Prediction of PM_{10} concentrations using Fuzzy c-Means and ANN.

M.G. Cortina-Januchs, J. Quintanilla-Domínguez
and D. Andina
Technical University of Madrid Madrid, Spain
Email:{januchs, joelq}@salamanca.ugto.mx,
d.andina@upm.es

Antonio Vega-Corona
Universidad de Guanajuato
Salamanca, Mexico
Email: tono@salamanca.ugto.mx

***Abstract*—Salamanca has been considered among the most polluted cities in Mexico. The vehicular park, the industry and the emissions produced by agriculture, as well as orography and climatic characteristics have propitiated the increment in pollutant concentration of Particulate Matter less than 10 $\mu g/m^3$ in diameter (PM_{10}). In this work, a Multilayer Perceptron Neural Network has been used to make the prediction of an hour ahead of pollutant concentration. A database used to train the Neural Network corresponds to historical time series of meteorological variables (wind speed, wind direction, temperature and relative humidity) and air pollutant concentrations of PM_{10}. Before the prediction, Fuzzy c-Means clustering algorithm have been implemented in order to find relationship among pollutant and meteorological variables. These relationship help us to get additional information that will be used for predicting. Our experiments with the proposed system show the importance of this set of meteorological variables on the prediction of PM_{10} pollutant concentrations and the neural network efficiency. The performance estimation is determined using the Root Mean Square Error (RMSE) and Mean Absolute Error (MAE). The results shown that the information obtained in the clustering step allows a prediction of an hour ahead, with data from past 2 hours.**

I. Introduction

Urban pollution is one of several major atmospheric pollution problem currently confronting the world's population. The problem is growing because of rapidly increasing urban population, unchecked urban and industrial expansion, traffic density and meteorological and topographical properties of the region [1], [2]. The Air pollution continues to be detrimental to human health despite these emission standards [3]. The inhalation of air polluted with particulate matter (PM_{10}) or irritant gases such as Nitrogen oxides (NO_x) and Sulphur Dioxide (SO_2) are associated with both short-term and long-term health effects, most of which impact on the respiratory and cardiovascular systems [4].

Moreover ambient air pollution is also strongly influenced by meteorological factors, such as: wind speed, temperature, relative humidity, sunlight intensity [5], [6]. Examining air quality information is important in understanding possible human exposure and potential impacts in health and welfare [7]. Because of this air quality models require meteorological data to correctly predict air pollutant concentrations.

Over the years, several authors have applied Artificial Neural Networks (ANN) in order to predict pollutant concentration. The Multilayer Perceptron (MLP) and Elman ANN have been applied for the prediction of SO_2, O_3 and $PM_{2.5}$ [8]–[10], using traffic flow and meteorological variables as input data in ANN, obtaining predictions of 1, 2 or 3 hours ahead. The results show that the ANN models performed slightly better than the deterministic and linear statistical models. Ibarra *et al.* [11] predicted hourly level for five pollutants (SO_2, CO, NO_2, NO and O_3), using MLP, Radial Basis Function (RBF) and Generalized Regression Neural Network (GRNN), obtaining that in some cases, the GRNN and RBF can perform as well or even better than MLP. The Feed Forward Neural Network (FFNN) and linear regression have been applied for the prediction of PM_{10} in Grecee [12] in five monitoring station, in your ANN model data of 24 hours and the subsequent average of the data have been used to make the final prediction. Additionally the Mixing Layer Height, the temperature, wind direction component, relative humidity are used as ANN input.

In our previous works, we have applied MLP using time window in order to perform prediction of SO_2 [13] and PM_{10} [14], as well as meteorological variables. The main problem of long time windows occurs when data is missing. In this work, we implemented a method to extract additional information about relation among pollutant and meteorological variables, in order to reduce the size of time windows and improve prediction. A MLP has been used to predict an hour ahead of pollutant concentration.

