Modelling of Nonlinear Dynamic systems: Using surrogate data methods

Tanja Conradie

Thesis presented in partial fulfilment of the requirements
for the degree of Master of Science
at the University of Stellenbosch.

Dr M. Gerber

December 2000

Declaration:

I, the undersigned, hereby declare that the work contained in this thesis is my own original work, and that I have not previously in its entirety, or in part, submitted it at any university for a degree.

Signature: Date:.

ABSTRACT

This study examined nonlinear modelling techniques as applied to dynamic systems, paying specific attention to the Method of Surrogate Data and its possibilities. Within the field of nonlinear modelling, we examined the following areas of study: attractor reconstruction, general model building techniques, cost functions, description length, and a specific modelling methodology. The Method of Surrogate Data was initially applied in a more conventional application, i.e. testing a time series for nonlinear, dynamic structure. Thereafter, it was used in a less conventional application; i.e. testing the residual vectors of a nonlinear model for membership of identically and independently distributed (i.i.d) noise.

The importance of the initial surrogate analysis of a time series (determining whether the apparent structure of the time series is due to nonlinear, possibly chaotic behaviour) was illustrated. This study confirmed that omitting this crucial step could lead to a flawed conclusion.

If evidence of nonlinear structure in the time series was identified, a radial basis model was constructed, using sophisticated software based on a specific modelling methodology. The model is an iterative algorithm using minimum description length as the stop criterion. The residual vectors of the models generated by the algorithm, were tested for membership of the dynamic class described as i.i.d noise. The results of this surrogate analysis illustrated that, as the model captures more of the underlying dynamics of the system (description length decreases), the residual vector resembles i.i.d noise. It also verified that the minimum description length criterion leads to models that capture the underlying dynamics of the time series, with the residual vector resembling i.i.d noise. In the case of the "worst" model (largest description length), the residual vector could be distinguished from i.i.d noise, confirming that it is not the "best" model. The residual vector of the "best" model (smallest description length), resembled i.i.d noise, confirming that the minimum description length criterion selects a model that captures the underlying dynamics of the time series.

These applications were illustrated through analysis and modelling of three time series: a time series generated by the Lorenz equations, a time series generated by electroencephalograhpic signal (EEG), and a series representing the percentage change in the daily closing price of the S&P500 index.

OPSOMMING

In hierdie studie ondersoek ons nie-lineêre modelleringstegnieke soos toegepas op dinamiese sisteme. Spesifieke aandag word geskenk aan die Metode van Surrogaat Data en die moontlikhede van hierdie metode. Binne die veld van nie-lineêre modellering het ons die volgende terreine ondersoek: attraktor rekonstruksie, algemene modelleringstegnieke, kostefunksies, beskrywingslengte, en 'n spesifieke modelleringsalgoritme. Die Metode and Surrogaat Data is eerstens vir 'n meer algemene toepassing gebruik wat die gekose tydsreeks vir aanduidings van nie-lineêre, dimanise struktuur toets. Tweedens, is dit vir 'n minder algemene toepassing gebruik wat die residuvektore van 'n nie-lineêre model toets vir lidmaatskap van identiese en onafhanlike verspreide geraas.

Die studie illustreer die noodsaaklikheid van die aanvanklike surrogaat analise van 'n tydsreeks, wat bepaal of die struktuur van die tydsreeks toegeskryf kan word aan nie-lineêre, dalk chaotiese gedrag. Ons bevesting dat die weglating van hierdie analise tot foutiewelike resultate kan lei.

Indien bewyse van nie-lineêre gedrag in die tydsreeks gevind is, is 'n model van radiale basisfunksies gebou, deur gebruik te maak van gesofistikeerde programmatuur gebaseer op 'n spesifieke modelleringsmetodologie. Dit is 'n iteratiewe algoritme wat minimum beskrywingslengte as die termineringsmaatstaf gebruik. Die model se residuvektore is getoets vir lidmaatskap van die dinamiese klas wat as identiese en onafhanlike verspreide geraas bekend staan. Die studie verifieer dat die minimum beskrywingslengte as termineringsmaatstaf wel aanleiding tot modelle wat die onderliggende dinamika van die tydsreeks vasvang, met die ooreenstemmende residuvektor wat nie onderskei kan word van indentiese en onafhanklike verspreide geraas nie. In die geval van die "swakste" model (grootse beskrywingslengte), het die surrogaat analise gefaal omrede die residuvektor van indentiese en onafhanklike verspreide geraas onderskei kon word. Die residuvektor van die "beste" model (kleinste beskrywingslengte), kon nie van indentiese en onafhanklike verspreide geraas onderskei word nie en bevestig ons aanname.

Hierdie toepassings is aan die hand van drie tydsreekse geïllustreer: 'n tydsreeks wat deur die Lorenz vergelykings gegenereer is, 'n tydsreeks wat 'n elektroënkefalogram voorstel en derdens, 'n tydsreeks wat die persentasie verandering van die S&P500 indeks se daaglikse sluitingsprys voorstel.

CONTENTS

ACKNOWLEDGEMENTS

I wish to thank my supervisor, Dr M Gerber firstly, for stimulating my interest in the field of nonlinear modelling. His contributions in terms of guidance and explanations, as well as new insights and applications, were invaluable. Lastly, I want to thank him for his patience in supervising me through this study.

I am indebted to the Mathematics Department of the University of Western Australia for providing me with the sophisticated software program that was used to construct models from the selected time series.

Finally, I want to thank my family and friends. Without their support this study would never have seen the light. I owe them much for their interest and encouragement. Their faith in this study, and me made me persevere even when I could not see the light.

LIST OF TABLES

LIST OF FIGURES

INTRODUCTION

The main purpose of this study was to gain exposure to the vast discipline of nonlinear, dynamic modelling, and to apply the knowledge and insights gained to selected time series. The expectation was to emphasise existing principles and concepts, and to supplement these with new insights and conclusions by means of the analyses of the applications. In order to break this daunting study into manageable tasks, certain guidelines relating to specific areas of study within the field of nonlinear, dynamic modelling were identified. These were:

- Attractor reconstruction,
- General model building techniques,
- Cost functions used in model building techniques,
- The Method of Surrogate Data,
- Correlation Dimension,
- Description Length, and
- A Specific Modelling Methodology.

In Chapter One, the fundamental principles of nonlinear, dynamic modelling are discussed. We look at attractor reconstruction [1] in Section 1.1, defining what is meant by an embedding, based on Takens' embedding theorem [35]. The choice of embedding dimension, as suggested by Takens [35] and Mañé [20], are examined and we investigate the method of False Nearest Neighbours [19] as an indication of the optimal embedding dimension for a specific time series. The choice of the value of the time lag is also discussed and we look at the method of Average Mutual Information [11] as a means of choosing the optimal time lag for the embedding of a specific time series. In Section 1.2, we explain general modelling building techniques, briefly examining polynomials [1, 18], neural networks [1, 18], genetic algorithms [1, 12] and radial basis functions [16, 18, 21]. Cost functions are discussed in Section 1.3, where we investigate the loglikelihood function [18], Akaike information criterion [3] and the Schwarz criterion [27]. We briefly mention another cost function known as description length [24] to be discussed in a later section.

The Method of Surrogate Data [36, 38, 39] is explained in Chapter Two, studying primarily the work of Theiler et al. [37, 38, 39, 40] and Takens [36]. General terminology pertaining to the Method of Surrogate Data is reviewed in Section 2.1. Section 2.2 defines linear surrogates [38], and Section 2.3 defines nonlinear surrogates [29, 31]. The Method of Surrogate Data is in essence statistical hypothesis testing (also referred to as surrogate analysis), and an important factor of this method is the discriminating statistic that determines whether the null hypothesis is rejected or not. Section 2.4 discusses some important issues to consider when choosing a discriminating statistic, which depend on the nature of the surrogate analysis and the selected time series.

In Chapter Three our methodology is discussed, firstly examining a modelling algorithm defined by Judd and Mees [16, 21] in Section 3.1. Judd and Mees use a variable embedding technique [17] to embed the time series, and then build a radial basis model using an iterative algorithm [19, 21] with minimum description length [24] as the stop criterion. In Section 3.2 we explain the Method of Surrogate Data as applied in this study. Firstly, we use the Method of Surrogate Data to find evidence of any nonlinear structure in the specific time series based on similar applications by Theiler and Rapp [40], and Small and Judd [20, 31]. Secondly, the Method of Surrogate Data is used to verify that the model, generated by Judd and Mees's modelling algorithm [16, 21], does indeed model the dynamics of the underlying system represented by the time series. If that is the case, then the model prediction error should resemble identically and independently distributed (i.i.d) noise. The discriminating statistic we used for both surrogate analyses is correlation dimension [13, 14] based on Judd's [15] methodology.

Section 3.3 gives details of the time series selected and the experiments performed in this study. We selected three time series from the data library of the Chaos Data Analyzer software program [33]. The time series generated by the Lorenz equations was selected because it serves as a prototypical example of a dissipative chaotic flow. A time series representing an electroencephalogram (EEG) was our second choice, primarily due to research done by Theiler and Rapp [37, 40], and because the research done by Theiler et al. [38, 39] on the Method of Surrogate Data, was examined extensively in this study. The time series of Standard and Poor's Composite Index of 500 stocks (S&P500) was selected as the third time series, due to our active interest in financial time series based on our work environment.

The results of our applications of the Method of Surrogate Data are discussed in Chapter Four. In Section 4.1.1, the results of the surrogate analysis of the Lorenz time series are examined. After

performing the surrogate analysis, the Lorenz time series was modelled using Judd and Mees's modelling algorithm [16, 21]. The model was tested by means of one-step and free-run predictions in Section 4.1.2. Using the Judd and Mees's iterative modelling algorithm, a model was constructed with every iteration, and the residual vectors of these models were analysed using the Method of Surrogate Data. The argument followed was that once the "best" model has been selected according to the minimum description length criterion, the resulting residual vector should resemble identically and independently distributed (i.i.d) noise. The results of the surrogate analysis of the models' residual vectors are examined in Section 4.1.3.

Section 4.2 deals with the analysis and modelling of the EEG time series. The outcome of the surrogate analysis of this time series is discussed in Section 4.2.1. In Section 4.2.2 we look at the model that was built for the EEG time series, using Judd and Mees's modelling algorithm [16, 21]. The results of the surrogate analyses of the models' residual vectors are examined in Section 4.2.3.

In Section 4.3 we discuss the application of our methodology to the S&P500 time series. The surrogate analysis of the S&P500 time series is investigated in Section 4.3.1.

The final part of this thesis is the Conclusion, summarising the insights gained through applying nonlinear modelling techniques and the Method of Surrogate Data to the selected time series.

CHAPTER 1

NONLINEAR MODELLING

The question that lies at the heart of time series analysis is the following:

> What can we infer of the dynamic laws governing a system, given a sequence of observations of one, or a few, time variable characteristics of such a system?

Our reason for posing this question is the desire to predict and control the particular system from which the time series was collected. There are different sources of predictability in a time series:

1. the existence of linear correlation in time, and
2. determinism.

If the first source of predictability is all the information we want to include in the model, then the best choice is either a linear stochastic model, a moving average (MA), an autoregressive (AR), or a combination of a moving average and an autoregressive model (ARMA). The reader is referred to [18] for a general introduction to these stochastic models.

If the second source of predictability (determinism) is of interest, then we would attempt to construct a deterministic model. In its purest form, determinism implies that equal states have an equal future, or similar states will evolve similarly, at least over short time periods. Our main interest is deterministic models, and they will be discussed together with the fundamental principles of nonlinear modelling. If a more detailed discussion is of interest, the reader is referred to the vast amount of literature that is available on this subject. References [1] and [18] are two examples. Many time series will exhibit a combination of both the mentioned sources of predictability, particularly in the case of experimental data. Currently, there only exist methods to optimally exploit either the linear correlations or the nonlinear determinism and we have to decide whether we want to model the process as a linear stochastic one or as a nonlinear deterministic one.

In this chapter we explore attractor reconstruction (Section 1.1) and specifically the choice of embedding dimension and time lag. In Section 1.2, model building is briefly discussed, as well as

an introduction of four types of basis functions, that can be used in modelling the attractor. Once it has been determined which basis functions to use for building the model, we need to establish a criterion to determine the quality of the model (as is discussed in Section 1.3).

1.1 Attractor Reconstruction

In order to model deterministic dynamics, or a dominant deterministic part of some mixed system, the scalar measurements must be embedded to reconstruct the attractor. An attractor is the sub-set of the phase-space that the system is attracted to, after some transient time, and is invariant under the dynamic evolution of the system. In the following section the basic theory underlying the method of attractor reconstruction is discussed. For a more detailed discussion on attractor reconstruction, Abarbanel et al. [1] may be consulted.

Assume (1) M is a compact m-dimensional manifold of a time series z;

(2) $Z: M \to M$ is a C^2 vector field on M ; and

(3) . $s: M \to \Re$ is a C^2 function, known as the measurement function. $s: M \to \Re$

The vector field Z gives rise to an associated evolution-operator (flow) $\phi_t : M \alpha M$ that evolves $z_t \in M$ to $z_{t+\tau} \in M$. If $z_t \in M$ is the state of the system at time t, then the state of the system at time $t+\tau$ is given by $z_{t+\tau} = \phi_\tau(z_t)$. Based on the definition of the measurement function above, the observation at time t is $s(z_t) \in \Re$ and at time $t+\tau$ the observation is defined as $s(z_{t+\tau}) = s(\phi_\tau(z_t))$.

According to Takens' embedding theorem [35], given the situation depicted above, the system that is generated by the map $\Phi_{Z,s} : M \to \Re^{2m+1}$ where

$$\begin{aligned}\Phi_{Z,s}(z_t) &:= (s(z_t), s(\phi_\tau(z_t)),\ldots, s(\phi_{2m\tau}(z_t))) \\ &= (s(z_t), s(z_{t+\tau}),\ldots, s(z_{t+2m\tau}))\end{aligned} \qquad (1.1)$$

is an "embedding". By embedding we mean that the asymptotic behaviour of $\Phi_{Z,s}(z_t)$ and z_t are diffeomorphic.

An attractor can now be reconstructed using this result. Given a time series of experimental observations $\{y_t\}_{t=1}^{N}$ (where $y_t = s(z_t)$), a d_e-dimensional vector time series is generated using the embedding (1.1):

$$y_{t-\tau} \rightarrow v_t = (y_{t-\tau}, y_{t-2\tau}, \ldots, y_{t-d_e\tau}) \qquad \forall t > d_e\tau .$$

To proceed with the reconstruction of the attractor, values for the parameters τ (time lag) and d_e (embedding dimension) must be estimated. A good embedding, defined as a system that should be asymptotically diffeomorphic to the underlying dynamics, is a function of the choice of the embedding parameters τ and d_e. According to Takens's embedding theorem [35], any embedding with $d_e > 2d_c$ is asymptotically diffeomorphic to the underlying dynamics as $N \rightarrow \infty$.

1.1.1 Embedding dimension

Takens [35] and Mañé [20] provided an answer to the question as to the value to choose for the embedding dimension, d_e. They found that the desired value for d_e is a function of the dimension of the attractor, d_A. They argue that if d_e, which is an integer, is larger than $2d_A$, which can be fractional, then the attractor, as seen in the space with lagged coordinates, will be smoothly related to the attractor, as viewed in the original, unknown physical coordinates. However, calculating the dimension of the attractor is no trivial task.

One way of estimating the embedding dimension is to start with an arbitrary value for d_e and then to embed using increasing (subsequent) values of d_e, until consistent results are produced, i.e. the attractor is unfolded. This is the method that is most often applied in practice. Although this method produces acceptable results, we would prefer a method that estimates the embedding dimension with a certain accuracy. False Nearest Neighbours [19] is one such method. As an example, assume a d_0+1-dimensional delay reconstruction is an embedding, but a d_0-dimensional delay reconstruction is not. The reason for this occurrence is that, in the passage from d_0 dimensions to d_0+1 dimensions, the "false" neighbours disappear, and we are left with "true" neighbours, i.e. points lie close to one another because it is a property of the attractor, not because the embedding dimension is too small. Therefore, two points will also be neighbours in a delay reconstruction in dimensions greater than d_0+1, while in the d_0-dimensional delay reconstruction

there will exist at least two points that are neighbours in that dimension, but not in the higher dimensional delay reconstruction.

The method of False Nearest Neighbours [19] is based on the following methodology. We embed a scalar time series $\{y_t\}_{t=1}^{N}$ in increasingly higher dimensions, while at each stage comparing the number of pairs of vectors v_t and v_t^{NN} (the nearest neighbour of v_t) which are close when embedded in R^{d_0} but not close in R^{d_0+1}. Each point

$$v_t = (y_{t-\tau}, y_{t-2\tau}, \ldots, y_{t-d_0\tau})$$

has a nearest neighbour

$$v_t^{NN} = (y_{t'-\tau}, y_{t'-2\tau}, \ldots, y_{t'-d_0\tau}).$$

When there is a large amount of data the distance (Euclidean norm is sufficient) between v_t and v_t^{NN} should be small. If these two points are genuine neighbours, they became close due to the system dynamics and should separate slowly. However, these two points may have become close because the embedding in R^{d_0} has produced trajectories that cross (or become close) due to the embedding, and not the system dynamics. For each pair of neighbours v_t and v_t^{NN} in R^{d_0} one can increase the embedding dimension by 1 so that

$$\hat{v}_t = (y_{t-\tau}, y_{t-2\tau}, \ldots, y_{t-d_0\tau}, y_{t-(d_0+1)\tau})$$

and

$$\hat{v}_t^{NN} = (y_{t'-\tau}, y_{t'-2\tau}, \ldots, y_{t'-d_0\tau}, y_{t'-(d_0+1)\tau})$$

may or may not still be close. The increase in the distance between these two points is given only by the difference between the last components:

$$\left\|\hat{v}_t - \hat{v}_t^{NN}\right\|^2 - \left\|v_t - v_t^{NN}\right\|^2 = (y_{t-(d_0+1)\tau} - y_{t'-(d_0+1)\tau})^2.$$

In order to establish whether neighbours are true or false, we still require some threshold size, R_T, to indicate that R^{d_0+1} is large enough. Kennel et al. [19] recommend using the criterion

$$\frac{\left|y_{t-(d_0+1)\tau} - y_{t'-(d_0+1)\tau}\right|}{\left\|v_t - v_t^{NN}\right\|} > R_T ,$$

and claim that $R_T \geq 10$ is a good indication that neighbours are "false".

We can choose the embedding dimension, d_e, as the minimum value of d_0, for which the proportion of points that satisfy the above condition, is below some small threshold. We expect that the percentage of neighbours should decrease as the embedding dimension increases, thereafter stabilising at a low value. Figure 1.1 serves as a typical example of the relationship between the percentage of false nearest neighbours and embedding dimension.

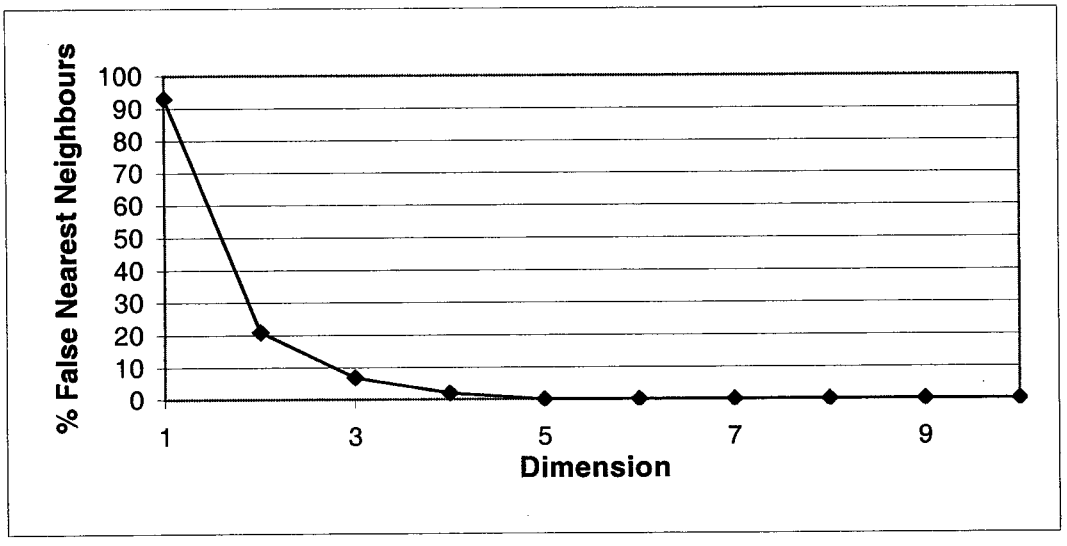

Figure 1.1. **False Nearest Neighbours:** The method of False Nearest Neighbours was applied to a time series of 2000 data points, generated by a Lorenz attractor and embedded with a time lag of $\tau = 3$. (The method used for determining the value of τ is discussed in Section 1.1.2.) The figure illustrates the relationship between the percentage of false nearest neighbours and the embedding dimension. With a threshold of 10 (R_T = 10), we can choose the embedding dimension, d_e, as 5.

1.1.2 Time lag

It is important to notice that the time lag, τ, is not an element of the embedding theorem, (although it is a hidden element), and therefore, from a mathematical viewpoint, the choice of the value of the time lag is arbitrary. In applications, however, it is important to consider the value of τ seriously, as the shape of the embedded times series is influenced by the value of the time lag.

A variety of criteria can be used to influence the choice of τ, i.e. the first zero of the autocorrelation function [18], the first minimum of the mutual information criteria [23], and average mutual information [11] to name but a few. We used the method of Average Mutual Information (AMI) as implemented by the software package "Contemporary Signal Processing for Windows". For a detailed discussion of this method, see [11]. A brief discussion of the rationale of the AMI method follows.

