Statistical test for dynamical nonstationarity in observed time-series data

Matthew B. Kennel*
Engineering Technology Division, MS 8088, Oak Ridge National Laboratory, Oak Ridge, Tennessee 37831-8088
(Received 30 October 1996)

Information in the time distribution of points in a state space reconstructed from observed data yields a test for ''nonstationarity.'' Framed in terms of a statistical hypothesis test, this numerical algorithm can discern whether some underlying slow changes in parameters have taken place. The method examines a fundamental object in nonlinear dynamics, the geometry of orbits in state space, with corrections to overcome difficulties in real dynamical data which cause naive statistics to fail. [S1063-651X(97)09907-8]

PACS number(s): 05.45.+b

Since the discovery of time-delay embedding for state-space reconstruction [1–3] a significant effort has been devoted to the development of techniques to extract information in observed time-series data from a geometrical, dynamical viewpoint. Underlying nearly all of these techniques (Ref. [4] is a review) is an assumption of *stationarity*: the dynamical process, and hence the geometrical attractor containing the orbits, has not changed on long time scales the order of the length of the dataset. If not true, there may be significant behavior on time scales longer than may be reliably resolved with the given data; or perhaps experimental parameters, presumed fixed, have actually changed during the run.

Despite its nearly universal assumption, there is little previous literature on reliably testing for stationarity in physical situations. This work demonstrates a statistic and associated hypothesis test which sensitively detects nonstationary behavior given broadband and potentially chaotic data. A stationary dataset is presumed to to be sufficiently long to trace out a good approximation to the invariant measure. The algorithm quantifies ''how much has the invariant measure, as inferred from the observations, changed over long time scales,'' and whether ''this change is statistically significant.''

It is important to recognize the difference between a statistic and a test. A statistic is a quantity deterministically computable from a particular dataset; common examples include the sample mean or the value of a correlation integral. A test combines a statistic with some additional knowledge and assumptions concerning the distribution of that statistic expected under some particular hypothesis of interest. With any nontrivial dynamics, the value of a particular statistic will differ for separate finite-sized samples of the time series generated by the same physical process, and thus one needs a probability distribution to describe the ''expected fluctuations'' in addition to the ''expected value'' of a statistic over some class of processes. Armed with such knowledge, a test places the observed value of the statistic in the expected distribution: if it lies well outside the bulk of the expected distribution, one can ''reject the null hypothesis''—and, one presumes, its underlying physical basis—under which one computed the expected distribution. A test is more an algorithm or procedure than a number. For illustration, consider testing for chaos by using the largest Lyapunov exponent as a statistic. A positive Lyapunov exponent implies chaos, but one can only practically compute finite-time approximations to the exponent. There is a distribution of finite-time exponents over samples of the orbit, and thus even if the true system has a zero or negative largest exponent, it is possible that any particular finite-time sample will register a positive Lyapunov exponent statistic. A true statistical test would account for this source of fluctuations.

A useful test is obviously substantially more difficult to design than a statistic, especially if one wishes to entertain realistic and interesting hypotheses and account for potentially confounding effects likely to occur in experimentally observed data. There is a real tension between tests which are accompanied by rigorous analytical results and acceptance-rejection criteria from the classical statistical literature, and the undeniably heuristic statistics invented by physicists and mathematicians to quantify particular properties of a time series. The former may measure properties rather uninteresting or irrelevant to a dynamicist, and make assumptions likely to be massively violated by actual time series (thus rendering their acceptance-rejection procedures void), while the latter do not usually come with sufficiently reliable test procedures.

For the purpose of testing for stationarity, one's first thought is to imagine measuring any number of simple statistics, such as the mean or standard deviation, from the two halves of the time series, and using a standard stastistical hypothesis test based on their presumed equality, but such an approach is not particularly good. First, the statistic is arbitrary and not related to any natural geometrical properties of the attractor, which is the interesting object when analyzing dynamical data. Unless the particular statistic estimates a parameter deemed physically or dynamically important, such arbitrary choices are not particularly enlightening, and their power against various sorts of nonstationarity vary greatly.

