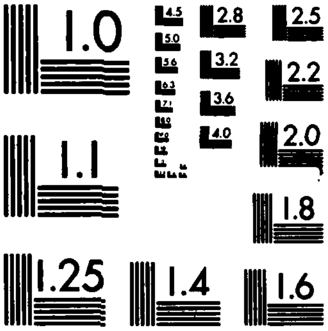
AD-A184 352 THREE-HALVES LAW IN SUNSPOT CYCLE SHAPE(U) STANFORD 1/1
UNIV CA CENTER FOR SPACE SCIENCE AND ASTROPHYSICS
R N BRACEWELL JUL 87 CSSA-ASTRO-87-8 N00014-85-K-0111

UNCLASSIFIED F/G 3/2 NL


1.0
4.5
5.0
5.6
6.3
7.1
8.0
2.8
3.2
3.6
4.0
2.5
2.2
2.0
1.1
1.8
1.25
1.4
1.6

11

C S S A

THREE-HALVES LAW IN SUNSPOT CYCLE SHAPE

R.N. Bracewell

CSSA-ASTRO-87-8
July 1987

(to appear in Monthly Notices of the Royal Astronomical Society)

CENTER FOR SPACE SCIENCE AND ASTROPHYSICS
STANFORD UNIVERSITY
Stanford, California

11

THREE-HALVES LAW
IN SUNSPOT CYCLE SHAPE

R.N. Bracewell

CSSA-ASTRO-87-8
July 1987

(to appear in Monthly Notices
of the Royal Astronomical Society)

NASA Grant NGL 05-020-272
ONR Contract N00014-85-K-0111

(+ or -)

Three-Halves Law in Sunspot Cycle Shape

R. N. Bracewell

Space, Telecommunications and Radioscience Laboratory, Stanford University, Stanford, CA94305, USA

April 29, 1987; July 31, 1987

Summary. Annual mean sunspot numbers $R(t)$ since 1700 show evidence of a nonlinear effect, first evidenced by the detection of third harmonic in $R_{\pm}(t)$, the alternating representation of the magnetic (22-year) cycle of solar activity. The form of the nonlinearity proves to be a three-halves law $R(t) = (100[|R_{lin}(t)|/83])^{3/2}$, where $R_{lin}(t)$, also an alternating quantity, is a presumed underlying or "linearized" sunspot number. The nonlinearity is of such a nature as to cause strong semicycles to be sharper than sinusoidal and to produce the inflexion in $R_{\pm}(t)$ noted at sunspot minimum. The difference $R(t) - |R_{lin}(t)|$ is sufficiently like third harmonic, for semicycles of average strength, to explain the band around 22/3 years which is noticeable in the spectrum of $R_{\pm}(t)$. However, third harmonic alone is not sufficient to account for the observed dependence of semicycle shape on amplitude, whereas the three-halves law accounts economically for a range of effects. A search for a physical explanation of a three-halves law reveals that just such a law results because large sunspot groups, such as occur around strong sunspot maxima, enter into the sunspot number, as conventionally defined, over more days than small groups, simply because large groups last longer.

Semicycle asymmetry, which cannot result from a simple nonlinear law, is here ascribed to magnetic buoyancy acting preferentially on the antinodal layers of a travelling wave. Profiles for semicycles of different strengths have been constructed on the assumption that the underlying influence is sinusoidal. Each sinusoid is distorted by the three-halves law, and then made unsymmetrical by applying a buoyancy theory for magnetized plasma rising against viscous drag. The majority of past semicycles, including those of the 18th century, can be matched quite well by the artificial profiles, a conclusion that supports the idea of an underlying influence that is sinusoidal, and the hypothesis that sunspots result from an upward traveling wave from a submerged 22-year oscillator.

3/2

Introduction

The shape of the sunspot cycle has been of importance for many years in connection with ionospheric physics, magnetospheric physics, and interplanetary phenomena, including solar proton events, partly because of the impact on such diverse practical matters as the scheduling of high-frequency radio communication, over-the-horizon radar, magnetic mines, ephemerides of low-level satellites, and the safety of astronauts.

The eleven-year recurrence tendency has been variously attributed to an oscillator subject to disturbances or alternatively to a resonator excited by random impulses; one sees that these proposals range from an active source at the one extreme to a passive structure at the other. Among mathematical models proposed in the past, some have been based on these two not very definite but nonetheless physical approaches, while other models have been frankly empirical with no physical basis at all. An understanding of the shape of the solar cycle would be helpful for prediction in practical circumstances.

Features to explain

The traditional solar cycle begins at sunspot minimum, lasts about 11 years, and ends at the following minimum, but since magnetic polarities reverse at sunspot minimum, an ambiguity has developed in the meaning of the word cycle. For example, one may read that the solar "cycle" is really 22 years. For this reason the terms "semicycle" and "magnetic cycle" will be used to distinguish the 11 ± 1.55-year episodes from the 22.2 ± 2.08-year ones.

The course of an "11-year" semicycle, as represented by annual means, has a distinctive nonsinusoidal shape that varies from occasion to occasion, apparently in dependence upon at least two parameters, the duration T and the maximum R_{max} reached by the annual mean sunspot number R. Starting from zero, or a small value, at $t = 0$, there is an initial fillet where $d^2R/dt^2 > 0$, a convex segment peaking at $t = t_m$, followed by a more or less gentle decline toward zero at $t = T$ that is often separated from the convex segment by a noticeable break in downward slope. Usually the maximum occurs earlier than at $t = \frac{1}{2}T$, the more so as R_{max} is larger; roughly speaking, the product $R_{max}t_m$ tends to be constant. In contrast, the declining phase tends to be longer as R_{max} is larger, the ratio $R_{max}/(T - t_m)$ tending to be constant. The unsymmetrical aspects are less pronounced when the semicycle is a weak one. In Fig. 1, on the left, these features are exhibited, for the strong semicycle, by the average of the semicycles beginning 1775, 1784, 1833, 1843, 1867, 1933, 1944, 1954, 1976, all of which had $R_{max} > 114$. The shape of a weak semicycle is represented by the average for 1755, 1798, 1810, 1823, 1878, 1889, 1901, 1923, where $R_{max} < 86$.

- Fig. 1. Features of the shape of a strong semicycle (left) include a leading fillet, a convex segment

A-1

peaking early, and a relatively long decline, sometimes with a noticeable break in slope. A weak semicycle (right) tends to be much more symmetrical.

