1994ApJ...429..415S

THE ASTROPHYSICAL JOURNAL, 429:415–418, 1994 July 1
© 1994. The American Astronomical Society. All rights reserved. Printed in U.S.A.

CROSS-CORRELATION OF POINT SERIES USING A NEW METHOD

RICHARD B. STOTHERS
Institute for Space Studies, NASA/Goddard Space Flight Center, 2880 Broadway, New York, NY 10025
Received 1993 July 16; accepted 1993 December 29

ABSTRACT

Traditional methods of cross-correlation of two time series do not apply to point time series. Here, a new method, devised specifically for point series, utilizes a correlation measure that is based on the rms difference (or, alternatively, the median absolute difference) between nearest neighbors in overlapped segments of the two series. Error estimates for the observed locations of the points, as well as a systematic shift of one series with respect to the other to accommodate a constant, but unknown, lead or lag, are easily incorporated into the analysis using Monte Carlo techniques. A methodological restriction adopted here is that one series be treated as a template series against which the other, called the target series, is cross-correlated. To estimate a significance level for the correlation measure, the adopted alternative (null) hypothesis is that the target series arises from a homogeneous Poisson process. The new method is applied to cross-correlating the times of the greatest geomagnetic storms with the times of maximum in the undecennial solar activity cycle.

Subject headings: methods: numerical — methods: statistical — solar-terrestrial relations

1. INTRODUCTION

Cross-correlation of two ordinary time series (or any two one-dimensional series of measurements) proceeds by making a comparison of the serial data at common positions within overlapped segments of the two series. Suppose, however, that the two series are point series, consisting only of the recorded times (or positions) of point events. (For convenience, each event can be assigned a unit weight.) If the times are accepted as being infinitely precise and the two series are not identical, there can be no common matchups of points, and the correlation coefficient will necessarily be zero. This result is unrealistic, however, because timing errors and variable lags and leads between the two series always exist in nature. Although devices like uniform-sized time bins can be used to group the events and thereby to introduce some artificial smearing into the series, potential information may be lost because of the coarsened resolution, and some false matchups may occur. In practice, the data themselves ordinarily impose a minimum spacing—for example, one year for annually summed, averaged, or sampled data. Going beyond this natural limit, however, involves a usually arbitrary choice of bin size. Too large a bin may cover up a possible lag or lead of one series with respect to the other, and may even induce an apparent regularity in both series that is not inherently present. This could create a spuriously high serial correlation (Quenouille 1952).

It would be useful, therefore, to have available a method of cross-correlation of two point series that completely avoids the arbitrariness and uncertainties of binning. This paper presents a new method that does not compute a traditional correlation coefficient, but computes instead an rms difference between time points that are the nearest counterparts of each other. The method is very simple, but seems not to have been suggested before. When the points possess measurement errors or when the statistical significance of the correlation is to be estimated, Monte Carlo techniques can be easily adapted to the problem. A necessary restriction is that one of the two series be treated as a fixed, template series against which the other series is cross-correlated. In many applications, however, this is not a drawback, and can even be the most natural way of approaching the problem. The present idea is: Given one series, is the other series possibly causally related to it? A reasonable null hypothesis might be that the purportedly dependent series arises from a random, unrelated process.

2. METHOD OF ANALYSIS

2.1. *Point Series*

Point series of events arise in many physical contexts, and several methods of statistically analyzing the structure of a single series of point events have been developed (e.g., Cox & Lewis 1966). Much less studied are double point series and multiple point series of events; however, two important problems connected with double and multiple series have been looked at. If the events arising from two or more periodic sources form a pooled output, various ways of disentangling the individual outputs have been devised (Cox & Lewis 1966). This problem is of some interest in astrophysics (e.g., a multiple time series of pulsed events from several spatially unresolved pulsing objects), but it will not be discussed here. The second problem concerns the quantitative assessment of near coincidences of events of two or more different types. The basic reference on this subject is Cox (1955), who used observed numbers of coincidences in specified time intervals and simple parametric tests for significance. Since more than this can be done with modern computer power, the following nonparametric method has been developed for specifically the case of two point series of events, subject to the restriction mentioned in § 1.

