# PM10 Classification with SOM Neural Networks for Air Quality Levels Estimation

**M.G. Cortina [1], J.M. Barron [1], S. Ledesma [1], D. Andina [2] and A. Vega-Corona [1]**
[1] Universidad de Guanajuato
Facultad de Ingeniería Mecánica, Eléctrica y Electrónica.
Salamanca, Guanajuato. México
januchs,badamem,selo,tono@salamanca.ugto.mx
[2] Universidad Politécnica de Madrid
Departamento de Señales, Sistemas y Radiocomunicaciones, E.T.S.I.
Telecomunicación, Madrid, Spain
andina@gc.ssr.upm.es
*(Paper received on September 04, 2006, accepted on September 27, 2006)*

**Abstract.** This paper presents a new methodology to detect and classify PM10 Particles concentration according to Air Quality Levels. In this work, Meteorological Variables are analyzed to make a classification decision. The method consists of tree steps. In first step, we group using a SOM Neural Networks the Pollutant concentrations in two classes, these classes are noise data and validated data. In second step, we create a Representative Feature Vector using the information of contingency levels that we know a priori. In third step, a new SOM Neural Network is trained with the Representative Feature Vector built in second step, and the pollutant concentrations and meteorological variables (Validated Data) are self-organized in fourth classes. Finally, we obtained the Air Quality Level. Our experiments with this methodology exhibit good results in Air Quality Classification Levels. In this case a time series obtained from the Environmental Monitoring Network of the Salamanca city, Guanajuato, México is used.

## 1. Introduction

Air pollution is a broad term applied to any chemical, physical, or biological agent that modifies the natural characteristics of the atmosphere. The atmosphere is a complex, dynamic natural system that is essential to support life on planet earth.

Air pollution is one of the most important environmental problems, pollution is caused by both natural and man-made sources. Major man-made sources of ambient air pollution include industries, automobiles, domestic activities and power generation [1]. Air pollution has both acute and chronic effects on human health. Health effects range anywhere from minor eyes irritations and the upper respiratory system to chronic respiratory disease, heart disease, lung cancer, and death.

Nowadays, many countries make big efforts to minimize air pollution. In polluted countries like Mexico a continuous monitoring of Air Quality to measure pollutant concentrations to reduce possible negative effects in population health is necessary. A special case with great pollution is Salamanca, Guanajuato in Mexico. Salamanca city is catalogued as one of the most polluted cities in Mexico [2]. The main causes of pollution in Salamanca are due to fixed emission sources such as Chemical Industry

© H. Sossa and R. Barrón (Eds.)
Special Issue in Neural Networks and Associative Memories
Research in Computing Science 21, 2006, pp. 97-104

and Electricity Generation, being the most important pollutants in Air, Sulphur Dioxide ($SO_2$), measured in Parts Per Billion (PPB), and Particulate Matter less than 10 micrometers in diameter ($PM_{10}$), measured in micrometers ($\mu m$). This article focuses on $PM_{10}$ concentration.

In 1999, an Environmental Monitoring Network (EMN) was established, this network provided time series about criteria pollutant [3] among other meteorological variables. In an effort to fight pollution on the region, in July 2005, the Environmental Contingency Program was launched, the purpose of it being to protect population health, especially that of vulnerable groups. This program contemplates the urgent and immediate reduction of $SO_2$ and $PM_{10}$ emissions when measurements of these pollutants register levels above those established by Health Authorities. To accomplish it, 3 phases were established: Pre-contingency, Phase I Contingency and Phase II Contingency for $SO_2$, $PM_{10}$ particles and for a combination of both [2, 4].

### 1.1. Particulate Matter

Particulate Matter consists of solid or liquid aerosol particles suspended in the air and has a diverse chemical composition related to its sources. Under normal ambient conditions of sampling and analysis, particulate matter exists almost exclusively in solid phase but can include liquid aerosols such as the heavier components of diesel combustion products and nitric acid.

Some particles are emitted directly into the air from a variety of sources that are either natural or related to human activity. Natural sources include bushfires, dust storms, pollens and sea spray. Those related to human activity include motor vehicle emissions, industrial processes, unpaved roads and wood heaters. Particulate Matter is commonly designated as $PM_{2.5}$ or $PM_{10}$ they are, refereed as particles with aerodynamic diameters less than 2.5 μm and 10 μm, respectively.

The statistical correlation between high levels of inhalable particulate matter and increased mortality has been widely reported. The particles in the $PM_{2.5}$ and $PM_{10}$ fractions can be inhaled into the lungs, causing damage to the alveolar tissues and inducing various health problems [5,6]. The adverse effects may range from the irritation of the lung tissues resulting in coughing to severe respiratory problems for individuals with asthma or heart disease. The mechanic details of how the constituents of the particulate matter induce adverse health effects are currently areas of intense scientific research. The polycyclic aromatic hydrocarbons and heavy metals present in the PM2.5 and PM10 samples have been studied extensively with regard to their roles in inducing toxicity effects [7].