The most important issue of ANN in pollutant forecasting is generalization, which refers to their ability to produce reasonable predictions on data sets other than those used for the estimation of the model parameters [15]. This issue has an important parameter that should be accounted for, it is data preparation. In this work, the preparation of the data is performed by applying a clustering algorithm. Clustering involves the task of dividing data sets, which assigns the same label to members who belong to the same group, so

978-1-61284-972-0/11/$26.00 ©2011 IEEE

that each group is more or less homogeneous and distinct from the others. In hard clustering (K-means), data is divided into crisp clusters, where each data set belongs to exactly one cluster. In fuzzy clustering, the data points can belong to more than one cluster, and associated with each of the points are membership grades which indicate the degree to which the data sets belong to the different clusters. For this reason the Fuzzy c-Means clustering algorithm (FCM) was used in order to find relationship among pollutant and meteorological variables. These relationship help us to get additional information that will be used for predicting. Unlike hard classification methods which force data to belong exclusively to one class, FCM allows data to belong to multiple classes with varying degrees of membership.

The FCM was initially development by Dunn [16] and later generalised by Bezdek [17]. This algorithm is based on optimising the objective function given by the equation

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

where the matrix $U = [\mu_i]\epsilon M_{fcm}$ is a fuzzy partition of the data set Z, and $V = [v_1, v_2, ...v_c]$ is the vector of prototypes of the clusters, which are calculated according to $D_{ikA} = \|z_k - v_k\|^2$, it is a square inner-product distance norm. $m\epsilon[1,\infty]$ is a weighting exponent that determines the fuzziness of the resulting clusters. μ_{ik} and v_i are obtained by the following equations

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

$$v_i = \sum_{k=1}^{N}\mu_{ik}^m z_k \Big/ \sum_{k=1}^{N}\mu_{ik}^m \tag{3}$$

The optimal partition U^* of Z using the Fuzzy c-Means algorithm is reached by implementing the couple (U^*, V^*) to locally minimise the objective function J_{fcm} according to an alternating optimisation method [18].

II. Study Area

Salamanca city is located in the state of Guanajuato, Mexico, and it has an approximate population of 234,000 inhabitants [19]. The city is 340 km northwest from Mexico City, with coordinates 20° 34' 09" North latitude, and 101° 11' 39" West longitude. The population growth, car park, industry, the refinery and thermoelectric activities, the emissions produced by agriculture, as well as orography and climatic characteristics have propitiated the increment in pollutant concentration of Sulphur Dioxide (SO_2) and Particulate Matter less than 10 micrometers in diameter (PM_{10}). Salamanca is ranked as one of the most polluted cities in Mexico [20].

The PM_{10} concentrations frequently exceed the legislated air quality standards in this city. Particulate Matter (PM) is a complex mixture of airborne particles that differ in size ($PM_{2.5}$ and PM_{10}), origin and chemical composition. PM is released from natural and anthropogenic sources, such as soils, car exhausts, industry, and power plants therefore increasing PM concentrations in many locations. As the molecules are very large to mix between the water molecules, some pollutants do not dissolve in water [21]. This material is termed as particulate matter and can lead to water pollution PM has been linked to a range of serious respiratory and cardiovascular health problems.

The Automatic Environmental Monitoring Network (AEMN) was installed, it consists of three fixed and one mobile stations [22]. The fixed stations cover approximately 80 % of the urban area while the mobile station covers the remaining 20 %. Each station has the necessary instrumentation to measure concentration of criteria pollutants and meteorological variables.

III. Methodology

The forecasting methodology includes the following operational steps:

- *Data preparation* for forecasting.
- *Network architecture determination.*
- *Design of network training strategy.*
- *Evaluation* of forecasting results.

A. Data Preparation for Forecasting

The data base obtained by the AEMN, represents the essential information to be used for the determination and prediction of environmental situations. The available data set refers to a monitoring station located in a residential area of the city. Due the conditions involved in air pollutants measurements it is necessary to revise and refine the gathered information from the AEMN. The validation of data base was done according to the INE manual [23]. PM_{10} pollutant concentrations, Wind Direction, Wind Speed, Temperature, and Relative Humidity were used to form patterns. The patterns were formed as follow:

$$P = [C_{PM10}, WS, WDI, T, HR] \tag{4}$$

where C_{PM10} is PM_{10} concentration, WS is wind speed, WDI is the Wind Direction Index [24], T is temperature and HR is the relative humidity. The WDI is defined according to the following expression:

$$WDI = 1 + \sin(WD + \pi/4) \tag{5}$$

The Fuzzy c-Means clustering algorithm was implemented to obtain the relationships between variables and get a better prediction. With the obtained result of clustering step, each pattern is labeled and this label is consider a feature.

