LA-UR -91-3343

LA-UR--91-3343

DE92 002463

TITLE: TESTING FOR NONLINEARITY IN TIME SERIES: THE METHOD OF SURROGATE DATA

AUTHOR(S): James Theiler

SUBMITTED TO: Proceedings of the Symposium on Interpretation of Time Series from Mechnanical Systems, held 25-30 August 1991 in Warwick, England. (Proceedings to appear in Physica D)

Testing for nonlinearity in time series: the method of surrogate data

James Theiler,[a,b] Bryan Galdrikian,[a,b] André Longtin,[a,b]
Stephen Eubank,[a,b,c] and J. Doyne Farmer[a,c]

(a) *Theoretical Division, and* (b) *Center for Nonlinear Studies, Los Alamos National Laboratory, Los Alamos, NM 87545;*
and
(c) *The Santa Fe Institute, 1120 Canyon Road, Santa Fe, NM 87501.*

ABSTRACT

We describe a statistical approach for identifying nonlinearity in time series; in particular, we want to avoid claims of chaos when simpler models (such as linearly correlated noise) can explain the data. The method requires a careful statement of the null hypothesis which characterizes a candidate linear process, the generation of an ensemble of "surrogate" data sets which are similar to the original time series but consistent with the null hypothesis, and the computation of a discriminating statistic for the original and for each of the surrogate data sets. The idea is to test the original time series *against* the null hypothesis by checking whether the discriminating statistic computed for the original time series differs significantly from the statistics computed for each of the surrogate sets. We present algorithms for generating surrogate data under various null hypotheses, and we show the results of numerical experiments on artificial data using correlation dimension, Lyapunov exponent, and forecasting error as discriminating statistics. Finally, we consider a number of experimental time series — including sunspots, electroencephalogram (EEG) signals, and fluid convection — and evaluate the statistical significance of the evidence for nonlinear structure in each case.

1 Introduction

The inverse problem for a nonlinear system is to determine the nature of the underlying dynamics (is it chaos or is it noise?) in the practical situation where all that is available is a time series of data. Algorithms are available which can in principle make this distinction, but they are notoriously fussy, and usually involve considerable human judgement. Particularly for experimental data sets, which are often short and noisy, simple autocorrelation can fool dimension and Lyapunov exponent estimators into signalling chaos where there is none. Most authors agree that the methods contain many pitfalls, but it is not always easy to avoid them. While some data sets very cleanly exhibit low-dimensional chaos, there are many cases where the evidence is sketchy and difficult to evaluate. Indeed, it is possible for one author to claim evidence for chaos, and for another to argue that the data is consistent with a simpler explanation [1–4]. Our aim is to provide a framework within which claims of nonlinearity can be evaluated. We describe our approach as first introduced in Ref. [5].

The problem is inherently statistical, for it is always *possible* for any finite length time series to be a particular realization of a noise process, just as it is possible for an effectively random time series to come from a low-dimensional deterministic process (witness the pseudorandom number generator). But the real complication arises because these two extremes — chaos and noise — are not the only available alternatives. The erratic fluctuations that are observed in an experimental time series owe their dynamical variation to a mix of various influences: chaos, nonchaotic but still nonlinear determinism, linear correlations, and noise — always noise, both in the dynamics and in the measuring apparatus. While we are motivated by the prospect of ultimately disentangling these influences, we take as a more modest goal the detection of nonlinear structure in a time series. Detecting nonlinearity is easier than describing it; we need not exhibit the underlying nonlinear dynamics, merely demonstrate the inadequacy of a linear model.

The *hard way* to detect low-dimensional behavior, for instance, is to attempt to estimate the dimension and then see if this value is small. With a finite amount of data, and especially if the data are noisy, the dimension estimated by the algorithm will at best be approximate — and at worst, wrong. One can guard against this by attempting to identify the various sources of error (both systematic and statistical), and then putting error bars on the estimate (see, for example, Refs. [6–12]). But this can be problematic for nonlinear algorithms like dimension estimators: first, assignment of error bars requires some model of the underlying process, and that is exactly what is not known; further, even if the underlying process were known, the computation of an error bar may be analytically difficult if not intractable.

Our approach, by contrast, is to determine the distribution of the quantity we are interested in (dimension, say) for an ensemble of "surrogate" data sets which are just different realizations of the particular noise process that is our null hypothesis. Then, instead of putting error bars on the estimated dimension, we put error bars around the value that we wish to distinguish it from (the value that noise gives). This can be done reliably because we know the model of the null hypothesis, and furthermore we bypass the issue of analytic tractability by computing the error bar numerically (from the standard deviation of all the numerically estimated dimensions of all the surrogate data sets).

In Section 2, we express the problem of detecting nonlinearity in terms of statistical hypothesis testing. We introduce our measure of significance, develop various null hypotheses and discriminating statistics, and describe algorithms for generating surrogate data. Section 3 demonstrates the technique for several computer-generated examples; we investigate the method of surrogate data under a variety of conditions: large and small data sets, high and low dimensional attractors, and various levels of observational and dynamical noise. We also argue that "bleaching" a chaotic time series degrades the the utility of the test. In Section 4, we illustrate the application of the method to several real data sets, including sunspots, electroencephalograms (EEG), and fluid dynamics data. With real data, there is always room for human judgement, and we argue that besides formally rejecting a null hypothesis, the method of surrogate data can also be useful in an informal way, providing a benchmark, or control experiment, against which the actual data can be compared.

2 Statistical Hypothesis Testing

The formal application of the method of surrogate data is expressed in the language of statistical hypothesis testing. This involves two ingredients: a null hypothesis against which observations are tested, and a discriminating statistic. The null hypothesis is the too-simple explanation that we seek to show is inadequate for explaining the data; and the discriminating statistic is a *number* which quantifies some aspect of the time series. If this number is different for the observed data than would be expected under the null hypothesis, then the null hypothesis can be rejected.

It is possible in some cases to derive analytically the distribution of a given statistic under a given null hypothesis, and this approach is the basis of many conventional tests for nonlinearity (*e.g.*, see Tong [13]). The method of surrogate data estimates this distribution by direct Monte-Carlo simulation. An ensemble of surrogate data sets are generated which share given properties of the observed time series (such as mean, variance, and Fourier spectrum) but are otherwise random as specified by the null hypothesis. For each surrogate data set, the discriminating statistic is computed, and from this ensemble of statistics, the distribution is approximated.

While this approach is computationally intensive, it avoids the analytical derivations which can be difficult if not impossible. This leads to increased flexibility in the choice of null hypotheses and discriminating statistics; in particular, the hypothesis and statistic can be chosen independently of each other. Efron [14] has argued persuasively in favor of computationally intensive statistics, permitting CPU time to take the place of simplifying assumptions and asymptotic results that are inevitable with classical analytical statistics.

2.1 Computing significance

Let Q_D denote the statistic computed for the original time series, and Q_{H_i} for the ith surrogate generated under the null hypothesis. Let μ_H and σ_H denote the (sample) mean and standard deviation of the distribution of the Q_{H_i}'s.

If multiple realizations are available for the observational data, then it may be possible to compare the two distributions (observed data and surrogate) directly, using for instance the Kolmogorov-Smirnov statistic which compares the full distributions, or possibly the Student-t statistic which only compares their means.

For our purposes, however, we consider that only one experimental data set is available (Of course, it is always possible to create several realizations out of that single set by chopping up the data; we haven't tried this approach, but we suspect that the numerical algorithms we work with would be severely handicapped by such short data sets.)

