Analysing the Effect of Recent Anti-Pollution Policies in Madrid City through Soft-Computing

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

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

Abstract. This study presents the application of dimensionality reduction and clustering techniques to episodes of high pollution in Madrid City (Spain). The goal of this work is to compare two scenarios with similar atmospheric conditions (periods of high NO_2 concentration): one of them when no actions were taken and the other one when traffic restrictions were imposed. The analyzed data have been gathered from two acquisition stations from the local air control network of Madrid City. The main pollutants recorded at these stations along four days during two time intervals are analyzed in order to determine the effectiveness of the anti-pollution measures. Dimensionality-reduction and clustering techniques have been applied to analyse the pollution public datasets.

Keywords. Clustering, *k*-means, air quality, time evolution, dimensionality reduction, Principal Components Analysis, Locally Linear Embedding.

1 Introduction

In recent years, our knowledge of atmospheric pollution and our understanding of its effects have advanced greatly. Systematic measurements in any country are fundamental due to the health risks caused by high levels of atmospheric pollution. Measurement stations acquire data continuously and, in the case of Spain, these data are available for further study and analysis thanks to the open data policy of public institutions [1]. In the City of Madrid, an Integral Air Quality System [2] was developed in order to monitor the levels of emissions of the main pollutants. This Control System comprises policies with associated actions to be taken during episodes of high pollution by Nitrogen Dioxide (NO_2) [3]. According to European regulation, the maximum value of concentration for this pollutant is 200 μg/m³ averaging in a period of an hour and 40 μg/m³ averaging in a period of a year [4]. In the city centers of many European capital cities (such as Paris, London, etc.) these limits are exceeded when there is no rain and wind, and there are high emissions from road traffic. European capitals are developing protocols and defining actions to restrict traffic in large cities in periods of high air pollution, in order to protect the health of citizens. The city of

Madrid is trying to control the high levels of air pollution specially produced by the traffic emissions, developing an Integral System [2] aimed at knowing the levels of atmospheric pollution in the city in real time. A part of this integral plan is the set of measures to be adopted during episodes of high levels of NO_2 [3]. According to the severity of the situation, two scenarios are defined in the Madrid polices; the Scenario I consist on informing the population and the agents involved, the speed limit in the M-30 (one of Madrid ring-roads) and the accesses to the city (both directions) from the M-40 (another Madrid ring-road) are reduced to 70 km/h, and the use of public transport is promoted. Scenario II comprises the activation of the Environmental Health Alert System and the prohibition of vehicles from non-residents to park in the areas of the SER (Regulated Parking Service) all over the city.

Scant attention has been devoted to this issue in previous work; [5] proposes a network air quality diagnosis of Madrid city center by taking into account both transport exhaust emissions and population exposure levels. The paper aims at identifying air pollution network hotspots in Madrid city center, but does not assess the effectiveness of the anti-pollution measures implemented, as present work does.

Present study focuses on the comparative analysis of the environmental contamination in the center of Madrid City, during two time periods with similar meteorological conditions. Protocols for the control of pollution during episodes of high NO_2 emissions had not yet been approved or activated during the first time period, while in the second period these protocols were applied. The underlying idea is to assess the effect of such protocols by comparing the pollutant levels in similar conditions. In order to do that, two dimensionality-reduction techniques are applied, with different projections, in a first step in order to visualize the internal structure of the data set. In a subsequent step, *k*-means [6] with the most widely-used distance measures is applied and the results are combined with the previously obtained projections for the analysis of the evolution of air quality.

The rest of this paper is organized as follows. Section 2 presents the techniques and methods that are applied. 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 main conclusions and future work.

2 Techniques and Methods

In order to analyse pollution data, two dimensionality reduction techniques and clustering methods are applied. Those methods are described in this section.

2.1 Neural Dimensional-Reduction Techniques

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

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

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.

From the wide range of dimensional-reduction techniques, two different ones have been applied in present work. Principal Component Analysis has been applied as a standard projection technique. At the same time, a more advanced technique (Locally Linear Embbeding) is also applied for comparison purposes.

Principal Component Analysis

Principal Component Analysis (PCA) [7] 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 it was proposed as an statistical method, it has been proved that it can be implemented by an Artificial Neural Network [8].

Locally Linear Embedding

Locally Linear Embbeding (LLE) [9] is an unsupervised learning algorithm that computes low-dimensional, neighborhood-preserving embedding of high-dimensional inputs [10]. 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 x_i from its neighbors minimizing the cost in by constrained linear fits:

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

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

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

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.

