Nonlinearity and Nonstationarity: The Use of Surrogate Data in Interpreting Fluctuations

Daniel T. Kaplan
Centre for Nonlinear Dynamics and Dept. of Physiology
McGill University, Montréal, Québec, Canada

"Surrogate data" is the basis for a technique for testing a time series for nonlinear dynamics and process nonstationarities. The theory behind surrogate data is briefly described, along with an algorithm for generating it. Examples are given of its use in detecting nonlinearity in heart rate signals, detecting nonstationarity, and estimating the sampling distribution of complicated statistics of heart rate variability.

1. Introduction

In studying heart rate and blood pressure variability, one is faced with a dilemma. The mechanisms regulating the cardiovascular system are known theoretically to contain many nonlinearities, and the system as a whole is known to be nonstationary. Yet the dominant techniques for analyzing time series — for example, spectral analysis — are based in the assumptions of linear dynamics and stationarity. The ultimate resolution to this dilemma may come in the form of improved time series analysis techniques that can cope optimally with nonlinearities and nonstationarities. This report has a much more modest goal: to describe newly developed techniques using "Surrogate Data" to detect nonlinearities and nonstationarities in data. Detecting nonlinearities — or failing to detect them — allows us to know when linear analysis techniques are and are not capturing all of the information in the time series. Detecting nonstationarities allows us to make informed decisions about issues such as whether collecting longer runs of data provides better estimates of physiological variables, or about which are the best analysis techniques that can allow us to track changes in the physiological system without unnecessarily increasing the variance of the estimates.

1.1 Nonlinearity

There are many sources of nonlinearity in cardiovascular regulation. One of the earliest to be given a mathematical formulation is the interaction between sympathetic and parasympathetic innervation of the SA node, as described by (Rosenblueth and Simeone, 1934). Other commonplace physiological mechanisms also correspond directly to mathematical nonlinearities: *adaptation* of the baroreceptors to changes in blood pressure; *saturation* of receptors; *changes in gain* of feedback systems with changing baseline levels of blood pressure; *reduction in cardiac output* at high heart rates. *Delays* are ubiquitous in physiological systems. Coupled with *high gains* in feedback loops, delays cause instability. Such instabilities are always associated with nonlinearities; in a linear system instability leads to a physically impossible blow-up to infinity. Guyton's textbook introduction (Guyton, 1991) to cardiovascular control is practically a catalogue of nonlinear mechanisms.

It is now widely appreciated that nonlinear systems can show irregular oscillations without any random input. This is called "chaos." Linear systems without an input are uninteresting; they decay

to a steady state. To produce irregular oscillations in a linear system requires an input that is itself irregular, such as white noise. The existence of nonlinear mechanisms in physiological control systems raises the possibility that irregular-looking physiological variability may not be entirely the response to a changing external environment or changing demands on the system (i.e., a changing "input") but may be in part a "self-oscillation." For instance, it has been suggested (Goldberger et al., 1990) that enabling physiological systems to adapt rapidly to external changes requires that they be unstable; containing the instability within physiologically acceptable bounds requires nonlinearity and leads to self-oscillations.

Self-oscillations are not obscure and esoteric things; they are as familiar as a pendulum clock. In the cardiovascular system, the well known 10-second oscillation likely reflects a self-oscillation arising from delays and high-gain in the baroreceptor feedback loop. Other self-oscillations are seen in vasomotion, renal vascular control (Holstein-Rathlou and Marsh, 1995), Cheyne-Stokes respiration, and of course the rhythmic firing of the SA-node itself.

Despite the existence of nonlinear mechanisms and observed nonlinear phenomenon, it might be that linear time series analysis techniques are adequate. An oscillation with a regular period will appear as a peak (or series of peaks) in a power spectrum, regardless of whether the oscillation is caused by a linear or nonlinear mechanism. The interaction of many cardiovascular feedback loops, together with random environmental influences, may eliminate detectable traces of nonlinearity from the time series, even though the time series is caused by nonlinear mechanisms. So, our theoretical knowledge about the existence of nonlinear mechanisms is not in itself sufficient to justify using nonlinear analysis techniques on time series — we need tests to know when nonlinearities appear in the time series itself, and when they can be safely disregarded. Surrogate data provides us with such tests.

1.2 Nonstationarity

Nonstationarity is also an important problem in studying signals from the cardiovascular system. Obviously, it would be unreasonable to expect unchanging signals from a subject whose level of physical activity is changing, or whose posture is changing, or whose mental or emotional state is changing. The solution to the problem of nonstationarity might seem to be to hold the subjects at a constant level of physical activity, of posture, and a constant mental or emotional state. In addition to the experimental difficulties in doing this, there are more subtle and less controllable difficulties. Many physiological control systems adapt over a period of minutes, hours or days. Having a subject rest quietly for one-half hour before the beginning of measurements does not eliminate this process of adaptation. In a nonlinear system, it should not be expected that rapidly adapting systems will rapidly reach a steady state; the more slowly adapting systems may force a re-adaption by the rapid systems. In addition, normal humans have a circadian rhythm which can be expected to cause changes over the course of hours in core body temperature and mental state. Hormones such as growth hormone are released in a pulsatile manner.

A challenging problem in studying long-term heart rate and blood pressure concerns $1/f$ noise. The heart rate and blood pressure are known to display variability that is approximately described by a power law over time scales of tens of minutes to days. This power-law variability means, for example, that there is no well defined mean or variance; the measured mean or variance will depend on the time scale used to make the measurement. Taking more and more data will not improve the quality of the measurement.

