Comparison of Self Organizing Maps Clustering with Supervised Classification for Air Pollution Data Sets

Ilias Bougoudis, Lazaros Iliadis, Stephanos Spartalis

Comparison of Self Organizing Maps Clustering with Supervised Classification for air pollution data sets

Ilias Bougoudis*[1], Lazaros Iliadis*[2], Stephanos Spartalis[#3]

Democritus University of Thrace
* Department of Forestry & Management of the Environment & Natural Resources, 193 Pandazidou st., 68200 N Orestiada, Greece, [#] Department of Production Management Engineering, Xanthi, Greece

Email: ibougoudis@yahoo.gr, liliadis@fmenr.duth.gr, sspart@pme.duth.gr

Abstract. Air pollution is a serious problem of modern urban centers. The objective of this research is to investigate the problem by using Machine Learning techniques. It comprises of two parts. Firstly, it applies a well established Unsupervised Machine Learning approach (UML) namely Self Organizing Maps (SOM) for the clustering of Attica air quality big data vectors. This is done by using the concentrations of air pollutants (specific for each area) for a period of 13-years (2000-2012). Secondly, it employs a Supervised Machine Learning methodology (SML) by using multi layer Artificial Neural Networks (ML-ANN) to classify the same cases. Actually, the ANN models are used to evaluate the SOM reliability. This is done, because there is no actual and well accepted clustering of the related data to compare with the outcome of the SOM and this adds innovation merit to this paper.

Keywords: Self Organizing Maps, Artificial Neural Networks, Classification, Air Pollution

1. Introduction

Air pollution has been defined as the release of any substances affecting the normal cycle of any vital process or degrade infrastructure [10]. The target of this research effort is to reveal existing air pollution patterns by performing distinct clusterings for each selected measuring station of the Attica area. The clusterings are based on a vast volume of air quality data vectors related to specific air pollutants for each site. The motivation to use Self Organizing Maps was mainly related to their potential to produce a low-dimensional (typically two-dimensional) discretized representation of the input space of the training samples. Another reason was the unsupervised learning mode of SOM [6], [13]. It is well known, that the effectiveness of such an effort is obtained by estimating Sensitivity and Specificity indices as a result of the comparison between the actual clusters to the obtained ones [2]. However, in the case examined here, there is no well accepted clustering either from the literature or from a study of the civil protection authorities. For this reason, Multi Layer Feed Forward (MLFF) ANNs were developed for each measuring location which classify the

available data of each site. The input vectors of the MLFF approach comprised of air pollutants, plus seven meteorological and five daily factors. The outputs of the MLFF ANNs were considered as a comparison metric for the validation of the SOM's efficiency. It is the first time that SML is employed to support the validation of the UML classification. Herein, we will designate by classification the process of supervised learning on labeled data and by clustering the process of unsupervised learning on unlabeled data.

1.1. Literature review

There are several research efforts in the literature that use ANN to model pollution of the atmosphere [1], [5]. [11]. However, only recently some researchers have used SOM clustering to analyze air pollution of major cities. Patterns of air quality have been searched for Mexico City by Neme and Hernandez [10], whereas Karatzas and Voukantsis have done the same for the city of Thessaloniki [7]. Skön et al., have analyzed indoor air quality using SOM [12]. Li and Chou have investigated air pollution spatial variation with SOM [9]. Glorenec applied SOM to forecast Ozone peaks [3]. Unfortunately though, due to different pollution measurements and different approach methods, any comparison of these works is inaccurate. Moreover, due to station's malfunctions in our case, there were a lot of records missing, which made our effort even more challenging. Also, we had only daily measurements for PM pollutants, while daily for every other pollutant. To the best of our knowledge there is no similar research for the wider area of Attica.

