# Application of Least Squares to the Search for Periodicities[1]

RAYMOND SNEYERS

*Institut Royal Météorologique de Belgique, Brusseles*

(Manuscript received 10 April 1975, in revised form 5 January 1976).

ABSTRACT

Recalling that the method of estimation by generalized least squares enables testing of hypotheses on the parameters under estimation, advantage is taken from the fact that the solution given by the classical harmonic analysis is identical with the one given by least squares to select those components significantly different from zero. The necessary levels of significance are found through a generalization of Walker's criterion. The same idea is applied to the search for periodicities in time series. In this case, special properties of the series of sample autocovariances and of autoregressive series are used as auxiliary tools.

The illustrations given concern the estimation of daily normals of the mean outdoor temperature at Uccle (Brussels) and the establishment that the sunspot activity is an essentially periodic phenomenon. Errors of estimation are computed in both cases.

## 1. Introduction

The selection of significant components in harmonic analysis is an important problem in climatology, since it makes it possible to improve the accuracy of hourly, daily or monthly normals. The most trivial case is the one in which, in the absence of any seasonal effect, the average of the 12 monthly normals gives a more accurate estimate of the common true mean than does each of the 12 monthly normals.

For this problem, the available methods are those introduced by Schuster (1898), Walker (1914) and Fisher (1929), which are discussed in Anderson (1971). However, being developed under the assumption of a residual independently distributed with the same normal distribution, these methods are more especially suited to the detection of individual significant amplitudes. Some generalization of the technique may thus seem desirable.

On the other hand, the search for periods in time series is generally made through spectral analysis, and from the results obtained in this way it is not always clear whether the detected periods should be ascribed to an autoregressive damping process or to the existence of some pure periodic component of the series.

In reality, it seems advisable to remember here that the equations used in harmonic analysis are precisely those given by the method of least squares, which is a method giving a complete solution of the problem of estimation.

[1] Revised version of an invited paper presented at the Third Conference on Probability and Statistics in Atmospheric Science (Boulder, Colo., 19–22 June 1973).

In particular (see, e.g., Kendall and Stuart, 1963, and Sneyers, 1975), if $x$ is a series of observations, $f(\theta)$ a linear model depending on a parameter $\theta$ or on a set (vector) $\boldsymbol{\theta}$ of parameters, and $e$ a random residual with zero mean, this method gives estimates $\hat{\theta}$ and $\hat{e}$ of $\theta$ and $e$ for which the relation

$$x = f(\hat{\theta}) + \hat{e} \tag{1}$$

between the corresponding values of $x$, $f(\hat{\theta})$ and $\hat{e}$ leads to a relation of the type

$$Q(x) = Q_1[f(\hat{\theta})] + Q_2(\hat{e}), \tag{2}$$

where $Q$, $Q_1$ and $Q_2$ are quadratic functions of the values of $x$, $f(\hat{\theta})$ and $\hat{e}$ among which $Q_2(\hat{e})$ is minimized by the estimate $\hat{\theta}$ and is independent of $Q_1$. Moreover, under relatively general conditions, the components of the estimate $\hat{\theta}$ have, according to the case, an exact or an approximate joint normal distribution from which the covariance matrix may be directly derived from the covariance matrix of the values of $e$. One may thus imagine some statistic $X(\hat{\theta})$ having at least an approximate $\chi^2$ distribution to test the null hypothesis $\hat{\theta}=0$. For nonlinear models, these properties remain approximately true.

Another interesting feature occurs when the covariance matrix of the values of $e$ is known. In this case, under the assumptions of normality of $e$ and of linearity of $f(\theta)$ in $\theta$, the method of least squares gives to $Q_2(\hat{e})$ a form such that it has an exact $\chi^2$ distribution.

It follows that, by working on $Q_1$ or on $Q_2$, it must be possible either to try to simplify an adjusted model

Unauthenticated | Downloaded 04/02/23 11:36 AM UTC

$f(\theta)$ or to test its goodness of fit, that is, to proceed to some kind of selection.

## 2. Selection of significant probabilities in a set of independent probabilities: Jointly significant probabilities

The best way of introducing the method of selection is to consider the case of a multiple test based on $k$ independent probabilities of exceedance.

