

Time-Series analysis of biological data: a case study involving periodicity of coral spawning.

F. J. Martibelli  
555-K Piholo Rd.  
Olinda, Maui, HI 96768

Abstract

A time-series data set of solar radiation and planula production in two spawning types of the coral *Pocillopora damicornis* was analyzed using adjustable, data-smoothing digital filters to separate the time-series into cyclic components. Annual and monthly cycles were described and separated from the noise component by a rigorous mathematical technique. This exemplary analysis demonstrates the usefulness of the method in biological applications on coral reefs.

Introduction

Investigations in the biological sciences frequently involve the measurement of one or more experimental variables at regular intervals. These measurements are often contaminated with random "noise" that obscures the behavior of the underlying phenomenon under study. Furthermore, if the true (i.e. noise-free) behavior is governed by several different influences that vary with time, the contribution of each driving force may be difficult to independently assess. In spite of these factors, patterns within and between data sets are often discernible. This article describes a set of data analysis tools and associated methodology for locating and measuring these patterns.

The data for each measurement is called a time-series and consists of a sequence of observed values of an experimental variable measured at equal intervals over a fixed period of time. The time-series analysis procedure described here (hereafter referred to as TSAP) has been designed to eliminate observation noise, resolve each time-series into a sum of component time-series, then compare the components using a lagged correlation technique. The TSAP can be installed on microcomputers for use in a laboratory environment. It provides working scientists with visual displays of their data, a procedure for modeling the data, and a means of measuring and displaying the correlation between any two time-series.

The TSAP uses an array of adjustable, data-smoothing digital filters to separate time-series into cyclic components whose frequencies, or periods, lie in specified ranges. (Note that frequency is the reciprocal of period length.) Each actual filtering computation is a running-weighted-average of the input time-series. That is, each computed output of the filter is a weighted average of input measurements during a short interval, or "data window", about the time of interest. Weights are calculated based on a prescribed data model so that the filter "passes" model components fitting the window segment to a model by the least-squares, then, based on the fit, calculating a filtered value for the midpoint of the data segment. The window then moves to the next time point and the process is repeated, thereby generating a filtered time-series of averaged segments. A slight modification of the process is used at the endpoints.

Each raw time-series is assumed to be the sum of one or more deterministic components, such as a trend or cycle, and a random or noise component. Analysis of this latter component is the subject of a vast literature (e.g. Koopmans, 1974; Chatfield, 1975; Anderson, 1976) and will not be discussed here. We are concerned instead with removing the noise (smoothing) and isolating each deterministic component for further analysis. The TSAP provides a setting in which trends, periodic components and noise are all modeled by a single family of functions. A trend, such as a linear trend, is interpreted as a slowly-varying or long period cyclic component while noise is interpreted as high frequency cycles. The digital filters may be thought of as "black boxes" which input a time-series, decompose it into cyclic components, discard selected components, combine the remaining components and output the resultant time-series.

In the next section, time-series models, filtering and correlation are discussed using several examples of real data to illustrate procedures. The last sections discuss the mathematics of filtering and correlation.

Example Data

Three time-series are used here as examples. The first is the average daily solar intensity (calories/cm) at Kaneohe Bay, Oahu, Hawaii between September 1981 and December 1982 (Fig. 1a); the second and third are planula release rates for two types of the coral *Pocillopora damicornis* in the bay (Figs 1b and 1c). The data were transformed and plotted as the natural log of the quantity one plus the daily mean number of larvae released. Transformation of these data is useful because of the high variability encountered in the counts (three orders of magnitude) and the presence of zero values (log 0 is undefined). We are interested in what influences, if any, solar intensity has on the corals' reproduction rate. A complete description of methods and various results of these experiments can be found in Richmond and Jokiel (1984), Jokiel et al. (1985) and Jokiel (in press).

Clearly present in the solar data (Fig. 1a) is a cycle with a period of approximately 1 year corresponding to seasons of the northern hemisphere. This trend is not obvious in either of Figs. 1b or 1c. Similarly, Figs. 1b and 1c each contain an approximately monthly cycle thought to be related to the lunar cycle (Jokiel et al., 1985). If the light data contains an additional monthly

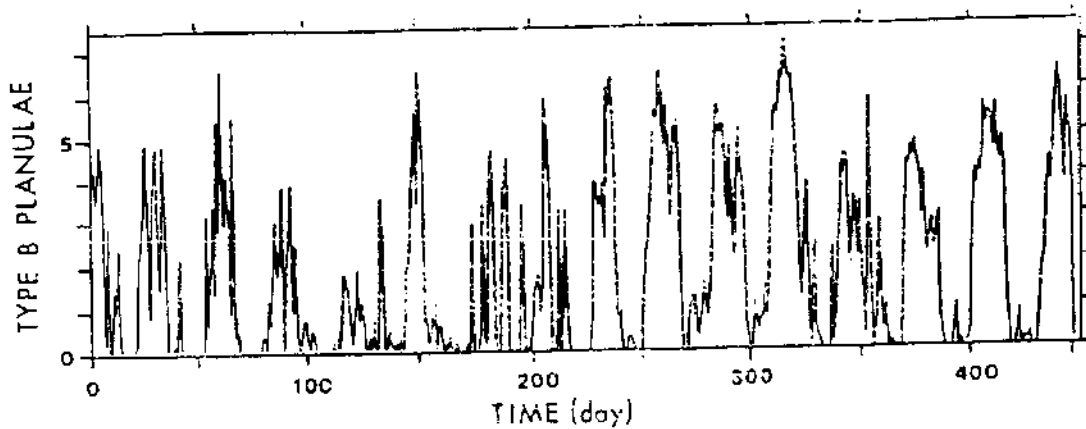


Fig. 1c. Logarithm of average daily production of Type B planula.

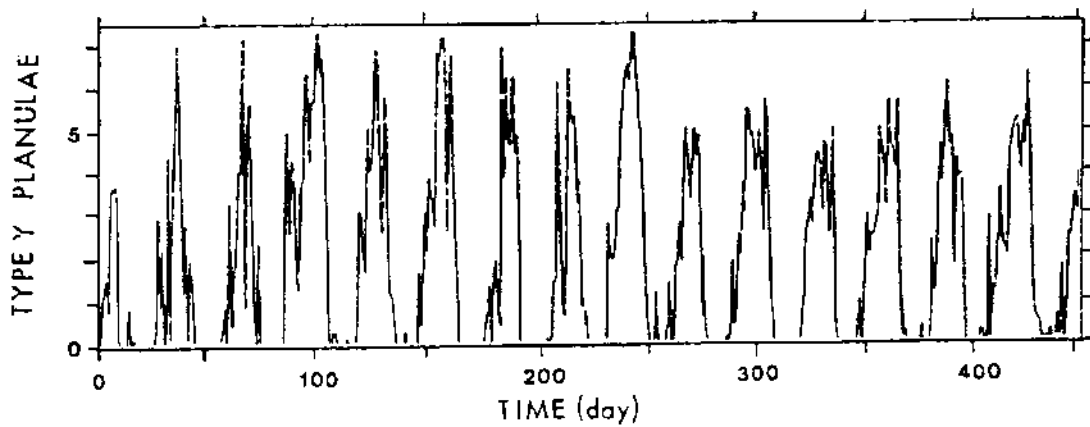


Fig. 1b. Logarithm of average daily production of Type Y planula.