2.2. *Cross-Correlation of Two Point Series*

Consider two point series of events. In the segments of overlap between series A and B, designate the event times as $t_{A1} \leq t_{A2} \leq \cdots \leq t_{AM}$ and $t_{B1} \leq t_{B2} \leq \cdots \leq t_{BN}$. (In general, M will not be equal to N.) For definiteness, series A and B will be called here the *template* and *target* series, respectively. If the target series B is being cross-correlated with the template (or comparison) series A, the overlapped segment of the target series must start at $t_{B1} \geq t_{A1}$ and must end at $t_{BN} \leq t_{AM}$.

© American Astronomical Society • Provided by the NASA Astrophysics Data System

1994ApJ...429..415S

The new method consists of finding in the template series the nearest event time t_{Ai} to each event time t_{Bi} of the target series, and of forming the residual difference $d_i = t_{Bi} - t_{Ai}$. The lumped variance of the observed times (Broadbent 1955) is then

$$s^2 = \left(\sum_{i=1}^{N} d_i^2\right) \Big/ N .$$

The estimate of the standard deviation, s, follows from this and serves as a formal measure of correlation. It quantifies the extent to which the observed times in the two series coincide; for a perfect match of all the target series points with template series points, $s = 0$. If the target chain is fairly short and the overlapped segment of the template series lacks one or more expected events, it may happen that s yields a very biased estimate of the standard deviation, owing to the presence of one or more highly deviant values of d_i. In that case, it would be preferable to work with the median value of the absolute residuals $|d_i|$.

It is obvious that, if the roles of the two series are reversed, the d_i^2 remain invariant because of symmetry. For $N \neq M$, however, the number of d_i^2 that contribute to s^2 is different in the two cases, being N in one case and M in the other. Accordingly, the value of s^2 in the reversed case is not necessarily (or even usually) the same as that in the original case. When looking for a correlation between point series where series role matters, the target and template series should be clearly labeled. In cases of physical interest, information is often available to decide, on the basis of either known or probable physical cause and effect, which series should be designated which.

An illustrative example that underscores the importance of series role consists of the observed times of maximum sunspot numbers during the sunspot cycle and the observed times of the greatest geomagnetic storms, over a selected observation period. Since it is not true in general that only one large storm occurs per sunspot cycle (none may occur and sometimes more than one occurs), N and M are not necessarily equal. How such an inequality affects s^2 can be seen as follows. Suppose that three storms occur close together during a particular sunspot cycle represented by an interval of time T. Then, if the template series is taken to be the series of times of solar maximum, three events contribute to s^2 in the interval T. On the other hand, if the series roles are inverted, only one event contributes, namely, the time of the solar maximum, and it is correlated with the time of the nearest one of the three storms; the two other storms are not counted. In this example, the physically plausible choice of series roles is the first one, because logic suggests that solar activity might cause geomagnetic storms, and not vice versa. If an attempt were made to cross-correlate *every* point in both series, an absurd situation would arise for any sunspot cycle during which no great storm occurred: the method would proceed to match that cycle's time of maximum with the time of the nearest storm in some neighboring cycle, possibly even a preceding cycle. The spuriously large value of s^2 that results might then falsely suggest a poor series correlation.

In some potential applications where series role is irrelevant (e.g., a comparison of radio and optical observations of pulse times from a noisy, pulsing object), it is reasonable to correlate every point in each series with its nearest neighbor in the other series. The total number of d_i^2 that can contribute to s^2 then lies between N or M (whichever is larger) and $N + M - 1$. It is worth emphasizing, however, that if series role is important this kind of two-way bookkeeping makes no sense, because some type A events may not lead to type B events, and some type B events may not be associated with type A events. In fact, in certain applications where the template series is known to be seriously incomplete, it may even be desirable to reverse the series roles.

