We are IntechOpen, the world's leading publisher of Open Access books Built by scientists, for scientists

4,800
Open access books available

122,000
International authors and editors

135M
Downloads

Our authors are among the

154
Countries delivered to

TOP 1%
most cited scientists

12.2%
Contributors from top 500 universities

CLARIVATE ANALYTICS BOOK CITATION INDEX INDEXED

WEB OF SCIENCE™

Selection of our books indexed in the Book Citation Index in Web of Science™ Core Collection (BKCI)

Interested in publishing with us?
Contact book.department@intechopen.com

Numbers displayed above are based on latest data collected.
For more information visit www.intechopen.com

Chapter 7

Chemometrics: Theory and Application

Hilton Túlio Lima dos Santos, André Maurício de Oliveira,
Patrícia Gontijo de Melo, Wagner Freitas and Ana Paula Rodrigues de Freitas

Additional information is available at the end of the chapter

http://dx.doi.org/10.5772/53866

1. Introduction

This chapter aims to present a chemometrics as important area in chemistry to be able to help work with many among of data obtained in analysis. The term *chemometrics* was introduced in initial 70th years by Svant Wold (Swede) and Bruce Kowalski (USA). According International Chemometrics Society, founded in 1974, the accept definition to chemometrics is (i) the chemical discipline that uses mathematical and statistical methods to design or select optimal measurement procedures and experiments (ii) to provide maximum chemical information by analyzing chemical data [1]. When the study involving many variable became the study in a multivariate analysis, so it is necessary to building a typical matrix and is normal to do a pre-processing. Pre-processing is a procedure to adjust the different factors with different units in values than allow give for each factor the same change to contribute to the model. After, next step is usually the Pattern Recognition method, to find any similarity in your data. In This method is common using the unsupervised group where there are the HCA and PCA analysis and the supervised group where there is the KNN. The HCA analysis (Hierarchical Cluster Analysis) is used to examine the distance among the samples in two dimensional plot (dendogram) and cluster samples with similarity. (Figure 1). Now PCA analysis (Principal Component analysis) is used to try decrease the size data set, without lost information about samples (Figure 2) and KNN used to classify samples using cluster previously know [2].

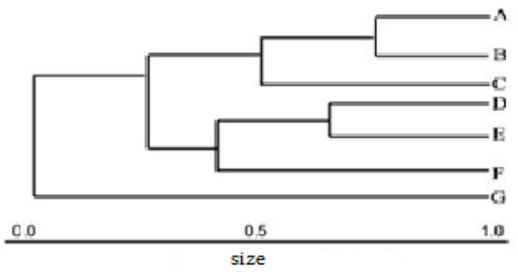

Figure 1. Example of dendogram

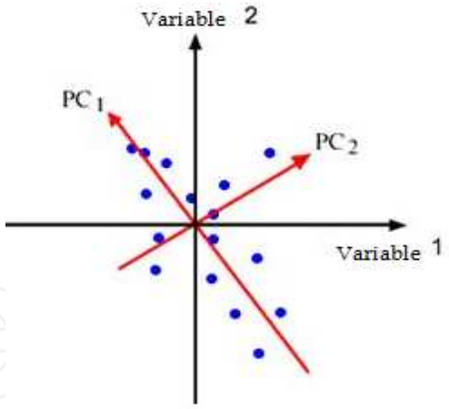

Figure 2. Clustering by PCA

Thus, the chemometrics show to be wide may be used in several area of knowledge.

2. Pattern recognition

In analytical chemistry when we have the data set, it is important find similarities and differences between samples based on measurements. For this is necessary to use methods according with information about the samples. And can be: Unsupervised (HCA and PCA) and Supervised methods (KNN)

2.1. Unsupervised methods

In this group there are two methods: Hierarchical Cluster Analysis (HCA) and Principal Components Analysis (PCA), and the goal is to evaluate if there is any clustering in data set without using the class about samples.