We define our measure of "significance" by the difference between the original and the mean surrogate value of the statistic, divided by the standard deviation of the surrogate values.

$$\mathcal{S} \equiv \frac{|Q_D - \mu_H|}{\sigma_H} \tag{2.1}$$

The significance is properly a dimensionless quantity, but it is natural to call the units of $\mathcal{S}$ "sigmas." Thus, one might speak of a two sigma effect as not especially significant, but ten sigmas as extremely significant. If the distribution of statistic values is gaussian (and numerical experiments indicate that this is often a reasonable approximation), then the p-value associated with a significance $\mathcal{S}$ is given by $p = \text{erfc}[\mathcal{S}/\sqrt{2}]$; this is the probability of observing a significance $\mathcal{S}$ or larger if the null hypothesis is true.

If computational effort really were not a consideration, then a more robust way to define significance would be directly in terms of p-values with rank statistics. In particular, if the observed time series has a statistic which is in the lower one percentile of all the surrogate statistics (and at least a hundred surrogates would be needed to make this determination), then a (two-sided) p-value of $p = 0.02$ could be quoted.

2.2 Hierarchy of null hypotheses

The null hypothesis defines the nature of the candidate process which may or may not adequately explain the data. Our null hypotheses usually specify that certain properties of the original data are preserved — such as mean and variance — but that there is no further structure in the time series. The surrogate data is then generated to mimic these preserved features but to otherwise be random. There is some latitude in choosing which features ought to be preserved: certainly mean and variance, and possibly also the Fourier power spectrum. If the raw data is discretized to integer values, then the surrogate data should be similarly discretized.

Ultimately we envision a hierarchy (perhaps even a hierarchical tree) of null hypotheses against which time series might be compared. Beginning with the simplest hypotheses, and increasing in generality, the following sections outline some of the possibilities that we have considered.

2.2.1 Temporally uncorrelated noise

The null hypothesis of no temporal correlations is of particular interest in circumstances (*e.g.,* stock market returns, or outcomes on a roulette wheel) where ***any correlation at all*** can potentially be exploited for profit. The simplest null hypothesis in this case is that the observed data is fully described by independent and identically distributed (IID) gaussian random variables. Surrogate data in this case are readily generated from a standard pseudorandom number generator, normalized to the mean and variance of the original data.

A clever extension of this approach was used by Scheinkman and LeBaron [15] in an analysis of stock market returns. To test the hypothesis of IID noise ***with arbitrary amplitude distribution***, they generated surrogate data by shuffling the time-order the original time series. This more closely mimics the original data, but it destroys any temporal correlations that may have been in the data.

2.2.2 Ornstein-Uhlenbeck noise

For most physical systems, it is usually obvious that there *are* temporal correlations, but the nature of these correlations may not be so clear. The simplest case of non-IID noise is given by the Ornstein-Uhlenbeck process [16]. For a discrete time series, this can be produced by

$$x_t = a_o + a_1 x_{t-1} + \sigma e_t \tag{2.2}$$

where e_t is uncorrelated gaussian noise of unit variance. The coefficients a_o, a_1, and σ collectively determine the mean, variance, and autocorrelation time of the time series. In fact, the autocorrelation function is exponential in this case:

$$A(\tau) \equiv \frac{\langle x_t x_{t-\tau}\rangle - \langle x_t\rangle^2}{\langle x_t^2\rangle - \langle x_t\rangle^2} = e^{-\lambda|\tau|} \tag{2.3}$$

where $\langle\rangle$ denotes an average over time t, and $\lambda = -\log a_1$.

To make surrogate data sets, the mean μ, variance v, and first autocorrelation $A(1)$ are estimated from the original time series; from these the coefficients are fit: $a_1 = A(1)$, $a_o = \mu(1 - a_1)$, and $\sigma^2 = v(1 - a_1^2)$. Finally, one generates the surrogate data by iterating Eq. (2.2), using a pseudorandom number generator for the unit variance gaussian e_t.

2.2.3 Linearly correlated noise

The null hypothesis of Ornstein-Uhlenbeck noise is arguably restrictive, and can readily be generalized to higher order. This can be done by fitting coefficients a_k and σ to a process

$$x_t = a_o + \sum_{k=1}^{q} a_k x_{t-k} + \sigma e_t \tag{2.4}$$

which mimics the original time series in terms of mean, variance, and the autocorrelation function for delays of $\tau = 1, \ldots, q$. This is an auto-regressive (AR) model; a more general model includes a moving average (MA) of time delayed noise terms as well, and the combination is called an ARMA model. For large enough q, the models are equivalent. The null hypothesis in this case is that all the structure in the time series is given by the autocorrelation function, or, equivalently, by the Fourier power spectrum. Using surrogate data based on this hypothesis has previously been advocated in Refs. [10, 11, 17, 18].

One algorithm for generating surrogate data under this null hypothesis is again to iterate Eq. (2.4), where the coefficients have been fit to the original data. An alternative algorithm is described in Section 2.4.1. This algorithm does not directly employ Eq. (2.4), but the generated surrogate data is guaranteed to have the same Fourier spectrum as the original data.

We remark that this is the null hypothesis that is associated with residual-based tests for nonlinearity; we argue in Section 3.3 that it is generally preferable to use the method of surrogate data on the raw data directly, rather than working with residuals.

2.2.4 Static nonlinear transform of linearly correlated noise

We have also considered a slightly more general null hypothesis, that the dynamics is linear, but the observation function may be nonlinear. In particular, we suppose that there is an "underlying" time series $\{y_t\}$, consistent with the null hypothesis of linearly correlated noise, and an observed time series $\{x_t\}$ given by

$$x_t = h(y_t). \tag{2.5}$$

Since x_t depends only on the current value of y_t and not on derivatives or past values, the filter $h()$ is said to be "static." The null hypothesis further assumes that $h()$ is an essentially invertible (monotonic suffices) function.

In Section 2.4.2, an algorithm for generating surrogate data corresponding to this null hypothesis is described. It effectively shuffles the data but in such a way as to preserve the linear correlations of the underlying time series $y_t = h^{-1}(x_t)$. An advantage of shuffling over, for example, a smooth fit to the function $h()$, is that any discretization that was present in the original data will be reflected in the surrogate data.

Note that time series in this class are strictly speaking nonlinear, but that the nonlinearity is not in the dynamics. Most conventional tests for nonlinearity would (correctly) conclude that the time series is nonlinear, but would not indicate whether the nonlinearity was in the dynamics or in the amplitude distribution. By using surrogate data that have been tailored to this specific null hypothesis, it becomes possible to make such fine distinctions about the underlying dynamics.

2.2.5 More general null hypotheses

Ultimately, we would like to extend this list to include more general null hypotheses. Foremost in our minds is the noisy limit cycle, which cannot be described by a linear process, even if viewed through a static nonlinear transform. Yet it is often of great interest, particularly in systems driven by seasonal cycles, to determine the nature of the inter-seasonal variation.

2.3 Battery of discriminating statistics

Since we are motivated by the possibility that the underlying dynamics may be chaotic, our first choices for discriminating statistics are just the conventional discriminants of nonlinear dynamics: correlation dimension, Lyapunov exponent, and forecasting error. Indeed, one of our eventual interests in this project is to outline the conditions in which one or the other of these methods will be more effective.

