Small-Particle Pollution Modeling Using Fuzzy Approaches

Àngela Nebot and Francisco Mugica

Soft Computing Research Group, Technical University of Catalonia,
Jordi Girona 1-3, Barcelona, Spain
{angela,fmugica}@lsi.upc.edu

Abstract. Air pollution caused by small particles is a major public health problem in many cities of the world. One of the most contaminated cities is Mexico City. The fact that it is located in a volcanic crater surrounded by mountains helps thermal inversion and imply a huge pollution problem by trapping a thick layer of smog that float over the city. Modeling air pollution is a political and administrative important issue due to the fact that the prediction of critical events should guide decision making. The need for countermeasures against such episodes requires predicting with accuracy and in advance relevant indicators of air pollution, such are particles smaller than 2.5 microns ($PM_{2.5}$). In this work two different fuzzy approaches for modeling $PM_{2.5}$ concentrations in Mexico City metropolitan area are compared with respect the simple persistence method.

Keywords: Air Pollution Modeling, $PM_{2.5}$ Pollution, Fuzzy Inductive Reasoning, ANFIS, Persistence, Time Series Analysis.

1 Introduction

The high levels of particulate matter in the air are of high concern since they may produce severe public health effects and are the main cause of the attenuation of visible light. There are very high levels of particles in North Africa, much of the Middle East, Asia and Latin America as well as in the large urban areas. Comparing it with population density maps, the WHO concluded that more than 80% of the world population is exposed to high levels of fine particles ($PM_{2.5}$) [1]. Likewise, identifies $PM_{2.5}$ as an important indicator of risk to health and might also be a better indicator than PM_{10} for anthropogenic suspended particles in many areas [2]). According to the WHO Guidelines, concentrations at this level and higher are associated with an approximately 15% increased risk of mortality, relative to the Air Quality Guideline (AQG) of 10 $\mu g\ m^{-3}$ [1].

Regarding the $PM_{2.5}$, it has not yet been identified a threshold below which damage to health does not occur, this has motivated that the limits for the protection of public health are getting lower every year. The geographical characteristics of the Mexico City metropolitan area, i.e. its height, average temperature and terrain, added to the pressure exerted by the growth and intensification of urban activities cause high air pollution episodes that constitute a permanent challenge to the health of its

M.S. Obaidat et al. (eds.), *Simulation and Modeling Methodologies, Technologies and Applications*, Advances in Intelligent Systems and Computing 256,
DOI: 10.1007/978-3-319-03581-9_17,

inhabitants. Although the measures taken over the past 15 years to reduce the impact of air pollution have managed to significantly decrease pollutants such as SO2, CO or the Pb, the concentrations of ozone and fine particles exceed quite often air quality standards.

The monitoring of $PM_{2.5}$ from 2004 to date shows that around 20 million people in Mexico city are exposed to annual average concentrations of this contaminant in between 19 and 25 μg m^{-3}, exceeding by more than double the WHO standard of 10 μg m^{-3} and substantially exceeding the Mexican norm of 15 μg m^{-3}.

The increase of the concentration of particles in Mexico City is strongly associated with the meteorology of the Valley. During the days of intense wind, resuspension of dust from the ground produces significant increases in the concentrations of total suspended particles (PST) and particles lower than 10 μm (PM_{10}). The presence of surface thermal inversions can contribute to the increase in the concentration of particles smaller than 10 μm and fine particles, due to the lack of dispersion and the accumulation in the atmosphere of the particles emitted by vehicles and industry. Higher concentrations usually occur when the layer trapped under the inversion is not very high and the duration of the thermal inversion is maintained throughout the morning.

The national weather service reported a total of 107 days with surface thermal inversions during 2010, the highest in the past 13 years. The largest part was recorded during the winter months, when the long and cold nights favor its formation. In the dry season months it has been reported a 40% of days with thermal inversion. The months of April and December had the largest number of events with 16 and 17 days, respectively. The influence of high pressure systems during the months of March to May was responsible for the formation of surface thermal inversions [3].

Fuzzy logic-based methods have not been applied extensively in environmental science, however, some interesting research can be found in the area of modeling of pollutants [4-10], where different hybrid methods that make use of fuzzy logic are presented for this task.

In this research we propose prediction models of hourly concentrations of $PM_{2.5}$, based on data registered at downtown Mexico City. In a first study, the concentration of $PM_{2.5}$ is used as input variable, becoming a time series modeling. In a second study, the daily maximum temperature is added to the input data in order to obtain prediction models of $PM_{2.5}$ concentrations.

The fuzzy approaches chosen to perform these tasks are the Fuzzy Inductive Reasoning (FIR) methodology and the Adaptive Neuro Fuzzy Inference System (ANFIS). These two fuzzy approaches for modeling small particles are presented and the prediction results obtained are compared with the results of the persistence simple method.

