Soft computing models to identify typical meteorological days

Emilio Corchado[1], Ángel Arroyo[2,*] and Verónica Tricio[3]

[1]*Departmento de Informática y Automática, Universidad de Salamanca, Salamanca, Spain,* [2]*Department of Civil Engineering, University of Burgos, Burgos, Spain and* [3]*Department of Physics, University of Burgos, Burgos, Spain*

Abstract

Soft computing models are capable of identifying patterns that can characterize a 'typical day' in terms of its meteorological conditions. This multidisciplinary study examines data on six meteorological parameters gathered in a Spanish city. Data on these and other variables were collected for over 6 months, in 2007, from a pollution measurement station that forms part of a network of similar stations in the Spanish Autonomous Region of Castile–Leon. A comparison of the meteorological data allows relationships to be established between the meteorological variables and the days of the year. One of the main contributions of this study is the selection of appropriate data processing techniques, in order to identify typical days by analysing meteorological variables and aerosol pollutants. Two case studies are analysed in an attempt to identify a typical day in summer and in autumn.

Keywords: Artificial neural networks, soft computing, meteorology, atmospheric pollution, statistical models.

1 Introduction

In recent years, our knowledge of atmospheric pollution and our understanding of its effects have advanced greatly. It has now been accepted for some years that air pollution not only represents a health risk, but that it also reduces, e.g., food production and vegetative growth due to its effects on photosynthesis. Other serious consequences may be mentioned, such as acid rain, corrosion, climate change and global warming. Thus, all efforts that are directed towards studying these phenomena may improve our understanding and help us to prevent the serious problematic nature of atmospheric pollution.

Finding solutions to current environmental problems constitutes a fundamental step towards life with a sense of sustainability. Ensuring that we have a clean atmosphere is clearly an important factor, given its impact on the dynamics of the biosphere. An understanding of the mechanisms by which pollutants are emitted into the air is therefore indispensable, as is the knowledge of their atmospheric life cycles, combination reactions and removal paths, among other points, bearing in mind that the approaches to the problem vary according to their spatial and temporal contexts.

Systematic measurements in Spain, which are usually taken within large cities, are fundamental due to the health risks caused by high levels of atmospheric pollution. Recent trends point to the benefits of continuing to extend the network of atmospheric pollution measurement stations. European legislation, as well as setting certain target values with

*E-mail: aarroyop@ubu.es

doi:10.1093/jigpal/jzq035 Advance Access published 21 July 2010

regard to ozone levels will, in the long term, establish how and where such pollutants should be measured.

The basis of this study is the application of a series of statistical and soft computing models to identify what may be called 'Typical Days' in terms of previously selected meteorological variables.

The rest of this study is organized as follows. Section 2 presents the statistical and soft computing methods applied throughout this research. Section 3 details the various case studies and Section 4 describes the experiments and results. Finally, Section 5 sets out the conclusions and future lines of work.

2 Soft computing and statistical models

Several statistical and soft computing models are used in this study to analyse data taken from cases studies on meteorological parameters and aerosol pollutants, in order to assess their performance.

First, a review of principal component analysis (PCA) [1] is presented, followed by an outline of the theory behind exploratory projection pursuit (EPP) [2, 3]. A description is then made of cooperative maximum likelihood Hebbian learning (CMLHL) [4, 5] and the way it can be derived from a PCA neural version is outlined—i.e. the negative feedback network [6]. Finally, local linear embedding (LLE) [7] is reviewed.

2.1 PCA

PCA [1] gives the best linear compression of the data in terms of least mean square error and can be implemented by several artificial neural networks [8, 9]. The basic PCA network [10] applied in this study is described by the following three equations ((1)–(3)): an N-dimensional input vector at time t, $x(t)$, and an M-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:

$$y_i = \sum_{j=1}^{N} W_{ij} x_j, \forall i \tag{1}$$

Feedback step:

$$e_j = x_j - \sum_{i=1}^{M} W_{ij} y_i \tag{2}$$

Change weights:

$$\Delta W_{ij} = \eta e_j y_i \tag{3}$$

This algorithm is equivalent to Oja's Subspace Algorithm [9]:

$$\Delta W_{ij} = \eta e_j y_i = \eta (x_j - \sum_{k} W_{kj} y_k) y_i \tag{4}$$

EPP [2, 3] projects the data onto a low-dimensional subspace that allows its structure to be examined by eye. This is done by means of an index that measures the 'interestingness' of a given projection, the data for which is then represented by projections that maximize the most 'interesting' vectors. 'Interesting' structure is usually defined with respect to the fact that most projections of high-dimensional data onto arbitrary lines through most multidimensional data give almost Gaussian distributions [2, 11]. Therefore, to identify 'interesting' features in data, we should look for those directions onto which the data projections are as far from the Gaussian as possible.