### 1.2. Artificial Neural Network

Artificial Neural Networks (ANN) are biologically inspired network based on the organization of neurons and decision making process in the human brain [8]. In other words, it is the mathematical analogue of the human nervous system.

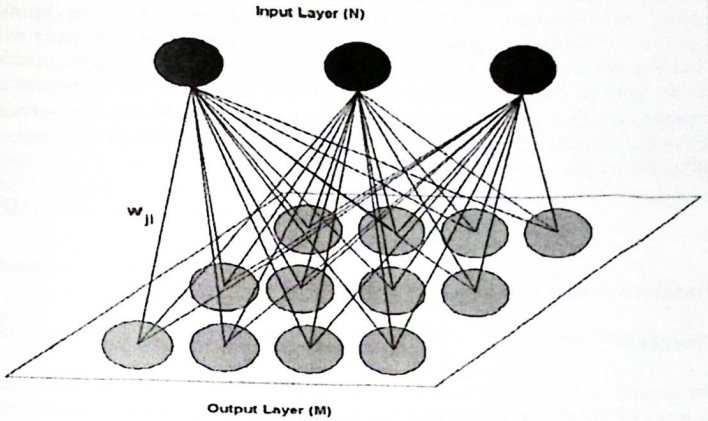


**Fig. 1.** SOM Neural Network structure

**SOM Neural Network** The Self-organizing map [9] basically provides a form of cluster analysis by producing a mapping of high-dimensional input data x, $x \in \Re^n$ onto a usually 2-dimensional output space while preserving the topological relationships between the input data items as faithfully as possible. It consists of a set of units, which are arranged in some topology where the most common choice is a two dimensional grid [10,11]. Each of the units *i* is assigned a weight vector mi of the same dimension as the input data, $m_i \in \Re^n$ . In the initial setup of the model prior to training, the weight vectors are filled with random values.

During each learning step, the unit *c* with the highest activity level, the winner *c* with respect to a randomly selected input pattern *x*, is adapted in a way that it will exhibit an even higher activity level at future presentations of that specific input pattern. Commonly, the activity level of a unit is based on the Euclidean distance between the input pattern and that unit's weight vector. The unit showing the lowest Euclidean distance between its weight vector and the presented input vector is selected as winner. Hence, the selection or the winner *c* may be written as given in expression (1)

$$c : \|x - m_c\| = min_i \|x - m_i\| \tag{1}$$

Adaptation takes place at each learning iteration and is performed as a gradual reduction of the difference between the respective components of the input vector and the weight vector. The amount of adaptation is guided by a learning-rate $\alpha$ that is gradually decreasing in the course of time. As an extension to standard competitive learning, units in a time-varying and gradually decreasing neighborhood around the

winner are adapted, too. This strategy enables the formation of large clusters in the beginning and fine-grained input discrimination towards the end of the learning process. In combining these principles of self-organizing map training, we may write the learning rule as given in Expression 2. We make use of a discrete time notation with $t$ denoting the current learning iteration. The other variables of this expression are $\alpha$ representing the time-varying learning-rate, $h_{ci}$ representing the time-varying neighborhood-kernel, $x$ representing the currently presented input pattern, and mi denoting the weight vector assigned to unit $i$.

$$m_i(t+1) = m_i(t) + \alpha(t)h_{ci}[x(t) - m_i(t)] \tag{2}$$

## 2. Database and Variables Definition

### 2.1. Database

In this research, a real and historical time series database from the EMN has been used. Data series of the months from March to May in the year 2005 have been analyzed. These time series consist of a total of 129,600 multidimensional patterns of pollutants and meteorological variables.

### 2.2 Variables Definition

It consists of normalized pollutants concentration values ($PM_{10}$) and normalized meteorological values (Wind Speed and Relative Humidity). In Table 1, variables are defined in order to build a Representative Feature Vector $\mathbf{x}_j$ and to define a pattern set $\mathbf{X}_* = \{x_1, x_2, \ldots x_j, \ldots, x_n\}$. Let $\mathbf{X}_{PM10}$ be a particles concentration set, thus their corresponding pattern is defined as $x_j = \{x_1, x_2, x_3\}$.