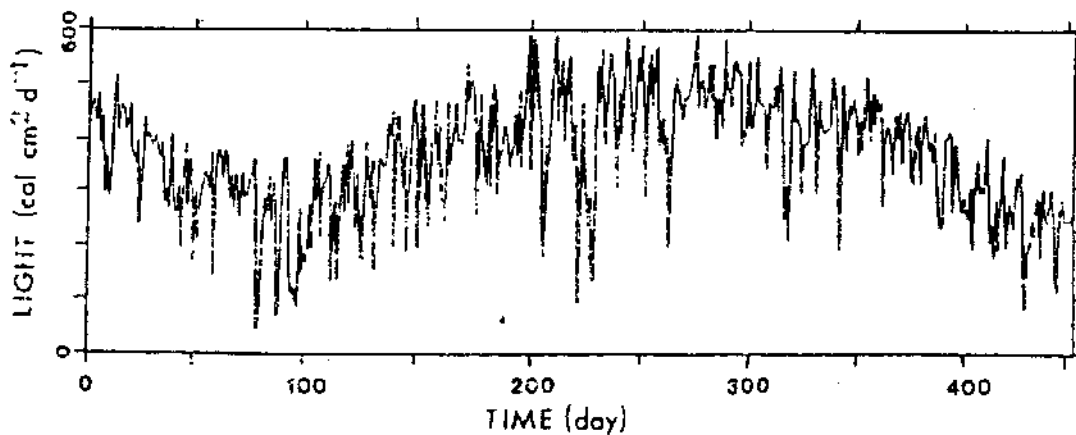


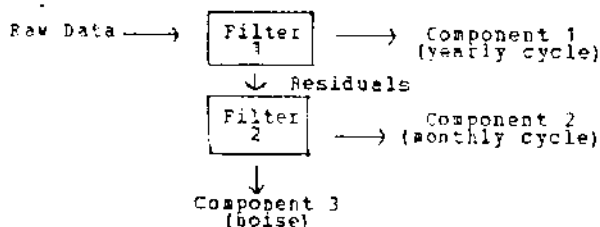
Fig. 1a. Average daily solar intensity in calories/cm<sup>2</sup>.

Fig. 1. Raw Data

cycle, possibly due to sunspot activity, its presence has been obscured by the relatively high noise intensity, due for the most part to cloud cover. In the following section we will develop a general model assumed to hold for all three time-series that is based on the yearly cycle in Fig. 1a and the monthly cycles in Figs. 1b and 1c.

#### Data Modeling

From the above considerations, we are guided to separate each time-series into three components as indicated in the following diagram:



Residuals from Filter 1 are obtained by subtracting daily values of the yearly cycle from corresponding values of the raw data. Component 3 is obtained similarly. Notice that each filtering operation separates its input time-series into exactly two components and that components 1 and 2 in the diagram are simply smoothed versions of their respective input time-series, with residuals in each case being regarded as noise with respect to the filter. Restricting filters to be smoothers simplifies the specification of filter parameters described below. Also, more components can be separated by simply adding more filtering operations.

We now have a general model of the input data, namely:

$$\text{raw data} = \text{yearly cycle} + \text{monthly cycle} + \text{noise}.$$

The next step is to match this to a mathematical model. TSAP's modeling procedure makes tacit use of a basic mathematical fact from linear algebra: any time-series of length, say,  $N$  measurements can be represented exactly (i.e. modeled exactly) by a linear combination of  $N$  other time-series, provided these latter time-series are independent of each other (Draper and Smith, 1966). Among the simplest classes of independent modeling functions are the polynomials and trigonometric functions. While the different classes allow the investigator greater modeling flexibility, our experience has been that final results are nearly the same, provided equivalent choices from each class are made. In the present case, the general model contains periodic components and so trigonometric functions are used as the modeling blocks. In other cases, such as stock market data that contains linear trends, polynomials may be more appropriate.

Let  $N$  represent the number of observations in the data window and let  $\#$  represent a class of modeling functions defined by

$$C_k(n) = \cosine(\theta) \{ (k-1)n/N \} \quad n=1,2,\dots,N$$

for  $k = 1,2,\dots,N$ . For each  $k$ ,  $C_k$  is a discrete, periodic function having frequency  $(k-1)/2N$  cycles-per-day and period  $2N/(k-1)$  days. When  $k=1$ ,  $C_1 = 1$  for each  $n$ . This is a degenerate trigonometric function of infinite period, i.e., a horizontal line, and is required to model the overall mean of the data. Function  $C_2$  has a period of  $2N$  days,  $C_3$  a period of  $N$  days and so on; finally  $C_N$  has a period of 2 days. This is the minimum detectable period of a daily sampling rate.

For each component we must select a data window length  $N$  and then choose those members of  $\#$  which best model the trend we are seeking to isolate. Since input time-series are always smoothed, only the first few functions in  $\#$  are used to model the trend. If the first  $P$  functions are used we say the filter has order  $P$ . Filters are completely characterized by the window length  $N$  and the order  $P$  which are the only inputs required by the filtering algorithm, besides the input data. Selection of  $N$  and  $P$  may require a trial-and-error process. As a general guide, superimposed plots of raw and smoothed data should compare reasonably well with what the investigator might draw freehand. If  $P=1$ , that is if only the function  $C_1$  is used, the filtering process is identical with a simple moving average. As it contains very little model structure, this filter tends to oversmooth. Larger values of  $P$  mean the model is able to capture finer features in the data. However, care must be taken not to choose  $P$  too large, or portions of the noise may be included in the trend. Table 1 indicates appropriate choices for the present situation.

Fig. 2 shows the output from Filter 1 for each time-series. The similarity between 2a and 2c is now more apparent. Also, the lack of similarity between 2a and 2b is apparent. Residuals from this first filtering are shown in Fig. 3, where monthly cycles are still obvious in 3b and 3c, but not in 3a. Figs. 4 and 5 are the trend and noise, respectively, from Filter 2. Type Y planula exhibits a fairly regular cyclic pattern with a period of about 30 days, while Type B shows some irregularities during late winter and early spring. Also, maximum planula production for Type Y seems to occur later than Type B.

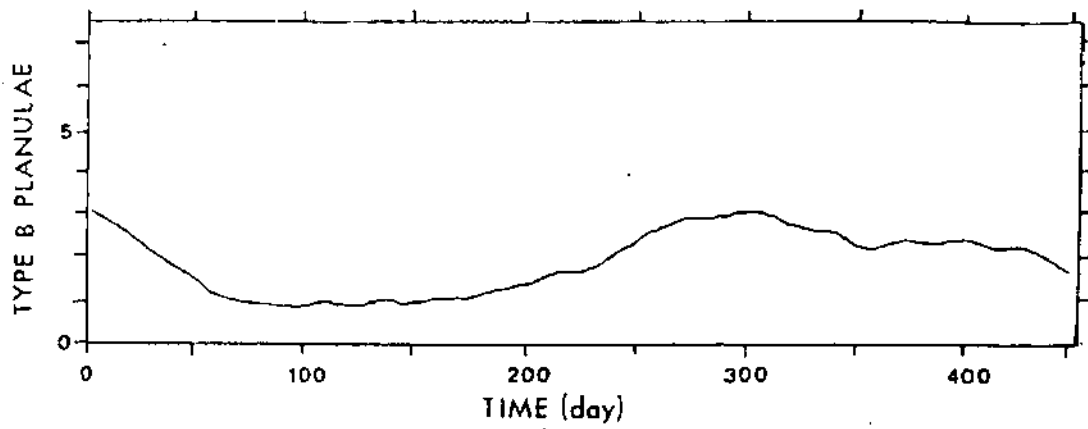


Fig. 2c. Type B planula yearly cycle.

