Hybrid Unsupervised Modeling of Air Pollution Impact to Cardiovascular and Respiratory Diseases

Lazaros Iliadis, Civil Engineering Department, School of Engineering, Democritus University of Thrace, Xanthi, Greece

Vardis-Dimitris Anezakis, Department of Forestry, School of Agriculture and Forest Sciences, Democritus University of Thrace, Orestiada, Greece

Konstantinos Demertzis, Civil Engineering Department, School of Engineering, Democritus University of Thrace, Xanthi, Greece

Georgios Mallinis, Department of Forestry, School of Agriculture and Forest Sciences, Democritus University of Thrace, Orestiada, Greece

ABSTRACT

During the last few decades, climate change has increased air pollutant concentrations with a direct and serious effect on population health in urban areas. This research introduces a hybrid computational intelligence approach, employing unsupervised machine learning (UML), in an effort to model the impact of extreme air pollutants on cardiovascular and respiratory diseases of citizens. The system is entitled Air Pollution Climate Change Cardiovascular and Respiratory (APCCCR) and it combines the fuzzy chi square test (FUCS) with the UML self organizing maps algorithm. A major innovation of the system is the determination of the direct impact of air pollution (or of the indirect impact of climate change) to the health of the people, in a comprehensive manner with the use of fuzzy linguistics. The system has been applied and tested thoroughly with spatiotemporal data for the Thessaloniki urban area for the period 2004-2013.

KEYWORDS

Fuzzy Chi-Square Test, Fuzzy Logic, Meteorological Factors, Morbidity, Mortality, Photochemical Cloud, Self-Organizing Maps, Smog Cloud, Thessaloniki, Unsupervised Machine Learning

INTRODUCTION

The increase of primary air pollutants (CO, NO, NO_2, SO_2) or secondary ones (O_3), has caused serious degradation in the quality of life of urban areas' residents. Moreover, changes in the heating methods of Greek houses due to the financial crisis, has influenced the concentration of Particulate Matter (PM) in the cities. Extended exposition of the urban population to high concentrations of pollution, increase the percentages of morbidity and mortality due to Cardiovascular (CARD) and Respiratory (RES) problems. Especially, people who live in areas with high levels of air pollution are phasing not only risks of cardiological and respiratory problems, but they are also risking narrowing the arteries and specifically the carotid one. This fact increases the possibility of stroke due to low levels of brain oxygen. Patients with severe disease history and young children or elder population (sensitive groups)

DOI: 10.4018/IJISCRAM.2017070102

are more vulnerable to atmospheric pollution (APO) and they should avoid transportation in the city during the days with high concentration of pollutants.

The combination of meteorological conditions and APO levels plays a major role for the determination of Morbidity and Mortality Risk index (MOMORI). The fluctuation of the MOMORI in an urban center, is mainly influenced by the meteorological conditions that favor the development of smog (CO, SO_2, PM_x) during the winter and the development of Photochemical Cloud (PHOC) in the summer (NO_x, O_3). The topography of an area, the hours of the population's employment in external activities and the percent of the people who live near industrial zones are crucial for the level of risk estimation.

The analysis and continuous monitoring of the APO levels as well as timely forecasting of the conditions that can cause high concentrations, result to the impose of preemptive actions and thus to an effective management of the problem and to the estimation of the impact on related diseases.

This research paper presents the *Air Pollution and Climate Change Cardiovascular and Respiratory Modeling* (APCCCR) hybrid intelligent system. APCCCR considers the effect of atmospheric pollution parameters to Cardiovascular-Respiratory hospitalization incidents and it determines the interdependencies between them in the wider Thessaloniki urban area.

The system is developed in two discrete phases. The first one uses the UML *Self Organizing Maps* algorithm (SOM) to cluster the values of the involved features and to determine the meteorological values that directly affect the pollutants' concentrations which have a serious effect on the considered diseases. The second phase uses the Fuzzy Chi-Square Test (FUCS) to determine the interdependency between the parameters in a rational and comprehensive mode by using proper Linguistics. This is achieved by fuzzifying the P-Value of the Chi-Square test. This process produces Linguistics that express Low, Medium or High dependency by employing fuzzy Membership functions (FMF).

The FUCS application is performed for each cluster in order to determine which atmospheric parameters determine the level of the hospital treatment incidents in the prefecture of Thessaloniki. The testing of the APCCCR was based on a comparative performance analysis between four UML algorithms namely: Self-Organizing Maps, Expectation Maximization, Sequential Information Bottleneck, and Simple K-Means.

Wide use of this approach can enforce the mechanisms of civil protection authority by acting as a means of warning the public hospitals regarding the days of bad meteorological conditions that favor high pollutants' concentrations.

Literature Review-Related Work

To the best of our knowledge, the Sequential Information Bottleneck algorithm has not been used in modeling and assessment of environmental risks. The Expectation Maximization algorithm, Self-Organizing Maps and Simple k-Means have been used for the classification of meteorological and air quality data. The following lines present some cases of classification met in the literature.

(Hernawati, Insani, Bambang, Nur Hadi, & Sahid, 2017) used an unsupervised SOM approach. They considered data related directly or indirectly to pollution (e.g. demographic and social data, air pollution water and soil pollution levels) as well as the geographical situation of each province.

(Štrbová, Štrba, Raclavská, & Bilek, 2018) used SOM to find association between PM concentrations, elevation, selected meteorological variables, and GPS location coordinates.

(Cortina-Januchs, Quintanilla-Dominguez, Andina, & Vega-Corona, 2012) used a Multilayer Perceptron Neural Network (MPNN) to make the prediction of pollutant concentrations for the next hour. A database used to train the ANN based on historical time series of meteorological variables and air pollutant concentrations of SO_2. Before the prediction, Fuzzy C-Means (FCM) and k-Means Clustering (k-MC) algorithms were employed in order to find relationship among pollutant and meteorological variables.

The EM, SOM and SKM Algorithms have been used in the literature to cluster and correlate the cardiovascular and respiratory (CARE) health problems with air pollution.

(Almeida et al., 2013) performed a robust data mining approach for cardiac risk patterns identification. Eight classifiers were tested. As for clustering procedures, k-MC and EM were the chosen algorithms. The clustering techniques were used for the analysis of a dataset that represent different risk groups in terms of cardiovascular function. (Pearce et al., 2015) applied SOM on daily air quality data for 10 pollutants, to define a collection of multipollutant day types. Next, they conducted an epidemiologic analysis of ambient air quality and daily counts of emergency department (ED) visits for asthma or wheeze among children aged 5 to 17 as the health endpoint. They estimated rate ratios for the association of multipollutant day types and pediatric asthma ED visits using a Poisson generalized linear model controlling for long-term, seasonal, and weekday trends and weather. (Basara &Yuan, 2008) used the intrinsic distributions of 92 environmental variables to classify 511 communities into five clusters. SOM determined clusters were reprojected to geographic space and compared with the distributions of several health outcomes. (Requia, Koutrakis, Roig, Adams, & Santos, 2016) assessed the association between vehicle emissions and cardiorespiratory disease risk in Brazil and its modification by spatial clustering of socio-economic conditions.

According to the international literature review, multiple statistical models have been used that reveal the existence of correlations between the meteorological conditions, the level of atmospheric pollution and the consequences in public health (Kalantzi et al., 2011; Xie et al., 2015). However, the use of hybrid approaches and supervised or unsupervised Machine Learning is capable to model multiparametric environmental problems and it also offers optimization mechanisms in order to produce reliable results (Anezakis, Demertzis, Iliadis, & Spartalis, 2016a, 2017; Iliadis, 2007; Iliadis & Papaleonidas, 2016).

Machine Learning algorithms and hybrid approaches have been used in the literature to model air pollution in Athens (Anezakis, Dermetzis, Iliadis, & Spartalis, 2016b; Bougoudis, Dermetzis, & Iliadis, 2016a; Bougoudis, Dermetzis, & Iliadis, 2016b; Bougoudis, Demertzis, Iliadis, Anezakis, & Papaleonidas, 2017; Bougoudis, Iliadis, & Papaleonidas, 2014; Iliadis, Bougoudis, & Spartalis, 2014). Also, important research efforts have been carried out during the last ten years related to the modeling and air pollution forecasting in Thessalonki (Karatzas & Kaltsatos, 2007; Kyriakidis, Karatzas, Papadourakis, Ware, & Kukkonen, 2012; Voukantsis et al., 2011)

Regarding the FUCS test it has been published in the literature recently (Grzegorzewski & Szymanowski, 2015; Lin, Wu, & Watada, 2012) but it has not been used for this purpose.

INNOVATION OF THE PROPOSED METHODOLOGY

From the literature review, it is clearly shown that there is lack of an integrated approach to the environmental problem, rather than fragmentary. Specifically, there is a serious gap in the development of periodic classifications related to health risks due to the combination of meteorological and air pollution data. The innovation of this research is mainly based on the development of an intelligent information system using unsupervised Machine learning algorithms, to cluster atmospheric and air pollution data, related to public health. More specifically, the proposed system provides important information on the dependency of extreme atmospheric and pollution conditions with the morbidity and mortality of the residents of the Thessaloniki city wider area. The clustering of available data accurately attributes the values of the atmospheric parameters that maximize or minimize the Cardiovascular and Respiratory diseases incidents in public hospitals.

The main objective of this research is to inform hospitals in time about potential serious negative effects of the combination "air pollution-meteorological data" in public health. This could lead to better management of increased cardiovascular and respiratory incidents and more generally, hospitals or civil protection authority could reach sound decisions regarding public safety.

DATA

The clustering of atmospheric conditions was based on 3270 daily meteorological and daily air pollution data from the wider urban area (2004-2013). In particular, we have obtained data from 12 air pollution stations of the city whereas the meteorological data were collected from the Thessaloniki Airport station.

Totally the following eight meteorological parameters have been measured: Max Temperature (MaxT), Average Temperature (AvT), Min Temperature (MinT), Average Relative Humidity (AvRH), Rainfall (R), Atmospheric Pressure (AP), Wind Speed (WS), Sunshine Hours (SUN).

Also, data have been obtained for seven air pollutants, namely: Carbon Monoxide (CO), Nitrogen Monoxide (NO), Nitrogen Dioxide (NO_2), Ozone (O_3), Sulfur Dioxide (SO_2), Particulate Matter (PM_{10}) and ($PM_{2.5}$).

The extreme meteorological and pollution values are very important as they significantly contribute to the maximization of MOMO incidents, so they were not excluded from our dataset. It should be clarified that the *circulatory system* diseases (I00-I99) and the *respiratory* ones (J00-J99) were categorized on the basis of the 10th review of the International Statistical Classification of Diseases and Related Health Problems (ICD-10) (http://apps.who.int/classifications/icd10/browse/2016/en#).