2.2 ε-Insensitive Hebbian learning

It has been shown [12] that the non-linear PCA rule:

$$\Delta W_{ij} = \eta \left(x_j f(y_i) - f(y_i) \sum_k W_{kj} f(y_k) \right) \tag{5}$$

can be derived as an approximation of the best non-linear compression of the data. Initially, therefore, there is a cost function:

$$J(W) = 1^T E \left\{ \left(\mathbf{x} - Wf\left(W^T \mathbf{x} \right) \right)^2 \right\} \tag{6}$$

which may be minimized to give, (5) [12] using the residual in the linear version of (6) to define a cost function of the residual:

$$J = J_1(e) = J(\mathbf{x} - W\mathbf{y}) \tag{7}$$

where $f_1 = ||.||^2$ is the (squared) Euclidean norm in the standard linear or non-linear PCA rule. With this choice of $f_1()$, the cost function is minimized with respect to any set of samples from the data set on the assumption that the residuals are chosen independently and identically distributed from a standard Gaussian distribution. The minimization of J is equivalent to minimizing the negative log probability of the residual, if $\mathbf{e}$ is Gaussian.

$$\text{Let } p(e) = \frac{1}{Z} \exp(-e^2) \tag{8}$$

then, p will denote a general cost function associated with this network:

$$J = -\log p(e) = (e)^2 + K \tag{9}$$

where K is a constant. A gradient descent is performed on J:

$$\Delta W \propto -\frac{\delta J}{\delta W} = \frac{\delta J}{\delta e} \frac{\delta e}{\delta W} \approx \mathbf{y}(2e)^T \tag{10}$$

in which a less important term has been discarded [2].

In general [10], the minimization of such a cost function may be understood to increase the residual probability depending on the probability density function (pdf) of the residuals.

Thus, if the pdf of the residuals is known, this knowledge could be used to determine the optimal cost function; a possibility that was investigated by [14] using the (one-dimensional) function:

$$p(e)=\frac{1}{2+\varepsilon}\exp(-|e|_{\varepsilon}) \tag{11}$$

where:

$$|e|_{\varepsilon}=\begin{cases} o & \forall |e|<\varepsilon \\ |e|-\varepsilon & \text{otherwise} \end{cases} \tag{12}$$

with ε being a small scalar ≥ 0.

With this model of the residual pdf, the optimal $f_1()$ function is the ε-insensitive cost function:

$$f_1(e)=|e|_{\varepsilon} \tag{13}$$

In the case of the negative feedback network, the learning rule is shown:

$$\Delta W \propto -\frac{\delta J}{\delta W}=-\frac{\delta f_1(e)}{\delta e}\frac{\delta e}{\delta W} \tag{14}$$

which gives:

$$\Delta W_{ij}=\begin{cases} o & \text{if} \left|e_j\right|<\varepsilon \\ \text{otherwise} & \eta y(\text{sign}(e)) \end{cases} \tag{15}$$

2.3 CMLHL

CMLHL [4, 5] is an extended version of maximum likelihood Hebbian learning (MLHL) [4, 15], adding lateral connections that have been derived from the rectified Gaussian distribution [12].

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 is shown in [5, 8], the CMLHL can be carried out as a four-step procedure:

Feedforward step, outputs are calculated:

$$y_i=\sum_{j=1}^{N} W_{ij}x_j, \forall i \tag{16}$$

Lateral activation passing step:

$$y_i(t+1)=[y_i(t)+\tau(b-Ay)]^{+} \tag{17}$$

Feedback step:

$$e_j=x_j-\sum_{i=1}^{M} W_{ij}y_i, \forall j \tag{18}$$

Weights update step:

$$\Delta W_{ij} = \eta . y_i . \text{sign}(e_j) |e_j|^{p-1} \tag{19}$$

Where t is time, $[\ ]^+$ is necessary to ensure that the y-values remain in the positive quadrant, 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.