Fourier analysis, derectification and linearisation

Fourier analysis of the annual mean sunspot number series $R(t)$ has been frequently carried out but it has always been difficult to interpret the results. If one agrees that $R(t) = |R_{\pm}(t)|$, where $R_{\pm}(t)$ is an alternating series that reverses sign in successive semicycles (Bracewell, 1953) and whose period is the full magnetic cycle, then it is easy to see where much unsignificant spectral detail has originated: $R(t)$ is related to $R_{\pm}(t)$ in the same way that the output voltage of a full-wave rectifier is related to its input; thus rectification of the alternating series $R_{\pm}(t)$ results in the appearance in $R(t)$ of sum and difference frequencies without physical relevance. Fourier analysis of $R_{\pm}(t)$, by contrast, gives a much simpler spectrum (Cole, 1973). The factitious band of spectral lines around 11.1^{-1} c/a, together with the harmonics of this band, practically disappears, while a cleaner 22.2^{-1} c/a band is seen. Not only does the naïve spectrum of $R(t)$ contain artifacts of rectification, but other significant spectral features, such as a third harmonic (period $\sim 22/3$ years) which appears in Cole's spectrum, are suppressed. This particular feature has recently assumed considerable physical importance.

Such a third harmonic was independently noticed in the course of an attempt to determine an envelope for $R_{\pm}(t)$ (Bracewell, 1985), where it was found that removal of the harmonic by filtering was an essential prerequisite to success of the Hilbert transform envelope procedure. Since then a similar third-harmonic phenomenon was noticed in the application of Hilbert transform analysis to varve data (Bracewell and Williams, 1986). For the purpose of those two papers it was sufficient to reduce the unwanted harmonic component to a tolerable level by filtering; but it was observed that the waveform being suppressed was not strictly a monochromatic third harmonic of the 22.2-year fundamental. Instead, the zero-crossings of the unwanted perturbation were so related to each "11-year" semicycle that the perturbation was ascribable to a nonlinear effect operating year by year on the oscillating quantities, annual sunspot number and varve thickness respectively. Strong semicycles were accompanied by a perturbation of amplitude greater than in direct proportion to the strength of the semicycle, while the zero-to-zero spacing of the perturbation was not fixed, as would be true of a harmonic, but changed in accordance with the duration of that semicycle. Thus the phenomenon which originally revealed itself through the appearance of a 7-year peak in the computed spectrum received a compact description. The spectral description, on the other hand, while mathematically equivalent, proved to be dependent upon the duration and epoch of the time segment chosen for spectral analysis.

Since both the sunspot series and the varve series exhibit this phenomenon of cycle shape distortion, a source in the sun is indicated.

Subsequently it was found that the sunspot number curve could be retrodicted by the introduction of a quadratically nonlinear function

$$R_{\pm} = (R_{lin} + \beta R_{lin}^2)\,\mathrm{sgn}\,R_{lin}$$

in which the coefficient was adjusted empirically at $\beta = 0.0035$ (Bracewell, 1986). If the linearized sunspot number R_{lin} is taken to vary quasisinusoidally then the enhancement at large R will cause the sharpened profile familiar in strong semicycles of observed sunspot number. Linear Fourier synthesis with inclusion of third harmonic is unable to mimic the observed semicycle shapes as well as the nonlinear formula allows.

The motivation for nonlinearity correction was that R_{lin} might be in some sense physically more fundamental than the annual sunspot number $R(t)$, a presumption that is supported to the extent that some features of $R(t)$ are accounted for and the spectrum becomes simpler.

The form of the nonlinear law

Instead of assuming the nonlinear relation represented by an arbitary quadratic law with the one disposable parameter β it would be nice to determine the law explicitly point by point. Suppose that measurements $y(t)$ were available of a nonlinear function $\mathcal{H}$ of a quasisinusoidally varying function $x(t) = A\sin(wt+\alpha)$, i.e. $y(t) = \mathcal{H}[x(t)]$. The situation could be represented as in Fig. 2. Now the function $\mathcal{H}$ could be determined by plotting $y(t)$ versus $x(t)$ with t as a parameter, if $x(t)$ were known.

- Fig. 2. A nonlinear characteristic (top left) relates a stimulus $x(t)$ (lower left) to a response $y(t)$ (upper right). In terms of the corresponding points (x_m, t_m) and (y_m, t_m) the reduction factor K is x_m/y_m.

In this present problem $x(t)$ is not known, we only know $y(t)$, and it would seem hopeless to ask for $\mathcal{H}$. But we do know that A and α are slowly varying functions of time, a property that provides a toehold.

A remarkable iterative approach has been found that makes use of the quadratic approximation to get an immediate improvement in the law. In Fig. 2 suppose that $y(t)$ is drawn in but that $x(t)$ and the desired law $\mathcal{H}$ are, to begin with, missing. We can start by marking the zero crossings of $x(t)$ on the down-going time axis, and we can also mark the direction of crossing. This is because we are looking for a monotonic law that passes through the origin. The slopes $x'(t)$ at $x(t) = 0$ are, however, not known, contrary to the assumption of limiting linearity, built into the quadratic law, according to which R is proportional to R_{lin} at small values. If the law were linear at small sunspot numbers, as the quadratic expression implies, then the slope of $x(t)$ at a zero crossing would be ± 1. Now, however, we would like to avoid this prejudgment and determine the actual law from the data. The reasoning proceeds as follows.

In the time interval $[-\phi/\omega, (\pi-\phi)/\omega]$ between adjacent zero crossings let $y(t)$ have an observed maximum y_m. The artificial half-period sinusoid $y_m \sin(\omega t + \phi)$ can be constructed and compared in shape with $y(t)$ in that time interval to give some information about the nonlinearity in the range 0 to y_m. If such information from all the semicycles could be combined, knowledge of the nonlinear characteristic could be sharpened. To make results from all the semicycles superposable, imagine a graph of $y(t)$ against $y_m \sin(\omega t + \phi)$ in which the abscissae for any one semicycle are reduced by the constant factor $K = [-1 + (1 + 4\beta y_m)^{1/2}]/2\beta y_m$ for that semicycle. The reduction factor comes from solving $y_m = x + \beta x^2$ and dividing by y_m. Superposing points from the 286 available years since 1700 would confirm the quadratic expression, if it were correct, because all the points would fall on the one curve. Otherwise a systematic departure would be seen which would be the basis for a corrected law.