The method of Average Mutual Information was developed in information theory. It is an estimation of the information available regarding $y(t+\tau)$ with the knowledge of $y(t)$. In order to calculate the mutual information, we must establish the dependency of the values of $y(t+T)$, on the values of $y(t)$ for $T>0$.

Suppose we have two systems; X and Y, which generated the data $x(t)$ and $y(t)=x(t+T)$. We assume that there is a probability distribution associated with each system, that governs the possible outcomes of observations of these values. The amount we can learn (in binary bits) about a measurement of y_j from a measurement of x_i, is given by the arguments of information theory as

$$I_{X,Y}(x_i, y_j) = \log_2\left[\frac{P_{X,Y}(x_i, y_j)}{P_X(x_i)P_Y(y_j)}\right],$$

where the probability of observing x_i out of the set of all X is $P_X(x_i)$ and the probability of observing y_j in the set of all of Y, is $P_Y(y_j)$. The joint probability of the measurement of x_i and y_j is $P_{X,Y}(x_i, y_j)$. This quantity, $I_{X,Y}(x_i,y_j)$, is called the "mutual information" of the two

measurements x_i and y_j and is equivalent to the amount we learn about y_j, from measuring x_i.

The average mutual information between measurements of any value x_i from system X, and y_j from system Y, is the average over all possible measurements of $I_{X,Y}(x_i, y_j)$:

$$I_{X,Y}(T) = \sum_{x_i y_j} P_{X,Y}(x_i, y_j) I_{X,Y}(x_i, y_j).$$

The average mutual information between observations at time step n and at $n+T$, i.e. the average amount of information we have about $y(n+T)$ when we make an observation of $y(n)$, is then

$$I(T) = \sum_{n=1}^{N-T} P(y_n, y_{n+T}) \log_2 \left[\frac{P(y_n, y_{n+T})}{P(y_n) P(y_{n+T})} \right].$$

Fraser and Swinney [11], suggest as a general guideline, that we choose the time lag, τ, as that value of T which corresponds to the first minimum of the function $I(T)$. Figure 1.2 illustrates the relationship between the average mutual information, $I(T)$ and the time delay, T.

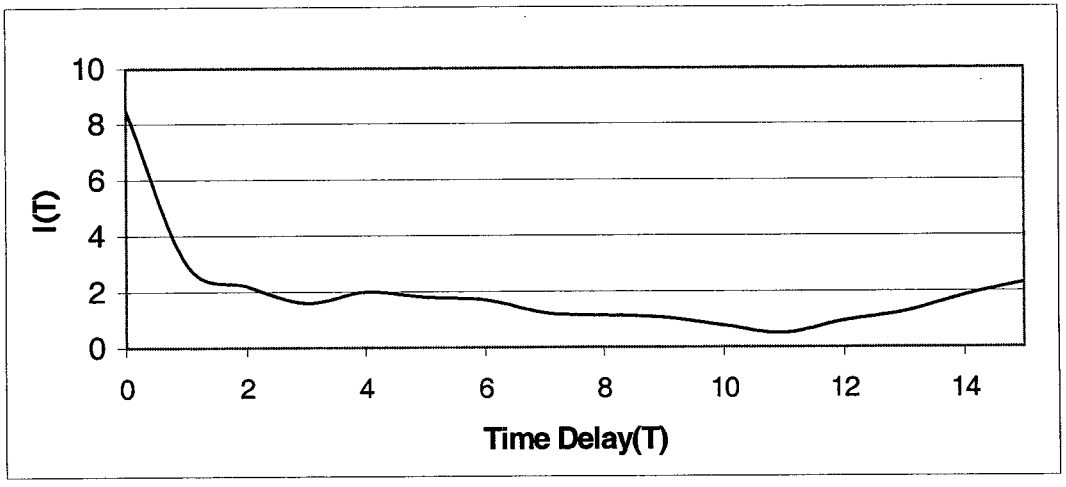

Figure 1.2 **Average Mutual Information:** The method of Average Mutual Information was applied to a time series of 2000 data points, generated by a Lorenz attractor with the embedding, $d_e = 5$. A closer study reveals that the first minimum occurs at $T = 3$, therefore 3 is a good choice for the time lag, τ.

1.2 Model building

Since the time series data are discretely sampled over time, a deterministic model is always a map. In delay embedding space with time lag, τ, embedding dimension, d_e, and the embedding $v_t = (y_{t-\tau}, y_{t-2\tau}, \ldots, y_{t-d_e\tau})$, it reads:

$$v_{t+1} = (F(v_t), y_{(t+1)-2\tau}, \ldots, y_{(t+1)-d_e\tau}).$$

Thus all we need for a forecast is a prediction of $y_{(t+1)-\tau}$, that is we want to find a scalar function $F(v_t)$ such that:

$$\hat{y}_{(t+1)-\tau} = F(v_t).$$

We need to select a general form for the function F containing enough freedom so that it is capable of representing the data. The idea of modelling consists of choosing an appropriate functional form for F, which is flexible enough to model the true function on the whole attractor.

Taking F as a linear superposition of basis functions is one method of modelling the attractor:

$$F(v) = \sum_{i=1}^{k} \alpha_i \Phi_i. \qquad (1.2.1)$$

During the fitting process the k basis functions, Φ_i are kept fixed, while the coefficients α_i are varied. Polynomials, neural networks and radial basis functions are examples of the types of basis functions that are used in practice. Each of these three examples will be discussed briefly.

1.2.1 Polynomials

Using a polynomial to model the time series is inferred by the observation that global linear models can be regarded as the first approximation in a Taylor series expansion of $F(v)$ [18]. Since F acts on an m-dimensional space, it has to be a multivariate polynomial, which, for order l, has $k = {}^{(m+l)!}/_{m!l!}$ independent coefficients. Although the large number of parameters may seem

daunting, they can usually be readily determined as long as the time series does not contain too much noise. An advantage of polynomials is that there properties and behaviour are well known, which facilitates an understanding of the result. A shortcoming of such a model is its discontinuities (basis functions diverge for large arguments and trajectories may escape to infinity).

1.2.2. Neural networks

Because prediction is one of the primary goals of nonlinear modelling, it makes sense to use neural networks, which are supposed to have the ability to learn complex tasks, such as prediction. Neural networks use interconnected elementary units (neurons) with a specific architecture. After an initial learning phase, during which the parameters are calculated, the interconnections between the neurons are established, and the result is a nonlinear model comprising of, for example, sigmoid functions instead of polynomials or radial basis functions. The reader is referred to [1] and [18] for an introductory discussion on neural networks.

One particular class of neural networks that has been used for time series modelling comprises of the so-called "feed-forward networks" with one hidden layer. This means that there is one layer of input units, one layer of neurons, and one layer of output units. In the case of scalar predictions, the layer of output units is a single unit summing up the output of the neurons. The input layer consists of m units if we work in an m-dimensional embedding space, and it does nothing but distribute the m components of the delay vectors to the neurons. Since the whole structure of the net is inspired by the nervous system, the function of a neuron is usually a smooth step function or a sigmoid function such as $\Phi = 1/\left(1+e^{bv+c}\right)$. The whole net is thus nothing other than the function

$$F(v) = \sum_{i=1}^{k} \frac{a_i}{1+\exp(b_i v - c_i)},$$

where the parameters a_i, b_i, and c_i have to be determined using fitting techniques. The parameters are determined by the learning (or training) phase and iteratively modified. The most popular fitting algorithm is error backpropagation which is nothing more than a gradient descent method, where a cost function is minimised by presenting all learn pairs (v_{t+1}, v_t) of the one-step prediction error to the function F. Learning algorithms do not all seem to be equally effective at finding the

global minimum, and therefore the successful training of neural nets often requires a considerable amount of experience.

1.2.3. Radial basis functions

In the case of radial basis functions [16, 18, 21], one defines a scalar function $\Phi(r)$ with only positive arguments r. The general form of the function F is:

$$F(v) = \lambda_0 + \sum_{i=1}^{k} \lambda_i \Phi(\|v - c_i\|),$$

with λ_0 a constant function as the zeroth basis function, λ_i, i =1,..,k as parameters and c_i, i=1,..,k as the centre of the i^{th} basis function. The selection of the parameters λ_i and centres c_i is a complex nonlinear optimisation problem. Typical basis functions Φ are bell-shaped with a maximum at $r = 0$ and a rapid decay factor towards 0 with increasing r. The function F is modelled by adjusting the coefficients λ_i of the functions Φ. If the centres, c_i are reasonably well distributed on the attractor, the linear superposition yields a well-behaved surface.

1.3 Cost functions

Once a decision has been made on which basis functions to use for building the model, we need to find a criterion for the quality of the model. The usual way is to commence with a cost function, which is then minimised. The choice of the proper cost function depends on the purpose of the model.

A cost function measures the quality of the fit of the model to the data. One approach is to maximise the likelihood of the observations with respect to the model [18]. If the prediction errors are assumed to follow a Gaussian distribution, maximising the likelihood of the observations, amounts to minimising the mean squared prediction error,

$$e = \sqrt{1/N \sum_{t=1}^{N} (y_{t+1} - F(v_t))^2} \quad , \qquad (1.3.1)$$

where N is the number of data points of the scalar time series $\{y_t\}_{t=1}^{N}$ and $F(v_t)$ is a linear superposition of basis functions as detailed in equation 1.2.1. However, this is not always the optimal cost function. Overfitting can occur since the more free parameters a model contains, the better it can be adapted to the data. In order to avoid this problem it is necessary to add a term for the model costs to the minimisation problem. This is done by adding an appropriate function of the number of adjustable parameters to the likelihood function. Suppose we start from a general function for the dynamics, $F_{a_1,\ldots,a_k}(v_t)$, depending on k adjustable parameters a_j, j=1,…,k. We want to find the particular set of parameters $\{\hat{a}_j\}$ which maximises the probability $p(\{y_t\},\{a_j\})$ of observing the data sequence $\{y_t\}$ when the underlying model is $F_{a_1,\ldots,a_k}(v_t)$, i.e. maximise the loglikelihood function

$$L = -\ln p(\{y_t\},\{a_j\}) \tag{1.3.2}$$

with respect to the parameters $\{a_j\}$. If the errors $y_{t+1} - F_{\{a_j\}}(v_t)$ are Gaussian distributed with variance σ^2, the probability of observing the data is

$$p(\{y_t\},\{a_j\}) = \frac{1}{(\sigma\sqrt{2\pi})^N}\exp\left[-\frac{\sum_{t=1}^{N}(y_{t+1} - F_{\{a_j\}}(v_t))^2}{2\sigma^2}\right]. \tag{1.3.3}$$

Estimating the variance σ^2 by $\hat{\sigma}^2 = \sum_{t=1}^{N}(y_{t+1} - F(v_t))^2 / N$ we obtain

$$L_{Guass} = \frac{N}{2}\ln\hat{\sigma}^2 + \frac{N}{2}\ln 2\pi + \frac{N}{2}, \tag{1.3.4}$$

which is maximal when the variance estimator is minimal. This implies that the mean squared prediction error, e, will also be minimal due to the relationship $\hat{\sigma}^2 = e^2$ (this follows from equation 1.3.1 and the equation for the variance estimator, $\hat{\sigma}^2$).

Equation 1.3.2 lacks a term that takes the complexity of model F into account. Therefore a term $-\alpha k$ is added to L. The value L is calculated for different values of k (number of parameters) and the model is selected based on the maximum value of L.

Depending on the goal of the analysis, there are two suggested values of the constant, α. If minimising the expected value of the out-of-sample prediction error is required, then the choice of α should be 1. This suggestion is based on information theoretic arguments and is known as the Akaike information criterion [3].

If the main interest is the model itself, it is preferable to use $\alpha = \frac{1}{2}\ln N$ which is known as the Schwarz criterion [27].

Thus the loglikelihood functions including the complexity term are:

$$L' = -\ln p(\{y_t\};\{a_j\}) - \begin{cases} k & \text{(Akaike)} \\ \frac{k}{2}\ln N & \text{(Schwarz).} \end{cases}$$

The corresponding changes to the mean squared cost function for Gaussian errors are:

$$e'^2 = \frac{1}{N}\sum_{t=1}^{N}(y_{t+1} - F(v_t))^2 \times \begin{cases} e^{\frac{k}{N}} & \text{(Akaike)} \\ e^{\frac{k}{2N}\ln N} & \text{(Schwarz).} \end{cases}$$

A discussion of modelling criteria would be incomplete without mention of description length based on the ideas by Rissanen [24]. This criterion is also derived from information technology and is based on the optimal encoding of the data. The data is encoded in two parts:

- a description of the model including the parameters, and
- the model prediction errors.

The basic premise is that for a given resolution, fewer bits are needed to encode the residual errors if these errors are smaller than the original data. However, this gain must be weighed up against the encoding of the model. For a more detailed discussion on description length, the reader is referred to Section 3.1.1.

CHAPTER 2

THE METHOD OF SURROGATE DATA

The Method of Surrogate Data [36, 38, 39] is often employed to identify the presence of nonlinear, possibly chaotic behaviour in a time series. Whilst nonlinear measures such as correlation dimension, Lyapunov exponents and nonlinear prediction error are often applied with the intention of identifying the presence of nonlinear behaviour in a time series (see for example [2, 7, 9, 25]), quantitative estimations and subsequent classifications are difficult and uncertain.

Broadly, the Method of Surrogate Data [36, 38, 39] compares the value of a statistic for the time series to the approximate distribution of the statistic for a class of nonlinear processes. From this comparison we can make assumptions whether the time series has some characteristics which are distinct from stochastic linear systems. Statistical hypothesis testing is essentially the basic procedure of surrogate data methods, as described by Theiler et al. [38, 39] as well as Takens [35]. We specify a well-defined underlying linear process (the null hypothesis). We then determine the distribution of the quantity of interest (e.g. correlation dimension) for an ensemble of surrogate data sets, which are simply different realisations of the hypothesised linear stochastic process. Error bars are put on the value given by the surrogates, rather than on the correlation dimension of the original data. This can be done reliably, because one has a model for the underlying dynamics (the null hypothesis itself). As we have many realisations of the null hypothesis, we can estimate the error bar numerically from the standard deviation of all estimated correlation dimensions of the surrogate data sets. In this chapter we introduce some terminology relating to the Method of Surrogate Data in Section 2.1 and review common methods of generating linear surrogates in Section 2.2. In Section 2.3 a method for generating nonlinear surrogates is discussed, and in Section 2.4 we explore the choosing of a discriminating statistic.

2.1. Terminology

When defining the null hypothesis, certain assumptions are made as to the nature of the system underlying the data. The discriminating statistic provides a criterion on which to base the acceptance, or rejection, of the null hypothesis: if this criterion is different for the observed data, than would be expected under the null hypothesis, the null hypothesis can be rejected.

The distribution of the statistic is estimated by direct Monte Carlo simulation [5]. An ensemble of surrogate data sets, which share given properties of the observed time series (such as mean, variance, and Fourier spectrum), but are otherwise random as specified by the null hypothesis, are generated. For each surrogate data set the discriminating statistic is computed, and from this ensemble of statistics the distribution is approximated.

Let ϕ be a specific hypothesis and F_ϕ the set of all processes consistent with that hypothesis. Let $z \in \Re^N$ be a time series consisting of N scalar measurements, and let $T:\Re^N \rightarrow U$ be a statistic which we will use to test the hypothesis ϕ that z was generated by some process $F \in F_\phi$. Surrogate data sets z_i, $i = 1,2,3\ldots$ that are consistent with the hypothesis ϕ being tested and of the same length as z are generated from z. Generally U will be in $\Re$, and one can discriminate between the time series z and surrogates z_i consistent with the hypothesis, given the approximate probability density $\rho_{F,T}(t) = \text{Prob}(T(z_i) < t)$, i.e. the probability density of T given F.

Theiler and Prichard [39] point out that it is important to distinguish two types of null hypotheses: "simple" and "composite".

Definition 2.1 A "simple" null hypothesis asserts that the given time series is a random realisation of a specified unique process, i.e. the set of all processes consistent with the hypothesis F_ϕ is singleton. The "composite" null hypothesis specifies a family of processes.

They also suggest that there are two fundamentally different types of test statistic: "pivotal" and "non-pivotal".

Definition 2.2 A test statistic T is "pivotal" if the distribution of T is the same for all members F of the family F_ϕ of processes consistent with the null hypothesis; otherwise it is "non-pivotal".

For example, Lyapunov exponent and correlation dimension are pivotal test statistics if F_ϕ belongs to the class of noise processes while autocorrelation, nonlinear prediction error and rank distribution statistics (standard deviation or higher moments) are non-pivotal statistics.

When the null hypothesis is composite, it is difficult knowing which process (in the family of processes covered by the null hypothesis) we should be simulating. This complication was pointed out by Theiler and Prichard [39]. Demanding that T is pivotal (which means that the distribution of T is the same for all members F of the family F_ϕ of processes consistent with the null hypothesis) solves this problem.

There are situations when using a non-pivotal statistic is preferable, e.g. using nonlinear prediction error to distinguish deterministic chaos from stochastic processes by its predictability. Theiler and Prichard [39] suggest that in the case of a composite hypothesis and a non-pivotal statistic, a constrained realisation scheme should be employed.

Definition 2.3 Let $\hat{F} \in F_\phi$ be the process estimated from the time series z, and let z_i be a surrogate data set generated from $F_i \in F_\phi$. Let $\hat{F}_i \in F_\phi$ be the process estimated from z_i, then a surrogate z_i is a constrained realisation if $\hat{F}_i = \hat{F}$. Otherwise it is non-constrained.

We will use the time series to estimate the parameters of the underlying process, and then generate surrogates that are typical realisations of a model of the time series. We then focus on those realisations that give identical estimates of the parameters, compared to the parameter estimates from the time series.

For example, let ϕ be that hypothesis that z is generated by independently and identically distributed (i.i.d) noise. To generate non-constrained surrogates, one would estimate or guess the best linear model (from z), and generate realisations from this assumed model.

By shuffling the phases of the Fourier transform of z, constrained realisation surrogates are generated. This method produces a random data set with the same power spectra, and hence autocorrelation, as the data. As previously mentioned, autocorrelation, nonlinear prediction error, and rank distribution statistics (standard deviation or higher moments) are examples of non-pivotal test statistics. The probability distribution of these statistic values depends on the form of the noise source and type of linear filter. Correlation dimension or Lyapunov exponents are pivotal test statistics, but producing a pivotal estimate of these quantities, is difficult. Due to the pivotal characteristic of these statistics, the probability distribution of these quantities will be the

same for all processes, and the exactness of the estimation of the linear model and i.i.d noise source, is less important. See Theiler and Prichard [39] for a complete discussion on constrained realisations of a time series and the implications for the test statistic used in the surrogate analysis.

2.2. Linear surrogates

In the original work by Theiler et al. [38] on surrogate methods, a "hierarchy" of hypotheses that test for membership of particular classes of linear systems, was suggested. More specifically, the null hypothesis defines the nature of the candidate process, which may or may not adequately explain the data. This process is defined as belonging to a specific dynamic class, and different types of surrogate data are generated to test this defined membership. The "hierarchy" as suggested by Theiler et al. [38], consists of three types of linear surrogates known as algorithms 0, 1 and 2 and address the following three hypotheses:

Algorithm 0: linearly filtered noise;
Algorithm 1: linear transformation of linearly filtered noise; and
Algorithm 2: monotonic, nonlinear transformation of linearly filtered noise.

Constrained realisations consistent with each of these hypotheses can be generated by:

Algorithm 0: shuffling the data;
Algorithm 1: randomising (or shuffling) the phases of the Fourier transform of the data; and
Algorithm 2: reordering the data to have the same rank distribution as phase randomised, amplitude adjusted Gaussian noise.

The methods for generating each of the three linear surrogate types are explained below.

Algorithm 2.1 Algorithm 0 surrogates:

1. Generate an i.i.d Gaussian data set y.
2. Reorder z so that it has the same rank distribution as y.
3. The surrogate z_i is the reordering of z.

Algorithm 2.2 Algorithm 1 surrogates:

1. Calculate the Fourier transform Z of z.
2. Either randomise the phases of Z or shuffle them by applying algorithm 2.1.
3. Take the inverse Fourier transform to produce the surrogate z_i.

Algorithm 2.3 Algorithm 2 surrogates (Also referred to as Amplitude Adjusted Fourier Transformed (AAFT) surrogates):

1. Generate an i.i.d Gaussian data set y.
2. Reorder y so that it has the same rank distribution as z.
3. Create an Algorithm 1 surrogate y_i of y. (Either by shuffling or randomising the phases of the Fourier transform of y.)
4. Reorder the original data z to create surrogate z_i which has the same rank distribution as y_i.

If the data was generated by a nonlinear system, each of these hypotheses should be rejected. Rejecting these hypotheses does not necessarily indicate the presence of a nonlinear system, but only that it is unlikely that the data is generated by a monotonic, nonlinear transformation of linearly filtered noise.

In a recent article by Screiber and Schmitz [26], it was pointed out that the algorithm for generating AAFT surrogates might be flawed. Using Theiler's algorithm the data and surrogates have the same distributions of amplitudes by construction, but do not usually have the same sample power spectra. There are two reasons for this. Firstly, step 3 of algorithm 2.3 preserves the Gaussian distribution only on average; the fluctuations of y and y_i will differ in detail. Futhermore, the nonlinearity contained in the amplitude adjustment procedure (step 4 of algorithm 2.3) will turn these into a bias in the empirical power spectrum. Such systematic errors can lead to false rejections of the null hypothesis if a statistic is used which is sensitive to autocorrelations. Secondly, step 4 of algorithm 2.3 is not the inverse of the nonlinear measurement function, and instead of recovering the time series $\{z_n\}$ we will find some other Gaussian series. Schreiber and Schmitz suggest an alternative algorithm [26] for generating Algorithm 2 surrogates, and verify that false rejections are indeed avoided by this scheme.

