4

Air Pollution Analysis with a Possibilistic and Fuzzy Clustering Algorithm Applied in a Real Database of Salamanca (México)

B. Ojeda-Magaña,[1] R. Ruelas[1], L. Gómez-Barba[1], M. A. Corona-Nakamura[1], J. M. Barrón-Adame[2], M. G. Cortina-Januchs[2], J. Quintanilla-Domínguez[2] and A. Vega-Corona[2]

[1]*University of Guadalajara*
[2]*University of Guanajuato*
México

1. Introduction

Air pollution is one of the most important environmental problems in developed and undeveloped countries and it is associated with significant adverse health effects. Air pollution is characterized by the presence of a heterogeneous, complex mixture of gases, liquids and particulate matter in air. Pollution is caused by both natural and man-made sources, and it may greatly vary from one region to another according to the geography, demography, climate, and topography of these ones. For example, pollutant concentrations decrease significantly when the urban area meets certain characteristics as topography or large rain season (Celik & Kadi, 2007). Forest fires, volcanic eruptions, wind erosion, pollen dispersal, evaporation of organic compounds, and natural radioactivity are among natural causes of air pollution. Major man-made sources of air pollution include: industries, transportation, agriculture, power generation, and unplanned urban areas (Fenger, 2009).

Air pollutants exert a wide range of impacts on biological, physical, and ecosystems. Their effects on human health are of particular concern. The World Health Organization (WHO) consider air pollution as the mayor environmental risk to health and is estimated to cause approximately 2 million premature deaths worldwide per year (WHO, 2008).

This type of pollution is classified in criterio and non-criterio pollutants, the firsts are considered dangerous to human and animal health, its name was given after the result of various evaluations regarding air pollution published by the United States of America (EPA, 2008). Six criteria of pollutants are defined: Nitrogen Dioxide (NO_2), Sulfur Dioxide (SO_2), Carbon Monoxide (CO), Particulate Matter (PM), Lead (Pb), and Ozone (O_3). The objective of this classification is to establish permissible levels to protect human and animal health and for the preservation of the environment. Human health is one of the most important concerns due to the short-term consequences of air pollution, especially in metropolitan areas, health effects are dependent on the type of pollutant, its concentration in air, length of exposure to the pollutant and individual susceptibility. Several groups of individuals react differently to air pollution, Children and elderly people are the most affected by this kind of pollution. Global warming and the greenhouse effect are among long term consequences of the global climate.

Examine and study air pollutant information is very important for a better understanding of the human exposure and its potential impacts in health and welfare.
In recent years, the city of Salamanca has been catalogued as one of the most polluted cities in Mexico (Zuk et al., 2007). Sulphur Dioxide (SO_2), and Particular Matter (PM_{10}) are the criteria for searching air pollutants with the highest concentration in Salamanca, where three monitoring stations have been installed in order to know the level of air pollution; measure records of each monitoring station are handled separately. Actually an environmental contingency alarm is activated when the daily average pollutant concentration exceeds an established threshold (in a single monitoring station).
In this work, we propose to apply the PFCM (*Possibilistic Fuzzy c Means*) clustering algorithm to the measured data obtained from three monitoring stations so that a local environmental contingency alarm can be taken, according to the pollutant concentration reported by each monitoring station, general (or city) environmental contingency alarms will depend on the levels provided by the combined measure. So, the PFCM algorithm is used to find the prototypes of patterns that represent the relation between SO_2 and PM_{10} air pollutants. For this relation analysis we use records from January 2007.
Once the prototypes have been estimated, a comparison is made between the average pollution of each monitoring station and the prototypes. In the analysis is used a data set from January to December 2007. The analysis include pollutant concentration as SO_2 , PM_{10}, meteorological variables, wind speed, wind direction, temperature, and relative humidity.
It is also analyzed the impact of meteorological variables on the dispersion of pollutants, this is done through the calculus of correlation coefficients. This important correlation analysis is very simple and it is intended for improving decision making in environmental programs. Only the data gathered by the *Nativitas monitoring* station is used for the correlation analysis.
This paper is organized as follow: In Section 2 is presented the features, and explain the air pollution problem in Salamanca. In Section 3 is introduced the PFCM (*Possibilistic Fuzzy c Means*) clustering algorithm and the correlation coefficients. Section 4 presents the obtained results. And finally, in Section 5 we present our conclusions.

2. Study case

Salamanca is located in the state of Guanajuato, Mexico, and it has an approximate population of 234,000 inhabitants INEGI (2005). The city is 340 km northwest from Mexico City, with coordinates 20°34'22" North latitude, and 101°11'39" West longitude. It is located on a valley surrounded by the *Sierra Codornices*, where there are elevations with an average height of 2,000 meters Above Mean Sea Level (AMSL).
Salamanca has been one of the Mexican cities with more important industrial development in the last fifty years. Refinery and Power Generation Industries settled down in the fifty and seventy decades, respectively. These industries constitute the main and most important energy source for local, regional and national economy. However, the increase of population, quantity of vehicles, and the industry, refinery and thermoelectric activities, as well as orography and climatic characteristics have propitiated the increment in SO_2 and PM_{10} concentrations INE (2004). The existent orography difficults the dispersion of pollutants by the wind, which produces the worst pollutant concentrations. SO_2 emissions are bigger than those in the Metropolitan area of Mexico City or Guadalajara city, the two biggest cities of Mexico, even when these ones have a bigger population than the city of Salamanca Cortina-Januchs et al. (2009). Orography hinders the dispersion of the worst pollutants by winds.