Sections 2 and 3 introduce the basic concepts of FIR and ANFIS methodologies, respectively. Section 4 presents the methods, i.e. the data, the fuzzy models development and the models evaluation. Section 5 describes the results obtained. Finally the conclusions of this research are given.

2 Fuzzy Inductive Reasoning (FIR)

The conceptualization of the FIR methodology arises of the General System Problem Solving (GSPS) approach proposed by Klir [11]. This methodology of modeling and simulation is able to obtain good qualitative relations between the variables that compose the system and to infer future behavior of that system. It has the ability to describe systems that cannot easily be described by classical mathematics or statistics, i.e. systems for which the underlying physical laws are not well understood.

FIR methodology, offers a model-based approach to predicting either univariate or multi-variate time series [12, 13]. A FIR model is a qualitative, non-parametric, shallow model based on fuzzy logic. Visual-FIR is a tool based on the Fuzzy Inductive Reasoning (FIR) methodology (runs under Matlab environment), that offers a new perspective to the modeling and simulation of complex systems. Visual-FIR designs process blocks that allow the treatment of the model identification and prediction phases of FIR methodology in a compact, efficient and user friendly manner [14].

The FIR model consists of its structure (relevant variables) and a set of input/output relations (history behavior) that are defined as if-then rules. Feature selection in FIR is based on the maximization of the models' forecasting power quantified by a Shannon entropy-based quality measure. The Shannon entropy measure is used to determine the uncertainty associated with forecasting a particular output state given any legal input state. The overall entropy of the FIR model structure studied, H_s, is computed as described in equation 1.

$$H_s = -\sum_{\forall i} p(i) \cdot H_i \tag{1}$$

where $p(i)$ is the probability of that input state to occur and H_i is the Shannon entropy relative to the i^{th} input state. A normalized overall entropy H_n is defined in equation 2.

$$H_n = 1 - \frac{H_s}{H_{\max}} \tag{2}$$

H_n is obviously a real-valued number in the range between 0.0 and 1.0, where higher values indicate an improved forecasting power. The model structure with highest H_n value generates forecasts with the smallest amount of uncertainty.

Once the most relevant variables are identified, they are used to derive the set of input/output relations from the training data set, defined as a set of if-then rules. This set of rules contains the behaviour of the system. Using the five-nearest-neighbors (5NN) fuzzy inference algorithm the five rules with the smallest distance measure are selected and a distance-weighted average of their fuzzy membership functions is computed and used to forecast the fuzzy membership function of the current state, as described in equation 3.

$$Memb_{out_{new}} = \sum_{j=1}^{5} w_{rel_j} \cdot Memb_{out_j} \tag{3}$$

The weights w_{rel_j} are based on the distances and are numbers between 0.0 and 1.0.

Their sum is always equal to 1.0. It is therefore possible to interpret the relative weights as percentages. For a more detailed explanation of the FIR methodology refer to [14].

3 Adaptive Neuro-Fuzzy Inference System (ANFIS)

The Adaptive Neuro-Fuzzy Inference System (ANFIS), developed by Jang, is one of the most popular hybrid neuro-fuzzy systems for function approximation [15]. ANFIS represents a Sugeno-type neuro-fuzzy system. A neuro-fuzzy system is a fuzzy system that uses learning methods derived from neural networks to find its own parameters. It is relevant that the learning process is not knowledge-based but data-driven.

The main characteristic of the Sugeno inference system is that the consequent, or output of the fuzzy rules, is not a fuzzy variable but a function, as shown in equation 4.

$$\begin{aligned} &\text{Rule}_1\text{: If A is } A_1 \text{ and B is } B_1 \text{ then } Z = p_{1*}a + q_{1*}b + r_1 \\ &\text{Rule}_2\text{: If A is } A_2 \text{ and B is } B_2 \text{ then } Z = p_{2*}a + q_{2*}b + r_2 \end{aligned} \tag{4}$$

Figure 1 describes graphically how a Sugeno model composed by the two rules described in equation 4 works.

The first step of the Sugeno inference is to combine a given input tuple (in the example of figure 1, a double is used (a=3,b=2)) with the rule's antecedents by determining the degree to which each input belongs to the corresponding fuzzy set (left panel of Fig. 1). The **min** operator is then used to obtain the weight of each rule, w_i, which are used in the final output computation, **Z** (right panel of Fig. 1). Notice that the Sugeno inference has two differentiated set of parameters. The first set corresponds to the membership functions parameters of the input variables. The second set corresponds to the parameters associated to the output function of each rule, i.e. p_i, q_i and r_i.

