Pramana – J. Phys. (2021) 95:141
https://doi.org/10.1007/s12043-021-02181-2

Improved tests for non-linearity using network-based statistics and surrogate data

M C MALLIKA[1], K ASOKAN[2], K S ANIL KUMAR[3], T R RAMAMOHAN[4] and K SATHEESH KUMAR[1],*

[1]Department of Futures Studies, University of Kerala, Kariavattom 695 581, India
[2]Department of Mathematics, College of Engineering, Trivandrum 695 016, India
[3]University of Kerala, Palayam, Thiruvananthapuram 695 034, Kerala, India
[4]Department of Chemical Engineering, M.S. Ramaiah Institute of Technology, MSR Nagar, Bengaluru 560 054, India
*Corresponding author. E-mail: kskumar@keralauniversity.ac.in

MS received 10 March 2021; revised 26 April 2021; accepted 28 April 2021

Abstract. We report the results of studies of improved tests for non-linearity based on time series-induced network statistics and surrogate data. We compare results from the network-based statistics with the earlier tests available in the literature and demonstrate the superiority of these tests over the previous tests for several systems. The method we propose is based on constructing a network from a time series and using easily computable parameters of the resulting network such as the average path length, graph density and clustering coefficient as test statistics for the surrogate data test. These statistics are tested for their ability to distinguish between nonlinear processes and linear noise processes, using surrogate data tests on time series obtained from the Rössler system, the Lorenz system, the Henon map, the logistic map and an actual experimental time series of wind speed data, and compared with popularly used time series associated statistics. The network-based statistics are found to distinguish between the nonlinear time series and surrogates derived from the data to a higher degree than the commonly used time series-based statistics, even in the presence of measurement noise and dynamical noise. These statistics may thus prove to be of value in distinguishing between time series derived from nonlinear processes and time series obtained from linearly correlated stochastic processes even in the presence of measurement noise and dynamical noise. The results also show that the efficiency of the network parameters is not exacerbated by the presence of outliers in the given time series.

Keywords. Nonlinearity; surrogate; time-series-induced network; test statistics.

PACS Nos **05.45.−a; 05.45.Tp; 05.45.Pq; 05.45.Ac**

1. Introduction

An essential step in analysing time series data, which contain persistent apparent random fluctuations, is to judge whether the source of apparent randomness in the data is purely stochastic or deterministic chaos or a mixture of deterministic and stochastic effects. While the output of a noise process could well be a random time series, it is equally possible for a low-dimensional deterministic system also to generate a time series that appears random [1]. Deterministic low-dimensional chaotic systems have many characteristic features such as attractors with small fractional dimension, positive Lyapunov exponents and broad power spectrum, and there exist various tools to quantitatively estimate these discriminating measures from a single time series originating from such a system using the idea of phase-space reconstruction [2–4]. However, it has been demonstrated that many of these features could also be shown by linear stochastic systems [5,6] and as such it is difficult to distinguish a deterministic chaotic time series from a purely random series based solely on numerical values of these discriminating measures unless there is *a priori* knowledge about the underlying system. Thus, before values of these measures can be accepted as evidence for chaos, it is essential to establish the nonlinearity of the data to preclude the possibility that the data come from a linear noise process.

Published online: 19 August 2021

The surrogate data test is widely held to be a reliable technique for discriminating a nonlinear process from a noise process [7,8]. This is basically a statistical procedure for testing the null hypothesis that the given time series is a linear Gaussian process, possibly distorted by a nonlinear measurement function. One then generates an ensemble of random data sets (surrogates) that are consistent with the null hypothesis but preserve the linear properties of the original data. Thus, the surrogates are like independent realisations of the process, which generates the original data if that process satisfies the null hypothesis. Deviation from the null hypothesis is estimated by comparing the value of some discriminating statistic on the data with the distribution of the corresponding values from the surrogates. If there is a significant difference in the values, the hypothesis is rejected, and the original data are considered nonlinear at the specified level of significance. The statistical interpretation of the surrogate test depends on the choices of the null hypothesis, the test statistic used and the method of generating surrogates. Various attempts have been made to improve the efficiency of the surrogate data test by using better test statistics and improving the methodology [9–11]. Chavez and Cazelles [12] have recently proposed a method combining wavelet transforms and non-stationary surrogates to detect short-lived spatial coherent patterns from multivariate time series. While surrogate data tests are generally used to distinguish nonlinear deterministic systems from the linear stochastic system, Hirata and Shiro [13] recently proposed a method for testing the nonlinear stochasticity of a given system.

