Pollutant concentrations and Meteorological Data Classification by Neural Networks

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

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

***Abstract*—This paper present an environmental contingency forecasting tool based on Neural Networks (NN). Forecasting tool analyzes every hour and daily Sulphur Dioxide (SO_2) concentrations and Meteorological data time series. Pollutant concentrations and meteorological variables are self-organized applying a Self-organizing Map (SOM) NN in different classes. Classes are used in training phase of a General Regression Neural Network (GRNN) classifier to provide an air quality forecast. In this case a time series set obtained from Environmental Monitoring Network (EMN) of the city of Salamanca, Guanajuato, México is used. Results verify the potential of this method versus other statistical classification methods and also variables correlation is solved.**

I. Introduction

In polluted countries like México a continuos monitoring of the air quality to take forecast measures on possible negative effects in the population health is necessary. Air pollution is one of the most important environmental problems and is the result of human activities. Pollution has diverse causes and sources, such as industrial, commercial, agricultural and domestic activities. Combustion, used to generate heat, electricity or movement, is the process in which many pollutants are produced. Other activities like foundry and chemical production can induce to a deterioration of the air quality if it isn't controlled. Now a days, many first world countries make big efforts to minimize the effects of this activity (Kyoto) [1]. 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. A special case with great pollution is the city of Salamanca. A great quantity of industries are located in Salamanca, in many cases are chemical industries and also of electricity generation. A pollution alert has been recorded in last years in Salamanca, when in several times the Ecological Mexican Standard NOM-085, has been surpassed [2]. Nine years ago, an Environmental Monitoring Network (EMN) was installed in which time series about pollutant concentrations like Sulphur Dioxide ($\mathbf{SO_2}$) and Particulate Matter less than 10 micrometers in diameter $\mathbf{PM_{10}}$ among other meteorological variables are obtained. This article focuses the analysis on SO_2 concentrations.

A. Case of Study

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 [3]. In recent years, the city of Salamanca has been catalogued as one of the most polluted cities in Mexico [4]. 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.

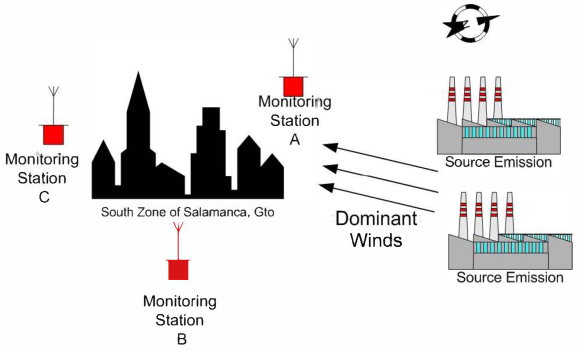

Fig. 1. Monitoring distribution

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 emissions when measurements of these pollutants register levels above those established by Health Authorities. Local Air Quality Index (AQI), provides daily information in a simple and uniform way on the air pollution concentration. The AQI is a value to inform at the population on the actions to reduce the air pollution or environmental forecasting. The AQI is a simple number into a scale from 0 to 500 [5]. Intervals in AQI scale related to the health concerns on the population are defined in Table I, and explained as follow:

Good: AQI units between 0 and 50 are considered satisfactory and the air pollution possesses little or few risk.
Moderate: AQI units between 51 and 100 are acceptable; However, for some pollutants it can have a health concern for a small number of population.
Unhealthy for Sensitive Groups: When AQI units are between 101 and 150, members of sensitive groups can suffer effects in their health.
Unhealthy: All the population can suffer dangerous effects in the health when AQI values are between 151 and 300.
Dangerous: For a superior AQI values of 300 a warning alarm is emitted for health conditions, the whole population will be more probably affected.

TABLE I
HEALTH LEVELS AND AIR QUALITY INDEX (AQI)

Health Concern Levels	
Air Quality	AQI values
Good	0 to 50
Moderate	51 to 100
Unhealthy for sensitive groups	101 to 150
Unhealthy	151 to 300
Dangerous	301 to 500

In the air quality evaluation for SO_2 concentrations, a daily mean estimation (24 hrs.) of 340 $\mu g/m^3$ (0.13 ppm) is considered and equivalent to 100 AQI units.