2. Materials and Methods

2.1. Area of research and Data

The area of research included four characteristic air pollution and meteorological stations. They were chosen as follows: "Athinas" is located exactly in the heart of the city centre. "Piraeus" has a unique position by the seaside, "Peristeri" is an urban site at Northwest and "Agia Paraskevi" is a suburb in the Northeastern part of Attica, close to a more mountainous area. So each measuring station is characterized by different topographic and geographic attributes. Figure 1 depicts the four locations under study. The available data sets that were used were related to the wider area of Athens, for a period of 13-years (2000-2012) and they varied from station to station. Overall, the pollutants of interest were hourly concentrations of monoxides plus dioxides like CO, NO, NO_2, SO_2, and daily concentrations of Particulate Matter like PM_{10}, ($\frac{\mu g}{m^3}$). The meteorological data were hourly values of air temperature (C^o), solar radiation (Wm^{-2}), wind speed ($\frac{m}{\sec}$) and direction (rad), pressure (mbar), Illumine and relative humidity. Finally, for every record we added its daily attributes; Year, Month, Day, Day_ID (1 for Monday, 7 for Sunday) and Hour. However not all stations

measure the same pollutants. Thus, the inputs fed to both SOMs and MLFF ANNs were dependent on the sensors of each station. The meteorological parameters were the same everywhere and obtained from "Penteli" station, except from "Agia Paraskevi" station, where, because it is located between "Penteli" and "Thiseion" stations, we averaged the meteorological values for it.

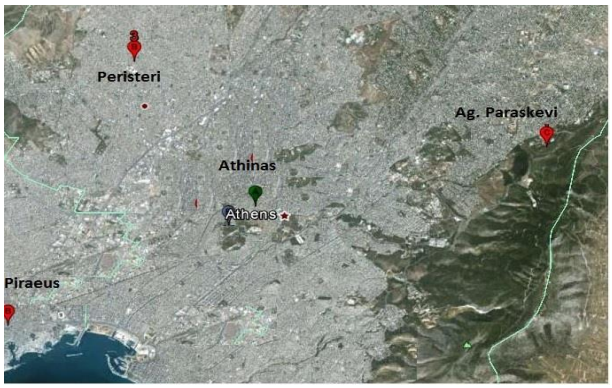

Fig. 1. The four measuring stations

Data were obtained from the Greek ministry of Environment [4]. Totally 1,017,733 vectors without missing values were available, whereas an average as high as 18.82% of the data are missing with the station of Piraeus having the worst percentage equal to 33.67%.

The following table 1 presents the types of parameters measured in each station and percentage of missing values [14].

Table 1. Description of the stations employed for this research

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	Athinas	ATH	21.86%	89,058	O_3, NO, NO_2, CO, SO_2
3	Peristeri	PER	33.61%	75,668	O_3, NO, NO_2, CO, SO_2
4	Piraeus	PIR	33.67%	75,600	O_3,NO, NO_2, CO, SO_2
5	Penteli	PEN	3.66%	109806	Meteorological
6	Thiseion	THI	0.30%	113,632	Meteorological

2.2. Unsupervised Learning

In UML we let the algorithm decide how to group samples into classes that share common properties. Some examples of unsupervised learning are Kohonen's Self Organizing Maps, K-Means Clustering and Neural Gas Networks.

Self Organizing Maps (SOMs) are a well established unsupervised ML approach, based on competitive learning. Their main advantage is their ability to isolate clusters

in high dimensional spaces [9]. The Self Organizing Map as all ANN consists of neurons (nodes). A weight vector is assigned to each neuron. This vector is of the same dimension as the input data vectors. The amount of nodes is the number of the clusters that will be used to group the input data. The obtained Map is a NxN space, where the data are scattered and arranged. The number of neurons is set as the square of the map. Their function can be summarized in four steps [8]:

Initialization: All of the connection weights of each cluster are initialized

Competition: In this stage for each input pattern, the neurons compete to each other in order to "win" this input. The neuron which adapts its value closest to the input "wins". We can define the discriminant function to be the squared Euclidean distance between the input vector x and the weight vector w_j for each neuron j as:

$$d_j(X) = \sum_{i=1}^{D} \left(X_i - W_{ji}\right)^2 \quad (1)$$

Cooperation: Here follows the creation of a neighbourhood located close to the previously winning neuron. In this way, the winning neuron creates a neighbourhood with other neurons, in order to cooperate with each other and win future inputs. If S_{ij} is the lateral distance between neurons i and j on the grid of neurons, we define a topological neighbourhood $T_{j,I(X)}$ where I(x) is the index of the winning neuron.

$$T_{j,I(X)} = \exp[-\frac{S_{j,I(X)}^2}{2\sigma^2}] \quad (2)$$

Adaption: In this last stage, each neuron creates a neighbourhood or becomes a member of a neighbourhood and self - organizes, so that the feature map between inputs is formed. In practice, the appropriate weight update equation is:

$$\Delta W_{ji} = n(t).T_{j,I(X)}(t),(X_i - W_{ji}) \quad (3)$$ [8].

In every step, all neurons adapt their weights to the current input, but not as much as the winner neuron and its neighbourhood [8]. In this way, each neighbourhood is suitable for certain values, and so the map is ordered and shaped.

3. SOM Clustering

In this study, the number of neurons was the only subject of experimentation, in the development of the feature maps. The main target was to isolate the extreme pollutant values of every station and then to record the meteorological and spatiotemporal characteristics for these particular cases. Through this research we have experimented with 2, 3 and 4 neurons. As a result, the obtained maps have 4, 9 and 16 clusters accordingly. Often, the selection of a small number of clusters may lead to loss of the pattern (if there is one). However in this effort, the numbers mentioned above have proven to be suitable. Larger numbers produced larger maps, separating our input data

in a chaotic level. For every map produced, we kept the hits of each cluster, the cluster of each record, the neighbour weight distances and the weights of each input.

The following images show the results of the SOM clustering. They can represent quite efficiently almost every other case. The Neighbour Weight Distance Figure provides information about the neighbourhoods created; the darker colours represent larger distances, and the lighter colours represent smaller distances. In fact we have separated the neurons which were isolated and not part of any neighbourhood. These neurons had the extreme pollutant values and thus they correspond to very interesting cases.

The Input Weights Figure shows the weight assigned to each node from every input. It must be specified that input1 to input5 correspond to CO, NO, NO_2, O_3, SO_2 respectively. Lighter colour means larger value. Here, wherever we have bright colours (yellow or orange), we have the isolated neurons who are assigned the extreme values. Another clue for the most extreme neuron was that this neuron usually had less record than the others (Figure 4). The combination of these three figures helped us to figure out where the most hazardous values were. We are seeking yellow or orange from the Input's Weights Figure, while dark connections in the Distances Figure. The neuron in the bottom - left corner is the first, the one next to it is the second, while the last one is on the top - right corner.

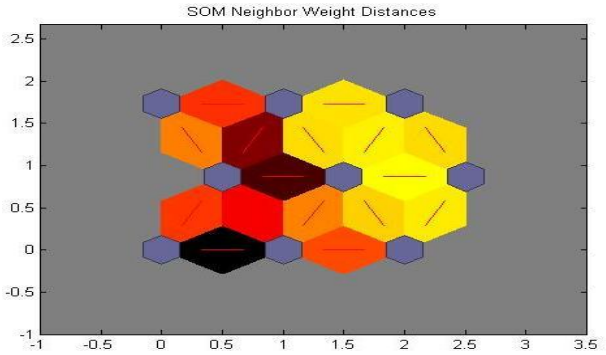

Fig. 2. Neighbor Distances for "Athinas" Hourly pollutants values, period 00-04 (the 1st neuron seems to be the extreme one)