One approach to dealing with nonstationarity is to transform the time series so that it becomes statistically stationary. For example, taking the first difference of an 24-hour RR-interval time series will eliminate the statistical nonstationarity of the mean and variance. The motivation behind such

statistical tricks is that the underlying process may be unchanging, but our measurement may not reflect this. A simple example is a random walk: the walker takes a sequence of steps in random directions. Over time, even though the mechanism is unchanging, the walker's position will drift. Taking a first difference transforms the measurement to one that reflects the statistically unchanging random steps. Unfortunately, techniques such as first-differencing cannot magically transform any nonstationary signal into a stationary one.

In physiology, nonstationarity may stem from changing mechanisms, and we may desire to track these changes. This can be done by dividing up a long data record into short segments (or using more sophisticated techniques such as time-spectral analysis or wavelets). The basic question is, "How long a segment is appropriate?" If the segment is too short, the estimates made from the data may be poor. If the segment is too long, the estimates may be unacceptably influenced by nonstationarity. Surrogate data can help us to decide whether data are nonstationary, and what length segment is appropriate for analysis.

In the following, the original time series, the one we want to test for nonlinearity or nonstationarity, will be termed the *test data*.

2. Surrogate data

Surrogate data was initially developed to deal with the question: "Can we demonstrate the existence of deterministic chaos in a time series?" The answer, it turns out, is no, at least unless some additional assumptions are made.

The problem with using data to demonstrate chaos is that any measured signal is consistent with a linear system driven by some input, and linear systems can never be chaotic. For example, Figure 1a shows test data from the famous chaotic Lorenz equation. Given a linear model for the dynamics underlying the time series, we can deduce what the input to the linear dynamics must be — these deduced inputs are termed "residuals." Figure 1b shows the residuals when the posited linear model is the "optimal" 2nd-order linear autoregressive model. Note that the variance of the residuals is 1.6, much lower than the variance of the test data (44.2). This might suggest that the model is explaining more than 96% of the energy in the test data. Use of higher-order models (Fig. 1c) suggests that linear models might "explain" more than 99.5% of the energy in the test data.

In this case, the linear models actually explain very little of the signal. We know this because the equations that produced the signal are nonlinear, and involve unstable fixed points in an essential way. But if we did not know the equations of the system, how could we tell just from the data that a linear model is not explaining the data?

If the linear model itself were doing a good job of explaining the data, then the output of the model should not depend critically on the particular inputs used. So, we could ask either of two questions:

1. Do the residuals themselves contain some structure that we do not expect them to? For example, we might expect that the residuals from a successful linear model should be random white noise with a gaussian distribution.

2. If we run the model with random inputs (e.g., gaussian white noise), do we produce an output that is similar in important ways to the measured signal?

Which of these two questions is the proper one to ask? Traditionally, the approach has been to examine things such as the whiteness of the residuals using tests for which the sampling distribution is known. For instance, Jenkins and Watts (Jenkins and Watts, 1968) give a test for the whiteness of a time series based on the sample integrated periodogram, and give confidence intervals for white noise;

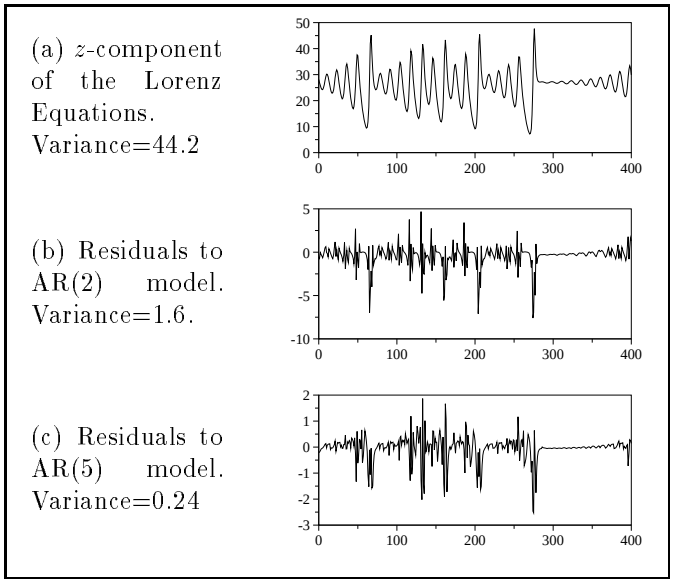

Figure 1: Output of the Lorenz equations, and the deduced inputs to a linear system producing this output.

to test a given time series for whiteness, one need merely see if the sample integrated periodogram falls outside of the confidence intervals. Such tests rely on the ability to construct such confidence intervals, and it is not always known how to do this.

Suppose that the characteristic of the inputs that we want to probe involves a statistic for which we do not know the confidence intervals. For example, we might want to test the residuals from the linear model to find out if they are themselves chaotic. It is well known that many chaotic systems produce time series that pass conventional tests for whiteness, and so we need to use a statistic that tests for characteristic other than whiteness. In section we will describe briefly three nonlinear statistics that are useful in testing for chaos; the confidence intervals are unknown for these statistics for random white noise.

In such a situation, we can use *bootstrapping* to estimate the confidence intervals for the statistic. We simply generate many realizations of white noise by, for example, randomly shuffling the order of points in the residual time series. Then we calculate the value of our favorite statistic on each of these time series. The distribution of these values indicates the confidence intervals for our statistic.

The bootstrapping technique can also be used to address question (2). First, we pick a statistic that quantifies some important aspect of the test data. Then, we generate many realizations of white noise, and pass these through our linear model to produce realizations of the output of the linear process. The statistic can then be calculated for each of these linear outputs, and we can directly compare the value of the statistic for the test data to the distribution of the statistic for the output realizations.

