Nonlinearity and prediction of air pollution

Milan Paluš, Emil Pelikán, Kryštof Eben, Pavel Krejčíř and Pavel Juruš
Institute of Computer Science, Academy of Sciences of the Czech Republic,
Pod vodárenskou věží 2, 182 07 Prague 8, Czech Republic

A presence of nonlinearity in time series of concentrations of air pollutants and in their relations to time series of meteorological variables is tested using information-theoretic functionals and the surrogate data approach. The results are discussed in relation to predictability of the pollutant concentrations aimed to alert smog episodes.

In: Artificial Neural Nets and Genetic Algorithms. Proceedings of the International conference. (Ed.: Kurkova V., Steele N.C., Neruda R., Karny M.) - Wien, Springer 2001, pp. 473-476 (ISBN: 3-211-83651-9). This study was supported within the European Union Fifth Framework Programme project APPETISE (IST–99–11764).

I. INTRODUCTION

The most used types of air quality models are either deterministic models or models given by simple regression-based statistics. Their success, however, is limited either by their failure to capture the nonlinear behaviour of air pollutants, or our incomplete understanding of the physical and chemical processes involved. The APPETISE project (Air Pollution ePisodes: modElling Tools for Improved Smog managEment, see `http://www.uea.ac.uk/env/appetise/`) aims to develop and test the suitability of novel nonlinear statistical methods to improve our ability to accurately forecast variations in air quality. The work is being carried out over a period of 2 years by a consortium from 9 institutions from 5 European countries and is founded under the European Union Fifth Framework Programme. The project will work towards the construction of a prototype air quality prediction and warning system and is concentrated on 4 key pollutants: nitrogen oxides, particulates, sulphur dioxide and ground level ozone. The latter is the main research topic for the group of investigators represented by the authors of this paper.

Before trying to enhance the existing linear models by nonlinear ones it is suitable to test a presence of nonlinearity in the dynamics of time series of the ground level ozone (GLO) concentration as well as in relations of these data to time series of the most influential meteorological variables and to concentrations of other pollutants.

II. TESTING NONLINEARITY

In this section we briefly review a method for detection and characterization of nonlinear relations in multivariate as well as in univariate time series. The method employs the technique of uni- and multivariate surrogate data and information-theoretic functionals called redundancies. The test for nonlinearity based on the redundancy – linear redundancy approach, combined with the surrogate data is described in detail in Ref. [3], its multivariate version in Ref. [4]. The surrogate data have been introduced in Ref. [6], and their multivariate version in Ref. [5]. More details about the information-theoretic functionals can be found in Ref. [1].

Consider n discrete random variables $X_1, \ldots, X_n$ with sets of values $\Xi_1, \ldots, \Xi_n$, and probability distribution functions (PDF) $p(x_1), \ldots, p(x_n)$, respectively, and the joint PDF $p(x_1, \ldots, x_n)$. The redundancy $R(X_1; \ldots; X_n)$, in the case of two variables also known as mutual information (MI) $I(X_1; X_2)$, quantifies average amount of common information, contained in the n variables $X_1, \ldots, X_n$:

$$R(X_1; \ldots; X_n) = \sum_{x_1 \in \Xi_1} \cdots \sum_{x_n \in \Xi_n} p(x_1, \ldots, x_n) \log \frac{p(x_1, \ldots, x_n)}{p(x_1) \ldots p(x_n)}. \quad (1)$$

Now, let the n variables $X_1, \ldots, X_n$ have zero means, unit variances and correlation matrix $\mathbf{C}$. Then, we define the *linear redundancy* $L(X_1; \ldots; X_n)$ of $X_1, X_2, \ldots, X_n$ as

$$L(X_1; \ldots; X_n) = -\frac{1}{2} \sum_{i=1}^{n} \log(\sigma_i), \quad (2)$$

where σ_i are the eigenvalues of the $n \times n$ correlation matrix $\mathbf{C}$.

If $X_1, \ldots, X_n$ have an n-dimensional Gaussian distribution, then $L(X_1; \ldots; X_n)$ and $R(X_1; \ldots; X_n)$ are theoretically equivalent (see [3] and references therein). The general redundancies R detect all dependences in data under study, while the linear redundancies L are sensitive only to linear structures [3].

