Air Pollution Data classification by SOM Neural Network

J. M. Barrón-Adame, O. G. Ibarra-Manzano and A. Vega-Corona
División de Ingenierías
Universidad de Guanajuato
Salamanca, Gto., México
Email: badamem@ugto.mx

M. G. Cortina-Januchs and D. Andina
Technical University of Madrid
Madrid, Spain
Email: andina@gc.ssr.upm.es

***Abstract*—Over the last ten years, Salamanca has been considered among the most polluted cities in México. This paper presents a Self-Organizing Maps (SOM) Neural Network application to classify pollution data and automatize the air pollution level determination for Sulphur Dioxide (SO_2) in Salamanca. Meteorological parameters are well known to be important factors contributing to air quality estimation and prediction. In order to observe the behavior and clarify the influence of wind parameters on the SO_2 concentrations a SOM Neural Network have been implemented along a year. The main advantages of the SOM 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. The results show a significative correlation between pollutant concentrations and some environmental variables.**

I. Introduction

Nowadays, many countries make big efforts to minimize air pollution [1], [2], [3]. 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 [4]. The main causes of pollution in Salamanca are due to fixed emission sources such as Chemical Industry and Electricity Generation, being Sulphur Dioxide (SO_2) (measured in Parts Per Billion, (PPB), and Particulate Matter less than 10 micrometers in diameter PM_{10} (measured in micrometers, (μm) the most important air pollutants. This article focuses the analysis on Sulphur Dioxide (SO_2) concentrations. SO_2 is one air pollutants with the highest concentration in Salamanca, where three monitoring stations have been installed in order to know the level of air pollution; the measure records of each monitoring station are handled separately. Actually, an environmental contingency alarm is activated when daily average pollutant concentration, in a single monitoring station, exceeds a established threshold.

Meteorology is well known to be an important factor contributing to air quality [5], [6], [7], [8], [9]. 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 [10], [11]. 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. [12].

In recent years, the considerable progress has been in the developing of Artificial Neural Network (ANN) models for air quality [13], [14]. The Self-Organizing Maps (SOM) [15], an ANN with unsupervised learning is the other commonly used clustering algorithm in environmental data [16]. SOM is suitable for data classification because of its visualization property [17]. For example, the SOM has been used to identify patterns in satellite imagery in oceanography [18]; to visualize and cluster volcanic ash [19]; or to estimate the risk of insect species invasion associated with geographic regions [20]. In this paper the application of the Self-organizing Maps as clustering algorithm is presented in order to classify the SO_2 pollutant concentration.

The paper is organized as follow, in section II the case of study is presented were the conditions of the city of Salamanca are introduced. Section III introduce the Artificial Neural Networks with especial emphasis in the Self-organizing Maps. Section IV shows the obtained results and finally, section V presents the generated conclusions.

II. Case of study

In recent years, the city of Salamanca has been catalogued as one of the most polluted cities in Mexico [21]. Salamanca is a city in the state of Guanajuato with a population of approximately 234,000 inhabitants and located around 350 km to the northwest of Mexico city [22]. Currently, an

Environmental Monitoring Network (EMN) is installed in Salamanca. EMN is composes for three monitoring stations. Time series of criteria pollutants among other meteorological variables are obtained in each monitoring station. Figure 1 shows the EMN distribution.

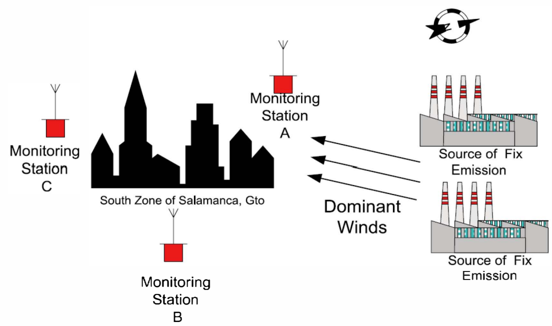


Fig. 1. Monitoring distribution.

