Fast and low cost prediction of extreme air pollution values with hybrid unsupervised learning

Ilias Bougoudis, Konstantinos Demertzis and Lazaros Iliadis*
Department of Forestry and Management of the Environment and Natural Resources, Democritus University of Thrace, Orestiada, Greece

Abstract. Air pollution is the problem of adding harmful substances or other agents into the atmosphere and it is caused by industrial, transport or household activities. It is one of the most serious problems of our times and the determination of the conditions under which we have extreme pollutants’ values is a crucial challenge for the modern scientific community. The innovative and effective hybrid algorithm designed and employed in this research effort is entitled Easy Hybrid Forecasting (EHF). The main advantage of the EHF is that each forecasting does not require measurements from sensors, other hardware devices or data that require the use of expensive software. This was done intentionally because the motivation for this work was the development of a hybrid application that can be downloaded for free and used easily by everyday common people with no additional financial cost, running in devices like smart phones. From this point of view it does not require data from sensors or specialized software and it can offer people reliable information about extreme cases.

Keywords: Fuzzy C-means, neural gas, self-organizing maps, feed forward neural networks, random forest, air pollution, feature selection

1. Introduction

The high concentration of chemicals or particles in the atmosphere results in the alteration of its structure, composition and characteristics and causes serious health problems to the humans and generally to all living creatures and ecosystems. The sources of this problem are man-caused activities concentrated mainly in urban areas (e.g. energy production from solid or liquid fuel, transport, industries, building heating systems and dust). There are primary air pollutants directly emitted (e.g. CO, NO, NO₂, SO₂) and secondary ones (e.g. O₃) that are caused as results of chemical reactions. Though the atmosphere has its own mechanisms to alienate the pollutants, the extreme cases are mainly due to unfavorable meteorological conditions that impede the attenuation of the contaminants. Some of these conditions even accelerate the creation of air pollution.

The analysis and forecasting of high pollution levels in the atmospheric air is one of the most important tasks of the environmental science and research, due to its impact in the fauna and flora and in the human health. This task is even more imperative for densely populated urban areas [29]. Though environmental science has evolved during the last decade, there is still great need of more reliable models, in order to develop reliable prevention and control strategies.

This research paper presents an innovative, accurate and effective fast and low cost forecasting system which allows the prediction of extreme air pollutants values.

1.1. State of the art

Some computational research efforts have been done in the past towards air pollution modeling. Iliadis et
al. [19] have developed FFNNs in order to model Ozone concentrations in Athens. Patterns of air quality have been developed for Mexico City by Neme and Hernandez [26], whereas Karatzas and Voukantsis have done the same for the city of Thessaloniki [20]. Skön et al. have analyzed indoor air quality using SOM [28]. Li and Chou have investigated air pollution spatial variation with SOM [22]. Glorennec applied SOM to forecast Ozone peaks [13].

Also Yong et al. [31] have built a forecasting system based on Support vector machines (SVMs), Xiao et al. [12] proposes a novel hybrid model combining air mass trajectory analysis and wavelet transformation to improve the artificial neural network (ANN) forecast accuracy of daily average concentrations of PM2.5, while Zabkar et al. [32] have applied methods of machine learning to the problem of ground level ozone forecasting. All this proposals used measured data and data calculated by the numerical weather prediction model or stations. On the other hand Lopez-Rubio et al. [23] introduce Bregman divergences in self-organizing models, which is based on stochastic approximation principles, so that more general distortion measures can be employed. A procedure is derived to compare the performance of networks using different divergences. Moreover, a probabilistic interpretation of the model is provided, which enables its use as a Bayesian classifier. Experimental results show the advantages of these divergences with respect to the classical Euclidean distance. Also Menéndez et al. [25] proposes a new algorithm, named genetic graph-based clustering (GGC), takes an evolutionary approach introducing a genetic algorithm (GA) to cluster the similarity graph. The experimental validation shows that GGC increases robustness of spectral clustering and has competitive performance in comparison with classical clustering methods. Donos et al. [11] have presented a study to provide a seizure detection algorithm that is relatively simple to implement on a microcontroller, so it can be used for an implantable closed loop stimulation device. The classification of the features is performed using a random forest classifier. Finally Quiros et al. [27] have extended the traditional definitions of k-anonymity, l-diversity and t-closeness of fuzzy sets as a way to improve the protection of privacy in microdata. The performance of these new approaches is checked in terms of the risk index.

In this paper we propose a new, highly efficient, hybrid model for the prediction of extreme values of air pollutants, which uses as input variables only data which do not include pollutant values or measurements from specialized hardware or software. More specifically, the prediction of a future pollutant’s value is determined by the specific attributes of the time period we want to predict (Year, Month, Day, Hour), temperature (Air Temperature) and values Cluster_Id and Station_Id, which are calculated automatically by geolocation based services.

This is done intentionally because the motivation for this work is the development of a hybrid application that could be downloaded for free and used easily by everyday common people with no additional financial cost, running in devices like smartphones and tablets. This would result in low cost applications that would offer people wide access, regardless of space and time and without the requirement of a specialized expensive device.

