Atmospheric Environment 98 (2014) 98–104

Contents lists available at ScienceDirect

Atmospheric Environment

journal homepage: www.elsevier.com/locate/atmosenv

Prediction of ozone concentration in tropospheric levels using artificial neural networks and support vector machine at Rio de Janeiro, Brazil

A.S. Luna [a, *], M.L.L. Paredes [a], G.C.G. de Oliveira [a], S.M. Corrêa [b]

[a] Rio de Janeiro State University, Institute of Chemistry, Rua São Francisco Xavier, 524, Maracanã, Rio de Janeiro 20550-013, Brazil
[b] Rio de Janeiro State University, Faculty of Technology, Rodovia Presidente Dutra Km 298, Pólo Industrial, Resende, Rio de Janeiro 27537-000, Brazil

HIGHLIGHTS

- The tropospheric ozone concentration was predicted using chemometric tools.
- The ANN and SVM were used in predicting the O_3 with R^2 up to 0.95.
- The predictive model is linked with the interaction of local-level meteorological.

ARTICLE INFO

Article history:
Received 8 April 2014
Accepted 25 August 2014
Available online

Keywords:
Air pollution
Artificial neural networks
Support vector machine
Ozone
Troposphere

ABSTRACT

It is well known that air quality is a complex function of emissions, meteorology and topography, and statistical tools provide a sound framework for relating these variables. The observed data were contents of nitrogen dioxide (NO_2), nitrogen monoxide (NO), nitrogen oxides (NO_x), carbon monoxide (CO), ozone (O_3), scalar wind speed (SWS), global solar radiation (GSR), temperature (TEM), moisture content in the air (HUM), collected by a mobile automatic monitoring station at Rio de Janeiro City in two places of the metropolitan area during 2011 and 2012. The aims of this study were: (1) to analyze the behavior of the variables, using the method of PCA for exploratory data analysis; (2) to propose forecasts of O_3 levels from primary pollutants and meteorological factors, using nonlinear regression methods like ANN and SVM, from primary pollutants and meteorological factors. The PCA technique showed that for first dataset, variables NO, NO_x and SWS have a greater impact on the concentration of O_3 and the other data set had the TEM and GSR as the most influential variables. The obtained results from the nonlinear regression techniques ANN and SVM were remarkably closely and acceptable to one dataset presenting coefficient of determination for validation respectively 0.9122 and 0.9152, and root mean square error of 7.66 and 7.85, respectively. For these datasets, the PCA, SVM and ANN had demonstrated their robustness as useful tools for evaluation, and forecast scenarios for air quality.

1. Introduction

Tropospheric ozone can have a negative impact on the environment and public health when present in the lower atmosphere, in sufficient quantities. Regulations have been introduced to set limits on the emissions of pollutants in such a way that they cannot exceed prescribed maximum values (EPA, 1999). Attention was given to mathematical and computer modeling of air quality to achieve these limits.

To trace and predict ozone, one must know the conditions that contribute to its formation. Besides, ozone concentrations are strongly connected to meteorological conditions. Land–sea breezes also affect ozone concentrations at coastal sites. Therefore, it is necessary to develop a model that encompasses the description and understanding relationships between ozone concentrations and the many variables that cause or inhibit ozone production to predict ozone concentrations (Abdul-Wahab and Al-Alawi, 2002).

Recently, a group of researchers (Wilson et al., 2012) analyzed trends in ozone levels in the European troposphere from 1996 to 2005. They indicated that average levels have been increasing despite reductions in pollutants which impact ozone formation. However, they also identified year-by-year variations, caused by

* Corresponding author.
E-mail addresses: asluna@uerj.br, adsluna@gmail.com (A.S. Luna).

http://dx.doi.org/10.1016/j.atmosenv.2014.08.060
1352-2310/

climate and weather events, and suggested they could be masking the impact of emission reductions on long-term ozone trends. This study confirmed that the relationship between ozone and its precursors is complicated. It is caused by the fact that meteorological and chemical processes can interact over a remarkably wide range of temporal and spatial scales (e.g., Adame et al., 2008). For this reason, the use of statistical tools provides a sound framework for the analysis of such data.

Multiple linear regression (MLR), partial least squares (PLS), as well as principal component regression (PCR) methods, were carried out to assess ambient air quality in Miskolc, in Hungary. Ozone concentration was modeled by MLR and PCR with the same efficiency if the conditions of meteorological parameters were not changed (i.e. morning and afternoon). Without night data, PCR and PLS suggest that the main process is not a photochemical but a chemical one (Lengyel et al., 2004). Abdul-Wahab et al. (2005) used multiple linear regression to predict the ozone concentration through chemical and meteorological factors, but as the presence of multicollinearity makes the models obtained improperly, the authors used principal component regression analysis (PCR), applying the stepwise regression option in the choice of the principal components to enter the regression equation, with the ozone as the dependent variable (Abdul-Wahab et al., 2005). To overcome the presence of multicollinearity, Álvarez et al. (2000) used one of the most powerful and regular techniques in regional studies on atmospheric pollution, called rotated principal component analysis (RPCA). The main advantage of RPCA over other techniques is that it offers the possibility of analyzing the spatial and temporal variability of pollutants simultaneously at regional level (Álvarez et al., 2000).