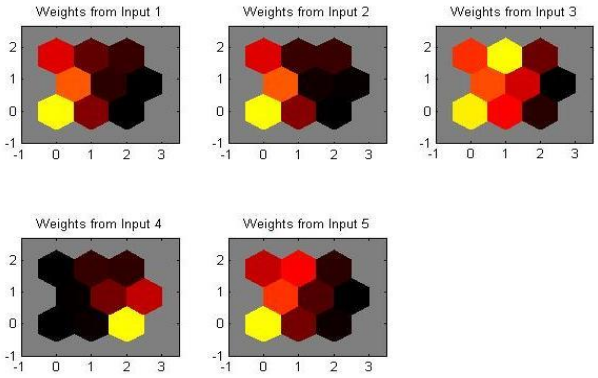

Fig. 3. Input weights for "Athinas" Hourly pollutants values, period 00-04 (the 1st neuron seems to be the extreme one for CO, NO, NO_2 and SO_2, while the third for O_3)

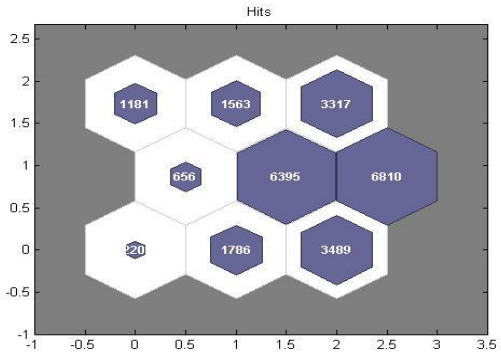

Fig. 4. Sample hits for "Athinas" Hourly pollutants values, period 00-04 (the first neuron has won 220 records, the forth 656, while the ninth 3317)

4. Pattern Recognition with MLFF ANNs

Classification is a form of Supervised Learning. After we obtained the Clusters for each station, we used the pattern recognition tool of MATLAB to classify the available data records by developing MLFF ANNs (three for each site plus a daily one for the stations were we had PM pollutants measures). The purpose was to figure out if obtained grouping was suitable. For the development of the optimal ANNs the data were divided in training, evaluation and testing sets by 70%-30% and 30% respectively. The evaluation metrics were mean square error and confusion matrices. The training function was trainscg (Scaled conjugate gradient backpropagation) and the number of hidden neurons 10, for all networks.

As it has already been mentioned, the input data for every station comprised of the measured pollutants plus seven meteorological parameters (as shown in table 1) and five daily parameters (year, month, day, day identification and hour). By this

approach, evaluation became stricter, as we had 17 input parameters totally. As it is shown in the following Figures, the performance of the ANN is very accurate for the "Peristeri" station for the period 2000-2004.

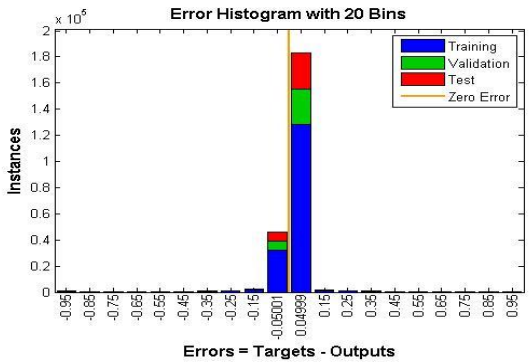

Fig. 5. Error Histogram for the MLFF ANN for the "Peristeri" station (00-04)

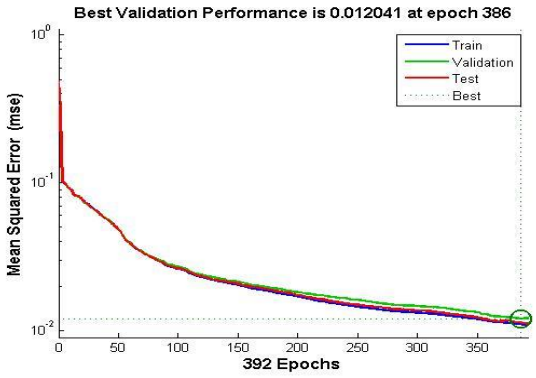

Fig. 6. Performance for the MLFF ANN for the "Peristeri" station (00-04)