Second, naively applying such procedures may greatly overestimate the significance of differences: observed dynamical data are far from uncorrelated, yet the simple, classical statistical estimations of confidence rely heavily on the notion of independent observations. For example, measuring empirical means of first and second halves of a chaotic dataset and performing the classical t test for their equality [5] will quite often spuriously (and vehemently) reject the

*Electronic address: kennel@yapunov.ucsd.edu

null hypothesis of stationarity, even when the data come from clean stationary experiments or well-known simple models such as the Lorenz attractor. Such methods do not reliably diagnose the intuitive concept of dynamical stationarity that a typical physicist imagines.

The present work attempts to rectify these two issues. I sidestep *direct* estimations of the invariant measure from observed data. Counting points in boxes of coarse-grained state space, as used for computing mutual information [6], for instance, introduces potentially problematic issues such as the arbitrary choice of box size, quantization artifacts, and poor scaling with the embedding dimension. Kernel density estimators are very computationally intensive in higher dimensions, and functionals or statistics on such estimates may require difficult multidimensional integrals. Further, their formalism does not naturally offer clear tests for significance. Instead, the solution adopted here quantifies nonstationarity using properties of *nearest neighbors* in state space. Neighbor searching is efficient and the estimates of consequent properties do not have a *prima facie* exponential ''curse of dimensionality.'' Neighbor statistics were used in Refs. [7,8] to determine minimum embedding dimension for reconstruction, and to quantify predictability of observed chaotic data [9]. In any case, accounting for the serial correlation present in experimental data is a major focus of this work, and neighbor information seems suited to this task.

As background motivation, suppose we have two empirical probability distributions ρ_1 and ρ_2, the measures in the first and second halves of the dataset. These distributions are over the full multidimensional state space, not scalar histograms. One wonders whether the observations are compatible with the hypothesis that $\rho_1=\rho_2$. Rewrite these distributions as

$$\rho_1=\frac{\rho_1+\rho_2}{2}+\frac{\rho_1-\rho_2}{2}=\rho_0+\delta\rho, \tag{1}$$

$$\rho_2=\frac{\rho_1+\rho_2}{2}-\frac{\rho_1-\rho_2}{2}=\rho_0-\delta\rho. \tag{2}$$

Given some $\mathbf{x}$ from the first half, consider the probability that its nearest neighbor in state space, $\mathbf{x}^{\mathrm{nn}}$, is also in the same half. Assuming $\rho(\mathbf{x}^{\mathrm{nn}})\approx\rho(\mathbf{x})$, $p_{\mathrm{same}}=\rho_1/(\rho_1+\rho_2)=(\rho_0+\delta\rho)/2\rho_0$. The expected proportion of matches is thus

$$\int d\mathbf{x}\rho_1(\mathbf{x})p_{\mathrm{same}}=\left\langle\frac{(\rho_0+\delta\rho)^2}{2\rho_0}\right\rangle_{\text{first half}}.$$

With the same argument for the second half, we find the overall expectation of seeing same-half matches is

$$E(\mathrm{same})=\left\langle\frac{(\rho_0-\delta\rho)^2+(\rho_0-\delta\rho)^2}{2\rho_0}\right\rangle=\left\langle\frac{\rho_0^2+\delta\rho^2}{2\rho_0}\right\rangle$$
$$=\frac{1}{2}+\left\langle\frac{\delta\rho^2}{2\rho_0}\right\rangle. \tag{3}$$

Nonstationarity, i.e., $\delta\rho\neq 0$, always increases this quantity, meaning that neighbors in state space are especially close in *time* when the distribution drifts over time.

The actual statistic ''feels'' the same effect but is more subtle: one collects the distribution of $D\equiv|\Delta(\mathbf{x})|\equiv|(T(\mathbf{x}^{\mathrm{nn}})-T(\mathbf{x})|$ for all observed $\mathbf{x}$, where $T(\ )$ denotes the time index of the point. Nonstationarity induces an excess number of small values of D than otherwise expected. The data need not be partitioned into sections.