Although environmental management in Mexico began in 1971 with the Law to Prevent and Control Environmental Pollution, in the last decade Mexico began its efforts to generate and compile environmental information [23]. In Salamanca, the Program to Improve the Air Quality (ProAire) is composed of measures that affect transportation, industry, the service sector, natural resources, health, and education. The ProAire program integrate 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. As was previously mentioned, this work focus the analisys in SO_2.

SO_2 is a colorless gas with a sharp, irritating odor. It is produced from the burning of fossil fuels (coal and oil) and the smelting of mineral ores that contain sulfur. When sulfur dioxide combines with water, it forms sulfuric acid, which is the main component of acid rain. When acid rain falls it can cause deforestation, acidify waterways to the detriment of aquatic life and corrode building materials and paints. SO_2 can affect the respiratory system, the functions of the lungs and irritate eyes. When SO_2 irritates the respiratory tract it causes coughing, mucus secretion, aggravates conditions such as asthma and chronic bronchitis and makes people more prone to respiratory tract infections [24].

III. Artifcial Neural Networks

Artificial Neural Networks (ANNs) are biologically inspired networks based on the neuron organization and decision making process in human brain [25]. One advantage of ANN approach is that most of the intense computation takes place during the training process. Once ANNs are trained for a particular task, operation is relatively fast and unknown samples can be rapidly identified in the field. ANNs can be classified as supervised and unsupervised. A kind of unsupervised ANN is the Self-organizing Map (SOM) [15].

A. Self-Organizing Maps (SOM)

The basic SOM Neural Network consists of the input layer, and the output (Kohonen) layer which is fully connected with the input layer by the adjusted weights (prototype vectors). The number of units in the input layer corresponds to the dimension of the data. The number of units in the output layer is the number of reference vectors in the data space. In SOM, the high-dimensional input vectors are projected in a nonlinear way to a low-dimensional map (usually a two-dimensional space), and SOM can perform this transformation adaptively in a topologically ordered fashion. Therefore, the neurons are placed at the nodes of a two-dimensional lattice. Every neuron of the map is represented by an n dimensional weight vector (prototype vector), $\theta = [\theta_1, ..., \theta_n]$, where n denotes the dimension of the input vectors. The prototype vectors together form a codebook. The units (neurons) of the map are connected to adjacent ones by a neighborhood relation, which indicates the topology of the map. The rectangular topology was used in this study. SOM can adjust the weight vectors of adjacent units in the competitive layer by competitive learning. Figure 2 shows a hexagonal SOM topology.

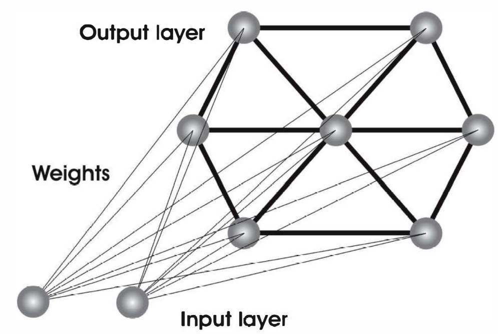


Fig. 2. A basic hexagonal SOM topological neighbourhood.

In the training (learning) phase, the SOM forms an elastic net that folds onto the “cloud” formed by the input data. Similar input vectors should be mapped close together on the nearby neurons, and group them into clusters. SOM is an unsupervised classification which is used to cluster a data set based on statistics only, and can be trained by an unsupervised learning algorithm in which the network learns to form its own classifications of training data without external help. The SOM is trained iteratively. The learning steps are as follows [26]:

Step 1. Initialize randomly the weight vectors, $\theta_j(0)$, drawn from the input dataset and set $t = 0$.

