Self-Organizing Maps to Validate Anti-Pollution Policies

ÁNGEL ARROYO[*], *Grupo de Inteligencia Computacional Aplicada (GICAP), Departamento de Ingeniería Civil, Escuela Politécnica Superior, Universidad de Burgos, 09006 Burgos, Spain.*

CARLOS CAMBRA[**], *Grupo de Inteligencia Computacional Aplicada (GICAP), Departamento de Ingeniería Civil, Escuela Politécnica Superior, Universidad de Burgos, 09006 Burgos, Spain.*

ÁLVARO HERRERO[†], *Grupo de Inteligencia Computacional Aplicada (GICAP), Departamento de Ingeniería Civil, Escuela Politécnica Superior, Universidad de Burgos, 09006 Burgos, Spain.*

VERÓNICA TRICIO[††], *Departmento de Física, Facultad de Ciencias, Universidad de Burgos, 09002 Burgos, Spain.*

EMILIO CORCHADO[§], *Departamento de Informática y Automática, University of Salamanca, Salamanca, Spain.*

Abstract

This study presents the application of self-organizing maps to air-quality data in order to analyze episodes of high pollution in Madrid (Spain's capital city). The goal of this work is to explore the dataset and then compare several scenarios with similar atmospheric conditions (periods of high Nitrogen dioxide concentration): some of them when no actions were taken and some when traffic restrictions were imposed. The levels of main pollutants, recorded at these stations for eleven days at four different times from 2015 to 2018, are analyzed in order to determine the effectiveness of the anti-pollution measures. The visualization of trajectories on the self-organizing map let us clearly see the evolution of pollution levels and consequently evaluate the effectiveness of the taken measures, after and during the protocol activation time.

Keywords: Air quality, time evolution, self-organizing maps, trajectories, data visualization

1. Introduction

In recent years, knowledge of atmospheric pollution and understanding of its effects have advanced greatly. Systematic measurements are key in every country due to the health risks caused by high

[*]E-mail: aarroyop@ubu.es
[**]E-mail: ahcosio@ubu.es
[†]E-mail: ccbaseca@ubu.es
[††]E-mail: vtricio@ubu.es
[§]E-mail: escorchado@usal.es

Vol. 28, No. 4, © The Author(s) 2019. Published by Oxford University Press. All rights reserved. For permissions, please e-mail: journals.permissions@oup.com.
Advance Access published 9 December 2019 doi:10.1093/jigpal/jzz049

levels of atmospheric pollution [14, 27]. 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 [15]. In the City of Madrid, an Integral Air Quality System (IAQS) [7] was developed in order to monitor the levels of emissions of the main pollutants. To fulfill its objective, the IAQS is constituted by three subsystems: surveillance subsystem, prediction subsystems and information subsystem. The IAQS comprises policies with associated actions to be taken during episodes of high pollution by Nitrogen dioxide (NO_2) [8]. According to the European regulation, the maximum values of concentration for this pollutant are 200 μg/m^3 (averaging in a period of an hour) and 40 μg/m^3 (averaging in a period of a year) [13]. 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. Most European countries 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. This is the case of Madrid, which is trying to control the high levels of air pollution by the IAQS, 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 [8]. According to the severity of the situation, four scenarios are defined: the Scenario I consists 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 owned by non-residents to park in the areas of the regulated parking service all over the city. When the warning level exceeds during two consecutive days, the Scenario III is activated; in addition to the measures adopted during Scenario II, it is added the restriction of circulation in the interior area of the M-30 road for 50% of all vehicles. Furthermore, the non-circulation of empty taxis (except Ecotaxis and Eurotaxis) in the interior area of the M-30 road is recommended. The Scenario IV is considered when the warning level exceededs during three consecutive days or when the alert level is reached. The measures associated with this scenario are the most restrictive ones, comprising the mandatory restriction on the circulation of taxis (except Ecotaxis and Eurotaxis) in the interior area of the M-30 road and a set of measures to promote public transport.

Although meteorological conditions have been previously analyzed by means of machine learning techniques [2], scant attention has been devoted to the problem of forecasting and analyzing short periods of high air pollution by NO_2 in big cities in previous work, [28] 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. In [3], the application of dimensionality reduction [31, 32] and clustering techniques [6, 19, 33] to episodes of high pollution in Madrid City (Spain) is presented in order to visually check the effectiveness of the protocols for traffic restrictions during episodes of high NO_2 levels. In [4], the convenience of quantile regression to predict extreme concentrations of NO_2 is investigated. Using data from the city of Madrid, including NO_2 concentrations as well as meteorological measures, models of quantile regression to predict extreme NO_2 concentrations are built. In [24], the multilayer perceptron is used in order to forecast the maximum daily value of the European Regional Pollution Index as well as the number of consecutive hours with at least one of the pollutants above a limit concentration. The prediction concerns seven different places within the Greater Athens Area, Greece. The air pollution data used in this study have been recorded by the network of the Greek Ministry of the Environment, Physical Planning and Public Works over a 5-year period. In [26], 10 years of ozone hourly concentrations are analyzed. Data were collected from 2000 to 2009 in 11 places located in the Iberian Peninsula. Quantile regression and clustering techniques were applied in order to describe the temporal variability of different quantiles of the ozone distribution. Quantile regression computed the trends at different quantiles of the ozone data distribution while clustering was applied to summarize the resulting distributions of sample quantile slopes. In [1], it is proposed a modeling framework based on the Bayesian Maximum Entropy method that integrates monitoring data and outputs from existing air quality models based on Land Use Regression and Chemical Transport Models. It is proposed to estimate the yearly average NO_2 concentrations over the region of Catalunya (Spain). In [5], multivariate statistical techniques were applied to uncover existing relationships between meteorology and air pollutant (aerosol and trace gas) concentrations and also to reduce data dimensions in Chicago (Illinois) during 2010–2012. Data were explored by principal component analysis (PCA) and canonical correlation analysis (CCA). PCA and CCA brought forth multivariable relationships, not represented in descriptive statistics, useful in understanding pollution variability. In [11], an innovative framework for supporting intelligent analytics over big atmospheric data via clustering-based spatio-temporal analysis is proposed. This paper investigated the interesting applicative setting represented by greenhouse gas emissions (GGEs) in 32 countries. The *k*-means clustering algorithm on data from GGEs was applied.

The self-organizing maps (SOMs) [22] have been used for the analysis of air pollution in many studies. In [23], an air quality modelling that can forecast urban air quality for the next day using airborne pollutant is developed, considering meteorological and timing variables. Hourly airborne pollutant and meteorological averages collected during the years 1995–1997 were analysed in order to identify air quality episodes having typical and the most probable combinations of air pollutant and meteorological variables. This modelling was done using the SOM, the Sammon's mapping [30] and fuzzy distance metrics. In [20], the effects of long-range transport patterns of air masses to the regional PM profile in Istanbul (Turkey) are studied. Five-day hourly backward trajectories were obtained by the HYSPLIT [18] model for selected episodic events in 2008. The SOM was also used to cluster these trajectories.

