On Nonlinear Processing of Air Pollution Data

Rob Foxall*[1], Igor Krcmar†, Gavin Cawley*, Stephen Dorling‡ and Danilo P. Mandic*

*School of Information Sytems, University of East Anglia, Norwich, UK, e-mail: {rjf, gcc, mandic}@sys.uea.ac.uk †Faculty of Electrical Engineering, University of Banjaluka, Banjaluka, e-mail: ikrcmar@etf-bl.rstel.net ‡School of Environmental Sciences, University of East Anglia, Norwich, UK, e-mail: s.dorling@uea.ac.uk

Abstract

Three methods – DVS plots, attractor reconstruction, and variance analysis of delay vectors – for detecting nonlinearities in time series are compared on an air pollution dataset. For rigour each method is also used on a surrogate dataset, based on a high-order linear fit to the original data. Finally, a comparison of a standard linear analysis to a neural network model analysis of the air pollution dataset is provided.

1 Introduction

Air pollutants such as surface Ozone (O_3), Nitrogen Oxides (NO_X), Sulphur Dioxide (SO_2) and Particulates have significant health effects associated with them at high concentrations. A rigorous analysis of pollutant data requires consideration of a number of meteorological variables (e.g. wind speed) and non-meteorological variables (e.g. traffic density). To obtain an insight into the underlying structure, however, it is worthwhile to look initially at each pollutant time series individually with the standard linear methods, and to do nonlinearity analysis only if it appears that a linear model is inadequate. A NO_2 time series of hourly measurements taken over a four-year period from the Leeds meteo station is used throughout this paper.

2 Linear Analysis of Time Series

A standard model of linear time series, the ARIMA(p, d, q) model popularised by Box and Jenkins [1], assumes that the time series $x(j)$ is generated by a succession of "random shocks" $\epsilon(j)$, drawn from a distribution with zero mean and variance σ_ϵ^2. If $x(j)$ is non-stationary, then successive differencing of $x(j)$ via the differencing operator, $\nabla x(j) = x(j) - x(j-1)$ can provide a stationary process. A stationary process $z(j) = \nabla^d x(j)$ can be modelled as an autoregressive moving average

$$z(j) = \sum_{i=1}^{p} a_i z(j-i) + \sum_{i=1}^{q} b_i \epsilon(j-i) + \epsilon(j) \quad (1)$$

Of particular interest are pure autoregressive (AR) models, which have an easily understood relationship to the nonlinearity detection technique of DVS (Deterministic Versus Stochastic) plots. The order of the AR model can be chosen by the point where the *autocorrelation function* (ACF) essentially vanishes for all subsequent lags, other methods, such as AIC or BIC, can also be used. Figure 1 shows

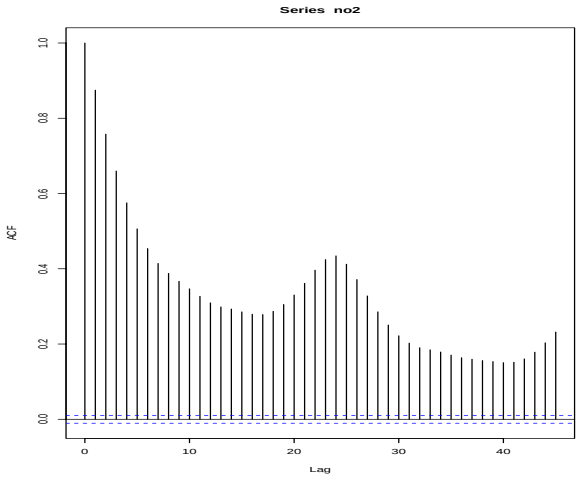

Fig. 1. ACF plot of NO_2 series.

the autocorrelation function for 40 lags for the NO_2 dataset; the ACF does not vanish and a high-order AR model is necessary.