2.1.1. Hierarchical Cluster Analysis (HCA)

The Hierarchical Cluster Analysis is a technique to evaluate the distance between de samples and group in a plot calling dendogram. Theses distance can be calculated utilizing different methods as Euclidean or Mahalanobis or Manhattan distance, for example. For the Euclidean distance is using the equation 1, for Mahalanobis distance is using the equation 2 and for Manhattan distance is using equation 3:

$$\text{Distance} = \sqrt{(X_1 - Y_1)^2 + (X_2 + Y_2)^2 + \cdots + (X_n + Y_n)^2} \quad (1)$$

Where:
X_n and Y_n are the coordinates of sample X and Y in the n^{th} dimension of row space.

$$\text{Distance} = \sqrt{(X_i - Y_j)^T C^{-1} (X_i + Y_j)} \quad (2)$$

Where:
X_i and Yj are column vectors for objects *i* and *j*, respective and *C* is the covariance matrix.

$$\text{Distance} = \sum_{i=1}^{p} |X_i - Y_i| \quad (3)$$

Where:
X_i and Y_i are vectors.

When performed the estimate for distance, so is possible plot the dendogram. A general dendogram is showing below (Figure 3). In this dendogram is possible to see the samples (letters) and the distances (numbers). Samples belonging to clusters A, has a distance of 0,2 from one another. Same time the sample B has a distance 0,5 from cluster A. The value of distance can change according with the distance used to calculate.

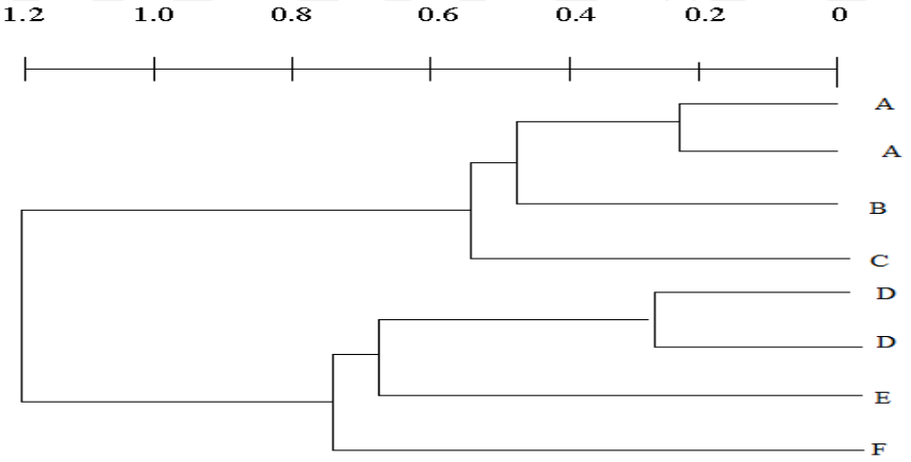

Figure 3. The general dendogram where above are the distances and right side are the samples

2.1.2. Principal Components Analysis (PCA)

The Principal Components Analysis (PCA) has the goal available the distances between the points using few axes in the row plot. In a matrix, each row is the point in the graphic below (Figure 2). So the aim is study the relationship between these samples to find the similarity and differences. In this general example are using two principal components (PC1 and PC2). The first PC (PC1) describes the major points in the graph and the maximum amount of variance, while the PC2 explain the remaining points. It is important to know that the sum of percentage described by PC′s must be close 100%. Another propriety of PC′s is about de position. The PC′s are always perpendiculars one with another.

The PCA technical can be used to define which variables are more important in a process. For this analysis is necessary use the factors (column in the matrix) and objects (row in the matrix). When the aim is to determine which variable are more important for the process is used *loading* and when want studying the relationship between objects is used *scores*

2.2. Supervised methods

The Supervised methods are using when want to construct a model using the class membership for future samples. In this group, KNN is a technical widely used when the goal is this.