Present study focuses on the comparative analysis of the environmental pollution in the center of Madrid, during four periods from 2015 to 2018 with similar meteorological conditions. These meteorological conditions are characterized by high stability, due to the practical absence of wind and rain, together with a very dry climate. This set of conditions causes high concentrations of pollutants such as NO_2, PM10, carbon monoxide (CO) and sulphur dioxide (SO_2) among other usual pollutants related to high volumes of traffic. Protocols for the control of pollution during episodes of high NO_2 emissions had not yet been approved during the first time period in year 2015, while they were in force in the other three periods these protocols were applied (years 2017 and 2018). The underlying idea of present study is to assess the effect of such protocols by comparing the pollutant levels in similar conditions.

Unlike the time window in a previous work [3] by the authors, a wider time window is used for data analysis in this study (from 2015 to 2018), and hence four episodes of high NO_2 concentrations are studied and with a larger number of samples in each one of them. In order to do that, the SOM with the extension of trajectory data is applied to the air quality data acquired in the two locations described in Section 3. With the application of the aforementioned SOM trajectory extension (that has not been applied in any of the previous works presented in this section), the study aims to analyze the evolution of the air pollution over the elapsed time: during the days prior to the activation of the protocols described above, during the days of its activation and on subsequent days. It can be observed how the data samples are grouped according to different levels of pollution throughout the eleven days analysed for each one of the four episodes. In previous study, through the application of dimensionality reduction [32] and clustering techniques [19], it was observed the data grouping with similar levels of air quality, but without considering the time evolution.

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. Self-organizing maps

The SOM [21, 22] is a biologically plausible method for visualizing high-dimensional data onto a low dimensional display. It consists of components called nodes or neurons. Associated to each neuron, there is a weight vector of the same dimension as the input data vectors and a position in the map space. The usual arrangement of neurons is a regular spacing in a hexagonal or rectangular grid. The SOM is composed of a discrete array of L nodes arranged on an N-dimensional lattice and it maps these nodes into a D-dimensional data space while maintaining their ordering. The dimensionality (N) of the lattice is usually smaller than that of the input data.

Typically, the array of nodes is one or two-dimensional, with all nodes connected to the N inputs by an N-dimensional weight vector. The self-organization process is commonly implemented as an iterative on-line algorithm. An input vector x is presented to the network and a winning node c is chosen whose weight vector W_c has the smallest Euclidean distance [12] from the input

$$c = \arg\min_i \left(\|x - W_i\|\right). \tag{1}$$

The SOM is a vector quantizer, and data vectors are quantized to the reference vector in the map that is closest to the input vector. The weights of the winning node and the nodes close to it are then updated to move closer to the input vector. The neighbourhood of node i is the set of nodes denoted by $N(i)$ that are close enough to be influenced by the node i whenever it is the winner. Therefore, if the winner is c, then the weights of the nodes $i \in N^{(c)}$ will be updated during training. The amount by which the neighbours are updated is determined by the neighbourhood function (h_{ci}), which is a function that takes into account the Euclidean distance between the winner node (c) and the other nodes in its neighbourhood i. This function is usually a Gaussian function [17]. There is also a learning rate parameter (η) that usually decreases as the training process progresses. The weight update rule is defined as follows:

$$\Delta W_i = \eta h_{ci} \left[x - W_i\right], \forall i \in N^{(c)}. \tag{2}$$

When this algorithm is sufficiently iterated, the map self-organizes to produce a topology-preserving mapping of the lattice of weight vectors to the input space based on the statistics of the training data. Each weight vector lies approximately at the centre of its Voronoi region [34], which holds the subset of points in the data space that are closer to this vector than any other in the map.

2.1. Neighborhood functions

The neighborhood functions applied in the experiments are the following ones:

- Gaussian function

$$h_{ci}(t) = e^{-d_{ci}^2/2\sigma_t^2} \tag{3}$$

Where σ_t is the neighborhood radius at time *t, and* $d_{ci} = |r_c - r_i|$ is the distance between map units *c* and *i*.

- Bubble function

$$h_{ci}(t) = 1\,(\sigma_t - d_{ci}) \tag{4}$$

Where σ_t is the neighborhood radius at time *t, and* $d_{ci} = \left| r_c - r_i \right|$ is the distance between map units *c* and *i*, and *1(x)* is the step function.

2.2. *Unified distance matrix*

The unified distance matrix (U-matrix**)** visualizes distances between neighboring map units, and thus shows the cluster structure of the map: high values of the U-matrix indicate a cluster border while uniform areas of low values indicate the clusters themselves. Each component plane shows the values of one variable in each map unit. On top of these visualizations, additional information can be shown: labels, data histograms and trajectories [16].

The SOM with trajectories facility is available in the SOM MATLAB toolbox [25]. This function launches a ‘comet’ trajectory visualization. It also makes the visualization of the distribution of samples in neurons over time possible. This distribution is graphically shown in the U-matrix as can be seen in Section 4 of present paper.

3. Real-life case study

In present study, pollutant data recorded in two different places in the city of Madrid (Spain) are analyzed (See Figure 1). Hourly data from two different time intervals (with similar conditions of high air-pollution) have been selected. In the first one (comprising 11 days) no actions against pollution were taken while in the other three (comprising 11 days each one of them), the previously explained Scenarios were activated.

The two stations selected for this study are as follows.

Madrid 1. ‘Plaza del Carmen’ Station. 657 meters above sea level (masl). It is a data acquisition station characterized as background urban station.

Madrid 2. ‘Escuelas Aguirre’ Station. 672 masl. It is a data acquisition station characterized as urban traffic station.

These stations have been selected from the Madrid network of measurement stations due to two main reasons: both of them are located close to 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

- NO_2—μg/m^3, primary pollutant. From the standpoint of health protection, exposure to nitrogen dioxide has been limited in the long and short term [29].
- SO_2—μg/m^3, primary pollutant. It is a gas that smells like burnt matches and 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 [29].
- CO—μg/m^3, primary pollutant. It is an odorless and 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 [29].

- Ozone (O_3)—μg/m^3, secondary pollutant. It is an odorless and 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 [29].

From the timeline point of view, data are selected from four different time intervals; in [9] the list of high NO_2 episodes in Madrid city is available. The data from the four days prior to the entry into force of the protocols and the data of some days after the end of the protocols have been selected, in order to study eleven days for each episode. The four episodes analysed in present study are

1. 5th–15th January 2015. During these days, there were some characteristics of high environmental pollution determined by a very dry meteorology and the lack of wind [10]. The protocols in the IAQS were not activated as they were approved in March 2015.
2. 6th–16th March 2017. During these days, the environmental conditions were very similar to those in the 2015 period [10]. Protocol actions associated to Scenarios I and II (above described) were activated on Friday (10th March) and Saturday (11th March).
3. 20th to 30th October 2017. The atmospheric conditions that characterize these days were of high stability [10]. Protocol actions were activated from Tuesday (24th October) to Saturday (30th October). Both Scenarios (I and II), were applied in this episode of high NO_2 concentration.
4. 20th–30th January 2018. Again, a high stability in the climate (low wind and very low humidity) led to a high atmospheric pollution. Protocol actions were activated from Tuesday (23th January) to Wednesday (24th October). Only the Scenario I was activated on this episode.