**Table 1.** Variables Definition: $PM_{10}$ Concentration, RH; Relative Humidity, WS; Wind speed.

| | $\mathbf{X}_{PM_{10}}$ |
|---|---|
| Variables $x_i$ | $\mathbf{x}_j$ |
| $x_1$ | $PM_{10}$ |
| $x_2$ | HR |
| $x_3$ | WS |

## 3. Proposed Methodology

Figure 2 shows the flow diagram of the methodology that was followed for the classification process. This process has three steps. Steps which will be explained in detail:

- The Cleaning Clustering Method
- The Building Representative Feature Vector Method
- The Clustering Classification Method

Cleaning Clustering Method: This method uses a SOM Neural Network to classify the patterns to be analyzed. Because time series have noise, we trained the SOM Neural Network using time series in two classes, the first class is noisy data and the second class is validated data. Validated data will be used in the Clustering Classification Method, meanwhile noisy data will be deleted.

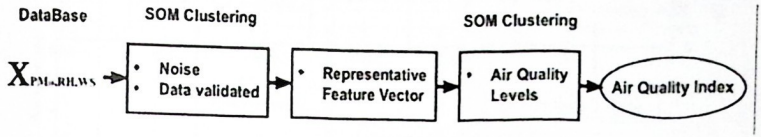


**Fig. 2.** Block diagram of our methodology.

**Building Feature Vector** Because of data nature of the variables, it is possible to know the variables limits. In this section a Representative Feature Vector (RFV) is built [12], with the centers of the involved variables, these centers are for Non Contingency, Pre-Contingency, Phase I Contingency and Phase II Contingency. This RFV was used to train a new SOM Neural Network with a [4 1] line topology, and try to locate the SOM Neural Neural node in the appropriate contingency level.

**Clustering Classification Method** After the SOM Neural Network has been trained with a [4 1] line topology [9,13] in order to have four clusters and therefore four prototypes according to contingency levels, we proceed to group the new patterns set in four classes according to the contingency levels (see Table 2).

**Tabla 2.** Contingency Concern Levels and Pattern centers vector

| Contingency Level Center | | |
|---|---|---|
| Contingency Levels | Pattern Center WS-$PM_{10}$ | Pattern Center RH-$PM_{10}$ |
| Non Contingency | 75.10 | 75.50 |
| Pre-Contingency | 200.5.10 | 200.5.50 |
| Phase I Contingency | 304.5.10 | 304.5.50 |
| Phase II Contingency | 450.10 | 450.50 |

## 4 Experimental Results

Figure 3 shows the SOM Neural Network trained with the RFV shown in Table 3. Figure 4 shows a real time series for PM10 concentrations analyzed every minute.

They are classified according to the process followed by our methodology. The meteorological variables used were Wind Speed and Relative Humidity, creating 129,600 three dimensional pattern vectors $x_i$, for pollutant ($X_{PM10}$), as it is shown in Table 1. Pollutant concentrations and meteorological variables are provided by the EMN from Salamanca. The clustering method, four clusters have been performed from a new feature vector.

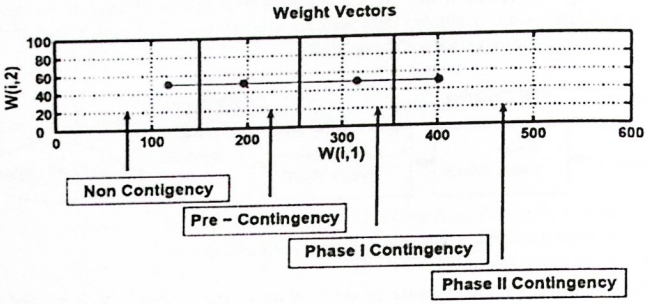


**Fig. 3.** SOM Neural Network structure trained using a Representative Feature Vector, with Linear [4 1] topology

**Tabla 3.** Contingency Concern Levels and SOM center clusters

| Contingency Level Center | | |
|---|---|---|
| Contingency Levels | Cluster Center WS-$PM_{10}$ | Cluster Center RH-$PM_{10}$ |
| Non Contingency | 117.47.10 | 117.47.50 |
| Pre-Contingency | 196.14.10 | 196.14.50 |
| Phase I Contingency | 315.35.10 | 315.35.50 |
| Phase II Contingency | 401.53.10 | 401.53.50 |

In Table 4 the error percent obtained with the proposed methodology is shown, we can observe that the error depends on the number of data; more data, more error.

Noisy pattern is an inconsistent element in the time series and it is caused by blasts of wind. Noisy elements can cause bad estimation, so with this method a better estimate is obtained.

## 5. Conclusions

In this work, A time series was obtained from the EMN, this time series contain noisy data and valid data, with a SOM Neural Network this time series was clustered in two classes, the class with noisy data was deleted and the valid data was using to train a

new SOM Neural Network. This SOM classify the valid data according to the Air Contingency Levels. This methodology presented good results, because the Contingency Levels are known, allowing to create a Representative Feature Vector for each level. Thus, less patterns are required to train a SOM Neural Network. The classification error depends only on the number of data. Our method produces errors that are less than 1%.

