A Hybrid Intelligent System for the Analysis of Atmospheric Pollution: a Case Study in two European Regions

Ángel Arroyo[1], Álvaro Herrero[1], Emilio Corchado[2], Verónica Tricio[3],

Department of Civil Engineering, University of Burgos, Burgos, Spain.
{aarroyop, ahcosio}@ubu.es
[2]Departamento de Informática y Automática, University of Salamanca, Salamanca, Spain.
escorchado@usal.es
[3]Department of Physics, University of Burgos, Burgos, Spain.
vtricio@ubu.es

Abstract. The combined application of several soft-computing and statistical techniques is proposed for the characterization of atmospheric conditions in two European regions: Madrid (Spain) and Prague (Czech Republic). The resulting Hybrid Artificial Intelligence System (HAIS) combines projection models for dimensionality reduction and clustering, combining neural and fuzzy paradigms, in a decision support tool. In present paper, this proposed HAIS is applied in order to analyze the air quality in these two geographical regions and get a better understanding of its circumstances and evolution. To do so, real-life data from six data-acquisition stations are analyzed. The main pollutants recorded at these stations between 2007 and 2014, their geographical locations and seasonal changes are all studied, in a research that shows how such factors determine variations in air-borne pollutants. Furthermore, neural projections of the clustering results from data on atmospheric pollution are studied.

Keywords. Hybrid systems, clustering techniques, air quality, projection models, artificial neural networks.

1 Introduction

In recent years, our knowledge of atmospheric pollution and our understanding of its effects have advanced greatly. Air pollution has, for some years now, represented more than a health risk. Other serious consequences may be mentioned such as acid rain, corrosion, climate change, extreme weather events and global warming. So, all efforts directed towards the study of these matters [33], will improve our understanding and help us to prevent the serious problematic hazards associated with atmospheric pollution. Systematic measurements, usually recorded within large cities in Spain, are fundamental, due to the health risks caused by high levels of atmospheric pollution. Recent high-impact weather events justify the continued enlargement of the network of weather stations that are continuously recording data. Thanks to the open data policy that the public institutions [2] have adopted, these data are available for further study and analysis.

Dimensionality reduction techniques [15] transform high-dimensional data into meaningful representations of reduced dimensionality. These techniques have previously been widely applied to the field of Environmental Conditions (EC) [10, 11, 18]. A wide range of dimensionality reduction techniques, such as Principal Component Analysis (PCA) [6], Local Linear Embedding (LLE) [25], Isometric Mapping (ISOMAP) [34] and Cooperative Maximum Likelihood Hebbian Learning (CMLHL) [13] have previously achieved very good results in the EC field [7, 8]. In [7], statistical and neural models are presented for analyzing data on the emissions of atmospheric pollution in urban areas. The main target was to classify the levels of atmospheric pollutants according to the day of the week, differentiating between working days and non-working days. In [8], atmospheric pollution conditions at two different places in the Czech Republic were analyzed. Seven variables with atmospheric pollution information were considered and dimensionality reduction techniques (PCA, LLE, and CMLHL) were applied, to show the variability of atmospheric pollution conditions between each place, as well as the significant variability of air quality over time. One characteristic of the proposed techniques is that clusters may be identified in the graphical representation with the naked eye, without any label or assignment of samples to a certain group of data. These techniques are very useful as pre-processors for the application of other machine learning paradigms, or as the final step in the visualization task.

Clustering or grouping techniques [23] divide a dataset into groups of similar objects, according to one of several "similarity" measures. In most cases, the number of desired clusters and the function determines the assignment of each sample to a certain cluster. These sets of techniques have previously been applied to EC [24, 31]. PCA and Cluster Analysis (CA) were applied in [24] for the analysis of the mass concentrations of Sulfur Dioxide (SO_2) and Particulate Matter (PM10) in Oporto over 3 years. In [31], a method for the clustering of multidimensional non-stationary meteorological time series is presented and the results are compared with standard fuzzy clustering techniques for a dataset with temperatures in Europe ranging over forty years. Most of the previous studies apply clustering techniques and only a few of them have considered dimensionality reduction techniques. So, to the best of the authors' knowledge, this is the first time that those techniques have been applied in unison and their results compared.

In this study, differentiating from previous works, dimensionality reduction techniques and clustering techniques are combined to categorize the air quality in two capital cities in Europe.

Unlike previous work, [9] a wider time window is used for data analysis in this study, running between 2007 and 2014. Additionally, two case studies are analyzed and two paradigms of soft-computing techniques are combined in the proposed Hybrid Artificial Intelligent System (HAIS), described in Section 3. The main idea of this work is to analyze data describing air pollution from two case studies in the regions of Madrid (Spain) and Prague (Czech Republic). Two objectives are set for this analysis: firstly, carry out an air-quality comparison between these two European capitals and then, validate the proposed HAIS for the analysis of multidimensional data on air pollution. This HAIS analyzes high dimensional datasets in a graphical and numerical way, finding relationships and special situations, supporting human decision making.

The rest of this study is organized as follows. Section 2 presents the combined and applied techniques. Section 3 presents the proposed HAIS. Section 4 details the real-life case studies addressed in the present work, while Section 5 describes the experiments and results. Finally, Section 6 sets out the conclusions, discussion and future work.

2 Transformation Techniques and Clustering Methods

2.1 Transformation Techniques

The problem of dimensionality reduction can be expressed as follows: for each sample *i* determine a selection or transformation of attributes so that:

$$x_{ij} \longrightarrow y_{ik}, j = 1,...,n; k = 1,...,l, l < n \quad (1)$$

Where: x_{ij} represents each data in the input space, y_{ik} represents each data in the output space, and n and l are the number of dimensions in the input and output spaces, respectively.

From among the wide range of neural projection techniques, three different ones are applied here: Principal Component Analysis, as a standard projection technique, Locally Linear Embedding and Cooperative Maximum Likelihood Hebbian Learning, as advanced techniques, also applied for comparative purposes.

Principal Component Analysis. Principal Component Analysis (PCA) [30] is a well-known method that gives the best linear data compression in terms of least mean square error by addressing the data variance. Although proposed as a statistical method, its implementation by several Artificial Neural Networks has been proven [28, 29]. The basic PCA network is described by "Eq. (2)" and "Eq. (3)": an n-dimensional input vector at time t, $x(t)$, and an l-dimensional output vector, y, with W_{ij} being the weight linking input j to output i, and η being the learning rate. Its activation and learning may be described as follows:

Feedforward step:

$$y_i = \sum_{j=1}^{N} W_{ij} x_j, \forall i \quad (2)$$

Feedback step:

$$e_j = x_j - \sum_{i=1}^{M} W_{ij} y_i \quad (3)$$

Locally Linear Embedding. Locally Linear Embedding (LLE) [32] is an unsupervised learning algorithm that computes low-dimensional, neighborhood-preserving embedding of high-dimensional inputs [35]. LLE attempts to discover nonlinear structures in high dimensional data by exploiting the local symmetries of linear reconstructions. Notably, LLE maps its inputs into a single global coordinate system of lower dimensionality, and its optimizations - though capable of generating highly nonlinear embedding - do not involve local minima.

Suppose the data consist of N real-valued vectors x, each of dimensionality n, sampled from some smooth underlying manifold. Provided there is sufficient data (i.e. the manifold is well-sampled), it is expected that each data point and its respective neighbors will lie on or close to a locally linear patch of the manifold. The method can be defined as follows:

1. Compute the neighbors of each vector, x.
2. Compute the weights W_{ij} that best reconstruct each vector x from its neighbors (the number of neighbors is the only free parameter in the implementation) minimizing the cost by constrained linear fits:

$$\varepsilon(W) = \sum_i \left| x_i - \sum_j W_{ij} x_j \right|^2 \quad (4)$$

3. Finally, find point y_i in a lower dimensional space to minimize:

$$\Phi(Y) = \sum_i \left| y_i - \sum_j W_{ij} y_j \right|^2 \quad (5)$$

This cost function in (5), as the previous one in (4), is based on locally linear reconstruction errors, but here the weights W_{ij} are fixed while optimizing the coordinate y_i. The embedding cost in (5) defines a quadratic form in the vectors y. Data points are reconstructed from their K nearest neighbors, as measured by Euclidean distance or normalized dot products. For such implementations of LLE, the algorithm has only one free parameter: the number of neighbors, K.

Subject to constraints that make the problem well-posed, it can be minimized by solving a sparse n_x_n eigenvector problem, whose bottom d non-zero eigenvectors provide an ordered set of orthogonal coordinates centered on the origin.

Low-dimensional embedding in the dimensional embedding space is computed to best preserve the local geometry represented by the reconstruction weights.

Cooperative Maximum Likelihood Hebbian Learning. Cooperative Maximum Likelihood Hebbian Learning (CMLHL) [12] is an extended version of Maximum Likelihood Hebbian Learning (MLHL) [14] that incorporates lateral connections, derived from the Rectified Gaussian Distribution. The resultant net can find the independent factors of a data set, but does so in a way that captures some type of global ordering in the data set.

