A COMPARATIVE EVALUATION OF NON-LINEAR TIME SERIES ANALYSIS AND SINGULAR SPECTRUM ANALYSIS FOR THE MODELLING OF AIR POLLUTION

ANTHONY FRANCIS DIAB

Thesis submitted to the University of Stellenbosch
in partial fulfilment of the requirements
for the degree of Master of Science in Engineering

Supervisor: Dr. A.B. Taylor

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.

A.F. Diab

ABSTRACT

Air pollution is a major concern in the Cape Metropole. A major contributor to the air pollution problem is road transport. For this reason, a national vehicle emissions study is in progress with the aim of developing a national policy regarding motor vehicle emissions and control. Such a policy could bring about vehicle emission control and regulatory measures, which may have far-reaching social and economic effects.

Air pollution models are important tools in predicting the effectiveness and the possible secondary effects of such policies. It is therefore essential that these models are fundamentally sound to maintain a high level of prediction accuracy. Complex air pollution models are available, but they require spatial, time-resolved information of emission sources and a vast amount of processing power. It is unlikely that South African cities will have the necessary spatial, time-resolved emission information in the near future. An alternative air pollution model is one that is based on the Gaussian Plume Model. This model, however, relies on gross simplifying assumptions that affect model accuracy.

It is proposed that statistical and mathematical analysis techniques will be the most viable approach to modelling air pollution in the Cape Metropole. These techniques make it possible to establish statistical relationships between pollutant emissions, meteorological conditions and pollutant concentrations without gross simplifying assumptions or excessive information requirements. This study investigates two analysis techniques that fall into the aforementioned category, namely, *Non-linear Time Series Analysis* (specifically, the method of delay co-ordinates) and *Singular Spectrum Analysis* (*SSA*).

During the past two decades, important progress has been made in the field of Non-linear Time Series Analysis. An entire "toolbox" of methods is available to assist in identifying non-linear determinism and to enable the construction of predictive models. It is argued that the dynamics that govern a pollution system are inherently non-linear due to the strong correlation with weather patterns and the complexity of the chemical reactions and physical transport of the pollutants. In addition to this, a statistical technique (the method of surrogate data) showed that a pollution data set, the oxides of Nitrogen (NO_x), displayed a degree of non-linearity, albeit that there was a high degree of noise contamination. This suggested that

a pollution data set will be amenable to non-linear analysis and, hence, Non-linear Time Series Analysis was applied to the data set.

SSA, on the other hand, is a linear data analysis technique that decomposes the time series into statistically independent components. The basis functions, in terms of which the data is decomposed, are data-adaptive which makes it well suited to the analysis of non-linear systems exhibiting anharmonic oscillations. The statistically independent components, into which the data has been decomposed, have limited harmonic content. Consequently, these components are more amenable to prediction than the time series itself. The fact that SSA's ability has been proven in the analysis of short, noisy non-linear signals prompted the use of this technique.

The aim of the study was to establish which of these two techniques is best suited to the modelling of air pollution data. To this end, a univariate model to predict NO_x concentrations was constructed using each of the techniques. The prediction ability of the respective model was assumed indicative of the accuracy of the model. It was therefore used as the basis against which the two techniques were evaluated. The procedure used to construct the model and to quantify the model accuracy, for both the Non-linear Time Series Analysis model and the SSA model, was consistent so as to allow for unbiased comparison. In both cases, no noise reduction schemes were applied to the data prior to the construction of the model. The accuracy of a 48-hour step-ahead prediction scheme and a 100-hour step-ahead prediction scheme was used to compare the two techniques.

The accuracy of the SSA model was markedly superior to the Non-linear Time Series model. The paramount reason for the superior accuracy of the SSA model is its adept ability to analyse and cope with noisy data sets such as the NO_x data set. This observation provides evidence to suggest that Singular Spectrum Analysis is better suited to the modelling of air pollution data. It should therefore be the analysis technique of choice when more advanced, multivariate modelling of air pollution data is carried out.

It is recommended that noise reduction schemes, which decontaminate the data without destroying important higher order dynamics, should be researched. The application of an effective noise reduction scheme could lead to an improvement in model accuracy. In addition to this, the univariate SSA model should be extended to a more complex multivariate

model that explicitly encompasses variables such as traffic flow and weather patterns. This will explicitly expose the inter-relationships between the variables and will enable sensitivity studies and the evaluation of a multitude of scenarios.

OPSOMMING

Die hoë vlak van lugbesoedeling in die Kaapse Metropool is kommerwekkend. Voertuie is een van die hoofoorsake, en as gevolg hiervan word 'n landswye ondersoek na voertuig-emissie tans onderneem sodat 'n nasionale beleid opgestel kan word ten opsigte van voertuig-emissie beheer. Beheermaatreëls van so 'n aard kan verreikende sosiale en ekonomiese uitwerkings tot gevolg hê.

Lugbesoedelingsmodelle is van uiterste belang in die voorspelling van die effektiwiteit van moontlike wetgewing. Daarom is dit noodsaaklik dat hierdie modelle akkuraat is om 'n hoë vlak van voorspellingsakkuraatheid te handhaaf. Komplekse modelle is beskikbaar, maar hulle verg tyd-ruimtelike opgeloste inligting van emmissiebronne en baie berekeningsvermoë. Dit is onwaarskynlik dat Suid-Afrika in die nabye toekoms hierdie tyd-ruimtelike inligting van emissiebronne gaan hê. 'n Alternatiewe lugbesoedelingsmodel is dié wat gebaseer is op die "Guassian Plume". Hierdie model berus egter op oorvereenvoudigde veronderstellings wat die akkuraatheid van die model beïnvloed.

Daar word voorgestel dat statistiese en wiskundige analises die mees lewensvatbare benadering tot die modellering van lugbesoedeling in die Kaapse Metropool sal wees. Hierdie tegnieke maak dit moontlik om 'n statistiese verwantskap tussen besoedelingsbronne, meteorologiese toestande en besoedeling konsentrasies te bepaal sonder oorvereenvoudigde veronderstellings of oormatige informasie vereistes. Hierdie studie ondersoek twee analise tegnieke wat in die bogenoemde kategorie val, naamlik, Nie-lineêre Tydreeks Analise en Enkelvoudige Spektrale Analise (ESA).

Daar is in die afgelope twee dekades belangrike vooruitgang gemaak in die studieveld van Nie-lineêre Tydreeks Analise. 'n Volledige stel metodes is beskikbaar om nie-lineêriteit te identifiseer en voorspellingsmodelle op te stel. Dit word geredeneer dat die dinamika wat 'n besoedelingsisteem beheer nie-lineêr is as gevolg van die sterk verwantskap wat dit toon met weerpatrone asook die kompleksiteit van die chemiese reaksies en die fisiese verplasing van die besoedelingstowwe. Bykomend verskaf 'n statistiese tegniek (die metode van surrogaatdata) bewyse dat 'n lugbesoedelingsdatastel, die okside van Stikstof (NO_x), nie-lineêre gedrag toon, alhoewel daar 'n hoë geraasvlak is. Om hierdie rede is die besluit geneem om Nie-lineêre Tydreeks Analise aan te wend tot die datastel.

ESA daarenteen, is 'n lineêre data analise tegniek. Dit vereenvoudig die tydreeks tot statistiese onafhanklike komponente. Die basisfunksies, in terme waarvan die data vereenvoudig is, is data-aanpasbaar en dit maak hierdie tegniek gepas vir die analise van nie-lineêre sisteme. Die statisties onafhanklike komponente het beperkte harmoniese inhoud, met die gevolg dat die komponente aansienlik makliker is om te voorspel as die tydreeks self. ESA se effektiwitiet is ook al bewys in die analise van kort, hoë-graas nie-lineêre seine. Om hierdie redes, is ESA toegepas op die lugbesoedelings data.

Die doel van die ondersoek was om vas te stel watter een van die twee tegnieke meer gepas is om lugbesoedelings data te analiseer. Met hierdie doelwit in sig, is 'n enkelvariaat model opgestel om NO_x konsentrasies te voorspel met die gebruik van elk van die tegnieke. Die voorspellingsvermoë van die betreklike model is veronderstel om as 'n maatstaf van die model se akkuraatheid te kan dien en dus is dit gebruik om die twee modelle te vergelyk. 'n Konsekwente prosedure is gevolg om beide die modelle te skep om sodoende invloedlose vergelyking te verseker. In albei gevalle was daar geen geraasverminderings-tegnieke toegepas op die data nie. Die akuraatheid van 'n 48-uur voorspellingsmodel en 'n 100-uur voorspellingsmodel was gebruik vir die vergelyking van die twee tegnieke.

Daar is bepaal dat die akkuraatheid van die ESA model veel beter as die Nie-lineêre Tydsreeks Analise is. Die hoofrede vir die ESA se hoër akkuraatheid is die model se vermoë om data met hoë geraasvlakke te analiseer.

Hierdie ondersoek verskaf oortuigende bewyse dat Enkelvoudige Spektrale Analiese beter gepas is om lugbesoedelingsdata te analiseer en gevolglik moet hierdie tegniek gebruik word as meer gevorderde, multivariaat analises uitgevoer word.

Daar word aanbeveel dat geraasverminderings-tegnieke, wat die data kan suiwer sonder om belangrike hoë-orde dinamika uit te wis, ondersoek moet word. Hierdie toepassing van effektiewe geraasverminderings-tegniek sal tot 'n verbetering in model-akkuraatheid lei. Aanvullend hiertoe, moet die enkele ESA model uitgebrei word tot 'n meer komplekse multivariaat model wat veranderlikes soos verkeersvloei en weerpatrone insluit. Dit sal die verhoudings tussen veranderlikes ten toon stel en sal sensitiwiteit-analises en die evaluering van menigte scenarios moontlik maak.

ACKNOWLEDGEMENTS

I am most grateful to my supervisor, Dr. A.B. Taylor, for introducing me to the concept of air pollution modelling and for making this thesis possible through his guidance, encouragement and financial support.

The study was partly funded by the National Research Foundation and this is gratefully acknowledged.

I would like to thank Grant Ravenscroft of the Cape Metropolitan Council for providing the pollution data.

J.P. Barnard of the Institute of Mineral Processing and Intelligent Process Systems was most helpful in setting up the neural network model and his insight into the field of Non-linear Time Series Analysis was invaluable.

Thank-you to my parents, friends and girlfriend for their added support and encouragement throughout the duration of this study.

TABLE OF CONTENTS

LIST OF FIGURES

Chapter 6

LIST OF TABLES

LIST OF COMMONLY USED SYMBOLS

Chapter 4

$C(k)$	autocorrelation function
d_c	correlation dimension
$I(k)$	average mutual information
k	time lag
m	embedding dimension
Δn	step-ahead prediction horizon
N	number of observations in the data set
NO_x	oxides of Nitrogen
R	Pearson product moment correlation co-efficient
$\Re^m$	phase space of dimension m
s_n	the n^{th} measured observation in a time series
$\boldsymbol{s}_n$	the n^{th} embedding vector
$\hat{s}_{N+\Delta n}$	predicted value of s, Δn steps ahead of s_N
$\boldsymbol{x}_n$	the n^{th} state vector

Chapter 5

a^k	the k^{th} principal component
ARMA	auto-regressive moving average
$\mathbf{e}_i$	the i^{th} eigenvector
$\mathbf{E}$	diagonalising matrix
E^k	the k^{th} empirical orthogonal function
EOF	empirical orthogonal function
k	step-ahead prediction horizon
M	window length
PC	principal component
RC	reconstructed component
$\mathbf{S}$	lagged-covariance matrix
$\mathbf{X}$	trajectory matrix
x_i	the i^{th} measured observation in a time series

1 INTRODUCTION

Air pollution is rapidly becoming a major concern in the Cape Metropole. Not only does this pollution contribute to the formation of unsightly smog that detracts from the region's natural beauty, but it is also a cause of concern regarding health risks.

A national vehicle emissions study is in progress with the aim of developing a national policy regarding motor vehicle emissions and control (Terblanche, 1995). Such a policy could bring about vehicle emission control and regulatory measures, which may have far-reaching social and economic effects. Thus, considering the importance of such legislation, it is imperative that the best possible information be available when drafting a bill of this nature.

Air pollution models are important tools in predicting the effectiveness and the possible secondary effects of such policies. They are also useful for performing important sensitivity analyses in demographic and metropolitan planning and management. It is therefore essential that these models are fundamentally sound to maintain a high level of prediction accuracy.

Complex models that address the physical and chemical processes of air pollution from first principles are in existence. However, these models require large quantities of spatial, time-resolved information of emission sources and a vast amount of processing power. The most frequently applied dispersion model is one based on the Gaussian Plume Model. Although relatively straightforward to implement, it relies on a few gross simplifying assumptions that restrict the accuracy of the model.

The use of statistical and mathematical techniques enables the analysis of experimental data with the aim of building an empirical model that is an accurate representation of the underlying physical situation. This approach offers an optimal compromise between the need to have spatial, time-resolved data and the ability to represent all the primary effects that influence air quality in a region. Numerous statistical and mathematical techniques can be applied. This study will specifically look at *Non-linear Time Series Analysis* (specifically the method of delay co-ordinates) and *Singular Spectrum Analysis* (*SSA*). **The aim of the study is to establish which of these two techniques is best suited to the modelling of air pollution data.**

A complete air pollution model will be a multivariate model that uses statistical and mathematical techniques to establish relationships between pollutant emission sources, meteorological conditions and pollutant concentrations. This study is restricted to the prediction of a single pollutant species, namely, the oxides of Nitrogen (NO_x). Prediction models are constructed using both the Non-linear Time Series Analysis techniques and Singular Spectrum Analysis. The prediction ability of the respective model is assumed to be indicative of the accuracy of the model. It will therefore be used as the basis against which the two techniques are evaluated.

Prediction of a single time series using only that time series to build the model may, *prima facie*, seem unfounded. It is however important to bear in mind that a single record of a variable from a dynamic system (e.g. the pollution system) is the outcome of all interacting variables. Thus, theoretically, the single record should implicitly contain information about the dynamics of all the important variables involved in the evolution of the system. Statistical and mathematical techniques are applied to the data in the hope of extracting the "embedded" information that each variable contains pertaining to the pollution system. This information can then be used to build a model which has the ability to predict future values of the time series. Such a model will be of limited practical value in that sensitivity analysis cannot be performed since the relationships between the interacting variables are not known explicitly. However, in this study, the prediction of a single time series provides a means of comparing the analysis techniques under investigation.

Once the most accurate and robust technique is identified, further work could involve the extension of the univariate model to a more complex multivariate model that explicitly encompasses variables such as traffic flow and weather patterns. This type of model will have the ability to perform sensitivity analyses and evaluate the impact of a multitude of scenarios. At this point, it should be emphasised that at the time when this research was initiated, there was a lack of useful traffic data. This study will present evidence of the strong correlation between pollutant concentrations and traffic flow. For this reason, it is imperative that one of the inputs to a comprehensive air pollution model should be an indication of traffic activity. Cape Town's present traffic control system has the ability to capture real time data. The data is however not recorded at reasonable rates or for a sufficient duration of time. Another important consideration is the fact that the traffic data should be synchronised with the pollutant concentration data. To construct a comprehensive air pollution model, a system will

have to be devised by which traffic data can be recorded and archived properly for use in computerised analysis.

A limitation of this study is the fact that it is based on a single time series. This limitation stemmed from the lack of complete air pollution data sets. The NO_x data set was the most complete and it was therefore used in this study. Although this limitation precludes drawing conclusive evidence, the research does provide a solid foundation for future work.

The majority of the computation performed in this study is done using the MATLAB® programming language. Computation is carried out using a desktop computer with a 500 MHz Intel® Celeron™ processor and 192 Mb of RAM.

The study starts out with a discussion of the experimental data that was used for the construction of the modes. NO_x is identified as the pollutant that will be used as the data set since it is the most complete record available with the least number of spurious data points. There is a brief discussion of the characteristics of the NO_x pollutant. This is followed by a summary of the various methods used in the construction of air pollution models in Chapter 3. Chapter 4 details the technique of Non-linear Time Series Analysis. The theory, its application in modelling the NO_x data set and a discussion of the results are contained in this chapter. An analogous approach is taken in Chapter 5 with the method of Singular Spectrum Analysis. A discussion, focussing on the comparison of the results obtained using the two different techniques, is to be found in Chapter 6. Following this, conclusions are made and, finally, recommendations for future research are outlined.

2 EXPERIMENTAL DATA

The foundation of any model is the data upon which the model is built. The measured observations describe the system. It is the model's task to extract the information from the measured observations and in so doing provide an accurate representation of the underlying physical process. This suggests that the accuracy of the model, irrespective of its complexity, will be dependent on the quality of the data set on which the model is built. This chapter details the selection of the pollution data set and the characteristics of the selected pollutant.

2.1 POLLUTION DATA

The Cape Metropolitan Council's (CMC) Scientific Services Department measures pollution data for Cape Town City Centre at the City Hall. Pollution data, including NO_x, particulates and ozone, was obtained from the CMC for the period spanning 1 January 1995 to 31 December 1995. The pollutant concentrations are reported as hourly averages on the hour.

The data sets were scanned to identify missing data points, negative values and obvious excursions above a maximum. Maximum values for the pollutant concentrations were suggested by Grant Ravenscroft (CMC Scientific Services, personal communication). As would be expected, the data sets were incomplete because of downtime due to equipment failure etc. The NO_x data for the period 1 January 1995 to 31 December 1995 was the most complete data set with the least number of spurious measurements and missing data. This was one of the primary reasons for using the NO_x data to evaluate which of the two time series analysis techniques provided the most accurate representation of a pollution process.

There were only 53 single data points missing and they were replaced simply by means of linear interpolation with the neighbouring data points. Outliers – values that are "far" removed from the middle of the distribution – were investigated by means of a box plot constructed with the statistical computer package, Statistica®. The NO_x data set was divided into 7 subsets. Each subset was analysed for outliers.

A data point was deemed an outlier if the following conditions were met:

$$\text{data point value} > UBV + o.c * (UBV - LBV) \tag{2.1}$$

or

$$\text{data point value} < LBV - o.c * (UBV - LBV) \tag{2.2}$$

where:

UBV is the upper box value – the 75th percentile.

LBV is the lower box value – the 25th percentile.

o.c is the outlier co-efficient which was chosen as 1.5 in accordance with "accepted practice" (hypertext reference 1).

The box plot of the NO_x data set is shown in Figure 2.1. As can be seen from the plot, no data points lie outside the non-outlier maximum and minimum. This indicates that no significant outliers were detected. The value of the outlier co-efficient was reduced to *o.c* = 1 and only then were 14 points identified as outliers. It was felt that a value of *o.c* = 1 was excessively stringent and it was therefore decided to accept that the data set was free of any significant outliers.

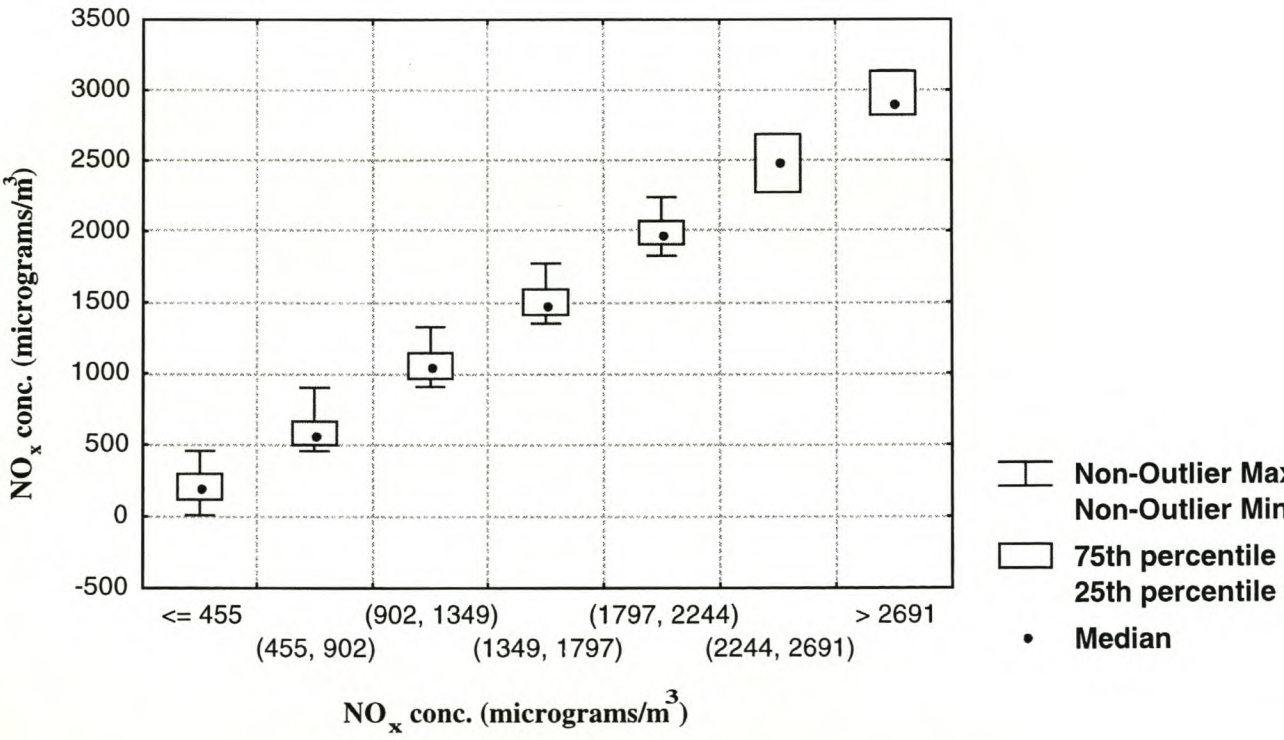


Figure 2.1 Box plot of the NO_x data set to identify outliers

The complete data set of NO_x concentrations, in micrograms per cubic metre ($\mu g/m^3$), is shown in Figure 2.2. The data set comprises 8760 hourly average values spanning the period 1 January to 31 December 1995.

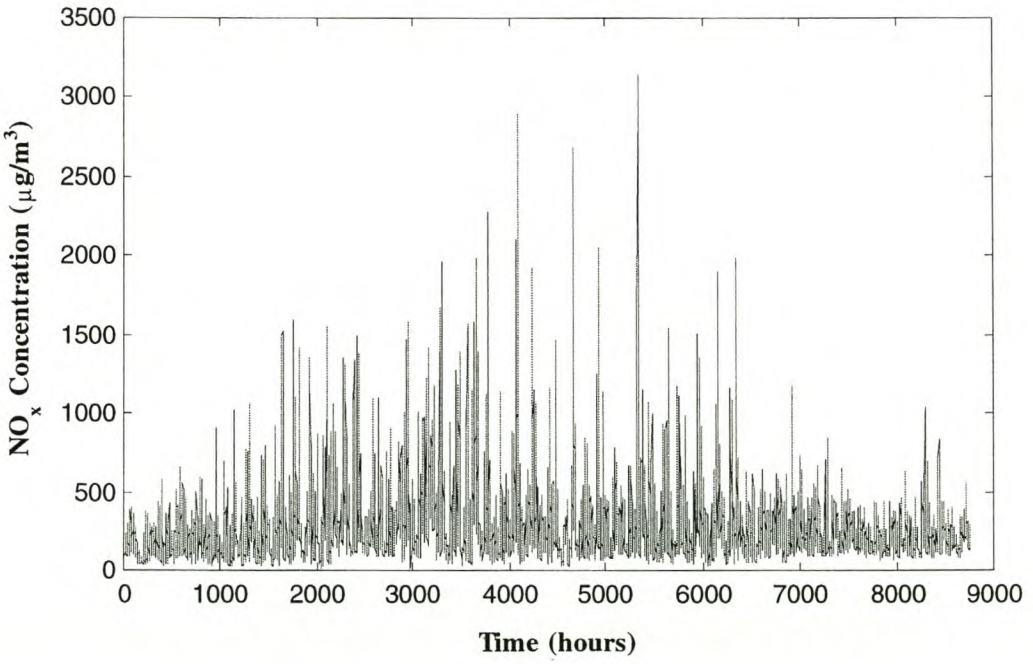


Figure 2.2 NO_x Data for 1995 (Hourly averages)

2.2 CHARACTERISTICS OF NO_x

NO and NO_2, collectively known as NO_x, are products of high-temperature, aerobic combustion. They are formed by the oxidation of atmospheric nitrogen. Combustion sources of NO_x are, among others, spark-ignition and compression-ignition engines, oil-fuelled power plants and tyre burning.

Previous studies of air pollution in the Cape Town region (Wicking-Baird *et al.*, 1997) have shown that vehicles contribute approximately 66% to NO_x emissions. NO_x data, obtained from the Cape Metropolitan Council's (CMC) Scientific Services Department, supports this statistic (Figure 2.3). As can be seen from the figure, the NO_x concentration appears to be correlated with traffic flow, with peaks at 09h00 and 18h00. Each peak occurs approximately

one hour after the corresponding peak traffic flow. This lag is probably associated with the transport and the chemical reaction of the NO / NO_2 system.

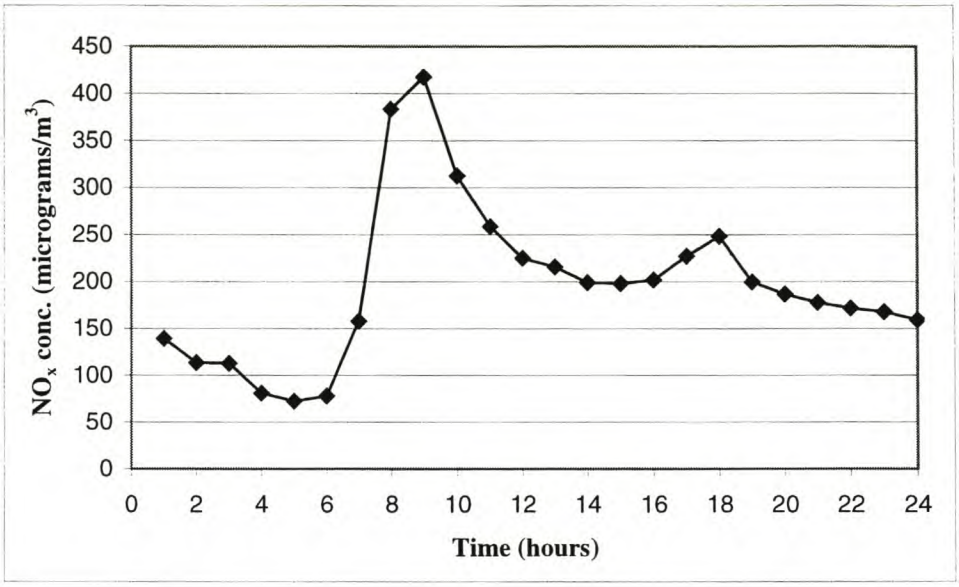


Figure 2.3 NO_x concentrations – Average hourly values for 1995

In the internal combustion engine, NO and small amounts of NO_2 are formed in the cylinder's combustion zone during the combustion stage. The high temperature in the combustion zone facilitates these reactions (Heywood, 1988). To exacerbate the problem, modern spark-ignition (petrol) engines without catalytic converters are designed to run lean of stoichiometric air/fuel mixtures to attain maximum fuel efficiency. The evaporative cooling effect of excess fuel is diminished, peak combustion temperatures are increased and excess air is present. These higher temperatures, in the presence of excess air, favour the formation of NO_x. Catalytic converters fitted to the exhaust system of a vehicle will reduce these NO_x emissions. In South Africa, however, they are found on only a small number of modern cars – mainly the luxury vehicle category that accounts for only a small percentage of the total vehicle population. NO_x emissions from compression-ignition (diesel) engines are less than, but comparable to the emissions from spark-ignition engines (Heywood, 1988). Catalytic converters have not yet been fitted to diesel vehicles on a commercial scale.

