Acta Scientiarum

http://www.uem.br/acta
ISSN printed: 1806-2563
ISSN on-line: 1807-8664
Doi: 10.4025/actascitechnol.v34i1.10892

Self-organizing maps as a chemometric tool for aromatic pattern recognition of soluble coffee

Evandro Bona[1*], Rui Sérgio dos Santos Ferreira da Silva[2], Dionísio Borsato[3] and Denisley Gentil Bassoli[4]

*[1]Programa de Pós-graduação em Tecnologia de Alimentos, Universidade Tecnológica Federal do Paraná, 87301-006, Cx. Postal 271, Campo Mourão, Paraná, Brazil. [2]Departamento de Ciência e Tecnologia de Alimentos, Universidade Estadual de Londrina, Londrina, Paraná, Brazil. [3]Departamento de Química, Universidade Estadual de Londrina, Londrina, Paraná, Brazil. [4]Companhia Iguaçu de Café Solúvel, Cornélio Procópio, Paraná, Brazil. *Author for correspondence. E-mail: ebona@utfpr.edu.br*

ABSTRACT. The electronic nose (EN) is an instrument very used for food flavor analysis. However, it is also necessary to integrate the equipment with a multivariable pattern recognition system, and to this end the principal component analysis (PCA) is the first choice. Alternatively, self-organizing maps (SOM) had been also suggested, since they are a nonlinear and reliable technique. In this study SOM were used to distinguish soluble coffee according to EN data. The proposed methodology had identified all of the seven coffees evaluated; in addition, the groups and relationships detected were similar to those obtained through PCA. Also, the analysis of network weights allowed gathering the e-nose sensors into 4 groups according to the behavior regarding the samples. Results confirm SOM as an efficient tool to EN data pos-processing, and have showed the methodology as a promising choice for the development of new products and quality control of soluble coffee.

Keywords: self organizing maps, soluble coffee, electronic nose.

Utilização dos mapas auto-organizáveis como uma ferramenta quimiométrica para o reconhecimento de padrões aromáticos de café solúvel

RESUMO. Ao nariz eletrônico, um instrumento muito utilizado na análise de voláteis alimentares, é necessário acoplar uma interface multivariada para reconhecer padrões. O mapa auto-organizável (SOM), uma técnica robusta e não-linear, vem sendo sugerido como alternativa à análise de componentes principais (ACP). Neste trabalho, foi empregado o SOM para avaliar café solúvel a partir dos dados de um nariz eletrônico. Por meio de testes estatísticos, foi avaliada a confiabilidade dos resultados que também foram comparados aos obtidos pela ACP. O SOM identificou os sete cafés avaliados, além disso, os grupos formados e as relações entre eles foram semelhantes às obtidas pela ACP. A análise dos pesos da rede permitiu, também, reunir os sensores em quatro grupos de acordo com o comportamento em relação às amostras. Os resultados confirmam a potencialidade da metodologia para a análise multivariada dos dados de um nariz eletrônico. E, confirmam o SOM como uma metodologia para o controle da qualidade aromática de café solúvel.

Palavras-chave: mapas auto-organizáveis, café solúvel, nariz eletrônico.

Introduction

For coffee, the aroma is the most important criterion in quality assessment and one determinant parameter for the consumer choice (FARAH et al., 2006). The coffee aroma is made up by numerous compounds with varied functional groups, and this composition depends on factors such as species and variety, conditions of growth and harvest, storage, roasting intensity and type of roaster, besides other process conditions (MELLO; TRUGO, 2003). From the roasted coffee, extraction columns are used in the industrialization of soluble coffee to obtain the water-soluble coffee extract (CLARKE, 2001). The research for the approximation of the soluble coffee aroma to the brewed coffee has been constant and, in general, an important goal of the process of increasing the intensity of the aromas in the instant coffee or even of its aromatization. In recent decades, the electronic nose has been frequently used in analysis of volatiles in foods (DEISINGH et al., 2004; GHASEMI-VARNAMKHASTI et al., 2010) and even in roast and ground coffee (PARDO; SBERVEGLIERI, 2002). It is up to the equipment to emulate the functioning of a human nose, but is also necessary a multivariate and "intelligent" system to play a role

similar to the cerebral cortex for aromatic pattern recognition. One of the most traditional multivariate methods (CRAVEN et al., 1996) is the principal component analysis (PCA), but some disadvantages of the method are well known (MELSSEN et al., 2006). First, the data need to be described by linear combinations; hence, non-linear systems will not be well represented. In addition, the visualization power of these transformation methodologies deteriorates considerably if the number of relevant factors in the multivariate space remains high after a PCA analysis. The artificial neural network (ANN) is a set of techniques of nonlinear mapping, robust and tolerant to punctual discrepancies (HUANG et al., 2007). Among the several types of ANN, the self-organizing maps are highly recommended for data mining (MELSSEN et al., 2006). In a self-organizing map (SOM) or feature map, the artificial neurons are, generally, placed in nodes of a two-dimensional grid. A SOM is therefore characterized by a topological map of input patterns in which the spatial localizations of the neurons on the grid are indicative of intrinsic statistical characteristics contained in the input patterns. By presenting a non-linear characteristic, a SOM can be seen as a non-linear generalization of the principal component analysis, with the advantage of being a simpler approach from the mathematical point of view (HAYKIN, 2001). In the food industry, the "intelligent" computational models have become very important and used in all steps, from the production to the market segmentation (CORNEY, 2002). In this context, several applications for the SOM have been found, for instance: study on the ripening of the Port Salut Argentino cheese (VERDINI et al., 2007); characterization of strawberry varieties (BOISHEBERT et al., 2006); identification of binary blends of Italian olive oils (MARINI et al., 2007); characterization of rosemary samples according to their geographical origins (TIGRINE-KORDJANI et al., 2007). However the results obtained from neural networks are always subjected to variability due to the sensitivity to initial conditions, and convergence to a local minimum through the learning algorithm. For a more rational use of the SOM, it is important the employment of statistical tools to evaluate both the preservation of topology inherent to the data, the variability of the quantization error and the stability of neighborhood relationships (BODT et al., 2002).