Consider an n-dimensional input vector x, an l-dimensional output vector y and a weight matrix W, where the element W_{ij} represents the relationship between input x_j and output y_i, then as shown in [12], CMLHL can be performed as a four-step procedure:

Feed-forward step, where outputs are calculated according to:

$$y_i = \sum_{j=1}^{n} W_{ij} x_j, \forall i \quad (6)$$

Lateral activation passing step, where lateral connections of output neurons are given by:

$$y_i(t+1) = \left[y_i(t) + \tau\left(b - Ay\right) \right]^+ \quad (7)$$

Feedback step:

$$e_j = x_j - \sum_{i=1}^{l} W_{ij} y_i, \forall j \quad (8)$$

The function to be approximated is clearly sufficiently smooth, the learning rate is acceptable and the rule is as follows in the weight update step, where the following learning rule is applied:

$$\Delta W_{ij} = \eta . y_i . sign\left(e_j\right) | e_j |^{p-1} \quad (9)$$

Where η is the learning rate, τ is the "strength" of the lateral connections, b the bias parameter, p a parameter related to the energy function which relates the *pdf* to the learning rule, e_j is the activation of the neuron and A is a symmetric matrix used to modify the response to the data. The effect of this matrix is based on the relation between the distances separating the output neurons.

2.2 Clustering Methods

Cluster analysis organizes data by abstracting underlying structures either as a grouping of individuals or as a hierarchy of groups. The representation can then be investigated, to observe patterns in the data group that will confirm preconceived ideas or suggest new experiments [23] [12]. Intuitively, two elements in a valid cluster will share greater similarity than those in different groups. Some clustering methods, from the wide range of clustering techniques, have been applied in the present work and are described below.

***K*-means.** The standard *k*-means [16] is an algorithm for grouping data points into a given number of clusters. Its application requires two input parameters: the number of clusters, *k*, and their initial centroids, which can be chosen by the user or obtained through some pre-processing. Each data element is assigned to the nearest group centroid, thereby obtaining the initial composition of the groups. Once these groups are obtained, the centroids are recalculated and a further reallocation takes place. The process is repeated until there are no further changes in the centroids. Given the heavy reliance of this method on initial parameters, a reliable measure of the goodness of the grouping is simply the sum of the proximity Error Sums of Squares (SSE) that it attempts to minimize:

$$SSE = \sum_{j=1}^{k} \sum_{x \in G_j} \frac{p(x_i, c_j)}{n} \tag{10}$$

where, p is the proximity function, k is the number of groups, c_j are the centroids, and n the number of rows. In the case of working with Euclidean distance, the expression is equivalent to the global mean square error.

Fuzzy c-means (fcm). Conventional clustering approaches (such as *k*-means) assume that an object can belong to only one cluster. In practice, the separation of clusters is a fuzzy notion. Fuzzy clustering algorithms are an extension of traditional clustering techniques applying Fuzzy-Sets theory, which outperform conventional clustering algorithms by allowing each object to be assigned to one or more clusters. According to Fuzzy Logic, the membership function is not a crisp value (0, 1) but it is now defined as a percentage.

The *c*–means clustering algorithm [17] is a clustering method that lets one piece of data belong to two or more clusters. Let $X = [x_1, x_2, \ldots, x_m]$ be a set of numerical data in $\mathbf{R}^N$. Let c be an integer, $1 < c < M$. Given X, it can be said that c-fuzzy subsets $\{u_k: X \rightarrow [0, 1]\}$ are partitions of X, if the following conditions are satisfied:

$$0 \leq u_{kj} \leq 1 \forall\, k, j \tag{11}$$

$$\sum_{k=1}^{c} u_{kj} = 1 \forall\, j \tag{12}$$

$$0 < \sum_{j=1}^{M} u_{kj} < n \forall\, k \tag{13}$$

Where $u_{kj} = u_k(X_j)$, $1 \leq k \leq c$ and $1 \leq j \leq M$.

3 Proposed Hybrid Artificial Intelligent System

A hybrid system combining dimensionality reduction techniques and clustering techniques is proposed for the analysis of data sets with atmospheric pollution information.

A HAIS combines different paradigms in Artificial Intelligence for solutions to real world problems. In this case, different ways of combining and testing clustering dimensionality reduction techniques and clustering techniques are investigated, for the analysis of air pollution records in the two case studies described in Section 4. The hybrid system is described in the following subsections. In Figure 1, the three steps for the proposed hybrid model are shown alongside a graphic example of each step to its right.

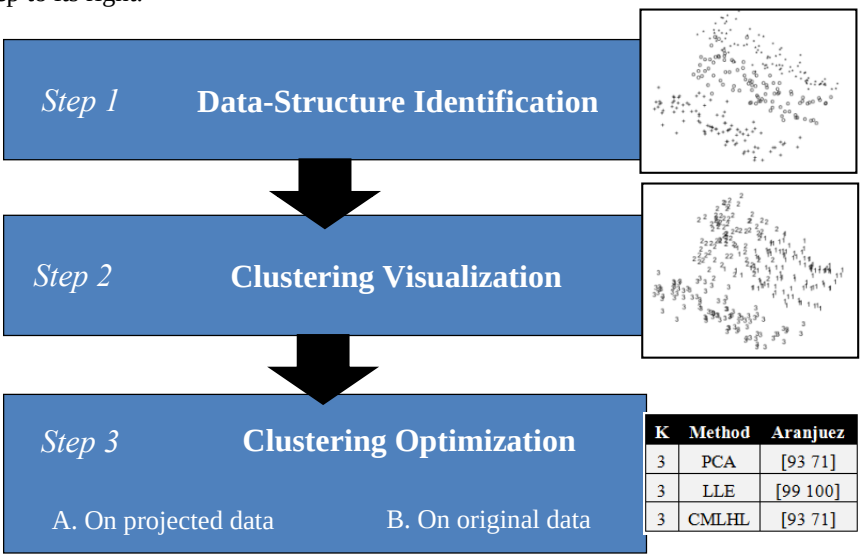


Figure 1: Proposed HAIS.

Step 1: Data Structure Identification. Several dimensionality reduction techniques are applied to the original dataset, in order to discover the internal structure of the data under analysis. In addition, experimentation with different graphical representations sought to capture the main directions and data groups in this structure.

Step 2: Clustering Visualization. In the second step, clustering techniques are applied to the original data set. Different values of the parameter k (number of target clusters) are tested, according to the number of previously identified groups from the projections obtained by the dimensionality reduction methods applied in Step 1.

After that, the dimensionality reduction techniques are once again applied, taking the sample assignment results of the clustering method as the labels in the projection. By doing so, deep knowledge is gained about the criterion that determines how the original data will be grouped, being possible to observe the coupling or similarity between the results obtained by the clustering techniques and the graphical view offered by the dimensionality reduction techniques.

The main idea behind the present study is to characterize the atmospheric pollution of different locations by applying clustering techniques, so the final step is mainly focused on the optimization of such clustering. To do so, various techniques are applied to different data as described below (Steps 3.A and 3.B).

Step 3.A: Clustering on projected data. In this step, the clustering techniques presented in Section 2.2 are applied to a low-dimensionality data set. To do so, PCA, LLE and CMHL reduce the dimensionality of the original data set to a number of dimensions equal to two and three. After that, k-means and fcm are applied to the low-dimensionality data set, with values of k similar to those that achieved the best results in Step 2. The main idea of this step is to check the performance of clustering techniques working on (previously projected) datasets of reduced dimensionality.

Step 3.B: Clustering on original data. Clustering techniques (described in Section 2.2) are applied to the original data set. The range of values for the k parameter in Step 3.B is similar to those previously selected in Step 3.A.

By comparing the results from Step 3.A and Step 3.B, interesting conclusions can be obtained about the differences in the sample allocation process of clustering techniques, depending on the input data (original datasets versus low-dimensionality datasets).

4 Data for the Case Studies

In this Section, the two real case studies analyzed in this study are presented. Three reasons motivate the selection of the stations, in both cases of study: the selected pollutants are the most interesting ones, in order to study the environmental conditions. The second reason is that each station is in an area quite unlike the others in relation to traffic density, population and geographical location. The third reason is that both regions are important capital cities in Europe, with relevant air-pollution monitoring networks and with air-pollution problems, both in the past and in the present. These air-pollution monitoring networks are in many ways similar to the classification of the stations (urban, suburban, high-traffic density, vegetation protection…), the recorded parameters and the number of available data acquisition points. Figure 2 shows the six selected locations (three from the region of Madrid and three from the region of Prague)