2.4 LLE

LLE [7, 13] is a recently proposed unsupervised procedure for non-linear mapping of high-dimensional data onto a lower dimensional space. In contrast to ISOMAP [16, 17], it attempts to preserve solely local properties of the data, making LLE less vulnerable to short circuiting than ISOMAP. One virtue of LLE is that it avoids the need to solve large dynamic programming problems. LLE also tends to accumulate very sparse matrices, whose structure can be exploited to save time and storage space. The LLE [18] algorithm is based on simple geometric intuitions. 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 neighbours will lie on or close to a locally linear patch of the manifold. The method can be defined as follows:

(1) Compute the neighbours of each vector, X_i.
(2) Compute the weights W_{ij} that best reconstruct each vector X_i from its neighbours, minimizing the cost in (20) by constrained linear fits.

$$\varepsilon(W) = \sum_i \left| x_i - \sum_j W_{ij} x_j \right|^2 \tag{20}$$

(3) Finally, find a point y_i in a lower dimensional space to minimize (21)

$$\Phi(Y) = \sum_i \left| y_i - \sum_j W_{ij} y_j \right|^2 \tag{21}$$

This cost function in (21)—like the previous one in (20)—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 (20) defines a quadratic form in the vectors y_i. Subject to constraints that make the problem well posed, it can be minimized by solving a sparse NxN 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.

3 Case studies: identification of a 'Typical Day' in summer and in autumn seasons

This interdisciplinary study analyses the evolution of different meteorological parameters using the records of a meteorological measurement station (made available by the Department of the Environment – Directorate of Environmental Quality of the Regional Government of the Spanish Autonomous Region of Castile–Leon) [19]. The methods applied in the experimental process are based on data collected at the aforementioned station that is situated in an urban area of the city of Burgos. The study was conducted over approximately half a year in 2007.

Meteorological parameters influence atmospheric pollution levels in various ways. For example, two parameters are very representative of that influence on levels of ozone pollution [20]: solar radiation that directly affects photochemical reactions taking place in the atmosphere; and temperature, because all chemical reactions depend to some extent on temperature and need favourable levels for the reaction to take place. In this study, the following variables have been analysed: wind direction (degrees), wind speed (m/s), dry temperature (°C), relative humidity (%HR), atmospheric pressure (mbar) and solar radiation (W/m^2).

The general characteristics of the geographical zone where the measurement station is placed for this study are as follows: Burgos, a city in north-western Spain with a population of approximately 170,000 inhabitants and a total municipal area of $\sim$107 km^2. The geographic coordinates of the city of Burgos are 854 masl (meters above sea level), latitude (N) 42°20' and longitude (W) 3°42'. The measurement station is an urban station within the city.

The main purpose of this study is to examine the performance of several statistical and soft computing methods when analysing the above-mentioned meteorological variables, in order to identify the existence of 'typical' meteorological days or at least some kind of associated pattern.

4 Experiments and results

This study, which forms part of a more ambitious project, analysed a data set containing meteorological and aerosol pollutant parameters recorded at 15-min intervals: a daily total of 96 records for all data in 2007, referring solely to six variables. All data were previously normalized for the study. The process of selecting a 'typical day' was as follows: initially 1 day in each week was selected randomly for the study, for which the three soft computing models described in Section 2 were applied. The graphical results obtained for every day were analysed and compared with the rest of the days in the data set. For each model, the Typical Day is determined as one which fits a similar pattern in a high percentage of the cases, at all times >70%.

4.1 Analysis of the behaviour of variables

In the first place, the behaviour of all the meteorological variables in relation to the 15-min intervals data acquisition was studied. Figure 1a shows the temporal evolution of the normalized variables for a Typical Day in summer and Figure 1b for a Typical Day in autumn.