The initial conditions for this clustering method were as follows:

- The cluster number took values from 2 to 7
- Prototypes were initialised as random values
- The number of membership degrees was set to 2
- The maximum number of iterations was set to 100
- The minimum amount of improvement was set to $1 \text{ x } 10^{-3}$

B. Network architecture determination

The proposed system is based on MLP. After trying with some others ANN structures, a MLP structure with two hidden layer was used. The MLP model consists of N inputs (concentration of PM_{10}, WS, WDI, T and RH) in time $t = 0...N$, where $N\epsilon[2,6]$, two hidden layers of N neurons and $N/2$ neurons respectively, obtaining as in the output the next-hour concentration. The MLP with the same structure was used for all our experiments in order to find the time-window size necessary to make the best prediction using meteorological variables and labeled of clustering step.

The input patterns P_{in_t} and output patterns P_{out_t} of neural network were formed as follow:

$$P_{in_t} = [C_t, C_{t-1}, ..., C_{t-n}, WS_t, WDI_t, T_t, HR_t, L_t] \quad (6)$$

$$P_{ou_t} = [C_{t+1}] \quad (7)$$

where C concentration observed in time t, n is hours before we need to make the prediction, L is the clustering label. The training set was formed with 70 % of patterns and the remaining patterns were used as test set. In addition, training and test sets were normalized in the range [0 1]. All the mathematical computations were performed using Neural Network Toolbox in Matlab©.

C. Design of network training strategy

The network structures used are as follows:

- Input layer: N neurons, where each neuron is feature.
- Hidden layer: one hidden layer with $N/2$ neurons.
- Output layer: one output layer, where in the output the next-hour concentration is obtained
- Learning rate: 1
- The used activation function: the log-sigmoid function.
- Training set: 490 pattern.
- Training conditions: epoch = 200.
- Performance function:
  Mean Squared Error (MSE) = 0.01.
- Test set: 210 patterns.

TABLE I
THE BEST OBTAINED RESULTS OF 30 ANN TRAINED

No. Hours	No.Cluster	MAE	MSE	RMSE
2	3	0.0558	0.0061	0.0783
3	7	0.0539	0.0062	0.0787
4	5	0.0648	0.0092	0.0958
5	5	0.0662	0.0088	0.0940
6	3	0.0751	0.0102	0.1011

D. Evaluation of forecasting results

The ANN model performance was evaluated though the following two parameters: Mean Absolute Error *(MAE)*, Mean Square Error *(MSE)* and Root Mean Square Error *(RMSE)* defined as follow:

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|X_i - Y_i| \quad (8)$$

$$MSE = \left[\frac{1}{N}\sum_{i=1}^{N}(X_i - Y_i)^2\right] \quad (9)$$

$$RMSE = MSE^{1/2} \quad (10)$$

where X_i is the observed value at time i, Y_i is the predicted value at time i and N is the total number of observations.

IV. Experimental Results

The figure 1 shows the obtained results for 7 clusters, each cluster is labeled and represents a homogeneous group. These labels are used as input vector of ANN. In the figure 1, we observe that the groups are overlapping, because the small variation range of temperature and wind velocity in the study area. The cluster 6 represents the highest pollutant concentration and the highest wind speed. The cluster 7 contains high pollutant concentrations but a mean wind speed. The cluster 3 represents the lowest pollutant concentration with a mean wind speed. After applied clustering algorithm and labeled each pattern, P_{in_t} was formed using different time-windows, where $t = 2...6$ in order to compare with previous works.

In this work, 30 neural networks were trained, the table I shows the summary of the best obtained results for each time-window. In table the first column is the size time-window, second column is the number of cluster, and the remaining columns represent the prediction errors of PM_{10} concentrations. The best prediction was obtained with size-time-window=2 and 3 clusters, with MAE=0.0558 and RMSE=0.0783. The worst results were obtained with size-time-window=6 and 3 clusters, with MAE=0.0751 and RMSE=0.1011. Figure IV the obtained results of with size-time-window=2 and 3 clusters.

