# The Mean Angular Distance Among Objects and Its Relationships with Kohonen Artificial Neural Networks

Jorge F. Magallanes,*,†,‡ Jure Zupan,§ Darío Gomez,†,‡ Silvia Reich,|| Laura Dawidowski,† and Neva Groselj§

Comisión Nacional de Energía Atómica, Av. Gral, Paz 1499, San Martín - B1650KNA, Provincia de Buenos Aires, Argentina, Universidad de Buenos Aires, Ciudad Universitaria de Nuñez, CP 1428, Capital Federal, Argentina, National Institute of Chemistry, Hajdrihova 19, SLO-1000 Ljubljana, Slovenia, and Universidad Nacional de Gral, San Martín, Calle Alem 3901, 1653 Villa Ballester Argentina

Received April 3, 2003

This job refers to classification of multidimensional objects and Kohonen artificial neural networks. A new concept is introduced, called the mean angular distance among objects (MADO). Its value can be calculated as the cosine of the mean centered vectors between objects. It can be expressed in matrix form for any number of objects. The MADO allows us to interpret the final organization of the objects in a Kohonen map. Simulated examples demonstrate the relationship between MADO and Kohonen maps and show a way to take advantage of the information present in both of them. Finally, a real analytical chemistry case is analyzed as an application on a big data set of an air quality monitoring campaign. It is possible to discover in it a subgroup of objects with different characteristics than those of the general trend. This subgroup is linked to the existence of an unidentified $SO_2$ source that, a priori, has not been taken into account.

## INTRODUCTION

Many chemometric problems require the interpretation of a large quantity of complex measurements. Usually, "a single measurement" or object demands the recording of many different variables, the number of which may be greater than several hundreds or even thousands of them. Such measurements are expressed in vectorial form $\mathbf{x_i} = (x_{i1},x_{i2},...,x_{im})$. Due to the fact that measurements are made at different conditions, or they represent a time series of measurements of a complex process which are repeated in consecutive time intervals; the complete set of measurements is collected in a big data matrix, **X**. Examples of this kind of measurements are as follows: environmental monitoring campaigns, where a number of chemicals and meteorological variables are monitored simultaneously, studies of climate patterns, quantitative structure−activity relationship (QSAR) data, that links hundreds of structure descriptors with biological or pharmaceutical properties of their respective compounds, etc.[1,2]

Within the scope of multivariate statistics there are many mathematical techniques to analyze these large data sets. Dimensionality reduction is often one of the first goals to simplify data interpretation when dealing with such large matrices.[1,2] Usually the next step is the estimation of the linear or nonlinear relationships among the variables and responses, followed by the selection of the most appropriate method to obtain the best results. One example is the selection of the appropriate modeling technique between multiple linear regression (MLR)[3] and artificial neural networks (ANN).[4]

The Kohonen artificial neural network (KANN) or formation of self-organizing maps (SOMs) is a well-established technique to map the multivariate data into 2D space.[5] Furthermore, the ability of KANN for classification tasks is widely recognized. The detailed analysis of KANN can yield valuable information for further data handling.

In the present work, the connection between the positions of objects within the multivariate space and Kohonen SOMs is analyzed, and the correlation between variables also plays an important role in this development. The study focuses on the interpretation of the mean angular distance among objects in the actual m-dimensional measurement space and the topological distances between the projected objects, i.e., their relative positions in the Kohonen map.

## THE CORRELATION BETWEEN VARIABLES AND THE MEAN ANGULAR DISTANCE AMONG OBJECTS

The symbols for vectors and matrices are hereinafter in bold characters, while regular type is kept for scalar magnitudes.

We should consider under study an experimental data set composed of $n$ objects. Each object $\boldsymbol{x_i}$ is composed of $m$ individual measurements of m variables, $\boldsymbol{x_i} = (x_{i1},x_{i2},...,x_{im})$ for $i = 1...n$. All the measured data can be collected in a data matrix, $\boldsymbol{X}$, of $n$ rows and $m$ columns as follows:

$$\boldsymbol{X} = \begin{vmatrix} x_{1,1} & x_{1,2} & \dots & x_{1,m} \\ x_{2,1} & x_{2,2} & \dots & x_{2,m} \\ . & . & \dots & . \\ . & . & \dots & . \\ x_{n,1} & x_{n,2} & \dots & x_{n,m} \end{vmatrix}$$

* Corresponding author: phone: 54-11-6772-7891; fax: 54 11 6722-7886; e-mail: magallan@cnea.gov.ar.
† Comisión Nacional de Energía Atómica.
‡ Universidad de Buenos Aires.
§ National Institute of Chemistry.
|| Universidad Nacional de Gral, San Martín.

10.1021/ci034062v CCC: $25.00 
Published on Web 08/05/2003

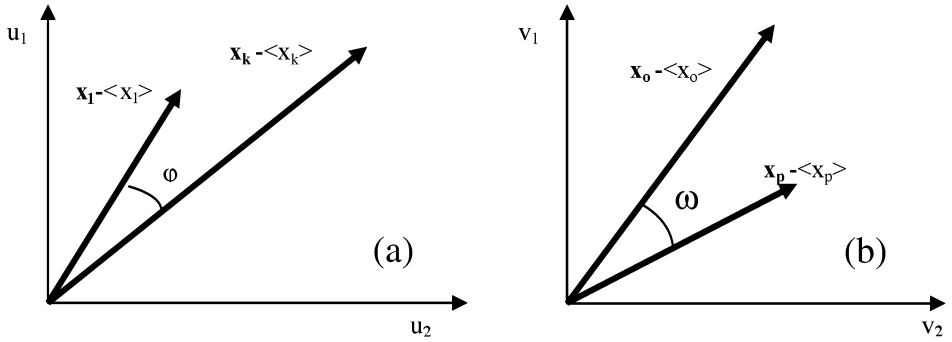


**Figure 1.** (a) Geometrical interpretation of the correlation between "l" and "k" variables. (b) Geometrical interpretation of MADO (eq 6) between objects "o" and "p".