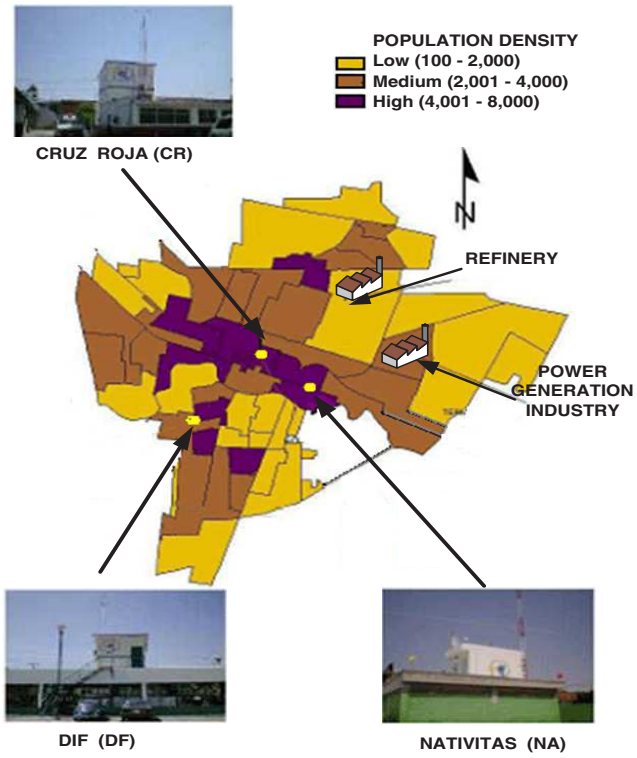


Fig. 1. Location of monitoring stations in the city of Salamanca.

Sulfur dioxide is produced fundamentally by the combustion of fossil fuels, and it has the energy generation sector as the main source of pollution. That is, the industrial sector generates 99.3 % of this pollutant, and only an approximate percentage of 0.06 % is generated by the transport sector. Particles produced by electric power generation represent 29 % of the total emissions, it follows the vehicular traffic in the roads without paving with 27 %, next the agriculture burns with 17 %, transport sector with 10 %,and the remaining 17 % is emitted by other sub-sectors.

Authorities of the city have made important efforts to measure and record on concentrations of pollutants Zamarripa & Sainez (2007). In 1999 the *Air Quality Monitoring Patronage* (AQMP) was formed. Since then the AQMP has been in charge of running the *Automatic Environmental Monitoring Network* (AEMN), and disseminate information. This information is validated by the *Institute of Ecology* (IE), which constantly analyzes the levels of pollutants INE (2004). The AEMN consists of three fixed and one mobile stations. The fixed stations are: *Cruz Roja* (CR), *Nativitas* (NA), and *DIF*.

The fixed stations cover approximately 80 % of the urban area while the mobile station covers the remaining 20 %. Fig. 1 illustrates the location of the three fixed stations. Each station has the necessary instrumentation to automatically track concentration of pollutants and meteorological variables every minute. Table 1 contains a sample of the concentration of pollutants and meteorological variables in each of the three fixed stations.

Pollutants

	Cruz Roja	Nativitas	DIF
Ozone (O_3)	√	√	√
Sulfur Dioxide(SO_2)	√	√	√
Carbon Monoxide (CO)	√	√	√
Nitrogen Dioxide (NO_x)	√	√	√
Particulate Matter less than 10 micrometer in diameter (PM_{10})		√	√

Meteorological variables

	Cruz Roja	Nativitas	DIF
Wind Direction (WD)	√	√	√
Wind speed (WS)	√	√	√
Temperature (T)		√	√
Relative Humidity (RH)		√	√
Barometric Pressure (BP)		√	√
Solar Radiation (SR)		√	√

√ Measured

Table 1. Pollutants concentrations and meteorological variables recorder in the monitoring stations

3. Clustering algorithms

In this work we take advantage of the qualities of fuzzy and possibilistic clustering algorithms in order to find c groups in a set of unlabeled data set $Z = \{z_1, z_2, \ldots, z_k, \ldots, z_N\}$ in an M-dimensional space, where the nearest z_k to a prototype, or group center v_i, belong to the group i among c possible groups. The membership of each z_k to the different groups depends on the kind of partition of the M-dimensional space where data set is defined. This way, a *c-partition* can be either: hard (or crisp), fuzzy, and possibilistic Bezdek et al. (1999). The hard c-partition of the space for a data set $Z(k) = \{z_k | k = 1, 2, ..., N\}$, of finite dimension and c groups, where $2 \leq c < N$, is defined by (1), (2) defines the fuzzy c-partition, whereas (3) defines the possibilistic c-partition.