Algorithm 2.4 Schreiber and Schmitz's iterative algorithm for generating type 2 surrogates:

1. Store a sorted list of values $\{y_t\}$ and the squared amplitudes of the Fourier transform of $\{y_t\}$, $S_k^2 = \left|\sum_{t=0}^{N-1} y_t e^{\frac{i2\pi kt}{N}}\right|^2$.
2. Begin with a random shuffle, $\{y_t^0\}$ of the data.
3. Bring $\{y_t^{(i)}\}$ to the desired sample power spectrum: take the Fourier transform of $\{y_t^{(i)}\}$, replace the squared amplitudes $\{S_k^{2,(i)}\}$ by $\{S_k^2\}$ and transform back. The phases of the complex Fourier components are kept. This step enforces the correct spectrum but usually the distribution will be modified.
4. Rank order the resulting series in order to assume exactly the values taken by $\{y_t\}$. The spectrum of the resulting $\{y_t^{(i+1)}\}$ will be modified again, and therefore one needs to repeat step 3.
5. At each iteration stage one can check the remaining discrepancy of the spectrum and iterate until an acceptable accuracy is reached.

Convergence to the same Fourier spectra is also not guaranteed under this method, but their results seem to indicate a closer agreement between power spectra. Small et al. [28, 31, 32] have found that, although estimates of the power spectra may not agree very closely for small to moderately large values of τ, the autocorrelation using standard AAFT surrogate generation techniques does. Additionally, linear surrogates are all forms of filtered noise and effectively infinitely dimensional. This relieves the concerns raised by Schreiber and Schmitz [26], since if the test statistic is pivotal it is irrelevant whether a constrained or non-constrained realisation scheme was used.

In [37], Theiler also addressed the problem of determining the presence of temporal correlation between cycles, in the case of periodic cycles. He proposed that the choice of surrogates should also be periodic. Theiler decomposed the signal into cycles and shuffled the individual cycles. The hypothesis tested is that there is no dynamic correlation between the cycles. The algorithm used to generate these surrogates is depicted in algorithm 2.5.

Algorithm 2.5 Cycle shuffled surrogates

1. Split the signal, z, into its individual cycles by identifying the location of the peak or some other convenient point within each cycle.
2. Randomly reorder the cycles and form a new time series, z_i, by concatenating the individual cycles. If the original time series, z, is even slightly non-stationary then the individual cycles will almost certainly have to be shifted vertically to preserve the "continuity" of the original time series, z.

In some respect this algorithm is analogous to algorithm 0, except that it tests temporal correlation between cycles and not data points. In [28] Small, after having examined the correlation between cycles directly by reducing each cycle to a single measurement, points out that decomposing the time series into cycles, can result in substantial loss of information, which might complicate the modelling process.

2.3. Nonlinear surrogates

The surrogate methods discussed in the previous section are non-parametric, and are concerned with rejecting the hypothesis that some form of linear system generates a given data set. An obvious extension of this would be to classes of nonlinear parametric models.

In [28, 29, 30, 31] Small and Judd suggest a method for generating surrogates that are parametric and nonlinear. They address the hypothesis that the data comes from a noise driven nonlinear system of a particular form. A cylindrical basis model is applied to the data, and then noise driven simulations are generated from that model. These noise driven simulations are the surrogates. Rejection of the hypothesis could imply that a cylindrical basis model cannot describe the data, or that the modelling algorithm failed to build an accurate model.

It is important to note that nonlinear surrogates generated by this method will be non-constrained surrogates. Therefore attention should be paid to the "pivotalness" of the chosen test statistic. In [32] Small and Judd demonstrate that correlation dimension as estimated by Judd's algorithm [15] provides a suitable statistic, and apply the aforementioned methods to respiratory data from sleeping infants.

2.4. Choice of Statistic

Once a number of surrogate data sets have been generated, a statistic T on each of the surrogate data sets, as well as on the original data set, has to be calculated. The statistic should be independent of the generation of the surrogates. In principle virtually any discriminating statistic can be used, but choosing a statistic that actually provides a good estimate of a physically interesting quantity, makes the method more useful. If the null hypothesis is rejected, the statistic can be used to characterise the nature of the nonlinearity informally. In more recent work by Theiler and Rapp [37, 40] it was demonstrated that the quality of test statistics vary, and that the test statistic should be selected with caution.

Theiler et al. [38] suggest a "battery" of discriminating statistics. Their original choices for the discriminating statistic were correlation dimension, Lyapunov exponent and forecasting error, since they were motivated by the possibility that the underlying dynamics may be chaotic. Correlation dimension counts degrees of freedom, Lyapunov exponent quantifies the sensitivity to initial conditions, and forecasting error tests for determinism.

In [39], Theiler and Prichard further debate the choice of a test statistic, depending on the approach used to generate the surrogates (constrained or typical realisations), and the nature of the hypothesis (simple or composite). In the instance of a composite null hypothesis and typical realisations of the specific process, T should be pivotal. The distribution of T will be the same for all members F of the family of processes consistent with the null hypothesis. In practice this often turns out to be a very stringent criterion and in many cases is satisfied only in the asymptotic limit as the size of the data set approaches infinity.

If the surrogate data sets are generated using the constrained-realisation approach, then the test statistic can be pivotal or non-pivotal. Theiler and Prichard [39] showed that the method behaves as if the discriminating statistic was pivotal, even in those cases where it is not. Since the statistic need not be pivotal, the data analyst has more flexibility in designing a test that may be powerful against relatively exotic alternatives.

Small and Judd [31] argue that any dynamic measure is a pivotal statistic for a wide range of standard (linear) and nonlinear hypotheses addressed by surrogate data analysis; however, we must be able to estimate this quantity reliably from data. They used correlation dimension as a test statistic, and showed that it is pivotal for a broad enough range of processes to be an effective

pivotal test statistic for surrogate analysis, for both linear surrogates (Algorithm 0, 1, and 2 surrogates) and nonlinear surrogates, generated by simulations of cylindrical basis models.

The importance of the choice of a test statistic should be clear from this discussion. Insufficient attention to whether the test statistic should be pivotal or non-pivotal, and to the estimation of the chosen statistic, could lead to incorrect results from the surrogate analyses.

CHAPTER 3

METHODOLOGY

The objective of this study (as articulated in the introduction), was to explore the world of nonlinear dynamic systems through the study of nonlinear modelling techniques and the Method of Surrogate Data. This knowledge was utilised to perform surrogate analyses on three time series, with the expectancy of finding evidence of nonlinear structure. When evidence was found that the underlying system represented by the particular time series does appear to have nonlinear structure, the time series was modelled using Judd and Mees's modelling algorithm [16, 21]. The Method of Surrogate Data was also applied within the modelling algorithm, as a means of verifying that the final model successfully models the underlying dynamics of the system.

Chapter Two illustrated that surrogate analysis is, in essence, statistical hypothesis testing, and that it is most often employed to identify the presence of nonlinear, possibly chaotic behaviour in a time series. In this study, surrogate analysis was used to test two hypotheses. Firstly, the hypothesis that the apparent structure in the particular time series was most likely due to nonlinearity, was explored. Once this had been ascertained, it was possible to achieve reasonable success with modelling the time series using nonlinear modelling techniques. Secondly, the hypothesis tested whether the model prediction error, ε_t, corresponding to the best model that had been selected according to the minimum description length criterion, belonged to the dynamic class described as identically and independently distributed (i.i.d) noise. Intuitively one would expect that once the best model of the underlying dynamics of the system (of which the time series is representative), had been selected, then the model prediction error should resemble i.i.d noise. If this is not the case, then we can question whether the model did indeed capture the underlying dynamics of the system. Therefore our study had two main components: a modelling algorithm and the surrogate analyses. In Section 3.1 the modelling algorithm and its components are explained. Section 3.2 discusses the Method of Surrogate Data as applied in this study. An explanation of the applications is detailed in Section 3.3.

3.1. Modelling algorithm

A sophisticated software program developed at the University of Western Australia (UWA) and based on Judd and Mees's algorithm [16, 21] was used to model the time series after evidence of

nonlinearity was established. A brief overview of the approach followed by Judd and Mees are discussed, but the reader is referred to [16] and [21] for a detailed discussion of the methodology.

All models to be considered are of the following form. Consider a random process generating $(X_t, Y_t) \in R \times R^d, t \in Z$, for which there is a relationship

$$Y_t = F(X_t) + \varepsilon_t$$

where the random variables ε_t are independent of Y_γ and X_s, for $\gamma \geq t$ and $s \geq t$. Building a model means, given a realisation of this process (y_t, x_t), $t = 1,2,..,n$, find an approximation $\hat{F}$ of F.

Judd and Mees take the embedding as given, and assume that the difficult choice of embedding dimension, d_e, and time lag, τ, have been dealt with. Judd and Mees make use of a given embedding or variable embedding strategies [17] which results in the embedding being different for each part of the phase space, i.e. locally optimal embeddings. The local embeddings are found and optimised as part of the modelling algorithm.

To visualise the usefulness of this approach, consider modelling the Lorenz system with its butterfly-shaped attractor. When the system state is out on the "wings" of the attractor, a two-dimensional embedding is sufficient to model and predict motions. However, near the origin, where the cross over of the wings occur, a three-dimensional embedding is essential. One can imagine constructing a perfectly adequate model that does not use a global embedding, but rather uses appropriate local embeddings, as the system state varies. The reader is referred to Judd and Mees's article [17] for a detailed discussion on embedding as a modelling problem.

When modelling a time series one restricts attention to a subclass of models $\hat{F}(X) = G(X, \lambda)$ parameterised by $\lambda \in \Re^m$ for some m. The fitting of a model $\hat{F}$ corresponds to an estimation of the parameters λ subject to the constraint that only a finite number of components of λ are non-zero. The model selection problem asks: given the possibly infinite number of models resulting from the different combinations of non-zero parameters, which is the best model?

Judd and Mees restrict themselves to a class of models that they call "strong pseudo-linear" models. "Strong" refers to methods that are well behaved with increasing dimension, and of

which the fitting process is inherently nonlinear. Strong methods include neural nets, radial basis functions and wavelets. The software program used in this study uses radial basis functions when modelling a time series. "Pseudo-linear" refers to the appearance of parameters in a way that allows easy fitting, i.e. the models are linear combinations of nonlinear functions. Rissanen's use of minimum description length [24] to characterise model quality is used to select the most appropriate model from a class of models.

3.1.1 Description Length

The description length criterion is based on ideas by Rissanen [24]. It is derived from information theory and based on the optimal encoding of data to some fixed finite precision. Rissanen argued that we should regard good models as those that compress the data best. The data is encoded in two parts:

- a description of the model including the parameters, and
- the model prediction error.

If the time series does not suit the class of models being considered, then the most economical method to compress the data would be simply to transmit the time series. If, however, there is a model that fits the data efficiently, it is more effective to transmit the model, and then to transmit the deviations of the time series from that predicted by the model, i.e. the model prediction errors. Judd and Mees [16, 21] utilise this idea in their approach. Broadly, the description length is given by an expression of the form

(Description length) ≈(number of data samples) × log(Sum of squares of prediction errors) + (Penalty for number and accuracy of parameters)

The method employed by Judd and Mees to encode integers is that suggested by Rissanen. According to Rissanen the optimal encoding for a floating point binary number $\overline{\lambda}_j = 0.a_1a_2...a_{n_j} \times 2^{m_j}$ is as follows. If λ_j is the jth model parameter and $\overline{\lambda}_j$ is that parameter truncated to n_j binary bits, then the difference between λ_j and $\overline{\lambda}_j$ will be at most $\delta_j = 2^{-n_j}$, which is called the precision of the parameter λ_j. Therefore, to encode $\overline{\lambda}_j$ we need to encode the binary mantissa $a_1a_2...a_{n_j}$ and the exponent m_j, which are two integers. An integer p may be

encoded in $\log_2 p$ bits, but we have to find a way to indicate where the codeword ends. This is achieved by presenting the codeword's length before the codeword, i.e. to send the value of $\lceil \log_2 p \rceil$ first. This requires $\lceil \log_2 \lceil \log_2 p \rceil \rceil$ bits, which is an integer that needs $\lceil \log_2 \lceil \log_2 \lceil \log_2 p \rceil \rceil \rceil$ bits to encode, and so on. Therefore, the length of code must be encoded as a codeword of length

$$L^*(p) = \lceil \log_2 c \rceil + \lceil \log_2 p \rceil + \lceil \log_2 \lceil \log_2 p \rceil \rceil + \lceil \log_2 \lceil \log_2 \lceil \log_2 p \rceil \rceil \rceil + \ldots$$

bits. This sequence continues until the last term is either 0 or 1, $\lceil x \rceil$ is the smallest integer not less than x, and $\lceil \log_2 c \rceil$ is an additional cost associated with small integers. The cost of encoding λ_j is given by

$$L(\bar{\lambda}_j) = L^*\left(\frac{1}{\delta_j}\right) + L^*\left(\left\lceil \log_2\left(2\max\left\{\bar{\lambda}_j, \frac{1}{\bar{\lambda}_j}\right\}\right)\right\rceil\right) \text{bits.}$$

Making a substitution of nats (natural logarithm) for bits, one arrives at the cost of encoding all the parameters as

$$L(\bar{\lambda}) = \sum_{j=1}^{k} L^*\left(\left\lceil \frac{1}{\delta_j} \right\rceil\right) + \sum_{j=1}^{k} L^*\left(\left\lceil \ln\left(2\max\left\{\bar{\lambda}_j, \frac{1}{\bar{\lambda}_j}\right\}\right) \right\rceil\right) \text{nats.} \quad (3.1.1.1)$$

The factor of 2 is the additional cost of the sign of $\bar{\lambda}_j$.

To simplify equation 3.1.1.1, Judd and Mees argue that the repeated loglog.... terms are slowly varying and so the $-\sum \ln \delta_j$ terms dominate. Assuming that the parameters only take values within some fixed range, the exponent cost is fixed. One then has

$$\tilde{L}(\bar{\lambda}) = \sum_{j=1}^{k} \ln \frac{\gamma}{\delta_j} \quad (3.1.1.2)$$

as an approximation to equation 3.1.1.1. The factor γ is a constant related to the assumed range of the exponent.

The description length of a time series z, given a model described by the parameters λ, is

$$L(z,\bar{\lambda}) = L(z|\bar{\lambda}) + L(\bar{\lambda}) \qquad (3.1.1.3)$$

where $L(z|\bar{\lambda}) = -\ln P(z|\bar{\lambda})$ is the data code length. This code length is simply the negative log likelihood of the data under the assumed distribution. If we assume Gaussianity ($\varepsilon_t \approx N(o,\sigma^2)$) then the data code length is given by $-\ln\left(\frac{1}{(2\pi\sigma^2)^{n/2}} e^{-\varepsilon^T\varepsilon/2\sigma^2}\right)$.

Judd and Mees assumed that the optimal values of δ_j are small and $\bar{\lambda}$ will not be too far from the maximum likelihood value $\hat{\lambda}$ which optimises $L(z|\lambda)$ over λ. Furthermore

$$L(z|\bar{\lambda}) \le L(z|\hat{\lambda}) + \frac{1}{2}\delta^T Q \delta \qquad (3.1.1.4)$$

where $Q = D_{\lambda\lambda}\left(L(z|\hat{\lambda})\right)$ is the second derivative matrix corresponding to the maximum likelihood solution. From equation (3.1.1.2), (3.1.1.3) and (3.1.1.4) one gets

$$L(z,\bar{\lambda}) \le L(z|\hat{\lambda}) + \frac{1}{2}\delta^T Q \delta + k\ln\gamma - \sum_{j=1}^{k}\ln\delta_j \qquad (3.1.1.5)$$

as the approximation to be minimised. This minimisation yields

$$(Q\delta)_j = \frac{1}{\delta_j} \qquad (3.1.1.6)$$

for every j. Let $\hat{\delta}_j$ denote the values of δ_j corresponding to the solution of equation (3.1.1.6), then as an approximation to the description length of a given model we have

$$L\left(z\middle|\hat{\lambda}\right)+\left(\frac{1}{2}+\ln\gamma\right)k-\sum_{j=1}^{k}\ln\hat{\delta}_j\,. \qquad (3.1.1.7)$$

Note that $L\left(z\middle|\hat{\lambda}\right)$ is the description length of the model prediction errors and will decrease with increasing model size. The last two terms of equation (3.1.1.7) are the description length of the model parameter and will increase with model size.

3.1.2. Pseudo-linear models and subset selection

Pseudo-linear models are models, which are linear combinations of linear and nonlinear functions, and have the general form:

$$f(v)=\sum_{j=1}^{m}\lambda_j\phi_j(v) \qquad (3.1.2.1)$$

for some arbitrary scalar functions ϕ_i (called basis functions), and embedding $v_t=(y_{t-\tau}, y_{t-2\tau},\ldots, y_{t-d_e\tau})\ \forall t > d_e\tau$. As mentioned before, radial basis functions are used for the functions ϕ_i in the software program of UWA. The equation (3.1.2.1) can be rewritten as

$$f(v)=\sum_{j=1}^{m}\lambda_j\phi\left(\left\|v-c_j\right\|\right) \qquad (3.1.2.2)$$

where $\phi: R^{+}\ \alpha\ R$ is a fixed function. Judd and Mees use a slight generalisation of the functional form described above by including an additional scaling factor, r_j:

$$f(v)=\sum_{j=1}^{m}\lambda_j\phi\left(\frac{\left\|v-c_j\right\|}{r_j}\right), \qquad r_j>0. \qquad (3.1.2.3)$$

In general the selection of the parameters λ_j, c_j (the centres of the jth basis function) and r_j (the scaling factor of the basis function) is a complex nonlinear optimisation problem. However, these parameters can be selected to minimise the mean squared prediction error of the model G. In the software program the scaling factors r_j are selected randomly. For the centres c_j, a large set of

what Judd and Mees call "chaperons" [16], are selected. Chaperons are randomly selected datapoints, or a subset thereof, that lie outside the region occupied by the data, therefore enlarging the convex hull of the data. It is argued, and has been found by experiment, that it is often advantageous to use centers somewhat outside the region occupied by the data. The other proposed methods for selecting centres, e.g. local centroids, k-means or bounding box methods, all select centres that lie within the convex hull of the data, and often within a region less than the convex hull. For this reason Judd and Mees use the random selection method.

The selection problem is to choose the smallest subset of radial basis functions that adequately models the time series, i.e. select which components of λ should be nonzero. The following algorithm can solve this problem. Define:

$y = (y_{d_e\tau+1},\ldots, y_n)^T$,

$\lambda = (\lambda_1,\ldots,\lambda_m)$,

V is a n' x m matrix with i^{th} column $v_i = (\phi_i(v_{d_e\tau+1}),\ldots,\phi_i(v_n))^T$ and $n' = n - d_e\tau$,

$|\cdot|$ is a norm chosen to suit the assumed distribution of the errors ε_t, and

$\mathrm{N}(\lambda)$ is the number of non-zero components of λ.

Then, for $k = 0,1,\ldots$

$$\text{minimise } |y - V\lambda| \text{ subject to } \mathrm{N}(\lambda) = k. \qquad (3.1.2.4)$$

Continue to increase k until there is no significant improvement in the model according to the description length criterion.

Judd and Mees employ a modified quadratic programming technique that uses sensitivity analysis to solve equation (3.1.2.4) for a succession of values of k. If the noise process is i.i.d Gaussian, the maximum likelihood problem reduces to a least squares fitting, so that the restricted selection problem involves choosing the optimal subset of size k at each stage so as to minimise $(y - V\lambda)^T (y - V\lambda)$.

The problem can then be written as:

Minimise $\frac{1}{2}e^T e$

over e, λ

subject to $V\lambda + e = y, \mathrm{N}(\lambda) = k$.

Using sensitivity analysis, Judd and Mees derive the following algorithm for selecting the k non-zero λ_i, $i = 1,2,\ldots,k$.

Let B be the set of basic indices, namely $B = \{j : \lambda_j \neq 0\}$ so $\mathrm{N}(\lambda) = |B|$, $\mathrm{V_B}$ the n' x k matrix formed from the columns of V with indices in B, let λ_B be the least squares approximation to $y = \mathrm{V_B}\lambda$ and let $e_B = y - \mathrm{V_B}\lambda_B$.

With the above assumptions, Judd and Mees illustrated in [16] that the description length is bounded by

$$\left(\frac{\check{N}}{2} - 1\right)\ln\frac{e^T e}{\check{N}} + (k+1)\left(\frac{1}{2} + \ln\gamma\right) - \sum_{j=1}^{k}\ln\delta_j + C,$$

where γ is related to the scale of the data, $\check{N} = n - d_e\tau$ is the number of embedded vectors, and C is a constant independent of the model parameters.

The model selection algorithm used by Judd and Mees and implemented in the software program, is summarised below.

Algorithm 3.1: Model Selection Algorithm.

1. Normalise the columns of V to have unit length.
2. Let $S_0 = (\frac{\check{N}}{2} - 1)\ln(\frac{y^T y}{\check{N}}) + \frac{1}{2} + \ln\gamma$. Let $e_B = y$ and $B = \varnothing$.
3. Let $\mu = V^T e_B$ and j be the index of the component of μ with maximum absolute value. Let $B' = B \cup \{j\}$.

