CHAPTER 2

METHODS FOR DATA CHARACTERIZATION

2.1 **Introduction**

A considerable range of techniques have been developed to characterize nonlinear time series data. Each with distinct theoretical background and significance, contribute complementary information regarding signal characteristics. Characteristic patterns of variations over time represent a defining feature of complex systems. Despite the intrinsic dynamic, interdependent and nonlinear relationships of their parts, complex systems exhibit robust systemic stability. Nonlinear analysis methods provide a novel tool with which to evaluate the overall properties of a complex system.

The data analysis basically measures variations, including time domain analysis, frequency distribution, spectral power), frequency contribution (spectral analysis), scale invariant (fractal) behavior (detrended fluctuation and power law analysis) and regularity (approximate and multiscale entropy), sensitivity to initial conditions (Lyapunov exponent) and detection of principal trend (Principal component analysis).

Nonlinear systems are complex systems; specifically, they are systems that are spatially and temporally complex, built from a dynamic web of interconnected feedback loops marked by interdependence, and redundancy. Complex systems have properties that cannot wholly be understood by understanding the parts of the system [1]. The properties of the system are distinct from the properties of the parts, and they depend on the integrity of the whole; the systemic properties vanish when the system breaks apart, whereas the properties of a linear system can be described in terms of its any part.

2.2 Methods of Characterization

2.2.1 Sampling

The analysis of patterns of change over time or variability is performed on a series of data collected continuously or semi-continuously over time. In order to reconstruct the underlying signal without error, the Nyquist Theorem is applied, which states that the sampling frequency must be at least twice the highest frequency of the signal being sampled.

2.2.2 Stationarity

Stationarity defines a limitation in techniques designed to characterize a time series. It requires that statistical properties such as mean and standard deviation of the signal remain the same throughout the period of recording, regardless of measurement epoch. Stationarity does not preclude variability of data, but it provides boundaries for variations such that variability does not change with time or duration of measurement. If this requirement is not met, as is the case with most if not all financial time series when socioeconomic conditions change, then the impact of trends with change on the mean of the data set must be considered in the interpretation of the analysis of variations.

2.2.3 Artifact

Analysis should be performed on data that are free from artifact, with a minimal noise to signal ratio. Noise is measurement error, or imprecision secondary to measurement technology. Several techniques, such as a Poincaré Plot of the difference between consecutive data points, have been developed to facilitate automated identification and removal of artifact [2-4]. Different techniques are more or less sensitive to artifact, which again should be taken into account.

2.2.4 Time domain analysis

Time series analysis represents the simplest means of evaluating variations, identifying its measures of variation over time such as standard deviation and range. In addition, a visual representation of data collected as a time series may be obtained by plotting a frequency distribution, plotting the number of occurrences of values in selected ranges of values or bins.

Considered the simplest means of measuring variations, time domain analysis involves performing a statistical analysis of data expressed as a sequence in time. For example, the standard deviation (SD) has been used as a measure; greater variation yields higher standard deviation. Standard deviation of the longer intervals within the entire period of recording is a measure of longer term variation because the averaging process removes tick to tick variations. As a measure of global variation, standard deviation is altered by the duration of measurement; longer series will have greater Standard deviation.

Various permutations of measurement of standard deviation, in an effort to isolate short-term, high frequency fluctuations from longer term variation, are possible. In order to characterize a frequency distribution, it may be fitted to a normal distribution, or rather a log-normal distribution – one in which the log of the variable in question is normally distributed. The skewness or degree of symmetry may be calculated, with positive and negative values indicating distributions with a right-sided tail and a left-sided tail, respectively. Kurtosis may also be calculated to identify the peakedness of the distribution; positive kurtosis (leptokurtic) indicates a sharp peak with long tails, and negative kurtosis (platykurtic) indicates a flatter distribution.

Time domain analysis involves the statistical evaluation of data expressed as a series in time. Time series of parameters derived from financial

systems are known to follow log-normal frequency distributions, and deviations from the log-normal distribution have been proposed to offer a means with which to characterize fluctuations [5].

Statistical measures of variations are easy to compute and provide valuable prognostic information about the data. Frequency distributions also offer an accurate, visual representation of the data, although the analysis may be sensitive to the arbitrary number of bins chosen to represent the data. Time domain measures are susceptible to bias secondary to nonstationary signals.