After the emission of the NO_x pollutant from the source, it is subjected to solar radiation. The molecules absorb light and convert this energy into molecular energy. This photochemical reaction results in the formation of NO_2 from NO (Grobliki *et al.*, 1981). Meteorological

factors cause the NO_x pollutant to be transported and dispersed in the atmosphere. This reaction cycle is one of the precursors to photochemical smog (Dzubay, 1982) – the familiar brown haze which plagues Cape Town.

In addition to NO_x contributing to photochemical smog, it is of direct concern to human health. Exposure to NO_x can cause a variety of effects including alterations of lung metabolism, lung structure and lung function, and increased susceptibility to pulmonary infections and emphysema-like effects. As a matter of interest, the recommended hourly mean NO_x level not to be exceeded, is $935\mu g/m^3$ (Bailie *et al.*, 1994).

3 ATMOSPHERIC DISPERION MODELS

Atmospheric dispersion models are constructed to obtain a representation of the physical process that governs pollutant concentration and dispersion. Knowledge of this physical process will enable a multitude of analyses to be carried out. Sensitivity analysis of the various pollution sources, forecasting the outcome of numerous scenarios and pollutant concentration prediction (so as to anticipate high pollution episodes) are some of the studies which could be undertaken. Necessarily, these models have to maintain a high level of accuracy since it is conceivable that important social and economic decisions could be based on results obtained from such models.

Essentially, there are three main approaches to dispersion modelling – the Eularian approach, the Lagrangian approach, and the statistical approach (Hassounah & Miller, 1994). These are discussed briefly in the sections that follow.

3.1 EULARIAN APPROACH

This approach uses the continuity equation to develop a description of the physical and chemical processes that govern the relationships between emissions and the resulting concentrations. It is a rigorous model that addresses physical and chemical processes from first principles. It requires a large amount of spatial, time-resolved information of emission sources. This approach is extremely complex and requires vast amounts of input data and processing power. In the near future, it is unlikely that South African cities will have the required spatial, time-resolved information of emission sources to apply the rigorous Eularian approach to dispersion modelling.

3.2 LAGRANGIAN APPROACH

This is the most frequently applied method to date due to the ease of application. The motion of the pollutant particles in the atmosphere is modelled using a probabilistic description, and this in turn is used to derive expressions for pollutant concentration. The most commonly used probabilistic description is the Gaussian plume model. This model has a few simplifying

assumptions that are restrictive (Turner, 1994). It assumes steady state conditions – the rate of emission is constant and the probability of the wind velocity is independent of time and location. Furthermore, the concentration of a pollutant along the vertical and crosswind axes is assumed to be normally distributed. These excessive simplifying assumptions restrict the accuracy of this method.

3.3 STATISTICAL AND MATHEMATICAL APPROACH

With this method, statistical and mathematical techniques are used to establish relationships between pollutant emissions, meteorological conditions and pollutant concentrations. The data requirements of this method are not as demanding as the Eularian approach's requirements and it requires substantially less computation power. The model could be run on a modern desktop computer whereas dispersion models that addresses physical and chemical processes from first principles often employ supercomputers to run the model. In addition, the statistical and mathematical approach is not subject to the excessive simplifications of the Lagrangian approach.

These methods allow single data records to be analysed to extract implicit information pertaining to the pollution system. This information, regarding the pollution system as a whole, is "embedded" in that single time series. The reasoning behind this is that a single record of a variable from a dynamic system (e.g. the pollution system) is the outcome of all interacting variables. Thus, theoretically, the single record should implicitly contain information about the dynamics of all the important variables involved in the evolution of the system (Elsner & Tsonis, 1996).

This study will make use of statistical and mathematical methods to analyse a single time series, the NO_x data. Two analysis techniques, namely, Non-linear Time Series Analysis and Singular Spectrum Analysis (SSA) will be applied to the NO_x data set to construct prediction models. The prediction ability of the respective model is assumed to be indicative of the accuracy of the model and will be used as the basis against which the two techniques are evaluated. The construction of these prediction models from a single time series provide a means of achieving the aim of this study – to identify the technique which is best suited to the modelling of air pollution data.

Once the most accurate and robust analysis technique has been identified, further work could involve the extension of the univariate model to a more complex multivariate model that explicitly encompasses variables such as traffic flow and weather patterns. Such a model will be of great practical value in that it will have the ability to perform sensitivity analyses and predict the outcome to a multitude of scenarios.

The two analysis techniques will be presented in separate chapters. The theory, and the application of the technique on the NO_x data set, will be presented and discussed within the respective chapters.

4 NON-LINEAR TIME SERIES ANALYSIS

Many systems do not obey the linear paradigm of small causes lead to small effects. These non-linear systems often display irregular behaviour that cannot be quantified using linear analysis methods. Linear methods will attribute this irregular behaviour to a random external input whereas, in fact, this irregular behaviour is often an inherent part of the non-linear system dynamics. The use of linear methods could lead to a signal being classified as stochastic although it displays determinism i.e. it is not random and it can be predicted. During the past two decades, important progress has been made in the field of Non-linear Time Series Analysis. An entire "toolbox" of methods is available to assist in identifying non-linear determinism and to enable the construction of a predictive model. The framework of this technique constitutes an approach to analysing and extracting information from systems that display non-linear determinism.

It is not too far a stretch of the imagination to assume that the pollution system is governed by non-linear dynamics. The process is strongly correlated to weather patterns that are by no means linear. In addition, complex chemical reactions play a vital role in pollutant formation and destruction. For these reasons, Non-linear Time Series Analysis was selected as one of the techniques to model the NO_x data set.

4.1 INTRODUCTION TO NON-LINEAR TIME SERIES ANALYSIS

To introduce the idea of a non-linear system and its characteristics, the non-linear set of Lorenz equations (4.1) is examined (Lorenz, 1963). Lorenz formulated these three ordinary differential equations as an approximation to the partial differential equations describing thermal convection in the lower atmosphere as derived by Salzmann (1962).

$$\begin{aligned} \dot{x} &= -\sigma(x-y) \\ \dot{y} &= -xz + rx - y \\ \dot{z} &= xy - bz \end{aligned} \tag{4.1}$$

where x, y and z are the state variables representing convective overturning, horizontal temperature variation and vertical temperature variation respectively. The values b, σ and r are control parameters.

The Lorenz model is an example of a non-linear *deterministic* system, i.e. there is no randomness associated with it and it can be predicted. However, a different conclusion is reached if this system is analysed using linear techniques. The time series of the x-state from the Lorenz model is plotted in Figure 4.1a. Notice the irregular behaviour displayed by this equation set. Figure 4.1b shows the Fourier spectrum (a linear analysis technique) of the signal.

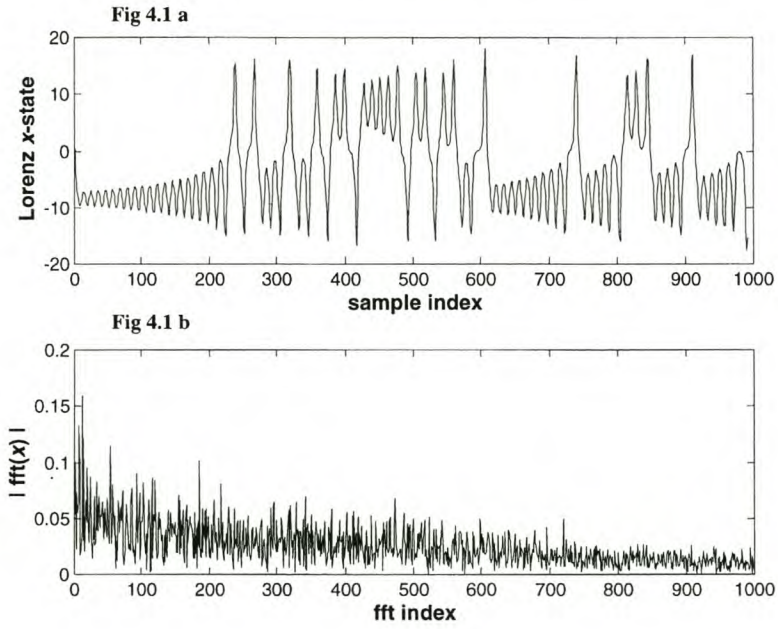


Figure 4.1 Broadband Fourier spectrum of the x-state of the non-linear Lorenz system (Diab & Barnard, 1999)

The broadband spectrum would lead one to classify the data as *stochastic* i.e. random (Addison, 1997). However, it is known that the Lorenz model, described by three finite differential equations, is not stochastic but displays non-linear determinism – it is not a random system and it can be predicted. The reason for the stochastic appearance is that non-linear systems do not have a finite range of fundamental periods. The state vector that describes the system does not follow a closed trajectory in phase space. This results in the lack of a finite range of fundamental periods (an example of a state vector, which does not follow a closed trajectory, can be seen in Figure 4.3 on page 16). It is for this reason that the Fourier analysis of the time series shows a broadband spectrum. Such a spectrum would generally be indicative of a stochastic signal (noise).

The fact that a linear analysis technique identifies a non-linear deterministic signal as stochastic emphasises the need to apply non-linear analysis techniques to non-linear systems. In recent years, techniques have been developed that correctly identify the apparent randomness in these non-linear time series as non-linear determinism (Abarbanel, 1996). These techniques can be applied to measured observations of a system to extract the important underlying dynamics that govern the system. Having obtained a handle on the dynamics, a model can be built which is a representation of the underlying physical situation.

The same broadband spectrum observed for the Lorenz equations is obtained for the Fourier spectrum of the NO_x data. See Figure 4.2 below (actually, this is the power spectrum – *periodogram* – which is the square of the Fourier values).

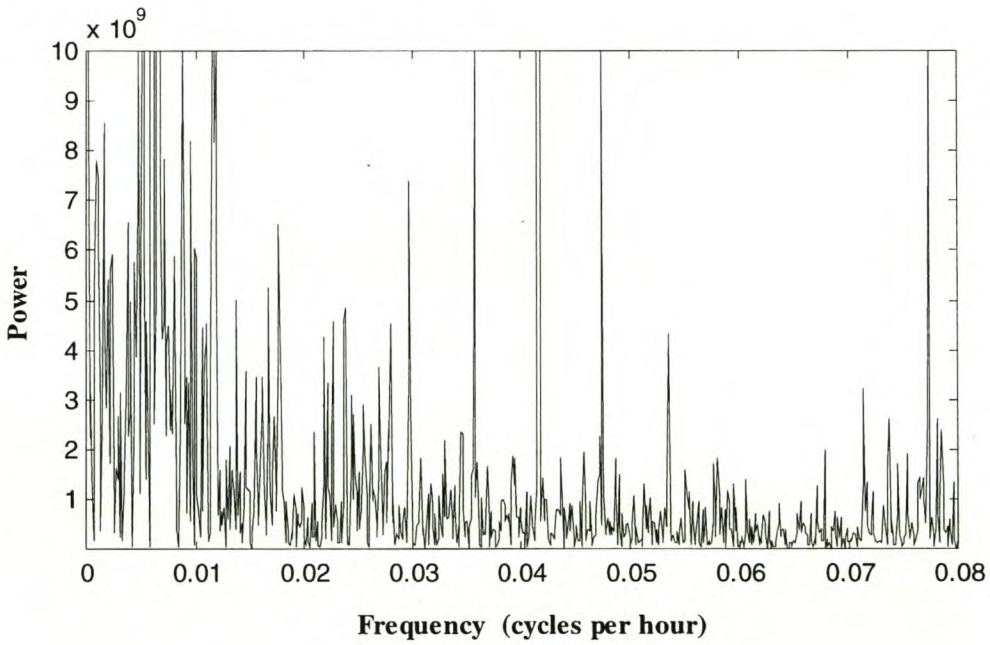


Figure 4.2 Periodogram of the NO_x Data

Although this spectral broadening can be considered a hallmark of a non-linear system, it is by no means sufficient proof. As mentioned before, broadband spectral components may also arise from stochastic signals. Appropriate non-linear techniques have to be applied to the data to identify and analyse non-linear systems correctly.

4.2 THE DYNAMIC ATTRACTOR

To understand non-linear systems, it is necessary to understand their properties – a very important one being the *dynamic attractor*. The system dynamic attractor is the cornerstone of Non-linear Time Series Analysis. This geometrical construction of the system's state vectors, plotted in phase space, reveals the dynamics of the system. In other words, a *phase space* is created where specifying a point in the space specifies the state of the system, and vice versa.

A dynamic system can be represented mathematically by a state equation in a number of state variables, i.e. a state vector $\boldsymbol{x}$ in a finite-dimension phase space $\Re^m$:

$$\boldsymbol{x}_{n+1} = \boldsymbol{F}(\boldsymbol{x}_n) \quad n \in \text{ length of time series} \tag{4.2}$$

where $\boldsymbol{x}_n$ is the state vector, $\boldsymbol{x}_{n+1}$ is the observed output and $\boldsymbol{F}(\,)$ is the measurement function. Starting from some initial conditions, a sequence of points $\boldsymbol{x}_n$ satisfying the above equation (4.2), follows a trajectory that is confined to some closed subspace of the total available state space (Farmer *et al.*, 1983). These trajectories, in their entirety, constitute an m-dimensional map in $\Re^m$ that describes the dynamics of the system – *the attractor*. The behaviour that is observed depends on $\boldsymbol{F}$ and on the initial conditions. However, after transients have died out, all solutions are drawn to the same closed subspace – the *basin of attraction* – which forms the system's *dynamic attractor*.

As an example, refer to the Lorenz system of equations (4.1). By setting $b = 2.67$, $\sigma = 10.00$ and $r = 28.00$, and solving the equations using a 4^{th} order Runge-Kutta procedure, the state variables x, y, z can be calculated. Refer to Figure 4.3. The state variables x, y, z of the Lorenz system are plotted against each other, rather than against time.

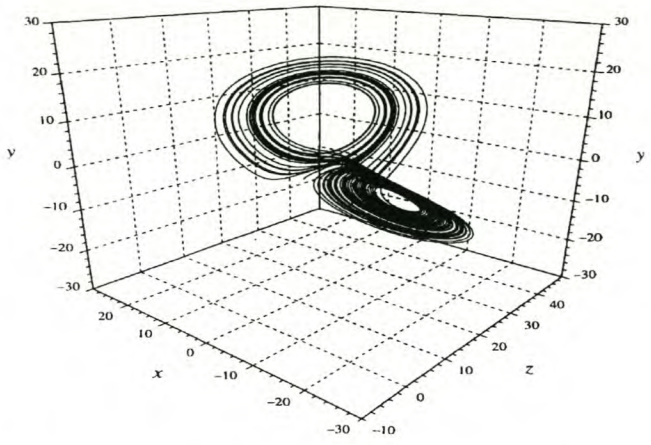


Figure 4.3 The Dynamic Attractor of the Lorenz System (Addison, 1997)

Starting from many initial conditions, the solutions of the system are attracted to the same closed subspace in the total available state space. This is a plot of the long-term behaviour of the Lorenz system and is known as the *dynamic attractor*. The attractor, to which the trajectory converges, is a smooth non-linear manifold of this state space and defines the true dynamics of the system. Identification of the system involves determining the functional relationship of points on the attractor with the observed states of the system (Sauer *et al.*, 1991). Referring to equation (4.2) above, this means determining a suitable measurement function $\boldsymbol{F}(\,)$.

4.3 DIMENSION OF THE ATTRACTOR

The dimension of an attractor is a way of quantifying the properties of a signal by representing a sequence of data as a single number. The dimension is the most basic property of an attractor and is the first level of knowledge necessary to characterise its properties. It can be seen as an indication of the amount of information necessary to specify the position of a point on the attractor to within a certain accuracy (Farmer *et al.*, 1983). In other words, it is a statistic that should enhance knowledge of the underlying system. It does not depend significantly on the measurement procedure, chosen co-ordinates etc. and is thus referred to as an *invariant* (Kantz & Schreiber, 1997).

A characteristic dimension often used is the *correlation dimension*, d_c, defined as

$$d_c = \lim_{\varepsilon \to 0} \lim_{N \to \infty} \frac{\log C_N}{\log \varepsilon} \tag{4.3}$$

N is the number of observations in the data set. C_N is the correlation function defined as the fraction of all possible pairs of points that are closer than a given distance ε in a particular norm:

$$C_N(\varepsilon) = \binom{N}{2}^{-1} \sum_{0 \le i < j \le N} I\left(\left\| s_i - s_j \right\| < \varepsilon\right) \tag{4.4}$$

s is the scalar measurement, I(||·||) is a Heavyside function that returns one if the distance between point i and j is within a box of size ε, and zero otherwise. The size of ε is limited from below by the finite accuracy of the data and by the inevitable lack of near neighbours at small length scales. N is limited by the sample size.

The conventional approach to estimating the correlation dimension is the Grassberger-Procaccia algorithm (Grassberger & Procaccia, 1983 and 1983a). The correlation function is calculated for many different box sizes ε. A plot of $log\ C_N(\varepsilon)$ versus $log(\varepsilon)$ is constructed and, in theory, the slope of the graph in the limit as $\varepsilon \to 0$ should approach the correlation dimension. This presents a dilemma when working with finite data sets. As ε approaches a minimum value, the graph line begins to fluctuate as a result of the small number of points contributing to the correlation sum. For this reason, the graph is analysed at slightly larger values of ε where the slope of $log\ C_N(\varepsilon)$ remains relatively constant. This region is known as the *scaling region*. The correlation dimension is obtained by finding the gradient of a best-fit line fitted to the scaling region. See Figure 4.4.

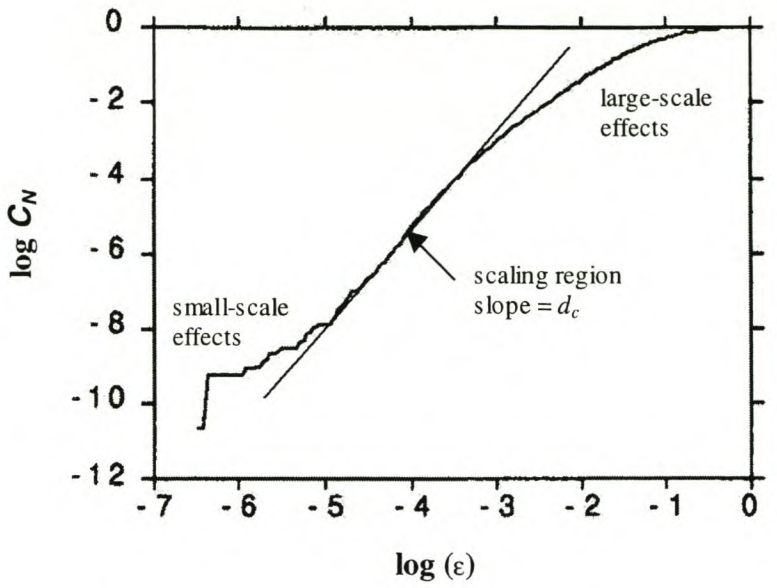


Figure 4.4 Determination of the Correlation Dimension of a data set (Judd, 1992 and Addison, 1997)

Judd (1992), however, has deemed this method problematic and points out certain shortcomings. He points out that often a scaling region is not straight but curved. Since a small change in the scaling region significantly changes the dimension estimate, the estimate will be very dependent on the choice of the scaling region on the curved surface. This necessarily leads to uncertainty in the correlation dimension estimate.

Judd (1992 and 1994) has proposed an algorithm that eliminates the necessity to choose a scaling region. Rather, a polynomial of the order of the topological dimension is fitted in that region. Furthermore, to provide a reliable dimension estimate, the algorithm requires fewer data points than the Grassberger-Procaccia approach (Judd, 1994). The Judd algorithm will be utilised in this study.

As a matter of interest, the correlation dimension calculated for the Lorenz attractor in Figure 4.3 is $d_c \approx 2.06$. Note that this value is not an integer. Such fractal dimensions are typical of non-linear systems.

4.4 ATTRACTOR RECONSTRUCTION FROM EXPERIMENTAL DATA

It was emphasised that the dynamic attractor in phase space is the cornerstone of non-linear analysis since it represents the underlying dynamics of the system. When dealing with experimental data, however, there is no phase space object that represents the dynamics. The measured values are not the actual states, $\boldsymbol{x}_n$, but rather a scalar sequence of measurements s_n which functionally depend on these states. This can be represented as:

$$s_n = s(\boldsymbol{x}_n) \qquad n = 1,2,\ldots,N \tag{4.5}$$

where $\boldsymbol{x}_n$ is the actual state vector, s_n is the scalar measurement and $s(\ )$ is the measurement function analogous to $\boldsymbol{F}(\ )$ in equation (4.2).

The task at hand is to reconstruct an attractor from state vectors $\boldsymbol{s}_n$ derived from the scalar measurements s_n. This reconstructed attractor should be a suitable representation of the system's underlying dynamic attractor. The method, by which the measured observations s_n are converted into state vectors $\boldsymbol{s}_n$ to enable attractor reconstruction, is known as the *method of delay co-ordinates*.

According to Takens (1981), in the absence of noise and with a sufficient amount of data, an equivalent representation of the system state space can be reconstructed from a time series observation of a single observed output. Such a reconstruction is called an *embedding* of the observed time series by way of delay co-ordinates (equivalent state variables $\boldsymbol{s}_n$). The number of these co-ordinates is the embedding dimension, m, and the time delay, k, is the delay or lag between each co-ordinate. A discussion of the theoretical background to the embedding of time series can be found in a paper by Osborne & Provenzale (1989).

If the attractor of the data is reconstructed in a phase space of suitable embedding dimension m, and a suitable time delay k is selected, then the differential information of the underlying dynamics will be preserved (Sauer *et al.*, 1991). The fundamental idea is to ensure that the attractor has unfolded itself completely and that there are no parts of it that overlap (a trajectory crossing itself on return) which would result in some of the dynamics being hidden. This reconstructed attractor is essentially a pseudo-attractor of dimension m that gives us insight into the dynamic properties of the underlying attractor of the system. The underlying

attractor has dimension d_c. Note that m is an integer, although d_c need not be. The requirement for an accurate embedding dimension estimate is $m > d_c$ (Takens, 1981).

The time lag k between the state variables is usually determined by the method of Average Mutual Information (Fraser & Swinney, 1986), while the embedding dimension m is typically calculated using the method of False Nearest Neighbours (Kennel *et al.*, 1992). Both of these techniques will be discussed in ensuing sections.

After embedding the time series, the m dimensional embedding vector, s_n, constructed from the time series, $s_n(n = 1,2,3\ldots N)$, is defined as follows:

$$s_n = (s_{n-(m-1)k} , s_{n-(m-2)k} , s_{n-(m-3)k} ,\ldots, s_{n-k} , s_n) \qquad n = (m-1)k+1, (m-1)k+2,\ldots,N \tag{4.6}$$

As an example, consider the following time series sampled every 60 seconds:

$$[2.5;\ 1.3;\ 5.6;\ 12.3;\ 4.2;\ 7.8;\ 9.1;\ 4.2;\ 2.3;\ 4.1;\ 8.9;\ \ldots]$$

If the embedding dimension m was estimated as 3 and the time delay k was estimated as 2 sampling periods (120 seconds), then n would run from $n = 5,\ldots,N$ with s_5 being the first reconstructed co-ordinate. The first three co-ordinates of the reconstructed attractor would be the set:

$$\{s_5 = (2.5, 5.6, 4.2); s_6 = (1.3, 12.3, 7.8); s_7 = (5.6, 4.2, 9.1)\}$$

These embedding co-ordinates, plotted in $\Re^m$ phase space $(m = 3)$, constitute the reconstructed attractor.

4.5 SELECTING A TIME DELAY *k*

From a mathematical point of view, the time delay k can be chosen as any arbitrary value if an infinite amount of noise-free data is used (Takens, 1981). In practice, however, it has been shown that the quality of the phase space reconstruction is dependent on the choice of k (Fraser & Swinney, 1986).

If the time delay is too short, the co-ordinates s_n and s_{n+k} which make up the reconstructed state vector $\boldsymbol{s}_n$, will not be independent enough. Not enough time will have evolved between s_n and s_{n+k} to have explored enough of the system's state space and produce new information about the system's dynamic attractor (Abarbanel, 1996).

On the other hand, if the time delay is too long, there will be very little or no connection between the values s_n and s_{n+k}. The values will essentially be random with respect to each other and attempts to construct a meaningful attractor will be futile.

Numerous techniques to identify the optimum time delay have been proposed. Two methods, which are frequently used, are discussed.

4.5.1 The Autocorrelation Function

This function compares two data points in a time series separated by a delay k. It is defined as follows:

$$C(k) = \frac{\sum_{n=1}^{N-k} (s_n - \bar{s}_n)(s_{n+k} - \bar{s}_n)}{\sum_{n=1}^{N-k} (s_n - \bar{s}_n)^2} \tag{4.7}$$

where $\bar{s}_n$ denotes the average value of s over time. $C(k)$ is a dimensionless value.

The delay is chosen as the value of k at some threshold value of $C(k)$. The most popular approach has been to choose k as the first time lag where $C(k)$ is equal to zero (Addison, 1997). This is equivalent to requiring linear independence between the two data points (Fraser & Swinney, 1986). The application of this technique to non-linear systems has been criticised since the autocorrelation function measures the linear dependence of two variables. Abarbanel (1996) illustrates that the use of the autocorrelation function for determining the time delay can be misleading.

Application to NO_x Data

A plot of the autocorrelation function of the NO_x data set is shown in Figure 4.5. The first time lag where $C(k)$ is equal to zero occurs at a time lag of $k = 41$ hours. Other choices could be $k = 19$ hours or $k = 30$ hours since the value of $C(k)$ is close to zero. The first minimum is sometimes used as an indication of a suitable time lag and this would be at $k = 7$ hours (Addison, 1997). It should be remembered, though, that the autocorrelation function is not the ideal measure for a suitable time lag.

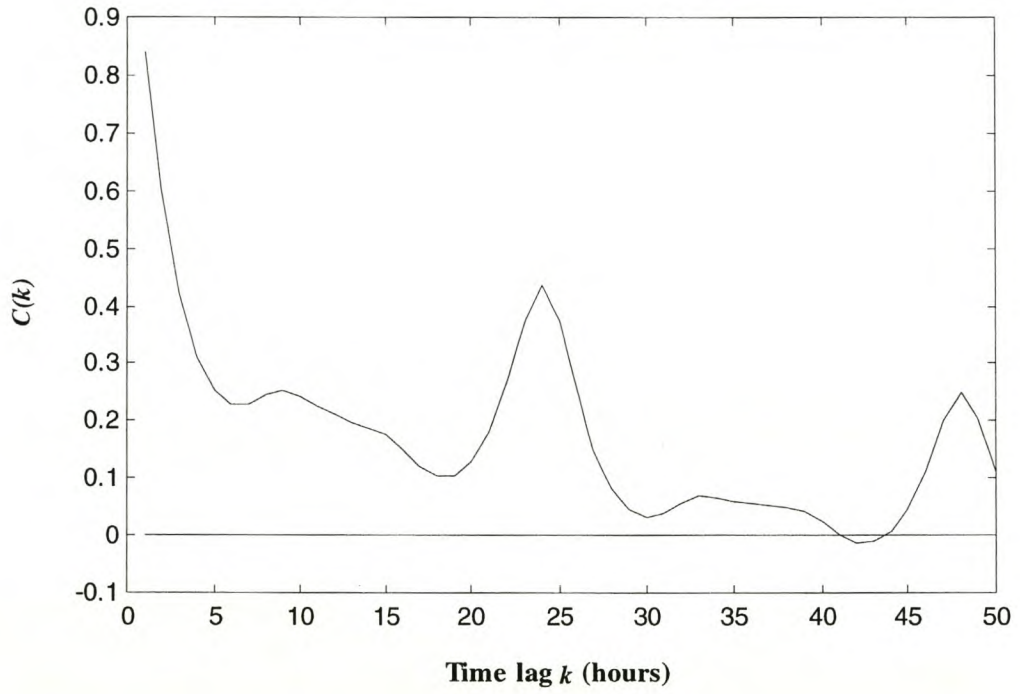


Figure 4.5 The Autocorrelation Function for the NO_x Data

4.5.2 The Method of Average Mutual Information (AMI)

Whereas the autocorrelation function measures the linear dependence of two variables, the mutual information function measures the general dependence of two variables. It is an indication of the amount of information that is possessed about the value of s_{n+k}, k time steps later, if s_n is known. The AMI is calculated using an algorithm proposed by Fraser & Swinney (1986).