Step 2. Present an input vector x to the network and compute the Euclidean distance, d_j, between a sample of input vectors and all the prototype vectors at iteration t.

$$d_j = \|X_j - \theta_j(t)\| \quad (1)$$

Step 3. Find the winner unit c (best matching unit, BMU) which has the minimum Euclidean distance:

$$U_c = min\{d_j\} \tag{2}$$

Step 4. Update the connecting weight vectors of all neurons:

$$\theta_{i(t+1)} = \theta_j(t) + \eta(t)h_{cj}(t)[x(t) - \theta_j(t)] \tag{3}$$

Step 5. Increase time t to $t+1$. If $t < T$ then go to step 2, otherwise stop the training.

Here, t is the time of iteration and T is a predefined number of iterations, respectively; $x(t)$ is an input vector randomly chosen at time t; $\eta(t)$ is the learning rate and is a decreasing function of time; $h_{cj}(t)$ is called the neighborhood function.

The neighborhood function will decrease in time. The topological distance $r = \|r_j - r_c\|$ is calculated between unit j and winner unit c. The most commonly used neighborhood function is the Gaussian:

$$h_{cj}(t) = exp\left(-\frac{\|r_j - r_c\|^2}{2\sigma^2(t)}\right) \tag{4}$$

where $\sigma(t)$ is called the neighborhood radius.

Both the learning rate and neighborhood radius decrease monotonically during training, and the $\eta(t)$ will converge towards 0. The learning is broken down into two phases: the ordering phase and tuning phase. In the ordering phase, the neighborhood radius decreases linearly from 5 to 1, and the value of 1 was maintained over the tuning phase.

Figure 3 illustrates the clustering process. In the first step, the data are separated into two groups: training and testing. A SOM with four neurons is created and trained using a training dataset. Clustering results are compared with pollution levels established by Health Autorities. Another SOM is created with an additional neuron and trained. The evaluation criterion is compared. The number of neurons in SOM is increased until the evaluation criterion is achieved. The SOM with the best evaluation results is selected and the testing dataset is clustered using the best SOM. We stop the training process when all neurons in SOM structure have a difference of 1 % of each variable in the feature space and will be considered in the error classification. Finally, the evaluation criterion values are reported.

Each pattern (pollutant concentration and meteorological variables) can be represented as a point in a 3-dimension space and its projection on the 1D lattice using an SOM has been used to detect similar or different behavior among patterns during the analysis period. Patterns with a similar behavior can be expected to be projected onto the same neuron, while patterns with different behavior will tend to be assigned to different neurons in the SOMs. An optimal mapping would be the one that preserves on the 1D lattice, in the most faithful fashion, the exist ing distances in the 3-dimensions space.

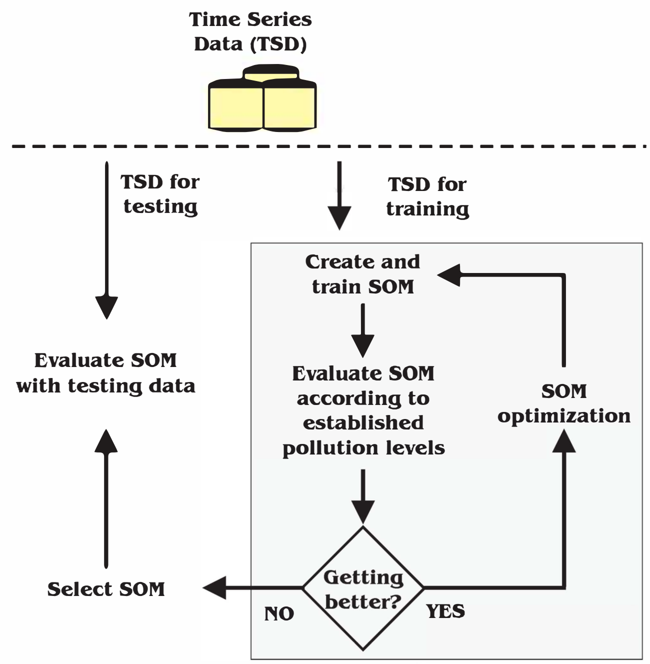


Fig. 3. Clustering process.