2.3. *Time Lag*

If, as is the case in many applications, an unknown lag or lead is suspected to exist between the target series and the template series, the preceding analysis should be modified by choosing an arbitrary total lag range and applying a small lag step to the target series, holding the template series fixed. Although both N and M may change with the stepping up of the lag, this change is unimportant in practice if N and M are much greater than unity. Out of the complete set of correlation measures generated by stepping up the lag, the smallest correlation measure indicates the best estimate of the lag, in the same manner as the largest correlation coefficient indicates the lag for a traditional cross-correlogram. If the two series are identical, the analog of an ordinary auto-correlogram is obtained.

2.4. *Significance Testing*

An estimate of the statistical significance of the derived serial correlation measure, if one is wanted, must finally be made in some way. Here the observed series that is selected as the template series is simply accepted as given; no assumption is made about the process giving rise to it. The template series is therefore "assigned." Consequently, any estimated significance of the correlation between it and the observed target series will refer here only to the assumed null hypothesis of a homogeneous Poisson process (i.e., a rectangular distribution of points) for the target series. Properly calibrated, this choice of hypothesis test for a point series is known to be valid (Broadbent 1955; Stigler & Wagner 1987). Although no significance test can ever be "best" in a nonparametric situation, the present test possesses the virtue of applying generally to the cause-and-effect type of paired time series that often occur in physical problems. Ideally, all possible null hypotheses should be tested, but, as this is impossible, the hypothesis that is probably most used in time series analysis has been adopted. In special cases, another hypothesis could be substituted.

To perform the test, a random target series, lying inside the actual (or potential) range of the observed target series and containing, like the observed series, N points, is generated from a uniform distribution of independent random variables, and cross-correlated with the template series. This process is repeated a large number of times. The frequency with which the correlation measure in the simulations equals or falls below the correlation measure derived for the two original series yields an estimate of the probability that the original series are not correlated, in the sense that the null hypothesis as formulated above is not rejected. Expressed as a percentage, this quantity represents the estimated significance level. It indicates how often a random series outperforms the observed target series in matching the observed template series.

The adopted Monte Carlo simulations are capable of handling, in principle, any values of N and M in the derivation of estimates of the significance level. Even if N and M are very small, consistent results should be obtained because the significance tests are conducted by performing random simulations in exactly the same manner as for the observed series. And even

© American Astronomical Society • Provided by the NASA Astrophysics Data System

1994ApJ...429..415S

if M in the template series is relatively large so that each of the N points in the observed target series has a high probability of accidentally being a close neighbor of some point in the template series, this probability applies also, to the same extent, for the N points in the random target series. In practice, the number of simulations needed for sufficient accuracy can be found empirically by using an increasingly large number of simulations until the estimated significance level becomes effectively constant.

For applications where series role is not relevant, *pairs* of random series consisting of N and M points can be generated and cross-correlated.

2.5. *Measurement Errors*

Observed times (in both series) that contain quantifiable uncertainty need not be used at their face value. They can be partially randomized in various ways. Cross-correlation of a large number of such pseudo-randomized time series yields an *average* correlation measure that will not, in general, be equal to the correlation measure of the two original series. This is true even if the estimated errors of the observed times are symmetrically distributed around the time peaks. Since it is possible, whenever estimated errors are large, for randomly perturbed times to stray out of order, the reordering of times in some of the perturbed series may occasionally be necessary, if the numerical algorithm that is adopted requires ordered times.

To test the average correlation measure for statistical significance, Monte Carlo simulations should be performed by cross-correlating each fully random target series with *all* of the pseudo-randomized template series. In the case of N_A template series and N_B target series computed per template series, the total number of serial correlations that are needed for the significance test, $N_A N_B$, can be very large. If, however, the estimated errors in the observed times are not very big, it may be sufficient to adopt $N_A \ll N_B$.