Variables are often associated among them and usually it is important to study the nature of such relationships. One way to express the relationship between two of any these variables, $x_k$ and $x_l$, is the *covariance*, expressed as follows

$$\mathrm{cov}(x_k,x_l) = \frac{1}{n-1}\sum_{h=1}^{n}(x_{h,k} - \langle x_k\rangle)(x_{h,l} - \langle x_l\rangle) \qquad (1)$$

where $\langle x_i\rangle$ represents the mean value of the column $i$. The *correlation coefficient* or Pearson correlation coefficient, $r(x_k,x_l)$, similarly describes the same relationship

$$r(x_k,x_l) = cov(x_k,x_l)/(s_k s_l) \qquad (2)$$

taking into account the estimator of the standard error of the mean, $s_k$ and $s_l$, for each of both variables, respectively. It is possible and more convenient to describe the correlation coefficient as a vectorial expression:

$$r(x_k,x_l) = \frac{(\boldsymbol{X}_\mathbf{k} - \langle x_k\rangle)^T(\boldsymbol{X}_\mathbf{l} - \langle x_l\rangle)}{||\boldsymbol{X}_\mathbf{k} - \langle x_k\rangle||\ ||\boldsymbol{X}_\mathbf{l} - \langle x_l\rangle||} = \cos\varphi \qquad (3)$$

Here, the bold letters represent column vectors of $\boldsymbol{X}$ containing the $k$th and $l$th variables, while suffix $T$ denotes a transpose. The eq 3 is slightly different compared to the previous one (2), because for the calculation of $s$ in eq 3, the division by $n$ rather than by $(n-1)$ is applied, as it is in conventional statistics. Then, the correlation coefficient of two sets of numbers is equal to the scalar product of the normalized mean centered vectors.[3] The vectorial expression gives a geometrical interpretation of the correlation coefficient. In eq 3 $r$ equals to $\cos\varphi$, where $\varphi$ represents the angular distance between the mean centered vectors as it is shown in Figure 1(a).

For the original matrix $\boldsymbol{X}$, the correlation coefficients among all the variables can be calculated simultaneously through matrix operation. The column-standardized matrix $\boldsymbol{X}_\mathbf{s}$ is obtained by subtracting the mean value of its respective column to each element of the matrix $\boldsymbol{X}$ and dividing the difference by the standard deviation of the column elements. Then, the correlation matrix among all the variables is as follows:

$$\boldsymbol{R}_\boldsymbol{n} = 1/n\boldsymbol{X}_\boldsymbol{s}^\boldsymbol{T}\boldsymbol{X}_\boldsymbol{s} \qquad (4)$$

$\boldsymbol{R}_\boldsymbol{n}$ is a symmetric matrix whose main diagonal elements are equal to 1.

Using the same data matrix $\boldsymbol{X}$, one can define the two following equations which are similar to those shown at (1) and (3), respectively.

$$\mathrm{cov}r(x_o,x_p) = \frac{1}{m-1}\sum_{i=1}^{m}(x_{o,i} - \langle x_o\rangle)(x_{p,i} - \langle x_p\rangle) \qquad (5)$$

$$rr(x_o,x_p) = \frac{(\boldsymbol{X}_\mathbf{o} - \langle x_o\rangle)^T(\boldsymbol{X}_\mathbf{p} - \langle x_p\rangle)}{||\boldsymbol{X}_\mathbf{o} - \langle x_o\rangle||\ ||\boldsymbol{X}_\mathbf{p} - \langle x_p\rangle||} = \cos\omega \qquad (6)$$

In the above equation, the operations in our $\boldsymbol{X}$ matrix (the mean centered vectors) are carried out on rows instead of columns. Subindices $o$ and $p$ refer to rows of "objects" instead of columns of "variables". In this case $\omega$ represents the angular distance between the mean centered vectors of two objects (Figure 1b) that we call the **"mean angular distance among objects**" (**MADO**). Let us show a simple example: if we have two objects in a three-dimensional space whose coordinates are $\mathbf{x_1} = (1,3,5)$ and $\mathbf{x_2} = (-1,3,7)$, then the averages of the vector elements are $x_{1av} = (1+3+5)/3 = 3$ and similarly $x_{2av} = 3$. The mean centered vectors are $\mathbf{v_1} = \mathbf{x_1} - x_{1av} = (1-3,3-3,5-3) = (-2,0,2)$ and $\mathbf{v_2} = (-4,0,4)$. Finally, the cosine of the mean angle between $\mathbf{v_1}$ and $\mathbf{v_2}$ is $\cos\omega = 1$. It should be noted that $\mathbf{v_1}$ and $\mathbf{v_2}$ are collinear while $\mathbf{x_1}$ and $\mathbf{x_2}$ are not.

The same way of operation on $\boldsymbol{X}$ that has been used to calculate the covariance or the correlation between the variables, that is "column-wise", can be performed "row-wise".[8] Algebraically, the same result is obtained if the calculation is performed row-wise on $\boldsymbol{X}$ or column-wise on $\boldsymbol{X}^\mathbf{T}$.

It is worth remarking that the covariance and the correlation refer always to the relationships between two variables, but here the relationships between two objects will be obtained by doing the operations on $\boldsymbol{X}^\mathbf{T}$. It is also possible to obtain the MADO through a matrix operation with the following expression

$$\boldsymbol{R}_\boldsymbol{m} = 1/m\boldsymbol{X}_\boldsymbol{r}\boldsymbol{X}_\boldsymbol{r}^\boldsymbol{T} = 1/m\boldsymbol{X}_{\boldsymbol{TS}}{}^\boldsymbol{T}\boldsymbol{X}_{\boldsymbol{TS}} \qquad (7)$$

where $\boldsymbol{X}_\boldsymbol{r}$ is the row-standardized matrix of $\boldsymbol{X}$ and $\boldsymbol{X}_{\boldsymbol{TS}}$ is the column-standardized of $\boldsymbol{X}^\boldsymbol{T}$. In this calculation $rr$ becomes the elements of the $\boldsymbol{R}_\boldsymbol{m}$ matrix. It should be noted that while $\boldsymbol{R}_\boldsymbol{n}$ is a symmetric matrix of size $m \times m$ representing all correlations among $m$ variables, $\boldsymbol{R}_\boldsymbol{m}$ is a symmetric matrix of size $n \times n$ because it relates to $n$ objects. The first column of $\boldsymbol{R}_\boldsymbol{m}$ represents the cosine function of the angles of object

**Table 1.** Relationship between $\boldsymbol{R_m}$ and Neighborhoods for the SOM Trained with Points of a Surface of a Sphere

| objects location | no. of objects in the place | av of the $\boldsymbol{R_m}$ elements | variances of $\boldsymbol{R_m}$ elements | av of the $\boldsymbol{R_m}$ elements among neurons | av of $rr(x_o,x_p)$ among central cell and neighborhoods |
|---|---|---|---|---|---|
| central cell: row 7, column 13 | 5 | 1.000 | 0.0032 | | |
| first neighborhood | 13 | 0.9343 | 0.067 | 0.9727 | 0.9870 |
| second neighborhood | 23 | 0.8252 | 0.179 | 0.9426 | 0.9704 |
| third neighborhood | 27 | 0.6982 | 0.310 | 0.6322 | 0.7883 |

1 with the rest of them, the second one relates to the angles of object 2 with the rest of the objects, and so on. The elements of the main diagonal of $\boldsymbol{R_m}$ are equal to 1; these values represent the *cosine* of the angle of each object with itself.