B. Pollutant concentrations and meteorological data

Clean air is a gassy mixture composed by Nitrogen (78%), Oxygen (21%), Argon, Carbon Dioxide, Ozone and other gases in small quantities (1%). Therefore, the atmospheric pollution can be defined as the emission of great quantities of substances that perturb the physical and chemical air properties. Pollutants are classified in primary and secondary. Primary pollutants are in the atmosphere when they are originally emitted by the source. Secondary pollutants are those that experience chemical changes as a result of the meteorological effects or combination with other pollutants (as photochemical oxidizers) and some radicals like Ozone. 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 [6], [7]. 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 [8], [9]. 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. [10].

II. MODEL AND THEORETICAL FUNDAMENT

The proposed method considers an automatic multivariate data analysis of time series obtained from EMN on monitoring points A, B and C geographically distributed in dominant winds direction and bigger population concentration (see Figure 1). The problem is to determine the correlation among all the variables involved in the decision making exercise on health risk for the population. Each monitoring point is considered like a sample point build with different sensors and also with its own perception field. Therefore, each point showed can be seen as a sensors fusion of where the time series are obtained. In problem solution a self-organized method that uses a Self-Organized Map (SOM) Neuronal Network has been proposed in order to build an automatic noise suppression method. Neural Networks (NN) are computational structures and they can learn from examples [5]. In some multi-dimensional engineering problems (like air pollution) is necessary to recognize certain patterns without the necessity of knowing of data nature or their statistical distribution. Some patterns recognition techniques apply NN to solve problems without the necessity of a prior data distribution knowledge or to make statistical suppositions. Other techniques in pattern recognition have the necessity to make statistical assumptions about data nature like Bayes theorem [11], [12]. Consequently, NN is an ideal tool to solve the problem here exposed due to their operation which is analyzed like a black box that minimizes the energy function [13], [14].

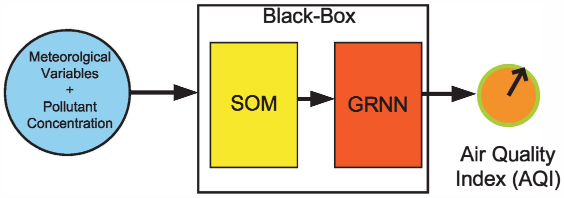

Fig. 2. Proposed Model to estimate the AQI

A. Variables definition

Variables definition is considered like normalized concentration values about pollutants and normalized meteorological values (Wind Speed, Temperature and Relative Humidity). In Table II, variables are defined in order to build a feature vector x_j and to define a pattern set $\mathbf{X_*} = \{\mathbf{x_1}, \mathbf{x_2}, .., \mathbf{x_j}, .., \mathbf{x_n}\}$. Let $\mathbf{X_{SO_2}}$ be a Sulphur Dioxide set concentration and their corresponding pattern is defined as $\mathbf{x_j} = \{x_1, x_2, x_3, x_4\}$.

TABLE II
VARIABLES DEFINITION: $\mathbf{SO_2}$ CONCENTRATION, $\mathbf{T}$; TEMPERATURE, $\mathbf{RH}$; RELATIVE HUMIDITY, $\mathbf{WS}$; WIND SPEED.

	$\mathbf{X_{SO_2}}$
Variables x_i	$\mathbf{x}_j$
x_1	SO_2
x_2	T
x_3	RH
x_4	WS

B. Proposed Model

1) Data Base and Pre-processing: In this work, a real and historical time series database from the EMN has been used. Data series of three months from December to February in three years from 2002 to 2005 have been analyzed. During Winter the pollutant and the meteorological conditions have major health concern. Time series consider a total of 6,480 multidimensional patterns about pollutant and meteorological variables. In Figure 3, a typical day data concentrations for $\mathbf{SO_2}$ and its correlation with meteorological variables (Temperature and Relative Humidity) is shown. In this figure it is possible to appreciate the complicated nature of this problem.