Naively counting D from pointwise neighbors does not render a successful algorithm. As with the correlation dimension [10], one must exclude neighbors close in time because they are not independent of the reference point. If a prospective neighbor would result in $D=|T(\mathbf{x}^{\mathrm{nn}})-T(\mathbf{x})|<W$, ignore it and continue searching instead. The interval W is set to a characteristic autocorrelation time, perhaps 3–5 times the first minimum of mutual information.

Equally important, but less obvious, is accounting for serial correlation of neighboring trajectories: iterates of nearest neighbors often remain nearest neighbors, but this does not give *new* information. The present algorithm gathers multiple pairs of points and their neighbors which share the same Δ, i.e., are iterates of previously seen pairs, into the same *strand*. If the Δ associated with $\mathbf{x}(i)$ is the same as that for $\mathbf{x}(i-k)$ for any $k\in[1,W]$, append $\mathbf{x}(i)$ and its nearest neighbor to the strand associated with $\mathbf{x}(i-k)$. Otherwise, start a new strand with $\mathbf{x}(i)$, and its nearest neighbor with the as-yet sole element. Note that all pairs of a single strand need not be contiguous in time; there is allowed a gap up to W time steps long, though this large an upper bound is rarely realized in practice. Allowing such gaps prevents noise from damaging the proper accounting of neighbor correlations. Strands are collections of pairs of points which are nearest neighbors to one another, and which all share the same Δ. Their key utility is that they automatically correct for substantial serial correlation in the time series by bundling ''similar'' observations into a single unit (see Ref. [8] for details).

The final correction culls strands which share any underlying points, whether in the reference or neighbor part, because the information that they contribute is not completely independent. If any two strands share any underlying point, one is randomly deleted until no remaining strands have any points in common.

Without the corrections, the N used in the statistical test would be substantially larger than it ought to be, and would cause spurious null violations. With the corrections, the observed D values from the final strand set are nearly independent of one another, and one can apply standard techniques of statistical inference assuming that they are so.

We test the observed distribution of D for the final set of strands against the distribution expected under the null hypothesis of stationarity. If the process were complete stationarity, the time index of a neighbor would be independent from that of the reference; this fact may be used constructively to generate the expected distribution. For each observed strand, keep the reference portion as it is, but imagine that the neighboring portion could have started at any time index in $[1,N]$ with uniformly equal probability, excluding the interval W steps before the start and W steps after the end of the reference portion of the strand, with starting index j and length l. Each potential ''neighboring location'' has an associated D which contributes to the expected distribution. That is, for each potential neighboring location k

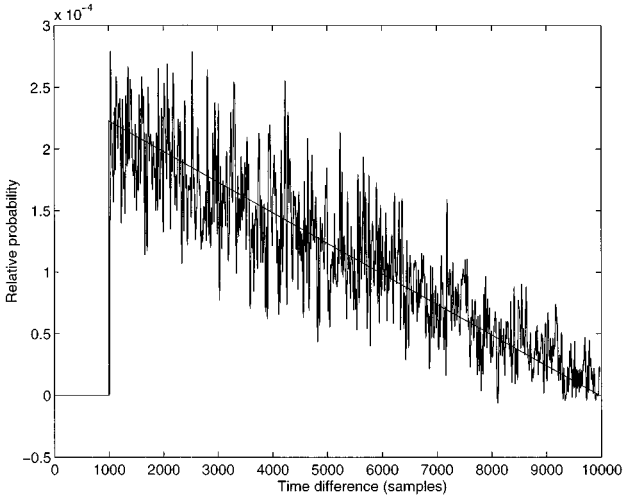

FIG. 1. An example of expected and observed probability density functions for D under the null hypothesis.