Figure 1a shows the great variability of meteorological conditions in Burgos, such as solar radiation and wind speed. On a Typical Day in summer, it can be seen how solar radiation

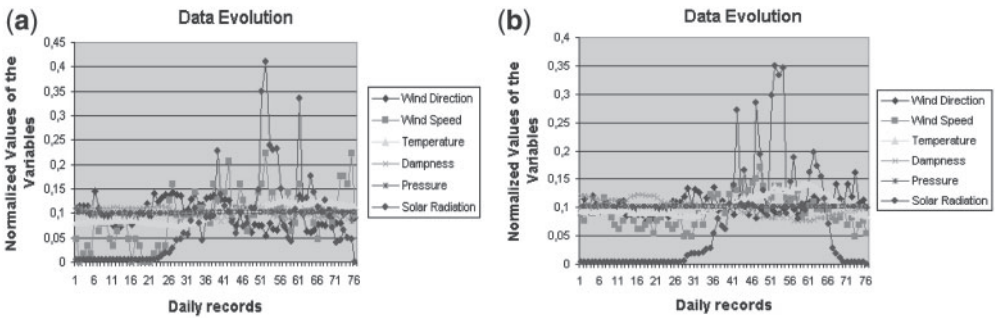

FIG. 1. Temporal evolution of the input variable values for a Typical Day in summer (**a**) and for a Typical Day in autumn (**b**). Each day is described by 96 records.

repeatedly registers high levels throughout long periods of the day. Wind fluctuates around an average value in direction and intensity.

Figure 1b shows that the values of all variables, except for solar radiation, do not fluctuate so much throughout a Typical Day in autumn. Solar radiation is a very significant variable, although it is not as present as on a Typical Day in summer (Figure 1a). It starts increasing earlier in the morning in summer than it does in autumn and it decreases earlier in the afternoon in autumn.

4.2 Analysis of results in the case of study

4.2.1 First case study: a Typical Day in summer

The graphical results obtained in this study for a Typical Day in summer are presented (Figure 2) and analysed as follows.

A Typical Day in Burgos, summer 2007, according to the results obtained by PCA once applied to the meteorological variables is shown in Figure 2a. Two data clusters are identified. Cluster $\mathbf{C}_2$ is related to samples with the highest values, which correspond to the records taken around midday and the early afternoon. Those moments correspond to high levels of solar radiation and temperature. Cluster $\mathbf{C}_1$ is related to samples with the lowest values, which correspond to the earliest as well as to the latest records in the day. Cluster $\mathbf{C}_2$ contains fewer samples than $\mathbf{C}_1$. These are the general characteristics of a Typical Day in summer: variations between the different Typical Days are explained by the lowest values of the most representative variables—temperature and solar radiation—for the day, which are found among the earliest or the latest records of the day. In general, fewer samples register high levels of solar radiation in comparison to samples that register low levels of solar radiation.

Figure 2b shows the results obtained by applying LLE. In this case, the results are really similar to those in Figure 2a, although they improved the results obtained by PCA method, simply because it is easier to analyse the samples contained into the clusters.

Figure 2c shows the results obtained by applying CMLHL. This model is able to improve the response of PCA, (Figure 2a) and LLE, (Figure 2b) as CMLHL is able to identify three clusters instead of two, achieving a sparser representation. Cluster $\mathbf{C}_1$ obtained by PCA and LLE methods contains the same samples as cluster $\mathbf{C}_1$ in CMLHL, but this time the cluster