4. Calculate $\lambda_{B'}$ so that $y - V_{B'}\lambda_{B'}$ is minimised. Let $\mu' = V^T e_{B'}$. Let o be the index in B' corresponding to the component of u' with the smallest absolute value.
5. If $o \neq j$, then set $B = B' \setminus \{o\}$, calculate λ_B so that $y - V_B\lambda_B$ is minimised, let $e_B = y - V_B\lambda_B$, and go to step 3, else go to step 6.
6. Define $B_k = B$, where $k = |B|$. Find δ such that $\left(V_B^T V_B \delta\right)_j = \frac{1}{\delta_j}$ for each $j =$ $\{1,2,\ldots,k\}$ and calculate $S_k = \left(\frac{\overset{\vee}{N}}{2} - 1\right)\ln\frac{e_B^T e_B}{\overset{\vee}{N}} + (k+1)(\frac{1}{2} + \ln\gamma) - \sum_{j=1}^{k}\ln\hat{\delta}_j$.
7. If some stopping condition has not been met, then go to step 3.
8. Take the basis B_k such that S_k is a minimum as the optimal model.

Typically one will continue increasing k until it is clear that the minimum of S_k has been reached. Usually S_k decreases rapidly as k increases, then slowly increases. Around the minimum the value of S_k usually fluctuates. To allow for this fluctuation Judd and Mees suggest not to terminate the algorithm immediately when $S_k < S_{k+1}$, but to wait until $S_k < S_{k+l}$ for $l = 1,\ldots,p$, the number p being arbitrary.

3.2 Surrogate Analysis

We use the method of surrogate analysis to find evidence of any nonlinear structure in a time series. If evidence is found, we determine whether the model prediction error, ε_t, resulting from algorithm 3.1, belongs to the dynamic class described as i.i.d noise.

Before we attempt any nonlinear modelling techniques on a time series, it is necessary to ascertain whether the apparent structure of the time series is due to nonlinear dynamics within the system that the time series represents, or to linear correlations. Only nonlinear deterministic signals are characterised by finite-dimensional attractors, finite positive Lyapunov exponents and good nonlinear short-term predictability. Estimated finite dimensions and Lyapunov exponents alone are not suitable for proving nonlinear determinism, as linear structures can be mistaken for determinism due to linear correlations. For this reason we use the method of surrogate analysis to test for nonlinear determinism.

Hypothesis testing [42] forms the basis of the Method of Surrogate Data. We therefore need a null hypothesis, and the expected rate of Type I error (α), which will determine the sample size, i.e. the number of surrogate data sets to generate, and a discriminating statistic, T.

Recall from section 2.2 that Theiler et al. [38] suggested a "hierarchy" of hypotheses that test for membership of particular classes of linear systems known as Algorithm 0 surrogates, Algorithm 1 surrogates, and Algorithm 2 surrogates. In order to ascertain whether the apparent structure of the time series is due to nonlinear dynamics, we test the time series for membership of each of the three classes. The first hypothesis is that the time series belongs to the class of dynamic systems described by linearly filtered noise (Algorithm 0 surrogates). The second hypothesis is that the time series belongs to the dynamic class described as a linear transformation of linearly filtered noise (Algorithm 1 surrogates), and the third hypothesis is that the time series belongs to the class of dynamic systems described as a monotonic, nonlinear transformation of linearly filtered noise (Algorithm 2 surrogates). If all three of these hypotheses are rejected, i.e. we cannot assume that the time series is described by any of the modelled dynamic systems, we then have reason to suspect that the time series might belong to a nonlinear, possibly chaotic system.

In considering the hypothesis that the model prediction error, ε_t, belongs to the dynamic class described as identically and independently distributed (i.i.d) noise, recall that Judd and Mees [16, 21] consider models where there is a random process generating $(X_t, Y_t) \in R \times R^d, t \in Z$, for which there is a relationship

$$Y_t = F(X_t) + \varepsilon_t,$$

where the random variables ε_t are independent of Y_γ and X_s, for $\gamma \geq t$ and $s \geq t$. Their method renders a strong approximation $\hat{F}$ of F where ε_t is identically and independently distributed with finite variance. This study used the Method of Surrogate Data to test the hypothesis that ε_t is i.i.d noise. If the strong approximation $\hat{F}$ of F does model the dynamics of the underlying system, then ε_t should belong to the dynamic class described as i.i.d noise, which is described by Theiler et al. [40] as Algorithm 0 surrogates. Their method for generating constrained realisations of Algorithm 0 surrogates (algorithm 2.1) was used to test membership of this dynamic class. Therefore, in this instance the null hypothesis is that ε_t belongs to the dynamic class described as i.i.d noise, which is modelled by Algorithm 0 surrogates.

The hypotheses tests are carried out as follows. A null hypothesis is defined for a particular time series, x_t, i.e. that the time series belongs to the dynamic class described by Algorithm 0,1 or 2 surrogates. The value of a discriminating statistic t_0 for the time series x_t is then calculated. Next, the value of the discriminating statistic t_k for each of the $k = 1,\ldots,B$ surrogate data sets is calculated. For a two-sided test, the null hypothesis is rejected at the level α, if t_0 is observed among the largest $(B+1)\alpha/2$, or the smallest $(B+1)\alpha/2$ in the sorted list that includes t_0 as well as $t_1,\ldots,t_B$. The question remains as to how many of these surrogate data sets (value of "B") to generate.

A "Type I error" exists when a test rejects the null hypothesis when in fact, the null hypothesis is true. The sample size, B, is usually taken as large as $\frac{2}{\alpha}$. With $\alpha = 0.05$, the sample size should be a multiple of 40. In this study a sample size of 40 was used which means that for every hypothesis, 40 surrogate data sets were generated. We aimed for $\alpha = 0.05$ by rejecting the null hypothesis when the orginal time series, x_t, had a discriminating statistic which is either larger, or smaller, than all the surrogates' discriminating statistic[1]. If we aimed for a $\alpha = 0.33$ (67% confidence level), the sample size should have been a multiple of 7. Choosing a sample size, B, of 35 in this instance, would have lead to the following rejection criteria: if discriminating statistic was observed among the largest 6[2], or the smallest 6 in the sorted list that includes the time series' discriminating statistic as well as the surrogates' discriminating statistic, the null hypothesis would have been rejected.

The only outstanding issue is the choice and calculation of the discriminating statistic, T.

3.2.1 Correlation Dimension

As discussed in Section 2.1, the discriminating statistics can be classified into two groups: pivotal and non-pivotal. A statistic is "pivotal" if its distribution is the same for all processes consistent with the null hypothesis; otherwise it is "non-pivotal". Theiler and Prichard [39] showed that when using the constrained-realisation approach to generate surrogate data sets, it is not required that the discriminating statistic be pivotal. This allows one more flexibility in the choice of a

[1] With $B = 40$ and $\alpha = 0.05$, $(B+1)\alpha/2 \cong 1$.
[2] With $B = 35$ and $\alpha = 0.33$, $(B+1)\alpha/2 \cong 6$.

discriminating statistic. In this study the constrained-realisation approach is used to generate the surrogate data sets, and therefore it is not necessary that a pivotal discriminating statistic is used.

Correlation dimension was chosen as the test statistic due to its great significance in nonlinear time series analysis and its topical interest [4, 8, 9, 10, 13, 14, 15, 32, 40]. Correlation dimension is a generalisation of the concept of integer dimension for fractal objects with non-integer dimension. In one, two or three dimensions it is intuitive that a measure of volume V (length, area or volume) varies as

$$V \propto \varepsilon^{d}, \qquad (3.2.1.1)$$

where ε is a length scale and d is the dimension of the object. For a general fractal we can assume a relation similar to equation (3.2.1.1) holds true, in which case its dimension is given by,

$$d \approx \frac{\log V}{\log \varepsilon}.$$

Let $\{v_t\}_{t=1}^{N}$ be an embedding of a time series in R^{d_e}. Define the correlation function $C_N(\varepsilon)$ by

$$C_N(\varepsilon) = \binom{N}{2}^{-1} \sum_{0 \le i < j \le N} I\left(\left\| v_i - v_j \right\| < \varepsilon\right).$$

Here $I(X)$ is a function whose value is 1 if condition X is satisfied, and 0 otherwise, and $\|\cdot\|$ is the usual distance function in R^{d_e}. The sum $\sum_i I\left(\left\| v_i - v_j \right\| < \varepsilon\right)$ is the number of points within a distance ε of v_j. If the points v_i are distributed uniformly within an object, then this sum is proportional to the volume of the intersection of a sphere of radius ε with the object, and $C_N(\varepsilon)$ is proportional to the average of such volumes. Therefore one expects that

$$C_N(\varepsilon) \propto \varepsilon^{d_c},$$

where d_c is the dimension of the object. The correlation integral is defined as $\lim_{N\to\infty} C_N(\varepsilon)$. Define the correlation dimension d_c by

$$d_c = \lim_{\varepsilon \to 0} \lim_{N \to \infty} \frac{\log C_N(\varepsilon)}{\log \varepsilon}.$$

The curious normalisation of $C_N(\varepsilon)$ is chosen so that, rather than $C_N(\varepsilon)$ being an estimate of the expected number of points of an object within a radius ε of a point, it is an estimate of the probability that two points chosen at random on the object are within a distance ε of each other. The reason for choosing the probability rather than the expectation, is that the concept of dimension still makes sense. It generalises situations where the sample points v_i are not distributed uniformly within the object. For a more detailed discussion of the general situation, the reader is referred to [15].

The method most often employed to estimate the correlation dimension is the Grassberger-Procaccia algorithm [13, 14]. In this method one calculates the correlation function and plots $\log C_N(\varepsilon)$ against $\log \varepsilon$. The gradient of this graph in the limit as $\varepsilon \to 0$ should approach the correlation dimension. Among the pitfalls of this method is that when using a finite amount of data, the graph will jump about irregularly for small values of ε. To avoid this, one should look at the behaviour of this graph for moderately small ε. A typical correlation integral plot will contain a "scaling region" over which the slope of $\log C_N(\varepsilon)$ remains relatively constant. A common method to examine the slope in the scaling region is to numerically differentiate the plot of $\log \varepsilon$ against $\log C_N(\varepsilon)$. This ought to produce a function which is constant over the scaling region, and its value on this region should be the correlation dimension.

Judd [15] points out that there are several problems with this procedure, the most obvious of which is that the choice of the scaling region is entirely subjective. For many data sets a slight change in the region used can lead to substantially different results. Judd demonstrates that for many objects, including many fractals, a better description of $C_N(\varepsilon)$ is that for ε less than some ε_0

$$C_N(\varepsilon) \approx \varepsilon^{d_c} q(\varepsilon),$$

where $q(\varepsilon)$ is a polynomial of order t, the topological dimension of the attractor, taking variations of the slope within and outside of a scaling region, into account.

The Grassberger-Procaccia method assumes $C_N(\varepsilon) \propto \varepsilon^{d_c}$, but allows for the presence of a further polynomial term that takes into account variations of the slope within and outside of a scaling region. Judd's method dispenses with the need for a scaling region and substitutes a single scale parameter ε_0, which has an interesting benefit. For many natural objects the dimension is not the same at all length scales. Consequently, it is appropriate to consider dimension d_c as a function of ε_0 and write $d_c(\varepsilon_0)$.

By allowing the dimension to be a function of scale, one produces estimates that are both more accurate and informative. Some of the approximation necessary to define correlation dimension as a single number, is avoided, and one can extract more detailed information about the changes in dimension with scale.

There are several acceptable methods to compare correlation dimension curves. Due to the definition of correlation dimension as a function of ε_0, one can compare the value of dimension for some fixed values of ε_0. This is the method used in this study. Other possibilities include the mean value of the dimension estimate, or the slope of the line of best fit. More sophisticated methods include statistical testing such as the χ^2-test, or the Kolmogorov-Smirnov statistic applied to the distribution of inter-point distances to determine whether the distributions are the same. Details of these tests can be found in most introductory statistics texts, as in [5].

3.3 Experimental design

The modelling algorithm and the application of the Method of Surrogate Data used in this study have been discussed, but no attention has been given to the particular time series used. Although the focus is nonlinear modelling, the Method of Surrogate Data and its analysis, needs a time series to which to apply these techniques. Three time series were selected from the data library of the Chaos Data Analyzer software package [33].

We selected the time series generated by the Lorenz equations (figure 3.1) because it serves as a prototypical example of a dissipative chaotic flow. We considered it best to apply our algorithm and hypotheses tests to a time series with known properties initially, before applying it to other time series, where ambiguities regarding certain of the time series' properties do exist.

Secondly, we selected the time series representing an electroencephalogram (figure 3.2) primarily due to research done by Theiler and Rapp [37, 40], using surrogate analysis to re-examine the evidence for low-dimensional, nonlinear structure in the human electroencephalogram. Furthermore, electroencephalographic signals are the object of many studies [8, 25, 37, 40] researching evidence of nonlinear, low-dimensional structure in the signals. Depending on the method and discriminator used in the specific study, the results can vary substantially. We deemed it useful to analyse the available electroencephalographic signal, expecting results comparable to that of Theiler and Rapp [40], given the similar methodologies applied.

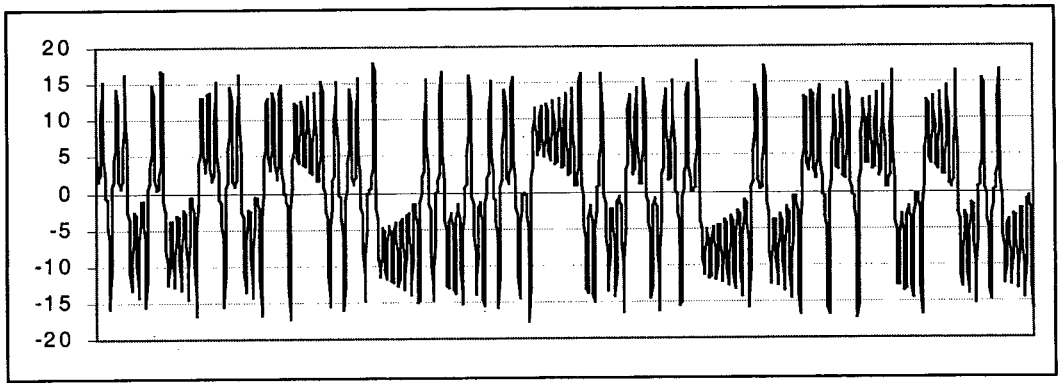

Figure 3.1 **Time series generated by the Lorenz equations**: A file of 2000 data points for $X(t)$ at intervals of Δt = 0.05 for the Lorenz attractor $dX/dt = 10(Y - X), dY/dt = 28X - Y - XZ, dZ/dt = XY - 8Z/3$.

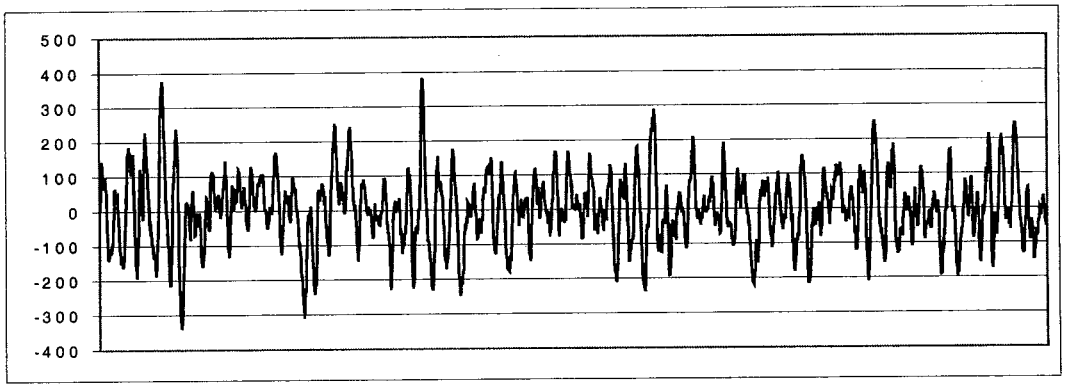

Figure 3.2 **EEG time series:** A file of 2,000 data points representing 8 seconds of an electroencephalogram from a 19-year old, healthy male subject, seated and relaxed with eyes closed, sampled at 250 Hz with electrodes on the left front of the head. Units are 0.1 microvolts.

As our third time series, we selected Standard & Poor's Composite Index of 500 stocks (figure 3.3). In recent years, nonlinear systems have received a great deal of attention in the financial

community [2, 6, 9, 22, 41]. The potential of nonlinear modelling techniques in demystifying seemingly random price and interest rate movements makes it particularly intriguing to those (like us) involved in forecasting the markets. We therefore had an active interest in the analysis of the S&P500 time series.

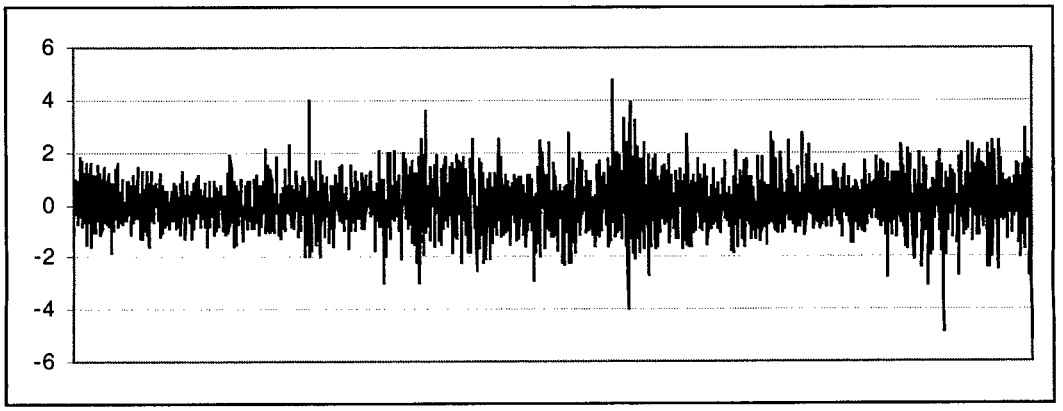

Figure 3.3 **S&P500 time series:** A file of 3,000 values of daily percentage change in the Standard & Poor's Composite Index of 500 Stocks leading up to the 20% drop that occurred on October 19, 1987.

The experimental design comprised of two phases corresponding to the two hypotheses. The first phase dealt with the hypothesis that the apparent structure of the particular time series is most likely due to nonlinearity. The second phase dealt with the hypothesis that the model prediction error, ε_t, belongs to the dynamic class described i.i.d noise, after selecting the best model according to the minimum description length criterion.

To test the hypothesis that the apparent structure of the particular time series is most likely due to nonlinearity, we developed software that read the time series from disk, and generated surrogates from the time series. Algorithm 0, 1, and 2 surrogates were generated for the three time series and the results were analysed to ascertain whether there is reason to reject the different null hypotheses. Recall that Algorithm 0 surrogates (Algorithm 2.1) test the hypothesis that the time series belongs to the dynamic class described as linearly filtered noise. Algorithm 1 surrogates (Algorithm 2.2) test the hypothesis that the time series belongs to the dynamic class described as a linear transformation of linearly filtered noise. Finally, Algorithm 2 surrogates (Algorithm 2.3) test the hypothesis that the time series belongs to the dynamic class described as a monotonic, nonlinear transformation of linearly filtered noise.

Correlation dimension was used as the test statistic for the hypotheses testing. Sophisticated software developed by the Mathematics Department of the University of Western Australia (UWA) and based on Judd's method [15] (see Section 3.2.1), was used to calculate the correlation dimension of the time series and their surrogates. The results were written to text files for analysis and hypothesis testing.

The time lag, τ and the embedding dimension, d_e needed to be specified as input parameters for the correlation dimension software program. The method of Average Mutual Information (AMI) was used to give an indication of the optimal time lag, and the method of False Nearest Neighbours (FNN) was used to calculate the embedding dimension.

A typical experimental run for the surrogate analysis of the time series comprised of the following steps:

1. Read the time series from disk.
2. Calculate the time lag, τ and the embedding dimension, d_e.
3. Generate Algorithm 0, 1 and 2 surrogates of the time series.
4. Calculate the correlation dimension, d_c, of the time series and the surrogates, using the parameters calculated in step 2.
5. Perform the surrogate analysis based on correlation dimension as the test statistic, T, and accept or reject the specific null hypothesis.

Depending on the surrogate analyses of the time series, the time series was modelled using Judd and Mees's algorithm [16, 21] (see Section 3.1), and the model prediction error, ε_t, was tested for membership of the dynamic class described as i.i.d noise. The reader is reminded that Judd and Mees's algorithm (algorithm 3.1) is an iterative process that terminates once the stop criterion is met. Every iteration of the program (i.e. modelling algorithm) strives to build a "better" model, modelling more of the dynamics of the underlying system and less of the noise. This should lead to the residual vectors, e, bearing a closer resemblance to i.i.d noise with every improved model. It is these residual vectors, e, that are critical to the study. Residual vectors are the model prediction errors of the various models that are generated with the iterations of the modelling algorithm. We refrain from using the term "model prediction errors", as it is generally used for the difference between a time series and the final model of an iterative modelling algorithm, i.e. a model that fits the time series relatively well. It was necessary to make a few changes to one of the software modules so that these residual vectors were written to disk for analyses.

A typical experimental run for these hypothesis tests comprised of the following steps:

1. Read data file from disk.
2. Generate residual vectors by modelling the time series using UWA's software program.
3. For every residual vector:
   - Calculate the time lag, τ and embedding dimension, d_e.
   - Generate Algorithm 0 surrogates from the residual vector.
   - Calculate the correlation dimension, d_c of the residual vector and the surrogates.
   - Perform the surrogate analyses based on correlation dimension as the test statistic, T, and accept or reject the null hypotheses.