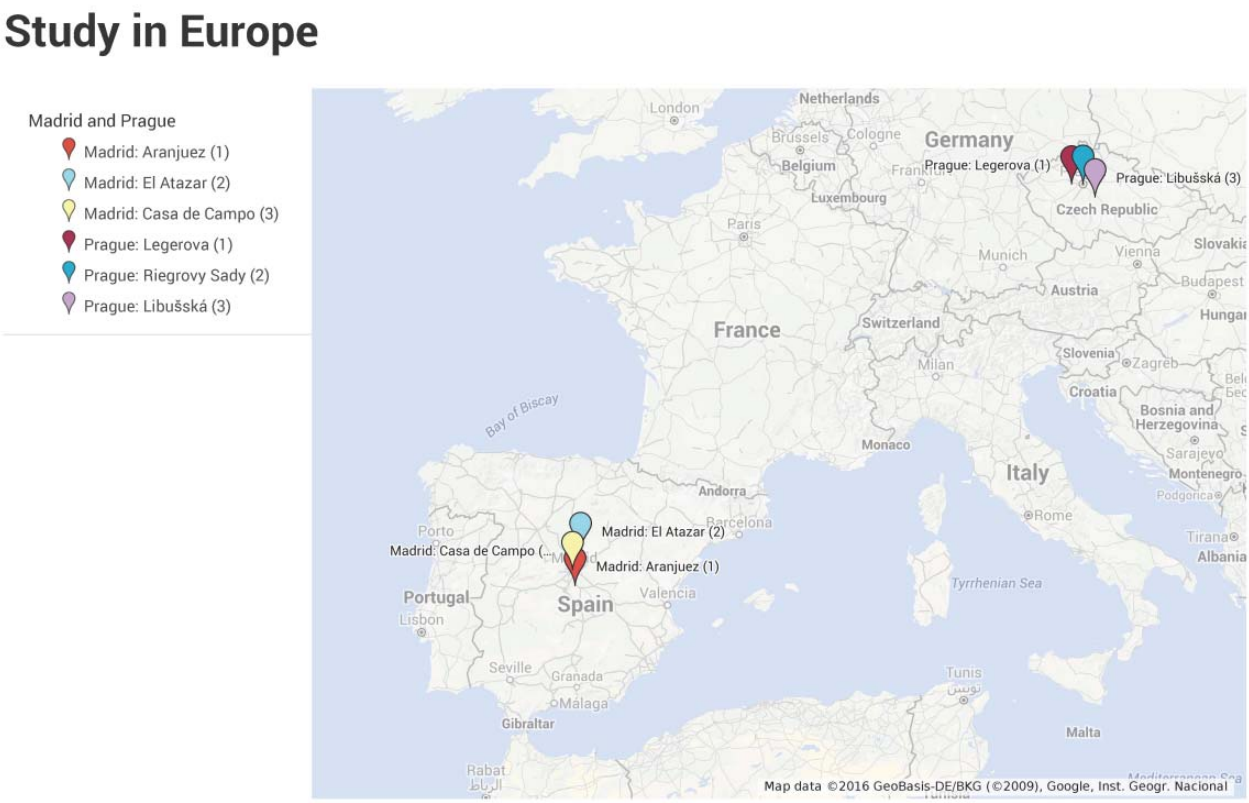


Figure2: Location of the two case studies in Europe, by Google Maps.

4.1 Case Study I: Region of Madrid (Spain)

The first study is focused on the analysis of air pollution data recorded in the region of Madrid (central Spain) [1] [26]. According to Köppen's climate classification criteria [5], the climate of Madrid is classified as a Mediterranean climate type (mild with dry and hot summer), in a highly urban environment. The Gross Domestic Product (GDP) of Madrid city in 2011 was estimated at 124,780 million Euros, which represents 65.9% of the total GDP of the region of Madrid and 11.6% of Spanish GDP, of which its industrial contribution was 7.9% and construction accounted for 6.1%. However, it is the service sector, which accounts for 85.9% of total economic activity, which defines the city of Madrid's production structure [3]. Sources of atmospheric pollutants are not the same everywhere and pollution varies greatly in the atmosphere, depending on the characteristics of each area. More precisely, the main sources of pollution in the city are: road transport, combustion plants (industrial and non-industrial), other means of transport and mobile machinery, and the use of non-industrial solvents and products [26].

The data were selected from between 2007 to 2014, at the start of the economic crisis in Spain; a fact that implies a significant variation in the levels of air pollution compared with other periods of time. Over these years, economic activity in strategic fields, such as building, was significantly reduced. The data were gathered from stations located in the following areas:

1. Casa de Campo in Madrid city center. Longitude: 3º 44' 50" W, latitude: 40º 25' 68" N, meters above sea level (masl): 645. Full station specification can be found in [1].
2. Aranjuez: an urban background station. Longitude: 3º 32' 46" W, latitude: 40º 2' 2" N, masl: 501. The full station specification can be found in [26].
3. El Atazar: a vegetation protection station. Longitude: 3° 28' 13" W, latitude: 40° 56' 4" N, masl: 995. Full station specification can be found in [26].

There are a total of 287 samples, as one sample per month (monthly daily average) was collected for the twelve months of every year, between 2007 and 2014, and the 3 stations analyzed in this study. Some corrupted values were omitted because any value was missing (leaving 287 samples rather than 288). The following parameters (four air pollutants and four meteorological variables) were analyzed:

1. Nitric Oxide (NO) - µg/m^3, primary pollutant. NO is a colorless gas which reacts with ozone undergoing rapid oxidation to NO_2, predominant in the atmosphere.
2. Nitrogen Dioxide (NO_2) - µg/m³, a primary pollutant. From the standpoint of health protection, nitrogen dioxide has set exposure limits for long and short duration.
3. Particulate Matter (PM10) - µg/m³, primary pollutant. These particles remain stable in the air for long periods of time without falling to the ground and can be moved significant distances by the wind. Defined by the ISO as follows: *"particles which pass through a size-selective inlet with a 50 % efficiency cut-off at 10 µm aerodynamic diameter. PM10 corresponds to the 'thoracic convention' as defined in ISO 7708:1995, Clause 6"* [4].
4. Ozone (O_3) - µg/m³, secondary pollutant. Ozone is an odorless, colorless gas composed of three oxygen atoms. It occurs both in the Earth's upper atmosphere and at ground level. It can be "good" or "bad" for people's health and for the environment, depending on its location in the atmosphere.
5. Relative Humidity (RH) - %.
6. Atmospheric Pressure - mbar.
7. Wind Speed Module - m/s.
8. Air Temperature - °C.

4.2 Case Study II: Region of Prague (Czech Republic)

The second study is focused on the analysis of air pollution data recorded in the region of Prague (Czech Republic) [19]. The city of Prague has a Continental climate with marked temperature differences between winter and summer and between day and night.

Approximately, a quarter of the Czech Republic's GDP (24.6% in 2012) is generated in Prague. GDP per capita in Prague reached 20% of the Czech Republic's average. At present, values in Prague are well above the average values for the entire EU-28 (GDP per capita in Prague was 68.9% higher). The city is one of the most affected European regions from the perspective of low air quality. Its air quality is mainly influenced by traffic flows, electricity production and heat generation. The heating plant in Malešice and the cement factory in Radotín are the main sources of emissions. The limit values fixed by the European Union in all cases, exceeded in air-quality measurements by inmission: especially sulfur dioxide concentrations, suspended PM10 particulate matter and benzopyrene [27]. As in the case study presented in Section 4.1, data were selected from between 2007 and 2014. The data were gathered from stations located in the following areas:

1. Aleg: a traffic/urban/residential/commercial station. Longitude: 50° 4' 20" W, latitude: 14° 25' 50" N, masl: 219. Full station specification in [20].
2. Arie: a background/urban/residential station. Longitude: 50° 4' 53" W, latitude: 14° 26' 33" N, masl: 256. Full station specification in [22].
3. Alib: a background/suburban/residential station. Longitude: 50° 0' 26" W, latitude: 14° 26' 45" N, masl: 301. Full station specification in [21].

There were a total of 281 samples as one sample per month (monthly daily average) was collected for the twelve months of every year between 2007 and 2014 and the 3 stations analyzed in this study. Some samples were removed as they contained missing values. The following parameters (five air pollutants) were analyzed:

1. Nitric Oxide (NO) - µg/m^3, secondary pollutant
2. Nitrogen Dioxide (NO_2) - µg/m³, secondary pollutant.
3. Nitrogen Oxides (NO_x) - µg/m³, secondary pollutant. Consisting of nitric oxide (NO), nitrogen dioxide (NO_2) and nitrous oxide (N_2O), formed when nitrogen (N_2) combines with oxygen (O_2).
4. Particulate Matter (PM10) - µg/m³, primary pollutant.
5. Particulate Matter (PM2.5) - µg/m³, primary pollutant. Defined by the ISO as follows: *"particles which pass through a size-selective inlet with a 50 % efficiency cut-off at 2.5 µm aerodynamic diameter. PM2.,5 corresponds to the 'high-risk respirable convention' as defined in ISO 7708:1995, 7.1. Stands for Particulate Matter of less than 2.5 millionths of a metre in diameter"* [4].

5 Evaluation of the Case Studies using the Proposed HAIS

The Steps described in Section 3 were applied to both case studies presented in Section 4 and the results are discussed below.

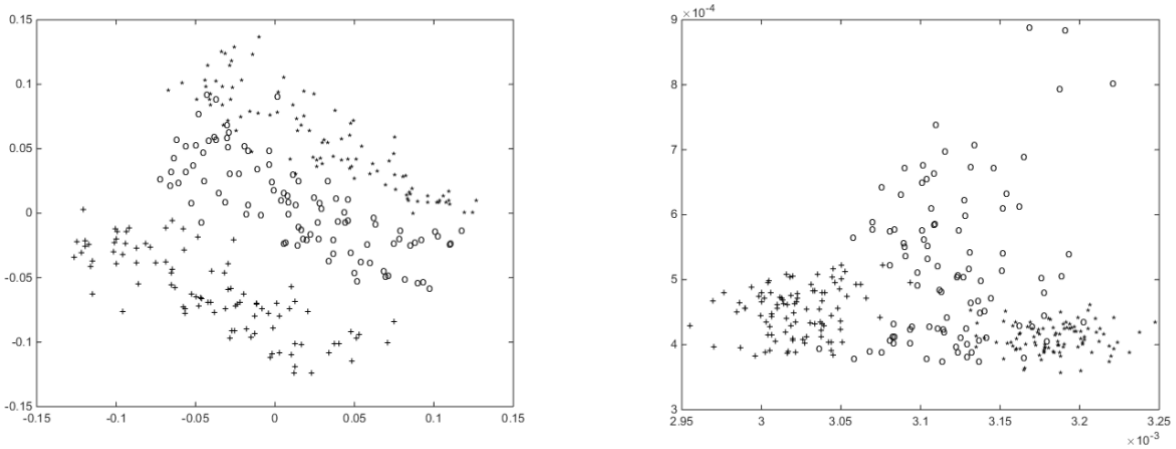
5.1 Case Study I: Region of Madrid (Spain)

Step 1. In this first step the three dimensionality reduction techniques described in Section 2.1 are applied to the original data set, seeking a possible structure in the data. In this step, the best possible data projection is aimed to reach and the different dimensionality projection technique are tried in order to discover the number of groups in the original dataset. For the sake of brevity, only the best results are shown. Table 1 shows the values of the parameters for LLE and CMLHL techniques, whose results are shown in the following figures (3 to 7).