$\in[1,j-W]\cup[j+l+W,N]$, increment the count of expected $D=|k-j|$: $C(D)\rightarrow C(D)+1$. Repeat for all strands, generating a total $C(D)$, and normalize:

$$\rho_{\text{expected}}(D)=C(D)\Big/\sum_k C(k). \tag{4}$$

This procedure takes computation time $O(N_{\text{points}}N_{\text{strands}})$, and so may be slow. An approximation good for reasonably large N is the triangular-shaped function derived by assuming all strands are but a single point long, and that all reference locations were used with uniform probability:

$$\rho(D)=0, \quad D\in[0,W],$$

$$\rho(D)=M^{-1}\frac{N-D}{N-(W+1)}, \quad D\in[W+1,N-1],$$

with M chosen to normalize $\rho(D)$. This can be derived by assuming a uniform probability density in $[1,N]$ of a random variable j and another variable k with uniform density in $[1,N]$ excluding $[j-W,j+W]$, and integrating to find the distribution of $|k-j|$. The triangular distribution would be exact for infinitely long datasets, or if strands were always but one point long. The analytical approximation improves with the increasing size of the dataset, the circumstance when the computational effort grows increasingly burdensome. In practical application, switching to the analytical formula when $N_{\text{points}}N_{\text{strands}}>10^7$ has proven to be a reasonable, and conservative, choice. Figure 1 shows an example observed and expected distribution as an illustration of the typical shape. Keep in mind that using this approximation for the null hypothesis probability does not mean that one should actually evaluate the distribution of D on the actual data considering only points and not strands.

This picture suggests using something like the Kolmogorov-Smirnov test [5] on observed and expected $\rho(D)$, a standard procedure which tests for statistically significant differences in an empirical and theoreticaly expected cumulative distribution of a independent scalar random variable. In practice, that statistic turned out overly sensitive to medium-time-scale dynamical fluctuations observed even in stationary attractors. That is, depending on the dynamical system, there could sometimes be excessive rejections of the null hypothesis even under stationary conditions. Instead, a ''sign test'' provides an even simpler and effective test which is most sensitive to the long-time-scale changes characteristic of nonstationarity. Denote the location at the median of the expected distribution $\rho(D)$ as D^*: that is,

$$\int_0^{D^*}\rho(z)dz=\int_{D^*}^{N}\rho(z)dz=\frac{1}{2}. \tag{5}$$

Then one counts the proportion of actually observed strands which had values of $D<D^*$: N_{observed}. Under the assumptions of the null, one expects to find $p=N_{\text{observed}}/N_{\text{strands}}$ near $p_0=0.5$, because that was the criterion which generated D^*. Testing the proportion of independent events against a presumed underlying value is an elementary procedure in classical statistical theory, often called the ''binomial test.'' Under the null hypothesis,

$$z=(p-p_0)[N_{\text{strands}}{}^{-1}p_0(1-p_0)]^{-1/2} \tag{6}$$

has a calculable distribution, which in turn is nearly $N(0,1)$ for reasonably large N. $N(0,1)$ is a Gaussian random variable with zero mean and unit standard deviation. Imagine accumulating one for each strand with $D<D^*$, and zero for the remainder. If these were independent random events, then this sum has a distribution which converges to Gaussian by the central limit theorem. The normal approximation is very with $N>50$, very likely to be the case with experimental data from physical experiments. Thus if one observes $z>2.36$, the location of the 99% cumulative area of a standard Gaussian, one rejects the null at the 99% confidence level. Unlike some generalized tests of the equality of distributions, this particular sign test ensures that the violation be in the proper direction to be caused by nonstationarity, which causes especially *large* values of p and hence z. Significant, but negative, values of z suggest important nonuniform neighbor time differences distinct from nonstationarity. Strong low-frequency periodic behavior, resulting in long-term neighbor correlations, seems able to produce such results.

