Available online at www.sciencedirect.com
ScienceDirect
Environmental Modelling & Software 23 (2008) 1056–1069
Environmental Modelling & Software
www.elsevier.com/locate/envsoft

Development and comparative analysis of tropospheric ozone prediction models using linear and artificial intelligence-based models in Mexicali, Baja California (Mexico) and Calexico, California (US)

E. Salazar-Ruiz [a], J.B. Ordieres [b,*], E.P. Vergara [b], S.F. Capuz-Rizo [c]

[a] Instituto Tecnológico de Mexicali, Av. Tecnológico, s/n Col. Elías Calles, 21396 Mexicali, B.C., Mexico
[b] Universidad de La Rioja, Edificio Departamental, c/Luis de Ulloa 20, E-26004 Logroño, La Rioja, Spain
[c] Universidad Politécnica de Valencia, Camino de Vera, s/n E-46022 Valencia, Spain

Received 13 August 2007; received in revised form 25 November 2007; accepted 28 November 2007
Available online 21 February 2008

Abstract

This study developed 12 prediction models using two types of data matrix (daily means and a selection of the mean for the first 6 h of the day). The Persistence parametric prediction technique was applied separately to these matrices, as well as semiparametric Ridge Regression and three non-parametric or artificial intelligence techniques: Support Vector Machine, Multilayer Perceptron and ELMAN networks. The target was the prediction of maximum tropospheric ozone concentrations for the next day in the Mexicali–Calexico border area. The main ozone precursors and meteorological parameters were used for the different models. The proposals were evaluated using specific performance measurements for the air quality models established in the Model Validation Kit and recommended by the US Environmental Protection Agency.

Results with similar margins of error were obtained in various models developed in this study, and some of them have provided smaller margins of error than similar prediction models existing in the literature developed in other regions. For this reason, we consider it feasible to apply the prediction models developed and they could be useful for supporting decisions in the matter of ozone pollution in the region under study, as well as for use in daily forecasting in this area.

Keywords: US–Mexico border; Ozone neural network modeling; Multilayer Perceptron (MLP); ELMAN neural network; Support Vector Machine; Ridge Regression; Model Validation Kit (MvK); Transboundary air quality

1. Introduction

One of the main problems of atmospheric pollution in urban areas is contamination caused by photochemical oxidants such as ozone (O_3) and nitrogen dioxide (NO_2) (Lee et al., 1996; Kongtip et al., 2006). Ozone is considered to be one of the main greenhouse gases and a component of photochemical smog with potentially harmful effects on human health, mainly in high-risk populations (Sousa et al., 2007; Filleul et al., 2006; Weschler, 2006), and on habitats and their vegetation (Davis and Orendovici, 2006; Scebba et al., 2006).

Due to the nature of ozone, its photolysis in the troposphere has been shown to be directly related to ultraviolet solar radiation at a wavelength of around 300 nm, followed by reaction with water molecules, sources of OH radicals, which take part in reactions responsible for the oxidation of other gases present in the atmosphere (Guicherit and Roemer, 2000). The nitrogen dioxide (NO_2) photo-dissociates to form nitrogen oxide (NO) and atomic oxygen (O), which immediately combines with oxygen (O_2) to form ozone (O_3). The nitrogen oxides (NO_x) act as a catalyst in the ozone formation process (Frost et al., 1998). Studies by Monks (2000), Kleinman (2000) and Trainer et al. (2000), all based on observations in rural environments, showed that ozone production was limited by NO_x availability. In the presence of a sufficient amount of NO_x, the main source of

* Corresponding author. Fax: +34 941 299 277.
E-mail address: joaquin.ordieres@unirioja.es (J.B. Ordieres).

1364-8152/$ - see front matter
doi:10.1016/j.envsoft.2007.11.009

ozone production is the oxidation of carbon monoxide (CO) and volatile organic compounds (VOCs).

Of particular interest is the chemical coupling between ozone and nitrogen oxides, prompting different authors to study the atmospheric relationship between O_3 and NO_x, in order to obtain further knowledge of this phenomenon (Clapp and Jenkin, 2001). Sillman (1999) affirms that two photochemical regimes may be differentiated in ozone production: an initial regime, before NO_x saturation takes place, when photochemical ozone production in urban areas increases with emissions of NO_x, but is less sensitive to emissions of VOCs; and a second regime, after NO_x saturation has occurred, when ozone levels rise as VOC levels increase, and fall when NO_x levels decrease.