Table 1: Parameter values for the models associated to the projections shown in Figs. 3 to 7.

Technique	Output Dimensions	Neighbors	Iterations	Learning Rate	*p*	τ
LLE	2	35				
CMLHL	2		1000	0.0008	2.5	2.5

In Figure 3, data samples are depicted according to their geographical location (data acquisition station).

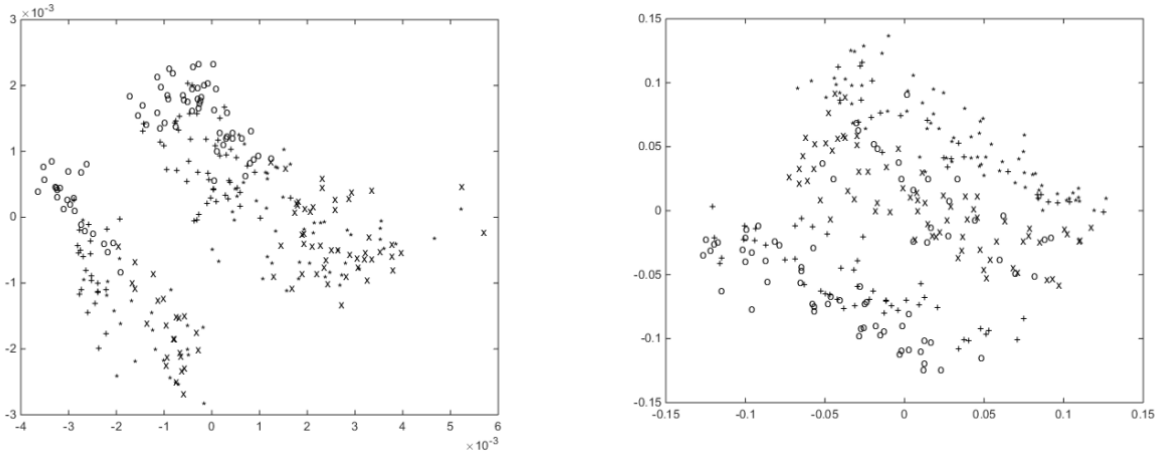


Legend: '*' Aranjuez, '+' El Atazar, 'o' Casa de Campo.

Figure 3: Projections of case study I, by station (Step 1): Left: LLE, Right: CMLHL.

In Figure 3, data samples are depicted according to their geographical location (data acquisition station). LLE depicts the structure of the data, where three main groups can be identified. Samples corresponding to the station from ‘El Atazar’, located at the bottom of the picture, are at a considerable distance from the samples corresponding to the other two stations, because the lowest levels of contamination are found at this location. The samples of 'Aranjuez' are at the top the projection, where pollution levels are the highest of the three locations under analysis. CMLHL groups most samples from 'Aranjuez' and 'El Atazar' into two compact clusters, while samples from 'Casa de Campo' are much more dispersed, outliers being distinguished that correspond to samples where levels of NO are specially high.

In Figure 4, data are projected according to their particular season of the year (winter, spring, summer and autumn).



Legend: ‘*’ Winter, ‘+’ Spring, ‘o’ Summer.

Figure 4: Projections of case study I, by season (Step 1): Left: PCA, Right: LLE.

In Figure 4, the data projections are by the season of the year to which they refer (winter, spring, summer and autumn). The projection is not as clear as the previous one in Figure 3, as data from the four seasons were mixed up. Analyzing the data, the evolution of atmospheric pollution related to the season may be observed; summer and spring contains the highest levels of O_3 and winter and autumn the highest registered levels of NO and NO_2. The four outliers on the right side of the PCA projection correspond to samples from winter in 2007 and 2008 where the level of NO reached values around 100 $\mu g/m^3$. This value is above the daily limit value for the protection of human health fixed at 50 $\mu g/m^3$ in 2005 by the European Union [36].

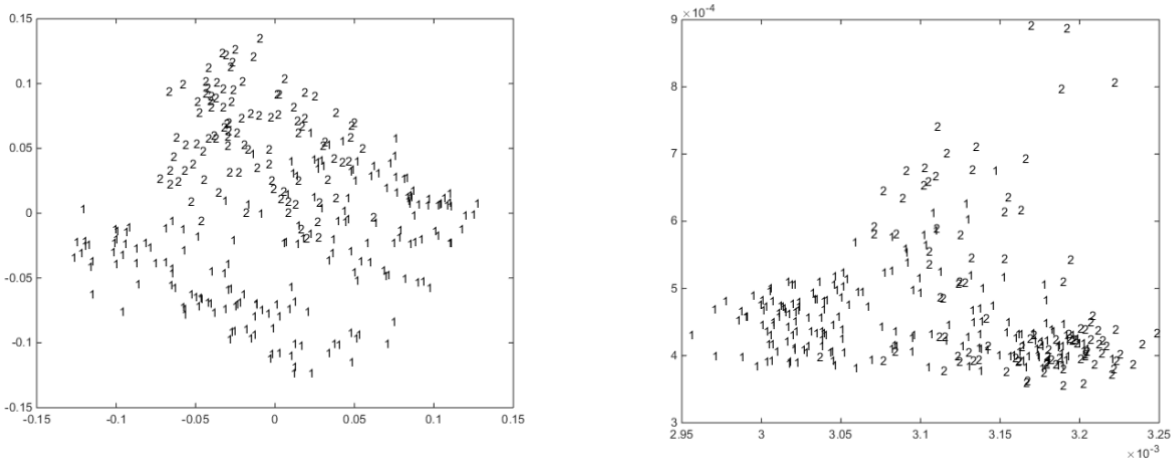
From the LLE projection, the existence of an important group of samples associated with the winter season may be observed at the top of the plot. These samples present the highest levels of PM10 and PM2.5 for the whole year. The absence of groups in the projection according to the seasons of the year is due in part to the use of the monthly averaged data set.

Step 2. In the first step, different dimensionality reduction techniques were applied and evaluated for the original dataset.

After this, the clustering technique *k*-means (described in Section 2.2) applied different values for the *k* parameter according to the numbers of clusters found in Step 1 (that is, 2 and 3). Then, the dimensionality reduction techniques were applied to visualize these results (sample process allocation obtained by *k*-means). The results for *k* equal to 2, 3, and 4 are shown in Figs. 5, 6, and 7, respectively.

In this step, the results for PCA are omitted, because PCA obtained no clear results for this case study in the first step.

In Figure 5, data are depicted according to the sample process allocation obtained by *k*-means, when *k* equals 2.

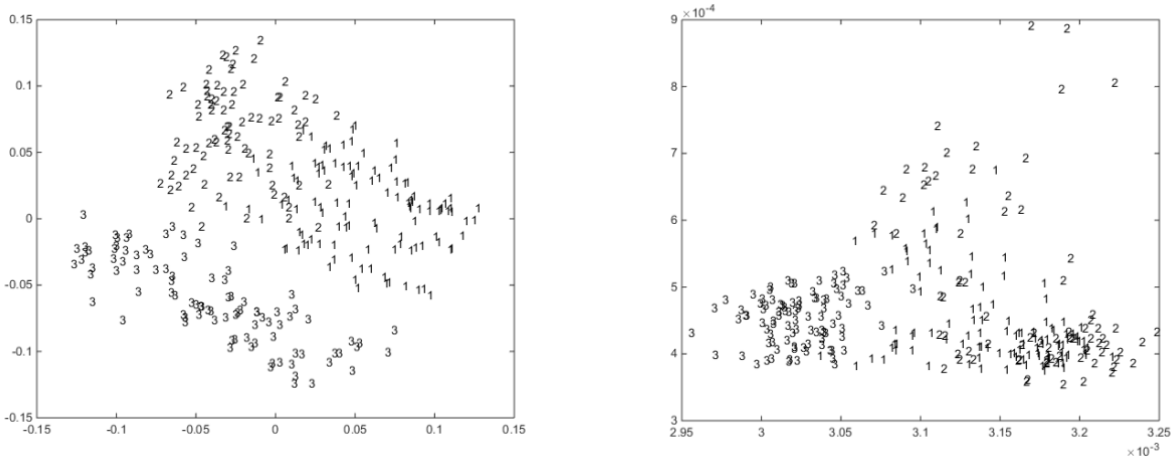


Data are labeled according to the cluster allocation (1 or 2).

Figure 5: Projections of case study I (Step 2): Left: LLE, Right: CMLHL. *k*=2.

In the LLE projection (Fig. 5), all the samples from 'El Atazar' are grouped in the same cluster, labeled as 1, while the samples from the other two stations are distributed in both clusters (1 and 2). The samples labeled as 2 are those from ‘Casa de Campo’ and ‘Aranjuez’ and in most cases from winter and autumn, as depicted in Figure 4. It can be seen that CMLHL obtained similar results; samples from 'El Atazar' are grouped in the cluster labeled as 1, while samples from the other two stations are located in clusters 2 and 3.

In Figure 6, the data are depicted according to the sample process allocation obtained by *k*-means, when *k* equals 3.