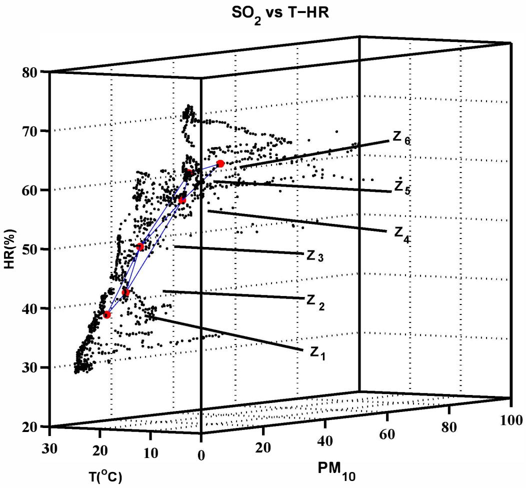

Fig. 3. Graph with $\mathbf{SO_2}$ correlation and Meteorological variables, and Self-Organized Map.

2) Clustering Method: In this research a prior knowledge about patterns is unavailable. Therefore, in the classifier design, the pattern classes obtained applying the clustering method are used. Unsupervised learning is popularly adopted in data clustering where a prior class information is unavailable. SOM Neural Network is good for mapping similar patterns in a high dimension feature space to a much lower dimension output map while preserving the topological order.

Due to data nature is unknown a Self-Organized method has been proposed. In order to group the patterns in six classes according to health concern levels and noisy patterns (as shown in Table III), a SOM Neural Network is apply as it is shown in Figure 4. The idea is to build different training pattern sets in order to design a classifier based on a NN. A SOM structure with euclidian distance function and hexagonal topology and 3:2:1:1 structure has been proposed [15]. In order to have six clusters and therefore six prototypes or weights ($\mathbf{Z}_i$), one per class or AQI (as shown in Figure 3), which have a representation that involved variables in NN training phase and which was mentioned in Section II-B1.

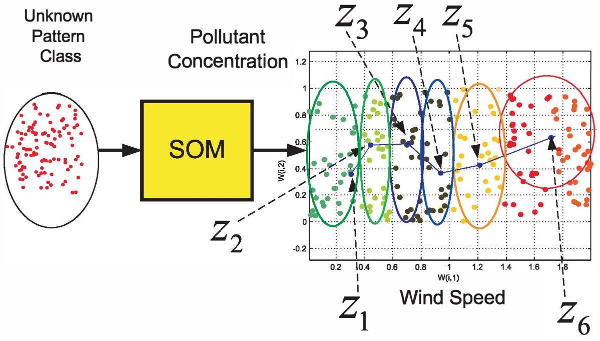

Fig. 4. Self-Organized Neural Network

In Table III, the classification for each health concern level of $\mathbf{SO_2}$ in function of their AQI intervals and their corresponding class $\mathbf{Z}_i$ (or category in a Self-Organized Map) are shown. Therefore, the center for each class is build as $\mathbf{Z}_i$=$\{\mu_{1i}$,μ_{2i},μ_{3i},$\mu_{4i}\}$, where $\mathbf{Z_i}$ is the class center i of each type of AQI pattern, μ_{1i} is the pollutant concentration level prototype ($\mathbf{SO_2}$ concentration if the pattern belong to $\mathbf{X_{SO_2}}$ set), μ_{2i} is the Temperature prototype, μ_{3i} is the Relative Humidity prototype and μ_{4i} is the Wind Speed prototype.