Therefore, additional, more sophisticated methods of analysis are necessary to characterize and differentiate financial signals

2.2.5 Frequency domain analysis

In this approach the data collected as a series in time, as with any time series, is considered a sum of sinusoidal oscillations with distinct frequencies. Conversion from a time domain to frequency domain analysis is made possible with various mathematical transformations like Fourier transformation, wavelet, and Hilbert transformation etc. The amplitude of each sine and cosine wave determines its contribution to the time series; frequency domain analysis displays the contributions of each sine wave as a function of its frequency. Facilitated by computerized data harvest and computation, the result of converting data from time series to frequency analysis is termed spectral analysis because it provides an evaluation of the power (amplitude) of the contributing frequencies to the underlying signal.

The power spectrum is a different representation of the same time series data, and the transformation may be made from time to frequency and back again. This provides an analysis of the relative contributions of different frequencies to the overall variation in a particular data series. Interpretation of the analysis must factor in the assumptions inherent to this calculation, namely

stationarity and periodicity. The square of the contribution of each frequency is the power of that frequency to the total spectrum, and the total power of spectral analysis (area under the curve of the power spectrum) is equal to the variance described above (they are different representations of the same measure). The fast Fourier transform or analysis represents a nonparametric calculation because it provides an evaluation of the contribution of all frequencies, not discrete or preselected frequencies. The fast Fourier transform is a discrete Fourier transform that reduces the number of computations. The result of the Fourier transform is a complex number (a number multiplied by the square root of -1) for each frequency, the square of which is considered the spectral power of that frequency. The whole process is called spectral analysis, because it provides an evaluation of the spectral power (amplitude) of the contributing frequencies of an underlying signal.

The power spectral density function, or power spectrum provides a characteristic representation of the contributing frequencies to an underlying signal. By identifying and measuring the area of distinct peaks on the power spectrum, it is possible to derive quantitative connotation to facilitate comparison between individual and different data sets.

In order to derive a valid and meaningful analysis using a fast Fourier transform and frequency domain analysis, the assumptions of stationarity and periodicity must be fulfilled. The signal must be periodic, namely it is a signal that is comprised of oscillations repeating in time, with positive and negative alterations [6]. In the interpretation of experimental data, periodic behavior may or may not exist when evaluating alterations in spectral power in response to intervention. The assumption of stationarity may also be violated with prolonged signal recording.

Thus, the performance and interpretation of spectral analysis must incorporate these limitations. Recommendations based upon the stationarity assumption include the following: short-term and long-term spectral analyses must be distinguished; long-term spectral analyses are felt to represent averages of the alterations present in shorter term recordings and may hide information; traditional statistical tests should be used to test for stationarity when performing spectral analysis.

2.2.6 Time spectrum analysis

Another means to address the stationarity assumption inherent in the Fourier transform is to evaluate the power spectral density function for short periods of time when stationarity is assumed to be present, and subsequently follow the evolution of the power spectrum over time. This combined time varying spectral analysis allows the continuous evaluation of change in variability over time. One can use sequential spectral approach, Wavelet analysis, the Wigner-Ville technique or Walsh transforms, all of which provide an analysis of frequency alteration over time.

2.2.7 Power law

Power law behavior describes the dynamics of widely disparate phenomena, from earthquakes, solar flares and stock market fluctuations to avalanches. These dynamics arise from the system itself; the theory of self-organized criticality has been suggested to represent a universal organizing principle in finance [65].

Power law behavior may be described by the equation:

$$F(x) = \alpha x^{\beta} \tag{2.1}$$

Where α and β are constants. Taking the logarithm of both sides, a straight line (graph log f [x] versus log x) with slope β and intercept log α is revealed:

$$\text{Log } f(x) = \log (\alpha x^{\beta}) = \log \alpha + \log x^{\beta}$$

$$= \log \alpha + \beta \log x \qquad (2.2)$$

Thus, power law behavior is scale invariant; if a variable x is replaced by Ax', where A is a constant then the fundamental power law relationship remains unaltered. If dynamics follow a power law, a log–log representation of the power spectrum (log power versus log frequency) reveals a straight line, always within a defined range consistent with the size and duration of the system. The straight line is fitted using linear regression, and the slope β and intercept can readily be obtained. When β = -1, the dynamics are described as 1/f noise.

