THE ASTROPHYSICAL JOURNAL, 513:990–996, 1999 March 10

ON THE SOLAR-CYCLE MODULATION OF THE HOMESTAKE SOLAR NEUTRINO CAPTURE RATE AND THE SHUFFLE TEST

GUENTHER WALTHER
Department of Statistics, 370 Serra Mall, Stanford University, Stanford, CA 94305; walther@stat.stanford.edu
Received 1998 July 6; accepted 1998 October 13

ABSTRACT

There exists no significant correlation between the Homestake neutrino data up to run 133 and the monthly sunspot number, according to a test that is based on certain optimality properties for this type of problem. It is argued that priorly reported highly significant results for segments of the data are due to a statistical fallacy: the usual methods for evaluating the significance of common tests for correlation are not applicable in the sunspot-neutrino context. Moreover, an appropriate evaluation of these tests gives results that are compatible with the hypothesis of no correlation. Some new methods are introduced for assessing the significance of common measures of correlation in a time series setting, with a special emphasis on the Spearman rank correlation coefficient.

Subject headings: methods: statistical — Sun: activity — Sun: particle emission — sunspots

1. INTRODUCTION

The Homestake solar neutrino experiment has been monitoring the flux of neutrinos produced in the core of the Sun for over 25 years. The results deduced from this experiment have been a puzzle in several ways: the inferred neutrino flux is several times smaller than those calculated from solar models (the "solar neutrino problem"). Further, an apparent association between the inferred solar neutrino flux and various indicators of solar activity such as the sunspot number has stimulated a considerable amount of research on that topic (see, e.g., Davis 1996; Bahcall, Field, & Press 1987; Bahcall & Press 1991; Bieber et al. 1990; Basu 1982; Delache et al. 1993; Dorman & Wolfendale 1991; Krauss 1990; Massetti & Storini 1993; McNutt 1995; Oakley et al. 1994; Raychaudhuri 1986, 1991). The perceived anticorrelation has motivated proposals in which neutrinos have a much larger magnetic moment than is implied by standard electroweak theory (Cisneros 1971; Okun 1986; Voloshin, Vysotskii, & Okun 1986; Voloshin & Vysotskii 1986). This article addresses the statistical significance of this anticorrelation. It has two goals: first, to expound the claim—briefly reported in Walther (1997)—that the reported statistically highly significant results concerning the anticorrelation are due to a statistical fallacy and that the data are in fact consistent with no correlation, and, second, to put forth some new statistical techniques for assessing the statistical significance of certain nonparametric measures of correlation in a time series context. A particular aim of this article is to show that the highly significant results found in earlier segments of the data disappear when the same statistical tests used in those analyses are evaluated in a proper way.

Section 2 describes the data used. Section 3 explains in nontechnical terms some important points concerning measures of correlation, which will be relevant when investigating the neutrino-sunspot correlation. In Section 4 a result for the Spearman rank correlation coefficient in a time series context is presented, and some known and new methods are introduced to evaluate the significance of nonparametric measures of correlation in this setting, focusing on the Spearman rank correlation coefficient. Section 5 investigates in some detail the evidence for a neutrino-sunspot correlation, both for all the Homestake runs up to run 133 and for segments of these data for which highly significant results have been reported. The conclusions are summarized in § 6.

The article is written in a nontechnical way and should be understandable to readers with a basic statistical knowledge. The only exception may be § 4, which can be skipped by a reader who is only interested in the neutrino-sunspot correlation.

2. THE DATA

The neutrino data of the Homestake experiment (Davis 1996) consist of a run number, and, for each run, of a start and stop time for the run, a mean ^{37}Ar production rate in units of atoms per day, and a lower and upper 68% confidence limit of the production rate. The confidence limits pose problems for the analysis of the data: They need not correspond to 1 σ error bounds, and the lower bound was set to zero in those cases where the data analysis provided a negative value (due to the subtraction of the background rate). Further, the confidence limits have recently been revised by the Homestake team. The effect of these issues on the statistical analysis will be addressed at the appropriate places below. The Homestake data used in this article were generously made available by K. Lande (1996, private communication) and consist of the following 108 runs: runs 18–133, except runs 23, 25, 26, 34, 90, 93, 117, and 123, for which no data are provided owing to various problems with the experiment.

The sunspot data used are the monthly sunspot numbers provided by the National Geophysical Data Center of the US Department of Commerce.