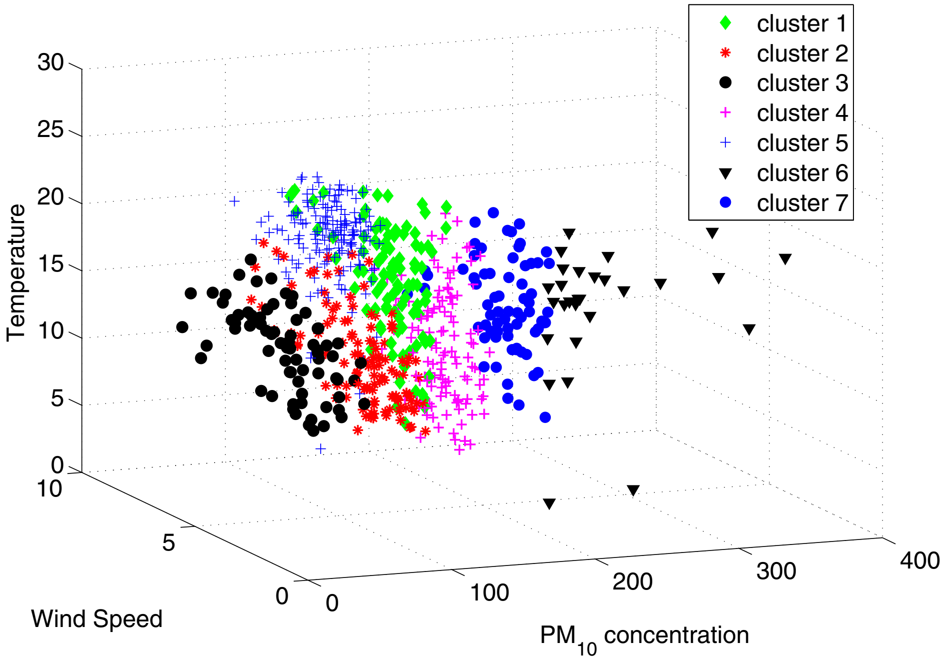

Fig. 1. Obtained results of FCM with 7 cluster for PM_{10} concentrations

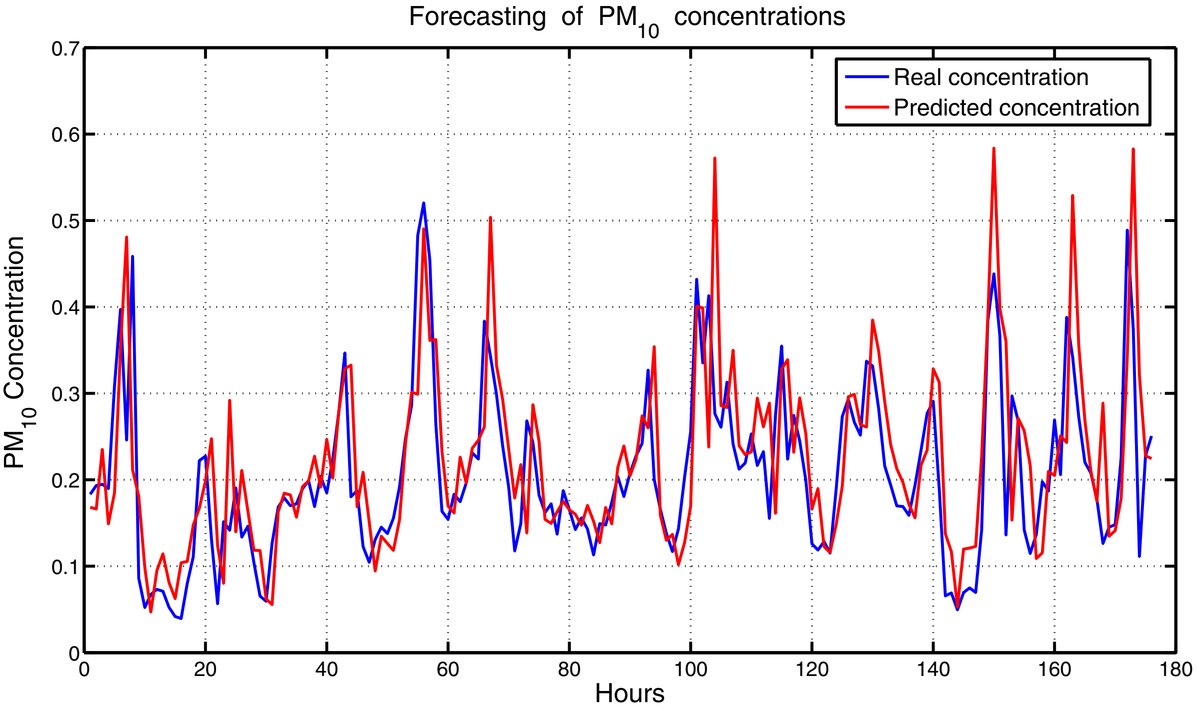

Fig. 2. Obtained results for prediction of PM_{10} concentrations

V. Discussion

In a previous work [14], we had applied MLP in order to predict PM_{10} concentrations, where the best results were obtained with patterns formed with time-window size of six hours. Each pattern was formed with six hours of pollutant concentration and four meteorological variables (WD, WS, T, HR).

In this work, we used the same meteorological variables but, we have applied FCM clustering algorithm before the prediction in order to find relationship among pollutant and meteorological variables. FCM allows data to belong to multiple classes with varying degrees of membership.

The results show that using the degrees of membership reduce the size of time windows and improve pollutant prediction. In table I show that the best result was obtained with time-windows size of 2 hours and 3 clusters. The addition of clustering step reduces the size of the time-window, being especially helpful when data is missing due to errors in measurement devices.

VI. Conclusions

In the presented work, we study the advantages of applying a soft clustering algorithm, as Fuzzy C-means, to extract information about relationship among PM_{10} pollutant concentration and meteorological variables (wind speed, wind direction, temperature and relative humidity) in a polluted environment prediction system. This information, plus different time-windows were probed in a MLP in order to predict an hour ahead of PM_{10} concentration. The best results in PM_{10} prediction were obtained with time-windows of 2 hours and 3 cluster, showing that it is possible to drastically reduce time window for a reasonable prediction of pollution concentration.