Cardiovascular Hospitalization (CH) and Respiratory Hospitalization (RH) data were gathered from all public hospitals of Thessaloniki city. At prefecture level, the collection of Cardiovascular Deaths (CD) and Respiratory Deaths (RD) data was conducted by the Hellenic Statistical Authority (ELSTAT).

THEORETICAL BACKGROUND

Self-Organizing Maps

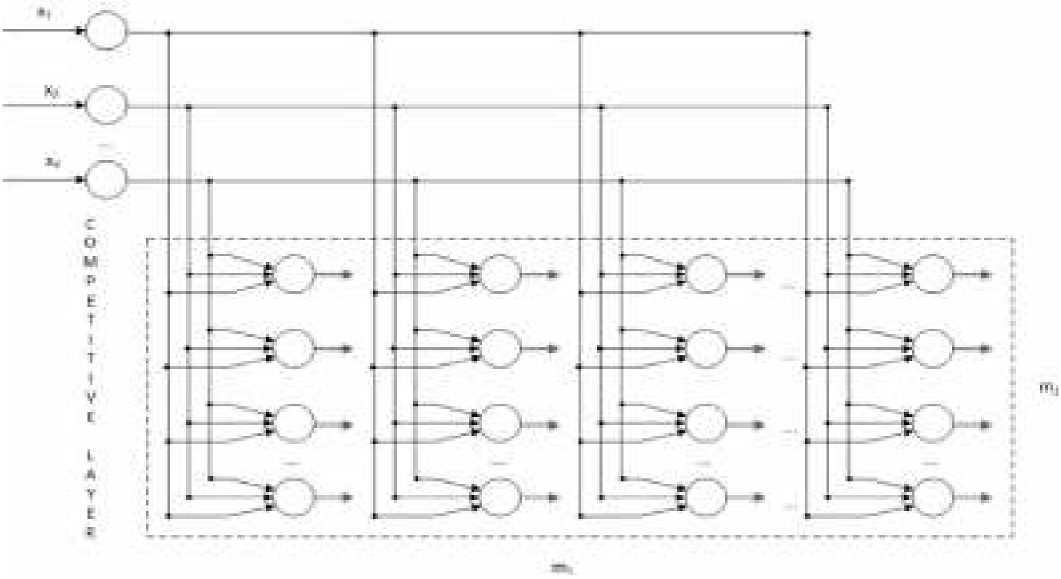
Self-Organizing Maps is an algorithm of competitive learning developed by Kohonen (Haykin, 2009). A SOM includes the input layer and the competing neurons layer, which are organized in a 2-D lattice like in Figure 1. Each one of them is characterized by a weight vector $W_i = (w_{il},\ldots,w_{id})^T$. When an input vector $X_i = (x_1,\ldots,x_d)^T \, x \in R$ is introduced, the lattice neurons compete each other and the winning neuron (WINE) *m* is obtained. Its vector W_m appears to have the highest similarity with vector X_i . Thus, SOM depict an input X_i of dimension d, at the coordinates of the grid $R_m = (z_{m1}, z_{m2})^T$ (Haykin, 2009; Kohonen, 1989).

In order to group the data, a self-organization map is formed, initializing the weights $W_i = (w_{il},\ldots,w_{id})^T$ with small values randomly produced by a random number generator function. The next algorithmic steps follow:

- **Competition:** For each training sample X_n the lattice neurons estimate the respective value of the similarity function using the Euclidean distance between the input vector $X_i = (x_1,\ldots,x_d)^T \, x \in R$ and the weight vector $W_i = (w_{il},\ldots,w_{id})^T$ of the competing neurons. The neuron with the highest similarity is the winner.
- **Cooperation:** The WINE i delimits its topological neighborhood $h_{j,i}$ (around the surrounding neurons) which is a function of $d_{j,i}$ which is the distance between the winning neuron i and neuron j . The neighborhood is symmetric to the WINE.

The following Gaussian function 1 is used to perform the clustering:

Figure 1. Flowchart of the proposed model Self-Organizing Map with d inputs and 2-dimensions lattice m1 x m2



$$h_{j,i}(x) = \exp\left(-\frac{d_{j,i}^2}{2\sigma^2}\right) \tag{1}$$

where σ is the effective width of the topological neighborhood which determines the extent to which the neurons in the neighborhood of the winner are involved in the process. This parameter is reduced exponentially in every epoch n, based on the following function 2.

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

The parameter σ_0 is the initial value of the active width and the value of τ_1 is given by the following equation 3:

$$\tau_1 = \left(\frac{n_0}{\ln(\sigma_0)}\right) \tag{3}$$

Considering a 2-D lattice, we have assigned the value of its radius to the initial value of the active width σ_0 whereas

$$\tau_1 = \left(\frac{1000}{\log(\sigma_0)}\right) \tag{4}$$

- **Synaptic Adaption:** At this last stage of the training process, we have been updating the weights of neurons on the competitive level. The change metric is given by equation 5:

$$\Delta wj = \eta h_{j,i(x)}(x - w_j) \tag{5}$$

The index i is used to denote the winner and j is a neuron in its neighborhood. Given the weight vector $W_j(n)$ for a certain point in time n, we estimated the new vector for the time stamp $n+1$ from the following function 6:

$$w_j(n+1) = w_j(n) + n(n)h_{j,i(x)}(n)(x(n) - w_j(n)) \tag{6}$$

From the above relationship it follows that the learning rate n starts from a value around 0.1 and decreases gradually to the value of 0.01. These values were achieved according to the function 7 below:

$$n(n) = \eta_0 \exp\left(-\frac{n}{\tau_2}\right) n = 0,1,2,..., \text{with } \eta_0 = 0.1 \text{ and } \tau_2 = 1000 \tag{7}$$

It should be stated that τ_2 is given by equation 8 accordingly:

$$\tau_2 = \frac{n_0}{\ln(100 * \eta_0)} \tag{8}$$

where n_0 is the number of iterations of the layout phase η_0 is the initial learning rate and σ_0 is given by the following equation 9.

$$\sigma_0 = \sqrt{w^2 + h^2} \tag{9}$$

It should be clarified that w and h are the length and the height of the 2-D lattice respectively.

Chi-Square Test and Fuzzy Chi-Square Test

The Chi-Squared hypothesis-testing is a non-parametric statistical test in which the sampling distribution of the test statistic is a chi-square distribution when the null hypothesis is true. The null hypothesis H_0 usually refers to a general statement or default position that there is no relationship between two measured phenomena, or no difference among groups. The H_0 is assumed to be true until evidence suggest otherwise (Corder & Foreman, 2014). The statistical control index used for this assessment is the test statistic X^2.

$$X^2 = \sum \frac{\left(f_0 - f_e\right)}{f_e} \tag{10}$$

where fe is the expected frequency and fo the observed one.

The degrees of freedom are estimated as follows (based on the rXc table of labeled categories). The degrees of freedom are calculated based in equation 11. The value of the degrees of freedom for the selected Confidence Interval (C_INT) is calculated from the statistical Chi-Square Test.

$$df = \left(r-1\right)\times\left(c-1\right) \tag{11}$$

For the H_0 the critical values for the test statistic X^2 are estimated by the X^2 distribution after considering the degrees of freedom. If the result of the test statistic is less than the value of the Chi-Square distribution, then we accept H_0 otherwise we reject it.

The produced p-values include the potential error magnitude in the range [0, 1]. Every error value is multiplied by 10^{-6}. The p-values are fuzzified by the use of FCST, according to the specified C_INT and to the significance level. The p-value a=0.05 (in the Confidence interval 0.95 with selected significance level a=0.05) is considered as critical for the dependence or independence of the variables. Variable dependence is determined with a $p-value<a$ and the independence with $p-value>a$.

The construction of the FIS-Mamdani membership functions for $p-values<a$ in the interval $[0,0.05)$ represent the dependence Linguistics of the two variables. Strong dependence with membership values (MV) close to 1 is assigned to p-values more distant from 0.05 approaching 0. On the other hand, p-values close to 0.05 (but never equal to it) exhibit Lower dependence and their MV will be close to 0. Totally 50,000 values were fuzzified all included in the interval $[0,0.05*10^{-6})$. Every p-value was multiplied by 10^{-6} . In Table 1 we present the Dependency MV of chosen p-values and their classification in dependency Linguistics with the corresponding MV. The characteristic Lower and Higher Boundary p-values (LBV_i and HBV_i) of the Fuzzy High Dependency Linguistic are LBV_1 =0 and HBV_1 =0.01250 where for the Medium and Low Dependency Linguistics they are LBV_2 =0.1245 and HBV_2 =0.0375 LBV_3 =0.0365 HBV_3 =0.0500 respectively, leaving space for overlapping as it is always the case in fuzzy logic approaches.

The $p-values>a$ are fuzzified in the interval $[0.05,1]$ by constructing membership functions to calculate the MV of the independent variables and their inclusion in classes of independence Linguistics. P-values close to 1, (the largest possible error) indicate that the variables will be independent with a MV equal to 1 (strongly independent) whereas for p-values close to 0.05 MV will be close to zero and the independence of the variables will decrease (Low independence).

Table 1. Sample of MV of the Dependency Linguistics for p-values<a

P-Value	Linguistics	High Dependence Degree of Membership	Medium Dependence Degree of Membership	Low Dependence Degree of Membership
0	High	1	0	0
0.000010	High	0.999450	0	0
0.012499	High	0.375050	0.374950	0
0.012501	Medium	0.3749500	0.375050	0
0.030000	Medium	0	0.750000	0
0.035000	Medium	0	0.500000	0.2500125
0.037499	Medium	0	0.375050	0.3749680
0.037500	Low	0	0.375000	0.3750187
0.049999	Low	0	0	1

We have fuzzified 950,000 $p - values$ (the values that are independent in the interval $[0.05,1]$. The characteristic LBV_i and HBV_i of the Fuzzy High Independence Linguistic are LBV_1 =0.05 and HBV_1 =0.287551 where for the Medium and Low Independence Linguistics they are LBV_2 =0.28755 and HBV_2 = 0.76252 LBV_3 =0.762517 HBV_3 =1 respectively, leaving space for overlapping as it is always the case in fuzzy logic approaches (see Table 2).

The FUCS is a hybrid methodology where the degree of dependency between two parameters is defined with accuracy by employing Soft Computing. The proposed model has been created by the authors aiming to fuzzify the P-values and to provide the Linguistic expression of the level of dependency (Anezakis, Demertzis, Iliadis et al., 2017; Anezakis, Iliadis, Demertzis et al., 2017; Dimou, Anezakis, Demertzis et al., 2017).

