Neural Networks for the Visual Analysis of Regional Pollution

Ángel Arroyo, Verónica Tricio, Álvaro Herrero
University of Burgos
Burgos, Spain
aarroyop@ubu.es, vtricio@ubu.es, ahcosio@ubu.es

Emilio Corchado
University of Salamanca
Salamanca, Spain
escorchado@usal.es

***Abstract*— This study presents the application of different neural models to a real-life problem of studying the environmental conditions of Castilla y León region (Spain). The goal of this work is to visually and intuitively analyze the level of air pollution in four points of this Spanish region between years 2007 and 2014. The analyzed data were provided by four data acquisition stations from the regional control network of air quality. The main pollutants measured by these stations are analyzed in order to study how the geographical location of these stations and the different seasons of the year are decisive in the behavior of air pollution. Different models for supervised and unsupervised dimensionality reduction have been applied, and subsequent interesting conclusions are obtained.**

Keywords—air quality; pollution; artificial neural networks; dimensionality reduction; exploratory projection pursuit

I. INTRODUCTION

In recent years, our knowledge about atmospheric pollution and our understanding of its effects have advanced greatly [1]. It has been accepted for some years now that air pollution represents a health risk. Systematic measurements in any country are fundamental due to the health risks caused by high levels of atmospheric pollution. The measurement stations acquire data continuously. Thanks to the recent open-data policy promulgated by the Spanish public institutions [2] these data are available for further study and analysis. As the data are ready, different tools to obtain interesting knowledge from there are to be proposed. Present study proposes neural projection techniques for the analysis of air-pollution data.

Artificial Neural Networks (ANN) are one of the main paradigms existing in the area of Artificial Intelligence (AI) [3]. An ANN [4] is a system designed to simulate the way in which the mammal brain performs a particular task or function of interest. In the same way, a task usually performed by ANN is that of dimensionality reduction. The idea is to transform high-dimensional data into a meaningful representation of reduced dimensionality. These methods have been previously applied to the field of Environmental Conditions (EC) [5-7]. There is a wide range of dimensionality reduction techniques based on unsupervised learning, such as Principal Component Analysis (PCA) [8], Local Linear Embedding (LLE) [9], and Cooperative Maximum Likelihood Hebbian Learning (CMLHL) [10], which have previously achieved very good results on the field of EC [11]. Additionally, supervised learning models, such as Linear Discriminant Analysis (LDA) [12] or Quadratic Discriminant Analysis (QDA) [13] are also promising models in such a field. One characteristic of these techniques is that they provide us with an intuitive visualization that let us identify natural clusters of data according to its meaningful information with the naked eye. This is very interesting in those cases in which there is not any label or assignment of each sample to a certain group of data.

According to this idea, in this study, some dimensionality reduction techniques are applied to analyse the atmospheric pollution in four different places in the region of Castilla y León: provinces of Salamanca and Burgos. The data are provided by the regional government of Castilla y León [14].

The remaining sections of this study are organized as follows. Section 2 presents the techniques and methods that are applied to analyse the data. Section 3 details the real-life case study that is addressed in present work, while Section 4 describes the experiments and results. Finally, Section 5 sets out the conclusions and future work.

II. NEURAL PROJECTION TECHNIQUES

To reduce the cost associated to the dimensionality of a representation space, the use of feature selection procedures has been proposed [15], with the aim of reducing this dimensionality. 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} represent each vector in the input space, y_{ik} represent each vector in the output space, n and l are the number of dimensions in the input and output spaces, respectively. To obtain such dimensionality-reduced vectors (y_{ik}), many different techniques can be applied. Following subsections described some of the techniques applied for such a problem that are applied in present study.

A. Unsupervised Learning

Unsupervised learning studies how systems can learn to represent particular input patterns in a way that reflects the statistical structure of the overall collection of input patterns when information about classes (or groups) is not available. Some unsupervised methods are described below.

1) Principal Component Analisys

Principal Component Analysis (PCA) [16] is a well-known method that gives the best linear data compression in terms of least mean square error by addressing the data variance.