ANFIS is the responsible of adjusting in an automatic way these two set of parameters by means of two optimization algorithms, i.e. back-propagation (gradient descendent) and least square estimation. Back-propagation is used to learn about the parameters of the antecedents (membership functions) and the least square estimation is used to determine the coefficients of the linear combinations in the rules' consequents. ANFIS is a function of the Fuzzy toolbox that runs under the Matlab environment. For a more detailed explanation of the ANFIS methodology refer to [15].

4 Methods

4.1 The Data

The data used for this study stems from the Atmospheric Monitoring System of Mexico City (SIMAT in Spanish) that measures contaminants and atmospheric

variables from 36 stations distributed through the 5 regions of the Mexico City metropolitan area [16]. The registered variables are the air pollutants, including $PM_{2.5}$, as well as other 10 contaminants, and meteorological variables, 24 hours a day, every day of the year. The web page of SIMAT [16] offers a data base with meteorological and contaminant registers since 1986 up to date, although $PM_{2.5}$ has been registered for the first time in 2004.

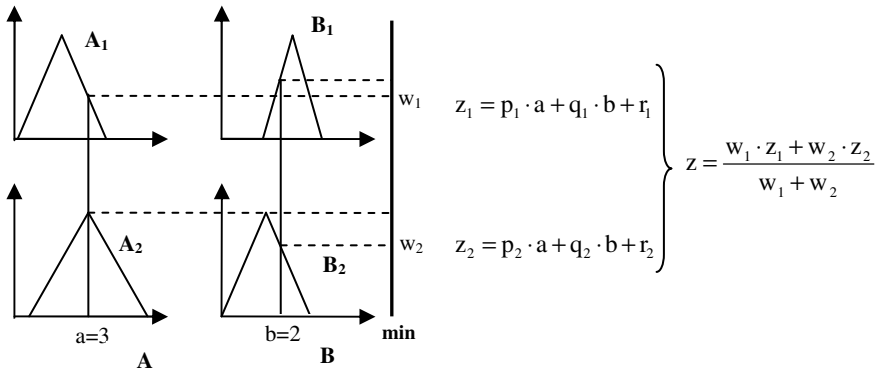

Fig. 1. Example of how a Sugeno model works (evaluation of two fuzzy rules with two input variables or antecedents, i.e. A and B)

A mechanically oscillated mass balance type instrument, TEOM 1400a, is used for the registration of the $PM_{2.5}$. This instrument is very sensitive to changes in concentrations of mass and can provide accurate measurements for samples with less than an hour in length.

This study is centered on the modeling and forecasting of particulate matter with diameter of 2.5 micrometers or less ($PM_{2.5}$) in the Merced station, located in the commercial and administrative district at the downtown of Mexico City Metropolitan Area (MCMA).

$PM_{2.5}$ values are hourly instantaneous observations, not the maximum or the mean of minute registered data. The typical pattern of $PM_{2.5}$ from some city areas, such as for example downtown, suggests that concentrations of this contaminant increase regularly between 8:00 and 16:00 hours, with maximum concentrations around 13:00 hours [17].

It has been decided to use, in this study, data from the half of the year that Mexico City suffers higher $PM_{2.5}$ concentrations, i.e. from December to May. We have used 4 data sets containing 6 month of hourly registers each one, i.e. from the 1st of December until de 31st of May, for years 2007-2008, 2008-2009, 2009-2010 and 2010-2011.

For the first data set, i.e. 1st December 2007 to 31st May 2008, the average concentration is 31.2 $\mu g\ m^{-3}$, the maximum is 147 $\mu g\ m^{-3}$ and the standard deviation is 15.6 $\mu g\ m^{-3}$. For the second data set, i.e. 1st December 2008 to 31st May 2009, the average concentration is 26.6 $\mu g\ m^{-3}$, the maximum is 102 $\mu g\ m^{-3}$ and the standard deviation is 14.3 $\mu g\ m^{-3}$.

For the third data set, i.e. 1st December 2009 to 31st May 2010, the average concentration is 20.8 $\mu g\ m^{-3}$, the maximum is 101 $\mu g\ m^{-3}$ and the standard deviation is

13.4 µg m^{-3}. For the last data set, i.e. 1st December 2010 to 31st May 2011, the average concentration is 32.5 µg m^{-3}, the maximum is 175 µg m^{-3} and the standard deviation is16.5 µg m^{-3}. Figure 2 shows the hourly concentrations of $PM_{2.5}$ during December, 2009.

The data available contains missing values that correspond to data that was not registered due to instrument problems. From the total number of 17496 hourly data registered of $PM_{2.5}$ concentration, 1316 are missing values.

