17th European Symposium on Computer Aided Process Engineering – ESCAPE17
V. Plesu and P.S. Agachi (Editors)

Air quality prediction in Uberlândia, Brazil, using linear models and neural networks

Taisa S. Lira, Marcos A. S. Barrozo, Adilson J. Assis

School of Chemical Engineering, Federal University of Uberlândia, Av. João Naves de Ávila, 2121, 38408-100, Uberlândia- MG, Brazil, e-mail: ajassis@ufu.br

Abstract

Particulate air pollution is associated with a range of effects on human health, including effects on the respiratory and cardiovascular systems, asthma and mortality. Hence, the development of an efficient forecasting and early warning system for providing air quality information towards the citizen becomes an obvious and imperative need. The objective of this work was to investigate that forecasting capability using linear models (such as ARX, ARMAX, output-error and Box-Jenkins), and neural networks. They were used meteorological variables and 24-h PM_{10} concentration of the present day as input data. As output foreseen by the models, the 24-h PM_{10} concentration is obtained, with horizon of prediction of up to three days ahead. The results showed that fairly good estimates can be achieved by all of the models, but Box-Jenkins model showed best fit and predictability.

Keywords

Air quality; Linear models; Neural networks; Particulate matter; Public health.

1. Introduction

In recent years, air quality has emerged as a major factor contributing to the quality of living in urban areas, especially in densely populated and industrialized areas. Particulate air pollution is associated with a range of effects on human health, including effects on the respiratory and cardiovascular

systems, asthma and mortality [1,2]. Short-term forecasting of air quality is needed in order to take preventive and evasive action during episodes of airborne pollution. In this way, by influencing people's daily habits or by placing restrictions on traffic and industry, it should be possible to avoid excessive medication, reduce the need for hospital treatment and even prevent premature deaths [3,4].
The trend in recent years has been to use more statistical methods instead of traditional deterministic modelling to forecast air pollution. Neural network (NN) models have been used for the forecasting of a wide range of pollutants and their concentrations at various time scales, with very good results [5-8]. In their overview of applications of NN in the atmospheric sciences, Gardner and Dorling [8] concluded that neural networks generally give as good or better results than linear methods. Linear models are being used here as a novelty, since air pollution forecast can be seen as similar as system identification.
The advantages of these models are that they do not require very exhaustive information about air pollutants, reaction mechanisms, meteorological parameters or traffic flow and that they have the ability of allowing nonlinear relationships between very different predictor variables. These facts and the quality of the results they have provided are the reasons that make them more attractive to apply than other models.
The objective of this work was to investigate the forecasting capability of the following methods: linear models (such as ARX, ARMAX, output-error and Box-Jenkins), and neural networks. The models used meteorological variables and 24-h PM_{10} concentration of the present day as input data. As output foreseen by the model, the 24-h PM_{10} concentration is obtained, with horizon of prediction of up to three days ahead.

2. Data and methodology

2.1. Data

This study is based on PM_{10} concentration data collected by School of Chemical Engineering of the Federal University of Uberlândia (UFU) during the years of 2003, 2004 and 2005. The samples were collected with Hi-Vol samplers in periods of 24 hours, every three days, in agreement with norms established by ABNT (Brazilian Association of Technical Norms). The equipment is located in the central bus station of Uberlândia city. More details concerning data collection methodology can be obtained elsewhere [9].
It is known that the concentration of pollutant atmospheric is strongly related to the meteorological conditions. Studies of the influence of meteorological conditions in the concentration of air pollutant can be seen in Elminir and Hien *et al.* [10,11]. The meteorological data used in study were obtained in the

climatic station of the Institute of Geography of the UFU located 2.07Km far from the place of PM_{10} sampling.

2.2. Neural networks

NN are mathematical structures which make use of a complex combination of weights and functions to convert input variables into an output (prediction). NN are capable of learning from the patterns presented to them and from the errors they commit in the learning process, so that finally they should identify patterns never seen before (generalization).

In the current study, the multilayer perceptron (MLP) was adapted. It is the most commonly used type of feedforward neural network in the atmospheric sciences [8]. MLP is composed of at least three layers of neurons: the input layer, the hidden layer(s) and the output layer. The input layer plays no computational role but merely serves to pass the input vector to the network. Each unit in the hidden layer sums its input, processes it with a transfer function and distributes the result to the output layer.

Training a MLP is the procedure by which the values for the individual weights are determined. Different training algorithms could be applied to minimize the error function, but the most widely used is the backpropagation algorithm [6]. This algorithm is nothing else than the application of the gradient descent method, using as objective function the sum square error among the net output and the training data.