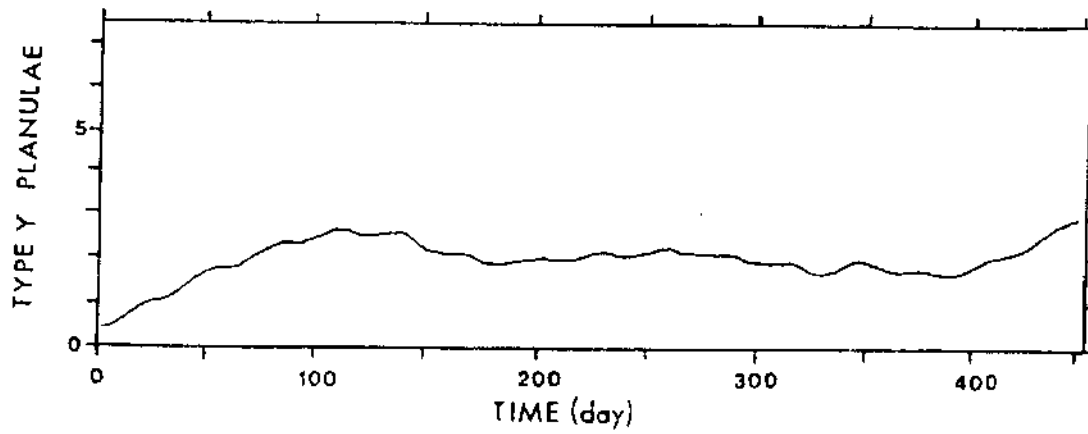


Fig. 2b. Type Y planula yearly cycle.

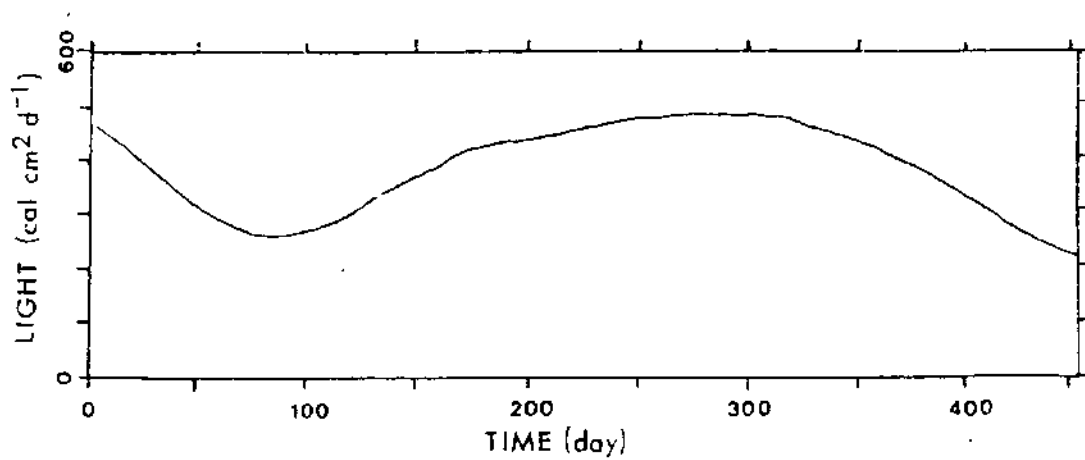


Fig. 2a. Solar yearly cycle.

Fig. 2. Component 1.

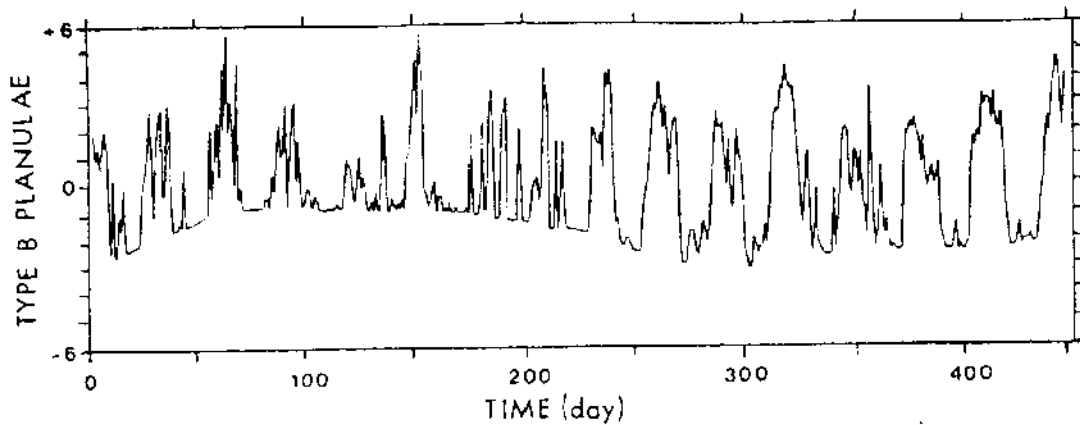


Fig. 3c. Type B planula residuals from Filter 1.

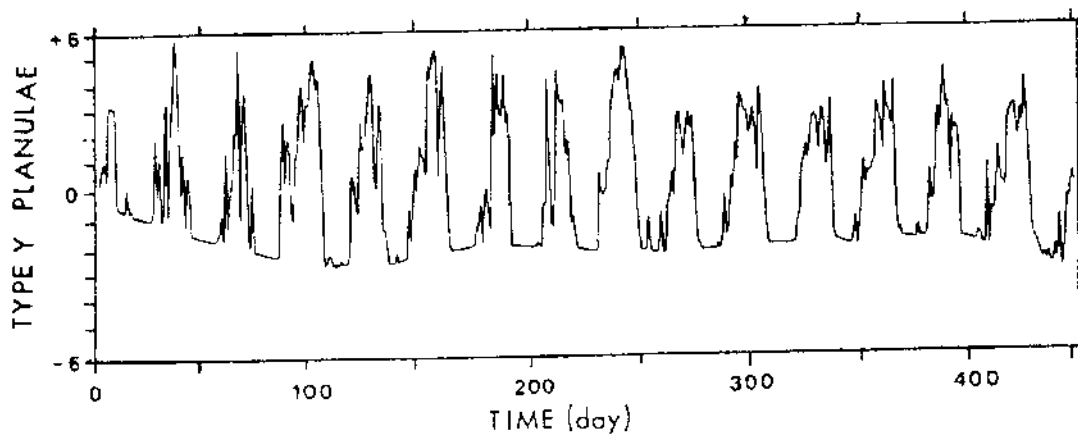


Fig. 3b. Type Y planula residuals from Filter 1.