The episode numbered as 1 has been selected as it was the last episode of high levels of NO_2 prior to the entry into force of the current regulations. The other episodes (2, 3 and 4) have been selected because high levels of NO_2 in the two selected stations were registered and Scenarios I and II were activated. Up to now there have been no episodes where Scenarios III and IV have been activated. Data about the four pollutants were recorded with an hourly frequency (from 1:00 to 24:00), so there is a total of 2,089 samples (24 samples per each one of the eleven days in each one of the four episodes). All data from these six variables have been normalized for the study and missing or corrupted data have been omitted. This data set has been divided into two subsets, one for each measuring station.

4. Results and discussion

The SOM, described in Section 2, has been applied to the case study presented in Section 3 and the results are discussed below. To obtain the results presented in this section, many experiments have been performed with different values for the SOM parameters:

- Initialization: random and linear.
- Training algorithm: batch and sequential.
- Number of neurons: 50, 80, 100, 120, 150 and 200.
- Neighborhood function: Gaussian, cut Gaussian, bubble and Epanechikov functions.

As a first step, the Hits and the U-matrix visualizations of the obtained SOM for the whole dataset (samples corresponding to both stations described in Section 3) are presented in Figure 2. This is

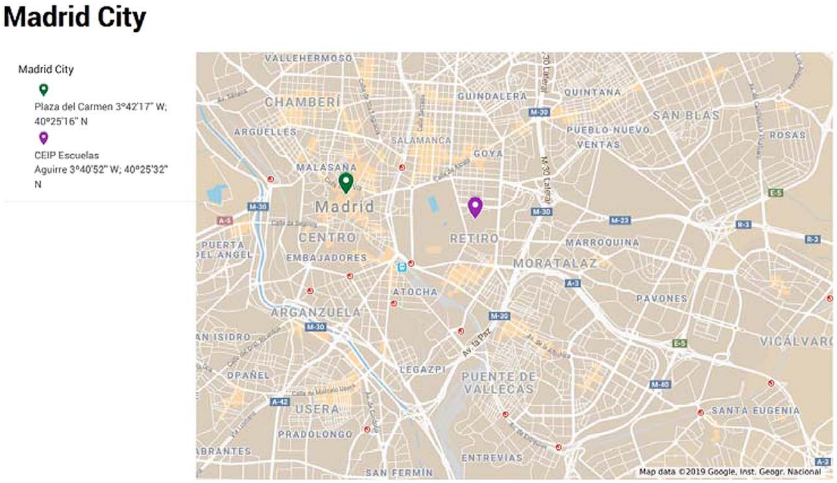

FIGURE 1. Location of the two selected stations in Madrid by Google Maps.

the best result, obtained with the following parameter values: random initialization, batch training algorithm, 200 neurons and Gaussian neighborhood function. In the following sections, data are split according to the station, for a fine-grained analysis.

In order to differentiate the three groups of samples labeled in Figure 2a, two criteria have been used: the variation in the air pollution levels according to Figure 2b and the number of hits associated to these neurons, as shown in Figure 2a. Data in these groups have been analyzed; Group 2 contains most of the samples with highest levels of air pollution, corresponding mainly to the days before and the first days of the protocol activation. Many of the neurons responding to data with the lowest values of pollution are found in Group 3 and correspond to the first days of the analyzed data set (where the maximum levels of NO_2 in the air had not been reached) and to the final days of episodes (because of the positive effect of the traffic control measures). Finally, Group 1 contains a high concentration of data, which have low-average levels of air pollution; these levels can be recorded in nightly shots during any of the analyzed days.

Figure 3 shows the SOM trajectories to the complete dataset (corresponding to both stations).

In Figure 3, the X-axis represents the time in hours for the eleven days analyzed (a total of 264 hours). The first sample at the beginning of the X-axis corresponds to the 1:00 on the first day and the last one corresponds to the 24:00 on the 11th day. The Y-axis represents each one of the 200 neurons in the SOM, numbered from bottom to top. The actions taken according to Scenarios I and II begin on the fourth day in the morning, except for the first episode in which these protocols had not been approved. The protocols were activated between Days 2 and 5, depending on the evolution of weather conditions and the effectiveness of traffic restriction protocols. According to this information, In Figure 3 the samples (in red) located to the left of the horizontal red line are previous to the beginning of the activation of the protocols and the samples at the right side (in green) correspond to the activation of the protocols, as well as a few days after it.

A high number of the green samples are associated to neurons at the upper side of Figure 2, while the red ones are much more associated to neurons in the lower part of the figure. This graphically illustrates the positive influence of the activation of the measures of traffic restrictions, showing the positive evolution of pollution levels.

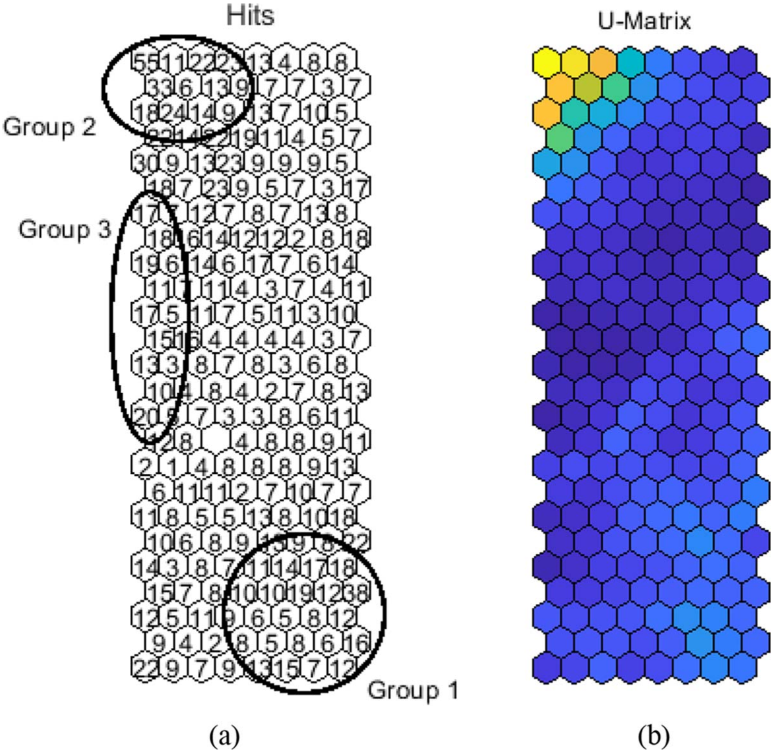

FIGURE 2. (a) Hits and (b) U-matrix visualizations for the complete dataset.

