Searching for low dimensionality in air pollution time series

Europhys. Lett., **40** (6), pp. 589-594 (1997)

Searching for low dimensionality in air pollution time series

M. Lanfredi and M. Macchiato

Istituto Nazionale per la Fisica della Materia, Dipartimento Scienze Fisiche
Università di Napoli "Federico II" - Mostra d'Oltremare, Pad. 20, 80125 Napoli, Italy

(received 6 January 1997; accepted in final form 6 November 1997)

PACS. 02.50−r – Probability theory, stochastic processes, and statistics.
PACS. 05.45+b – Theory and models of chaotic systems.
PACS. 92.60Sz – Air quality and air pollution.

Abstract. – Time series of atmospheric pollutants (NO_x, CO and O_3) have been analyzed and the possible presence of low-dimensional chaos has been checked. The analysis has been performed by exploiting the concept of short-term predictability of chaotic systems. Although the time series exhibit statistical characteristics which mimic those typically explained by complex systems, no sign of chaos has been evidenced.

Interest in studying and understanding deterministic, mathematical systems which appear to have random, unpredictable behavior (*e.g.*, Ruelle [1]) cuts across a large number of science fields. At present these systems, commonly classified as chaotic, could provide a proper dynamical description for many seemingly complex phenomena. In particular, chaos may have a dramatic impact on applied sciences. Indeed, models for chaotic systems may suggest a parsimonious representation for processes governed by few dynamical degrees of freedom. In addition, the chaotic nature of a system puts limits to its predictability from past history, even in the absence of any random component.

Evidence of chaotic behaviors has been provided by the analysis of experimental time series from well-controlled laboratory experiments (*e.g.*, see the meeting report [2]). On the other hand, for observational time series generated by a variety of natural systems which cannot often be controlled at all, the problem is still debated. In concerning atmospheric phenomena, there are not yet unquestionable results which show the presence of low-dimensional chaos in observational time series. Positive results obtained by some authors seem to be due to inappropriate use of analysis tools and the presence of low-dimensional attractors is likely to be an artificial result of the finite lengths of the time series examined (*e.g.*, Ruelle [3]).

This paper discusses the statistical analysis of air pollution time series based on estimating their short-term predictability. The time series were recorded in Bristol and New Castle (Pennsylvania, USA) in the decade 1983-1992 by the Environmental Protection Agency (EPA) Aerometric Information Retrieval System (AIRS). Data consist of ten years of one-hour averaged measures of the two following air pollution marker in urban areas: NO_x, CO. In addition to such two primary pollutants, the secondary pollutant, O_3, will be considered, too. The

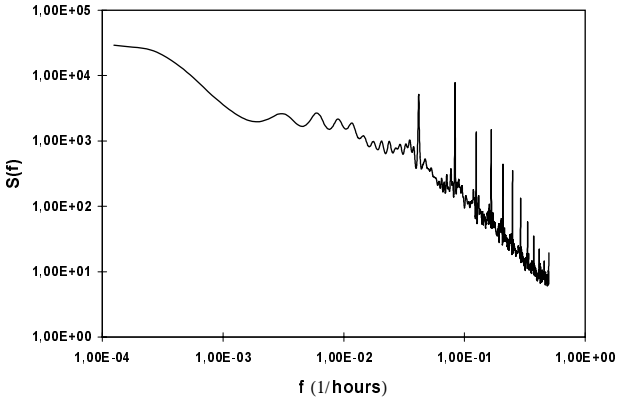

Fig. 1. – Variance spectrum of the series of CO concentration (Bristol) in log-log scale.

primary source of this kind of pollution is traffic. However, once originated, the pollution dynamics is regulated by many complicated physical, chemical and anthropogenic factors. The same troposphere is a very complex environment, especially in its boundary layer which is strongly affected by interactions with the Earth surface, so that the time evolution of air pollution is likely to exhibit non-trivial statistical structures. Figure 1 shows an example of variance spectrum estimated from our time series. As is possible to note, peaks at the daily frequency and subtones are overimposed on a continuous background with a predominance of low frequency. This behaviour is characteristic of processes which show persistence and is, very often, due to a relatively high dynamical correlation. Such a spectrum may agree with a complex dynamics, however, as a result of our analysis, no evident sign of low-dimensional chaos has been found.