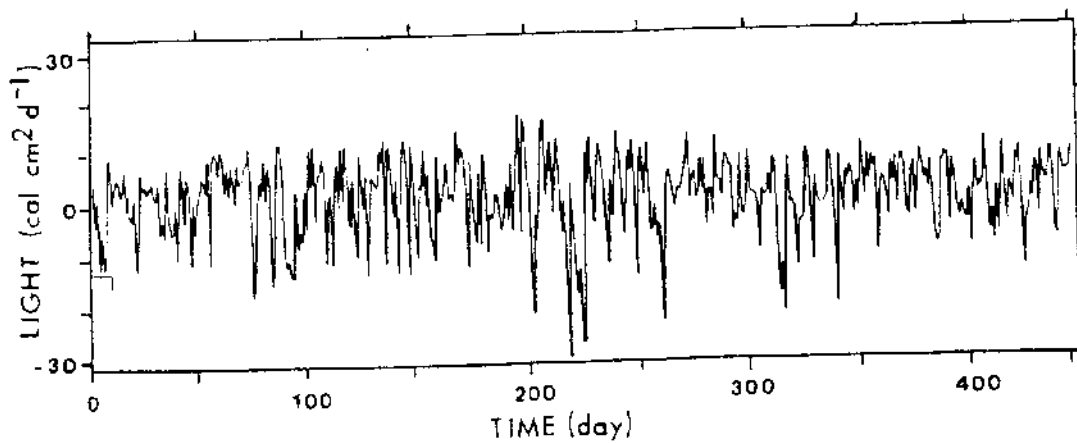


Fig. 3a. Solar residuals from Filter 1.

Fig. 3. Filter 1 Residuals

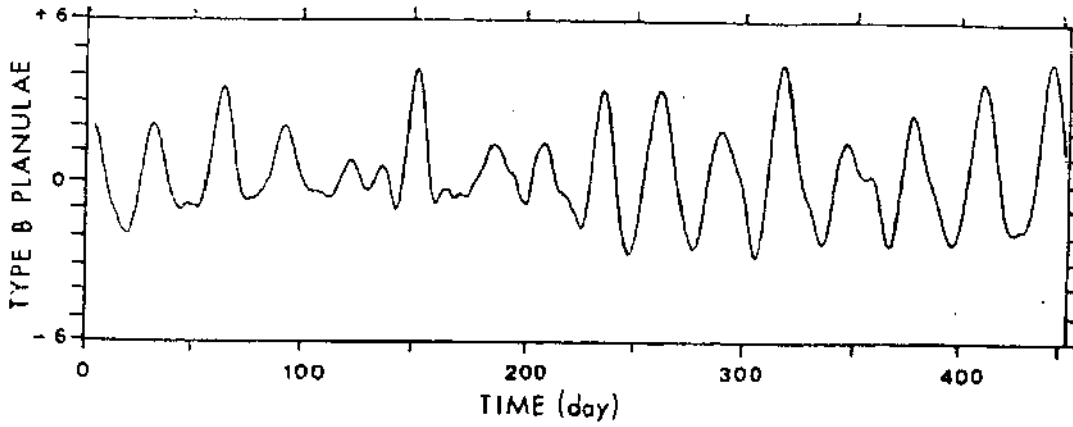


Fig. 4c. Type B planula monthly cycle.

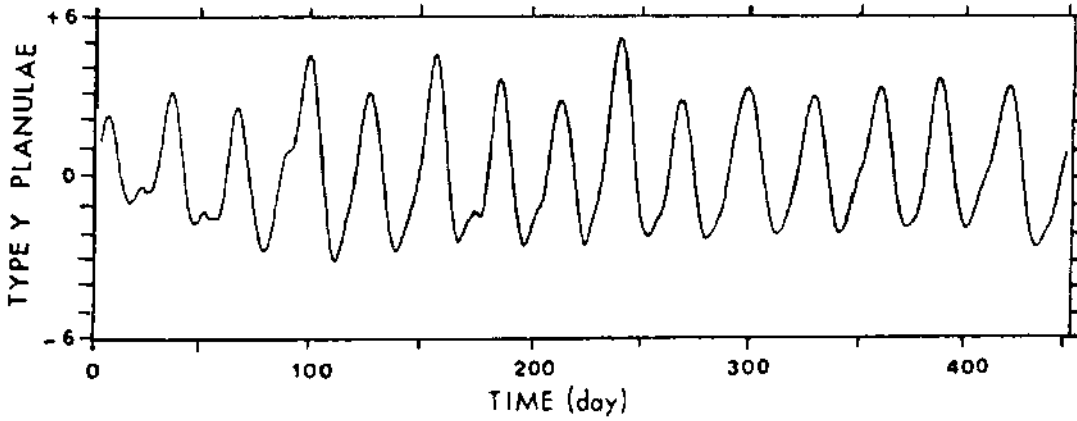


Fig. 4b. Type Y planula monthly cycle.

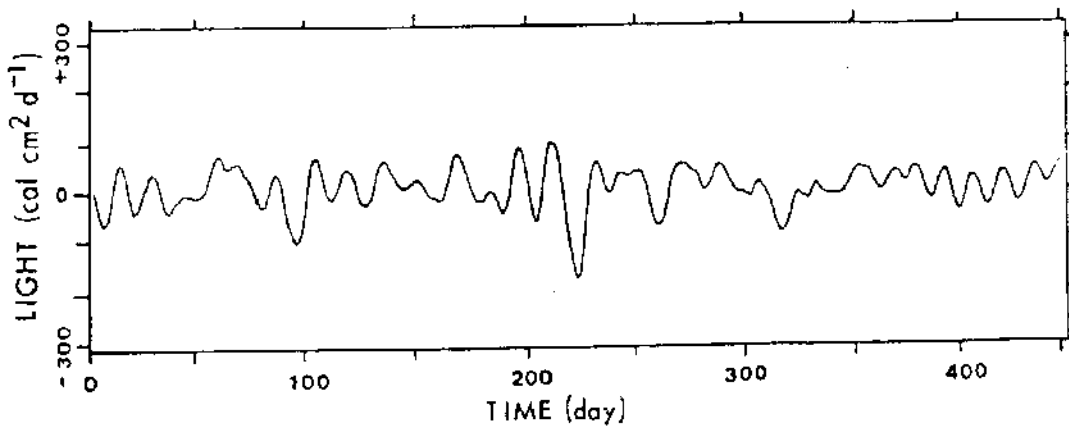


Fig. 4a. Solar monthly cycle.

Fig. 4. Component 2.

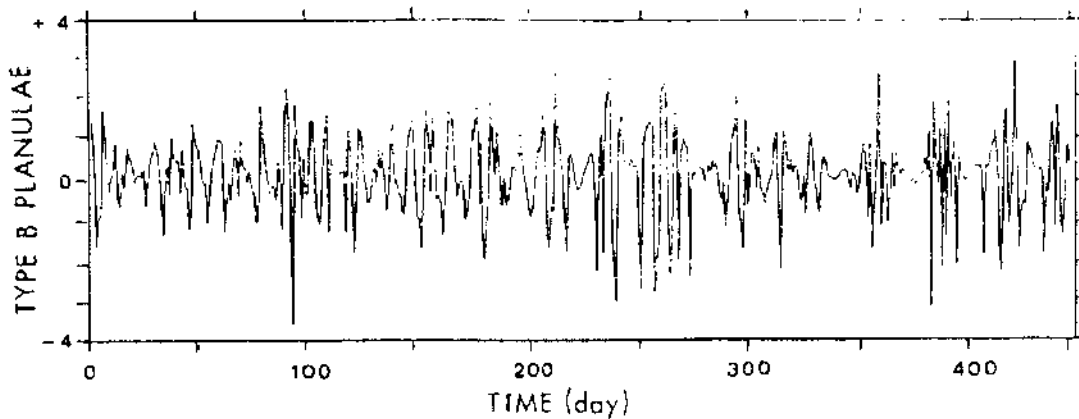


Fig. 5c. Type B planula data residuals from Filter 2.