The software program used to model the time series needs certain parameters to be specified other than the time series. The parameters are as follows:

- z: A 1-dimensional time series from which to build a model.
- 7v: Matrix with variable embedding strategies.
- z_test: A 1-dimensional time series for testing the model.
- fac: A factor that determines the number of iterations.
- r: The maximum radius with default $2\times\sigma(z)$ which is twice the standard deviation of the time series.
- n: The maximum number of functions to be used.
- f: The functions to be used for model building, in this case gaussian radial basis functions.

The primary objective of this study was to examine the Method of Surrogate Data and a few nonlinear modelling techniques by applying this method and techniques to the selected time series, rather than building an optimal model of the time series. However, the model generated by the software program, was used for one-step and free-run predictions of a "test" time series (z_test). One-step predictions forecast the time series for one time step, and then use the actual value of the time series for the forecasted time step, to forecast the next value of the series. Free-run predictions use the forecasted values of the time series at every time step, to forecast the time series, i.e. the forecasted time series is not adjusted by the actual time series values.

We used the method of surrogate data in the more conventional application to test for nonlinear, low-dimensional structure in a time series. In a less conventional application, we used it to verify, rather than determine, that the final model (using the specific modelling algorithm) does in fact capture the dynamics of the underlying system, by studying the residual vectors of the different models generated by every iteration of the modelling algorithm. The concept of applying the Method of Surrogate Data to test a model fit is novel and appears to have considerable promise.

CHAPTER 4

RESULTS

In this chapter we discuss the results of the surrogate analyses of the different time series. In the case where the initial surrogate analysis of a specific time series failed, i.e. the null hypotheses was rejected, indicating the possibility that the apparent structure of the time series might be due to nonlinear dynamics, the time series was modelled using algorithm 3.1, based on Judd and Mees's modelling methodology. The model of the particular time series is not the object of this study and consequently, is not discussed further. Of more concern are the model's residual vectors which are analysed using the Method of Surrogate Data. Not every residual vector generated by the modelling algorithm will be discussed due to the substantial amount of residual vectors generated in most cases. A few of the models' residual vectors generated by the modelling algorithm for a specific time series, were selected to illustrate our theory regarding the models' residual vectors and the final model prediction error. In Section 4.1 the Lorenz time series is discussed, in Section 4.2 the EEG time series and, in Section 4.3 the S&P500 time series.

4.1 Lorenz time series

The Lorenz time series that is the subject of this analysis, was generated by the following three equations (unfortunately the initial conditions are not supplied in CDA[33]):

$$dX/dt = 10(Y - X),$$
$$dY/dt = 28X - Y - XZ,$$
$$dZ/dt = XY - 8Z/3.$$

The reader is referred to figure 3.1 for a graphic representation of the time series.

4.1.1 Surrogate Analysis of Lorenz time series

As mentioned in Section 3.3, we used the method of Average Mutual Information (AMI) and False Nearest Neighbours (FNN) to obtain an indication of the optimal time lag, τ, and the embedding dimension, d_e, respectively. Figure 4.1.1 and 4.1.2 are graphic representations of the results of these two methods.

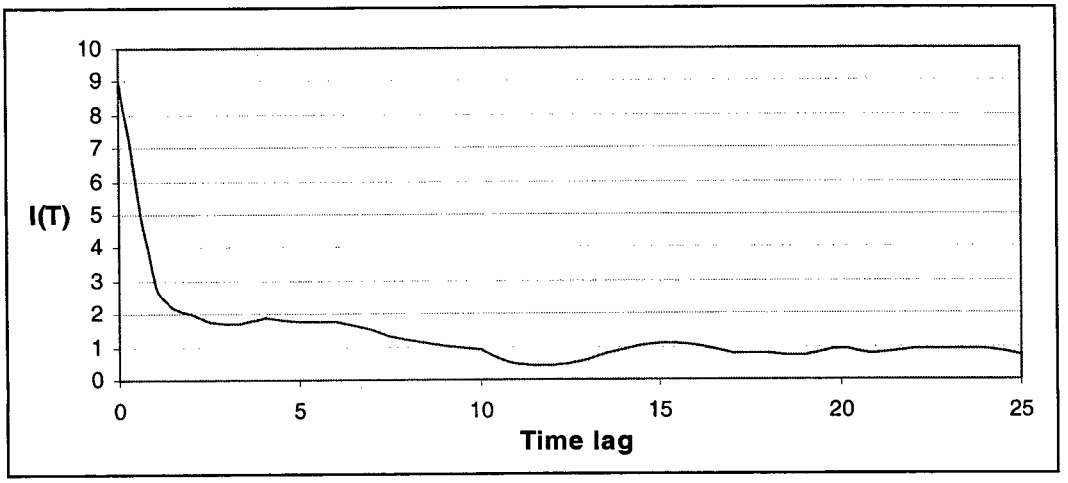

Figure 4.1.1 Average Mutual Information for Lorenz time series: The first minimum occurs at a time lag value of 3. According to the criterion suggested by Fraser and Swinney [11], we therefore choose the value for the time lag, τ, as 3.

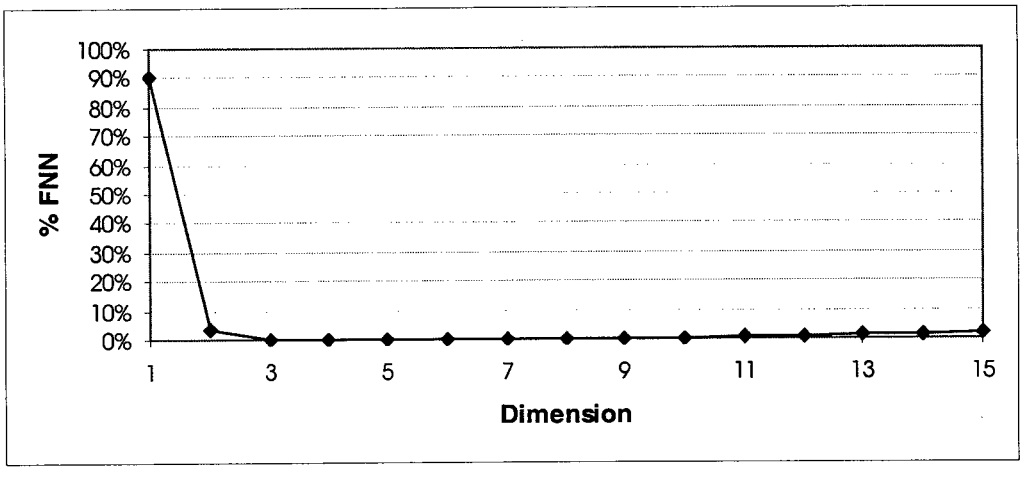

Figure 4.1.2 False Nearest Neighbours for Lorenz time series ($\tau = 3$): The percentage of false nearest neighbours fall to 0 when the embedding dimension is 3, which leads us to choose 3 as the embedding dimension, d_e.

Having decided on values for the time lag, τ, and the embedding dimension, d_e, we proceeded to test the time series for membership of the three dynamic systems described by each of the null hypotheses of Algorithm 0, 1 and 2 surrogates, respectively. Correlation dimension was used as the discriminating statistic in all three the tests.

Figure 4.1.3 is a graphic summary of the result of the first hypothesis test, testing the time series for membership of the dynamic class described as linearly filtered noise. The correlation

dimension curve of the time series can be distinguished from the correlation dimension curves of the surrogates, indicating that the correlation dimension values of the time series are smaller than the correlation dimension values of the surrogates, for every scale parameter value. Based on this evidence, the null hypothesis that the Lorenz time series belongs to the dynamic class modelled by Algorithm 0 surrogates, was rejected with a 95% confidence level. A numeric comparison of this result is summarised in table 4.1.4

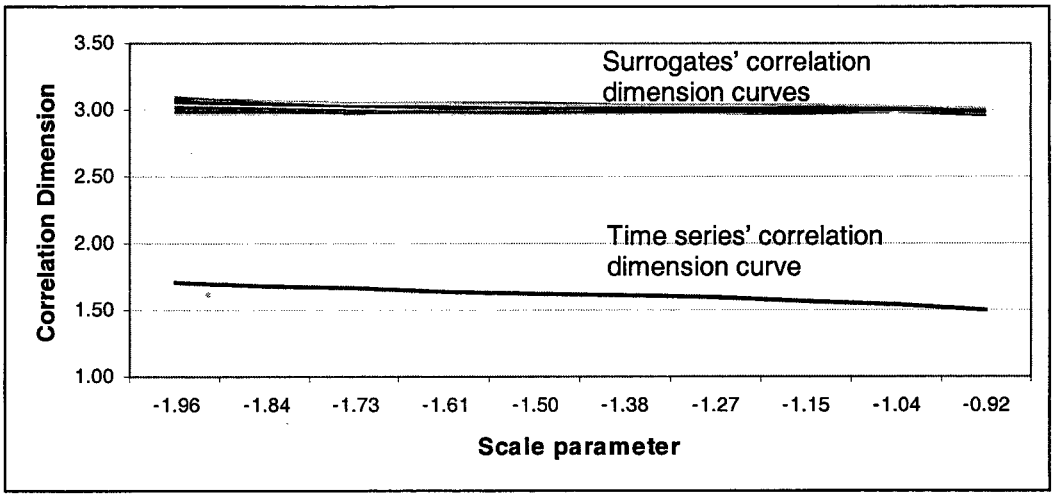

Figure 4.1.3 Correlation dimension curves of the Lorenz time series and Algorithm 0 surrogates.

Scale Parameter	Lorenz time series' Correlation Dimension	Surrogates' Minimum Correlation Dimension	Surrogates' Maximum Correlation Dimension
-1.9572	1.7089	2.9694	3.1007
-1.8421	1.6823	2.9733	3.0784
-1.7269	1.6601	2.9769	3.0723
-1.6118	1.6422	2.9770	3.0671
-1.4967	1.6290	2.9706	3.0608
-1.3816	1.6134	2.9768	3.0508
-1.2664	1.5933	2.9742	3.0535
-1.1513	1.5659	2.9702	3.0504
-1.0362	1.5347	2.9800	3.0445
-0.9210	1.4997	2.9611	3.0200

Table 4.1.4 Summary of the analysis for Algorithm 0 surrogates of the Lorenz time series.

A similar conclusion can be made from the analysis of Algorithm 1 surrogates (Algorithm 2.2), testing whether the time series belongs to the dynamic class described as a linear transformation of linearly filtered noise. The result of the surrogate analysis is graphically represented by figure 4.1.5. For the same reason mentioned in the discussion of the Algorithm 0 surrogate analysis, the

null hypothesis of the Algorithm 1 surrogates was rejected with a 95% confidence level. The correlation dimension values of the time series are smaller than the correlation dimension values of the surrogates, for every scale parameter value. The numeric evidence for this observation is summarised in table 4.1.6. We cannot therefore, assume that the time series belongs to the dynamic class described as a linear transformation of linearly filtered noise.

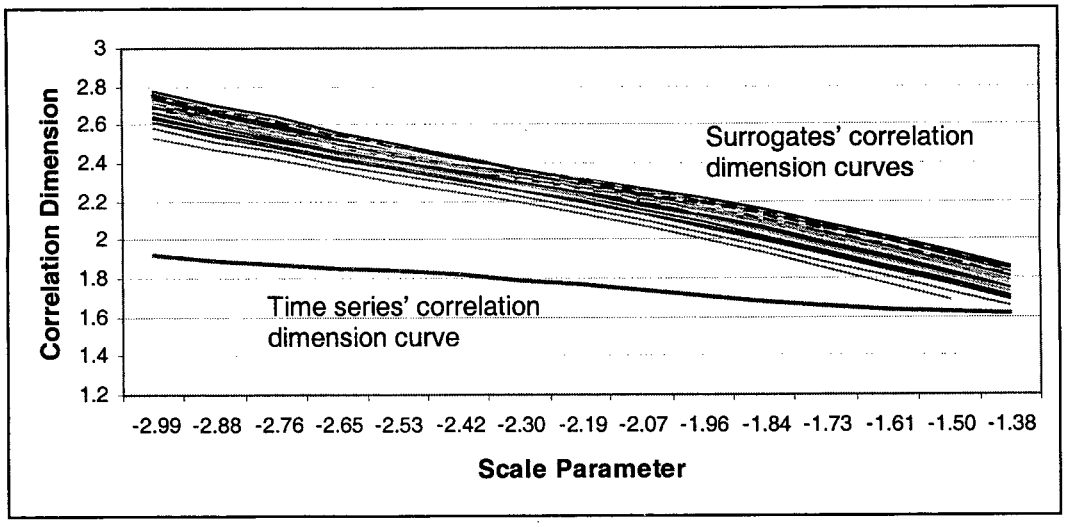

Figure 4.1.5 Correlation dimension curves of the Lorenz time series and Algorithm 1 surrogates.

Scale Parameter	Lorenz time series' Correlation Dimension	Surrogates' Minimum Correlation Dimension	Surrogates' Maximum Correlation Dimension
-2.9934	1.9196	2.5352	2.7761
-2.8782	1.8891	2.4736	2.7046
-2.7631	1.8683	2.4189	2.6477
-2.6480	1.8542	2.3545	2.5656
-2.5328	1.8395	2.3021	2.5033
-2.4177	1.8204	2.2460	2.4372
-2.3026	1.7946	2.1932	2.3742
-2.1875	1.7658	2.1308	2.3175
-2.0723	1.7365	2.0713	2.2636
-1.9572	1.7089	2.0021	2.2134
-1.8421	1.6823	1.9309	2.1535
-1.7269	1.6601	1.8519	2.0892
-1.6118	1.6422	1.7723	2.0171
-1.4967	1.6290	1.6926	1.9411

Table 4.1.6 Summary of the analysis for Algorithm 1 surrogates of the Lorenz time series.

The third hypothesis tests the time series for membership of the dynamic class described as a monotonic, nonlinear transformation of linearly filtered noise. Algorithm 2 surrogates

(Algorithm 2.3) model the mentioned dynamic class. Figure 4.1.7 is a graphic representation on the result of this hypothesis test. Again, the correlation dimension curve of the time series can be distinguished from the correlation dimension curves of the surrogates. Table 4.1.8 compares the correlation dimension values of the time series to the correlation dimension values of the surrogates, for every scale parameter value. The correlation dimension value of the time series is smaller than the correlation dimension values of the surrogates, for every scale parameter value, and the null hypothesis, that the Lorenz time series belongs to the dynamic class modelled by Algorithm 2 surrogates, was rejected with a 95% confidence level.

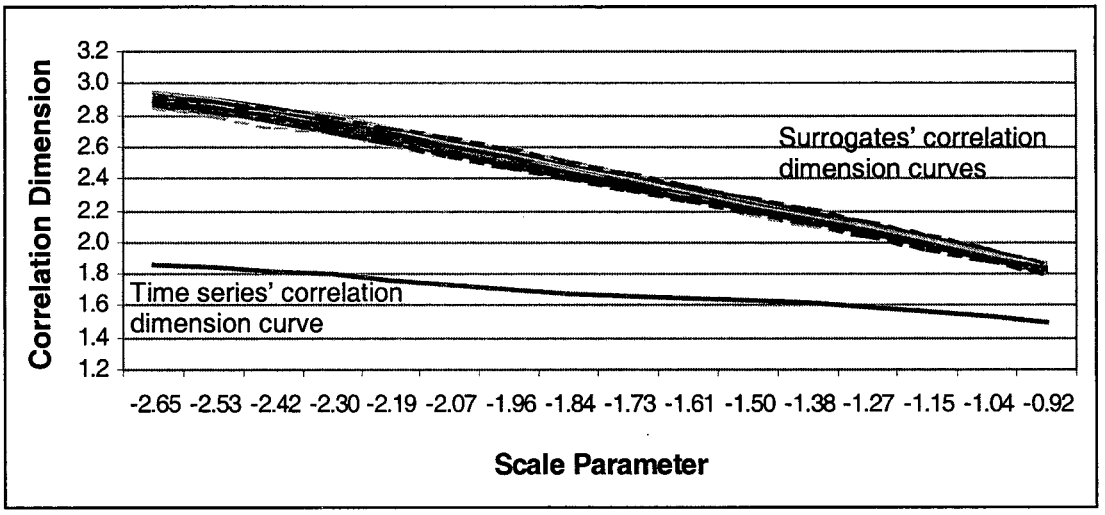

Figure 4.1.7 Correlation dimension curves of the Lorenz time series and Algorithm 2 surrogates.

The surrogate analyses failed for the Lorenz time series, i.e. we rejected all three null hypotheses with a 95% confidence level, using correlation dimension as the discriminating statistic. Therefore, it is unlikely that the time series was generated by a monotonic, nonlinear transformation of linearly filtered noise, and we have reason to suspect that the time series was generated by a nonlinear, possibly chaotic system. This confirms our initial "assumption", that the Lorenz time series was generated by a system exhibiting dissipative, chaotic flow. We did not expect any other result for the surrogate analyses, as the properties of the system are known.

Scale Parameter	Lorenz time series' Correlation Dimension	Surrogates' Minimum Correlation Dimension	Surrogates' Maximum Correlation Dimension
-2.6480	1.8542	2.8449	2.9483
-2.5328	1.8395	2.7948	2.9030
-2.4177	1.8204	2.7301	2.8493
-2.3026	1.7946	2.6866	2.8022
-2.1875	1.7658	2.6154	2.7295
-2.0723	1.7365	2.5389	2.6636
-1.9572	1.7089	2.4699	2.6004
-1.8421	1.6823	2.3970	2.5214
-1.7269	1.6601	2.3280	2.4415
-1.6118	1.6422	2.2594	2.3650
-1.4967	1.6290	2.1850	2.2913
-1.3816	1.6134	2.1114	2.2156
-1.2664	1.5933	2.0356	2.1366
-1.1513	1.5659	1.9591	2.0512
-1.0362	1.5347	1.8792	1.9647
-0.9210	1.4997	1.7975	1.8758

Table 4.1.8 Summary of the analysis for Algorithm 2 surrogates of the Lorenz time series.

4.1.2 Modelling of Lorenz time series

In Section 3.3 we discussed the software program used to model the time series and the required parameters. Table 4.1.9 is a brief explanation and the parameters and the specific values used.

Parameter	Description	Value
z	time series	first 1700 data points
v	variable embedding strategies	Matrix with variable embedding strategies, given that $\tau = 3$ and $d_e = 3$.
z_test	time series for testing	last 300 data points
fac	number of algorithm iterations	20
r	maximum radius for radial basis functions	2 × standard deviation of the time series
n	maximum number of basis functions	75
f	type of basis functions	Gaussian radial basis functions

Table 4.1.9 Parameter values for software program used to model the Lorenz time series.

The software program that implements algorithm 3.1 generated a model with a description length of –9025. Figure 4.1.10 is a graphic representation of the one-step predictions. The free-run predictions are displayed in figure 4.1.11. As expected the model's one-step predictions predicted the time series accurately, with the out-of-sample model prediction error being zero. The free-run predictions were less accurate, with the model resulting to greater prediction errors with time.

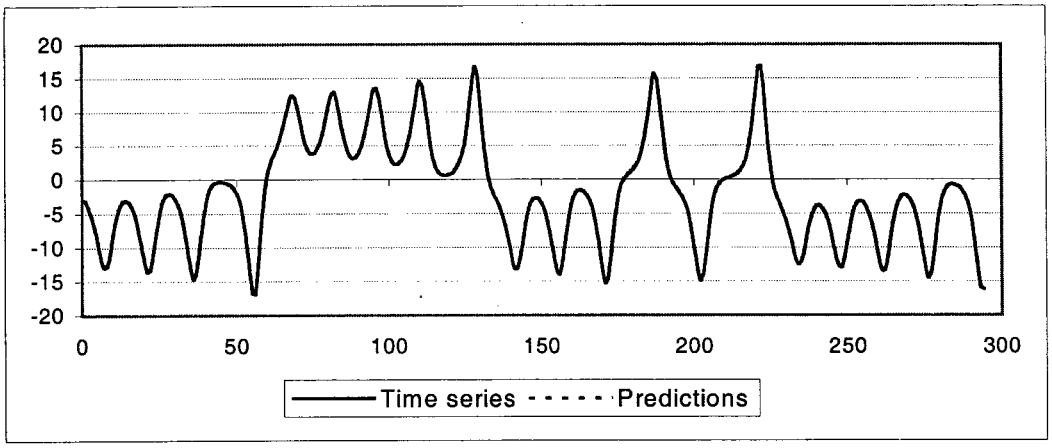

Figure 4.1.10 One-step predictions of the Lorenz time series using the model with a description length of –9025.

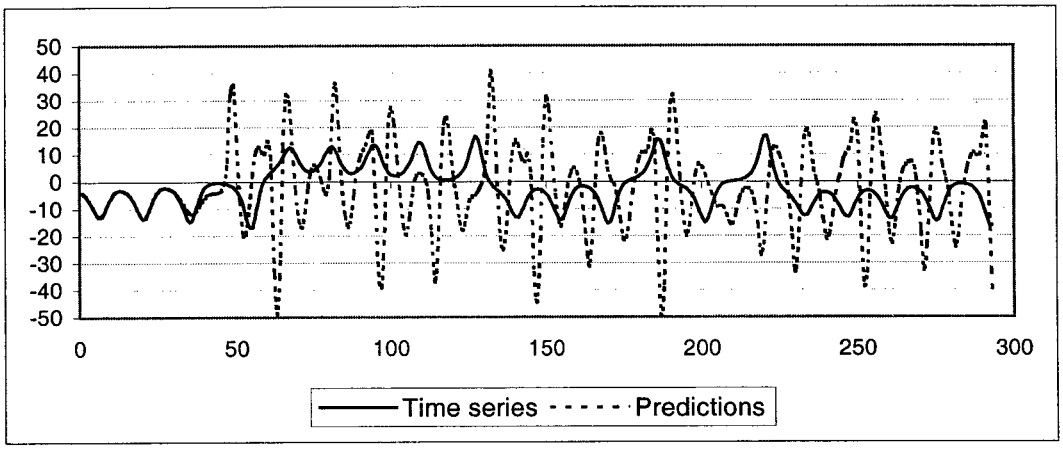

Figure 4.1.11 Free-run predictions of the Lorenz time series using the model with a description length of –9025.

