Available online at www.sciencedirect.com
SCIENCE DIRECT®
Atmospheric Environment 40 (2006) 1774–1780
ATMOSPHERIC ENVIRONMENT
www.elsevier.com/locate/atmosenv

Air quality forecasting using a hybrid autoregressive and nonlinear model

Asha B. Chelani*, S. Devotta

National Environmental Engineering Research Institute, Nagpur 440 020, India

Received 10 October 2005; received in revised form 27 October 2005; accepted 9 November 2005

Abstract

The usual practices of air quality time-series forecasting are based on applying the models that deal with either the linear or nonlinear patterns. As the linear or nonlinear behavior of the time series is not known in advance, one applies the number of models and finally selects the one, which provides the most accurate results. The air pollutant concentration time series contain patterns that are not purely linear or nonlinear and applying either technique may give inadequate results. This study aims to develop a hybrid methodology that can deal with both the linear and nonlinear structure of the time series. The hybrid model is developed using the combination of autoregressive integrated moving average model, which deals with linear patterns and nonlinear dynamical model. To demonstrate the utility of the proposed technique, nitrogen dioxide concentration observed at a site in Delhi during 1999 to 2003 was utilized. The individual linear and nonlinear models were also applied in order to examine the performance of the hybrid model. The performance is compared for one-step and multi-step ahead forecasts using the error statistics such as mean absolute percentage error and relative error. It is observed that hybrid model outperforms the individual linear and nonlinear models. The exploitation of unique features of linear and nonlinear models makes it a powerful technique to predict the air pollutant concentrations.

Keywords: Time-series forecasting; ARIMA; Nonlinear dynamics; Hybrid model

1. Introduction

In the air quality literature, time-series analysis is generally carried out to understand the cause and effect relationships, which in turn helps in forecasting the future concentrations. In this direction, a class of techniques including autoregressive integrated moving average (ARIMA) or Box–Jenkins models (Shi and Harrison, 1997; Milionis and Davies, 1994; Zennetti, 1990) and structural models (Schlink et al., 1997) have been applied to analyze air pollutant concentrations. These approaches are widely applied in the air-quality literature due to the lack of data on emissions of air pollutants. Although these models are quite flexible as they can represent several different types of time series, their major limitation is the pre-assumed linear form of the model. The approximation of linear models to real-world problems is not always satisfactory. For example, the air pollutant concentrations are influenced by several factors in the atmosphere and prediction using linear models may not always give reasonable results (Benarie, 1987). As an

*Corresponding author.
E-mail address: ashachelani@rediffmail.com (A.B. Chelani).

1352-2310/$ - see front matter
doi:10.1016/j.atmosenv.2005.11.019

alternative, nonlinear models have been proposed in the literature. Artificial neural networks are one of the potential examples of nonlinear models that are applied to model and predict air pollutant concentrations (Gardner and Dorling, 1998). These models are generally developed using the external inputs such as meteorology and emissions and output is the air pollutant concentration at a site (Gardner and Dorling, 1999; Chelani et al., 2002). The application of these models is, however, restricted to the some particular cases where the data on emissions and meteorological parameters are available.

Recently the time-series forecasting based on the nonlinear dynamical theory or chaos theory has been extensively studied and used in the forecasting of air pollutant concentrations (Raga and LeMoyne, 1996; Li et al., 1994; Chen et al., 1998; Kocak et al., 2000). The basic assumption involved in the application of these techniques is that the single air pollutant concentration time series contains the effect of all the influencing factors. The major advantage of these techniques over ARIMA is their ability to take into account the nonlinear dynamics involved in the time series. The information about the linearity or nonlinearity of the time series is however, not available in advance. So one applies the number of linear and nonlinear models and finally selects the one, which provides the most accurate results. Also, the real time series contains patterns that are not purely linear or nonlinear and applying either of the techniques may give inadequate results. Hence the models need to be developed that consider both linearity and nonlinearity involved in the time series. This helps in improving the forecasting ability of the model. Also, combining different models can increase the chance to capture different patterns in the data and improve forecasting performance (Clemen, 1989; Newbold and Granger, 1974).

In this study, a hybrid methodology is therefore proposed to tackle the problem of modeling the air pollutant time series with linear and nonlinear patterns. For this, the concepts from ARIMA model and nonlinear dynamical systems theory are utilized. The proposed technique is applied to the time series of nitrogen dioxide (NO_2) concentrations in ambient air measured at a site in Delhi during 1999–2003. In order to compare the forecasting efficiency of the proposed hybrid model, ARIMA and nonlinear models are also developed individually and the results of these models are then compared with the hybrid model.

2. Box–Jenkins ARIMA models

ARIMA linear models have dominated many areas of time series forecasting. As the application of these models is very common, it is described here briefly. In general, a nonseasonal time series, $x_{t,t=1\ldots n}$ (n being the number of observations) of air pollutant concentrations measured at an equal time intervals, can be modeled as a combination of past values and past errors as

$$x_t = a_1 x_{t-1} + a_2 x_{t-2} + \cdots + a_p x_{t-p} + e_t - b_1 e_{t-1} - b_2 e_{t-2} \cdots b_q e_{t-q}, \tag{1}$$

where a and b are the coefficients, p and q are the order of the autoregressive and moving average polynomials, respectively. The further details to estimate the parameters and order of the model are given in Box and Jenkins (1970).

3. Nonlinear dynamical modeling

The nonlinear dynamical modeling involves the reconstruction of phase space of the time series to describe the behavior of a nonlinear system. A phase space is an abstract construct whose coordinates are the components of the state (Cambel, 1993). In general, phase space is nothing but the collection of all possible variables underlying the system. The phase space portrait can be analyzed mathematically to demonstrate the presence of an attractor and its dimension. An attractor characterizes the long-term behavior of the system in the phase space (Martelli, 1999). If an attractor exists, then the minimum number of independent variables describing the system can be estimated by computing the dimension of the attractor.

The general nonlinear prediction method is to reconstruct the phase space from the set of data in a minimum embedding space and then predict the future using a local approximation function computed from the set of given data (Farmer and Sidorowich, 1987; Abarbanel et al., 1993; Takens, 1981). According to Takens' embedding theorem, the predictions can be obtained from the set of previous data points using the functional relationship

$$X_{n+T} = f(X_n), \tag{2}$$

where X_n is a vector of data points defined by

$$X_n = (x_n, x_{n-\tau}, x_{n-2\tau}, \ldots, x_{n-(m-1)\tau}), \tag{3}$$