3. MEASURES OF ASSOCIATION AND THEIR ASSUMPTIONS

This section explains the fallacy with simple examples. The relevance of the various issues raised here to the neutrino-sunspot context is discussed in § 5. For simplicity, this section concentrates on the most commonly used measures of correlation between random variables X and Y when n pairs $(X_1, Y_1), \ldots, (X_n, Y_n)$ of these quantities are observed: the Pearson's product-moment correlation coefficient r, and its nonparametric counterpart, the Spearman rank correlation coefficient r_S. Similar conclusions to those exhibited in the following apply also for other measures of

correlation and association, such as Kendall's τ or the χ^2 statistic for contingency tables in the case of nominal of grouped data. The Spearman coefficient r_S is attractive because it is based on the ranks of the data and hence its null distribution (the distribution of r_S when no correlation between X and Y is present) does not depend on the distribution of the data. Thus one can obtain significance levels in quite general situations with few assumptions. There are *some* assumptions, however, and those have to be checked in each situation at hand. Usually textbooks present the null distribution of r_S only for the case where the (X_i, Y_i) pairs are independent and identically distributed. While it will be seen shortly that this assumption can be somewhat weakened, a simple Monte Carlo study shows that, e.g., a serial dependence structure in each variable, typical for a time series analysis, can have drastic consequences: 10^5 simulations of 200 independent standard normal random variables $X_1, \ldots, X_{100}, Y_1, \ldots, Y_{100}$ were generated, and for each simulation the two random walks $S_k = \sum_{i=1}^{k} X_i$ and $T_k = \sum_{i=1}^{k} Y_i$, $k = 1, \ldots, 100$, were computed. Clearly, the two series S_k and T_k are independent, but the Spearman rank correlation coefficient rejected the null hypothesis of independence at the 1% significance level for 66.7% of the simulations!

The distribution of r_S under the null hypothesis of independence is derived using the assumption that each pairing of the ranks of the Xs with any permutation of the ranks of the Ys is equally likely (Bickel & Doksum 1977). This yields as minimal assumption for the applicability of the exact or asymptotic null distribution of r_S (see, e.g., Table 8 in Bickel & Doksum 1977 for the former, and the t-approximation following eq. [3] for the latter) that at least one of the two series, say Y, is permutation invariant, meaning that

$$\begin{aligned} &\text{the distribution of } (Y_1, \ldots, Y_n) \\ &\quad \text{is the same as the distribution of } (Y_{\pi(1)}, \ldots, Y_{\pi(n)}) \\ &\quad \text{for any permutation } \pi \text{ of the integers } 1, \ldots, n\,. \end{aligned} \tag{1}$$

The Monte Carlo study summarized in Table 1 shows how easily one is led to an erroneous claim of a significant correlation between the sunspot numbers and an independent random series that violates condition (1) in commonly encountered ways. Each row of Table 1 treats a different type of independent time series X and Y. Each of the random series is simulated 10^5 times, and the columns give the relative frequency of rejection of the null hypothesis of independence at nominal significance levels 5%, 1%, and 0.1%, using the t-approximation for the transformation (3). Other tests for correlation referred to in this section produce qualitatively similar results. All significance levels refer to two-sided tests. In rows 1 through 4, X is taken to be the series of monthly sunspot numbers of length 100 starting in 1970 January.

Row 1 uses 100 independent standard normal random variables for the series Y. Y satisfies condition (1), so the observed relative frequencies of rejection should be equal to the nominal levels, apart from Monte Carlo simulation error and the error due to the t-approximation for the transformation (3). Thus row 1 serves as a check on these approximations. In rows 2 and 3, Y is a three-point and six-point running mean of independent standard Gaussian random variables. The dependence structure in Y violates condition (1), and the simulation results show how this leads to erroneously significant results. (The time series of sunspot numbers clearly does not satisfy condition [1], a proposition corroborated by these simulations!)

A heuristic explanation of this aspect of the fallacy can be given as follows: the null distribution of r_S (the distribution of r_S when no correlation between the series is present) depends on the number n of observations. If more data are available then r_S will be concentrated more closely around zero. Indeed, the standard deviation of r_S is approximately $n^{-1/2}$; see § 4. If there is a serial dependence between the observations, then one can think of the n pairs of dependent data as containing only as much "information" as a smaller number, say $n/4$, of independent data. Hence when assessing the significance of r_S, one should compare it to the null distribution pertaining to a sample of size $n/4$, not n. The latter distribution is concentrated more closely around zero than the former, so r_S looks more significant than it really is.

The above heuristic explanation will provide the basic idea for developing a technique in § 4 for assessing the significance of r_S in such situations. Comparing rows 2 and 3 of Table 1 shows that a larger degree of smoothing leads to seemingly more significant results, as expected from the heuristic. This phenomenon will be relevant in the solar neutrino context; see § 5 below.