But the method in principle can be used with any discriminating statistic. We have had some success using the correlation integral ($C(r)$ for some value of r) directly as a discriminating statistic, instead of the dimension. Also, we

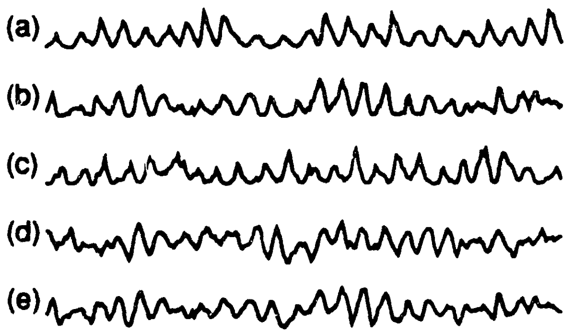

Figure 1: *The Wolfer sunspot numbers and some surrogates under various null hypotheses. Note that differences ~ often detectable by eye. The true series is in (a). The series in (b) and (c) are generated by Algorithm II. The series in (d) and (e) are generated by Algorithm I. Surrogate data generated by Algorithm I have a gaussian distribution, which is clearly distinguishable from the nongaussian amplitude distribution of the original data in (a).*

have considered but not implemented two-sided forecasting — predicting the "present" x_t from the "past" $x_{t-1}, \ldots$ *and* the "future" $x_{t+1}, \ldots$, instead of the usual forecasting which seeks to predict the future from the past. In our forecasting, we are careful to separate the "training" set from the "testing" set, to avoid overfitting; but this is not really necessary. The in-sample fitting error may also suffice as a discriminating statistic. Other candidates which we have not investigated include the embedding criterion of Liebert and Schuster [19], and the dimension statistic of Čenys and Pyragas [20]. The most exotic example we know is due to Brock, Lakonishok, and LeBaron [21], who used technical trading rules as discriminating statistics for financial data.

Below, we describe how we used the three particular discriminating statistics that we chose for the numerical experiments in this article. Ideally, dimension counts degrees of freedom, Lyapunov exponent quantifies the sensitivity to initial conditions, and forecasting error tests for determinism. These are three different aspects of low-dimensional chaotic systems. Now, we are not explicitly looking for low-dimensional chaos but just trying to detect nonlinearity, so any nonlinear statistic is a viable candidate. But our choice of statistic is *motivated* by the notion that the underlying process *might* be chaotic, and so (we hope) the statistics which characterize such processes might be most adept at detecting them. In general, we advocate using a battery of statistics, not only to increase the opportunity of rejecting the null hypothesis (since we expect some will be sensitive where others are not), but also to have some qualitative notion of "how" the data set differs from the surrogates.

2.3.1 Correlation dimension, ν

Dimension is an exponent which characterizes the scaling of some bulk measure with linear size. To compute a dimension, it is necessary to choose some range of sizes over which the scaling is to be estimated. Algorithms abound [11, 22] for estimating the dimension of an underlying strange attractor from a time series; we chose a box-assisted variation [23] (see Grassberger [24] for an elegant alternative) of the Grassberger-Procaccia-Takens algorithm [25–27] to compute a correlation integral, and the best estimator of Takens [8] for the dimension itself. The Takens estimator requires an upper cutoff size; we used one-half of the rms variation in the time series for this value. This is rather large if our aim is to make our best guess of the fractal dimension , but it gives us good values for statistical significance.

2.3.2 Lyapunov exponent, λ

Following Sano and Sawada [28] and Eckmann and Ruelle [29], we compute Lyapunov exponent by multiplying Jacobian matrices along a trajectory, with the matrices computed by local linear fits. We use the QR decomposition method of Eckmann *et al.* [30] to maintain orthogonality. For the results in this article, we consider only the largest exponent.

We have found that numerical estimation of Lyapunov exponents in the presence of noise can be problematic. Indeed, for our surrogate data sets, for which the linear dynamics is contracting, we often obtain positive Lyapunov exponents. It may be possible to remedy this by using more neighbors in the local fits, but remember that it is not the best estimate of the Lyapunov exponent itself that we are seeking, but only a statistic which distinguishes the original data from surrogate data. We would prefer to use a discriminating statistic which correctly quantifies some feature of the dynamics, as this provides more qualitative information, but the method of surrogate data does not formally require this.

2.3.3 Forecasting error, ϵ

A direct test for determinism comes from quantifying the forecasting errors obtained from nonlinear modeling. The method we use entails first splitting the time series into a fitting set of length N_f, and a testing set of length N_t, with $N_f + N_t = N$, the length of the time series; then fitting a local linear model [31] to the fitting set, locality given by the number of neighbors k; and finally, using this model to forecast the values in the testing set, and comparing them with the actual values.

If $e_t = x_t - \hat{x}_t$ is the difference between the actual value of x and the predicted value, $\hat{x}$, then we define our discriminating statistic as the mean log absolute prediction error.

$$\epsilon = \frac{1}{N_t} \sum_{t=N_f+1}^{N_f+N_t} \log |e_t|. \tag{2.6}$$

Several modeling parameters must be chosen, including the partitioning of the data set into fitting (N_f) and testing (N_t) segments, the number of steps ahead to predict (T), and number of neighbors (k) used in the local linear fit. We arbitrarily chose to divide the fitting and testing sets equally, with $N_f = N_t = N/2$, and to predict one step ahead, so $T = 1$. More important is the choice of k. For the results in this article, we set k to 1.5 times the minimum number needed for a fit, but we note that this is often not optimal. Indeed, Casdagli [32, 33] has advocated sweeping the parameter k in a local linear forecaster as an exploratory method to look for nonlinearity in the first place. For few neighbors, this models noise-free low-dimensional determinism; for many neighbors, the forecasting method is effectively a global linear predictor. When the optimal k is some intermediate value, this indicates nonlinearity with noise.

2.4 Algorithms for generating surrogate data

In this section, we describe two algorithms we use for generating surrogate data. The first is consistent with the hypothesis of linearly correlated noise described in Section 2.2.3, and the second considers the possibility of a static nonlinear transform as discussed in Section 2.2.4.

2.4.1 Algorithm I

The first algorithm is based on the null hypothesis that the data come from a linear stochastic process; the assumption is that there is no nonlinearity either in the dynamics or in the observation of the data. The surrogate data are designed to have the same Fourier spectra as the raw data.

1. Input the original data into an array `y[t]`, `t=0,...,N-1`
2. Compute the Discrete Fourier Transform:
   $z[n] = \sum_{t=0}^{N-1} e^{2\pi i nt/N} y[t]$.
   Note `z[n]` has real and imaginary components.
3. Randomize the phases: `z'[n] = z[n]`$e^{i\phi[n]}$
   Here, ϕ`[n]` is uniformly distributed between 0 and 2π.
4. Symmetrize the phases:
   `Re z''[n] = Re( z'[n] + z'[N-1-n] )/2`
   `Im z''[n] = Im( z'[n] - z'[N-1-n] )/2`

5. Invert the Discrete Fourier Transform:
$y'[t] = (1/N) \sum_{n=0}^{N-1} e^{-2\pi i nt/N} z''[n]$.
Note that because of the symmetry of the phases, the resulting time series y'[t] is real; this is the surrogate data.

2.4.2 Algorithm II

The second algorithm creates data that are realizations of the null hypothesis that the observed time series is a nonlinear static transformation of a linear stochastic process. Our approach is first to rescale values of the original time series so that they are gaussian, then to use the first algorithm to create a surrogate time series which has the same Fourier spectrum as the rescaled original. This surrogate is then rescaled to have the same values as the original time series.