## CLASSIFICATION OF OBJECTS AND KOHONEN NEURAL NETWORK

To establish the connection between several of the Kohonen neural network features and the concepts described above, the Kohonen artificial neural network is briefly described. For detailed descriptions, the reader is addressed to specialized literature.[4,5] Basically, the network architecture consists of a layer of assembled neurons, which in the simplest configuration, is a rectangle of $p$ rows and $q$ columns of them. Thus, each neuron represents one $m$-weight column, which is perpendicular to the $p \times q$ rectangle. The neurons are mapped in a $p \times q$ matrix of cells in the SOM. During the training stage, each object is presented sequentially to all the neurons of the network. This means that each variable of the object feeds only to one specific input weight of each neuron. All the neurons are fed in parallel with the variables of the object. When the training is carried out, the variables of the object $x_{si}$ are compared to the corresponding weights $w_{ji}$ of all neurons ($j=1...p \times q$). The neuron, whose weights are most similar to the variables of the object, is proclaimed as the winner, and its weights are adapted toward the input values. Apart from the weights of the winning neuron, in a prespecified neighborhood of it the weights of neurons are corrected too, albeit for a proportionally smaller amount. Both, the amount of correction and the size of the neighborhood in which the correction around the winning neuron is carried out are shrunk during the training stage. The algorithm that performs the correction of the weights causes the winning neurons to become more and more similar to the objects that "excites" them. When the network is trained with a set of $n$ objects, for instance, the $n$ objects of the matrix $\boldsymbol{X}$, said network is organized in such a way that it finally shows a map where the similarity of the objects are connected with the distance (reverse similarity) among them as it is shown in Figure 3. For classification purpose, this technique is more efficient than linear methods of clustering when mapping objects have strong nonlinear relationships among them.[6,7]

## COMBINING MADO WITH KOHONEN MAPS

Now, it is possible to explore what kind of relationships the objects keep among them in the self-organized Kohonen map. The position of objects, which is plotted in the map of objects of the Kohonen network (the so-called "top map" or SOM), is strongly related to $\boldsymbol{R_m}$. When $\boldsymbol{R_m}$ is calculated for the objects that excite the same neuron in the top map, its elements tend to 1, indicating that all these mean centered vectors have almost the same direction (hereinafter, the elements of the main diagonal are not included in the average value, because there are always equal to 1). That is, the mean centered vectors of objects within a SOM′ cell are almost collinear. For the objects exciting neurons located in the first neighborhood of the previously mentioned neuron, the elements of $\boldsymbol{R_m}$ have average values close to 1 but lower than that of the central neuron. This shows a larger dispersion (angular distance) among the objects. The average values of the elements of $\boldsymbol{R_m}$ successively decrease by repeating the calculations for the second, third, etc. neighborhoods. The same trend of the corresponding calculations can be seen for any neuron selected as the central one.

A numerical example of a $\boldsymbol{R_m}$ calculation on Kohonen map is shown below. The example consisting on *500* points on the surface of a sphere are taken from that given in ref 4, and the *20* × *20* Kohonen map has been recalculated. Five hundred points (objects) placed on the sphere surface are generated randomly. Each object is represented by three variables: *x*, *y*, and z, and according to their values and signs, each object is classified in one of the eight spherical triangles in which the sphere has been divided. After the KANN has been trained, the objects were mapped onto the Kohonen map. The calculation of $\boldsymbol{R_m}$'s for objects belonging to each neighborhood shows that the average of its elements decreases from the central cell to the periphery. Table 1 shows these averages and the corresponding standard deviations. The arithmetic means of $\boldsymbol{R_m}$ elements for a cell selected as the central one (row 7, column 13) and cells of successive neighborhoods can be compared among themselves through their variances and the *F* test, showing that these populations can be considered statistically distinguishable.

This interpretation of a KANN map leads to an interesting possibility of data handling. Usually, the original data to be classified by KANN should be scaled. The scaling of the variables has several advantages: All the variables become dimensionless and ranged between (*0, 1*) or (*−1, 1*), thus facilitating the comparison among them, and, furthermore, the scaling gives the same weight to all variables when distances among objects are calculated. One of these scaling transformations is $x_s = (x - \langle x \rangle)/(x_{max} - \langle x \rangle)$. In the following examples all data have been scaled in this way. It should be noted that this scaling involves a column-centered operation as those described previously for the column standardized matrix $\boldsymbol{X_s}$.

## SIMULATED EXAMPLES

To make the relationship between KANN and *rr* (or $\boldsymbol{R_m}$) more explicit, two simulated examples will be shown. The already obtained Kohonen network trained with the points lying on the surface of the sphere (Figure 2) is used to

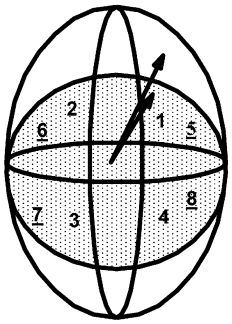


**Figure 2.** Spherical and ellipsoidal surfaces. The numbers identify spherical triangles, the underlining indicates backward faces. Vectors for two objects in different surfaces show that while Euclidean distance between them could be large, the angular distance *rr* may not be so.

evaluate the predictions on a new set of points belonging to a class of objects not previously included in the training stage.

The new set of points, which is employed to test the prediction ability (classification power) of the trained network, is taken from an ellipsoidal surface centered on the sphere. The labels of the points on the ellipsoid bear the same numbers as the corresponding spherical triangles, hence, it is possible to identify the cluster where the points of the ellipsoid will be classified.

Figures 1b and 2 show that points placed on the corresponding spherical and ellipsoidal triangles (bearing the same labels) could have a large Euclidean distance between them but similar (or possibly even equal) central angle.

Although the training was made only with the points of the spherical triangles, the predictions for cluster assignment of the points from the ellipsoid surface sections was 100% correct.