TABLE III
HEALTH CONCERN LEVELS RESPECT TO AIR QUALITY INDEX AND THEIR CATEGORY MAP REPRESENTATION FOR $\mathbf{SO_2}$ AND $\mathbf{PM_{10}}$

Index Classification levels for $\mathbf{SO_2}$		
Air Quality	AQI	Cluster Prototype
Good	0 to 50	$\mathbf{Z_1}$
Moderate	51 to 100	$\mathbf{Z_2}$
Unhealthy for sensitive groups	101 to 150	$\mathbf{Z_3}$
Unhealthy	151 to 300	$\mathbf{Z_4}$
Dangerous	301 to 500	$\mathbf{Z_5}$
* **Noise**	————-	$\mathbf{Z_6}$

3) Classifier Design: In the classifier design, the pattern classes of the clusters with center $\mathbf{Z}_i$, obtained by means of clustering method are used for this purpose. A General Regresión Neural Network (GRNN) with clustering structure which is shown in Figure 5, is trained to obtain a continuos estimation for the AQI in two models. A first model is trained for $\mathbf{SO_2}$ concentrations set ($\mathbf{X_{SO_2}}$). The GRNN was introduced by Donald F. Specht [16]. The main advantage of GRNN over a Multi-Layer Perceptron (MLP) is that, unlike the MLP which need a larger number of iterations to be performed in training phase to converge to a desired solution, the GRNN needs only a single learning pass to achieve optimal performance in classification.

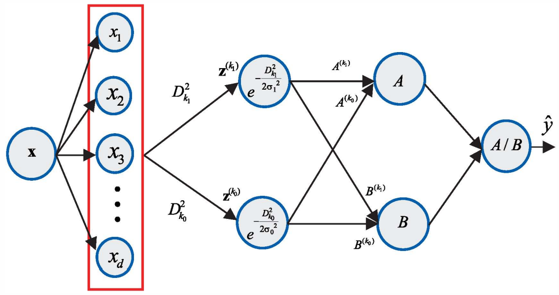

Fig. 5. GRNN cluster structure

In general, the GRNN operation is described. Let x be a feature vector and y be a scalar and $f(\mathrm{x}, y)$ the joint probability density function (**pdf**) of x and y. The expected value of y given x is defined as

$$E[y \mid \mathrm{x}] = \frac{\int_{-\infty}^{\infty} y f(\mathrm{x}, y) dy}{\int_{-\infty}^{\infty} f(\mathrm{x}, y) dy} \tag{1}$$

the **pdf** is unknown, therefore it must be estimated from sample values of x_i and y_i from a kernel function estimator proposed by Parzen, see [16]. The estimator is defined as a reduced gaussian kernel $\exp(-\frac{D_i^2}{2\rho^2})$. Thus, is possible to obtain a discrete conditional mean of y given x or an estimation of $\hat{y}$ as,

$$E[y \mid \mathrm{x}] = \hat{y}(\mathrm{x}) = \frac{\sum_{i=1}^{n} y_i \exp(-\frac{D_i^2}{2\rho_i^2})}{\sum_{i=1}^{n} \exp(-\frac{D_i^2}{2\rho_i^2})} \tag{2}$$

where ρ_i is the kernel width, n is the number of all the patterns in the $\mathbf{Z}_i$ clusters and D_i is the euclidian distance among the input pattern and the $i-$th training pattern or ${D_i}^2 = (\mathrm{x} - \mathrm{x}_i)^T(\mathrm{x} - \mathrm{x}_i)$. The GRNN operation is simple, the input layer simply passes the patterns x to all units in the hidden layers composed by kernels functions $\exp(-\frac{D_i^2}{2\rho^2})$ and computes the squared distances among the new pattern x and x_i training samples; the hidden-to-output weights are just the targets y_i, thus the output $\hat{y}(\mathrm{x})$, is simply a weighted average of the target values y_i of the training cases x_i close to the given input case x. The only parameters ρ that need to be learned are adjusted using the algorithm proposed for us in [17].