This study aimed to apply the SOM for aromatic pattern recognition of soluble coffee through data obtained by an electronic nose.

Material and methods

Soluble coffee and Electronic nose

Seven types of soluble coffee (Table 1) were tested in this study. The experimental samples were produced in the Companhia Iguaçú de Café Solúvel (Cornélio Procópio, Paraná State, Brazil). The other samples are commercial soluble coffee purchased in Brazil and England. It was used the Portable Electronic Nose Type PEN2 (Airsense Analytics, German), which monitors the variation of conductance in each of its sensors, according to the passage of gas flow with the sample. Then this variation is compared to that obtained with the room air generating a dimensionless signal. The electronic nose used has ten sensors of partial specificity, here named S1 to S10. A detailed description of the sensors is available in Zhang et al. (2007). It was analyzed 2.8 g of soluble coffee placed in specific vials for the headspace analysis. Each coffee (Table 1) was evaluated in three genuine replicates. The operational parameters of the equipment were: 1.0 s interval between the samples; 60 s of measurement time; 120 s for purging the sensors; 400 mL $min.^{-1}$ of flow injection.

Table 1. Short description of the samples of soluble coffee analyzed.

Sample*	Coffee	Description
1 2 3	1	Experimental sample of medium roast. Arabica coffee classified as a hard bean from the southern region of Minas Gerais State.
4 5 6	2	Experimental sample of medium roast Arabica coffee classified as a hard bean from the northern region of Paraná State.
7 8 9	3	Experimental sample of medium roast. Conilon coffee from the Rondônia State.
10 11 12	4	Brazilian commercial sample of dark roast, consisting of conilon and arabica coffee.
13 14 15	5	Brazilian sample; commercial; unknown composition. Dark intensity of roast, sensorially estimated.
16 17 18	6	English sample; commercial; unknown composition. Medium intensity of roast, sensorially estimated.
19 20 21	7	English sample; commercial; unknown composition. Medium/light intensity of roast, sensorially estimated.

*Samples in triplicate.

Integration and autoscaling

After acquiring the data, the area under the curve generated by each sensor was numerically integrated, forming an input vector with ten values for each sample. Before being fed into the neural network, the input vectors were autoscaled (mean equal to zero, and unit variance) to ensure the same metrics for all the variables (HAYKIN, 2001).

Self-Organizing Map (SOM)

In this study, it was implemented the two-dimensional self-organizing map (SOM) algorithm proposed and discussed in details in Haykin (2001). The goal of a self-organizing map is to transform a pattern of incident signal with arbitrary dimension into a two-dimensional discrete map and perform this transformation adaptively in a manner topologically ordered. Each variable is represented by a weight level that after the training phase can be used in the attempt to extract rules for the groups formed. The algorithm responsible for the formation of the SOM initializes randomly the synaptic weights, thus, no prior organization is imposed to the map of features. After the initialization, there are three essential processes: competition, cooperation and synaptic adaptation (MELSSEN et al., 2006).

In the competitive process, for each input pattern, the neurons of the grid calculate their respective values from a discriminant function (Euclidean distance). The neuron with the smallest Euclidean distance is called winner neuron, or centroid, for the input pattern considered. An important point of the SOM algorithm is the reduction in the neighborhood size along the training. At the beginning, the function should be wide enough to cover the entire map, but as the training continues the width should be reduced to ensure the formation of specialized topological regions. The way to implement the reduction is by the exponential decay (HAYKIN, 2001).

During the adaptive process, it is needed changes in the synaptic weight vector in relation to the input patterns. The correction process occurs through the modification in the Hebb's postulate of learning. The adaptation process can be divided into two steps: ordination and convergence. At the ordination stage, the learning rate should be higher to ensure to the topological map, initially disordered, has greater adjustments in their synaptic weights. In the convergence stage, a fine-tune of the map is performed using a lower learning rate. The decrease in the learning rate can also be done by the exponential decay (HAYKIN, 2001).

After the training, the result can be seen through a topological map and weights map. In the first, each input pattern is placed on the map of features according to the position of the winner neuron of the sample considered. The function of the topological map is to ease the visualization of groups and also the neighborhood relationship between the formed groups. Groups close share some similarity, as well as, the greater the distance, the greater the behavior difference. The weights map is a contour plot for each level of weights (MELSSEN et al., 2006). Together with the topological map, it is possible to extract behavior rules for each group formed and infer about the influence of each variable on the obtained result.