2.2 Clustering Techniques

Cluster analysis organizes data by abstracting underlying structure either as a grouping of individuals or as a hierarchy of groups. The representation can then be investigated to see if the data group according to preconceived ideas or to suggest new experiments [10]. Intuitively, two elements belonging to a valid cluster should be more similar to each other than those which are in different groups.

***K*-means**

The standard *k*-means [6] 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 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 is made. The process is repeated until the centroids do not change. Given the strong dependence of this method on initial parameters, a good measure of the goodness of the grouping is simply the sum of the proximity Sum of Squares of Errors (SSE) that it attempts to minimize:

$$SSE = \sum_{j=1}^{k} \sum_{x \in G_j} \frac{p(x_i, c_j)}{n} \tag{4}$$

where p is the proximity function, k is the number of the 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.

For comparison purposes, in present study different distance criteria have been applied, namely: Standardized Euclidean (Seuclidean), Cityblock, Cosine and Correlation.

3 Real-life Case Study

In present study, pollutant data recorded in two different places in the city of Madrid are analyzed. Hourly data from two different time intervals with similar conditions of high air pollution have been selected; in the first one (comprising 4 days) no actions against pollution were taken in Madrid City while in the second one (comprising 4 days), the previously explained Scenarios were activated.

The two stations selected for this study are:

1. Madrid 1. “Plaza del Carmen” Station. Geographical coordinates: 3º 42' 11, 42" W; 40º 25' 09, 15" N; 657 meters above sea level (masl). Data acquisition station characterized as background urban station.
2. Madrid 2. “Escuelas Aguirre” Station. Geographical coordinates: 3º 40' 56, 35" W; 40º 25' 17, 63" N; 672 masl. Data acquisition station characterized as urban traffic station.

They have been selected from the Madrid network of measurement stations due to two main reasons: both of them are located in the M-30 road where protocols for the air pollution control during episodes of high NO_2 are activated, and the two of them record information about the same pollutants, that are:

1. Nitrogen Dioxide (NO_2) - µg/m³, primary pollutant. From the standpoint of health protection, nitrogen dioxide has set exposure limits for long and short duration [11].
2. Sulphur Dioxide (SO_2) - µg/m³, primary pollutant. It is a gas. It smells like burnt matches. It also smells suffocating. Sulfur dioxide is produced by volcanoes and in various industrial processes. In the food industry, it is also used to protect wine from oxygen and bacteria [11].
3. Carbon Monoxide (CO) - µg/m³, primary pollutant. Is an odorless, colorless gas formed by the incomplete combustion of fuels. When people are exposed to CO gas, the CO molecules will displace the oxygen in their bodies and lead to poisoning [11].
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 [11].

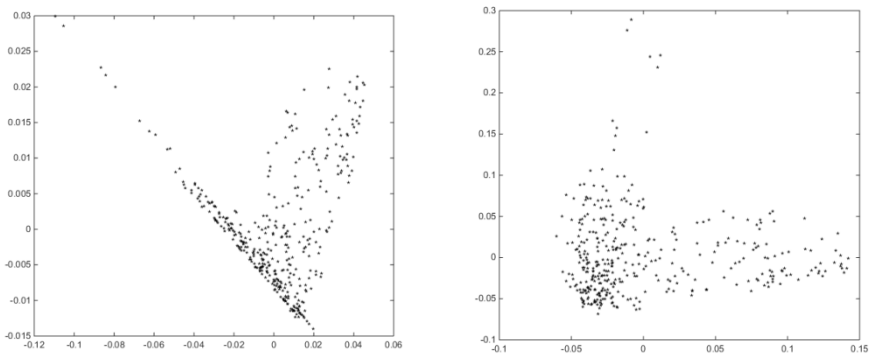
From the timeline point of view, data are selected from two different time intervals:

1. January 8th to 11th, 2015. During these days, some characteristics of high environmental pollution determined by a very dry meteorology and lack of wind were present [12]. The action protocols for episodes of high environmental pollution by emissions of NO_2 were not activated as they were approved in March 2015.
2. March 9th to 12th, 2017. During these days, the environmental conditions were very similar to those in the 2015 period. Protocol actions (above described) were activated on Friday (March, 10th) and Saturday (March, 11th). Apart from these two days, a day before and after the protocol activation have also been considered in order to analyse the consequences. For a fair comparison, the same days (Thursday to Sunday) were also considered in the 2015 period.

Data about the four pollutants were recorded with an hourly frequency so there are a total of 384 samples from the two selected stations (twenty four samples per each one of the four days in each one of the two time intervals).