978-1-4799-8569-2/15/$31.00 ©2015 IEEE

Although it was proposed as an statistical method, it has been proved that it can be implemented by several ANNs [17, 18].

2) Locally Linear Embedding

Locally Linear Embbeding (LLE) [19] is an unsupervised learning algorithm that computes low-dimensional, neighborhood-preserving embedding of high-dimensional inputs [15]. LLE attempts to discover nonlinear structure 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_i, each of dimensionality D, sampled from some smooth underlying manifold. Provided there is sufficient data (such that 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_i.
2. Compute the weights W_{ij} that best reconstruct each vector xi from its neighbors minimizing the cost in by constrained linear fits:

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

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

$$\Phi(Y) = \sum_i \left| y_i - \sum_j W_{ij} y_j \right|^2 \quad (3)$$

This cost function in (3) like the previous one in (2) is based on locally linear reconstruction errors, but here the weights W_{ij} are fixed while optimizing the coordinates y_i. The embedding cost in (3) defines a quadratic form in the vectors y_i.

3) Cooperative Maximum Likelihood Hebbian Learning

Cooperative Maximum Likelihood Hebbian Learning (CMLHL) [20] is an extended version of Maximum Likelihood Hebbian Learning (MLHL) [21] that incorporates lateral connections, which have been 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 M-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 it is shown in [20], CMLHL can be carried out as a four-step procedure:

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

$$y_i = \sum_{j=1}^{N} W_{ij} x_j, \forall i \quad (4)$$

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

$$y_i(t+1) = [y_i(t) + \tau(b - Ay)]^+ \quad (5)$$

Feedback step:

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

Weight update step, where the following learning rule is applied:

$$\Delta W_{ij} = \eta . y_i . sign(e_j) | e_j |^{p-1} \quad (7)$$

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, and A is a symmetric matrix used to modify the response to the data.

B. Supervised Learning

In supervised learning, the model defines the effect one set of observations, called inputs, has on another set of observations, called outputs.

1) Linear Discriminant Analisys

Linear Discriminant Analysis (LDA) is a linear transformation technique which is commonly used for dimensionality reduction. LDA is a supervised technique and computes the directions, linear discriminants that will represent the axes that maximize the separation between multiple classes [12].

Assume we have a set of D-dimensional samples $\{x^1, x^2, ..., x^N\}$, N^1 of which belong to class W_1, and N^2 to class W_2. We seek to obtain a scalar y by projecting the samples x onto a line $y = W^T$ .Of all the possible lines we would like to select the one that maximizes the separability of the scalars.

In order to find a good projection vector, we need to define a measure of separation – The mean vector of each class in x-space and y-space is:

$$u_i = \frac{1}{N_i} \sum_{x \in wi} \quad (8)$$

And:

$$\tilde{u}_i = \frac{1}{N_i} \sum_{y \in wi} \quad (9)$$

Where:

$$y = \frac{1}{N_i} \sum_{y \in wi} \quad (10)$$

And:

$$w^T x = w^T u_i \quad (11)$$

We could then choose the distance between the projected means as our objective function

$$J(W) = |\tilde{u}_1 - \tilde{u}_2| = |w^T (u_1 - u_2)| \quad (12)$$

III. A REAL CASE STUDY IN CASTILLA-LEÓN

This study is focused on the analysis of pollution data gathered in the Autonomous Community of Castilla-León, Spain. To do so, some representative stations for air quality have been selected. The main reason that determines the selection of the stations listed below is that two of the stations are assigned to the zone division oriented to the health protection, and the other two stations are assigned to the ozone study, according to the zoning process in Castilla y León for the assessment of air quality [22]. Another important reason is the geographical location of the stations. Two of them are located in an urban area and the other two are located in a small town or out of the city. This fact is important for comparing the levels of air pollution between areas with high population density and others with low population density.