Topology preservation

Considering that the SOM makes a representation of a multi-dimensional space into only two dimensions, it should be evaluated if the topological properties of the data were preserved during the transformation. For a qualitative evaluation of the topology preservation, Demartines and Hérault (1997) proposed using the graph *dy-dx*. In this scatterplot, is represented on the axis x (dx) the Euclidean distance between the positions defined by the input vectors for all possible pairs between the existing samples. And in the axis y (dy), the Euclidean distance between the centroids of each one of the possible pairs. If the topology is preserved, the points should be distributed along a line. In order to evaluate the linearity degree between the points of the graph *dy-dx*, in this study, we also calculated the Pearson correlation coefficient.

Reliability of the Self-Organizing Map

The goal of this step is to verify, through classical statistical inference, such as the coefficient of variation and test of hypothesis, if the neighborhood relationships between the groups are due to intrinsic properties of the data or just random. According to Bodt et al. (2002), the two main sources of variability for the SOM are the initialization and sampling. The training of several maps, always using the same data set, would allow only the analysis of variability through the process of random generation of initial weights. Thus, it is necessary to use a sampling technique that permits a different data set for each new training. With the bootstrap methodology is possible to use samples generated artificially with the same empirical distribution (mean and standard deviation) of the original data. Thereby, using the bootstrap technique, each SOM had been trained one hundred times; and in each repetition the initial weights and samples were different, allowing evaluating the two main sources of variability. After each training repetition, it was evaluated the mean quantization error (MQE) of the map in relation to the K input patterns, according to the equation (1). QE_k is the quantization error for each sample, which is defined as the Euclidean distance between the input vector and the weight vector of the respective winner neuron.

$$MQE = \frac{1}{K}\sum_{k=1}^{K} QE_k \quad (1)$$

With the MQE of each repetition, the coefficient of variation was estimated, which permitted to assess the stability of the representation of the data by SOM. Besides that, according to Bodt et al. (2002), a plot of the coefficient of variation against the number of used neurons provides an empirical way to select the most adequate size of the SOM for the analyzed data. According to the same author, the increase in the number of used neurons in each dimension leads to an increase in the coefficient of variation. So, the greater the number of units in the map the greater the possibilities of samples distribution, leading to instability in the results.

In order to evaluate the reliability of the groups and neighborhood relationships it is necessary to define a new function (2). In this equation, the radius r is defined for a square neighborhood function. Being neighbors means that two input patterns $\mathbf{x}^i$ and $\mathbf{x}^j$ are projected in the topological map into centroids located within the region of radius r. The index b indicated that the function (2) must be evaluated for each repetition of training. It is possible that in a given repetition, two samples are neighbors, and not in another. This variability is precisely the aim of evaluation of the neighborhood reliability.

$$NEIGH^b_{i,j}(r) = \begin{cases} 0 & \text{if } x^i \text{ and } x^j \text{ are not neighbors} \\ 1 & \text{if } x^i \text{ and } x^j \text{ are neighbors} \end{cases} \quad (2)$$

After the evaluation of neighborhood relationships in all the repetitions, it is calculated a mean value for the function (2). This mean value, here represented by $STAB_{i,j}(r)$, is the stability of the neighborhood relationship between a pair of samples $\mathbf{x}^i$ and $\mathbf{x}^j$ within the radius r considered. If the pair considered is neighbor in all the repetitions $STAB_{i,j}(r) = 1$ (one), otherwise, $STAB_{i,j}(r) = 0$ (zero).

The next step is to accomplish a statistical significance test for the value of $STAB_{i,j}(r)$. The number of neighbors in a region with radius r of a two-dimensional map is equal to $\nu = (2r+1)^2$. N being the total number of neurons in the map, the probability of a given pair of observations be neighbor at random is ν/N. As the neurons located in the ends of the map do not have the same number of neighbors, Bodt et al. (2002) suggest using the average number of neighbors. To construct a hypothesis test it was used a binomial distribution with probability of success ($\mathbf{x}^i$ and $\mathbf{x}^j$ are neighbors) equal to ν/N, in B independent repetitions. The null hypothesis (H_0) considers that the neighborhood relationship between $\mathbf{x}^i$ and $\mathbf{x}^j$ is at random. The alternative hypothesis (H_1) supposes that the pair can be statistically considered neighbor or not within a predetermined radius. In order to perform the calculation, it is necessary to approximate the binomial distribution using a Gaussian distribution. To validate the approximation, two criteria have to be met: (i) the number of (B) repetitions is great enough (> 30); (ii) the probabilities of success or failure are not close to zero or, in this case, $B\nu/N > 10$ and $B\left(1-\frac{\nu}{N}\right) > 10$. Using the second condition and determining the amount of repetitions, $B = 100$ in this study, it was possible to build an interval of valid size for the neighborhood, as a function of the total number of neurons of the map (3), as well as, an interval of ideal radius to evaluate the stability of neighborhood relationships.

$$\frac{N}{10} < \nu < \frac{9N}{10} \quad (3)$$

After the considerations above, it is possible to establish the interval for the acceptance of H_0 with 95% confidence (BODT et al., 2002), as presented in equation (4). If the value of $STAB_{i,j}(r)$ is higher than the upper limit of this interval, H_1 is accepted and the pair of samples is considered neighbor, at 2.5% level. If the value is less than the lower limit, H_1 is also accepted, and the pair is considered non-neighbor, at the same significance level.

$$B\frac{\nu}{N} \pm 1{,}96\sqrt{B\frac{\nu}{N}-\left(1-\frac{\nu}{N}\right)} \quad (4)$$

For an analysis that considers all possible pairs, two criteria can be examined. First, in the distribution used for the hypothesis test, the random probability of accepting H_1 is 5%. Therefore, if the percentage of neighbor or non-neighbor, statistically valid, for a given r is lower than 5%, the analysis can not be considered significant. The second alternative is the construction of cumulative probability plots that enable a qualitative analysis of significance. A discrepancy between the cumulative probability plot of the SOM and of the binomial distribution is an indicative that the neighborhood relationships built are not casual. For a quantitative analysis, in this study, the χ^2 test was employed to test the relationship between the distributions.