4 Results and Discussion

The techniques described in Section 2 were applied to the case study presented in Section 3 and the results are discussed below. As a first step, in order to discover a possible internal structure of the dataset, two dimensionality reduction techniques have been applied to the normalized original data set. Fig. 1 shows the results of applying PCA and LLE techniques to the data set.



(a) **(b)**

Fig. 1. Projections of the original data set. Number of output dimensions: 2. **(a)** PCA. **(b)** LLE (number of neighbors: 21).

Fig. 1.a and Fig. 1.b, shows similar results, a big number of samples gathered at the bottom of the projections, a cloud of samples more open in the right side and few samples located at the top. However, LLE provides us with clearer results as the main cloud of samples at the bottom of Fig. 1.b are more open than in those in Fig. 1.a.

Thanks to this initial step, it is known that there is an internal structure in the dataset, whose meaning is to be discovered. In order to do that, further experiments were performed on the same dataset. The result of one of these experiments (Fig. 2) consisted in visualizing samples in the LLE projection according to the day when the data were gathered. The associate legend is shown in Table 1.

Table 1. Sample labeling in Fig. 2

Year	Thursday	Friday	Saturday	Sunday
2015	**1** **8/1/2015**	**2** **9/1/2015**	**3** **10/1/2015**	**4** **11/1/2015**
2017	**5** **9/3/2017**	**6** **10/3/2017**	**7** **11/3/2017**	**8** **12/3/2017**

Fig. 2 shows the LLE projection of the original data according to the recording day.

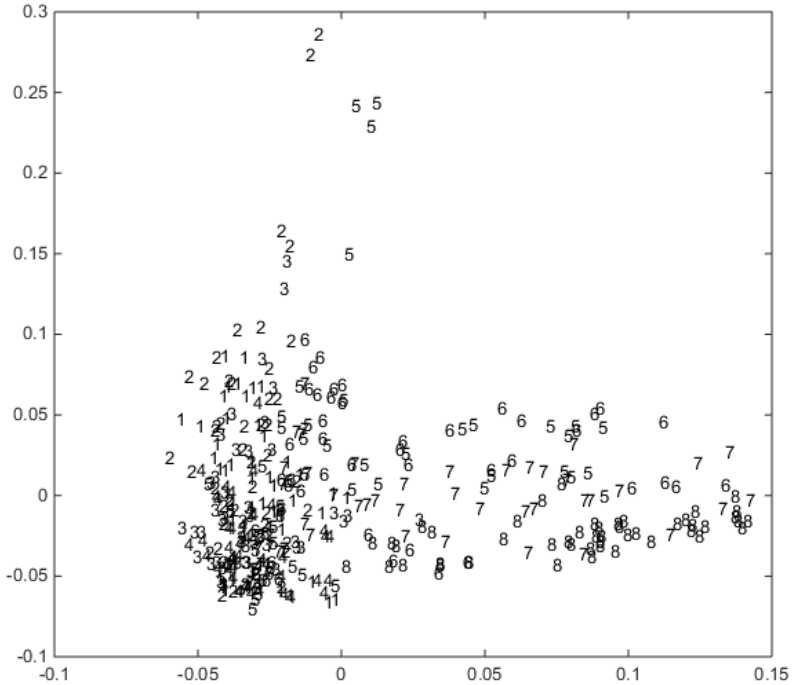


Fig. 2. LLE projection of the original data set showing the recording day.

Located in the lower left side of Fig. 2, most of the samples from day '8' can be observed. These are the samples with the lowest levels of air pollution, as a consequence of applying the anti-pollution Scenario I on day '6' from 6:00 onwards and the Scenario II on day '7' from 9:00 to 15:00. Just above this area, there are most of the samples corresponding to days '6' and '7', during which the two scenarios were applied, environmental pollution decreased during those two days. In the right side of Fig. 2 there is a cloud with many samples from days '1' to '5'. These days, the pollution levels are high and any scenario was still approved. At the top of the image there are some samples with the highest levels of air pollution, reaching values of NO_2 close to 299 ug/m^3.