$$M_{hcm} = \Big\{ \mathbf{U} \in \Re^{c \times N} | \mu_{ik} \in \{0,1\}, \forall i \ and \ k; \\ \sum_{i=1}^{c} \mu_{ik} = 1, \forall k; \quad 0 < \sum_{k=1}^{N} \mu_{ik} < N, \forall i \Big\}; \tag{1}$$

$$M_{fcm} = \Big\{ \mathbf{U} \in \Re^{c \times N} | \mu_{ik} \in [0,1], \forall i \ and \ k; \\ \sum_{i=1}^{c} \mu_{ik} = 1, \forall k; \quad 0 < \sum_{k=1}^{N} \mu_{ik} < N, \forall i \Big\}; \tag{2}$$

$$M_{pcm} = \Big\{ \mathbf{U} \in \Re^{c\times N} | \mu_{ik} \in [0,1], \forall i \ and \ k; \\ \forall k, \exists i, \mu_{ik} > 0; \quad 0 < \sum_{k=1}^{N} \mu_{ik} < N, \forall i \Big\}. \tag{3}$$

3.1 Fuzzy c-Means algorithm

The Fuzzy *c*-Means clustering algorithm (FCM) was initially developed by Dunn Dunn (1973), and generalized later by Bezdek Bezdek (1981). This algorithm is based on the optimization of the objective function given by (4),

$$J_{fcm}(\mathbf{Z}; \mathbf{U}, \mathbf{V}) = \sum_{i=1}^{c} \sum_{k=1}^{N} (\mu_{ik})^m \|z_k - v_i\|^2, \tag{4}$$

where the membership matrix $U = [\mu_{ik}] \in M_{fmc}$, is a fuzzy c-partition of the space where Z is defined, $V = [v_1, v_2, ..., v_c]$ is the vector of prototypes of the c groups, which are calculated according to $D_{ikA_i} = \|z_k - v_i\|^2$, a squared inner-product distance norm, and $m \in [1, \infty]$ is a weighting exponent which determines the fuzziness of the partition. The optimal c-partition for a Fuzzy c-Means algorithm, is reached through the couple (U^*, V^*) which minimizes locally the objective function J_{fcm}, according to the *alternating optimization* (AO).

Theorem FCM Bezdek (1981): If $D_{ikA_i} = \|z_k - v_i\| > 0$, for every $i, k, m > 1$, and Z contains at least c distinct data points, then $(U, V) \in M_{fcm} \times \Re^{c\times N}$ may minimize J_{fcm} only if

$$\mu_{ik} = \left(\sum_{j=1}^{c} \left(\frac{D_{ikA_i}}{D_{jkA_i}} \right)^{2/(m-1)} \right)^{-1} \tag{5}$$

$$1 \le i \le c; \quad 1 \le k \le N$$

$$v_i = \sum_{k=1}^{N} \mu_{ik}^m z_k \Big/ \sum_{k=1}^{N} \mu_{ik}^m \tag{6}$$

$$1 \le i \le c.$$

Following the previous equations of the FCM algorithm, the solution can be reached with the next steps:

FCM-AO-V

Given the data set Z choose the number of clusters $1 < c < N$, the weighting exponent $m > 1$, as well as the ending tolerance $\delta > 0$.

I Provide an initial value to each one of the prototypes $v_i, i = 1, .., c$. These values are generally given in a random way.

II Calculate the distance of z_k to each one of the prototypes v_i, using $D^2_{ikA_i} = (z_k - v_i)^T A_i (z_k - v_i)$, $1 \le i \le c$, $1 \le k \le N$.

III Calculate the membership values of the matrix $U = [\mu_{ik}]$, if $D_{ik\mathbf{A}} > 0$, using equation (5).

IV Update the new values of the prototypes v_i using equation (6).

V Verify if the error is equal or lower than δ,

$$\|V_{k+1} - V_k\|_{err} \leq \delta,$$

If this is truth, stop. Else, go to step **II**.

The FCM is an algorithm that calculates a membership value μ_{ik} for each point z_k in function of all prototypes v_i. The sum of the membership values of z_k to the c groups must be equal to one. However, a problem arises when there are several equidistant points from the prototypes of the groups, because the FCM is not able to detect noise points or nearest and furthest points from the prototypes. Pal *et al* Pal et al. (2004) show an example with two points located in the boundary of two groups, one point near to the prototypes and the other one far away from them. This must be handled with care, as both points are not *equally representative* of the groups, even if they have the same membership values. One way to overcome this inconvenience is to use a possibilistic algorithm.

3.2 Possibilistic c-Means algorithm

The Possibilistic c-Means clustering algorithm (PCM) Krishnapuram & Keller (1993) is based on *typicality values* and relaxes the constraint of the FCM concerning the sum of membership values of a point to all the c groups, which must be equal to one. Thus, the PCM identifies the similarity of data points with an alone prototype v_i using a typicality values that takes values in [0,1]. The nearest data points to the prototypes are considered *typical*, further data points are *atypical* and data points with zero, or almost zero, typicality values are considered *noise* Ojeda-Magaña et al. (2009a). The objective function J_{pcm} proposed by Krishnapuram Krishnapuram & Keller (1993) for this algorithms is given by

$$J_{pcm}(\mathbf{Z};\mathbf{T},\mathbf{V},\gamma) = \left\{ \sum_{k=1}^{N}\sum_{i=1}^{c}(t_{ik})^m \|z_k - v_i\|_A^2 + \sum_{i=1}^{c}\gamma_i \sum_{k=1}^{N}(1-t_{ik})^m \right\}, \tag{7}$$

where

$$T \in M_{pcm}, \qquad \gamma_i > 0, \qquad 1 \leq i \leq c. \tag{8}$$

The first term of J_{pcm} is identical to that of the FCM objective function, which is based on the distance of the points to the prototypes. The second term, that includes a penalty γ_i, tries to bring t_{ik} toward 1.