An efficient implementation is not difficult. Nearest neighbors in state space may be found in total $O(N\log N)$ time using data structures such as the k-d tree [11,12]. A reverse index from point indices to those strands which contain them is sortable in $O(N\log N)$ time, making the culling efficient in addition. Given a culled set of strands, the statistic may be evaluated rapidly. As with most algorithms of this type, the evaluation of near neighbors dominates all other aspects of the computation time. On a standard Pentium workstation, the author's implementation of the statistic takes approximately 5 s to evaluate, with $N=30\,000$ points embedded in R^3.

A computationally expensive but valuable confirmation protocol is to examine the proportion of z values which reject the null as computed using randomly selected contiguous subsegments of lengths $N^*<N$ of the original data: a ''poor-man's bootstrapping.'' With authentically nonstationary behavior, this proportion rises steadily with N^*. One may also examine the behavior with N^* of p averaged over

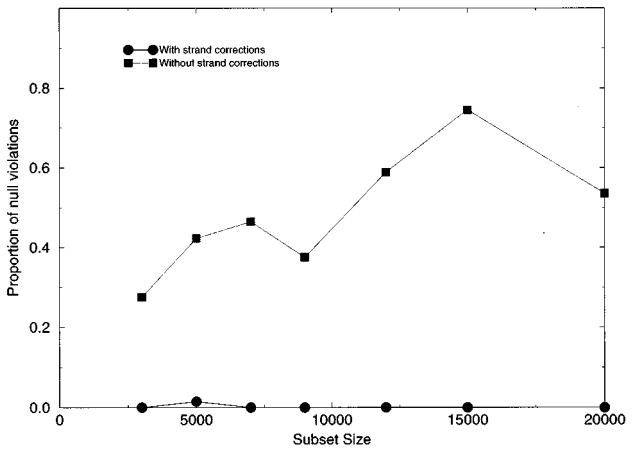

FIG. 2. Proportion of rejections as a function of subsample size when strand corrections are turned off and on. The data set is from a stationary low-dimensional chaotic circuit, but without strand corrections there is frequent suprious rejection of the null.

subsamples as well as its effective significance via Eq. (6) to check whether p increases consistently with largely N^*, and is not just wider than $N(0,1)$, with some resamples being anomalously large and some anomalously small.

Figure 2 demonstrates the importance of the strand corrections. The data come from an experimental nonlinear circuit used to investigate synchronization and chaotic communications. The dynamics are known to be low dimensional, and the data clean. In any useful sense the data are quite stationary, yet the uncorrected statistic shows large violations, as would naive tests found in statistical textbooks such as equality of means or variances tried on first and second halves. By constrast, the present method shows no spurious null violations above the expected proportion. Consistent with the bootstrapping analysis, the dataset *in toto* violates the null strand without corrections, but is consistent when those corrections are reinstated.

The next example demonstrates a more concrete engineering application. The dataset was the pressure drop across a 15-cm gap in a ''fluidized bed reactor,'' consisting of glass particles 2.7 mm in mean diameter in a 10-cm-diameter vertical cyclinder with air blown at constant flow from the bottom, a small-scale model of industrial chemical reactors. For some external parameter regimes, the mass of particles undergoes complex motion, which appears to be a combination of low-dimensional bulk dynamics and small-scale high-dimensional turbulence of the individual particles [13]. The observed variable was a pressure difference between two vertically separated taps. Figure 3 shows portions of time-delay embedding of orbital sections of the dataset taken at the same experimental parameters, and one when the air flow was boosted by 5%. Figure 4 shows time-series excerpts taken from the two conditions. The change in the attractor is rather subtle, and difficult to diagnose reliably by eye. The statistic distinguishes them easily: Figure 5 shows the bootstrapping result on three datasets; one under stationary conditions, one with a step change to the higher flow, and one with a slow ramp to that same flow. The lower right plot in Fig. 3 (not the upper left) is from the data taken at a different flow rate than the others.

The author applied the method to quite a variety of data sets, simulated and experimental, and it yields correct and appropriate results in all cases found so far. It is not highly sensitive to reconstruction parameters within a reasonably large range, and does not require that the data be known *a priori* to be clean and low dimensional: it is not clear whether the fluidized bed datasets analyzed herein are better described as ''chaos'' or ''very noisy periodicity.''