In the simple form described this approach fails on the sunspot data because of the relation between rise time and maximum discovered by Waldmeier, as a concomitant of which the sunspot maximum does not necessarily fall midway between sunspot minima. However, given the years t_1 and t_2 of two adjacent zero crossings and t_m the year of maximum, one can set up a modulated time-varying phase

$$\phi(t) = \alpha_1(t - t_1) + \alpha_2(t - t_1)^2,$$

such that

$$\phi(t) = \begin{cases} 0 & t = t_1 \\ \pi/2 & t = t_m \\ \pi & t = t_2. \end{cases}$$

One can show that the coefficients are given by

$$\alpha_2 = \pi[t_m - \tfrac{1}{2}(t_1 + t_2)]/(t_2 - t_1)(t_2 - t_m)(t_m - t_1)$$
$$\alpha_1 = \pi/2(t_m - t_1) - \alpha_2(t_m - t_1).$$

The points in Fig. 3 show the result of plotting $R \pm (t)$ against $[-1 + (1 + 4\beta R_{max})^{\frac{1}{2}}]$ $\sin \phi(t)/2\beta$ using applicable values of R_{max} and $\phi(t)$ for each semicycle. Since the diagram will be taken as the basis for a new law defining R_{lin}, the abscissa x is labeled $R_{lin}(t)$.

- Fig. 3. The nonlinear relation between $R \pm (t)$ and $R_{lin}(t)$ put in evidence by yearly points from 1700 to 1985. The curve is $R \pm (t) = 100[| R_{lin}(t) | /83]^{3/2} \operatorname{sgn} R_{lin}$.

Inflexion effect

The rather definite functional dependence put in evidence by this first step differs only a little in general magnitude from the quadratic approximation (which is not shown), but has a noticeable qualitative difference at low values of sunspot number. One notices a distinct inflexion rather than a linear passage through the origin. Following sunspot minimum the rate of growth of sunspot number accelerates for a year or two, a feature that is not normally thought of as related to the decline of the sunspots of the preceding semicycle. In the alternating view afforded by $R\pm(t)$, the phenomenon appears as an inflexion in what would otherwise be a more or less linear crossing through zero. The rate of onset of a new semicycle should then correlate with the rate of decline of the old, a correlation for which there was indeed evidence (Bracewell, 1953) but which the inflexion effect tended to obscure.

Curve fitting

To fit a curve through the points of Fig. 3 is a classical problem. In this case a single-sided logarithmic plot (Fig. 4) tells us that a simple power law will do very well. With this knowledge one could return to nonlogarithmic coordinates and determine parameters by least squares. However, it is not likely that one could do better than the straight line fit shown on Fig. 4 which leads to the revised law

$$R = 100(|R_{lin}|/83)^{3/2}\,\mathrm{sgn}\,R_{lin}.$$

The curve drawn on Fig. 3 has been constructed from this equation and clearly is about as good a fit as one could expect, given the spread of points, although in principle further iteration could be carried out. The straight line $\log y = 1.5 \log x + 0.88$ is rather sharply defined; if the three-halves exponent is changed by more that ± 0.1 an acceptable fit cannot be found. Consequently for practical purposes of data analysis we shall adopt the exponent 3/2.

- Fig. 4. The data of the previous figure combined in one quadrant by plotting $\log |y|$ against $\log |x|$. The straight line has a slope 3/2.

The method of solution of this technical problem has proved to be most satisfactory and could be productive in applications to other astronomical and geophysical time series.

Effects of linearization

To see the effect on sunspot semicycle shapes that is produced by the three-halves law consider Fig. 5 in which four artificial sinusoids of representative strengths (middle row) are subjected to distortion to generate simulated sunspot semicycles (top row). As the semicycles become stronger they develop leptokurtosis, or a departure in the direction of greater sharpness from the "normal" condition, taken as sinusoidal. In addition, the inflexion effect in $\mathcal{H}$ at low values strongly influences the shape of the weakest semicycle to produce a profile reminiscent of the records obtained around the years 1803, 1818, 1883 and 1907. These shapes could be described as platykurtic in the dictionary sense of the word. The thing that distinguishes the three-halves law from the quadratic law is that the latter produces only leptokurtosis whereas the three-halves law gives rise to platykurtosis also, exactly as observed in the annual mean sunspot record.

- Fig. 5. Artificial "11-year" semicycle shapes versus time (top) associated with fixed-duration semisinuoids (center) of four different strengths. The artificial shapes are derived from a three-halves nonlinear law without introduction of asymmetry. The difference curves (below) resemble third harmonic, in the middle range of strength.

The difference curves (bottom row) show that the perturbation is rather like third harmonic for semicycle strengths around $R_{max} = 100$ and one can understand how Fourier analysis of $R \pm (t)$ could show a band of spectral content around a period of 22.2/3 years. However, the shapes seen for very weak and very strong semicycles show that *linear* analysis into harmonics does not do justice to the phenomenon at hand.

Waldmeier effect

For the purposes of Fig. 5 symmetrical sine waves were used, and so the asymmetry that sets in with strong semicycles was neglected. Waldmeier expressed the asymmetry quantitatively in the form of a relation between the rise time and the maximum sunspot number reached. For the present purpose it is convenient to express the asymmetry in terms of how early the peak arrives with reference to the midpoint between zero crossings. In previous work (Bracewell, 1986) this lead time was taken as $1.5(R_{max}/150)^2$ years, a simple empirical rule that is quantitatively compatible with Waldmeier's results. However, this rule makes the lead time accelerate with increasing R_{max}, which is unreasonable because the lead time must have an upper limit (of half a semicycle). The deficiency becomes apparent when sets of semicycle shapes are calculated; the shapes, for large R_{max}, do not look right. Therefore, a dependence based on a theory of the asymmetry has been attempted, to replace the empirical relation. The following derivation is based on the idea (Bracewell, 1985) that asymmetry is due to magnetic buoyancy. The only other explanation previously offered for Waldmeier effect is in terms of the eruption hypothesis, according to which an initial impetus carries the sunspot number to a greater or lesser height from which it recovers with a time constant that is the same for all semicycles. This is more descriptive than explanatory and lacks a physical basis.

Buoyancy theory for semicycle shapes

Consider an oscillatory source of horizontal azimuthal magnetic field H_ϕ at a radius r_1 inside the sun, which launches waves upward with period 22 years and a phase velocity v. At a greater radius r, nodes of H_ϕ arrive with a delay appropriate to the velocity v, but the antinodes arrive early because of magnetic buoyancy.

According to the concept of magnetic buoyancy, introduced by E.N. Parker, solar plasma permeated by a magnetic flux density B experiences a vertical buoyancy force proportional to B^2, because the magnetic pressure relieves the particle number density needed for quasihydrostatic equilibrium. We now apply this concept to very long period wave motion, keeping in mind that a central wave source is only a working hypothesis (Bracewell, 1987), and that general opinion tends to favour the relaxation oscillation originally introduced by Babcock (1961) and subsequent developments invoking sonal vortices in the convective region (Ribes et al. and surface dynamo action (Wilson, 1987).