1. Input the original data into an array x[t], t=0,...,N-1
2. Sort the array Sx[k], k=0,...,N-1
3. Make ranked time series Rx[t], defined to satisfy Sx[Rx[t]] = x[t].
   Note Sx[k] is a monotonic function with a well-defined inverse; so Rx[t] = Sx^{-1}[t] is a static rescaling of x[t].
4. Create a random gaussian data set g[t], t=0,...,N-1
5. Sort the gaussian random numbers Sg[k], k=0,...,N-1
6. Define new time series: y[t] = Sg[Rx[t]]
   The new time series is a static rescaling of x[t] with the property that the amplitude distribution is gaussian.
7. Use Algorithm I to make a surrogate of this gaussian time series: y'[t].
8. Make a ranked time series for y'[t], call it Ry'[t].
9. The surrogate time series is then given by x'[t] = Sx[Ry'[t]].

Note that the surrogate time series x'[t] is just a shuffling of the original time series x[t], so it obviously has the same amplitude distribution.

3 Numerical Experiments

To properly gauge the utility of surrogate data will eventually require many tests with data from real laboratory experiments. To give a sense of how this approach ought to work in practice, however, we begin with some numerical examples.

First, we note that a time series which actually is generated by a linear process should by construction give a negative result (that is, the null hypothesis should *not* be rejected); this we checked and found to be the case.

The results presented in this section used Algorithm I for generating the surrogate data. Some of the experiments were repeated with Algorithm II; the significance was reduced slightly, but the qualitative effects remained.

3.1 Variation with number of data points and complexity of attractor

The significance with which nonlinearity can be detected in a chaotic time series increases with the number of points in the time series, and in general decreases with the complexity of the time series. This is shown in Fig. 2 for the attractor of Hénon [34], using dimension and forecasting error as the discriminating statistic. Here 'Hen-n' corresponds to the sum of n independent trajectories of the Hénon map; thus it is a time series whose underlying strange attractor will have dimension $n\nu$ where $\nu \approx 1.25$ is the dimension of a single Hénon trajectory.

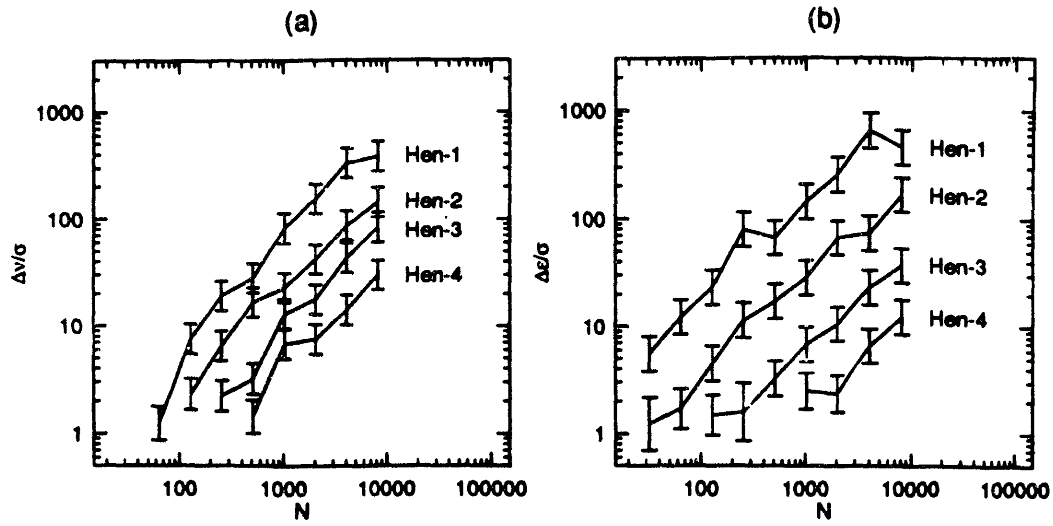

Figure 2: *Significance as a function of the number N of data points for time series generated by adding n independent trajectories of the Hénon map. The discriminating statistic is (a) correlation dimension, and (b) forecasting error. Note that significance increases with number of data points and decreases with the complexity of the system.*

3.2 Effect of observational and dynamical noise

To test whether nonlinear determinism can be detected even when it is mixed with noise, we added both dynamical (η) and observational (ε) noise to the cosine map: $y_t = \lambda\cos(\pi y_{t-1}) + \eta_t$; $x_t = y_t + \varepsilon_t$. In Fig. 3, we plot significance as a function of noise level for both dynamical and observational noise. As expected, significance decreases with increasing noise level, though we remark that the nonlinearity is still observable even with considerable noise. In the absence of noise, the rms amplitude of the signal is 0.36; thus we are able to detect significant nonlinearity even with a signal to noise ratio of one, using a time series of length $N = 512$.

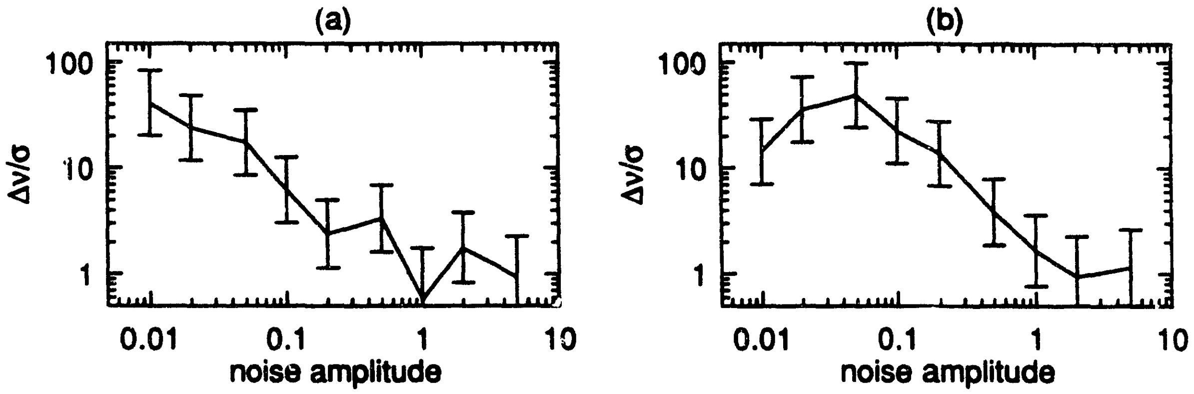

Figure 3: *Effect of noise on significance for a time series of* $N = 512$ *points, derived from the cosine map with* $\lambda = 2.8$*: (a) observational noise; (b) dynamical noise.*

3.3 Don't bleach chaotic data

A common approach to testing for nonlinearity involves first "subtracting out" the linear component, and then working with what is left, the "residuals." For instance, see Tong [13] for a review of conventional tests for nonlinearity, or Brock, Dechert, and Scheinkman [35] for a more recently proposed statistic based on the correlation integral. Given a time series x_t, the residuals are given by