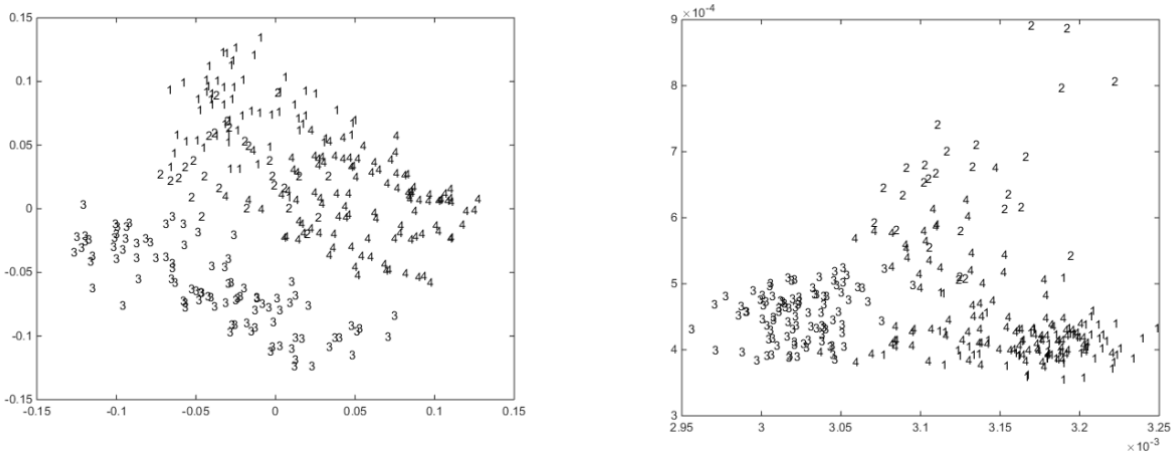


Data are labeled according to the cluster allocation (1, 2 or 3).

Figure 6: Projections of case Study I (Step 2): Left: LLE. Right: CMLHL. *k*=3.

In Figure 6, in the same way as in Figure 5, the 'El Atazar' samples are all grouped into the same cluster, labeled as 3, while samples from the other two stations are distributed in clusters labeled as 1 and 2. These samples are grouped according to the season of the year in most cases, while the samples from autumn and winter are labeled as 2.

In Figure 7, the data are depicted according to the sample process allocation obtained by *k*-means, when *k* equals 4.



Data are labeled according to the cluster allocation (1, 2, 3 or 4).

Figure 7: Projections of case study I (Step 2): Left: LLE. Right: CMLHL. *k*=4.

Showing similar results than in previous cases (Figs. 5 and 6), in Figure 7 the samples from 'El Atazar' are grouped into the same cluster, labeled as 3. Samples from the other two stations are distributed mainly in clusters labeled as 1 (samples from the seasons of autumn and winter) and 4 (samples from the seasons of summer and winter). Only a few samples are grouped in the cluster labeled as 2. The results are similar, whether the results are projected by LLE or by CMLHL.

Step 3. A. In this third step, the sample process allocation is the result of applying the clustering techniques to the low dimensionality data set. In Table 2, 'Dimensions' represents the number of dimensions of the dataset after applying the dimensionality reduction technique. 'Cluster Samples Allocation' represents the percentage of samples from each one of the stations (Casa de Campo, Aranjuez and Atazar) that are allocated to each one the clusters; e.g. [47 53] indicates that 47% of the samples from the station are allocated to the first cluster and 53% to the second one. Each sample is assigned to the cluster whose probability of allocation is closer to one.

Table 2: *k*-means clustering results after dimensionality reduction for case study I (Step 3.A).

k	**Dimensions**	**Method**	**Cluster Samples Allocation (%)**		
			Casa de Campo	**Aranjuez**	**Atazar**
2	3	PCA	[47 53]	[49 51]	[100 0]
3	3	PCA	[42 58 0]	[41 59 0]	[0 0 100]
4	3	PCA	[56 28 16 0]	[53 0 47 0]	[0 0 0 100]
2	4	PCA	[47 53]	[49 51]	[100 0]
3	4	PCA	[44 0 56]	[41 0 59]	[0 100 0]
4	4	PCA	[56 0 16 28]	[53 0 47 0]	[0 100 0 0]
2	3	LLE	[42 58]	[57 43]	[66 34]
3	3	LLE	[20 2 78]	[98 0 2]	[0 100 0]
4	3	LLE	[36 52 8 4]	[38 1 61 0]	[5 0 0 95]
2	4	LLE	[100 0]	[100 0]	[0 100]
3	4	LLE	[66 4 30]	[10 44 46]	[28 54 18]
4	4	LLE	[38 56 6 0]	[51 0 49 0]	[0 0 0 100]
2	3	CMLHL	[71 29]	[63 37]	[2 98]
3	3	CMLHL	[45 46 9]	[1 97 2]	[0 16 84]
4	3	CMLHL	[2 42 32 24]	[0 53 47 0]	[65 2 33 0]
2	4	CMLHL	[46 54]	[61 39]	[20 80]
3	4	CMLHL	[30 35 35]	[22 40 38]	[60 40 0]
4	4	CMLHL	[24 21 29 26]	[33 6 32 29]	[29 42 29 0]

According to the results obtained in Step 2 (Figs. 5, 6 and 7), Table 2 shows how samples from 'El Atazar' are assigned to the same data cluster with a percentage of 75%, when applying PCA and LLE as a pre-step before applying *k*-means. Regarding the distribution of the samples from the other two stations, the results are very similar to those obtained in Step 2. The samples are distributed by the season of the year. When applying CMLH before clustering, the number of samples assigned to the same cluster in the case of 'El Atazar' decreases significantly.

In Table 3, the performance of fcm applied to the low-dimensionality dataset is shown. Column 'Cluster Samples allocation' represents the percentage of samples assigned to a right cluster in a first and a second option. i.e: [93 71] 93% of the samples of 'Aranjuez' are assigned to the same cluster with the highest percentage according to the results returned by the method, 71% of samples that are not assigned to that cluster with the highest percentage area are assigned to the cluster with the second highest percentage.

Table 3: fcm clustering results after dimensionality reduction for case study I (Step 3.A).

k	**Dimensions**	**Method**	**Cluster Samples allocation (%)**		
			Casa de Campo	**Aranjuez**	**Atazar**
3	3	PCA	[67 16]	[93 71]	[100 0]
4	3	PCA	[51 13]	[61 30]	[100 0]
3	4	PCA	[54 16]	[61 30]	[100 0]
4	4	PCA	[54 7]	[63 8]	[71 100]
3	3	LLE	[75 54]	[99 100]	[99 100]
4	3	LLE	[59 49]	[55 93]	[96 100]
3	4	LLE	[71 75]	[70 100]	[95 100]
4	4	LLE	[69 83]	[55 49]	[83 88]
3	3	CMLHL	[74 76]	[93 71]	[100 0]
4	3	CMLHL	[84 7]	[97 33]	[100 0]
3	3	PCA	[67 16]	[93 71]	[100 0]
4	3	PCA	[51 13]	[61 30]	[100 0]
3	4	PCA	[54 16]	[61 30]	[100 0]
4	4	PCA	[54 7]	[63 8]	[71 100]
3	3	LLE	[75 54]	[99 100]	[99 100]
4	3	LLE	[59 49]	[55 93]	[96 100]
3	4	LLE	[71 75]	[70 100]	[95 100]
4	4	LLE	[69 83]	[55 49]	[83 88]

As happened in Table 2, samples from 'El Atazar' are mainly assigned to the same cluster in the first option. Regarding the samples belonging to the other two stations, the percentage of proper allocations in the first option is higher (k = 3) (Dimensions = 3). It is worth mentioning that the percentage of success in the allocation process for the second option is higher when the percentage of success for the first option is also high. Finally, it can be said that by applying CMLH before clustering, most of the samples are assigned to the same data cluster in the case of 'El Atazar'.

Step 3.B. Finally, *k*-means applied to the original dataset yields the results that are shown in Table 4.

Table 4: *k*-means clustering results for case study I on original data (Step 3.B).

k	**Cluster Samples Allocation (%)**		
	Casa de Campo	**Aranjuez**	**Atazar**
2	[47 53]	[49 51]	[100 0]
3	[57 43 0]	[59 41 0]	[0 0 100]
4	[16 29 0 55]	[47 0 0 53]	[0 0 100 0]

By applying *k*-means to the original data set, the results are similar to those obtained in Step 3.A (Table 2). 'El Atazar' samples are always assigned to the same cluster while samples from the other two stations are distributed (50%), with k = 2

and by a percentage close to 50% between two clusters when $k = 3$. In the case of $k = 4$, the samples are distributed between two clusters, in the case of 'Aranjuez', and between three clusters in the case of 'Casa de Campo', which is similar to the results obtained in Step 2 (Figure 7).

In Table 5, the results of applying fcm to the original dataset are shown. For each experiment, the percentage of samples assigned to the correct cluster is shown as either first or second options.

Table 5: fcm clustering results for case study I on original data (Step 3.B).

k	Cluster Samples allocation (%)		
	Casa de Campo	**Aranjuez**	**Atazar**
3	[57 37]	[42 100]	[97 0]
4	[43 10]	[93 0]	[100 0]

Applying fcm to the original dataset, the results obtained are similar to those shown in Step 3.A (Table 3). The samples belonging to 'El Atazar' are almost entirely assigned to the same cluster, while samples of the other two stations are distributed in clusters mainly by the season of the year.