If $\alpha_1, \alpha_2, \ldots, \alpha_k$ are the probabilities having under the null hypothesis the distribution function $(1-\alpha)$, $-2\log\alpha_i$ has a $\chi^2$ distribution with 2 degrees of freedom, from which it follows that $X=-2\sum\log\alpha_i$ is $\chi^2$ distributed with $2k$ degrees of freedom. The criterion introduced by Fisher in 1929 (see Fisher, 1958) consists then in considering the $k$ probabilities as being jointly significant at the level $\alpha_0$ as soon as the statistic $X$ is significant at that level. Moreover, the distribution function of the largest $(1-\alpha_i)$ or of the smallest $\alpha_i$ being $(1-\alpha)^k$, with Walkers' criterion $\alpha_i$ becomes significant at the level $\alpha_0$ as soon as it is less than $\alpha'$ taken from the equation

$$(1-\alpha')^k=1-\alpha_0. \tag{3}$$

The immediate generalization of this criterion is to consider the joint probability $\alpha$ of $k_1$ among $k$ independent probabilities as being significant at the level $\alpha_0$ as soon as $\alpha$ is less than $\alpha'$ taken from the equation

$$(1-\alpha')^{k/k_1}=1-\alpha_0. \tag{4}$$

When the statistic of a multiple test in a $k$-space is made through a statistic having a $\chi^2$ distribution with $k$ degrees of freedom and when this property remains true in every $k_1$-subspace, the properties of the $\chi^2$ distribution justify the calculation of the levels of significance through Eq. (4). It may be noted for this case that if $\alpha_0=0.05$, it appears from the $\chi^2$ tables that, for decreasing degrees of freedom, the successive critical values may roughly be derived from the initial one by subtracting one for each degree of freedom lower.

For this level of significance, the procedure is thus particularly simple.

## 3. Application to harmonic analysis—Example: estimation of the true mean of the daily average temperature from twelve daily normals equally distant in the year

Different applications of the method have been outlined in Sneyers (1975), so we shall confine ourselves to an example. The twelve normals are the values $x_i$ of Table 1 given in 0.1°C, the corresponding days being approximately placed in the middle of each month. The values $v_i$ are the variances of the error associated with the normals considered as estimations of the true means. The correlation between daily meteorological elements at a distance of one month or more being negligible, the covariances may be considered as being zero-valued. Moreover, normals being averages, the associated errors relative to the true means may be assumed to be normally distributed. If the model to be adjusted to the series is of the form given in Section 1 (Eq. 1), the method of least squares consists then in minimizing

$$\sum e_i^2/v_i=\sum_i [x_i-f_i(\theta)]^2/v_i, \tag{5}$$

which, by differentiation relative to the components of $\theta$, leads to the well-known normal equations. In addition, if $\hat{\theta}$ is the estimate of $\theta$ and if $\hat{e}_i=x_i-f_i(\hat{\theta})$ is the estimate of $e_i$, the statistic $X=\sum\hat{e}_i^2/v_i$ is $\chi^2$ distributed as soon as $f_i(\theta)$ is linear in the $k$ (functionally independent) components of $\theta$, the number of degrees of freedom being $12-k$. The procedure in Section 2 may thus be applied.

To test the null hypothesis that all the true means are zero-valued, that is, $x_i=e_i$, the statistic is

$$X_{12}=\sum x_i^2/v_i=10551.5. \tag{6}$$

For 12 degrees of freedom, the critical value being 21.0 at the 0.05 level, this hypothesis may be rejected. In the same way, the model $f_i(\theta)\equiv a_0$, leading to

$$\hat{a}_0=\sum(x_i/v_i)/\sum(1/v_i)=107.92. \tag{7}$$

With $\hat{e}_{1i}=x_i-107.92$, we have

$$X_{11}=\sum(x_i-107.92)^2/v_i=2003.9, \tag{8}$$

which again is more than the corresponding critical value $21.0-1=20.0$ for one degree of freedom lower. The true means have thus to be considered as being unequal and harmonic analysis appears to be justified.

The complete model then turns out to be

$$f_i(\theta)\equiv a_0+\sum_1^5(a_j\sin j\alpha i+b_j\cos j\alpha i)+a_6(-1)^i, \tag{9}$$

with $\alpha=2\pi/12=\pi/6$. Minimizing (5) in this case leads to the ordinary complete solution of harmonic analysis, since here $e_i\equiv 0$. Thus we have

$$\left.\begin{aligned}\hat{a}_0&=\sum x_i/12,\quad \hat{a}_j=(\sum x_i\sin j\alpha i)/6\\ \hat{b}_j&=(\sum x_i\cos j\alpha i)/6,\quad \hat{a}_6=[\sum x_i(-1)^i]/12\end{aligned}\right\}, \tag{10}$$

which gives the values in column $\hat{\theta}_i$ of Table 1.