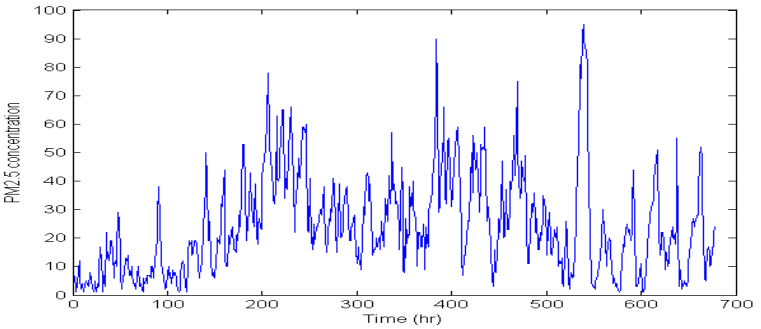

Fig. 2. Hourly concentrations of $PM_{2.5}$ data for December 2009. Units are µg m^{-3}. From the 720 data points, 42 are missing values that are not plotted.

4.2 Models Development

As mentioned before, our goal is to obtain ANFIS and FIR models capable of forecasting the $PM_{2.5}$ concentrations some time in advance, in such a way that efficient actions could be taken in order to protect the citizens of high concentrations episodes.

A study of autocorrelation, both causal and temporal, is first performed. To this end, we used the model structure identification process of the FIR methodology that carries out a feature selection based on the entropy reduction measure, described in section 2.

It has been found that it is possible to relate the concentration of $PM_{2.5}$ at a given time of the day to the sequence of 24 points corresponding to the hourly concentrations of the preceding day. Moreover, the structure of the FIR model has determined that there is a direct causal relation between the level of pollution at present time and the levels at 6 am, 12 pm, 18 pm and 24 pm of the preceding day. That is, there is a positive correlation at 12 pm and 24 pm and a negative correlation at 6 am and 18 pm.

With this information available we think that an interesting and useful approximation to modeling and forecasting $PM_{2.5}$ concentrations is to obtain a specific model for each of the most relevant hours of the day (i.e. 6 am, 12 pm, 18 pm and 24 pm), based on the values of the 6 am, 12 pm, 18 pm and 24 pm hours of the previous day, i.e. hourly models.

Two studies have been performed: the first one uses as models' inputs only the $PM_{2.5}$ concentrations (we refer to them as univariate models) and, the second one, uses also the daily maximum temperature (we refer to them as multi-variate models).

In the first study, we chose to work with the $PM_{2.5}$ scalar time series keeping in mind the idea that if we use a large enough window of data as input, the effect of other pollutants or meteorological data should be implicit in its structure [18].

In the second study, we decide to add a meteorological variable in order to try to enhance the results of the univariate models. Cobourn concludes that the meteorological variables that have a nonlinear relationship with $PM_{2.5}$ statistically significant are daily maximum temperature and wind speed. Moreover, the strongest single relationship between $PM_{2.5}$ and any meteorological variable is the relationship with daily maximum temperature [19]. This is the reason why this variable has been included as input in the multi-variate models.

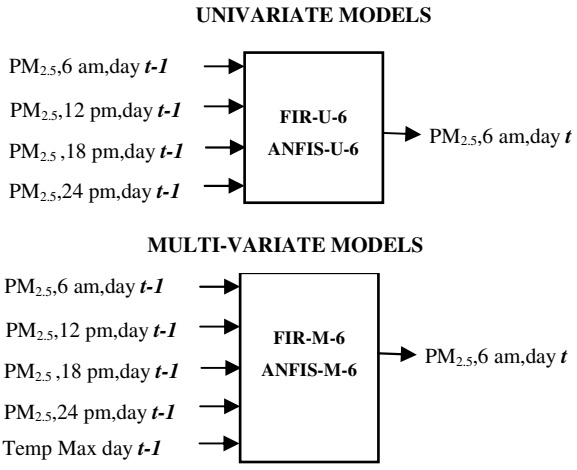

Fig. 3. Input and output variables of the univariate and multi-variate models that predict the $PM_{2.5}$ concentration at 6 am of the subsequent day. The models FIR-U-12, ANFIS-U-12, FIR-M-12 and ANFIS-M-12 have the same input variables described in this figure and as output variable the $PM_{2.5}$ concentration at 12 pm of the subsequent day. Idem for models FIR-U-18, ANFIS-U-18, FIR-M-18, ANFIS-M-18 that have as output variable the $PM_{2.5}$ concentration at 18 pm of the following day, and FIR-U-24, ANFIS-U-24, FIR-M-24, ANFIS-M-24 that have as output variable the $PM_{2.5}$ concentration at 24 pm of the subsequent day.

Both, the univariate and multi-variate models have as input variables the $PM_{2.5}$ concentration at 6 am, 12 pm, 18 pm and 24 pm. The multi-variate models have as additional input variable the daily maximum temperature.