Theorem PCM Krishnapuram & Keller (1993): if $\gamma_i > 0, 1 \leq i \leq c$, $m > 1$ and Z has at least c distinct data points, then $(T, V) \in M_{pcm} \times \Re^{c\times N}$ may minimize J_{pcm} only if

$$t_{ik} = \frac{1}{1 + \left(\frac{\|z_k - v_i\|^2}{\gamma_i}\right)^{1/(m-1)}}, \tag{9}$$

$$1 \leq i \leq c; \qquad 1 \leq k \leq N$$

$$v_i = \sum_{k=1}^{N} t_{ik}^m z_k \Big/ \sum_{k=1}^{N} t_{ik}^m, \tag{10}$$

$$1 \leq i \leq c; \qquad 1 \leq k \leq N.$$

Krishnapuram and Keller Krishnapuram & Keller (1993) Krishnapuram & Keller (1996) recommend to apply the FCM at a first time, such that the initial values of the PCM algorithm can be estimated. They also suggest the calculus of the penalty γ_i with equation (11)

$$\gamma_i = K \frac{\sum_{k=1}^{N} \mu_{ik}^m \|z_k - v_i\|_A^2}{\sum_{k=1}^{N} \mu_{ik}^m} \tag{11}$$

where $K > 0$, although the most common value is $K = 1$, and the membership values $\{\mu_{ik}\}$ are those calculated with the FCM algorithm in order to reduce the influence of noise.

The PCM algorithm is very sensitive to the $\{\gamma_i\}$ values, and the typicality values depend directly on it. For example, if the value of γ_i is small, the typicality values t_{ik} of T are also small, whereas if the value of γ_i is high, the t_{ik} are also high. For this work, the $\{\gamma_i\}$ values are obtained from equation (11).
In order to avoid a problem with the initial PCM algorithm, as sometimes the prototypes of different groups coincided Hoppener et al. (2000), even if the natural structure of data has well delimited different groups, Tim *et al* Timm et al. (2004); Timm & Kruse. (2002) have modified the objective function to include a constraint based on the repulsion among groups, thus avoiding identical groups when they must be different.
The objective of the fuzzy clustering algorithms is to find an internal structure in a numerical data set into n different subgroups, where the members of each subgroup have a high similarity with its prototype (centroid, cluster center, signature, template, code vector) and a high dissimilarity with the prototypes of the other subgroups. This justifies the existence of each one of the subgroups Andina & Pham (2007).
A simplified representation of a numerical data set into n subgroups, help us to get a better comprehension and knowledge of the data set Barron-Adame et al. (2007). Besides, the particional clustering algorithms (hard, fuzzy, probabilistic or possibilistic) provide, after a learning process, a set of prototypes as the most representative elements of each subgroups.
Ruspini was the first one to use fuzzy sets for clustering Ruspini (1970). After that, Dunn Dunn (1973) developed in 1973 the first fuzzy clustering algorithm, named Fuzzy c-Means (FCM), with a parameter of fuzziness m equal to 2. Later on Bezdek Bezdek (1981) generalized this algorithm. The FCM is an algorithm where the membership degree of each point to each fuzzy set A_i is calculated according to its prototype. The sum of all the membership degrees of each individual point to all the fuzzy sets must be equal to one.
Krishnapuram and Keller Krishnapuram & Keller (1993) developed the Possibilistic c-Means (PCM) clustering algorithm, where the principal characteristic is the relaxation of the restriction that gives the relative typicality property of the FCM. The PCM provides a similarity degree between data points and each one of the prototypes, value known as absolute typicality or simply typicality Pal et al. (1997). So, the nearest points to a prototype are identified as typical, whereas the furthest points as atypical, and noise Ojeda-Magaña et al. (2009a)Ojeda-Magaña et al. (2009b).

3.3 PFCM clustering algorithm

Pal *et al*. Pal et al. (1997) have proposed to use the membership degrees as well as the typicality values, looking for a better clustering algorithm. They called it *Fuzzy Possibilistic c-Means* (FPCM). However, the sum equal to one of the typicality values for each point was the origin of a problem, particularly when the algorithm uses a lot of data. In order to avoid this problem, Pal *et al* Pal et al. (2005) proposed to relax this constraint and they developed the PFCM clustering algorithm, where the function to be optimized is given by (12)

$$J_{pfcm}(\mathbf{Z};\mathbf{U},\mathbf{T},\mathbf{V}) = \sum_{i=1}^{c}\sum_{k=1}^{N}(a\mu_{ik}^{m} + bt_{ik}^{\eta}) \times \|z_k - v_i\|^2 + \sum_{i=1}^{c}\gamma_i \sum_{k=1}^{N}(1-t_{ik})^{\eta}, \tag{12}$$

and subject to the constraints $\sum_{i=1}^{c}\mu_{ik} = 1 \forall k$; $0 \leq \mu_{ik}, t_{ik} \leq 1$ and the constants $a > 0$, $b > 0$, $m > 1$ and $\eta > 1$. The parameters a and b define a relative importance between the membership degrees and the typicality values. The parameter μ_{ik} in (12) has the same meaning as in the FCM. The same happens for the t_{ik} values with respect to the PCM algorithm.