The average mutual information, $I(k)$ (measured in bits), between the observation s_n and the observation s_{n+k}, k time steps later, is given by:

$$I(k) = \sum_{s_n, s_{n+k}} P(s_n, s_{n+k}) \log_2 \left[\frac{P(s_n, s_{n+k})}{P(s_n) P(s_{n+k})} \right] \tag{4.8}$$

where $P(\ )$ is the probability function.

Fraser & Swinney (1986) suggest that the AMI function be used as a kind of non-linear autocorrelation function. The function can be used to determine when the values of s_n and s_{n+k} are independent enough of each other to be useful as co-ordinates in a reconstructed state vector. However they should not be so independent so as to have no connection with each other at all (Abarbanel, 1996). $I(k) = 0$ implies that the data is completely stochastic, whilst $I(k) = \infty$ indicates complete dependence between the data. It has been suggested by Shaw that a suitable choice of time delay requires the mutual information to be a minimum (in Fraser & Swinney, 1986). Specifically, the value of k at the first local minimum of mutual information should be used as the time delay.

Application to NO_x Data

The AMI plot is shown in Figure 4.6. The AMI was calculated using the algorithm from the TISEAN© computer package of Hegger *et al.* (hypertext reference 2). As can be seen, the AMI drops off rapidly as k increases to around $k = 10$ hours. At this stage it is near zero and gradually decreases until $k = 19$ hours, where after it increases again. The change in $I(k)$ from $k = 10$ hours to $k = 19$ hours is approximately 1.5%. Considering this small change in the magnitude of $I(k)$ and the fact that $I(k)$ is very near zero, it was decided that a time lag of $k = 10$ hours would be suitable.

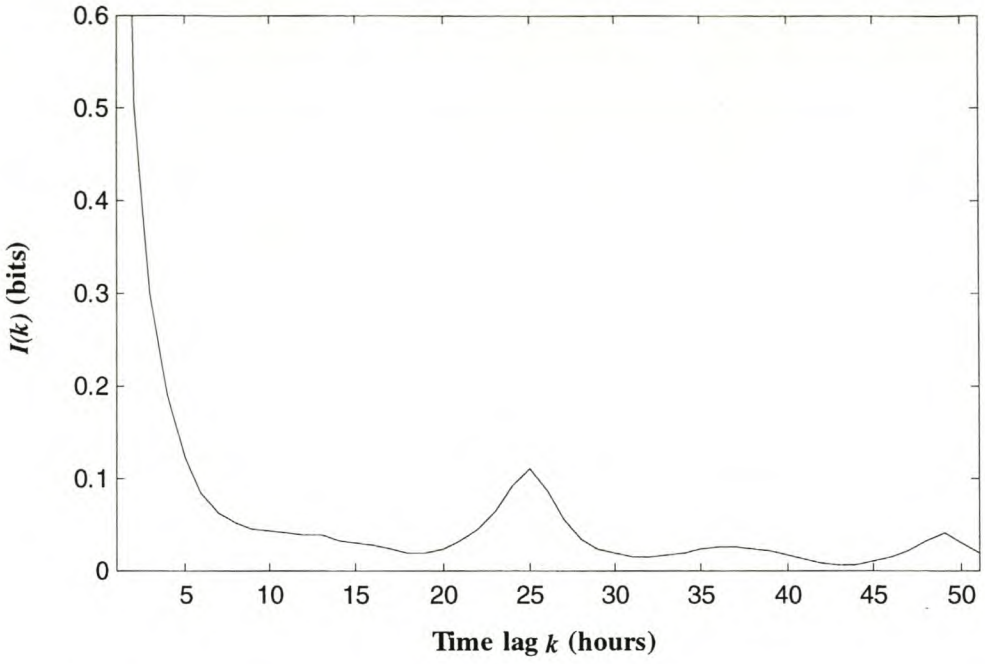


Figure 4.6 Average Mutual Information for the NO_x Data

It should be mentioned that these methods merely provide a practical indication as to the time delay that should be used. The time delay k has no relevance in the mathematical framework and hence there exists no rigorous way of determining its optimal value, or what properties the optimal value should have. In the mathematical sense, and for noise-free data, the embedding is not sensitive to the choice of the time delay (Kantz & Schreiber, 1997). In practice, however, a good choice of time delay facilitates analysis of the system. It is therefore sagacious to apply the available techniques to identify a time delay that will benefit the reconstruction of the dynamic attractor.

4.6 DETERMINING THE EMBEDDING DIMENSION *m*

A suitable embedding dimension, m, has to be determined to ensure that there are a sufficient number of co-ordinates in phase space to adequately contain the reconstructed attractor. If m is not sufficiently large, there will be too few co-ordinates to unfold the attractor and parts of the attractor will overlap. This overlapping of trajectories will result in some of the dynamics being hidden. The lowest dimension, which unfolds the attractor so that none of these overlaps remain, is known as the embedding dimension m (Abarbanel, 1996). Takens (1981) noted that if the underlying attractor has dimension d_c, the requirement for an accurate

embedding dimension estimate is $m > d_c$. From a mathematical point of view, any dimension $d \geq m$ will be a suitable embedding dimension since the attractor will be completely unfolded. However, from a practical perspective, working in a dimension that is larger than the required minimum dimension, results in the need for excessive computational power when further analysis of the system is carried out. Also, the extra d - m dimensions will be dominated by high dimensional system "noise" which further contaminates the system (Kennel *et al.*, 1992).

4.6.1 The Method of False Nearest Neighbours (FNN)

The method of false nearest neighbours, described by Kennel *et al.* (1992), determines a suitable minimum embedding dimension by looking at the behaviour of near neighbours (points near to each other) under changes in the embedding dimension. The state vector, to be used in the phase space reconstruction of the attractor in dimension d, is:

$$\boldsymbol{s}_n = (s_{n-(d-1)k} , s_{n-(d-2)k} , s_{n-(d-3)k} ,..., s_{n-k} , s_n) \tag{4.9}$$

Average mutual information was used to determine the time delay k. The nearest neighbour in phase space $\Re^m$ to the state vector $\boldsymbol{s}_n$ is:

$$\boldsymbol{s}_n^{NN} = (s_{n-(d-1)k}^{NN}, s_{n-(d-2)k}^{NN}, s_{n-(d-3)k}^{NN} ,..., s_{n-k}^{NN}, s_n^{NN}) \tag{4.10}$$

In dimension d, the possibility exists that $\boldsymbol{s}_n^{NN}$ was projected into the neighbourhood of $\boldsymbol{s}_n$ because of the attractor not being completely unfolded in that particular dimension. Moving from dimension d to dimension d+1 will further unfold the attractor and eliminate self-crossings of trajectories. This unfolding will result in $\boldsymbol{s}_n^{NN}$ being projected out of the neighbourhood of $\boldsymbol{s}_n$. It can thus be declared that $\boldsymbol{s}_n^{NN}$ is a *false neighbour* of $\boldsymbol{s}_n$. If $\boldsymbol{s}_n^{NN}$ is truly a neighbour of $\boldsymbol{s}_n$, it will remain in the neighbourhood of $\boldsymbol{s}_n$ through projection from dimension d to d+1. This is because $\boldsymbol{s}_n^{NN}$ arrived in the neighbourhood of $\boldsymbol{s}_n$ as a result of system dynamics and not because of false projection in too low a phase space dimension. There are two criteria that classify $\boldsymbol{s}_n^{NN}$ as a false neighbour.

First Criterion: To determine the minimum embedding dimension m, every state vector $\boldsymbol{s}_n$ is examined to determine in what dimension all false neighbours are removed. When $\boldsymbol{s}_n$ is only surrounded by true neighbours, the minimum embedding dimension m has been determined. The distance between the state vectors $\boldsymbol{s}_n$ and $\boldsymbol{s}_n^{NN}$ in dimension d has to be compared to the distance between the same vectors in dimension d+1 to ascertain whether these neighbours are true or false.

Note that in going from dimension d to d+1, the additional component of the state vector $\boldsymbol{s}_n$ is just $s_{n\text{-}dk}$ and the additional component of the vector $\boldsymbol{s}_n^{NN}$ is s_{n-dk}^{NN} (Abarbanel, 1996). Hence it is only necessary to compare $\left|s_{n\text{-}dk} - s_{n\text{-}dk}^{NN}\right|$ with the Euclidian distance $\left|\boldsymbol{s}_n - \boldsymbol{s}_n^{NN}\right|$ between nearest neighbours in dimension d. If the additional distance $\left|s_{n\text{-}dk} - s_{n\text{-}dk}^{NN}\right|$ in dimension d+1 is large compared to the distance $\left|\boldsymbol{s}_n - \boldsymbol{s}_n^{NN}\right|$ between nearest neighbours in dimension d, then $\boldsymbol{s}_n^{NN}$ can be considered a false neighbour. Mathematically, the square of the Euclidian distance between a state vector $\boldsymbol{s}_n$ and a nearest neighbour $\boldsymbol{s}_n^{NN}$ as a function of dimension d is (Kennel *et al.*, 1992):

$$R(d)_n^2 = \sum_{d=1}^{d} [s_{n-(d-1)k} - s_{n\text{-}(d\text{-}1)k}^{NN}]^2 \tag{4.11}$$

In dimension d+1 this distance is:

$$R(d+1)_n^2 = \sum_{d=1}^{d+1} [s_{n-(d-1)k} - s_{n\text{-}(d\text{-}1)k}^{NN}]^2 \;\; = R(d)_n^2 + \left|s_{n-dk} - s_{n-dk}^{NN}\right|^2 \tag{4.12}$$

The criterion for a false neighbour is defined by re-arranging equation (4.12) and dividing by $R(d)_n^2$:

$$\sqrt{\frac{R(d+1)_n^2 - R(d)_n^2}{R(d)_n^2}} = \frac{\left|s_{n-dk} - s_{n-dk}^{NN}\right|}{R(d)_n} > R_{threshold} \tag{4.13}$$

Equation (4.13) is a ratio of the distance between points in dimension d+1, relative to the distance in dimension d. Kennel *et al.* (1992) have shown that a false neighbour is identified for a threshold value of $R_{threshold} \geq 10$. Abarbanel (1996), after examining a large variety of systems, suggests that a value $R_{threshold} \geq 15$ defines a false neighbour.

Second Criterion: The above criterion, on its own, is not sufficient for determining a minimum embedding dimension m. When dealing with a finite amount of data, the problem exists that although $\mathbf{s}_n^{NN}$ is the nearest neighbour to $\mathbf{s}_n$, it may not necessarily be close to $\mathbf{s}_n$. As the embedding dimension increases, a situation could occur where the distance $R(d)_n$ between two points is comparable to the size of the attractor R_A. This is a direct result of trying to uniformly populate an object in d dimensions with a fixed number of points – the points necessarily move further and further apart as d increases (Kennel *et al.*, 1992).

The second criterion deals with the issue of a finite data size. The distance added between neighbours, in going from dimension d to dimension $d+1$, should not be larger than the nominal size of the attractor R_A.

Kennel *et al.* (1992) state this criterion as: "*If the nearest neighbour to* $\mathbf{s}_n$ *is not close* $[R(d)_n \approx R_A]$ *and it is a false neighbour, then the distance* $R(d+1)_n$ *resulting from adding on a* $(d+1)^{\text{th}}$ *component to the data vectors will be* $R(d+1)_n \approx 2R_A$."

Mathematically, this second criterion is written as:

$$\frac{R(d+1)_n}{R_A} > A_{threshold} \tag{4.14}$$

where $A_{threshold}$ is a number of order 2. It has been ascertained that the results of determining the embedding dimension m are fairly insensitive to the values of $R_{threshold}$ and $A_{threshold}$, as long as the data set is not too small (Abarbanel, 1996).

R_A is measured as follows:

$$R_A^2 = \frac{1}{N}\sum_{n=1}^{N}[s_n - \bar{s}_n]^2 \tag{4.15}$$

where

$$\bar{s}_n = \frac{1}{N}\sum_{n=1}^{N} s_n \tag{4.16}$$

If either of the above criteria is violated, a near neighbour is declared *false*.

4.6.2 The effect of noise

The subject of "noise" will be addressed more comprehensively in section 4.8, but it is necessary to mention it here because of its effect on determining the minimum embedding dimension. Experimental data always contains a random element known as noise. This may be *dynamical noise* – due to random fluctuations in the dynamical system – or *measurement noise* – due to random errors added to the system by the measurement system. By its very nature, this noise contamination is infinite dimensional since it is a sequence of random numbers. Because it is of a higher dimension than the dynamic signal, it would want to be unfolded in a higher dimension than the signal of interest. This would mean, that in the presence of noise, the percentage of false nearest neighbours will not drop to zero in any dimension where there is sufficient data to examine near neighbours.

What this implies is that, although the system dynamics may be unfolded in a dimension d, the high dimensional noise will still result in false neighbours being detected at that dimension. Kennel *et al.* (1992) and Abarbanel (1996) have shown that the method of False Nearest Neighbours is rather robust against noise contamination.

4.6.3 Application to the NO_x Data

The method of False Nearest Neighbours was applied to the NO_x data using the algorithm from the TISEAN© computer package of Hegger *et al.* (hypertext reference 2). Adjustable parameters were chosen in accordance with the methods suggested by Hegger *et al.*. The results of the computation are shown in Figure 4.7. The percentage of false neighbours drops off rapidly up to dimension 10, after which it appears that a "floor" is reached. The signal's relatively high noise content could be the result of this "noise floor". As mentioned above in section 4.6.2, the percentage of false nearest neighbours will not drop to zero in the presence of noise which is of a high dimension. Small & Judd (1998) encounter a similar situation in their analysis of infant respiration data. A plateau was reached when they performed a FNN calculation on the data set. The value they selected as the embedding dimension, was the lowest value of m at the point where the plateau started.

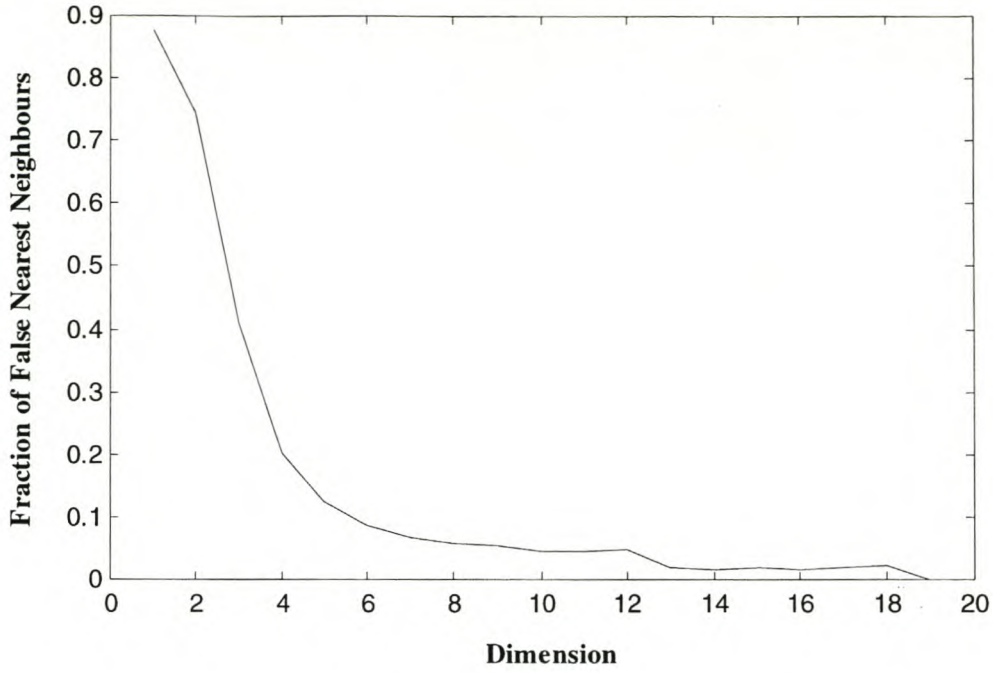


Figure 4.7 The Fraction of False Nearest Neighbours (FNN) for the NO_x Data

Looking at Figure 4.7, it appears as though the fraction of FNN decreases to a plateau with a minimum at $m = 11$. At dimension $m = 14$ the fraction of FNN forms another plateau which finally disappears to zero at dimension $m = 19$. In keeping with the rationale of Small & Judd, it was decided to opt for an embedding dimension of $m = 10$ to avoid unfolding too much of the noise content.

4.7 THE USE OF SURROGATE DATA IN TESTING FOR NON-LINEAR DETERMINISM

The application of Non-linear Time Series Analysis techniques to an observed series of measurements may seem an astute approach, but it is a futile exercise if the data does not possess non-linear determinism. It could be that the complex, irregular behaviour, which is initially thought to be a result of non-linear dynamics, is caused by linear stochastic processes. If this were the case, it would be judicious to rather use a linear stochastic modelling technique such as an Auto-regressive Moving Average (ARMA) model or a Markov model. It is, therefore, important to positively identify non-linear determinism in the data set.

The method of surrogate data, which has been well documented by Theiler *et al.* (Theiler *et al.*, 1992 and Theiler & Pritchard, 1996), serves to identify non-linearity in a time series. It is a Monte Carlo procedure, the fundamentals of which are as follows (Theiler & Pritchard, 1996):

i. Specify a *null hypothesis* against which the data set is to be tested. For example, in testing for non-linearity, a null hypothesis would be that the data set results from some Gaussian *linear* stochastic process.

ii. Generate numerous *surrogate data* sets which are consistent with the null hypothesis but which are comparable to the measured data in certain specified respects (Kantz & Schreiber, 1997)

iii. Compute a suitable *discriminating statistic* that characterises the original time series and the surrogate data sets.

iv. Compare the discriminating statistic calculated for the original data set with those calculated from the surrogate data sets. If the statistic calculated for the original data is markedly different from the set of statistics obtained for the surrogate data sets, the null hypothesis is rejected. Rejecting the null hypothesis amounts to the detection of non-linearity since it would mean, for example, that the original data set is *not* based on some Gaussian linear stochastic process as assumed by the null hypothesis.

A decision has to be made as to which null hypotheses are to be used, and which discriminating statistic will be calculated for comparison.

4.7.1 The Null Hypothesis

It is necessary to specify a null hypothesis against which the original data set can be tested. The object is to identify non-linearity and, thus, the data should be tested to see if it results from some linear stochastic process.

The null hypotheses for testing against linear systems can be divided into three groups:

Type(0): The data is temporally uncorrelated i.e. the data is identically and independent distributed (iid) noise. This is achieved by generating a random Guassian time series and re-ordering this time series so that it is of the same rank as the original data set.

Type(1): The data is linearly filtered noise. This entails phase randomisation of the Fourier spectrum of the data set.

Type(2): The data is a monotonic non-linear transformation of linearly filtered noise. This is essentially a combination of the procedures followed in generating type(0) and type(1) surrogates. The procedure involves generating a Guassian time series y_n and re-ordering it so that it has the same rank as the original time series s_n (type(0)). Next, phase-randomise the Fourier spectrum of y_n to obtain a time series y'_n. Finally, the original time series s_n is time re-ordered so that it has the same rank as y'_n. This time re-ordered time series is a surrogate of the original time series with a matching amplitude distribution. This type of surrogate is known as an *Amplitude Adjusted Fourier Transform* (AAFT) surrogate.

4.7.2 The Discriminating Statistic

Essentially, the choice of discriminating statistic is arbitrary. An ensemble of choices is available – the autocorrelation, standard deviation, non-linear prediction error and the correlation dimension, to mention a few. There are, however, choices that prove more prudent than others.

Since we are expecting to be dealing with a system governed by non-linear dynamics, a non-linear discriminating statistic would seem more appropriate (Theiler *et al.*, 1992a). Theiler & Pritchard (1996) also suggest that a distinction can be made between a *pivotal* and a *non-pivotal* statistic. The probability distribution of a pivotal statistic is independent of the processes involved in the construction of the null hypothesis. In other words, the type of

linear filter or noise source used in the construction of the null hypothesis will not influence the probability distribution of the discriminating statistic. The probability distribution of a non-pivotal statistic, on the other hand, is dependent on the processes involved in the construction of the null hypothesis. For this reason, the choice of null hypothesis has to be considered carefully when dealing with a non-pivotal statistic. It is more reliable to use a pivotal statistic since it will provide an accurate estimate for a variety of null hypotheses (Small & Judd, 1998a).

The correlation dimension is a pivotal statistic and it is a fundamental statistic in the field of Non-linear Time Series Analysis. It has been used as the discriminating statistic in many practical applications (Theiler *et al.*, 1992; Small & Judd, 1998; Small & Judd, 1998a and Barnard, 1999). Furthermore, the correlation dimension can be calculated by a reliable and well-understood algorithm – the Judd algorithm (Judd, 1992 and 1994). For these reasons, it was chosen as the discriminating statistic of choice in this study.

4.7.3 Application to NO_x Data

An AAFT surrogate data set was generated from the NO_x data using computer code written by JP Barnard (Institute for Mineral Processing and Intelligent Systems, University of Stellenbosch, personal communication). A set of 100 surrogates was constructed to ensure statistical significance. The original data set and the surrogates were embedded in dimension $m = 10$ with time delay $k = 10$ hours. Instead of comparing the actual value of the correlation dimension, the correlation dimension curves calculated by the Judd algorithm were compared (Judd, 1992 and 1994).

Results of the surrogate data test are shown in Figure 4.8. As can be seen, the correlation dimension curve for the NO_x data lies below the ensemble of correlation dimension curves for the surrogate data, albeit that it is not very far removed. The surrogate data set should have correlation dimension curves that lie above the tested data since the surrogates are constructed from a stochastic process. Stochastic systems are essentially high dimensional random noise systems.

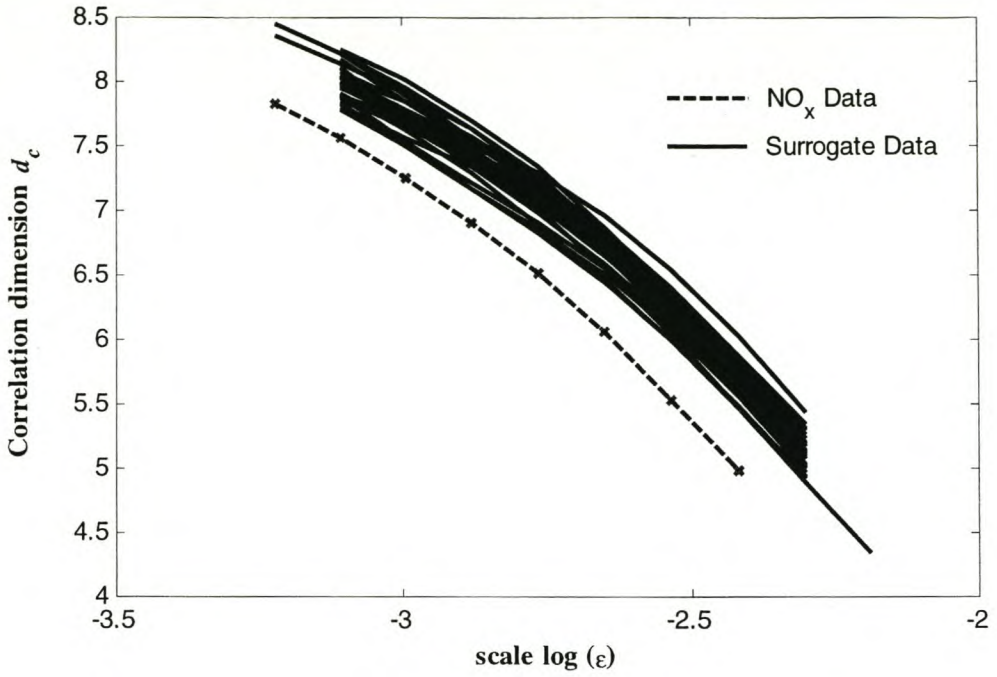


Figure 4.8 Surrogate Data Test for the NO_x Data

The fact that the original data set is not far removed from the surrogates implies that, although there is evidence of non-linear determinism in the original data set, there is quite a high random noise content. The subject of noise is discussed in the following chapter.

4.8 NOISE IN NON-LINEAR SYSTEMS

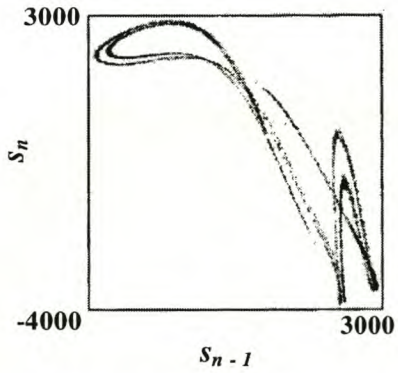
In the section addressing the method of false nearest neighbours, it was mentioned that experimental data always contains a random element known as "noise". As its name suggests, "noise" is the unwanted part of the data and contaminates the signal of interest. Noise in an experimental data set can be classified as follows:

- *Measurement noise* is due to random errors added to the system by the measurement system. It is independent of the system dynamics. Recall from equation (4.2) that a dynamic system can be represented mathematically by a state equation in a number of state variables $\boldsymbol{x}_{n+1} = \boldsymbol{F}(\boldsymbol{x}_n)$. The measured observations which describe the system are however scalars values $s_n = s(\boldsymbol{x}_n) + \eta_n$.

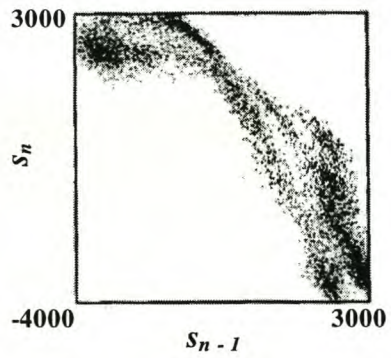
$s(\boldsymbol{x}_n)$ is a smooth function that maps the system's dynamic attractor to measurable real numbers (Kantz & Schreiber, 1997). The series of random numbers η_n is the measurement noise.

- *Dynamical noise* results from random fluctuations in the dynamical system. Mathematically, this can be represented in the state equation as $\boldsymbol{x}_{n+1} = \boldsymbol{F}(\boldsymbol{x}_n + \eta_n)$.

The effect of noise is that it densely fills the phase space that the system's attractor occupies. This filling of the phase space is well illustrated by an example that can be visualised in two dimensions. Under certain conditions, the amplitude of a Nuclear Magnetic Resonance (NMR) laser displays very irregular non-linear behaviour (Flepp *et al.*, 1991). An attractor can be reconstructed with embedding dimension $m = 3$ and time delay $k = 1$ sampling period. A 2-dimensional plot of the attractor is shown on the left in Figure 4.9. Now measurement noise, consisting of random numbers that have been filtered to have the same power spectrum as the data, is added to the data. The root mean square (rms) amplitude of the noise is taken as 10% of the amplitude of the observed data. The attractor is reconstructed with the contaminated data and is shown on the right in Figure 4.9. Note how the phase space has been filled up in the case of the contaminated data.