Any statistical inference is only as good as its assumptions, in this case, that all strands are completely independent, and that in stationary systems nearest-neighbor time delays are distributed completely uniformly. This is indeed true for unbiased stochastic draws from probability densities, but is only an approximation for real dynamical systems. In constrast to simple assumptions of classical statistics, the diversity of possible behaviors under nonlinearity makes it very difficult to construct any interesting test where *chaos* is the null, something surrogate data methods do not attempt. The present test does so by testing for one specific aspect of dynamical systems, and making an approximation that em-

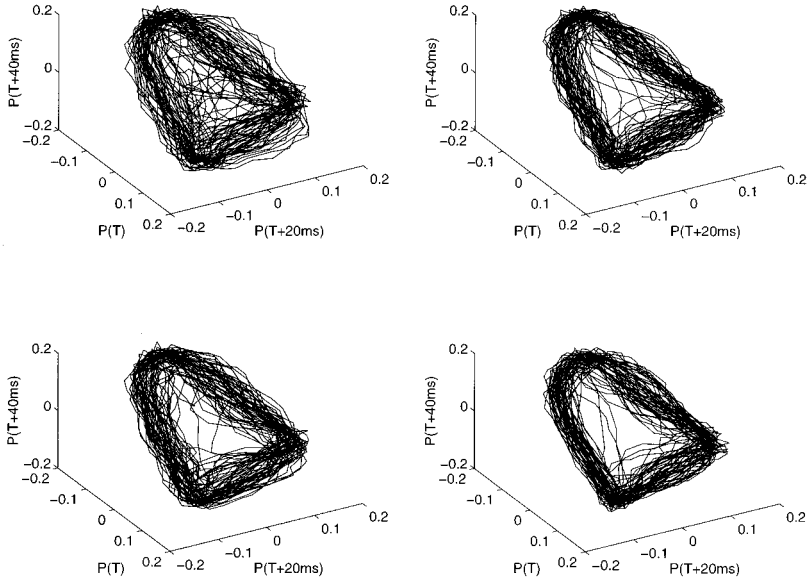

FIG. 3. Phase-space plots of the differential pressure signal from a fluidized bed reactor. Three are from the same parameters, one is different.

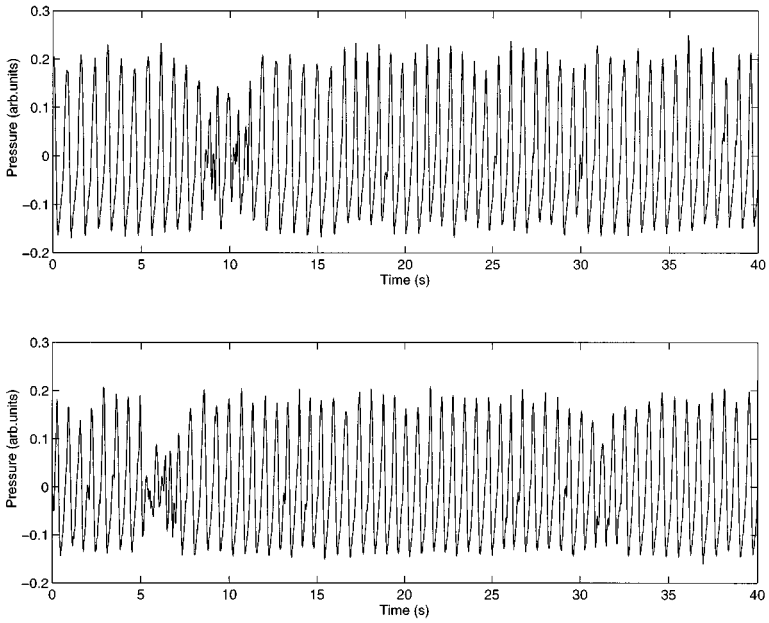

FIG. 4. Time series excerpts from a fluidized bed, with slightly different settings of the external parameter. The distinction is difficult to make by eye.