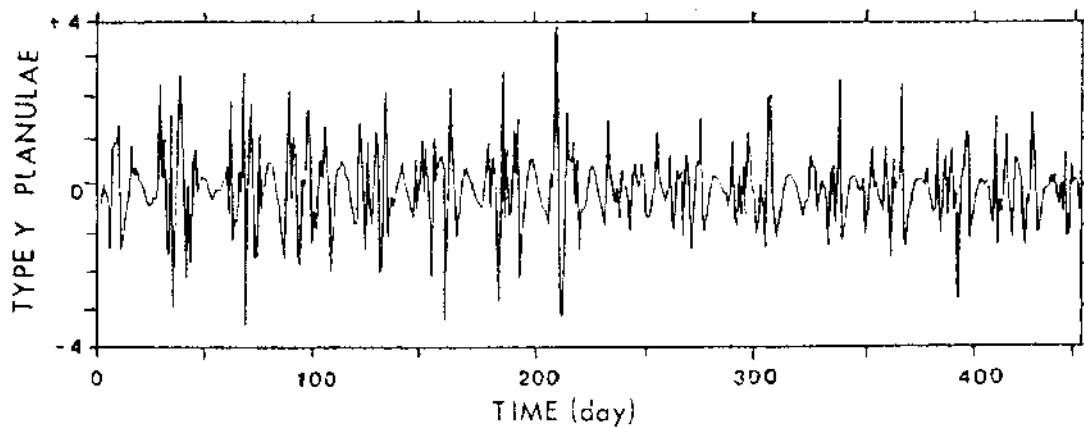


Fig. 5b. Type Y planula data residuals from Filter 2.

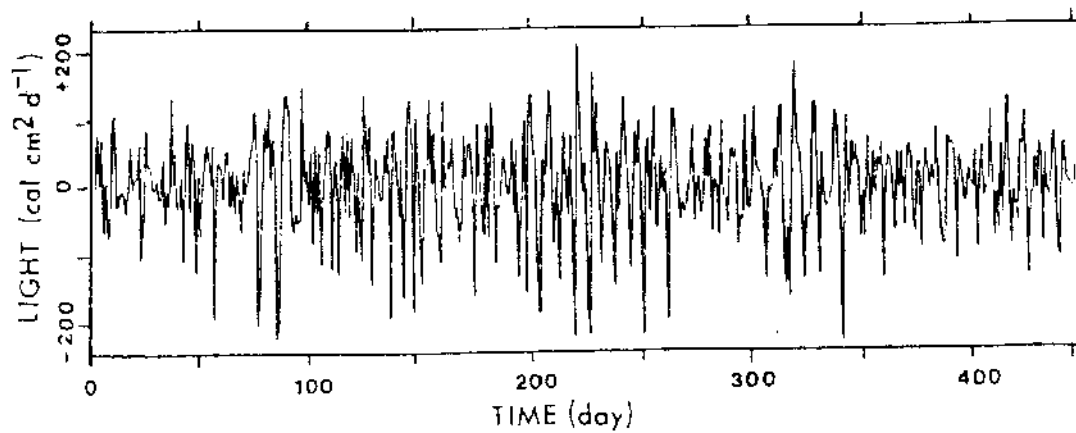


Fig. 5a. Solar residuals from Filter 2.

Fig. 5. Filter 2 residuals.

Table 1.

Component	Assumed Period	Window Length	k-values	Period of Ck	Filter Parameters
1	365 days	200 days	1 2 3	infinite 800 200	N=200 P=3
2	30 days	30 days	1 2 3 4 5	infinite 60 30 20 15	N=30 P=5

Some of these observations can be quantified using a statistic known as the (linear) correlation coefficient and denoted here by  $r$ . The correlation coefficient measures, in a sense, the similarity between two time-series. It can be regarded as a measure of how well one time-series can be approximated by a linear function of the other. Values for  $r$  range between  $-1$  and  $1$ . If  $r=1$  or  $r=-1$  we say the time-series are completely correlated; if  $r$  is negative they are negatively correlated. If  $r=0$  the time-series are uncorrelated or independent of each other. Another way to view  $r$  is that  $r^2$  can be thought of as the fraction of the variance of one time-series that is "explained" by the other.

Correlation values for pairs of smoothed time-series are given in Table 2, column 3. Correlations for pairs 6a and 6b appear to agree fairly well with visual inspection. Small correlations for the pairs in Fig. 7 indicate they are essentially uncorrelated even though the monthly cycles for the corals (4b and 4c) appear quite similar. The reason is that they are slightly shifted: if Fig. 4b and 4c are superimposed and 4c is moved to the right 8 days relative to 4b, the curves line up much better. Fig. 4b is said to "lag" 4c by 8 days. A plot of correlation versus lag time may be made and the lag time at maximum correlation, the so-called phase-shift, read from the plot. Figs. 6, 7 and 8 are lagged correlation plots for the pairs in Table 2. The last column in Table 2 indicates the maximum correlation for non-zero lag times. Fig. 7a reveals a fairly high correlation between planula monthly cycles at 8 d lag time as expected. For pair 6a the large phase shift (200 days) relative to a period of about 365 days, and the low correlation indicate the two curves are unrelated, while 6b reveals a high correlation at a lag of 40 days. Apparently Type B corals are more responsive to gradual changes in solar intensity than Type Y. From Figs. 7b and 7c we conclude that planula monthly cycles are essentially uncorrelated with the somewhat irregular solar monthly cycle.

Table 2.

Figure	Time-series Pair	Zero-lag Correlation	Phase Shift	Maximum Lagged Correlation
6a	2a-2b	-0.26	200	0.37
6b	2a-2c	0.67	40	0.86
7a	4b-4c	0.01	8	0.69
7b	4a-4b	-0.01	56	0.08
7c	4a-4c	-0.01	48	0.13
8a	4b-4b	1.00	30	0.97
8b	4c-4c	1.00	30	0.85

If a time-series is compared with a lagged version of itself, the resultant plot is called an auto-correlation. These have long been used to detect and measure the periods within a time-series. They have the property that zero-lag correlation is always one, as required by Equation 4, and that a phase shift equals the length of the period of a component. Auto-correlation plots for month cycles of both planula types are shown in Fig. 8a (Type Y) and 8b (Type B). Each reveals a phase shift of 30 days in agreement with the 29.5 day lunar cycle.