Figure 4 shows the U-matrix with the trajectories for both stations, where data are colored in the same way as in Figure 3. By means of the U-matrix, it is visually identified that the Group 2 in Figure 2a, corresponds in Figure 4 mainly whit data in red and samples assigned to Groups 1 and 3 in Figure 2a are most of them those in green.

4.1. *'Plaza del Carmen' station*

New experiments were run for the dataset that only contains data from the 'Plaza del Carmen' station. Table 1 shows a complete description of this dataset.

The Hits and the U-matrix visualizations corresponding to the best SOM mapping for the 'Plaza del Carmen' dataset are shown in Figure 5. It was obtained with the following parameter values: random initialization, batch training algorithm, 100 neurons, and bubble neighborhood function.

In Figure 5a, samples with the highest levels of air pollution are gathered in Group 2. On the contrary, Group 3 contains most of the samples with the lowest levels of air pollution. In Group 1a, a high number of samples with moderate levels of air pollution are grouped, which correspond to periods with low levels of traffic (at night) or to the days when the consequences of the protocols for the traffic control are starting to be perceived. It can be said that these results are similar to those shown in Figure 2a.

Figure 6 shows the results of applying the SOM trajectories to the subset of data corresponding to the 'Plaza del Carmen' station.

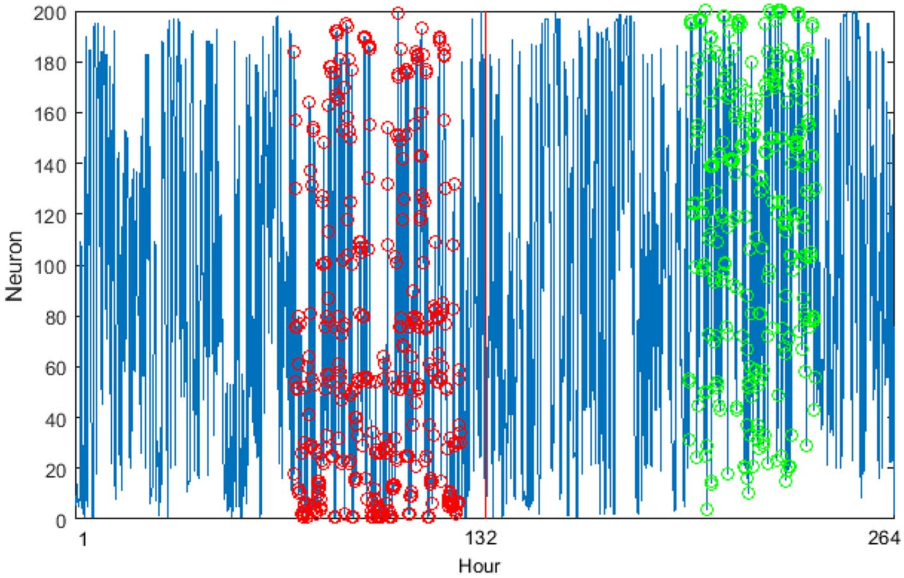

FIGURE 3. SOM trajectories for both stations.

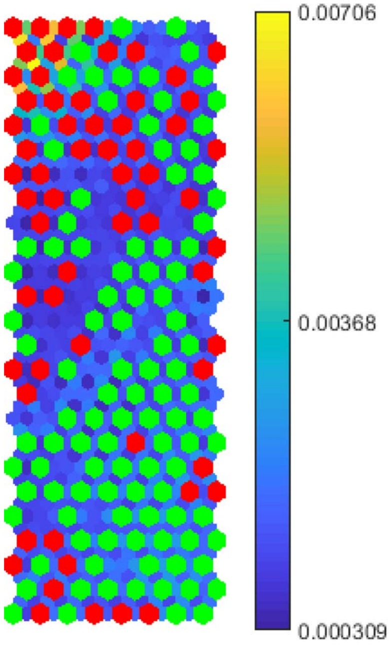

FIGURE 4. SOM trajectories on the U-matrix for both stations.

A high number of samples placed in the right side (in green) of Figure 6, were associated to neurons at the upper side of Figure 5. Complementarily, those in the left side of Figure 6 are much

TABLE 1. Description of the ‘Plaza del Carmen’ dataset.

	NO2	SO2	CO	O3
Units	μg/m^3	μg/m^3	μg/m^3	μg/m^3
Samples episode #1	253	253	253	253
Limit values episode #1 (min/max)	23/178	5/42	0.3/2.2	1/46
Samples episode #2	253	253	253	253
Limit values episode #2 (min/max)	11/137	1/27	0.2/2.9	2/82
Samples episode #3	253	253	253	253
Limit values episode #3 (min/max)	21/188	1/20	0.1/2.6	1/56
Samples episode #4	253	253	253	253
Limit values episode #4 (min/max)	26/112	6/16	0.1/0.8	5/35

TABLE 2. Description of the ‘Escuelas Aguirre’ station.

	NO2	SO2	CO	O3
Units	μg/m^3	μg/m^3	μg/m^3	μg/m^3
Samples episode #1	253	253	253	253
Limit values episode #1 (min/max)	21/299	8/45	0.3/3	4/43
Samples episode #2	253	253	253	253
Limit values episode #2 (min/max)	8/256	5/27	0.2/2.8	4/92
Samples episode #3	253	253	253	253
Limit values episode #3 (min/max)	12/349	2/15	0.2/4.9	5/83
Samples episode #4	253	253	253	253
Limit values episode #4 (min/max)	14/184	2/22	0.2/2	2/67

more distributed around neurons in the lower part of Figure 5. Once again, this graphically illustrates the positive consequences of the traffic-restriction measures. The pattern of time evolution visualized in Figure 3 (both stations) is repeated in Figure 6.

Figure 7 shows the U-matrix with the trajectories for the “Plaza del Carmen” dataset. It corresponds to the samples selected and colored in Figure 4. Thanks to this figure, it can be seen that the Group 2 in Figure 5a, corresponds in Figure 7 mainly whit data in red and samples assigned to Groups 1 and 3 in Figure 5a are most of them those in green.

Figure 8 shows the results of applying the SOM trajectories on each of the original features with normalized pollutant information (NO_2, SO_2, CO, O_3), to the subset of data corresponding to the ‘Plaza del Carmen’ station.

In Figure 8, the evolution of the four selected pollutants can be separately analyzed. It can be seen that the pollutants whose values decrease (in general terms) in the right part of Figure 8 are NO_2, SO_2 and CO. This is very positive because, in the center of the cities, these pollutants mostly depend on the traffic emissions. This fact, once again, demonstrates the effectiveness of the traffic-control protocols. In the case of the O_3 pollutant, it can be said that its evolution may depend more on the variation of weather conditions than on the air quality. Furthermore, only NO_2 influences the levels of O_3 and in a deferred way over time, which could be a reason for the O_3 levels increase in the right part of the image. The meteorological conditions usually change in the last days that are analyzed (right part of Figure 8), what usually causes the end of the episode and its traffic-control restrictions.