Power laws describe dynamics that have a similar pattern at different scales, namely they are 'scale invariant'. Detrended fluctuation analysis (DFA) is a technique that characterizes the pattern of variation across multiple scales of measurement. A power law describes a time series with many small variations, and fewer and fewer larger variations; and the pattern of variation is statistically similar regardless of the size of the variation. Magnifying or shrinking the scale of the signal reveals the same relationship that defines the dynamics of the signal, analogous to the self-similarity seen in a multitude of spatial structures found in finance [1, 65]. This scale invariant self-similar nature is a property of fractals, which are geometric structures pioneered and investigated by Benoit Mandelbrot [7]. Fractals represent structures that have no fixed length; their length increases with increased precision (magnification) of measurement, a property that confers a non-integer dimension to all fractals.

As with frequency domain analysis (discussed above), the first step in the evaluation of the power law is the calculation of the power spectrum. This calculation, based on the fast Fourier transform (defined above), yields the frequency components of a series in time. By plotting a log–log representation of the power spectrum (log power versus log frequency), a straight line is obtained with a slope of approximately -1. As the frequency increases, the size of the variation drops by the same factor, and this patterns exists across many scales of frequency and variation, within a range consistent with system size and signal duration. Mathematically, power law behavior is scale invariant.

Given that power law analysis is performed by plotting the log of spectral power versus the log of frequency using data derived from spectral analysis, the relationship between the two methods of characterizing variations can be compared.

Although derived using the same data, the two methods assess different characteristics of signals. Spectral analysis measures the relative importance or contribution of specific frequencies to the underlying signal, whereas power law analysis attempts to determine the nature of correlations across the frequency spectrum. These analyses may have distinct and complementary significance.

Because determining power law behavior requires spectral analysis, namely the determination of the frequency components of the underlying signal, the technique becomes problematic when applied to nonstationary signals. This limitation makes it difficult to draw conclusions regarding the mechanisms that underlie the pattern. In addition, because power law behavior measures the correlation between a large range of frequencies, it requires prolonged recording to achieve statistical validity. Nonetheless, as with the

time and frequency domain analysis, valid distinctions based on power law analysis can be demonstrated.

Specifically addressing the problem of nonstationarity, there is a problem in differentiating variations in a series of data that arise from the dynamics of a complex nonlinear system [88,89]. Both lead to a nonstationary variations but nonetheless represent distinct phenomena. The subsequent technique was developed to address this issue.

2.2.8 Detrended fluctuation analysis

Introduced by Peng and coworkers [8-10], DFA was developed specifically to distinguish between intrinsic fluctuations generated by complex systems and those caused by external or environmental stimuli acting on the system. Variations that arise because of extrinsic stimuli are presumed to cause a local effect, whereas variations due to the intrinsic dynamics of the system are presumed to exhibit long-range correlation. DFA attempts to quantify the presence or absence of long range scale-invariant (fractal) correlation. The first step in the technique to calculate DFA is to map a time signal, such as a series of stock price, to an integrated series. The integrated series is calculated by the sum of the random variable.

DFA is a second measure of scale invariant behavior because it evaluates trends of all sizes, trends that exhibit fractal properties (similar patterns of variation across multiple time scales). A component of the DFA calculation involves the subtraction of local trends (more likely related to external stimuli) in order to address the correlations that are caused by nonstationarity, and to help quantify the character of long-range fractal correlation representing the intrinsic nature of the system.

2.2.9 Multifractal analysis

DFA is a monofractal technique, in that the assumption is that the same scaling property is present throughout the entire signal. Multifractal techniques provide multiple, possibly infinite exponents, such that the analysis produces a spectrum rather than a discrete value. For example, wavelet analysis is a multifractal analysis technique similar to DFA.

2.2.10 Entropy analysis

Entropy is a measure of disorder or randomness, as embodied in the Second Law of Thermodynamics, namely the entropy of a system tends toward a maximum. In other words, states tend to evolve from ordered statistically unlikely configurations to configurations that are less ordered and statistically more probable and the spontaneous reverse occurrence is statistically improbable to the point of impossibility.

2.2.10.1 Approximate entropy

Related to time series analysis, approximate entropy (ApEn) provides a measure of the degree of irregularity or randomness within a series of data. It is closely related to Kolmogorov entropy, which is a measure of the rate of generation of new information [12]. ApEn was pioneered by Pincus [13] as a measure of system complexity; smaller values indicate greater regularity, and greater values convey more disorder, randomness and system complexity.