DESCRIPTION OF THE PROPOSED METHODOLOGY

The basic modeling methodology includes clustering of the considered features using the SOM algorithm. The choice of SOM was done after a comparison with four other unsupervised ML approaches. The SOM produced 4 clusters, EM 5 clusters, whereas both SIB and SKM gave 4. The SOM was chosen as the more appropriate for this case because its clusters were more homogeneous and more accurate.

In this research SOM was used to cluster the values of the smog and the photochemical cloud. Moreover, this approach has isolated in an outlier cluster the values of the atmospheric features that maximize the morbidity and mortality levels. Also, for each cluster obtained by the SOM, the FUCS test was applied to determine the degree of dependency between the fuzzy linguistics of the atmospheric parameters and the corresponding linguistics of the Cardiovascular and Respiratory MOMO indices. The whole algorithm includes 4 distinct steps described below.

(Anezakis, Iliadis, Demertzis, & Mallinis, 2017) have recently shown that the combined effect of low Temperatures with high Humidity values and low hours of Sunshine, compose the Meteorological Smog Index (MSMI). This increases the concentration of all air pollution indices except of O_3. Similarly, during the summer months, the PHOC risk is influenced by the effect of high temperatures together with the low humidity values and the high hours of sunshine. This leads to the maximization of O_3 concentrations. It is a fact that during Spring and Autumn the meteorological conditions contribute significantly either to the development of Smog (SM) or to the production of PHOC.

Table 2. Sample of Membership Values of Independence Linguistics in the interval [0.05-1]

P-Value	Linguistics	Low Independence Degree of Membership	Medium Independence Degree of Membership	High Independence Degree of Membership
0.050001	Low	1	0	0
0.050010	Low	0.999970	0	0
0.287550	Low	0.374871	0.374868	0
0.287551	Medium	0.374868	0.374871	0
0.430100	Medium	0	0.750000	0
0.715050	Medium	0	0.500000	0.250131
0.762518	Medium	0	0.375051	0.375047
0.762519	High	0	0.375048	0.375050
1	High	0	0	1

Step 1: Clustering Data Parameters (With Various Algorithms) That Favor Smog and Photochemical Clouds

EM, SIB, SKM and SOM clustering was performed. However, the SOM approach has offered the most meaningful clear and comprehensive results and conclusions. So, the discussion will focus mainly in the analysis and interpretation of the SOM output, although a comparative presentation of the results obtained by all four algorithms will be done in the following tables.

The training process of the SOM has employed the *trainbu* (trains a network with weight and bias learning rules with batch updates) and the *learnsomb* (employs batch self-organizing map weights) learning functions. The basic principle was the use of less complicated Maps. Thus, the SOM that have been developed consist of 2X2 neurons creating 4 clusters. The results, their analysis and their interpretation, are presented below in the following paragraphs.

Figure 2 shows the number of records that have been assigned (won) to each neuron. In this image, cluster4 with the fewest records (440) is the most important, since it contains the most extreme values. The numbering starts at the lower left corner (the cluster with 1264 cases is the first). The correct direction to follow is from left to right as we go up (the cluster containing 649 records is 2nd, the one comprising of 917 is 3rd. It is completed to the top right corner with the most extreme group of 440 entries.

Figure 3 shows the distances between the clusters. This figure plays a very important role in identifying the ones containing extreme values. Bright colors show close distances, whereas dark colors show long ones (data vectors located away from the majority). Thus, clusters 2 and 4 that use dark connections and contain the smallest number of records are the ones who contain high and extreme values.

In this step, the Morbidity and Mortality indices related to CARD and RES diseases were fuzzified by using Triangular membership functions (TRIMF). The values greater than or equal to Average+2σ

Figure 2. SOM_Sample_Hits_4Clusters

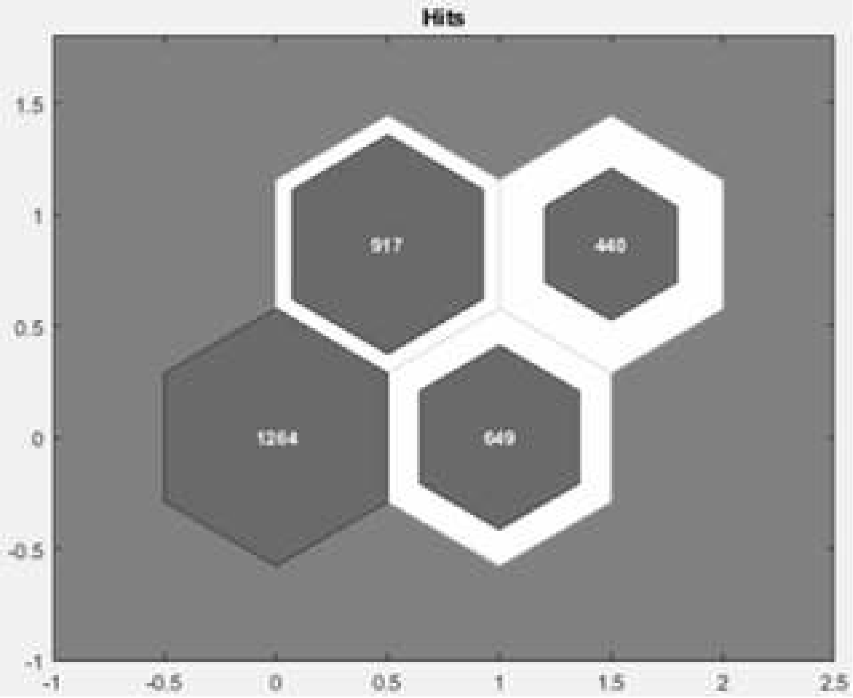
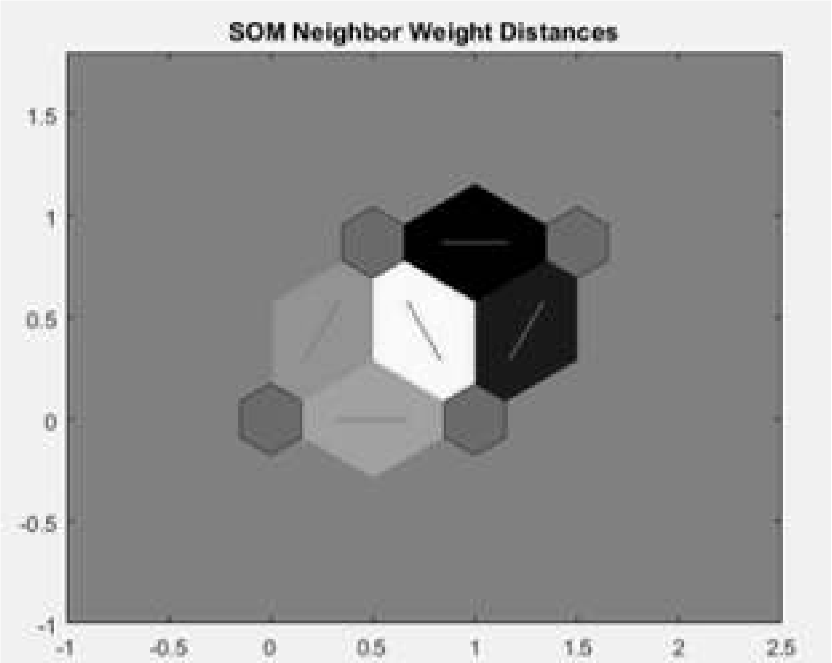


Figure 3. SOM_Neighbor weight Distances



(where σ is the standard deviation) were considered as extreme. The profile of each cluster was built based on the number of cases that were assigned the high and extremely high fuzzy Linguistics.

For all of the used algorithms, the meteorological conditions that favor Smog (NO, $NO_{2,}$ $SO_{2C,}$ $PM_{10,}$ $PM_{2.5,}$ CO) were distinguished properly from the ones that contribute towards the Photochemical cloud (O_3) by using two clusters (Tables 3 and 4). The Kohonen SOM has shown that the Smog cloud (cluster 4 comprising of 440 records) is favored by low or moderate daily temperatures, high daily moisture, few hours of sunshine, and very high daily concentrations of primary pollutants' emissions NO, NO2, SO2, PM10, PM2.5, CO. It corresponds to the period October to April. SOM has revealed that cluster 4 contains 224 incidents of high risk due to meteorological conditions (HRMC) and 59 cases of extreme Death risk due (EXDR) to Cardiovascular and Respiratory diseases. Cluster 1 (which contains 1264 records and is related to the period March to November) is characterized by high daily temperatures, low daily moisture values and many hours of sunshine. These conditions result in very high concentrations of O_3 and in Photochemical Cloud. Herein, the combination of atmospheric features has contributed to 384 records of HRMC and to 54 ones of EXDR (Tables 3 and 4).

Table 5 illustrates the most extreme cases of cardiovascular and respiratory Morbidity and Mortality for each month from 2004 to 2013. Each obtained cluster is related to a major subset which corresponds to a specific seasonal period. All four algorithms have proven that the most risky meteorological conditions that favor the development of the Smog and the most EXDR due to CARD/RES diseases are observed during January.

Also, all algorithms have shown that June is the month that favors the PHOC. June has the highest number (10 according to SOM) of extreme (CARD and RES) Morbidity risk values. In contrast, all algorithms agree that the most extreme Cardiovascular and Respiratory Mortality risk exists in July. The SOM algorithm recorded 11 extreme values in July. Table 6 presents the clusters: (EM-Cl2, SIB-Cl2, SKM-Cl3, SOM-Cl2) having the highest number of recorded extreme MOMO values (e.g. 59 of them only for March, according to SOM).