The data was recorded from stations located in the points:

- Nuestra Señora de Fátima street, Burgos. Labelled as **Burgos1** in [22]. Geographical coordinates: 03º40'27''W; 42º21'13''N; 929 meters above sea level. Data acquisition station oriented to the study of the ozone.
- Fuentes Blancas, Burgos. Labelled as **Burgos4** in [22]. Geographical coordinates: 03º38'10''W; 42º20'10''N; 929 meters above sea level. Data acquisition station oriented to the health protection.
- La Bañeza, Salamanca. Labelled as **Salamanca5** in [22]. Geographical coordinates: 05º39'55''W; 40º58'45''N; 797 meters above sea level. Data acquisition station oriented to the study of the ozone.
- Aldehuela, Salamanca. Labelled as **Salamanca6** in [22]. Geographical coordinates: 05º38'23''W; 40º57'39''N; 743 meters above sea level. Data acquisition station oriented to the health protection.

There are a total of 353 samples for the twelve months between 2007 and 2014 and the 4 stations analyzed in this study, 12 samples for each station and one sample per month (monthly daily average). Missing values are omitted. The following parameters were analyzed:

1. Nitric Oxide (NO) - $\mu g/m^3$, primary pollutant. NO is a colorless gas which reacts with ozone undergoing rapid oxidation to NO_2, which is the predominant in the atmosphere.
2. Nitrogen Dioxide (NO_2) - $\mu g/m^3$, primary pollutant. From the standpoint of health protection, nitrogen dioxide has set exposure limits for long and short duration.
3. Particulate Matter (PM10, PM2.5) - $\mu g/m^3$, primary pollutant. These particles remain stable in the air for long periods of time without falling to the ground and can be moved by the wind important distances.
4. Sulphur Dioxide (SO_2) - $\mu g/m^3$, primary pollutant. It is a gas. It smells like burnt matches. It also smells suffocating. Sulfur dioxide is produced by volcanoes and in various industrialprocesses. It is also used to protect wine from oxygen and bacteria.

IV. RESULTS AND DISCUSSIONS

The above mentioned dimensionality reduction methods have been applied to data described in section III. PCA is the first technique applied in order to find a possible structure in the data. In Fig. 1. data samples are depicted in the PCA projection, according to their geographical location (data acquisition station).

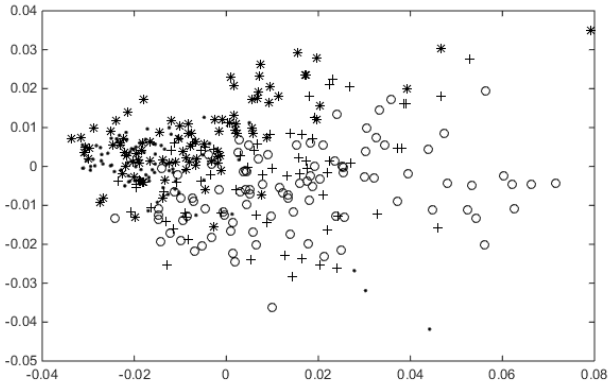

Fig. 1. PCA Projection - station. Number of output dimensions: 2. Label representation: '.' SALAMANCA6, '*' BURGOS4, 'o' SALAMANCA5, '+' BURGOS1.

PCA is not able to depict a clear data structure. Only a certain evolution of the samples according to its geographical location can be identified; especially those related to the control of the ozone levels, represented by '.' and '*' that are located in the upper section of the projection. In these locations the lowest values of pollution, especially NO_2, are given except for the outlier located in the upper right corner. This data point has been analyzed, representing a value of 60 $\mu g/m^3$ in BURGOS4 station in September of 2009, reaching a daily peak of 84 $\mu g/m^3$ the day 22/09/2009. This value is below the hourly limit value for the protection of the human health fixed at 200 $\mu g/m^3$, but above the daily limit value for the human health protection fixed at 50 $\mu g/m^3$since year 2005.It is also noteworthy the two samples belonging to SALAMANCA6, represented by '.', at the bottom of the projection. They are related to an abnormally PM2.5 value of 48 $\mu g/m^3$, registered in January of 2007, reaching a peak value of 89 $\mu g/m^3$ the day 12/01/2007. This value is above the daily limit value for the human health protection fixed at 25 $\mu g/m^3$ since year 2005.