IV. Results

In the experiments, The SOM structures start with 4 neurons, and the number of neurons is increased one by one. A 1-dimensional SOM structure is used and number of neurons is increased only in one direction, such as 4 x 1 x 1, 5 x 1 x 1, 6 x 1 x 1, ..., 20 x 1 x 1 (a total of 17 structures). Error classification level is computed by Mean Absolute Error (*MAE*) as:

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|X_i - Y_i| \tag{5}$$

where X_i and Y_i are the observed and estimated value at i time, and N is the total number of observations. The best SOM will be the one with the smallest error classification level.

Table I summarizes the SOM neuron position in the feature space. Continuous lines separate the established contingency levels by Health Authorities.

Figure 4 display the neuron positions in the feature space created with maximum and minimum daily SO_2 concentrations. A [16 x 1 x 1] SOM structure performs better to classify SO_2 pollutant concentrations correlated with wind parameters.

V. Conclusion

In this paper, a Self-organizing Maps (SOM) Neural Network were applied in order to classify the Sulphur Dioxide (SO_2) pollutant concentrations. In the experiments, the SOM Neural Network was trained iteratively allowing to

Neuron positions			
	Cruz Roja		
neuron	SO_2 ppb	Dir. ^{o}	Vel. m/s
1	20.5	1.9	354.8
2	40.8	1.8	346.3
3	47.2	1.9	295.3
4	22.7	2.2	248.7
5	20.1	2.0	201.8
6	24.6	1.7	127.8
7	19.7	2.0	76.7
8	12.1	2.2	28.2
9	15.0	2.0	7.1
10	37.6	1.9	12.3
11	103.2	2.7	42.1
12	171.8	3.4	58.8
13	270.2	3.3	66.7
14	379.2	3.1	69.5
15	493.0	2.8	64.7
16	578.9	2.8	63.1

TABLE I
SOM NEURON POSITION IN THE FEATURE SPACE FOR THE THREE MONITORING STATIONS. THE INVERTER ORDER IN NATIVITAS STATION IS DUE TO THE ALEATORY SOM NEURON ORDER INITIALIZATION.

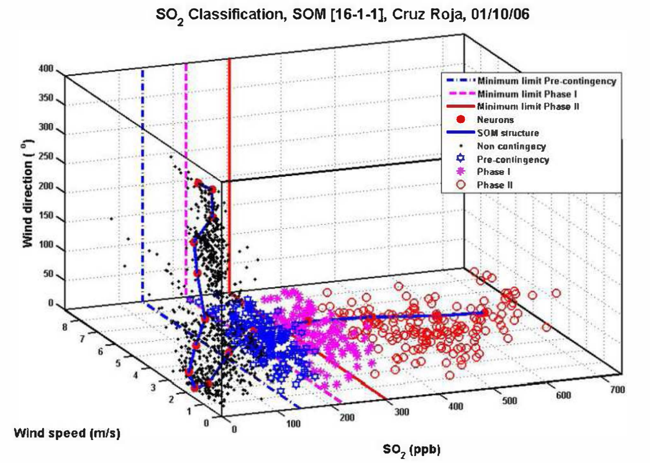
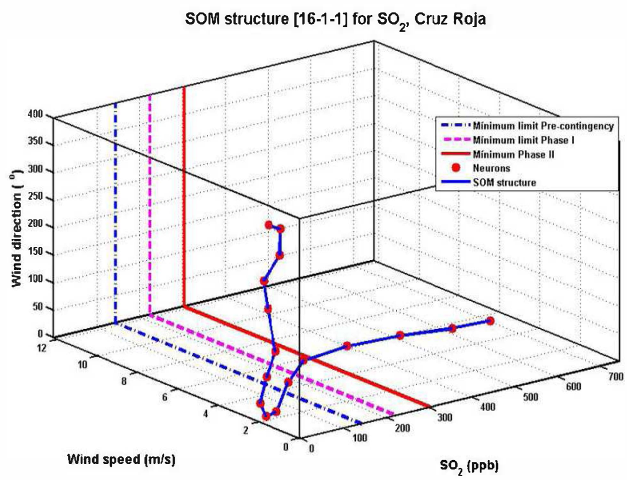