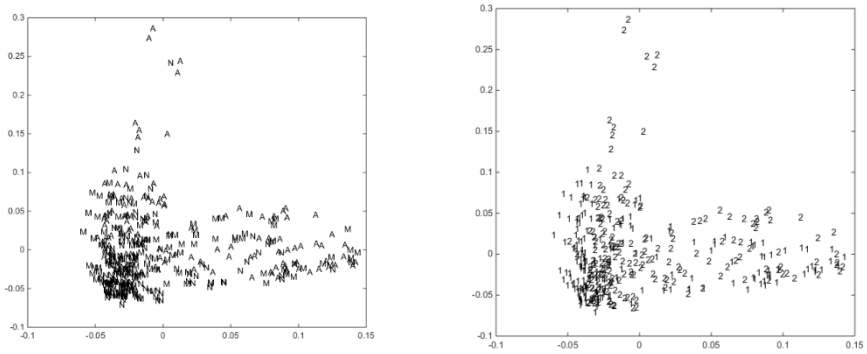
Table 2 shows the results obtained by applying *k*-means with different distance criteria and two suggested values for the *k* parameter (2 and 3). In this table, 'D' is the distance criterion applied: 1 – Squeclidean, 2 – Cityblock, 3 – Cosine, 4 – Correlation (for further details, see Section 2). The Cluster Samples Allocation represents the percentage of samples from each one of the 8 days (from 1 to 8) assigned to each one of the clusters; e. g. [48 52] in day 1 means that there 48% of samples from this day are assigned to the first cluster and the remaining ones (52%) are assigned to the second one.

Table 2. *k*-means clustering results according to the day.

D	*k*	Cluster Samples Allocation (%)							
		1	**2**	**3**	**4**	**5**	**6**	**7**	**8**
1	2	[48 52]	[62 38]	[31 69]	[15 85]	[27 73]	[25 75]	[10 90]	[0 100]
2	2	[90 10]	[98 2]	[98 2]	[90 10]	[69 31]	[67 33]	[54 46]	[0 100]
3	2	[0 100]	[0 100]	[0 100]	[0 100]	[13 87]	[21 79]	[38 62]	[85 15]
4	2	[0 100]	[0 100]	[0 100]	[0 100]	[10 90]	[15 85]	[33 67]	[83 17]
1	3	[4 40 56]	[0 54 46]	[0 25 75]	[0 4 96]	[10 21 69]	[19 19 62]	[34 6 58]	[90 0 10]
2	3	[4 50 46]	[0 38 62]	[0 69 31]	[0 85 15]	[17 58 25]	[25 50 25]	[38 54 8]	[92 8 0]
3	3	[4 0 96]	[0 0 100]	[2 0 98]	[10 0 90]	[25 2 73]	[27 8 65]	[29 21 50]	[31 69 0]
4	3	[96 0 4]	[100 0 0]	[98 0 2]	[96 0 4]	[73 2 25]	[65 8 27]	[52 21 27]	[0 69 31]

After analyzing the results shown in the Table. 2, it can be highlighted that for parameter *k* equal to 2 and 3 and applying specially Cityblock, Cosine and Correlation, most of the samples are located in the same cluster during the first four days (labeled as '1', '2', '3' and '4'), corresponding to the selected period in 2015. On the other hand, during the first day in 2017 (labeled as '5'), most of the samples are concentrated in the second cluster, not in the case of days '6' and '7' in which the air control protocols were applied and samples tend to be distributed between the two or three clusters of data. In the case of day '8' when atmosphere conditions changed, most of the samples are located in a different cluster than the one for previous days.

Fig. 3 shows the LLE projection of the original data according to the hour of the day and to the selected station.



(a) **(b)**

Fig. 3. LLE projections of the original data with different visualizations. **(a)** Visualization of the hours of the day: morning (M) – 8:00-15:00, afternoon (A) – 16:00-23:00, night (N) – 0:00-7:00. **(b)** Visualization of measurement station: 1 – "Plaza del Carmen" station, 2 – "Escuelas Aguirre" station.

In the Fig. 3 (a), to emphasize the difficulty of finding clusters of data according hours of the day, only 'N' samples keep together with similar pollution levels, except

in day '2' when peaks of 256 μg/m³ are reached at 0:00. The highest levels of pollution are presented in the afternoon ('A'). Regarding projection according to the selected stations, observe how both stations show similar levels of air pollution, being the peak values of pollution in the station labeled as '2' – "Escuelas Aguirre" station.

Table 3 shows the results obtained by *k*-means with different distance criteria and the two suggested values for the *k* parameter (2 and 3). Sample process allocation is according to the two selected stations and the hours of the day.

Table 3. *k*-means clustering results according to the selected station (**1** – "Plaza del Carmen" station, **2** – "Escuelas Aguirre" station) and to the hours of the day (M – 8:00-15:00, A – 16:00 – 23:00, N – 0:00-7:00).