Having $x_i\equiv f_i(\theta)$ when (10) is put into (9), a first selection is then based on the decrease of $X_{12}$ in (6) when in (9): $a_0=\hat{a}_0$, $a_j=b_j=0$ except for one value $j=j_0$ for which $a_{j_0}=\hat{a}_{j_0}$ and $b_{j_0}=\hat{b}_{j_0}$. The largest decrease of $X_{12}$ being realized with $j_0=1$, the first harmonic is selected. Moreover, having

$$X_9=\sum\hat{e}_{3i}^2/v_i=24.0 \tag{11}$$

with $\hat{e}_{3i}=x_i-\hat{a}_0-\hat{a}_1\sin i-\hat{b}_1\cos\alpha i$, which is larger

Unauthenticated | Downloaded 04/02/23 11:36 AM UTC

TABLE 1. Daily averages of the temperature of the air at Uccle (in 0.1°C). Selective harmonic analysis of the normals for twelve days equally distant in the year (reference period 1901–70).

| $i$ | $x_i$ | $v_i$ | $\hat{\theta}_i$ | $\hat{x}_i$ | var $\hat{x}_i$ |
|---|---|---|---|---|---|
| 1 | 23 | 26.4 | 95.17 | 18.1 | 10.1 |
| 2 | 24 | 26.4 | −40.90 | 32.4 | 9.4 |
| 3 | 64 | 17.5 | −64.60 | 60.0 | 8.0 |
| 4 | 92 | 17.5 | 2.454 | 93.1 | 7.4 |
| 5 | 125 | 18.5 | − 4.917 | 126.2 | 7.2 |
| 6 | 155 | 15.6 | 1.000 | 154.4 | 6.6 |
| 7 | 173 | 13.7 | 0.667 | 170.7 | 5.7 |
| 8 | 164 | 10.4 | 4.186 | 167.4 | 5.1 |
| 9 | 143 | 12.0 | 2.583 | 141.2 | 5.2 |
| 10 | 102 | 13.7 | 2.401 | 98.7 | 6.0 |
| 11 | 48 | 16.5 | 0.931 | 54.7 | 7.2 |
| 12 | 29 | 25.2 | − 0.833 | 24.9 | 9.0 |

than the critical value 21.0–3 = 18.0, the selection has to go on.

The same procedure with $a_0=\hat{a}_0$, $a_1=\hat{a}_1$, $b_1=\hat{b}_1$, shows that the second harmonic leads to the largest decrease of $X_9$ with

$$X_7=\sum \hat{e}_{5i}^2/v_i=10.7, \tag{12}$$

where $\hat{e}_{5i}=\hat{e}_{3i}-\hat{a}_2\sin 2\alpha i-\hat{b}_2\cos 2\alpha i$. But this time the critical value being 18.0−2 = 16.0, $X_7$ is no longer significant and the selection has to be stopped. On the other hand, having from $(1-\alpha')^{12}=0.95$, $\alpha'=0.0043$, and from $(1-\alpha')^6=0.95$, $\alpha'=0.0085$ and the corresponding critical values of $\chi^2$ with 1 or 2 degrees of freedom (8.4 or 9.6) being smaller than the decreases from $X_{12}$ to $X_{11}$, from $X_{11}$ to $X_9$ and $X_9$ to $X_7$, all the selected components may be considered as being separately significantly different from zero.

The model selected is thus

$$f_i(\theta)=a_0+a_1\sin\alpha i+b_1\cos\alpha i+a_2\sin 2\alpha i+b_2\cos 2\alpha i, \tag{13}$$

for which minimizing (1) gives the final estimates:

$$\left.\begin{aligned} \hat{a}_0=95.4,\quad \hat{a}_1=-40.58,\quad \hat{b}_1=-64.71 \\ \hat{a}_2=2.30,\quad \hat{b}_2=-5.48 \end{aligned}\right\}. \tag{14}$$

For the complete solution of the problem of estimation, we note that using matrix notations [see (Aitken, 1958) and (Sneyers, 1975)], if $\boldsymbol{\theta}$ is the vector with elements $a_0$, $a_1$, $b_1$, $a_2$, $b_2$, $\mathbf{x}$ the vector with elements $x_i$, $i=1, 2, \ldots, 12$, and $\mathbf{u}$ the matrix

$$\mathbf{u}=|1\ \sin\alpha i\ \cos\alpha i\ \sin 2\alpha i\ \cos 2\alpha i| \qquad i=1, 2, \ldots, 12, \tag{15}$$

the estimate $\hat{\mathbf{x}}$ of the vector of the true means is then given through

$$\hat{\mathbf{x}}=\mathbf{u}\hat{\theta} \tag{16}$$

and the covariance matrix var$\hat{\mathbf{x}}$ of elements $\hat{x}_i$ is

$$\text{var}\hat{\mathbf{x}}=\mathbf{u}(\text{var}\hat{\boldsymbol{\theta}})\mathbf{u}', \tag{17}$$

with $\text{var}\hat{\boldsymbol{\theta}}=(\mathbf{u}'\mathbf{v}^{-1}\mathbf{u})^{-1}$ where $\mathbf{u}'$ is the transposed matrix of $\mathbf{u}$ and $\mathbf{v}^{-1}$ is the inverse of the covariance matrix $\mathbf{v}$ of the normals $x_i$, which here is diagonal.

The values of var$\hat{x}_i$ (Table 1), which have been taken from the first diagonal of the matrix (17), are found to be in the mean 2.4 times smaller than the corresponding values $v_i$. On the other hand, var$\hat{\mathbf{x}}$ in (17) not being diagonal, the new estimates are no longer independent.