Fig. 5. SO_2 classification for October 1, 2006 in Cruz Roja station with a [16 x 1 x 1] SOM Neural Network structure where the • represent the SO_2 concentrations in Non- contingency, ⋆ represent the Pre-Contingency, ∗ represent the *Phase I* and ∘ represent the *Phase II* pollution levels respectively.



(a)

Fig. 4. Optimum SOM Neural Network with [16 x 1 x 1] structure to classify SO_2 pollutant concentrations correlated with wind parameters in Cruz Roja station

compare among the trained SOM networks. SOM clustering process, several SOM Neural Network structures (topologies) have been tested and trained in order to obtain a minimum SO_2 classification error. Presented methodology shows good results due to Contingency levels were knows, allowing to select the SOM network prototype. In order to obtain a robust SOM network, experiments can be extended in the future, using other variables such as Temperature, Relative Humidity, Rain, etc.

ACKNOWLEDGMENT

The authors wish to thank the Patronage for the air Quality Monitoring (Salamanca), the Institute of Ecology of Guanajuato (IEEG), The National Council for Science and Technology (*CONACyT*) in Mexico, The Computational Intelligence Laboratory (*LabInCo*). (University of Guanajuato. Mexico), The Group for Automation in Signals and Communications (*GASC*). (Technical University of Madrid. Spain) for the help provided to complete this study.

REFERENCES

[1] (In Spanish): Sistema de Información Medioambiental. “La contaminación atmosférica”. Madrid, España, Febrero 2007.
[2] Chinese Ambassy. “China SO_2 emissions reductions policies”. February 2007.
[3] UK Air Quality Archive, “Department for Environment Food, Rural Affairs, and the Devolved Administrations”. February 2007.
[4] (In Spanish): SIMA. “Sistema de información ambiental”, Red automática de monitoreo atmosférico, http://www.sima.com.mx/. 2004.
[5] N. L Seaman. “Meteorological modeling for air-quality assessments”, *Atmospheric Environment*, vol. 34, no. 12–14, pp. 2231–2259. 2000.
[6] S. I. V. Sousa, F. G. Martins, M. C. Pereira, M. C. M. Alvim-Ferraz, H. Ribeiro, M. Oliveira and I. Abreu. “Influence of atmospheric ozone, PM_{10} and meteorological factors on the concentration of airborne pollen and fungal spores”, *Atmospheric Environment*, vol. 42, no. 32, pp. 7452–7464, ISSN: 1352-2310. 2008.
[7] Hamdy K. Elminir, “Dependence of urban air pollutants on meteorology”, *Science of The Total Environment*, vol 350, no. 1-3, pp. 225-237, ISSN: 0048-9697. 2005.
[8] M. A. Arain, R. Blair, N. Finkelstein, J. R. Brook, T. Sahsuvaroglu, B. Beckerman, L. Zhang and M. Jerrett. “The use of wind fields in a land use regression model to predict air pollution concentrations for health exposure studies”, *Atmospheric Environment*, vol. 41, no. 16, pp. 3453–3464, ISSN: 1352-2310, 2007.
[9] C. Mandurino and P. Vestrucci. “Using meteorological data to model pollutant dispersion in the atmosphere”, *Environ. Model. Softw.*, Elsevier Science Publishers B. V., vol. 24, no. 2, pp. 270–278, ISSN: 1364-8152, February, 2009.
[10] I. J. Turias, F. J. González, M. L. Martín and P. L. Galindo. “A competitive neural network approach for meteorological situation clustering”, *Atmospheric Environment*, vol. 40, no. 3, pp. 532 – 541, ISSN: 1352-2310, 2006.
[11] I. Nadir and A. Selici. “Investigating the impacts of some meteorological parameters on air pollution in Balikesir, Turkey”, *Environmental Monitoring and Assessment*, Springer Netherlands, ISSN: 0167-6369, vol. 140, no. 1, pp. 267–277, 2008.