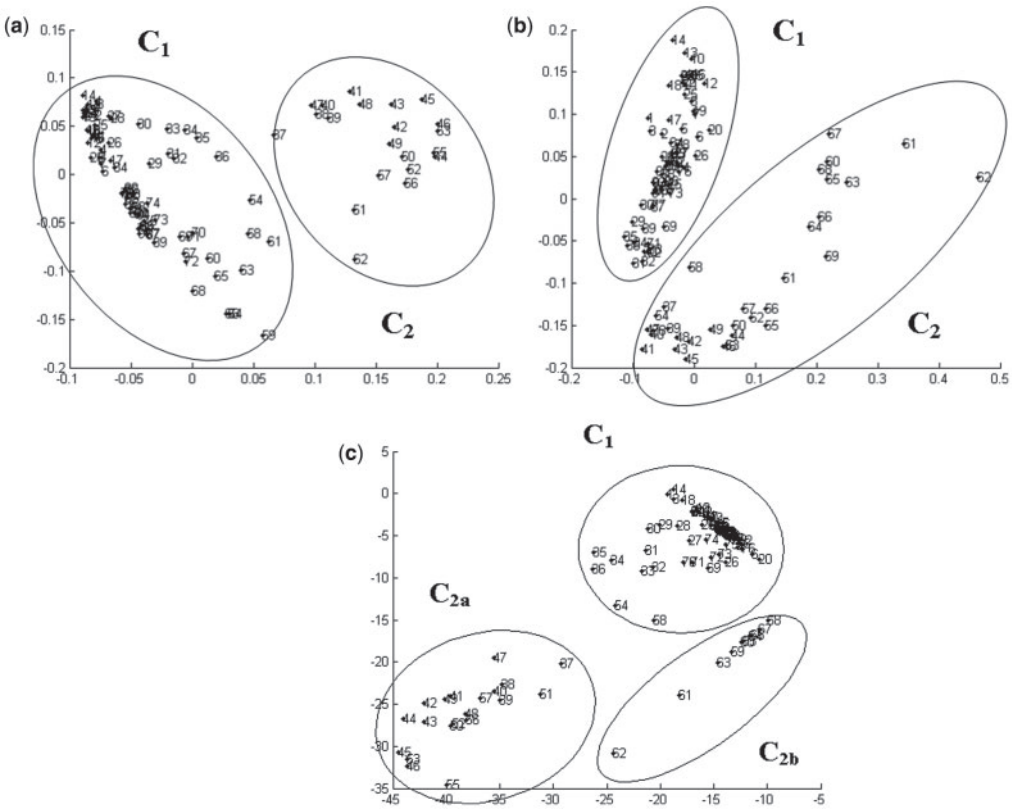

FIG. 2. Results of the three models applied to the problem of identifying a Typical Day in summer. (**a**) PCA—a Typical Day in summer, (**b**) LLE—a Typical Day in summer and (**c**) CMLHL—a Typical Day in summer.

is more spread out, helping to analyse the internal structure in an easier way. The samples in cluster $\mathbf{C}_2$, (Figure 2a and b), in the case of the CMLH model (Figure 2c), are located in Clusters $\mathbf{C}_{2a}$ and $\mathbf{C}_{2b}$. Samples in cluster $\mathbf{C}_{2a}$ correspond to the periods of the day with the highest levels of solar radiation. Samples in $\mathbf{C}_{2b}$ correspond to moments with high solar radiation, but in the late afternoon.

After applying the three methods, it can be concluded that they all are able to identify a certain degree of internal structure. They are all able to identify at least two clear clusters: cluster $\mathbf{C}_1$, which is related to records of day with low levels of solar radiation and low temperatures; and $\mathbf{C}_2$, with a lower number of samples, which groups the moments of the day with high solar radiation and temperature. Those records correspond to midday and afternoon. Nevertheless, CMLHL is able to provide more information as it provides a new cluster of samples and a finer response.

4.2.2 Second case study: a Typical Day in autumn

The graphical results obtained in this study (Figure 3) for the Typical Day in autumn are presented and analysed as follows.