3 The Method of Surrogate Data

Following the approach from [2], to gauge efficacy of the techniques for detecting nonlinearity, a surrogate dataset is simulated from a high order autoregressive model fit to the original series. The coefficients a_i from an $AR(45)$ model were used to generate the surrogate series, with surrogate residuals

[1]This work was supported by the European Commission, grant number IST-99-11764, as part of its Framework V IST programme.

$\epsilon(j)$ taken as a random permutation of the residuals from the original series. Evidence of nonlinearity from any method of detection is negated if the method gives a similar result when applied to the surrogate series, which is known to be linear [2].

4 Attractor Reconstruction

Existence and/or discovery of an attractor in phase space demonstrates whether the system is deterministic, purely stochastic, or somewhere in between. To reconstruct the attractor examine plots in m-dimensional space of $[x(j), x(j-\tau), \ldots, x(j-(m-1)\tau)]^T$. It is critically important for the dimension of the space, m, in which the attractor is to viewed, to be large enough to "untangle" the attractor. This is known as the *embedding dimension*. The value of τ, the *lag time* or *lag spacing*, is also important, particularly with noise present. The first inflection point on the autocorrelation function is a possible starting value for τ [3]. Alternatively, if the series is known to be sampled coarsely, the value of τ can be taken as unity [4].

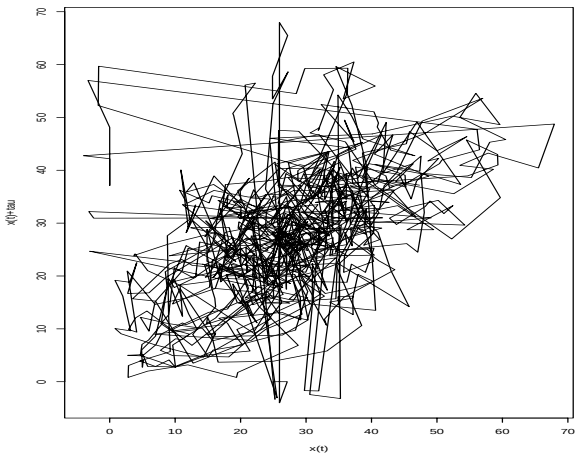

Fig. 2. NO_2 time series embedded into 2-dimensional phase-space with delay time $\tau = 22$.

Figure 2 shows the two-dimensional attractor reconstruction for the NO_2 time series after it has been passed through a linear filter to remove some of the noise present. Although the graph shows some regularity, if an attractor exists it is in a higher dimensional space.

5 DVS Plots

DVS plots [4] display the (robust) prediction error $E(k)$ for local linear models against the number of nearest neighbours, k, used to fit the model, for a range of embedding dimensions m. The last 500 values of the series are set aside for prediction purposes, these values are known as the *test* set. For each element in the test set $x(j)$, construct the delay vector $\mathbf{x}(j) = [x(j), x(j-\tau), \ldots, x(j-(m-1)\tau)]^T$. The k nearest neighbours are defined to be the k vectors $\mathbf{x}(j')$ from the series which have the shortest Euclidean distance to $\mathbf{x}(j)$, these k nearest neighbours are used to fit the local linear model.

If the optimal k, taken to be the value of k giving the lowest prediction error $E(k)$, is at or close to the maximum possible k, then globally linear models perform best and there is no indication of nonlinearity. In this case the model is equivalent to an AR model of order m when $\tau = 1$. Small optimal k suggests local linear models perform best, indicating nonlinearity and/or chaotic behaviour.