It should be noted that this result is independent of the ellipsoid elongation. Taking into account the expression of the revolution ellipsoid surface

$$\frac{z^2}{b^2} = 1 - \frac{x^2}{a^2} - \frac{y^2}{a^2} \tag{8}$$

it is possible to have a flattened $(b<a)$ or elongated $(b>a)$ ellipsoid or a sphere $(b=a)$. Substituting $b = f \cdot a$ into eq 8 the ratio between $z(x,y)$ variable for any ellipsoid with reference to a sphere, $(z_s)$, is $z/z_s = f$. Then, when the scaling is carried out, the factor $f$ is canceled and the scaled $z$-axis is independent of any elongation. It also shows that the scaling involving the "column center operation" is very important for the relationship between MADO and Kohonen SOMs.

For the analysis of a large volume of high-dimensional data it is possible to take the advantage of this property of the KANN. The conventional calculation of correlation through eqs 2 or 3 gives us the average angle between the variables corresponding to all the objects. Sometimes it is interesting to study a small number of tightly correlated objects within the entire data set. In these cases, it is difficult to analyze the data with the conventional methods. This difficulty arises because the number of possibilities to combine the objects in smaller subsets is very high. We will be back to this point in the real case example, later.

Let us suppose that this time we are interested in studying the relationship between any of two variables, but for different subgroups of objects instead of getting a general indicator as the correlation is. We start with the training of the KANN, by keeping one of these two variables out of the data set. As a result, the SOM of objects based on all variables, except one, is obtained. In the test stage, the data set used for training is modified so that both variables under study are switched. The one that was left out replaces the one used in the training set. Taking into account the scaling of the data, in those cases where similarities between both variables exist, the mean centered vector of the test object will not have a significant change from those of the training set. This implies that the classification (position) of test objects will remain unchanged. Consequently, if the replacement modifies the covr or the *rr* (eqs 5 and 6) only slightly, the object will be classified correctly on the same neuron as before the switching. On the other hand, if the modification is large, the object will be moved to another neuron. The relationship between these two variables can be studied in detail by analyzing objects that have moved from its original position (or those that remained in the same place). For those objects that have not been moved from its original position, we can state that the two variables in question are strongly correlated. For those groups of objects belonging to neighboring cells, that have been moved together to another sector of corresponding neighborhoods, we can assert that the variables used in the second test are weakly correlated to the variables used in the previous one. Additionally, a different subset of objects with different correlation coefficients can be found. Furthermore, because the correlation between variables can be seen on the Kohonen map, we have the possibility to analyze the conditions under which the rest of the variables will influence (improve or worsen) the clustering of the objects.

A second example is introduced to show the above-mentioned procedure. Four clusters, each containing 300 objects (1200 total objects) with five variables, have been generated. The positions of the four centroids in the multidimensional space have been selected purposely in accordance with the distances among them. These centroids were positioned in the range $(-1, 1)$ in four out of five variables. A label with a number was assigned to each centroid to identify the class of the objects belonging to it. The objects were randomly generated around the centroid positions with a dispersion $\pm 0.2$ for each variable. The fifth variable, $v$, in all 1200 objects have been intentionally correlated with the variable $y$ through the relation $v = 1.5(y \pm \mathrm{D1}) \pm \mathrm{D2}$, where D1 and D2 are random dispersions with a range of $\pm 0.5$.

This data set has a resultant correlation coefficient equal to 0.87 between $v$ and $y$. One of the correlated variables was kept out of the training data set, and a KANN has been trained using one-third of the total objects (that is 400 objects, 100 object of each cluster) that contain the remaining four variables. To test the performance of the generated KANN map, another third of 400 new objects was taken from the original set, and the classification was predicted without significant errors. The remaining set of 400 objects was taken from the original collection to analyze the effect of noncorrelation on single (nongrouped) "moving" objects. In this set, on 300 objects belonging to clusters $1-3$, the correlated

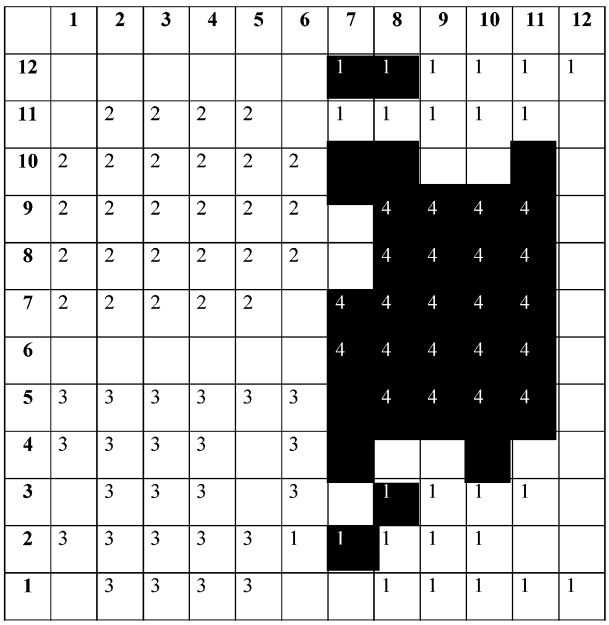


**Figure 3.** The SOM of a 12 × 12 KANN for classification of objects belonging to four different clusters. The numbers identify the clusters. The black area shows the shifting of objects of cluster 4 after changing a variable by another without correlation with the original one. For clusters 1−3, where the replacement kept a correlation of 0.87, their objects remain in the same place.

**Table 2.** Distances among Cluster Centroids

| | C1 | C2 | C3 | C4 |
|---|---|---|---|---|
| C1 | 0 | | | |
| C2 | 1.828 | 0 | | |
| C3 | 1.652 | 1.749 | 0 | |
| C4 | 1.417 | 1.706 | 2.238 | 0 |

variable omitted during the training, $v$, was switched with that one, $y$, used previously.

While in the remaining 100 objects belonging to the cluster 4, random numbers substituted the variable $y$. This means that only for objects belonging to cluster 4, the replaced variable was not correlated with that one for which the KANN was trained. The resulting top map is shown in Figure 3.

The objects in which the replacements have a correlation coefficient of 0.87 (clusters 1−3) still remain in their original area, while the objects of cluster 4 have been dispersed on the black area. From cluster 4, 62% of the objects has been moved far from their original positions. Due to the random nature of the substituted variable, it may well happen that, by chance, some objects can still excite the same neuron as before. Objects belonging to cluster 4 have moved to the neighborhood of the previously nonoccupied cells or cells belonging to cluster 1. The fact that objects of the cluster 4 have only invaded cells of cluster 1 can be explained through the distance among clusters in the multidimensional space. Table 2 shows the distances among cluster centroids, being the distance between cluster 1 and 4 the shortest one.