An initially sinusoidal wave-field variation $B_0 \sin(kr - \omega t)$ in the vicinity of $r = r_1$ should suffer distortion of its sinusoidal form as it slowly propagates upward. The general effect will be that the crest will rise faster relative to the preceding and following nodes, giving the waveform a forward-leaning appearance as required. Given an initial semicycle profile shape $B_1(r)$ near $r = r_1$ at $t = t_1$, we now ask for the distorted shape $B_2(r')$ near the greater height r_2 at a later time $t = t_2$.

The following discussion, which is offered in lieu of a full theory, assumes that the 22-year wave period is so long that buoyancy forces can assert a quasiequilibrium state within any one half-wavelength layer. From conservation of magnetic flux we expect that $B_1(r)dr = (r_2/r_1)B_2(r')$ or

$$B_2(r') = \frac{r_1}{r_2} B_1(r) \frac{dr}{dr'}.$$

The very long time scale means that viscous drag dominates inertial forces in the radial direction. The flux element $B(r)dr$ might then rise with an excess velocity proportional to pressure $[B_1(r)]^2$. Thus $r' = r + r_2 - r_1 + c[B_1(r)]^2$, where c is a constant, and $dr'/dr = 1 + 2cB_1(r)B_1'(r)$. So

$$B_2(r) = \frac{r_1}{r_2} \frac{B_1(r)}{1 + 2cB_1(r)B_1'(r)}.$$

Taking $B_1(r_1) = A \sin[k(r - r_1)]$, $r_1 < r < r_1 + \frac{1}{2}\lambda$, we find that an initially sinusoidal shape becomes

$$B_2(r) = \frac{r_1}{r_2} \frac{A \sin[k(r - r_2)]}{1 + 2cA \sin[k(r - r_2)kA \cos[k(r - r_2)}$$

$$= \frac{r_1}{r_2} \frac{A \sin[k(r - r_2)]}{1 + ckA^2 \sin[2k(r - r_2)]}, \qquad r_2 < r < r_2 + \tfrac{1}{2}\lambda.$$

In this derivation it is supposed that the distortion is controlled solely by the initial profile; but as propagation proceeded, the peak flux density would change and the velocity excess would alter. There would be a tendency to increased flux density due to bunching but this might be outweighed by an opposing tendency represented by the factor r_1/r_2 associated with spherical divergence. Since the total deformation called for is only about 1.5 years or so in 22, it has not seemed worthwhile in a first investigation to refine the theory for the waveshape. By taking excess velocity to be proportional to buoyancy forces, as though viscosity were the controlling resistance to motion, we reach results that will be seen below to be satisfactory. The proportionality constant c will in time be derivable from theory, pending development in assessing the magnitudes of r_1 and $B_{max}(r_1)$ and conversely will constrain the range of these magnitudes. Meanwhile, c has been adjusted to meet a coarse requirement of 1.5 years advance in the vicinity of 100 to 150 for R_{max}.

The general effect of buoyancy in a magnetized plasma somewhat resembles the waveshape effect on intense sound waves or shallow-water waves, where the crests travel faster than the nodes and produce wavefront steepening. However, unlike the troughs of the mechanical waves, which are retarded, both in-phase and antiphase semicycles of a magnetic wave are accelerated. The calculated profiles of Fig. 6 show the nature of the phenomenon described by the simple buoyancy theory presented. Noteworthy features are the wavefront steepening, peak sharpening and strengthening, and a break where the descent from the peak moderates into a gentler slope. For this calculation, $k = 1$, $c = 0.000048$, $r_1/r_2 = 1$, and $A = 25, 50, 75, 100$.

- Fig. 6. A sinusoidal profile $A \sin r$ becomes $A \sin r/(1 - cA^2 \sin 2r)$ when buoyancy is allowed for. As the initial amplitude A passes through increasing values (25, 50, 75, 100) the transmitted wave steepens in front, develops a break on the descending side, and the peak sharpens and intensifies substantially.

As may be seen, the calculated profiles do not closely resemble actual sunspot semicycles but this is because the three-halves law has not yet been applied. Fig. 7 (left) incorporates this law and also adjusts for the observed relation $\Delta t = -0.042(R_{max} - 72)$ for the departure Δt in semicycle duration from the mean in accordance with the strength of the semicycle (Bracewell, 1985). In addition, the maxima have been aligned in time, for comparison with the right hand part of the figure (due to Waldmeier, 1935).

The range of semicycle shapes provided by applying the nonlinear law and the buoyancy effect to semisinusoids is compared with the full range of historic semicycles in Fig. 8. The semicycles are arranged in horizontal rows according to amplitude. Although there is a correlation between duration and amplitude, considerable variation in duration is apparent within each amplitude group. This is because of other significant influences on duration, for example, strong odd-numbered semicycles in recent decades have been extended in duration by a noticeable negative offset in the 22-year mean of $R_{\pm}(t)$. This offset is analogous to the additive 350-year undulation (Williams, 1981, 1985) evidenced by zig-zag effect in the rock record. So a semisinusoid was chosen to pass through the start and finish points of each semicycle and then adjusted in accordance with $R_{\max}$ for that semicycle by applying the nonlinear law to the ordinate and the buoyancy shift to the abscissa.

In addition to the wavefront steepening and peak sharpening noted in the stronger semicycles, an initial fillet is produced in semicycles of all strengths. This feature used to present a difficulty for advocates of the alternating cycle because the crossovers were kinked, or inflected (see Bracewell, 1953 or Fig. 2 of Bracewell, 1985), but now the initial fillet is explained by the three-halves nonlinear law. The break in the descent is seen to result from buoyancy in strong semicycles (a bonus of the buoyancy calculations, which were introduced for another purpose). It is noticeable that many of the observed semicycles have stronger breaks than the simulations, which suggests a test for refinements to the buoyancy theory.

It is clear that the gross features of the semicycle shapes are in accord with the one-parameter theoretical shapes in the majority of cases. The discrepancies may originate in various ways. First of all the sunspot observations have their own systematic but unknown errors. Occasional semicycles have secondary peaks (more apparent in the monthly than in the annual numbers) which are not simulable with one adjustable parameter R_{max}. Also, any modern-day 350-year additive undulation $U(t)$, which would increase the semicycle length and enhance R_{max} where $U(t)$ and $R_{\pm}(t)$ had the same sign (Bracewell and Williams, 1986) has not been compensated, nor has account been taken of the tendency of sharp and blunt cycles to alternate (also attributable to undulation). On the whole, the calculated shapes, including the year of maximum, agree with the record rather well within the observational precision implied by the level of irregularities ("noise").