## 4. The search for periodicities in time series

When the elements of a time series are not independently distributed, the two models coming into competition are the periodic one and the autoregressive one. Moreover, for such series, the autocovariances having quite different properties, a first selection should be based on the sample autocovariances defined by

$$\bar{v}_j=\left[\sum_1^{n-j} x_ix_{i+j}-\left(\sum_1^{n-j} x_i\right)\left(\sum_1^{n-j} x_{i+j}\right)/(n-j)\right]/(n-j-1). \tag{18}$$

In fact (see Sneyers, 1975), if the time series is of the type

$$x_i=a_0+\sum_s c_s\sin(\alpha_s i+\phi_s)+e_i, \tag{19}$$

$T_s=2\pi/\alpha_s$ being then the period of the periodic component with semi-amplitude $c_s$, and if the residual $e_i$ is independently distributed, the series of autocovariances $v_j=\text{cov}(x_i, x_{i+j})$ is the sum of the corresponding components

$$v_j=\sum_s (c_s^2/2)\cos\alpha_s i. \tag{20}$$

Under the assumption of the normality of the distribution of $e_i$ with $v_0=\text{var}e$, we have then, as a first approximation,

$$\text{var}\bar{v}_j=v_0^2/(n-j-1) \quad\text{and}\quad \text{cov}(\bar{v}_j, \bar{v}_{j'})=0 \text{ when } j\neq j', \tag{21}$$

and the ratios

$$\left.\begin{aligned} r_s^2(x)&=v_0/c_s^2 \\ r_s^2(\bar{v}_j)&=[v_0^2/(n-j-1)]/(c_s^2/2)^2 \end{aligned}\right\} \tag{22}$$

show that for $n$ sufficiently large, the first values of $j$ lead to

$$r_s^2(v_j)<r_s^2(x). \tag{23}$$

Hence in the series of the sample autocovariances, the relative importance of the pure periodic components will be larger than in the original time series, at least in the first elements of the series. The calculation of the sample autocovariances thus becomes an excellent way for detecting the existence of periodicities.

On the other hand, if the time series is of the autoregressive type, that is, if we have a relation of

Unauthenticated | Downloaded 04/02/23 11:36 AM UTC

the type

$$x_{i+k}=a_0+a_1x_1+a_2x_{i+1}+\cdots+a_kx_{i+k-1}+e_i, \quad (24)$$

where $a_0, a_1, \ldots, a_k$ are constants with $a_1\neq 0$ and where $e_i$ is independently distributed with zero mean, the autocovariances $v_j$ verify the regression

$$v_{j+k}=a_1v_j+a_2v_{j+1}+\cdots+a_kv_{j+k-1}, \quad (25)$$

from which it follows that, for a stationary time series, the successive autocovariances are damping. Thus, for high values of $j$, $v_j$ tends to zero, which implies that the elements of an autoregressive time series are asymptotically independent when sufficiently distant one from another.

A useful constant which may be associated with an autoregressive series is the equivalent number of repetitions as defined by Bartels (1943) through

$$\omega_n=(n\ \mathrm{var}\bar{x}_n)/\mathrm{var}x, \quad (26)$$

where $\bar{x}_n$ is the average of the series $x_1, x_2, \ldots, x_n$.

In a first approximation it depends only on the constants $a_1, a_2, \ldots, a_k$ and for $k=1$ or $k=2$, it is given respectively by

$$\omega=(1+a_1)/(1-a_1)$$

or (27)

$$\omega=(1+a_1)(1-a_1+a_2)/[(1-a_1)(1-a_1-a_2)].$$

If $F(u)$ is the distribution function of the elements $x_i$, the distribution of the largest value of the series will then approximately be $[F(u)]^{n'}$ with $n'=n/\omega$.

## 5. The case of the series of annual values of the Wolfer sunspot numbers

The case of the Wolfer number is interesting because until very recent years this series was considered as having essentially an autoregressive character. Actually, this opinion was based on a very incomplete analysis of the sample autocovariances and the failure of the adjustment of a periodic model proceeded from the inadequacy of the periods that were tried.

To make a new analysis, the data that we used are the ones given in Waldmeyer (1961) for the years 1700 to 1960, completed by the official numbers for the years 1961 to 1972 kindly communicated by the services of the Royal Observatory of Belgium.

The sample auto-covariances computed through (18) are represented in Fig. 1. They give immediate evidence of a periodic phenomenon. In particular, they show for $j=1$ to 50 maximum values of $\bar{v}_j$ at multiples of a number between 10 and 11; afterwards, for $j=50$ to 140, minimum values at the following multiples of that number; then for $j$ larger than 150, again maximum values at the last multiples of the same number. The damping character of this series when limited from 1 to 60, as done by Craddock (1967), sustained the hypothesis introduced by Yule (1927) of an autoregressive process. However, the regularity of the next values of $\bar{v}_j$ until the very last ones suggests on the contrary variations of the form

$$v_j=f_1(2\pi j/T)\cdot f_2(2\pi j/T'), \quad (28)$$

where $f_1$ and $f_2$ are periodic functions of the cosine type with periods $T$ between 10 and 11 and $T'$ approaching 100.