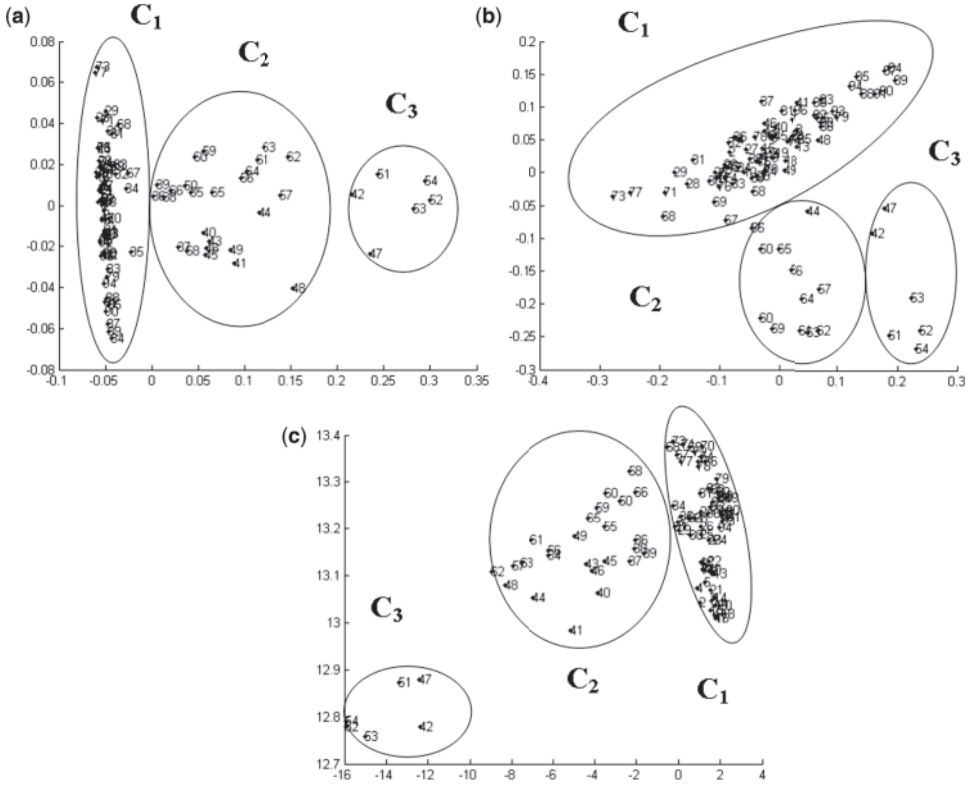

FIG. 3. Results of the four statistical and soft computing models applied to the problem of identifying a Typical Day in autumn. (**a**) PCA—a Typical Day in autumn, (**b**) LLE—a Typical Day in autumn and (**c**) CMLHL—a Typical Day in autumn.

Figure 3a shows the results of applying PCA to identify a Typical Day in autumn, which highlight three clusters ($\mathbf{C}_1$, $\mathbf{C}_2$ and $\mathbf{C}_3$). In this case, $\mathbf{C}_1$ has much more samples than $\mathbf{C}_2$ and $\mathbf{C}_3$. Cluster $\mathbf{C}_2$ once again corresponds to daily records with high levels of temperature and solar radiation, in the early afternoon, but in this case there are few samples with high solar radiation values. Cluster $\mathbf{C}_3$ groups together the highest records related to solar radiation throughout the day, around midday. This is a time of the day when solar radiation is at its highest values in comparison with the rest of the day.

Figure 3b shows the results of applying LLE. Three clusters ($\mathbf{C}_1$, $\mathbf{C}_2$ and $\mathbf{C}_3$) are identified once again. Clusters $\mathbf{C}_1$ and $\mathbf{C}_2$ do not contain the same samples as those found when applying PCA (Figure 3a). This is due to the fact that $\mathbf{C}_1$ contains samples that registered high levels of solar radiation throughout the day.

Finally, Figure 3c shows the results obtained by CMLHL, which are somehow quite similar to the results obtained by the other two previous models, PCA (Figure 3a) or LLE (Figure 3b). Again the same three clusters have been identified. Unlike the other models, in this case it is easier to analyse the samples contained in $\mathbf{C}_2$.

After applying the three models, it may be concluded that the methods provide very similar results allowing a Typical Day in autumn in the city of Burgos to be easily identified.

Nevertheless, the best results are obtained by applying CMLHL, as this model once again provides a sparser representation.

5 Conclusions and future works

The validity of different statistical and soft computing models used to identify a 'Typical Day' in summer and in autumn on the basis of the meteorological data that has been examined in this study.

After applying three different methods to these two case studies, a clear internal structure has been identified. PCA provides an approximation to the internal structure of the data, but the other two soft computing methods are able to provide an improved response. Both methods, LLE and CMLHL, yielded reasonable results for the identification of a Typical Day. CMLHL is the most sensitive method, capable of maximizing the variation in the information to obtain a major and clearer grouping of the samples forming the different clusters, which helps to achieve a better analysis of the results.

The 'Typical Day' in both summer and autumn, in the city of Burgos, resemble each other with regard to the importance of solar radiation in the graph that determines them. What differs between these two typical days is the major variation in environmental conditions in autumn and the difficulty of identifying a 'Typical Day' in that season.

Future work will be based on the analysis of more complex data sets that combine pollution and meteorological data, using soft computing methods, in order to identify the relationship between pollution and meteorological conditions throughout the week and over different time periods. Different soft computing models will also be applied such as Self-Organizing Maps or Spectral Clustering.