The analysis method we use to discriminate low-dimensional chaos from randomness was originally developed by Serio and co-workers [4]-[6]. In this approach, both linear (global) and non-linear (local) representation are used to forecast time series and short-term predictive skills are estimated and compared. If the series are generated by a low-dimensional chaotic system, predictive advantage of the local representation over the global one and therefore over the entire class of linear stochastic systems including regular attractors and colored noises, is expected. Conversely, if the data are generated by a high-dimensional dynamics the global representation will give better previsions. Indeed, if only a few degrees of freedom interact non-linearly to generate a deterministic chaotic signal, then a *local* (nonlinear) predictor can be constructed which approximates the dynamics of the low-dimensional signal better than any *global* (linear) predictor. Technically, predictors are built by locally and globally fitting to the data autoregressive processes. The *local approach* is equivalent to represent the data on the basis of a low-dimensional deterministic nonlinear system whereas the *global approach* is equivalent to represent the observations by means of a linear (infinite-dimensional) stochastic model. Comparison between both linear and nonlinear fit strategies is needed because chaotic and random systems may exhibit similar behaviors such as similar signatures in correlation and so the only non linear short-term prediction [7]-[11] may not be conclusive in identifying chaos [4].

Unlike methods based, *e.g.*, on correlation dimension and related parameters or Lyapunov exponents, which provide statistically inefficient estimators, *i.e.* the estimates are affected from a very large variance, even when a long realization of the phenomenon is known, the technique we use has been proved to be robust to observational noise and its very nature is

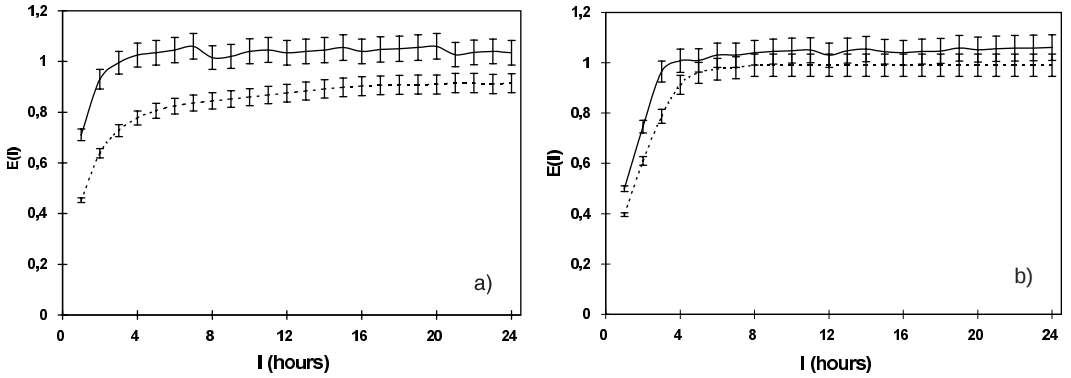

Fig. 2. – Comparison between local (solid line) and global (dashed line) forecast error function, $E(l)$, for the NO_x time series. a) New Castle ($p = 3$ for the local representation; $p = 335$ for the global one), b) Bristol ($p = 3$ for the local representation; $p = 335$ for the global one).

statistically consistent, that is the sampling variance tends to zero as the size of the series becomes large (see [6] and references therein). Furthermore, the technique does not need any specific assumption about the density distribution of the data which must not be necessarily Gaussian [6]. Finally, the procedure yields a predictive model directly from the series of observations and therefore it provides results of practical value.

In order to explain the mechanics and the related computational details of the technique, in the following we will discuss the main mathematical aspects of the procedure. For more detailed information we refer the reader to [4]-[6].