pirically appears to be reasonably good. The main problem is that the ''level'' of the test, the frequency of finding $z>2.36$ under stationary conditions, is not exactly calibrated to the supposed 1%. This does not seem to be resolvable in general unless has had large amounts of orbits on the specific attractor observed in stationary conditions, from which one could generate the actual distribution of z in Eq. (6) instead assuming the normality. If the data were truly drawn independently and randomly from probability functions ρ_1 and ρ_2, the approximation would be exact. The value of this method is an approach and approximation that works for many realistic datasets without requiring a large database of previously observed stationary orbits. This issue is only a major concern if one wishes to detect nonstationarity with the maximum power and know the proportion of false rejections; in many practical applications where data are reasonably copious the statistic detects the physically relevant nonstationarity at a very high confidence level, so that one can use a very conservative threshold. For instance, in the current fludized bed application the typical z value under nonstationarity was approximately 20, a definitive rejection of the null.

The exact power of the statistic depends, of course, on the particular sort of state-space reconstruction that might be used, as that choice can influence the existence and properties of neighbors used to compute the stationarity test. As with all dynamical statistics operating on state space, some external knowledge of the appropriate time scale and dimension is needed to ensure that the neighbor information in the reconstructed space is not wholly spurious. In the author's experience, the present method is less sensitive than other statistics such as correlation dimension or Lyapunov exponent, but admittedly this is a difficult notion to quantify. The particular z value may change with embedding parameters, but the presence or absence of a strong rejection usually does not change over a reasonably wide range of parameters.

It may conceivably be possible that a particular low-dimensional reconstruction might mask the effects of true physical nonstationarity, but this circumstance does not seem to be generic; most of the time different reconstruction parameters will reveal the nonstationarity to some degree or another. If the full probability distribution in the true state space is $\rho(\mathbf{y})$, then in a partially reconstructed state space $\rho(\mathbf{x})$ is generically the sum over regions of $\mathbf{y}$ mapping into $\mathbf{x}$ of $\rho(\mathbf{y})$ multiplied by the Jacobian determinant for the change of variables. Unless the choice of the projection from $\mathbf{y}\rightarrow\mathbf{x}$ is chosen carefully, part of the information in every variable of $\mathbf{y}$, and thus the nonstationary $\rho(\mathbf{y})$ is going to show its effects in the $\mathbf{x}$ space, if the variables are sufficiently coupled. The principle is similar to that governing the reconstruction of a dynamical system itself [3].

There is a whole class of related statistics that use the same neighbor principle. Instead of $|(T(\mathbf{x}^{\mathrm{nn}})-T(\mathbf{x})|$ one may use the distribution of any general function $f(\mathbf{x}^{\mathrm{nn}},\mathbf{x})$. For instance, $f(\cdot,\cdot)$ may be the ''indicator function,'' yielding 1 if both its arguments come from the same dataset, and 0 otherwise. This provides a test for equivalence of the two data sets, and can also yield a distance measure. The author has already done so to implement a ''change-point detector'' which accurately finds the particular moment in time when some underlying parameter is changed, and the statistical confidence of its authenticity. Choosing

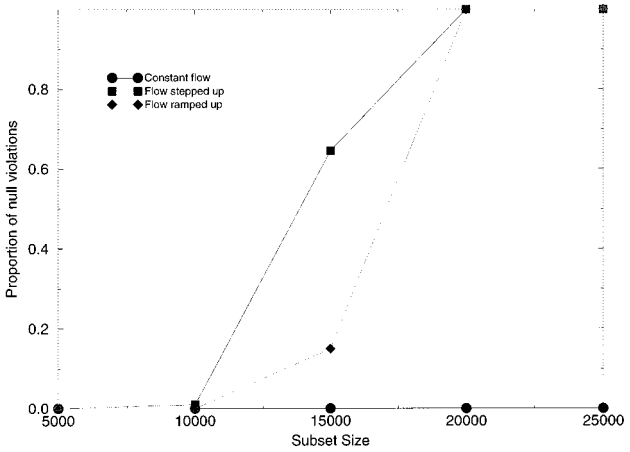

FIG. 5. Proportion of rejections for stationary, step change, and ramped air flow.