The output variable of each model is the PM2.5 concentration at its corresponding hour. For instance, the output of the 6 am models (i.e. univariate and multi-variate) is the $PM_{2.5}$ concentration at 6 am of the subsequent day. Therefore, for this prediction model, pollutant concentrations are given 6 hours in advance. Figure 3 clarify the inputs and outputs of each of the models developed in this work.

In order to obtain all the models it is necessary to arrange the data in such a way that we have a data stream for each day instead of 24 data streams (one for each hour of that day). The 4 data sets available, and described in section 4.1, have been

arranged accordingly, obtaining now a total number of 725 daily data, out of which 220 are missing values.

In this work a 10-fold cross validation is used to assess how the results of the obtained models generalize to an independent data set. The objective is to estimate how accurately the predictive models developed in this study will perform in practice.

FIR Models

The first step in order to obtain the FIR models is to convert quantitative values into fuzzy data. To this end, it is necessary to specify two discretization parameters, i.e. the number of classes for each system variable (granularity) and the membership functions (landmarks) that define its semantics. In this study the granularity and the clustering method used to obtain the landmarks are summarized in table 1. Many

Table 1. Granularity and clustering methods used to discretize the input and output variables in univariate (i.e. FIR-U-6, FIR-U-12, FIR-U-18, FIR-U-24) and multi-variate (i.e. FIR-M-6, FIR-M-12, FIR-M-18, FIR-M-24) FIR models

Number classes	Clustering method	UNIVARIATE MODELS	MULTI-VARIATE MODELS
2	Fuzzy C-means	FOLD 1, 5, 7, 8, 9	FOLD 1, 4, 8, 9
3	Equal Frequency Partition	FOLD 2, 6	FOLD 2, 6, 7
2	Median Linkage	FOLD 3, 10	FOLD 3
2	K-Means	FOLD 4	FOLD 5

folds are discretized into two classes. It is not possible to use more classes in this case because the number of training data (450 points) is not large enough. several clustering methods such are fuzzy c-means, median linkage, k-means and equal frequency partition are used in this study.

In a general way, the univariate FIR models structure can be described using equation 5.

$$y_i(d) = f_q(x_6(d-1), x_{12}(d-1), x_{18}(d-1), x_{24}(d-1)) \tag{5}$$

where $y_i(d)$ is the predicted $PM_{2.5}$ concentration at time i of day d; x_i represent the real concentration at time i of the preceding day (d-1); and f_q is the qualitative relation of the FIR model. For multi-variate FIR models equation 5 should include the daily maximum temperature as an additional parameter of the function f_q.

ANFIS Models

In order to obtain ANFIS models it is necessary to define the following five parameters: the granularity of each input variable (i.e. number of classes), the shape of the membership functions of the input variables, the type of the output function (i.e. constant or linear), the optimization method to train the fuzzy inference system and the number of training epochs. Several combinations of these parameters have been analyzed in this research and the best results are obtained when two classes with triangular shape membership functions are used for each input variable, a constant

output function is defined and a hybrid optimization method is used for training during 200 epochs.

It was expected than a higher granularity (3 for example) and a linear output function would be a better set of parameters to capture system's behavior, however this is not the case. The results obtained when using these parameter values are bad because some prediction points are very big or very low, distorting the whole prediction set. Analyzing the results, we think that this is due to the complex nature of the data. The bad predictions correspond to those real "extreme" situations that do not appear in the actual training data and, therefore, the obtained ANFIS model has not been adapted to this type of data. When the output function is simplified and the number of classes reduced, less partitioned is the space, and these extreme situations are softened.

4.3 Model Evaluation

The normalized root mean square error, described in equation 6, is used to evaluate the performance of each ANFIS, FIR and persistence models.

$$RNMSE = \frac{\sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}}{\sum_{i=1}^{N} y_i \Big/ N} \tag{6}$$

where $\hat{y}$ is the predicted output, y is the real output and N is the number of samples.

5 Results and Discussion

The persistence method consists on a very simple principle, i.e. tomorrow at time t the $PM_{2.5}$ mass concentration will be the same as today at time t, as described in equation 7. Therefore, there are no parameters to adjust.

$$y_t(d) = x_t(d-1) \tag{7}$$

The prediction results obtained by ANFIS, FIR and persistence univariate models of the $PM_{2.5}$ contaminant at 6 am, 12 pm, 18 pm and 24 pm of the subsequent day, for each of the 10 folds, are summarized in tables 2 and 3.

From tables 2 and 3 it can be seen that ANFIS and FIR models perform much better than persistence for all the four univariate models, i.e. 6 am, 12 pm, 18 pm and 24 pm. Moreover, both ANFIS and FIR models obtain better results than persistence fold by fold, except for fold 7 of model 12 pm.