An application that will be able to use the EHF model will be running machine learning algorithms directly from the user’s device. For example the operating systems Android-SDK [1] and the iOS-SDK [2] are running applications written in Java, Python and other programming languages and special libraries exist for this purpose like the MLP Neural Net [3] for iOS and the Neuroph [4] for Android which allow the development of machine learning applications in mobile devices. Also another development approach could operate based on the Cloud Software as a Service (SaaS), Cloud Platform as a Service (PaaS) or Cloud Infrastructure as a Service (IAAS). According to this methodology the application will be able to function with the interaction of the device that will provide the data vectors of the independent and depended variables and also by using the cloud service that will provide the machine learning mechanisms.

1.2. Innovation of the EHF project

In a previous work of our research team [16], SOM has been used in order to cluster the obtained pollutants’ concentrations. The ultimate goal was to find the most isolated group, where all the extreme values of pollutants were concentrated. This vital group should comprise of the meteorological and temporal characteristics of the extreme hazardous pollutants. In this clustering effort only the pollutants’ records were used as inputs.

The use of the EHF model capable of forecasting extreme air pollutants’ values will enable the development of an information system towards recording and assessing the level of air pollution in urban centers. The aim and the innovation are threefold. First, in each
I. Bougoudis et al. / Fast and low cost prediction of extreme air pollution values with hybrid unsupervised learning

Table 1: Statistical analysis for the Athinas station

<table>
<thead>
<tr>
<th></th>
<th>CO</th>
<th>NO</th>
<th>NO₂</th>
<th>O₃</th>
<th>SO₂</th>
</tr>
</thead>
<tbody>
<tr>
<td>Records</td>
<td>89057</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MAX</td>
<td>21.4</td>
<td>908</td>
<td>377</td>
<td>253</td>
<td>259</td>
</tr>
<tr>
<td>MIN</td>
<td>0.1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>MODE</td>
<td>0.8</td>
<td>8</td>
<td>60</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>COUNT</td>
<td>4925</td>
<td>2352</td>
<td>1500</td>
<td>6641</td>
<td>10434</td>
</tr>
<tr>
<td>AVERAGE</td>
<td>1.83</td>
<td>59.43</td>
<td>63.61</td>
<td>32.85</td>
<td>9.65</td>
</tr>
<tr>
<td>STDEV</td>
<td>1.48</td>
<td>90.12</td>
<td>27.09</td>
<td>28.61</td>
<td>9.39</td>
</tr>
</tbody>
</table>

Table 2: Statistical analysis for the Patision station

<table>
<thead>
<tr>
<th></th>
<th>CO</th>
<th>NO</th>
<th>NO₂</th>
<th>O₃</th>
<th>SO₂</th>
</tr>
</thead>
<tbody>
<tr>
<td>Records</td>
<td>102065</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MAX</td>
<td>24.6</td>
<td>953</td>
<td>533</td>
<td>145</td>
<td>445</td>
</tr>
<tr>
<td>MIN</td>
<td>0.1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>MODE</td>
<td>1.1</td>
<td>14</td>
<td>83</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>COUNT</td>
<td>3315</td>
<td>816</td>
<td>1175</td>
<td>7635</td>
<td>6412</td>
</tr>
<tr>
<td>AVERAGE</td>
<td>2.52</td>
<td>116.46</td>
<td>88.01</td>
<td>21.67</td>
<td>20.79</td>
</tr>
<tr>
<td>STDEV</td>
<td>1.96</td>
<td>101.24</td>
<td>37.79</td>
<td>21.94</td>
<td>21.48</td>
</tr>
</tbody>
</table>

Table 3: Statistical analysis for the Peiraias station

<table>
<thead>
<tr>
<th></th>
<th>CO</th>
<th>NO</th>
<th>NO₂</th>
<th>O₃</th>
<th>SO₂</th>
</tr>
</thead>
<tbody>
<tr>
<td>Records</td>
<td>75597</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MAX</td>
<td>13.3</td>
<td>902</td>
<td>296</td>
<td>217</td>
<td>293</td>
</tr>
<tr>
<td>MIN</td>
<td>0.1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>MODE</td>
<td>0.5</td>
<td>5</td>
<td>47</td>
<td>4</td>
<td>2</td>
</tr>
<tr>
<td>COUNT</td>
<td>6442</td>
<td>2879</td>
<td>1082</td>
<td>3382</td>
<td>8130</td>
</tr>
<tr>
<td>AVERAGE</td>
<td>1.19</td>
<td>46.26</td>
<td>60.25</td>
<td>38.06</td>
<td>16.64</td>
</tr>
<tr>
<td>STDEV</td>
<td>0.94</td>
<td>56.65</td>
<td>28.24</td>
<td>31.31</td>
<td>20.25</td>
</tr>
</tbody>
</table>