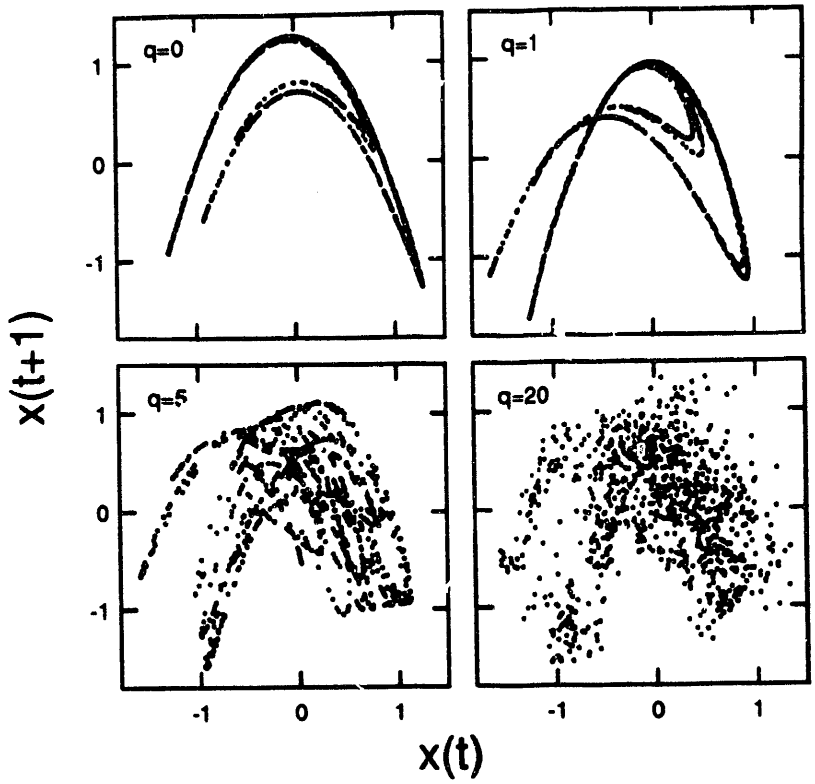

Figure 4: *Residuals of the Hénon map, as fitted with Eq. (3.1). As q increases, the deterministic nature of the map becomes less evident.*

$$\epsilon_t = x_t - \left[a_o + \sum_{k=1}^{q} a_k x_{t-k} \right] . \qquad (3.1)$$

where the coefficients a_k are chosen to minimize the sample variance $\sum_t \epsilon_t^2/N$ of the residuals. Here q is the order of the linear model.

Because the residuals ϵ_t are spectrally white (equal power at all frequencies), the process of determining residuals is sometimes called "pre-whitening" or "bleaching." However, linear filtering of chaotic data is not without its pitfalls. While the fit is based on the best auto-regressive (AR) *model*, the linear map that takes x_t to e_t in Eq. (3.1) is a moving-average (MA) *filter*, and will not formally change the structure of the attractor for finite q. For example, if x_t lies on a low-dimensional attractor, then ϵ_t will lie on an attractor of the same dimension. However, in practice, the distortion induced by the linear map can drastically affect the *appearance* of the attractor and can likewise affect *estimates* of its dimension. The effect of linear filtering on the Hénon attractor is shown in Fig. 4. The determinism which is obvious in the unfiltered data ceases to be so obvious in the filtered case.

We computed the significance of the nonlinearity in time series obtained from the Hénon map and then bleached with ever larger values of q. We show in Fig. 5 a decrease in significance, as quantified by the method of surrogate data, computed with statistics based on dimension, forecasting, and Lyapunov exponent. However, this result may only hold for chaotic data. Townshend [36] has described a situation with data from human speech in which nonlinear predictions of linearly filtered data were superior to direct nonlinear predictions of the original time series.

Following a suggestion of Brock (personal communication), we considered a time series generated by AR filtering the Hénon time series, and treating this time series as our raw data. We found in this case that a mild amount of MA filtering did improve the significance. The optimal value of q was never larger than 3, however. For larger values of q, the significance again decreased. In general, we do not recommend statistical tests for nonlinearity that are based on best estimates of the residuals, as these usually require high order (large q) filtering.

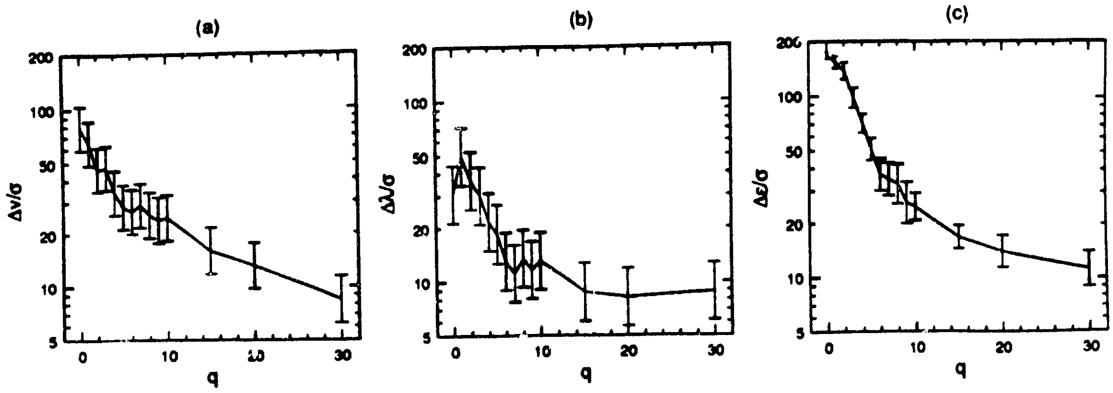

Figure 5: *Effect of bleaching on significance for a time series of length* $N = 1024$ *derived from the Hénon map. For* $q = 0$*, the raw data are used. For* $q > 0$*, the q-th order residuals (as computed by Eq. (3.1)) are used. It is apparent that attempting to "subtract out" the linear component only decreases the power of the test. The discriminating statistic is (a) correlation dimension, (b) largest Lyapunov exponent, and (c) forecasting error.*

4 Real data

We report preliminary results on some experimental time series from various sources. These results should be taken as anecdotal, and not necessarily typical of the class which they represent (the sunspot time series is an exception, of course). In particular, we have not yet attempted to "normalize" our findings with others that have previously appeared in the literature. Unless otherwise noted, the surrogate data for the results in this section were generated by Algorithm II, which corresponds to a static nonlinear filter of a linear time series.

4.1 Superfluid convection

Data from a superfluid convection cell [37] provides an example where the evidence for low-dimensional chaos is quite clear. Using discriminating statistics of dimension and forecasting error, we obtain about fifteen and forty sigmas of significance, respectively. This data was also analyzed by Farmer and Sidorowich [31], who found sizable increases in predictability using nonlinear rather than linear predictors.

4.2 Electroencephalogram (EEG)

That the brain should exhibit chaos is an idea that some authors have been unable to resist. Our own investigations so far have been mixed; some data sets exhibit nonlinear structure and some do not. A more systematic survey is clearly in order. In the meantime, we present two results, one positive and one negative. The two time series are from the same individual, eyes closed and resting; one is from a probe at the left occipital (O1), and the other from the left central (C3) part of the skull. The sampling rate is 150 Hz, and $N = 2048$ time samples are taken. The two time series are not necessarily contemporaneous. Using the dimension statistic, the first data set shows no significant evidence for nonlinearity, but the second data set exhibits about eight sigmas. Even in the significant case, we do not see any evidence that the time series is in fact low-dimensional (the correlation dimension ν does not converge with increasing embedding dimension m). We are formally able to reject the null hypothesis that the data arise from a linear stochastic process, but by comparing the surrogate data to the real data, we see no reason to expect that the "significant" data arises from a low-dimensional chaotic attractor.