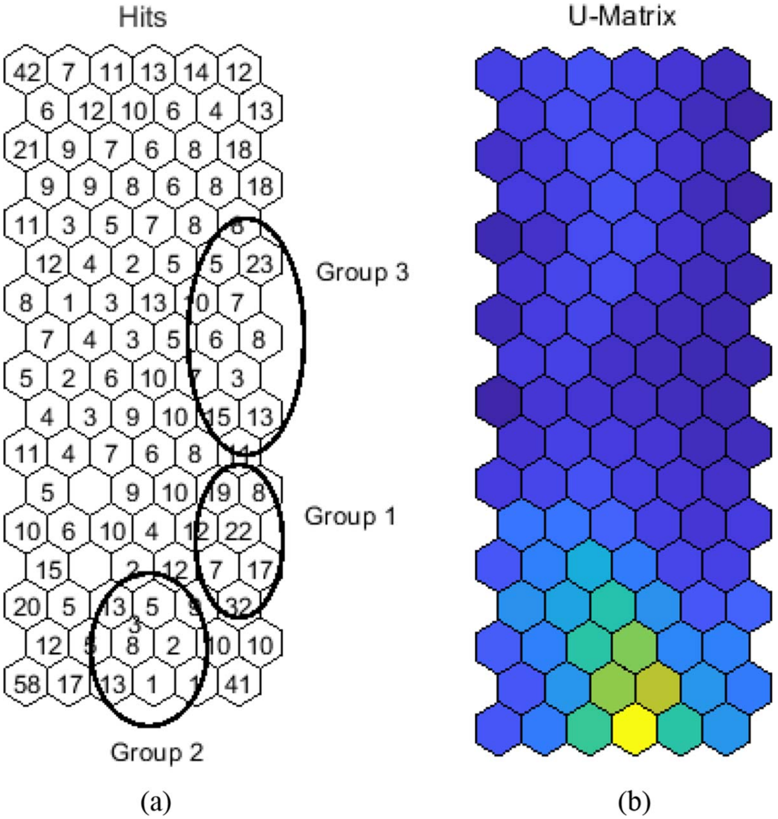

FIGURE 5. (a) Hits and (b) U-matrix for the 'Plaza del Carmen' dataset.

Figure 9 shows the SOM trajectories on the U-matrix corresponding to the 'Plaza del Carmen' subset of data, for each of the four components separately.

It is worth highlighting from Figure 9 that for NO_2, SO_2 and CO pollutants, the areas corresponding to Group 2 in Figure 5a (the set of samples with higher levels of air pollution) are in red. It is different in the case of the O_3 pollutant; as it has been previously mentioned, this pollutant does not depend on the air pollution as the other three ones.

4.2 "Escuelas Aguirre" station

The Hits and U-matrix visualizations corresponding to the best SOM mapping for the 'Escuelas Aguirre' dataset (Table 2 shows a complete description of this dataset.) are shown in Figure 10a. It was obtained with the following parameter values: random initialization, batch training algorithm, 100 neurons and Gaussian neighborhood function.

The U-matrix corresponding to the 'Escuelas Aguirre' dataset shows a mapping similar to those shown in Figures 3a and Figures 5a. Samples with higher levels of air pollution are located in Group 2 and correspond to the days before the entry into force of the pollution protocols. Many of the samples with the lowest levels of air pollution in most of the components are gathered Group 1,

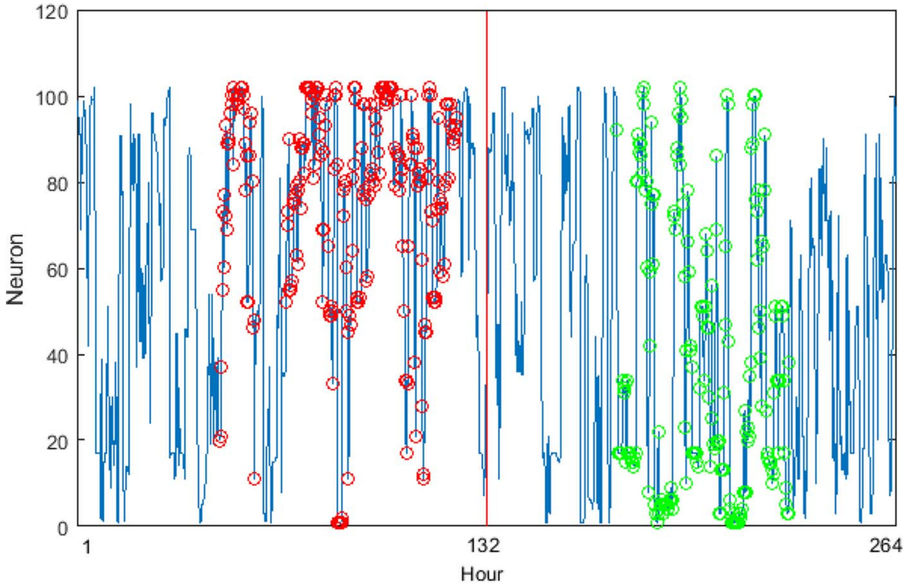

FIGURE 6. SOM trajectories for the ‘Plaza del Carmen’ dataset.

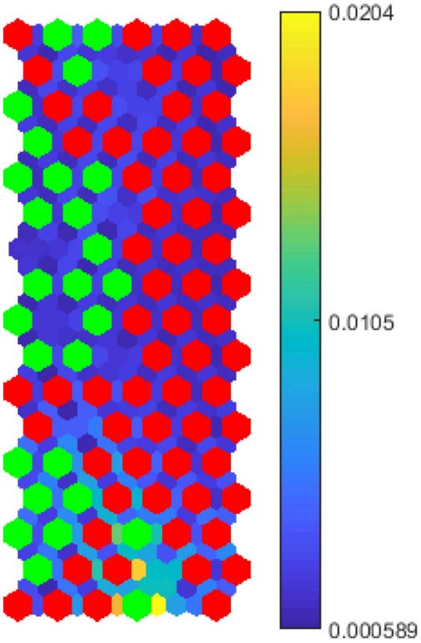

FIGURE 7. SOM trajectories on the U-matrix for the ‘Plaza del Carmen’ dataset.

while a great number of samples with low and medium levels of air pollution (mainly in NO_2, SO_2 and CO) are located in Group 3.