A problem with the technique just described is that we need to specify the linear model to use. This involves picking a model order, which is an arbitrary decision, and is called a "nuisance parameter." (See (Theiler and Pritchard, 1995).) If the selected model order is too small, then we unfairly limit the ability of the linear system to model the data. If the selected model order is too high, then we will be overfitting the data and will introduce spurious correlations in the bootstrapped outputs. We can avoid the problem of the nuisance parameter by using the surrogate data technique.

Surrogate data is constructed to have the same power spectrum and, hence, the same autocorre-

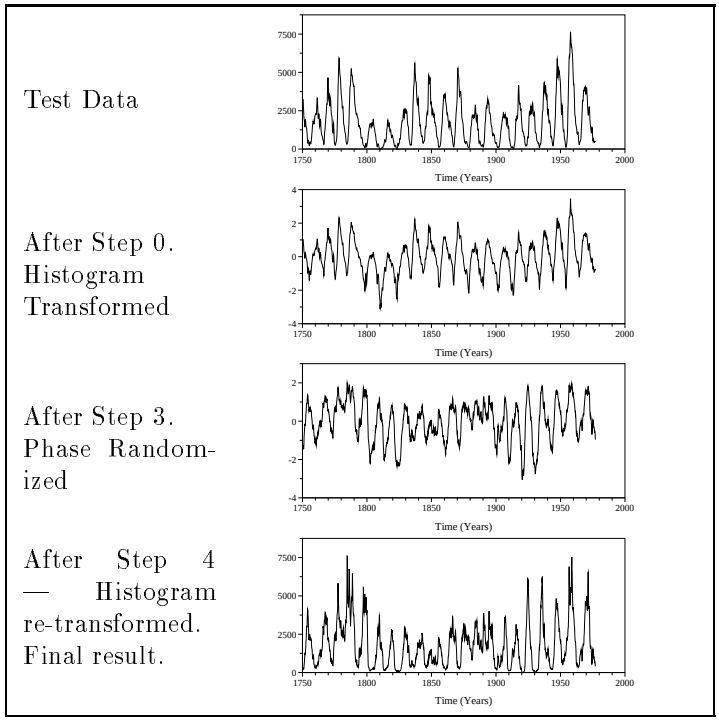

Figure 2: Steps in the generation of surrogate data. Steps 1 and 2, which produce a result in the Fourier domain rather than the time domain, are not shown here. (The test data is the Wolper sunspot numbers.)

lation function as the test data. Since optimal linear models are estimated from the autocorrelation function, surrogate data is described by the same optimal linear model *for any model order*. But we do not need to pick a model order to generate the surrogate data. Instead, a three-step procedure is used:

(1) Calculate the discrete fourier transform of the test data. This consists of an amplitude and a phase at each frequency.

(2) Randomize the phase at each frequency to be uniformly distributed in $[0, 2\pi]$. (Preserve the phase asymmetry around frequency 0.)

(3) Take the inverse fourier transform.

Since the amplitude of the fourier transform has not been changed, the power spectrum of the surrogate data constructed in this manner is identical to the power spectrum of the test data. (The generation of surrogate data is described in detail in (Theiler et al., 1992) and (Kaplan and Glass, 1995)

This procedure has one major flaw: the surrogate data has a normal (gaussian) histogram, whereas test data often has a distribution that is significantly non-normal. Non-normal distributions can be created by dynamically trivial and uninteresting mechanisms such as nonlinear measurement functions (e.g., we measure the square of a normally distributed variable rather than the variable itself). In order to avoid problems with comparing non-normal test data to normal surrogate data, we add two steps to the sequence for generating surrogate data, as proposed by (Theiler et al., 1992).

(0) Before taking the DFT of the test data, apply a static nonlinear transformation to give the test data's distribution a normal shape. This is effectively done through sorting algorithms. (See (Theiler et al., 1992) and (Kaplan and Glass, 1995).)

(4) After phase randomization and taking the inverse DFT, transform the gaussian surrogate data back to the distribution of the test data. Again, this can be done using sorting algorithms.

The sequence of steps is illustrated in Figure 2. We can generate different realizations of surrogate data from one test data set by using different random numbers in step (2) of the surrogate data generation process. Some examples are shown in Figure 3.

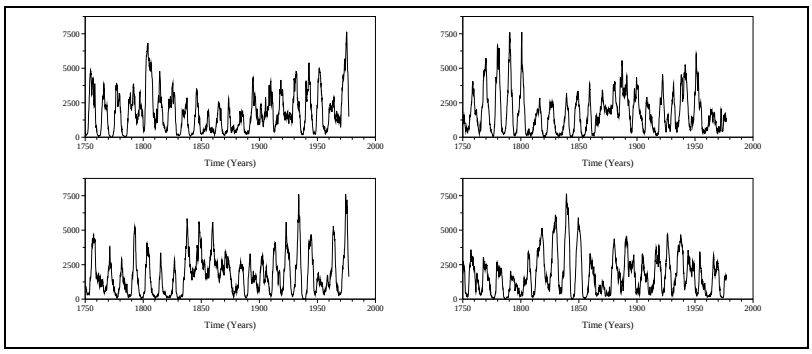

Figure 3: Four examples of surrogate data generated from the Wolper sunspot test data

2.1 Properties of Surrogate Data

Surrogate data has two important properties that concern us here:

1. Surrogate data has no dynamical nonlinearities. By construction, surrogate data is equivalent to passing gaussian white noise through a linear filter. This is practically the definition of a linear stochastic system.