Funding

This research has received partial support from the JCyL (projects BU006A08 and BU035A08), and from the Spanish Ministry of Education and Innovation (projects CIT-020000-2008-2 and CIT-020000-2009-12). The authors would also like to thank Grupo Antolin Ingenieria, S.A., which supported this research through the MAGNO2008 - 1028.- CENIT Project funded by the Spanish Ministry.

References

[1] H. Hotelling. Analysis of a complex of statistical variables into principal components. *Journal of Education Psychology*, **24**, 417–444, 1933.

[2] A. Hyvärinen. Complexity pursuit: separating interesting components from time series. *Neural Computation*, **13**, 898–883, 2001.

[3] A. Hyvärinen, J. Karhunen, and E. Oja. *Independent Component Analysis*. Wiley, 2002.

[4] 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**, 1447–1466, 2003.

[5] E. Corchado, Y. Han, and C. Fyfe. Structuring global responses of local filters using lateral connections. *Journal of Experimental & Theoretical Artificial Intelligence*, **15**, 473–487, 2003.

[6] C. Fyfe. PCA properties of interneurons. In *From Neurobiology to Real World Computing, Proceedings of International Conference on Artificial Neural Networks, ICAAN 93*, pp. 183–188. Springer.

[7] S. Roweis and L. Saul. Nonlinear dimensionality reduction by locally linear embedding. *Science*, **290**, 2323–2326, 2000.

[8] E. Oja, H. Ogawa, and J. Wangviwattana. Principal components analysis by homogeneous neural networks, part 1, the weighted Subspace Criterion. *IEICE Transaction on Information and Systems*, **E75D**, 375–366, 1992.

[9] E. Oja. Neural networks, principal components and subspaces. *International Journal of Neural Systems*, **1**, 68–61, 1989.

[10] C. Fyfe and R. Baddeley. Non-linear data structure extraction using simple Hebbian networks. *Biological Cybernetics*, **72**, 541–533, 1995.

[11] S. Seung, N. D. Socci, and D. Lee. The rectified Gaussian distribution. *Advances in Neural Information Processing Systems*, **10**, 350–356, 1998.

[12] P. L. Lai, D. Charles and C. Fyfe. Seeking independence using biologically inspired artificial neural networks. In *Developments in Artificial Neural Network Theory: Independent Component Analysis and Blind Source Separation*, M. A. Girolami, ed. Springer, 2000.

[13] H. Chang and D.Y. Yeung. Robust locally linear embedding. *Pattern Recognition*, **39**, 1053–1065, 2006.

[14] C. Fyfe and D. MacDonald. ε-Insensitive Hebbian learning. *Neurocomputing*, **47**, 35–57, 2002.

[15] E. Corchado, D. MacDonald, and C. Fyfe. Maximum and minimum likelihood Hebbian learning for exploratory projection pursuit. *Data Mining and Knowledge Discovery*, **8**, 203–225, 2004.

[16] J. A. Lee, A. Lendasse, and M. Verleysen. Nonlinear projection with curvilinear distances: ISOMAP versus curvilinear distance analysis. *Neurocomputing*, **57**, 49–76, 2004.

[17] O. Samko, A. D. Marshall, and P. L. Rosin. Selection of the optimal parameter value for the Isomap algorithm. *Pattern Recognition Letters*, **27**, 968–979, 2006.

[18] P. Perona and M. Polito. Grouping and dimensionality reduction by locally linear embedding. In *Advances in Neural Information Processing Systems 14*, T. G. Dietterich, S. Becker, and Z. Ghahramani, eds, pp. 1255–1262. MIT Press, 2002.

[19] V. Tricio, R. Viloria, and A. Minguito. Evolución del ozono en Burgos y provincia a partir de los datos de la red de medida de contaminación atmosférica. Los retos del desarrollo sostenible en España. In *Informe CONAMA 2006*. http://www.conama8.org/modulodocumentos/documentos/CTs/CT86.pdf (Last accessed date June 29, 2010)

[20] V. Tricio, R. Viloria, and A. Minguito. Ozone measurements in urban and semi-rural sites at Burgos (Spain). In *Geophysical Research Abstracts*, vol. 5. EGS-AGU-EUG Joint Assembly, 2003.

Received 10 December 2009