Neural Comput & Applic
DOI 10.1007/s00521-015-1927-7

EANN

HISYCOL a hybrid computational intelligence system for combined machine learning: the case of air pollution modeling in Athens

Ilias Bougoudis[1] · **Konstantinos Demertzis**[1] · **Lazaros Iliadis**[1]

Received: 8 January 2015 / Accepted: 16 May 2015

Abstract The analysis of air quality and the continuous monitoring of air pollution levels are important subjects of the environmental science and research. This problem actually has real impact in the human health and quality of life. The determination of the conditions which favor high concentration of pollutants and most of all the timely forecast of such cases is really crucial, as it facilitates the imposition of specific protection and prevention actions by civil protection. This research paper discusses an innovative threefold intelligent hybrid system of combined machine learning algorithms HISYCOL (henceforth). First, it deals with the correlation of the conditions under which high pollutants concentrations emerge. On the other hand, it proposes and presents an ensemble system using combination of machine learning algorithms capable of forecasting the values of air pollutants. What is really important and gives this modeling effort a hybrid nature is the fact that it uses clustered datasets. Moreover, this approach improves the accuracy of existing forecasting models by using unsupervised machine learning to cluster the data vectors and trace hidden knowledge. Finally, it employs a Mamdani fuzzy inference system for each air pollutant in order to forecast even more effectively its concentrations.

✉ Lazaros Iliadis
liliadis@fmenr.duth.gr

Ilias Bougoudis
ibougoudis@yahoo.gr

Konstantinos Demertzis
kdemertz@fmenr.duth.gr

[1] Democritus University of Thrace, 193 Pandazidou St., 68200 Orestiada, Greece

Keywords Ensembles learning · Ensembles of classifiers · Fuzzy inference systems · Feedforward neural network · Random forest · Air pollution

1 Introduction

The increase in the human population and the growth of the productive process during the years led to a series of negative environmental consequences. This fact is the cause of several health problems of human beings and living creatures in general. Air pollution is one of the most characteristic examples of environmental burden caused by human activity. This research effort deals with the following primary air pollutants (which are directly emitted by human actions) CO, NO, NO_2, SO_2 and with one secondary pollutant (caused by chemical reactions) the ozone O_3. The chemical composition and the characteristics of all pollutants cause well-known problems in the human respiratory system and hospitalization for heart or lung diseases, and also, they are favoring the development of various types of cancer. Due to their dissimilarity and to the distinct mechanisms that they are using to enter the atmosphere, it is difficult to model their concentrations and to estimate their exact consequences in human health. An effective quantitative estimation of their impact requires an integrated spatiotemporal analysis of the conditions that favor their concentration and the determination of the relations between the air pollutants (APOL) and between the pollutants and meteorological factors. Of course the problem is monitored and watched mainly in major urban centers.

Forecasting VAP is really important so that civil protection authorities can impose specific prevention or warning protection measures aiming to protect the population.

Published online: 10 June 2015

It is very positive that modern computational intelligence and machine learning technologies offer the proper mechanisms capable of forecasting APOL values.

This research proposes an innovative ensemble and fuzzy inference system entitled HISYCOL that forecasts the concentrations of air pollutants, and it reaches proper decisions toward protection of the urban centers people. Obviously, it is based on combined various computational intelligence methodologies.

More specifically, this paper proposes a new effective and reliable hybrid system that is based on the combination of unsupervised clustering, ANN and random forest ensembles and fuzzy logic.

The general framework of the proposed model comprises of the following stages: (a) Unsupervised clustering of the initial dataset is executed in order to re-sample the data vectors. (b) Ensemble ANN modeling is performed using combination of machine learning algorithms. (c) Finally, the last stage comprises of the optimization of the modeling performance with a Mamdani rule-based fuzzy inference system that exploits the relations between the parameters affecting the concentrations of APOL. More specifically, self-organizing maps (SOM) are used to perform dataset re-sampling, then ensembles of feedforward artificial neural networks (FFANN) and random forests (RAF) are trained on the clustered data vectors, and finally, the obtained models are optimized by using a fuzzy inference system.

1.1 Literature review: motivation

In an earlier research effort of our team [1], we have made an effort to get a clear and comprehensive view of the air quality in the wider urban center of Athens and also in the Attica basin. This study was based on data that were selected from the air pollution measuring stations of the area during the temporal periods (2000–2004, 2005–2008 and 2009–2012). This method was based on the development of 117 partial ANNs whose performance was averaged by using an ensemble learning approach. The system used also fuzzy logic in order to forecast more efficiently the concentration of each pollutant. The results showed that this approach outperforms the other five ensemble methods.

There are other similar studies in the literature that are trying to forecast the air pollution values [16, 18–20]. However, they have certain limitations that do not guarantee their generalization ability. More specifically, they train ANN models with data related to a narrow area (e.g., city center), and they consider this data sample as representative of a wider area that covers locations varying from a topographic, microclimate or population density point of view. For example, paper [23] predicts particulate matter concentrations in India, using data from only two stations, paper [4] uses data from ten stations in order to figure out an air pollution picture for the whole country of Belgium, while paper [6] uses only four stations for the city of Istanbul. Also there are important seasonal studies in the literature that do not offer more generalized annual solutions. For instance, paper [25] uses only summer records in order to train the developed neural networks. Moreover, paper [17] describes a model that estimates ozone concentrations, based on a limited data volume. Kunwar et al. [11] used a hybrid approach which selects a subset of the involved features by employing principal components analysis (PCA). It is a model combining three ensemble learning methods applied in the area of Lucknow, India. It is worth to mention that they tried to interpolate the output to cases with different climate conditions with limited results.

An interesting approach [21] blending time series with multi-linear regression ANNs in order to achieve acceptable forecasting accuracy based on limited air quality and meteorological data vectors was proposed for the case of Temuco, Chile. In this place, residential wood burning is a major pollution source during cold winters. The described model considered a limited volume of surface meteorological and PM_{10} primitive data [21].

This paper aims to overcome the above limitations, by providing more generalized models that have emerged after considering reasonable and representative amount of data from all types of measuring stations. It is rational that such ANNs can be effectively applied in wider areas. Furthermore, a main objective was to combine machine learning techniques, in order to achieve better convergence for the developed models.

Paper [8] was an inspiration to use ensemble neural networks (ENNs). More specifically, in [8], it is stated that ensemble methods may be more effective than single ANN approaches. The research described in [8] was held in China, and it introduced ENNs for pollutant's estimation.

Additionally, in our previous work [1], we had already created ENNs for this purpose. In this research, due to the individuality and particularity of each residential area of Athens, separate local ANNs had to be developed, capable of performing reliable interpolation of missing data vectors on an hourly basis. Also due to the need for hourly overall estimations of pollutants in the wider area of a major city, ANN ensembles were additionally developed by employing four existing methods and an innovative fuzzy inference approach.

In paper [12], the relationships between the ensembles and their comprising ANNs are analyzed aiming to create a set of nets with the use of a sampling technique. This technique is such that each net in the ensemble is trained on a different subsample of the training data. Also [22]

Fig. 1 Measuring stations in the Attica basin

performs a review of the existing ensemble techniques, and it can serve as a tutorial for practitioners who are interested in building such systems (e.g., ENNs). As a result, papers [13, 27] were very useful, as they provided the theoretical background for our research.

Summarizing all of the above, it is a fact that the motivation for this research was the development of a hybrid model capable of absorbing and overcoming the problem of bad local behaviors of the existing ones. The main idea was that such an approach would require ANN ensembles applied on homogenous data clusters and not in randomly divided datasets. This could add much more efficiency to an air pollution forecasting system. Additionally, a fuzzy inference system could act as an optimizer to improve further the reliability of the model. The design, development and application of this model are described in the following paragraphs.