Table 4: Statistical analysis for the Peristeri station

<table>
<thead>
<tr>
<th></th>
<th>CO</th>
<th>NO</th>
<th>NO₂</th>
<th>O₃</th>
<th>SO₂</th>
</tr>
</thead>
<tbody>
<tr>
<td>Records</td>
<td>75665</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MAX</td>
<td>11.5</td>
<td>447</td>
<td>353</td>
<td>284</td>
<td>272</td>
</tr>
<tr>
<td>MIN</td>
<td>0.1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>MODE</td>
<td>0.3</td>
<td>1</td>
<td>12</td>
<td>4</td>
<td>2</td>
</tr>
<tr>
<td>COUNT</td>
<td>15002</td>
<td>13337</td>
<td>1979</td>
<td>1926</td>
<td>14126</td>
</tr>
<tr>
<td>AVERAGE</td>
<td>0.68</td>
<td>14.63</td>
<td>37.63</td>
<td>57.52</td>
<td>10.94</td>
</tr>
<tr>
<td>STDEV</td>
<td>0.68</td>
<td>30.92</td>
<td>26.75</td>
<td>38.41</td>
<td>13.86</td>
</tr>
</tbody>
</table>

Table 5: Overall statistical analysis

<table>
<thead>
<tr>
<th></th>
<th>CO</th>
<th>NO</th>
<th>NO₂</th>
<th>O₃</th>
<th>SO₂</th>
</tr>
</thead>
<tbody>
<tr>
<td>Records</td>
<td>342384</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MAX</td>
<td>24.6</td>
<td>953</td>
<td>533</td>
<td>284</td>
<td>445</td>
</tr>
<tr>
<td>MIN</td>
<td>0.1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>MODE</td>
<td>0.4</td>
<td>1</td>
<td>60</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>COUNT</td>
<td>21942</td>
<td>14854</td>
<td>3979</td>
<td>15732</td>
<td>39101</td>
</tr>
<tr>
<td>AVERAGE</td>
<td>1.58</td>
<td>86.73</td>
<td>35.77</td>
<td>32.65</td>
<td>17.75</td>
</tr>
<tr>
<td>STDEV</td>
<td>1.58</td>
<td>86.73</td>
<td>35.77</td>
<td>32.65</td>
<td>17.75</td>
</tr>
</tbody>
</table>

Month, Day, Day ID (which is equal to 1 for Monday, 2 for Tuesday and so on) plus Hour of measurement.

Moreover, it contains 7 meteorological factors namely: Air Temperature (T), Relative Humidity (RH), Air Pressure (PR), Solar Radiation (SR) available for all years except 2013, Percentage of Sunshine (SUN) (up to 2010), Wind Speed (WS) and Wind Direction (WD).

2. Data

The data used herein were related to the following four measurement stations located in the wider Attica area: Athinas, Patision, Peristeri and Piraeus. The data records contained hourly measurements of the following air pollutants: CO, NO, NO₂, O₃, SO₂. CO was measured in mg/m³ whereas all other measures were measured in µg/m³. For this research and for each station the data records cover the period from 2000 till 2012. 2013 data vectors were used to test the developed forecasting framework.

The following tables contain a statistical analysis of the air pollution concentrations for the period 2000–2012.

Each data record contains the concentrations of the five available air pollutants and also the current Year, forecasting the parameters do not require measurements from sensors or other hardware devices. Moreover, the most important target of such a system will be its application and execution in everyday low cost devises like smart-phones that will use their embedded applications to forecast the extreme air pollutants’ values in a real time mode. The potential users of such an application will have easy and continuous access to the above forecasting and they will have the potential to phase serious air pollution incidents. On the other hand these devices will be functioning in an alternative way with ecological consciousness. Finally it will enable the determination of immediate prevention and protection measurements and actions.

2.1. Contribution of clustering

Initially before the application of any type of clustering, there was an effort to forecast the actual values of each air pollutant (depended feature), without the input of any other individual pollutant in the independent parameters. This effort was not successful and the performance was quite low even though several algorithms were tried namely: (Feed Forward Neural Networks, Random Forest, and Support Vector Machines-SVM).

The following Fig. 1 presents the performance evaluation of one of these unsuccessful efforts for CO and for Athinas station.

The clustering was done in order to determine the conditions under which extreme pollutant values
emerge. Due to the specific physicochemical composition and to the adverse benign conditions related to each pollutant we have developed two types of groups of extreme values for each measuring station. The first category was related to the extreme values of the four primary pollutants (CO, NO, NO\(_2\), SO\(_2\)) and the second one to the extreme values for Ozone (O\(_3\)) which is a secondary pollutant. The final result was the creation of the EDF\(_{ijpr}\) (\(i\) to the number of primary pollutants and \(j\) is the number of measuring stations) files containing all of the records related to high concentrations of primary pollutants for each station and the creation of the EDF\(_{ijsec}\) (\(i\) to the number of secondary pollutants and \(j\) is the number of measuring stations) files related to the O\(_3\).

2.1.1. Clustering

Clustering has been performed with using Fuzzy C-means, Neural Gas Artificial Neural Networks (NGANN), Unsupervised Self Organizing Maps (UNSM) and Semi Supervised Self Organizing Maps (SEMSOM).

When SOM’s are employed the following three basic procedures are executed:

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

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

(a) It should be symmetric to the point of the local minimum (point of winning neuron) where \(d_{ji} = 0\).