emphTheorem PFCM Pal et al. (2005): If $D_{ikA} = \|z_k - v_i\| > 0$, for every $i, k, m, \eta > 1$, and Z contains at least c different patterns, then $(U, T, V) \in M_{fcm} \times M_{pcm} \times \Re^p$ and J_{pfcm} can be minimized if and only if

$$\mu_{ik} = \left(\sum_{j=1}^{c} \left(\frac{D_{ikA_i}}{D_{jkA_i}} \right)^{2/(m-1)} \right)^{-1} \tag{13}$$
$$1 \leq i \leq c; \quad 1 \leq k \leq n$$

$$t_{ik} = \frac{1}{1 + \left(\frac{b}{\gamma_i} D_{ik_{A_i}}^2 \right)^{1/(\eta-1)}} \tag{14}$$
$$1 \leq i \leq c; \quad 1 \leq k \leq n$$

$$v_i = \sum_{k=1}^{N}(a\mu_{ik}^{m} + bt_{ik}^{m})z_k \Big/ \sum_{k=1}^{N}(a\mu_{ik}^{m} + bt_{ik}^{m}), \tag{15}$$
$$1 \leq i \leq c.$$

The membership degrees are calculated with equation (13), the typicality values with (14) and for the prototypes the equation (15) is used.

The iterative process of this algorithm follows the next steps:

PFCM-AO-V

Given the data set Z choose the number of clusters $1 < c < N$, the weighting exponents $m > 1$, $\eta > 1$, and the values of the constants $a > 0$, and $b > 0$.

I Provide an initial value to each one of the prototypes $v_i, i = 1, .., c$. These values are generally given in a random way.

II Run the FCM-AO-V algorithm.

III With these results, calculate the penalty parameter γ_i for each cluster i. Take $K = 1$.

IV Calculate the distance of z_k to each one of the prototypes v_i using $D^2_{ikA_i} = (z_k - v_i)^T A_i (z_k - v_i)$, $1 \leq i \leq c$, $1 \leq k \leq N$.

V Calculate the membership values of the matrix $U = [\mu_{ik}]$ if $D_{ik\mathbf{A}} > 0$, use equation (13).

VI Calculate the typicality values of the matrix $T = [t_{ik}]$, if $D_{ik\mathbf{A}} > 0$, use equation (14).

VII Update the value of the prototypes v_i using equation (15).

VIII Verify if the error is equal or lower than δ,

$$\|V_{k+1} - V_k\|_{err} \leq \delta,$$

if this is truth, stop. Else, go to step **IV**.

3.4 PFCM clustering algorithm in the AEMN

As it is known, in the partition clustering algorithms is necessary a minimum of two groups. However, in our problem we only have one group, this group is formed by patterns $[SO_2;PM_{10}]$ pollutant concentrations. Therefore, is proposed a synthetic cloud of patterns with the following covariance matrix and vector of centers:

$$\Sigma_1 = \begin{bmatrix} 400 & 0 \\ 0 & 400 \end{bmatrix}, \quad [v_1] = [100 \ -600].$$

In this case, the number of patterns (4320) is the same in the synthetic cloud and the pollutant concentration.

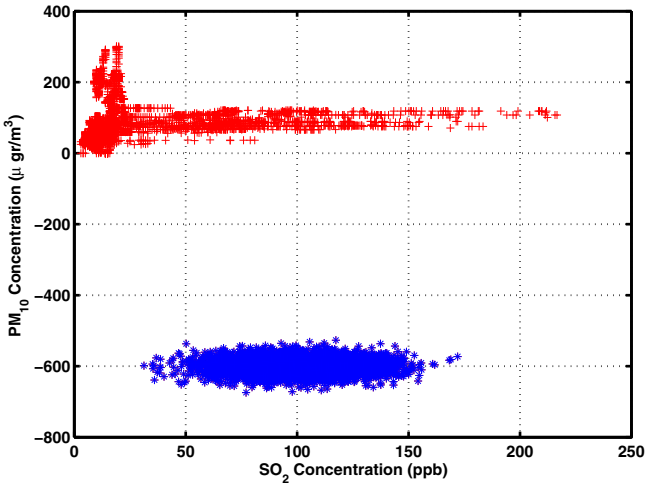


Fig. 2. Air pollution and synthetic cloud patterns.

Fig. 2 shows clearly the synthetic cloud (located in the lower part) and the pollutant concentration patterns (located in the superior part). Once the groups are identified, we apply the PFCM clustering algorithm.

3.5 Correlation coefficient

The correlation coefficient r (also called Pearson's product moment correlation after Karl Pearson Pérez et al. (2000)) is used to determine the strength and direction of the relationship between two variables. This form of correlation requires that both variables are normally distributed, interval or ratio variables. The correlation coefficient is calculated by eq.(16):

$$r = \frac{n\sum x_i y_i - (\sum x_i)(\sum y_i)}{\sqrt{n(\sum {x_i}^2) - (\sum x_i)^2}\sqrt{n(\sum {y_i}^2) - (\sum y_i)^2}} \tag{16}$$

where n is the number of data points. The numerical values of correlation coefficient range from +1 to -1. If two variables move exactly together, the value of the correlation coefficient is 1. This indicates perfect positive correlation. If two variables move exactly opposite to each other, the value of the correlation coefficient is -1. Low numerical values indicate little relationship between two variables, such as -0.10 or +0.15 indicate little relationship between on two variable.