1.2 Data

The data used are related to nine air pollution measuring stations and two meteorological ones located in the Attica basin (as seen in Table 2). Every station counts hourly values for CO, NO, NO_2, O_3 and SO_2. All the values are counted in $\mu g/m^3$, except from CO which is measured in mg/m^3. The time period of this research starts from 2000 and finishes in 2012. Additionally, every record in each measuring station includes five temporal data, namely *Year*, *Month*, *Day*, *Day_Id* (1 for Monday, 2 for Tuesday and so on) and *Hour* value. Moreover, six meteorological factors are considered, namely *Air Temperature* (Air_Temp), *Relative Humidity* (RH), *Atmospheric Pressure* (PR), *Solar Radiation* which is not included for 2013 (SR), *Wind Speed* (WS) and *Wind Direction* (WD), and finally the measuring stations code *Station*. As it is seen in Table 2, the meteorological data are related to the "Penteli" and "Thiseion" stations. Figure 1 shows the location of the measuring station in the basin.

Table 1 Statistical analysis of the whole SOM dataset

SOM (5,12,971 records)	CO	NO	NO_2	O_3	SO_2
MAX	24.6	953	533	320	445
MIN	0.1	1	0	1	2
MODE	0.4	4	32	3	2
COUNT_MODE	45,532	39,003	5786	21,660	75,081
AVERAGE	1.41	51.62	57.36	41.70	13.21
STANDARD_DEV	1.46	81.04	35.24	36.26	16.44

The selected data were stored in an integrated dataset that comprises of the vectors related to all measuring stations except the ones of "Agia Paraskevi" and "Aristotelous" for which there is a serious problem of missing data for the whole period under research. Table 1 presents a descriptive statistical analysis of the dataset on which this research was based.

1.3 Data preprocessing

Data preprocessing aims to phase various problems that emerge during their gathering like the manipulation of missing values, the tracking of extreme values and the transformation of data so that they can be proper input for the learning algorithms.

1.3.1 Missing data

Missing data is one of the most serious problems when trying to develop a rational and effective model. The dispersion of missing values was estimated, and after confirming their random appearance, the following approaches toward overcoming this problem were studied, by taking into consideration their advantages and disadvantages.

Replace missing values with sample mean or mode

- Advantage:
  - Can use complete case analysis methods
- Disadvantages:
  - Reduces variability
  - Weakens covariance and correlation estimates in the data (because ignores relationship between variables)

Dummy variable adjustment

- Advantage:
  - Uses all available information about missing observation
- Disadvantages:
  - Results in biased estimates
  - Not theoretically driven

Replacement of missing values with predicted scores from a regression equation

- Advantage:
  - Uses information from observed data
- Disadvantages:
  - Overestimates model fit and correlation estimates
  - Weakens variance

Identification of the set of parameter values that produces the highest likelihood

- Advantages:
  - Uses full information (both complete cases and incomplete cases) to calculate likelihood
  - Unbiased parameter estimates with missing at random data
- Disadvantage:
  - Complexity of model.

Discarding all missing values

- Advantages:
  - Simplicity
  - Comparability across analyses
- Disadvantages:
  - Reduces statistical power
  - Does not use all information

Such malfunctions are divided into two basic categories. The first type is the so-called partial deficiencies where measuring stations malfunction for a long time that may last up to some months. The second is known as "total deficiencies" which occur when a station does not measure a pollutant for a long scaled time period which may last for years, e.g., 2000–2012 or even for a wider period, e.g., from 2005 till today. In both cases, missing data records were excluded for the whole related time period.

Table 2 shows a brief presentation of the measuring stations with statistical data related to missing air pollutants' values.

1.3.2 Extreme values

The determination of the extreme air pollution concentrations with the inter quartile range (IQR) method is a purely statistical data preprocessing approach, which locates the divergent dataset values. In fact IQR detects extreme values that can potentially cause "noise" and lead to generalization incapability. For example, there might be some CO values much higher than the upper statistical boundary of average $+3\sigma$ (where σ is the standard deviation). These values are considered outliers and moreover extreme ones.

Table 2 Statistics of measuring stations

ID	Station's name	Code	Missing values (%)	Correct data vectors	Station's data
1	Ag. Paraskevi	AGP	12.32	99.936	O_3, NO, NO_2, SO_2
2	Amarusion	MAR	21.58	89.371	O_3, NO, NO_2, CO
3	Peristeri	PER	33.61	75.668	O_3, NO, NO_2, CO, SO_2
4	Patision	PAT	10.45	102.068	O_3, NO, NO_2, CO, SO_2
5	Aristotelous	ARI	16.76	94.873	NO, NO_2
6	Geoponikis	GEO	26.84	83.381	O_3, NO, NO_2, CO, SO_2
7	Piraeus	PIR	33.67	75.600	O_3, NO, NO_2, CO, SO_2
8	N Smyrnh	SMY	26.06	84.272	O_3, NO, NO_2, CO, SO_2
9	Penteli	PEN	3.66	109.806	Meteorological station
10	Thiseion	THI	0.30	113.632	Meteorological station
11	Athinas	ATH	21.86	89.058	O_3, NO, NO_2, CO, SO_2

An extreme value (outlier) is a point that lies far away from the mean value of a feature. The distance is usually measured as a multiple of the standard deviation (SD). For a parameter that follows normal distribution, a distance equal to twice the SD covers 95 % of the expected values, whereas this percentage grows to 99 % when we are dealing with a distance three times the SD. Data records with values far away from the mean value are the cause of serious errors in the training phase, and they have destructive results. This bad influence gets even worse when the extreme values are due to noise that has emerged during measuring. If the number of the extreme values is small, then the corresponding records can be removed from the dataset and they can be analyzed independently. The IQR approach was used to trace the extreme values. The IQR method locates outliers based on the inter quartile scales. The quartiles divide the dataset to four equal parts. The IQR is the difference between the third ($Q3$) and the first ($Q1$) quartile, IQR = $Q3-Q1$, which includes the intermediate 50 % of the data, whereas the rest 25 % is less than $Q1$ and the other 25 % is greater than $Q3$. The calculation process of the extreme values is presented by the following equations below [26]:

Outliers:

$$\begin{aligned} &Q3 + \text{OF} \times \text{IQR} < x \leq Q3 + \text{EVF} \times \text{IQR} \quad \text{or} \\ &Q1 - \text{EVF} \times \text{IQR} \leq x < Q1 - \text{OF} \times \text{IQR} \end{aligned} \tag{1}$$

Extreme value:

$$x > Q3 + \text{EVF} \times \text{IQR} \quad \text{or} \quad x < Q1 - \text{EVF} \times \text{IQR} \tag{2}$$

Key:

$Q1 = 25$ % quartile, $Q3 = 75$ % quartile, IQR = interquartile range (difference between $Q1$ and $Q3$), OF = outlier factor, EVF = extreme value factor.

It should be mentioned that the extreme values in this case are what we are looking for; because based on their determination civil protection authorities should be activated to take all necessary actions. For this reason, the EXDV were not removed or isolated from the dataset, but they were used to create objective training data samples that would enable the development of models capable of generalizing. In this way, the developed models would respond to new data from other measuring stations or other cities quite efficiently. After using the above method, 31,857 vectors were characterized as outliers and 7459 ones were found to be related to extreme values.

1.3.3 Data normalization

Data normalization was performed for the concentrations of air pollutants, in order to phase the problem of prevalence of features with wider range over the ones with a narrower range, without being more important. The result was to keep all of their values in the closed interval [−1, +1] by using Eq. 3:

$$x_{1_{\text{norm}}} = 2 \times \left(\frac{x_1 - x_{\min}}{x_{\max} - x_{\min}} \right) - 1, \quad x \in R \tag{3}$$

2 Theoretical background

2.1 Ensemble learning

Ensemble methods [22] use multiple learning algorithms to obtain better predictive performance than could be obtained from any of the constituent learning algorithms. Usually, they refer only to a concrete finite set of alternative models, but typically they allow for much more flexible structures to exist between those alternatives. Also, they are primarily used to improve the performance of a model, or to reduce the likelihood of an unfortunate selection of a poor one. Other applications of ensemble learning include assigning a confidence to the decision made by the model, selecting optimal (or near optimal) features, data fusion, incremental learning, non-stationary learning and error correcting.