An important consequence of the above heuristic is illustrated in Figure 1: the scatter plot in Figure 1*a* shows the first 100 of 109 typical independent observations $(X_1, Y_1), \ldots, (X_{109}, Y_{109})$ from a standard bivariate normal distribution. Figure 1*b* shows the plot of the 100 running

TABLE 1

RELATIVE FREQUENCIES OF REJECTION OF NULL HYPOTHESIS OF INDEPENDENCE AT VARIOUS NOMINAL SIGNIFICANCE LEVELS[a]

TIME SERIES	RELATIVE FREQUENCY OF REJECTION AT NOMINAL SIGNIFICANCE LEVEL (%)		
	5%	1%	0.1%
$X =$ sunspot numbers, $Y = (Z_1, \ldots, Z_{100})$	4.9	1.0	0.12
$X =$ sunspot numbers, $Y_k = \sum_{i=k}^{k+2} Z_i$ $(k = 1, \ldots, 100)$	23.9	12.0	4.7
$X =$ sunspot numbers, $Y_k = \sum_{i=k}^{k+5} Z_i$ $(k = 1, \ldots, 100)$	39.7	26.6	15.6
$X =$ sunspot numbers, $Y_k = 1 + Z_k$ if $T_{2i} < k \leq T_{2i+1}$, $Y_k = 3 + Z_k$ otherwise $(k = 1, \ldots, 100)$	35.7	22.4	11.7

[a] Results of a Monte Carlo study using the nominal null distribution of Spearman's correlation coefficient. The random series in each row were simulated 10^5 times. X is the series of the 100 monthly sunspot numbers starting in 1970 January. The Z_i are independent standard normal random variables. T_i is the sum of the first i terms of a sequence of independent exponential random variables with mean 10 months.

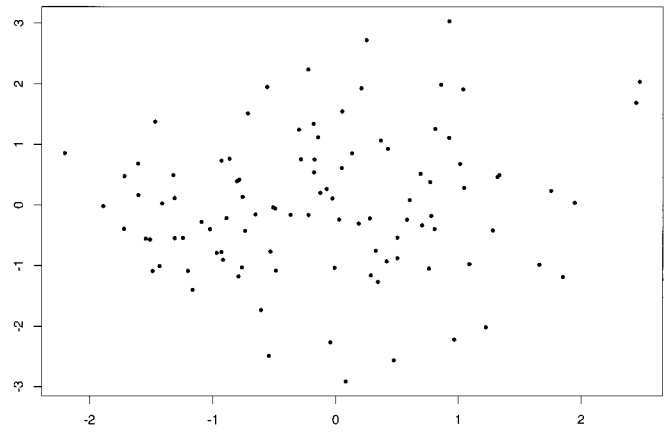

FIG. 1*a*

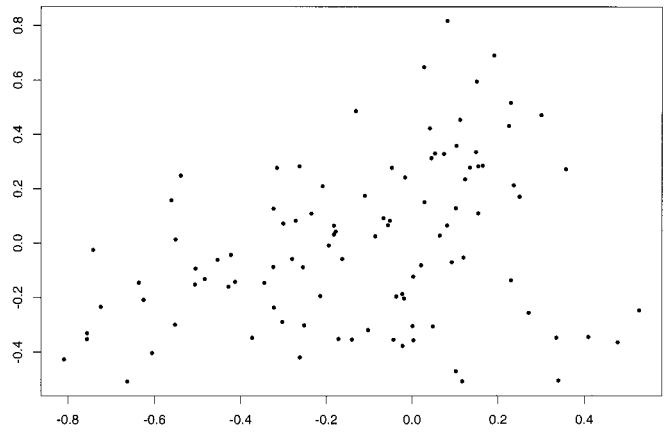

FIG. 1*b*

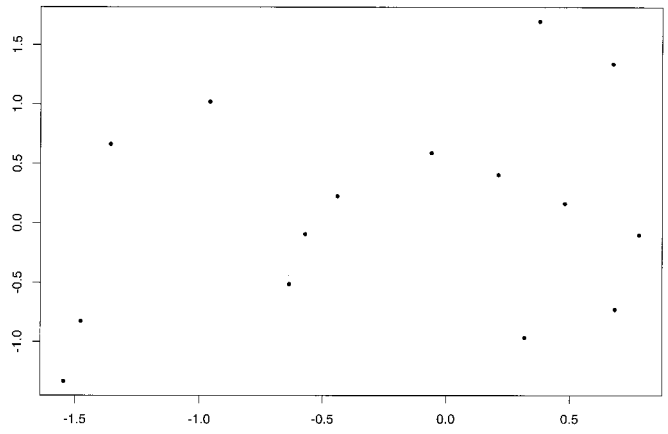

FIG. 1*c*

FIG. 1.—(*a*) Scatter plot of 100 independent standard bivariate normal observations. (*b*) Scatter plot of the running means of length 10 of the data in (*a*). (*c*) Scatter plot of 15 independent standard bivariate normal observations. Pearson's correlation coefficient is 0.12 for the plot in (*a*) and 0.30 for plots in (*b*) and (*c*).

means of length 10, $(\frac{1}{10}\sum_{i=k}^{k+9} X_i, \frac{1}{10}\sum_{i=k}^{k+9} Y_i)$, $k = 1, \ldots, 100$. The correlation is visibly larger in Figure 1*b* (Pearson's r equals 0.30) than in Figure 1*a* ($r = 0.12$). When confronted with Figure 1*b*, most people would consider it quite unlikely that such a correlation is due to chance alone. But this is just a psychological effect: the brain is calibrated by scatter plots consisting of independent pairs. But an appropriate comparison plot would consist of only about 15 such pairs, not 100, as will be seen below. Figure 1*c* shows such a sample of 15 independent Gaussian pairs with $r = 0.30$. Now the correlation looks much less convincing. Indeed, the probability of obtaining values of $|r|$ of at least the observed size is *larger* for the situations in Figures 1*b* and 1*c* (about 27%) than for that in Figure 1*a* (24%), as can be verified by simulations. This example shows that the scatter plot, usually the first tool used to examine bivariate data, can be quite misleading in commonly encountered situations.