The purpose of this paper is as follows. First, we present some new test statistics for a surrogate data test, which are based on the idea of constructing a complex network from the time series. Second, the new test statistics are compared with a couple of other statistics in terms of their efficiency and robustness when the original data are possibly distorted by some measurement noise or are affected by the dynamical noise. Further, there is always room for improvement in the detection of nonlinearity in a time series, especially when it is corrupted by measurement noise or dynamical noise. Keeping this in mind, we report results of network-based statistics used on surrogate data as an improved test for nonlinearity in the presence of measurement noise and dynamical noise. The concept of generating a complex network from a time series was first introduced by Zhang and Small [14], and since then, many different methods for constructing networks have been proposed [15–18] and used to derive qualitative information about the original time series. Various applications of this approach have also been reported in [19–27]. We demonstrate that statistics derived from the network constructed from the time series can be used in surrogate tests to yield better results in distinguishing linear noise processes from nonlinear processes that are possibly corrupted by measurement noise or affected by a systematic source of dynamical noise. The new test statistics are compared with time invariance and time-reversal statistics on several different sets of time series data at different noise levels. Time series generated by the Lorenz equations, the Rössler system, the Henon map and the logistic map, as well as a field data of wind speed – all of which are known to be deterministic systems exhibiting chaotic behaviour – are used in the simulations at various levels of noise. In all the cases, the network parameters yield larger values for the significance of variation at all levels of noise, making it possible to judge the nature of the data accurately. While there are many studies on the effect of measurement noise in surrogate data tests, the effect of dynamical noise has not been addressed sufficiently, except in a few cases [28].

2. The method of surrogates

The Fourier transform method [7,29] of creating surrogates for the null hypothesis that the time series is a linear Gaussian process, is rather straightforward. Given the time series $x_t,\ t = 1, 2, \ldots, N$, we first take the discrete Fourier transform

$$z_n = \sum_{t=1}^{N} \mathrm{e}^{2\pi int/N} x_t. \tag{1}$$

Keeping the absolute values of z_n fixed, the phases are now randomised, by choosing K values ϕ_k uniformly distributed between 0 and 2π, and constructing a set of Fourier transformed data

$$z_n^k = |z_n|\mathrm{e}^{i\phi_k}, \quad k = 1, 2, \ldots, K. \tag{2}$$

Inverting the Fourier transform now gives the K surrogates

$$y_t = (1/N) \sum_{n=1}^{N} \mathrm{e}^{-2\pi int/N} z_n^k. \tag{3}$$

Constructed this way, the surrogates have the same power spectrum (or equivalently the same autocorrelation function) as the original data. Linear Gaussian processes are essentially determined by their mean and autocorrelation functions and their realisations should differ only in their Fourier phases. Hence, the surrogates can be considered as independent realisations of the original data, if it were consistent with the null hypothesis.

A slight modification of this procedure is used to generate surrogates for the extended null hypothesis that the observed time series is a static non-linear monotonic function of a linear Gaussian process [7,29]. In this case, the given time series is first rescaled so that its values are Gaussian, and then surrogates are constructed as before and finally, the surrogates are rescaled to match the original series.

Deviations from the null hypothesis are now estimated by comparing the value of some discriminating statistic of the given data with the distribution of values obtained from the surrogates. The null hypothesis is accepted or rejected based on the value of the significance of difference given by [30]

$$S = \frac{\mu_{\mathrm{orig}} - \mu}{\sigma}, \tag{4}$$

where μ and σ are the mean and standard deviations of the distribution of the statistic as computed from the surrogates and μ_{orig} is the mean of the statistic of the original data. The null hypothesis is rejected at 95% confidence level if $S > 2$.

There are many discriminating statistics that are commonly used to test nonlinearity using surrogate data. Maiwald *et al* [31] presented a comparison of about a dozen of these statistics under different combinations of null hypotheses and surrogate types, and demonstrated that the *time-invariance* statistic introduced by them leads to better results in all the cases considered. For a set of data x_n, time invariance is given by

$$\mathrm{TIV} = \max\left\{\frac{\#\{|x_{i+1}| > |x_i|\}}{\#\{|x_{i+1}| < |x_i|\}}, \frac{\#\{|x_{i+1}| < |x_i|\}}{\#\{|x_{i+1}| > |x_i|\}}\right\}. \tag{5}$$

Another powerful statistic is the *time-reversal asymmetry*, which measures deviations from time reversibility which characterises linear systems, given by [29]

$$\mathrm{TR} = \frac{\sum (x_i - x_{i-\tau})^3}{\sum (x_i - x_{i-\tau})^2}. \tag{6}$$

In the next section, we introduce some new test statistics based on a network representation of time series and compare their performance with TIV and TR in identifying non-linearity in time series data.

3. Network from time series

Complex networks are generally used to study (usually spatially extended) complex systems by representing the basic elements or subsystems within the complex system by nodes and the relations or interactions among them by edges connecting the nodes. The network representation allows one to analyse the relations among the system elements on local or global scales by using several network parameters and statistical descriptors developed from classical graph theory. The method has been largely successful in understanding the critical structural and dynamical properties of complex systems.

In recent years several researchers have attempted to study the dynamics of time series by transforming them into complex networks using different techniques to construct nodes and define edges. Zhang and Small [14] studied a pseudoperiodic time series by representing each cycle as a node and connecting them by an edge if the phase-space distance or correlation coefficient between them is less than a threshold value. Yang and Yang [15] and Gao and Jin [16] extended this method by using individual state vectors in the phase-space embedding of the time series as nodes and the correlation coefficient as the threshold parameter for defining edges. Lacasa *et al* [17] used individual observations of the time series data as nodes and a visibility condition among the nodes to construct connections. Other methods include using phase-space vectors and their neighbourhood relations for defining networks [32] or the recurrence matrix of time series for defining adjacency in networks [18]