The basic idea in the surrogate-data based nonlinearity test is to compute a *nonlinear* statistic from data under study and from an ensemble of realizations of a linear stochastic process, which mimics "linear properties" of the studied data. If the computed statistic for the original data is significantly different from the values obtained for the surrogate set, one can infer that the data were not generated by a linear process; otherwise the null hypothesis, that a linear model fully explains the data, is accepted. For the purpose of such test the surrogate data

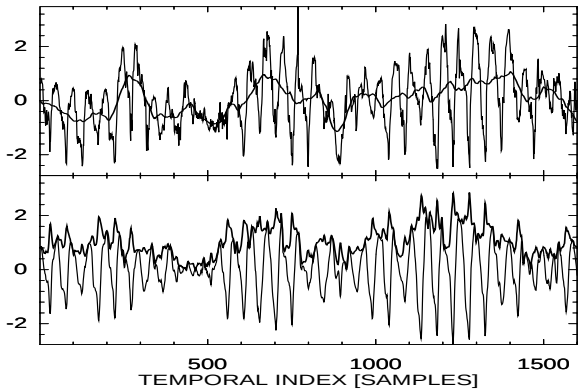

FIG. 1. Top panel: A segment of the ground level ozone concentration time series – the raw data (thin line) and the MA trend (thick line). Bottom panel: The MA filtered diurnal oscillations of the above GLO data (thin line) and their instantaneous amplitude (thick line).

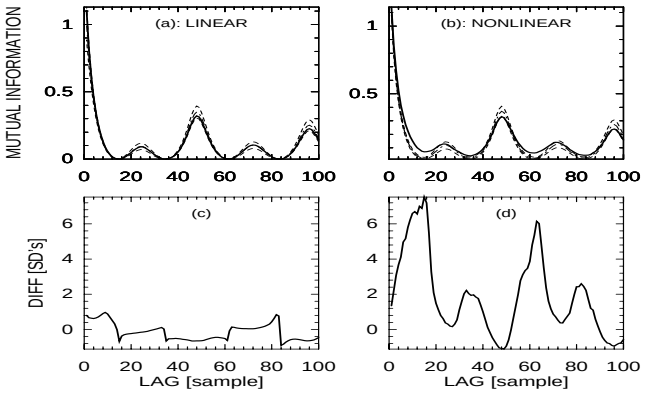

FIG. 2. Linear (a) and general (nonlinear) (b) mutual information (bivariate redundancy) for the raw ground level ozone concentration data (full line) and its surrogate data – mean (dotted line) and mean±SD (dashed lines) of the 30 surrogate realizations. Differences ("significances") obtained from linear (c) and nonlinear (d) redundancy.

must preserve the spectrum and consequently, the autocorrelation function of the series under study [6]. In the multivariate case also cross-correlations between all pairs of variables must be preserved [5].

Like in [3] we define the test statistic as the difference between the redundancy obtained for the original data and the mean redundancy of a set of surrogates, in the number of standard deviations (SD's) of the latter. The result is considered significant if the difference is clearly larger than 2 SD. In this study only 2-variable mutual information $I(X;Y)$ was applied: the univariate version $I(X(t);X(t+\tau))$ when dynamical properties and nonlinearity of individual series (variables) were studied, and the bivariate version $I(X(t);Y(t+\tau))$ when dynamical relations between two variables were investigated. The mutual information $I(X;Y)$[o] from the scrutinized data and the mean mutual information $I(X;Y)$[s] from the surrogates, as well as the test statistics, defined above, were plotted as functions of lag τ. Significant differences found between $I(X;Y)$[o] and $I(X;Y)$[s] were used to infer nonlinearity in dynamics of a variable (in univariate case), or in a relation between two variables (in bivariate case). The values of $I(X;Y)$[o] indicate a "coherence" or predictability of a variable, i.e., the dependence between $x(t)$ and $x(t+\tau)$ (in univariate case), or a strength of the link between two variables (in bivariate case), both as a function of the lag τ.

III. DATA

Time series of GLO, NO_2 and NO_x concentrations, as well as air temperature, wind speed and relative humidity were selected from a database of 37 Czech stations and several stations from UK, Germany, Finland and Italy. The sampling time is 30 min. in the data from the Czech stations and 1 hour otherwise. The lengths of processed data segments were 2048 and 4096 samples, according to data availability. Since all the data are dominated by the diurnal cycle, in addition to raw data also the following two slow components were extracted from the data and processed: a) *trends,* (Fig. 1, top panel) obtained from the raw data by simple moving average (MA, window length equal to 49 samples), b) *instantaneous amplitude* (Fig. 1, bottom panel) of the diurnal cycle. The latter has been obtained from the MA filtered data by using the analytic signal concept of Gabor [2]. For any signal $s(t)$, the analytic signal $\psi(t)$ is a complex function of time defined as

$$\psi(t) = s(t) + j\hat{s}(t) = A(t)e^{j\phi(t)}, \tag{3}$$

where the function $\hat{s}(t)$ is the Hilbert transform of $s(t)$

$$\hat{s}(t) = \frac{1}{\pi}\ \mathrm{P.V.}\int_{-\infty}^{\infty}\frac{s(\tau)}{t-\tau}d\tau. \tag{4}$$