(b) The amplitude of the function should be reduced monotonically as the distance \(d_{ji}\) from the winning neuron increases. The function that satisfies the above limitations and it was used in this research is the following Gaussian

\[
h_{ji}(x) = \exp\left(-\frac{d_{ji}^2}{2\sigma^2}\right) \tag{1}\]

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

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

It should be mentioned that \(\sigma_0\) is the initial value of the effective width and

\[
\tau_1 = \frac{n_0}{\ln(\sigma_0)} \tag{3}\]

(c) Synaptic Weight Adaption. In this last training stage the weight vectors of the competitive neurons are updated. The value of this change is given by the following Eq.:

\[
\Delta w_j = \eta h_{ji}(x) (x - w_j), \tag{4}\]

where \(i\) is the winning neuron and \(j\) is a neuron in its neighborhood. Given the weight vector \(w_j(n)\) for a specific time point \(n\), we estimate the new vector for the moment \(n + 1\) from the following Eq. (5):

\[
w_j(n + 1) = w_j(n) + \eta(n) h_{ji}(x) (n) (x(n) - w_j(n)). \tag{5}\]
whereas the SEMSOM creates maps with dimensions λNGANN require the monitoring of the parameters and it is gradually reduced to 0.01 by using the above creates automatically the.

The learning rate \( \eta \) controls how much weight is given to the closest points, weighted by their degree of belonging to the cluster \([8,10]\): 

\[
C_k = \frac{\sum x w_k(x)^m}{\sum w_k(x)^m}
\]

The degree of belonging, \( w_k(x) \), is related inversely to the distance from \( x \) to the cluster center as calculated on the previous pass. It also depends on a parameter \( m \) that controls how much weight is given to the closest center.

### 2.2. Clustering analysis

The statistical data of the extreme air pollution groups created from NGANN algorithm for the Athinas station are presented in the Tables 6 and 7. It should be clarified that in all of the tables, M stands for month, D for date, D_Id is a code used to identify each day where 1 stands for Monday and H for the hour of the day.

During the study of the extreme pollutants’ values (EXPV) we have reached the following general and environmental conclusions:

(i) The EXPV are divided in two basic groups namely: The group with the primary pollutants (CO, NO, NO\(_2\), SO\(_2\)) and the one with the secondary pollutants (O\(_3\)).
(ii) There is negative correlation between the extreme values of O₃ with the rest of the pollutants.

(iii) The EXPV of the primary team appear usually during the winter months from 8–10 in the morning, when the temperature is low, relative moisture is high, solar radiation is low and the wind speed is low.

(iv) The EXPV of the secondary pollutants are recorded during the spring or summer months either from (16–18) hours, or very early in the morning (e.g. 4 in the morning) and when temperature is high relative moisture is low, solar radiation is high and wind speed is high.

(v) There are much more EXPV recorded for O₃ compared to the EXPV related to primary pollutants.

These conclusions emerged from clustering, during which the pollutants’ values were grouped and the correlation between the conditions that create these values was identified. They illustrate the physicochemical mechanisms that contribute to the occurrence of extreme pollutant values, while the statistical analysis assists in evaluating the hazardousness of each area.

Moreover, because it is a real problem with real data, this analysis represents an evaluation of the method, as it can explain and verify whether the clustering of the values with machine learning methods is consistent and in line with the analyzes of the environmental scientists for these specific conditions, but also with common sense.

2.3. Datasets used

After finishing the clustering process, four (4) datasets were created with the EXPVs, one corresponding to each clustering algorithm. Also, two additional characteristics were added that define the correlation of each case to the cluster type. More specifically, the attribute “cluster parameter” was added, whose value equals to 1 if the data vector is related to an extreme case of primary pollutant and it is equal to 2 if the record is related to EXPV for secondary pollutant (O₂).

In this way the homogeneity of the cases’ clustering was achieved and the problem of the inverse correlation between some of them was resolved. For example, in the cases when we had extreme values of O₃, we had extremely low values for CO and NO and vice versa. Another added attribute was the station code, a unique value corresponding to each measuring station. The codes are the following: 1 for “Athinas” station, 2 for “Patission”, 3 for “Piraeus” and finally 4 for “Peristeri”.

The independent variables of the dataset were the following: five temporal (Year, Month, Day, Day_Id, Hour), six meteorological (AirTemp, RH, PR, SR, WS, WD), the station code, (Station) and the cluster number where the record belongs to (Cluster).

The depended variable was each time one of the following five air pollutants (CO, NO, NO₂, O₃, SO₂).

The emerged datasets are:
(i) SOM Weka Dataset: 213,058 records.
(ii) SOM Matlab Dataset: 30,076 records.
(iii) Fuzzy Dataset: 91,440 records.
(iv) Neural Gas Dataset: 53,589 records.

2.3.1. Data preprocessing

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

2.3.1.1. Data normalization

Data normalization was performed for the concentrations of air pollutants, in order to phase the problem of prevalence of features with wider range over the ones with a narrower range, without being more important.

The result was to keep all of their values in the closed interval [−1, +1] by using the following Eq. (16):

\[
x_{1_{\text{norm}}} = 2 \times \left( \frac{x_1 - x_{\text{min}}}{x_{\text{max}} - x_{\text{min}}} \right) - 1 \quad x \in \mathbb{R}
\]
2.3.1.2. Outliers and extreme values analysis

An outlier is a point that lies far away from the mean value of a feature. The distance is usually measured as a multiple of the standard deviation (stdev). For a parameter that follows normal distribution a distance equal to twice the stdev covers 95% of the expected values, whereas this percentage grows to 99% when we are dealing with a distance three times the stdev. Data records with values far away from the mean value are the cause of serious errors in the training phase and they have destructive results. Also, extreme values are very specific kinds of outliers, in which it is assumed that the values which are either too large or too small are outliers.