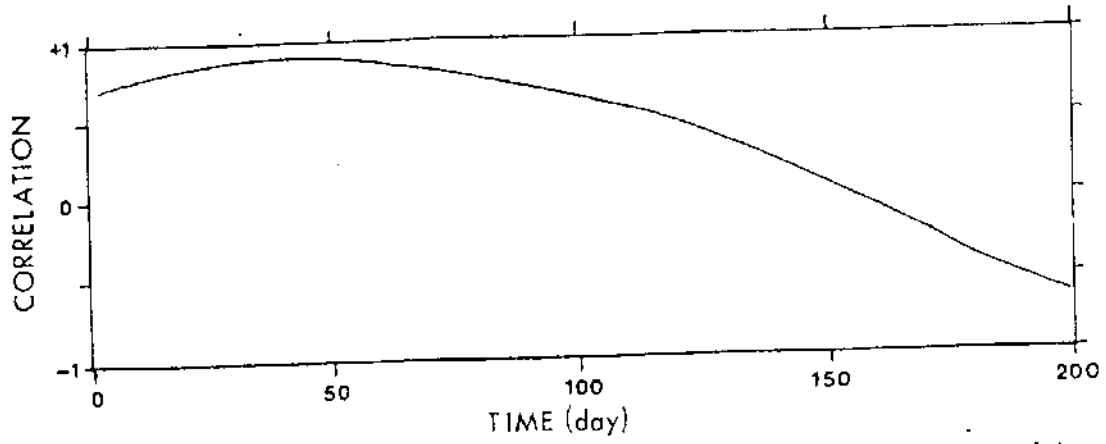


Fig. 6h. Correlation between yearly cycles of solar and Type B planula data.

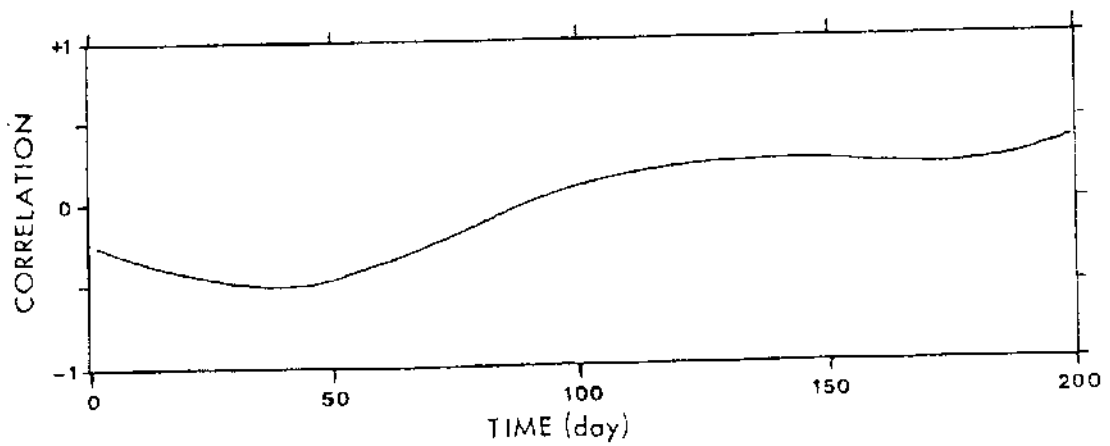


Fig. 6a. Correlation between yearly cycles of solar and Type Y planula data.

Fig. 6. Correlations for yearly cycles.

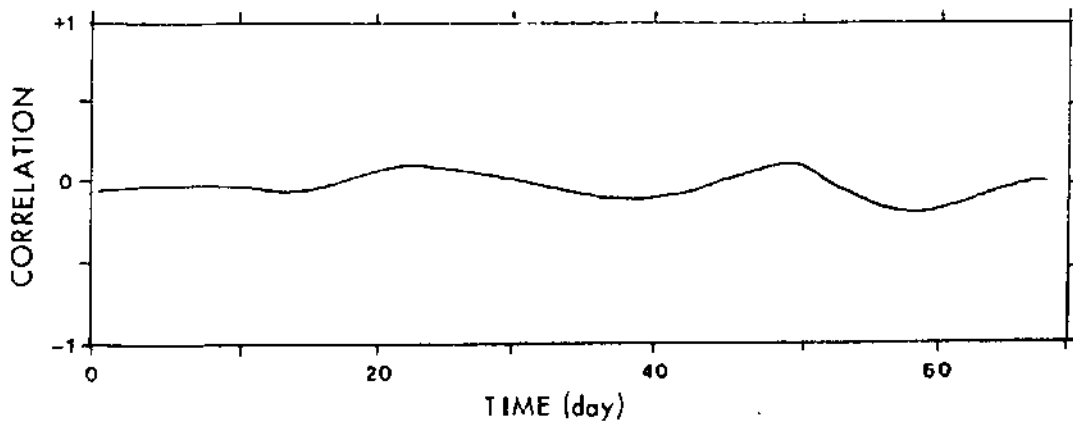


Fig. 7c. Correlation between monthly cycles of solar and Type B planula data.

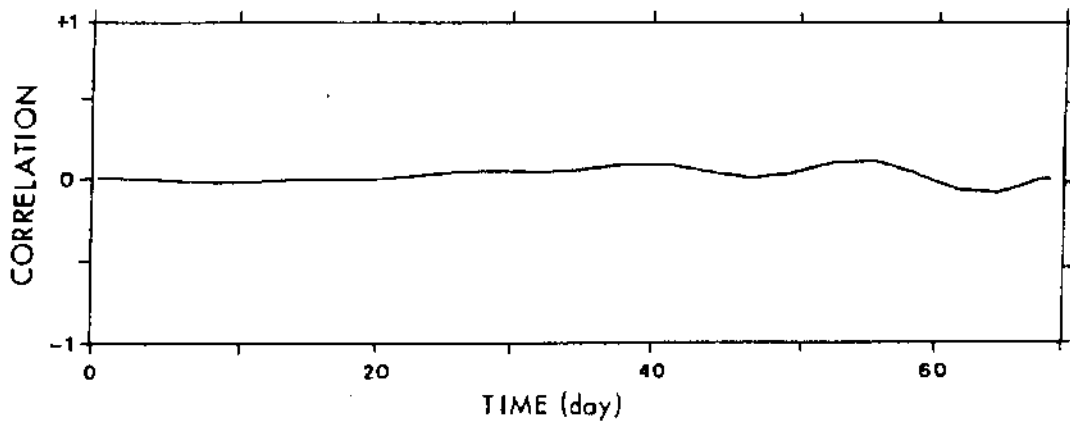


Fig. 7b. Correlation between monthly cycles of solar and Type Y planula data.

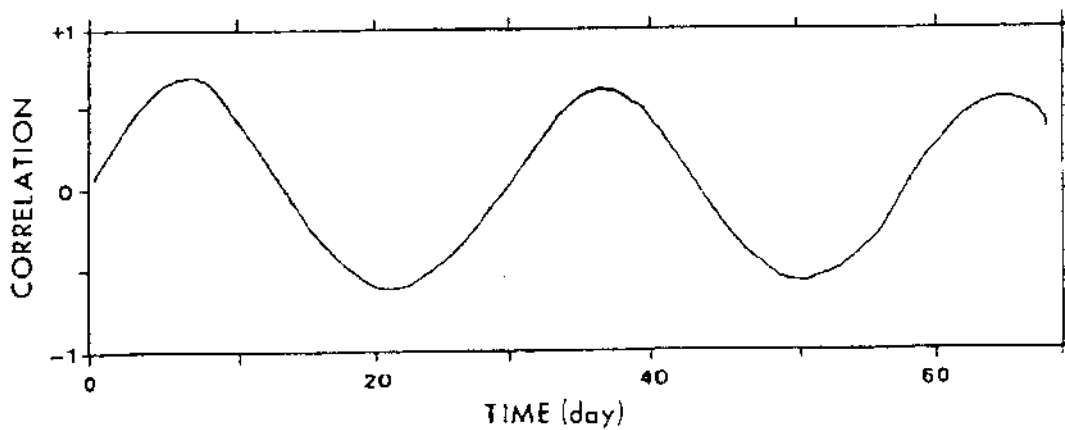


Fig. 7a. Correlation between monthly cycles of Type Y and Type B planula data.