The ANFIS models are between 12.5% and 30% better than the persistence models while FIR models between 9.6% and 24%.

If we compare the results of ANFIS vs. FIR models it can be concluded that ANFIS obtains lower average errors than FIR, however the differences are really

small. Therefore, both fuzzy methodologies can be considered appropriate approaches to deal with this complex modelling problem.

ANFIS and FIR 6 am models obtain very good results, with average errors of 0.36 and 0.38, respectively. These are low errors if we compare with the errors obtained by the rest of the models, i.e. 12 pm, 18 pm and 24 pm. This makes sense, since models at 6 am are predicting values only 6 hours in advance whereas the rest of the models predict the $PM_{2.5}$ concentration 12, 18 and 24 hours in advance. It is interesting to mention that the higher average error obtained with univariate ANFIS and FIR models in this research are of the order of 0.5. In general, the results obtained are quite good for the problem at hand if we compare them with other results found in the literature that deal with the same problem and use also univariate $PM_{2.5}$ concentration. For instance in [20], the best model obtained that is based on neural networks, has an RNMSE of 0.5, i.e. their better results correspond to the worse results obtained in this research. It should be noted that this comparison is only to point out the complexity of the problem. Both studies use different data and, therefore, it is not possible to perform a rigorous comparison of the different methods used.

The prediction results obtained by ANFIS, FIR and persistence multi-variate models of the $PM_{2.5}$ contaminant at 6 am, 12 pm, 18 pm and 24 pm of the subsequent day, for each of the 10 folds, are summarized in tables 4 and 5. It is important to clarify that the errors of the univariate persistence models are not exactly the same than the errors of the multi-variate persistence models because the inclusion of the daily maximum temperature means an increase in the number of missing values in the test data.

From tables 4 and 5 it can easily be concluded that no enhancement has been pro duced when daily maximum temperature is included as additional input variable to the ANFIS and FIR models. The RNMSE are the same or almost the same for univariate models (tables 2 and 3) and multi-variate models (tables 4 and 5). Therefore, in this case, the use of meteorological information does not help to obtain more accurate and reliable models.

Table 2. Prediction errors (RNMSE) of each fold separately and its average for the $PM_{2.5}$ concentration series. Predictions correspond to 6 am and 12 pm of the subsequent day using ANFIS, FIR and persistence univariate models.

	Univariate Models at 6 am			**Univariate Models at 12 pm**		
	ANFIS-U-6	**FIR-U-6**	**PERS.-U-6**	**ANFIS-U-12**	**FIR-U-12**	**PERS.-U-12**
FOLD 1	0.51	0.54	0.74	0.44	0.46	0.55
FOLD 2	0.37	0.41	0.51	0.54	0.52	0.60
FOLD 3	0.38	0.40	0.54	0.30	0.35	0.42
FOLD 4	0.29	0.34	0.42	0.45	0.44	0.50
FOLD 5	0.26	0.28	0.31	0.56	0.56	0.62
FOLD 6	0.33	0.33	0.52	0.71	0.70	0.92
FOLD 7	0.28	0.33	0.50	0.80	0.82	0.75
FOLD 8	0.48	0.47	0.54	0.53	0.57	0.74
FOLD 9	0.31	0.34	0.39	0.40	0.42	0.51
FOLD 10	0.33	0.40	0.53	0.49	0.50	0.55
AVERAGE	**0.35**	**0.38**	**0.50**	**0.52**	**0.54**	**0.62**

Table 3. Prediction errors (RNMSE) of each fold separately and its average for the $PM_{2.5}$ concentration series. Predictions correspond to 18 pm and 24 pm of the subsequent day using ANFIS, FIR and persistence univariate models.

	Univariate Models at 18 pm			Univariate Models at 24 pm		
	ANFIS-U-18	FIR-U-18	PERS.-U-18	ANFIS-U-24	FIR-U-24	PERS.-U-24
FOLD 1	0.44	0.48	0.55	0.42	0.43	0.52
FOLD 2	0.54	0.52	0.54	0.48	0.46	0.62
FOLD 3	0.41	0.39	0.43	0.35	0.40	0.42
FOLD 4	0.52	0.54	0.68	0.42	0.45	0.46
FOLD 5	0.45	0.48	0.52	0.60	0.65	0.69
FOLD 6	0.65	0.66	0.74	0.55	0.55	0.62
FOLD 7	0.53	0.53	0.63	0.44	0.45	0.45
FOLD 8	0.55	0.57	0.65	0.49	0.54	0.60
FOLD 9	0.38	0.40	0.45	0.36	0.39	0.42
FOLD 10	0.42	0.44	0.44	0.35	0.35	0.42
AVERAGE	**0.49**	**0.50**	**0.56**	**0.45**	**0.47**	**0.52**

On the other hand, ANFIS and FIR multi-variate models obtain similar results and it is not possible to conclude which one has a better performance. Again the prediction errors obtained with 6 am models are much lower than the ones obtained with the rest of the models. ANFIS and FIR models perform much better than persistence models, as already happened in the univariated case. The ANFIS multi-variate models are between 1.9% and 24% better than the persistence models while FIR multi-variate models between 9.4% and 22%.