2. The process that generates surrogate data is stationary. If we think of the linear filter that is coloring the gaussian white noise inputs as the dynamical process, then that process is not changing during the duration of the surrogate data.

Note that process stationarity is not the same thing as stationarity of the data itself. We can, for example, generate surrogate data from $1/f$ noise — the original data will be nonstationary (the variance is unbounded as the signal length increases) and the surrogate data will also be nonstationary. However, the process that generates the surrogate data does not change between, say, the first half of the data set and the second half.

There are a few situations that can cause surrogate data to be a misleadingly poor match to the test data, even though the power spectrum and histogram are the same in the surrogate and test data. One situation that is particularly common in studying heart rate data concerns the presence of sharp spikes in the RR-interval series that might arise from ectopy or false R-wave detection. The surrogate data technique turns these spikes into white noise. Even though the power spectrum of the surrogate and test data are identical, the spikes are highly localized in time in the test data but spread throughout the surrogate data. Technically, such spikes are themselves evidence for nonlinearity or nonstationarity. But insofar as one is not interested in the dynamics underlying the spikes, one should be careful to remove them from the data set; even a single spike can have a statistically significant effect on surrogate data, as shown in Figure 4.

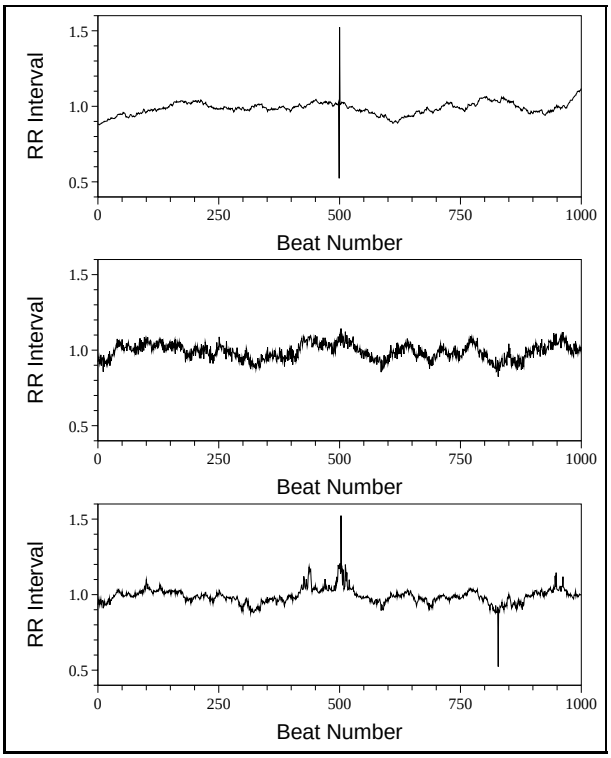

Figure 4: (top) Simulated RR-interval data with a single false QRS detection. (middle) Surrogate data generated without amplitude adjustment. (bottom) Surrogate data with amplitude adjustment.

3. Testing for dynamical nonlinearities in data

We can exploit the fact that surrogate data contains no dynamical nonlinearities in order to test for dynamical nonlinearities in data. The first step is to choose some discriminating statistic that summarizes some information of interest that pertains to the dynamics of the system. (Three such statistics are described briefly in the next section of this report.) We calculate this statistic on our test data, and also on a large number of surrogate time series. Our notation is as follows: $\mathcal{S}$ refers to the value of the discriminating statistic on our test data; $\tilde{\mathcal{S}}$ refers to the value of the discriminating statistic on one realization of surrogate data.

The surrogate time series allow us to estimate how the statistic is distributed in the absence of dynamical nonlinearities. If the value of the statistic for the test data falls into the distribution for surrogate data, then our statistic does not allow us to distinguish the test data from the surrogates and we have no evidence for dynamical nonlinearities. On the other hand, if the test data has a value for the statistic that is outside of the distribution for the surrogate data, then we can conclude that the test data is somehow different from the surrogates. Now, *if we assume* that the dynamics underlying the test data are stationary, we then have evidence that there is some dynamical nonlinearity in the test data. Chaos is one example of a dynamical nonlinearity, but it is not the only possibility. Other possibilities include nonlinear stochastic dynamics, such as for example amplitude or frequency modulation with a stochastic modulating input. And, of course, nonstationarity is just an assumption at this point.

There are two practical ways to assess whether the statistic for the test data falls into the distribution for the surrogate data: parametric and nonparametric. In the nonparametric approach, we

generate very many realizations of surrogate data. We count how many of these realizations have a value $\tilde{\mathcal{S}}$ that is more extreme than the test data's $\mathcal{S}$. Doubling this number gives us an estimate of the probability that a realization of the null hypothesis can generate a value as extreme as $\mathcal{S}$. (Or, we might prefer to do a one-sided test, in which we don't need to double the number.) For example, suppose we find that $\mathcal{S} = 15.4$ and that 12 out of 100 realizations of surrogate data have $\tilde{\mathcal{S}} \geq 15.4$. Then we can estimate that the probability that a realization of the null hypothesis would have $\tilde{\mathcal{S}}$ more extreme than 15.4 is $2 \times 12\%$. In this case, we are not justified in rejecting the null hypothesis.