Fig. 7. Correlations for monthly cycles.

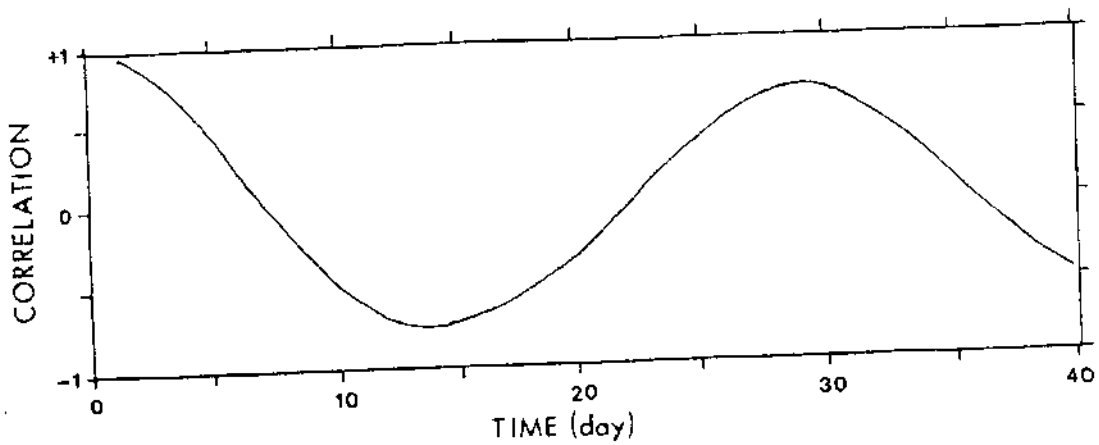


Fig. 8b. Auto-correlation for Type B planula data.

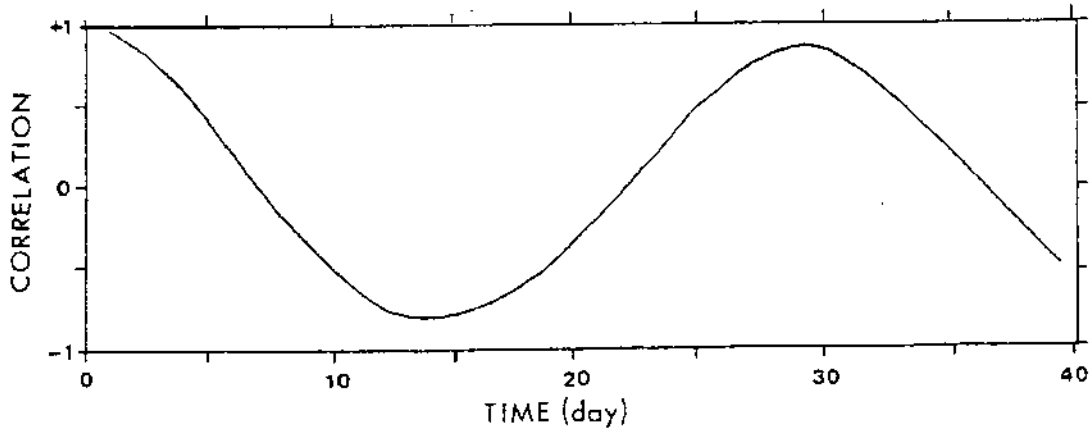


Fig. 8a. Auto-correlation for Type I planula data.

Fig. 8. Auto-correlations for planula monthly cycles.

### The Filtering Algorithm

In this section we present the mathematics underlying the filtering algorithm. Let

$$X(1), X(2), X(3), \dots, X(N)$$

represent time-series  $X$  of  $N$  observations, or measurements, and let  $X(n)$  represent the  $n$ th observation of  $X$ . For filtering purposes,  $X$  is assumed to have only two components:

$$X(n) = I(n) + Z(n), \quad n = 1, 2, \dots, N$$

or, more simply:  $X = Y + Z$ . Here  $Y$  represents the trend in  $X$  and  $Z$  represents filter residuals. Since  $Z = X - Y$ , and  $X$  is known, we only need to estimate  $Y$ . Let  $N$  be the data window length where  $1 \leq N \leq N$  and let  $\mathcal{C} = \{C_k, k=1, 2, \dots, P\}$  represent a family of modeling functions for this window length. As described in Section 3, only the first few members of  $\mathcal{C}$  are used to model a given data component. Let  $P$  be the number of required modeling functions. Then  $Y$  is estimated by a linear combination of the first  $P$  modeling functions:

$$(1) \quad \hat{Y} = A_1 C_1 + A_2 C_2 + \dots + A_P C_P = \sum_{j=1}^P A_j C_j$$

where  $\hat{Y}$  is the estimate of  $Y$  and the  $A_j$ 's are regression coefficients. The  $A_j$ 's are required to minimize the sum of squared differences between the estimate  $\hat{Y}$  and the input data  $X$ . Using the method of least-squares, we find the  $A_j$ 's that minimize

$$(2) \quad S = \sum_{n=1}^N (\hat{Y}(n) - X(n))^2$$

Substituting for  $\hat{Y}(n)$  in (2) gives

$$(3) \quad S = \sum_{n=1}^N \left( \sum_{j=1}^P A_j C_j(n) - X(n) \right)^2$$

and the minimizing  $A_j$ 's are found by differentiating  $S$  with respect to each  $A_k$ :

$$(4) \quad \frac{\partial S}{\partial A_k} = 2 \sum_{n=1}^N \left( \sum_{j=1}^P A_j C_j(n) - X(n) \right) C_k(n) = 0, \quad k = 1, \dots, P$$

Reversing the summation orders in (4) gives

$$(5) \quad \sum_{j=1}^P A_j D_{jk} = \sum_{n=1}^N X(n) C_k(n) \quad k = 1, \dots, P$$

where

$$(6) \quad D_{jk} = \sum_{n=1}^N C_j(n) C_k(n)$$

Equations (5) are known as the Normal Equations for this particular problem and are usually solved by inverting the matrix  $D_{jk}$ . However, if  $P$  is larger than 2, matrix inversion at each data point requires too much computation time to be practical. Instead, using a technique known as the Gram-Schmidt orthogonalization procedure (Lancaster, 1969) the modeling functions  $C_k$  can be adjusted so that

$$(7) \quad D_{jk} = \begin{cases} 0 & \text{if } j \neq k \\ 1 & \text{if } j = k \end{cases}$$

This adjustment has no effect on the modeling capabilities of the  $C_k$ 's.