Table 4. Prediction errors (RNMSE) of each fold separately and its average for the $PM_{2.5}$ concentration series. Predictions correspond to 6 am and 12 pm of the subsequent day using ANFIS, FIR and persistence multivariate models.

	Multi-variate Models at 6 am			Multi-variate Models at 12 pm		
	ANFIS-M-6	FIR-M-6	PERS.-M-6	ANFIS-M-12	FIR-M-12	PERS.-M-12
FOLD 1	0.44	0.46	0.62	0.43	0.48	0.56
FOLD 2	0.34	0.36	0.48	0.46	0.48	0.50
FOLD 3	0.27	0.31	0.36	0.31	0.32	0.43
FOLD 4	0.35	0.41	0.50	0.53	0.47	0.52
FOLD 5	0.35	0.38	0.39	0.58	0.61	0.63
FOLD 6	0.59	0.48	0.73	0.71	0.70	0.94
FOLD 7	0.39	0.44	0.70	0.79	0.66	0.75
FOLD 8	0.42	0.40	0.41	0.47	0.48	0.65
FOLD 9	0.36	0.39	0.44	0.40	0.45	0.50
FOLD 10	0.32	0.29	0.36	0.31	0.34	0.42
AVERAGE	**0.38**	**0.39**	**0.50**	**0.50**	**0.50**	**0.59**

Table 5. Prediction errors (RNMSE) of each fold separately and its average for the $PM_{2.5}$ concentration series. Predictions correspond to 18 pm and 24 pm of the subsequent day using ANFIS, FIR and persistence multivariate models.

	Multi-variate Models at 18 pm			Multi-variate Models at 24 pm		
	ANFIS-M-18	FIR-M-18	PERS.-M-18	ANFIS-M-24	FIR-M-24	PERS.-M-24
FOLD 1	0.46	0.48	0.55	0.45	0.49	0.53
FOLD 2	0.55	0.49	0.56	0.55	0.50	0.62
FOLD 3	0.31	0.38	0.42	0.59	0.39	0.44
FOLD 4	0.52	0.54	0.68	0.43	0.47	0.43
FOLD 5	0.51	0.50	0.53	0.68	0.62	0.70
FOLD 6	0.68	0.68	0.73	0.69	0.56	0.62
FOLD 7	0.57	0.50	0.62	0.46	0.48	0.45
FOLD 8	0.57	0.61	0.66	0.60	0.54	0.63
FOLD 9	0.37	0.40	0.44	0.40	0.42	0.46
FOLD 10	0.39	0.41	0.51	0.35	0.34	0.38
AVERAGE	**0.49**	**0.50**	**0.57**	**0.52**	**0.48**	**0.53**

$PM_{2.5}$ is a difficult contaminant to be predicted due to the fact that there are significant variations of the concentrations of this pollutant from one day to the subsequent day, and, from one hour to the subsequent one, even with similar weather conditions.

Previous works have been focused on the modelling and prediction of mean [21] or maximum [19] $PM_{2.5}$ concentrations. Also, there are studies that perform binary predictions, i.e. if a dangerous level has been reached [22]. Contrarily, we have focused on a short-term $PM_{2.5}$ forecast, although uncertainties in hourly registers pose enormous challenges for developing accurate models.

6 Conclusions

This paper studies the performance of two fuzzy modelling approaches in a complex problem, i.e. the prediction of $PM_{2.5}$ concentration in downtown Mexico City metropolitan area. The first is a neuro-fuzzy approach, i.e. ANFIS, and the second is a hybrid fuzzy-pattern recognition approach, i.e. FIR.

Two studies have been performed: the first one uses as models input only the $PM_{2.5}$ concentrations (called univariate models) and, the second one, uses also the daily maximum temperature (called multi-variate models).

Our approach is based on hourly models. The idea is to obtain a specific model for each of the most relevant hours of the day (i.e. 6 am, 12 pm, 18 pm and 24 pm), based on the values of the 6 am, 12 pm, 18 pm and 24 pm of the previous day. Therefore, eight ANFIS and FIR models have been developed (4 univariate and 4 multi-variate) and its performance compared with persistence models.