In this work, we propose to use the parameters related to the structural properties of the complex network constructed from a time series as test statistics for a surrogate data test. To construct the network, the individual state vectors in the reconstructed phase space are used as nodes and the Euclidean distance between them selected as the criterion for defining edges. More specifically, given the time series $x_1, x_2, x_3, \ldots, x_n$, we construct the m-dimensional state vectors $\mathbf{X}_t$ with time delay τ as

$$\mathbf{X}_t = (x_t, x_{(t+\tau)}, x_{(t+2\tau)}, \ldots, x_{(t+(m-1)\tau)}),\quad t = 1, 2, 3, \ldots. \tag{7}$$

Taken's embedding theorem and its extensions [3,4,33] assert that the dynamics of $\mathbf{X}_t$ in the m-dimensional space will be topologically identical to the dynamics of the original system which generates the time series x_i, under very general conditions. There is no restriction, in principle, on the value of τ, but a sufficient (but not necessary) condition on the embedding dimension is $m \geq 2d + 1$, where d is the fractal dimension of the underlying attractor [4]. However, in practice, embedding dimensions less than $2d + 1$ work in many cases.

To construct the network, we assume the state vectors $\mathbf{X}_t$, $t = 1, 2, \ldots$ as the nodes. Connections between nodes are established based on the Euclidean distance d_{ij} between nodes $X_i = (x_{i1}, x_{i2}, \ldots, x_{im})$ and $X_j = (x_{j1}, x_{j2}, \ldots, x_{jm})$ given by

$$d_{ij} = \sqrt{|x_{i1} - x_{j1}|^2 + |x_{i2} - x_{j2}|^2 + \cdots + |x_{im} - x_{jm}|^2}. \tag{8}$$

Nodes X_i and X_j are connected if $d_{ij} < \epsilon$ for a fixed value of the critical length ϵ_c. Gao and Jin [34] proposed a method to fix the critical value of ϵ by computing the normalised maximum size of subgraph distribution for varying length. In this paper, we computed the number of edges (NNE), normalised to the number of edges of the complete graph with the same number of nodes, by varying the cut-off length to form an edge. The critical value of ϵ corresponds to the maximum rate of increase of NNE with respect to the varying cut-off length.

Several structural properties are associated with the network thus constructed, which can serve as a test statistic for the surrogate test. In this paper, we propose to use three such metrics, which are easily computable from the network, namely the graph density (GD), the average path length (APL) and clustering coefficient (CC). The graph density of a network is the ratio of the number of edges to the maximum number of edges and is given by

$$\mathrm{GD} = \frac{2E}{N(N-1)}, \tag{9}$$

where E is the number of edges and N is the number of nodes in the network. The average path length is the average of the shortest path length, averaged over all pairs of nodes, and is computed as

$$\mathrm{APL} = \frac{1}{N(N-1)} \sum_{i \neq j} d_{ij}, \tag{10}$$

where d_{ij} is the length of the shortest path between nodes i and j, which is taken as zero if i and j are disconnected. The (global) clustering coefficient is defined by

$$\mathrm{CC} = \frac{3 \times \text{Number of triangles}}{\text{Number of triplets}}, \tag{11}$$

where a set of three nodes $\{i, j, k\}$ is called a triangle if every two of them are connected by an edge and a triplet if i is connected to j and j is connected to k. The clustering coefficient is a measure of the degree to which nodes tend to cluster together within a graph.

4. Results and discussion

We compare the performance of the parameters calculated from the complex network resulting from a time series for a surrogate data test against the null hypothesis that a linear Gaussian process generates the time series. We are mainly concerned with testing the robustness of the statistic for correctly identifying nonlinearity in the data even when it may be corrupted by the noise that enters through the measurement process or through a systematic dynamical source. This aspect is significant when testing determinism in experimental data as such data are usually affected by the measurement noise while the actual dynamics might be deterministic and nonlinear. Besides, noise may also be generated in the system by neglecting higher-dimensional processes of low amplitude, which is common in realising a nonlinear process. Even numerical solutions of a nonlinear system have these contributions from roundoff error.

4.1 *Comparison of network and time series-based parameters*

We consider five different sets of chaotic time series for the analysis, including those generated by continuous dynamical systems (Rossler system and Lorenz system), discrete systems (logistic and Henon maps) and also a field data of observed wind speed measured at regular intervals. We tested the procedure first on the Rössler system. The Rössler system is governed by [35]

$$\begin{aligned} \dot{x} &= -y - z \\ \dot{y} &= x + ay \\ \dot{z} &= b + (x - c)z \end{aligned} \tag{12}$$

which is chaotic for $a = 0.1$, $b = 0.1$ and $c = 18$.

To carry out the analysis on the Rössler system, we integrated eqs (12) numerically and sampled the output of integration at an interval of $\delta t = 0.4$ time units to obtain the temporal sequence of the solution. The sequence of the first component of the solution x_i of length 2500 has been considered as a typical time series generated by the purely nonlinear deterministic dynamics of the Rössler system. By adding white noise to this series we obtain the noisy deterministic series