Dynamic attractor of the NMR laser data



Dynamic attractor of the NMR laser data plus 10% artificial noise

Figure 4.9 NMR Attractor (Kantz & Schreiber, 1997)

This filling of the phase space results in an increase in the embedding dimension m of the attractor. In addition, it is apparent that the contamination of the data will lead to difficulty in further analysis of the system. Determination of quantities such as the dimension of the

attractor is hindered and the prediction horizon of any model fitted to the data will be limited. For these reasons, it may seem that applying some noise reduction scheme would be beneficial to the analysis of the system under observation. It is important though to realise that there are distinct difficulties and disadvantages to applying an effective noise reduction algorithm (Mees & Judd, 1993 and Abarbanel, 1996). The principal drawback of applying noise reduction schemes without *a priori* knowledge of the noise dynamics or the system dynamics is that higher order dynamics could be removed along with the noise.

From the surrogate data tests, it was ascertained that the NO_x data set lies near to the noise regime. This would indicate that there are high order dynamics at play in the NO_x data set. For this reason, it was decided not to perform noise reduction on the data for fear of destroying some of the higher dimensional dynamics.

4.9 STATIONARITY

It is important to ascertain whether the amount of data collected is in fact sufficient to perform analysis accurately with a certain measure of statistical significance. In general, it is necessary to capture enough data so that all the system dynamics are enclosed in the data set. Deterministic rules governing a system should not change during the time span in which this data set was sampled. Stated in a formal manner (Kantz & Schreiber, 1997):

A signal is referred to as stationary if all transition probabilities from one state of the system to another state are independent of time within the observation period.

This requires consistency of statistical parameters throughout the time series. In addition, phenomena belonging to the dynamics of the system should be contained in the time series sufficiently frequently.

4.9.1 Method to Determine Non-stationarity

A foremost requirement for stationarity is that the time series should cover a sufficient stretch of time which is longer than the longest characteristic time scale that is relevant for the

evolution of the system (Kantz & Schreiber, 1997). For example, the photosynthesis activity of a plant is driven by solar intensity. This would mean that it roughly follows a 24-hour cycle. If this system is under investigation and data is recorded for less than 24 hours, the data set will be non-stationary, irrespective of the sampling frequency of the data.

Rigorous stationarity tests are based on the following idea:

i. Divide the time series into a number of segments S_i of length l.
ii. Estimate a certain statistic for each of the segments of the time series.
iii. Compare the statistic calculated for each segment.
iv. If the observed variations are found to be significant, outside the expected statistical fluctuations, the time series is regarded as non-stationary.

Statistics that are generally used for comparison are the mean, the variance, or the power spectrum. The test adopted here uses a non-linear cross prediction error as the discriminating statistic (Schreiber, 1997).

4.9.1.1 Detecting Non-stationarity using Non-linear Cross Predictions

The non-linear prediction error is robust and provides a stable estimate on relatively short segments S_i, S_j (a few 100 points or so). The procedure is as follows:

i Divide the time series into I ($i = 1,\ldots,I$) segments S_i of length l.
ii Use a simple non-linear prediction algorithm to predict S_j, using S_i as the training set. Compute the root mean square (rms) error for the prediction, using S_j as the test set. The prediction error as a function of i and j reveals which segments differ in their dynamics.
iii If the prediction error for a set of segments S_i, S_j is larger than the average, the data in S_i obviously provides a bad model for the prediction of data in S_j. This implies that the dynamics have changed over time, and thus the particular time series is non-stationary.

As would be expected, the diagonal entries $i = j$ will be systematically smaller since the training set and the test set are identical.

4.9.1.2 The Simple Non-linear Prediction Algorithm and the Prediction Error

Firstly, for a given scalar time series $s_1,\ldots,s_N$, it is necessary to determine a suitable time delay k and an embedding dimension m so as to form delay vectors $\boldsymbol{s}_{(m-1)k+1},\ldots,\boldsymbol{s}_N$ in $\Re^m$. To predict a time Δn ahead of N, choose a neighbourhood size ε and form a neighbourhood $U_\varepsilon(\boldsymbol{s}_N)$ of radius ε around the point $\boldsymbol{s}_N$. The maximum norm is used to determine whether a point belongs to the neighbourhood $U_\varepsilon(\boldsymbol{s}_N)$ i.e. a point belongs to $U_\varepsilon(\boldsymbol{s}_N)$ if none of its co-ordinates differs by more than ε from the corresponding co-ordinate of $\boldsymbol{s}_N$ (Kantz & Schreiber, 1997).

For all points $\boldsymbol{s}_N \in U_\varepsilon(\boldsymbol{s}_N)$ (i.e. all points closer than ε to $\boldsymbol{s}_N$), the future values $s_{n+\Delta n}$ are looked up. The final prediction, $\hat{s}_{N+\Delta n}$, is the average of all these future values:

$$\hat{s}_{N+\Delta n} = \frac{1}{|U_\varepsilon(\boldsymbol{s}_N)|} \sum_{s_n \in U_\varepsilon(s_N)} s_{n+\Delta n} \tag{4.17}$$

$|U_\varepsilon(\boldsymbol{s}_N)|$ denotes the number of elements in the neighbourhood $U_\varepsilon(\boldsymbol{s}_N)$. In the event that no neighbours closer than ε are found, the value of ε should be increased until neighbours are found. The simple non-linear prediction algorithm is used to predict future values in segment j using segment i as the training set (i.e. find the neighbours in S_i and use these values to predict the future value in S_j).

For segments i and j, the root mean squared (rms) prediction error, γ, is computed as follows:

$$\gamma(S_i, S_j) = \sqrt{\frac{1}{N_{S_j} - m} \sum_{n=m}^{N_{S_j}} (\hat{s}_{n+1}^{S_i} - s_{n+1}^{S_j})^2} \quad \text{for all } i, j \tag{4.18}$$

$\hat{s}_{n+1}^{S_i}$ is the predicted value of $s_{n+1}^{S_j}$ using S_i as the training set. The embedding dimension is given by m and N_{S_j} is the number of points in segment S_j. The root mean squared prediction error γ is calculated for all combinations of i and j and a surface plot is generated to observe the prediction error as a function of i and j.

4.9.2 Detecting Non-Stationarity in the NO_x data

The method of non-linear cross prediction was applied to the NO_x data. Two thousand four hundred points were used in the first test, then 4800, 6400 and finally 8000 points (which equate to 100, 200, 267 and 333 days respectively). The time series was divided into 8 segments. Delay vectors in $\Re^{10}$ were formed using embedding dimension m = 10 and time delay k = 10 hours. The neighbourhood size ε was set at one thousandth the size of the data interval. This size was incrementally magnified by a factor of 1.2 until a minimum of 30 neighbours was found. A plot of the prediction error as a function of i and j was generated to detect non-stationarity (Figure 4.10).

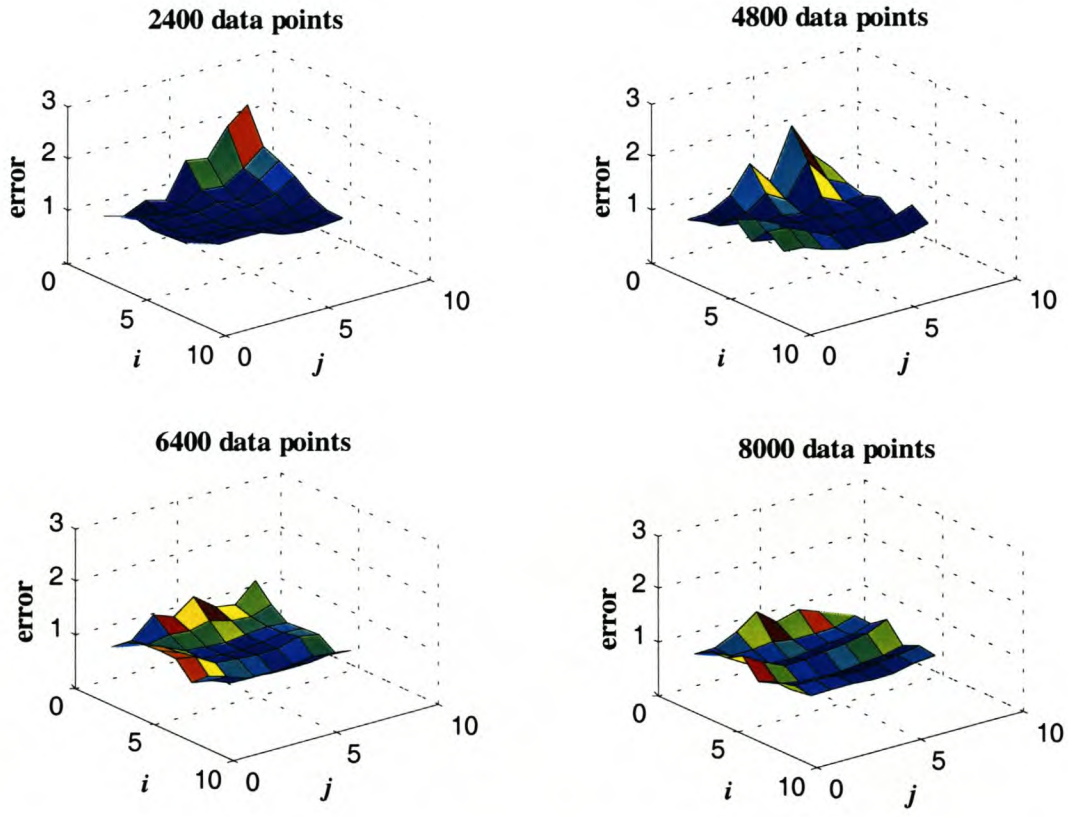


Figure 4.10 Non-linear Cross Prediction Errors for the NO_x Data set

Using only 2400 points, it can be seen that the prediction error deviates significantly from the average when trying to predicted later segments using early segments. This is also the case for 4800 data points, implying that the system is non-stationary for this length of data. The

prediction error flattens off to a more uniform, average value when the time series length is increased. At 6400 points it can be seen that the error is uniform. As would be expected, increasing the sample size to 8000 points also results in a uniform prediction error. This indicates that the system under investigation does not display non-stationary after 6400 points are used in the time series. Based on this analysis, 7000 data points (approximately 292 days of hourly data) were used for further investigation of the NO_x data so as to ensure that modelling would not be carried out on a non-stationary system.

4.10 CONSTRUCTING THE MODEL FOR NO_x PREDICTION

A prediction model can be built once the measured scalar time series $s_1,\ldots,s_N$ has been embedded in a suitable embedding dimension m with time delay k. The embedding procedure forms the delay vectors $\boldsymbol{s}_{(m-1)k+1},\ldots,\boldsymbol{s}_N$ in $\Re^m$ which are used in the reconstruction of the dynamic attractor. A prediction model maps this reconstructed attractor onto the observed time series allowing future values of the time series to be predicted from the established relationship of the time series with the attractor. This is equivalent to determining the function $\boldsymbol{F}(\,)$ in the equation:

$$s_{n+\Delta n} = \boldsymbol{F}(\boldsymbol{s}_n) \quad n \in \text{length of time series} \tag{4.19}$$

where $s_{n+\Delta n}$ is the value of the time series Δn steps ahead of s_n, Δn is the step-ahead prediction horizon, $\boldsymbol{s}_n$ is the m dimensional embedding vector and $\boldsymbol{F}(\,)$ is the measurement function to be determined. A value of Δn is selected prior to mapping the reconstructed attractor onto the observed time series.

The mapping is achieved by using methods such as a polynomial fit, radial basis functions or neural networks (Casdagli, 1989). This study made use of a neural network. The neural network model was set up by JP Barnard (Institute for Mineral Processing and Intelligent Systems, University of Stellenbosch, personal communication). It was established that the NO_x data set did not displayed non-stationarity after approximately 6400 observations and, therefore, 7000 points (292 days of data) were used to construct the prediction model. The model structure was a multilayer perceptron network with an input layer of m nodes (m = embedding dimension = 10), a hidden layer of six bipolar sigmoidal nodes, and a single

output node. Network parameters were estimated using the Levenberg-Marquardt algorithm (Levenberg, 1944 and Marquardt, 1963). Model order was optimised using the Schwarz Information Criterion (Schwarz, 1978).

Three models of different step-ahead prediction capabilities were constructed – a one-hour step-ahead, a 48-hour step-ahead and a 100-hour step-ahead model.

To quantify the model accuracy, the square of the Pearson product moment correlation co-efficient, R^2, was calculated for the predicted points 7001 to 8500 (approximately day 292 to day 354, which is roughly half-way through September to the end of the year). This constitutes an *out-of-sample* model validation since the model was constructed from data outside this validation set. The co-efficient R^2 is given by (Box *et al.*, 1978):

$$R^2 = \left(\frac{N\sum xy - \sum x \sum y}{\left[\left(N\sum x^2 - \left(\sum x\right)^2\right)\left(N\sum y^2 - \left(\sum y\right)^2\right)\right]^{\frac{1}{2}}} \right)^2 \tag{4.20}$$

where x is the observed data and y is the data produced by the model. As before, N is the number of data points.

Figure 4.11, on the following page, is a graphical representation of the procedure used for the construction of the model and the validation.

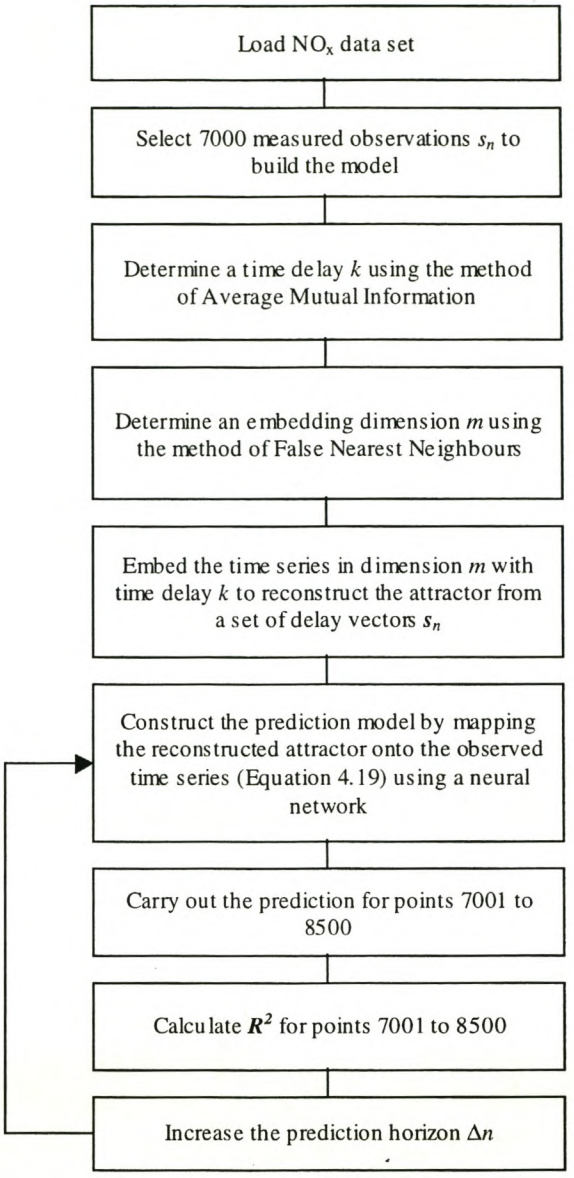


Figure 4.11 Block Diagram representing the Procedure used for Model Construction and Validation

4.11 RESULTS AND DISCUSSION OF NO_x PREDICTION

The prediction accuracy of each of the three Non-linear Time Series models is shown in Table 4.1.

Table 4.1 R^2 values for Step-ahead Prediction Δn

Step-ahead Prediction Δn	R^2
1	0.730
48	0.161
100	0.042

To obtain a visual representation of each model's prediction accuracy, an arbitrary range of points in the validation set is shown for the one-hour step-ahead prediction (Figure 4.12), the 48-hour step-ahead prediction (Figure 4.13) and the 100-hour step-ahead prediction (Figure 4.14).

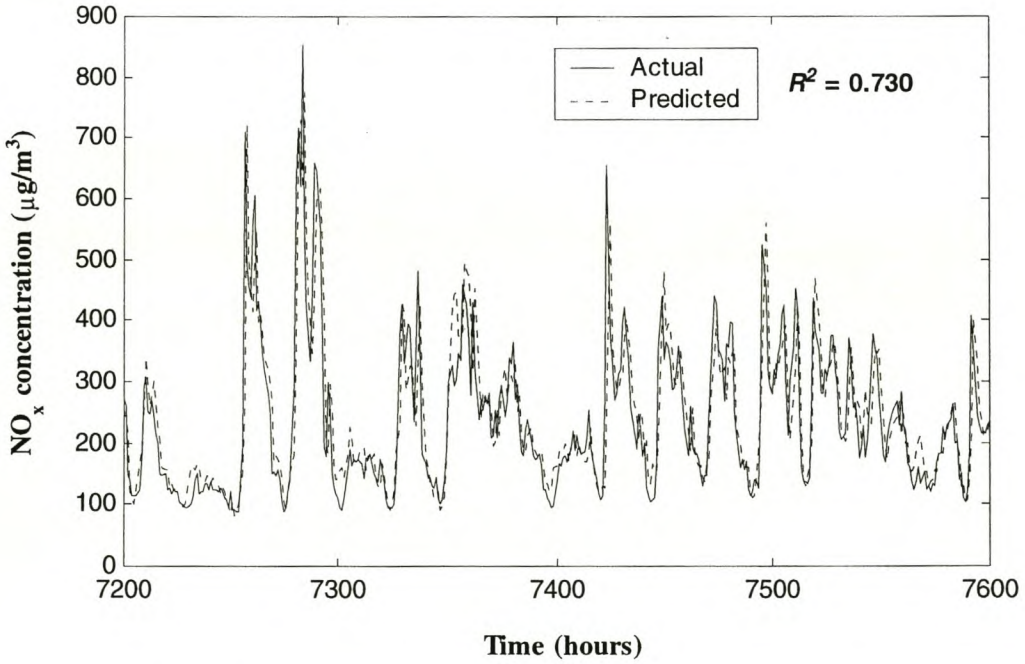


Figure 4.12 One-hour Step-ahead Prediction

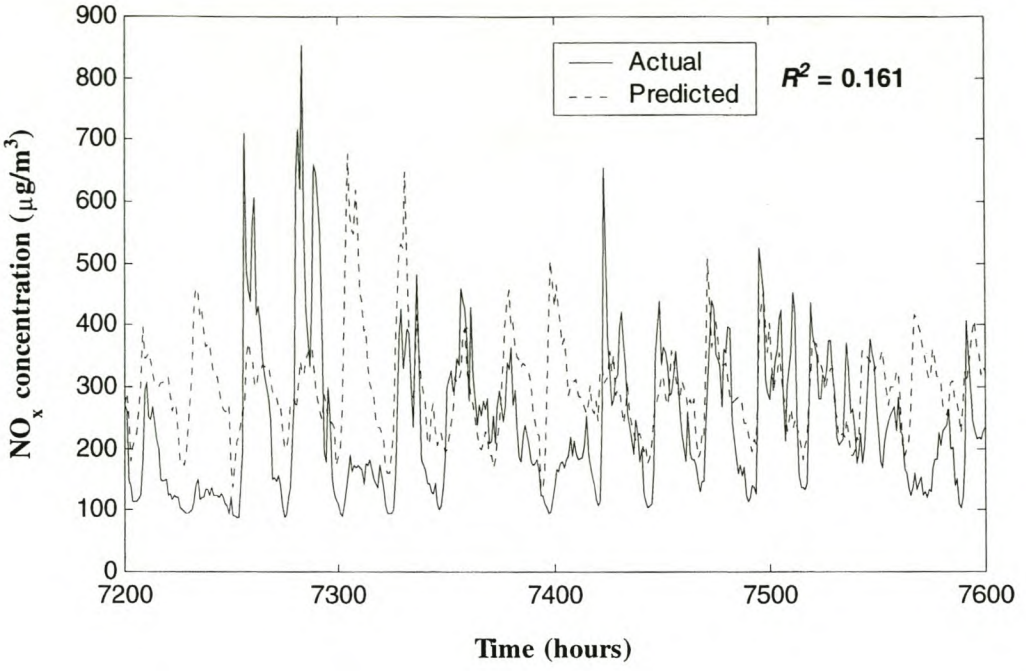


Figure 4.13 Forty-eight hour Step-ahead Prediction

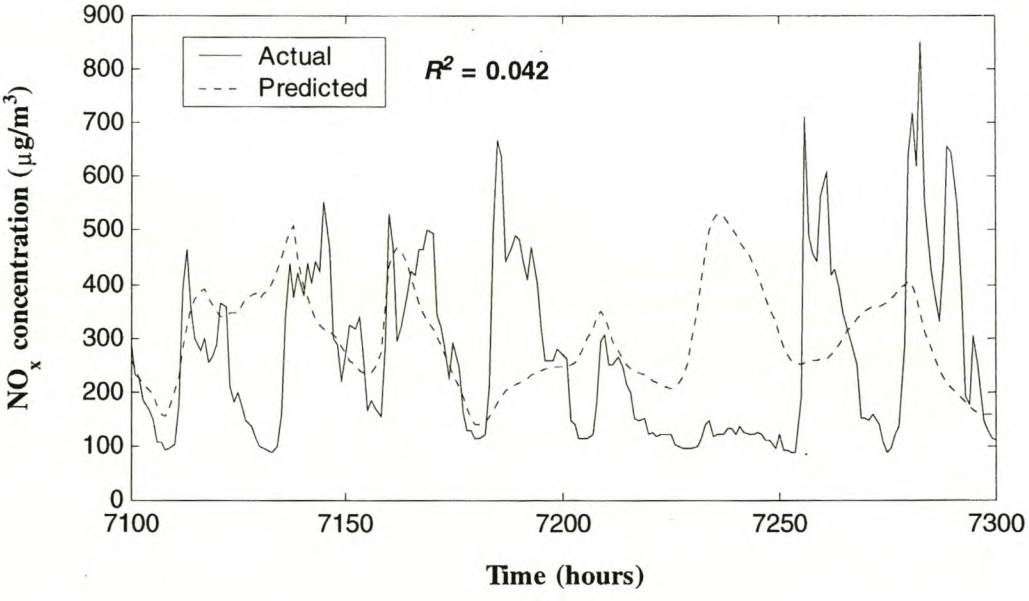


Figure 4.14 One-Hundred hour Step-ahead Prediction

The one-hour step-ahead prediction scheme, with an R^2 value of 0.730, can be considered relatively accurate. Often the one step-ahead prediction is quoted as being a reliable representation of the model accuracy (Kantz & Schreiber, 1997). It is however not indicative of the model suitability since it is not particularly groundbreaking to predict one step ahead in

time. A continuous time process (such as the NO_x data set) is autocorrelated and a one step-ahead prediction is therefore not too trying – extrapolating the trend of the last two measurements could provide a reasonably accurate one step-ahead prediction. A more robust and reliable indication of model suitability is the R^2 measure of a larger step-ahead prediction horizon. For this reason, a 48-hour and a 100-hour step-ahead prediction scheme was employed. The low R^2 values for these two predictions show that the prediction scheme failed to provide an accurate description of the underlying physical process. Figure 4.13 and Figure 4.14 bear witness to the lack of model accuracy. The prediction schemes do on occasion pick up the general trend of the pollution data, but the schemes are littered with aberrations.

The principal reason for the inadequacy of the Non-linear Time Series Analysis technique is the high noise content of the NO_x data set. The surrogate data test (section 4.7.3) indicated that, although there is evidence of non-linear determinism in the original data set, there is a high random noise content. The noise content densely fills the phase space that the system's attractor occupies (as discussed in section 4.8). This contamination of the dynamic attractor clouds the system's underlying dynamics, thereby hindering the reconstruction of a representative dynamic attractor using Non-linear Time Series Analysis techniques. Failure to reconstruct a representative dynamic attractor, and hence uncover the system's underlying dynamics, results in a limited prediction horizon. Careful application of a specialised noise reduction scheme (beyond the scope of this study) could be beneficial to the analysis of the system under observation. This would require thorough research of noise reduction schemes so as to ensure that the data is decontaminated without discarding too much of the higher order dynamics. Proper noise filtering would enhance the analysis procedure and therefore improve the prediction horizon.

.

5 SINGULAR SPECTRUM ANALYSIS (SSA)

SSA is a data analysis technique that is used to extract useful information from a time series. The information can be used to construct a prediction model for the system under investigation. The method is borrowed from digital signal processing (Kumaresan & Tufts, 1980) and is based on the univariate application of Principal Component Analysis (PCA) in the time domain (Joliffe, 1986). SSA is essentially a linear analysis technique, but the data-adaptive character of the basis functions, in terms of which the data is decomposed, allow this method to be useful in non-linear dynamics (Elsner & Tsonis, 1996 and Vautard *et al.*, 1992).

SSA has successfully been used in the analysis of short, noisy chaotic signals. Atmospheric Angular Momentum data has been analysed (Penland *et al.*, 1991) as well as paleoclimatic records (Vautard & Ghil, 1989). Keppene & Ghil (1992) successfully applied SSA in the prediction of the Southern Oscillation Index that has been connected with the seasonally recurring El Niño phenomenon. Success was also achieved by Vautard *et al.* (1992) who utilised SSA for the analysis and prediction of globally averaged surface air temperatures. These successful applications prompted the use of SSA in the analysis of the NO_x pollution data that exhibits a high noise content (discussed in section 4.7.3). The primary goal is to set up a predictive model so that the prediction capabilities of SSA and the Non-linear Time Series Analysis techniques can be compared. In this way it can be decided which route to follow when more advanced, multivariate modelling of air pollution data is required.

5.1 INTRODUCTION TO SSA

SSA is based on the idea of sliding a window of length M down a time series of length N and determining the orthogonal patterns which account for a high proportion of the variance in the time series (Allen & Smith, 1996). The method develops a set of data-adaptive filters that spectrally decompose the time series into statistically independent components with no presumption as to their functional form. Of particular interest is that the basis functions, in terms of which the data is decomposed, are determined from the time series itself i.e. they are not given *a priori* such as in Fourier Analysis which uses bases of sines and cosines. As mentioned before, these basis functions are determined from the data itself. This data-

adaptive characteristic makes SSA more flexible and better suited for the analysis of non-linear, anharmonic data (Vautard *et al.*, 1992).

In SSA, the time series is decomposed into the form:

$$x_{i+j} = \sum_{k=1}^{M} a_i^k E_j^k \qquad 1 \le i \le N,\ 1 \le j \le M \tag{5.1}$$

where x_{i+j} is the $(i+j)^{\text{th}}$ value of the time series. The basis function E^k is the k^{th} *Empirical Orthogonal Function* (EOF) and a^k is the k^{th} *Principal Component* (PC). The index i denotes a moment in time and the index j denotes a lag from time i. M is the *window length*. It is important to note that for a chosen window length M, there will be M principal components (PCs) and M empirical orthogonal functions (EOFs).

The vectors E^k are obtained from the observed data set as normalised eigenvectors of the *lagged-covariance matrix* of the data. This eigenvalue-eigenvector decomposition of the lagged-covariance matrix is achieved via spectral decomposition of the aforementioned matrix. Once these eigenvectors have been determined, they can be used to calculate the principal components, a^k, by projecting the original time series, x, onto the EOFs:

$$a_i^k = \sum_{j=1}^{M} x_{i+j} E_j^k \qquad i = 1,2,\ldots, N - M \tag{5.2}$$

and as before, E_j^k represents the j^{th} component of the k^{th} EOF.

The individual principal components have a very limited harmonic content and are thus more amenable to prediction than the time series itself. Once the individual principal components have been predicted, they can be used to recover a prediction of the original time series.