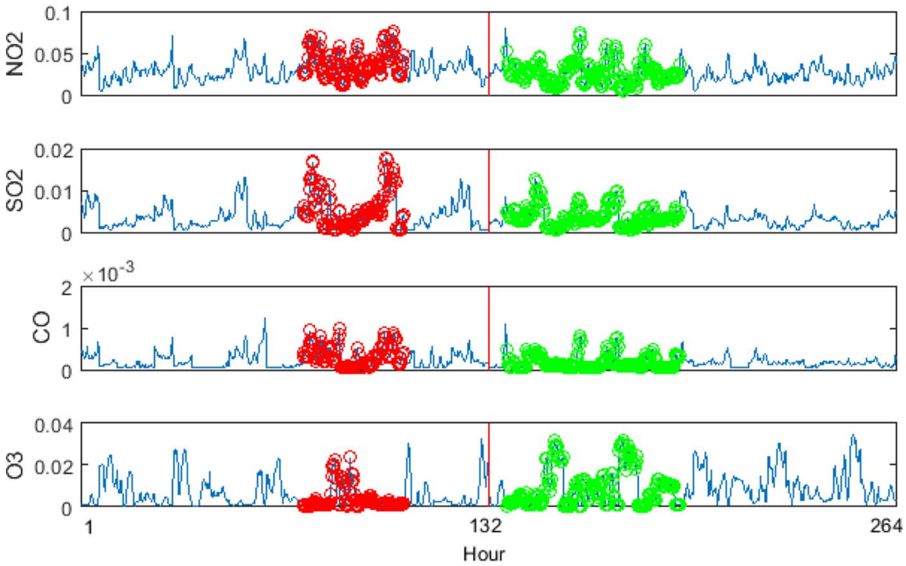

FIGURE 8. SOM trajectories for each one of the four pollutant features ('Plaza del Carmen' dataset).

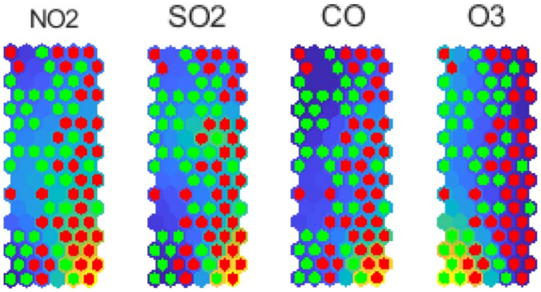

FIGURE 9. SOM trajectories on the U-matrix for each one of the four features ('Plaza del Carmen' dataset).

Figure 11 shows the results of the SOM trajectories when applied to the 'Escuelas Aguirre' dataset. In Figure 11, it can be highlighted that in the right part of the figure, the data samples associated to the days corresponding to the entry into force of the protocols tend to be distributed in neurons different from those of the first days (left part of the image). This fact is consistent with the results from the previous station (Figure 6).

Figure 12 shows the U-matrix corresponding to the SOM trajectory of "Escuelas Aguirre" subset of data.

The trajectories generated in the U-matrix for the "Escuelas Aguirre" station show positive results, since in the upper-left part of the Figure 12 the samples with the highest levels of air pollution are concentrated as can be seen from the neurons in red.

Figure 13 shows the results of applying the SOM trajectories to each one of the four components with pollutant normalized information (NO_2, SO_2, CO, O_3) from the 'Escuelas Aguirre' dataset.

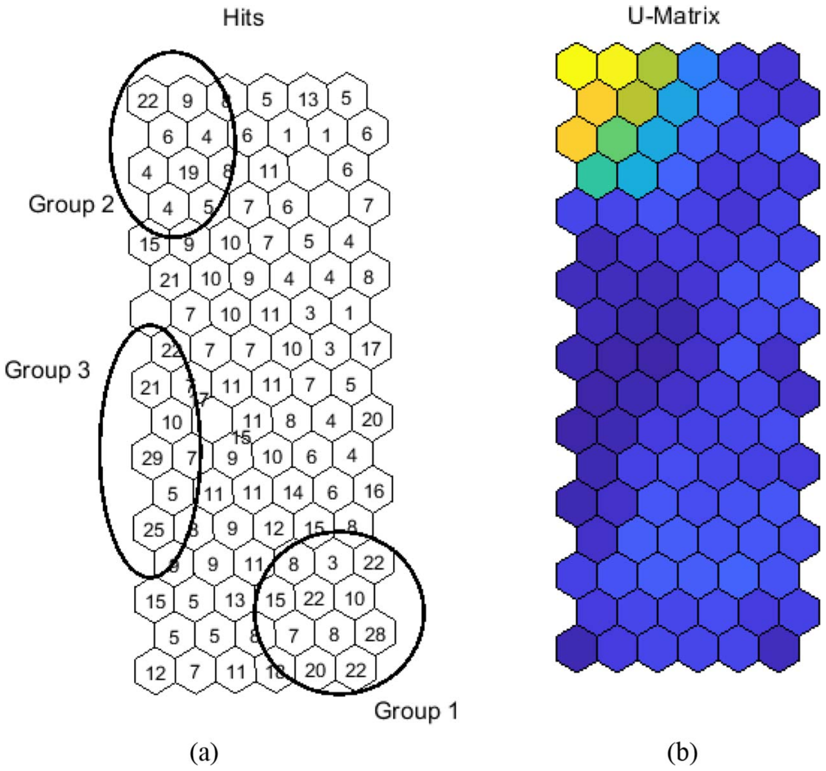

FIGURE 10. (a) Hits and (b) U-matrix for the ‘Escuelas Aguirre’ dataset.

Analyzing the evolution over time of each one of the four components, excluding some pollution peaks which happened in moments of high road traffic in the city center, a clear improvement is observed in the right part of the image for levels of NO_2, SO_2 and CO pollutants. It can be seen a significant decrease in the level of NO_2, which is the main target of these measures involving traffic control, especially considering that it is a station categorized as ‘urban traffic’ station. As in Figure 8, the O_3 level barely changes or increases, this is because it is more susceptible to changes in meteorology than air pollution levels (only influenced by the NO_2 pollutant), and this is consistent with the results analyzed in previous case study (Figure 8).

Figure 14 shows the U-matrix corresponding to the SOM trajectories when applied to the ‘Escuelas Aguirre’ dataset for each one of the four components separately.

In the case of NO_2, SO_2 and CO pollutants, the neurons associated with the highest pollution values (in red) are located in the upper left part of the Figure 14, labelled as Group 2 in Figure 10.

5. Conclusions and future work

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

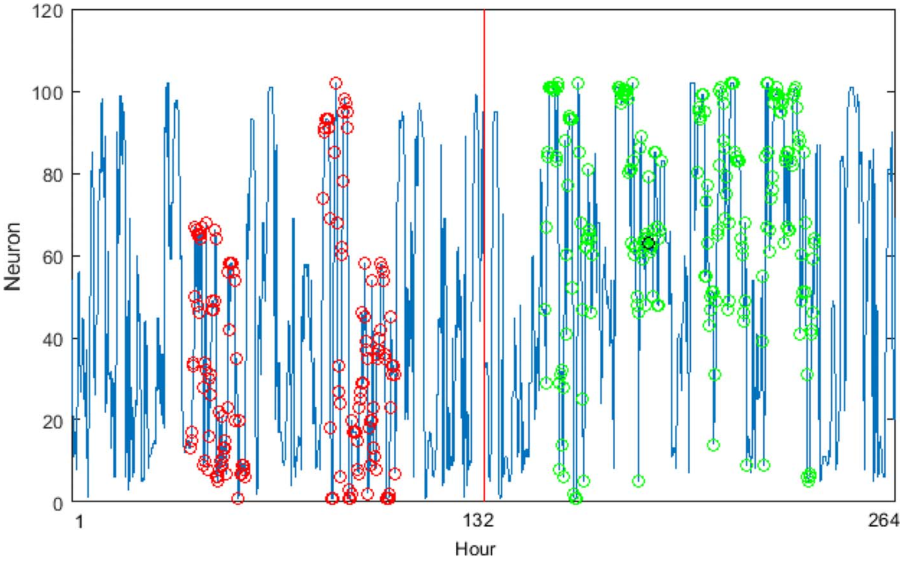

FIGURE 11. SOM trajectories for the ‘Escuelas Aguirre’ dataset.