In some problems like this research, the number of observations obtained can be sufficiently large that it is no longer practical to assign a separate node (or neuron) to each ith sample. Clustering method is used to group samples so that the group can be represented by only one node or prototype $\mathbf{Z}_i$ that measures distance of the input vector x from the cluster center $\mathbf{Z}_i$. However the cluster prototypes are determined, let us assign a new variable, m, to indicate the number of samples that are represented by the *i*th cluster center $\mathbf{Z}_i$. The estimation equation can then be rewritten as

$$\hat{y}(\mathbf{x}) = \frac{\sum_{i=1}^{m} A_i \exp(-\frac{D_i^2}{2\rho^2})}{\sum_{i=1}^{m} B_i \exp(-\frac{D_i^2}{2\rho^2})} \tag{3}$$

and

$$\begin{cases} A_i(k) = A_i(k-1) + y_j \\ B_i(k) = B_i(k-1) + 1 \end{cases} \tag{4}$$

where $m < n$ is the number of clusters, and $A_i(k)$ and $B_i(k)$ are the values of the coefficients for cluster *i* after *k* observations. $A_i(k)$ is the sum of the y_i values and $B_i(k)$ is the number of samples assigned to cluster i. The method of clustering can be as simple as establishing a single radius of influence, r. Starting with the first sample point $(\mathbf{x}_i, y_i)$, establish a cluster center, $\mathbf{x}_i$ at $\mathbf{x}$. All future samples for which the distance $|\mathbf{x} - \mathbf{x}_i|$, is less than the distance to any other cluster center and is also $\leq r$ would update equations (4) for this cluster. A sample for which the distance to the nearest cluster is $> r$ would become the center for a new cluster. Note that the numerator and denominator coefficients are completely determined in one pass through the data and no iteration is required to improve the coefficients [16].

4) Classifier Optimization: A problem in the estimations based in a GRNN is the adjustment of Perception Parameter (PP) of the neurons. Perception parameter is controlled by the ρ parameter to obtain a minimum classification error. It is solved using a multidimensional gradient algorithm and has been proposed by authors in [17].

III. Results

A. Data Clustering for $\mathbf{SO}_2$ concentrations

In the experiments, a real time series data base for $\mathbf{SO}_2$ concentrations have been analyzed every minute. The meteorological variables used were Wind Speed, Temperature and Relative Humidity, creating 6,480 four dimensional pattern vectors x_i, for the SO_2 set pollutant ($\mathbf{X}_{\mathbf{SO}_2}$), as is shown in Table II. Both pollutant concentrations like meteorological variables are provided by the EMN from Salamanca. In the clustering method, six clusters have been performed from time series of years 2002, 2003, 2004. The clusters centers $\mathbf{Z}_i$ (or weights in the SOM) are analyzed according its features to determine the label or class of each prototype like in Table III. Each prototype $\mathbf{Z}_i$, is used to build a GRNN in a cluster structure. Table III, also shows the classification for each AQI level for $\mathbf{SO}_2$ concentrations in a Self-Organized Neural Network, where the center for each class is $\mathbf{Z}_i$ and as well is defined like $\mathbf{Z}_i$=$\{\mu_{1i}$,μ_{2i},μ_{3i},$\mu_{4i}\}$, where, $\mathbf{Z}_i$ is the class center i, μ_{1i} is the $\mathbf{SO}_2$ concentration level according set, μ_{2i} is the Temperature, μ_{3i} is the Relative Humidity and μ_{4i} is Wind Speed.

B. AQI estimation for $\mathbf{SO}_2$ and Meteorological variables

The complicated interpretation of the correlation among $\mathbf{SO}_2$ pollutant concentrations and meteorological variables like Temperature (T) and Relative Humidity (RH) is shown in Figure 3. In contrast, in Figure 7, is easy to appreciate and to classify the health concern levels of $\mathbf{SO_2}$ concentrations. Figure 7, shows that the complexity is minimized using the classes or categories related to each prototype $\mathbf{Z}_i$. In this case a GRNN estructure has been applied. GRNN is trained using the cluster prototypes $\mathbf{Z}_i$. Time series from 2005 in AQI estimation (direct mode of GRNN) have been used. When a pattern is presented in the GRNN input the AQI estimation it is immediate and is necessary only one pass. When a noisy pattern $\mathbf{x}$ is detected, the GRNN output is inhibited. Noise suppression is an innovation in this research because in the current method it is not considered.