- Fig. 8. Comparison between recorded annual mean sunspot numbers (thin) and artificial profiles (thick), whose shapes, including the timing of the maximum, are determined by the peak amplitude.

Eighteenth century data

It would take a bold critic to claim that the 18th century profiles, in the left hand column, are less well represented than the more recent profiles. In fact, both aspects of kurtosis, the peakiness of strong semicycles and the inflexion effect, as well as the buoyancy effect, are all clearly exhibited, and speak in support of the general reliability of the Zürich compilation, confirming and extending an earlier discussion by Gleissberg (1960).

The semicycles of 1766, 1775, and 1798 are reproduced rather well. The semicycle starting in 1755 is generally symmetrical, but had a peak in 1761 that stood well above the preceding and following years; this peak, to which the simulated profile is matched, forces an early simulated maximum. If the data reports are erroneous, one possibility is that the sunspot numbers for the one year 1761 were less, by about 40 per cent, than reported. If that single year were altered, the simulated profile would fit the general run well. However, it is not unreasonable to accept that 1761 was a spotty year; other years have been seen when the sunspot number exceeded the mean of the preceding and following years by comparable amounts. A second possible explanation in terms of erroneous reports is that the two years 1758-1760 were spottier than reported by about 40 per cent.

The semicycle starting in 1784 is unusual because of its association with the great phase anomaly centered approximately on that year (Dicke, 1970, 1978; Bracewell, 1985). The reported maximum came conspicuously early but it is hard to see how this could be due to observational error. Possibly the unveiling of each semicycle proceeds at a rate that is advanced or retarded on occasion by convective motion in the subphotospheric medium. The shape described by the reports is quite classical and would closely fit a profile lasting 12 years instead of 14 years. The crossover year between semicycles 4 and 5 is subject to interpretation and in particular would be subject to correction for the 350-year undulation. The sense of the correction would be to make the crossover year of start later and of finish earlier. Perhaps the discrepant simulation for semicycle 4 is evidence for the presence of a positive undulation, which is expected, but not clearly evident

(Fig. 2 of Bracewell, 1985) in those years.

Origin of the nonlinearity

Because the nonlinearity discussed here was found in both the annual mean sunspot series (Bracewell, 1985) and also in the varve series (Bracewell and Williams, 1986) it is reasonable to look for a physical explanation in the sun. Terrestrial atmospheric explanations could be devised for either case taken alone but such explanations would call for an influence of astronomical seeing on the counting of sunspots and sunspot groups on the one hand and a compatible influence of climate on varve thickness on the other. Turning therefore to the sun we look first at the definition of the annual sunspot number. It is the mean of the daily values of $k(10g + f)$, where g is the number of groups and f is the number of spots counted by an observer whose personal factor k is chosen to establish historical consistency.

There is no reason to believe that sunspot number as conventionally defined should be proportional to its cause as measured in fundamental units of physics. It is of course conceivable that, if the underlying influence that causes sunspots were to double, then the number of groups would double in response. On the other hand, the groups might enlarge but not increase in number; and since it is known that the number of spots per group increases rapidly with area of the group (Kiepenheuer, 1953), here is a possible source of nonlinearity built into the sunspot number definition. This mechanism could be quantified. Meanwhile, there is a different nonlinearity that appears to be sufficient to explain the three-halves law.

Suppose that there is an underlying physical influence that manifests itself as a center of activity on the solar surface and that the "strength" of this center is directly proportional to the influence I, by definition. Since active centers erupt at intervals in various solar latitudes and longitudes and have finite lifetimes, the influence is sporadic. Suppose that the ith sporadic episode is quantified by influence I_i so that ΣI_i, summed over a year, could be the physical strength measure of that year's activity. Now consider how solar activity is characterized by counting sunspots.

Each active center is accompanied by a sunspot group whose temporal development is measured by $r_i(t)$, the day-by-day contribution of that sunspot group to the daily sunspot number $R_d(t) = \Sigma_i r_i(t)$.

Call a graph of $r_i(t)$ versus t in days a sunspot development curve. Such curves have been much studied; for example the mean duration is 6 days and $\langle r_i(t)\rangle = 12$. Detailed shapes of the curves have been sorted into categories, the sunspot groups have been classed, and the time dependence on class has been examined (Kiepenheuer, 1953).

The annual mean sunspot number $R(t)$ is defined by $(1/365)\sum_1^{365} R_d(t)$. Now $R_d(t)$ is itself a sum over the active centers; thus

$$R(t) = (1/365)\sum_{t=1}^{365}\sum_i r_i(t).$$

By inverting the order of summation we can express this definition in terms of the areas $A_i = \sum_{t_{1_i}}^{t_{2_i}} r_i(t) = \sum_1^{365} r_i(t)$, where t_{1_i} and t_{2_i} are the day numbers for the birth and death of group i and $r_i(t)$ is deemed to be zero before t_{1_i} and zero after t_{2_i}. Thus

$$R(t) = (1/365)\sum_i\sum_{t=1}^{365} r_i(t) = (1/365)\sum_i A_i.$$

The right-hand side, through the summation over the groups i, represents a mean that is weighted, very reasonably, in accordance with the frequency of each class of sunspot group; in addition weight is given both to sunspot size and to longevity through the contribution to area A_i under the development curve. We see that $R(t)$, while it is certainly a measure of the year's spottiness, is not at all a measure of the year's activity as expressed by ΣI_i. To take an analogy, ΣI_i is like the annual rainfall, while ΣA_i is like the annual mean depth of rainwater standing in a pond; one heavy fall of rain will contribute more to the mean depth of water than the same amount of rain falling in showers. In other words, when an episode is big as quantified by I_i then not only its size, but also its duration is factored into A_i. For this reason alone, $R(t)$ in years of large annual sunspot number will overestimate ΣI_i simply because, by and large, big sunspots are more durable.

To estimate the quantity ΣI_i we would need to take all the sunspot group development curves for the year and sum their amplitudes, i.e. their contribution to the sunspot number on the day when that contribution to the sunspot number was a maximum. However, since much statistical analysis is on record, we can simply plot the areas $\overline{A}$ of different sunspot development curves against $\overline{R}_{\max}$ the maximum sunspot number reached, where the bars indicate averages taken within the different categories. The results deduced from Kiepenheuer (1953) are plotted in Fig. 9.