Let $\{x(n)\}, n = 1, \ldots, N$ a discrete time series which we suppose sampled at equal interval of time. According to linear prediction, the l-step ahead forecast ($l \geq 1$) $\hat{x}(m+l)$ of $x(m+l)$, standing at the origin m, is obtained as a linear function of $x(m)$ and previous observations. If L is the maximum step ahead at which forecasts are computed and p is the filter order, let us introduce the L-dimensional vector $\hat{\mathbf{x}}_m(x(m+1), \ldots, x(m+L))$ of the forecasts from the origin m and the p-dimensional basis vector $\mathbf{x}_m(x(m), \ldots, x(m-p+1))$ of the observations. In the *global autoregressive representation* the predictor is constructed as

$$\hat{\mathbf{x}}_m^{\mathrm{T}} = \Phi \mathbf{x}_m^{\mathrm{T}}\,, \tag{1}$$

where T indicates transposition and both the order of the filter p and the coefficients of the $L \times p$ matrix Φ have to be determined. For a given p the coefficients matrix can be computed by means the well-known Yule-Walker estimation procedure (*e.g.*, [12]). As a consequence of the linear structure of the stochastic process with which the series are modeled in the global representation (*e.g.*, [12]), Φ is invariant under shift of the time origin (*the coefficients do not depend on* m).

Conversely, in the *local autoregressive representation*, the forecasts vector from m is obtained as

$$\hat{\mathbf{x}}_m^{\mathrm{T}} = \Phi_m \mathbf{x}_m^{\mathrm{T}}, \tag{2}$$

where Φ_m is the autoregressive coefficients matrix which is not shift invariant (*the coefficients depend on the time origin* m). A strategy to compute the autoregression coefficients for the local case is discussed by Serio [4], [6].

Let us define the forecast error for the l-step ahead prediction from an origin m as $e_m(l) = \hat{x}(m+l) - x(m+l)$, and the normalized root mean-square forecasting error, $E(l)$, as the

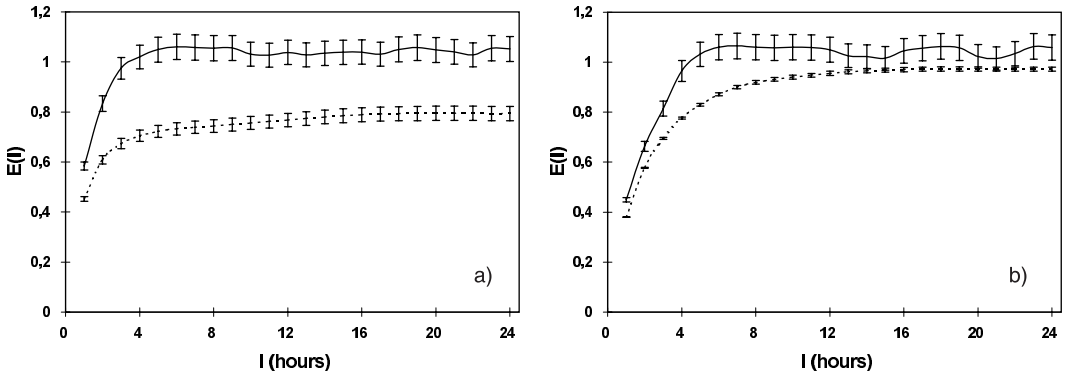

Fig. 3. – Comparison between local (solid line) and global (dashed line) forecast error function, $E(l)$, for the CO time series. a) New Castle ($p = 3$ for the local representation; $p = 336$ for the global one), b) Bristol ($p = 2$ for the local representation; $p = 336$ for the global one).

expectation value of the forecast error estimated by

$$E^2(l) = \frac{1}{\sigma_x^2} \sum_{m=1}^{M-l} \frac{e_m^2(l)}{M-l}\,, \tag{3}$$

where M is the number of test points for which we construct predictors and σ_x is the standard deviation of the series. $E(l)$ provides a quantitative measure of the predictive skill for both local and global autoregressive representation.