The parametric approach involves assuming that $\tilde{\mathcal{S}}$ has a normal (gaussian) distribution. We calculate the mean and standard deviation of $\tilde{\mathcal{S}}$. Call these $< \tilde{\mathcal{S}} >$ and $\sigma(\tilde{\mathcal{S}})$ respectively. The "standard score" for $\mathcal{S}$ is then

$$Z = \left| \frac{\mathcal{S} - < \tilde{\mathcal{S}} >}{\sigma(\tilde{\mathcal{S}})} \right|$$

and the probability of finding a Z as extreme as that measured from the test data, given that the null hypothesis is true, is specified by the complementary error function $erfc(Z/\sqrt{2})$. (For example, the familiar 4.5% when $Z = 2$, and 0.27% when $Z = 3$, and so on.) It should be cautioned that the distribution of $\tilde{\mathcal{S}}$ is commonly non-gaussian, and so the interpretation of the standard score ought to be taken with a grain of salt. Further dietary caution is indicated because statistics calculated from surrogate data may be biased by, for example, the amplitude adjusting steps (Kaplan and Theiler, in preparation); prudence suggests avoiding drawing dramatic conclusions based on a standard score of 2 or 3. Despite the problems with the parametric approach, it has the advantage of giving a rough indication of the significance of differences between test and surrogate data with only a few realizations of surrogate data. In addition, very large standard scores (e.g., $Z > 10$) can suggest that differences between test and surrogate data may be very strong; the nonparametric test cannot practically point to such possibilities.

3.1 Three statistics for detecting nonlinearity

There is a potentially infinite number of discriminating statistics that can be used to test for nonlinearity using surrogate data. The only absolute requirement is that the discriminating statistic not be directly derivable from the histogram of the time series, or from the autocorrelation function (or, equivalently, the power spectrum). This rules out the use of statistics such as the mean, variance, skewness, kurtosis, and percentile ranges, which can be calculated from the histogram, and statistics such as energy in a particular frequency band, locations of polls and zeros, or energy of the derivative function, which can be calculated from the autocorrelation function. I will term such statistics *linear* statistics, since they do not allow test data to be discriminated from surrogate data. (However, linear statistics are useful in using surrogate data to test for stationarity. See below.)

Beyond the injunction not to use linear statistics, not much is known about which statistics are good for detecting nonlinearity in data. In the following, I will use three statistics that have all been motivated by chaos, but there is no reason in principle not to use statistics motivated by other concerns:

Nonlinear predictability constructs an *ad hoc* nonlinear model of future values of the time series as a function of past values. The mean of the logarithm of the absolute value of the difference between predicted and measured values provides the discriminating statistic. The most widely cited implementation of nonlinear predictability is (Sugihara and May, 1990); note that the version used here differs in that a piecewise-constant model is used as opposed to the piecewise linear in (Sugihara and May, 1990) and the use of the mean log absolute predition error as opposed to the correlation coefficient used in (Sugihara and May, 1990).

δ-ϵ continuity attempts to estimate the noise level in the data without constructing a model (Kaplan, 1994), but assuming that the underlying dynamics are described by a continuous function.

Correlation dimension is the most widely known nonlinear dynamics statistic. Here, the "rule of five" method of estimation is used (Theiler and Lookman, 1993).

3.2 Results on heart rate data

To illustrate the surrogate data technique, I provide two anecdotes of the analysis of heart rate data. Figure 5 (top) shows a clear difference between the surrogate data and the test data using both δ-ϵ and nonlinear predictability. The difference is also seen in the return maps: the test data looks somewhat like a flower bouquet and has a narrow handle on the bottom right, while the surrogate data has a more rotund shape. In fact, differences between the test and surrogate data are apparent by eye in the time series themselves — the test data has four smooth dips that are absent in the surrogate time series. Using techniques beyond the scope of this report, we can trace back the differences between the discrimating statistics for the surrogate and test data to the dips. From the time series itself, we can only speculate about the origin of these dips; perhaps the dips are a manifestation of deterministic chaos, but this seems an unnecessarily dramatic conclusion. In this case, the surrogate data analysis simply points to the dips as the origin of the statistically detected nonlinearity — the statistics say nothing about the origin or physiological significance of the dips.

Figure 5 (bottom) again shows differences between the test and surrogate data. In this case, the difference is partly due to the nonstationarity of the data; this will be analyzed in Section 4.

3.3 Testing for nonlinearity of statistics

Researchers attempting to find useful ways of characterizing heart rate variability have found their motivation in a wide variety of areas. One recently proposed heart rate variability statistic is Detrended Fluctuation Analysis (Peng et al., 1995), which has been motivated (in my view) by the study of random fractals, and the power-law correlations seen in physics in critical phenomena and self-organized criticality. Peng et al. were able to use DFA to discriminate between normal subjects and subjects in congestive heart failure — this suggests that DFA is quantifying physiologically meaningful patterns in heart rate variability. Given the origin of the interest in power-law correlations in heart rate, there is an interesting possibility that the HR patterns quantified by DFA arise from nonlinear dynamics. But whatever the dynamical origin of the patterns, it is useful to know whether the patterns themselves are indicative of nonlinearity.

In their study, Peng et al. took 24-hour Holter records of RR intervals, and divided them into non-overlapping segments of 8192 RR intervals. For each segment, the DFA analysis provides two numbers: a power-law slope over a time scale of 16 to 64 beats, and a power-law slope over a shorter time scale of 4-16 beats. Figure 6a shows the DFA statistic calculated from many segments of 24-hour data graciously provided by Peng. Normal subjects are shown as circled numbers, and congestive heart failure subjects are shown as uncircled numbers. The number itself identifies the individual subject. Figure 6b shows the same analysis applied to surrogate data generated from each individual segment. The overall pattern is remarkably the same for the test data as for the surrogate data. This suggests that the HR patterns that are different between normals and congestive heart failure subjects are *linear* patterns. Of course, the possibility exists that the reason why there are linear differences between normals and CHF subjects is based in some physiological nonlinear dynamics, but the data themselves do not provide direct evidence for this.