$$y_n = x_n + nl\ N(0.1), \tag{13}$$

where $N(0, 1)$ denotes an uncorrelated Gaussian noise process with zero mean and unit variance and nl is a parameter to control the noise level. Thus, y_n represents a chaotic time series evolving under nonlinear deterministic laws but distorted by the measurement noise. The surrogate test is carried out on y_n at different levels of noise determined by the value of the parameter nl, at each level constructing several surrogates, using the amplitude-adjusted Fourier transform method [7,36], consistent with the null hypothesis that the given series is from a linear Gaussian process, possibly affected by a nonlinear measurement function. Thus, for a fixed value of nl, 40 surrogates are generated from y_n, and the null hypothesis was tested using as test statistics time invariance (TIV), time reversal asymmetry (TR), also the graph density (GD), average path length (APL) and clustering coefficient (CC) of the networks induced by the time series and the surrogates. In order to fix the critical value of the cut-off length ϵ for forming an edge, we constructed networks for increasing values of ϵ and computed the number of edges normalised to maximum possible edges (NNE), which is that of the complete graph with the same number of nodes. The value of ϵ corresponding to the maximum rate of increase is taken to be the critical value to construct the network. However, for the surrogate data test, a smaller value is also found to perform equally well. For a time series of noise level nl, the network structural properties and time series-based statistics were computed for both the original data and the distribution of surrogates, and the value of the significance of difference S, given by eq. (4), is calculated. The whole procedure is then repeated for increasing values of noise level (nl) to assess the robustness of the network-based test statistics and their sensitivity to measurement noise.

The values of NNE for increasing values of ϵ are plotted in figure 1. The maximum rate of increase of NNE occurs at $\epsilon = 20.6$, which was used to construct the networks. The average values of S with standard error, obtained by repeating ten sets of noise values, for a set of different values of nl are plotted in figure 2. We can make two observations from this figure. First, in the absence of noise ($nl = 0$), values of S for network-based parameters are greater than the values of S for the time series parameters TIV and TR, indicating that network-based test statistics are more capable of distinguishing nonlinearity and determinism present in a given time series compared to the best traditional time series-based parameters. Secondly, values of S for the time series parameters decrease rapidly with increasing values of the noise level nl. In contrast, the values of S computed from network-based parameters maintain their superiority with values above 2 for a significantly wider range of noise levels nl. These observations show the robustness and stability of the network-based test statistics in the presence of noise in identifying nonlinearity when it is dominant and noise is relatively weak. The results are more pronounced in the case of network-based test statistics suggesting the rejection of the null hypothesis for relatively higher levels of measurement noise with larger values for the significance of difference compared to the time series-based statistics. This observation shows that the structural properties of induced networks can identify inherent nonlinearity more faithfully, notwithstanding the presence of relatively higher levels of spurious noise in the data.

The structural properties of the network constructed may vary with the choice of the critical length ϵ_c and to evaluate how this might affect the values of S, we computed S for a range of values of ϵ less than 20.6 and found similar performance. The values of S are greater than 2 for a wide range of values of $\epsilon_c \leq 20.6$, showing that computations of S for the network parameters are robust with respect to minor variations in the value of the critical length ϵ, and would not affect the test results based on the values of S.

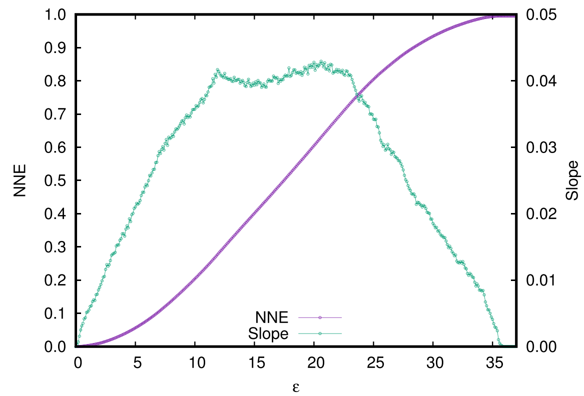

Figure 1. Rössler system: Plot of the normalised number of edges (NNE) vs. threshold value (ϵ) and its slope. The critical value $\epsilon_c = 20.6$ corresponding to the maximum slope is selected to construct the network.

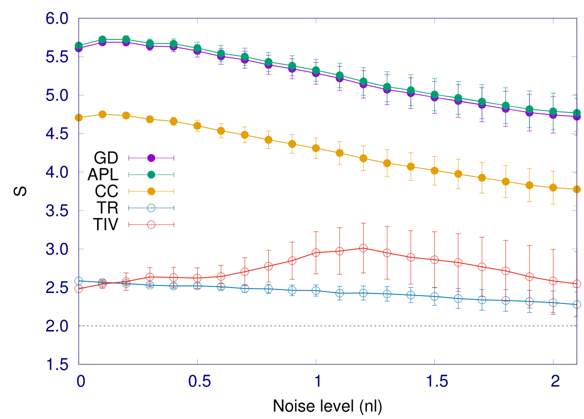

Figure 2. Rössler system: Plot of the significance of difference S based on network and time series base test statistics vs. noise level nl.