Observe in figure 4.1.10 and figure 4.1.11, that the one-step predictions were more accurate than the free-run predictions. This phenomenon is typical of a chaotic system due its sensitive dependence on initial conditions.

4.1.3 Surrogate Analysis of residual vectors, *e*

For all 20 executions of the algorithm, the description lengths ranged between the values 65.66 and –9025. Three residual vectors were selected for the surrogate analyses: the first one corresponds to a model with the largest description length (65.66), i.e. the "worst" model using description length as our criterion. The second residual vector corresponds to a model with a description length of -4082, and the third residual vector corresponds to a model with the smallest description length (-9025), i.e. the "best" model according to the description length criterion. The importance of the first and third selected residual vectors for our analysis, should be evident. The second residual vector was selected as it represents an "average" model, i.e. a model that is neither the worst, nor the best model.

We again used the methods of AMI and FNN to obtain an indication of the optimal time lag, τ, and embedding dimension, d_e. Table 4.1.12 gives the values of the time lag, τ, and embedding dimension, d_e, that was used for each of the residual vectors for the surrogate analysis.

Description Length	Embedding Dimension (d_e)	Time lag (τ)
65.66	3	5
-4082	3	3
-9025	3	3

Table 4.1.12 Embedding parameters for the residual vectors belonging to the models of the Lorenz time series with corresponding description lengths.

In Section 3.2 we discussed the surrogate analyses performed in this study, explaining reasons for testing the residual vectors for membership of the dynamic class described as i.i.d noise. This dynamic class is modelled by Algorithm 0 surrogates.

The null hypothesis that the residual vector corresponding to the model with a description length of 65.66, belongs to the dynamic class modelled by Algorithm 0 surrogates, was rejected with a 95% confidence level. Figure 4.1.13 visually summarises the surrogate analysis, indicating that the correlation dimension value of the residual vector is smaller than the correlation dimension values of the surrogates, for every scale parameter value.

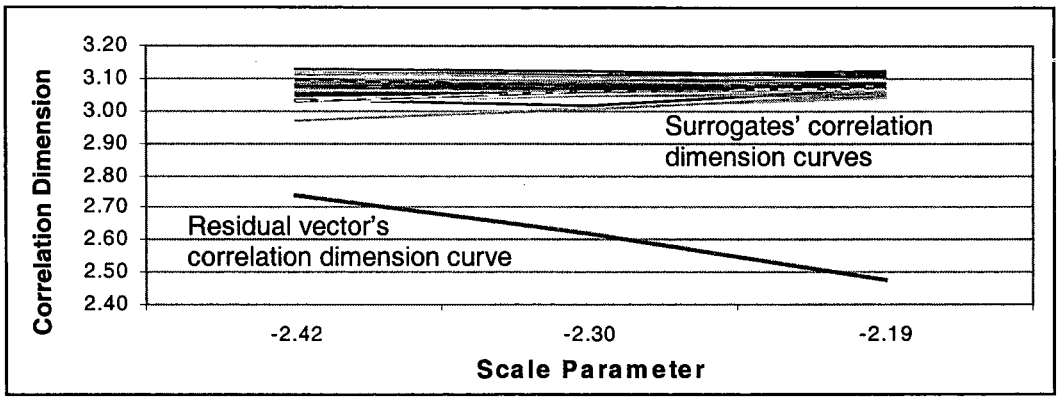

Figure 4.1.13 Correlation dimension curves of the residual vector corresponding to a Lorenz time series' model with a description length of 65.66, and Algorithm 0 surrogates.

The residual vector corresponding to the model with a description length of –4082 was also tested for membership of the dynamic class modelled by Algorithm 0 surrogates. Figure 4.1.14 indicates that the correlation dimension value of the residual vector is smaller than the correlation dimension values of the surrogates, for every scale parameter value. Based on this evidence, the null hypothesis that the residual vector, corresponding to the model with a description length of -4082, belongs to the class of dynamic systems described as i.i.d noise, was rejected with a 95% confidence level.

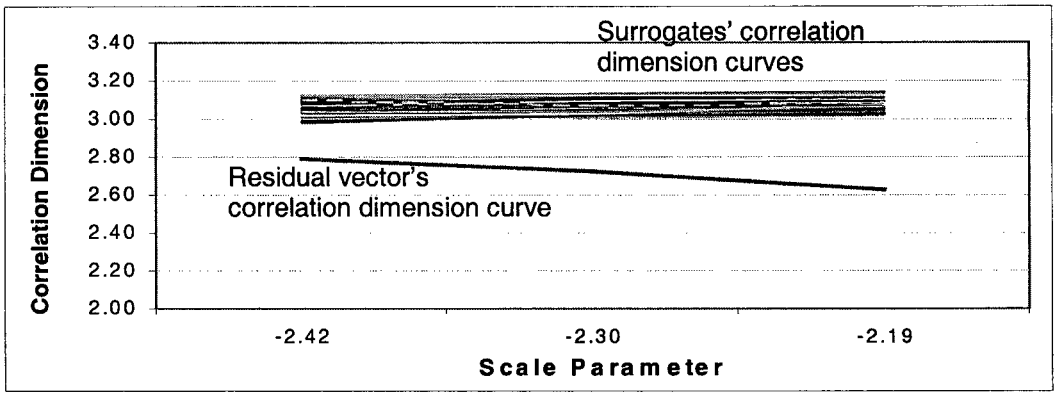

Figure 4.1.14 Correlation dimension curves of the residual vector corresponding to the Lorenz time series' model with a description length of –4082, and Algorithm 0 surrogates.

Figure 4.1.15 and table 4.1.16 gives a graphic and numeric summary of the surrogate analysis of the residual vector corresponding to the model with a description length of –9025. Observe that the correlation dimension value of the residual vector cannot be distinguished from the

correlation dimension values of the surrogates, for every scale parameter value. In this instance it is not immediately clear whether the null hypothesis should be rejected or not.

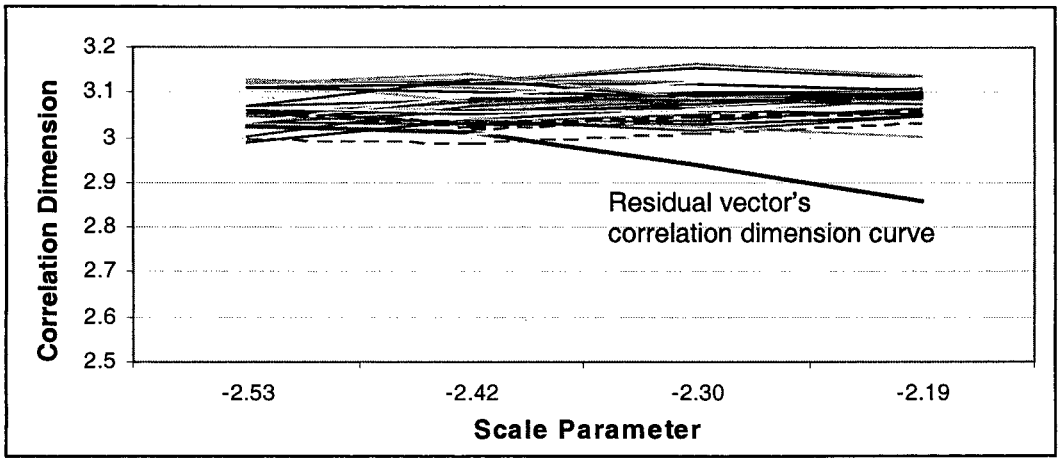

Figure 4.1.15 Correlation dimension curves of the residual vector corresponding to the Lorenz time series' model with a description length of –9025, and Algorithm 0 surrogates.

Scale Parameter	Residual vector's Correlation Dimension	Surrogates' Minimum Correlation Dimension	Surrogates' Maximum Correlation Dimension
-2.5328	3.0256	2.9086	3.1274
-2.4177	3.0098	3.0050	3.1284
-2.3026	2.9400	3.0314	3.1533
-2.1875	2.8599	3.0490	3.1337

Table 4.1.16 Summary of the analysis for Algorithm 0 surrogates and the residual vector corresponding to the Lorenz time series' model with a description length of –9025. The shaded rows indicate for which scale parameter values, the residual vector's correlation dimension value is not smaller than the surrogates' correlation dimension curves.

Initially we considered rejecting the null hypothesis based on the analysis of two scale parameter values (-2.5328 and -2.4177). The correlation dimension values of the residual vector for these scale parameter values, are greater than some of the correlation dimension values of the surrogates, and we could conclude that the residual vector corresponding to a model with a description length of –9025, belongs to the dynamic class described as i.i.d noise.

However, if we consider each correlation dimension value independent of the others (which is slightly incorrect in this case), we can apply the binomial distribution [42] to produce an estimate of the probability of producing more than a given number of false rejections (Type I error) at a

given significance level (α). According to table 4.1.16, the surrogate test yielded two rejections and two non-rejections at a 95% confidence level. The probability of two or more false rejections out of a possible four, given the null hypothesis and assuming the correlation dimension values are independent, is approximately 0.01[3]. Based on this evidence we could reject the null hypothesis (assuming independence).

Since the four correlation dimension values are not independent, the correct answer will be between these two situations. We suspect that the probability of two false rejections from four tests would still be fairly small, and rejected the null hypothesis that the residual vector corresponding to a model with a description length of –9025, belongs to the dynamic class described as i.i.d noise.

It is perhaps also not insignificant that the hypothesis is rejected for large scale parameter values, but not for small values. This could suggest that the time series and the surrogates are indistinguishable at small scales, but not at large scales, i.e. the model has adequately captured the small scale structure and most of the large scale dynamics of the underlying system.

This result partially supports our assumption that the "best" model selected according to the minimum description length criterion, does model the dynamics of the underlying system of which the time series is representative. Studying figures 4.1.13, 4.1.14 and 4.1.15 there does appear to be a "progression" in the level of noise of the residual vector. For example, the correlation dimension curve of the residual vector of figure 4.1.15 bears a closer resemblance to the correlation dimension curves of the surrogates, compared to figure 4.1.13. Although we rejected the null hypothesis for the residual vector of the "best" model, our results do support the assumption that improved models, selected according to the minimum description length criterion, capture more of the dynamics of the underlying system than the weaker models with a larger description length.

Development of mathematical techniques is needed to resolve these ambiguous issues and is beyond the scope of this study.

---

[3] $\binom{4}{4}(0.95)^0(0.05)^4 + \binom{4}{3}(0.95)^1(0.05)^3 + \binom{4}{2}(0.95)^2(0.05)^2 \approx 0.014.$

4.2 EEG time series

The electroencephalographic signal that was used in this study consists of 2000 data points representing 8 seconds of an electroencephalogram from a 19-year old, healthy male, seated and relaxed with eyes closed, sampled at 250 Hz with electrodes on the left front of the head. Figure 3.2 in Section 3.3, is a graphic representation of the time series.

4.2.1 Surrogate Analysis of EEG time series

Figure 4.2.1 is a graphic summary of the results of the AMI method, which was used to give an indication of the optimal time lag, τ.

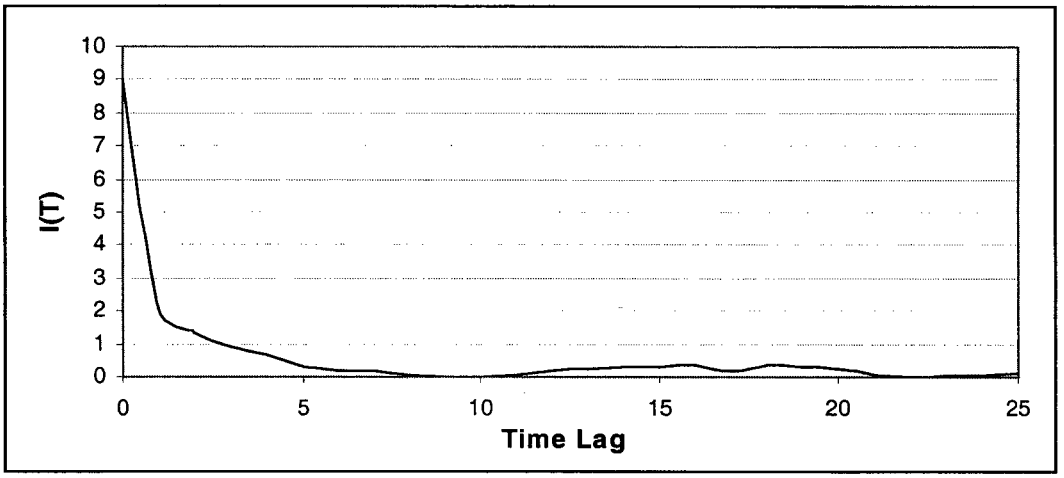

Figure 4.2.1 Average Mutual Information for EEG time series: The first minimum occurs at a time lag value of 9 and, therefore, according to the criterion suggested by Fraser and Swinney [11], we choose the value of the time lag, τ, as 9.

The method of FNN was used for an indication of the optimal embedding dimension, d_e. Figure 4.2.2 is a graphic representation of the aforementioned method's results when applied to the EEG time series.

With our chosen values for the time lag, τ, and the embedding dimension, d_e, we proceeded to test the time series for membership of the various dynamic classes modelled by the three types of surrogate algorithms. The objective of the surrogate analysis was to find evidence of nonlinear, possibly chaotic behaviour in the time series.

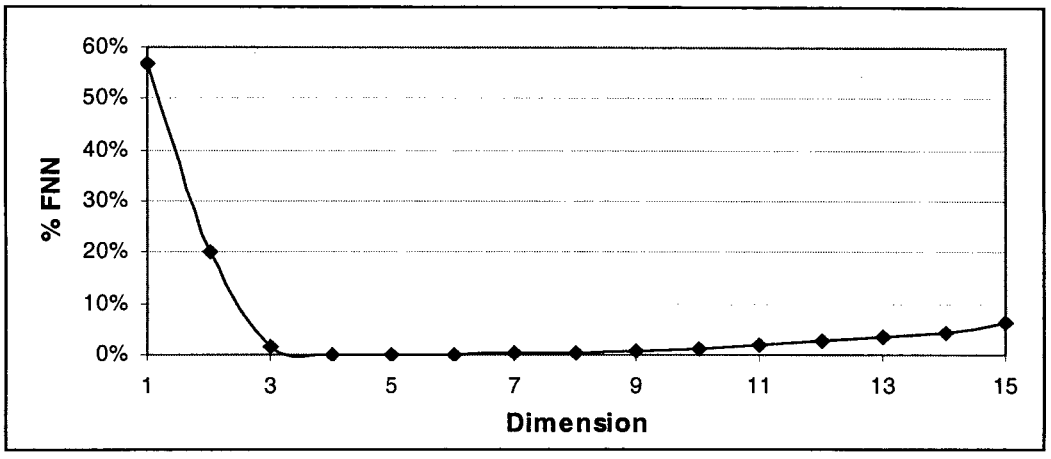

Figure 4.2.2 False Nearest Neighbours for EEG time series ($\tau = 9$): The percentage of false nearest neighbours falls to 0 when the embedding dimension is 4, which leads us to choose 4 as the embedding dimension, d_e.

From figure 4.2.3 it should be evident that the correlation dimension value of the time series is smaller than the correlation dimension values of the surrogates, for every scale parameter value. The null hypothesis of Algorithm 0 surrogates, was rejected with a 95% confidence level based on this comparison. The EEG time series does not therefore, belong to the dynamic class described as linearly filtered noise.

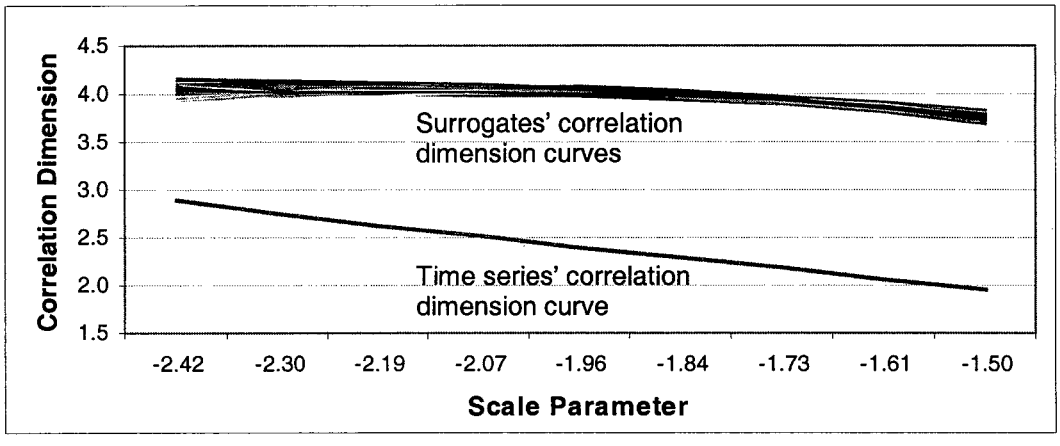

Figure 4.2.3 Correlation dimension curves of the EEG time series and Algorithm 0 surrogates.

Figure 4.2.4 is a visual summary of the surrogate analysis for Algorithm 1 surrogates. We cannot distinguish between the correlation dimension curve of the time series and the correlation dimension curves of the surrogates, for every scale parameter value. Table 4.2.5 indicates the

scale parameter values (eleven out of thirteen), where the correlation dimension value of the time series is greater than some of the correlation dimension values of the surrogates. The probability of two of more false rejections out of a possible thirteen, is 0.14[4]. Based on this result and the aforementioned comparisons, the null hypothesis that the time series belongs to the dynamic class described as a linear transformation of linearly filtered noise, could not be rejected with a 95% confidence level, although our conclusion is somewhat inconclusive.

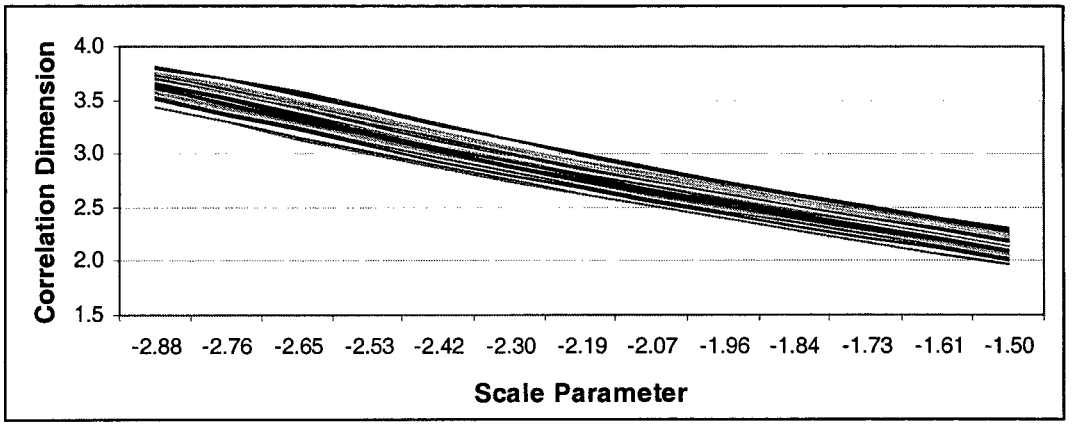

Figure 4.2.4 Correlation dimension curves of the EEG time series and Algorithm 1 surrogates.

Scale Parameter	EEG time series' Correlation Dimension	Surrogates' Minimum Correlation Dimension	Surrogates' Maximum Correlation Dimension
-2.8782	3.4361	3.4437	3.8161
-2.7631	3.3038	3.2986	3.7073
-2.6480	3.1599	3.1444	3.5897
-2.5328	3.0230	3.0018	3.4414
-2.4177	2.8864	2.8704	3.2807
-2.3026	2.7577	2.7395	3.1415
-2.1875	2.6332	2.6199	3.0033
-2.0723	2.5129	2.5057	2.8727
-1.9572	2.4000	2.3939	2.7414
-1.8421	2.2874	2.2849	2.6295
-1.7269	2.1787	2.1756	2.5159
-1.6118	2.0703	2.0660	2.4061
-1.4967	1.9671	2.0070	2.3018

Table 4.2.5 Summary of the analysis for Algorithm 1 surrogates of the EEG time series: The shaded rows indicate for which scale parameter values the residual vector's correlation dimension value is not smaller than the surrogates' correlation dimension values.

[4] $\sum_{i=2}^{13} \binom{13}{i} (0.95)^{13-i} (0.05)^{i} \approx 0.14.$

Testing the EEG time series for membership of Algorithm 2 surrogates, leads to a similar conclusion as in the previous test. The null hypothesis that the time series belongs to the dynamic class modelled by Algorithm 2 surrogates, could not be rejected. In figure 4.2.6 the correlation dimension curves of the time series and the surrogates are represented. The correlation dimension curve of the time series is indistinguishable from the correlation dimension curves of the surrogates, and the hypothesis test fails. Table 4.2.7 indicates that there are no scale parameter values where the correlation dimension value of the time series is either greater, or smaller, than the correlation dimension values of the surrogates. Based on this evidence, the null hypothesis that the time series belongs to the dynamic class described as a monotonic, nonlinear transformation of linearly filtered noise, could not be rejected with a 95% confidence level.