The null distribution of Pearson's r (see, e.g., [6.5.3] in Bickel & Doksum 1977) is derived under stronger assumptions than condition (1). Its asymptotic behavior is well known also when certain types of serial correlations are present, such as in the Gaussian running means case above. But these results are usually presented only in time series books (see, e.g., Theorem 11.2.2. in Brockwell & Davis 1987). The above sample-size calculation was done using that theorem.

Nonparametric methods such as r_S or Kendall's τ are usually not treated in a time series context, and the results in § 4 seem to be new. Common textbooks give no warning that the null distributions depend sensitively on assumptions such as condition (1). Indeed, conversations with established senior statisticians have shown that the effect of serial correlation on statistical procedures seems not well enough appreciated even within the statistical community.

There are other important ways in which condition (1) can be violated, even if the data are not smoothed. For example, different standard deviations of the Y_i or different means are incompatible with condition (1). An example for the latter case is given in row 4 of Table 1, which shows that changes in the mean neutrino flux that are completely unrelated to the solar cycle can lead to the appearance of a strong correlation: the neutrino flux is taken to be constant, equal to 1, for a random time which is distributed exponentially with mean 10 months; then the flux is set to 3 for a random time with the same distribution; then it is set back to 1; etc. Each month the flux is measured with independent standard Gaussian measurement errors. Table 1 shows how the null distribution of r_S erroneously gives highly significant results, even though the simulated observations are uncorrelated with the sunspot number by design (the choice of 10 months for the mean waiting time is not crucial; virtually any other time will produce similar results).

Even with a constant flux one may obtain erroneous results due to varying uncertainties for the measurements. As a consequence, such significant results are not even valid for testing whether the flux is constant. For an illustration, let $x = (3, 2, 1, 5)$ be a vector of four observations, and $Y_1, \ldots, Y_4$ be four independent Gaussian random variables with mean zero and standard deviations 1, 1, 1, and 4. The Ys represent observations of a constant quantity with measurement error. In 7.0% of 10^5 simulations of the Ys, the correlation r_S between x and y was equal to 1, whereas the table for the exact null distribution of r_S gives a value of 4.17%. Similar results obtain when the significance of the χ^2-statistic,

$$\chi^2 = \min_{a,b} \sum_{i=1}^{n} \left[\frac{Y_i - (a + bx_i)}{\sigma_i}\right]^2, \qquad (2)$$

or of the related F-statistic is evaluated by randomly shuffling the x_i (Bahcall et al. 1987; Bieber et al. 1990): the best

correlation (smallest value of χ^2, resp. largest value of F) is obtained by exactly one of the $4! = 24$ permutations of the data, yielding a significance level of $1/24 = 4.17\%$. However, this best correlation was obtained in 10.3% of the 10^5 simulations of the Ys. This effect becomes less pronounced with more data or more equal uncertainties, but it brings out an important point concerning this type of "shuffle test," which seems to have become popular in the astrophysics literature, apparently following its use in Bahcall et al. (1987): the intuitive physical concept of shuffling the data does not automatically provide a valid method for evaluating the significance of a test statistic. Rather, the validity of such a permutation test requires a careful justification of condition (1) for each situation at hand. An example concerning the neutrino data will be given in § 5.

The relevance of the examples given in this section for the neutrino-sunspot problem does not lie in the slogan "correlation does not imply causation." Rather, in the above examples there is not even a correlation present between the two series. In other words, seeing such large correlation coefficients in the given situations is *not unlikely*.

4. ACCOUNTING FOR SERIAL CORRELATION

It is well known (see, e.g., Sheskin 1997) that if the pairs $(X_1, Y_1), \ldots, (X_n, Y_n)$ are independent, identically distributed (or at least one of the series satisfies condition [1]), and n is large, then the null distribution of $\sqrt{n} r_S$ is approximately normal with mean zero and variance 1. Usually, one employs instead the transformation

$$T(n) = r_S \sqrt{\frac{n-2}{1-r_S^2}} , \tag{3}$$

which follows approximately a t_{n-2} distribution (Student's distribution with $n-2$ degrees of freedom). This approximation is better than the normal approximation to $\sqrt{n} r_S$ and gives excellent results for sample sizes as small as $n = 10$.