The conclusions are that no enhancement has been produced when daily maximum temperature is included as additional input variable to the ANFIS and FIR models. The accuracy of both ANFIS and FIR methodologies are almost the same, so both fuzzy methodologies can be considered appropriate approaches to deal with this

complex modelling problem. ANFIS and FIR models perform much better than persistence for all the univariate and multi-variate models.

As a future work we propose to:

- Include other meteorological variables into the model.
- Include additional information such are the day of the week or the hour of the day into the models.
- Use additional hybrid modelling techniques such as FIR with genetic algorithm, which will help to find in an efficient way the number of classes and landmarks parameters of FIR discretization process.

References

1. WHO World healt oranization. Air quality guidelines: the global update 2005 (2006)
2. van Donkelaar, A., Martin, R., Verduzco, C., Brauer, M., Kahn, R., Levy, R., Villeneuve, P.: 2010. A Hybrid Approach for Predicting PM2.5 Exposure: van Donkelaar et al. Respond. Environ. Health Perspect. 118(10), 425 (2010)
3. NWM: National Weather Service of Mexico (2012), http://smn.cna.gob.mx/
4. Mintz, R., Young, B.R., Svrcek, W.Y.: Fuzzy logic modeling of surface ozone concentrations. Computers & Chemical Engineering 29, 2049–2059 (2005)
5. Ghiaus, C.: Linear fuzzy-discriminant analysis applied to forecast ozone concentration classes in sea-breeze regime. Atmospheric Environment 39, 4691–4702 (2005)
6. Morabito, F.C., Versaci, M.: Fuzzy neural identification and forecasting techniques to process experimental urban air pollution data. Neural Networks 16, 493–506 (2003)
7. Heo, J.S., Kim, D.S.: A new method of ozone forecasting using fuzzy expert and neural network system. Sicence of the Total Environment 325, 221–237 (2004)
8. Yildirim, Y., Bayramoglu, M.: Adaptive neuro-fuzzy based modelling for prediction of air pollution daily levels in city of Zonguldak. Chemosphere 63, 1575–1582 (2006)
9. Peton, N., Dray, G., Pearson, D., Mesbah, M., Vuillot, B.: Modelling and analysis of ozone episodes. Environmental Modelling & Software 15, 647–652 (2000)
10. Onkal-Engin, G., Demir, I., Hiz, H.: Assessment of urban air quality in Istanbul using fuzzy synthetic evaluation. Atmospheric Environment 38, 3809–3815 (2004)
11. Klir, G., Elias, D.: Architecture of Systems Problem Solving, 2nd edn. Plenum Press, New York (2002)
12. Nebot, A., Mugica, F., Cellier, F., Vallverdú, M.: Modeling and Simulation of the Central Nervous System Control with Generic Fuzzy Models. Simulation 79(11), 648–669 (2003)
13. Carvajal, R., Nebot, A.: Growth Model for White Shrimp in Semi-intensive Farming using Inductive Reasoning Methodology. Computers and Electronics in Agriculture 19, 187–210 (1998)
14. Escobet, A., Nebot, A., Cellier, F.E.: Visual-FIR: A tool for model identification and prediction of dynamical complex systems. Simulation Modelling Practice and Theory 16, 76–92 (2008)
15. Nauck, D., Klawonn, F., Kruse, R.: Neuro-Fuzzy Systems. John Wiley & Sons (1997)
16. SIMAT (2012), http://www.sma.df.gob.mx/simat/
17. Muñoz, R., Carmona, M.R., Pedroza, J.L., Granados, M.G.: Data analysis of PM2.5 registered with TEOM equipment in Azcapotzalco (AZC) and St. Ursula (SUR) stations of the automatic air quality monitoring network (RAMA). In: National Congress of Medicine Engineering and Ambient Sciences, pp. 21–24 (2000) (in Spanish)

18. Pérez, P., Trier, A., Reyes, A.: Prediction of PM2.5 concentrations several hours in advance using neural networks in Santiago, Chile. Atmospheric Environment 34, 1189–1196 (2000)
19. Cobourn, W.G.: An enhanced PM2.5 air quality forecast model based on nonlinear regression and back-trajectory concentrations. Atmospheric Environment 44, 3015–3023 (2010)
20. Salini, G., Perez-Jara, P.: Time series analysis of atmosphere pollution data using artificial neural networks technique. Revista Chilena de Ingeniería 14(3), 284–290 (2006)
21. Dong, M., Yang, D., Kuang, Y., He, D., Erdal, S., Kenski, D.: PM2.5 concentration prediction using hidden semi-Markov model-based times series data mining. Expert Systems with Applications 36, 9046–9055 (2009)
22. Kang, D., Mathur, R., Trivikrama Rao, S.: Assessment of bias-adjusted PM2.5 air quality forecast over the continental United States during 2007. Geoscience Model Dev. 3, 309–320 (2010)