The novel concept of combining learning algorithms is proposed as a new direction of ensemble methods for the improvement of the performance of individual algorithms. These algorithms could be based on a variety of learning methodologies and could achieve different ratios of individual results. The goal of the ensembles of algorithms is to generate more certain, precise and accurate system results. Numerous methods have been suggested for the creation of ensembles of learning algorithms:

- Using different subsets of training data with a single learning method.
- Using different training parameters with a single training method (e.g., using different initial weights or learning methods for each neural network in an ensemble).
- Using different learning methods.

Herein the third approach was applied in order to develop the ANN ensembles. The ensemble learning is realized with feedforward neural networks and random forest algorithms, and it was applied in four clusters (subsets of the original dataset).

2.1.1 Feedforward artificial neural networks

FFNN are biologically inspired regression and classification algorithm. They consist of a (possibly large) number of simple neuron-like processing units organized in three

layers, namely input, hidden and output layer. The information moves forward, from the input nodes, through the hidden nodes and to the output nodes. Every unit in a layer is connected with all the units in the previous layer. These connections are not all equal; each connection between the *i*th and the *j*th neuron may have a different strength or weight w_{ij}. The weights on these connections encode the knowledge of a network. Actually, the weight coefficient reflects the degree of importance of the given connection in the ANN [3].

Often the units in a neural network are also called nodes. Data enters at the inputs and passes through the network, layer by layer, until it arrives at the outputs [5]. The output function is presented by Eq. 4, where x_i are the input vectors, y_j the output vectors, w_{ij} the synaptic weights, Θ_j the bias or external threshold and Ψ is the activation function.

$$y_j = \psi\left(\sum_{i=1}^{n} w_{ji}x_i + \Theta_j\right) \tag{4}$$

The architecture and the learning algorithm plus the transfer function used for the development of the MLFF ANN are presented clearly in Sect. 3.2.2.

2.1.2 Random forests

Random forests (RF) are an ensemble learning method for classification and regression that operate by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes output by individual trees. The algorithm for inducing a random forest was developed by Breiman and Adele Cutler in 2001 [2]. It is a popular and efficient algorithm very powerful for prediction, based on model aggregation ideas, for classification and regression. The principle of random forests is to combine many binary decision trees built using several bootstrap samples coming from the set of observations and choosing randomly at each node a subset of explanatory variables.

The training algorithm for RF applies the general technique of bootstrap aggregating, or bagging, to tree learners. More specifically, at each node, a given number of input variables are randomly chosen and the best split is calculated only within this subset. Clearly, no pruning step is performed so all the trees are maximal ones. In the random forest framework, the most widely used score of importance of a given variable is the increase in the mean of the error of a tree in the forest when the observed values of this variable are randomly permuted in the bag sample. The higher the importance is the stronger the variable influence [7].

So far to the best of our knowledge, only a limited number of research efforts have been reported in the literature on air quality modeling using random forests. More details about their related algorithms can be found in Sect. 3.2.1.

2.2 Competitive learning neural networks

Competitive learning neural networks (CLNN) include a competitive layer (CLA) comprising of competitive neurons (CNE) (Fig. 1). Every $\mathrm{CNE_i}$ is characterized by a weight vector $w_i = (w_{il},\ldots,w_{id})^{\mathrm{T}}$, $i = 1,\ldots,M$, and it estimates a similarity measure with the input data vector $x_i = (x_{il},\ldots,x_{id})^{\mathrm{T}}$ $x \in R$. For every input vector that is introduced to the network, there is a competition between the CNE for the determination of the winning neuron. The winner is the neuron that has the higher degree of similarity between the input vector and its assigned weight vector. The output of the winning CNE is set to $o_m = 1$, whereas for the rest of the neurons the output is $o_i = 0$, $i = 1,\ldots,M$, $i \neq m$. The default similarity function used is the inverse value of the actual Euclidean distance x–w_i between the input vector x^n and the weight vector w_i [3].

2.2.1 Self-organizing maps

Self-organizing map (SOM) is a widely used ANN architecture, which is based on competitive learning, and it was proposed by Kohonen [10] in the mid-1980s. It was proposed in an effort to model the self-organization which is performed in the human brain. A SOM network includes the input layer and the layer with the CNE which are organized in the form of a two-dimensional lattice (Fig. 2). Each competitive neuron is assigned a weight vector $w_i = (w_{il},\ldots,w_{id})^{\mathrm{T}}$. When an input $x = (x_1,\ldots,x_d)^{\mathrm{T}}$ $x \in R$ is

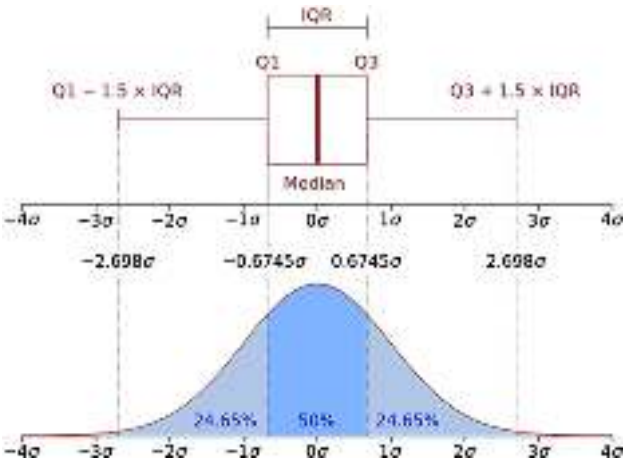

Fig. 2 Graphical display of the IQR method

applied, the lattice neurons compete each other and the result is the determination of the winning neuron m, whose vector w_m appears to have the best similarity with the vector x. Consequently, SOM implements a mapping of the input x that has dimension d to the coordinates of the lattice $r_m = (z_{m1}, z_{m2})^T$ [3, 10].

2.3 Fuzzy inference systems

A fuzzy inference system (FIS henceforth) [15] is a way of mapping an input space to an output space using fuzzy logic. FIS uses a collection of fuzzy membership functions and rules, instead of Boolean logic, to reason about data. The rules in FIS (sometimes may be called as fuzzy expert system) are fuzzy production rules of the form:

If p Then q, where p and q are fuzzy statements (*Linguistics*)

The antecedent describes to what degree the rule applies, while the conclusion assigns a fuzzy function to each of one or more output variables. The set of rules in a fuzzy expert system is known as knowledge base. There are two popular versions of fuzzy inference systems: Mamdani and Tagasi Sugeno type [24]. The first one was applied herein. The Mamdani-type FIS was proposed in 1975 [14] as an attempt to control a steam engine and boiler combination by synthesizing a set of linguistic control rules obtained from experienced human operators.

The algorithm of a Mamdani FIS is as follows:

1. Determining a set of fuzzy rules.
2. Fuzzifying the inputs using the input membership functions.
3. Combining the fuzzified inputs according to the fuzzy rules to establish rule strength.
4. Finding the consequence of the rule by combining the rule strength and the output membership function (if it is a Mamdani FIS).
5. Combining the consequences to get an output distribution.
6. Defuzzifying the output distribution [this step applies only if a crisp output (class) is needed].