This bad influence gets even worse when the extreme values are due to noise that has emerged during measuring [5,6].

If the number of the extreme values is small then the corresponding records can be removed from the dataset and they can be analyzed independently.

The Interquartile Range – IQR approach was used to trace the extreme values.

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

After using the above method the following numbers of data vectors were determined as having extreme values.

(a) SOM Weka Dataset: (30264 Outliers and 2205 Extreme Value).
(b) SOM Matlab Dataset: (5699 Outliers and 3191 Extreme Value).
(c) Fuzzy Dataset: (7639 Outliers and 9721 Extreme Value).
(d) Neural Gas Dataset: (5231 Outliers and 9184 Extreme Value).

2.3.1.3. Feature reduction

Feature reduction was performed in order to improve the regression results, by locating the cases with noise that reduces performance. In the feature selection process we tried to exclude the parameters that might contain information irrelevant to the case under analysis.

Principal Components Analysis (PCA) was used for the transformation of the vector feature space [6]. However, unfortunately after the application of PCA, the efficiency and the obtained results were not improved for the cases of Random Forests, k-nearest Neighbor (k_NN) and Multi-Layer ANN, whereas for the Linear Regression and for the e-SVR the performance remained very low.

2.3.1.4. Feature selection

Feature selection is the process where a subset with the most relative and correlated variables is chosen for the development of the optimal model suitable for a specific case study [6]. This process includes two parts [14]:

1) Attribute Evaluator. The Attribute Evaluator is a method in which a subset of attributes is assessed. In this research effort the Classifier Subset Eval method was used. According to this approach the assessment is done based on the regression algorithm used in each case.

2) Search Method. According to this method the most suitable subset is searched and chosen. In this study the Genetic Search method was used. This approach employs a simple Genetic algorithm to determine the proper attributes.

The best performance for each air pollutant was achieved from the following datasets and the use of their respective features.

1) CO, Neural Gas Dataset: (Year, Month, Day, Day_Id, Hour, Air Temp, Pressure, Cluster_Id, Station_Id).
2) NO, SOM Matlab Dataset: (Year, Month, Day, Day_Id, Hour, AirTemp, Cluster_Id, Station_Id)
3) NO₂, Neural Gas Dataset: (Year, Month, Day, Hour, Air Temp, Pressure, Wind Direction, Cluster_Id, Station_Id).
4) O₃, SOM Weka Dataset: (Year, Month, Day_Id, Hour, Air Temp, Pressure, Solar Radiation, Wind Speed, Wind Direction, Cluster_Id, Station_Id).
5) SO₂, Fuzzy C-means Dataset: (Year, Month, Day, Hour, Air Temp, Pressure, Cluster_Id, Station_Id).

2.3.1.5. A heuristic further feature selection

After the feature selection, we decided to reduce the input attributes even more: As a result, we created the “formula” vector, which consists of: Year, Month, Day, Hour, Air Temperature, Cluster_Id and Station_Id. This approach that uses the common features of the formula vector has been named as “Fore-
Table 8: Comparison of 6 Regressions’ Performances ($R^2$) (Dataset SOM Matlab)

<table>
<thead>
<tr>
<th></th>
<th>CO</th>
<th>NO</th>
<th>NO$_2$</th>
<th>O$_3$</th>
<th>SO$_2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>All</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Attr_Sel</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Random forest</td>
<td>0.934</td>
<td>0.945</td>
<td>0.846</td>
<td>0.857</td>
<td>0.926</td>
</tr>
<tr>
<td>K-NN</td>
<td>0.864</td>
<td>0.868</td>
<td>0.7651</td>
<td>0.7753</td>
<td>0.8458</td>
</tr>
<tr>
<td>Linear regression</td>
<td>0.451</td>
<td>0.4688</td>
<td>0.4196</td>
<td>0.4349</td>
<td>0.4514</td>
</tr>
<tr>
<td>RBF network</td>
<td>0.050</td>
<td>0.0550</td>
<td>0.0454</td>
<td>0.0495</td>
<td>0.0570</td>
</tr>
<tr>
<td>MultiLayer perceptron</td>
<td>0.882</td>
<td>0.884</td>
<td>0.842</td>
<td>0.849</td>
<td>0.9002</td>
</tr>
<tr>
<td>ε-SVR</td>
<td>0.379</td>
<td>0.3840</td>
<td>0.3092</td>
<td>0.3147</td>
<td>0.3641</td>
</tr>
<tr>
<td>FFNN</td>
<td>0.941</td>
<td>0.949</td>
<td>0.968</td>
<td>0.978</td>
<td>0.855</td>
</tr>
</tbody>
</table>

Table 9: Comparison of 6 Regressions’ Performances ($R^2$) (Dataset SOM Weka)