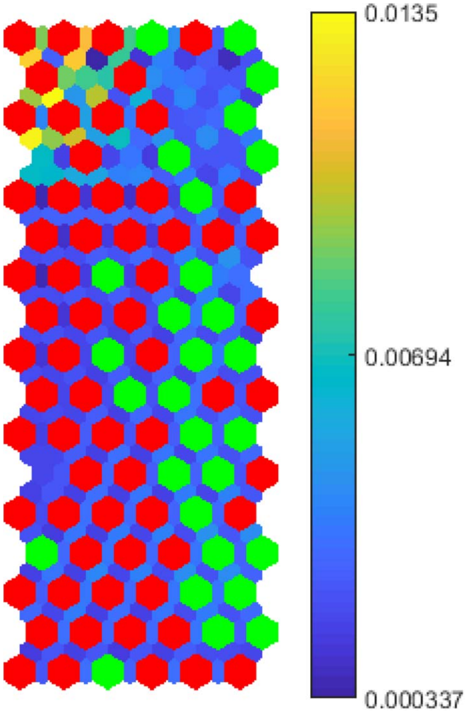

FIGURE 12. SOM trajectories on the U-matrix for the ‘Escuelas Aguirre’ dataset.

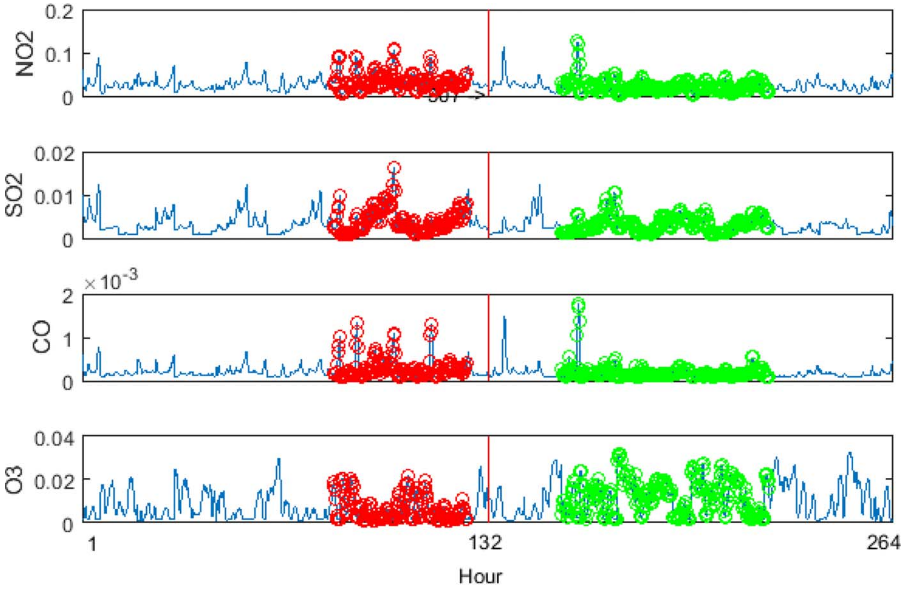

FIGURE 13. SOM trajectories for each one of the four pollutant features (‘Escuelas Aguirre’ dataset).

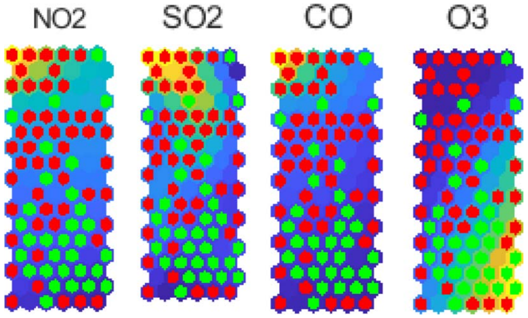

FIGURE 14. SOM trajectories on the U-matrix for each one of the four components (‘Escuelas Aguirre’ dataset).

Since the entry into force of the protocols for the control of road traffic in the center of Madrid, it has been necessary to activate these protocols in several occasions during the years 2017 and 2018 due to episodes of high levels of NO_2. Sometimes only the Scenario I was activated while in other episodes the Scenarios I and II were activated. The elapsed time of these measures is variable, from a couple of days to about a week. Considering the results presented in Section 4, it can be concluded that throughout the activation of these protocols the levels of air pollution are reduced (Figures 6 and 11) and especially the levels of NO_2, SO_2 and CO (Figures 8 and 13). It is important to highlight that in the results presented in Section 4 there is not a crystal-clear aggregation, this is due to the fact that there are data from an episode previous to the approval of the protocols, which are combined with those from three episodes in which the protocols were in force. Comparing the results of ‘Plaza del Carmen’ with those of ‘Escuelas Aguirre’, in the first case a greater decrease was obtained when applying the traffic control protocols (Figure 8 compared to Figure 13). After an analysis of the evolution of the components independently, it can be said that the results are equally positive for both locations (Figures 8 and 13), it is important to highlight the significant reduction in NO_2 levels in the ‘Escuelas Aguirre’ station, which is a very important hit as it is categorized as ‘urban traffic’ station. It is worth mentioning the different behavior of O_3 when compared to the other three pollutants, it is much less influenced by the air quality variability, except by the deferred influence over time of the NO_2 pollutant. This fact makes very interesting the independent analysis of components.

A conclusion derived from the application of the techniques presented in Section 2, it can be said that the results have been extremely satisfactory. The SOM and its trajectories extension have proved to be very useful analysis tools to visualize the evolution over time of a pollution dataset. Previous studies have applied techniques such as dimensionality reduction and clustering to analyze datasets about air quality, but the time variable was omitted. By applying SOM (Figures 5 and 10) it can be seen the samples distributions in neurons. This type of graphical grouping is useful but it only shows static information (at a given time). With the SOM trajectories extension, it is possible to analyze the moments in which there is a concentration of samples around a group of neurons. This is complementary to the information provided by the U-matrix. The extension can also be applied to individual variables, as shown in Figures 8 and 13. It allows of an analysis about which pollutants greatly affect air quality and at what times, which can be useful to deploy traffic control measures that lead to a better health of the population.