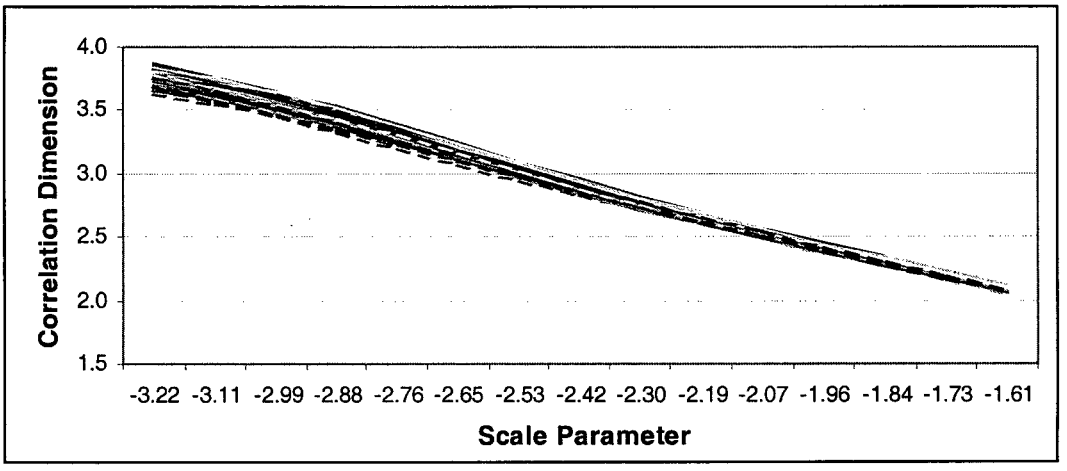

Figure 4.2.6 Correlation dimension curves for the EEG time series and Algorithm 2 surrogates.

Based on correlation dimension as the discriminating statistic, and on comparison with surrogate data, we found no evidence of nonlinearity in the EEG time series. Strictly speaking, our negative results apply only to the time series we examined and extrapolation to all EEG time series is certainly unwarranted. Our results agree with the results of Theiler and Rapp's study [40], although the origin of the EEG time series that they analysed, differs from the origin of the EEG time series we studied. The EEG time series we studied represents 8 seconds of an electroencephalogram from a 19-year old, healthy male subject, seated and relaxed with eyes closed. Theiler and Rapp [40] studied a total of 110 EEG records, obtained from 5 healthy adult subjects. With eyes closed, subjects were alternately resting or performing simple mental arithmetic (first counting backwards by seven and then counting forward by two).

Scale Parameter	EEG time series' Correlation Dimension	Surrogates' Minimum Correlation Dimension	Surrogates' Maximum Correlation Dimension
-3.2236	3.7243	3.6547	3.8682
-3.1085	3.6609	3.5847	3.7631
-2.9934	3.5557	3.4754	3.6435
-2.8782	3.4361	3.3584	3.5311
-2.7631	3.3038	3.2374	3.3935
-2.6480	3.1599	3.1053	3.2582
-2.5328	3.0230	2.9668	3.0984
-2.4177	2.8864	2.8365	2.9581
-2.3026	2.7577	2.7133	2.8230
-2.1875	2.6332	2.5985	2.6916
-2.0723	2.5129	2.4933	2.5700
-1.9572	2.4000	2.3875	2.4558
-1.8421	2.2874	2.2797	2.3415
-1.7269	2.1787	2.1732	2.2298
-1.6118	2.0703	2.0698	2.1176
-1.4967	1.9671	1.9632	1.9924

Table 4.2.7 Summary of the analysis of Algorithm 2 surrogates of the EEG time series: The shaded rows indicate for which scale parameter values the residual vector's correlation dimension value is not smaller than the surrogates' correlation dimension values.

4.2.2 Modelling of EEG time series

Although the surrogate analysis of the EEG time series failed, i.e. we could find no evidence of nonlinearity, we modelled the time series, hoping to illustrate the importance of this initial step (determining whether the apparent structure of the time series is due to nonlinearity). Omitting this step, and modelling the particular time series using nonlinear modelling techniques can lead to flawed conclusions, illustrated in the following discussion.

The EEG time series was modelled using the software program developed by UWA and based on Judd and Mees's modelling algorithm [16, 21]. Table 4.2.8 details the software program's required parameters and their values as used in the modelling of the EEG time series.

The detail of the generated model will not be discussed, as it was not the object of this study. The one-step predictions are presented in figure 4.2.9 and the free-run predictions are presented in figure 4.2.10.

Parameter	Description	Value
z	time series	first 1700 data points
v	variable embedding strategies	Matrix with variable embedding strategies, given that $\tau = 9$ and $d_e = 4$.
z_test	time series for testing	last 300 data points
fac	number of algorithm iterations	20
r	maximum radius for radial basis functions	2 × standard deviation of the time series
n	maximum number of basis functions	75
f	type of basis functions	Gaussian radial basis functions

Table 4.2.8 Parameter values for software program used to model the EEG time series.

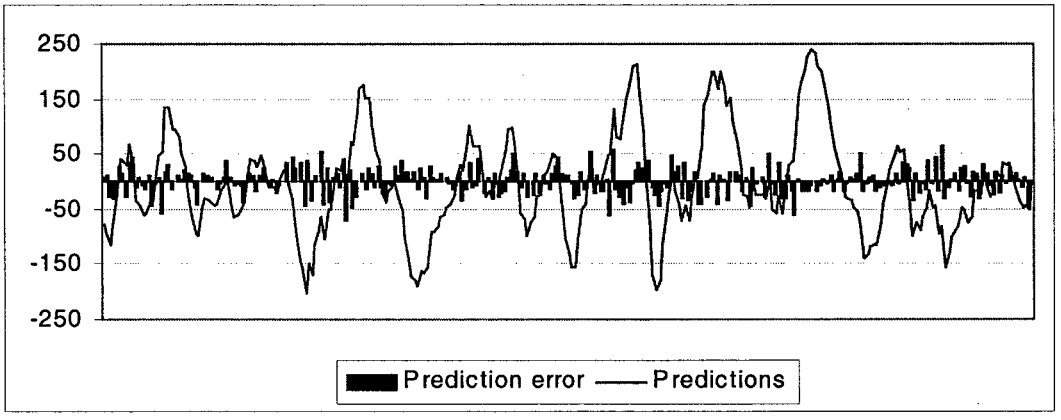

Figure 4.2.9 One-step predictions of the EEG time series using a model with a description length of 7692.

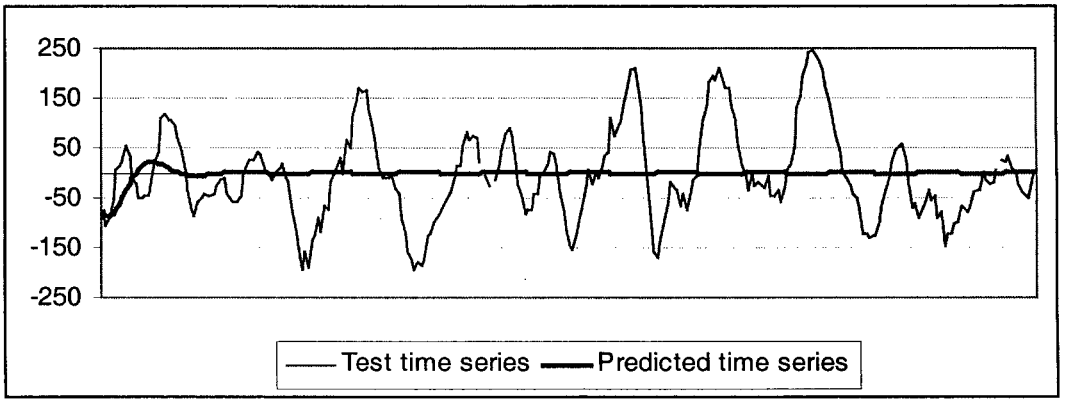

Figure 4.2.10 Free-run predictions of the EEG time series using a model with a description length of 7692.

Clearly, the one-step predictions are more accurate than the free-run predictions. This observation can be ascribed to sensitive dependence on initial conditions, which is a property of low-dimensional, chaotic systems. The danger of omitting the initial step (testing the time series for nonlinearity), becomes apparent, as we succeeded in building a nonlinear model of the time series, and using the model for predictions, without any indication that our initial assumption is wrong.

4.2.3 Surrogate Analysis of residual vectors, *e*

The modelling algorithm was executed 20 times and for the 20 models generated their description lengths ranged between the values 10200 and 7692. Two residual vectors were selected for the surrogate analysis. The one is the residual vector of the model with the largest description length (10200), i.e. the "worst" model using description length as our criterion, and the other is the residual vector of the "best" model, i.e. the model with the smallest description length (7692).

The methods of AMI and FNN was used to obtain an indication of the optimal time lag, τ, and the embedding dimension, d_e. The results of these methods are summarised in table 4.2.11.

Description Length	Embedding Dimension (d_e)	Time lag (τ)
10200	3	8
7692	3	1

Table 4.2.11 Embedding parameters for the residual vectors belonging to the models of the EEG time series with the corresponding description lengths.

These two residual vectors were tested for membership of the dynamic class described as i.i.d noise. The vectors were modelled by Algorithm 0 surrogates and correlation dimension was used as the discriminating statistic. The surrogate analyses performed were similar to that of the residual vectors of the Lorenz time series' model.

Figure 4.2.12 is a graphic representation and table 4.2.13 a summary of the surrogate analysis of the residual vector corresponding to the model with a description length of 10200. Although the correlation dimension curve of the residual vector appears indistinguishable from the correlation dimension curves of the surrogates, the probability of eight or more false rejections out of a

possible thirteen, is almost 0^5. Based on this evidence we rejected the null hypothesis that this residual vector belongs to the dynamic class described as i.i.d noise with a 95% confidence level.

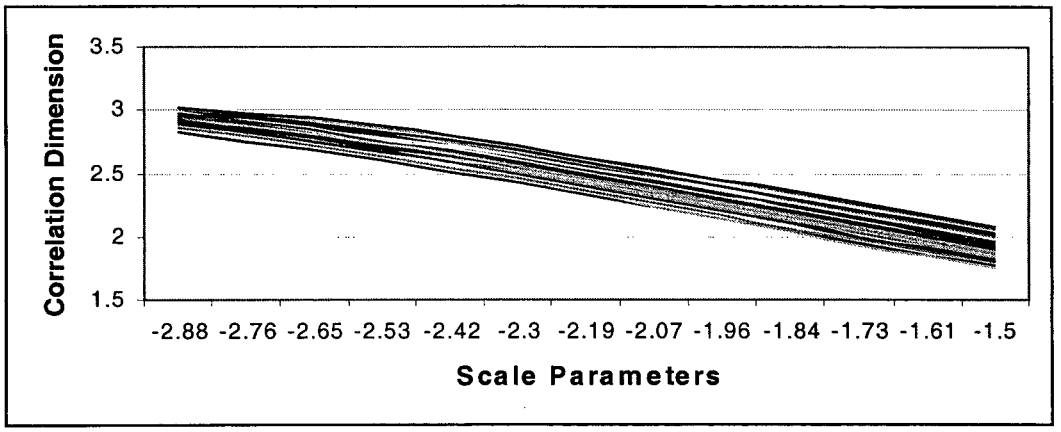

Figure 4.2.12 Correlation dimension curves of the residual vector corresponding to the EEG time series' model with a description length of 10200, and Algorithm 1 surrogates.

Scale Parameter	Residual vector's Correlation Dimension	Surrogates' Minimum Correlation Dimension	Surrogates' Maximum Correlation Dimension
-2.8782	2.8246	2.8570	3.0370
-2.7631	2.7559	2.7832	2.9934
-2.6480	2.6804	2.7123	2.9367
-2.5328	2.5962	2.6315	2.8843
-2.4177	2.5119	2.5408	2.8083
-2.3026	2.4283	2.4402	2.7296
-2.1875	2.3357	2.3406	2.6446
-2.0723	2.2398	2.2401	2.5625
-1.9572	2.1427	2.1349	2.4765
-1.8421	2.0454	2.0379	2.3882
-1.7269	1.9520	1.9394	2.2989
-1.6118	1.8588	1.8482	2.2029
-1.4967	1.7711	1.7585	2.1028

Table 4.2.13 Summary of the analysis for Algorithm 1 surrogates and the residual vector corresponding to the EEG time series' model with a description length of 10200: The shaded rows indicate for which scale parameter values the residual vector's correlation dimension value is not smaller than the surrogates' correlation dimension values.

The surrogate analysis of the residual vector corresponding to the model with a description length of 7692, leads to a different conclusion, i.e. that the residual vector belongs to the dynamic class described as i.i.d noise. The analysis is represented in figure 4.2.14 and table 4.2.15.

---

[5] $\sum_{i=8}^{13} \binom{13}{i} (0.95)^{13-i} (0.05)^{i} \approx 4 \times 10^{-8}$.

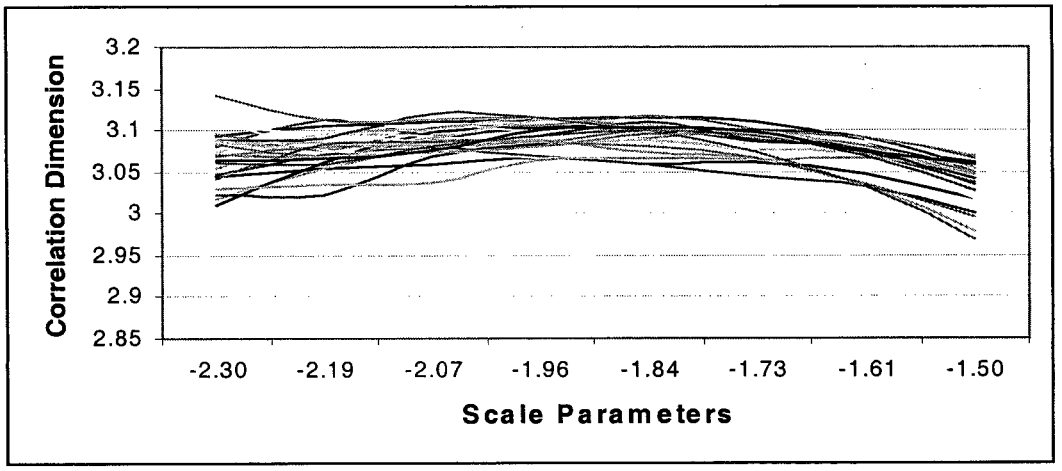

Figure 4.2.14 Correlation dimension curves of the residual vector corresponding to the EEG time series' model with a description length of 7692, and Algorithm 1 surrogates.

The correlation dimension curve of the residual vector in figure 4.2.14, cannot be distinguished from the correlation dimension curves of the surrogates. Table 4.2.15 indicates that the correlation dimension value of the residual vector is not smaller, than the correlation dimension values of the surrogates, for any of the scale parameters values. Based on this evidence, the null hypothesis that the residual vector belongs to the dynamic class described as i.i.d noise, could not be rejected with a 95% confidence level.

Scale Parameter	Residual vector's Correlation Dimension	Surrogates' Minimum Correlation Dimension	Surrogates' Maximum Correlation Dimension
-2.3026	3.0452	3.0103	3.1409
-2.1875	3.0539	3.0227	3.1270
-2.0723	3.0588	3.0378	3.1316
-1.9572	3.0654	3.0630	3.1318
-1.8421	3.0594	3.0579	3.1173
-1.7269	3.0609	3.0444	3.1089
-1.6118	3.0469	3.0129	3.1026
-1.4967	3.0172	2.9415	3.0783

Table 4.2.15 Summary of the analysis for Algorithm 1 surrogates and the residual vector corresponding to the EEG time series' model with a description length of 7692: The shaded rows indicate for which scale parameter values the residual vector's correlation dimension value is not smaller than the surrogates' correlation dimension values.

The residual vector corresponding to the "worst" model (description length = 10200), could be distinguished from i.i.d noise, based on the binomial probability. This indicates that the model is inadequate. The surrogate analysis of the EEG time series indicated that the time series does not

belong to the dynamic class described as i.i.d noise, but is indistinguishable from a monotonic, nonlinear transformation of linearly filtered noise. Hence there is deterministic structure, but this "worst" model does not capture the dynamics of the time series, leading us to suspect that it is possible to construct a "better" model.

The residual vector corresponding to the "best" model (description length = 7692) could not be distinguished from i.i.d noise and we rejected the null hypothesis. This confirmed our previous suspicion and we conclude that the deterministic structure of the EEG time series is captured by this "best" model.

This result supports our assumption that the "best" model selected according to the minimum description length criterion, does model the dynamics of the underlying system of which the time series is representative.

4.3 S&P500 time series

The time series that was used for this analysis consists of 3000 data points that represents the daily percentage change in the Standard & Poor's Composite Index of 500 stocks (S&P500) leading up to the 20% drop (classified as a market crash) that occurred on 19 October 1987. Figure 3.3 in Section 3.3 is a graphic representation of the time series.

4.3.1 Surrogate Analysis of S&P500 time series

Figure 4.3.1 is a graphic representation of the method of Average Mutual Information (AMI) that was used to obtain an indication of the optimal time lag, τ.

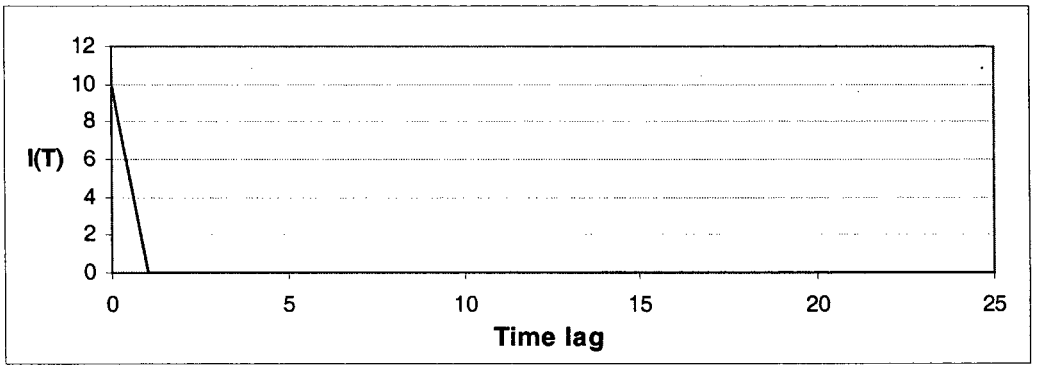

Figure 4.3.1 Average Mutual Information for S&P500 time series: The first minimum occurs at a time lag value of 1, and therefore, according to the criterion suggested by Framer and Swinney [11], we choose the value for the time lag, τ, as 1.

The method of False Nearest Neighbours (FNN) was used for an indication of the optimal embedding dimension, d_e. The results of this method applied to the S&P500 time series, is represented in figure 4.3.2.

Having decided on values for the time lag, τ, and the embedding dimension, d_e, for the S&P500 time series, we proceeded with the surrogate analysis hoping to find evidence of low-dimensional, nonlinear structure in the S&P500 time series.

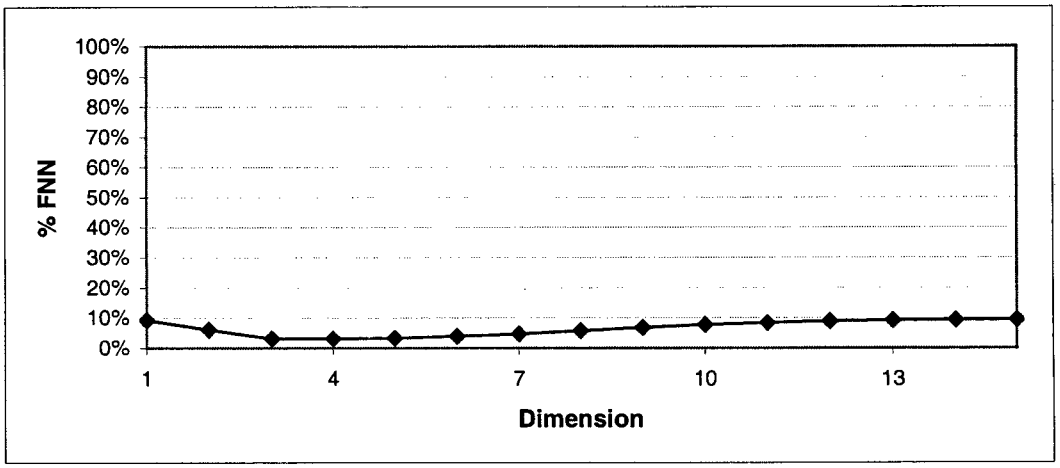

Figure 4.3.2 False Nearest Neighbours for S&P500 time series (τ=1): The percentage of false nearest neighbours has a minimum when the embedding dimension is 4, which leads us to choose 4 as the embedding dimension, d_e.

The time series was tested for membership of Algorithm 0 surrogates and the result of this hypothesis test is represented in figure 4.3.3.

The correlation dimension curve of the time series cannot be distinguished from curves of the surrogates. Table 4.3.4 indicates the scale parameter values, where the correlation dimension value of the time series is not smaller than the correlation dimension values of the surrogates. For two scale parameter values (-2.5328 and –2.3026), the correlation dimension values of the time series are greater than some of the correlation dimension values of the surrogates. Based on this evidence, we could consider to not reject the null hypothesis, but based on the probability of 0.086^6 of two or more false rejections out of a possible ten, the null hypothesis was rejected. Therefore, we can not assume that the time series belongs to the dynamic class modelled by Algorithm 0 surrogates.