4.3 Sunspots

The sunspot cycle has attracted perhaps more attention than any other time series, due to its interesting mixture of regularity and irregularity [38–43, 33].

Using both dimension and forecasting error, we can quite confidently reject the null hypothesis that the time series itself is linear stochastic; this is in agreement with the numerous authors [39–43, 33] who obtained better agreement using nonlinear models instead of linear models. However, when we expand the null hypothesis to include a static

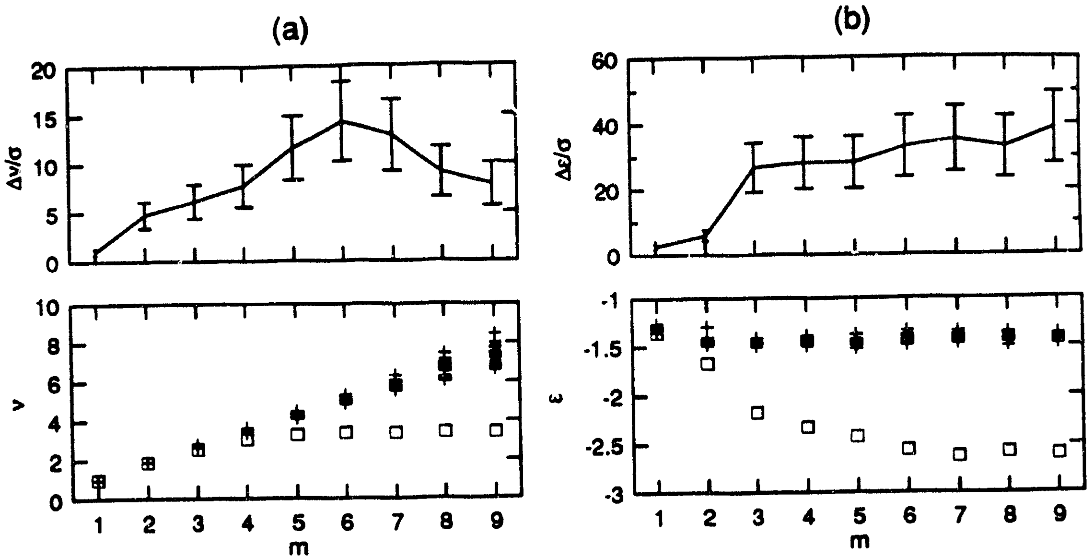

Figure 6: *Data from a fluid convection experiment exhibits very significant nonlinear structure, using (a) dimension, and (b) forecasting error. The top panel in these figures show the significance, measured in "sigmas," and the bottom panel shows the values of the statistics, with squares (□) for the original data and pluses (+) for the surrogates. Both panels plot these statistics against the embedding dimension m. Not only is the evidence for nonlinear structure statistically significant, but the estimated dimension of about* $\nu = 3.8$ *suggests that the underlying dynamics is in fact low-dimensional chaos.*

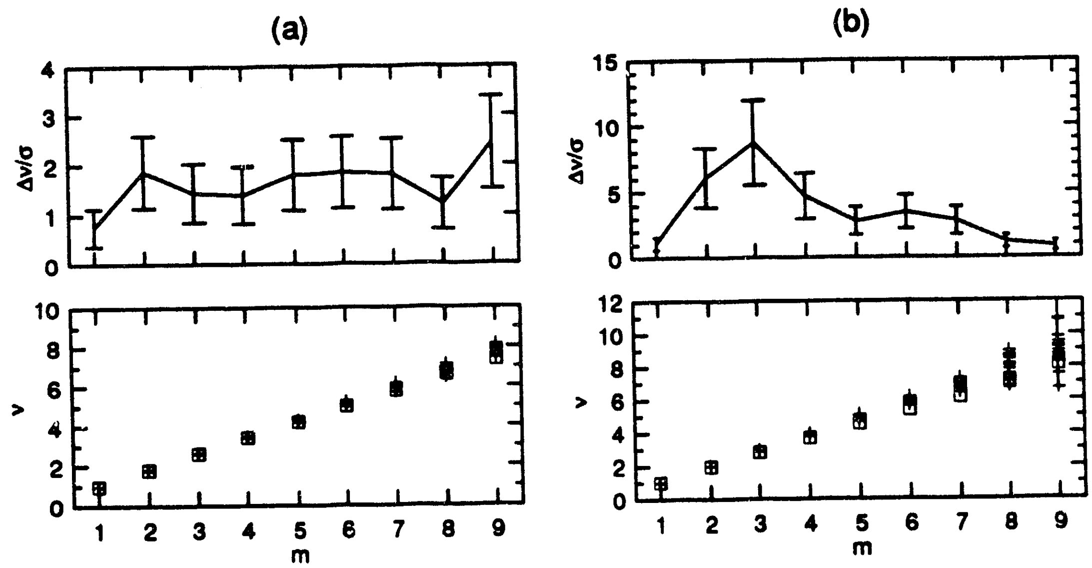

Figure 7: *Data from two electroencephalogram (EEG) time series. Using the dimension statistic, the first (a) shows no nonlinear structure, while the second (b) exhibits significant nonlinear structure at the eight sigma level. The evidence for low-dimensional chaos, however, is weak, since the estimated dimension increases almost as rapidly with embedding dimension for the original time series as it does for the surrogates.*

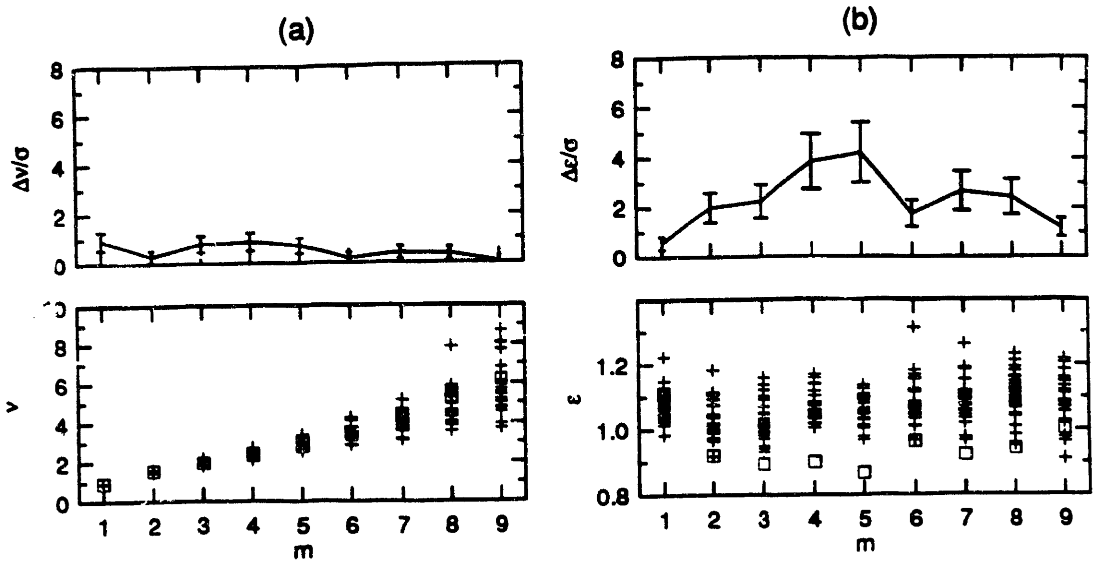

Figure 8: *Significance of nonlinearity in the Wolfer sunspot series. (a) using correlation dimension, and (b) using forecasting error. As with all the experimental time series reported here, Algorithm II was used to generate the surrogate data.*

nonlinear observation of an underlying linear stochastic process, the evidence (for dynamical nonlinear structure) is less dramatic. Using the dimension statistic, there is virtually no significance (of order one sigma). Using a forecasting algorithm, on the other hand, we do see significantly more predictability in the sunspot data than in surrogates, at about the five sigma level. This illustrates the advantage of having a battery of tests: the kind of nonlinear structure that one statistic is not sensitive to, another statistic may quite efficiently find.