If there is serial dependence in X and Y, then the scaling in these approximations changes: the variance of the normal distribution for $\sqrt{n} r_S$ now becomes

$$\sigma^2 = 1 + 18 \sum_{k=1}^{\infty} E\{[2F(X_1) - 1][2F(X_{1+k}) - 1]\} \times E\{[2G(Y_1) - 1][2G(Y_{1+k}) - 1]\} , \tag{4}$$

where F and G denote the cumulative distribution functions of X_1 and Y_1, respectively, and E denotes expected value. Analogously, the t_{n-2} approximation for the statistic (3) is now for $T(n)/\tau$, where τ is some constant. These theoretical results make precise the heuristic given above. The author can provide a proof in the case where both series are weakly dependent, e.g., if X is α-mixing or stationary with $\mathrm{corr}(X_1, X_k) \to 0$, and Y is β-mixing with $\sum_k \beta(k) < \infty$ (see, e.g., Doukhan 1994 for a definition of these mixing coefficients). These conditions are difficult to check and still somewhat restrictive, but computer simulations show that the t_{n-2} approximation to $T(n)/\tau$ in fact seems to apply in many commonly encountered situations and also for small sample sizes, similarly to the independent case. As a simple example, both X and Y of length $n = 12$ were simulated 1000 times as five-point running means of independent standard Gaussians. Figure 2 shows a plot of the percentiles of the resulting $T(12)/1.9$ versus the percentiles of the t_{10} distribution, which demonstrates that the approximation is excellent here.

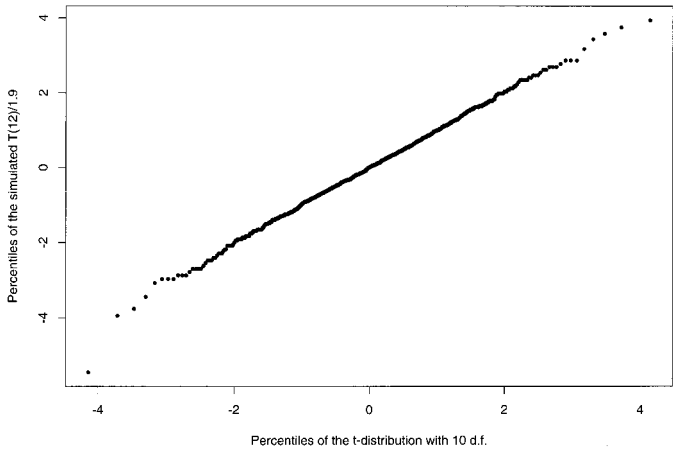

FIG. 2.—Percentiles of $T(12)/1.9$ from 1000 simulations of two series of independent five-point Gaussian running means of length 12 vs. the percentiles of the t_{10} distribution.

Clearly, expression (4) is not very usable as it stands, but modern statistical techniques allow for the rescaling in r_S or $T(n)$ to be done in an automatic way:

The moving block bootstrap (Künsch 1989) samples blocks of length l at random from the $n - l + 1$ such blocks contained in the original series X and concatenates them to a bootstrap series X^* of length n. Likewise, a bootstrap series Y^* is resampled, independently of X^*. The bootstrap then estimates the sampling null distribution of $T(n)$ by simulating a large number of bootstrap series (X^*, Y^*) and computing $T(n)$ for each simulation. Comparing the original $T(n)$ to the resampled values of this statistic yields the bootstrap significance value. One may expect to improve on this procedure by using the fact stated above that $T(n)/\tau$ follows a t_{n-2} distribution. Now only τ needs to be estimated by the bootstrap. This can be done by equating the variance of $T(n)/\tau$ to the variance of the t_{n-2} distribution, which equals $(n-2)/(n-4)$:

$$\frac{n-2}{n-4} = \frac{\mathrm{Var}\ T(n)}{\tau^2} .$$

The variance of $T(n)$ can be estimated by the sample variance v_n of the bootstrap replications of $T(n)$. Then one can solve the above equation to obtain the following estimate for τ:

$$\hat{\tau} = \sqrt{v_n (n-4)/(n-2)} . \tag{5}$$

The null hypothesis of independence is then tested by comparing $T(n)/\hat{\tau}$ to the t_{n-2} distribution.

Alternatively, the shuffle argument can be rescued by permuting blocks of the series instead of individual observations: it can be shown that asymptotically, as n and the block length become large, this kind of permutation test will give the right results.