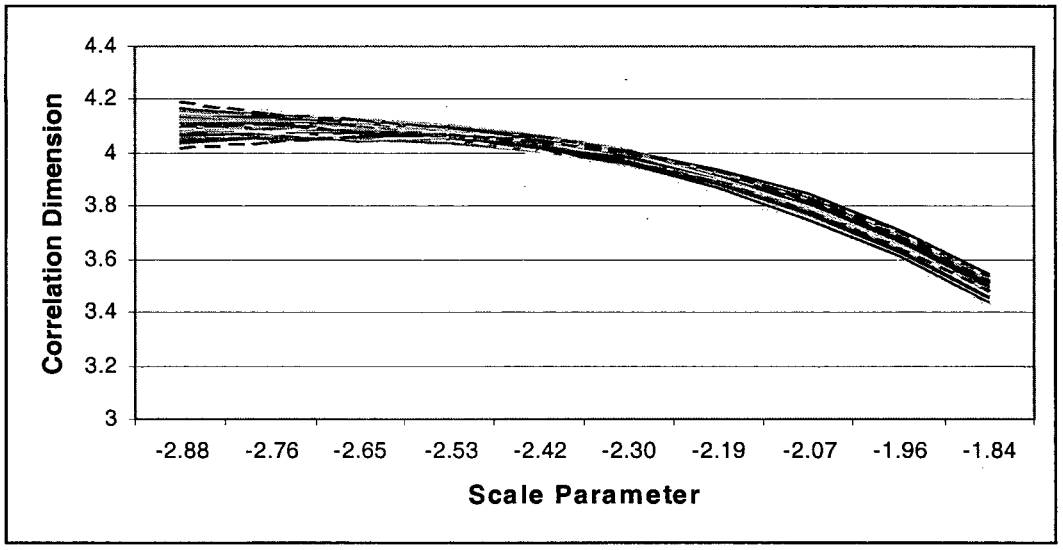

Figure 4.3.3 Correlation dimension curves of the S&P500 time series and Algorithm 0 surrogates.

[6] $\sum_{i=2}^{10} \binom{10}{i} (0.95)^{10-i} (0.05)^{i} \approx 0.086.$

Scale Parameter	Time Series' Correlation Dimension	Surrogates' Minimum Correlation Dimension	Surrogates' Maximum Correlation Dimension
-2.8782	4.1297	4.0340	4.1667
-2.7631	4.1145	4.0406	4.1547
-2.6480	4.1115	4.0325	4.1301
-2.5328	4.1054	4.0317	4.1027
-2.4177	4.0647	4.0004	4.0697
-2.3026	4.0170	3.9525	4.0146
-2.1875	3.9326	3.8667	3.9418
-2.0723	3.8293	3.7513	3.8467
-1.9572	3.6909	3.6125	3.7111
-1.8421	3.5183	3.4337	3.5446

Table 4.3.4 Summary of analysis for Algorithm 0 surrogates of the S&P500 time series: The shaded rows indicate the scale parameter values for which the residual vector's correlation dimension value is not smaller than the surrogates' correlation dimension values.

Figure 4.3.5 is a graphic representation and table 4.3.6 is a summary of the surrogate analysis. For the scale parameter values –2.3026 and –2.1875, the correlation dimension value of the S&P500 time series, cannot be distinguished from the values of the surrogates. This implies that for these scale parameter values, correlation dimension value of the time series is greater than some of the correlation dimension values of the surrogates. However, the probability of three false rejections out of a possible five is 0.0012[7]. Based on this evidence the null hypothesis is rejected at a 95% confidence level and we can not assume that the time series resembles Algorithm 1 surrogates.

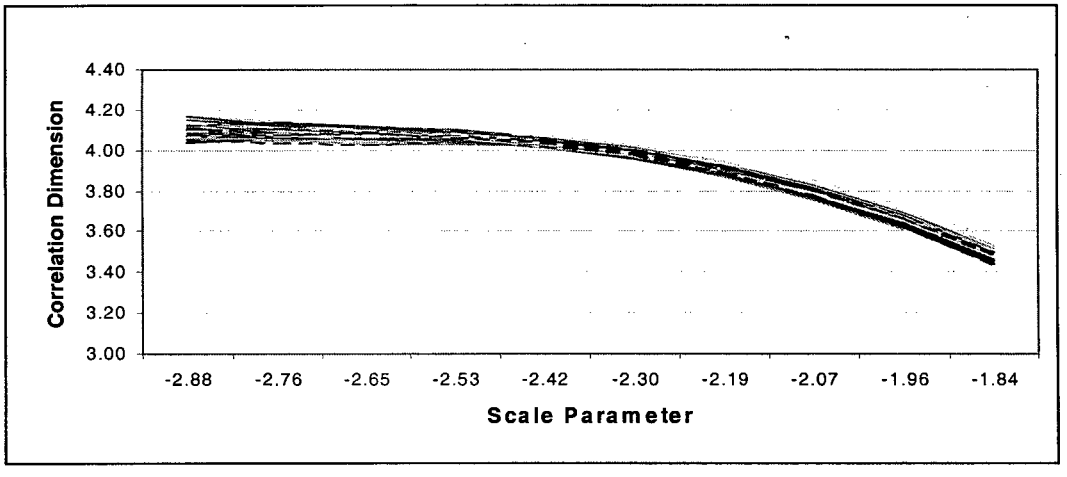

Figure 4.3.5 Correlation Dimension curves of the S&P500 time series and Algorithm 1 surrogates.

[7] $\sum_{i=3}^{5} \binom{5}{i} (0.95)^{5-i} (0.05)^{i} \approx 0.0012.$

Scale Parameter	Time Series' Correlation Dimension	Surrogates' Minimum Correlation Dimension	Surrogates' Maximum Correlation Dimension
-2.3026	4.0359	3.9810	4.1461
-2.1875	4.0170	4.0160	4.2012
-2.0723	3.9646	4.0666	4.1630
-1.9572	3.8793	4.0408	4.1424
-1.8421	3.6909	4.0109	4.1073

Table 4.3.6 Summary of analysis for Algorithm 1 surrogates of the S&P500 time series: The shaded rows indicate for which scale parameter values the residual vector's correlation dimension value is not smaller than the surrogates' correlation dimension values.

Testing the S&P500 time series for membership of Algorithm 2 surrogates lead to a different conclusion as in the case of Algorithm 0 and 1 surrogates. Figure 4.3.7 shows a graph of the correlation dimensions of the time series and the surrogates. We cannot distinguish between the correlation dimension curve of the time series and the curves of the surrogates. The null hypothesis that the time series belongs to the dynamic class described as a monotonic, nonlinear transformation of linearly filtered noise, could not be rejected with a 95% confidence level.

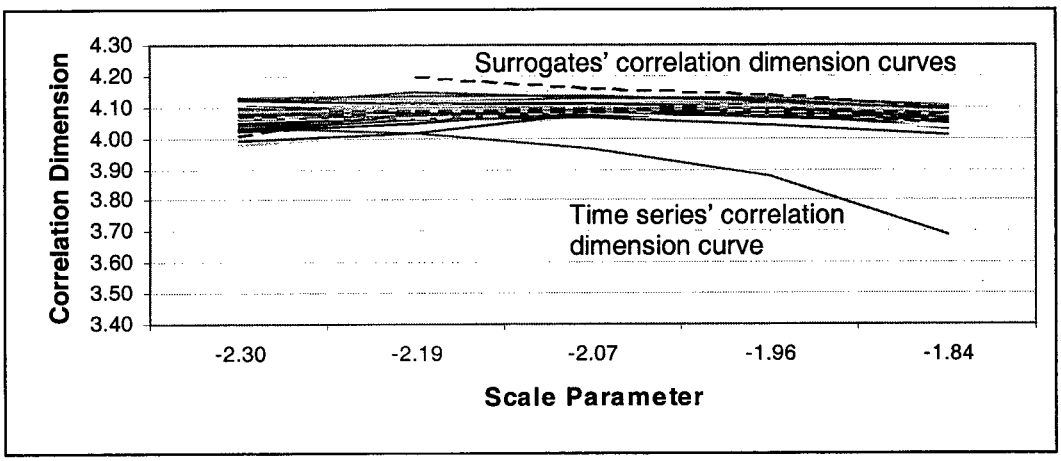

Figure 4.3.7 Correlation Dimension curves of the S&P500 time series and Algorithm 2 surrogates.

Observe in table 4.3.8 that for scale parameter value –2.5328, the correlation dimension value of the time series, is greater than the values of the surrogates. For all the other scale parameter values, the correlation dimension values of the time series is not smaller, than the values of the

surrogates. The probability of 1 or more false rejections out of a possible ten, is 0.40^8, and based on this evidence and the aforementioned comparison, we could not reject the null hypothesis at a 95% confidence level.

Scale Parameter	Time Series' Correlation Dimension	Surrogates' Minimum Correlation Dimension	Surrogates' Maximum Correlation Dimension
-2.8782	4.1297	4.0307	4.1692
-2.7631	4.1145	4.0486	4.1441
-2.6480	4.1115	4.0449	4.1148
-2.5328	4.1054	4.0366	4.0956
-2.4177	4.0647	4.0110	4.0669
-2.3026	4.0170	3.9551	4.0185
-2.1875	3.9326	3.8706	3.9415
-2.0723	3.8293	3.7633	3.8493
-1.9572	3.6909	3.6208	3.7228
-1.8421	3.5183	3.4461	3.5622

Table 4.3.8 Summary of analysis for Algorithm 2 surrogates of the S&P500 time series: The shaded rows indicate for which scale parameter values residual vector's correlation dimension value is not smaller than the surrogates' correlation dimension values.

In the case of the S&P500 time series, we rejected the null hypotheses that the S&P500 time series was modelled by either of the dynamic classes described by Algorithm 0 and 1 surrogates. We could not reject the null hypothesis that the S&P500 time series belongs to the dynamic class described as a monotonic, nonlinear transformation of linearly filtered noise. In other words, the S&P500 time series is distinguishable from linear noise at a large scale, but not at a small scale. We can conclude that the time series contains large scale static nonlinear (and dynamic linear) structure, with the small scale structure being dominated by noise.

Our conclusion corresponds to that of Abhyankar et al. [2] and Brorsen and Yang [6], who also found evidence of nonlinearity in the S&P500 time series they examined. Willey [41] and Eldridge et al. [9], on the other hand, found no evidence of nonlinearity in the S&P500 time series they examined.

[8] $\sum_{i=1}^{10} \binom{10}{i} (0.95)^{10-i} (0.05)^{i} \approx 0.40.$

CONCLUSION

This study examined nonlinear modelling techniques applied to dynamic systems, paying specific attention to the Method of Surrogate Data and its possibilities. It investigated the value that this method adds to the process of analysing and modelling a time series. The Method of Surrogate Data was initially applied in a more conventional application, i.e. testing a time series for nonlinear, dynamic structure. Thereafter, it was used in a novel application, i.e. testing the residual vectors of a nonlinear model for membership of i.i.d noise.

These applications were illustrated by analysis and modelling of three time series: a time series generated by the Lorenz equations, a time series generated by electroencephalograhpic signal (EEG), and a time series representing the percentage change in the daily closing price of the S&P500 index.

The Lorenz time series was utilised as our "base case", because the properties of this time series are well documented. As expected, the surrogate analyses rejected the null hypotheses, and led us to the conclusion that the apparent structure of the time series is due to nonlinearity. Based on this evidence, the time series was modelled using Algorithm 3.1, and the residual vectors generated by this modelling algorithm, tested for membership of the dynamic class described as i.i.d noise, using surrogate analysis. The results of this application illustrated that, as the model captures more of the underlying dynamics of the system (description length decreases), the residual vector bears closer resemblance to i.i.d noise. Unfortunately it only partially verified that the minimum description length criterion, leads to models that capture the underlying dynamics of the time series, with the residual vector resembling i.i.d noise. In the case of the "worst" model (largest description length), the residual vector could be distinguished from i.i.d noise, verifying that it is not the "best" model. However, in the case of the "best" model (smallest description length), the residual vector could also be distinguished from i.i.d noise, assuming independence of the correlation dimension values for the different scale parameter values. This assumption is not completely valid, but it does provide us with a means on which to base the rejection of the null hypothesis in the instance where the correlation dimension values of the residual vector, is smaller than the surrogates' correlation dimension values for some of the scale parameter values, but not for every scale parameter value. It is perhaps also not insignificant that the hypothesis for the residual vector corresponding to the "best" model was rejected for large scale parameter values, but not for small values. This could suggest that the residual vector and

the surrogates are indistinguishable at small scales, but not at large scales, i.e. the model has adequately captured the small scale structure and most of the large scale dynamics of the underlying system. Development of mathematical techniques is needed to resolve these ambiguous issues and is beyond the scope of this study.

The surrogate analysis of the EEG time series failed, i.e. we could not reject the null hypothesis that the time series belongs to the dynamic class described by Algorithm 2 surrogates. Although the EEG time series failed the initial surrogate analysis, we modelled the time series, using Algorithm 3.1 and analysed the residual vectors. Based on the time series' resemblance to a monotonic, nonlinear transformation of linearly filtered noise, we expected Algorithm 3.1 to construct a model that captures this static nonlinear (and dynamic linear) structure of the time series.

The residual vector corresponding to the "worst" model (description length = 10200), could be distinguished from i.i.d noise, as opposed to the residual vector corresponding to the "best" model (description length = 7692) which could not be distinguished from i.i.d noise. This result supports our assumption that the "best" model selected according to the minimum description length criterion, does model the dynamics of the underlying system of which the time series is representative.

Using surrogate analysis, we found evidence that the S&P500 time series resembles Algorithm 2 surrogates (a monotonic, nonlinear transformation of linearly filtered noise), but not Algorithm 0 or 1 surrogates. In other words, the S&P500 time series is distinguishable from linear noise at a large scale, but not at a small scale. We can conclude that the time series contains large scale static nonlinear (and dynamic linear) structure, with the small scale structure being dominated by noise. This conclusion corresponds to that of Abhyankar et al. [2] and Brorsen and Yang [6], who also found evidence of nonlinearity in the S&P500 time series they examined, but conflicts with the results of Willey [41] and Eldridge et al. [9]

The Method of Surrogate Data can be used in conjunction with the minimum description length criterion, to ascertain whether a model is "good", i.e. captures the underlying dynamics of the time series. It is a satisfactory first order test for the fit of a specific model of a time series. The simplicity of the method makes it particularly attractive. It is trivial to program, whilst the calculation of the description length of a model is complex and is complicated to program.

A significant advantage of the description length is that two description lengths can be compared numerically. This is not possible with the surrogate analysis of the residual vectors. We can determine whether the residual vectors resemble i.i.d noise or not, but we cannot determine whether the one residual vector resembles i.i.d noise more than another. The development of methodologies to perform such comparisons would enhance the value of surrogate analysis of residual vectors. This can be explored in further studies of the subject.

BIBLIOGRAPHY

[1] H.D.I. Abarbanel, R. Brown, J.J. Sidorowich, and L.S. Tsimring, 'The analysis of observed chaotic data in physical systems', *Review of Modern Physics* **65** (1993), 1331-1392.

[2] A. Abhyankar, L.S. Copeland, and W. Wong, 'Uncovering Nonlinear Structure in Real-Time Stock-Market Indexes: The S&P500, the DAX, the Nikkei 225, and the FTSE-100', *Journal of Business & Economic Statistics* **15** (1997), 1-14.

[3] A. Akaike, 'A new look at the statistical model identification', *IEEE transactions on Automatic Control* **19** (1974), 716-723.

[4] A.M. Albano, J. Muench, C. Schwartz, A.I. Mees, and R.E. Rapp, 'Singular-value decomposition and the Grassberger-Procaccia algorithm', *Physical Review A* **38** (1988), 3017-3026.

[5] D.A. Berry and B.W. Lindgren, *Statistics: Theory and methods*, (Brooks/Cole publishing company, 1990).

[6] B.W. Brorsen, and S-R Yang, 'Nonlinear Dynamics and the Distribution of Daily Stock Index Returns', *The Journal of Financial Research* **XVII** (1994), 187-203.

[7] M.C. Casdagli, L.D. Iasemidis, J.C. Sackellares, S.N. Roper, R.L. Gilmore and R.S. Savit, 'Characterizing nonlinearity in invasive EEG recordings from temporal lobe epilepsy', *Physica D* **99** (1996), 381-399.

[8] M. Ding, C. Grebogi, E. Ott, T. Sauer, and J.A. Yorke, 'Plateau onset for correlation dimension: When does it occur?', *Physical Review Lettes* **70** (1993), 3872-3875.

[9] R.M. Eldridge, C. Bernhardt, and I. Mulvey, 'Evidence of Chaos in the S&P500 Cash Index', *Advances in Futures and Options Research* **6** (1993), 179-192.

[10] J.D. Farmer, E. Ott, and J.A. Yorke, 'The dimension of chaotic attractors', *Physica D* **7** (1983), 153-180.

[11] A.M. Fraser and H.L. Swinney, 'Independent coordinates for strange attractors from mutual information', *Physical Review A* **33** (1986), 1134-1140.

[12] D.E. Goldberg, *Genetic Algorithms in Search, Optimization, and Machine Learning*, (Addison-Wesley , 1989).

[13] P. Grassberger and I. Procaccia, 'Characterization of strange attractors', *Physical Review Letters* **50** (1983), 346-349.

[14] P. Grassberger and I. Procaccia, 'Measuring the Strangeness of Strange Attractors', *Physica D* **9** (1983), 189-208.

[15] K. Judd, 'An improved estimator of dimension and some comments on providing confidende intervals', *Physica D* **56** (1992), 216-228.

[16] K. Judd and A.I. Mees, 'On selecting models for nonlinear time series', *Physica D* **82** (1995), 426-444.

[17] K. Judd and A.I. Mees, 'Embedding as a modeling problem', *Physica D* **120** (1998), 273-286.

[18] H. Kantz and T. Schreiber, *Nonlinear time series analysis*, (Cambridge University Press, 1997).

[19] M.B. Kennel, R. Brown, and H.D.I. Abarbanel, 'Determining embedding dimension for phase space reconstruction using a geometrical construction', *Physical Review A* **45** (1992), 3403-3411.

[20] R. Mañé, 'On the dimension of the compact invariant sets of certain nonlinear maps', *Lecture Notes in Mathematics* **898** (1981), 230-243

[21] A.I. Mees, 'Reconstructing Chaotic Systems in the Presence of Noise', *Proceedings of the 7th Toyota Conference, 1993: Towards Harnessing of Chaos*, (Elsevier, 1994).

[22] G.C. Philippatos, and E. Pilarinu, 'Instabilities and Chaotic Behavior of Stock Prices in International Capital Markets', *Managerial Finance* **20** (1994), 14-42.

[23] D. Prichard and J. Theiler, 'Generalized redundancies for time series analysis', *Physica D* **84** (1995), 476-493.

[24] J. Rissanen, *Stochastic complexity in statistical inquiry*, (Singapore: World Scientific, 1980).

[25] G.B. Schmid and R.M. Dünki, 'Indications of nonlinearity, intraindividual specificity and stability of human EEG: The unfolding dimension', *Physica D* **93** (1996), 165-190.

[26] T. Schreiber and A. Schmitz, 'Improved surrogate data for nonlinearity tests', *Physical Review Letters* **77** (1996), 635-638.

[27] G. Schwarz, 'Estimating the dimension of a model', *Annals of Statistics* **6** (1978), 461-464.

[28] M. Small, *Nonlinear dynamics in infant respiration*, (PhD Thesis, University of Western Australia, 1998).

[29] M. Small and K. Judd, 'Using surrogate data to test for nonlinearity in experimental data', *International Symposium on Nonlinear Theory and its Applications* **2**, 1133-1136 (Research Society for Nonlinear Theory and its Applications, IEICE, 1997).

[30] M. Small and K. Judd, 'Comparisons of new nonlinear modeling techniques with applications to infant respiration', *Physica D* **117** (1998), 283-298.

[31] M. Small and K. Judd, 'Detecting nonlinearity in experimental data', *International Journal of Bifurcation and Chaos* **8** (1998), 1231-1244.

[32] M. Small and K. Judd, 'Pivotal statistics for non-constrained realizations of composite null hypotheses in surrogate data analysis', *Physica D* **120** (1998), 386-400.

[33] J.C. Sprott and G.Rowlands, *User's Manual, Chaos Data Analyzer: The Professional Version*, (American Institute of Physics).

[34] S.H. Strogatz, *Nonlinear Dynamics and Chaos with applications in Physics, Biology, Chemistry and Engineering*, (Addison-Wesley Publishing Company, 1998).

[35] F. Takens, 'Detecting strange attractors in turbulence', *Lecture Notes in Mathematics* **898** (1981), 366-381.

[36] F. Takens, 'Detecting nonlinearities in stationary time series', *International Journal of Bifurcation and Chaos* **3** (1993), 241-256.

[37] J. Theiler, 'On the evidence for low-dimensional chaos in an epileptic electroencephalogram', *Physica Letters A* **196** (1995), 335-341.

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

[39] J. Theiler and D. Prichard, 'Constrained-realization Monte Carlo method for hypothesis testing', *Physica D* **94** (1996), 221-235.

[40] J. Theiler and P. Rapp, 'Re-examination of the evidence for low-dimensional, nonlinear structure in the human electroencephalogram, *Electroencephalography and clinical Neurophysoilogy* **98** (1996), 213-222.

[41] T. Willey, 'Testing for Nonlinear Dependence in Daily Stock Indices', *Journal of Economics and Business* **44** (1992), 63-76.

[42] 'Statistics for Scientists', *Lecture notes in Statistics for Scientists 314*, (University of Stellenbosch, 1997)

[43] 'The analysis of chaotic time series from experimental data: an introduction', *Lecture notes in Chaos Theory*, (University of Stellenbosch, 1998).