- Fig. 9. The number of sunspot-days $\overline{A}$ representing the mean area under different categories of sunspot group development curves as a function of $\overline{R}_{\max}$, the maximum contribution to sunspot number reached in the course of development averaged over the same category. The straight line has a slope of 3/2.

We see that on a log-log plot of $\overline{A}$ versus $\overline{R}_{\max}$ the 12 points lie near a straight line whose equation is

$$\overline{A} = 3\overline{R}_{\max}^{3/2}.$$

This result could be refined by weighting the points in accordance with the frequency distribution of the different categories of development curves. Meanwhile, it is extremely striking that the same three-halves law emerges directly from the available statistical studies as was arrived at from the third-harmonic investigations. One concludes that the gross phenomenon of overcounting sunspots in accordance with the duration of their groups is sufficient to explain the bulk of the nonlinear phenomenon that was discovered.

It ought to be added that the linearized variable R_{lin}, that has been dignified by terms such as "physical" and "underlying", and which is the goal of a deliberate quest for something "fundamental," by no means damages the physical status of the integrated quantity $R(t)$, which takes account of duration. The latter is after all the quantity which correlates with ionospheric characteristics for the simple reason that sustained solar ultraviolet radiation from a long-lived centre of activity continues to produce ionization as long as it is present. Assume that varve thickness represents the annual run-off from a periglacial region into a lake, and that the run-off in turn measures the integrated solar ultraviolet which, not being impeded by today's oxygen atmosphere, could reach down to the troposphere in Precambrian times. Then it is the sunspot number $R(t)$, not $R_{lin}(t)$, with which varve thickness would be expected to correlate; the presence of the nonlinear effect in varve thickness is consistent with this expectation. Similarly, annual rainfall, which is like $R(t)$, and the nonlinearly related mean depth of a rain-fed pond, which is like $R_{lin}(t)$, are each valid in their own applications.

Sunspot area

Authors have often spoken respectfully of sunspot area, a measurement that possesses long term stability, while nevertheless using sunspot number because it has been available in real time, whereas the *Greenwich Photoheliographic Results* used to appear many years late. For the years 1784 to 1938 it was found that sunspot area F, measured in millionths of the solar hemisphere, was related to the sunspot number R by $F = 16.7R$. But in fact F is not proportional to R. Data from both Greenwich and the U.S. Naval Observatory (reproduced by Waldmeier, 1968) respectively fit the following power laws: $F = 1600(R/120)^{1.25}$ and $3400(R/180)^{1.2}$ (Fig. 10). Consequently sunspot area should be proportional to $R_{lin}^{1.85}$ approximately and exhibit even greater departure from the underlying influence than does the sunspot number itself. The explanation for the nonlinear relation between sunspot area and sunspot number clearly lies with the way in which sunspot groups of different total area are divided into separate spots. One could rederive the linearized sunspot number $R_{lin}(t)$ from sunspot area records, and possibly obtain a better result than working from sunspot numbers allows.

- Fig. 10. Data for annual mean sunspot area $\bar{F}$ versus annual mean sunspot number $\bar{A}$ (Greenwich below, Washington above) showing a clearly nonlinear relation.

Conclusion

A nonlinearity entering into the definition of sunspot number has been demonstrated and explained. For purposes of physical theory it will be desirable in future to correct sunspot numbers for the nonlinear relationship that is now known to exist. As an example of such a correction, combined with allowance for buoyancy in a magnetized plasma, it has been possible to reproduce theoretically the known Waldmeier relation between rise time and maximum sunspot number, and other features of the 11-year sunspot curve. A break in the decline of strong 11-year cycles, which has been demonstrated to exist, is predicted by the same theory. Eighteenth century sunspot numbers fit theoretical semicycle shapes quite well; their reputed unreliability is not confirmed by the present work. The physical model envisages an upward travelling 22-year wave launched by a reciprocating magnetic oscillator deep in the sun. While this model is at present incomplete, it is qualitatively compatible with the fast rise times associated with strong 11-year cycles, and with other features of the cycle shape discussed in earlier papers, and to this extent is supported by observation.

Acknowledgments

Preparation of this manuscript was supported by NASA grant 05-020-272 and ONR contract N00014-85-K-0111.

References

Bracewell, R.N., (1953). *Nature*, 171, 649.

Bracewell, R.N., (1985). *Aust. J. Phys.*, 38, 1009.

Bracewell, R.N. & Williams, G.E., (1986). *Mon. Not. Roy. Astron. Soc.*, 223, 457.

Bracewell, R.N., (1986). *Nature*, 323, 516.

Bracewell, R.N., (1987). *Solar Physics*, (in press).

Cole, T.W., (1973). *Solar Physics*, 30, 103.

Dicke, R.H. (1970), "The rotation of the Sun," in *Stellar Rotation*, A. Slettebak ed., Reidel, Dordrecht-Holland.

Dicke, R.H., (1978). *Nature*, 276, 676.

Gleissberg, W., (1960). *Naturwiss.*, 47, 197.

Kiepenheuer, K.O., (1953). *The Sun*, G. Kuiper, ed., Chicago University Press, p. 322.

Waldmeier, M., (1935). *Astr. Mitt. Zürich*, 133, 105.

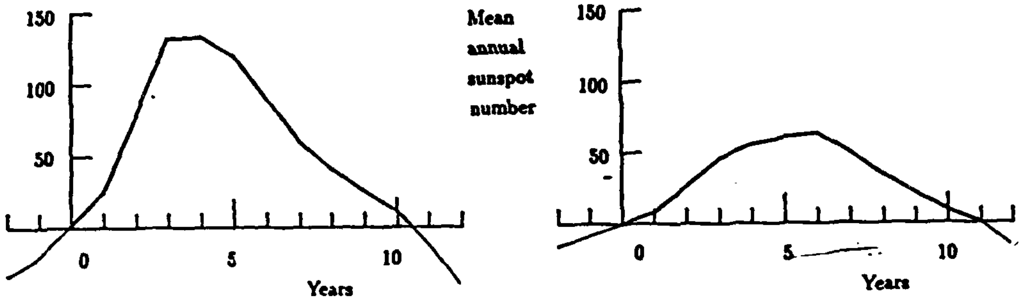
Waldmeier, M., (1968). *Astr. Mitt. Zürich*, 358, 23.

Williams, G.E., (1981). *Nature*, 291, 624.