3 Description of the HISYCOL

A hybrid machine learning system entitled HISYCOL has been developed which had the following targets: (a) to study in depth the conditions parameters under which high concentrations of air pollutants emerge, (b) to determine the correlation degree between these features and (c) to have the optimal forecasting and decision making efficiency. The algorithmic steps of this system are presented below (Fig. 3):

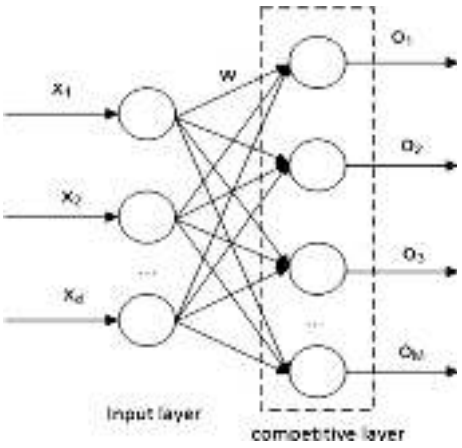

Fig. 3 General architecture of competitive learning

Step 1: Re-sampling εἶναι is one of the techniques used to achieve ensample learning. The first and innovative step was the segmentation of the general dataset in clusters by the employment of SOM. This resulted in the clustering of the data, based on the concentrations of air pollutants and the creation of hierarchical chains of correlated data structures. In this way, re-sampling of the data was performed not in a random manner but using unsupervised machine learning. Consequently, instead of using the whole dataset in the regression that was performed in the next step, or dividing it randomly, clustering was employed. The aim was to divide the dataset in four clusters, which means that the data vectors of each group would be related to each other and they would share common characteristics. Afterward, separate regressions were performed for each cluster by employing the ANN ensemble approach.

The SOM clustering algorithm is discussed in Sect. 3.1. From the clustering, we obtained four new datasets named as SOM-i (where $i = 0$–3) which are described in details in Sect. 3.1.1.

Step 2: Then for every new dataset, ANN regression was performed in order to develop models capable of forecasting the values of each air pollutant separately. More specifically, the following features were the independent parameters: *Year*, *Month*, *Day*, *Day_Id*, *Hour*, *AirTemp*, *Relative Humidity*, *Atmospheric Pressure*, *Solar Radiation*, *Wind Speed*, *Wind Direction*, *Station_Id*, including the air pollution ones CO, NO, NO_2, O_3, SO_2, whereas one of the air pollutants was excluded each time from the independent parameters vector and it was considered as the depended parameter. For example, the first time CO was not in the depended feature, the second time NO was the depended and so on.

Feedforward neural networks (FFNN), random forest ANN (RAF), ε-support vector regression, linear regression, k-nearest neighbors and radial basis function ANN were used for the regression. However, the FFNN and the RAF ones were chosen to be applied in the final model because they outperform by far the other approaches, and they generalize much better in testing. As it was mentioned in step 1, distinct regressions were performed for each cluster of data vectors.

The *tenfold cross-validation* and the "repeat random subsampling validation" were employed to enhance generalization of the emerged optimal models. Regression with the FFNN and with the RAF approach are described in Sects. 3.2.1 and 3.2.2, and the results of the regressions are presented in Sect. 3.2.3.

Step 3: The second innovation of the HISICOL is in the way it assigns and handles a new fuzzy value after the process of combined learning is fulfilled. More specifically, the RMSE and R^2 results from the above regression algorithms for every air pollutant were used as input vectors to a FIS, from which we obtain fuzzy values with the process described in Sect. 3.3, whereas the results are presented in Sect. 4.

The overall algorithmic approach that was proposed herein is described clearly and in details in Fig. 4.

3.1 SOM algorithm

A SOM network clustered the available data in order to create partitions of correlated data vectors and thus to perform re-sampling. The algorithm comprises of the following steps (Fig. 5):

First, the weight vectors $w_i = (w_{il},\ldots,w_{id})^{\mathrm{T}}$ are randomly initialized to small values with a random generator function. Then, the following procedures are executed as follows:

The application of the SOM algorithm starts with the initialization of the weight vector $w_i = (w_{il},\ldots,w_{id})^{\mathrm{T}}$ by using random generation functions. Afterward, the following three basic procedures are executed:

A. Competition: For every training vector sample x^n, the neurons calculate the similarity function value, where the neuron with the highest value is the winner. The Euclidean distance between the input vector $x = (x_1,\ldots,x_d)^{\mathrm{T}}$ $x \in R$ and the weight vector $w_i = (w_{il},\ldots,w_{id})^{\mathrm{T}}$ of the competing neurons is the similarity function.

B. Cooperation: The winning neuron i defines its topological $h_{j,j}$ from the surrounding neurons who adjusted their weights to the input vector. The distance between the winning neuron i and neuron j is symbolized as $d_{j,i}$ so that the topological neighborhood $h_{j,i}$ is a function of $d_{j,i}$ which satisfies two conditions:

   B1. It should be symmetric to the point of the local minimum (point of winning neuron) where $d_{j,i} = 0$.

   B2. The amplitude of the function should be reduced monotonically as the distance $d_{j,i}$ from the winning neuron increases. The function that satisfies the above limitations and was used in this research is the following Gaussian

$$h_{j,i}(x) = \exp\left(-\frac{d_{j,i}^2}{2\sigma^2}\right) \tag{5}$$

   where σ is the effective width of the topological neighborhood, which defines the degree of participation of the winning neuron neighborhood neurons to the training phase. The value of this parameter is reduced in every epoch according to the function below

$$\sigma(n) = \sigma_0 \exp\left(-\frac{n}{\tau_1}\right), \quad n = 0, 1, 2, \ldots \tag{6}$$

   It should be mentioned that σ_0 is the initial value of the effective width and

$$\tau_1 = \frac{n_0}{\ln(\sigma_0)} \tag{7}$$

C. Synaptic weight adaption: In this last training stage, the weight vectors of the competitive neurons are updated. The value of this change is given by the following equation:

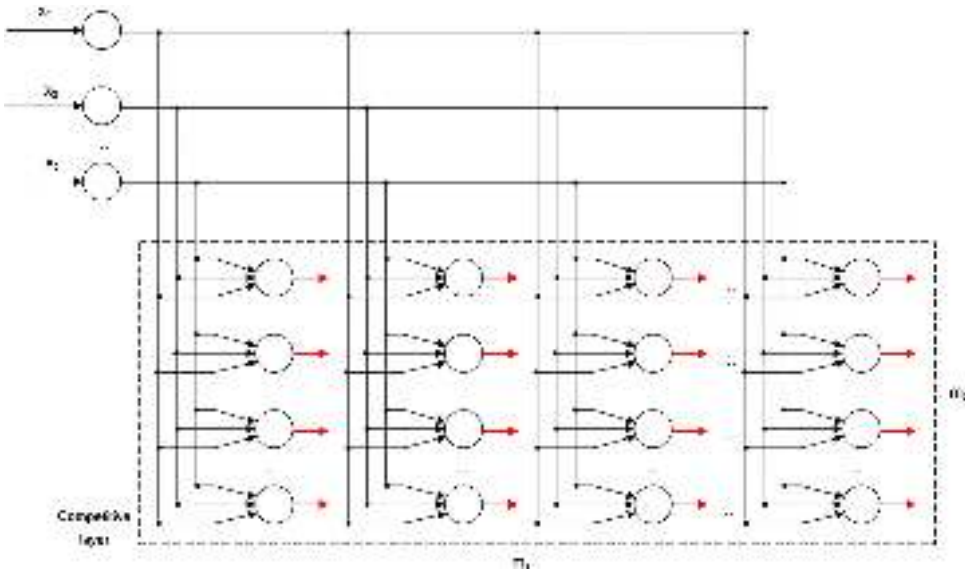

Fig. 4 Self-organizing map (d inputs and two-dimensional lattice $m_1 \times m_2$)

$$\Delta w_j = \eta h_{j,i(x)}(x - w_j) \tag{8}$$

where i is the winning neuron and j is a neuron in its neighborhood. Given the weight vector $w_j(n)$ for a specific time point n, we estimate the new vector for the moment $n + 1$ from the following equation:

$$w_j(n+1) = w_j(n) + \eta(n)\, h_{j,i(x)}(n)\big(x(n) - w_j(n)\big). \tag{9}$$

The learning rate $\eta(n)$ starts from the value around 0.1 and it is gradually reduced to 0.01 by using the above relation.