The resultant correlation between the variable $y$ used in the training and the values randomly generated is 0.14. The correspondence between correlation and the effect on moving objects, which is specific of each application, can be calculated in this example. Enlarging the range of parameters D1 and D2, mentioned previously, allows us to modify the correlation between the variables $v$ (for prediction data set) and $y$ (used in training data set) for all four clusters. The results are shown in Table 3. It should be taken into account that the percentages reported in this table refer only to objects moved to a far distance from their original positions. An additional percentage of objects, the amount of which depends on the value of the correlation coefficient, are moved from their original cell to other areas of the same class.

**Table 3.** Correspondence between Correlations and Movement of Objects

| correlation coefficient | objects moved to previously unoccupied cells | objects moved to areas of another class | percentage of total objects moved outside of its original class (on 400 total objects) |
|---|---|---|---|
| 0.87 | 41 | 0 | 10.25 |
| 0.61 | 95 | 1 | 24 |
| 0.46 | 115 | 2 | 29.25 |
| 0.34 | 124 | 3 | 31.75 |
| 0.26 | 138 | 14 | 38 |

In this table it is possible to see that there is approximately a linear relationship between the dispersion of objects on the Kohonen map from its original area and the correlation between the variables used for training and predictions. This fact demonstrates the connection that exists between correlation (eqs 2 or 3) and MADO (eq 6). In cases where only one variable out of a set of a thousand (or more) is replaced, the relative change made to the objects is very little. In such cases, the probability of a switch in location is small if the number of neurons in the SOM is remarkably smaller than the number of objects in the data set. Thus, a great number of objects for which the two variables show a large difference would not switch places. For these big multivariate systems, this particular use of MADO should be applied carefully.

Taken into account the previous discussions, it should be remarked that $\boldsymbol{R_m}$ matrix can perform many functions which are helpful to evaluate SOMs. It can be calculated for objects, neurons, or a combination of both of them. Suggested potential applications of $\boldsymbol{R_m}$ are presented below.

(a) It can give the information about the differences between SOM points in different directions, by calculating the $\boldsymbol{R_m}$ around a cell, taken as the central one, and all the cells of successive neighborhoods.

(b) It is possible to appreciate the boundaries of clusters indicated by gaps of low *rr* values calculated along a complete row or column of the SOM.

(c) It can determine the tendency of empty cells to be filled with objects of a specific class, calculating $\boldsymbol{R_m}$ among empty cells and their populated neighborhoods.

(d) $\boldsymbol{R_m}$ can appreciate the relative distances between the borders of clusters.

(e) $\boldsymbol{R_m}$ among the input objects and the exited neuron will give the information about how good the approach between them is (that means how good the training stage has been).

(f) $\boldsymbol{R_m}$ can describe the position of new "unknown" objects in the SOM.

(g) $\boldsymbol{R_m}$ can describe the characteristics of empty cells that represent objects not present in the training set.

For the most demonstrative cases, examples are given below.

Case (a) is shown in Figure 4; it is a partial area of Figure 3 showing a MADO surface of the SOM where it is possible

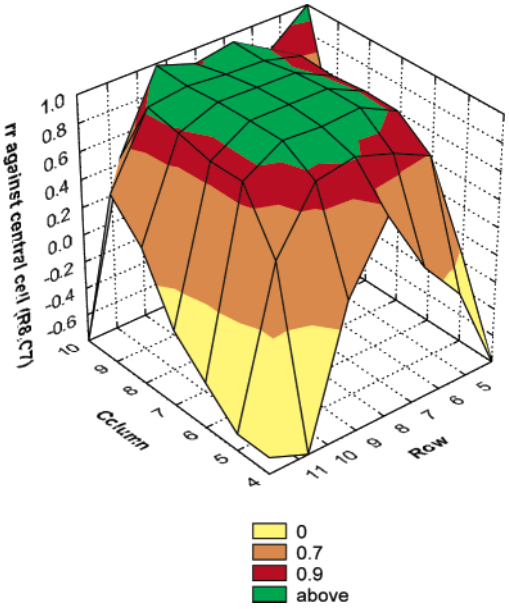


**Figure 4.** Partial area of Figure 3 taking the cell at row 8, column 7 as the central position. Vertical axis shows the decreasing values of *rr* (MADO) from the central cell to successive neighborhoods in different directions.

**A**

| Column boundaries | 1-2 | 2-3 | 3-4 | 4-5 | 5-6 | 6-7 | 7-8 | 8-9 | 9-10 | 10-11 | 11-12 | 12-1 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| MADO | 1.00 | 0.98 | 0.99 | 0.99 | 0.99 | 0.61 * | 0.68 * | 1.00 | 0.99 | 1.00 | 0.38 * | 0.92 |

**B**

| | | | | | |
|---|---|---|---|---|---|
| Row boundary 6-7 | 0.39 | 0.40 | 0.49 | 0.53 | 0.52 |
| Row boundary 5-6 | 0.80 | 0.81 | 0.75 | 0.9 | 0.81 |
| Columns of figure 3 | 1 | 2 | 3 | 4 | 5 |

**C**

| Columns of figure 3 | 8 | 9 | 10 | 11 | 12 | |
|---|---|---|---|---|---|---|
| MADO between Rows 9 and 11 | 0.23 | 0.23 | 0.20 | -0.02 | | |
| Cells of cluster 4 | | | | | -0.03 | 9 |
| | | | | | 0.03 | 8 |
| | | | | | -0.06 | 7 |
| | | | | | MADO Between columns 3 and 5 | ↑ Rows |

**Figure 5.** Examples of $R_m$ functions (b)−(d) described in the text. Parts A to C show $R_m$ calculations of MADO around borders of clusters and empty cells. Asterisks in part A indicate gaps between clusters.

to visualize different approaches to the central cell depending on the arriving direction.

Other functions are shown in Figure 5: parts A, B, and C of it are examples of calculations related to Figure 3. Part A is an example of the case (b) listed before. $R_m$ is calculated between each pair of neighboring cells along the row 9 of the SOM. On this line, there are two gaps corresponding to boundaries of clusters 2 and 4. Part B of Figure 5 refers to case (c); it is a calculation of $R_m$ between the empty cells of row 6 (columns 1−5) and their neighboring lines of cluster 2 and 3. In this case, the result shows that the empty cells have a bigger tendency to receive objects of cluster 3 than those of cluster 2, because of the shorter distance from empty cells to cluster 3. Finally, part C of the figure shows an example of case (d). Here, $R_m$ is calculated for the borders of cluster 4 limiting with clusters 2 and 1. The MADO between cluster 4 and 2 is bigger than the one between cluster 4 and 1. As it was previously mentioned the clusters were built in this way, being the distance between cluster 4 and 1 the shortest one.