4. Results

Fig. 3 shows the distribution of pollutant patterns $[SO_2;PM_{10}]$ at the three monitoring stations (CR, DF and NA). The mesh in Fig. 3 corresponds to the thresholds established by the program to improve the air quality in Salamanca (*ProAire*) INE (2004). Thresholds are Pre-contingency, Phase-I contingency and Phase-II contingency. For example, for SO_2 concentrations equal to or bigger than 145 ppb and smaller than 225 ppb (average per day), a level of environmental pre-contingency is declared. Therefore the spaces between lines in the mesh represent the levels of environmental contingency for SO_2 and PM_{10} concentrations.
In Fig. 3 each symbol (*, • and $\bigtriangledown$) represent the pollutant patterns at each monitoring station. At Nativitas monitoring station we observe that the highest PM_{10} and SO_2 pollutant concentrations are not present at the same time. On other hand, at the Cruz Roja monitoring

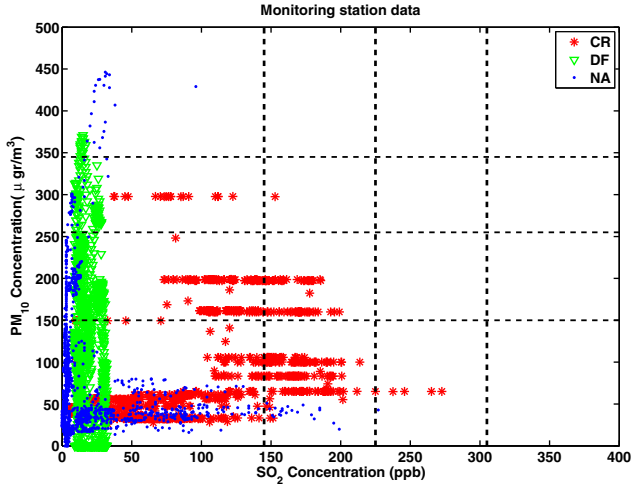
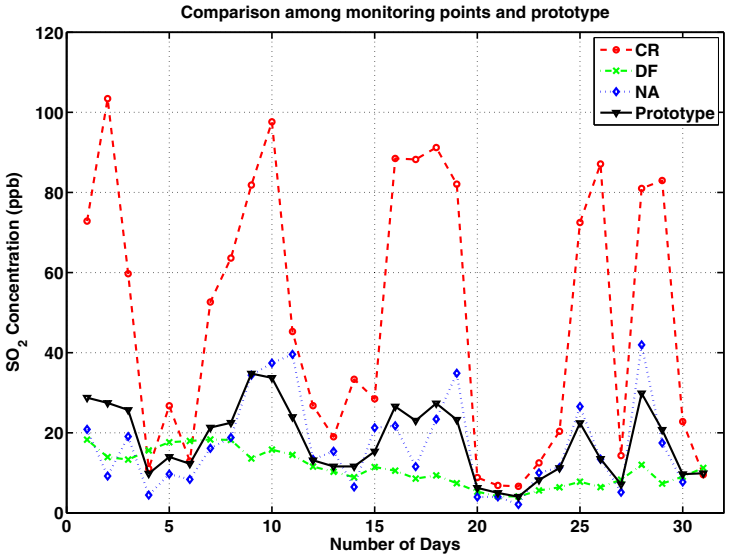


Fig. 3. Monitoring Network per minute.