Hence, the major part of the sunspot number may be considered as consisting of different pure periodic components and more especially as resulting from a beat of two waves with periods close to 10.5 years.

The first model which has to be adjusted to the data is thus of the type

$$x_i=a_0+\sum_1^4 c_s\sin(\alpha_s i+\phi_s)+e_i, \quad (29)$$

with

$$\alpha_1=2\pi/T,\quad \alpha_2=2\pi/T',\quad \alpha_3=2\pi\left(\frac{1}{T}+\frac{1}{T'}\right),$$

and

$$\alpha_4=2\pi\left(\frac{1}{T}-\frac{1}{T'}\right).$$

In fact, $T'$ being much larger than $T$, the model has been simplified by putting $c_2=0$. Successive trials with least squares adjustments of model (2) with $T$ between 10.4 and 10.6 and $T'$ between 150 and 250 lead then to the estimates:

$$\hat{T}=10.494 \quad \text{and} \quad \hat{T}'=200.5. \quad (30)$$

The subtraction of this adjusted periodic component from the initial series gives the first residual $e_1$, which has been analyzed in the same manner. Having found in $e_1$ evidence of a wave with a period of about 100 years, a model of the type

$$x_i=a_0+c\sin(\alpha i+\phi)+e_i, \quad (31)$$

with $\alpha=2\pi/T$, has been used. Trials with least squares adjustments give $\hat{T}=95.50$ and the subtraction of this new adjusted periodic component from $e_1$ leads to the residual $e_2$.

This procedure has been followed for each new residual and the successive results found in this way are given in Table 2. Moreover, in order to allow comparisons, autoregressions of second order have been adjusted to each residual and the corresponding period $T_r$ of the damping waves defined by the autoregression, the corresponding equivalent number $\omega$ computed according to (27) with $k=2$, and the variance $v(e_r)$ of the residual of the regression are given in Table 2. It should be noted that the residuals of these regressions have been found to be independently distributed. Hence, the parameters associated with these regressions are well suited to characterize the statistical