2.2.1. K- Nearrest Neighbor (K-NN)

The KNN technical allows use the samples or clusters to identify another samples or clusters. For this is necessary to calculate the distances between them, using a Euclidean or Mahalanobis or Manhattan distance, for example. The minimum distance is calculated and the object is assigned to the corresponding class. A classification is dependent on the number of objects in each class.

3. Chemometrics in medicinal chemistry

3.1. The QSAR principle: Hansch analysis

The development of new drugs is a continuous challenge, before uncountable diseases the lack an adequate pharmaceutical approach. The modern medicinal chemists concern specially with methods based upon rational and quantitative procedures, aiming to focus on potentially efficient candidates. In that context, the use of chemometric methods is very important, in quantitative structure-activity relationship (QSAR) studies, and it presupposes that the biological activity (BA), measured through a biological response (BR), keeps a relationship with chemical structure (CS):

$$BR = f\left(CS\right) \tag{4}$$

The first attempt to quantitatively relate chemical structure to chemical behavior in a series of structuraly kindred compounds remounts to 1940´s, with Hammett [3] who, studying the meta- and para-substituted benzoic acids at 25°C, stablished linear relationships between the R = X substituted benzoic acid ionization constant (KX) and the ionization constant of the non-substituted benzoic acid (R = H):

$$\begin{gathered}\left(m-/p-R\right)C_6H_4COOH \rightarrow \left(m-/p-R\right)C_6H_4COO^- + H^+ \\ \sigma = \log\left(\frac{K_X}{K_H}\right) = \log\left(K_X\right) - \log\left(K_H\right)\end{gathered} \tag{5}$$

The σ constant is group-specific, and represents the electronic effect (inductive and resonance type) pursuit by R group. In 1964, Corwin Hansch [4] combined the use of the electronic constants to the lipophylic parameter (π), which represents the contribution of each R group to the overall lipophylicity:

$$\pi = \log\left(\frac{P_X}{P_H}\right) = \log\left(P_X\right) - \log\left(P_H\right) \tag{6}$$

where P_X is the X-substituted compound octanol-water partition coefficient, and P_H, the partition coefficient for a non-substituted compound. Thus, a QSAR equation evolves some kind of RB, for example, the negative logarithm of the minimal inhibitory concentration (MIC) for am antimicrobial compounds series (-log(MIC)), and the electronic (σ) and

lipophylic effect (π) of the R groups, the makes distinction among the several series representatives, can be expressed as

$$-\log(MIC) = \log\left(\frac{1}{MIC}\right) = a \cdot \sigma + b \cdot \pi + c \tag{7}$$

where *a*, *b* and *c* are the multiple regression coefficients.

The Hansch´s hypothesis that RB may be related to specific physico-chemical to each substituent present in the basic skeleton in a congener series of similar BA led to the proposition of numerous descriptors, of different kinds, useful to the identification of the principal effects that show up in drug action.

3.2. Physico-chemical descriptors

There are several physico-chemical descriptors, useful in QSAR studies that can be divided in categories: constitutional, topological, stereochemical and electronic ones, beside the so called indicator variables.

3.2.1. Constitutional descriptors

This kind of descriptor is related to the presence of structural characteristics that can affect the BA, such as: amount of unsaturated bonds, amount of hydrogen-bond donors, average ring size, etc.

3.2.2. Topological descriptors

These are descriptors that represent shape and connectivity, such as: ramifications, spacing groups, unsaturations, etc. The Kier [5] and Wiener [6] descriptors are typical.

3.2.3. Steric (or stereochemical) descriptors

Steric descriptors exist to describe effects related to the size of chemical groups and hindrance behavior. Taft steric descriptor, *Es*, [7] is a common example.

3.2.4. Eletronic descriptors

These variables are related to molecular electronic densities, and are used to be calculated by quantum methods. One can mention as examples: dipole moments, atomic partial charges, highest occupied molecular orbital energy (HOMO) and lowest unoccupied molecular orbital energy (LUMO).