3.1.1 Clustering results

After the termination of the clustering process (following the procedure described above), four datasets were created. Each one contained the data vectors that were assigned to each cluster, namely (SOM-0, SOM-1, SOM-2 and SOM-3). The final shape of the data set includes the following independent parameters: (a) temporal attributes (***Year***, ***Month***, ***Day***, ***Day_Id***, ***Hour***), (b) the meteorological features values (***AirTemp***, ***RH***, ***PR***, ***SR***, ***WS***, ***WD***), (c) the Station_Id (*Station*) and (d) the values of all the air pollutants, except for the one pollutant that was left each time out to serve as the depended parameter. Tables 3, 4, 5 and 6 present a descriptive statistical analysis of the four related datasets.

The basic conclusions that were obtained from the clustering process were related to the high concentrations of the secondary air pollutant O_3 and to the determination of the conditions that favor its high values. In the cluster SOM_0, we had the highest O_3 concentrations. It is worth mentioning that the average O_3 value for the SOM_0 cluster was 57.8 μg/m^3, whereas the average value for the whole dataset was 41.69. In the SOM_0 dataset, we had quite high temperature values (almost 30 % higher than the average of the other three clustered datasets) with predominant value 24.9° Celcium and an average value equal to 23.21°. Also the same data cluster had a moderate average relative humidity equal to 50.55 %. As long as the calendar characteristics are concerned, the dominant month for high O_3 values was July and the time was 13:00 hours pm.

The rest of the clusters comprised of data vectors with conversely considering characteristics. For example, December was the most common month of the data vectors and night hours were the most dominant. Interesting characteristics with good practical value were located, and they plead the application of a model that exports the hidden knowledge of the related data. More specifically, SOM_1 cluster was characterized by the 16 hpm which was the dominant time of the days. Data vectors of the SOM_3 cover a time period from 6 o'clock in the morning till the first afternoon hours. The main attribute of the SOM-2 is that its data vectors comprise of characteristics related only to week days.

3.2 Regression

Regressions were performed in order to develop ANN ensemble models capable of estimating each depended variable (each separate air pollutant) efficiently with the two approaches FFNN and RAF. The four datasets obtained from the SOM clustering were used to perform these regressions.

So for each pollutant two ANN ensembles were developed (one with FFNN and one with RAF) each ensemble comprising of four ANN using the same algorithm, but

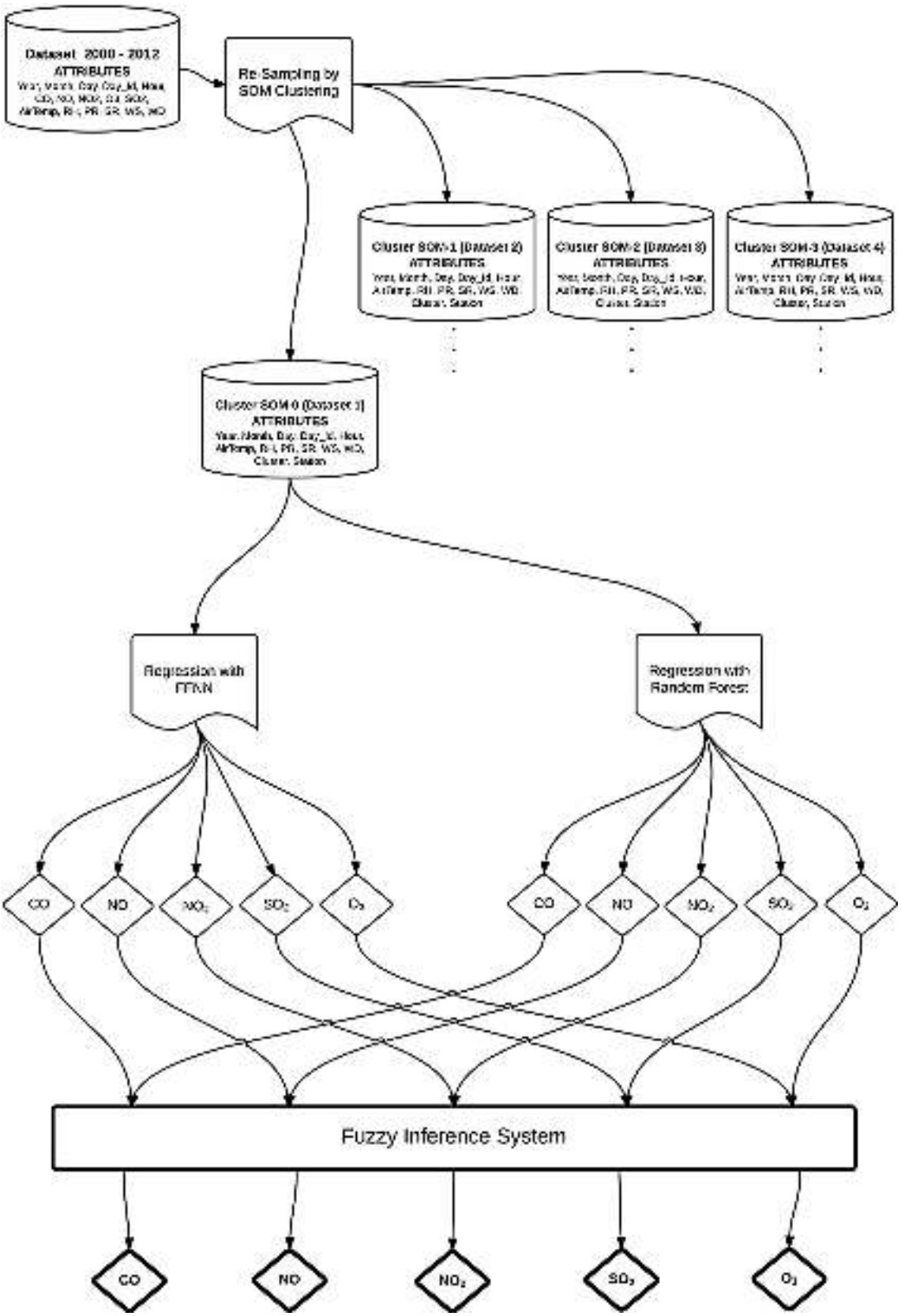

Fig. 5 Structure of the hybrid system of combined machine learning (HISYCOL) algorithm

Table 3 Statistical analysis of the dataset SOM-0

SOM-0 (1,63,384 records)	CO	NO	NO_2	O_3	SO_2
MAX	21.4	911	533	320	293
MIN	0.1	1	0	1	2
MODE	0.3	4	19	4	2
COUNT_MODE	18,613	13,743	2166	2431	21,877
AVERAGE	1.23	40.82	58.19	57.81	14.57
STANDARD_DEV	1.29	65.70	41.57	40.11	17.47

Table 4 Statistical analysis of the dataset SOM-1

SOM-1 (1,46,334 records)	CO	NO	NO_2	O_3	SO_2
MAX	22.9	914	264	164	290
MIN	0.1	1	0	1	2
MODE	0.3	1	48	3	2
COUNT_MODE	11,574	9251	1832	11,874	23,708
AVERAGE	1.51	67.84	56.70	26.46	11.45
STANDARD_DEV	1.59	98.28	30.21	27.47	13.50

Table 5 Statistical analysis of the dataset SOM-2

SOM-2 (1,06,308 records)	CO	NO	NO_2	O_3	SO_2
MAX	24.6	953	319	243	380
MIN	0.1	1	1	1	2
MODE	0.4	4	60	4	2
COUNT_MODE	8585	8248	1279	4366	14,363
AVERAGE	1.55	53.83	60.24	37.65	13.80
STANDARD_DEV	1.56	82.67	33.62	32.83	17.47

Table 6 Statistical analysis of the dataset SOM-3

SOM-3 (96,945 records)	CO	NO	NO_2	O_3	SO_2
MAX	19.6	825	350	223	445
MIN	0.1	1	0	1	2
MODE	0.4	4	20	3	2
COUNT_MODE	8925	8162	1231	3249	15,133
AVERAGE	1.38	42.89	53.77	41.98	12.94
STANDARD_DEV	1.40	68.74	31.93	33.42	17.26

applied on a different dataset namely (SOM-0, SOM-1, SOM-2, SOM-3).