If $\{x(n)\}$ is a realization of a process generated by a chaotic system and $p > D$, where D is the dimension of the underlying attractor, then, subject to very loose assumptions [6] we have that

$$E_{\rm gl}(l) > E_{\rm lc}(l)\,, \tag{4}$$

where $E_{\rm gl}$ denotes the root mean-square forecasting error affecting the global prediction (eq. (1)) and $E_{\rm lc}(l)$ is the corresponding value for the local prediction (eq. (2)).

Finally we need a suitable criterion to estimate the optimal order, p, of the autoregressive filter for both local and global representation. To this end we divide the time series into two parts and for the last M of the total N points we construct predictors by globally and locally fitting autoregressive models to the previous $N - M$ data. The forecast error we do when we substitute each one of the M test points with the predicted values is obtained for different values of p, say for $p = 1, \ldots, p_{\rm u}$, where $p_{\rm u}$ is an upper bound specified from the user. If L is the maximum step ahead at which forecasts are computed (in practice obtained by the condition $E(l) \approx 1$), then we select as optimal order of the process that value of p for which the norm $K(p) = \sqrt{E_p^2(1) + \cdots + E_p^2(L)}$ is minimized.

The procedure is used either in the case of the global predictors or in the case of the local ones. In both cases we obtain *true* optimal mean-square predictions. This point is important since for arbitrary p the representation (1) is not necessarily *the best* in the least-square sense. This point is not explicitly recognized, *e.g.*, by Casdagli [9] which uses an approach similar to ours.

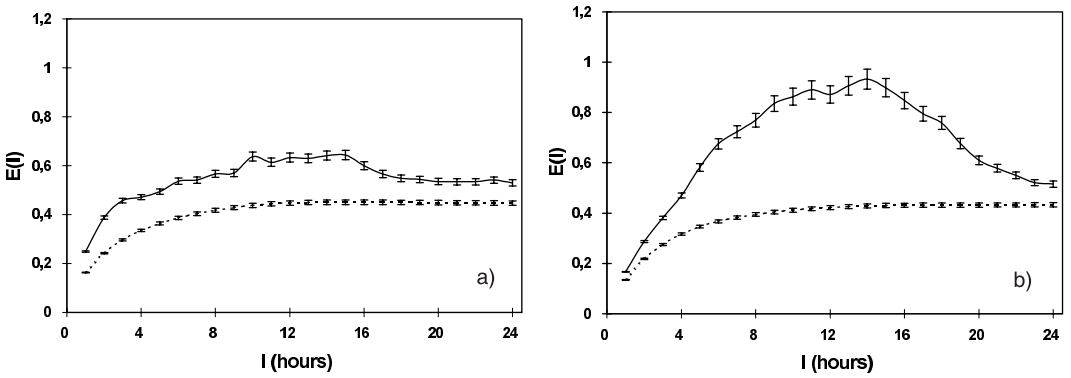

Fig. 4. – Comparison between local (solid line) and global (dashed line) forecast error function, $E(l)$, for the O_3 time series. a) New Castle ($p = 3$ for the local representation; $p = 383$ for the global one), b) Bristol ($p = 2$ for the local representation; $p = 387$ for the global one).

The error bar for $E(l)$ can be estimated as

$$\mathrm{var}[E^2(l)] \approx E^4(l) \left(\frac{2}{M-L} + \frac{\gamma_2}{M-L} \right) ,$$

where γ_2 is the *kurtosis* index of the population $\{e(l)\}$ ($\gamma_2 = 0$ for a Gaussian population).

The results concerning the primary pollutants (NO_x, CO) are summarized in fig. 2 and fig. 3 respectively. For both approaches the forecast error function $E(l)$ *vs.* the lead time l is shown. Note that the lead time, l, covers one diurnal cycle (*i.e.* $l = 1, \cdots, 24$) that is a very significant period in modulating anthropic uses and meteorological phenomena.