Acknowledgment

The authors wish to thank the National Council for Science and Technology (CONACyT), the Secretariat of Public Education (SEP), the Government of Mexico and the Group of Automation in Signal and Communications (GASC) of the Technical University of Madrid for their contributions.

References

[1] Celik, M.B., Kadi, I.: The relation between meteorological factors and pollutants concentration in Karabuk City. G.U. Journal of Science, 20(4), 89-95, (2007).

[2] Sousa, S.I.V., Martins, F.G.,Pereira M.C., Alvim-Ferraz, M.C.M, Ribeiro, H., Oliveira, M., Abreu, I.: Influence of atmospheric ozone, PM_{10} and meteorological factors on the concentration of airborne pollen and fungal spores. Atmospheric Environment 42, 7452–7464 (2008)

[3] World Health Organizatio (WHO). Air Quality Guidelines Global Update 2005. www.euro.who.in (Accessed 2011).

[4] Moreno, T., Lavín, J., Querol, X., Alastuey, A., Viana, M., Gibbons, W.: Controls on hourly variations in urban background air pollutant concentrations. Atmospheric Environment, 43(27): 4178-4186, (2009).

[5] Demuzere, M., Trigo, R.M., Vila-Guerau-de Arellano, J., Van Lipzig, N.P.M.: The impact of weather and atmospheric circulation on O_3 and PM_{10} levels at a rural mid-latitude site. Atmospheric Chemistry and Physics, 9(8): 2695-2714, (2009).

[6] Moussiopoulos, N., Louka, P., Finzi, G.,Volta, M., Colbeck, I., Diéguez, J.J., Palau, J.L., Pérez-Landa, G., Salvador, R., Millán, M.M.: Meteorological aspects of air pollution episodes in southern European cities. Meteorology applied to urban pollution problems, 119-133, (2002).