5.2 Case Study II: Region of Prague (Czech Republic)

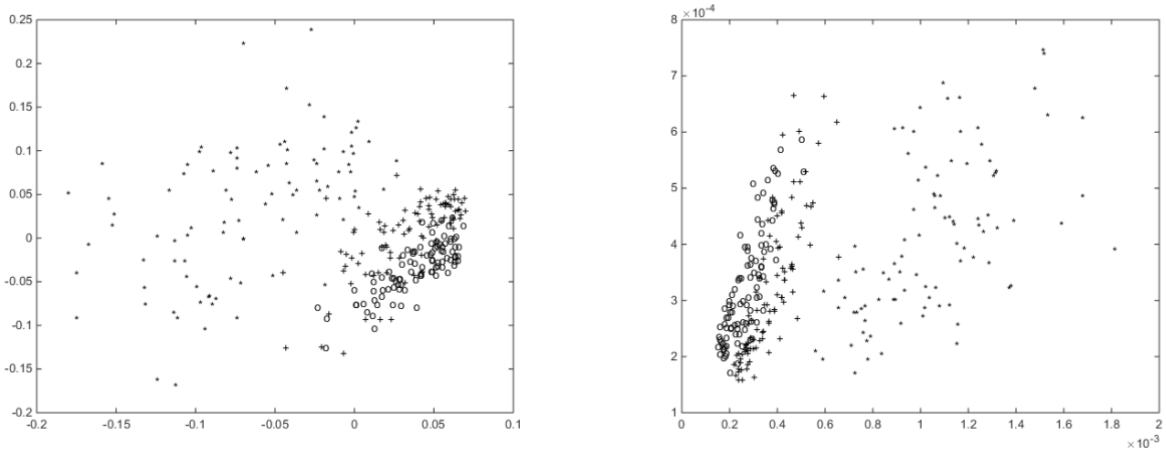
Step 1. In this first step, the three dimensionality reduction techniques described in Section 2.1 are applied to the original data set in order to reveal the structure of the data.

Table 6 shows the values of the parameters for LLE and CMLHL techniques, whose results are shown in the following figures (8 to 12).

Table 6: Parameter values for the models, associated to the projections shown in Figs. 8 to 12.

Technique	Output Dimensions	Neighbors	Iterations	Learning Rate	*p*	τ
LLE	2	33				
CMLHL	2		1000	0.002	1	1

In Figure 8, data samples are depicted according to their geographical location (data acquisition station).



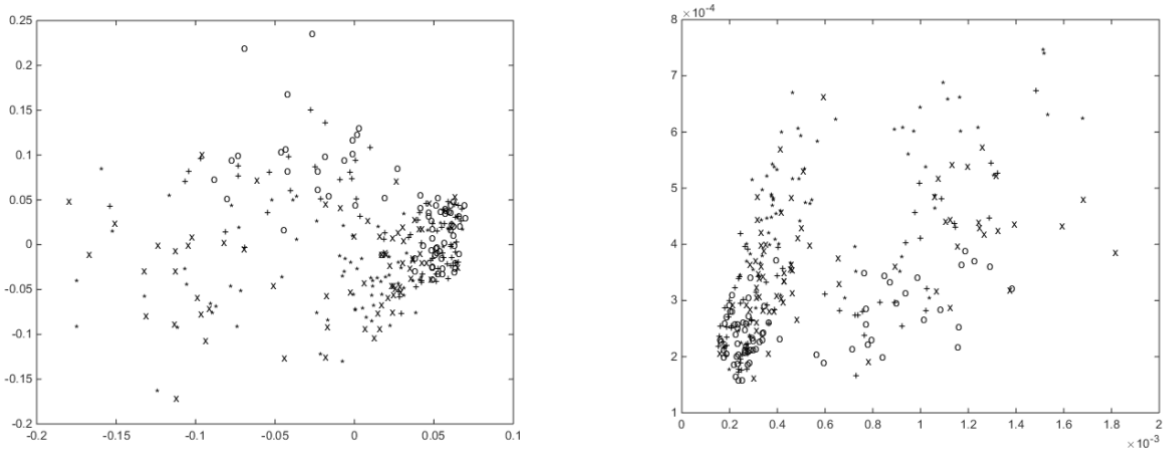
Legend: '*' Aleg, '+' Arie, 'o' Alib

Figure 8: Projections of case study II, by station (Step 1): Left: LLE, Right: CMLHL.

LLE depicts a group of data (not so clearly separated) and a big cloud of sparse samples. The group of data, located in the right-bottom corner of the projection, corresponds to the stations of 'Arie' and 'Alib', presenting similar levels of air

pollution. The rest of the points correspond to the ‘Aleg’ station, presenting high levels of air pollution, especially in NO_2 and PM10. This large point cloud reflects the great variability in NO_X, reaching peaks of 285 µg/m³, in August of 2011, and values of over 200 µg/m³, in the winter of 2009. These values exceed the limit value of 200 µg/m³ established in 2010 by the European Union [36]. By applying CMLHL, the samples from 'Arie' and 'Alie' are clustered into a group of data, but an even sparser one than if LLE had been applied. Note that the samples from the station located in ‘Aleg’ are more clearly differentiated from the samples belonging to the two data clusters than in the case of the LLE projection.

In Figure 9, the data are projected according to the season in which they were collected (winter, spring, summer and autumn).



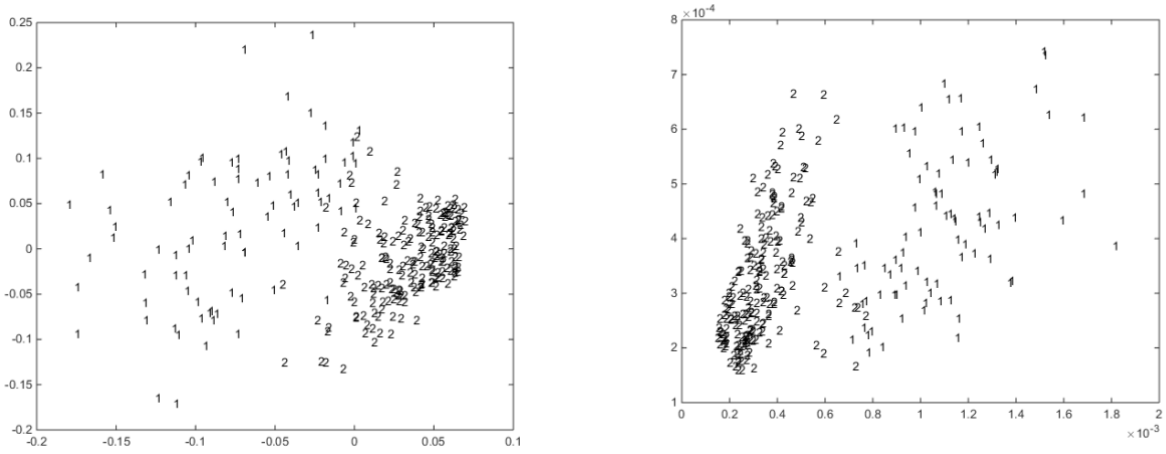
Legend: ‘*’ Winter, ‘+’ Spring, ‘o’ Summer

Figure 9: Projections of case study II, by season (Step 1): Left: LLE, Right: CMLHL.

The projection is even less clear than the previous one in Figure 8. The only thing that can be observed is an evolution of atmospheric pollution throughout the year, as winter and autumn show the highest levels of pollution in comparison with spring and summer, and are located (Fig. 9) at the bottom of the LLE projection and at the upper part in the case of CMLHL. The air pollution values are similar between the seasons of spring and summer and the seasons of winter and autumn.

Step 2. The results obtained by applying PCA are omitted, as PCA offered no clear results in Step 1 in this case study.

In Figure 10, data are depicted according to the sample process allocation obtained by *k*-means, when *k* equals 2.

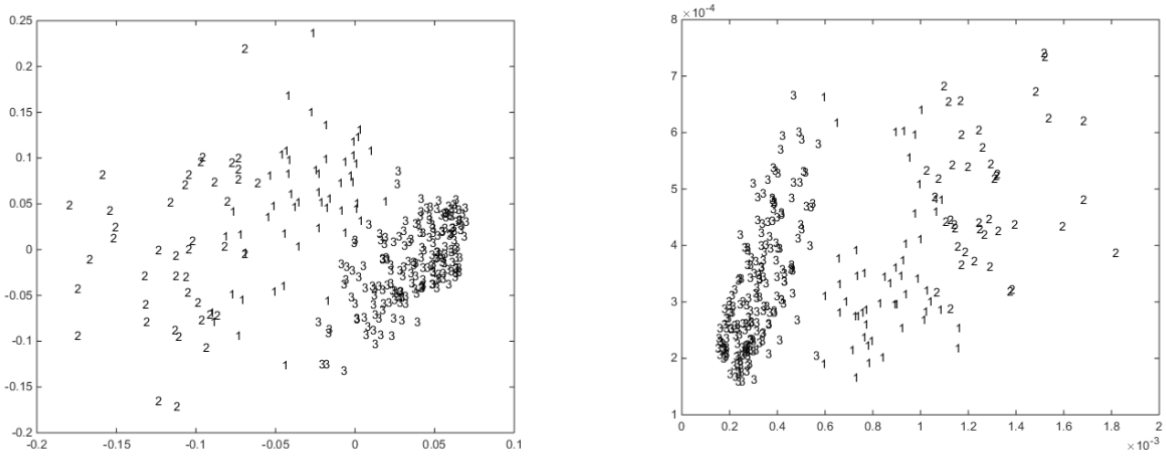


Data are labeled according to the cluster allocation (1 or 2).

Figure 10: Projections of case study II (Step 2): Left: LLE, Right: CMLHL. *k*=2.

In Figure 10, by applying LLE most of the samples from 'Aleg' form the open point cloud labeled as 1, while the samples belonging to the other two stations are distributed in the cluster labeled as 2. Similar results can be observed from the CMLHL projection.

In Figure 11, data are depicted according to the sample process allocation obtained by *k*-means, when *k* equals 3.