If a pollutant was missing values (for example in Agia Paraskeui, the SO_2 pollutant has value only from 2000 to 2005), we created a separate SOM for it. Also, if 9 neurons could not group the extreme values of a pollutant, we developed a second SOM with more neurons (16), in order to manage to cluster the specific pollutant. Unfortunately, the confusion matrix of the networks with 16 clusters is not visible, as it contains 256 elements. For every network we obtained the confusion matrix and the percentages of correct and incorrect classifications. Figure 7 shows the confusion matrix which specifies the high compatibility level between clustering and classification for the "Peristeri" site. Unfortunately there is not enough space for all confusion matrices. However the high agreement of the two approaches is shown in the following tables.

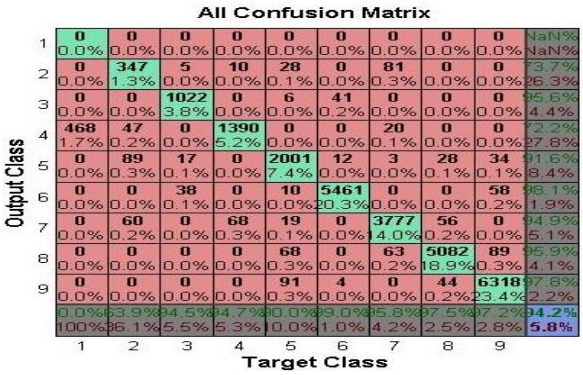

Fig. 7. Overall Confusion Matrix for the MLFF ANN for the "Peristeri" station (00-04)

5. Comparative analysis between SOMs and MLFF ANNs

The tables below show the results from both the clustering and the classification procedures. Columns 2 to 4 are the average, minimum and maximum values for each pollutant. Columns 5 to 8 are the same values extracted from each cluster which contained the extreme values for each pollutant. The last column is the correct percentage of the classification process.

Table 2. Comparative analysis for "Agia Paraskevi" Station

AGP	ALL RECORDS			SOM			CLASS	
	AVG	MIN	MAX	AVG	MIN	MAX	RMSE	%
NO (00-04)	2,75	1	411	25,84	1	411	0,1	93,6
NO_2 (00-04)	20,15	1	347	95,25	45	347		
O_3 (00-04)	91,38	1	330	190,67	155	330		
SO_2 (00-04)	7,35	2	244	25,98	2	244		
NO (05-08)	2,19	1	141	47,96	22	141	0,14	82,9
NO_2 (05-08)	22,37	2	198	73,44	36	165		
O_3 (05-08)	76,34	1	279	166,9	143	279		
SO_2 (05-08)	6,02	2	100	32,2	24	100	0,06	98,1
NO (09-12)	2,46	1	111	11,71	1	111	0,1	92,5
NO_2 (09-12)	13,23	1	287	49,94	22	287		
O_3 (09-12)	86,91	2	251	146,01	129	251		
PM_{10}	26,13	6	292	127,75	91	292	0,19	87,9
$PM_{2,5}$	17,17	4	74	45,87	23	67		