[7] Chen, H., Namdeo A., and Bell M.: Classification of road traffic and roadside pollution concentrations for assessment of personal exposure. Environmental Modelling and Software, 23(3): 282-287, (2008).

[8] Kukkonen, J., Partanen, L.: Extensive evaluation of neural network models for the prediction of NO_2 and PM_{10} concentrations, compared with a deterministic modelling system and measurements. Atmospheric Environment, 37: 4539-4550 (2003).

[9] Pérez, P., Trier, A.: Prediction of $PM_{2.5}$ concentrations several hours in advance using neural networks in Santiago, Chile. Atmospheric Environment, 34: 1189-1196, (2000).

[10] Kurt, A., Gulbagci, B.: An online air pollution forecasting system using neural networks. Environmental International, 34: 592-598, (2008).

[11] Ibarra-Berastegui, G., Elias, A., Barona, A., Saenz, J., Ezcurra, A., Diaz de Argandoña, J.: From diagnosis to prognosis for forcasting air pollution using neural network: Air pollution monitoring in Bilbao. Environmental Modelling and Software, 23: 622-637, (2008).

[12] Sfetsos, A., Vlachogiannis, D.: A new methodology development for the regulatory forecasting of $PM_1$0. Application in the Greater Athens Area, Greece. Atmospheric Environment, 44(26): 3159-3172, (2010).

[13] Cortina-Januchs, M. G., Barron-Adame, J. M., Vega-Corona, A., Andina, D.: Prevision of industrial SO_2 pollutant concentration applying ANNs. Environmental Modelling and Software Informatics, INDIN 2009, 510-515, (2009).

[14] Cortina-Januchs, M. G., Barron-Adame, J. M., Vega-Corona, A., Andina, D.: Pollution Alarm System in Mexico. Bio-Inspired Systems: Computational and Ambient Intelligence, IWANN 2009, (5517): 1336-1343, (2009).

[15] Andina, D. and Sanz-Gonzalez, J.L.: On the problem of binary detection with neural networks. Proceedings of the 38th Midwest Symposium on Circuits and Systems:(1), 554-557, (1995).

[16] Dunn, J.C.: A Fuzzy relative of isodata process and its use in detecting compact well-separated clusters. Journal of Cybernetics, (3): 32-57, (1973).

[17] Bezdek, J. C.: Pattern Recognition with Fuzzy Objective Function Algorithms Plenum Press, New York, (1981).

[18] Ojeda-Magaña, B., Quintanilla-Dominguez, J., Ruelas, R., Andina, D.: Images sub-segmentation with the PFCM clustering. 7th IEEE international conference INDIN 2009, 499-503, (2009).

[19] (In Spanish) INEGI: Population and Housing Census 2. National Institute of Geography and Statistics, (2005). *www.inegi.org.mx (Accessed 2011).*

[20] (In Spanish) Zamarripa, A., Sainez, A. :Medio Ambiente: Caso Salamanca, Instituto de Investigación Legistativa, Apuntes Legistativos, (2007).

[21] Grau, J. B., Anton, J. M., Tarquis, A. M., and Andina, D.: Election of water resources management entity using a multi-criteria decision (MCD) method in Salta province (Argentina). Journal of Systemics, Cybernetics and Informatics: 7(5), 17, (2009).

[22] (In Spanish) Instituto de Ecología del Estado de Guanajuato: Programa para la mejora de la calidad del aire en Salamanca 2007-2012. Secretaría de medio ambiente y recursos naturales.

[23] (In Spanish) INE, Instituto Nacional de Ecología, Dirección General de Investigación sobre la Contaminación Urbana y Regional: Investigación sobre la calidad del aire: contaminantes criterio. *www.ine.gob.mx* (Accessed 2010)

[24] Ordieres, J. B., Vergara, E. P., Capuz, R. S., Salazar, R. E.: Neural network prediction model for fine particulate matter ($PM_{2.5}$) on the US-Mexico border in El Paso (Texas) and Ciudad Juárez (Chihuahua). Environmental Modelling and Software, 20(5): 547-559, (2004).