5.2 THE TRAJECTORY MATRIX AND ITS RELATION TO THE LAGGED-COVARIANCE MATRIX

Previously, it was mentioned that SSA is based on the idea of sliding a window of length M down a time series of length N and looking for patterns that account for a high proportion of the variance. In sliding this window of length M down the time series, a $M \times M$ *trajectory matrix*, **X**, is formed which contains the complete record of patterns occurring in the window size M. The trajectory matrix is used in the calculation of the lagged-covariance matrix **S**. As an illustrative example of the trajectory matrix, consider the time series of length $N = 8$:

$$x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8$$

Using a window length of $M = 4$, the trajectory matrix is constructed as follows:

$$\mathbf{X} = \frac{1}{\sqrt{N}} \begin{pmatrix} x_1 & x_2 & x_3 & x_4 \\ x_2 & x_3 & x_4 & x_5 \\ x_3 & x_4 & x_5 & x_6 \\ x_4 & x_5 & x_6 & x_7 \\ x_5 & x_6 & x_7 & x_8 \end{pmatrix} \tag{5.3}$$

where division by the square root of N is a convenient normalisation. Notice that each row of the trajectory matrix **X** is a snapshot of size $M = 4$ of the time series. There are a total of N-M+1 snapshots of the time series, in this case 8-4+1=5 snapshots. Converting a univariate time series into a multivariate set of observations is essentially what the trajectory matrix achieves (Elsner & Tsonis, 1996).

An interjection is necessary at this point. The snapshots of the time series seen in the trajectory matrix should not be completely unfamiliar if the chapter covering Non-linear Time Series Analysis has been read. In essence, it is a similar process to the *method of delay co-ordinates* which is used to reconstruct the dynamic attractor (see section 4.4). It can be seen as an embedding of the time series with an embedding dimension m equal to the window length M. The fundamental approach to determining the embedding dimension m in Non-linear Time Series Analysis is, however, not the same as the determination of the window length M in SSA. Also, the *time delay* k is not a variable subject to calculation (by the method

of average mutual information, for example) and is always equal to one in the subject field of SSA.

Returning to the trajectory matrix **X**, its value in SSA is that it forms the basis for the construction of the lagged-covariance matrix **S**. The lagged-covariance matrix is the product of the trajectory matrix and its transpose:

$$\mathbf{S} = \mathbf{X}^{\mathbf{T}}\mathbf{X} \tag{5.4}$$

S is a real, symmetric, square matrix of size $M \times M$. The elements of **S** are proportional to the linear correlation between the patterns that appear in the window of length M.

There are a number of variations in the algorithms used to calculate the lagged-covariance matrix of a time series. This study will make use of the method originally implemented by Broomhead & King (1986). The algorithm for the matrix is:

$$S_{ij} = \frac{1}{N_t - M + 1} \sum_{t=1}^{N_t - M + 1} x_{i+t-1} x_{j+t-1} \quad i = 1,2,\ldots,M \quad j = 1,2,\ldots,M \tag{5.5}$$

for a set of observations x_t, $t = 1,2,\ldots,N_t$.

Its virtues were extolled in a paper by Allen (1992), but so as not to glibly presume the accuracy of this method, a comparison of the various methods was carried out. The lagged-covariance matrix for the NO_x data set was calculated by the Broomhead and King algorithm, the Yule-Walker algorithm (Vautard *et al.*, 1992) and the Burg algorithm (Press *et al.*, 1989). The Yule-Walker algorithm is given by:

$$S_{ij} = \frac{1}{N_t} \sum_{t=1}^{N_t - j} x_{i+t-1} x_{j+t-1} \quad i = 1,2,\ldots,M \quad j = 1,2,\ldots,M \tag{5.6}$$

for a set of observations x_t, $t = 1,2,\ldots,N_t$. The difference between the Yule-Walker algorithm and the Broomhead and King algorithm is obvious from a comparison of the two equations.

By contrast to the above methods, the Burg algorithm fits an auto-regressive (AR) model with M terms to the time series. This is equivalent to finding the parameters α and γ in equation (5.16). The auto-regressive model can then be used to develop estimates of the lagged-covariance matrix. The method, along with a computer algorithm, is detailed in Press *et al.* (1989).

SSA was carried out on the data set using the three methods of calculating the lagged-covariance matrix independently. Three predictive models were constructed and then step-ahead prediction was carried out. The square of the Pearson product moment correlation coefficient, R^2, was used to quantify model accuracy (R^2 is defined in section 4.10). Results are presented in Figure 5.1.

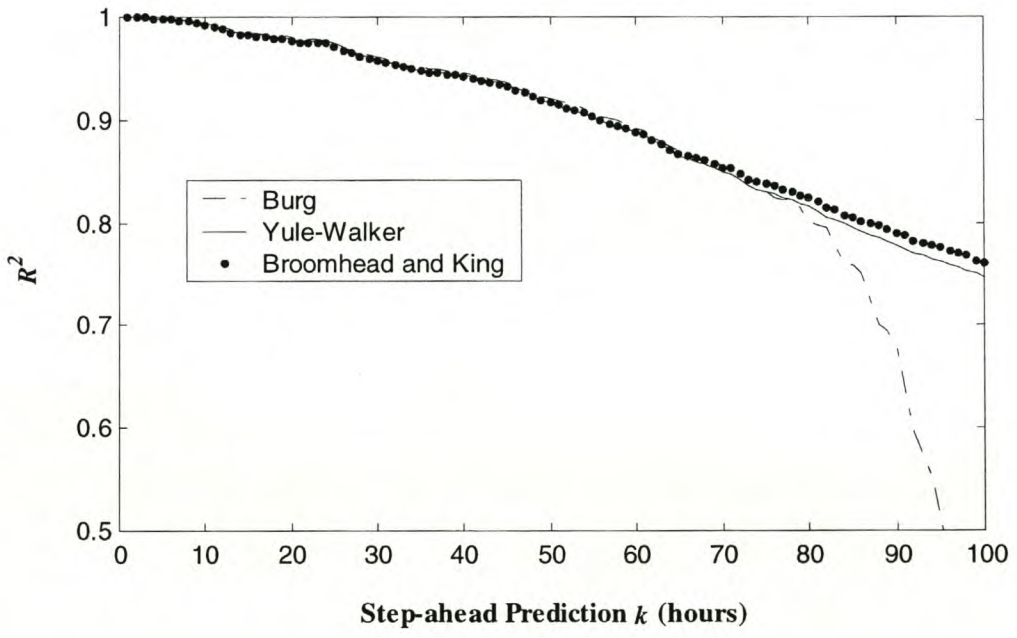


Figure 5.1 Prediction Accuracy for three models based on different Lagged-Covariance Matrices

As can be seen, the Broomhead and King algorithm achieved the best results followed closely by the Yule-Walker algorithm. The Burg estimate of the lagged-covariance matrix results in the accuracy of the prediction model falling away at large step-ahead predictions. For future reference, the above prediction schemes make use of a window length M = 170 hours. Another important point to take note of is that a simplified prediction scheme (described in Figure 5.14) is employed in the above test. The simplified prediction scheme allows for

parsimonious computation times. There is, however, a slight elevation in prediction accuracy compared to the results obtained using a fundamentally rigorous, comprehensive prediction algorithm.

5.3 SPECTRAL DECOMPOSITION AND THE EMPIRICAL ORTHOGONAL FUNCTIONS (EOFs)

The EOFs are determined by spectral decomposition of the lagged-covariance matrix. This entails decomposing the lagged-covariance matrix into its eigenvalue-eigenvector groups. The method of spectral decomposition will be described without proof since rigorous proofs are available in spectral analysis texts (Stoica & Moses, 1997 and Greenberg, 1988).

It was stated that the lagged-covariance matrix $\mathbf{S}$ is a real, symmetric matrix of square dimension $M \times M$ (where M is the chosen window length). Therefore $\mathbf{S} = \mathbf{S}^T$ and every eigenvalue of $\mathbf{S}$ is real. For each distinct eigenvalue λ_i, there is a corresponding eigenvector $\mathbf{e}_i$ which is orthogonal (in other words, $\mathbf{e}_i^T\mathbf{e}_j = 0$ for all $i \neq j$) and linearly independent (Greenberg, 1988). The eigenvector $\mathbf{e}_i$ should also be normalised so that $\mathbf{e}_i^T\mathbf{e}_j = 1$ for $i = j$. This is achieved by dividing the eigenvector by its magnitude $\|\mathbf{e}_i\|$. This will result in a set of orthonormal eigenvectors. Note that for an $M \times M$ lagged-covariance matrix $\mathbf{S}$, there will be M distinct eigenvalues and M corresponding eigenvectors.

Recall that a diagonal matrix is one in which the only non-zero elements lie on the main diagonal. A real, symmetric matrix $\mathbf{S}$ can be diagonalised by an orthogonal matrix $\mathbf{E}$. The columns of $\mathbf{E}$ will be the orthonormal eigenvectors of $\mathbf{S}$. Due to orthogonality, $\mathbf{E}^T\mathbf{E} = \mathbf{I}$ ($\mathbf{I}$ being the identity matrix) and hence $\mathbf{E}^T = \mathbf{E}^{-1}$. Now, since the matrix $\mathbf{S}$ is real and symmetric, there is a diagonalising matrix $\mathbf{E}$ such that $\mathbf{E}^{-1}\mathbf{S}\mathbf{E}$ is diagonal (Greenberg, 1988):

$$\mathbf{E}^{-1}\mathbf{S}\mathbf{E} = \Lambda \tag{5.7}$$

where Λ is a diagonal matrix, and the k^{th} diagonal element of Λ is equal to the k^{th} eigenvalue, λ_k, of $\mathbf{S}$.

From the fact that $\mathbf{E}^T = \mathbf{E}^{-1}$, the above expression can be written as:

$$\mathbf{S} = \mathbf{E}\Lambda\mathbf{E}^T \tag{5.8}$$

or

$$\mathbf{S} = \lambda_1\mathbf{e}_1\mathbf{e}_1^T + \lambda_2\mathbf{e}_2\mathbf{e}_2^T + \ldots + \lambda_M\mathbf{e}_M\mathbf{e}_M^T \tag{5.9}$$

This is called the *spectral decomposition* of a matrix **S**. It expresses **S** as a summation of the one-dimensional projections $\mathbf{e}_i\mathbf{e}_i^T$ (Elsner & Tsonis, 1996).

The diagonalising matrix **E** in equation (5.8) is composed of orthogonal eigenvectors which are the Empirical Orthogonal Functions (EOFs). The k^{th} EOF is denoted as E^k. There will be M EOFs corresponding to the chosen window length M. The diagonal matrix Λ consists of M eigenvalues, ordered with respect to magnitude i.e. the k^{th} eigenvalue is the k^{th} largest eigenvalue. The k^{th} eigenvalue is associated with the k^{th} column of **E** which is the k^{th} EOF. The k^{th} EOF has a corresponding k^{th} PC. The k^{th} eigenvalue gives the variance that the corresponding k^{th} PC accounts for. Adhering to standard practice, a *high ranked* EOF is one whose corresponding eigenvalue lies early in the rank-order i.e. its eigenvalue is larger than most. In this study, the magnitude of the eigenvalues will be presented in log base 10 format.

The square roots of the eigenvalues are called the singular values of **S**. These ordered singular values are referred to collectively as the *singular spectrum*, and hence the subject field of *Singular Spectrum Analysis* (SSA).

5.4 PRINCIPAL COMPONENTS AND RECONSTRUCTED COMPONENTS

As mentioned before, once the Empirical Orthogonal Functions (EOFs) have been determined, they can be used to calculate the corresponding principal components a^k by projecting the original time series onto the EOFs:

$$a_i^k = \sum_{j=1}^{M} x_{i+j}E_j^k \qquad i = 1,2,\ldots,N-M \tag{5.2}$$

The term a_i^k denotes the k^{th} PC at the moment i in time. There will be a total number of M PCs corresponding to the chosen window length. PCs are processes of length N-M+1 that can be regarded as a weighted moving average of the time series (Vautard *et al.*, 1992). Individual PCs are not pure sine waves, but have a very limited harmonic content. It is for this reason that linear Guassian models, such as Auto-regressive Moving Average (ARMA) models, perform better in predicting the individual PCs than the time series itself.

The k^{th} eigenvalue is associated with the k^{th} EOF, which in turn is associated with the k^{th} PC. It is very important to note that the eigenvalue associated with a particular EOF gives the variance of the corresponding PC (Keppene & Ghil, 1992).

Until now, no mention has been made of the term *"reconstructed component"* (*RC*). Vautard *et al.* (1992) introduced the RC as an alternative to the use of the PC (Appendix A contains the mathematical motivation of Vautard *et al.* (1992) for the reconstructed components). RCs are analogous to PCs and they can be used in place of PCs without loss of generality.

RCs are of length N and they carry both the contributions of the individual EOFs and the PCs implicitly. Recall that PCs are of length N-M+1. The consequence of this is that when the time series is recovered from the PCs, it will only be of length N-M+1 instead of its original length N.

The main advantage of the RC is that it is of length N and, therefore, the complete time series will be recovered with the use of RCs. Another advantage is that the RCs are additive and their complete sum recovers the original time series. The advantage of this will be clarified in the following section.

5.5 PREDICTION AND THE RECOVERY OF THE TIME SERIES

To recapitulate, the RCs (and PCs) are filtered versions of the original time series and they have limited harmonic content. Appendix B contains the first fourteen Reconstructed Components for a segment of the NO_x data set. The raw data set is displayed along with the RCs for comparative purposes. This serves to illustrate that the behaviour of the RCs (and

PCs) is more regular and, therefore, more predictable than the raw time series. Intuitively then, an improved method of predicting the time series would follow these lines:

i Decompose the original time series into EOFs and PCs.
ii Fit an ARMA model to each individual PC and do the required step-ahead predictions for each PC individually.
iii Recover the time series from the predicted PCs. This will be the prediction of the original time series.

The time series can be recovered by reconstructing a signal from a convolution of the PCs and their corresponding EOFs. Mathematically, the convolution of two functions $h(t)$ and $g(t)$ is defined as (Elsner & Tsonis, 1996):

$$\int_{-\infty}^{+\infty} g(\tau)h(t-\tau)d\tau \tag{5.10}$$

The numerical computation of the convolution of the PCs and their corresponding EOFs is discretely formulated as follows:

$$x_{i+j} = \sum_{k=1}^{M} a_i^k E_j^k \quad 1 \le i \le N,\ 1 \le j \le M \tag{5.11}$$

This is equivalent to equation (5.1) that governs the decomposition of the time signal. There is however a problem that arises when recovering the time series from the sum of the PCs and the EOFs – there is not a unique expansion of the signal. Notice from the above equation that at moment i, the recovered term of the original time series x_{i+j} is dependent on the index j. This means that there are M different ways of reconstructing the components of the signal and they do not, in general, give the same results (Vautard *et al.*, 1992).

RCs, on the other hand, are additive and provide a unique reconstruction of the original time series. The reconstruction of the time series is not dependent on the index j (see Appendix A):

$$x_i = \sum_{k=1}^{M} r_i^k \quad i = 1,2,\ldots,N \tag{5.12}$$

The term r_i^k in equation (5.12) is the value of the k^{th} RC at time i. As mentioned before, RCs allow the complete time series of length N to be recovered uniquely and not only a time series of length N-M+1 as is the case when using PCs. For these reasons, RCs will be used in this study.

5.6 CHOICE OF WINDOW LENGTH M

The choice of a suitable window length M is an important step in SSA. Essentially, the choice of M is a trade-off between the amount of information that is to be retained and the degree of statistical significance that is required. Numerous techniques for the judicious selection of an appropriate window length M are presented in SSA literature. A selection on these techniques will be applied to the NO_x data set to assist in selecting a value of M that will benefit further analysis of the data.

SSA does not resolve periods longer than the window length. For this reason, a larger window length has to be used to resolve longer period oscillations present in a signal. These longer period oscillations represent additional information over and above the high frequency components of the signal. The trade-off for the added information is a loss of statistical significance (the concept of statistical significance is given further attention in section 5.9). Conversely, a smaller window length allows a greater degree of statistical significance since the high frequency components do not "compete" with the low frequency components for the finite available variance (Elsner & Tsonis, 1996). There is, however, a limit to the amount of information that can be retained (e.g., the longer period oscillations will not be resolved).

Problems do arise in both cases if M is chosen to be an extreme value. Note that the spectral resolution which is achieved in SSA is $1/M$ (Vautard *et al.*, 1992). If M is too large (high spectral resolution), spurious spectral peaks are obtained which can be confused with physically valid peaks. To prevent this, Vautard *et al.* (1992) suggest that the window length should not exceed $M = N / 3$. If M is too small, neighbouring spectral peaks lying close together in frequency space will not be resolved because of the coarse spectral resolution.

It has been suggested by Penland *et al.* (1991), that the results from SSA are not significantly influenced by the size of the window length M as long as M is substantially smaller than the

number of observations in the raw data set N. "Substantially smaller" is quantified as $N / 4$. Yiou *et al.* (1994) observed that varying the window length about a sufficiently large M only served to stretch or compress the spectrum of eigenvalues. The relative magnitudes of the individual eigenvalues remained unchanged. To illustrate this point, eigenvalues for the NO_x data were computed for a number of window lengths. M was varied from 24 hours to 336 hours, in increments of 24 hours. The eigenvalues are plotted in Figure 5.2.

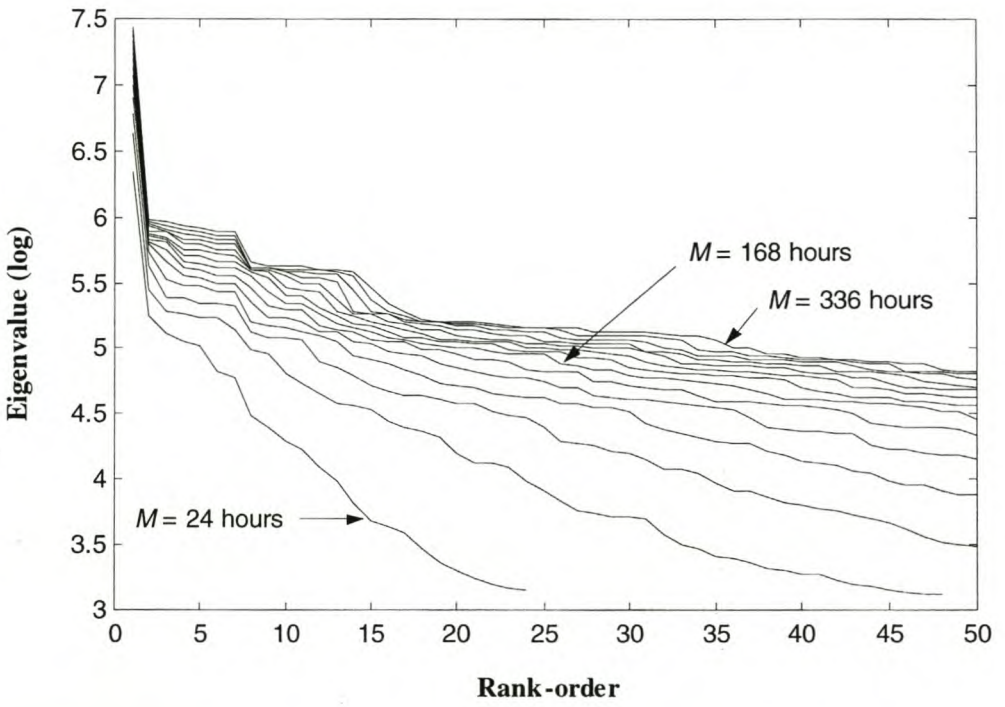


Figure 5.2 Eigenvalues for NO_x Data with varying Window Length M

The graph shows the eigenspectrums at various window lengths with the x-axis limited to the 50^{th} eigenvalue of each spectrum. It is evident that the eigenvalue spectrums approach some form of plateau as M is increased. At this plateau, the relative magnitude of the eigenvalues does not change significantly and the eigenvalues are merely stretched out across the spectrum. The line marked with an arrow, corresponding to a window length of M = 168 hours, is arguably the start of this plateau.

Vautard & Ghil (1989) suggest that no optimal M exists, but the way to decide on M is to evaluate stable features of the eigenvalue-eigenvector set over a reasonable range of M. This is essentially what has been done in Figure 5.2 above. The eigenspectrum was investigated over a range of window lengths to determine when the eigenvalues displayed stability with

respect to their magnitude. This stability was indicated by the formation of a plateau at around $M = 168$ hours.

Penland *et al.* (1991) examine the correlation function to decide on a value for M. They select a value which is sufficiently large to contain characteristic high frequency and low frequency oscillations. A plot of the correlation function for the NO_x data is displayed in Figure 5.3.

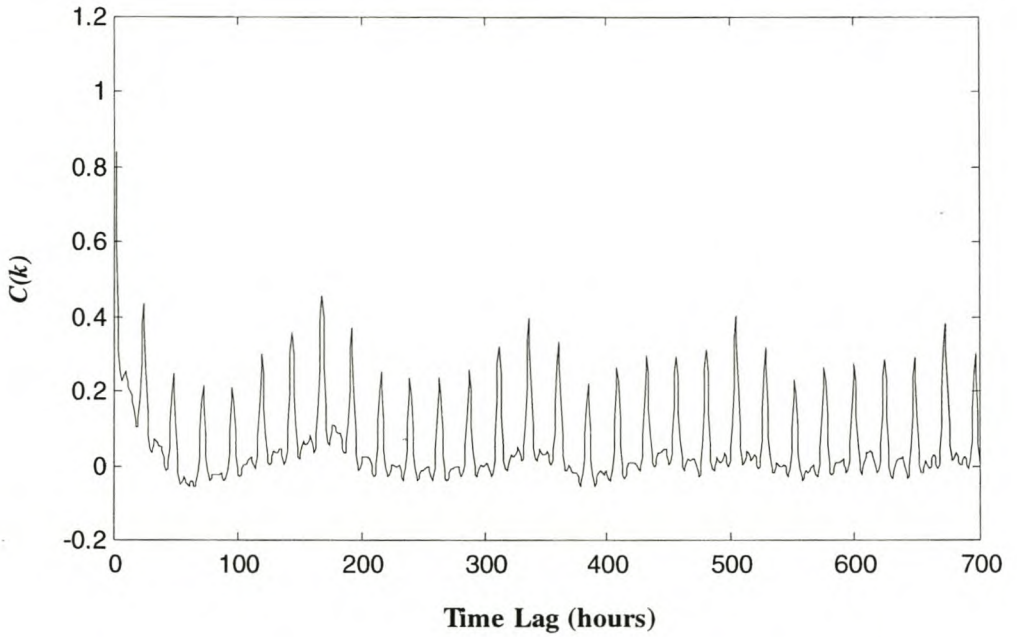


Figure 5.3 Autocorrelation for NO_x Data

The autocorrelation function consists of high frequency oscillations superimposed on a low frequency structure. The high frequency oscillations are regular and the low frequency structure appears to repeat itself at time lags of roughly 170 hours.

Furthermore, a look at the power spectrum of the NO_x data plotted against the period, reveals that there are distinct cycles at 8, 12, 24 and 168 hours (Figure 5.4). These distinct cycles corroborate the correlation of the NO_x cycle to traffic flow (Figure 2.3). It should also not come as too much of a surprise to notice that 168 hours is the period of a week.

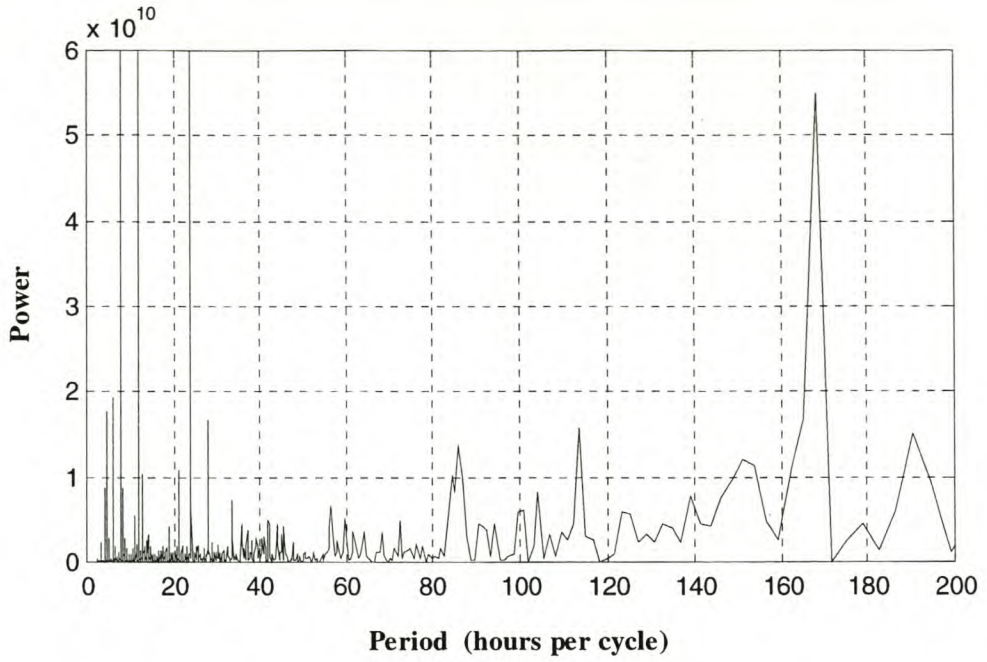


Figure 5.4 Power Spectrum versus Period for the NO_x Data

Since these periods display such high power, it will be judicious to capture these periods of importance. Ensuring that the window length exceeds these values will allow these periods to be resolved.

Practically, computational limitations have to be taken into account. Processing time is proportional to M and since model accuracy (in terms of the square of the Pearson product moment correlation co-efficient, R^2) has to be calculated for a range of prediction horizons k at various values of M, processing power is a limitation. For M = 336 hours, the computation time of the R^2 values for a prediction horizon ranging from 1 hour to 100 hours, took in the region of 38 hours. This is limiting if various different calculations and tests are to be carried out at that window length. For M = 170 hours, the processing time was 16 hours which, although not ideal, is manageable. It is however important to put the computational demands of the SSA model into perspective. From the above statistics, it seems as though the model requires an excessive amount of computational power. This was mainly because of the interpreted MATLAB® programming language used to construct the model and the limitations of the desktop computer. The model does in fact require less computational power than the Non-linear Time Series Analysis techniques.

A graph of R^2 for different window lengths at a range of prediction horizons is shown in Figure 5.5. Note that the results were produced by using the comprehensive prediction algorithm (as opposed to the simplified prediction scheme) and therefore the R^2 data is a true representation of model accuracy.

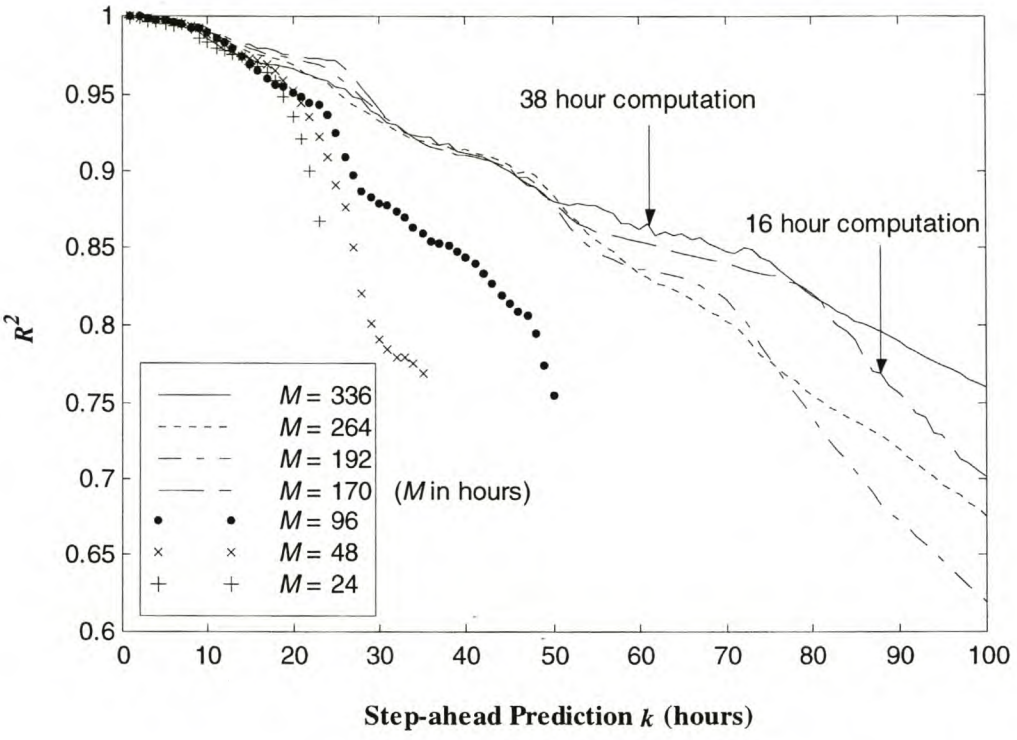


Figure 5.5 Comparison of R^2 for varying Window Length M

The prediction model for M = 170 hours outperforms all the models except for the model of window length M = 336 hours. However, the additional computational requirement to run the model for M = 336 hours does not justify the increase in accuracy. The models for the other window lengths perform well initially, but fall away once the prediction horizon exceeds the window length. As mentioned before, SSA does not resolve periods longer than the window length. For this reason, there is a lack of information regarding the longer period oscillations outside the window length and hence a drop in prediction accuracy.