Both of these methods concatenate pseudo-independent blocks and thus introduce artificial independence into the resampled series. This impairs the performance of these methods because the computation of the ranks used in r_S is based on the whole series, not just on the blocks. Therefore, a different method is proposed here: If a block X^l of length l is chosen at random from the $n - l + 1$ blocks contained in series X, and independently a block Y^l is chosen from Y by

the same method, then one can think of (X^l, Y^l) as a realization of length l of the series (X, Y), where X and Y are independent. Using the result stated above, one obtains that the resulting $T(l)/\tau$ follow approximately a t_{l-2} distribution. Resampling the (X^l, Y^l) many times and computing the sample variance v_l of the resulting $T(l)$ allows then to estimate τ by equating variances as introduced above. One obtains

$$\hat{\tau} = \sqrt{v_l(l-4)/(l-2)}\,. \tag{6}$$

The significance of the statistic $T(n)$ for the whole series is then evaluated by comparing $T(n)/\hat{\tau}$ to a t_{n-2} distribution. This approach of resampling blocks and rescaling is related to the procedures of Carlstein (1989) and Politis & Romano (1994), who employ such methods in a nonparametric way.

A simulation study was run to compare these procedures. One thousand pairs of independent series X and Y were simulated, each of length 108 and a five-point running mean of Gaussians. Table 2 gives the frequencies of rejection of the independence hypothesis at various significance levels for the four methods described above: comparing $T(108)$ to the moving block bootstrap distribution of that statistic ("MB bootstrap"), using the moving block bootstrap distribution to estimate τ and comparing $T(108)/\hat{\tau}$ to the t_{106} distribution ("parametric MB bootstrap"), permuting blocks instead of individual observations in the permutation test, and estimating τ by subsampling and comparing $T(108)/\hat{\tau}$ to the t_{106} distribution ("parametric subsampling"). For each method, block lengths of 12 and 18 were used and 500 resamples (resp. permutations) were generated in each of the 1000 simulations. The parametric subsampling method shows the best results, but the use of only 1000 simulation mandates some caution in such a comparison.

5. EVIDENCE FOR A CORRELATION BETWEEN NEUTRINOS AND SUNSPOTS

First the complete set of 108 experimental runs will be considered, and it will be shown that a proper test for correlation that is based on certain optimality criteria yields a clearly nonsignificant result. Then earlier segments of the data for which highly significant results have been reported will be revisited. It will be shown that the tests used there give results that are in fact compatible with the null hypothesis of no correlation, once these tests are evaluated in a way that accounts for the uncertainties and guards against a possible serial correlation under the null hypothesis.

The neutrino flux meeting the Earth at time t is denoted by flux(t) and is observed by the Homestake experiment at times t_i with random errors, yielding the reported flux N_i:

$$N_i = \text{flux}(t_i) + \sigma_i \epsilon_i, \quad i = 1, \ldots, 108\,. \tag{7}$$

Here the scalars σ_i denote the uncertainties in the neutrino measurements, which are usually taken to be "average uncertainties" (half the difference between upper and lower confidence limit) or "upper uncertainties" (the difference between upper confidence limit and measured neutrino flux); see Bahcall et al. (1987). The ϵ_i denote the random measurement errors after dividing those by the σ_i. Hence equation (7) does not require that the σ_i be exactly equal to the standard deviations of the error distributions. We make only the more reasonable assumption that the σ_i are proportional to those standard deviations, and that the ϵ_i are identically distributed (not necessarily with standard deviation 1).

A test for correlation can now be developed by examining how linear functions $a + bs_i$ of the sunspot numbers s_i explain flux(t_i), i.e., using regression techniques. The null hypothesis H_0 will be that $b = 0$, i.e., flux(t_i) $= a$ for all $i = 1, \ldots, 108$. The sunspot numbers used pertain to the month into which the mean time of the corresponding neutrino run falls.

A simple test applies if one is willing to assume that under H_0 the measurements for different runs are independent. This assumption is probably violated as will be explained below, and a more appropriate test will be presented promptly. However, assuming independence will make the results err in favor of an (anti)correlation and allows us to use a test with certain optimality properties that can be evaluated exactly by way of a permutation procedure, so it is informative to make this assumption for a moment. Then the ϵ_i are independent and identically distributed. Hence, by

TABLE 2

RELATIVE FREQUENCIES OF REJECTION OF NULL HYPOTHESIS OF INDEPENDENCE AT VARIOUS NOMINAL SIGNIFICANCE LEVELS, USING FOUR DIFFERENT TESTS[a]

METHOD	RELATIVE FREQUENCY OF REJECTION AT NOMINAL SIGNIFICANCE LEVEL (%)		
	5%	1%	0.5%
Block Length 12			
MB bootstrap	7.3	1.1	0.2
Parametric MB bootstrap.....	7.3	1.5	0.6
Permuting blocks	6.2	1.0	0.1
Parametric subsampling.....	4.2	0.9	0.4
Block Length 18			
MB bootstrap	5.6	1.5	0.6
Parametric MB bootstrap.....	5.9	1.9	1.0
Permuting blocks	5.7	1.6	1.1
Parametric subsampling.....	4.7	1.1	0.6

[a] Results of a Monte Carlo study using four different tests based on resampling. Results are based on 1000 simulations of two independent five-point Gaussian running means of length 108.