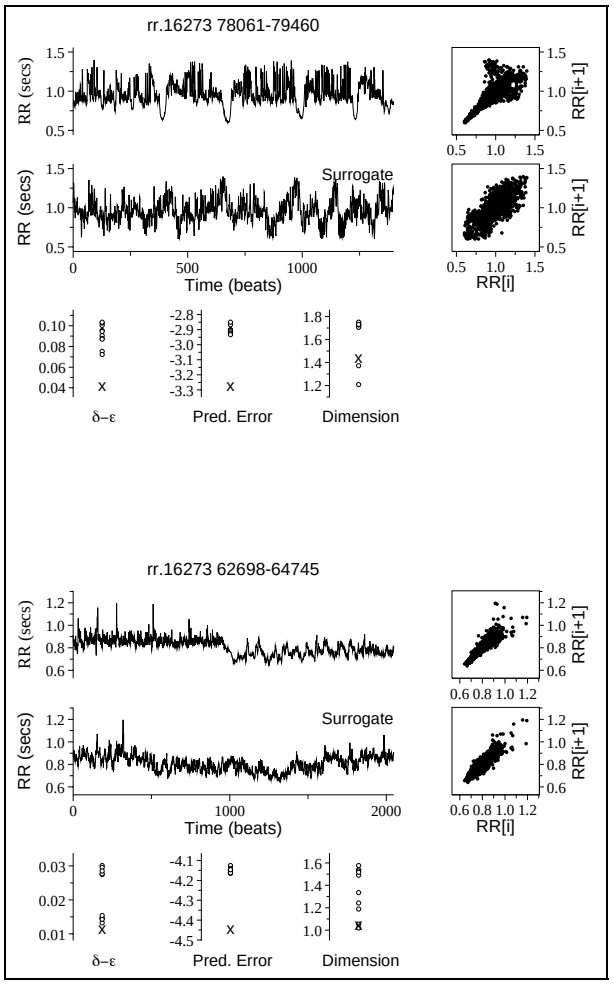

Figure 5: Testing for nonlinear dynamics in two RR-interval time series. In each case, a short time series of RR intervals is shown, as well as one realization of surrogate data. Beside the time series is shown the "return plot" of the data, that is, a scatter plot of RR$[i+1]$ versus RR$[i]$. Underneath the time series are shown the values of the three test statistics for the test data ($\times$) and 10 realizations of the surrogate data ($\circ$).

4. Process Nonstationarities

In testing for nonlinearity, we used nonlinear statistics and assumed that the data were generated by a stationary process. If we use linear statistics, we are unable to discriminate differences between test data and surrogate data; all the surrogate data sets will have the same value for the linear statistics as the test data. However, if we divide the surrogate and test data into segments, we will in general find that linear statistics differ from segment to segment. The surrogate data reflects a process that is both linear and stationary, and the segment-by-segment variability in the statistics reflects the sampling distribution of the statistics.

If the test data are consistent with the null hypothesis, that is if the test data come from stationary linear dynamics, then the segment-by-segment values of the statistic for the test data should fall in the sampling distribution given by the segment-by-segment statistics of the surrogate data. If there are differences, then they can be ascribed either to nonlinearity in the data, or to nonstationarity. By using linear statistics, we limit the sensitivity of the analysis to nonlinearity. So, if differences between