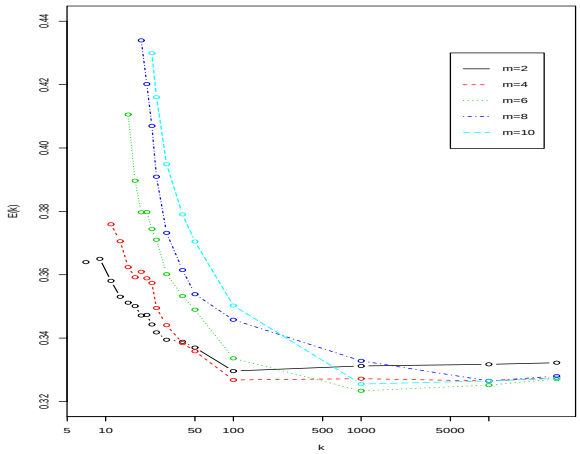

Fig. 3. DVS plot for the NO_2 time series with lead time $T = 1$ and lag delay $\tau = 22$.

In Figure 3 the DVS plot for $m = \{2, 4, 6, 8, 10\}$ is shown. For each value of m, the optimal k is less than the maximum, but the difference in the prediction error is minimal.

Figure 4 displays the equivalent DVS plot for a surrogate dataset simulated from the $AR(45)$ model fit to the series. The behaviour for the surrogate data is similar to the original data, suggesting that the underlying structure of the series has only a small nonlinear component.

6 Variance Analysis of Delay Vectors

Closely related to DVS plots is the nonlinearity technique introduced in [5]. For each observation $x(i), i \geq m+1$ construct the group, Ω_i, of nearest neighbours by

$$\Omega_i = \{\mathbf{x}(j) : j \neq i \quad \& \quad d_{ij} \leq \alpha A_x\}$$

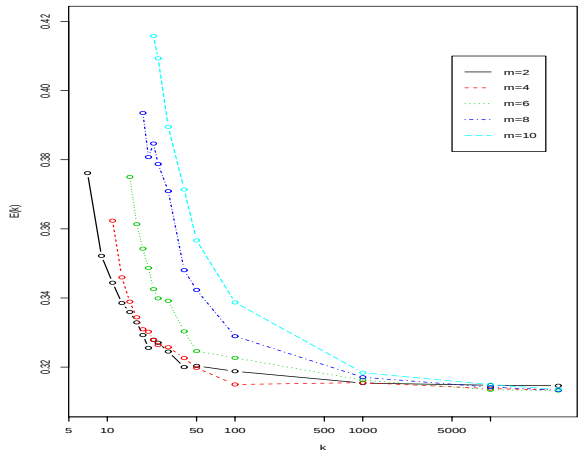

Fig. 4. DVS plot for the simulated NO_2 time series with lead time $T = 1$ and lag delay $\tau = 22$.

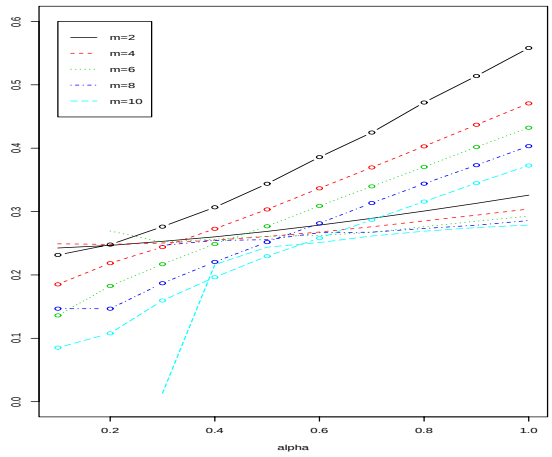

Fig. 5. Normalised variance plot. The lines with points correspond to the actual NO_2 series; the lines without points to the simulated linear series.

where $\mathbf{x}(j) = \{x(j-1), x(j-2), \ldots, x(j-(m-1))\}$, $d_{ij} = \|\mathbf{x}(j) - \mathbf{x}(i)\|$ is the Euclidean norm, $0 < \alpha \leq 1$, $A_x = \frac{1}{N} \sum_{i=m+1}^{N} |x(i)|$ and N is the length of the time series. If the series is linear, then the similar patterns $\mathbf{x}(j)$ belonging to a group Ω_i will map onto similar $x(j)$s. For nonlinear series, the patterns $\mathbf{x}(j)$ will not map onto similar $x(j)$s. This is measured by the variance σ^2 of each group Ω_i

$$\sigma_i^2 = \frac{1}{|\Omega_i|} \sum_i (x(j) - \mu_i)^2, \qquad \mathbf{x}(j) \in \Omega_i.$$