Future work will focus on extending the proposed analysis to other European big cities such as Barcelona, Paris or London where similar episodes of high pollution are happening. On the other hand, the SOM trajectories tool will be compared to some other visualization techniques that can deal with the time component.

References

[1] Y. Akita, J. M. Baldasano, R. Beelen, M. Cirach, K. De Hoogh, G. Hoek and A. De Nazelle. Large scale air pollution estimation method combining land use regression and chemical transport modeling in a geostatistical framework. *Environmental Science & Technology*, **48**(8), 4452–4459, (2014).

[2] Á. Arroyo, Á. Herrero, V. Tricio and E. Corchado. Analysis of meteorological conditions in Spain by means of clustering techniques. *Journal of Applied Logic*, **24**, 76–89, 2017.

[3] Á. Arroyo, V. Tricio, Á. Herrero and E. Corchado. (2017). Analysing the effect of recent anti-pollution policies in Madrid City through soft-computing. In *International Joint Conference SOCO’17-CISIS’17-ICEUTE’17 León, Spain, September 6–8, 2017, Proceeding*, pp. 286–295. Springer, Cham.

[4] J. L. Aznarte. Probabilistic forecasting for extreme NO_2 pollution episodes. *Environmental Pollution*, **229**, 321–328, 2017.

[5] K. Binaku and M. Schmeling. Multivariate statistical analyses of air pollutants and meteorology in Chicago during summers 2010–2012. *Air Quality, Atmosphere & Health*, **10**, 1227–1236, 2017.

[6] J. L. Casteleiro-Roca, A. J. Barragán, F. Segura, J. L. Calvo-Rolle and J. M. Andújar. Fuel cell output current prediction with a hybrid intelligent system. *Complexity*, **2019**, 2019.

[7] Council of Madrid City—Air Quality Integral System. http://www.mambiente.munimadrid.es/opencms/opencms/calaire/SistemaIntegral/concepto.html [Accessed 9 April 2019].

[8] Council of Madrid City—Scenarios for the control of emissions during high periods of NO2 concentration in the air. http://www.mambiente.munimadrid.es/opencms/opencms/calaire/ServCiudadanos/ProtocoloNO2.html [Accessed 9 April 2019].

[9] Council of Madrid City—List of Episodes of High NO_2. http://www.mambiente.munimadrid.es/opencms/opencms/calaire/Episodios/Informes_episodios.html [Accessed 9 April 2019].

[10] Council of Madrid City—Air quality annual reports. http://www.mambiente.munimadrid.es/opencms/opencms/calaire/Publicaciones/Memorias.html [Accessed 9 April 2019].

[11] A. Cuzzocrea, M. M. Gaber, E. Fadda and G. M. Grasso. An innovative framework for supporting big atmospheric data analytics via clustering-based spatio-temporal analysis. *Journal of Ambient Intelligence and Humanized Computing*, 1–16, 2018.

[12] P. E. Danielsson. Euclidean distance mapping. *Computer Graphics and Image Processing*, **14**, 227–248, 1980.

[13] European Union—European Commission Environment. http://ec.europa.eu/environment/legal/implementation_en.htm [Accessed 9 April 2019].

[14] IIIB. A. Franklin, R. Brook and C. A. Pope. Air pollution and cardiovascular disease. *Current Problems in Cardiology*, **40**, 207–238, 2015.

[15] Government of Spain—Aporta Project, http://administracionelectronica.gob.es [Accessed 13 March 2019].

[16] I. Horenko. On clustering of non-stationary meteorological time series. *Dynamics of Atmospheres and Oceans*, **49**, 164–187, 2010.

[17] Wolfram MathWorld—Gaussian function. http://mathworld.wolfram.com/GaussianFunction.html [Accessed 9 April 2019].

[18] Air Resource Laboratoy—HYSPLIT project. https://www.arl.noaa.gov/hysplit/hysplit/ [Accessed 9 April 2019].

[19] A. K. Jain and S. Maheswari. Survey of recent clustering techniques in data mining. *International Journal of Computer Science and Management Research*, **3**, 68–75, 2012.

[20] F. Karaca and F. Camci. Distant source contributions to PM10 profile evaluated by SOM based cluster analysis of air mass trajectory sets. *Atmospheric Environment*, **44**, 892–899, 2010.

[21] T. Kohonen. *Self Organization and Associative Memory*. Springer, Berlin Germany, 1988.

[22] T. Kohonen. The self-organizing map. *Proceedings of the IEEE*, **78**, 1464–1480, 1990.

[23] M. Kolehmainen, H. Martikainen, T. Hiltunen and J. Ruuskanen. Forecasting air quality parameters using hybrid neural network modelling. *Environmental Monitoring and Assessment*, **65**, 277–286, 2000.

[24] A. Kurt and A. B. Oktay. Forecasting air pollutant indicator levels with geographic models 3 days in advance using neural networks. *Expert Systems with Applications*, **37**, 7986–7992, 2010.

[25] Laboratory of Computer and Information Science—SOM Toolbox. http://www.cis.hut.fi/projects/somtoolbox/ [Accessed 9 April 2019].

[26] A. Monteiro, A. Carvalho, I. Ribeiro, M. Scotto, S. Barbosa, A. Alonso, J. M. Baldasano, M. T. Pay, A. I. Miranda and C. Borrego. Trends in ozone concentrations in the Iberian Peninsula by quantile regression and clustering. *Atmospheric Environment*, **56**, 184–193, 2012.

[27] C. M. Pintea, A. Calinescu, C. Pop Sitar and P. C. Pop. Towards secure & green two-stage supply chain networks. *Logic Journal of the IGPL*, **27**, 137–148, 2018.

[28] F. P. Prada and A. Monzon. Identifying Traffic Emissions Hotspots for Urban Air Quality Interventions: The Case of Madrid City (No. 17-05015), 2017.

[29] PubChem—PubChem compounds. https://pubchem.ncbi.nlm.nih.gov/ [Accessed 9 April 2019].

[30] J. W. Sammon. A nonlinear mapping for data structure analysis. *IEEE Transactions on Computers*, **100**, 401–409, 1969.

[31] F. S. Tsai. Comparative study of dimensionality reduction techniques for data visualization. *Journal of Artificial Intelligence*, **3**, 119–134, 2010.

[32] L. Van Der, E. Postma and J. Van den. Dimensionality reduction: a comparative. *Journal of Machine Learning Research*, **10**, 66–71, 2009.

[33] M. Verma, M. Srivastava, N. Chack, A. K. Diswar and N. Gupta. A comparative study of various clustering algorithms in data mining. *International Journal of Engineering Research and Applications (IJERA)*, **2**, 1379–1384, 2012.

[34] G. Voronoi. Nouvelles applications des paramètres continus à la théorie des formes quadratiques. Premier mémoire. Sur quelques propriétés des formes quadratiques positives parfaites. *Journal für die Reine und Angewandte Mathematik*, 97–178, 1908.

Received 4 July 2018