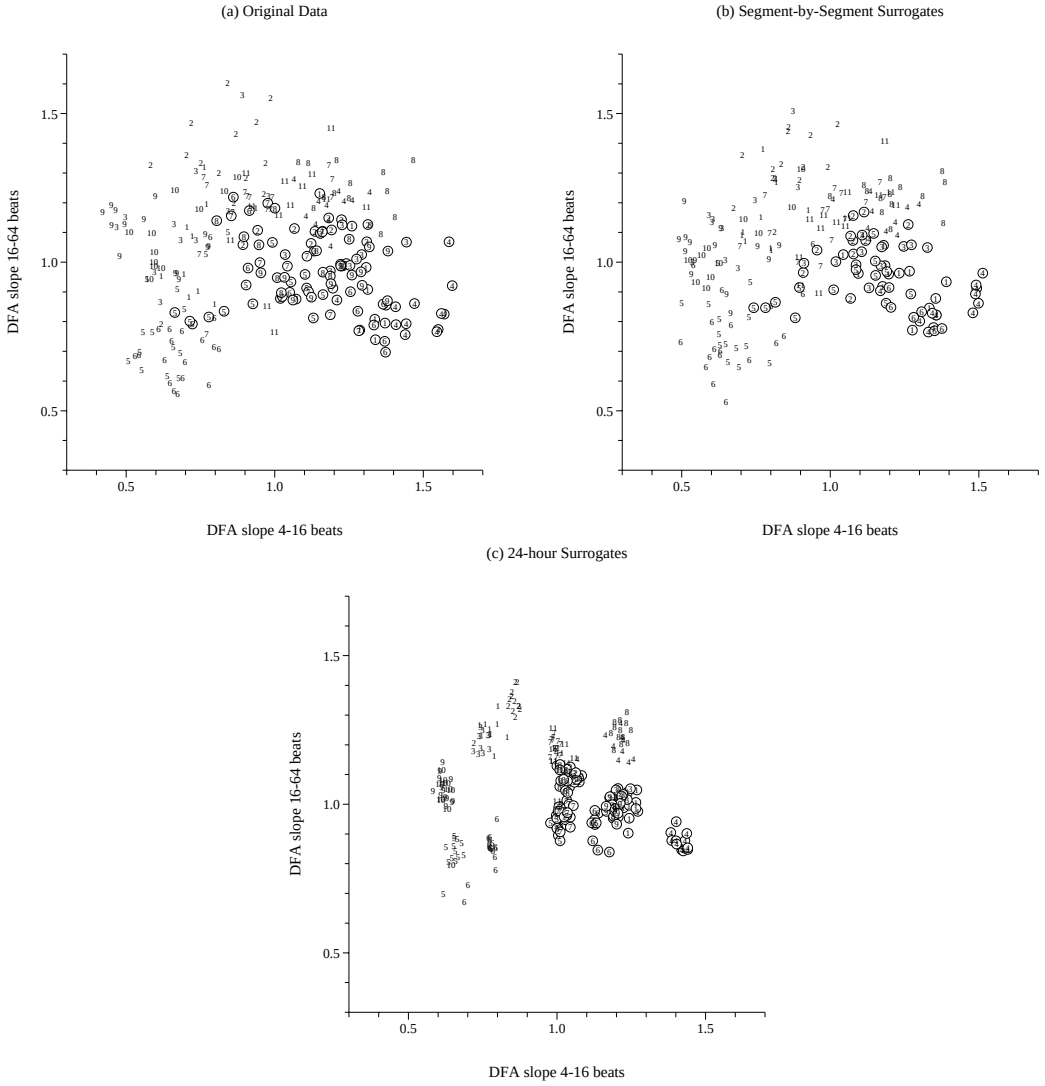

Figure 6: DFA Analysis of Holter data. The DFA analysis produces two numbers that characterize each time series. There are plotted here for segments of length 8192 beats out of a 24-hour Holter record. The code number for each subject is plotted for each segment; circled numbers indicate normal subjects, while uncircled numbers are congestive heart failure subjects. (a) Original data. (b) Surrogate data constructed separately for each segment. (c) Surrogate data constructed from the entire 24-hour record, and then divided into segments.

the test and surrogate data are found using linear statistics, we have evidence for nonstationarity of the (linear) process that generated the data. (Alternatively if we wanted to use a nonlinear statistic to test for nonstationarity, we could test for nonlinearity on the segment-by-segment level by calculating surrogates for each segment individually, and using any nonlinear statistic we choose.)

An informal example of a test for nonstationity is given in Figure 6c. Here, the surrogates have been generated for the whole 24-hour Holter record (as opposed to the segment-by-segment surrogates shown in Figure 6b). For each subject, there is a narrower distribution of the DFA statistics for the segments of the 24-hour surrogates than for the segments of the original data. Insofar as we have already tested the DFA statistic for nonlinearity, and found it to be effectively a linear statistic on this

data, we have evidence that the 24-hour data is nonstationary in terms of DFA. This is not altogether surprising, since human activity levels etc. changes considerable over a 24-hour period.

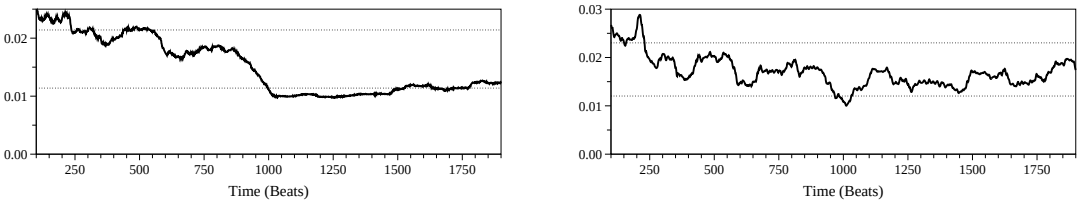

Figure 7: (left) Amplitude versus time of the respiratory-band component of the RR-interval time series shown in the lower panel of Figure 5. The two dotted lines ihere ndicate the 95% confidence intervals for surrogate data. (right) Same but for the low frequency component. A pass band of 0.15 to 0.30 beats/cycle was used for the respiratory band, and 0.05 to 0.10 beats/cycle was used for the low frequency band. The output of these FIR filters (length 65 beats) was squared and low-pass filtered with a boxcar of length 151 beats.

Another example of testing for nonstationarity is provided by the RR-interval data shown in Figure 5 (lower panel). The first and last halves of the time series seem by eye to be qualitatively different. Suppose that we want to find out whether the sympatho-vagal activity, as measured by RMS amplitudes in respiratory and low frequency bands, changes during the course of the recording. We can easily measure the RMS amplitudes using a band-pass filter, squaring the output, low-pass filtering, and taking the square root of the result. This is shown in Figure 7.

In order to establish whether the differences during the course of the time series are statistically significant, we generate surrogate data, and apply the same processing to them. The dotted lines in the figure show the 95% confidence intervals for the surrogate data (i.e., the 2.5 and 97.5 percentiles). Insofar as the test data spends more than 5% of the time outside of the confidence intervals, and given that the statistic used is linear, we have evidence for nonstationarity in the signal. Clearly this is the case for the respiratory-frequency-band amplitude, but the low-frequency-band signal shows no such pattern.

The surrogate data technique allows us to explore different parameters to use in detecting nonstationarity. For instance, if we wanted to check whether there are short bursts of respiratory frequency activity, we might want to use quite short band-pass and low-pass filters. The surrogate data technique would allow us to see how likely any detected bursts are to arise purely from chance.