2.3. Linear models

A general input-output linear model for a single-output system with input *u* and output *y* can be written [12]:

$$A(q)y(t) = \frac{B(q)}{F(q)}u(t) + \frac{C(q)}{D(q)}e(t) \tag{1}$$

where *e(t)* is white-noise and with

$$\begin{aligned} A(q) &= 1 + a_1 q^{-1} + \ldots + a_{n_a} q^{-n_a} \\ B(q) &= b_1 q^{-1} + \ldots + b_{n_b} q^{-n_b} \\ C(q) &= 1 + c_1 q^{-1} + \ldots + c_{n_c} q^{-n_c} \\ D(q) &= 1 + d_1 q^{-1} + \ldots + d_{n_d} q^{-n_d} \\ F(q) &= 1 + f_1 q^{-1} + \ldots + f_{n_f} q^{-n_f} \end{aligned} \tag{2}$$

The general structure may have up to 32 different model sets, depending on which of the five polynomials *A*, *B*, *C*, *D*, and *F* are used. However, only four possibilities were used here, and they are summarized in Table 1.

Table 1. Some models as special cases of Eq.(1).

Name of model structure	ARX	ARMAX	OE (output-error)	BJ (Box-Jenkins)
Polynomials used in Eq.(1)	A,B	A,B,C	B,F	B,C,D,F

3. Results and discussion

First at all, a multiple regression analysis (significance level: $p<0.05$) was applied to reveal atmospheric parameters controlling the day-to-day variations of PM_{10}. Temperature (^oC), relative humidity (%), rainfall (mm), wind speed (m/s), wind direction (degrees, 0 for N) and sunshine (h), in addition to the day of the week, were shown to be the most important parameters. They could explain 64% of the variances of 24-h PM_{10} concentrations.

The data set (total = 341) already normalized (mean = 0 and standard deviation = 1) was divided as follows: 2/3 for training/estimation and 1/3 for validation. Note that the wind direction and day variables were dichotomised using the sine and cosine functions. This enabled the neural algorithms to work properly despite the discontinuities in the original cyclic signals [7].

The software Matlab was used. Linear models were adjusted using the System Identification toolbox. For the neural net model (MLP), the learning algorithm used was Levenberg-Marquardt back-propagation (Neural Network toolbox). The transfer functions selected for the layers were hyperbolic tangent for the hidden layer and linear for the output layer. The number of neurons in the hidden layer and input-delay for MLP, and the orders of the polynomials and delays for linear models were the optimum found by cross validation.

For the evaluation of the models performance, three statistical measures, that are most frequently used in literature, were selected, namely the root mean square error (RMSE), the coefficient of determination (R^2) and the index of agreement (d). Even though the R^2 has its known defects in certain situations [13], this measure was used in order to maintain compatibility with other studies. The performance of all models was compared on the basis of the predicted and the observed PM_{10} concentration. The results are summarised in Table 2.

Table 2. Performance indicators for the models

	ARX	ARMAX	OE	BJ	MLP
RMSE	0.5078	0.5044	0.6495	0.4039	0.5424
d	0.9320	0.9318	0.8795	0.9629	0.9140
R^2	0.7799	0.7842	0.6394	0.8120	0.7591

Further comparisons can be found in Fig. 1 where scatter plots are made for the performances of the two models: BJ and MLP. For a scatter plot, the perfect case of prediction versus observation should be shown of a zero intercept and a unit slope. According to the plots shown in Fig.1 and the performance indicators in Table 2, all the five models showed a good forecast capability to the measured PM_{10} concentration, but Box-Jenkins model clearly gave the best results.

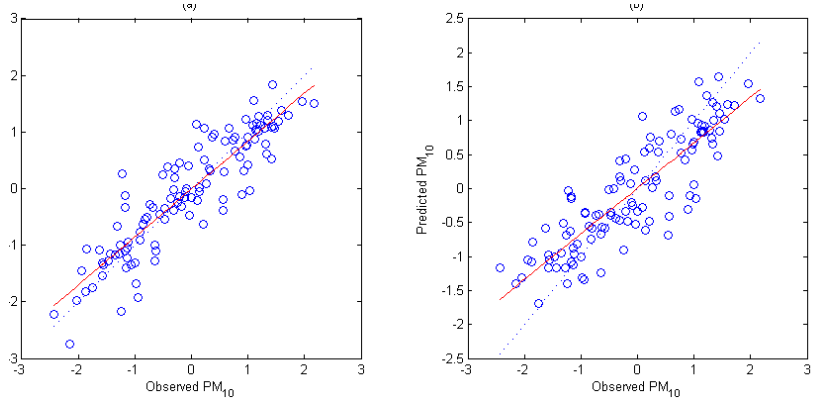

Fig.1 Scatter plots of observed and predicted PM_{10} concentrations: (a) BJ and (b) MLP.