3. APPLICATION USING SOLAR AND GEOMAGNETIC TIME SERIES

As a trial application, a statistical correlation is sought between the times of the 14 greatest geometric storms recorded at Greenwich during the period 1855–1955 (Chapman 1957) and the observed times of solar cycle maximum and minimum, as defined by the reported relative numbers of sunspots (Waldmeier 1961). Since sunspot extrema in the undecennial solar cycle are not sharply defined if one examines day-to-day or even month-to-month data, a decision was made to use only the *years* of maximum and minimum. (Observed intervals between successive maxima, or successive minima, are not regular, but range from 9 to 13 years.) For consistency, only the *years* in which the greatest geomagnetic storms occurred are actually employed as the target series; 4 years therefore appear twice in the series. The data used are listed in Table 1.

The derived cross-correlogram is shown in Figure 1. Notice that the smallest correlation measure occurs at, or very close to, the time of solar maximum. Drops in the correlation measure also occur at match positions of 10–11 years before and after solar maximum, which is not surprising because solar activity is quasi-periodic with a mean period of 10–11 years. The statistical significance level of the correlation at a time lag of zero is found to be 1%. If the lag is not considered known a priori and significance tests are performed by progressively shifting the random target series backward by 5 years and forward by 5 years in steps of 1 year, the significance level of the correlation becomes about 7%.

TABLE 1

SOLAR AND GEOMAGNETIC TIME SERIES

Geomagnetic Storms	Sunspot Maxima	Sunspot Minima
1859	1848	1843
1872	1860	1856
1882	1870	1867
1882	1883	1878
1903	1893	1889
1909	1905	1901
1921	1917	1913
1938	1928	1923
1938	1937	1933
1940	1947	1944
1941	1957	1954
1941		
1946		
1946		

A superposed epoch plot, presented in Figure 2, illustrates the spread of the residuals. Notice that although a remarkable gap occurs at exactly zero lag the data generally cluster, more or less equally, around both sides of the gap. Such a quasi-symmetrical clustering produces an average effect that causes, in a reasonable way, the maximum correlation to show up at zero lag. The strength of the maximum correlation, however, is somewhat weakened by the presence of the gap, as the standard deviation of the fit amounts to 2 years (Fig. 1). An earlier display of very similar data in tabular form (Newton & Milsom 1954) exhibits the same double-peak behavior seen in Figure 2. This feature appears to correspond to the bimodal maximum shown by the most outstanding active phenomena at the Sun's surface (Das Gupta & Basu 1965; Gnevyshev 1967; Goswami et al. 1988). The cause of the bimodality is not known.

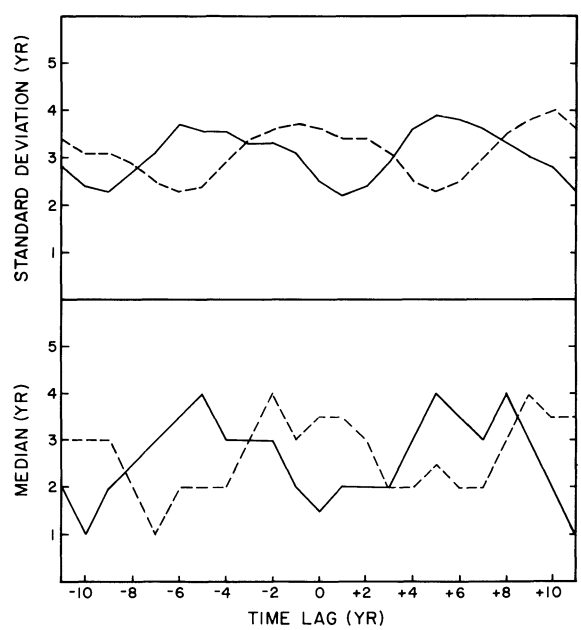

FIG. 1.—Cross-correlation of the years of the 14 greatest geomagnetic storms recorded at Greenwich during 1855–1955 with the years of sunspot maximum (*solid line*) or sunspot minimum (*dashed line*). The lag step is one year. The correlation measure shown is the standard deviation and the median, in the upper panel and lower panel, respectively.