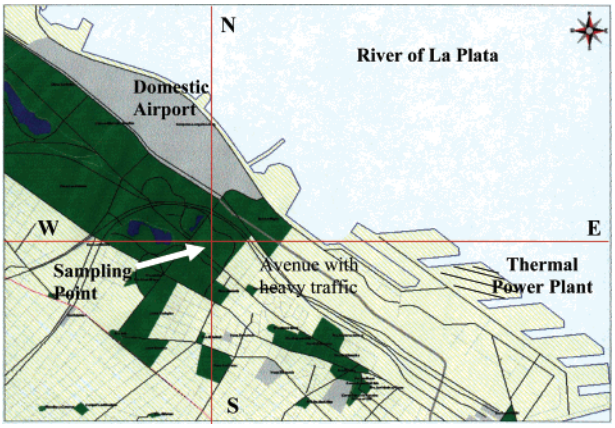


**Figure 6.** Map of the sampling point and main emitting sources.

## APPLICATION TO A REAL PROBLEM

The method described previously has been used as complement of other methods in order to prove its performance on a real case. The data set has been obtained from an air quality monitoring campaign. Five gaseous compounds (NO, $NO_2$, CO, $SO_2$, and $O_3$) plus $NO_x$ and suspended particulate matter with an aerodynamic particle diameter less than 10 $\mu$m ($PM_{10}$) were the chemical variables. Air pollutants were measured minute by minute during 45 days. The $PM_{10}$ and 11 meteorological variables were obtained hourly during the same period of time. The sampling point was located at Palermo Park in Buenos Aires City, Argentina. The most important sources around that point are as follows: the domestic airport, a thermal power plant (with a low dispatch during the sampling period), and an avenue with heavy traffic. The effects of these emitting sources have been detected through more conventional methods. The zone of sampling is shown at Figure 6. More detailed information about this sampling place and methods will be published elsewhere.

For air pollutants, hourly average concentrations were taken as the average of the last 10 min of each hour to match hourly values of meteorological variables. This method is used for measuring the meteorological variables, then both kinds of variables have been obtained in the same way.

For this exercise, we will focus the analysis on the primary gaseous pollutants. Therefore, only concentration of NO, CO, and $SO_2$ will be considered together with the meteorological variables interacting with them, that is, wind direction (WD) and wind intensity (WI), atmospheric stability (S), cloudiness (CL), and height of the lowest cloud layer (HLC).

## SCALING OF THE VARIABLES AND CLASSIFICATION

The frequency distributions of the concentration of pollutants usually have a log-normal shape, as it happens for the gases considered here. To take into account a similar

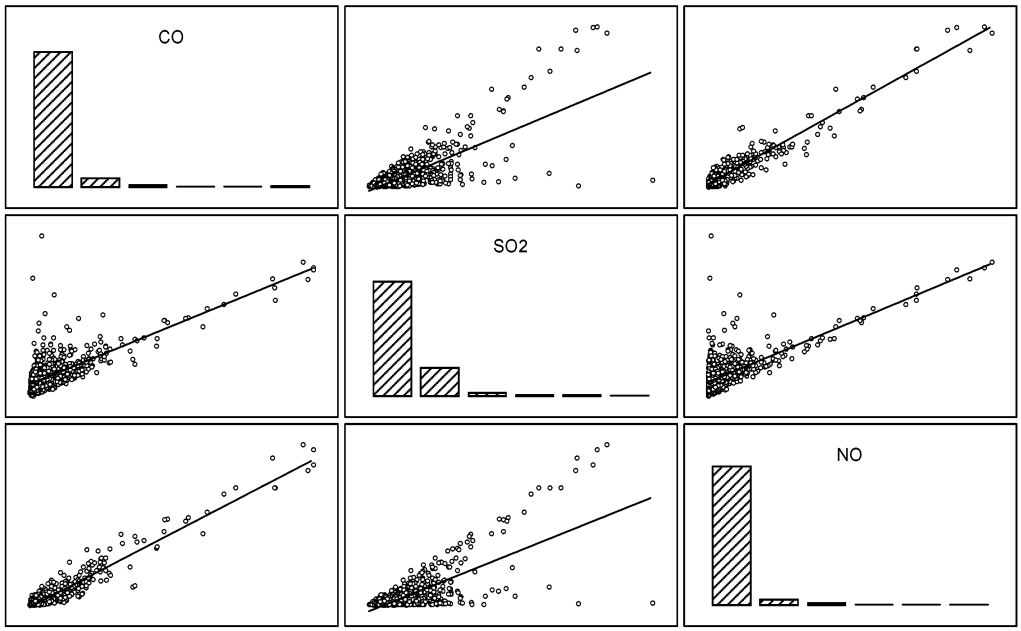


**Figure 7.** Existent correlation among primary gaseous pollutants.

discrimination among the lowest and the highest concentration values, the scaling involving the column center operation has been slightly modified to $x_s = (x - \langle x \rangle)/(\langle x \rangle - x_{min})$. The wind intensity has been scaled according to the previous examples in the range (−1,1). Cloudiness and HLC meteorological measurements (whose ranges are 0−10) have been divided by 10. The wind direction has been classified in five classes: northeast (NE), northwest (NW), southeast (SE), southwest (SW), and calm (no direction registered). Atmospheric stability has been classified in three categories: stable (S), unstable (U), and neutral (N) associating to them the values 1, −1, and 0, respectively. The CO was classified in three levels: low (less than 0.77 ppm), medium (≥0.77 ppm and < 1.5 ppm), and high (≥1.5 ppm).

To classify the objects we choose three relevant variables: the CO level, the stability, and the wind direction. By combining these levels, a total of 45 classes ($3 \times 3 \times 5$) of objects would be possible, but five of these classes have no objects at all. Eleven more classes having a percentage of objects less than 0.4% have been included in the training set, despite their low population to achieve a good training.