equation (7), the scaled differences,

$$d_i = \frac{N_i - a}{\sigma_i}, \quad i = 1, \ldots, 108, \tag{8}$$

satisfy condition (1). Thus the significance of the test statistic $T = \sum_{i=1}^{108} s_i d_i$ can be evaluated exactly (apart from the Monte Carlo approximation error) by randomly permuting the d_i. The statistic T has certain optimality properties for the problem at hand; see Maritz (1995). Extreme values of T indicate that there is a trend in the d_i that varies in concert with the s_i. The permutation distribution of T will in general not be symmetric about zero, so to obtain two-sided observed significance levels, one has to double the appropriate tail probability obtained by comparing T to the permutation distribution. If one does not want to hypothesize a value a for the flux, then one has to use an appropriate estimate $\hat{a}$. One can then argue that shuffling should still produce approximately valid results. The usual estimate for a is $\hat{a}_1 = (\sum_{i=1}^{108} N_i/\sigma_i^2)/\sum_{i=1}^{108} \sigma_i^{-2}$. It is argued in Cleveland et al. (1996) that the estimate provided by the combined Homestake maximum likelihood analysis is more appropriate, and hence we use in the following that estimate: $\hat{a}_2 = 0.482$ atoms day^{-1}. Using $\hat{a}_1$ instead further weakens the evidence for a correlation for most of the results reported below, and changes none of the conclusions drawn. Likewise, "average errors" result usually in somewhat less significant results than "upper errors," and so only results using the latter will be reported. The test just described was evaluated with 10^4 random permutations and resulted in an observed significance level of 8.4%. Applying the less powerful Spearman rank correlation test to the s_i and d_i yields 18.1%, a less significant result, as expected.

The evidence becomes even weaker if one cannot rule out a serial correlation within the observed neutrino flux under the null hypothesis. One example of a likely contributor to such an effect is a serial correlation within the background event rate. Background events are caused by cosmic rays and other sources (Bahcall 1989). The data analysis algorithm of the Homestake experiment estimates the background rate (assumed to be constant) jointly with the solar neutrino rate (Davis 1996), and the separation of these two is necessarily imperfect. For example, several lower confidence bounds or event rates were set to zero because subtracting the estimated background rate would otherwise lead to impossible negative values. As was explained in the previous sections, a simple permutation argument is no longer applicable if the background rate or the solar neutrino rate or the analysis apparatus exhibits some kind of serial correlation under the null hypothesis, e.g., due to a periodicity in cosmic rays (which does not have to be related to the solar cycle), or due to the presence of ^{37}Ar atoms in the tank that were not extracted after the previous run. Then the significance of the statistic T needs to be evaluated instead with the methods presented in the previous section, which guard against the effects of a serial correlation. Permuting blocks or using the moving blocks bootstrap with a block length of 18 yields significance levels of 13.8% and 15.4%, respectively. Applying these procedures to the Spearman correlation based on s_i and d_i gives again less significant results.

Most of the highly significant results in the literature were reported for earlier stretches of the data, which clearly look more correlated. However, an appropriate evaluation of the respective statistical tests will now show that the resulting evidence is well compatible with the hypothesis of no correlation, as was found for the whole series. While it is unavoidable to refer to some of these published results in order to make that point, one needs to keep in mind that these misinterpretations are not really due to those authors but to the negligent treatment of this issue in textbooks by the statistics community, as was explained in § 3.

As shown in § 3, there are at least three issues that have to be accounted for. First, one needs to be careful not to smooth the data, as that will make the two series seemingly more correlated. While this point will not be pursued further here, it seems at least plausible that the widely reported improved correlation of various quantities with smoother functions of the sunspot numbers (Bahcall & Press 1991; Delache et al. 1993; Massetti & Storini 1993; Oakley et al. 1994) is due to that effect. Second, one needs to take care of the uncertainties in the analysis. Third, the significance of the test statistic has to be evaluated in an appropriate way to guard against a possible serial correlation in the two series under the null hypothesis of no cross-correlation. Section 4 provided four methods for that using the example of the Spearman correlation coefficient. In the following these methods will always be used with a block size of 18, unless a short series necessitates a shorter block length for the permutation test, which will then be noted.

The F-statistic used by Bieber et al. (1990) incorporates the uncertainties, which are taken there to be the maximum of the upper and lower confidence width. The significance levels reported there for the run numbers between 18 and 108 are 0.08% and 0.4%, as evaluated by a F-table and the shuffle test, respectively. It was shown in § 3 that the use of the shuffle test cannot be justified here, because of the differing uncertainties. Furthermore, when the permutation test is based on blocks of length 12 to guard against a possible serial correlation, then the significance drops to 3.4%. Using the improved set of uncertainties in the same way as in the cited reference gives 8.6%.