So why not make the window length extremely long and obtain a large prediction horizon? Well, the answer has already been given. Increasing M, and hence increasing the number of EOFs, results in a loss of statistical significance. If M is too large, spurious spectral peaks are obtained which can be confused with physically valid peaks. In addition, excessive

computational requirements could be a limiting factor if a standard desktop computer is employed to do the computation.

If the above criteria are considered globally, a choice of M = 170 hours would seem to be the most feasible window length (subject to significance testing, which will be discussed in section 5.9). The final decision was to use a window length of M = 170 hours to construct the prediction model. Some of the test calculations which were carried out on the NO_x data set were done with shorter window lengths so as to minimise computation time.

5.7 DETRENDING THE DATA

Elsner & Tsonis (1996) pointed out that it might be important to employ trend removal from a raw data set before proceeding with SSA. For a time series of air temperatures, they showed that the eigenvalues of the lagged-covariance matrix could vary significantly in magnitude depending on whether the raw data had been detrended or not. This variation in eigenvalue magnitude could affect the resulting analysis.

The detrending of the data was done by subtracting the least-squares linear regression line from each value in the data set. The linear regression line is given by $y = ax + b$ where a and b are constants determined from the data. Figure 5.6 (taken from Elsner & Tsonis) is a comparison of the ten largest eigenvalues of the detrended data set and the raw data set. It can be seen that there is quite a substantial difference in the log magnitude of the three leading eigenvalues: approximately 900% difference for eigenvalue 1, 100% difference for eigenvalue 2 and 75% difference for eigenvalue 3.

The NO_x data was detrended by subtracting the least-squares linear regression line $y = -0.0005x + 277$ from each value in the data set (y is the NO_x concentration in $\mu g/m^3$ and x is the time in hours). The leading eigenvalues were calculated for the detrended data and these eigenvalues were plotted against the leading eigenvalues of the raw data set. This is shown in Figure 5.7. Note the small variation in eigenvalue magnitude.

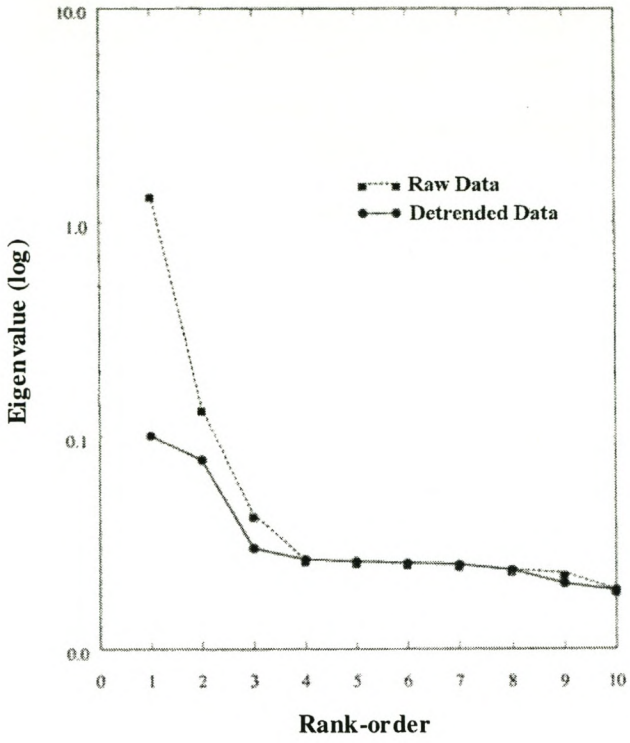


Figure 5.6 Comparison of the Leading Eigenvalues of an Air Temperature Data Set (Elsner & Tsonis, 1996)

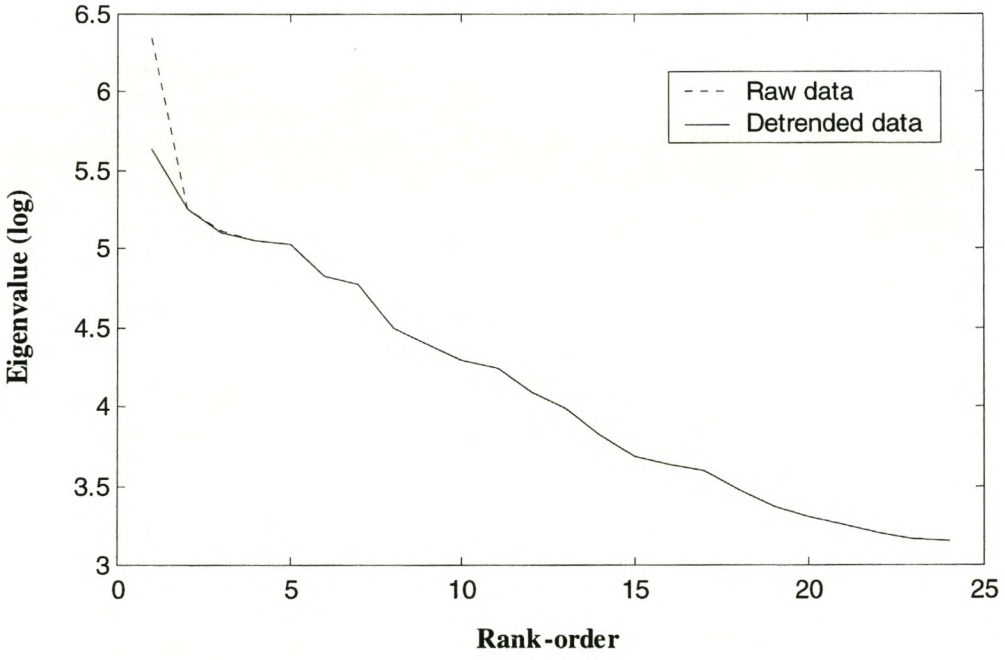


Figure 5.7 Comparison of the Eigenvalues for the NO_x Data

From Figure 5.7, it is obvious that the only variation in magnitude occurs at the leading eigenvalue – a minimal variation in the region of 10%. To ensure that this slight variation did not affect results, a model was built based on the detrended data. The accuracy of the model was compared to the accuracy of a model constructed from the raw data. The square of the Pearson product moment correlation co-efficient R^2 was used to quantify model accuracy. The model used a window length of M = 96 hours (as opposed to M = 170 hours) to reduce computation time and R^2 was plotted for a range of prediction horizons – for k = 1 hour to k = 20 hours. The results are presented in Figure 5.8.

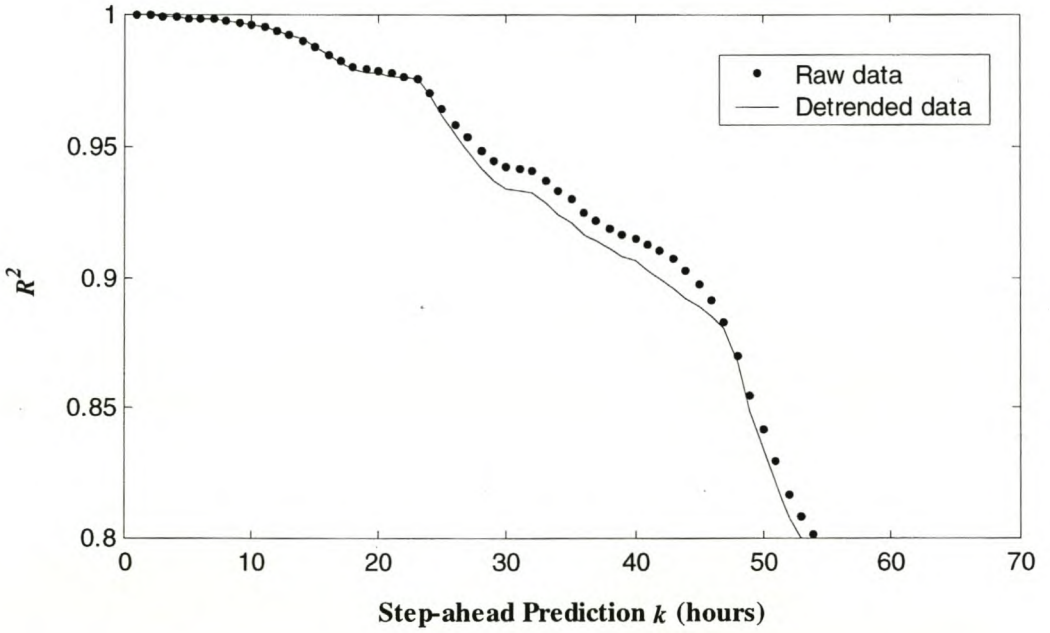


Figure 5.8 Comparison of Prediction Ability for Raw and Detrended Data (Window Length M = 96 hours)

There is very little difference between the two models. There is a slight variation in prediction accuracy, but not sufficient to draw any firm conclusions. Thus, for model construction, no detrending of the NO_x data was carried out.

5.8 NOISE

The subject of noise was discussed in the chapter on Non-linear Time Series Analysis (section 4.8). It will be re-addressed in the context of SSA.

Two noise processes have to be considered, namely, *white noise* and *autocorrelated* (*red*) *noise*.

- *White noise* has a Gaussian distribution and displays a flat power spectrum over the frequency range f (Rosie, 1966). In other words, it has a power spectrum that is proportional to $1/f^0$. The noise component is stationary and linearly independent of the signal-bearing component and thus the lagged-covariance matrix of the underlying process can be written as:

$$\mathbf{S} = \mathbf{S}^{\text{signal}} + \mathbf{S}^{\text{noise}} \tag{5.13}$$

where $\mathbf{S}^{\text{signal}}$ and $\mathbf{S}^{\text{noise}}$ are the lagged-covariance matrices of the signal and noise components respectively (Elsner & Tsonis, 1996). Since white noise is independent and identically distributed (iid), the lagged-covariance matrix is $\mathbf{S}^{\text{noise}} = \sigma^2\mathbf{I}$ where σ^2 is the noise variance and $\mathbf{I}$ is the rank-M identity matrix. So,

$$\mathbf{S} = \mathbf{S}^{\text{signal}} + \sigma^2\mathbf{I} \tag{5.14}$$

Adding $\sigma^2\mathbf{I}$ to $\mathbf{S}^{\text{signal}}$ only serves to increase all the eigenvalues of $\mathbf{S}$ by σ^2 without altering the eigenvectors (Allen & Smith, 1996). Therefore, if the time series consists only of signal and white noise, the EOFs of $\mathbf{S}$ still have a clear physical meaning. If $\mathbf{S}$ has M eigenvalues and $\mathbf{S}^{\text{signal}}$ has n non-zero eigenvalues (where $n<M$), the high-ranked EOFs of $\mathbf{S}$ provide a consistent estimate of the EOFs of $\mathbf{S}^{\text{signal}}$ (Allen & Smith, 1997). These high-ranked EOFs of $\mathbf{S}$ will appear as n eigenvalues of $\mathbf{S}$ lying above a flat *noise floor*. This is best illustrated with the aid of an example presented by Penland *et al.* (1991). White noise of unit variance is added to a quasiperiodic time series. The eigenvalues for the quasiperiodic time series with added white noise are shown for various window lengths in Figure 5.9.

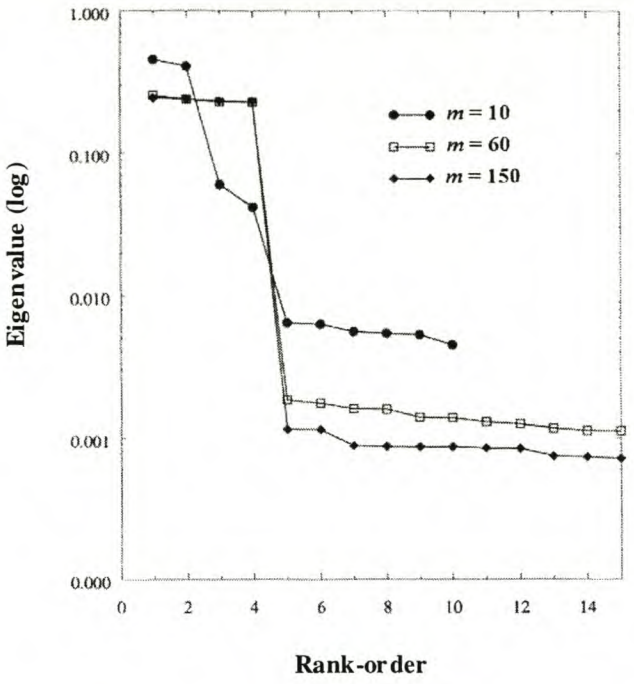


Figure 5.9 Singular Spectra of a Quasiperiodic Time Series with additive white noise for various Window Lengths (Elsner & Tsonis, 1996)

Note that for M = 60 hours and M = 150 hours, there is a clear break after the fourth eigenvalue to a set of eigenvalues which spread out into a nearly flat noise floor. A substantial percentage of the system's total variance can be described by the first four eigenvalues. The eigenvalues making up the noise floor correspond to the part of the variance that is unlikely to be explained by a deterministic model. The above reasoning is the basis for the standard practice of filtering a time series by truncating the eigenspectrum. This entails retaining only a certain number of the highest ranked eigenvalues and EOFs that are considered *significant*, and then reconstructing the time series using only the significant part of the truncated eigenspectrum. Although this method is effective in separating the signal from the noise, it has to be stressed that it is only of use if the signal is contaminated with white noise.

- *Red noise* is noise contamination that is autocorrelated i.e. noise in a particular observation is related to the noise in temporally near observations. Its power spectrum decreases with frequency according to $1/(f^2 + a^2)$ where a is some constant (Addison, 1997). This form of noise is particularly prevalent in natural phenomena such as weather patterns – the current weather conditions are strongly influenced by the most recent conditions (Elsner & Tsonis, 1996). Since weather has a strong influence on air pollution

concentrations, it would be expected that pollution concentrations would also have a degree of red noise content. As in the white noise process, the signal and the noise are linearly independent so that

$$\mathbf{S} = \mathbf{S}^{\text{signal}} + \mathbf{S}^{\text{noise}} \tag{5.15}$$

The difference is that the high-ranked EOFs of **S** are no longer a suitable approximation to the eigenvectors of $\mathbf{S}^{\text{signal}}$. The eigenvectors of **S** depend on $\mathbf{S}^{\text{signal}}$, $\mathbf{S}^{\text{noise}}$ and the signal to noise ratio (Elsner & Tsonis, 1996). Thus, using the truncation of the eigenspectrum to separate the signal from the noise becomes unreliable when red noise is present.

Red noise can be approximated by a first order auto-regressive model – AR(1) noise. The AR(1) model is given by:

$$x_t - \bar{x} = \gamma(x_{t-1} - \bar{x}) + \alpha\varepsilon_t \tag{5.16}$$

where $\bar{x}$ is the process mean, α and γ are process specific parameters and ε_t is Guassian, unit-variance white noise. The eigenvalue spectrum of a process contaminated by red noise has a sloping noise floor which, if it were at all possible, makes the identification of a break to the noise floor difficult (Elsner & Tsonis, 1996). However, as mentioned before, placing *significance* purely on rank-order position of an eigenvalue is only effective when dealing with white noise. The assumption that significance decreases with position in the eigenvalue rank-order is false when dealing with systems contaminated with red noise, or for non-linear systems in general (Allen & Smith, 1996).

For a process that is contaminated with red noise, the Monte Carlo approach can be employed to determine the significant eigenvalues. This procedure is explained in the following section dedicated to significance testing.

In the subject field of SSA, caution has to be exercised when dealing with the noise components of a signal. Referring back to section 4.7.3, the use of surrogate data identified that the NO_x data displayed non-linear characteristics. Palus & Dvorak (1992) noted that truncation of the eigenvalues reduces non-linear system dynamics rather than noise. They

observed that, for a non-linear system, dynamics could be suppressed onto the noise floor due to lower variance. The non-linear dynamics could be indistinguishable from red noise by standard tests that are based on linear theory.

5.9 SIGNIFICANCE TESTING USING A MONTE CARLO APPROACH

The aim of a significance test is to reduce the signal to its essentials by establishing which of the eigenvalues are significant. This involves applying a statistical test to determine which of the system's eigenvalues are significant in that they differ from a noise process. To this end, the Monte Carlo approach of generating a set of surrogate realisations based on a null hypothesis and testing these surrogates against the original data set, is one of the most efficient methods. This is essentially the same procedure described in section 4.7 to establish non-linear determinism with the use of surrogate data sets.

In the subject field of SSA, the above method of significance testing is known as Monte Carlo SSA. The general idea was proposed by Broomhead & King (1986), and the implementation was carried out by Allen (1992). The procedure is as follows:

i. Generate an ensemble of surrogate data sets that are consistent with the null hypothesis that the data is AR(1) noise. These surrogates are generated from equation (5.16). The parameters α and γ should be selected so that the surrogates are comparable to the measured data in certain respects and so that they maximise the likelihood that the null hypothesis is not rejected. It should however not allow the null hypothesis to be so stringent that no significant eigenvalues are found. The selection of these parameters is detailed in Allen & Smith (1996) and in Elsner & Tsonis (1996). The accuracy to which significance can be assessed is dependent on the number of surrogates that are generated. To ensure accuracy in the order of 1%, it is necessary to generate a surrogate data set that comprises of 1000 realisations (Elsner & Tsonis, 1996).

ii. For a predetermined window length M, compute the lagged-covariance matrices $\mathbf{S}^{surr}$ for each surrogate data set. Project each $\mathbf{S}^{surr}$ of the surrogate realisations onto the EOFs of the data as in equation (6.6) to obtain a diagonal matrix of surrogate eigenvalues:

$$\Lambda^{surr} = \mathbf{E}^{T}\mathbf{S}^{surr}\mathbf{E} \tag{5.17}$$

The diagonal matrix Λ^{surr} contains the surrogate eigenvalues λ_k^{surr}. These eigenvalues will serve as the discriminating statistic.

iii. Compute the 2.5^{th} and the 97.5^{th} percentile significance level of the surrogate eigenvalues at each rank-order. Plot these percentiles against the eigenvalues for the original data.

iv. Eigenvalues, which lie above the 97.5^{th} percentiles of the corresponding surrogate distributions, are indicative of eigenvalues that contain more variance than expected on the null hypothesis. This amounts to rejection of the null hypothesis and, hence, detection of a significant eigenvalue.

Application to NO_x Data

Monte Carlo SSA was carried out on the NO_x data for window lengths of M = 24 hours and M = 170 hours. A MATLAB® routine, written by Eric Breitenberger (Geophysical Institute, University of Alaska Fairbanks, personal communication), was modified to generate the surrogate data sets and calculate the percentiles. Figure 5.10 shows the results of the Monte Carlo SSA technique when applied to the NO_x data set with a window length M = 24 hours.

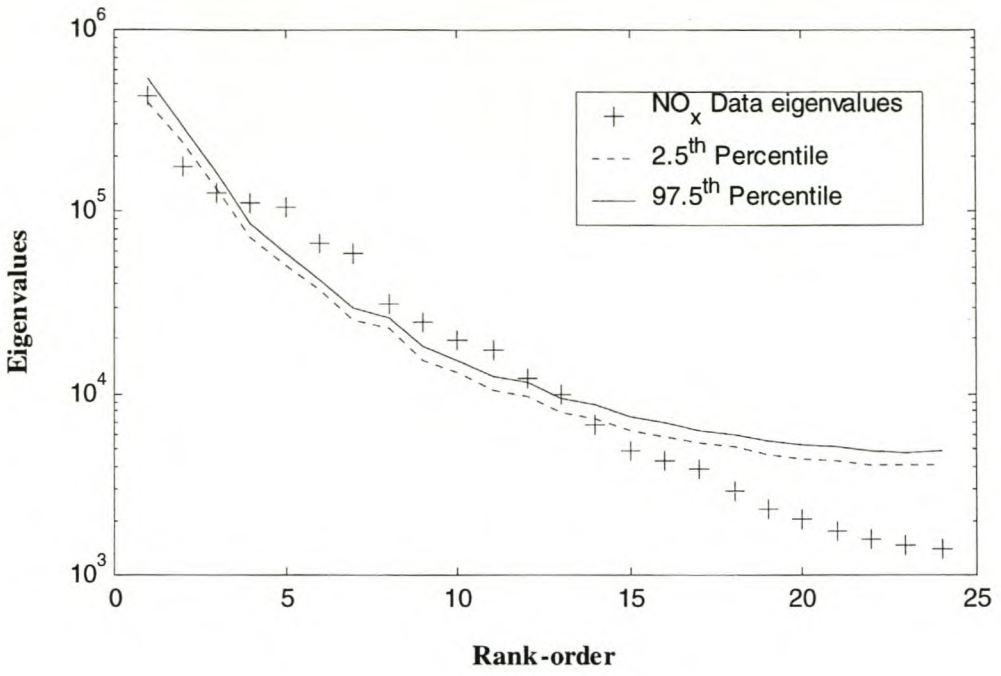


Figure 5.10 Eigenspectrum of NO_x Data and Surrogate Data Sets (Window Length M = 24 hours)

Although the leading eigenvalue is appreciably higher than the rest, the surrogate data test shows that it does not contain more variance than would be expected from a series of AR(1) noise. Hence, it does not exhibit any significance if it is assumed that the background noise is red. This leads to the important maxim that dominance is not sufficient for significance (Elsner & Tsonis, 1996). Eigenvalues 4 to 13 carry more variance than expected when judged against the null hypothesis and this would classify them as significant. Observe the sloping noise floor which is characteristic of a red noise process. Also, there are no pronounced "breaks" to a distinct noise floor which could also indicate that the mixing between signal and noise is smooth. This mixing increases the complexity of signal to noise separation (Vautard & Ghil, 1989). Only at eigenvalue 5 and 7 does there seem to be a slight break from the spectrum to a lower plateau. This, however, would only be of importance if white noise were assumed the only contamination.

The Monte Carlo SSA technique was also applied to the NO_x data set with a window length of M = 170 hours. The results are shown in Figure 5.11. Only the first 90 rank eigenvalues are shown to enable detail at the higher rank to be seen. Eigenvalues 91 to 170 fall below the 2.5^{th} percentile.

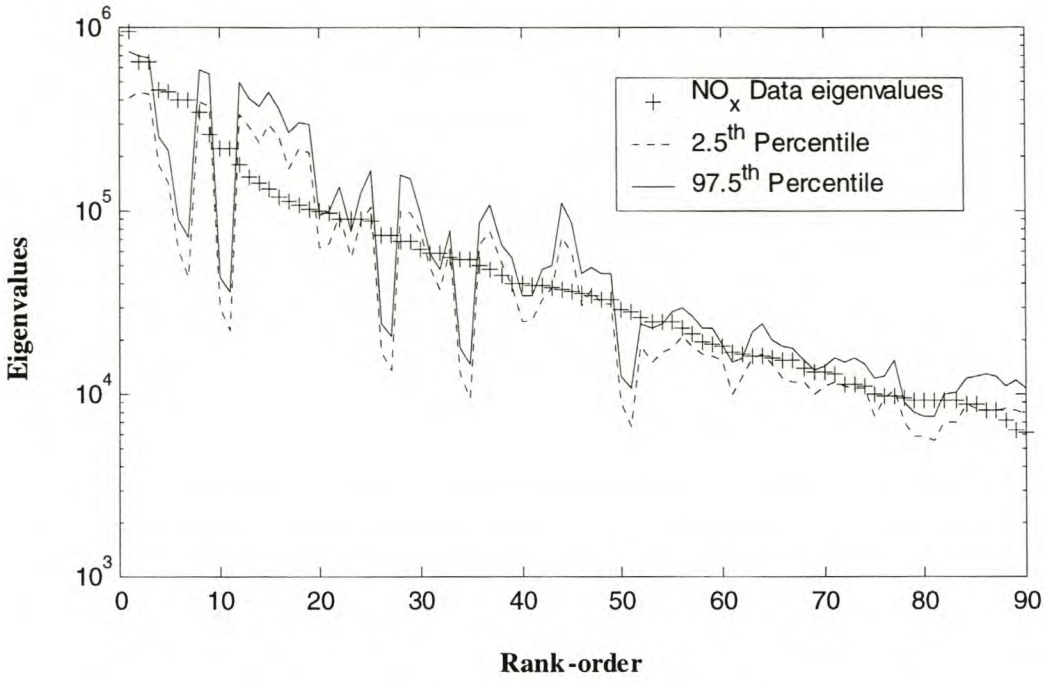


Figure 5.11 Eigenspectrum of NO_x Data and Surrogate Data Sets (Window Length M = 170 hours)

This eigenspectrum also displays the characteristic sloping noise floor of a red noise process. An interesting feature is that, in contrast to the case where M = 24 hours, the first eigenvalue is now identified as being significant. Recall that increasing the window length M increases the spectral resolution of SSA, but it also has the added advantage of increasing the potential signal to noise enhancement (Vautard *et al.*, 1992). The re-classification of the first eigenvalue as being significant is possibly due to the fact that the larger window length affords the analysis greater signal to noise enhancement. It is, however, judicious to take into consideration the fact that a larger M increases the number of individual excursions which are expected to occur above the given confidence level purely by chance.

To explain this, Allen & Smith (1996) have shown that even when testing a segment of pure noise against the null hypothesis, the average number of excursions above the 97.5^{th} percentile will be $0.025M$. This indicates that the average number of excursions that occur purely by chance scales linearly with M, simply because more tests are being done against the null hypothesis. Since an increase in the window length results in an increase in the number of EOFs and the corresponding eigenvalues, the statistical significance of individual

excursions above the 97.5th percentile is reduced. This explains the fact that an increase in M results in a loss of statistical significance.

A two-pass Monte Carlo procedure can be used to quantify the significance level explicitly (Livezy & Chen, 1982). This involves making a second pass through the ensemble of surrogates to estimate the probability of an excursion occurring purely by chance. This procedure would have been pursued if the contamination were thought to consist of only white and red noise. However, in section 4.7.3, it was shown that the NO_x data displayed signs of non-linear determinism. Monte Carlo SSA testing against a AR(1) noise hypothesis is a linear test procedure and will therefore identify significance against red noise, but non-linear dynamics may be indistinguishable from red noise (Palus & Dvorak, 1992). As mentioned before, the non-linear dynamics could be suppressed onto the noise floor and, thus, discarding eigenvalues of lower rank may result in the loss of dynamic information. For this reason, the full eigenspectrum was used to reconstruct the time series.