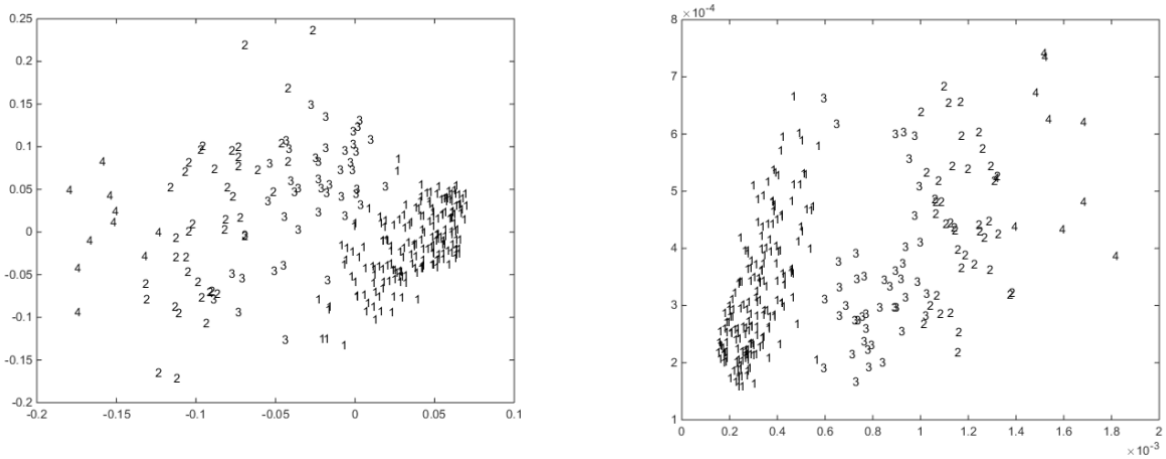


Data are labeled according to the cluster allocation (1 or 2).

Figure 11: Projections of case study II (Step 2): Left: LLE. Right: CMLHL. *k*=3.

In Figure 11, samples from 'Arie' and ‘Alib’ are all grouped in the same cluster, labeled as 3 in this case, which corresponds to the samples with lower levels of air pollution, while samples from the other two stations are distributed in clusters labeled as 1 and 2. Samples labeled as 2 correspond to those with higher levels of pollutants such as NO and NO_2. The results projected by LLE and CMLHL are similar in this case.

In Figure 12, data are depicted according to the sample process allocation obtained by *k*-means, when *k* equals 4.



Data are labeled according to the cluster allocation (1 or 2).

Figure 12: Projections of case study II (Step 2): Left: LLE. Right: CMLHL. *k*=4

In the same way as in Figure 10 and Figure 11, in Figure 12 the samples from 'Arie' and ‘Alib’ are grouped into the same cluster, labeled as 1. Samples from the ‘Alib’ station are labeled as 2, 3 and 4 depending on the levels of air pollution, the samples labeled as 4 being those with higher levels of NO_x. Results are similar whether we apply LLE and CMLHL.

Step 3.A. In this third step, the results obtained from the sample allocation process, by applying the clustering technique to the low dimensional data sets are shown.

In Table 7, the results of *k*-means applied to the low dimensionality datasets obtained by applying the three dimensionality reduction techniques are shown.

Table 7: *k*-means clustering results after dimensionality reduction for case study II (Step 3.A).

k	Dimensions	Method	Cluster Samples Allocation (%)		
			Aleg	Arie	Alib
2	3	PCA	[20 80]	[100 0]	[100 0]
3	3	PCA	[53 1 46]	[3 97 0]	[0 100 0]
4	3	PCA	[3 42 0 55]	[48 0 52 0]	[28 0 72 0]
2	4	PCA	[10 90]	[100 0]	[100 0]
3	4	PCA	[1 45 54]	[97 0 3]	[100 0 0]
4	4	PCA	[3 42 0 55]	[48 0 52 0]	[28 0 72 0]
2	3	LLE	[35 65]	[33 67]	[32 68]
3	3	LLE	[38 33 29]	[67 0 33]	[72 0 28]
4	3	LLE	[4 30 11 55]	[66 0 32 2]	[71 0 29 0]
2	4	LLE	[46 54]	[77 23]	[77 23]
3	4	LLE	[80 15 5]	[0 27 73]	[0 25 75]
4	4	LLE	[29 51 6 14]	[0 0 73 27]	[0 0 75 25]
2	3	CMLHL	[90 10]	[0 100]	[0 100]
3	3	CMLHL	[54 1 45]	[3 97 0]	[0 100 0]
4	3	CMLHL	[3 42 55 0]	[48 0 0 52]	[28 0 0 72]
2	4	CMLHL	[20 80]	[100 0]	[100 0]
3	4	CMLHL	[54 45 1]	[3 0 97]	[0 0 100]
4	4	CMLHL	[3 0 42 55]	[48 52 0 0]	[28 72 0 0]

The results shown in Table 7 are coherent with those obtained in Step 2, specially for the ‘Aleg’ station, due to the fact that the samples are allocated in different data clusters when *k*=2 and *k*=3. With regard to the sample process allocation corresponding to both the ‘Arie’ and the ‘Alib’ stations, the results differ from those obtained in Step 2. Though an important number of samples of these two places are assigned to the same cluster (especially in the case of ‘Alib’), in other cases the samples from these two locations are distributed in more than one data cluster. It is also interesting to highlight that when applying CMLHL to reduce the dimensionality of data, all the samples from the ‘Alib’ station are in most cases placed in the same data cluster, while the same was not true in the first case study (Table 2).

In Table 8, the results of fcm applied to the low dimensionality datasets are shown, obtained by applying the three dimensionality reduction techniques.

Table 8: fcm clustering results after dimensionality reduction for case study II (step 3.A).

k	Dimensions	Method	Cluster Samples allocation (%)		
			Aleg	Arie	Alib
3	3	PCA	[51 100]	[97 43]	[100 0]
4	3	PCA	[55 98]	[52 93]	[71 100]
3	4	PCA	[49 57]	[97 50]	[100 0]
4	4	PCA	[55 98]	[51 93]	[71 100]
3	3	LLE	[76 95]	[63 74]	[66 100]
4	3	LLE	[57 40]	[62 64]	[68 100]
3	4	LLE	[78 90]	[60 59]	[66 100]
4	4	LLE	[66 47]	[49 26]	[56 31]
3	3	CMLHL	[49 21]	[97 75]	[100 0]
4	3	CMLHL	[46 92]	[84 88]	[98 100]

3	4	CMLHL	[62 6]	[73 63]	[74 14]
4	4	CMLHL	[66 25]	[95 67]	[77 18]

In Table 8, samples from ‘Alib’ are assigned to the same data cluster in the first option in high percentages, but this percentage is lower in the case of the samples from ‘Arie’. This means that samples from ‘Alib’ show similar levels of air pollution throughout the year, very different from those in ‘Arie’ and ‘Aleg’. The percentage allocation in the first option of the samples from ‘Aleg’ is very low; these samples have a high variability in the levels of pollution, as previously mentioned in Steps 1 and 2.

Step 3.B. Finally, *k*-means and fcm are applied to the original dataset.

In Table 9, the results for *k*-means are shown in application to the original dataset.

Table 9: *k*-means clustering results for case study II on original data (Step 3.B).

k	**Cluster Samples Allocation (%)**		
	Aleg	**Arie**	**Alib**
2	[10 90]	[100 0]	[100 0]
3	[1 45 54]	[97 0 3]	[100 0 0]
4	[0 46 32 22]	[97 0 3 0]	[100 0 0 0]

By applying *k*-means to the original data set, the results are similar to those obtained in Step 2. ‘Arie’ and ‘Alib’ samples are mainly assigned to the same data cluster (first data cluster in this case), while samples from ‘Aleg’ are distributed to the rest of the data clusters. One difference between these results and the results obtained in Step 2 is that when $k = 4$, in Table 9, no sample from ‘Aleg’ is assigned to the first cluster but, in Figure 12, the four clusters contain no samples from this station. It is worth mentioning that all the samples from ‘Alib’ are assigned to the same cluster in the three results. This result resembles those shown in Table 4 in the case of ‘El Atazar’; in both studies one of the stations displays very different characteristics from the other two stations, as most of the samples from ‘Aleb’ are assigned to a different cluster than the samples from ‘Arie’ and ‘Alib’.

In Table 10, the results from fcm are shown applied to the original dataset.

Table 10: fcm clustering results for case study II on original data (Step 3.B).

k	**Cluster Samples allocation (%)**		
	Aleg	**Arie**	**Alib**
3	[49 57]	[97 50]	[100 0]
4	[54 6]	[51 6]	[74 88]

By applying fcm to the original data set, the results obtained are similar to those obtained by applying *k*-means (Table 9) when $k=3$. When $k=3$, in ‘Arie’ and ‘Alib’, almost all the samples are assigned to the same data cluster, this means that the samples from these two places showed similar pollution values over the years, and very different values to those of ‘Aleg’. Something similar happened when applying fcm to the reduced datasets (Table 8), when PCA and MLHL were applied and the number of dimensions was 3.

6 Conclusions, Discussion and Future Work

There are two sets of main conclusions: firstly, those associated with air pollution in the case studies under analysis; secondly, those on the behavior of the proposed HAIS.

Discussing air quality in the region of Madrid, the station at ‘El Atazar’ shows the lowest levels of contamination of the three locations. The highest levels of air pollution correspond to the station at 'Aranjuez' where pollution levels are the highest of the three locations. Summer and spring seasons are linked to the highest levels of O_3 and winter and autumn registered the highest levels of NO and NO_2.