Next, we repeated similar computations for the Lorenz system. The Lorenz system is a nonlinear deterministic system with governing equations [1]

$$\begin{aligned} \dot{x} &= \sigma(y - x) \\ \dot{y} &= -y + x(r - z) \\ \dot{z} &= xy - bz. \end{aligned} \tag{14}$$

The parameters are set at values $\sigma = 10$, $b = 8/3$ and $r = 40$ for which the system is chaotic. We considered the first component of the solution vectors as the chaotic time series for our analysis. Plots of the values of S for different noise level values and the different test statistics for the Lorenz system are given in figure 3. This plot provides further evidence for the superiority of the

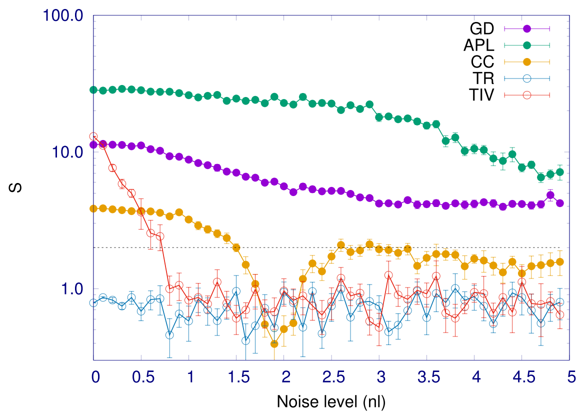

Figure 3. Lorenz system: Plot of S vs. noise level nl for different test statistics, with $\epsilon = 10.0$, $m = 7$, $\tau = 1$ and time series length 1024.

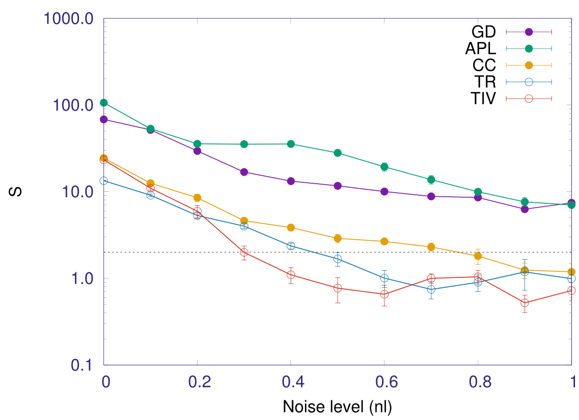

Figure 4. Logistic map: Plot of S vs. noise level nl for different statistics, for the parameter values $\epsilon = 0.9$, $m = 3$, $\tau = 1$ and time series length 512.

test statistics based on the structure of the induced networks in distinguishing nonlinear processes from linear Gaussian processes.

We also carried out a similar analysis for the time series generated from two discrete systems, viz., the logistic and Henon maps. The logistic map, given by

$$x_{n+1} = \mu x_n (1 - x_n) \tag{15}$$

generates a time series x_n which is chaotic for most values of $\mu > 3.57$. The two-dimensional Henon map defined by [37]

$$x_{n+1} = 1 - a x_n^2 + y_n$$
$$y_{n+1} = b x_n \tag{16}$$

is chaotic for the values $a = 1.4$ and $b = 0.3$. It generates the sequence (x_n, y_n) and the first coordinate values x_n constitute the time series for our analysis.

Figure 4 summarises the results for the surrogates of the noisy logistic time series at various noise levels for

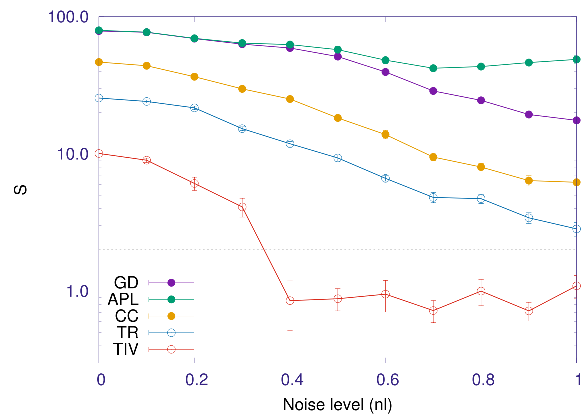

Figure 5. Henon map: Plot of S vs. noise level nl for different test statistics, with $\epsilon = 0.9$, $m = 5$, $\tau = 1$ and time series length 1024.

different test statistics. Figure 5 is for the Henon map time series. As in the case of continuous systems, we see that the test statistics APL, CC and GD are more reliable and powerful in identifying the underlying nonlinearity even in the presence of higher levels of noise.

To understand the effect of using the new statistics on actual experimental data, we selected a wind speed time series from a location that was earlier shown to have deterministic and chaotic underlying dynamics [38–42]. The data comprise measurements of wind speed taken at 10-min intervals of 1-week duration from a location given by latitude: 34.98420°N and longitude: −104.03971°W available from the National Renewable Energy Laboratory (http://www.nrel.gov), USA. Traditional methods of generating surrogate data, such as the amplitude-adjusted Fourier transform, preserve linear stochastic variations and amplitude distribution of the given time series. As these methods are devised to generate stationary surrogates, the existing non-stationarity of the given time series could be misinterpreted as dynamic nonlinearity [43,44]. Hence, the time series was detrended and rescaled to the range [0, 1] before carrying out further analysis [39,45]. The results obtained after treatment with the noise followed by surrogate analysis as earlier are plotted in figure 6. It is clearly seen that, as in other systems, the statistics APL, CC and GD perform much better than TIV and TR in detecting nonlinearity at all levels of noise considered.