Unauthenticated | Downloaded 04/02/23 11:36 AM UTC

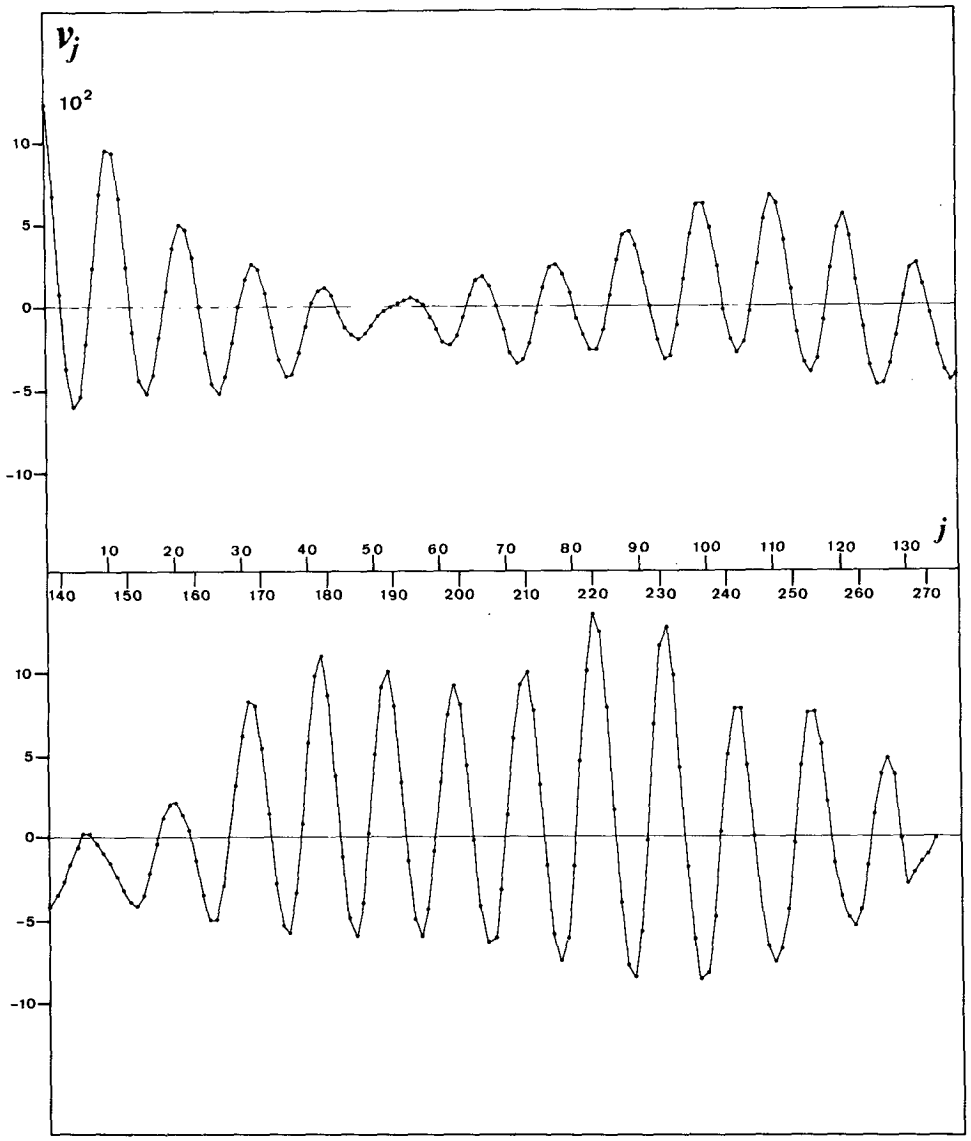


FIG. 1. Auto-covariance of Wolfer sunspot numbers from 1700–1960 (after Waldmeyer, 1960) computed using Eq. (18).

behavior of the residual. In particular, it is of some interest to follow the effect on the values of $T_r$, $\omega$, and $v(e_r)$ of the successive subtractions of the adjusted sinusoidal components from the initial sunspot series.

For the error of estimation as well as for the significance of all these components it should be noted that our procedure of estimating $c$ and $\alpha$ is equivalent to an estimation by least squares in the nonlinear case. From the calculation of the covariance matrix in this case (see Sneyers, 1975) it follows that the estimates are approximately independent and that we have in a first approximation:

$$\left.\begin{aligned} \operatorname{var}\hat{c} &= 2\sigma^2/n \\ \operatorname{var}\hat{\alpha} &= \sigma^2/[c^2 \textstyle\sum i^2 \cos^2(\alpha i+\phi)] \end{aligned}\right\}, \tag{32}$$

where $\sigma^2$ is the variance of the final (independently distributed) residual from which no further significant periodic component can be subtracted. Moreover, in the case where in (31) we have $c=0$, the ratio

$$\hat{c}^2/(2\sigma^2/n) \tag{33}$$

has an approximate $\chi^2$ distribution with 2 degrees of freedom.

To find the variances given in (32), we have proceeded as follows. For the set of 100 approximately equally spaced values of $\alpha$ given by $T=4(0.1)6(0.2)10(0.5)20(2.5)70(10)270$, the value of $c$ given by adjustment of (31) to $e_{24}$ has been computed and the value of $\omega$ for this series of estimates of $c$ has been computed through the adjustment of an autoregression.

Finding $\omega=4.96$ it may be assumed that the extremes of this series are distributed as those of a simple

Unauthenticated | Downloaded 04/02/23 11:36 AM UTC

TABLE 2. Analysis of the series of annual values of the Wolfer sunspot number.