Table 3. Meteorological-CARD Hospitalization/Deaths-RES Hospitalization/Deaths Range for each cluster

	MaxT daily interval	AvT daily interval	MinT daily interval	AvRH daily interval	SUN daily interval	CH daily interval / Linguistics	RH daily interval / Linguistics	CD daily interval / Linguistics	RD daily interval / Linguistics
EM-Smog Cloud-Cluster3 October-April 2004-2013	-1.4-25.6	-2.95-21.65	-7.8-18	45.66-97.65	0-10.9	3-96 High =65 Extreme =21	0-53 High =84 Extreme =23	0-12 High =88 Extreme =16	0-5 High =23 Extreme =9
EM-Photochemical Cloud-Cluster0 April -October 2004-2013	21.6-44	16.55-34.7	0.2-27.1	28.25-76.29	2.9-14	1-90 High =82 Extreme =19	0-46 High =73 Extreme =10	0-13 High =147 Extreme =18	0-5 High =48 Extreme =11
SIB-Smog Cloud-Cluster0 October-April 2004-2013	-1.4-24.6	-3.5-19.8	-7.8-17.6	39.74-97.65	0-12	0-108 High =73 Extreme =34	0-54 High =94 Extreme =36	0-13 High =123 Extreme =22	0-5 High =40 Extreme =15
SIB-Photochemical Cloud-Cluster1 March-November 2004-2013	10.4-40.3	5.25-32.85	0.2-27.1	28.25-82.38	0-14	1-90 High =64 Extreme =12	0-46 High =68 Extreme =8	0-14 High =181 Extreme =20	0-5 High =51 Extreme =13
SKM-Smog Cloud-Cluster0 October- April 2004-2013	-1.4-24.4	-2.95-19.8	-7.8-15.4	52.46-97.65	0-10.2	3-96 High =61 Extreme =22	0-62 High =67 Extreme =25	0-12 High =91 Extreme =15	0-5 High =24 Extreme =9
SKM-Photochemical Cloud-Cluster1 April-November 2004-2013	17.2-44	14.7-34.7	0.2-27.1	28.25-82.38	1-14	1-90 High =89 Extreme =25	0-46 High =90 Extreme =12	0-14 High =176 Extreme =22	0-5 High =54 Extreme =14
SOM-Smog Cloud-Cluster4 October-April 2004-2013	2-25.6	-2.45-20.45	-7.8-17.6	48.4-97.65	0-10.4	4-96 High =56 Extreme =16	0-53 High =66 Extreme =19	0-12 High =81 Extreme =14	0-5 High =21 Extreme =10
SOM-Photochemical Cloud-Cluster1 March-November 2004-2013	8-40.3	4.3-32.85	-0.6-27.1	28.25-82.11	0-14	1-90 High =66 Extreme =13	0-46 High =75 Extreme =11	0-13 High =187 Extreme =16	0-5 High =56 Extreme =14

* Max Temperature (MaxT), Average Temperature (AvT), Min Temperature (MinT), Average Relative Humidity (AvRH), Sunshine Hours (SUN), Cardiovascular Hospitalization (CH), Respiratory Hospitalization (RH), Cardiovascular Deaths (CD), Respiratory Deaths (RD)

Step 2: Clustering Data Features Based on High Rainfall and High Windspeed

In the clusters of Table 6 we can see that the highest daily rainfall value (131.6 mm) is connected to the appearance of high daily wind speed (32.5 knots). Maximization of the values of the above parameters causes dispersion and diffusion of air pollutants. Cluster 3 that has been produced by SOM (September to May) is characterized by high or moderate temperatures and high levels of moisture. SOM has detected 304 incidents of high MOMO risk and 51 records of extreme risk due to CARD/RES diseases. This is mainly due to the seasonal unfavorable meteorological conditions (high levels of rainfall and wind speed).

Step 3: Clustering Data Based on the Highest and Extreme MOMO Rates

Step 3 has produced the following results. Cluster 2 of SOM (649 records) has proven to be the most appropriate for the Morbidity and Mortality study, because it includes the high risky (477) and extremely risky (196) incidents (see Tables 7 and 8). This is very useful information for civil protection

Table 4. Range for Air Pollutants and CARD/RES diseases for each cluster

	CO daily interval	NO daily interval	NO_2 daily interval	O_3 daily interval	SO_2 daily interval	PM_{10} daily interval	$PM_{2.5}$ daily interval	CH daily interval / Linguistics	RH daily interval / Linguistics	CD daily interval / Linguistics	RD daily interval / Linguistics
EM-Smog Cloud-Cluster3 October-April 2004-2013	0.6-3.37	14.43-199.9	24.53-93.4	13-73.3	2.57-70.9	23.28-203	11-121.5	3-96 High =65 Extreme =21	0-53 High =84 Extreme =23	0-12 High =88 Extreme =16	0-5 High =23 Extreme =9
EM-Photochemical Cloud-Cluster0 April -October 2004-2013	0.25-1.18	4.2-43.6	4.83-65.51	43.53-127.9	1-20.4	13.77-126	7-64	1-90 High =82 Extreme =19	0-46 High =73 Extreme =10	0-13 High =147 Extreme =18	0-5 High =48 Extreme =11
SIB-Smog Cloud-Cluster0 October-April 2004-2013	0.5-2.75	9.44-199.9	15.27-93.4	12.94-74.46	1.33-70.9	16.35-203	11-121.5	0-108 High =73 Extreme =34	0-54 High =94 Extreme =36	0-13 High =123 Extreme =22	0-5 High =40 Extreme =15
SIB-Photochemical Cloud-Cluster1 March-November 2004-2013	0.25-2.89	4.2-39.23	4.83-58.78	43.53-127.9	1-20.4	13.45-74.28	6-60	1-90 High =64 Extreme =12	0-46 High =68 Extreme =8	0-14 High =181 Extreme =20	0-5 High =51 Extreme =13
SKM-Smog Cloud-Cluster0 October- April 2004-2013	0.66-2.75	30.3-199.9	24.6-93.4	13-63.01	2.62-70.9	28.83-195	12-121.5	3-96 High =61 Extreme =22	0-62 High =67 Extreme =25	0-12 High =91 Extreme =15	0-5 High =24 Extreme =9
SKM-Photochemical Cloud-Cluster1 April-November 2004-2013	0.25-1.21	4.2-51.29	4.83-67.2	41.37-127.9	1-22.4	13.45-158.6	6-64	1-90 High =89 Extreme =25	0-46 High =90 Extreme =12	0-14 High =176 Extreme =22	0-5 High =54 Extreme =14
SOM-Smog Cloud-Cluster4 October-April 2004-2013	0.66-2.75	33.99-199.9	24.6-93.4	12.94-73.3	2.6-70.9	29.4-203	13-121.5	4-96 High =56 Extreme =16	0-53 High =66 Extreme =19	0-12 High =81 Extreme =14	0-5 High =21 Extreme =10
SOM-Photochemical Cloud-Cluster1 March-November 2004-2013	0.25-2.89	4.21-42.2	4.83-58.8	50.69-127.9	1-21	12.66-68.15	6-57	1-90 High =66 Extreme =13	0-46 High =75 Extreme =11	0-13 High =187 Extreme =16	0-5 High =56 Extreme =14

* Cardiovascular Hospitalization (CH), Respiratory Hospitalization (RH), Cardiovascular Deaths (CD), Respiratory Deaths (RD)

authority. It can help towards the design of emergency readiness plans for hospitals. It has been shown that the CARD/RES MOMO risk in cluster 2 is mainly influenced by meteorological parameters and by air pollution concentrations PM_{10}, $PM_{2.5,}$ O_3 for the period September-June.

Figure 4 shows the influence of each parameter to the determination of the SOM. The open colors show high weights and dependency from the corresponding input parameter whereas the dark ones the opposite. Yellow colors express the highest level of dependency and they show which cluster is depended on which parameter. Totally the following 19 input features were used.

Input 1: Max Temperature (MaxT), Input2: Average Temperature (AvT), Input3: Min Temperature (MinT), Input4: Average Relative Humidity (AvRH), Input5: Rainfall (R), Input6: Sunshine Hours (SUN), Input7: Wind Speed (WS), Input8: Atmospheric Pressure (AP), Input9: Sulfur Dioxide (SO_2),

Table 5. Distribution of extreme hospitalization incidents and CV RES deaths over months

	Extreme CH incidents per month	**Extreme RH incidents per month**	**Extreme CD incidents per month**	**Extreme RD incidents per month**
EM-Smog Cloud-Cluster3 October-April 2004-2013	November: Extreme =4 December: Extreme =3 **January: Extreme =9** February: Extreme =3 March: Extreme =1	November: Extreme =3 December: Extreme =1 **January: Extreme =9** February: Extreme =8 March: Extreme =2	November: Extreme =3 December: Extreme =2 **January: Extreme =7** February: Extreme =3 March: Extreme =1	November: Extreme =2 December: Extreme =1 **January: Extreme =5** February: Extreme =1 March: Extreme =0
EM-Photochemical Cloud -Cluster0 April -October 2004-2013	April: Extreme =3 May: Extreme =6 **June: Extreme =4** July: Extreme =2 August: Extreme =0 September: Extreme =2	April: Extreme =2 May: Extreme =1 **June: Extreme =5** July: Extreme =2 August: Extreme =0 September: Extreme =0	April: Extreme =0 May: Extreme =2 June: Extreme =4 **July: Extreme =7** August: Extreme =3 September: Extreme =2	April: Extreme =0 May: Extreme =1 June: Extreme =2 **July: Extreme =3** August: Extreme =4 September: Extreme =1
EM-Cluster2- The most extreme risk incidents September-May 2004-2013	September: Extreme =1 October: Extreme =5 November: Extreme =3 December: Extreme =4 January: Extreme =9 February: Extreme =9 **March: Extreme =12** April: Extreme =3 May: Extreme =1	September: Extreme =0 October: Extreme =4 November: Extreme =1 December: Extreme =3 January: Extreme =9 February: Extreme =9 **March: Extreme =20** April: Extreme =5 May: Extreme =2	September: Extreme =1 October: Extreme =0 November: Extreme =2 December: Extreme =3 January: Extreme =3 February: Extreme =6 **March: Extreme =10** April: Extreme =3 May: Extreme =1	September: Extreme =0 October: Extreme =1 November: Extreme =1 December: Extreme =1 January: Extreme =1 February: Extreme =6 **March: Extreme =5** April: Extreme =1 May: Extreme =1
SIB-Smog Cloud-Cluster0 October-April 2004-2013	November: Extreme =3 December: Extreme =5 **January: Extreme =15** February: Extreme =8 March: Extreme =3	November: Extreme =3 December: Extreme =2 **January: Extreme =15** February: Extreme =10 March: Extreme =6	November: Extreme =3 December: Extreme =3 **January: Extreme =11** February: Extreme =4 March: Extreme =1	November: Extreme =2 December: Extreme =2 **January: Extreme =6** February: Extreme =3 March: Extreme =2
SIB- Photochemical Cloud- Cluster1 March-November 2004-2013	March: Extreme =0 April: Extreme =1 May: Extreme =3 **June: Extreme =5** July: Extreme =2 August: Extreme =0 September: Extreme =0 October: Extreme =1	March: Extreme =1 April: Extreme =0 May: Extreme =1 **June: Extreme =4** July: Extreme =2 August: Extreme =0 September: Extreme =0 October: Extreme =0	March: Extreme =1 April: Extreme =1 May: Extreme =3 June: Extreme =3 **July: Extreme =7** August: Extreme =3 September: Extreme =1 October: Extreme =0	March: Extreme =0 April: Extreme =0 May: Extreme =1 June: Extreme =2 **July: Extreme =4** August: Extreme =4 September: Extreme =2 October: Extreme =0
SIB-Cluster2- The most extreme risk incidents September-May 2004-2013	September: Extreme =4 October: Extreme =10 November: Extreme =5 December: Extreme =2 January: Extreme =3 February: Extreme =10 **March: Extreme =14** April: Extreme =8 May: Extreme =7	September: Extreme =0 October: Extreme =5 November: Extreme =1 December: Extreme =2 January: Extreme =4 February: Extreme =9 **March: Extreme =24** April: Extreme =10 May: Extreme =5	September: Extreme =3 October: Extreme =0 November: Extreme =2 December: Extreme =0 January: Extreme =0 February: Extreme =3 **March: Extreme =8** April: Extreme =2 May: Extreme =1	September: Extreme =0 October: Extreme =1 November: Extreme =0 December: Extreme =0 January: Extreme =0 February: Extreme =4 **March: Extreme =3** April: Extreme =1 May: Extreme =2
SKM-Smog Cloud-Cluster0 October- April 2004-2013	October: Extreme =1 November: Extreme =4 December: Extreme =2 **January: Extreme =11** February: Extreme =4 March: Extreme =0	October: Extreme =0 November: Extreme =3 December: Extreme =1 **January: Extreme =11** February: Extreme =9 March: Extreme =1	October: Extreme =0 November: Extreme =3 December: Extreme =1 **January: Extreme =8** February: Extreme =3 March: Extreme =0	October: Extreme =0 November: Extreme =2 December: Extreme =1 **January: Extreme =5** February: Extreme =1 March: Extreme =0
SKM-Photochemical Cloud- Cluster1 April-November 2004-2013	April: Extreme =3 May: Extreme =6 **June: Extreme =7** July: Extreme =2 August: Extreme =0 September: Extreme =3 October: Extreme =4	April: Extreme =2 May: Extreme =2 **June: Extreme =5** July: Extreme =2 August: Extreme =0 September: Extreme =2 October: Extreme =1	April: Extreme =0 May: Extreme =3 June: Extreme =5 **July: Extreme =7** August: Extreme =4 September: Extreme =3 October: Extreme =0	April: Extreme =0 May: Extreme =2 June: Extreme =2 **July: Extreme =4** August: Extreme =4 September: Extreme =1 October: Extreme =1