Both NO_x and CO series display a similar behavior. The local approach gives poorer predictions than the global one and in about 4 steps the variance of the predictions becomes comparable with the variance of the series. Overall the series appear to be largely unpredictable since $E(l)$ quickly reaches values close to 1 therefore indicating the presence of a very short-time correlation. However, note how the global approach is always superior over the local one, the difference being always statistically significant.

Note also that compared to Bristol, New Castle is characterized by slightly better predictions. Because of the relatively short time scale involved, these differences do not have a meteorological origin, but rather they are likely to reflect the different traffic organization or intensity.

Figure 4 shows the results concerning ozone. Ozone is not a primary result of traffic pollution. It is, indeed, the secondary result of photochemical reaction which involves mainly NO_x,CO and other pollutant species. It is interesting to note the relatively good predictive skill as it appears to a global analysis. Such a behavior reveals the presence of a significant background correlation which goes beyond the twenty-four hours. This effect can be explained by the presence of a periodic or quasi-periodic component which is well described by both the local and global techniques. This behavior would have been hardly inferred by inspecting the statistical properties of the primary pollutants. The two primary pollutants from one side and the secondary pollutant from the other, indeed, are characterized by quite different time scales. In addition, note that the global approach is always superior over the local one, therefore suggesting an underlying high-dimensional dynamics.

To sum up, multi-years samples of primary and secondary pollutants in two urban areas have been analyzed. Their statistical properties have been assessed by exploiting the concept

of short-term predictability for complex phenomena. The analysis has revealed interesting correlation time scales. The primary pollutants are characterized by typical time scales which are less than 4 hours and appear to be largely unpredictable. As shown by the ozone analysis, secondary pollutants may have a dynamics completely different with typical time scales greater than 1 day. To rightly understand such a behavior, it should be considered that the production of ozone is mainly driven from the sun light which has, of course, a dynamic quite different from the traffic one. Thus, while the primary pollutants mainly reflect the traffic dynamics, the ozone behavior seems to reflect mainly the time scale of meteorological concern.

However, as far as the deep structure of the series is concerned, the random (high-dimensional) behavior predominates over the chaotic (low-dimensional) one.

***

We are grateful to Prof. C. SERIO for a critical reading of the manuscript and his helpful discussions. We would like to thank Drs. N. PIRRONE and J. MILLER which gave us data. This paper is inserted in the framework of a research line promoted and coordinated by one of us (Prof. M. MACCHIATO) and partially financed by INFM and CNR funds.

REFERENCES

[1] RUELLE D., *Chaotic Evolution and Strange Attractors* (Cambridge University Press, Cambridge) 1989.
[2] ABRAHAM N. B., GOLLUB J. P. and SWINNEY H. L., *Physica D*, **11** (1984) 252.
[3] RUELLE D., *Proc. R. Soc. London, Ser. A*, **427** (1990) 241.
[4] SERIO C., *Nuovo Cimento B*, **107** (1992) 681.
[5] SERIO C., *Europhys. Lett.*, **27** (1994) 103.
[6] DRAHOŠ J., PUNČOCHÁŘ M., NINO E. and SERIO C., *Nuovo Cimento B*, **110** (1995) 1415.
[7] FARMER J. D. and SIDOROWICH J. J., *Phys. Rev. Lett.*, **59** (1987) 845.
[8] CASDAGLI M., *Physica D*, **35** (1989) 335.
[9] CASDAGLI M., *J. R. Stat. Soc. B*, **54** (1991) 303.
[10] SUGIHARA G. and MAY R. M., *Nature*, **344** (1990) 734.
[11] KENNEL M. B. and ISABELLE S., *Phys. Rev. A*, **6** (1992) 3111.
[12] BOX G. E. P. and JENKINS G. M., *Time Series Analysis* (Holden Day, S. Francisco) 1976.