Distance	**K**	**Cluster Samples Allocation (%)**				
		1	**2**	**M**	**A**	**N**
Squeclidean	2	[28 72]	[27 73]	[80 20]	[58 42]	[80 20]
Cityblock	2	[57 43]	[60 40]	[77 23]	[64 36]	[71 29]
Cosine	2	[80 20]	[84 16]	[15 85]	[30 70]	[13 87]
Correlation	2	[80 20]	[85 15]	[87 13]	[72 28]	[88 12]
Squeclidean	3	[19 59 21]	[23 59 18]	[14 14 72]	[29 35 36]	[16 14 70]
Cityblock	3	[21 28 51]	[22 26 52]	[18 16 66]	[41 31 28]	[19 19 62]
Cosine	3	[74 10 16]	[69 22 9]	[74 16 10]	[62 16 22]	[79 16 5]
Correlation	3	[10 16 74]	[21 9 70]	[77 13 10]	[62 16 22]	[79 16 5]

Some issues from the results in Table 3 are worth mentioning. Most of the samples from both stations keep together in the same cluster, excepts those obtained when applying Cityblock, as these samples are distributed in all the clusters. This means that the selected stations present similar air pollution conditions along the periods selected. Regarding the hours of the day, most of the samples fill the same cluster of data in the case of the 'M' and 'N', but not in the case of the 'A' samples. This means that in the afternoon ('A'), the air pollution conditions vary from days to days during the two periods. This is probably due to the change of traffic in Friday afternoon, Saturday and Sunday.

5 Conclusions and Future Work

Main conclusions derived from obtained results can be divided into two groups; firstly, those regarding the analysis of air quality conditions in the case study considered. Secondly, those related to the performance of the soft computing techniques applied in the case study.

Concerning the air quality conditions in the two selected places during the two times periods under analysis, it can be said that present study validates the effectiveness of the actions taken for the control of emissions during high periods of NO_2 concentration in the air. In March 10th 2017, Scenario I was activated until the evening of March, 11th 2017 when the Scenario II was activated. The activation of these scenari-

os was decisive in the reduction of emissions of NO_2 with respect to the episode in 2015. During March 12th 2017, the levels of air pollution were even lower due to the change weather conditions.

Regarding the applied dimensional reduction techniques, LLE with the appropriate selection of the number of neighbors and projecting by different concepts, is a very useful tool to discover the internal structure of the data and its possible connections. Once the structure of the data is known and the best projections are identified, *k*-means supports a complementary analysis to see how the samples are gathered in the desired number of groups.

Future work will consist of extending proposed analysis to other European capital cities.

References

[1] Government of Spain – Aporta Project, http://administracionelectronica.gob.es.
[2] Council of Madrid City – Air Quality Integral System, http://www.mambiente.munimadrid.es/opencms/opencms/calaire/SistemaIntegral/concepto.html.
[3] Council of Madrid City – Scenarios for the control of emissions during high periods of NO_2 concentration in the air, http://www.mambiente.munimadrid.es/opencms/opencms/calaire/ServCiudadanos/ProtocoloNO2.html.
[4] European Union – European Commission Environment, http://ec.europa.eu/environment/air/quality/standards.htm.
[5] Prada, F. P., Monzon, A.: Identifying Traffic Emissions Hotspots for Urban Air Quality Interventions: The Case of Madrid City (No. 17-05015) (2017).
[6] Jain, A. K., Murty, M. N., Flynn, P. J.: Data Clustering: A Review. ACM computing surveys (CSUR) 31(3), 264-323 (1999)
[7] Abdi, H., Williams, L. J.: Principal component analysis. Wiley Interdisciplinary Reviews: Computational Statistics. 2(4), 433-459 (2010).
[8] Corchado, E., Han, Y., Fyfe, C.: Structuring global responses of local filters using lateral connections. Journal of Experimental & Theoretical Artificial Intelligence. 15(4), 473-487 (2003).
[9] Roweis, S. T, Saul, L. K.: Nonlinear dimensionality reduction by locally linear embedding. Science. 290 (5500), 2323-2326 (2000).
[10] Li, X., Lin, S., Yan, S., Xu, Y.: Discriminant locally linear embedding with high-order tensor data. Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on. 38(2), 342-352 (2008).
[11] PubChem – PubChem Compounds, https://pubchem.ncbi.nlm.nih.gov/.
[12] Council of Madrid City – Air Quality Report 2015, http://www.mambiente.munimadrid.es/opencms/export/sites/default/calaire/Anexos/Memoria2015.pdf.