Principal Component Analysis (PCA)

This analysis consists of the linear combination of original variables that produce orthogonal principal components (PC). According to Hair Junior et al. (2005), for the choice of the adequate number of PC, there are several criteria, among them: (i) eigenvalue criterion – only the PC with eigenvalue higher than 1 are considered; (ii) percentage variance criterion – the number of PC chosen must be enough to explain at least 95% of data variability; (iii) scree test criterion – the point at which the scree plot starts to get constant is considered indicative of the maximum number of components to be extracted. After choosing the number of PC to be used, it is possible to analyze the factor score plot that allows visualizing clusters in the data. In the factor loadings plot it is possible to observe the influence of each original variable on the formed clusters.

Computational implementation

In order to execute the SOM algorithm, bootstrap, topology preservation tests, and reliability of neighborhood relationships, a program in FORTRAN 90 was developed. As a matter of presentation quality, all the graphs of the results obtained were generated using the software STATISTICA 7.1 (StatSoft Inc., Tulsa, OK, USA). The principal component analysis (using the correlation matrix) and the basic statistical tests (correlation, ANOVA, and Tukey's test) were also performed with the aid of STATISTICA version 7.1.

Results and discussion

Pre-processing

The mean areas of each sensor for each coffee (Table 1) are presented and statistically compared in Table 2. There was significant difference ($p < 0.05$) between the coffees for the ten sensors. Thus, for the data analysis using the SOM and PCA, all the sensors were used as input variables.

Table 2. Comparison by the Tukey's test ($p < 0.05$) between the mean areas obtained by the electronic nose for each sensor and coffee.

	Coffee						
Sensor	1	2	3	4	5	6	7
S1	36.03^{d}	46.70^{b}	54.84^{a}	52.78^{a}	48.98^{b}	41.71^{c}	34.84^{d}
S2	427.08^{a}	172.18^{d}	165.00^{d}	114.96^{c}	174.08^{c,d}	214.93^{c}	358.12^{b}
S3	32.86^{d}	45.27^{b}	52.39^{a}	51.04^{a}	47.03^{b}	40.47^{c}	33.60^{d}
S4	59.37^{a,b}	57.00^{c}	57.37^{c}	59.00^{a,b}	58.77^{b}	59.20^{a,b}	60.10^{a}
S5	37.88^{d}	48.05^{b}	55.74^{a}	53.70^{a}	50.27^{b}	42.98^{c}	36.54^{d}
S6	109.88^{b}	84.03^{c}	62.56^{c}	70.40^{d,c}	77.35^{c,d}	109.86^{b}	137.56^{a}
S7	141.15^{a}	70.39^{b}	65.07^{b}	61.31^{b}	70.27^{b}	68.22^{b}	134.53^{a}
S8	85.41^{c}	67.87^{d}	52.80^{c}	65.35^{d}	69.31^{d}	93.30^{b}	114.51^{a}
S9	314.45^{a}	117.74^{d}	96.18^{e,f}	79.86^{f}	107.14^{d,e}	149.10^{c}	232.04^{b}
S10	59.61^{a,b}	54.14^{c}	53.68^{c}	60.06^{a,b}	58.75^{a,b}	59.74^{a,b}	62.22^{a}

*Mean values in the same row with the same letter have no significant difference at 5% level, by the Tukey's test.

Principal component analysis

As previously mentioned, the PCA was performed using the matrix of correlation of the integrated and auto scaled signals of each one of the ten sensors used by the electronic nose. According to the criteria proposed by Hair Junior et al. (2005), the number of PC to be retained for interpretation should be: two according to the eigenvalue criterion, three for the criterion of explained variance, and four for the scree test. Considering that the objective of this study is not to make a detailed analysis of the data using the PCA, but rather to use these results to compare and validate the utilization of the SOM, only two components were selected (eigenvalue criterion), adding up 93.4% of the total variance of the data.

The dispersion of the factor scores (Figure 1a) on the coordinates system, formed by the first two principal components selected, showed a separation of the samples, according to the type of coffee. From the factor loadings plot (Figure 1b), it was verified that the input variables can be divided into four groups: group 1 (S1, S3 and S5), group 2 (S2, S7 and S9), group 3 (S4 and S10) and group 4 (S6 and S8). A similar distribution was obtained by Zhang et al. (2007), when an electronic nose with the same sensors, mark, and model was used for the wheat. Thus it is possible to infer that the correlation pattern is result from the partial specificity of each sensor.