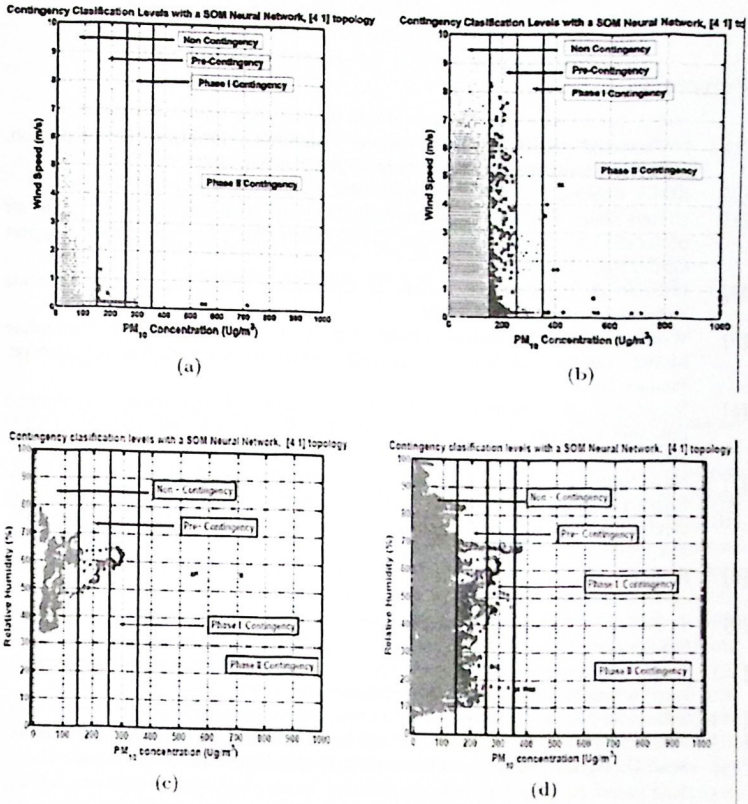


**Fig. 4.** (a) $PM_{10}$ vs WS Patterns Classification for March 5, 2005. (b) PM10 vs WS Patterns Classification for March, 2005. (c) PM10 vs RH Patterns Classification for March 5, 2005. (d) PM10 vs RH Patterns Classification for March, 2005.

**Tabla 4.** Error percent using a SOM Neural Network trained with a Representative Feature Vector

| % Error | | | |
|---|---|---|---|
| Data Analyzed | Number of Data | % Error WS-$PM_{10}$ | % Error RH-$PM_{10}$ |
| one day | 1440 | 0.28 | 0.28 |
| one month | 32971 | 0.4 | 0.4 |

## References


[1] Environmed Research Inc. Alpha Nutrition. Problem air pollution. http://www.nutramed.com/environment/particles.htm, 2004.

[2] SIMA. Sistema de información ambiental. http://www.sima.com.mx/, 2004.

[3] United States Environmental Protection Agency. Risk assessment for toxic air pollutants: A citizen's guide-epa 450/3-90-024. Air Risk Information Support Center (MD-13), March 1991.

[4] Instituto de Ecologia del Estado de Guanajuato. Programa de contingencias ambientales atmosféricas, 2005.

[5] Wordl Health Organization. Health Aspect of Air Pollution with Particulate Matter, Ozone and Nitrogen Dioxide. WHO Regional Office for Europe, January 2003.

[6] Secretariat of Commission for Environmental Cooperation. Continental Pollutant Pathway. Communication and Public Outreach Deppartment of CEC Secretariat, 1997.

[7] Ngee-Sing Chong, Kavitha Sivaramakrishnan, Marion Well, and Kathy Jones. Characrerization of inhalable particulate matter in ambient air by scanning electron macroscopy and energy-dispersive x-rau analysis. EJEAFChe, (3):145–164, 2002.

[8] R. Setiono and Huan Liu. Neural-network feature selector. IEEE Transactions on Neural Networks, 8(3):654–662, May 1997.

[9] T. Kohonen. The self organization map. Proceedings. IEEE, 78(9):1464–1480, September 1990.

[10] James A. Freman and David M. Skacura. Neural Networks Algorithms, Application and Programming Techniques. 1991.

[11] S.Haykin. Neural Networks. Prentice Hall, 2nd edition, 1999.

[12] Jain A.K. Pattern Recognition. John Wiley and Sons, Inc, 1988. pp.1052–1063.

[13] MatLab. Neural networks toolbox. Math Works, September 2005.