Expert Systems with Applications 39 (2012) 107–116

Contents lists available at SciVerse ScienceDirect

Expert Systems with Applications

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

Unsupervised system to classify SO_2 pollutant concentrations in Salamanca, Mexico

J.M. Barrón-Adame [a,*], M.G. Cortina-Januchs [a], A. Vega-Corona [a], D. Andina [b]

[a] Universidad de Guanajuato, Campus Irapuato-Salamanca, División de Ingenierías, Salamanca, Mexico
[b] Universidad Politécnica de Madrid, Escuela Técnica Superior de Ingenieros de Telecomunicación, Madrid, Spain

ARTICLE INFO

Keywords:
Air pollution
Meteorological variables
Artificial neural networks
Self-Organizing Maps (SOM)
Clustering

ABSTRACT

Salamanca is cataloged as one of the most polluted cities in Mexico. In order to observe the behavior and clarify the influence of wind parameters on the Sulphur Dioxide (SO_2) concentrations a Self-Organizing Maps (SOM) Neural Network have been implemented at three monitoring locations for the period from January 1 to December 31, 2006. The maximum and minimum daily values of SO_2 concentrations measured during the year of 2006 were correlated with the wind parameters of the same period. The main advantages of the SOM Neural Network is that it allows to integrate data from different sensors and provide readily interpretation results. Especially, it is powerful mapping and classification tool, which others information in an easier way and facilitates the task of establishing an order of priority between the distinguished groups of concentrations depending on their need for further research or remediation actions in subsequent management steps. For each monitoring location, SOM classifications were evaluated with respect to pollution levels established by Health Authorities. The classification system can help to establish a better air quality monitoring methodology that is essential for assessing the effectiveness of imposed pollution controls, strategies, and facilitate the pollutants reduction.

1. Introduction

Air pollution is a very complex phenomenon that poses significant threats to human health and the environment throughout the developed and developing countries (Chak & Xiaohong, 2008). Air pollution is caused by both natural and man-made sources. Major man-made sources of ambient air pollution include industries (Pal, Kim, Hong, & Jeon, 2008), transportation (Bignal, Ashmore, & Headley, 2008; Joumard, 2009), power generation (Younger, Morrow-Almeida, Vindigni, & Dannenberg, 2008), unplanned urban areas (Joumard, Lamure, Lambert, & Tripiana, 1996), etc. Therefore, the issue of air quality is receiving more attention as an increasing fraction of the countries population are now living in urban areas and are in demand of a cleaner environment (EPA, 2000; PNUMA, 2007; WHO, 2006).

Meteorology is well known to be an important factor contributing to air quality (Arain et al., 2007; Elminir, 2005; Mandurino et al., 2009; Seaman, 2000; Sousa et al., 2008). It is extremely important to consider the effect of meteorological conditions on atmospheric pollution, since they clearly influence dispersion capability in the atmosphere. It is well known that severe pollution episodes in the urban environment are not usually attributed to sudden increases in the emission of pollutants, but to certain meteorological conditions which diminish the ability of the atmosphere to disperse pollutants (Nadir & Selici, 2008; Turias, Gonzlez, Martfn, & Galindo, 2006). The frequency distribution of air pollutant concentration is useful in understanding the characteristics of air quality. It can be used to estimate how frequently a critical concentration level is exceeded (Ozden, Dogeroglu, & Kara, 2008). However, the concentrations of air pollutants usually vary randomly and are correlated with several factors such as types of fuels consumed, geographical and topographical peculiarities, town planning and meteorological factors, etc. (Demirci & Cuhadaroglu, 2000). Air quality management and information systems are required to control air pollutants and provide proper actions, controlling strategies and a better and safe environment for future generation (Bhanarkar, Goyal, Sivacoumar, & Chalapati Rao, 2005; Kurt, Gulbagci, Karaca, & Alagha, 2008; Lumbreras, ValdTs, Borge, & Rodrfguez, 2008). Thus, a thorough understanding of the meteorological field is fundamental to predicting and understanding air pollution in urban areas (Lee, Kim, Kim, & Lee, 2007).

Many clustering techniques can be used to determine the nature groups of similar objects (Du, 2010; Warren Liao, 2005). On atmospheric science, Hanna et al. (2001) showed ground ozone (O_3) concentrations in association with the mixing depth and wind field patterns and documented that meteorological fields may increase uncertainty in air quality (Hanna & Davis, 2002; Hanna, 2000). Yu and Chang (2001) analyzed the PM_{10} time series in

* Corresponding author.
E-mail address: badamem@salamanca.ugto.mx (J.M. Barrón-Adame).

0957-4174/$ - see front matter © 2011 Elsevier Ltd. All rights reserved.
doi:10.1016/j.eswa.2011.05.083