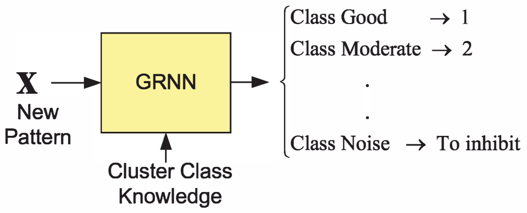

Fig. 6. GRNN operation in direct phase

Noisy pattern is an inconsistent element in the time series and it is caused by blasts of wind. Noisy elements can cause bad estimates about AQI, so with this method a better estimate is obtained. The GRNN operation in direct phase is shown in Figure 6.

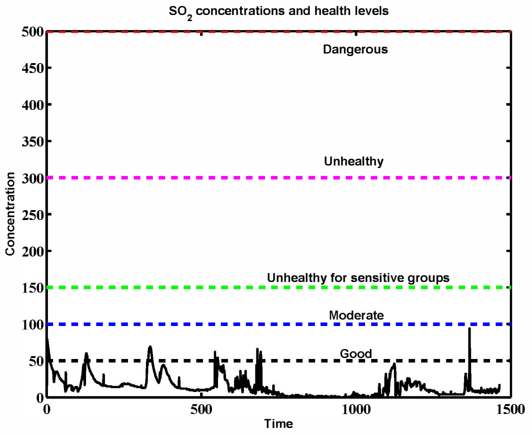

Fig. 7. AQI estimation for $\mathbf{SO_2}$ concentrations.

The complexity is solved with this method using the classes corresponding to $\mathbf{Z}_i$.

IV. Conclusions

In this work a vector set for $\mathbf{SO_2}$ pollutant concentrations and meteorological variables has been built. In order to solve the AQI estimation considering $\mathbf{SO_2}$ pollutant concentrations and meteorological variables a combination between Self-Organized Neuronal Network (SOM) and a General Regression Neural Network (GRNN) has been proposed. Both Neural Networks were trained and proven with multidimensional patterns of pollutants and meteorological variables. A correlation problem solution is given and results show the easy interpretation using the discrete classes for the AQI estimation. Concluding that the representation of values shown in Figure 7, allow to classify the patterns in function of their prototypes in a simple way than the multidimensional representation, therefore it is a good tool for making decisions in environmental forecasting.

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

[2] Instituto Nacional de Ecología. "Normas oficiales mexicanas para la protección ambiental". http://www.ine.gob.mx/ueajei/publicaciones/normas/, Diciembre 1994.

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

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

[5] S. Haykin, "Neural Networks". *Prentice Hall, 2nd edition*, 1999.

[6] M. G. Cortina-Januchs, J. M. Barron-Adame, A. Vega-Corona and D. Andina, "Prevision of industrial $\mathrm{SO_2}$ pollutant concentration applying ANNs" *IEEE 7th International Conference on Industrial Informatics*, pp. 510 – 515, ISSN: 1935-4576, 2009.

[7] 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.

[8] M. A. Arain, R. Blair, N. Finkelstein, J. R. Brook, T. Sahsuvaroglu, B. Beckerman, L. Zhang and Jerrett, M., "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] 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.

[10] 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.

[11] R. O. Duda and P. E. Hart. "Pattern Classification and Scene Analysis". *John Wiley and Sons, Inc., 2nd edition*, 1973.

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

[13] R. P. Lippmann. "An introduction to computing with neural net". *IEEE ASSP Magazine*, pages 4–22, April 1987.

[14] D. Andina and A.Vega. "Detection of microcalcifications in mammograms by the combination of a neural detector and multiscale feature enhancement". *Bio-Inspired Applications of Connectionism. Lecture Notes in Computer Science. Springer-Verlag*, 2085:385–392, 2001.

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

[16] D. F. Specht. "A general regression neural network". *IEEE Transactions on Neural Networks*, 2(6):568–576, 1991.

[17] F. S. Buendía Buendía., J. M. Barrón-Adame, A. Vega-Corona, and D. Andina. "Improving grnns in cad systems". *Proceedings on Independent Component Analysis and Blind Signal Separation: Fifth International Conference*, ICA 2004. Granada, Spain. Springer-Verlag.