In order to visualise the performance of the Box-Jenkins model on the estimation of the polynomial coefficients and MLP on the training, comparison between predicted versus observed data are given in Fig.2. The graphical presentation shows a quite good agreement between the predicted and the observed PM_{10} concentration, both for the estimation/training data and for the validation data.

4. Conclusions

Five models for air quality forecasting purposes were evaluated here using 24-h PM_{10} concentrations and basic meteorological variables from the city of Uberlândia (Brazil) collected during the years of 2003-2005. The results showed that fairly good estimates can be achieved by all of the models, but Box-Jenkins model presented best performance. The proposed models can be used, among other purposes, for the local public government, as a control tool

of urban traffic and also as a mechanism of formulation of preventive public politics in the areas of health and urban mobility.

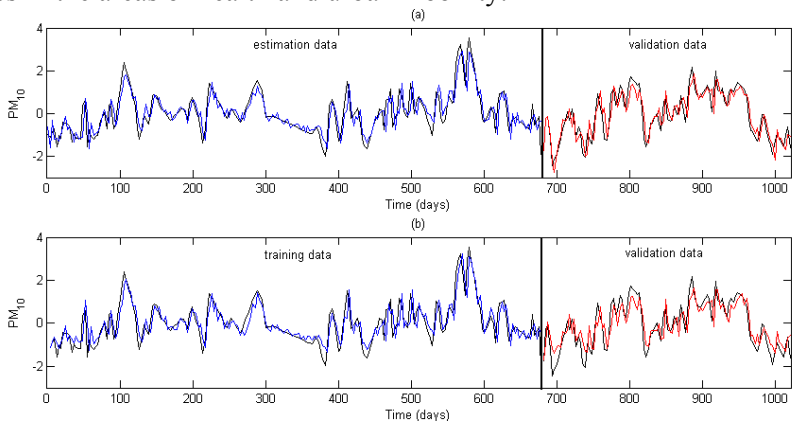

Fig.2 Predicted (blue line) versus observed (black line) data on the training/estimation and predicted (red line) versus observed (black line) data on the validation: (a) BJ and (b) MLP.

Acknowledgements

The authors wish to thank the Institute of Geography of the UFU for providing the meteorological data, and also the Euclides Antônio Pereira Lima for providing the PM_{10} concentration data.

References

1. J. Wordley , S. Walters, J .G. Ayres , Occupational Environment Medicine, 54 (1997) 108.
2. C. A. Pope, Journal of Aerosol Medicine, 55 (2000)1350.
3. J. Schwartz, Epidemiology, 7 (1997) 20.
4. P. Tiittanen, K. L. Timonen, J. Ruuskanen, A. Mirme, J. Pekkanen. European Respiratory Journal, 13 (1999) 266.
5. G. Grivas , A. Chaloulakou, Atmospheric Environment, 40 (2006) 1216.
6. E. Agirre-Basurkoa, G. Ibarra-Berastegib, I. Madariagac, Environmental Modelling & Software, 21(2006) 430.
7. M. Kolehmainen, H. Martikainen, J. Ruuskanen, Atmospheric Environment, 35 (2001) 815.
8. M.W. Gardner and S.R. Dorling, Atmospheric Environment, 32 (1998) 2627.
9. D. R. Cioqueta, B. C. Silvério, R. R. Reis, E. A. P. Lima, M. A. S. Barrozo, Proceedings of the second Mercosur Congress on Chemical Engineering and fourth Mercosur Congress on Process Systems Engineering , Rio de Janeiro, Brazil (2005).
10. H. K. Elminir, Science of the Total Environment, 350 (2005) 225.
11. D. Hien, T. Bac, C. Tham, D. Nhan, D. Vinh, Atmospheric Environment, 36 (2002) 3473.
12. L. Ljung, System identification: theory for the user (2nd ed.), Prentice-Hall PTR, Upper Saddle River, NJ (1999).
13. C. Willmott, Bulletin of the American Meteorological Society, 63 (1982) 1309.