continued on following page

Table 5. Continued

	Extreme CH incidents per month	Extreme RH incidents per month	Extreme CD incidents per month	Extreme RD incidents per month
SKM-Cluster3- The most extreme risk incidents September-June 2004-2013	September: Extreme =1 October: Extreme =5 November: Extreme =1 December: Extreme =4 January: Extreme =7 February: Extreme =6 **March: Extreme =12** April: Extreme =4 May: Extreme =2	September: Extreme =0 October: Extreme =4 November: Extreme =1 December: Extreme =3 January: Extreme =7 February: Extreme =8 **March: Extreme =21** April: Extreme =5 May: Extreme =3	September: Extreme =0 October: Extreme =0 November: Extreme =1 December: Extreme =0 January: Extreme =3 February: Extreme =6 **March: Extreme =9** April: Extreme =2 May: Extreme =1	September: Extreme =1 October: Extreme =0 November: Extreme =0 December: Extreme =1 January: Extreme =0 February: Extreme =5 **March: Extreme =4** April: Extreme =1 May: Extreme =1
SOM-Smog Cloud Cluster4 October-April 2004-2013	October: Extreme =1 November: Extreme =3 December: Extreme =3 **January: Extreme =8** February: Extreme =1 March: Extreme =0 April: Extreme =1	October: Extreme =0 November: Extreme =3 December: Extreme =0 **January: Extreme =10** February: Extreme =5 March: Extreme =1 April: Extreme =0	October: Extreme =0 November: Extreme =3 December: Extreme =2 **January: Extreme =7** February: Extreme =2 March: Extreme =0 April: Extreme =0	October: Extreme =0 November: Extreme =2 December: Extreme =1 **January: Extreme =5** February: Extreme =2 March: Extreme =0 April: Extreme =0
SOM-Photochemical Cloud- Cluster1 March-November 2004-2013	March: Extreme =0 April: Extreme =1 May: Extreme =1 **June: Extreme =6** July: Extreme =2 August: Extreme =0 September: Extreme =0 October: Extreme =1	March: Extreme =2 April: Extreme =2 May: Extreme =2 **June: Extreme =4** July: Extreme =2 August: Extreme =0 September: Extreme =0 October: Extreme =0	March: Extreme =0 April: Extreme =1 May: Extreme =1 June: Extreme =2 **July: Extreme =7** August: Extreme =2 September: Extreme =1 October: Extreme =0	March: Extreme =0 April: Extreme =0 May: Extreme =3 June: Extreme =2 **July: Extreme =4** August: Extreme =4 September: Extreme =1 October: Extreme =0
SOM-Cluster2- The most extreme risk incidents August-June 2004-2013	August: Extreme =0 September: Extreme =4 October: Extreme =9 November: Extreme =6 December: Extreme =4 January: Extreme =12 February: Extreme =16 **March: Extreme =18** April: Extreme =9 May: Extreme =8 June: Extreme =1	August: Extreme =0 September: Extreme =1 October: Extreme =5 November: Extreme =1 December: Extreme =2 January: Extreme =9 February: Extreme =14 **March: Extreme =28** April: Extreme =8 May: Extreme =5 June: Extreme =1	August: Extreme =2 September: Extreme =3 October: Extreme =0 November: Extreme =0 December: Extreme =1 January: Extreme =2 February: Extreme =3 **March: Extreme =9** April: Extreme =1 May: Extreme =1 June: Extreme =3	August: Extreme =0 September: Extreme =0 October: Extreme =1 November: Extreme =0 December: Extreme =0 January: Extreme =1 February: Extreme =1 **March: Extreme =4** April: Extreme =1 May: Extreme =2 June: Extreme =0

Table 6. Clustering of the data based on the high rainfall values and on high wind speed

	R daily interval	WS daily interval	MaxT daily interval	AvT daily interval	MinT daily interval	AvRH daily interval	CH daily Interval / Linguistics	RH daily Interval / Linguistics	CD daily Interval / Linguistics	RD daily Interval / Linguistics
EM-Cluster1 September-June 2004-2013	0.1-131.6	0-32.5	0.8-25.8	-3.5-25.5	-4.8-19.6	45.3-97.4	1-82 High =37 Extreme = 17	1-51 High =52 Extreme= 17	0-16 High =72 Extreme =12	0-5 High =27 Extreme =5
SIB-Cluster3 September-June 2004-2013	0-131.6	0-32.5	1.8-30.2	0.7-25.5	0-20.6	33.4-97.4	1-82 High =22 Extreme =10	0-51 High =38 Extreme =9	0-19 High =98 Extreme=15	0-5 High =28 Extreme=8
SKM-Cluster2 September-June 2004-2013	0-131.6	0-22.88	1-28.4	-3.5-25.6	-4.8-22.4	36.4-97.4	1-91 High =83 Extreme =31	0-59 High =110 Extreme =25	0-16 High =131 Extreme=20	0-5 High =38 Extreme=11
SOM-Cluster3 September-May 2004-2013	0-131.6	0-26.5	1-30.2	-3.5-26	-6-22	33.4-97.3	0-69 High =31 Extreme =4	0-51 High =82 Extreme =10	0-19 High =147 Extreme=24	0-7 High =44 Extreme=13

* Rainfall (R), Wind Speed (WS), Max Temperature (MaxT), Average Temperature (AvT), Min Temperature (MinT), Average Relative Humidity (AvRH), Cardiovascular Hospitalization (CH), Respiratory Hospitalization (RH), Cardiovascular Deaths (CD), Respiratory Deaths (RD)

Table 7. Range for Meteorological values and CARD/RES MOMO rates for each cluster

	MaxT daily interval	AvT daily interval	MinT daily interval	AvRH daily interval	SUN daily interval	R daily interval	WS daily interval	CH daily interval / Linguistics	RH daily interval / Linguistics	CD daily interval / Linguistics	RD daily interval / Linguistics
EM-Cluster2 September-May 2004-2013	1.8-31.4	2-26.4	-6.8-21	33.38-94.74	0-12.5	0	0.5-27.75	0-108 High =112 Extreme = 49	0-62 High =124 Extreme = 53	0-19 High =144 Extreme = 29	0-7 High =51 Extreme =17
SIB-Cluster2 September-May 2004-2013	2.4-44	1.9-34.7	0-25.4	30.94-96.27	0-13.9	0-8.3	0.5-27.75	5-91 High =174 Extreme =64	0-62 High =182 Extreme = 61	0-13 High =110 Extreme = 22	0-7 High =43 Extreme =11
SKM-Cluster3 September-June 2004-2013	0.8-28.4	-2.15-22.3	-6.8-19	33.38-89.89	2.4-12.3	0-15.7	1.6-32.5	0-108 High =100 Extreme =42	2-60 High =115 Extreme = 52	0-19 High =114 Extreme =22	0-7 High =46 Extreme = 13
SOM-Cluster2 September-June 2004-2013	-1.4-44	-2.95-34.7	-6.8-25.4	30.94-96.27	0-13.9	0-35.9	0.67-27.75	9-108 High = 180 Extreme =87	3-62 High = 159 Extreme =74	0-14 High = 97 Extreme =25	0-4 High =41 Extreme =10

* Max Temperature (MaxT), Average Temperature (AvT), Min Temperature (MinT), Average Relative Humidity (AvRH), Sunshine Hours (SUN), Rainfall (R), Wind Speed (WS), Cardiovascular Hospitalization (CH), Respiratory Hospitalization (RH), Cardiovascular Deaths (CD), Respiratory Deaths (RD)