$f=|\sin[\Omega T(\mathbf{x}^{\mathrm{nn}})+\Phi]-\sin[\Omega T(\mathbf{x})+\Phi]|$ yields a test for the presence, and statistical significance, of a slow periodic modulation of the underlying attractor. If one has measured some other slowly varying signal $y(t)$, then the choice of $f=|y[T(\mathbf{x}^{\mathrm{nn}})]-y[T(\mathbf{x})]|$ provides a test of whether there is any statistically significant dynamical correlation betwen y and the pattern of orbits traced out by $\mathbf{x}$. For instance, one might wish to test the hypothesis that some high-frequency weather patterns in $\mathbf{x}$ are significantly correlated with a slow variable such as historical CO_2 levels. These variations, alternative stationarity algorithms based on the correlation integral, as well as more extensive experimental results will be investigated in the author's forthcoming research.

Isliker and Kurths [14] proposed testing the one-dimensional marginal distribution of the data for stationarity, but this ignores dynamical time-domain information, and their method did not appear to account for serial correlation. Brown, Rulkov, and Tracy [15] synchronized empirical ordinary differential equation models to time series, and proposed using a long-term increase in synchronization error as a measure of nonstationarity. This method appears powerful, and relies on nontrivial dynamical information, but requires clean low-dimensional data and does not provide an obvious statistical test. Kantz [16] quantified attractor differences through the correlation integral, but did not provide a hypothesis test.

The author is indebted to discussions with C. Stuart Daw, Charles Finney, Ke Nguyen, and Martin Casdagli. This research was supported by the U.S. Department of Energy.

[1] N. H. Packard, J. P. Crutchfield, J. D. Farmer, and R. S. Shaw, Phys. Rev. Lett. **45**, 712 (1980).
[2] F. Takens, in *Dynamical Systems and Turbulence, Warwick 1980*, edited by D. Rand and L.-S. Young, Lecture Notes in Mathematics No. 898 (Springer, Berlin, 1981), p. 366.
[3] T. Sauer, J. A. Yorke, and M. Casdagli, J. Stat. Phys. **65**, 579 (1991).
[4] H. D. I. Abarbanel, R. Brown, J. Sidorowich, and Lev Sh. Tsimring, Rev. Mod. Phys. **65**, 1331 (1993).
[5] W. H. Press, S. T. Teukolsky, W. T. Vetterling, and B. P. Flannery, *Numerical Recipes in C*, 2nd ed. (Cambridge University Press, Cambridge, 1992).
[6] A. M. Fraser and H. L. Swinney, Phys. Rev. A **33**, 1134 (1986).
[7] Matthew B. Kennel, R. Brown, and H. D. I. Abarbanel, Phys. Rev. A **45**, 3403 (1992).
[8] Matthew B. Kennel and H. D. I. Abarbanel (unpublished).
[9] M. C. Casdagli, R. Stat. Soc. B **54**, 303 (1991).
[10] J. Theiler, Phys. Lett. A **155**, 480 (1991).
[11] J. H. Friedman, J. L. Bentley, and R. A. Finkel, ACM Trans. Math. Softw. **3**, 209 (1977).
[12] R. F. Sproull, Algorithmica **6**, 579 (1991).
[13] C. S. Daw *et al.*, Phys. Rev. Lett. **75**, 2308 (1995).
[14] H. Isliker and J. Kurths, Int. J. Bifurcation Chaos Appl. Sci. Eng. **3**, 1573 (1993).
[15] R. Brown, E. R. Rulkov, and N. F. Tracy, Phys. Rev. E **49**, 3784 (1994).
[16] H. Kantz, Phys. Rev. E **49**, 5091 (1994).