINS

by

W. C. Krumbein
Northwestern University

ABSTRACT

Given a sequence of observations, $x_1, x_2, \ldots, x_t, \ldots, x_{t+k}$, at equal intervals of time or equally spaced along a line, methods are available on the one hand for analyzing these data by autocorrelation and spectral analysis; and on the other by arranging the data into discrete states, leading to a transition probability matrix. The first path treats the observations as continuous, whereas the second requires discretization of the observations, unless the p_{ij} are sought in terms of autoregressive functions.

Several questions arise here of possible relations among these paths, and it is interesting to ask whether the concept of statistical entropy, either discrete as $E = -\Sigma p_i \log p_i$ or continuous as $E = -\int f(x) \log f(x) dx$, can be used as a unifying concept. The following scheme is presented as a possible search area for connecting threads among the diverse paths:

	STATIONARY PROCESS	
AUTOREGRESSION		AUTOCORRELATION
		FOURIER TRANSFORM
TRANSITION PROBABILITIES		SPECTRUM
	STATISTICAL ENTROPY(?)	

Formal presentation at the Colloquium will raise questions rather than provide answers.

This is the third colloquium which has been held on the campus of the University of Kansas. The first was concerned with classification procedures, the second with trend analysis. Proceedings from these colloquia have been published as COMPUTER CONTRIBUTIONS 7 and 12.

It is my pleasure to acknowledge several people who have helped with the preparations of this meeting. Prof. R.K. Moore and Dr. Adrian Fung of CRES (Center for Research in Engineering Science) assisted in the planning; Mr. R.F. Treece of University Extension made the arrangements; Dr. J.C. Davis, Mr. O.T. Spitz, and Dr. W. Schwarzacher of the Kansas Geological Survey read the manuscripts; Mrs. Nan C. Cocke, Mrs. Jo Anne Crossfield and Mrs. Laquetta Karch assisted in the preparation of the proceedings.

An up-to-date list of other COMPUTER CONTRIBUTIONS and related publications may be obtained by writing the Editor, COMPUTER CONTRIBUTION Series, Kansas Geological Survey, The University of Kansas, Lawrence, Kansas, 66044.

COMPUTER CONTRIBUTIONS

Kansas Geological Survey
University of Kansas
Lawrence, Kansas

Computer Contribution

1. Mathematical simulation of marine sedimentation with IBM 7090/7094 computers, by J.W. Harbaugh, 1966 . . . . . $1.00
2. A generalized two-dimensional regression procedure, by J.R. Dempsey, 1966 . . . . . $0.50
3. FORTRAN IV and MAP program for computation and plotting of trend surfaces for degrees 1 through 6, by Mont O'Leary, R.H. Lippert, and O.T. Spitz, 1966 . . . . . $0.75
4. FORTRAN II program for multivariate discriminant analysis using an IBM 1620 computer, by J.C. Davis and R.J. Sampson, 1966 . . . . . $0.50
5. FORTRAN IV program using double Fourier series for surface fitting of irregularly spaced data, by W.R. James, 1966. . . . . $0.75
6. FORTRAN IV program for estimation of cladistic relationships using the IBM 7040, by R.L. Bartcher, 1966 . . . . . $1.00
7. Computer applications in the earth sciences: Colloquium on classification procedures, edited by D.F. Merriam, 1966 . . . . . $1.00
8. Prediction of the performance of a solution gas drive reservoir by Muskat's Equation, by Apolonio Baca, 1967. . . . . $1.00
9. FORTRAN IV program for mathematical simulation of marine sedimentation with IBM 7040 or 7094 computers, by J.W. Harbaugh and W.J. Wahlstedt, 1967 . . . . . $1.00
10. Three-dimensional response surface program in FORTRAN II for the IBM 1620 computer, by R.J. Sampson and J.C. Davis, 1967 . . . . . $0.75
11. FORTRAN IV program for vector trend analyses of directional data, by W.T. Fox, 1967 . . $1.00
12. Computer applications in the earth sciences: Colloquium on trend analysis, edited by D.F. Merriam and N.C. Cocke, 1967 . . . . . $1.00
13. FORTRAN IV computer programs for Markov chain experiments in geology, by W.C. Krumbein, 1967 . . . . . $1.00
14. FORTRAN IV programs to determine surface roughness in topography for the CDC 3400 computer, by R.D. Hobson, 1967 . . . . . $1.00
15. FORTRAN II program for progressive linear fit of surfaces on a quadratic base using an IBM 1620 computer, by A.J. Cole, C. Jordan, and D.F. Merriam, 1967 . . . . . $1.00
16. FORTRAN IV program for the GE 625 to compute the power spectrum of geological surfaces, by J.E. Esler and F.W. Preston, 1967 . . . . . $0.75
17. FORTRAN IV program for Q-mode cluster analysis of nonquantitative data using IBM 7090/7094 computers, by G.F. Bonham-Carter, 1967 . . . . . $1.00
18. Computer applications in the earth sciences: Colloquium on time-series analysis, edited by D.F. Merriam, 1967 . . . . . $1.00