Table 3. Comparative analysis for "Athinas" Station

ATH	ALL RECORDS			SOM			CLASS	
	AVG	MIN	MAX	AVG	MIN	MAX	RMSE	%
CO (00-04)	2,43	0,1	21,4	11,03	5	21,4	0,1	95,0
NO (00-04)	72,86	1	908	659,62	545	908		
NO_2 (00-04)	73,83	1	377	184,14	128	377	0,1	82,5
O_3 (00-04)	33,22	1	253	124,51	93	253		
SO_2 (00-04)	13,29	2	259	41,13	3	259		
CO (05-08)	1,78	0,2	11,8	7,93	4,1	11,8	0,09	95,8
NO (05-08)	57,57	1	787	536	451	787		
O_3 (05-08)	30,97	1	199	-	-	-		
SO_2 (05-08)	8,76	2	126	34,55	2	122		
NO_2 (05-08)	63,92	4	275	147,13	117	275	0,15	43,8
CO (09-12)	1,39	0,1	10,4	5,81	1,7	10,4	0,09	97,5
NO (09-12)	50,72	1	678	431,76	362	678		
NO_2 (09-12)	55,35	3	323	94,4	39	263		
O_3 (09-12)	34,33	1	186	-	-	-		
SO_2 (09-12)	7,63	2	86	19,85	8	60		

Table 4. Comparative analysis for "Piraeus" Station

PIR	ALL RECORDS			SOM			CLASS	
	AVG	MIN	MAX	AVG	MIN	MAX	RMSE	%
CO (00-04)	1,58	0,1	13,3	5,8	0,1	6,6	0,12	90,3
NO (00-04)	55,81	1	590	343,09	246	590		
NO_2 (00-04)	65,41	1	243	108,09	47	224		
O_3 (00-04)	35,05	1	217	114,17	81	217		
SO_2 (00-04)	23,45	2	293	93,57	37	293		
CO (05-08)	1,23	0,1	9,8	4,82	1,2	7,9	0,16	73,6
NO (05-08)	50,66	1	902	377,05	300	902		
NO_2 (05-08)	66,26	1	296	129,25	12	198		
O_3 (05-08)	36,2	1	190	118,54	95	190		
SO_2 (05-08)	18,07	2	275	124,23	57	275		
CO (09-12)	0,88	0,1	6,4	2,75	0,6	6,4	0,14	81,6
NO (09-12)	34,67	1	504	279,9	210	504		
NO_2 (09-12)	49,77	1	277	118,97	61	277		
O_3 (09-12)	42,23	1	192	109,32	92	192		
SO_2 (09-12)	10,36	2	279	50,4	2	207		
PM_{10}	41,89	11	185	104,93	72	185	0,22	91,6
$PM_{2,5}$	29,48	5	157	60,7	12	157		

Table 5. Comparative analysis for "Peristeri" Station

PER	ALL RECORDS			SOM			CLASS	
	AVG	MIN	MAX	AVG	MIN	MAX	RMSE	%
CO (00-04)	0,79	0,1	11,5	4,15	0,3	11,5	0,1	94,2
NO (00-04)	17,72	1	427	198,97	133	427		
NO_2 (00-04)	42,74	1	289	106,38	22	289		

O_3 (00-04)	55,93	1	257	146,76	110	257		
SO_2 (00-04)	14,14	2	272	86,83	6	272		
CO (05-08)	0,71	0,1	8,3	3,6	0,9	8,3		
NO (05-08)	16,27	1	447	202,81	137	447	0,2	60,6
NO_2 (05-08)	40,76	1	353	86,98	42	287		
SO_2 (05-08)	11,48	2	163	36,3	4	156		
O_3 (05-08)	52,98	1	284	224	194	284	0,09	2,3
CO (09-12)	0,52	0,1	8	2,55	0,7	8		
NO (09-12)	9,31	1	284	104,8	60	284		
NO_2 (09-12)	28,46	1	201	79,4	44	201	0,1	96,6
O_3 (09-12)	64,15	1	246	137,31	112	246		
SO_2 (09-12)	6,7	2	106	14,92	2	77		

6. Conclusions – Discussion

In almost every station we managed to isolate the extreme pollutants' values. From the clusterings we have obtained certain pollution patterns comprising of specific temporal, meteorological and air pollutant concentrations. While observing these conditions, we came to the conclusion that high values of CO and NO are present, when we have low temperature, high humidity, low solar radiation and low wind speed, between 8-12 AM or 6-10 PM. On the other hand, high levels of O_3 appear under high temperature, low humidity and high solar radiation and between 12 - 6 PM.