Figure 7 is a first screening of the data set, where the correlation plots among primary gaseous pollutants are shown. In the figure it is noted that there is a good correlation between NO and CO, which leads to the conclusion that the monitoring site is primarily impacted by the same source, namely traffic.[9] However, the correlation between $SO_2$ and CO or NO is not so evident. The points show something like two branches, and the regression between both variables traces a line whose slope approximately bisects the angle between the two branches. It will be especially interesting to investigate why $SO_2$ reaches high concentrations for low values of NO and CO.

We train a Kohonen net in the same way that in the previous examples, keeping out of the data set the $SO_2$ concentration. To train the net, 494 objects have been used, and the rest (472 objects) were kept to check the training. After the training, the NO column variable has been replaced with the $SO_2$ variable, and new predictions have been done to compare the positions of objects before and after the replacement. Due to the fact that in this data set there exists a significant correlation among the variables, most of the objects have remained in their original position. However, it is possible to detect two important groups of movements: (1) the movement to cells (5,7) and (4,6) of objects coming from two different zones of the net and (2) the movement to cells (20,8) and (20,9) of objects coming from three different zones. To be brief we will discuss in detail only the first case. It will help to see, after training, the distribution of the weights of neurons for the CO and NO variables in the Figure 8 (a),(b).

The cells (5,7) and (4,6) are in a flat region of a valley corresponding to low concentrations of CO, whereas the NO values are on the top of a low hill that emerges from a flat zone. After the replacement of NO by $SO_2$, object groups move from areas of lower concentrations to this place, that means that this class of objects has a relative high value of $SO_2$ which does not correlate with the general trend of NO concentration. It should be remembered that levels of these three variables and not only the NO level determine the class. Now, we can analyze the levels for the rest of the variables in which these particular movements of objects occur. Figure 9 (a),(b) shows the scaled levels of the chemical variables and those of the meteorological ones.

Figure 9(a) shows high values of $SO_2$ relative to NO and CO concentrations. Figure 9(b) shows that this fact always occurs with the same atmospheric stability class (neutral), and mostly when wind direction corresponds to SE. Likewise, cloudiness takes values of 8 or is in the range 0−3. Wind intensity is irrelevant because it takes any value. Finally, the HLC takes the value 9 or wide range between 2 and 6. The SE wind direction indicates that wind is blowing neither from the airport nor from the heavy traffic road, confirming that these relative high concentrations of $SO_2$ do not came from these sources. These objects correspond to 87.2% of nightly hours. Several of them are grouped in periods of 3−5 nightly hours, and these periods agree with the persistence in the wind direction. Due to these considerations, and furthermore,

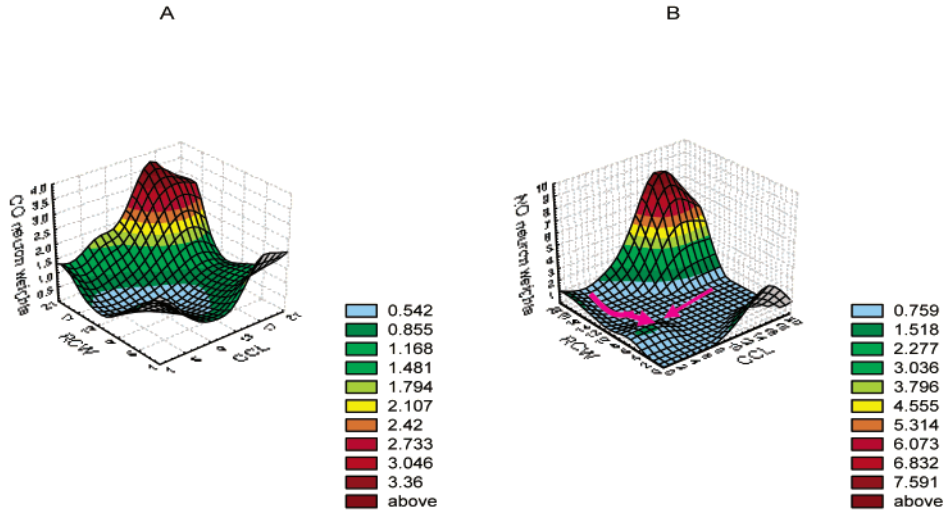


**Figure 8.** Neuron weights of the Kohonen top map for (a) CO variable level and (b) NO variable level. The arrows show the shifting of objects from two sectors to cells (4,6) and (5,7).