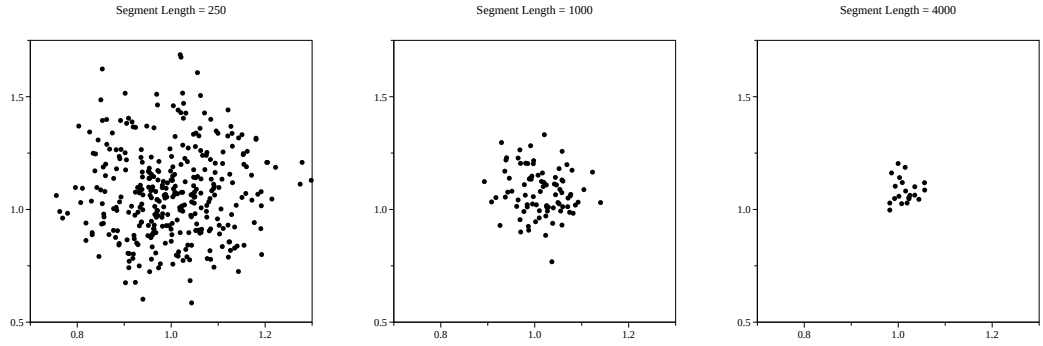

Figure 8: Sampling distribution of the DFA statistic for segment sizes of 250 beats, 1000 beats, and 4000 beats, based on surrogates generated from one subject.

Having identified a nonstationarity in the data, we now need to deal with it. A standard technique

in heart rate variability is to divide the data into segments. But how large a segment to choose? The classical trade-off is between using short segment size in order to track the nonstationarity, and using long segment size in order to minimize the extent of the sampling distribution of the statistic.

In the case of DFA, we might for example try to make the segment size as short as possible consistent with being able to discriminate between normals and congestive heart failure subjects. In order to do this, it is necessary to estimate the sampling distribution for each segment size. Surrogate data allows us to do this in an easy way.

Start by generating surrogate data for the entire 24-hour record. Then divide this surrogate data into segments of the desired length, and calculate the DFA statistic on each segment. Multiple realizations of surrogate data can be used to increase the number of segments. Results for one subject in the DFA study are shown in Figure 8. As expected, the distribution of the DFA statistic becomes narrower as the segment length is increased. If our task is to distinguish normals from CHFs, then we want the size of the DFA distribution to be smaller than the clouds for normals and CHF shown in Figure 6a. This appears to be satisfied for a segment length of 4000 beats.

5. Summary

Surrogate data offers a straightforward method to detect nonlinearity and nonstationarity in data. Detecting nonlinearity and nonstationarity is important even if one is not interested in chaos and nonlinear dynamics, since it points out situations where linear analysis (such as power spectrum analysis) may not be picking out all of the information in the time series.

Surrogate data allows the sampling distribution of even complicated statistics to be readily estimated so that it is no longer necessary to use rules of thumbe to pick parameters such as the length of data segment to be used, and so that formal tests of statistical significance can be used.

Given the computational ease of generating surrogate data, and the ever-present conflict in heart rate and blood pressure between the nonlinear/nonstationary physiological mechanisms and the linear/stationary analysis methods in use, surrogate data is a valuable method that should be added to the toolkit of those analyzing variability in heart rate and blood pressure.

Acknowledgements

This work was supported from grants from the Natural Sciences and Engineering Research Council of Canada, the Fonds de la recherche en santé du Québec, and the Fonds pour la formation de chercheurs et l'aide à la recherche. I would like to thank C-K Peng and A Goldberger for their generosity in supplying the RR-interval data used for examples in Figures 5, 6, 7, and 8.

References

[1] AL Goldberger, DR Rigney, BJ West (1990) "Chaos and Fractals in Human Physiology" *Scientific American* bf 262(2):42-49

[2] AC Guyton (1991) *Textbook of medical physiology* 8th edition, W.B. Saunders Co, Philadelphia ISBN 0-7216-3087-1

[3] N-H Holstein-Rathlou and DJ Marsh, (1995) "Renal blood flow regulation and arterial pressure fluctuations: a case study in nonlinear dynamics" *Physiol. Review* **74**:637-681

[4] G.M. Jenkins and D.G. Watts, (1968) *Spectral Analysis and its Applications* Holden-Day, San Francisco, pp. 243-238

[5] DT Kaplan (1994) "Exceptional events as evidence for determinism" *Physica D* **73**:38-48

[6] D.T. Kaplan and L. Glass (1995) *Understanding Nonlinear Dynamics* Springer-Verlag, New York.

[7] C-K Peng, S Havlin HE Stanley and AL Goldberger (1995) "Detrended Fluctuation Analysis", *Chaos* **5**(1):82-87

[8] A Rosenblueth and FA Simeone (1934), "The interrelations of vagal and accelerator effects on the cardiac rate" *Amer. J. Physiology* **110**:42-55

[9] J Theiler and T Lookman (1993) "Statistical error in a chord estimator of the correlation dimension: the 'rule of five' *Int. J. Bifurc. Chaos* **3**:765-71

[10] J Theiler, S Eubank, A Longtin, B Galdrikian, and JD Farmer (1992) "Testing for nonlinearity in time series: the method of surrogate data" *Physica D* **58**: 77-94

[11] J Theiler and D Pritchard (1995) "Constrained-realization Monte-Carlo method for hypothesis testing" preprint

[12] G Sugihara and RM May (1990), "Nonlinear forecasting as away of distinguishing chaos from measurement error in time series" *Nature* **344**:734-41