Following essentially the same procedure, we have also compared the performance of the statistics TIV, TR, APL, CC and GD in terms of the statistical power, defined as the probability of rejecting the null hypothesis when it is in fact false. As shown in figure 7, at 95% confidence level, CC, APL and GD maintain power close to 1 for the widest range of noise levels. In comparison, TIV and TR are seen to be equally good at

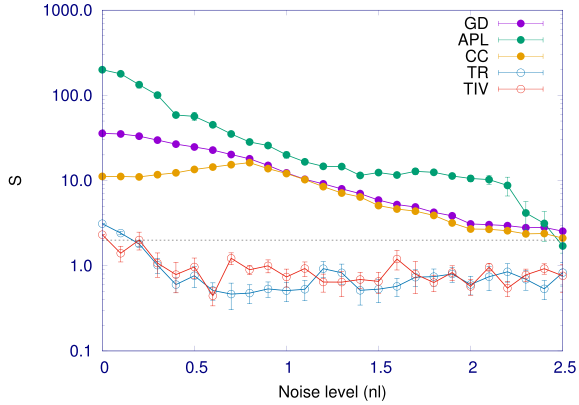

Figure 6. Wind speed data: Significance vs. noise level for different test statistics.

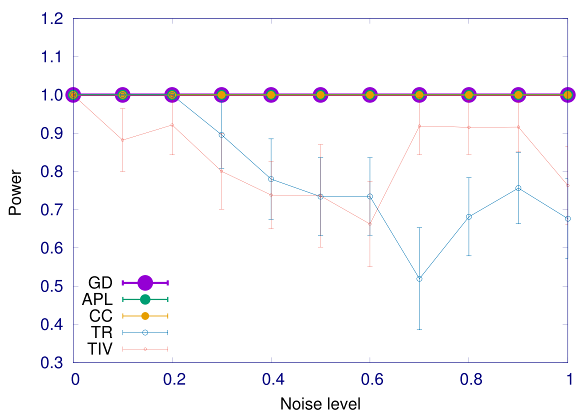

Figure 7. Wind speed data: Power of test vs. noise level for different test statistics.

lower noise levels but become less powerful as the noise level becomes six or more. Similar trends of power versus noise level were obtained for other systems as well. These results complement the earlier observations and suggest that the tests become more powerful when APL, CC or GD is used as a statistic for wider ranges of noise levels in this case also.

The edges mainly decide the structure of any network. In this work, the time series segments with delay τ represent the nodes, and the Euclidean distance determines the formation of the edge between nodes. Since the noise process added to the signal has zero mean, the distance between two segments more or less remains unaffected by the presence of noise as the effect of noise averages out. Thus, the addition of moderate levels of noise does not significantly affect the structure of the induced network, which could explain the better performance of network metrics as test statistics.

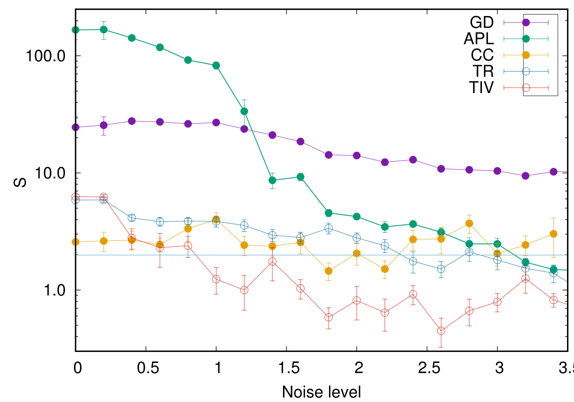

Figure 8. Rössler: Plot of S vs. dynamical noise level nl for different test statistics, with $\epsilon_c = 1.34$, $m = 7$, $\tau = 1$ and time series length 2500 sampled at $\delta t = 0.4$. The dynamical noise was at every $\delta t = 0.1$.

4.2 *Dynamical noise*

As a next step, we tested the efficiency of the networks-based test statistics against the dynamical noise. We first experimented with the Rössler system. To incorporate the effect of dynamical noise, we added a noise term $nl \times N(0, 1)$ on the right side of eqs (12) and then integrated the system using a differential equation solver. In order to avoid the solution blowing up, we used a reflecting boundary fixed by the attractor of the system without dynamical noise. The time series of length 2500 data points comprising the first component of the output of integration, sampled at time steps of 0.4 units, was used for the analysis. We computed the values of S for network and time series-based test statistics for increasing noise level nl. The result is plotted in figure 8. It can be observed that network parameters perform significantly better than time series-based parameters as test statistics. Similar plots for the logistic map and Lorenz system are given in figures 9–11. While TR performs well for Rössler system, even in the presence of dynamical noise, it is seen to be less efficient in the case of Lorenz x-series. This fact has been reported earlier [46, p. 117] and is attributed to the presence of large outliers in the time series data. However, the results indicate that the network-based statistics do not suffer from such shortcomings and provide another motivation for using network metrics as test statistics for surrogate data test.

4.3 *Effect of sampling interval*

The time interval at which the data points of the time series is sampled out from a system solution can also have an effect on the significance level S. For a chaotic solution, data points sampled at large interval may not capture the determinism because of the exponential

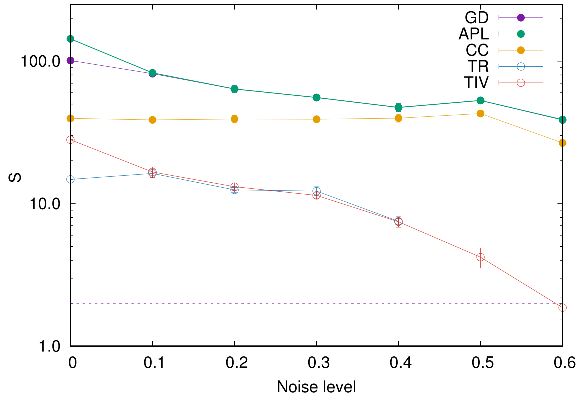

Figure 9. Logistic map: Plot of S vs. dynamical noise level nl for different test statistics, with $\epsilon_c = 0.9$, $m = 3$, $\tau = 1$ and time series length 512.