So keeping in mind that the air pollutants of interest were five, totally ten ANN ensembles were developed for each ANN algorithm.

3.2.1 Regression with random forests

This methodology applies the general technique of bootstrap aggregating (also known as bagging) to tree learners. Actually, bootstrap aggregating is a machine learning ensemble meta-algorithm which is designed to improve the stability and the accuracy of machine learning algorithms that are used in regression. Moreover, it is really important that it reduces variance and helps to avoid over-fitting.

To estimate the extreme air pollutants' values with the use of the random forests algorithm and for a given training set $X = x_1,\ldots, x_n$ with responses $Y = y_1,\ldots,y_n$, bagging repeatedly it selects a random sample with replacement of the training set and fits trees to these samples.

More specifically, the algorithm goes as follows:

For $b = 1,\ldots,B$:

- Sample, with replacement, n training examples from X, Y call these X_b, Y_b.
- Train a decision or regression tree f_b on X_b, Y_b.
- After training, predictions for unseen samples x' can be made by averaging the predictions from all the individual regression trees on x' by using Eq. 10:

$$\hat{f} = \frac{1}{B}\sum_{b=1}^{B}\hat{f}_b(x') \tag{10}$$

This can be done also by taking the majority vote in the case of decision trees [2].

It should be specified that B is a free parameter corresponding to the number of samples/trees.

3.2.2 Regression with FFNN

When FFNN are used, the training process includes the following steps [3]:

- The weighted sums of the inputs are calculated based on function 11:

$$s_j = \sum_{i=1}^{n} \left(W_{ij} X_i\right) - \theta_j \quad j = 1, 2, \ldots, h \tag{11}$$

where n, h and m are the number of input, hidden and output nodes, respectively, and W_{ij} is the connection weight from the ith node of the input layer to the jth node of the hidden layer. Also θ_j *is the* bias (threshold).

The output for each hidden node is estimated with Eq. 12:

$$S_j = \text{sigmoid}_{(s_j)} = \frac{1}{\left(1 + \exp\left(-s_j\right)\right)} \quad j = 1, 2, \ldots, h \tag{12}$$

The final output is estimated based on Eqs. 11 and 12 [3, 5]:

$$o_k = \sum_{j=1}^{h} \left(W_{jk} S_j\right) - \theta'_k \quad k = 1, 2, \ldots, m \tag{13}$$

$$O_k = \text{sigmoid}_{(o_k)} = \frac{1}{(1 + \exp(-o_k))} \quad k = 1, 2, \ldots, m \tag{14}$$

The learning technique which used is the back propagation in order to calculate the gradient of the loss function with respect to all the weights in the ANN. The gradient is fed to the optimization method which uses it to update the weights, in an attempt to minimize the loss function. In this research, we have developed the FFNN by employing the following parameters:

- Training function: *TRAINBR*
- Learning function: *LEARNGDM*
- Transfer function: *TANSIG*
- Performance function: *MSE*

3.2.3 Regression results

Root mean square error (RMSE) and the coefficient of determination (regression R^2) were used as metrics to evaluate the performance of the ANN regression ensemble models. Each of the above criteria is represented by an index which is obtained by comparing the forecasted values of each air pollutant to the actual ones in training, validation and testing phases. The tenfold partitions average values of these indices were produced by the use of *tenfold cross-validation* in each one of the various architectures that were tried in order to minimize the error.

Root mean square error (RMSE) or root mean square deviation (RMSD) is a frequently used measure of the differences between value (sample and population values) predicted by a model or an estimator and the values actually observed. Basically, it represents the sample standard deviation of the differences between predicted values and observed values. These individual differences are called residuals when the calculations are performed over the data sample that was used for estimation, and are called prediction errors when computed out of sample. The RMSE serves to aggregate the magnitudes of the errors in predictions for various times into a single measure of predictive power. RMSE is a good measure of accuracy, but only to compare forecasting errors of different models for a particular variable and not between variables, as it is scale dependent.

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{N} (O_i - Pi)^2}{N}} \tag{15}$$

Coefficient of determination (regression R^2) in linear least squares regression with an estimated intercept term equals the square of the Pearson correlation coefficient between the observed and modeled (predicted) data values of the dependent variable. Under more general modeling conditions, where the predicted values might be generated from a model different from linear least squares regression, the R^2 value can be calculated as the square of the correlation coefficient between the original and modeled data values. In this case, the value is not directly a measure of how good the modeled values are, but rather a measure of how good a predictor might be constructed from the modeled values (by creating a revised predictor of the form $\alpha + \beta f_i$). This usage is specifically the definition of the term "coefficient of determination": the square of the correlation between two (general) variables.

$$R^2 = \sqrt{\frac{\sum_{i=1}^{N} (O_i - Pi)^2}{\sum_{i=1}^{N} (O_i - \overline{O_i})^2}} \tag{16}$$

The values of the regression evaluation indices used are presented in Table 7. From this table, it is concluded that the random forests algorithm outperforms the others, as it offers ANN ensembles with much better average performance in testing and better generalization ability. However, the FFNN approach has also a good and quite reliable performance.

3.3 Fuzzy inference ensemble model (FIEM)

The innovation of this paper is enhanced by the employment of a flexible approach based on fuzzy logic. For each pollutant, a Mamdani fuzzy inference system (FIS) has been developed. The FIS considers the range from the two evaluation metrics (correlation coefficient and mean square error). This consideration leads to the development of corresponding Mamdani rule sets. From the execution of the system, two outputs are obtained, namely fuzzy correlation coefficient and fuzzy mean square error after performing de-fuzzification with the centroid approach.

Table 7 Ensemble regression results

	Random forests		Neural networks	
	R^2	RMSE	R^2	RMSE
CO				
Som0	0.93	0.34	0.90	0.42
Som1	0.94	0.40	0.92	0.46
Som2	0.94	0.40	0.92	0.45
Som3	0.94	0.35	0.92	0.40
Som All	0.94	0.36	0.90	0.46
NO				
Som0	0.92	18.64	0.90	20.89
Som1	0.94	24.31	0.92	27.47
Som2	0.94	20.93	0.92	23.67
Som3	0.94	17.09	0.92	18.94
Som All	0.94	19.62	0.91	23.84
NO_2				
Som0	0.90	13.36	0.84	16.58
Som1	0.87	10.74	0.79	13.88
Som2	0.88	11.48	0.81	14.51
Som3	0.90	10.36	0.83	13.18
Som All	0.90	11.31	0.81	15.47
O_3				
Som0	0.89	13.42	0.79	18.17
Som1	0.90	8.67	0.84	11.03
Som2	0.88	11.29	0.79	15.18
Som3	0.89	11.17	0.81	14.55
Som All	0.91	10.91	0.81	16.01
SO_2				
Som0	0.66	10.24	0.50	12.26
Som1	0.74	6.85	0.63	8.18
Som2	0.76	8.57	0.64	10.31
Som3	0.78	8.16	0.69	9.58
Som All	0.75	8.30	0.55	11.06

The FIEM model takes under consideration the accuracy of each network in a flexible manner. Moreover, the new values are produced through machine learning, filtered by fuzzy logic. In this way, the outputs are unbiased.

As a result, this method offers a more objective approach. A distinct fuzzy inference system has been developed for every pollutant. Each FIS has an Inference mechanism comprising of the following Heuristic rule set. The differentiation between the separate systems (corresponding to each pollutant) lies in the determination of the fuzzy membership functions.