Self-organizing map

It was tested a SOM with a square grid containing from four to nine neurons at each dimension. Each map was trained over 5,000 times, and the training was repeated 100 times using, for each repetition, different weights and samplings, aiming to evaluate the stability and reliability of obtained results. As observed in Figure 2, the more neurons, the higher the coefficient of variation of the MQE; this indicates that a map with four units at each dimension (16 neurons) would be more proper for the data analyzed. Another fact that corroborates is the correlation coefficient for the graphs *dy-dx* (Figure 2). The network with four neurons had the highest correlation ($r = 0.804$), thus there was greater preservation in the data topology. Furthermore, an increase in the map only causes a greater dispersion among the samples, but the neighborhood properties are maintained. Boishebert et al. (2006) also reported this effect, i.e., the application of larger maps would be a waste of computational time.

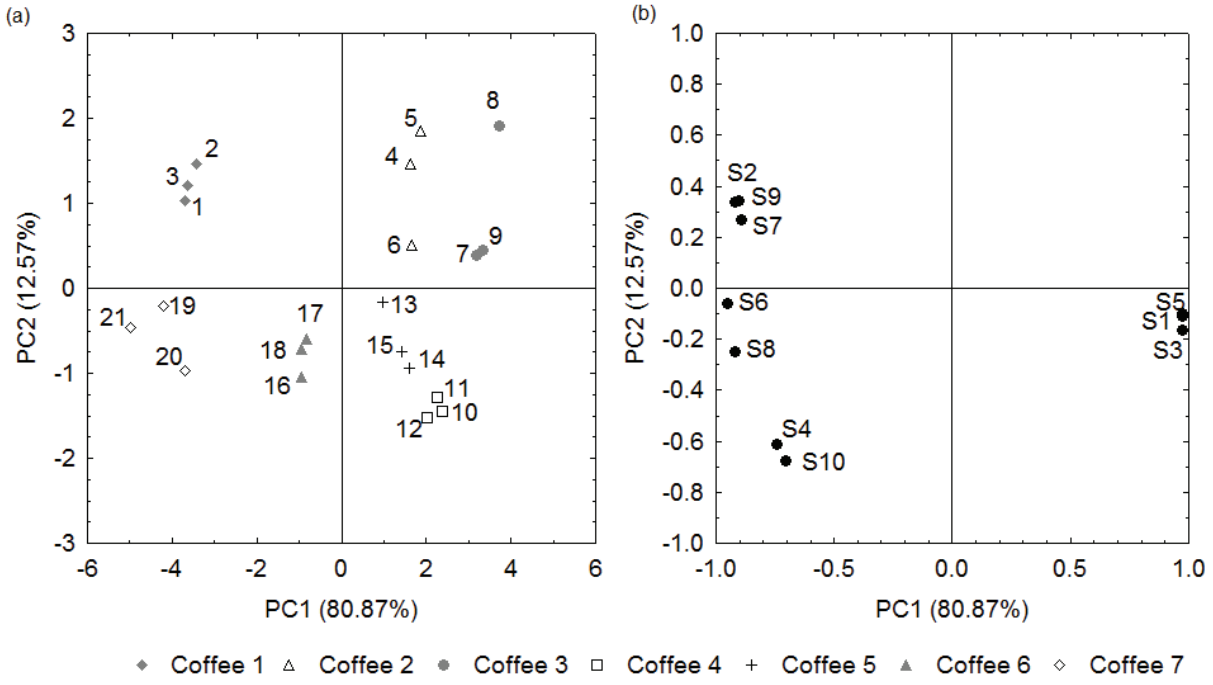

Figure 1. Scatter plot of the scores (a) and of the loadings (b) along the first two principal components.

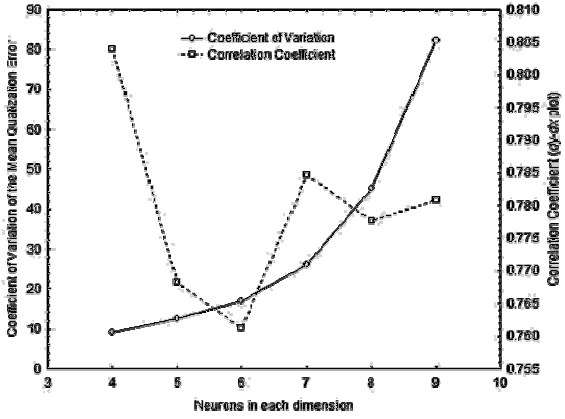

Figure 2. Coefficient of variation of the MQE and correlation coefficient of the *dy-dx* plot as a function of the number of neurons at each dimension of the self-organizing map.

The evaluation of the reliability of neighborhood relationships was carried out using square regions with radii equal to 0 and 1. For $r = 1$, the samples were considered neighbors if they are in adjacent centroids, so, they are similar, but can not be considered repetitions of a same behavior pattern. These values are within the ideal range of neighbors ($1.6 < \nu < 14.4$), which obeys the conditions necessary to perform the hypothesis test in a map 4 x 4. According to the criteria established by Bodt et al. (2002), the analyses of stability of neighborhood relationships for $r = 0$ and $r = 1$ can be considered statistically valid (Table 3). The statistically significant percentage of neighbors or non-neighbors is much greater than 5%. In the adhesion test, the value observed for the χ^2 is high and shows that the cumulative probability distribution of the maps is statistically different from that observed for the binomial distribution, at 1% level.

Table 3. Analysis of the reliability of neighborhood relationships of the topological map with four neurons per dimension.