Table 8. Range for Air Pollutants and CARD/RES MOMO rates for each cluster

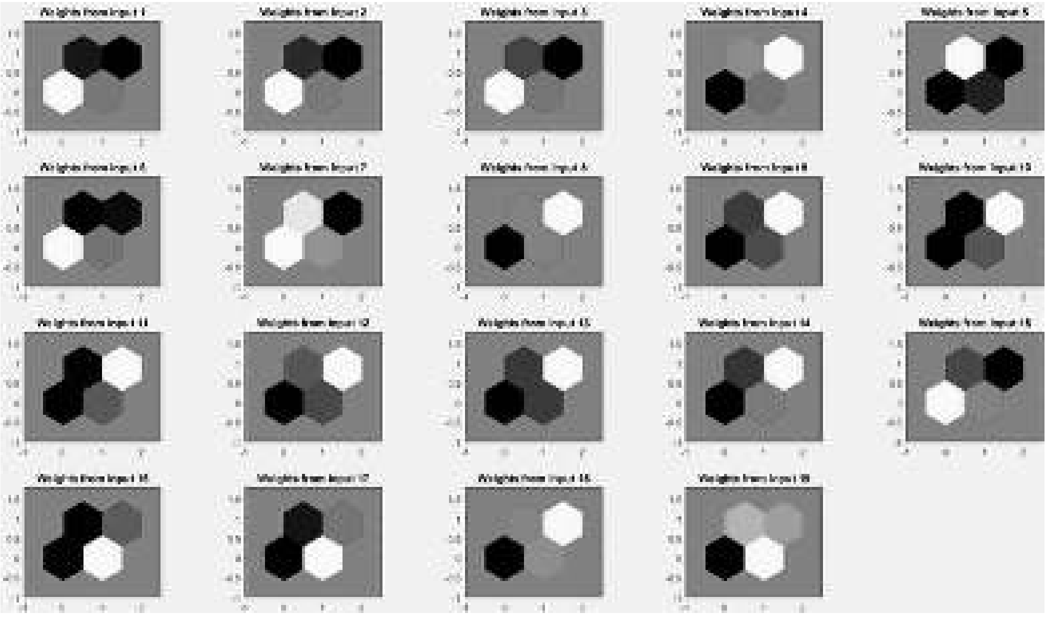
	CO daily interval	NO daily interval	NO_2 daily interval	O_3 daily interval	SO_2 daily interval	PM_{10} daily intervals	$PM_{2.5}$ daily interval	CH daily interval / Linguistics	RH daily interval / Linguistics	CD daily interval / Linguistics	RD daily interval / Linguistics
EM-Cluster2 September-May 2004-2013	0.41-2.23	5.62-86.05	7.65-71.4	21.35-99.44	1.4-41.6	11.18-85.53	2-60	0-108 High =112 Extreme = 49	0-62 High =124 Extreme = 53	0-19 High =144 Extreme = 29	0-7 High =51 Extreme =17
SIB-Cluster2 September-May 2004-2013	0.32-3.37	5.79-90.35	8.1-71.4	20.65-105.83	1-43.36	11.27-158.6	6-67.66	5-91 High =174 Extreme =64	0-62 High =182 Extreme = 61	0-13 High =110 Extreme = 22	0-7 High =43 Extreme =11
SKM-Cluster3 September-June 2004-2013	0.32-3.37	5.51-90.35	9.88-73.6	30.01-93.18	1.4-35.27	11.03-86.04	2-57	0-108 High =100 Extreme =42	2-60 High =115 Extreme = 52	0-19 High =114 Extreme =22	0-7 High =46 Extreme = 13
SOM-Cluster2 September-June 2004-2013	0.32-2.23	5.68-71.6	9.08-71.4	23-105.8	1.4-43.4	15.26-158.6	10.5-100	9-108 High = 180 Extreme =87	3-62 High = 159 Extreme =74	0-14 High = 97 Extreme =25	0-4 High =41 Extreme =10

* Cardiovascular Hospitalization (CH), Respiratory Hospitalization (RH), Cardiovascular Deaths (CD), Respiratory Deaths (RD)

Input10: Particulate Matter (PM_{10}) Input11: Particulate Matter ($PM_{2.5}$), Input12: Carbon Monoxide (CO), Input13: Nitrogen Monoxide (NO), Input14: Nitrogen Dioxide (NO_2), Input15: Ozone (O_3), Input16: Cardiovascular Hospitalization (CH), Input17: Respiratory Hospitalization (RH), Input18: Cardiovascular Deaths (CD), Input19: Respiratory Deaths (RD).

The following conclusions are obtained regarding the data clustering after analyzing te results presented in Figure 4.

Figure 4. SOM Input Planes of the 4 Clusters



- **Cluster 1:** The maximum, moderate and minimum temperatures, the sunshine hours, the wind speed and O_3 are characterizing mainly the 1st cluster. For the records of cluster1, clustering of the atmospheric data is related to the PHOC.
- **Cluster 2:** The maximum and minimum temperatures, $NO_{2,}$ O_3 and Cardiovascular Mortality are the 2nd most important group of characteristics of the 2nd cluster, whereas CARD/RES hospitalization and RES Mortality are the members of the most characteristic vector. This cluster has the most recorded high and extreme MOMO values related to the above diseases.
- **Cluster 3:** Rainfall has the highest influence in cluster#3 whereas wind speed and average moisture are second. This cluster is characterized mostly by the above mentioned two parameters.
- **Cluster 4:** In cluster#4 average moisture, SO_2, $PM_{10,}$ $PM_{2.5}$, CO, NO, NO_2 and CARD Hospitalization constitute the vector of the most influential parameters whereas meteorological data and Ozone are not related. This is due to the fact that this cluster is mainly related to the Smog.

The following Tables 9 and 10 present the profile of each cluster and the range of their parameters.

Step 4: Fuzzy Chi Square Test

The fuzzy chi square test (described in a previous chapter) is performed and applied for all four clusters, in the fourth step in order to determine the level of dependence between the atmospheric features and the CARD/RES MOMO indices at confidence interval of 95%. The basic principles of the algorithm have already been presented. The results of the FUCS are discussed thoroughly in the following chapter (Table 11) in order to enhance the final conclusions. Figure 5 shows the flowchart of the proposed model.

Table 9. Range of the meteorological factors based on the SOM cluster

	MaxT daily interval	AvT daily interval	MinT daily interval	AvRH daily interval	SUN daily interval	R daily interval	WS daily interval	CH daily interval / Linguistics	RH daily interval /Linguistics	CD daily interval / Linguistics	RD daily interval / Linguistics
SOM-Photochemical Cloud-Cluster1 March-November 2004-2013	8-40.3	4.3-32.85	-0.6-27.1	28.25-82.11	0-14	0-23	1.75-32.5	1-90 High =66 Extreme =13	0-46 High =75 Extreme =11	0-13 High =187 Extreme =16	0-5 High =56 Extreme =14
SOM-The most extreme risk incidents-Cluster2 September-June 2004-2013	-1.4-44	-2.95-34.7	-6.8-25.4	30.94-96.27	0-13.9	0-35.9	0.67-27.75	9-108 High = 180 Extreme =87	3-62 High = 159 Extreme =74	0-14 High = 97 Extreme =25	0-4 High =41 Extreme =10
SOM-based on the high rainfall values and on high wind speed-Cluster3 September-May 2004-2013	1-30.2	-3.5-26	-6-22	33.4-97.3	0-11.9	0-131.6	0-26.5	0-69 High =31 Extreme =4	0-51 High =82 Extreme =10	0-19 High =147 Extreme=24	0-7 High =44 Extreme=13
SOM-Smog Cloud-Cluster4 October-April 2004-2013	2-25.6	-2.45-20.45	-7.8-17.6	48.4-97.65	0-10.4	0-28	0-15	4-96 High =56 Extreme =16	0-53 High =66 Extreme =19	0-12 High =81 Extreme =14	0-5 High =21 Extreme =10

Table 10. Range of the air pollution factors based on the SOM cluster

	CO daily interval	NO daily interval	NO_2 daily interval	O_3 daily interval	SO_2 daily interval	PM_{10} daily interval	$PM_{2.5}$ daily interval	CH daily interval / Linguistics	RH daily interval / Linguistics	CD daily interval / Linguistics	RD daily interval / Linguistics
SOM-Photochemical Cloud-Cluster1 March-November 2004-2013	0.25-2.89	4.21-42.2	4.83-58.8	50.69-127.9	1-21	12.66-68.15	6-57	1-90 High =66 Extreme =13	0-46 High =75 Extreme =11	0-13 High =187 Extreme =16	0-5 High =56 Extreme =14
SOM-The most extreme risk incidents-Cluster2 September-June 2004-2013	0.32-2.23	5.68-71.6	9.08-71.4	23-105.8	1.4-43.4	15.26-158.6	10.5-100	9-108 High = 180 Extreme =87	3-62 High = 159 Extreme =74	0-14 High = 97 Extreme =25	0-4 High =41 Extreme =10
SOM-Cluster3-based on the high rainfall values and on high wind speed-September-May 2004-2013	0.33-3.37	5.63-89.2	7.49-62.9	11.9-74.9	1-41.6	11.03-72.2	2-56	0-69 High =31 Extreme =4	0-51 High =82 Extreme =10	0-19 High =147 Extreme=24	0-7 High =44 Extreme=13
SOM-Smog Cloud-Cluster4 October-April 2004-2013	0.66-2.75	33.99-199.9	24.6-93.4	12.94-73.3	2.6-70.9	29.4-203	13-121.5	4-96 High =56 Extreme =16	0-53 High =66 Extreme =19	0-12 High =81 Extreme =14	0-5 High =21 Extreme =10

RESULTS AND DISCUSSION

After extensive testing has been done on all SOM algorithm clusters, we have examined all possible cases of dependence between the Linguistics assigned to the atmospheric parameters and to the CARD/RES MOMO indices. The most important dependences and the highest correlations for each cluster of the SOM algorithm are presented below.

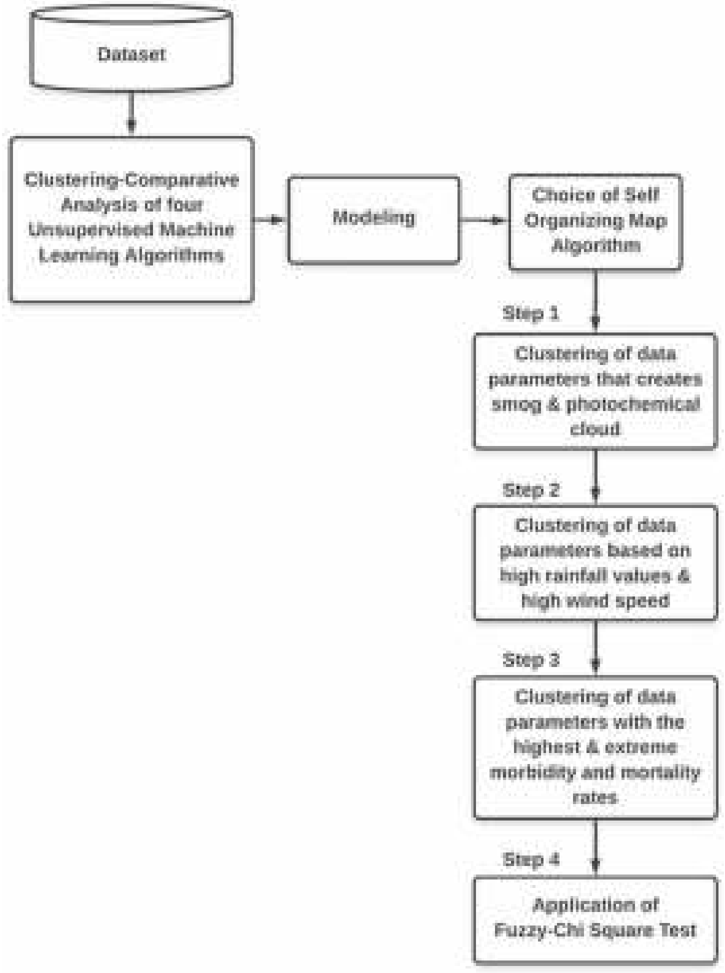
Table 11. Fuzzy chi square test application between involved features in SOM algorithm