The measure of nonlinearity is taken to be the mean of σ_i^2 over all the Ω_i, denoted $\bar{\sigma_N^2}$, normalised by dividing through by σ_x^2, the variance of the entire time series $\bar{\sigma^2} = \frac{\bar{\sigma_N^2}}{\sigma_x^2}$. The larger the value of $\bar{\sigma^2}$ the greater the suggestion of nonlinearity [5].

The results are shown in Figure 5. Apart from a few exceptions for $\alpha < 0.5$, the normalised variance of similar delay vectors for the simulated series is much lower than for the real series, an indication that the series is nonlinear.

7 Neural Adaptive Filters in the Air Pollution Time Series Prediction

As stated above prediction of air pollution time series is a difficult task due to the complex and cyclic nature of the underlying process that generates atmospheric pollutants. In addition, some results, i.e. DVS plots and attractor reconstruction, indicate inherent nonlinearity of the air pollution time series. Thus, in order to obtain good prediction of the future value of the time series, based on the past measurements, an efficient algorithm should be employed, an algorithm that is inherently nonlinear and/or adaptive. Gradient-descent (GD) based neural adaptive filters, due to inherent simplicity and nonlinearity, are adequate choices for the prediction of time series that represents atmospheric pollution data. Furthermore, the structure of neural adaptive filters could be chosen to reflect the nature of the underlying process, i.e. it could be feedforward or recurrent.

The adaptation of a GD based neural adaptive filter can be described by the following set of equations

$$v(k) = \mathbf{w}^T(k)\mathbf{u}(k) \quad (2)$$

$$e(k) = d(k) - \Phi(v(k)) \quad (3)$$

$$\mathbf{w}(k+1) = \mathbf{w}(k) - \eta \nabla_{\mathbf{w}} E(e(k)) \quad (4)$$

where $d(k)$ is some training (desired) signal, $e(k)$ is the instantaneous error at the output neuron, $E(\cdot)$ is the filter cost function, η denotes the learning rate parameter, $\mathbf{w}(k) = [w_1(k), ..., w_N(k)]^T$ is the weight vector, and $\Phi(\cdot)$ represents a nonlinear activation function of a neuron. Definition of the vector $\mathbf{u}(k)$ depends on the structure of a neural adaptive filter. In the case of the feedforward filter $\mathbf{u}(k)$ contains only samples of the input signal $x(k)$, and it is defined as $\mathbf{u}(k) = [x(k-1), ..., x(k-N)]^T$. The most common choice for the cost function $E(\cdot)$ is

$$E(e(k)) = \frac{1}{2} e^2(k). \quad (5)$$

Computation of the gradient of the cost function, denoted by $\nabla_{\mathbf{w}} E(e(k))$, depends on the structure of

a neural adaptive filter. For the feedforward type of a filter, this gradient is given by

$$\nabla_{\mathbf{w}} E(e(k)) = e(k)[x(k-1), ..., x(k-N)]^T. \quad (6)$$

The algorithm described by equations (3) – (6) is usually referred to as the nonlinear gradient-descent (NGD) algorithm. The gradient of the cost function for a nonlinear ARMA(p, q) recurrent perceptron is defined as

$$\nabla_{\mathbf{w}} E(e(k)) = e(k)\mathbf{\Pi}(k) \quad (7)$$

where $\mathbf{\Pi}(k) = [\frac{\partial y(k)}{\partial w_1(k)}, \ldots, \frac{\partial y(k)}{\partial w_N(k)}]$ represents the gradient at the output of the neuron. The normalized nonlinear gradient-descent (NNGD) algorithm exhibits optimal behaviour in the sense that it minimizes instantaneous prediction error, thus providing an adaptive learning rate η [6]. In the case of the linear activation function of an output neuron, the NNGD algorithm reduces to the normalized least mean squares (NLMS) algorithm.