	Neighborhood radius (r)	
	0	1
Mean number of neighbors (ν)	1.00	6.25
Probability of neighborhood at random ν/N	6.25	39.06
Acceptance range of H_0 [a]	$0.0151 < EVIZ_{i,j}(r) < 0.1099$	$0.2950 < EVIZ_{i,j}(r) < 0.4863$
Neighbors[b]	20	51
Non-neighbors[b]	182	155
% Total[c]	96.19	98.10
χ^2 (observed)	484.06	261.23
$\chi^2_{0.99}$ (table)	36.2	29.1

[a]The hypothesis H0 supposes that the neighborhood relationship is due to chance; equation (6). [b]Statistically significant, at 2.5% level. [c]Statistically significant percentage of neighbors or non-neighbors in relation to the total of possible pairs (210) for 21 samples.

In the Figure 3a it is presented the neighborhood relationships for the sample 1 (coffee 1). As expected, the samples 2 and 3 are located within the same centroid, since they are repetitions of the same coffee. The samples 19, 20 and 21, all belonging to the coffee 7, are located at a centroid adjacent, indicating a similarity of aromatic pattern. The analysis of the factor scores plot (Figure 1a) confirms the proximity between the groups and also justifies the lowest value of

$EVIZ_{1,20}(1)$, because is visible that the sample 20 is farther away from the samples 19 and 21, of the group of samples of the coffee 1. As the same behavior described for the sample 1 was observed for the repetitions 2 and 3, the respective graphs have been omitted. The neighborhood relationships for the coffee 7, represented by the sample 19 and illustrated in the Figure 3b, confirm the aromatic similarity with the coffee 1. Moreover, it is evident that the SOM has grouped the repetitions 20 and 21 in the same centroid and that the sample 20 is a little farther away from the group, because the value $EVIZ_{19,20}(0)$ is lower than $EVIZ_{19,21}(0)$. For the coffee 6, sample 16 in the Figure 3c, we only observed the group of the repetitions within the same centroid. Therefore, there was no statistically similar aromatic pattern to the coffee 6, considering the analyzed samples. The above statements are corroborated by careful analysis (the scales of each component are different) of the factor scores plot (Figure 1a).

Graphs similar to those presented in Figure 3 have been built for the other coffees. The correct grouping of the repetitions was verified in a same centroid for all the coffees. Besides that, there was a great similarity between the neighborhood relationships described by the SOM and those indicated by the PCA. Verdini et al. (2007) also reported a similarity between the methodologies. According to the authors, only small variations were obsrved by the different weights that each method provides for the input variables. On the other hand, Díaz et al. (2003) argued that the SOM improved the separation of the samples in relation to the PCA, since it is a method that allows the representation of non-linear relationships in the data. It is possible to search among the performed training repetitions, one with the topological behavior inferred by the analysis of stability of neighborhood relationships. In Figure 4, it was represented the topological maps of the training repetitions 20 and 98. It is observed a difference in the position of the groups, but the neighborhood relationships are maintained.

The repetition 20 was selected for the construction of the weight maps, but the same rules were identified for the repetition 98. The contour plots of the Figure 5 were obtained by interpolation using spline functions (StatSoft Inc., Tulsa, OK, USA), of the weight plans correspondent to each sensor. As with the PCA (Figure 1b), the maps of weights can be placed into the same four groups according to the behavior presented. Thus, through the SOM it was also possible to identify the high positive correlation between some sensors. In the Figure 5 only one sensor of each group is illustrated for the analysis of the topological distribution of the samples. The examination of the Figures 4a and 5a indicates that the coffees 1 and 7 are those with the smaller signal for the sensors S1, S3 and S5 (group 1), the coffee 6 is located in an intermediate region and the rest of the coffees had higher values. For the group 2 of sensors (S2, S7 and S9) in the Figures 4a and 5b, it is possible to see that the coffees 1 and 7, and the coffee 4 are, respectively, those with the highest and lowest signals. The other coffees are located in regions of intermediate value for these sensors. The group 3 (S4 and S10) separates the coffees (Figures 4a and 5c) into three categories according to the intensity of signal: larger (1, 4, 6 and 7); intermediate (5); and small (2 and 3). For the group 4 (S6 e S8), the grouping (Figures 4a and 5d) according to the values of the input variables would be: higher for 1; 6 and 7 (the last is slightly larger); intermediate for 2 and 5; smaller for 3 and 4. All the observations described above are in accordance to the PCA and the means comparison test (Table 2).

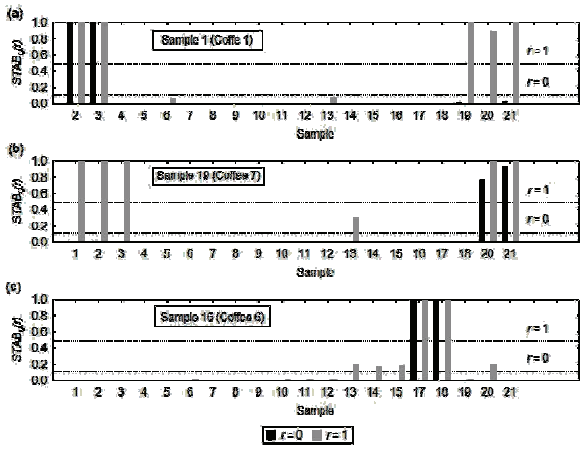

Figure 3. Graphs with values of $STAB_{i,j}(r)$ for the samples 1 (a), 19 (b) and 16 (c). Horizontal dotted lines indicate the limit above which a sample is considered statistically neighbor at 2.5% level, for a given radius.