Parameters-Diseases-Cluster	Statistic Test	P-Value	Linguistics of P-Value	Degree of membership of linguistics
MinT-CH Cluster1	13.424	0.0094	High Dependence	0.53
MinT-CD Cluster1	9.901	0.0071	High Dependence	0.645
MinT-RD Cluster1	7.945	0.0188	Medium Dependence	0.69
PM_{10}-CH Cluster1	9.447	0.0089	High Dependence	0.555
PM_{10}-CD Cluster1	6.455	0.0111	High Dependence	0.445
NO-RH Cluster1	4.400	0.0359	Medium Dependence	0.455
NO_2-RH Cluster1	19.142	0.0001	High Dependence	0.995
MinT-CH Cluster2	21.221	0.0017	High Dependence	0.915
MinT-CD Cluster2	13.189	0.0042	High Dependence	0.79
R-CD Cluster2	11.875	0.0078	High Dependence	0.61
WS-CH Cluster2	20.606	0.0022	High Dependence	0.89
WS-RH Cluster2	13.103	0.0414	Low Dependence	0.57
AP-RH Cluster2	18.375	0.0054	High Dependence	0.73
AP-RD Cluster2	8.184	0.0423	Low Dependence	0.615
SO_2-CD Cluster2	8.749	0.0328	Medium Dependence	0.61
PM_{10}-CH Cluster2	18.073	0.0061	High Dependence	0.695
NO_2-CH Cluster2	16.645	0.0107	High Dependence	0.465
NO_2-RH Cluster2	14.944	0.0207	Medium Dependence	0.785
NO_2-RD Cluster2	11.001	0.0117	High Dependence	0.415
O_3-RH Cluster2	13.934	0.0075	High Dependence	0.625
AvRH-RH Cluster3	45.664	<0.00001	High Dependence	1
R-CH Cluster3	9.374	0.0247	Medium Dependence	0.985
SO_2-CD Cluster3	12.883	0.0449	Low Dependence	0.745
O_3- RH Cluster3	7.377	0.025	Medium Dependence	1
CO-CH Cluster4	20.408	0.0023	High Dependence	0.885
SO_2-CH Cluster4	19.927	0.0029	High Dependence	0.855

Cluster 1: Clustering the atmospheric parameters that create the Photochemical Cloud

The conditions related to the development of the PHOC belong to the season March to November (2004-2013). The linguistics of the minimum temperature (MinT) have high dependency (HIDE) (MV= 0.645) with the linguistics of the CARD Mortality (CAMO). Also, MinT has significant dependency (SIGDE) to the RES Hospitalizations (RESHO). The PM_{10} have HIDE to the RESHO and to the CAMO. During photochemical reactions the NOx participate in the development of the PHOC (O_3). The linguistics of NO have Moderate dependency MV= 0.455 with the Respiratory hospitalizations. NO_2 has very high dependency (MV=0.995) to the RES hospitalizations. A previous research of (Anezakis, Iliadis, Demertzis, & Mallinis, 2017) has shown that the Pollution due to PHOC (O_3) is high depended (MV=0.83575) with the Cardiovascular deaths on the same day.

Figure 5. Flowchart of the proposed model


Dataset
Clustering-Comparative Analysis of four Unsupervised Machine Learning Algorithms
Modeling
Choice of Self Organizing Map Algorithm
Step 1
Clustering of data parameters that creates smog & photochemical cloud
Step 2
Clustering of data parameters based on high rainfall values & high wind speed
Step 3
Clustering of data parameters with the highest & extreme morbidity and mortality rates
Step 4
Application of Fuzzy-Chi Square Test

Cluster 2: Clustering of the atmospheric parameters that are related to the fuzzy Linguistics of High and Extreme CARD/RES MOMO

We have performed clustering of data related to linguistics of high and extreme Cardiovascular - Respiratory hospitalization risk. This data belongs to the season September to June 2004-2013. In cluster 2 the atmospheric parameters are significantly influencing the CARD/RES Morbidity/Mortality. There is high dependency between minimum temperature and wind speed linguistics to MOMO due to Cardiovascular diseases (MV equal to 0.915 and 0.89 respectively). Also, there is an influence of Atmospheric Pressure to RES Morbidity (MV=0.73). Finally, the RES MOMO is depended to NO_2 concentrations and SO_2 plays a moderate role (MV=0.61) to CARD Mortality whereas O_3 has an influence in respiratory morbidity (high with MV=0.625).

(Anezakis, Iliadis, Demertzis, & Mallinis, 2017) have proven that SO_2, CO, PM_{10}, $PM_{2.5}$, O_3 have a high contribution (MV=0.9916) to the CARD hospitalizations of the same day, in spring time.

Cluster 3: Clustering of the atmospheric parameters based on high rainfall and high wind speed

Data clustering related to high levels of rainfall and high wind speed belongs to the period September to May 2004-2013. A main attribute of this cluster is the lack of high and extreme pollution values (moderate ones are included).

AvRH linguistics are influencing (MV=1) the Respiratory Morbidity and they also had a moderate influence (MV=0.985) to CARD Morbidity. Linguistics related to O_3 had a moderate influence (MV=1) to RES Morbidity.

Cluster 4: Clustering of the atmospheric parameters responsible for Smog

In cluster 4 (related to winter period) the hospitalizations of CARD had a high dependency to the fuzzy linguistics of CO and SO_2 (MV=0.855). (Anezakis, Iliadis, Demertzis, & Mallinis, 2017) have shown that during the winter the Smog index (CO, SO_2, PM_{10}, $PM_{2.5}$) highly influences the Cardiovascular deaths of the same day.

CONCLUSION

This research proposes a hybrid computational intelligence approach, employing Unsupervised Machine Learning, which combines the SOM learning algorithm for clustering data. At the same time the innovative FUCS algorithm determines the degree of dependence/independence of air pollution and meteorological parameters under consideration, to the number of serious Cardiovascular/Respiratory hospitalizations or deaths. The wider urban area of Thessaloniki city has been chosen for the selection of data and for the application of the proposed system.

The validity of the SOM algorithm was confirmed by in-depth testing and comparisons with four unsupervised Machine Learning algorithms in order to select the algorithm that best responds to the nature of the problem.

Summarizing the profile and characteristics of each cluster we can propose the following:

It has been proven that in cluster1 (related to the Photochemical cloud) the minimum temperature and PM_{10} linguistics have a serious influence to the MOMO of Cardiovascular diseases. Also, Respiratory incidents are significantly depended on NO and NO_2.

Cluster2 is characterized by the highest fuzzy linguistics of RES/CARD risk which is influenced by minimum temperature and SO_2. Linguistics of O_3 and NO_2 are contributing to the serious increase of respiratory incidents.

In cluster 3 (which is related to high rainfall and wind speed) (AvRH) and O_3 have a significant contribution to Res Morbidity. On the other hand, SO_2 does not have a serious influence to the Cardiovascular deaths for this season.

In cluster4 (related to the Smog favoring data) CARD hospitalization incidents are depended on SO_2 and CO concentrations.

This research aims to trace the specific role of each air pollutant in the Morbidity and Mortality of the citizens of a major urban center. The intelligent information system that has been developed, can be a very useful tool in the hands of civil protection. It can provide warnings that can contribute towards the better management of the hospitals, when the conditions are critical. Also, for every season the authorities would expect specific types of threats to human health (depending on the cluster) and they will be able to impose proper actions (e.g. alerting hospitals or advising people to avoid traveling depending on their health problem. Overall, the system has the potential to contribute towards the improvement of the quality of life in urban areas.

ACKNOWLEDGMENT

This research was supported by the resources of the second level graduate scholarship Program “Development of human potential, Education and Lifelong Learning” of the National Strategic Framework for Development ESPA 2014-2020 co-funded by the European Social Fund (E.K.T) and the Hellenic Republic [grant number 2017-050-0504-10726].

REFERENCES

Almeida, V. G., Borba, J., Pereira, T., Pereira, H. C., Cardoso, J., & Correia, C. (2013). Data mining based methodologies for cardiac risk patterns identification. In P. Fernandes, J. Solé-Casals Ana, L.N. Fred, & H. Camboa (Eds.), *Proceedings of the International Conference on Bioinformatics Models, Methods and Algorithms*. Barcelona, Spain: SciTePress.

Anezakis, V.-D., Demertzis, K., Iliadis, L., & Spartalis, S. (2016a). A hybrid soft computing approach producing robust forest fire risk indices. In L. Iliadis & I. Maglogiannis (Eds.), *Proceedings of Artificial Intelligence Applications and Innovations, AIAI 2016, IFIP Advances in Information and Communication Technology,* Thessaloniki, Greece, *LNCS* (Vol. 475, pp. 191–203). Cham: Springer. doi:10.1007/978-3-319-44944-9_17

Anezakis, V.D., Demertzis, K., Iliadis, L., & Spartalis, S. (2017). Hybrid intelligent modeling of wild fires risk. *Evolving Systems,* 1-17. doi:10.1007/s12530-017-9196-6

Anezakis, V.-D., Dermetzis, K., Iliadis, L., & Spartalis, S. (2016b). Fuzzy cognitive maps for long-term prognosis of the evolution of atmospheric pollution based on climate change scenarios: The case of Athens. In N.T. Nguyen, L. Iliadis, Y. Manolopoulos & B. Trawiński (Eds.), Proceedings of Computational Collective Intelligence, ICCCI 2016, Halkidiki, Greece, LNCS (Vol. 9875, pp. 175-186). Cham: Springer. doi:10.1007/978-3-319-45243-2_16

Anezakis, V.-D., Iliadis, L., Demertzis, K., & Mallinis, G. (2017). Hybrid Soft Computing Analytics of Cardiorespiratory Morbidity and Mortality Risk Due to Air Pollution. In I. Dokas, N. Bellamine-Ben Saoud, J. Dugdale & P. Díaz (Eds.), Proceedings of Information Systems for Crisis Response and Management in Mediterranean Countries, ISCRAM-med 2017, Xanthi, Greece, LNCS (Vol. 301, pp. 87-105). Cham: Springer. doi:10.1007/978-3-319-67633-3_8