Rule set:

If(R^2 *is max) and (**RMSE** is min)* ***Then*** *(**R^2fuzzy** is max)*
AND *(**RMSEfuzzy** is min)*
If(R^2 *is min) and (**RMSE** is max)* ***Then*** *(**R^2fuzzy** is min)*
AND *(**RMSEfuzzy**is max)*
If(R^2 *is med) and (**RMSE** is med)* ***Then*** *(**R^2fuzzy** is med)*
AND *(**RMSEfuzzy** is med)*
If(R^2 *is min)* ***Then***(R^2 *fuzzy is min)*
If(R^2 *is max)* ***Then*** *(R^2 fuzzy is max)*
If(***RMSE*** *is min)* ***Then*** *(**RMSE fuzzy** is min)*
If(***RMSE*** *is max)* ***Then***(***RMSE fuzzy*** *is max)*

The first three rules were given a weight value of 0.5, whereas the last four a value of 1. This was done because in many cases the overall performance of a network is not defined by both the correlation coefficient and the root mean square error. For example, there were networks where we had a high correlation coefficient value (great for the overall performance), but also at the same time a high root mean square error (inadequate for the overall performance). So we decided that the outputs of the system should be influenced from each input separately (fuzzy correlation coefficient and fuzzy mean square error), rather than from both at the same time.

For the fuzzy membership functions (FMFs), the range of each input was the range of the values for each pollutant. Finally, the FMFs used in the inputs were Triangular (*trimf* for MATLAB) FMF for the minimum and maximum Linguistics and Trapezoidal (*trapmf* in MATLAB) for the medium Linguistic, whereas the FMFs employed in the output functions were Triangular for the minimum and maximum Linguistics and Gaussian (*Gausmf* in MATLAB) for the medium one.

4 FIS results

Two distinct regression algorithmic approaches were employed for each one of the four datasets, so totally eight R^2 and RMSE values were produced. The R^2 and the RMSE values that were obtained as regression results for each air pollutant were used as input to each FIS. Actually, one FIS was developed for each air pollutant. The FIS applied values and relations congruent to the individual R^2 and RMSE values based on the rules described in Sect. 3.3. This was done in order to phase rationally the correlation between the conditions that favor high concentrations of air pollutants. The fuzzy values emerged from the above process and the average values of the regression methods before and after the application of the FIS and also the average overall values for each pollutant are presented in the following Tables 8, 9, 10, 11, 12 and 13.

It should be clarified that RFSom0 stands for the performance results of the random forest application on the dataset som0. The RF_Som_All and NN_Som_All are the methods applied in the whole dataset (with all records participating) before dividing it to clusters, and it is the opponent of the ensemble and the FIS approaches. Also the

term "Actual" stands for the performance of the ensembles networks and the FIS for the performance of the fuzzy inference system. The term "Total" stands for the average performance values of the models where the whole dataset was used.

In Table 8 (the CO case), it is shown that if we compare the average values of the three approaches they are more or less similar to a slightly better R^2 value for the ensembles, whereas in Table 9 (the NO case) we see that the FIS has the best average performance for both indices and in Table 10 (NO_2) the ensembles have the smallest average RMSE (Table 11).

Generally speaking, the hybrid approach developed here performs more or less the same with the holistic approaches of the RAF and FFNN if we pay attention to the average values. However, as it was mentioned in the beginning of the paper, the aim was to develop local models related to homogenous data vectors with the same characteristics and to absorb the bad local behaviors, and this has been achieved. The benefit from this approach is that if we wish to forecast the air pollution values for a specific type of concentration (e.g., for extreme cases), then we can perform the forecasting not only for the whole dataset but for the extreme cluster separately, and the regression will be behaving more smoothly (Table 12).

Finally, Table 13 presents a comparison between the results of the three average value approaches (ensembles) emerging after the process of combinatorial learning. Two kinds of comparisons are done to approaches found in the literature: (a) a comparison by using the average value (AVGV) to the FIS average performance and (b) a comparison by using the FIS average performance to the highest potential AVGV that might emerge. The proposed hybrid model shows its validity and its reliability.

Table 8 FIS ensemble vs simple ensemble regression efficiency for CO

	Actual		FIS		Total	
	R^2	RMSE	R^2	RMSE	R^2	RMSE
RF Som0	0.93	0.34	0.94	0.33		
RF Som1	0.94	0.40	0.95	0.40		
RF Som2	0.94	0.40	0.95	0.40		
RF Som3	0.94	0.35	0.95	0.33		
NN Som0	0.90	0.42	0.87	0.45		
NN Som1	0.92	0.46	0.91	0.46		
NN Som2	0.92	0.45	0.91	0.46		
NN Som3	0.92	0.40	0.91	0.40		
RF Som All					0.94	0.36
NN Som All					0.90	0.46
Averages	0.92	0.40	0.92	0.40	0.92	0.41

Table 9 FIS ensemble versus simple ensemble regression efficiency for NO

	Actual		FIS		Total	
	R^2	RMSE	R^2	RMSE	R^2	RMSE
RF Som0	0.92	18.64	0.91	19.82		
RF Som1	0.94	24.31	0.95	27.81		
RF Som2	0.94	20.93	0.95	20.23		
RF Som3	0.94	17.09	0.95	20.03		
NN Som0	0.90	20.89	0.87	20.22		
NN Som1	0.92	27.47	0.91	28.48		
NN Som2	0.92	23.67	0.91	24.15		
NN Som3	0.92	18.94	0.94	19.83		
RF Som All					0.94	19.62
NN Som All					0.91	23.84
Averages	0.92	21.49	0.92	22.57	0.93	21.73

Table 10 FIS ensemble versus simple ensemble regression efficiency for NO_2

	Actual		FIS		Total	
	R^2	RMSE	R^2	RMSE	R^2	RMSE
RF Som0	0.90	13.36	0.91	13		
RF Som1	0.87	10.74	0.90	8.60		
RF Som2	0.88	11.48	0.90	8.80		
RF Som3	0.90	10.36	0.91	8.56		
NN Som0	0.84	16.58	0.84	17.44		
NN Som1	0.79	13.88	0.77	13		
NN Som2	0.81	14.51	0.81	13.81		
NN Som3	0.83	13.18	0.84	13		
RF Som All					0.90	11.31
NN Som All					0.81	15.47
Averages	0.85	13.01	0.86	12.03	0.85	13.39

It is really important that the FIS method is based on rational rules which are either of heuristic nature, related to the variance of the pollutants concentrations, or they are emerging from hidden knowledge coming from the machine learning analysis of the dataset.

5 Discussion and conclusions

An innovative ensemble learning modeling approach was discussed in this paper. The method was tested successfully in the forecasting of air pollutants values for the case of the wider Attica area with data gathered from several measuring stations distributed all over the basin. More specifically, this research paper presents the design

Table 11 FIS ensemble versus simple ensemble regression efficiency for O_3

	Actual		FIS		Total	
	R^2	RMSE	R^2	RMSE	R^2	RMSE
RF Som0	0.89	13.42	0.91	13		
RF Som1	0.90	8.67	0.91	7.08		
RF Som2	0.88	11.29	0.90	8.34		
RF Som3	0.89	11.17	0.91	7.65		
NN Som0	0.79	18.17	0.77	18.92		
NN Som1	0.84	11.03	0.84	10.75		
NN Som2	0.79	15.18	0.77	18.17		
NN Som3	0.81	14.55	0.80	13.16		
RF Som All					0.91	10.91
NN Som All					0.81	16.01
Averages	0.85	12.93	0.85	12.14	0.86	13.46

Table 12 FIS ensemble versus simple ensemble regression efficiency for SO_2

	Actual		FIS		Total	
	R^2	RMSE	R^2	RMSE	R^2	RMSE
RF Som0	0.66	10.24	0.64	9.08		
RF Som1	0.74	6.85	0.75	4.56		
RF Som2	0.76	8.57	0.76	5.13		
RF Som3	0.78	8.16	0.76	4.93		
NN Som0	0.50	12.26	0.51	13.44		
NN Som1	0.63	8.18	0.63	7.25		
NN Som2	0.64	10.31	0.64	9.17		
NN Som3	0.69	9.58	0.65	9		
RF Som All					0.75	8.30
NN Som All					0.55	11.06
Averages	0.68	9.27	0.67	7.82	0.65	9.68

implementation and testing of an innovative hybrid model capable of forecasting the concentrations of air pollutants.