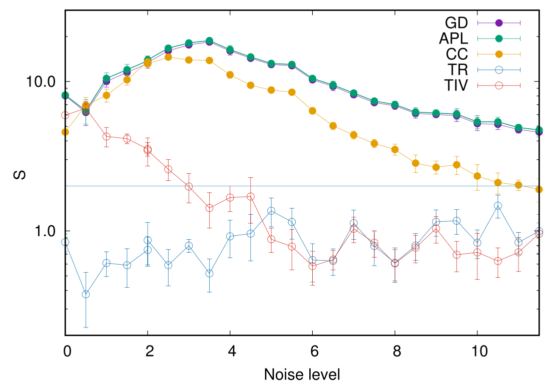

Figure 10. Lorenz system: Plot of S vs. dynamical noise level nl for different test statistics, with $\epsilon_c = 34.6$, $m = 7$, $\tau = 1$ and time series length 1024.

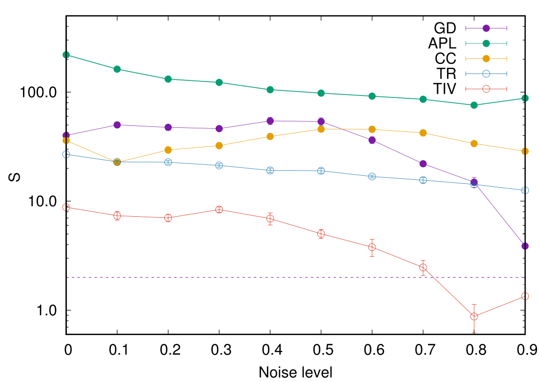

Figure 11. Henon map: Plot of S vs. time step or sample interval for different test statistics, with $\epsilon_c = 0.9$, $m = 5$, $\tau = 1$ and time series length 512.

divergence. In order to assess this impact, we computed the values S for the time series of Henon map generated

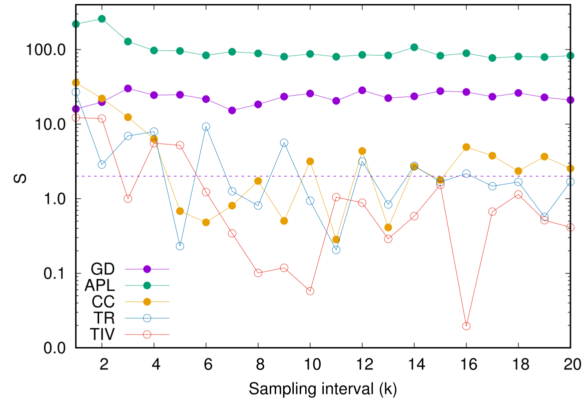

Figure 12. Henon map: Plot of S vs. the sampling interval of time series for different test statistics, at noise level $nl = 0$, with $\epsilon_c = 0.9$, $m = 5$, $\tau = 1$ and time series length 1024.

with the same parameters as in figure 11 but for different sampling intervals keeping the number of points at 1024. The values of S are plotted in figure 12. It is evident from this figure that the network-based statistics yield superior performance compared to the traditional measures at all sampling intervals. Other systems considered in this work also show similar results.

5. Conclusions

This paper has introduced some new test statistics for conducting surrogate data tests to differentiate a nonlinear time series from a linear stochastic time series and demonstrated their use in numerically simulated and actual experimental data. The method is based on the construction of a network from a time series and using easily computable parameters of the network as test statistics for the surrogate test. These parameters, namely the average path length, the graph density and the clustering coefficient of the network, are compared with commonly used statistics such as time invariance and time reversibility and are tested for their ability to correctly identify an underlying determinism in data even in the presence of a possible measurement noise affecting the data or a systematic source of dynamical noise. Numerical simulations conducted on the time series generated from the Rössler system, Lorenz system, logistic map, Henon map and actual experimental data of wind speed show that the new statistics are more reliable and robust in the presence of noise in identifying nonlinearity and determinism in the data. Also, the network parameters do not seem to underperform in the presence of outliers in the time series, while some of the time series-based statistics tend to worsen its efficiency, as noted in the case of the Lorenz series.

Acknowledgements

The authors wish to thank the National Renewable Energy Laboratory (http://www.nrel.gov), USA, for making their data available and the campus computing facility of the University of Kerala under the DST-PURSE programme for providing computational facilities. They are also grateful to Dr Drisya Alex Thumba for computational assistance. The first author (MCM) would like to acknowledge the financial assistance through the e-grants scheme of the Government of Kerala.