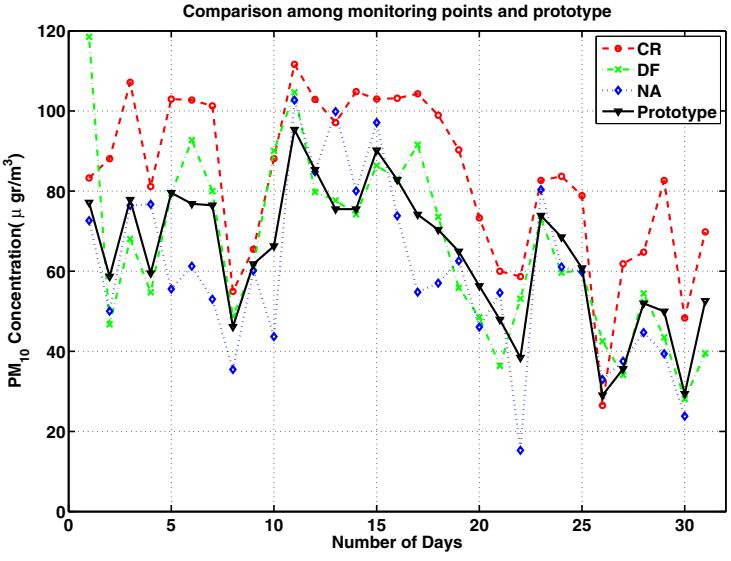
station we observe that either SO_2 or PM_{10} pollutant concentrations are highest. At the DIF monitoring station we observe the highest PM_{10} concentrations in the AEMN network.
The main proposal in this work is to apply the PFCM clustering algorithm to the AEMN in Salamanca as well to integrate the pollutant measures from the three monitoring stations.
The PFCM initial parameters (a, b, m and η) are very important in order to reduce the outlier effects in the pattern prototypes. Pal *et al*, in Pal et al. (2005) recommend of b parameter value larger than the a parameter value in order to reduce the mentioned effects. On the other hand, a small value for η and a value greater than 1 for m are recommended. nevertheless, choosing a too high of a value of m reduces the effect of membership of data to the clusters, and the algorithm behaves as a simple PCM.
Taking into account the previous recommendations, the initial parameters for the PFCM clustering algorithm were set as follows: $a = 1$, $b = 5$, $m = 2$ and $\eta = 2$. The found prototypes (a and b) are shown in Fig. 4.
In Fig. 4(a) the daily averages of SO_2 concentrations are presented for each monitoring station together with the corresponding prototypes. It is observed also that Cruz Roja monitoring station receives the highest emissions of SO_2 concentrations: this is due to its location near to the refinery. The prototypes in this case were very low in comparison with the observed SO_2 concentrations, because only one station observed high SO_2 concentrations (Cruz Roja). According with the analyzed patterns the emitted pollutant is only measured by the Cruz Roja monitoring station (see Fig. 4).
Fig. 4(b) shows the daily averages of PM_{10} concentrations and result prototypes. In this case, the observed averages are very similar at the three monitoring stations. The PM_{10} pollutant dispersion is more uniform then the SO_2 pollutant dispersion in the city.
Table 2 shows the correlation results among SO_2 and PM_{10} pollutants and the meteorological variables. The database used in the correlation analysis correspond to year 2004 of Nativitas. This period was taking because contains more meteorological registrations. The obtained results of the SO_2 correlation coefficient show a high positive correlation between SO_2 pollutant and Wind Speed, also a high and negative correlation between SO_2 pollutant and Wind Direction is observed. The other meteorological variables have not impact. For the PM_{10} pollutant, the meteorological variable with more impact is the Relative Humidity. We observe, when the Relative Humidity increases the pollutant concentration decreases. The PM_{10} particles are caught and fall to the ground during rain.

	SO_2	PM_{10}
SO_2	1	0.0731
PM_{10}	0.0731	1
WS	0.4756	-0.1385
WD	-0.6151	0.1478
T	-0.0329	-0.0007
RH	-0.0322	-0.4416
BP	0.1462	0.1806
SR	-0.021	-0.1207

Table 2. Correlation Coefficient between pollutant concentration and meteorological variables.



(a) SO_2



(b) PM_{10}

Fig. 4. Comparison between air pollutant averages and estimated prototypes.

5. Conclusions

Nowadays, there is a program to improve the air quality in the city of Salamanca, Mexico. Besides, this program has established thresholds for several levels of contingencies depending on the SO_2 and PM_{10} pollutant concentrations. However, a particular level of contingency for the city is declared taking into account the highest pollutant concentration provided by one of the three monitoring stations. For example, if a pollutant concentration exceeds a given threshold in a single monitoring station, the alarm of contingency applies to the whole city. This value is normally provided by the Cruz Roja station, due to its proximity to the refinery and power generation industries.

Looking for local and general contingency levels in the city, we have proposed to estimate a set of prototypes such that they can represent a calculated measure of pollutant concentrations according to the values measured in the three fixed stations. In such a way, a local alarm of contingency can be activated in the area of impact of the pollution depending on each station, and a general alarm of contingency according to the values provided by the prototypes. Nevertheless, the last case requires adjusting the thresholds, as the actual values would be only used for local contingency because they depend on the measured values of pollutant concentrations, and the general contingency requires thresholds as a function of calculated values.

6. References

Andina, D. & Pham, D. T. (2007). *Computational Intelligence*, Springer.

Barron-Adame, J. M., Herrera-Delgado, J. A., Cortina-Januchs, M. G., Andina, D. & Vega-Corona, A. (2007). Air pollutant level estimation applying a self-organizing neural network, *Proceedings of the 2nd international work-conference on Nature Inspired Problem-Solving Methods in Knowledge Engineering. IWINAC-07*, pp. 599–607.

Bezdek, J. C. (1981). *Pattern Recognition With Fuzzy Objective Function Algorithms*, Kluwer Academic.

Bezdek, J. C., Keller, J., Krishnapuram, R. & Pal, N. R. (1999). *Fuzzy Models and Algorithms for Pattern Recognition and Image Processing*, first edn, Boston, London.

Celik, M. B. & Kadi, I. (2007). The relation between meteorological factors and pollutants concentration in karabuk city, *G.U. Journal of science* 20(4): 87–95.

Cortina-Januchs, M. G., Barron-Adame, J. M., Vega-Corona, A. & Andina, D. (2009). Prevision of industrial so2 pollutant concentration applying anns, *Proceedings of The 7th IEEE International Conference on Industrial Informatics (INDIN 09)*, pp. 510–515.

Dunn, J. (1973). A fuzzy relative of the isodata process and its use in detecting compact well-separated clusters, *Journal of Cybernetics* 3(3): 32–57.

EPA (2008). Air quality and health, chapter Environmental Protection Agency, National Ambient Air Quality Standards (NAAQS).