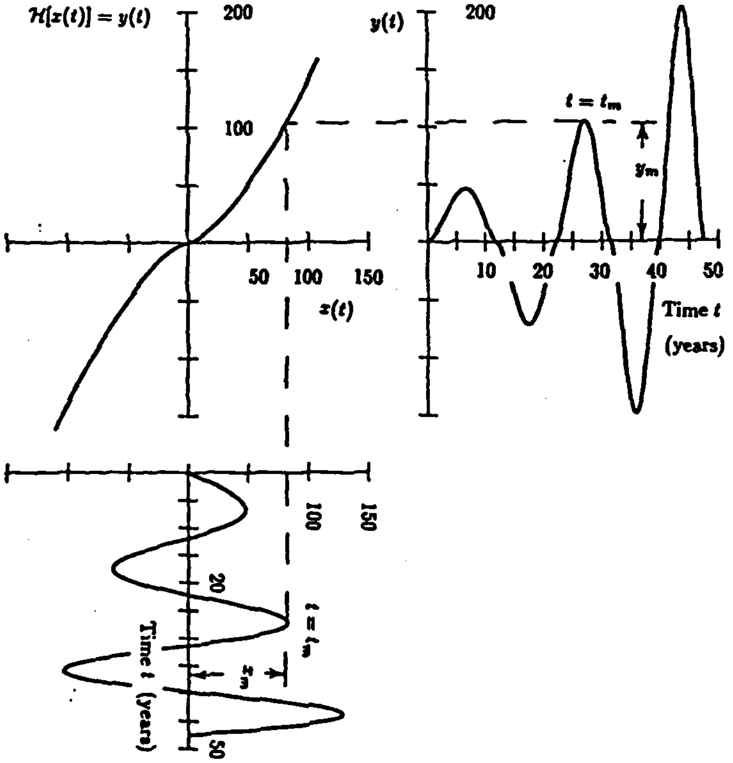
Williams, G.E., (1985). *Aust. J. Phys.*, 38, 1027.

Wilson, P.R., (1987). *Solar Physics*, (in press).

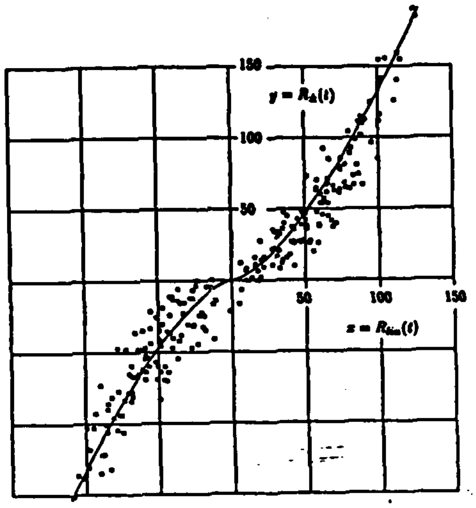
Ribes, E., Mein, P.& Mangeney, P.,(1985) Nature, 318, 170



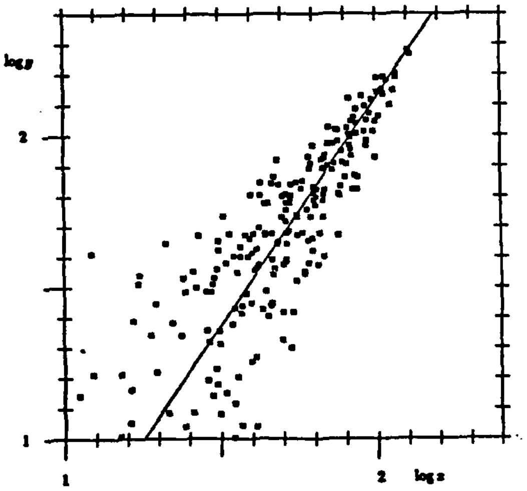
• Figure 1. Features of the shape of a strong semicycle (left) include a leading fillet, a convex segment peaking early, and a relatively long decline, sometimes with a noticeable break in slope. A weak semicycle (right) tends to be much more symmetrical.



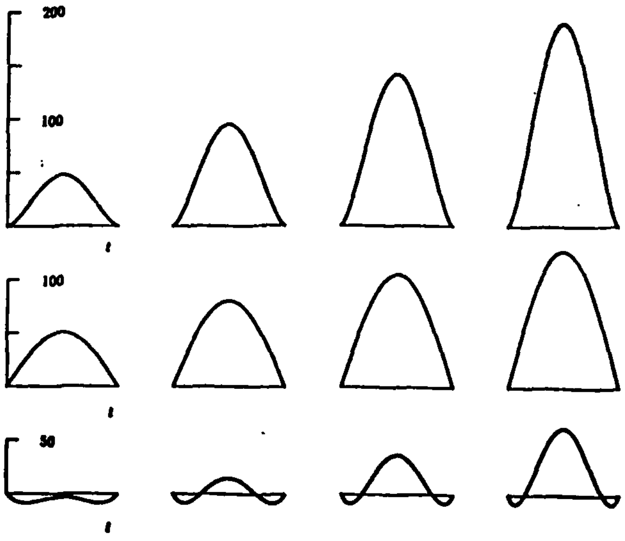
• Figure 2. A nonlinear characteristic (top left) relates a stimulus $x(t)$ (lower left) to a response $y(t)$ (upper right). In terms of the corresponding points (x_m, t_m) and (y_m, t_m) the reduction factor K is x_m/y_m.



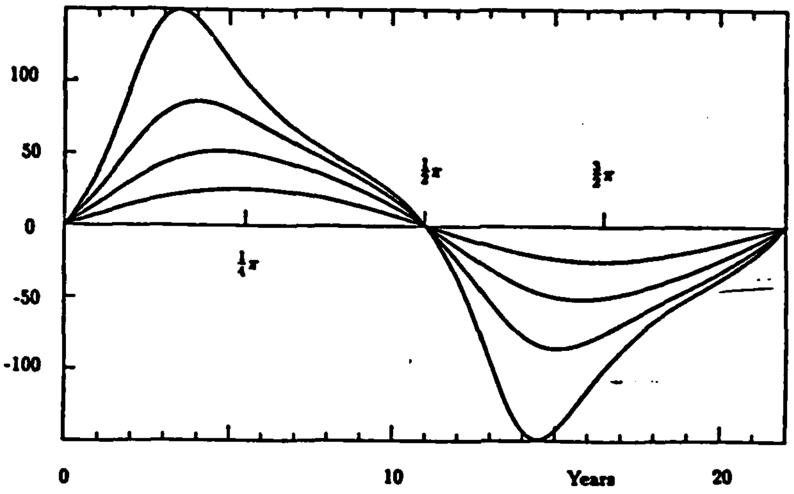
• Figure 3. The nonlinear relation between $R \pm (t)$ and $R_{lin}(t)$ put in evidence by yearly points from 1700 to 1985. The curve is $R \pm (t) = 100[| R_{lin}(t) | /83]^{3/2}$ sgn R_{lin}.