References

[1] E N Lorenz, *J. Atmos. Sci*. **20**, 130 (1963)
[2] N Packard, J Crutchfield, J Farmer and R Shaw, *Phys. Rev. Lett.* **45**, 712 (1980)
[3] F Takens, in *Lecture notes in mathematics* (Springer, Berlin Heidelberg, 1981) pp. 366–381
[4] T Sauer, J A Yorke and M Casdagli, *J. Stat. Phys*. **65**, 579 (1991)
[5] A Osborne and A Provenzale, *Physica D* **35**, 357 (1989)
[6] J Theiler, *Phys. Lett. A* **155**, 480 (1991)
[7] J Theiler, S Eubank, A Longtin, B Galdrikian and J Doyne Farmer, *Physica D* **58**, 77 (1992)
[8] G Lancaster, D Iatsenko, A Pidde, V Ticcinelli and A Stefanovska, *Phys. Rep.* **748**, 1 (2018)
[9] S Chen and P Shang, *Physica A* **537**, 122716 (2020)
[10] Y Hirata, M Shiro and J M Amigó, *Entropy-switz.* **21**, 713 (2019)
[11] Y Sun, P Shang, J He and M Xu, *Fluct. Noise Lett*. **17**, 1850035 (2018)
[12] M Chavez and B Cazelles, *Sci. Rep*. **9**, 1 (2019)
[13] Y Hirata and M Shiro, *Phys. Rev. E* **100**, 022203 (2019)
[14] J Zhang and M Small, *Phys. Rev. Lett.* **96**, 238701 (2006)
[15] Y Yang and H Yang, *Physica A* **387**, 1381 (2008)
[16] Z Gao and N Jin, *Phys. Rev. E* **79**, 066303 (2009)
[17] L Lacasa, B Luque, F Ballesteros, J Luque and J C Nuño, *Proc. Natl. Acad. Sci.* **105**, 4972 (2008)
[18] N Marwan, J F Donges, Y Zou, R V Donner and J Kurths, *Phys. Lett. A* **373**, 4246 (2009)
[19] C W Kulp, J M Chobot, H R Freitas and G D Sprechini, *Chaos* **26**, 073114 (2016)
[20] L Lacasa and R Toral, *Phys. Rev. E* **82**, 036120 (2010)
[21] W J Bosl, H Tager-Flusberg and C A Nelson, *Sci. Rep.* **8**, 6828 (2018)
[22] Y Zou, R V Donner, N Marwan, J F Donges and J Kurths, *Phys. Rep.* **787**, 1 (2019)
[23] B A Gonçalves, L Carpi, O A Rosso, M G Ravetti and A Atman, *Physica A* **525**, 606 (2019)
[24] G Iacobello, S Scarsoglio and L Ridolfi, *Phys. Lett. A* **382**, 1 (2018)
[25] W Fang, X Gao, S Huang, M Jiang and S Liu, *Open Phys.* **16**, 346 (2018)
[26] K Liu, T Weng, C Gu and H Yang, *Physica A* **538**, 122952 (2020)
[27] R Jacob, K Harikrishnan, R Misra and G Ambika, arXiv:1508.02724 (2015)
[28] X Luo, T Nakamura and M Small, *Phys. Rev. E* **71**, 026230 (2005)
[29] T Schreiber and A Schmitz, *Physica D* **142**, 346 (2000)
[30] G Pavlos, M Athanasiu, D Kugiumtzis, N Hatzigeorgiu, A Rigas and E Sarris, *Nonlin. Processes Geophys*. **6**, 51 (1999)
[31] T Maiwald, E Mammen, S Nandi and J Timmer, in *Understanding complex systems* (Springer, Berlin Heidelberg, 2008) pp. 41–74
[32] X Xu, J Zhang and M Small, *Proc. Natl. Acad. Sci*. **105**, 19601 (2008)
[33] T Sauer and J A Yorke, *Int. J. Bifurc. Chaos* **03**, 737 (1993)
[34] Z-K Gao and N-D Jin, *Nonlinear Anal. Real World Appl*. **13**, 947 (2012)
[35] O Rössler, *Phys. Lett. A* **57**, 397 (1976)
[36] T Schreiber and A Schmitz, *Phys. Rev. Lett*. **77**, 635 (1996)
[37] M Hénon, in *The theory of chaotic attractors* (Springer, New York, 1976) pp. 94–102
[38] R Sreelekshmi, K Asokan and K Satheesh Kumar, *Ann. Geophys.* **30**, 1503 (2012)
[39] G Drisya, D C Kiplangat, K Asokan and K Satheesh Kumar, *Ann. Geophys.* **32**, 1415 (2014)
[40] G Drisya, K Asokan and K S Kumar, *Renew. Energ*. **119**, 540 (2018)
[41] G Drisya, P Valsaraj, K Asokan and K Kumar, *Int. J. Comput. Sci. Eng*. **9**, 362 (2017)
[42] P Valsaraj, D A Thumba, K Asokan and K S Kumar, *Appl. Energ*. **260**, 114270 (2020)
[43] M Mannattil, H Gupta and S Chakraborty, *ApJ* **833**, 208 (2016)
[44] J Lucio, R Valdés and L Rodríguez, *Phys. Rev. E* **85**, 056202 (2012)
[45] K S Kumar, C Kumar, B George, G Renuka and C Venugopal, *J. Geophys. Res.* **109**, https://doi.org/10.1029/2002ja009768 (2004)
[46] S J Guastello and R A Gregson, *Nonlinear dynamical systems analysis for the behavioral sciences using real data* (CRC Press, 2016)