Fenger, J. (2009). Air pollution in the last 50 years - from local to global, *Journal of Atmospheric Environment* 43(1): 13–22.

Hoppener, F., Klawonn, F., Kruse, R. & Runkler, T. (2000). *Fuzzy Cluster Analysis, Methods for classification, data analysis and image recognition*, Chistester, United Kingdom.

INE (2004). *Programa para mejorar la calidad del aire en Salamanca*, 2 edn, Instituto de Ecología del Estado de Guanajuato, Calle Aldana N.12, Col. Pueblito de Rocha, 36040 Guanajuato, Gto.

INEGI (2005). *National Population and Housing Census 2*, National Institute of Geography and Statistics. www.inegi.org.mx.

Krishnapuram, R. & Keller, J. (1993). A possibilistic approach to clustering, *International Conference on Fuzzy Systems* 1(2): 98–110.

Krishnapuram, R. & Keller, J. (1996). The possibilistic c-means algorithm: Insights and recommendations, *International Conference on Fuzzy Systems* 4, no 3: 385–393.

Ojeda-Magaña, B., Quintanílla-Dominguez, J., Ruelas, R. & Andina, D. (2009b). Images sub-segmentation with the pfcm clustering algorithm, *Proceedings of The 7th IEEE International Conference on Industrial Informatics (INDIN 09)*, pp. 499–503.

Ojeda-Magaña, B., Ruelas, R., Buendía-Buendía, F. & Andina, D. (2009a). A greater knowledge extraction coded as fuzzy rules and based on the fuzzy and typicality degrees of the GKPFCM clustering algorithm, *In Intelligent Automation and Soft Computing* 15(4): 555–571.

Pal, N. R., Pal, S. K. & Bezdek, J. C. (1997). A mixed c-means clustering model, *IEEE International Conference on Fuzzy Systems, Spain*, pp. 11–21.

Pal, N. R., Pal, S. K., Keller, J. M. & Bezdek, J. C. (2004). A new hybrid c-means clustering model., *Proceedings of the IEEE International Conference on Fuzzy Systems, FUZZ-IEEE04, I. Press, Ed.*

Pal, N. R., Pal, S. K., Keller, J. M. & Bezdek, J. C. (2005). A possibilitic fuzzy c-means clustering algorithm, *IEEE Transactions on Fuzzy Systems* 13(4): 517–530.

Pérez, P., Trier, A. & Reyes, J. (2000). Prediction of pm 2.5 concentrations several hours in advance using neural networks in santiago, chile, *Atmospheric Environment* 34(8): 1189–1196.

Ruspini, E. (1970). Numerical method for fuzzz clustering, *Information Sciences* 2(3): 319–350.

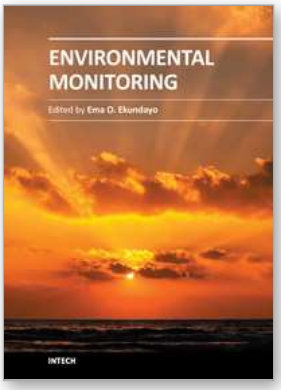
Timm, H., Borgelt, C., Döring, C. & Kruse, R. (2004). An extension to possibilistic fuzzy cluster analysis, *Fuzzy Sets and systems* 147, no 1: 3–16.

Timm, H. & Kruse., R. (2002). A modification to improve possibilistic fuzzy cluster analysis., *Conference Fuzzy Systems, FUZZ-IEEE, Honolulu, HI, USA*.

WHO (2008). *Air quality and health*, chapter World Health Organization.

Zamarripa, A. & Sainez, A. (2007). *Medio Ambiente: Caso Salamanca*, Instituto de Investigación Legistativa, H. Congreso del Estado de Guanajuato, LX legislatura.

Zuk, M., Cervantes, M. G. T. & Bracho, L. R. (2007). Tercer almanaque de datos y tendencias de la calidad del aire en nueve ciudades mexicanas, *Technical report*, Secretaría de Medio Ambiente, Recursos Naturales Instituto Nacional de Ecología, México, D.F.



Environmental Monitoring
Edited by Dr Ema Ekundayo

ISBN 978-953-307-724-6
Hard cover, 528 pages
Publisher InTech
Published online 04, November, 2011
Published in print edition November, 2011

"Environmental Monitoring" is a book designed by InTech - Open Access Publisher in collaboration with scientists and researchers from all over the world. The book is designed to present recent research advances and developments in the field of environmental monitoring to a global audience of scientists, researchers, environmental educators, administrators, managers, technicians, students, environmental enthusiasts and the general public. The book consists of a series of sections and chapters addressing topics like the monitoring of heavy metal contaminants in varied environments, biolgical monitoring/ecotoxicological studies; and the use of wireless sensor networks/Geosensor webs in environmental monitoring.

INTECH
open science | open minds

InTech Europe
University Campus STeP Ri
Slavka Krautzeka 83/A
51000 Rijeka, Croatia
Phone: +385 (51) 770 447
Fax: +385 (51) 686 166
www.intechopen.com

InTech China
Unit 405, Office Block, Hotel Equatorial Shanghai
No.65, Yan An Road (West), Shanghai, 200040, China
中国上海市延安西路65号上海国际贵都大饭店办公楼405单元
Phone: +86-21-62489820
Fax: +86-21-62489821