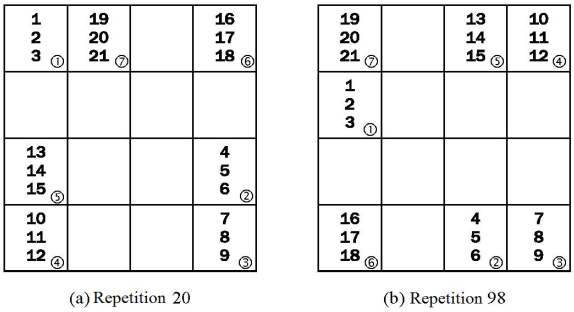

Figure 4. Topological maps 4 x 4 of two training repetitions. Each square represents one neuron, the numerals, the position of the respective sample, and those circulated, the coffee.

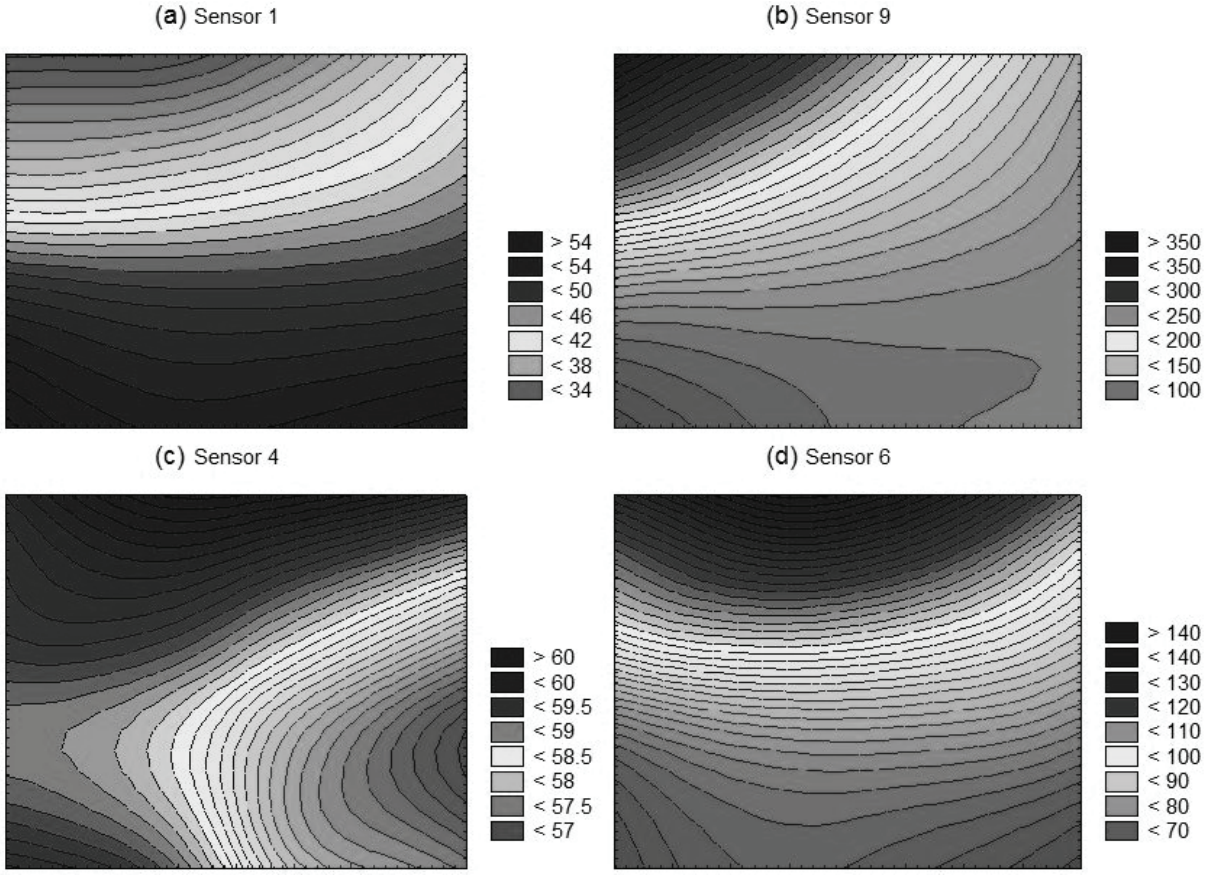

Figure 5. Weight maps of some selected sensors (20th training repetition).

Through the weight and topological maps, a great similarity was registered between the results of PCA and SOM. This fact points out that the aromatic pattern obtained by the electronic nose for the seven coffees analyzed can be described by linear relationships. Among the advantages of employing SOM, we can highlight the mathematical simplicity, the quantity and quality of results as graphs, which facilitates the extraction of rules about the data without a thorough understanding of the methodology. Moreover, the SOM allows the representation of non-linear relationships through two-dimensional maps without any increase in the complexity of the method. Thus, when the data have non-linear behaviors, the utilization of the SOM is more recommended than the PCA. On the other hand, the training of a neural network can be computationally more costly than the classical multivariate methodologies.

Conclusion

It was proved that, both the PCA and SOM are able to distinguish the seven types of coffee. The neighborhood relationships and the influence of original variables were also, similarly, represented by both methodologies. The combination of the electronic nose with the methodology of applied analysis in cthis study enables the reliable use of SOM in the quality control of soluble coffee.