Furthermore, it should not be forgotten that daily variations in ozone are controlled not only by variations in precursor gases and VOCs but also by local weather conditions (Fischer et al., 2003). Factors such as temperature, wind behavior and relative humidity have an important influence on daytime ozone levels. The specific meteorological characteristics of the area studied here make the development of ozone prediction models all the more interesting. The study was performed in two neighboring cities: Calexico, in California (US), and Mexicali, in Baja California (Mexico). These cities are situated in semi-desert regions, with high summertime temperatures being a natural characteristic of the area. Moreover, air quality in both cities has binational features due to their geographical location and the natural, two-way flow of contaminants.

Legislation in both countries has laid down regulations to govern acceptable national emission levels in the first instance, and regional levels in agreement with local governments. A National Ambient Air Quality Standard (NAAQS) is in force in the US, but California has established its own standards: California Ambient Air Quality Standards (CAAQS). Californian standards differ substantially as regards ozone. The national standard for ozone is 0.080 ppm average concentration over 8 h and 0.12 ppm over 1 h as maximum values (US-EPA, 2006). In contrast, the maximum average concentrations in California are 0.070 ppm over 8 h and 0.090 ppm over 1 h (CARB, 2005, California).

Using the various mechanisms, techniques and methodologies available, numerous authors have proposed different strategies for resolving air contaminant prediction problems of highest peaks, both short-term and mid-term prediction accuracy, etc (McCullagh and Nelder, 1989; Hastie and Tibshirani, 1990; Salcedo et al., 1999; Ho et al., 2002; Gardner and Dorling, 2000a; Podnar et al., 2002; Sousa et al., 2007). Sokhi et al. (2006) and Han (2007) used Eulerian grid models with good results. Thompson et al. (2001) and Gardner and Dorling (2000b) made an interesting review of statistical methods for the meteorological adjustment of ozone. Along the same lines, Schlink et al. (2006) confirm the efficient performance of non-linear multivariate tools, such as generalized additive and neural network models for application in warning systems of high ozone concentrations. A detailed review of prediction techniques can be found in Gardner and Dorling (1998). The first Position Paper of this journal, Jakeman et al. (2006), contains a comprehensive set of guidelines for evaluating environmental models including quantitative and qualitative measures of performance.

Among the different strategies reported in the literature for the development of models, neuronal networks and, specifically, the MultiLayer Perceptron (MLP) are being increasingly used in applications for predicting contamination levels or for estimating meteorological adjustments in ozone trends (Hornik, 1993; Bishop, 1997; Gardner and Dorling, 1998; Flake, 1998; Haykin, 1999; Kolehmainen et al., 2001; Ordieres et al., 2005; Dudot et al., 2007). In particular, nitrogen oxide (NO_x and NO_2) concentrations were predicted by Gardner and Dorling (1999) applying a Multilayer Perceptron-based model and other statistical models, and the comparison of the different results revealed the benefit of using a Multilayer Perceptron. Additionally, Artificial Neural Networks have recently been used to predict SO_2 levels and have proven to be of greater efficiency than linear methods (Chelani et al., 2002). Multilayer Perceptrons, in particular, have provided better results than statistical linear methods. Artificial Neural Networks (ANNs) are mathematical models capable of determining a non-linear relationship between two data sets (Haykin, 1999). ANNs are universal functions of approximation that can be applied to problems, in which there is *a priori*, no knowledge of the relevance of the input variables (Hornik et al., 1989; Hornik, 1993; Pernía-Espinoza et al., 2005; Martínez-De-Pisón et al., 2006). Since the mapping carried out by ANNs is non-linear, it is complex to understand; nevertheless, certain simple methods can be used to explore input relevance. Recently, Pires et al. (2008) used Multiple Linear Regression (MLR) and Principal Components Regression (PCA) for meteorological and environmental parameter validation for tropospheric ozone forecasting models.

This study has two main objectives: firstly, to provide an advanced model for predicting maximum ozone levels 1 day ahead in order to establish a strategic decision-making process; and secondly, to explore the capability of recurrent neural networks, such as ELMAN, in order to test their capabilities. As regards the first objective, an in-depth analysis was performed for the types of models, and an exhaustive search was performed for each model to identify relevant variables, since the aim was to build a model with the lowest possible margins of error. Although there were no such models in the region studied, there are those that have been developed for predicting maximum ozone levels in other parts of the world; hence, we also sought to improve on models described in the prior literature or, in their absence, on strategies that were not hitherto useful. As regards the second objective, in the case of the MLP and ELMAN networks, the aim with the recurrent networks was to try to improve the slow learning times displayed by the neuronal networks with MLP, without any noticeable loss of quality in the solution.

To begin this paper, Section 1.1 describes the geographical area where the construction of the models was validated, as well as the data available and the data management strategy. This is followed by a description of the models used, from the most traditional models to the non-parametric models, as well as the different error criteria used to measure the quality