The Normal Equations (5) are now automatically solved for the  $A_j$ 's:

$$(8) \quad A_j = \sum_{n=1}^N X(n) C_j(n) \quad j = 1, \dots, P$$

The actual filtering algorithm is now easily derived from equations (1) and (8). For any time  $n$ ,  $1 \leq n \leq N$ , the estimate of  $Y$  is given by

$$\begin{aligned} \hat{Y}(n) &= \sum_{k=1}^P A_k C_k(n) \\ &= \sum_{k=1}^P \sum_{n=1}^N X(n) C_k(n) C_k(n) \\ &= \sum_{n=1}^N \sum_{k=1}^P C_k(n) C_k(n) X(n) \end{aligned}$$

and finally

$$(9) \quad \hat{Y}(m) = \sum_{n=1}^M F_n(m) X(n) \quad m = 1, \dots, M$$

where

$$(10) \quad F_n(m) = \sum_{k=1}^M C_k(n) C_k(m) \quad m, n = 1, \dots, M$$

Equation (9) is the filtering algorithm with filter coefficients (weights)  $F_n(m)$ . If the window length is chosen equal to data length  $M$ , the algorithm is equivalent to a least-squares fit of the input data to the model. However, note that, as it stands, the algorithm requires a different set of filter weights for each filtered point (each  $m$ ), resulting in little, if any, computational savings over the matrix inversion procedure. To avoid this problem, and to provide a broader class of data models, the following approach is used. Assume  $M$  is odd and let  $u = (M + 1)/2$ ; then  $\hat{Y}(u)$  is the estimate of  $Y$  at the midpoint of the window. If the window moves forward in time by one unit, the midpoint filter weights do not change. Therefore, neglecting the first and last  $u$  points in the input time-series, estimates for  $\hat{Y}$  can be written

$$(11) \quad \hat{Y}(k) = \sum_{j=-u}^u G(j) X(k+j) \quad k = u, \dots, M-u$$

where

$$(12) \quad G(j) = F_u(u+j) \quad j = -u, \dots, 0, \dots, u$$

Equation (11) is known as a convolution of the data  $X$  with filter weights  $G$ . This algorithm is easily programmed on micro computers and gives reasonably fast execution times. For large data sets, (i.e. long time-series), the algorithm is available "hard-coded" in devices known as array processors with extremely fast execution times.

The first and last  $u$  data points of the input time-series must be treated differently. The simplest and statistically most accurate approach is to use equation (10) to compute new filter weights for each of the endpoints. This, however, may significantly increase running times, especially if  $M$  is large. An alternative approach is to fit straight lines to the first and last  $u$  data points by least-squares and extrapolate these lines forward or backward as necessary to produce  $u$  additional points at each end of the input time-series. The moving window can then be used on the augmented input time-series. The fitting algorithm and resulting filtering algorithm are computationally very fast. This latter approach has been adopted here.

#### The Correlation Algorithm

As described in Section 3, the correlation coefficient  $r$  is a statistical measure of the degree of linear dependence between two time-series. We will first give a computational formula for  $r$ , then attempt to motivate its usage.

Let  $X$  and  $Y$  be two time-series of  $N$  observations each. The (sample) mean of  $X$  is defined by

$$(1) \quad \bar{X} = (1/N) \sum_{n=1}^N X(n)$$

and the (sample) variance of  $X$  by

$$(2) \quad V(X) = (1/N) \sum_{n=1}^N (X(n) - \bar{X})^2$$

$V(X)$  measures, in a sense, the size of  $X$ . Similarly, the (linear) covariance between  $X$  and  $Y$  is given by

$$(3) \quad \text{Cov}(X, Y) = (1/N) \sum_{n=1}^N (X(n) - \bar{X})(Y(n) - \bar{Y})$$

which measures, in a sense, how closely  $Y$  resembles  $X$ . If the covariance is normalized by the variances of  $X$  and  $Y$ , we get the linear correlation coefficient between  $X$  and  $Y$ :

$$(4) \quad r = \frac{\text{Cov}(X, Y)}{\sqrt{V(X)} \sqrt{V(Y)}}$$

Motivation for this formula comes from the following considerations. Suppose we plot values of  $X$  versus  $Y$  on a Cartesian coordinate system, as in the diagram, and let

$$(5) \quad \hat{Y} = A + B X$$

represent the best linear fit of  $Y$  to  $X$ , i.e. the linear least-squares regression line. If all points fall on the line, that is, if  $Y = A + B X$ , then  $\hat{Y}$  and  $Y$  are completely (linearly) correlated and it is easily shown that  $r = 1$  when  $B$  is positive and  $r = -1$  when  $B$  is negative. On the other hand, if  $X$  and  $Y$  are completely independent of each other, then  $\hat{Y} = \bar{Y}$  which implies  $B = 0$  and from equations (3) and (4) we find  $r = 0$ . It should be pointed out that a high correlation between variables does not necessarily indicate a high degree of dependence. For example, the rainfall in Tokyo may be highly correlated with new home sales in San Francisco even though the two are obviously unrelated.

The notion of lagged correlation arises when one time-series is displaced in time relative to the other. Define a new time-series  $W$  by  $W(n) = Y(n+k)$ . Then  $W$  is said to lag  $Y$  by  $k$  time units. If  $k$  takes values between zero and some maximum lag time, both  $Cov(X,W)$  and the associated correlation become functions of  $k$ . A plot of  $r$  versus  $k$  reveals the lag time at maximum correlation, or the so-called phase shift between  $X$  and  $W$ .

#### Discussion

The time-series analysis procedure described and demonstrated in this paper provides a useful tool to the coral reef biologist. Time-series data are notoriously difficult to analyze objectively. The TSAP procedure was designed to eliminate observation noise and resolve each time-series into a sum of component time-series. The components are compared using a lagged correlation technique. The TSAP is especially useful because it can be installed on microcomputers for use in a laboratory environment. The technique provides the coral reef biologist with a visual display of his data, a procedure for modelling the data, and a means of measuring and displaying the correlation between any two time series. The exemplary data analysis demonstration shows the usefulness of this approach.

#### Literature Cited

- Anderson, O. D. 1976. Time Series Analysis and Forecasting: The Box-Jenkins Approach. Butterworth, London, 182 pgs.
- Chatfield, C. 1975. The Analysis of Time Series: Theory and Practice. Chapman and Hall, London. 263 pgs.
- Draper, N. R. and H. Smith. 1966. Applied Regression Analysis. John Wiley and Sons, Inc. New York. 407 pgs.
- Jokiel, P. L. in press. Lunar periodicity of planula release in the reef coral Pocillopora damicornis in relation to various environmental factors. Proc. 5th Int. Symp. Coral Reefs, Moorea, Tahiti.
- Jokiel, P. L., R. Y. Itz, and P. M. Liu. 1985. Night irradiance and synchronization of lunar release of planula larvae in the reef coral Pocillopora damicornis. Marine Biology 88:167-174.
- Koopmans, L. H. 1974. The Spectral Analysis of Time Series. New York, Academic Press. 366 pgs.
- Lancaster, P. 1969. Theory of Matrices. Academic Press, New York 316 pgs.
- Richmond, F. H. and P. L. Jokiel. 1984. Lunar periodicity of larva release in the reef coral Pocillopora damicornis at Enewetak and Hawaii. Bull. mar. Sci. 34:280-287.