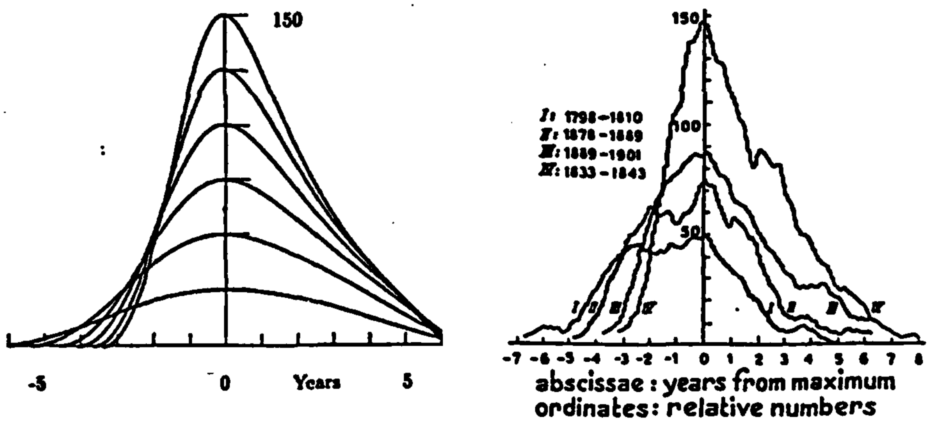
• Figure 4. The data of the previous figure combined in one quadrant by plotting log $|y|$ against log $|x|$. The straight line has a slope 3/2.



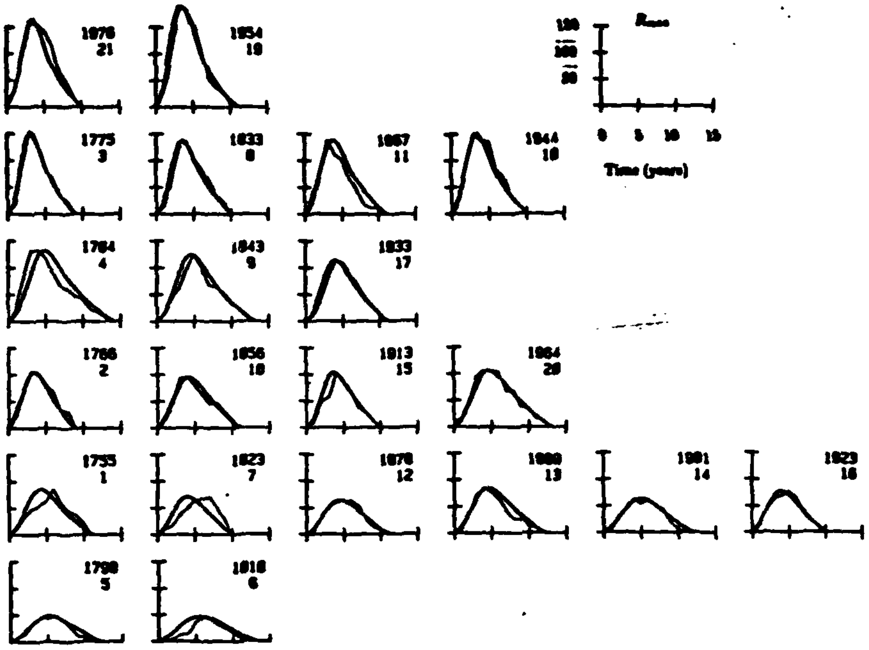
• Figure 5. Artificial "11-year" semicycle shapes versus time (top) associated with fixed-duration semisinuoids (center) of four different strengths. The artificial shapes are derived from a three-halves nonlinear law without introduction of asymmetry. The difference curves (below) resemble third harmonic, in the middle range of strength.



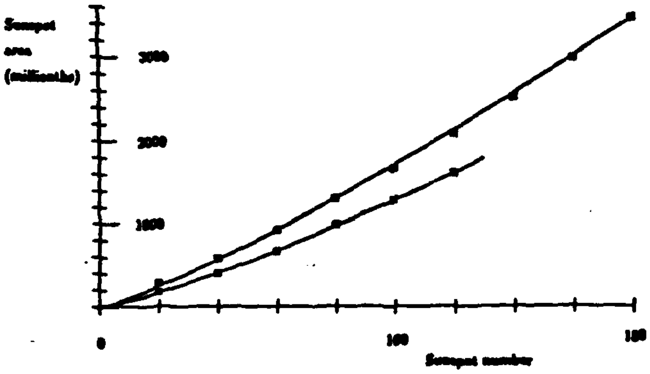
• Figure 6. A sinusoidal profile $A \sin \tau$ becomes $A \sin \tau/(1 - cA^2 \sin 2\tau)$ when buoyancy is allowed for. As the initial amplitude A passes through increasing values (25, 50, 75, 100) the transmitted wave steepens in front, develops a break on the descending side, and the peak sharpens and intensifies substantially.



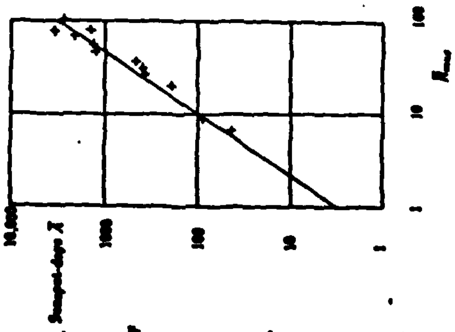
• Figure 7. Artificial semicycle shapes incorporating (a) nonlinearity, (b) an advance in maximum associated with magnetic buoyancy, and (c) a shortening of the semicycle duration inversely correlated with semicycle strength (left). For comparison, some selected semicycles from history (right). Note that the historical semicycles fail to reach zero because spots from the preceding or following semicycles are included.



• Figure 8. Comparison between recorded annual mean sunspot numbers (thin) and artificial profiles (thick), whose shapes, including the timing of the maximum, are determined by the peak amplitude.



• Figure 9. The number of sunspot-days $\overline{A}$ representing the mean area under different categories of sunspot group development curves as a function of $\overline{R}_{max}$, the maximum contribution to sunspot number reached in the course of development averaged over the same category. The straight line has a slope of 3/2.



• Figure 10. Data for annual mean sunspot area F versus annual mean sunspot number R (Greenwich below, Washington above) showing a clearly nonlinear relation.

END

10-87

DTIC