In almost every case, the cluster which included the extreme values was in one corner of the map (in most cases, in the lower right cluster). This was because this cluster was isolated and not part of a neighbourhood.

Although most cases had their extreme values put together in the same cluster, the values of O_3 were separated from the others. What is more, the cluster which contained the extreme values of O_3 was across the one which contained the other pollutants.

Almost all percentages of correct classification were above 80%, showing that the SOM clustering can be considered reliable and compatible to the MLFF ANN classification.

In most cases, when we had two SOMS (with 9 and 16 clusters) the correct percentage of the clustering with 9 clusters is higher from the one with 16 clusters.

Most of the times, the network failed to assign the extreme records in the right neuron. Instead, it put every record in the closest cluster. Although this may sound like a malfunction of the classification, it may be used as a specific pattern recognition tool. This occurs because meteorological parameters act in a catalytic manner and they are not the primary source of pollutants.

This research is quite innovative and it has proven the ability of SOM to reveal clusters with specific attributes in real world problems and moreover their potential contribution for quality of life.

Future work will include the comparison and evaluation of SOM with other unsupervised methods (like fuzzy k- means and Neural Gas Networks), for the same case.

In the matrices below we have the daily attributes modes for every station's extreme pollutant values. Inside the parenthesis we have the counter for each mode (for example, most extreme records for NO_2 appeared in 2009 511 times, while they occurred on Friday (DAY_ID is 5)). PM_{10} and $PM_{2.5}$ have given the same results, so we combined them into one column.

Table 6. Daily attributes for extreme pollutant values for "Agia Paraskevi" Station

AGP	CO	NO	NO_2	O_3	SO_2	PM
YEAR	2009 (511)	2009 (511)	2009 (511)	2010 (453)	2003 (247)	2008 (5)
MONTH	1(253)	1(253)	1 (253)	8(499)	5(147)	4(5)
DAY	16(74)	16(74)	16(74)	6(81)	30(47)	21(2)
DAY_ID	5(210)	5(210)	5(210)	6(277)	4(160)	3(4)
HOUR	8(159)	8(159)	8(159)	14(291)	10(127)	-

Table 7. Daily attributes for extreme pollutant values for "Athinas" Station

ATH	CO	NO	NO_2	O_3	SO_2	PM
YEAR	2009 (139)	2009 (139)	2009 (139)	2002 (444)	2009 (139)	-
MONTH	1(194)	1(194)	1(101)	7(363)	1(194)	-
DAY	3(28)	3(28)	3(28)	4(61)	3(28)	-
DAY_ID	5(72)	5(72)	5(72)	7(184)	5(72)	-
HOUR	9(298)	9(298)	11(165)	16(154)	11(90)	-

Table 8. Daily attributes for extreme pollutant values for "Piraeus" Station

PIR	CO	NO	NO_2	O_3	SO_2	PM
YEAR	2002 (145)	2002 (145)	2002 (294)	2002 (893)	2002 (294)	2007(11)
MONTH	12(76)	12(76)	3(151)	5(294)	3(151)	12(7)
DAY	30(27)	30(27)	26(36)	17(73)	26(36)	1(4)
DAY_ID	3(59)	3(59)	1(95)	6(500)	1(95)	2(9)
HOUR	9(85)	9(85)	10(136)	17(577)	10(136)	-

Table 9. Daily attributes for extreme pollutant values for "Peristeri" Station

PER	CO	NO	NO_2	O_3	SO_2	PM
YEAR	2011 (282)	2011 (282)	2011 (436)	2003 (479)	2011 (282)	-
MONTH	12(203)	12(203)	5(535)	7(579)	12(330)	-
DAY	2(43)	2(43)	13(71)	24(61)	20(44)	-
DAY_ID	5(113)	5(113)	5(231)	5(196)	5(113)	-
HOUR	9(210)	9(210)	8(177)	15(310)	11(103)	-