In the region of Prague, the three locations are not so clearly differentiated by air quality, as in the case of the region of Madrid. The stations located in ‘Arie’ and ‘Alib’ have similar levels of air pollution and the station at ‘Aleg’ presents the

highest levels of air pollution especially in NO_2 and PM10. The NO_X presents great variability reaching peaks of 285 μg/m³ in August of 2011 and values over 200 μg/m³ in winter of 2009.

Air quality in the two selected regions behaves in a similar way, reaching very high values in urban areas and decreasing considerably in residential or rural areas. However, it is important to note that in the region of Madrid values do not reach significantly high values, while in the region of Prague peaks (which widely exceeds the maximum levels stated by the European Union) are observed, especially in NOx and PM10. It can also be seen how the levels of environmental pollution in the region of Madrid are strongly related to the placement of the measuring station (city center, background station, vegetable protection station), while in the region of Prague the pollution levels do not depend that much on the location of the measuring station.

The application of dimensionality reduction techniques as the first step in the proposed hybrid system allow us to identify the structure of the data and determine the best way of approaching the characterization of data. Projection of the dataset by more than one criterion yields different views upon it. The dimensionality reduction methods that offer the best results are LLE and CMLHL. As a result of applying this first step, an approximate number of clusters can be obtained. This information is useful, in order to determine the parameters (number of clusters and the number of output dimensions) in Step 2. All in all, at the end of Step 2, the relation between the process of clustering and dimensionality reduction can be visually observed. Finally, in Step 3 (3.A and 3.B) two important issues can be checked. Firstly, the comparison of the results obtained by applying clustering techniques to the original data set or applying it to reduced dimensionality data sets. The most reliable results are obtained by applying clustering to the original data, as in the other case the results vary considerably depending on the dimensionality reduction technique that is applied. In a special situation where dimensionality reduction is applied before *k*-means, the clustering in the original data can be observed in Table 3 (Dimensions=4 and k=2), as 100% of the samples from ‘Aranjuez’ and ‘Casa de Campo’ are assigned to a different cluster than the samples from ‘El Atazar’. The second conclusion is the significant correspondence of these results with those obtained in Steps 1 and 2.

In future work, the proposed hybrid model will be applied to other case studies with larger datasets and further clustering and projection techniques will be investigated.

References

[1] Madrid City Council (2015). *Air Quality.* Available at: http://www.mambiente.munimadrid.es/opencms/opencms/calaire/ContaAtmosferica/portadilla.html

[2] Spanish Government (2015). *El proyecto Aporta como impulsor de la reutilización de la información del Sector Público en España*. available at: http://datos.gob.es/content/proyecto-aporta-como-impulsor-de-reutilizacion-de-informacion-del-sector-publico-espana

[3] Council of Madrid City (2016). *Index - Council of Madrid.* Available at: http://www.madrid.es/portal/site/munimadrid

[4] ISO - International Organization for Standardization (2016). ISO 13271:2012. Available at: http://www.iso.org/iso/iso_catalogue/catalogue_tc/catalogue_detail.htm?csnumber=53581

[5] Encyclopaedia Britannica (2016). *Koppen climate classification.* Available at: http://global.britannica.com/science/Koppen-climate-classification

[6] H. Abdi and L. J. Williams. Principal component analysis. *Wiley Interdisciplinary Reviews: Computational Statistics.* 2(4): pp. 433-459. 2010.

[7] A. Arroyo, E. Corchado and V. Tricio. Atmospheric Pollution Analysis by Unsupervised Learning. *Intelligent Data Engineering and Automated Learning - IDEAL 2009*. 5788: pp. 767-772. Springer, 2009.

[8] Á. Arroyo, E. Corchado, V. Tricio, L. García-Hernández and V. Snášel. Soft Computing techniques applied to a case study of air quality in industrial areas in the Czech Republic. *Soft Computing Models in Industrial and Environmental Applications*. 188: pp. 537-546. Springer, 2013.

[9] Á. Arroyo, V. Tricio, E. Corchado and Á. Herrero. Neuro-Fuzzy Analysis of Atmospheric Pollution. *Hybrid Artificial Intelligent Systems*. 9121: pp. 382-392. Springer, 2015.

[10] G. Chattopadhyay, S. Chattopadhyay and P. Chakraborthy. Principal component analysis and neurocomputing-based models for total ozone concentration over different urban regions of India. *Theoretical and Applied Climatology*. 109(1): pp. 1-11. 2011.

[11] E. Corchado, A. Arroyo and V. Tricio. Soft computing models to identify typical meteorological days. *Logic Journal of the IGPL*. 19(2): pp. 373-383. 2010.

[12] E. Corchado and C. Fyfe. Connectionist techniques for the identification and suppression of interfering underlying factors. *International Journal of Pattern Recognition and Artificial Intelligence.* 17(8): pp. 1447-1466. 2003.

[13] E. Corchado, Y. Han and C. Fyfe. Structuring global responses of local filters using lateral connections. *Journal of Experimental & Theoretical Artificial Intelligence.* 15(4): pp:¡. 473-487. 2003.
[14] E. Corchado, D. MacDonald and C. Fyfe. Maximum and Minimum Likelihood Hebbian Learning for Exploratory Projection Pursuit. *Data Mining and Knowledge Discovery.* 8(3): pp. 203-225. 2004.
[15] E. Corchado and J. C. Perez. A three-step unsupervised neural model for visualizing high complex dimensional spectroscopic data sets. *Pattern Analysis and Applications.* 14(2): pp. 207-218. 2011.
[16] C. Ding and X. He. K-means clustering via principal component analysis. *Proceedings of the twenty-first international conference on Machine learning.* pp: 29. 2004.
[17] X. Gao and W. Xie. Advances in theory and applications of fuzzy clustering. *Chinese Science Bulletin.* 45(11): pp. 961-970. 2000.
[18] T. J. Glezakos, T. A. Tsiligiridis, L. S. Iliadis, C. P. Yialouris, F. P. Maris and K. P. Ferentinos. Feature extraction for time-series data: An artificial neural network evolutionary training model for the management of mountainous watersheds. *Neurocomputing.* 73(1–3): pp. 49-59. 2009.
[19] Czech Hidrometeorological Institute (2015). *Information about air quality in the Czech Republic.* Available at: http://portal.chmi.cz/files/portal/docs/uoco/web_generator/locality/pollution_locality/active_region_district_2731_GB.html
[20] Czech Hydrometeorological Institute. (2015). *Information about air quality in the Czech Republic - ALEG.* Available at: http://portal.chmi.cz/files/portal/docs/uoco/web_generator/locality/pollution_locality/loc_ALEG_GB.html
[21] Czech Hydrometeorological Institute. (2015). *Information about air quality in the Czech Republic - ALIB*. Available at: http://portal.chmi.cz/files/portal/docs/uoco/web_generator/locality/pollution_locality/loc_ALIB_GB.html.
[22] Czech Hydrometeorological Institute. (2015). *Information about air quality in the Czech Republic - ARIE*. Available at: http://portal.chmi.cz/files/portal/docs/uoco/web_generator/locality/pollution_locality/loc_ARIE_GB.html.
[23] A. K. Jain and S. Maheswari. Survey of Recent Clustering Techniques in Data Mining. *International Journal of Computer Science and Management Research*. 1(01). 2013.
[24] P. Kassomenos, S. Vardoulakis, R. Borge, J. Lumbreras, C. Papaloukas and S. Karakitsios. Comparison of statistical clustering techniques for the classification of modelled atmospheric trajectories. *Theoretical and applied climatology.* 102(1-2): pp. 1-12. 2010.
[25] X. Li, S. Lin, S. Yan and D. Xu. Discriminant locally linear embedding with high-order tensor data. *Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on.* 38(2): pp. 342-352. 2008.
[26] Region of Madrid. (2015). *Air Quality Net Stations.* Available at: http://gestiona.madrid.org/azul_internet/html/web/ListaEstacionesAccion.icm?ESTADO_MENU=3_2
[27] Czech Statistical Office (2016). *Czech Statistical Offic.* Available at: https://www.czso.cz/csu/czso/home
[28] E. Oja. Neural networks, principal components, and subspaces. *International Journal of Neural Systems.* 1(1): pp. 61-68. 1989.
[29] E. Oja. Principal components, minor components, and linear neural networks. *Neural Networks.* 5(6): pp. 927-935. 1992.
[30] K. Pearson. On lines and planes of closest fit to systems of points in space. *Philosophical Magazine.* 2(2): pp. 559-572. 1901.
[31] J. C. M. Pires, S. I. V. Sousa, M. C. Pereira, M. C. M. Alvim-Ferraz and F. G. Martins. Management of air quality monitoring using principal component and cluster analysis—Part I: SO2 and PM10. *Atmospheric Environment.* 42(6): pp. 1249-1260. 2008.
[32] S. T. Roweis and L. K. Saul. Nonlinear dimensionality reduction by locally linear embedding. *Science*. 290 (5500): pp. 2323-2326. 2000.
[33] R. San José, J. L. Pérez and R. M. González. An operational real-time air quality modelling system for industrial plants. *Environmental Modelling & Software.* 22(3): pp. 297-307. 2007.
[34] C. Shao and H. Hu. Extension of ISOMAP for Imperfect Manifolds. *Journal of Computers.* 7(7): pp. 1780-1785. 2012.
[35] A. J. Smola and B. Schölkopf. A tutorial on support vector regression. *Statistics and Computing.* 14(3): pp. 199-222. 2004.
[36] European Union. (2016). *European Commission Environment.* Available at: http://ec.europa.eu/environment/air/quality/standards.htm