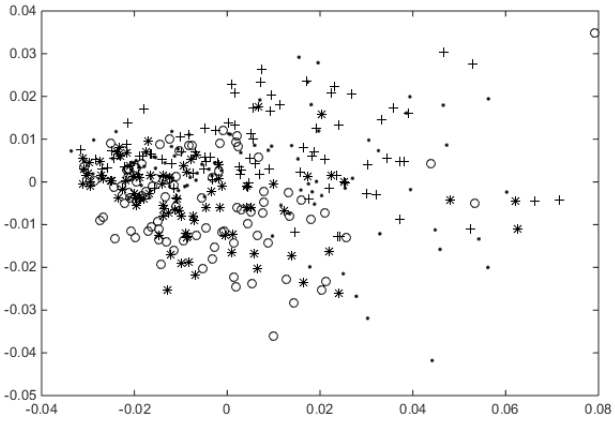

Fig. 2 PCA Projection - season. Number of output dimensions: 2. Label representation: '.' WINTER, '*' SPRING, 'o' SUMMER, '+' AUTUMN.

PCA projection is also depicted in Fig. 2. In this case, data are projected according to the season of the year (winter, spring, summer and autumn). As in the case of Fig. 1, there cannot be seen any clear structure of the samples, except for some differentiation between samples belonging to the spring and summer on the one hand, and those belonging to the other two stations on the other hand.

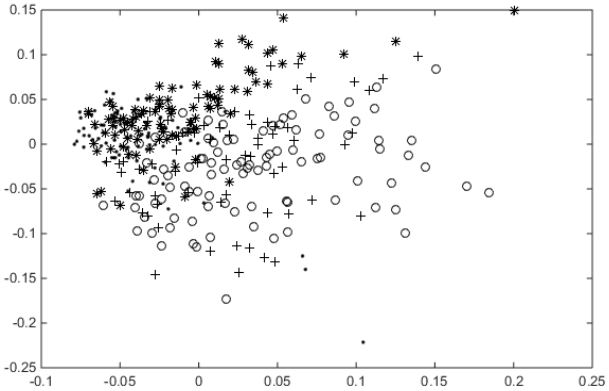

Fig. 3 LLE Projection - stations. Number of output dimensions: 2, Number of Neighbors: 36. Label representation: '.' SALAMANCA6, '*' BURGOS4, 'o' SALAMANCA5, '+' BURGOS1.

Fig. 3. shows the LLE projection where data are depicted according to their geographical location (place of the data acquisition station). LLE outperforms PCA as it is able to separate the samples belonging to SALAMANCA6 and BURGOS4 from the stations oriented to the health protection (SALAMANCA5 and BURGOS1) in a clearer way. The outliers located in the upper right corner and at the bottom of Fig. 3 correspond to the same outliers described for Fig. 1.

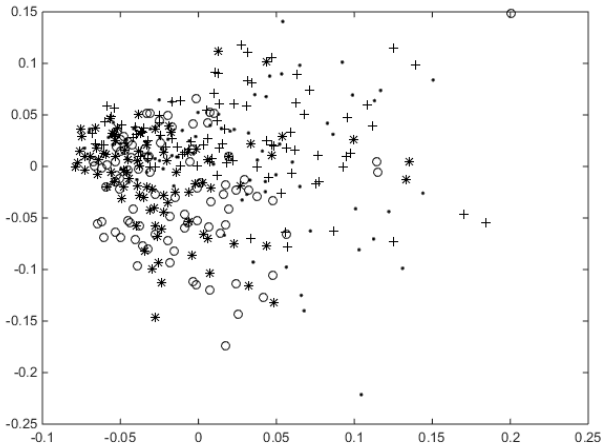

Fig. 4 LLE Projection - season. Number of output dimensions: 2, Number of Neighbors: 36. Label representation: '.' WINTER, '*' SPRING, 'o' SUMMER, '+' AUTUMN.