(P.V. means that the integral is taken in the sense of the Cauchy principal value.) The instantaneous amplitude is then

$$A(t) = \sqrt{s(t)^2 + \hat{s}(t)^2}. \tag{5}$$

IV. RESULTS

The results from the above described nonlinearity test obtained from the raw GLO concentration data (the Czech Station Tušimice, 4096 half-hour samples from the 1997 season) are presented in Fig. 2. Results from the linear MI (Fig. 2a,c) show no significant differences between the data and the surrogates, i.e., the surrogates correctly reflect the linear properties of the studied data

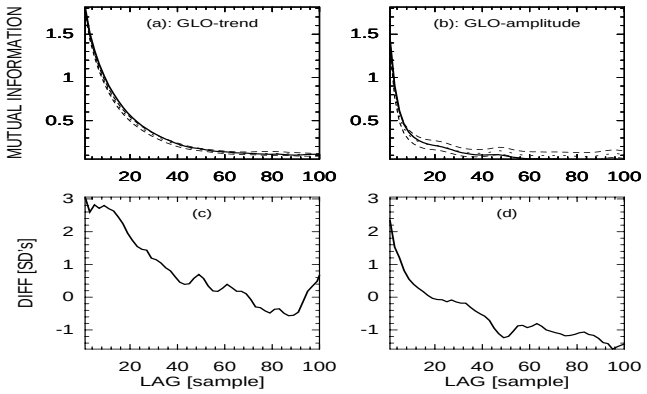

FIG. 3. The (nonlinear) mutual information (a,b) and the related difference statistics (c,d) for the trend (a,c) and the instantaneous amplitude (b,d) obtained from the ground level ozone concentration data (full line) and its surrogate data (in a,b: mean (dotted line) and mean±SD (dashed lines) of the 30 surrogate realizations).

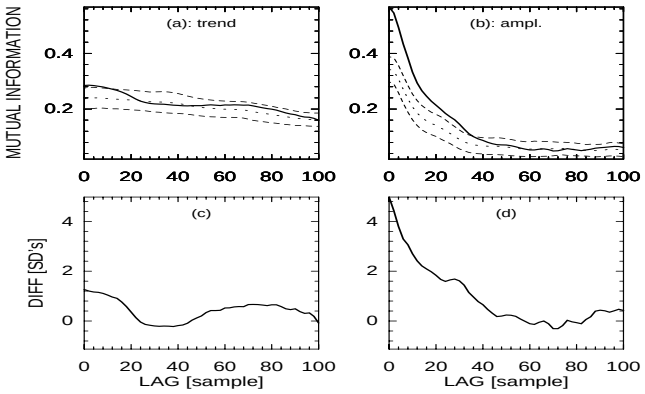

FIG. 4. The (nonlinear) mutual information (a,b) and the related difference statistics (c,d) between the trends (a,c) and the instantaneous amplitudes (b,d) of the air temperature and the ground level ozone concentration data (full line) and their bivariate surrogate data (in a,b: mean (dotted line) and mean±SD (dashed lines) of the 30 surrogate realizations).

and the significant differences, detected by the nonlinear MI (Fig. 2b,d) should not be due to flawed surrogates. On the other hand, the formal rejection of the null hypothesis of a linear stochastic process is not a conclusive evidence for nonlinear character of the process underlying the GLO concentration time series. It is clear that the diurnal cycle is externally driven and could bring a formal statistical long-term dependence on linear or nonlinear level which, however, does not represent any causal connection between $x(t)$ and $x(t+\tau)$. For the purpose of predicting the GLO concentration we should study the series of the trends and amplitudes, obtained from the raw data as described above. In the following tests the surrogate data were constructed from the raw data, and their trends and amplitudes were obtained from the surrogates in the same way as from the raw data. The results for the trend and amplitude of the above GLO concentration data (the Czech Station Tušimice, 4096 half-hour samples from the 1997 season) are presented in Fig. 3. Without the diurnal cycle the serial dependence (and predictability) of this series falls quickly, esp. in the case of amplitudes where it lasts less then 20 hours (Fig. 3b), while the serial dependence of the trend spreads to lags of approx. 30 hours (Fig. 3a). This dependence is predominantly linear, a nonlinear dependence can be detected for short lags up to 10 hours for the trend (Fig. 3c), while for the amplitude it is practically negligible (Fig. 3d).