<table>
<thead>
<tr>
<th></th>
<th>CO</th>
<th>NO</th>
<th>NO$_2$</th>
<th>O$_3$</th>
<th>SO$_2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>All</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Attr_Sel</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Random forest</td>
<td>0.792</td>
<td>0.811</td>
<td>0.799</td>
<td>0.809</td>
<td>0.943</td>
</tr>
<tr>
<td>K-NN</td>
<td>0.738</td>
<td>0.7415</td>
<td>0.671</td>
<td>0.686</td>
<td>0.795</td>
</tr>
<tr>
<td>Linear regression</td>
<td>0.521</td>
<td>0.5408</td>
<td>0.5790</td>
<td>0.606</td>
<td>0.655</td>
</tr>
<tr>
<td>RBF network</td>
<td>0.194</td>
<td>0.207</td>
<td>0.175</td>
<td>0.184</td>
<td>0.221</td>
</tr>
<tr>
<td>MultiLayer perceptron</td>
<td>0.790</td>
<td>0.792</td>
<td>0.772</td>
<td>0.796</td>
<td>0.909</td>
</tr>
<tr>
<td>ε-SVR</td>
<td>0.319</td>
<td>0.324</td>
<td>0.438</td>
<td>0.457</td>
<td>0.496</td>
</tr>
<tr>
<td>FFNN</td>
<td>0.67</td>
<td>0.7</td>
<td>0.615</td>
<td>0.65</td>
<td>0.72</td>
</tr>
</tbody>
</table>

cast Framework” to forecast the extreme air pollutants’. The overall algorithmic approach that was proposed herein is described clearly and in details in the Fig. 3.

3. Regression

In the next step the obtained four data sets were used to perform regression where one air pollutant was the depended parameter. What is very important here is the fact that (as it has already been mentioned) no pollutant measurements were used as inputs deliberately just to keep the application as simple as possible and easy to use. The following Machine Learning approaches were chosen to be employed for regression herein: (Random Forests, k-Nearest Neighbor (k-NN), Feed Forward ANN, Radial Basis Function ANN and Support Vector Machines like ε-SVR). Also Linear Regression was used. However after the performance of regression only the Random forests and the Multi-layer ANN had a reliable performance. As a result, we are going to deal only with them.

The above modeling algorithms are quite easy to be developed and fast to run, without requiring extra computational power or resources. They produce reliable results and they are ideal to be used by machines of low cost.

3.1. Regression with Random Forests

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

For $b = 1, \ldots, B$:
### Table 10
Comparison of 6 Regressions’ Performances ($R^2$) (Dataset Fuzzy C-means)

<table>
<thead>
<tr>
<th></th>
<th>CO</th>
<th>NO</th>
<th>NO₂</th>
<th>O₃</th>
<th>SO₂</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>All Attr Sel</td>
<td>All Attr Sel</td>
<td>All Attr Sel</td>
<td>All Attr Sel</td>
<td>All Attr Sel</td>
</tr>
<tr>
<td>Random Forest</td>
<td>0.923</td>
<td>0.934</td>
<td>0.923</td>
<td>0.939</td>
<td>0.899</td>
</tr>
<tr>
<td>K-NN</td>
<td>0.905</td>
<td>0.921</td>
<td>0.916</td>
<td>0.919</td>
<td>0.804</td>
</tr>
<tr>
<td>Linear Regression</td>
<td>0.733</td>
<td>0.744</td>
<td>0.788</td>
<td>0.795</td>
<td>0.750</td>
</tr>
<tr>
<td>RBF Network</td>
<td>0.284</td>
<td>0.299</td>
<td>0.338</td>
<td>0.348</td>
<td>0.373</td>
</tr>
<tr>
<td>MultiLayer Perception</td>
<td>0.895</td>
<td>0.919</td>
<td>0.920</td>
<td>0.919</td>
<td>0.822</td>
</tr>
<tr>
<td>ε-SVR</td>
<td>0.375</td>
<td>0.389</td>
<td>0.425</td>
<td>0.434</td>
<td>0.318</td>
</tr>
<tr>
<td>FFNN</td>
<td>0.934</td>
<td>0.934</td>
<td>0.937</td>
<td>0.945</td>
<td>0.843</td>
</tr>
</tbody>
</table>

### Table 11
Comparison of 6 Regressions’ Performances ($R^2$) (Dataset Neural Gas)

<table>
<thead>
<tr>
<th></th>
<th>CO</th>
<th>NO</th>
<th>NO₂</th>
<th>O₃</th>
<th>SO₂</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>All Attr Sel</td>
<td>All Attr Sel</td>
<td>All Attr Sel</td>
<td>All Attr Sel</td>
<td>All Attr Sel</td>
</tr>
<tr>
<td>Random Forest</td>
<td>0.940</td>
<td>0.948</td>
<td>0.956</td>
<td>0.963</td>
<td>0.869</td>
</tr>
<tr>
<td>K-NN</td>
<td>0.634</td>
<td>0.646</td>
<td>0.809</td>
<td>0.823</td>
<td>0.786</td>
</tr>
<tr>
<td>Linear Regression</td>
<td>0.349</td>
<td>0.362</td>
<td>0.375</td>
<td>0.394</td>
<td>0.474</td>
</tr>
<tr>
<td>RBF Network</td>
<td>0.170</td>
<td>0.183</td>
<td>0.161</td>
<td>0.170</td>
<td>0.196</td>
</tr>
<tr>
<td>MultiLayer Perception</td>
<td>0.896</td>
<td>0.915</td>
<td>0.925</td>
<td>0.938</td>
<td>0.840</td>
</tr>
<tr>
<td>ε-SVR</td>
<td>0.30</td>
<td>0.314</td>
<td>0.212</td>
<td>0.220</td>
<td>0.263</td>
</tr>
<tr>
<td>FFNN</td>
<td>0.942</td>
<td>0.95</td>
<td>0.960</td>
<td>0.967</td>
<td>0.873</td>
</tr>
</tbody>
</table>