More importantly, the aim of this study is to evaluate two approaches to air pollution modelling and to establish which of these analysis techniques would be the most efficient in modelling air pollution. Therefore, it is essential that the two approaches that were used be investigated over a common denominator. No filtering was done in the Non-linear Time Series Analysis and, hence, no filtering should be carried out when analysing the signal with SSA. Reconstructing the signal using only selected eigenvalues amounts to filtering and, therefore, all the eigenvalues were deemed significant in reconstructing the time series.

5.10 CONSTRUCTING THE MODEL FOR NO_x PREDICTION

The same parameters that were used in the construction of the Non-linear Time Series Analysis model were used for the construction of the SSA model. This consistency allows for equivalent comparison of the two techniques. To recapitulate, non-stationarity was attained after approximately 6400 measured observations. Seven thousand points were used to construct the prediction model. To quantify the model accuracy, the square of the Pearson product moment correlation co-efficient, R^2, was calculated for points 7001 to 8500 (approximately day 292 to day 354, which is roughly halfway through September to the end of the year). This constitutes an *out-of-sample* model validation.

It was mentioned in previous sections that a simplified prediction scheme was used to perform some of the tests. The sole reason for this was to save computation time. The simplified algorithm is based on simplifying assumptions that are not fundamentally rigorous. The simplified algorithm performs SSA on the entire data set once, and then uses the results for model prediction. This saves having to re-calculate the computationally costly SSA routine at each update step, but it does mean that the simplified model is "cheating" since it has some prior knowledge of future events. The simplified algorithm does however produce results which are representative of the results obtained using the comprehensive (and fundamentally accurate) prediction algorithm at $1/5^{th}$ the computation time. Figure 5.12 is a comparative representation of the performance of the two algorithms.

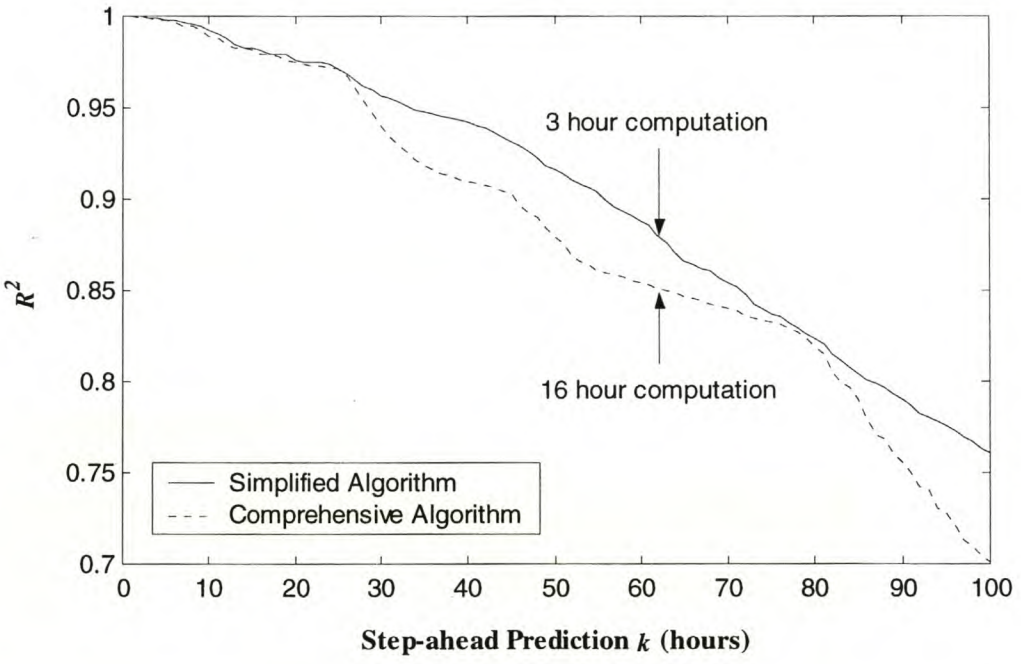


Figure 5.12 Comparison of R^2 values for prediction using the Comprehensive Algorithm and the Simplified Algorithm (M = 170 hours)

The prediction accuracy of the simplified algorithm is elevated. This is acceptable in light of the fact that the algorithm was only used to compare methods within the SSA subject field, namely, comparison of the different algorithms used to compute the lagged-covariance matrix (section 5.2) and the effect of detrending (section 5.7). The simplified algorithm provides a means of computing an estimate of the R^2 value at a fraction of the computation requirement of the comprehensive algorithm.

The comprehensive algorithm was used to generate the prediction models that are compared to the models obtained using Non-linear Time Series Analysis techniques. Prediction models for seven different window lengths were constructed. Each model has the ability to make predictions for a range of pre-determined step-ahead values.

Figure 5.13, on the following page, represents the comprehensive algorithm used for model construction and validation.

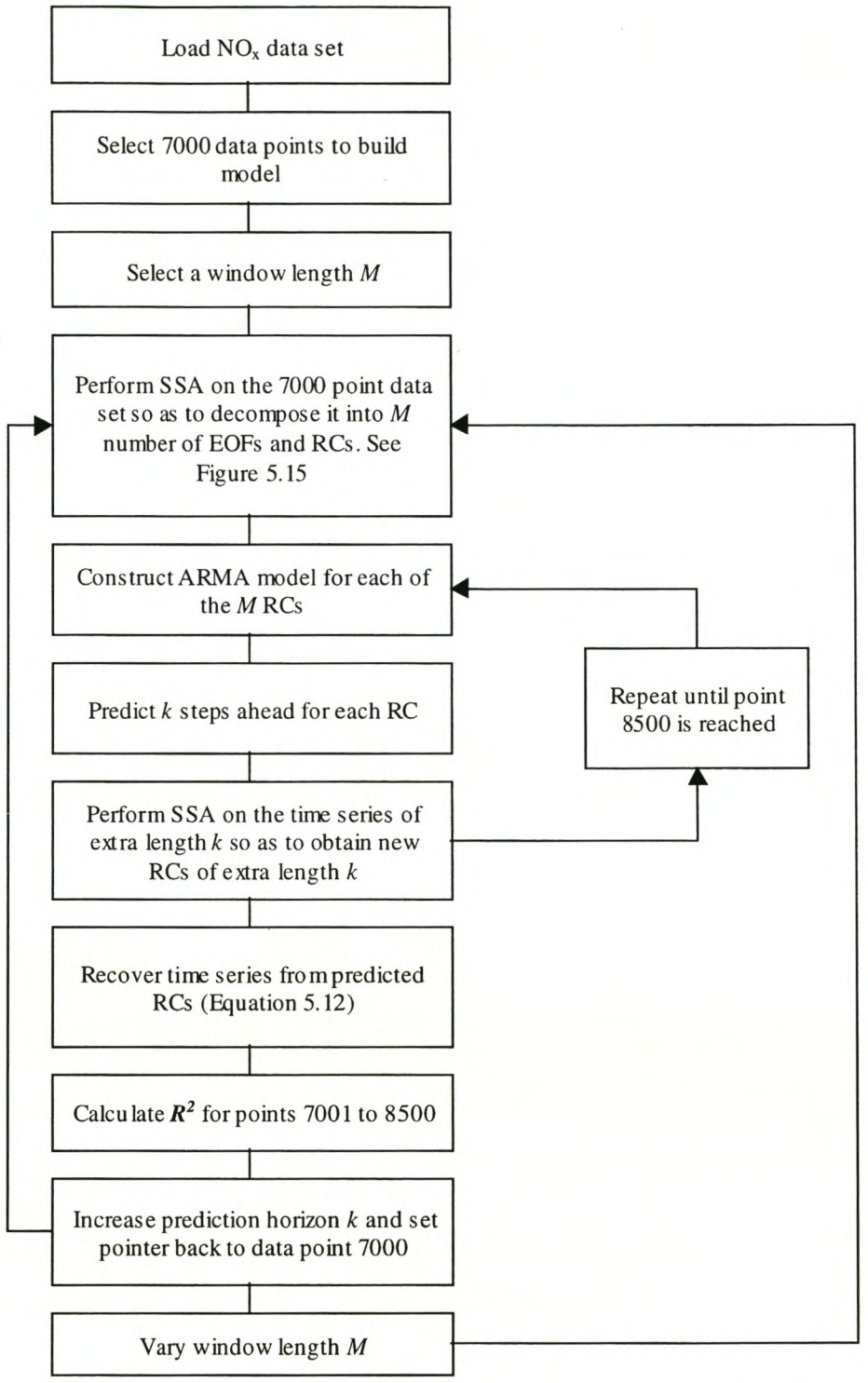


Figure 5.13 Block Diagram representing the Comprehensive Algorithm used for Model Construction and Validation

Figure 5.14 represents the simplified algorithm used to compute an estimate of the R^2 value with minimal computation requirement.

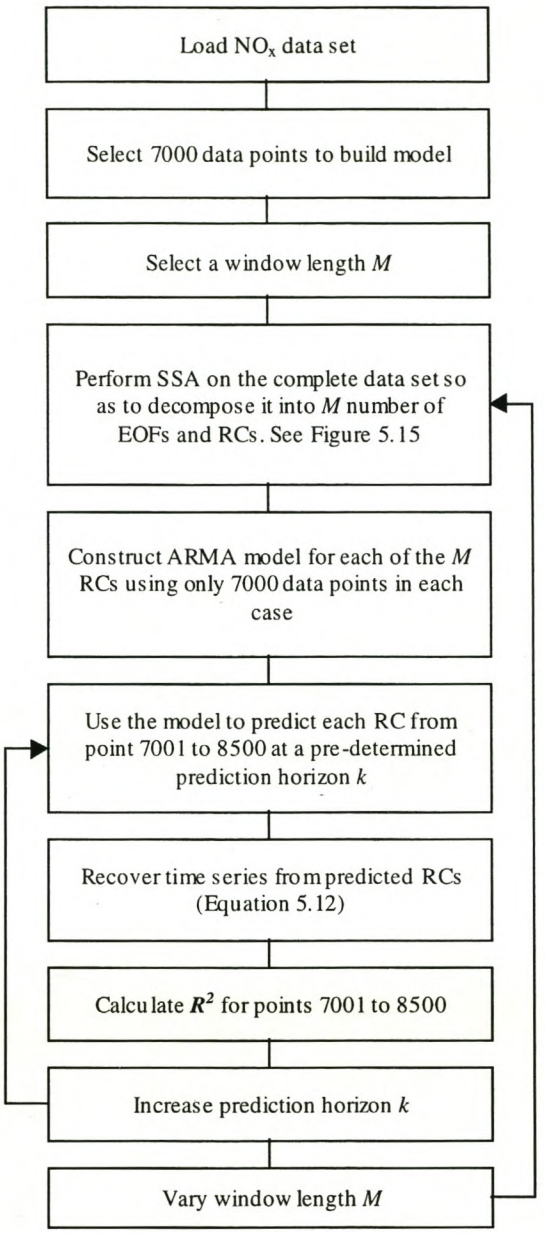


Figure 5.14 Block Diagram representing the Simplified Algorithm

The algorithm to perform SSA on the time series follows the process depicted in Figure 5.15.

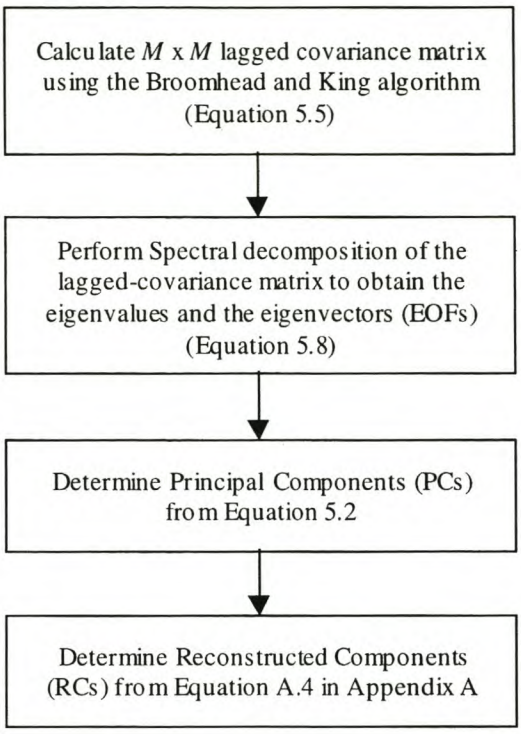


Figure 5.15 Block Diagram representing the Singular Spectrum Analysis Algorithm

An *Auto-regressive Moving Average* (*ARMA*) *Model* was fitted to each RC to effect the prediction of the RC. The ARMA process is of the form:

$$x_n = a_0 + \sum_{i=1}^{P} a_i x_{n-i} + \sum_{j=0}^{Q} b_j \eta_{n-j} \tag{5.18}$$

where a_i and b_j are real constants (a_P and $b_Q \neq 0$), η_n are zero-mean Guassian random numbers (*white noise*), P is the order of the AR part of the model and Q is the order of the MA part of the model. The theory of ARMA modelling is detailed *ad nauseum* in numerous texts (Box & Jenkins, 1976; Cryer, 1986; Graupe, 1984; Ljung, 1987; Stoica & Moses, 1997 and Tong, 1990).

Judicious selection of the model order, P and Q, is important in model optimisation. The model accuracy for the data set under investigation will improve with an increase in the model order. This, however, results in greater model complexity and, hence, an increase in

computational power. More importantly though, too high a model order can result in *overfitting*. The object is to find a model that reflects the general information of the underlying process described in the data set under investigation. If additional parameters are used in the ARMA model, these parameters adjust themselves to the specific features of the noise in that particular data set which is used to build the model (Ljung, 1987). This *overfit* will result in a model that describes the finite data set (used to build the model) and not the general underlying process as intended. Box & Jenkins (1976) strongly advocate the principle of *parsimony* – using the least number of parameters as possible. They stipulate that orders of ARMA models should rarely be above $P = 3$, $Q = 3$.

There are methods available for determining model order that operate under the premise of parsimony - simpler models are favoured over ones that are more complex. These methods include Akaike's Information Criterion (AIC) (Akaike, 1974) and Rissanen's Minimum Description Length (MDL) (Rissanen, 1980). They operate on the principle that proportionately more penalty is assigned to models of increasing complexity, to reflect the cost of obtaining added accuracy. As Ljung puts it: "*If I am going to accept a more complex model (according to my own complexity measure) it has to prove to be significantly better!*" (Ljung, 1987).

The AIC method was applied to a few of the ARMA models which predicted the RCs (A routine in the MATLAB® System Identification Toolbox was used to calculate AIC). This method suggested, in some cases, that the model order could go as high as $(P, Q) = 10$. It was decided that these values were far to high since, considering the high noise content, the parameters could be fitting themselves to the specific features of the noise content in the data set. A more pragmatic approach, used by Graupe (1984), was followed to select the model order. A prediction model with a window length of $M = 24$ hours was constructed from the first 7000 points of the NO_x data set. Each of the RCs was modelled with the same order of ARMA model. The model order, P and Q, of the ARMA model was increased from values of $(P, Q) = 1$ to $(P, Q) = 8$. Twenty-hour step-ahead predictions were carried out from point 7001 to point 8500. Model accuracy, measured out-of-sample by R^2, was plotted for these different model orders. The results are presented in Figure 5.16.

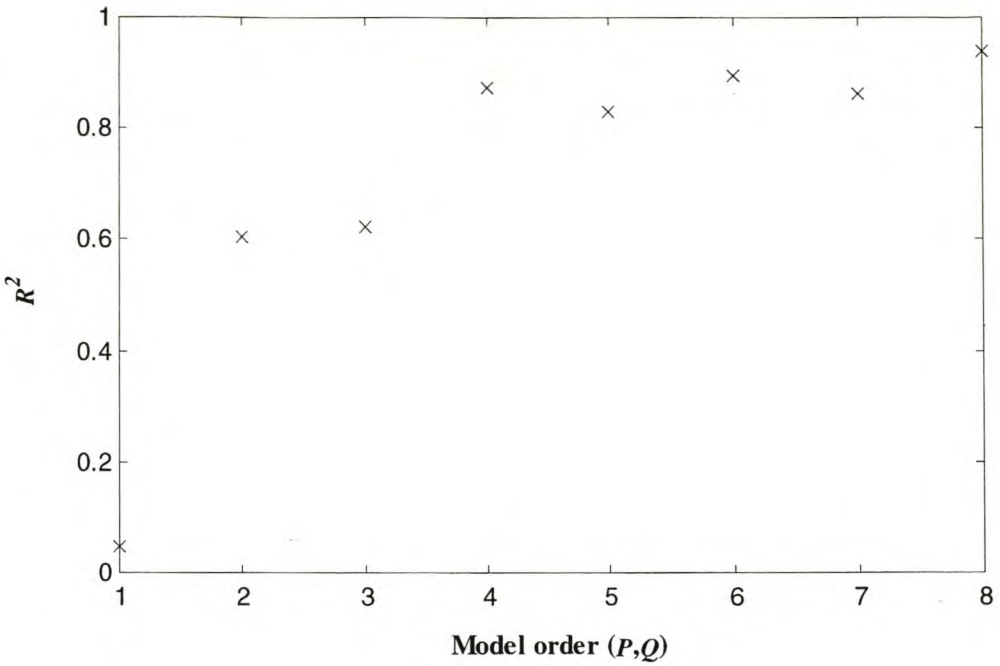


Figure 5.16 R^2 values for various ARMA model orders P, Q (Window Length M = 24 hours)

There is a significant increase in accuracy as the order moves from (P, Q) = 1 to (P, Q) = 2. The following significant increase is from (P, Q) = 3 to (P, Q) = 4 where after the accuracy remains relatively constant with increasing model order. With the principle of parsimony in mind, the most judicious choice of model order would seem to be (P, Q) = 4. The increase in model accuracy moving from (P, Q) = 4 to (P, Q) = 8 is not worth the added complexity of the model. More importantly, with (P, Q) = 8, the risk of overfitting is greatly increased. ARMA models of order (P, Q) = 4 were used to model the RCs in this study. In principle, an optimum model order can be determined for the each individual ARMA model fitted to the corresponding RC, but the added benefit may not be worth the extra effort.

5.11 RESULTS AND DISCUSSION OF NO_x PREDICTION

A graph of prediction accuracy R^2, for different window lengths at a range of step-ahead prediction horizons, is shown in Figure 5.17. The results were produced by using the comprehensive prediction algorithm (as opposed to the simplified prediction scheme) and, therefore, the R^2 data is a true representation of model accuracy. Figure 5.17 was discussed in

section 5.6 pertaining to the choice of the window length M. This discussion is repeated and extended upon for completeness.

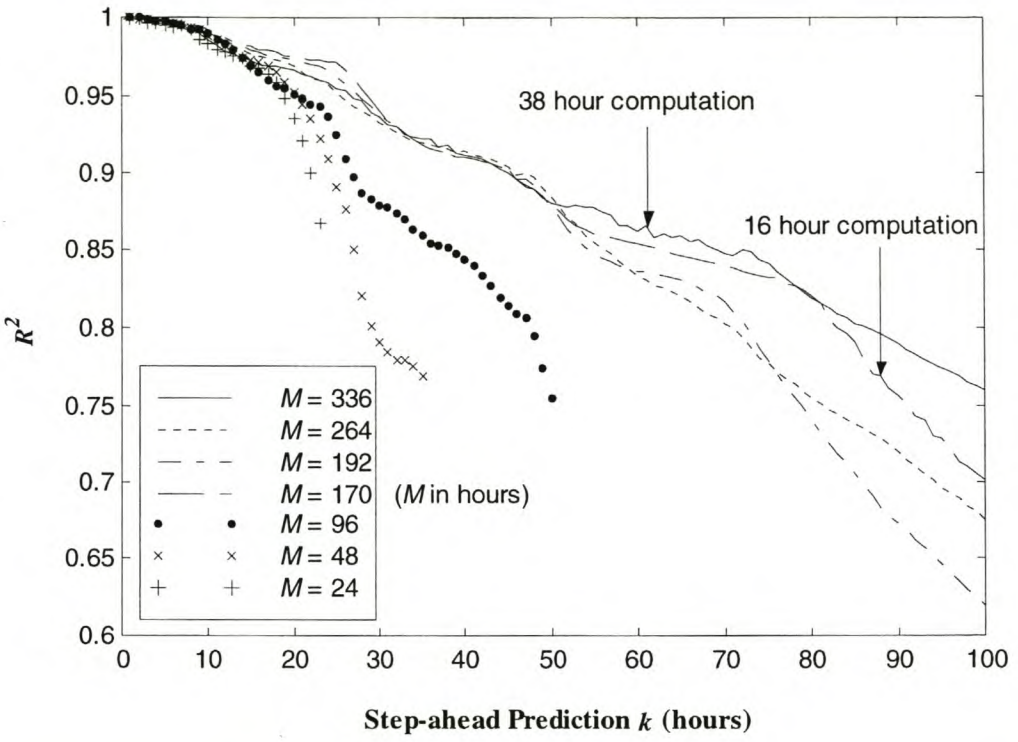


Figure 5.17 Comparison of R^2 for varying Window Length M

The prediction model for M = 170 hours outperforms all the models except for the model of window length M = 336 hours. However, the additional computational requirement to run the model for M = 336 hours does not justify the increase in accuracy. The prediction model for M = 170 hours is, at first sight, also computationally demanding but this is as a result of the interpreted MATLAB® programming language and the limitations of the desktop computer. A compiled version of the code, in a programming language such as C++, and a modern desktop computer would greatly reduce computation time.

The models for the other window lengths perform well initially, but fall away once the prediction horizon exceeds the window length. As mentioned before, SSA does not resolve periods longer than the window length. For this reason, there is a lack of information regarding the longer period oscillations outside the window length and, hence, a drop in prediction accuracy.

So why not make the window length extremely long and obtain a large prediction horizon? Increasing M, and hence increasing the number of EOFs (and eigenvalues), results in a loss of statistical significance. If M is too large, spurious spectral peaks are obtained which can be confused with physically valid peaks. The subject of statistical significance was discussed in section 5.9 above, but a few of the observations are repeated to emphasise their importance.

Monte Carlo SSA can be applied to the time series to determine which of the system's eigenvalues are significant in that they differ from a noise process. This statistical test identifies the statistically significant eigenvalues. Once these significant eigenvalues are identified, the signal can be reconstructed using only these selected eigenvalues. This procedure amounts to filtering of the data set and raises concerns previously mentioned. Non-linear dynamics could be indistinguishable from red noise (Palus & Dvorak, 1992) and thus discarding eigenvalues of lower rank may result in the loss of dynamic information. For this reason, the full eigenspectrum was used to reconstruct the time series.

However, more importantly, the aim of this study is to evaluate two approaches to modelling air pollution and to establish which of these analysis techniques would be the most efficient in modelling air pollution. Thus, it is essential that the two approaches that were used be investigated over a common denominator. No filtering was done in the Non-linear Time Series Analysis and, hence, no filtering was carried out when analysing the signal with SSA. Reconstructing the signal using only selected eigenvalues, amounts to filtering and therefore all the eigenvalues were deemed significant in reconstructing the time series.

To obtain an idea of the prediction accuracy of the model, an arbitrary range of points in the validation set is shown for the 48-hour step-ahead prediction (Figure 5.18) and the 100-hour step-ahead prediction (Figure 5.19). The one-hour step-ahead prediction has an R^2 value of 0.999. Other than the fact that the one-hour step-ahead prediction accuracy is an inadequate indication of model accuracy, the predicted points track the measured observations almost identically and therefore the graph is not shown.

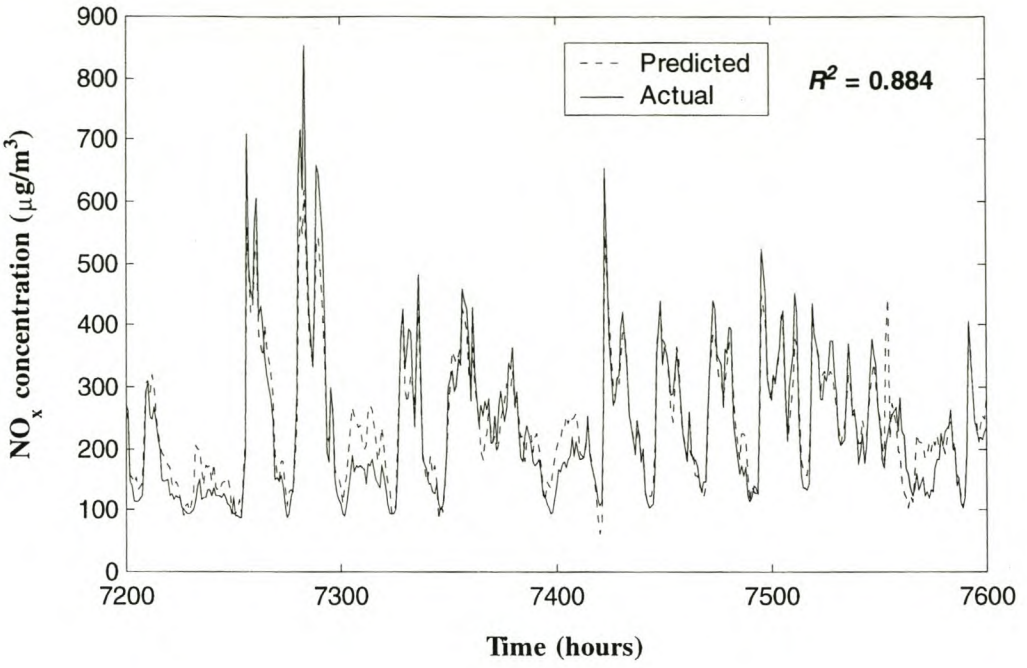


Figure 5.18 Forty-eight hour Step-ahead Prediction

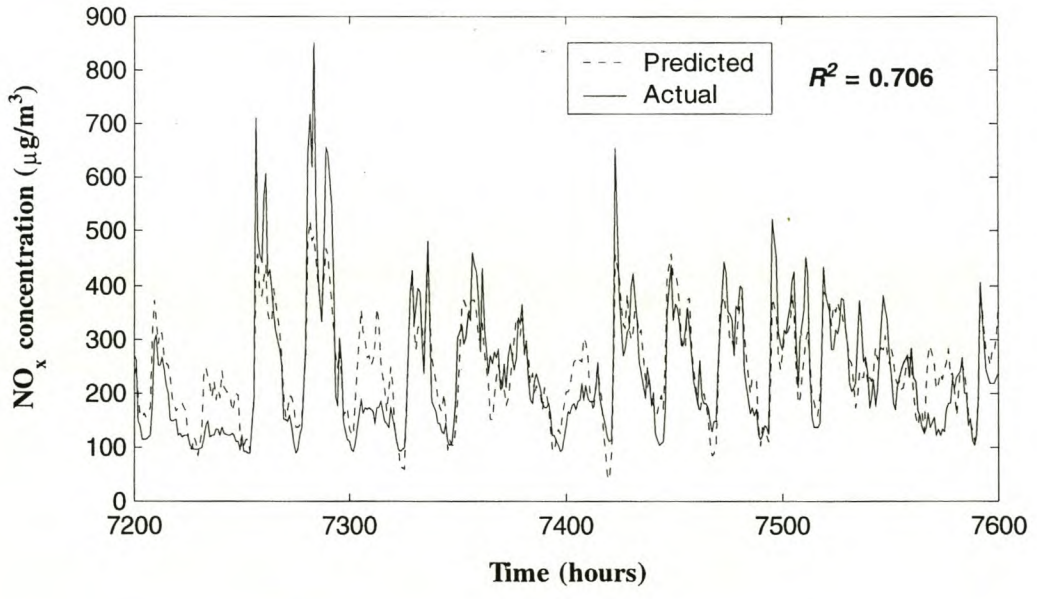


Figure 5.19 One-Hundred hour Step-ahead Prediction

The accuracy achieved with the SSA model is very promising. Even for the 100-hour step-ahead prediction, the predicted values track the actual data set with an acceptable degree of accuracy. A few points are under-predicted (e.g. at around points 7260 and 7290) and other points are overshot (e.g. at around points 7230, 7310 and 7410). However, the trend of the

NO_x concentration is predicted accurately almost without fail. This is relatively impressive considering that this scheme predicts the NO_x concentration a little more than 4 days in advance using only noisy, historic NO_x data to construct the model. What makes SSA so adept at handling this data is the fact that the basis functions (Empirical Orthogonal Functions), in terms of which the data is decomposed, are determined from the time series itself. This data-adaptive characteristic gives SSA greater flexibility and renders it better suited to handling noisy data sets such as the NO_x data.