[12] E. Demirci and B. Cuhadaroglu. "Statistical analysis of wind circulation and air pollution in urban Trabzon", *Energy and Buildings*, vol. 31, no. 1, pp. 49 – 53, ISSN: 0378-7788, 2000.

[13] M. W. Gardner and S. R. Dorling. "Artificial neural networks (the multilayer perceptron)–a review of applications in the atmospheric sciences", *Atmospheric Environment*, vol. 32, no. 14-15, pp. 2627 - 2636, 1998.

[14] M. G. Cortina-Januchs, J. M. Barrón-Adame, A. Vega-Corona and D. Andina. "Prevision of industrial SO_2 pollutant concentration applying ANNs", *7th IEEE International Conference on Industrial Informatics*, pp. 510-515, 2009.

[15] T. Kohonen. "The Self-Organizaing Map", *Proceedings of IEEE*, vol. 78, no. 9, pp. 1464-1480, 1990.

[16] D. Andina, A. Jevtić, A. Marcano and J.M. Barrón-Adame. "Error Weighting in Artificial Neural Networks Learning Interpreted as a Metaplasticity Model". *IWINAC '07: Proceedings of the 2nd international work-conference on The Interplay Between Natural and Artificial Computation, Part I*, pp. 244–252, ISBN: 978-3-540-73052-1, 2007.

[17] J. Vesanto and E. Alhoniemi. "Clustering of the self-organizing map", *IEEE Transactions on Neural Networks*, vol. 11, no. 3, pp. 586 –600. 2000.

[18] A. J. Richardson, C. Risien and F. A. Shillington. "Using self-organizing maps to identify patterns in satellite imagery", *Progress In Oceanography*, vol. 59, no. 2–3, pp. 223–239. DOI: 10.1016/j.pocean.2003.07.006, 2003.

[19] O. Ersoy, E. Aydar, A. Gourgaud, H. Artuner and H. Bayhan. "Clustering of volcanic ash arising from different fragmentation mechanisms using Kohonen self-organizing maps", *Computers Geosciences*. vol. 33, no. 6, pp. 821–828, DOI: 10.1016/j.cageo.2006.10.008. 2007.

[20] M. J. Watts and S. P. Worner. "Estimating the risk of insect species invasion: Kohonen self-organising maps versus k-means clustering", *Ecological Modelling*, vol. 220, no. 6, pp. 821 - 829. ISSN = 0304-3800. 2009.

[21] SEMARNAT, "Ministry of the Environment, Natural Resources and Fisheries", www.semarnat.gob.mx/Pages/inicio.aspx. 2008.

[22] INEGI, "National Institute of Geography and Statistics. Population and Housing Census 2, 2005," www.inegi.org.mx, 2005.

[23] INE, National Institute Ecology. General Direction for Research on the Urban and Regional Pollution. "Air Quality Research: Criteria Pollutants", www.ine.org.mx, 2009.

[24] EPA, Environmental Protection Agency. "Air Quality Index, AGuide to Air Quality and your helth". 2006.

[25] D. Andina and D. T. Pham, "Computational intelligence for engineering and manufacturing," 1st ed., Springer,2007.

[26] L. Hsin-Chung and F. Guor-Cheng. "Estimating the frequency distributions of PM_{10} and $PM_{2.5}$ by the statistics of wind speed at Sha-Lu, Taiwan", *The Science of The Total Environment*. vol. 298, no. 1-3, pp. 119 – 130, ISSN: 0048-9697. 2002.