5 Comparison to other work

Numerous authors have carefully compared their dimension estimates for real data against similar estimates for white noise. A few have extended this informal control to other forms of correlated noise; noteworthy are Grassberger [2], who showed that a reported dimension for climate data could be reproduced with data from an Ornstein-Uhlenbeck process; Kaplan and Cohen [17], who argued that fibrillation was not usefully described as chaotic, since data generated by randomizing the phases of the Fourier transform gave similar dimensions; and Ellner [44], who showed that a variety of "plausible alternatives" might adequately explain measles and chickenpox data, despite earlier claims of chaos.

Brock and coworkers in particular [21, 35, 45–47], and the economics community in general [15, 48–51], have been extremely active in the development of statistical tools for time series analysis. While the choice of null hypotheses for financial time series tends to be different than for more physical time series, the overall methodologies are quite similar. Classical statisticians are becoming increasingly aware of low-dimensional chaos (just as physicists are becoming increasingly aware of the importance of the statistical approach), and we cite Tong [13] as *the* review which most neatly and comprehensively ties these two fields together.

A slightly different approach, but with a very similar flavor, was applied by Chervin and Schneider [52] in assessing the statistical significance of predictions based on global climate models.

6 Conclusion

In this article, we have provided a framework for evaluating the statistical significance of evidence for nonlinearity in a stationary time series. (We do *not* seek to characterize *non*-stationary time series — see Refs. [53–56] for a discussion of some of the problems arising in the estimation of nonlinear statistics from nonstationary data.)

The test properly fails to find nonlinear structure in linear stochastic systems, and correctly identifies nonlinearity in several well-known examples of low-dimensional chaotic time series, even when contaminated with considerable noise. Our experiments with chaotic data found that using linear pre-processing to "bleach" out the linear correlations decreases the power of the test to detect nonlinearity. Consequently, we advocate a direct application of the method of surrogate data to the raw time series, instead of to a time series of residuals.

Finally, we illustrated the method with several experimental data sets, and confirmed the evidence for nonlinear structure in some systems, while failing to see such structure in other time series.

6.1 Discussion

We have described the method of surrogate data as a formal test for quantifying the statistical significance of the rejection of a particular null hypothesis. It is useful, however, to take a more informal approach, and view surrogate data as a control experiment. Having exhibited that some data gives a certain dimension, say, it is wise to compute the dimension for surrogate data as well, to make sure that the estimator is not being fooled by some feature (such as linear autocorrelation) that is also present in the random surrogate data.

In this case, there is some room for human judgement. For example, if the estimated dimension for the original data and the surrogate data are approximately equal and both small (or worse yet, if the surrogate data exhibit a *lower* dimension than the original data), then the conclusion that the data arises from low-dimensional dynamics is doubtful. It may be that the data is significantly nonlinear in the formal sense, so that the particular null hypothesis can be positively rejected, but that does not automatically imply low-dimensional chaos. In general, we advocate using a battery of statistical tests, not only to increase the chances of rejecting the null hypothesis, but to provide some intuitive insight into the nature of the nonlinearity. Eventually reducing the role of human judgement will require the development of more general null hypotheses, and the associated algorithms for generating surrogate data.

Any test which fails to reject the null hypothesis is strictly speaking inconclusive. Just because the original and surrogate time series have the same value for the discriminating statistic, that does not imply that they have the same underlying dynamics. On the other hand, if one observes evidence of low fractal dimension, but surrogate data shows the same low dimension, then claims based on the original evidence can be dismissed as not well-founded.

Considerable work remains to make the method of surrogate data a powerful and flexible tool for nonlinear analysis. In particular, we hope to expand the hierarchy of null hypotheses and to broaden the battery of discriminating statistics. The algorithms should also be extended to deal with multivariate time series and input-output systems. Further investigation of the effectiveness of various statistics for different null hypotheses in different situations will be valuable not only for increasing our ability to reject null hypotheses, but also for the more qualitative task of characterizing the nature of the nonlinearity that might be evidenced by one statistic but not another.

Acknowledgements

We are grateful to Martin Casdagli, Xiangdong He, William Brock, Blake LeBaron, Peter Grassberger, Mark Berge, Tony Begg, and Bette Korber for useful discussions. We also thank Bob Ecke for providing convection data, and Paul Nuñez and Arden Nelson for providing EEG data. This work was partially supported by the National Institute for Mental Health under grant 1-R01-MH47184-01, and performed under the auspices of the Department of Energy.

References

1. C. Nicolis and G. Nicolis, "Is there a climatic attractor?," *Nature* **311** (1984).
2. P. Grassberger, "Do climatic attractors exist?," *Nature* **323**, 609 (1986).
3. C. Nicolis and G. Nicolis, "Evidence for climatic attractors," *Nature* **326**, 523 (1987).
4. P. Grassberger, "Reply to: Evidence for climatic attractors," *Nature* **326**, 524 (1987).
5. J. Theiler, B. Galdrikian, A. Longtin, S. Eubank, and J. D. Farmer, "Using surrogate data to detect nonlinearity in time series," in *Nonlinear Prediction and Modeling*, M. Casdagli and S. Eubank, eds. (Addison-Wesley, 1991). To appear.
6. J. B. Ramsey and J.-J. Yuan, "Bias and error bars in dimension calculations and their evaluation in some simple models," *Phys. Lett. A* **134**, 287 (1989).
7. J. Theiler, "Statistical precision of dimension estimators," *Phys. Rev. A* **41**, 3038–3051 (1990).