There is however room for improvement since a greater prediction horizon would be beneficial for effective air pollution modelling. Further research to effectively identify significant eigenvalues without discarding non-linear dynamics will provide an effective means of filtering the data and thus improve the model.

6 A COMPARISON OF THE ANALYSIS TECHNIQUES

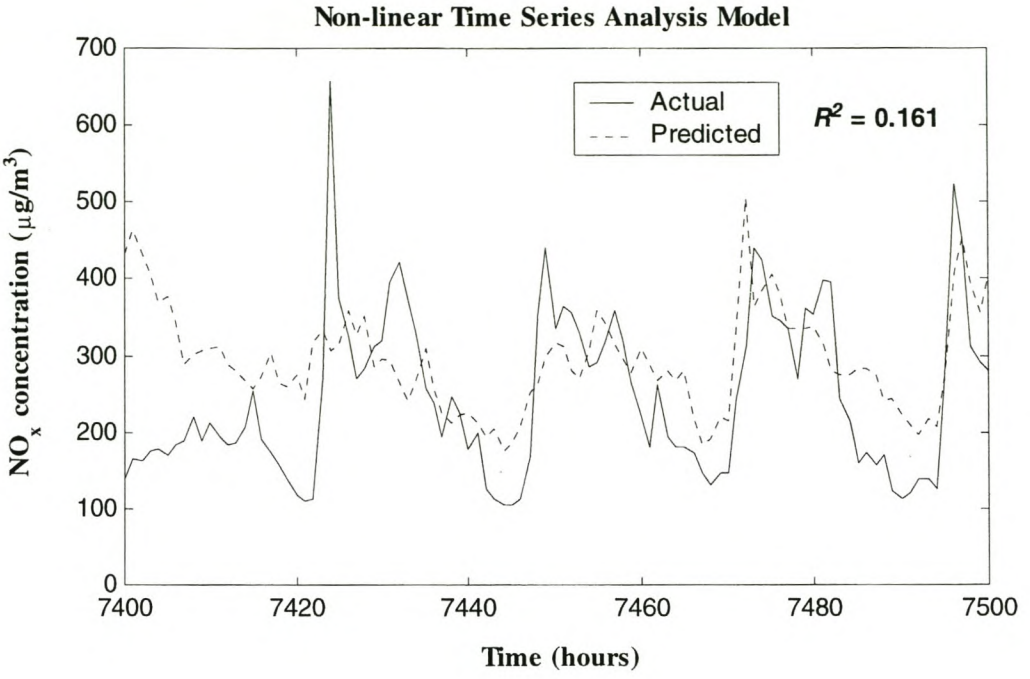
The procedure used to construct the model and to quantify the model accuracy for both the Non-linear Time Series Analysis model and the SSA model is consistent. In both cases 7000 data points were used to construct the model and no noise filtering was done. The R^2 values were calculated out-of-sample for the points 7001 to 8500 (approximately day 292 to day 354, which is roughly halfway through September to the end of the year). This consistency enables the R^2 value to be used as a measurement for the comparison of the two techniques. Table 6.1 presents the R^2 values for the two techniques at different prediction horizons.

Table 6.1 Comparison of R^2 values for the Non-linear Time Series Analysis Model and the SSA Model

Step-ahead Prediction	R^2	
	Non-linear Time Series Model	SSA Model
1	0.730	0.999
48	0.161	0.884
100	0.042	0.706

It is evident that the accuracy of the SSA model is markedly superior to the Non-linear Time Series model. The paramount reason for the superior accuracy of the SSA model is its adept ability to analyse and cope with noisy data sets such as the NO_x data set. The surrogate data test of section 4.7.3 showed that, although the NO_x data set displays non-linear determinism, it has a high random noise content. This high noise content posed a problem for the Non-linear Time Series Analysis technique. The noise contamination hinders the extraction of the system's underlying dynamics and, thus, the model is unable to provide an adequate representation of the underlying physical process.

A visual representation of the model comparison is shown below. An arbitrary range of points in the validation set is shown for the 48-hour step-ahead prediction (Figure 6.1) and the 100-hour step-ahead prediction (Figure 6.2).



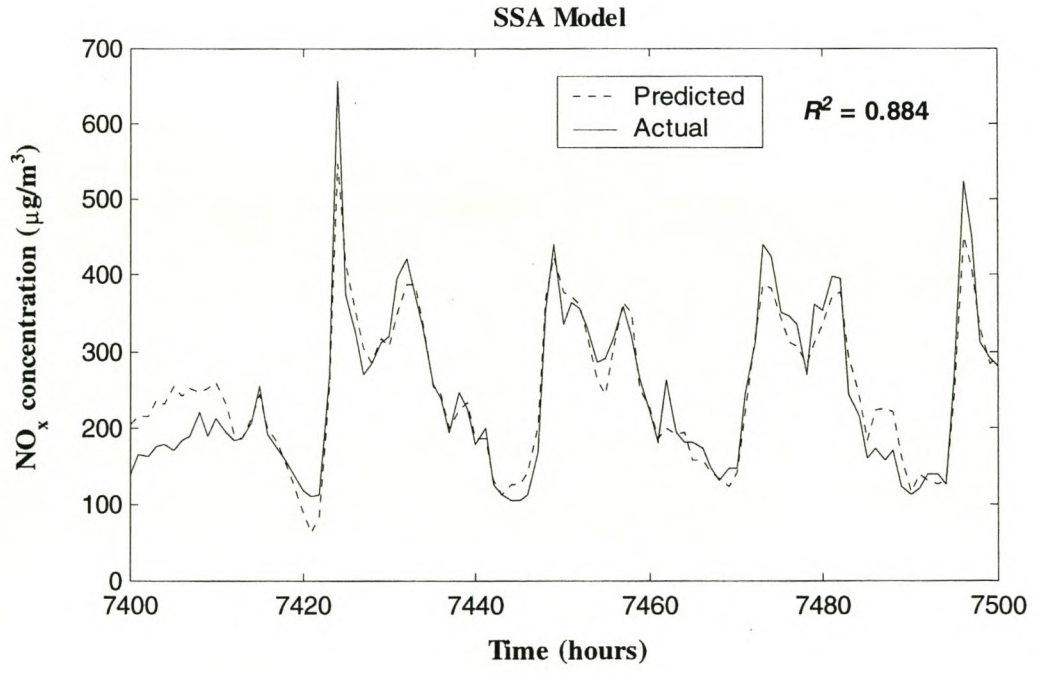


Figure 6.1 Comparison of the Models for a 48-hour Step-ahead Prediction

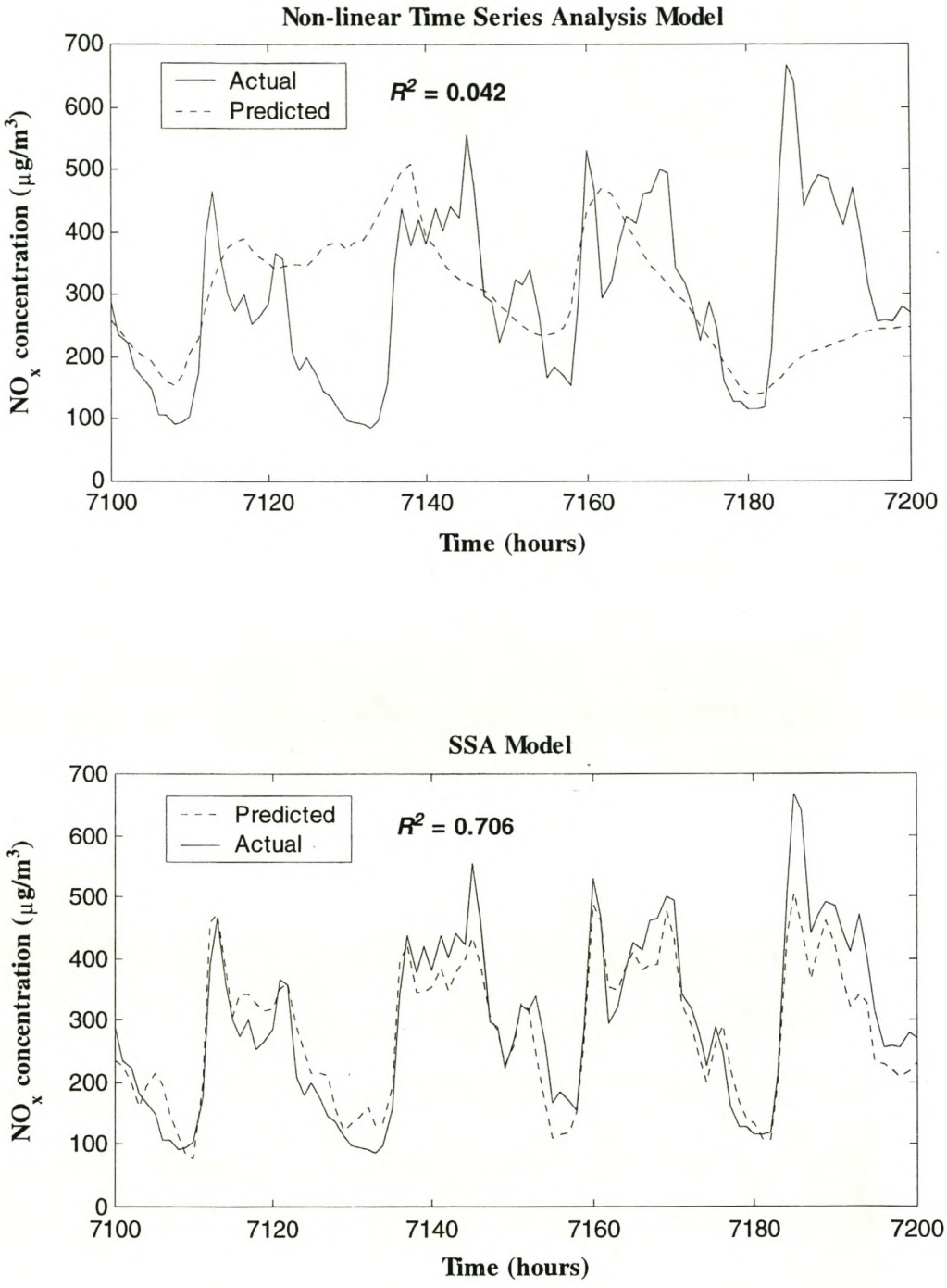


Figure 6.2 Comparison of the Models for a 100-hour Step-ahead Prediction

7 CONCLUSIONS AND RECOMMENDATIONS

Air pollution models are important tools for the management of air quality control. A fundamentally sound and accurate air pollution model can be useful for performing important sensitivity analyses in demographic and metropolitan planning and management. In the near future, it is unlikely that South African cities will have the spatial, time-resolved information of emission sources to apply complex dispersion models that address the physical and chemical processes of air pollution from first principles. On the other hand, models based on the Gaussian Plume Model are restricted by the excessive simplifying assumptions. Statistical and mathematical analysis techniques provide a solution to the problems encountered in using the aforementioned models. There are no simplifying assumptions that adversely affect the model accuracy and the model can be run on a modern desktop computer. For these reasons, it is evident that these techniques provide the most viable approach to modelling air pollution in the Cape Metropole.

This study investigated the techniques of Non-linear Time Series Analysis and Singular Spectrum Analysis (SSA) for the modelling of air pollution data – specifically, hourly NO_x data from the year 1995. The aim of the study was to establish which of these two techniques is best suited to the modelling of air pollution data. To this end, NO_x prediction models for each of the techniques were constructed and the prediction accuracy was compared.

7.1 CONCLUSIONS

The theory and application of Non-linear Time Series Analysis was discussed in Chapter 4. It was argued, from a pragmatic viewpoint, that the dynamics that governed the NO_x pollution system are inherently non-linear due to the strong correlation with weather patterns and the complexity of the chemical reactions and physical transport of the pollutant. In addition to this, the method of surrogate data showed that the NO_x data displayed a certain degree of non-linear determinism that made it amenable to non-linear analysis. The necessary analysis was carried out on the data and a prediction model was constructed with the use of a neural network. The prediction accuracy (computed on an out-of-sample validation set using the square of the Pearson product moment correlation co-efficient, R^2) for a 48-hour step-ahead prediction was $R^2 = 0.161$. For a 100-hour step-ahead prediction scheme, a prediction

accuracy of $R^2 = 0.042$ was attained. These poor results for the prediction accuracy indicate that the model was unable to provide an adequate representation of the underlying physical process. The principal reason for the inadequacy of the Non-linear Time Series Analysis technique is the high noise content of the NO_x data set. The surrogate data test of section 4.7.3 indicated that, although there is evidence of non-linear determinism in the original data set, there is a high random noise content. This noise content contaminates the system's underlying dynamics thereby hindering the effectiveness of the analysis and, thus, limiting the prediction horizon. Extremely careful application of a specialised noise reduction scheme (beyond the scope of this study) could filter the contaminated data without discarding too much of the higher order dynamics. This would enhance the analysis procedure and therefore improve the prediction horizon.

The method of Singular Spectrum Analysis (SSA) was discussed and applied to the NO_x data set in Chapter 5. Although SSA is a linear data analysis technique, the basis functions, in terms of which the data is decomposed, are data-adaptive which makes it well suited to the analysis of non-linear systems exhibiting anharmonic oscillations. The data-adaptive basis functions decompose the time series into statistically independent components that have limited harmonic content. Consequently, these components are more amenable to prediction than the time series itself. Once these components have been predicted, the time series can be recovered by reconstructing a signal from a convolution of the predicted components. SSA's ability has been proven in the analysis of short, noisy non-linear signals. Prompted by these successes and the flexibility of the technique, SSA was carried out on the NO_x data set. The predictive model was constructed using an Auto-regressive Moving Average (ARMA) model. Prediction accuracy of the model was determined for a range of step-ahead prediction horizons using precisely the same calculation procedure as was used for the Non-linear Time Series model. For a 48-hour step-ahead prediction, an accuracy of $R^2 = 0.884$ was achieved. For a 100-hour step-ahead prediction scheme, the prediction accuracy was $R^2 = 0.706$. These results are very promising and bode well for the success of an air pollution model based on Singular Spectrum Analysis. Further improvement of the model could be achieved via an effective filtering scheme that ensures that non-linear dynamics are not destroyed during the filtering process. Although less computationally demanding than Non-linear Time Series Analysis, the SSA model required substantial computation time. This was mainly because of the interpreted MATLAB® programming language used to construct the model and the limitations of the desktop computer.

The accuracy of the SSA model is markedly superior to the Non-linear Time Series model. The paramount reason for the superior accuracy of the SSA model is its adept ability to analyse and cope with noisy data sets such as the NO_x data set.

The study provides evidence to suggest that Singular Spectrum Analysis is better suited to the modelling of air pollution data. It should therefore be the analysis technique of choice when more advanced, multivariate modelling of air pollution data is carried out.

7.2 RECOMMENDATIONS

Based on the conclusions of this study, the following recommendations are made:

i. Noise reduction schemes, which decontaminate the data without destroying important higher order dynamics, should be researched. The application of an effective noise reduction scheme could lead to an improvement in model accuracy.

ii. The univariate SSA model should be extended to a more complex multivariate model that explicitly encompasses variables such as traffic flow and weather patterns. This will explicitly expose the inter-relationships between the variables and will allow a sensitivity study of the effects that the variables have on the pollution system. In addition to this, a multivariate model will have the ability to predict the outcome to a multitude of scenarios.

iii. The programming code that is used to construct the model should be compiled using a programming language such as C^{++}. This will guarantee an improvement in the computation time of the air pollution model. A modern state-of-the-art desktop computer will have to be employed to cope with a comprehensive, multivariate air pollution model.

REFERENCES

Abarbanel, H.D.I. (1996). *Analysis of Observed Chaotic Data*. Springer-Verlag, New York.

Addison, P.S. (1997). *Fractals and Chaos: An Illustrated Course*. Institute of Physics Publishing, Bristol and Philadelphia.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Trans. Automat. Contr.*, **AC-19**, No. 6, 716-723.

Allen, M.R. (1992). *Interactions between the Atmosphere and Oceans on Time Scales of Weeks to Years*. Ph.D. thesis, St. John's College, Oxford, United Kingdom.

Allen, M.R. & Smith, L.A. (1996). Monte Carlo SSA: Detecting irregular oscillations in the presence of coloured noise. *J. Climate*, **9**, No. 12, 3373-3404.

Allen, M.R. & Smith, L.A. (1997). Optimal Filtering in singular spectrum analysis. *Phys. Lett. A*, **234**, 419-428.

Bailie, R.S., Ehrlich, R.I. & Truluck, T.F. (1994). Trends in photochemical smog in the Cape Peninsula and the implications for health. *South African Medical Journal*, **84**, No. 11, 738-742.

Barnard, J.P. (1999). *Empirical State Space Modelling with Application in Online Diagnosis of Multivariate Non-linear Dynamic Systems*. Ph.D. thesis, University of Stellenbosch, South Africa.

Box, G.E.P., Hunter, W.G. & Hunter, J.S. (1978). *Statistics for Experimenters: An Introduction to Design, Data Analysis, and Model Building*. John Wiley and Sons, New York.

Box, G.E.P. & Jenkins, G.M. (1976). *Time Series Analysis, Forecasting and Control*. Holden-Day, San Francisco.

Broomhead, D.S. & King, G.P. (1986). Extracting qualitative dynamics from experimental data. *Physica D*, **20**, 217-236.

Casdagli, M. (1989). Nonlinear prediction of chaotic time series. *Physica D*, **35**, 335-356.

Cryer, J.D. (1986). *Time Series Analysis*. PWS Publishers, Boston.

Diab, A.F. & Barnard, J.P. (1999). Non-linear Time Series Analysis: An Approach to Modelling NO_x Concentrations. Paper presented at the National Association for Clean Air Annual Conference, Cape Town, October 1999.

Dzubay, T.G. (1982). Visibility and aerosol composition in Houston, Texas. *Envir. Sci. Technol.*, **16**, 514-525.

Elsner, J.B. & Tsonis, A.A. (1996). *Singular Spectrum Analysis: A New Tool in Time Series Analysis*. Plenum Press, New York.

Farmer, J.D., Ott, E. & Yorke, J.A. (1983). The dimension of chaotic attractors. *Physica D*, **7**, 153-180.

Flepp, L., Holzner, R., Brun, E., Finardi, M. & Badii, R. (1991). Model identification by periodic-orbit analysis for NMR-laser chaos. *Phys. Rev. Lett.*, **67**, 2244-2247.

Fraser, A.M. & Swinney, H.L. (1986). Independent coordinates for strange attractors from mutual information. *Phys. Rev. A*, **33**, 1134-1140.

Grassberger, P. & Procaccia, I. (1983). Measuring the strangeness of strange attractors. *Physica D*, **9**, 189-208.

Grassberger, P. & Procaccia, I. (1983a). Characterization of strange attractors. *Phys. Rev. Lett.*, **50**, 346-349.

Graupe, D. (1984). *Time series analysis, identification and adaptive filtering*. Robert E. Krieger Publishing Company Inc., Florida.

Greenberg, M.D. (1988). *Advanced Engineering Mathematics*. Prentice-Hall, New Jersey.

Grobliki, P.J., Wolff, G.T. & Countess, R.J. (1981). Visibility reducing species in the Denver "Brown Cloud". *Atmospheric Environment*, **15**, No. 12, 2473-2484.

Hassounah, M.I. & Miller, E.J. (1994). Modelling air pollution from road traffic: a review. *Traffic Engineering and Control*, September, 510-514.

Heywood, J.B. (1988). *Internal Combustion Engine Fundamentals*. McGraw-Hill, Singapore.

Hypertext reference 1: http://www.statsoft.com

Hypertext reference 2: http://www.mpipks-dresden.mpg.de/~tisean/TISEAN_2.0/

Joliffe, I.T. (1986). *Principal Component Analysis*. Springer, New York.

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

Judd, K. (1994). Estimating dimension from small samples. *Physica D*, **71**, 421-429.

Kantz, H. & Schreiber, T. (1997). *Nonlinear Time Series Analysis*. Cambridge University Press, United Kingdom.

Kennel, M.B., Brown, R. & Abarbanel, H.D.I. (1992). Determining minimum embedding dimension for phase-space reconstruction using a geometrical construction. *Phys. Rev. A*, **45**, 3403-3411.

Keppene, C.L. & Ghil, M. (1992). Adaptive filtering and prediction of the southern oscillation index. *J. Geophys. Res.*, **97**, 20449-20454.

Kumaresan, R. & Tufts, D.W. (1980). Data-adaptive principal component signal processing. In IEEE Proc. Conf. on Decision and Control, Albuquerque, 1980.

Levenberg, K. (1944). A method for the solution of certain nonlinear problems in least squares. *Quart. Applied Mathematics*, **2**, 164-168.

Livezey, R.E. & Chen, W.Y. (1982). Statistical field significance and its determination by Monte Carlo techniques. *Mon. Wea. Rev.*, **110**, 46-57.

Ljung, L. (1987). *System Identification: Theory for the User*. Prentice-Hall, New Jersey.

Lorenz, E.N. (1963). Deterministic non-periodic flow. *J. Atmos. Sci.*, **20**, 130-141.

Marquardt, D.W. (1963). An algorithm for least squares estimation of nonlinear parameters. *Journal of SAIM*, **11**, 431-441.

Mees, A.I. & Judd, K. (1993). Dangers of geometric filtering. *Physica D*, **68**, 427-436.

Osborne, A.R. & Provenzale, A. (1989). Finite correlation dimension for stochastic systems with power-law spectra. *Physica D*, **35**, 357-381.

Palus, M. & Dvorak, I. (1992). Singular-value decomposition in attractor reconstruction: pitfalls and precautions. *Physica D*, **55**, 221-234.

Penland, C., Ghil, M. & Weickmann, K.M. (1991). Adaptive filtering and maximum entropy spectra with application to changes in atmospheric angular momentum. *J. Geophys. Res.*, **96**, 22659-22671.

Press, W.H., Flannery, B.P., Teukolsky, S.A. & Vetterling, W.T. (1989). *Numerical Recipes – The Art of Scientific Computing* (*FORTRAN version*). Cambridge University Press.

Rissanen, J. (1980). Consistent order estimates of autoregressive processes by shortest description of data. In Jacobs, O. *et al.*, editors, *Analysis and Optimisation of Stochastic Systems*, Academic, New York.

Rosie, A.M. (1966). *Information and Communication Theory*. Gordon and Breach, New York.

Salzmann, B. (1962). Finite amplitude free convection as an initial value problem-I. *J. Atmos. Sci.*, **29**, 329-341.

Sauer, T., Yorke, J.A. & Casdagli, M. (1991). Embedology. *J. Stat. Phys.*, **65**, 579-616.

Schreiber, T. (1997). Detecting and analysing nonstationarity in a time series with nonlinear cross prediction. *Phys. Rev. Lett.*, **78**, 843-846.

Schwarz, G. (1978). Estimating the dimension of a model. *Ann. Stat.*, **6**, 461-464.

Small, M. & Judd, K. (1998). Comparison of new nonlinear modeling techniques with applications to infant respiration. *Physica D*, **117**, 283-298.

Small, M. & Judd, K. (1998a). Correlation Dimension: A pivotal statistic for non-constrained realizations of composite hypotheses in surrogate data analysis. *Physica D*, **120**, 386-400.

Stoica, P. & Moses, RL. (1997). *Introduction to Spectral Analysis*. Prentice-Hall, Upper Saddle River, New Jersey.

Takens, F. (1981). *Detecting strange attractors in turbulence*. Lecture Notes in Mathematics, **898**, 366-381, Springer, Berlin.

Terblanche, P. (1995). Motor Vehicles Emission Policy Development: Phase 1. Report no. EV9404, Department of Mineral and Energy Affairs, South Africa.

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

Theiler, J., Galdrikian, B., Longtin, A., Eubank, S. & Framer, J.D. (1992a). Using surrogate data to detect nonlinearity in time series. In Casdagli, M. & Eubank, S. (1992), *Nonlinear Modeling and Forecasting*. Sante Fe Institute Studies in the Science of Complexity, Proc. Vol. XII, Addison-Wesley Publishing Company.

Theiler, J. & Pritchard, D. (1996). Constrained-realization Monte-Carlo method for hypothesis testing. *Physica D*, **94**, 221-235.

Tong, H. (1990). *Non-linear Time Series*. Oxford University Press, Oxford.

Turner, D.B. (1994). *Atmospheric Dispersion Estimates – an introduction to dispersion modelling*. Lewis Publishers, Florida.

Vautard, R. & Ghil, M. (1989). Singular spectrum analysis in nonlinear dynamics with applications to paleoclimatic time series. *Physica D*, **35**, 395-424.

Vautard, R., Yiou, P. & Ghil, M. (1992). Singular spectrum analysis: A toolkit for short, noisy chaotic signals. *Physica D*, **58**, 95-126.

Wicking-Baird, M.C., De Villiers, M.G. & Dutkiewicz, R.K. (1997). *Cape Town Brown Haze Study*. Energy Research Institute, University of Cape Town, South Africa.

Yiou, P., Ghil, M., Jouzel, J., Paillard, D. & Vautard, R. (1994). Nonlinear variability of the climate system from singular and power spectra of the quaternary records. *Climate Dyn.*, **9**, 371-389.

APPENDIX A

Here follows the mathematical motivation of Vautard *et al.* for the Reconstructed Components (Vautard *et al.*, 1992):

Consider a subset A of eigenelements k over which the reconstruction is to be performed. In analogy with equation (6.1), it is necessary to seek a series, $y = R_A x$, of length N such that the quantity

$$H_A(y) = \sum_{i=0}^{N-M} \sum_{j=1}^{M} \left(y_{i+j} - \sum_{k \in A} a_i^k E_j^k \right)^2 \qquad \text{(A.1)}$$

is minimised. In other words, the optimal series y is the one whose augmented version Y is the closest, in the least square sense, to the projection of the augmented series X onto the EOFs with indices belonging to A. The solution $y = R_A x$ to this least-squares problem is given by

$$(R_A x)_i = \frac{1}{M} \sum_{j=1}^{M} \sum_{k \in A} a_{i-j}^k E_j^k \qquad \text{for } M \le i \le N\text{-}M\text{+}1 \qquad \text{(A.2)}$$

$$(R_A x)_i = \frac{1}{i} \sum_{j=1}^{i} \sum_{k \in A} a_{i-j}^k E_j^k \qquad \text{for } 1 \le i \le M\text{-}1 \qquad \text{(A.3)}$$

$$(R_A x)_i = \frac{1}{N-i+1} \sum_{j=i-N+M}^{M} \sum_{k \in A} a_{i-j}^k E_j^k \qquad \text{for } N\text{-}M\text{+}2 \le i \le N \qquad \text{(A.4)}$$

When A consists of a single index k, the series $R_A x$ is called the k^{th} RC, and will be denoted by r^k. RCs have additive properties:

$$R_A x = \sum_{k \in A} r^k \qquad \text{(A.5)}$$

In particular, the series can be expanded as the sum of its RCs:

$$x_i = \sum_{k=1}^{M} r_i^k \qquad i = 1,2,\ldots,N \tag{A.6}$$

where r_i^k is the value of the k^{th} RC at time i.

APPENDIX B

This is an illustration of the first fourteen reconstructed components (RCs) of a segment of the NO_x data set.