In the same way that in Fig. 2, LLE does not provide an interesting visualization when data are projected according to the season of the year they belong to (winter, spring, summer and autumn), as shown in Fig. 4. It can only can be highlighted a certain grouping of the samples belonging to the seasons of spring and summer at the bottom of Fig. 4.

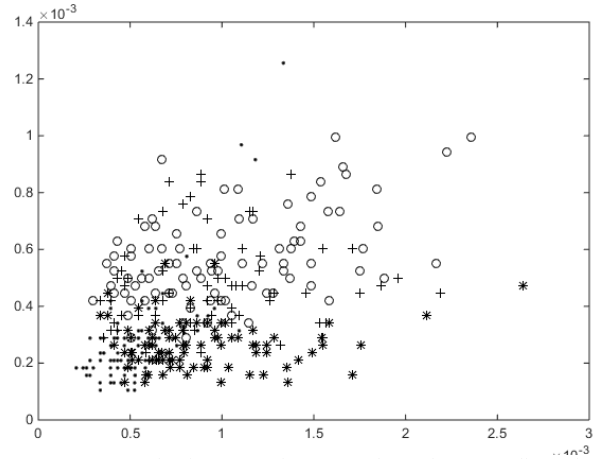

Fig. 5 CMLHL projection – stations. Number of output dimensions: 2, number of iterations: 1000, learning rate: 0.003, p parameter: 2, τ: 2.5. Label representation: '.' SALAMANCA6, '*' BURGOS4, 'o' SALAMANCA5, '+' BURGOS1.

In Fig. 5, the CMLHL projection of same data is shown. According to that, it can be said that this model is able to improve the results obtained by LLE (Fig. 3). While CMLHL is able to distinguish two main groups of data as LLE did in Fig. 3, it also gathers many of the samples belonging to SALAMANCA6 (identified by '.') in a same group, located at the bottom of the projection. The outlier found at the top of Fig. 5 is the same one that has been previously described.

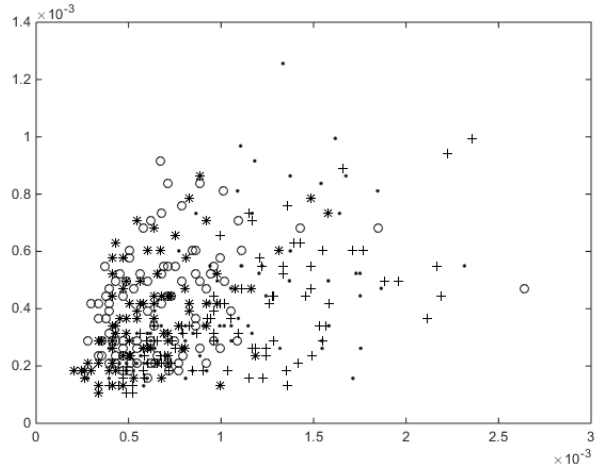

Fig. 6 CMLHL projection - seasons. Number of output dimensions: 2, number of iterations: 1000, learning rate: 0.003, p parameter: 2, τ: 2.5. Label representation: '.' WINTER, '*' SPRING, 'o' SUMMER, '+' AUTUMN.

According to the results obtained in Fig. 2 and Fig. 4, CMLHL is not able to offer good results when data are projected according to the season of the year they belong to (winter, spring, summer and autumn), as can be seen in Fig. 6.

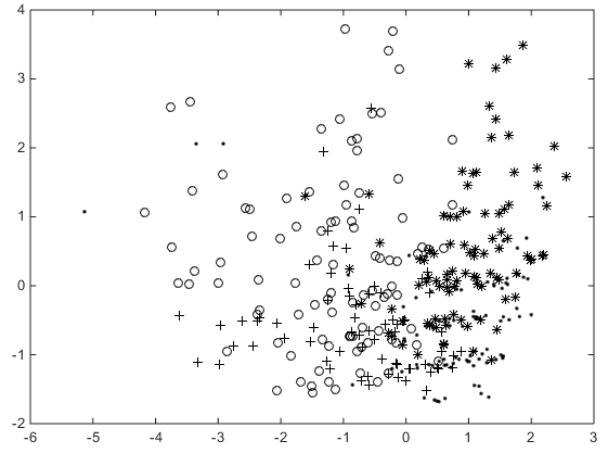

Fig. 7 LDA Projection - stations. Number of output dimensions: 2. Label representation: '.' SALAMANCA6, '*' BURGOS4, 'o' SALAMANCA5, '+' BURGOS1.