Acknowledgements

The authors wish to express their gratitude to Fundação Araucária, by financial support; to CAPES, by the doctoral degree scholarship; and to the Iguaçu Company of Soluble Coffee.

References

BODT, E.; COTTRELL, M.; VERLEYSEN, M. Statistical tools to assess the reliability of self-organizing maps. **Neural Networks**, v. 15, n. 8-9, p. 967-978, 2002.

BOISHEBERT, V.; GIRAUDEL, J. L.; MONTURY, M. Characterization of strawberry varieties by SPME-CG-MS and Kohonen self-organizing map. **Chemometrics and Intelligent Laboratory Systems**, v. 80, n. 1, p. 13-23, 2006.

CLARKE, R. J. Technology III: Instant coffee. In: CLARKE, R. J.; VITZTHUM, O. G. (Ed.). **Coffee**: Recent Developments. London: Blackwell Science, 2001. p. 125-139.

CORNEY D. Food bytes: intelligent systems in food industry. **British Food Journal**, v. 104, n. 10, p. 787-805, 2002.

CRAVEN, M. A.; GARDNER, J. W.; BARTLETT, P. N. Electronic noses – development and future prospects. **Trends in Analytical Chemistry**, v. 15, n. 9, p. 486-493, 1996.

DEISINGH, A. K.; STONE, D. C.; THOMPSON, M. Applications of eletronic noses and tongues in food analysis. **International Journal of Food Science and Technology**, v. 39, n. 6, p. 587-604, 2004.

DEMARTINES, P.; HÉRAULT, J. Curvilinear component analysis: A self-organizing neural network for nonlinear mapping of data sets. **IEEE Transactions on Neural Networks**, v. 8, n. 1, p. 148-154, 1997.

DÍAZ, C.; CONDE, J. E.; ESTÉVEZ, D.; OLIVERO, S. J. P.; TRUJILLO, J. P. P. Application of multivariate analysis and artificial neural networks for the differentiation of red wines from the Canary Islands according to the Island of origin. **Journal of Agricultural and Food Chemistry**, v. 51, n. 15, p. 4303-4307, 2003.

FARAH, A.; MONTEIRO, M. C.; CALADO, V.; FRANCA, A. S.; TRUGO, L. C. Correlation between cup quality and chemical attributes of Brazilian coffee. **Food Chemistry**, v. 98, n. 2, p. 373-380, 2006.

GHASEMI-VARNAMKHASTI, M.; MOHTASEBI, S. S.; SIADAT, M. Biomimetic-based odor and taste sensing systems to food quality and safety characterization: An overview on basic principles and recent achievements. **Journal of Food Engineering**, v. 100, n. 3, p. 377-387, 2010.

HAYKIN, S. **Redes neurais**: princípio e práticas. 2. ed. Porto Alegre: Bookman, 2001.

HAIR JUNIOR, J. F.; ANDERSON, R. E.; TATHAM, R. L.; BLACK, W. C. Análise fatorial. In: HAIR JUNIOR, J. F.; ANDERSON, R. E.; TATHAM, R. L.; BLACK, W. C. (Ed.). **Análise multivariada de dados**. 5. ed. Porto Alegre: Bookman, 2005. p. 89-127.

HUANG, Y.; KANGAS, L. J.; RASCO, B. A. Applications of artificial neural networks (ANNs) in food science. **Critical Reviews in Food Science and Nutrition**, v. 47, p. 113-126, 2007.

MARINI, F.; MAGRÌ, A. L.; BUCCI, R.; MAGRÌ, A. D. Use of different artificial neural networks to resolve binary blends of monocultivar Italian olive oils. **Analytica Chimica Acta**, v. 599, n. 2, p. 232-240, 2007.

MELLO, A. A.; TRUGO, L. C. Tipificação odorífera de compostos voláteis do café. In: FRANCO, M. R. B. (Ed.). **Aroma e sabor de alimentos**. São Paulo: Livraria Varela, 2003. p. 169-175

MELSSEN, W.; WEHRENS, R.; BUYDENS, L. Supervised Kohonen networks for classification problems. **Chemometrics and Intelligent Laboratory Systems**, v. 83, n. 2, p. 99-113, 2006.

PARDO, M.; SBERVEGLIERI, G. Coffee analysis with an electronic nose. **IEEE Transactions on Instrumentation and Measurement**, v. 51, n. 6, p. 1334-1339, 2002.

TIGRINE-KORDJANI, N.; CHEMAT, F.; MEKLATI, B. Y.; TUDURI, L.; GIRAUDEL, L.; MONTURY, M. Relative characterization of rosemary samples according to their geographical origins using microwave-accelerated distillation, solid-phase microextraction and Kohonen self-organizing maps. **Analytical and Bioanalytical Chemistry**, v. 389, n. 2, p. 631-641, 2007.

VERDINI, R. A.; ZORRILLA, S. E.; RUBIOLO, A. C.; NAKAI, S. Multivariate statistical methods for Port Salut Argentino cheese analysis based on ripening time conditions, and sampling sites. **Chemometrics and Intelligent Laboratory Systems**, v. 86, n. 1, p. 60-67, 2007.

ZHANG, H.; WANG, J.; TIAN, X.; YU, H.; YU, Y. Optimization of sensor array and detection of stored duration of wheat by electronic nose. **Journal of Food Engineering**, v. 82, n. 4, p. 403-408, 2007.

Received on August 14, 2010.
Accepted on October 15, 2010.