Taiwan from July 1993 to June 1998, and delineated the PM_{10} concentrations into five air quality basins by hierarchical clustering (Tai-Yi & Len-Fu, 2001). In another study presented by Turalloglu, Nuhoglu, and Bayraktar (2005), the relationship between daily average Total Suspended Particulate (TSP) and Sulphur Dioxide (SO_2) concentrations with meteorological factors for 1995–2002 winter seasons was statistically analyzed using the stepwise multiple linear regression analysis for Erzurum City. They have shown that, higher TSP and SO_2 concentrations are strongly related to colder temperatures, lower wind speed, higher atmospheric pressure and weakly correlated with rain and higher relative humidity (Turalloglu et al., 2005). More recently, Riccio, Giunta, and Chianese (2007) apply a trajectory classification of PM_{10} aiming to identify the role exerted by meteorology in the Naples urban area (Southern Italy). They identify and evaluate the effects of eight clusters on air quality (Riccio et al., 2007). Rimetz-Planchon, Perdrix, Sobanska, and BrTmard (2008) investigated also for PM_{10} polluted episodes with meteorological situations in an urban and industrialized coastal site of the southern part of the North Sea, representative of a typical harbor for trade. They explain the spatio-temporal variability of PM_{10} at the urban scale and identify Air Quality (AQ) regimes related to PM_{10} levels and local weather conditions applied to the air quality database of Dunkerque in 2002 (Rimetz-Planchon et al., 2008). Kim Oanh, Chutimon, Ekbordin, and Supat (2005) have also developed an automated scheme to classify the synoptic meteorological conditions governing over Northern Thailand. Because a quantitative approach utilizes a variety of meteorological variables for the classification of synoptic patterns, it involves intensive statistical data treatment, normally accomplished in the literature by a combination of the Principal Component Analysis (PCA) and clustering techniques (Kim Oanh et al., 2005). In the pre-mentioned reviews, traditional statistical clustering techniques were used for classification of environmental data. In recent years, the considerable progress has been in the developing of Artificial Neural Network (ANN) models for air quality (Cortina-Januchs, Barrón-Adame, Vega-Corona, & Andina, 2009; Gardner & Dorling, 1998). The Self-Organizing Maps (SOM) (Kohonen, 1990), an ANN with unsupervised learning is the other commonly used clustering algorithm in environmental data (Andina, Jevtić, Marcano, & Barrón-Adame, 2007). SOM is suitable for data classification because of its visualization property (Alvarez-Guerra, Gonzlez-Piuela, AndrTs, Galn, & Viguri, 2008; Seo & Obermayer, 2004; Vesanto & Alhoniemi, 2000). For example, the SOM has been used to identify patterns in satellite imagery in oceanography (Richardson, Risien, & Shillington, 2003); to visualize and cluster volcanic ash (Ersoy, Aydar, Gourgaud, Artuner, & Bayhan, 2007); or to estimate the risk of insect species invasion associated with geographic regions (Watts & Worner, 2009).

In this study, the suitability of SOM for classifying and interpreting the air quality and level of SO_2 concentrations in Salamanca city was investigated to implement an air pollution system. The results were compared to pollution levels for SO_2 established by Health Authorities. We have selected SOM as the best method for the following reasons:

- Unsupervised nature: Since the trajectory clusters are not known, unsupervised learning is required for trajectory clustering, which can be achieved by SOM.
- Classification/clustering power of neural networks: SOM is a neural network, which is a well-known powerful classification/clustering tool.
- Topological learning structure: In the training process, not only the winning neuron, but also neighboring neurons learn from the training data depending on their distance to the winning neuron, which is known as one of the most important aspects of SOM.

2. Material and method

2.1. Features of study area

Salamanca is a city in the Mexican state of Guanajuato with a population of approximately 234,000 inhabitants and located some 350 km to the northwest of Mexico city (INEGI, 2005).

Salamanca is cataloged as one of the most polluted cities in Mexico (Vega López, 2006). Although environmental management in Mexico began in 1971 with the Law to Prevent and Control Environmental Pollution, in the last decade Mexico began with true efforts to generate and compile environmental information. The National Institute of Ecology (INE), a decentralized organization of the Ministry of the Environment and Natural Resources (SEMARNAT, 2008), oversees policy-making decisions for air quality, solid and hazardous waste management, environmental impact assessment, global climate change, ozone depletion, wildlife management and natural reserves (INE, 2008).

In Salamanca, the Program to Improve the Air quality (ProAire) is composed of measures that affect transportation, industry, service sector, natural resources, health, and education. The ProAire program contemplates the urgent and immediate reduction of pollutant emissions when measurements of these pollutants register levels above those established by Health Authorities. When first ProAire concluded in 2000, environmental authorities undertook a longer, ambitious air quality improvement program ProAire 2002–2010. However, accurate measures were needed to determine how improving air quality would improve health and reduce health expenditures so that the new pollution control strategies could be evaluated (Gurjar, Butler, Lawrence, & Lelieveld, 2008; Jehng-Jung & Ming-Ru, 2006; Monteiro, Miranda, Borrego, & Vautard, 2007).

In our study the established ProAire limits by Health Authorities are taken as references to select the best SOM structure to classify the SO_2 concentrations correlated with wind fields. Table 1 shows the Established ProAire limits for SO_2.

2.2. Air pollutants and meteorological data

The main causes of pollution in Salamanca are due to fixed emission sources such as Chemical Industry and Power Generation, SO_2 being one the most important pollutant in air (IEEG, 2008; INE, 2008). Currently, in Salamanca an Automatic Environmental Monitoring Network (AEMN) is installed in which time series of criteria pollutant and meteorological parameters are obtained. Fig. 1 shows the AEMN distribution in Salamanca.

In our study, we have considered the maximum and minimum daily SO_2 concentration during the period of 2006 to train a SOM Neural Network in each monitoring station. Pollutant concentration have associated their correlated wind parameter measured simultaneously per minute. In this case, we considerer that Hourly and Daily mean do not represent an appropriate distribution of pollutants to train a Neural Network because this can be affected

Table 1
Established environmental contingency levels for SO_2 by Environmental Autorities in Salamanca, (a) *Pre-contingency* (b) *Phase I* and (c) *Phase II*.

Pollutant	Level activation	Level annulment
(a) Environmental Pre-contingency		
SO_2	$\geqslant$145 ppb and <225 ppb	<145 ppb
(b) Environmental Contingency Phase I		
SO_2	$\geqslant$225 ppb and <305 ppb	<145 ppb
(c) Environmental Contingency Phase II		
SO_2	$\geqslant$305 ppb	<145 ppb