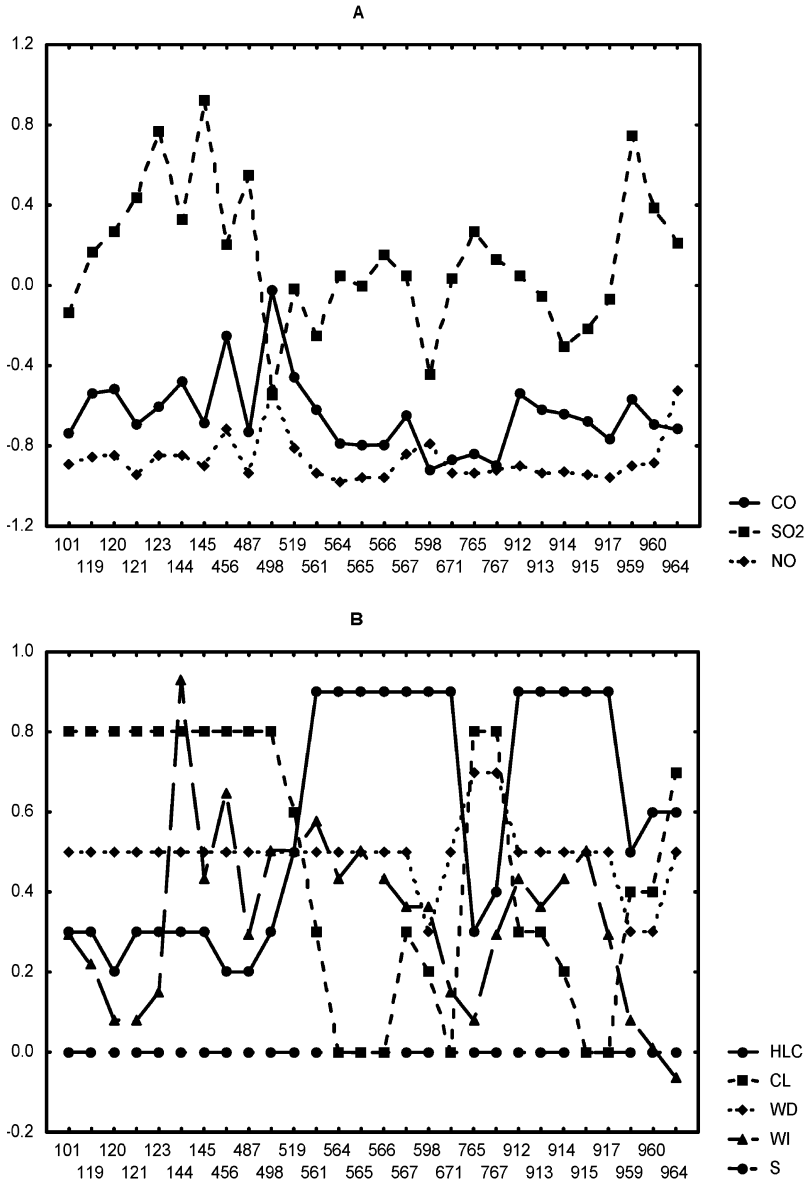


**Figure 9.** Scaled variables of objects moved to cells (4,6) and (5,7). (a) Chemical variables. (b) Meteorological variables. The numbers in horizontal scale identify objects; variables in the range −1,1 are represented in the vertical scale.

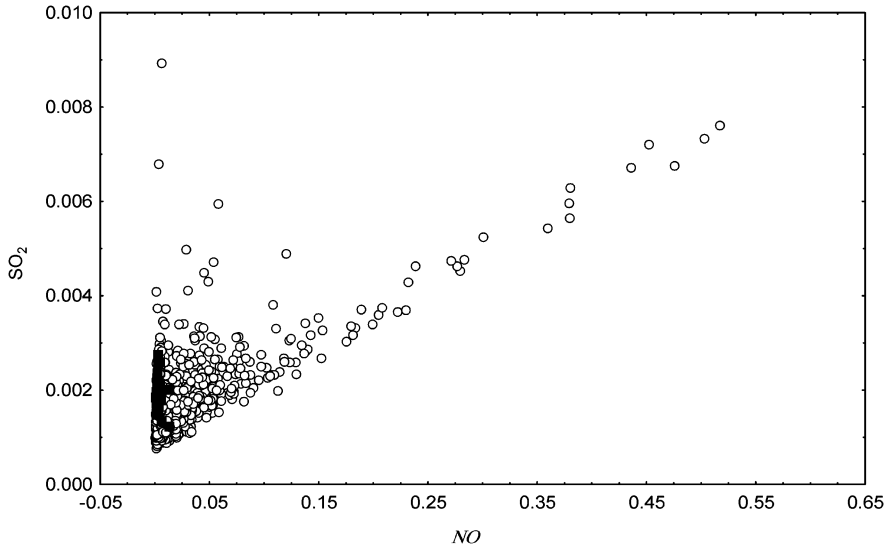


**Figure 10.** $SO_2$ and NO axes are measured concentration in $\mu g/m^3$ of air. Shaded points differentiate objects moved to cell (14,13) from the rest.

the relative high levels of $SO_2$, an additional $SO_2$ emission source may be postulated. Further samplings are required to confirm this hypothesis.

Figure 10 shows the correlation between $SO_2$ and NO concentrations for the data set. If we located the moving objects in this graph by shading the points, we observe that they all correspond to one of the two previously identified branches. The graph suggests that the moved objects mostly correspond to an independent source of $SO_2$. Figure 10 shows that we were able to identify through the MADO formalism, those objects diverging from the general trend.

## CONCLUSIONS

Starting from basic equations of statistics, we had introduced a new algorithm that we called the mean angular distance among objects, MADO. Its calculation for a group of n objects depending on m variables can be expressed as a matrix operation. This matrix, called $\boldsymbol{R_m}$, has been related to the organization of the objects in a Kohonen map obtained after its training. We have analyzed numerically simulated examples showing some consequences of the definition of $\boldsymbol{R_m}$ for objects with linear relationships among variables. Taking advantage of this knowledge, a procedure to investigate relationships between variables (total or partial correlation or non correlation) using a Kohonen net has been described. A real case is analyzed for objects of a big environmental data set belonging to an air quality monitoring campaign. In this case, we were able to identify a subgroup of them with two variables having a stronger different correlation than for those of the rest of the objects. The analysis of all the variables of this subgroup allows us to postulate the existence of another emission source, different than those with permanent emissions, previously found with standard methods. It is worth noting that, unlike the simpler simulated examples, in this data set several variables have nonlinear relationships.

Kohonen networks are currently used for classifications tasks, but their capabilities may be further extended through this algorithm and other new applications can be designed.

## ACKNOWLEDGMENT

This work was developed under the frame of the Bilateral Cooperation Agreement supported by SECYT of Argentina and MSZS of Slovenia. Project number ES/PA00/E03.

## REFERENCES AND NOTES

(1) Tenenbaum, J. B.; de Silva, V.; Langford, J. C. A Global Geometric Frameworks for Nonlinear Dimensionality reduction. *Science* **2000**, *290*, 2319−2323.
(2) Roweis, S. T.; Saul, L. K. Nonlinear Dimensionality Reduction by Locally Linear Embedding. *Science* **2000**, *290*, 2323−2326.
(3) Massart, D. L.; Vandeginste, B. G. M.; Buydens, L. M. C.; De Jong, S.; Lewi, P. J.; Smeyers-Verbeke, J. *Handbook of Chemometrics and Qualimetrics. Part A*; Elsevier: Amsterdam, 1997; p 240.
(4) Zupan, J.; Gasteiger, J. *Neural Networks in Chemistry and Drug Design*, 2nd ed.; Wiley-VCH: Weinheim, 1999; Chapter 6.
(5) Kohonen, T. *Self-Organizing Maps*; Springer-Verlag: Berlin, 1995.
(6) Ruisanchez, I.; Potokar, P.; Zupan, J. Classification of Energy Dispersion X-ray Spectra of Mineralogical samples by Artificial Neural Networks. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 214−220.
(7) Magallanes, J; Vazquez C. Automatic Classification of Steels by Processing Energy-Dispersive X-ray Spectra with Artificial neural Networks. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 605−609.
(8) Massart, D. L.; Vandeginste, B. G. M.; Buydens, L. M. C.; De Jong, S.; Lewi, P. J.; Smeyers-Verbeke, J. *Handbook of Chemometrics and Qualimetrics. Part B*; Elsevier: Amsterdam, 1997; p 47.
(9) Bogo, H.; Gómez, D. R.; Reich, S. L.; Negri, R. M.; San Román, E. Traffic Pollution in a Downtown Site of Buenos Aires. *Atmos. Environ.* **2001**, *35*, 1717−1727.

CI034062V