(i) Sample, with replacement, $n$ training examples from $X, Y$ call these $X_b, Y_b$.
(ii) Train a decision or regression tree $f_b$ on $X_b, Y_b$. $B$ is a free parameter. Typically, a few hundred to several thousand trees are used, depending on the size and nature of the training set. An optimal number of trees $B$ can be found using cross-validation.

After training, predictions for unseen samples $x'$ can be made by averaging the predictions from all the individual regression trees on $x'$:

$$\hat{f} = \frac{1}{B} \sum_{b=1}^{B} \hat{f}_b (x')$$  \hspace{1cm} (12)

or by taking the majority vote in the case of decision trees [9].

#### 3.2. Regression with FFNN

When the FFNN [17] are employed, the training process includes the following steps [15]:

(i) The weighted sums of the inputs are calculated based on Eq. (12):

$$s_j = \sum_{i=1}^{n} (W_{ij} X_i) - \theta_j \quad j = 1, 2, \ldots, h$$  \hspace{1cm} (13)

where $n, h, m$ are the number of input, hidden and output nodes respectively and $W_{ij}$ is the connection weight from the $i$th node of the input layer to the $j$th node of the hidden layer. Also $\theta_j$ is the bias (threshold).

(ii) The output for each hidden node is estimated with the following Eq. (13):

$$S_j = \text{sigmoid}(s_j) = \frac{1}{1 + \exp(-s_j)}$$  \hspace{1cm} (14)

$$j = 1, 2, \ldots, h$$
Table 12  
Training results with the Formula (2000–2012)

<table>
<thead>
<tr>
<th></th>
<th>CO</th>
<th>NO</th>
<th>NO2</th>
<th>O3</th>
<th>SO2</th>
<th>ALL</th>
</tr>
</thead>
<tbody>
<tr>
<td>2000–2012</td>
<td>R²</td>
<td>RMSE</td>
<td>R²</td>
<td>RMSE</td>
<td>R²</td>
<td>RMSE</td>
</tr>
<tr>
<td>SOM</td>
<td>0.86</td>
<td>0.75</td>
<td>0.92</td>
<td>36</td>
<td>0.74</td>
<td>19.2</td>
</tr>
<tr>
<td>Neural Gas</td>
<td>0.90</td>
<td>0.7</td>
<td>0.94</td>
<td>33</td>
<td>0.74</td>
<td>17.6</td>
</tr>
<tr>
<td>Fuzzy C-means</td>
<td>0.88</td>
<td>0.62</td>
<td>0.92</td>
<td>30.27</td>
<td>0.72</td>
<td>15.4</td>
</tr>
<tr>
<td>SOM_WEKA</td>
<td>0.42</td>
<td>1.29</td>
<td>0.37</td>
<td>76.39</td>
<td>0.54</td>
<td>23.63</td>
</tr>
</tbody>
</table>

Table 13  
Testing results with the Formula (2013)

<table>
<thead>
<tr>
<th></th>
<th>CO</th>
<th>NO</th>
<th>NO2</th>
<th>O3</th>
<th>SO2</th>
<th>ALL</th>
</tr>
</thead>
<tbody>
<tr>
<td>2013</td>
<td>R²</td>
<td>RMSE</td>
<td>R²</td>
<td>RMSE</td>
<td>R²</td>
<td>RMSE</td>
</tr>
<tr>
<td>SOM</td>
<td>0.77</td>
<td>0.53</td>
<td>0.83</td>
<td>40</td>
<td>0.48</td>
<td>17.9</td>
</tr>
<tr>
<td>Neural Gas</td>
<td>0.76</td>
<td>0.62</td>
<td>0.90</td>
<td>30.1</td>
<td>0.49</td>
<td>16.2</td>
</tr>
<tr>
<td>Fuzzy C-means</td>
<td>0.76</td>
<td>0.57</td>
<td>0.85</td>
<td>40.6</td>
<td>0.53</td>
<td>14.5</td>
</tr>
<tr>
<td>SOM_WEKA</td>
<td>0.19</td>
<td>0.98</td>
<td>0.38</td>
<td>58</td>
<td>0.25</td>
<td>25.1</td>
</tr>
</tbody>
</table>

(iii) The final output is estimated based on the Eqs (14) and (15) below [18,22]:

\[
o_k = \sum_{j=1}^{h} (W_{jk}S_j) - \theta'_k \quad k = 1, 2, \ldots, m \quad (15)
\]

\[
O_k = \text{sigmoid}(o_k) = \frac{1}{(1 + \exp(-o_k))} \quad k = 1, 2, \ldots, m \quad (16)
\]

4. Results

The algorithms used for regression and the results of their efficiency in the training process are presented in the tables below. It should be clarified that in all of the tables, All stands for All Attributes and Attr Sel for Attribute Selection.