In addition to this problem, a study was carried out to compare the artificial neural networks (ANNs) with multiple linear regression models to predict the next day's maximum hourly ozone concentration in the Athens basin. Results based on a wide array of forecast quality measures indicate that the ANNs provides better estimates of ozone concentrations at the monitoring sites, whilst the more commonly used linear models are less effective for accurately forecasting high ozone concentrations (Chaloulakou et al., 2003).

On the other hand, the support vector machine (SVM) can be used for time-series prediction and has been reported to perform well by some promising results. One group of researchers developed an online SVM model to predict air pollutant levels in an advancing time-series based on the monitored air pollutant database in Hong Kong downtown area (Wang et al., 2008). In another application, the SVM was used to the prediction of hourly ozone values in Madrid urban area. Using the modified SVM-r, based on reductions of the SVM-r hyper-parameters search space, they explore different influences, which may alter the ozone forecast, such as former ozone measurements in a given station, measurements in neighbors stations, and the impact of meteorological variables. A comparison study with the results using ANN (multilayer perception) was also carried out. The prediction tool based on SVM-r was flexible enough to incorporate any other prediction variable, such as city models, or traffic patterns, which may enhance the prediction obtained with the SVM-r (Ortiz-García et al., 2010).

The objective of this study was to get parsimonious prediction models (i.e., models that depend on as few variables as needed) for ozone as a function of other ambient air concentration data and meteorological parameters as predictor variables, using the support vector machine (SVM-r) and artificial neural networks regression (ANN-r) models.

2. Chemometric techniques concepts

In this section, we briefly describe the primary characteristics of the chemometric techniques used in this work. The reader interested in deepening their knowledge in this area should seek especially publications such as described by Vandeginste et al. (2003).

2.1. Principal Component Analysis

The Principal Component Analysis (PCA) is an unsupervised method of exploratory analysis. The aim of PCA is dimension reduction, which may be used for visualization of multivariate data by scatter plots transformation of highly correlating *x*-variables into a smaller set of uncorrelated latent variables that can be used by other methods separation of relevant information (by a few latent variables) from noise combination of many variables that characterize a chemical-technological process into a single or a few "characteristic" variables. PCA can be seen like a method to compute a new orthogonal coordinate system formed by latent variables (components), where only the most informative dimensions are used. Latent variables from PCA optimally represent the distances between objects in the high-dimensional variable space — the distance of objects is a measure of the similarity of the objects. PCA is successful for data sets with correlating variables as is often the case with data from chemistry (Häggblom, K.-E.).

The PCA matrix **X** is decomposed into two matrices, loadings and scores as shown in Eq. (1):

$$\mathbf{X} = \mathbf{T}\mathbf{P}^{\mathbf{T}} + \mathbf{E} \tag{1}$$

where **X** is the samples matrix versus variables (m × n), **T** is the matrix of scores (m × number of PCs), $\mathbf{P}^{\mathbf{T}}$ is the transposed matrix of loadings (number of PCs × n), and **E** is the residual matrix. The score is represented by a matrix (m × number of PCs) consisting of samples, and the loading is represented by a matrix (number of PCs × n) comprising variables (Otto, 2007; Jolliffe, 2002).

The variance of the PC scores, preferably given in percent of the total variance of the original variables, is significant indicators of how many PCs to include. Because of the way the PCs are determined, the scores of each new PC have a lower variance than the scores of the previous PC. If the score variance of certain PC is close to the noise variance of the data, the PC does not contain any useful information. Obviously, the number of PCs should be less than this number. However, in practice it may be difficult to apply this criterion. A simple method is to plot (not shown) the cumulative variance of scores (the variance for each new PC added to the variance of previous PCs) against the PC number. As a rule of thumb, the PCs should explain at least 80%, maybe 90%, of the total variance.

The PCA result and the data can be analyzed and evaluated in many ways. In this work, a biplot was chosen because it combines a scores plot and loadings plot; it gives info about clustering of objects and variables relationships between objects and variables. Objects and variables tend to be positively correlated if close to each other negatively correlated if far away from each other (Häggblom, K.-E.).

2.2. Support vector of machine regression (SVM-R)

The support vector machine (SVM) is a machine learning technique developed by Vapnik, which increasingly is gaining ground in many areas of knowledge. Originally this technique was developed for pattern recognition problems. The model consists of a number of support vectors (essentially samples selected from the