Basara, H. G., & Yuan, M. (2008). Community health assessment using self-organizing maps and geographic information systems. *International Journal of Health Geographics*, *7*(1), 1–8. doi:10.1186/1476-072X-7-67 PMID:19116020

Bougoudis, I., Demertzis, K., Iliadis, L., Anezakis, V.-D., & Papaleonidas, A. (2017). FuSSFFra, a fuzzy semi-supervised forecasting framework: The case of the air pollution in Athens. *Neural Computing & Applications, 29(7), 375-388.* doi:10.1007/s00521-017-3125-2

Bougoudis, I., Dermetzis, K., & Iliadis, L. (2016a). HISYCOL a hybrid computational intelligence system for combined machine learning: The case of air pollution modeling in Athens. *Neural Computing & Applications*, *27*(5), 1191–1206. doi:10.1007/s00521-015-1927-7

Bougoudis, I., Dermetzis, K., & Iliadis, L. (2016b). Fast and low cost prediction of extreme air pollution values with hybrid unsupervised learning. *Journal of Integrated Computer-Aided Engineering*, *23*(2), 115–127. doi:10.3233/ICA-150505

Bougoudis, I., Iliadis, L., & Papaleonidas, A. (2014). Fuzzy Inference ANN Ensembles for Air Pollutants Modeling in a Major Urban Area: The Case of Athens. In: V. Mladenov C. Jayne & L. Iliadis (Eds.), Proceedings of Engineering Applications of Neural Networks, EANN2014, Sofia, Bulgaria, CCIS (Vol.459, pp.1-14). Cham: Springer. doi:10.1007/978-3-319-11071-4_1

Corder, G. W., & Foreman, D. I. (2014). *Nonparametric Statistics: A Step-by-Step Approach* (2nd ed.). Hoboken, NJ: Wiley& Sons.

Cortina-Januchs, M. G., Quintanilla-Dominguez, J., Andina, D., & Vega-Corona, A. (2012). ANN and Fuzzy c-Means applied to environmental pollution prediction. In *Proceedings of World Automation Congress (WAC '12),* Puerto Vallarta, Mexico. IEEE.

Dimou, V., Anezakis, V. D., Demertzis, K., & Iliadis, L. (2017). Comparative analysis of exhaust emissions caused by chainsaws with soft computing and statistical approaches. *International Journal of Environmental Science and Technology*, 1–12. doi:10.1007/s13762-017-1555-0

Grzegorzewski, P., & Szymanowski, H. (2015). Chi-square test for homogeneity with fuzzy data. *Advances in Intelligent Systems and Computing*, *315*, 151–158. doi:10.1007/978-3-319-10765-3_18

Haykin, S. (2009). *Neural networks and learning machines* (3rd ed.). New York: Pearson Education.

Hernawati, K., Insani, N., Bambang, S. H. M., & Nur Hadi, W., & Sahid (2017). Mapping the Indonesian territory, based on pollution, social demography and geographical data, using self organizing feature map. In *Proceedings of the 4th International Conference on Research, Implementation, and Education of Mathematics and Sciences: Research and Education for Developing Scientific Attitude in Sciences and Mathematics, (ICRIEMS '17)*, Yogyakarta State, Indonesia. AIP Publishing. doi:10.1063/1.4995117

Iliadis, L. (2007). *Intelligent Systems and Application in Risk estimation*. Thessaloniki, Greece: Stamoulis A.

Iliadis, L. Bougoudis, l., & Spartalis, S. (2014). Comparison of Self Organizing Maps Clustering with Supervised Classification for Air Pollution Data Sets. In L. Iliadis, I. Maglogiannis & H. Papadopoulos (Eds.), *Proceedings of the Artificial Intelligence Applications and Innovations (AIAI2014),* Rhodes, Greece, *LNCS* (Vol.436, pp. 424-435). Berlin: Springer. doi:10.1007/978-3-662-44654-6_42

Iliadis, L., & Papaleonidas, A. (2016). *Computational Intelligence and Intelligent Agents*. Thessaloniki, Greece: Tziolas A.

International Statistical Classification of Diseases and Related Health Problems 10th Revision. Retrieved December 20, 2017 by http://apps.who.int/classifications/icd10/browse/2016/en#

Kalantzi, E. G., Makris, D., Duquenne, M. N., Kaklamani, S., Stapountzis, H., & Gourgoulianis, K. I. (2011). Air pollutants and morbidity of cardiopulmonary diseases in a semi-urban Greek peninsula. *Atmospheric Environment*, *45*(39), 7121–7126. doi:10.1016/j.atmosenv.2011.09.032

Karatzas, K., & Kaltsatos, S. (2007). Air pollution modeling with the aid of computational intelligence methods in Thessaloniki, Greece. *Simulation Modelling Practice and Theory*, *15*(10), 1310–1319. doi:10.1016/j.simpat.2007.09.005

Kohonen, S. (1989). *Self-organization and associative memory* (3rd ed.). Berlin: Springer. doi:10.1007/978-3-642-88163-3

Kyriakidis, I., Karatzas, K., Papadourakis, G., Ware, A., & Kukkonen, J. (2012). Investigation and forecasting of the common air quality index in Thessaloniki, Greece. In L. Iliadis, I. Maglogiannis, H. Papadopoulos et al. (Eds.), *Proceedings of the 8th Artificial Intelligence Applications and Innovations(AIAI2012), IFIP Advances in Information and Communication Technology (LNCS) (Vol.*382(II), pp.390-400). Halkidiki, Greece: Springer, Berlin, Heidelberg doi:10.1007/978-3-642-33412-2_40

Lin, P. C., Wu, B., & Watada, J. (2012). Goodness-of-fit test for membership functions with fuzzy data. *International Journal of Innovative Computing, Information, & Control*, *8*(10B), 7437–7450.

Pearce, J.L., Waller, L.A., Mulholland, J.A., Sarnat, S.E., Strickland, M.J., Chang, H.H., & Tolbert, P.E. (2015). Exploring associations between multipollutant day types and asthma morbidity: Epidemiologic applications of self-organizing map ambient air quality classifications. *Environmental Health: A Global Access Science Source, 14*(1), 1-12. doi:10.1186/s12940-015-0041-8

Requia, W. J., Koutrakis, P., Roig, H. L., Adams, M. D., & Santos, C. M. (2016). Association between vehicular emissions and cardiorespiratory disease risk in Brazil and its variation by spatial clustering of socio-economic factors. *Environmental Research*, *150*, 452–460. doi:10.1016/j.envres.2016.06.027 PMID:27393825

Štrbová, K., Štrba, R., Raclavská, H., & Bílek, J. (2018). Analysis of air pollution in vertical profile using self-organizing maps. In A. Abraham, A. Haqiq, A. Ella Hassanien et al. (Eds.), *Proceedings of the Third International Afro-European Conference for Industrial Advancement (AECIA '16). Advances in Intelligent Systems and Computing,* Marrakesh, Morocco. Cham: Springer. doi:10.1007/978-3-319-60834-1_17

Voukantsis, D., Karatzas, K., Kukkonen, J., Räsänen, T., Karppinen, A., & Kolehmainen, M. (2011). Intercomparison of air quality data using principal component analysis, and forecasting of PM10 and PM2.5 concentrations using artificial neural networks, in Thessaloniki and Helsinki. *The Science of the Total Environment*, *409*(7), 1266–1276. doi:10.1016/j.scitotenv.2010.12.039 PMID:21276603

Xie, W., Li, G., Zhao, D., Xie, X., Wei, Z., Wang, W., & Liu, J. et al. (2015). Relationship between fine particulate air pollution and ischaemic heart disease morbidity and mortality. *Heart (British Cardiac Society)*, *101*(4), 257–263. doi:10.1136/heartjnl-2014-306165 PMID:25341536

Lazaros S. Iliadis (BSc in Mathematics AUTh, MSc Computer Science, Wales UK, PhD Expert systems, AUTh) is Professor of Applied Informatics in the School of Engineering Department of Civil Engineering, Lab of Mathematics and Informatics of the Democritus University of Thrace. He has authored, co-authored 60 publications in international scientific journals, more than 100 publications in Proceedings of international conferences corresponding to 967 citations and he has been the organizer/chair of more than 18 scientific conferences. He is author and coauthor of 2 scientific books and a member of ACM and IEEE. He has supervised 3 PhD dissertations and 24 MSc ones. He has lectured as a visiting Professor in many Greek and foreign Universities (UOM, UTh, UOL, UEL, CRAN, COV).

Vardis Dimitris Anezakis is a PhD candidate (Forest Informatics) in the Department of Forestry and Management of the Environment and Natural Resources of the Democritus University of Thrace, Greece. His research interests include Hybrid Computational Intelligence modeling of environmental risks and threats. More specifically his PhD research is related to intelligent modeling of the impacts of atmospheric conditions-pollution on public health, considering the climate change contribution. He has published 7 research papers in international scientific journals and 7 in Proceedings of international conferences.

Konstantinos Demertzis is an officer of the Hellenic Army, serving in Research and Informatics Corps. He holds a PhD in Environmental Informatics and Computational Intelligence from the Democritus University of Thrace and an MSc in the Communication & Computer Networking Technologies, from the University of the Aegean. He is a Part-time Lecturer at the Computer and Informatics Engineering department of the Eastern Macedonia & Thrace Institute of Technology, Research Associate at the National and Kapodistrian University of Athens and Research Assistant at the University of Peloponnese. His research interests include Computational Intelligence (Big Data Analytics, Machine Learning), Environmental Informatics (Invasive Species, Climate Change) and Cyber Security (Critical Infrastructure Protection, Malware Analysis).

Georgios Mallinis (1976) obtained his Ph.D. in Remote Sensing and GIS of forest ecosystems, at the Aristotle University of Thessaloniki (2006). He is an Assistant Professor at the Democritus University of Thrace. His educational duties include teaching courses in remote sensing, chartography and GIS. His research is focused on the field of remote sensing and GIS applications in environmental mapping and monitoring as well as in disaster management. Dr. Mallinis has been involved in various projects, the most recent of which are Fire Management concerning major Natural and Cultural Heritage sites of the EUR-OPA countries (GFMC, Council of Europe), Conservation and sustainable capitalization of biodiversity in forested areas-BIOPROSPECT (Interreg BalkanMED) and Forest Roads for Civil Protection-FORCIP+) (DG ECHO). His work has been published in peer-reviews journals, books, and international conferences. Finally, during the last 3 years. Dr. Mallinis serves as the President of the Management Board in Pindos National Park-Greece's largest terrestrial national park.