| Variate | $v_0$ | Recurrence $T_r$ | Recurrence $\omega$ | $v(e_r)$ | Subtracted periods $\hat{T}$ | Subtracted periods $\hat{c}$ | Standard errors $\sigma(\hat{T})$ | Standard errors $\hat{T}/[2\sigma(\hat{T})]$ |
|---|---|---|---|---|---|---|---|---|
| $x_w$ | 1478.72 | 10.65 | 1.95 | 259.88 | 11.074 | 25.156 | $4.2\times10^{-3}$ | 1318 |
| | | | | | 9.972 | 20.552 | $4.2\times10^{-3}$ | 1187 |
| | | | | | 10.494 | 12.212 | $7.8\times10^{-3}$ | 673 |
| $e_1$ | 767.50 | 13.55 | 4.04 | 219.59 | 95.50 | 18.074 | 0.45 | 106 |
| $e_2$ | 596.59 | 10.33 | 2.76 | 203.73 | 56.98 | 13.856 | 0.20 | 142 |
| $e_3$ | 501.25 | 8.93 | 2.07 | 189.97 | 11.981 | 13.264 | $9.4\times10^{-3}$ | 637 |
| $e_4$ | 413.50 | 8.45 | 1.97 | 178.92 | 8.482 | 9.409 | $6.6\times10^{-3}$ | 643 |
| $e_5$ | 369.29 | 8.66 | 2.21 | 171.56 | 9.415 | 8.501 | $9.0\times10^{-3}$ | 523 |
| $e_6$ | 333.03 | 8.77 | 2.34 | 165.41 | 13.05 | 7.338 | $2.0\times10^{-2}$ | 326 |
| $e_7$ | 306.10 | 8.47 | 2.24 | 160.08 | 5.485 | 5.354 | $4.9\times10^{-3}$ | 560 |
| $e_8$ | 291.67 | 9.21 | 2.57 | 152.96 | 8.158 | 6.786 | $8.4\times10^{-3}$ | 486 |
| $e_9$ | 268.54 | 10.34 | 2.90 | 146.89 | 11.414 | 7.226 | $1.8\times10^{-2}$ | 317 |
| $e_{10}$ | 249.15 | 10.32 | 2.86 | 141.89 | 16,794 | 5.307 | $4.6\times10^{-2}$ | 183 |
| $e_{11}$ | 234.99 | 9.67 | 2.69 | 138.65 | 13,841 | 4.976 | $3.3\times10^{-2}$ | 210 |
| $e_{12}$ | 222.46 | 9.28 | 2.57 | 135.41 | 15.163 | 5.272 | $3.7\times10^{-2}$ | 205 |
| $e_{13}$ | 208.52 | 8.74 | 2.41 | 131.82 | 21.197 | 4.676 | $8.4\times10^{-2}$ | 126 |
| $e_{14}$ | 197.46 | 8.18 | 2.22 | 128.88 | 27.608 | 4.821 | $1.4\times10^{-1}$ | 97 |
| $e_{15}$ | 185.78 | 7.56 | 2.00 | 125.21 | 46,630 | 7.042 | 0.27 | 86 |
| $e_{16}$ | 160.47 | 6.38 | 1.46 | 114.91 | 133.70 | 5.565 | 2.7 | 25 |
| $e_{17}$ | 145.23 | 5.81 | 1.13 | 106.33 | 8.813 | 4.680 | $1.4\times10^{-2}$ | 315 |
| $e_{18}$ | 134.25 | 5.64 | 1.08 | 101.50 | 4.838 | 3.696 | $5.5\times10^{-3}$ | 440 |
| $e_{19}$ | 127.38 | 5.75 | 1.19 | 98.40 | 7.487 | 3.978 | $1.2\times10^{-2}$ | 312 |
| $e_{20}$ | 119.44 | 5.67 | 1.19 | 95.20 | 4.991 | 3.132 | $6.9\times10^{-3}$ | 362 |
| $e_{21}$ | 114.51 | 5.78 | 1.29 | 92.90 | 5.750 | 4.821 | $5.9\times10^{-3}$ | 487 |
| $e_{22}$ | 102.84 | 6.02 | 1.47 | 87.14 | 10.86 | 3.706 | $2.7\times10^{-2}$ | 201 |
| $e_{23}$ | 95.93 | 5.71 | 1.35 | 83.37 | 253.5 | 3.366 | 15.0 | 8.5 |
| $e_{24}$ | 90.00 | 5.26 | 1.12 | 79.01 | | | | |

random series of size $100/4.96=20.16$. At the 0.05 level, one element of this series is thus significant in a one-sided test if under the null hypothesis its value is beyond the $\beta$ quantile with $(\beta)^{20.16}=0.95$, that is $\beta=0.99746$. In the case of a $\chi^2$ distribution with 2 degrees of freedom, this leads to a critical value of 12.57. With $\omega=1.12$ for $e_{24}$ (Table 2), the value of $n$ to put into (33) becomes $n=273/1.12=243.75$, while equating (33) to 12.57, with $\sigma^2=v_0(e_{24})=90.00$, gives the critical value $c_0^2=9.28$ or $c_0=3.05$.

All the values of the series of 100 estimates of $c$ from $e_{24}$ having been found to be smaller than $c_0$ and those appearing in Table 2 being larger than $c_0$, it may be considered that the selection has been correctly stopped.

Hence, (32) gives similarly

$$\mathrm{var}\hat{c}=2\times90.00/243.75=0.75 \quad \text{or} \quad \sigma(\hat{c})=0.86, \tag{34}$$

while for $\hat{T}$, having from $T=2\pi/\alpha$, $\mathrm{var}\hat{T}=(\hat{T}/2\pi)^2$ with $\mathrm{var}\alpha$ as given in (32) multiplied by 1.12, we find the values of $\sigma(\hat{T})$ indicated in Table 2.

To avoid misinterpretation, one remark remains necessary relative to the estimation by means of the adjusted model. In fact, in the case of models of type (31), if we estimate the pure part of the series through the relation

$$\hat{x}_i=\hat{a}_0+\hat{c}\sin(\hat{\alpha}i+\hat{\phi}), \tag{35}$$

the variance of the error associated with this estimate has a different character according as it concerns the values for $i$ between 1700 and 1972, i.e., inside the period of observation or values for $i$ larger than 1972, i.e., extrapolation into the future.

For the first case, we note that the method of least squares leads here to a decomposition of the form

$$\sum e_i^2=\sum \hat{e}_i^2+\sum (d\hat{x}_i)^2, \tag{36}$$

with $\hat{e}_i=x_i-\hat{x}_i$ and $d\hat{x}_i$ being the error of $\hat{x}_i$ relative to its true value.