References

1. Alkasassbeh M., Sheta A.F., Faris H., and Turabieh H. (2013) “Prediction of PM10 and TSP Air Pollution Parameters Using Artificial Neural Network Autoregressive, External Input Models: A Case Study in Salt, Jordan” Middle-East Journal of Scientific Research 14 (7): 999-1009 ISSN 1990-9233
2. Fawcelt T. (2006) "An Introduction to ROC Analysis". Pattern Recognition Letters 27 (8): 861–874. DOI:10.1016/j.patrec.2005.10.010.
3. Glorennec P. Y. (2002) “Forecasting Ozone Peaks Using Self-organizing Maps and Fuzzy Logic” Air Pollution Modelling and Simulation Springer Verlag pp 544-550
4. Iliadis L., S. Spartalis, A. Paschalidou, P. Kassomenos (2007) “Artificial Neural Network Modelling of the surface Ozone concentration” INTERNATIONAL JOURNAL OF COMPUTATIONAL AND APPLIED MATHEMATICS Volume 2, No 2, pp.125-138
5. Jordan, M. I., C. M. Bishop (2004) “Neural Networks”. In Allen B. Tucker. Computer Science Handbook, Second Edition (Section VII: Intelligent Systems). Boca Raton, FL: Chapman & Hall/CRC Press LLC. ISBN 1-58488-360-X
6. Karatzas K., D. Voukantsis (2008) “Studying and predicting quality of life atmospheric parameters with the aid of computational intelligence methods” iEMSs 2008: International Congress on Environmental Modelling and Software Integrating Sciences and Information Technology for Environmental Assessment and Decision Making 4th Biennial Meeting of iEMSs, Proceedings M. Sànchez-Marrè, J. Béjar, J. Comas, A. Rizzoli and G. Guariso (Eds.) International Environmental Modelling and Software Society (iEMSs)
7. Kirt, T., E. Vainik., & L. Võhandu (2007) A method for comparing self-organizing maps: case studies of banking and linguistic data. In Y. Ioannidis, B. Novikov, & B. Rachev (Eds.), Proceedings of eleventh East-European conference on advances in databases and information systems (pp. 107–115). Varna, Bulgaria: Technical University of Varna.
8. Li S. T., S. W. Cho (2000) Multi-Resolution Spatio-temporal Data Mining for the Study of Air Pollutant Regionalization Proceedings of the 33rd Hawaii Conf. on System Sciences
9. Neme A., Hernández L., (2011) “Visualizing Patterns in the Air Quality in Mexico City with Self-Organizing Maps” Proceedings of the 8th WSOM Advances in Self-Organizing Maps Lecture Notes in Computer Science Volume 6731, pp 318-327
10. Paschalidou A., L. Iliadis, P. Kassomenos, C. Bezirtzoglou (2007) “Neural Modeling of the Tropospheric Ozone concentrations in an Urban Site” 10th International Conference Engineering Applications of Neural Networks p.p. 436-445
11. Skön J.P., M. Johansson, M. Raatikainen, U. Haverinen-Shaughnessy, P. Pasanen, K. Leiviskä, M. Kolehmainen (2012) “Analysing Events and Anomalies in Indoor Air Quality Using Self-Organizing Maps” International Journal of Artificial Intelligence Vol 9, N. A12
12. Yin, Hujun; “Learning Nonlinear Principal Manifolds by Self-Organising Maps” in Gorban, Alexander N.; Kégl, Balázs; Wunsch, Donald C.; and Zinovyev, Andrei (Eds.); Principal Manifolds for Data Visualization and Dimension Reduction, Lecture Notes in Computer Science and Engineering (LNCSE), vol. 58, Berlin, Germany: Springer, 2007
13. Greek ministry of Environment http://www.ypeka.gr/