In order to measure the degree of regularity of a series of data (of length N), the data series is evaluated for patterns that recur. This is performed by evaluating data sequences of length m, and determining the likelihood that other runs in the data set of the same length m are similar (within a specified tolerance r); thus two parameters, m and r, must be fixed to calculate ApEn. Once the frequency of occurrence of repetitive runs is calculated, a measure of

their prevalence (negative average natural logarithm of the conditional probability) is found. ApEn then measures the difference between the logarithmic frequencies of similar runs of length m and runs with the length m+1. Small values of ApEn indicate regularity, given that the prevalence of repetitive patterns of length m and m+1 do not differ significantly and their difference is small.

2.2.10.2 Sample and Multiscale entropy

An inherent bias within the ApEn calculation exists because the algorithm counts similar sequences to a given sequence of length m, including counting the sequence itself (to avoid the natural logarithm of 0 within the calculations). As a result, ApEn can be sensitive to the size of the data set, giving inappropriately low values when the total number of data points is low; this, and a lack of consistency in differentiating signals when m and r are altered, have led to the development of a new family of statistics named sample entropy (SampEn), in which self-matches are excluded in the analysis. SampEn has the advantage of being less dependent on the length of the data series in question. Finally, because both ApEn and SampEn are noted to evaluate differences between sequences of length m and m+1, they evaluate regularity on one scale only, the shortest one, and ignore other scales.

2.3 References

[1]. Gallagher R, Appenzeller T: Beyond reductionism. Science, 284:79(1999).

[2]. Cunningham S, Symon AG, McIntosh N: The practical management of artifact in computerised physiological data. Int J Clin Monit Comput, 11:211-216.(1994)

[3]. Sapoznikov D, Luria MH, Mahler Y, Gotsman MS: Computer processing ofartifact and arrhythmias in heart rate variability analysis. Comput Methods Programs Biomed, 39:75-84.(1992).

[4]. Berntson GG, Quigley KS, Jang JF, Boysen ST: An approach to artifact identification: application to heart period data. Psychophysiology, 27:586-598(1990).

[5]. Zhang CL, Popp FA: Log-normal distribution of physiological parameters and the coherence of biological systems. Med Hypotheses, 43:11-16(1994).

[6]. Mansier P, Clairambault J, Charlotte N, Medigue C, Vermeiren C, LePape G, Carre F, Gounaropoulou A, Swynghedauw B: Linear and non-linear analyses of heart rate variability: a minireview. Cardiovasc Res, 31:371-379(1996).

[7]. Mandelbrot B: The Fractal Geometry of Nature (French edition published 1975.) New York: Freeman; (1983).

[8]. Peng CK, Havlin S, Stanley HE, Goldberger AL: Quantification of scaling exponents and crossover phenomena in nonstationary heartbeat time series. Chaos, 5:82-87(1995).

[9]. Peng CK, Buldyrev SV, Goldberger AL, Havlin S, Mantegna RN, Simons M, Stanley HE: Statistical properties of DNA sequences. Physica A, 221:180-192(1995).

[10]. Peng CK, Buldyrev SV, Havlin S, Simons M, Stanley HE, Goldberger AL: Mosaic organization of DNA nucleotides. Phys RevAvailable online http://ccforum.com/content/8/6/R367

[11]. Willson K, Francis DP: A direct analytical demonstration of the essential equivalence of detrended fluctuation analysis and spectral analysis of RR interval variability. Physiol Meas, 24:N1-N7(2003).

[12]. Richman JS, Moorman JR: Physiological time-series analysis using approximate entropy and sample entropy. Am J Physiol Heart Circ Physiol, 278:H2039-H2049(2000).

[13]. Pincus SM: Approximate entropy as a measure of system complexity. Proc Natl Acad Sci USA, 88:2297-2301(1991).

[14]. Pincus SM, Goldberger AL: Physiological time-series analysis: what does regularity quantify? Am J Physiol, 266:H1643-H1656(1994).

[15]. Pincus S, Singer BH: Randomness and degrees of irregularity. Proc Natl Acad Sci USA, 93:2083-2088(1996).

[16]. Pincus S: Approximate entropy (ApEn) as a complexity measure. Chaos, 5:110-117(1995).

[17]. Costa M, Goldberger AL, Peng CK: Multiscale entropy analysis of complex physiologic time series. Phys Rev Lett, 89:068102(2002).