Dividing by the variance $\sigma^2$ of the random component $e_i$, $\sum (d\hat{x}_i)^2$ in (36) is transformed into an approximately $\chi^2$ distributed variate with $k$ degrees of freedom, $k$ being the number of estimated parameters. We have thus

$$\mathrm{var}(d\hat{x}_i)=k\sigma^2/n \tag{37}$$

uniformly for any $\hat{x}_i$.

In the case of extrapolation (prediction), on the contrary, the error of estimation associated with $\hat{a}_0$, $\hat{c}$ and $\hat{\phi}$ have to be treated separately from the one of $\hat{\alpha}$. The first error remains in fact the same whatever the range of the extrapolation may be and its variance is given by (37), where $k$ is the number of parameters of the type $a_0$, $c$ and $\phi$. For $\hat{\alpha}$, the error introduces a progressive shift of the estimated wave relative to the true wave, destroying the validity of

Unauthenticated | Downloaded 04/02/23 11:36 AM UTC

the extrapolation in the long run. More precisely, if we put $\Delta\hat{T}=2\sigma(\hat{T})$ for roughly the upper limit of the absolute value of the error on $\hat{T}$, at a confidence level of 0.95, the ratio $\hat{T}/(\Delta\hat{T})$ will be the range at which the shift will be less than one year. The values of this ratio (Table 2) show that short waves lead to large values for this range and large waves to relatively small values of this range.

This restriction being made, it follows from these values, however, that for short range prediction (tens of years) the relation (37) may be used to estimate the total variance of the error on $\hat{x}_i$ in the same manner as in the first case, that is, by taking one degree of freedom for $a_0$ and three degrees of freedom for each sinusoidal wave. With $k=1+27\times3=82$, (37) thus gives

$$\text{var}(d\hat{x}_i)=82\times90/243.75=30.26, \qquad (38)$$

or $\sigma(d\hat{x}_i)=5.5$ for the standard error.

## 6. Conclusions

Both examples have shown how the application of the method of least squares makes it possible to estimate and to select significant periodic components.

The importance of the result of the analysis of the series of sunspot numbers needs, however, further comments. It confirms in fact the existence of a beat phenomenon, as we suggested in Boulder, and converges with the results found independently by Cohen and Lintz (1974).

Two points in these results may be disappointing: first, the many components involved in the model and, second, the relatively high value of the standard deviation of the final residual (about 9.5 against 38.5 for the sunspot number).

For the first point, we have to note that the sunspot number being a non-negative number, by virtue of Fourier's theorem it can only be represented by means of several harmonics, the existence of the beat making the situation still more complex.

The second point may be related to the elementary character of the definition of the sunspot number itself, which makes an important "accidental error" unavoidable. It is even to be expected that this accidental error is still larger since our analysis does not make any distinction between physical effects and artificial effects such as eventual changes in the method of observation.

In any event, the main conclusion of this analysis remains, in our opinion, that sunspot activity must be considered as a generalized periodic phenomenon and as such has a long-range predictability.

## REFERENCES


Aitken, A. L., 1958: *Determinants and Matrices*. Oliver and Boyd.

Anderson, T. W., 1971: *The Statistical Analysis of Time Series*. Wiley.

Bartels, J., 1943: Gesetz und Zufall in der Geophysik. *Naturwiss.*, **31**, 421–435.

Cohen, T. J., and P. R. Lintz, 1974: Long term periodicities in the sunspot cycle. *Nature*, **250**, 398–400.

Craddock J. M., 1967: An experiment in the analysis and prediction of time series. *The Statistician*, **17**, 257–258.

Fisher, R. A., 1929: Test of significance in harmonic analysis. *Proc. Roy. Soc. London*, **A125**, 54–59.

——, 1958: *Statistical Methods For Research Workers*. Oliver and Boyd.

Kendall, M. G., and A. Stuart, 1963: *The Advanced Theory of Statistics*, 3 vols. Griffin.

Schuster, A., 1898: On the investigation of hidden periodicities with application to a supposed 26-day period of meteorological phenomena. *Terr. Mag. Atmos. Elect.*, **3**, 13.

Sneyers, R., 1975: Sur l'analyse statistique des séries d'observations. O.M.M., Note Technique No. 143.

Waldmeyer, M., 1961: The sunspot activity in the years 1610–1960. Technische Hochschule, Zurich.

Walker, G. T., 1914: Correlation in seasonal variation of weather. III. On the criterion for the reality of relationships or periodicities. *Mem. Indian Meteor. Dept.*, **21**, 13–15.

Yule, G. U., 1927: On a method for investigating periodicities in disturbed series, with special reference to Wolfer's sunspot numbers. *Phil. Trans. Roy. Soc. London*, **A226**, 267–298.


Unauthenticated | Downloaded 04/02/23 11:36 AM UTC