In Bahcall & Press (1991) a segment of the neutrino data with run numbers between 18 and 105 was considered, and using the Spearman correlation coefficient, a significance level of 5×10^{-5} obtains for the last two-thirds (54 data pairs) of that segment. Without prior reason why such a correlation should occur only in a certain segment, these results have to be adjusted for "fishing" for such a significant stretch. In the quoted reference an adjustment factor of 10 is given. (With 27 more runs available now which show hardly any correlation, one could argue in favor of using an even bigger adjustment factor. Likewise, some adjustment would be necessary for the results of the F-statistic above.) When basing the analysis on the d_i in equation (8) to account for the uncertainties, all four methods introduced in the previous section yield significance levels between 0.3% and 0.5% for the Spearman correlation coefficient (between 0.1% and 0.2% if one uses the obsolete set of uncertainties). Multiplying those numbers by some adjustment factor bigger than 10 makes them comparable to the range of 5%–10% found for the F-statistic after some adjustment. The point here is not to justify an exact value of such an adjustment factor. Rather, it should become clear that an appropriate evaluation increases the significance levels of these statistics by an amount that makes it plau-

sible that the results are compatible with the hypothesis of no correlation, when taking into account that only the most significant segments of the data were analyzed.

6. SUMMARY

Standard tests for correlation are built on strong assumptions that usually do not apply in a time series context. A violation of these assumptions can lead to erroneous, highly significant results. There exist statistical techniques to evaluate nonparametric measures of correlation in a time series context. An appropriate test based on certain optimality properties shows that there is no significant correlation between the sunspot number and the solar neutrino flux up to run 133. A proper reanalysis of reported highly significant correlations for earlier stretches of the data gives results that are compatible with the hypothesis of no correlation. These findings allow one to put together a coherent picture of the somewhat conflicting evidence reported in the literature: the correlation analysis does not contradict the periodogram analysis in Bahcall & Press (1991), where no significant 11 yr component was found in the neutrino data. And the widely reported improved correlation with smoother functions of solar activity is likely due to the statistical effect described in § 3.

I wish to thank Raymond Davis and Kenneth Lande for kindly making the Homestake data available, and Giorgio Gratta and Peter Sturrock for helpful discussions. This work was supported by NASA grants NAS 8-37334 and NAGW-2265, Air Force grant F49620-95-1-0008, and NSF grant DMS-9704557.

REFERENCES

Bahcall, J. N. 1989, Neutrino Astrophysics (New York: Cambridge Univ. Press)
Bahcall, J. N., Field, G. B., & Press, W. H. 1987, ApJ, 320, L69
Bahcall, J. N., & Press, W. H. 1991, ApJ, 370, 730
Basu, D. 1982, Sol. Phys., 81, 363
Bickel, P. J., & Doksum, K. A. 1977, Mathematical Statistics: Basic Ideas and Selected Topics (Oakland: Holden-Day)
Bieber, J. W., Seckel, D., Stanev, T., & Steigman, G. 1990, Nature, 348, 407
Brockwell, P.J., & Davis, R. A. 1987, Time Series: Theory and Methods (New York: Springer)
Carlstein, E. 1986, Ann. Stat., 14, 1171
Cisneros, A. 1971, Ap&SS, 10, 87
Cleveland, B. T., Daily, T., Davis, R., Jr., Distel, J.R., Lande, K., Lee, C. K., & Wildenhain, P. S. 1998, ApJ, 496, 505
Davis, R., Jr. 1996, Nucl. Phys. B (Proc. Suppl.), 48, 284
Delache, P. H., Gavryusev, V., Gavryuseva, E., Laclare, F., Regulo, C., & Roca Cortes, T. 1993, ApJ, 407, 801
Dorman, L. I., & Wolfendale, A. W. 1991, J. Phys. G, Nuclear Part. Phys., 17, 769
Doukhan, P. 1994, Mixing (New York: Springer)
Krauss, L. M. 1990, Nature, 348, 403
Künsch, H. R. 1989, Ann. Stat., 17, 1217
Maritz, J. S. 1995, Distribution-free Statistical Methods (London: Chapman & Hall)
Massetti, S., & Storini, M. 1993, Sol. Phys., 148, 173
McNutt, R. L., Jr. 1995, Science, 270, 1635
Oakley, D. S., Snodgrass, H. B., Ulrich, R. K., & VanDeKop, T. L. 1994, ApJ, 437, L63
Okun, L. B. 1986, Soviet J. Nucl. Phys., 44, 546
Politis, D. N., & Romano, J. P. 1994, Ann. Stat., 22, 2031
Raychaudhuri, P. 1986, Sol. Phys., 104, 415
———. 1991, Mod. Phys. Lett., A6, 2003
Sheskin, D. J. 1997, Handbook of Parametric and Nonparametric Procedures (Boca Raton: Chemical Rubber Company)
Voloshin, M. B., & Vysotskii, M. I. 1986, Soviet J. Nucl. Phys., 44, 544
Voloshin, M. B., Vysotskii, M. I., & Okun, L. B. 1986, Soviet J. Nucl. Phys., 44, 440
Walther, G. 1997, Phys. Rev. Lett., 79, 4522