8 Experimental Results

Air pollution data represent hourly measurements of the concentration of nitrogen dioxide (NO_2), in the period 1994 – 1997, provided by the Leeds meteo station. In the performed experiments the logistic function was chosen as the nonlinear activation function of an output neuron. The logistic function performs contraction mapping for the slope β set to be $0 \leq \beta \leq 4$ [7]. The quantitative performance measure was the standard prediction gain, a logarithmic ratio between the expected signal and error variances $R_p = 10\log(\hat{\sigma}_s^2/\hat{\sigma}_e^2)$. The slope of the nonlinear activation function of the neuron β was set to be $\beta = 4$, since this value makes Φ close to the linear function in the vicinity of the origin. The learning rate parameter η in the NGD algorithm, was set to be $\eta = 0.3$, and the constant C in the NNGD algorithm, was set to be $C = 0.1$. The order of the feedforward filter N was set to be $N = 10$ [8]. The order of the MA part q and the AR part p, of the nonlinear ARMA recurrent perceptron, were set to be $q = 3$ and $p = 1$. Due to saturation type of logistic nonlinearity, input data was prescaled to fit the range of an output neuron activation function. The summary of the performed experiments is given in Table 1.

It is obvious that nonlinear algorithms for adaptation of a neural adaptive filter have better performance comparing to the best linear adaptive algorithm (NLMS).

	NGD	NNGD	Rec.Perc.	NLMS
Pred. gain [dB]	5.78	5.81	6.04	4.75

Table 1. Performance of the algorithms employed in the prediction of the NO_2 time series

9 Conclusions

An insight into the dynamical properties of an air pollutant dataset has been provided. Nonlinear adaptive algorithms have been compared with the linear algorithms on the air pollution series and have provided better results. The time series is nonlinear and cyclic and therefore the recurrent perceptron has exhibited the best performance, corroborating the results given by the measures of nonlinearity.

References

[1] G. E. P. Box and G. M. Jenkins, *Time Series Analysis: Forecasting and Control.* San Fransisco: Holden-Day, Inc., 1970.

[2] J. Theiler, P. S. Linsay, and D. M. Rubin, *Time Series Prediction : Forecasting the Future and Understanding the Past*, ch. Detecting Nonlinearity in Data with Long Coherence Times, pp. 429–455. Addison Wesley, 1993.

[3] D. Beule, H. Herzel, E. Uhlmann, J. Kruger, and F. Becker, "Detecting nonlinearities in time series of machining processes," *Proceedings of the American Control Conference*, pp. 694–698, 1999.

[4] M. C. Casdagli and A. S. Weigend, *Time Series Prediction : Forecasting the Future and Understanding the Past*, ch. Exploring the Continuum Between Deterministic and Stochastic Modeling, pp. 347–366. Addison Wesley, 1993.

[5] A. A. M. Khalaf and K. Nakayama, "A hybrid nonlinear predictor: Analysis of learning process and predictability for noisy time series," *IEICE Trans. Fundamentals*, vol. E82-A, no. 8, pp. 1420–1427, 1999.

[6] D. P. Mandic, "NNGD algorithm for neural adaptive filters," *Electronic Letters*, vol. 36, no. 9, pp. 845–846, 2000.

[7] D. P. Mandic and J. Chambers, "Relationship between the slope of the activation function and the learning rate for the rnn," *Neural Computation*, vol. 11, no. 5, pp. 1069–1077, 2000.

[8] I. R. Krcmar, D. P. Mandic, and R. J. Foxall, "On predictability of atmospheric pollution time series," in *To appear in the Proceedings of the 5th ICANNGA*, 2001.