LDA projection is shown in Fig. 7. It offers different results from those obtained by PCA, LLE and CMLHL in (Fig. 1, Fig. 2 and Fig. 3 respectively). As LDA is a supervised classification technique, information about classes is required. Four classes are defined in this case, according to the four different locations (BURGOS1, BURGOS4, SALAMANCA5, SALAMANCA6). Two 'zones' of data can be identified in Fig. 7: one of them for SALAMANCA6 and BURGOS4 (on the right side of Fig. 7) and the other one for SALAMANCA5 and BURGOS1 (on the left side). This is quite consistent with the results of PCA, LLE and CMHL, but at this time the data projection are much sparser, providing us with a more intuitive visualization. Additionally, in the case of SALAMANCA6 and BURGOS4 small subsets of data are detected at the bottom of the projection, with samples from SALAMANCA6 and BURGOS4 at the top of Fig. 7.

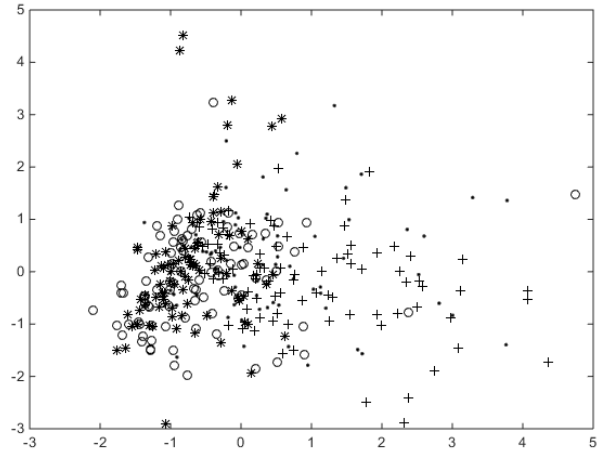

Fig. 8 LDA Projection - seasons. Number of output dimensions: 2. Label representation: '.' WINTER, '*' SPRING, 'o' SUMMER, '+' AUTUMN.

According to results obtained in Fig. 2, Fig. 4 and Fig. 6, LDA is not able to offer good results when data are projected according to the season of the year they belong to (winter, spring, summer and autumn). In the same way that in Fig. 7, four classes are selected, but this time according to the four different seasons of the year (winter, spring, summer and autumn). This time cannot be concluded anything about the results provided by Fig. 8. The selection of classes in this occasion has not been adequate for this case of study.

V. Conclusions and Future Work

Conclusions can be divided into two parts; with regard to the analysis of air pollution in the analysed case study, or related to the behaviour of the four dimensionality reduction methods applied in the case of study and described in Section III.

The analysis of air pollution in the four locations selected shows the fact of the higher levels of pollution in the stations labelled as SALAMANCA5 and BURGOS1 (health protection stations) when compared with SALAMANCA5 and BURGOS1 (ozone level control stations). The pollutants which determine this fact are NO, NO_2 and PM10. This is due to the urban location of SALAMANCA5 and BURGOS1 where traffic emissions originate these higher levels of concentration. The pollutant SO_2 suffers no major changes as it comes mainly from industrial production and none of the four stations located in industrial areas. Regarding the performed analysis on the four seasons, the results are less clear than those offered by projecting station locations. Only a certain grouping of the samples belonging to summer seasons and spring can be seen compared to autumn and winter. Thus, it can be concluded that, from the obtained projection, there are no strong differences from one season to the other one.