8. F. Takens, "On the numerical determination of the dimension of an attractor," in *Dynamical systems and Bifurcations, Groningen, 1984* vol. 1125 of *Lecture Notes in Mathematics*, (Springer-Verlag, Berlin, 1985).
9. J. Holzfuss and G. Mayer-Kress, "An approach to error-estimation in the application of dimension algorithms," in *Dimensions and Entropies in Chaotic Systems – Quantification of Complex Behavior* vol. 32 of *Springer Series in Synergetics*, (Springer-Verlag, Berlin, 1986), pp. 114–122.
10. J. Theiler, *Quantifying Chaos: Practical Estimation of the Correlation Dimension*. PhD thesis, Caltech (1988).
11. J. Theiler, "Estimating fractal dimension," *J. Opt. Soc. Am. A* **7**, 1055–1073 (1990).
12. R. L. Smith, "Optimal estimation of fractal dimension," in *Nonlinear Prediction and Modeling*, M. Casdagli and S. Eubank, eds. (Addison-Wesley, 1991). To appear.
13. H. Tong, *Non-linear Time Series: A Dynamical System Approach*. (Clarendon Press, Oxford, 1990).
14. B. Efron, "Computers and the theory of statistics: thinking the unthinkable," *SIAM Review* **21**, 460–480 (1979).
15. J. A. Scheinkman and B. LeBaron, "Nonlinear dynamics and stock returns," *J. Business* **62**, 311–338 (1989).
16. G. E. Uhlenbeck and L. S. Ornstein, "Theory of Brownian motion," *Phys. Rev.* **36**, 823–841 (1930). Reprinted in *Noise and Stochastic Processes*, N. Wax, ed. (Dover, New York, 1954).
17. D. T. Kaplan and R. J. Cohen, "Is fibrillation chaos?," *Circulation Res.* **67**, 886–892 (1990).
18. M. Casdagli, D. Des Jardins, S. Eubank, J. D. Farmer, J. Gibson, N. Hunter, and J. Theiler, "Nonlinear modeling of chaotic time series: theory and applications," in *EPRI Workshop on Applications of Chaos* (1991). To appear.
19. W. Liebert and H. G. Schuster, "Proper choice of the time delay for the analysis of chaotic time series," *Phys. Lett. A* **142**, 107–111 (1988).
20. A. Cenys and K. Pyragas, "Estimation of the number of degrees of freedom from chaotic time series," *Phys. Lett. A* **129**, 227 (1988).
21. W. A. Brock, J. Lakonishok, and B. LeBaron, "Simple technical trading rules and the stochastic properties of stock returns," Technical Report 9022, Social Systems Research Institute, University of Wisconsin, Madison, (1991).
22. E. J. Kostelich and H. L. Swinney, "Practical considerations in estimating dimension from a time series," in *Chaos and Related Natural Phenomena*, I. Procaccia and M. Shapiro, eds. (Plenum, New York, 1987), pp. 141–156.
23. J. Theiler, "Efficient algorithm for estimating the correlation dimension from a set of discrete points," *Phys. Rev. A* **36**, 4456–4462 (1987).
24. P. Grassberger, "An optimized box-assisted algorithm for fractal dimensions," *Phys. Lett. A* **148**, 63–68 (1990).
25. P. Grassberger and I. Procaccia, "Characterization of strange attractors," *Phys. Rev. Lett.* **50**, 346 (1983).
26. P. Grassberger and I. Procaccia, "Measuring the strangeness of strange attractors," *Physica D* **9**, 189–208 (1983).
27. F. Takens, "Invariants related to dimension and entropy," in *Atas do* 13° Colóqkio Brasiliero de Matemática, (1983).
28. M. Sano and Y. Sawada, "Measurement of the Lyapunov spectrum from chaotic time series," *Phys. Rev. Lett.* **55**, 1082 (1985).
29. J.-P. Eckmann and D. Ruelle, "Ergodic theory of chaos and strange attractors," *Rev. Mod. Phys.* **57**, 617 (1985).
30. J.-P. Eckmann, S. O. Kamphorst, D. Ruelle, and S. Ciliberto, "Liapunov exponents from a time series," *Phys. Rev. A* **34**, 4971–4979 (1986).
31. J. D. Farmer and J. J. Sidorowich, "Exploiting chaos to predict the future and reduce noise," in *Evolution, Learning and Cognition*, Y. C. Lee, ed. (World Scientific, 1988), pp. 277–330.
32. M. Casdagli, "Nonlinear forecasting, chaos and statistics," in *Modeling Complex Phenomena, Proceedings of the Third San Jose State University Woodward Conference, April, 1991*, L. Lam and V. Naroditsky, eds. (Springer-Verlag, 1991). To appear.
33. M. Casdagli, "Chaos and deterministic versus stochastic nonlinear modeling," *J. Roy. Stat. Soc.* (1991). To appear.
34. M. Hénon, "A two-dimensional mapping with a strange attractor," *Comm. Math. Phys.* **50**, 69 (1976).
35. W. A. Brock, W. D. Dechert, and J. Scheinkman, "A test for independence based on the correlation dimension," Technical Report 8702, Social Systems Research Institute, University of Wisconsin, Madison, (1986).
36. B. Townshend, "Nonlinear prediction of speech signals," in *Nonlinear Prediction and Modeling*, M. Casdagli and S. Eubank, eds. (Addison-Wesley, 1991). To appear.
37. H. Haucke and R. Ecke, "Mode locking and chaos in Rayleigh-Benard convection," *Physica D* **25**, 307 (1987).
38. G. U. Yule, "On a method of investigating periodicities in disturbed series with special reference to Wolfer's sunspot numbers," *Philos. Trans. Roy. Soc. London A* **226**, 267–298 (1927).
39. H. Tong and K. S. Lim, "Threshold autoregression, limit cycles and cyclical data," *J. Roy. Stat. Soc. B* **42**, 245–292 (1980).
40. J. Kurths and A. A. Ruzmaikin, "On forecasting the sunspot numbers," *Solar Physics* **126**, 407–410 (1990).
41. A. Weigend, B. Huberman, and D. Rummelhart, "Predicting the future: a connectionist approach," *Int. J. Neural Systems* **1**, 193–206 (1990).
42. M. Mundt, W. B. Maguire II, and R. B. P. Chase, "Chaos in the sunspot cycle: Analysis and prediction," *J. of Geophys. Res.* **96**, 1705–1716 (1991).
43. D. Currie. Personal Communication.
44. S. Ellner, "Detecting low-dimensional chaos in population dynamics data: a critical review," in *Does chaos exist in*

Ecological Systems, J. Logan and F. Hain, eds. (University of Virginia Press, 1991). To appear.

45. W. A. Brock and S. M. Potter, "Diagnostic testing for nonlinearity, chaos, and general dependence in time series data," in *Nonlinear Prediction and Modeling*, M. Casdagli and S. Eubank, eds. (Addison-Wesley, 1991). To appear.
46. W. A. Brock and W. D. Dechert, "Statistical inference theory for measures of complexity in chaos theory and nonlinear science," in *Measures of Complexity and Chaos*, N. Abraham *et al.*, eds. (Plenum, 1989).
47. W. A. Brock and C. L. Sayers, "Is the business cycle characterized by deterministic chaos?," Technical Report 8617, Social Systems Research Institute, University of Wisconsin, Madison, (1986).
48. D. A. Hsieh, "Testing for nonlinear dependence in daily foreign exchange rate changes," *J. Business* **62**, 339–368 (1989).
49. D. A. Hsieh, "Chaos and nonlinear dynamics: application to financial markets." Preprint, to appear in *J. Finance*, (1990).
50. T.-H. Lee, H. White, and C. W. J. Granger, "Testing for neglected nonlinearity in time series models: A comparison of neural network methods and alternative tests." Preprint, Department of Economics, UCSD, (1989).
51. R. S. Tsay, "Nonlinear time series analysis: diagnostics and modelling." Preprint, Dept. Statistics, Carnegie Mellon University, (1988).
52. R. M. Chervin and S. H. Schneider, "On determining the statistical significance of climate experiments with general circulation models," *J. Atmos. Sci.* **33**, 405–412 (1976).
53. G. Mayer-Kress, "Application of dimension algorithms to experimental chaos," in *Directions in Chaos, Vol. I*, H. Bai-Lin, ed. (World Scientific, 1988).
54. A. R. Osborne and A. Provenzale, "Finite correlation dimension for stochastic systems with power-law spectra," *Physica D* **35**, 357–381 (1989).
55. A. Provenzale, A. R. Osborne, and R. Soj, "Convergence of the K_2 entropy for random noises with power law spectra," *Physica D* **47**, 361–372 (1991).
56. J. Theiler, "Some comments on the correlation dimension of $1/f^{\alpha}$ noise," *Phys. Lett. A* **155**, 480–493 (1991).

END

DATE FILMED

12/23/91