Root Mean Square Error (RMSE) and Coefficient of Determination (R²) given in the following Eqs (19) and (20) were used as criterions to check the validity of the performed regressions. Each criterion is represented by indices which are determined by comparing the forecasted to the actual values of the EXPV. The 10-Fold Cross Validation approach was employed and thus the average values of the above indices were obtained. This was done in order to enhance the generalization ability of the developed models. In the Eqs (19) and (20) below, O_i stands for the actual values measured directly, stands for their mean values, and P_i stands for the produced values by the model.

\[
RMSE = \sqrt{\frac{\sum_{i=1}^{N} (O_i - P_i)^2}{N}} \quad (17)
\]

\[
R^2 = 1 - \frac{\sum_{i=1}^{N} (O_i - P_i)^2}{\sum_{i=1}^{N} (O_i - \bar{O})^2} \quad (18)
\]

From the above tables it is concluded that the FFNN algorithm offers the most reliable results and simulates the relation between the independent and depended parameters more efficiently. More specifically, FENN presents the smallest average error values for the RMSE indices, which show that it outperforms the other approaches. We are reaching the same conclusion by comparing the R² values. The Coefficient of Determination is much high effort the FFNN algorithm.

4.1. Application of the EHF for 2013

The proposed regression models were constructed by the use of extreme values data records related to the time interval 2000–2012. In order to show the validity of this algorithm, we performed four regression analysis (one for each data set namely: SOM Matlab, Neural Gas, Fuzzy C-means, SOM Weka) where the input vectors comprised of data vectors related to extreme pollutants from 2000–2012 and the output vectors were related to 2013. The developed Feed Forward ANN was applied to forecast the extreme air pollutants value for the year 2013. In the Fig. 4 below we see its performance by using the attributes of the Forecast Framework for 2000–2012 for the CO air pollutant.

The following Tables 12 and 13 present the performance of the Forecast Framework in training and testing respectively.

While scrutinizing Tables 8–13, we can draw several conclusions about our data and the methodologies used: Firstly, the SOM_WEKA dataset produces...
the worst results; no matter which algorithm we implemented upon this dataset, the results were always frustrating. We strongly believe that this is caused by the huge number of records included in that dataset. In Section 2.3, page 7, we see that SOM_WEKA consists of 213,058 records, way above the other clustered datasets. Although a fully-supervised clustering algorithm, it turned out to be inappropriate for our goals. Furthermore, we notice that from all regression algorithms we used, Random Forests, MultiLayer Perceptron and FFNN were the most efficient, regarding our efforts to predict pollutants efficiently. What is more, we see that pollutants CO, NO and O₃ have the best results, in training as well as testing.

We surmise that this is an immediate result of their importance in the formation of the 2 extreme categories, as explained in page 4.

5. Discussion – conclusions

The EHF innovating forecasting system which allows the prediction of extreme air pollutant values was introduced and tested with real data records in this paper. Its main advantage is that though it takes no pollutants as inputs it manages to operate quite efficiently.

Moreover, it uses a small number of inputs (7), which comprises of 4 temporal inputs, air Temperature, a station identification code and a cluster identification code (was determined automatically by geolocation based services).

In order to produce the EHF model, we have used four unsupervised learning algorithms namely: SOM, Neural Gas ANN, Fuzzy C-means and a fully unsupervised SOM algorithm. For every algorithm, we have searched for the most extreme cluster, which contained the most hazardous pollutant values EXPV. Thereafter, we gathered all the records from the extreme clusters, in order to create four datasets, one for each algorithm. These four datasets were used as inputs to the EHF model, which has given promising results in forecasting pollutants’ concentrations.

More specifically in the first stage, clustering has been performed by using Fuzzy C-means, NGANN, UNSOM and SEMSOM for the determination of the extreme values. Additionally, the proposed and developed hybrid system is carrying out forecasting by employing FFNNs.

EHF uses formula which called Forecast Frame-work, a small number of inputs, which has given promising results in forecasting future air pollutants. The most interesting things of the proposed system, is the ability to forecast air pollutants with parameters that not required measurements from sensors or other devices and its application in low cost devices like smart-phones.

Future research will include the implementation of the EHF algorithm in a wider area that will cover all measurement stations of Athens. Finally, we would like to apply this methodology to different cities that have different climate characteristics.

References


My Publications

Cyber Security informatics


Environmental informatics


47. Κωνσταντίνος Δεμερτζής. Ενίσχυση της Διοικητικής Ικανότητας των Δήμων Μέσω της Ηλεκτρονικής Διακυβέρνησης: Η Στρατηγική των «Έξυπνων Πόλεων» με Σκοπό την Αειφόρο Ανάπτυξη. Θέματα Δασολογίας και Διαχείρισης Περιβάλλοντος και Φυσικών Πόρων, 10ος Τόμος: Περιβαλλοντική Πολιτική: Καλές Πρακτικές, Προβλήματα και Προοπτικές, σελ. 84 - 100, ISSN: 1791-7824, ISBN: 978-960-9698-14-6, Νοέμβριος 2018, Εκδοτικός Οίκος: Δημοκρίτειο Πανεπιστήμιο Θράκης.