When analysing the results of the application of dimensionality reduction techniques, it can be concluded that LLE and specially CMLH improve the results obtained by PCA creating natural groups of data, by applying unsupervised

learning. Applying LDA (supervised learning), the results are very sensitive to the classes selected. When the classes correspond to the four locations the results are as good as in CMLHL or LLE, while if the selected classes are the four weather seasons the results are inconclusive.

In future work, will expand the locations selected covering most areas of the region of Castilla y Leon and stations aimed at the protection of vegetation.

REFERENCES

[1] R. San José, J. L. Pérez and R. M. González, "An operational real-time air quality modelling system for industrial plants," Environmental Modelling & Software, vol. 22, pp. 297-307, 2007.

[2] The Aporta project as a driver of the re-use of public sector information in Spain. Available: http://datos.gob.es/ 2015.

[3] S. J. Russell and P. Norvig, "Artificial intelligence: a modern approach", Prentice hall, 2010.

[4] A. Herrero and E. Corchado, "Mobile Hybrid Intrusion Detection: The MOVICAB-IDS System", Springer, 2011.

[5] E. Corchado, A. Arroyo, and V. Tricio, "Soft computing models to identify typical meteorological days," Logic Journal of IGPL, vol. 19, pp. 373-383, 2010.

[6] 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, vol. 109, pp. 1-11, 2012.

[7] 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, vol. 73, pp. 49-59, 2009.

[8] H. Abdi and L. J. Williams, "Principal component analysis," Wiley Interdisciplinary Reviews: Computational Statistics, vol. 2, pp. 433-459, 2010.

[9] 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, vol. 38, pp. 342-352, 2008.

[10] E. Corchado, Y. Han, and C. Fyfe, "Structuring global responses of local filters using lateral connections," Journal of Experimental & Theoretical Artificial Intelligence, vol. 15, pp. 473-487, 2003.

[11] Á. Arroyo, V. Tricio, E. Corchado, and Á. Herrero, "Neuro-Fuzzy Analysis of Atmospheric Pollution," in Hybrid Artificial Intelligent Systems, Springer International Publishing, pp. 382-392, 2015.

[12] B. Scholkopft and K.-R. Mullert, "Fisher discriminant analysis with kernels", Neural networks for signal processing IX, Vol. 1, pp. 1, 1999.

[13] S. Srivastava, M. R. Gupta, and B. A. Frigyik, "Bayesian Quadratic Discriminant Analysis", Journal of Machine Learning Research, vol. 8, pp. 1277-1305, 2007.

[14] Government of Castilla y León. Air Quality Network. Available: http://servicios.jcyl.es/esco/Login.do 2015

[15] A. J. Smola and B. Schölkopf, "A tutorial on support vector regression", Statistics and Computing, vol. 14, pp. 199-222, 2004.

[16] K. Pearson, "On lines and planes of closest fit to systems of points in space", Philosophical Magazine, vol. 2, pp. 559-572, 1901.

[17] E. Oja, "Principal components, minor components, and linear neural networks", Neural Networks, vol. 5, pp. 927-935, 1992.

[18] E. Oja, "Neural networks, principal components, and subspaces", International Journal of Neural Systems, vol. 1, pp. 61-68, 1989.

[19] S. T. Roweis and L. K. Saul, "Nonlinear dimensionality reduction by locally linear embedding", Science, vol. 290, pp. 2323-2326, 2000.

[20] E. Corchado and C. Fyfe, "Connectionist techniques for the identification and suppression of interfering underlying factors", International Journal of Pattern Recognition and Artificial Intelligence, vol. 17, pp. 1447-1466, 2003.

[21] E. Corchado, D. MacDonald, and C. Fyfe, "Maximum and Minimum Likelihood Hebbian Learning for Exploratory Projection Pursuit", Data Mining and Knowledge Discovery, vol. 8, pp. 203-225, 2004.

[22] Government of Castilla y León. Zoning process in Castilla y León for assessment Air Quality. Available: http://www.jcyl.es/ 2015