In the following we analyse pairs of simultaneously recorded time series using the bivariate surrogate data. The relation between the air temperature and the GLO concentration (in terms of trends and amplitudes) is analyzed in Fig. 4 using 4096 half-hour sample data from the Czech station Teplice, 1997 season. There is a slowly decreasing, long term, entirely linear dependence between the temperature and GLO trends (Fig. 4a,c), while the dependence between the temperature and GLO amplitudes is decreasing quickly, however, for the short lags (up to 10 hours) it is stronger than the dependence of the trends (Fig. 4b) and is also nonlinear (Fig. 4d). The linear dependence between the trends of the relative air humidity and the GLO concentration has a similar long-term slowly decreasing character as the air temperature - GLO trends dependence, however, unlike the latter case there is a strong nonlinear relation between the humidity and GLO trends, though confined to short time lags (up to approx. 20 hours, however, with a high significance again only to 10 hours, see Fig. 5a,c). The relation between the amplitudes is again linear, long-term slowly decreasing one (Fig. 5b,d). The above results were obtained from 4096 half-hour samples, recorded at the Czech Station Tušimice, in the 1997 season. Comparable results have been obtained from other Czech stations as well as from some UK and German stations.

In further analyses, the short-lag nonlinear dependence has also been found in the relations of the trends of the wind velocity and the GLO concentration in the data from several Czech stations, however, has not been confirmed in the German and UK data. Only a weak and short linear dependence has been found in the relations of the amplitudes of wind velocity and GLO data.

The analyses of GLO relations to other pollutants are in their introductory state, so only as a preliminary result we can state the short-lag nonlinearity in the GLO – NO_2 amplitudes relation, while the GLO – NO_2 trends and both the trends and amplitudes of the GLO – NO relations appear to be limited to a weak linear dependence.

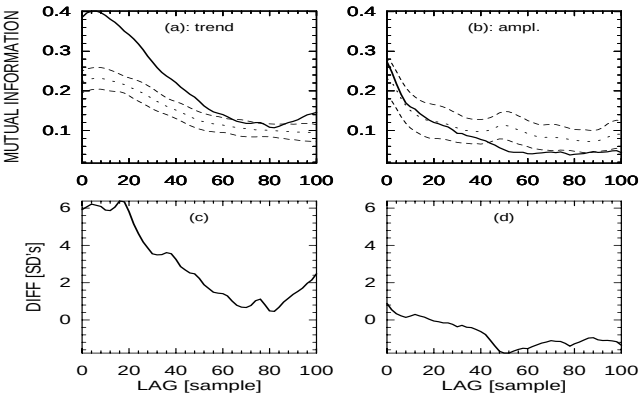

FIG. 5. The (nonlinear) mutual information (a,b) and the related difference statistics (c,d) between the trends (a,c) and the instantaneous amplitudes (b,d) of the air relative humidity and the ground level ozone concentration data (full line) and their bivariate surrogate data.

V. DISCUSSION

The presence of nonlinearity in the dynamics of time series of the ground level ozone (GLO) concentration as well as in relation of this data to time series of the most influential meteorological variables and to concentrations of other pollutants has been investigated by the test for nonlinearity employing the mutual information and the surrogate data method. The analysis of the raw data indicated long-term dependence and some formally proven nonlinearity caused by the diurnal cycle which dominates all the studied data sets. Since the externally driven diurnal cycle has practically no implications for predictability of the scrutinized time series, for further analyses the data had been preprocessed in order to obtain the long term trends and the instantaneous amplitude of the diurnal cycle. The GLO concentration has been found related to the influential meteorological variables (air temperature, relative humidity and wind velocity) by the slowly decreasing long-term linear dependence in some cases enhanced by a short-lag (up to 10 hours) nonlinearity. Thus nonlinear time series models, such as neural networks, can improve only the short term (several hours) GLO concentration forecasts. For predictions with day or longer horizons the statistical models should be combined with deterministic models and forecasted meteorological variables.

[1] T.M. Cover and J.A. Thomas, *Elements of Information Theory* (J. Wiley & Sons, New York, 1991).

[2] Gabor D., "Theory of Communication," *J. IEE London* vol. **93** pp. 429–457, 1946.

[3] M. Paluš, "Testing for nonlinearity using redundancies: Quantitative and qualitative aspects," *Physica D* vol. **80** pp. 186–205, 1995.

[4] M. Paluš, "Detecting nonlinearity in multivariate time series," *Phys. Lett. A* vol. **213** pp. 138–147, 1996.

[5] D. Prichard and J. Theiler, "Generating surrogate data for time series with several simultaneously measured variables," *Phys. Rev. Lett.* vol. **73** pp. 951–954, 1994.

[6] J. Theiler, S. Eubank, A. Longtin, B. Galdrikian and J.D. Farmer, "Testing for nonlinearity in time series: the method of surrogate data," *Physica D* vol. **58** pp. 77–94, 1992.