© American Astronomical Society • Provided by the NASA Astrophysics Data System

1994ApJ...429..415S

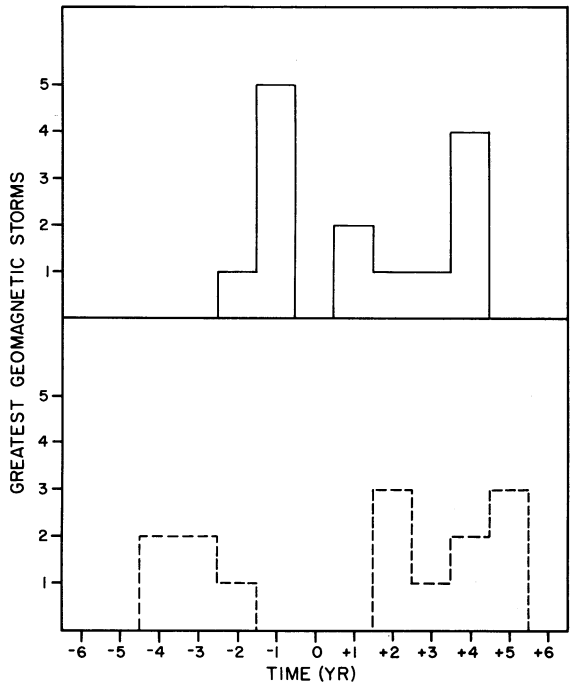

FIG. 2.—Superposed epoch diagram, showing the years of the 14 greatest geomagnetic storms recorded at Greenwich during 1855–1955, plotted with respect to the years of the nearest sunspot maximum (*solid line*) or the nearest sunspot minimum (*dashed line*).

Further applications of the present method, especially to time series in which the observed times have appreciable error (§ 2.5), have been published elsewhere (Stothers 1993a, b).

4. CONCLUSION

Two point series of events may be cross-correlated by matching nearest neighbors in overlapped segments of the two series and then computing an rms difference or, alternatively, a median absolute difference between all the corresponding points. Either correlation measure is easy to compute and interpret. Measurement errors in the locations of the points are easily handled by partially randomizing the two series and computing an average correlation measure. A uniform displacement between the two series can be detected by constructing an analog of the traditional cross-correlogram. An estimated significance level for the correlation measure is, finally, obtainable by cross-correlating many random point series with the prespecified one of the two observed series that is designated as the template series, bearing in mind that the size of the correlation measure may depend on which of the two observed series is treated as the template series and which as the target series. The null hypothesis adopted here is a homogeneous Poisson process for the target series. As is usual in statistics, it is necessary to analyze small number statistics by doing large number statistics.

Application of the present method of cross-correlation to a time series consisting of the dates of the greatest geomagnetic storms between 1855 and 1955 shows that these storms significantly correlate with the maxima in the undecennial solar activity cycle at the 1% level.

This work was supported by the NASA Climate Research Program. I am grateful to the late S. Lebedeff for preliminary discussions that led up to the present study, and to the referee for suggestions for clarifying various points.

REFERENCES

Broadbent, S. R. 1955, Biometrika, 42, 45
Chapman, S. 1957, Nature, 179, 7
Cox, D. R. 1955, J. Roy. Stat. Soc., B17, 129
Cox, D. R., & Lewis, P. A. W. 1966, The Statistical Analysis of Series of Events (London: Methuen)
Das Gupta, M. K., & Basu, D. 1965, J. Atm. Terr. Phys., 27, 1029
Gnevyshev, M. N. 1967, Sol. Phys., 1, 107
Goswami, J. N., McGuire, R. E., Reedy, R. C., Lal, D., & Jha, R. 1988, J. Geophys. Res., 93, 7195
Newton, H. W., & Milsom, A. S. 1954, J. Geophys. Res., 59, 203
Quenouille, M. H. 1952, Associated Measurements (London: Butterworths)
Stigler, S. M., & Wagner, M. J. 1987, Science, 238, 940
Stothers, R. B. 1993a, Geophys. Res. Lett., 20, 887
———. 1993b, Geophys. Res. Lett., 20, 1399
Waldmeier, M. 1961, The Sunspot-Activity in the Years 1610–1960 (Zurich: Schulthess)

© American Astronomical Society • Provided by the NASA Astrophysics Data System