This method uses a technique of combined learning which avoids bad local behaviors and contributes to a smoother forecasting for several homogenous clusters of data vectors. It employs exclusively computational intelligence techniques in order to re-sample the dataset in homogenous clusters that share common characteristics (in terms of pollutants concentrations, day, time, month, temperature, relative humidity and so on). In this way, it exploits local hidden knowledge.

Also this model proposes an effective system for the estimation of the common value of the depended parameter (air pollutant concentration) that is derived by the combined learning approach based on heuristic and rational rules of a fuzzy inference system. This paper presents a

Table 13 Comparison of ensembles versus FIS performance

	FIS versus actual	Actual versus total	FIS versus total
CO			
R^2	0.0001	0.002	0.002
RMSE	0.0002	0.007	0.007
NO			
R^2	0.001	−0.002	−0.003
RMSE	1.08	0.24	−0.84
NO_2			
R^2	0.007	−0.0007	0.006
RMSE	0.99	0.37	1.36
O_3			
R^2	0.002	−0.009	−0.007
RMSE	0.80	0.53	1.32
SO_2			
R^2	0.009	0.03	0.01
RMSE	1.45	0.41	1.86

model that has been developed by considering data vectors related to all involved factors obtained from various representative measuring stations with specific topographic and microclimate characteristics. Thus, it can be considered a rational modeling effort with good level of convergence and with a high practical merit. Testing (a rather difficult forecasting and decision making task) has been performed with reliable and rational results.

Other research efforts like [9] propose the use of an ANFIS ensemble scheme (adaptive neuro-fuzzy inference system) to forecast the air pollution in Macau. It should be clarified that though ANFIS combines the advantages of ANNs with the ones of Fuzzy Logic (which are related to the generalization learning ability and error tolerance of the ANN with the comprehensive knowledge representation of fuzzy inference) the choice of data is a crucial process for the good performance of the model. When the application is related to a wide area, the variations in the topographic and microclimate characteristics should be taken into consideration. Also a high volume of data vectors related to a long temporal period should be also used.

The hybrid methodology described in this paper can be used simultaneously with holistic ANN forecasting models in order to produce more rational forecasts related to specific types of homogenous data (e.g., extremely high or extreme low values, or values related to specific meteorological conditions or months).

It would be interesting to employ the general architecture and the framework of this model in order to perform the same task in other areas with different climate, topographic or urban characteristics. Also various development scenarios must be employed with the use of different

approaches of unsupervised and semi-supervised clustering algorithms and the performance of regressions under various types if ANN ensembles' learning. Finally, an interesting potential would be the use of genetic algorithms to further enhance and optimize the performance.

References

1. Bougoudis I, Iliadis L, Papaleonidas A (2014) Fuzzy inference ANN ensembles for air pollutants modeling in a major urban area: the Case of Athens. Eng Appl Neural Netw Commun Comput Inf Sci 459(2014):1–14
2. Breiman L (2001) Random Forests. Mach Learn 45(1):5–32. doi:10.1023/A:1010933404324
3. Haykin S (2009) Neural networks and learning machines, 3rd edn. New York, Pearson Education
4. Hooyberghs J, Mensink C, Dumont G, Fierens F, Brasseur O (2005) A neural network forecast for daily average PM10 concentrations in Belgium. Atmos Environ 39(18):3279–3289
5. Iliadis L (2007) Intelligent information systems and applications in risk estimation. Stamoulis publication, Thessaloniki. ISBN: 978-960-6741-33-3 A
6. Inal F (2010) Artificial neural network prediction of tropospheric ozone concentrations in Istanbul, Turkey. CLEAN Soil Air Water 38(10):897–908
7. Jollois FX, Poggi JM, Portier **B** (2009) Three non-linear statistical methods for analyzing PM10 pollution in Rouen area CSBIGS 3(1):1–17 CS http://www.bentley.edu/csbigs/documents/poggi.pd
8. Kadri C, Tian F, Zhang L, Dang L, Li G (2013) Neural network ensembles for online gas concentration estimation using an electronic nose. Int J Comput Sci 10(2):1
9. Lei KS, Wan F (2012) Applying ensemble learning techniques to ANFIS for air pollution index prediction in Macau. ISNN 2012, Part I, LNCS 7367. pp 509–516
10. Kohonen T (1989) Self-organization and associative memory, 3rd edn. Springer, Berlin
11. Singha KP, Guptaa S, Rai P (2013) Identifying pollution sources and predicting urban air quality using ensemble learning methods. Atmos Environ 80:426–437
12. Lopez M, Melin P, Castillo O (2007) A method for creating ensemble neural networks using a sampling data approach. Theor Adv Appl Fuzzy Log ASC42 pp. 772–780, Springer
13. Maclin R, Opitz D (1999) Popular ensemble methods: an empirical study. J Artif Intell Res 11:169–198
14. Mamdani EH, Assilian S (1975) An experiment in linguistic synthesis with a fuzzy logic controller. Int J Man Mach Stud 7(1):1–13
15. Mamdani EH (1974) Application of fuzzy algorithms for the control of a simple dynamic plant. In: Proceedings of IEEE, pp 121–159
16. Ordieres Meré JB, Vergara González EP, Capuz RS, Salaza RE (2005) Neural network prediction model for fine particulate matter (PM). Environ Modell Softw 20:547–559
17. Ozcan HK, Bilgili E, Sahin U, Bayat C (2007) Modeling of trophospheric ozone concentrations using genetically trained multi-level cellular neural networks. Adv Atmos Sci Springer 24(5):907–914
18. Ozdemir H, Demir G, Altay G, Albayrak S, Bayat C (2008) Environ Eng Sci 25(9):1249–1254
19. Paoli C (2011) A neural network model forecasting for prediction of hourly ozone concentration in Corsica. In: Proceedings IEEE of the 10th International Conference on Environment and Electrical Engineering (EEEIC)
20. Paschalidou A, Iliadis L, Kassomenos P, Bezirtzoglou C (2007) Neural modeling of the tropospheric ozone concentrations in an urban site. In: 10th ICEANN, pp 436–445
21. Robles LA, Ortega JC, Fu JS, Reed GD, Chow JC, Watson JG, Moncada-Herrera JA (2008) A hybrid ARIMA and artificial neural networks model to forecast particulate matter in urban areas: the case of Temuco, Chile. Atmos Environ 42(35):8331–8340
22. Rokach L (2010) Ensemble-based classifiers. Artif Intell Rev 33(1–2):1–39. doi:10.1007/s10462-009-9124-7
23. Roy S (2012) Prediction of particulate matter concentrations using artificial neural network. Resour Environ 2(2):30–36. doi:10.5923/j.re.20120202.05
24. Takagi T, Sugeno M (1985) Fuzzy identification of systems and its applications to modeling and control. IEEE Trans Syst Man Cybern SMC-15(1):116–133
25. Wahab A-SA, Al-Alawi SM (2002) Assessment and prediction of tropospheric ozone concentration levels using artificial neural networks. EM & Softw 17:219–228
26. www.cs.waikato.ac.nz/ml/weka/
27. Zhou ZH, Wu J, Wei T (2010) Corrigendum to "Ensembling neural networks: many could be better than all". Artif Intell 174(18):1570