3.2.5. Indicator variable and Taylor analysis

Indicator variables represent a useful way to convert a qualitative information into quantitative once, just as the occurrence of some kind of structural feature – setting 1 when

this feature is present, and 0 otherwise. The Taylor QSAR [8] approach employs indicator variables.

3.3. Chemometric methods applied to drug design

Chemometric statistical methods find in QSAR a large application field, considering that the multivariate problems are inherent to it.

3.3.1. Discriminatory and classificatory methods

Those methods aim the grouping and classification of compounds and variables in classes or categories that share resemblances, and are very interesting in pattern recognition situations and in dimensionality reduction of complex systems.

3.3.2. Principal Component Analysis (PCA)

Principal component (PCs) methods aim to combine correlated variables, projecting them in a new coordinate system, so that fewer variables are obtains, without any intercorrelation. The former coordinates are projects in a new axis system, in which the system variability is maximum along PC1, decreasing along the other axises (PC2, PC3...), all of the orthogonal each other, what allows one to deal just with the first components (usually PC1, PC2 and PC3). Thus, from a multi-variable universe, commonly multicolinear, one can obtain a simpler system with almost the same amount of information. Naming X the data matrix, with I×J dimension (I molecules and J descritors), a PCA generates two matrices, T e L, so that

$$X = TL^T \tag{8}$$

The matrix T is of scores, and represents the position of the compounds in a a novel coordinate system in which the components are its axises, and L is the loading matrix. Plotting the PCs instead of the original descriptors, one obtains groups governed by the similarities among the data.

3.3.3. Hierarchical Cluster Analysis (HCA)

This analysis is also useful to the classification of compounds, permitting visually distinguish the patterns and cluster. The plot resembling a tree, called dendogram, presents similar compounds at the same branches. Those branches are plotted based upon a similarity matrix, **S**, and each component of it is given by the similarity index between two samples *k* and *l*, *Skl*:

$$S_{kl} = 1.0 - \frac{d_{kl}}{d_{max}} \tag{9}$$

In this expression, d_{kl} is the Euclidian distance between *k* and *l*, and d_{max}, the maximum distance. Ferreira [9] describes a PCA/HCA analysis for a 25-compound series of 1,4-

naphtoquinones with antitumour activity. Using electronic descriptors, it was possible to distinguish active from inactive compounds (Figure 4). The loadings values indicate that the presence of high-density groups in side chain and terminal positions favours activity. The same profile arise from the dendogram analysis.

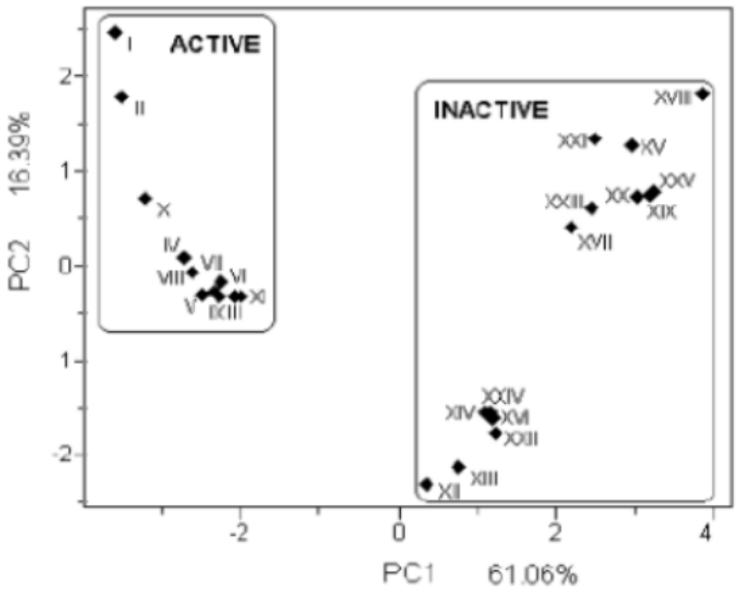

Figure 4. PC1 versus PC2 scores plot.

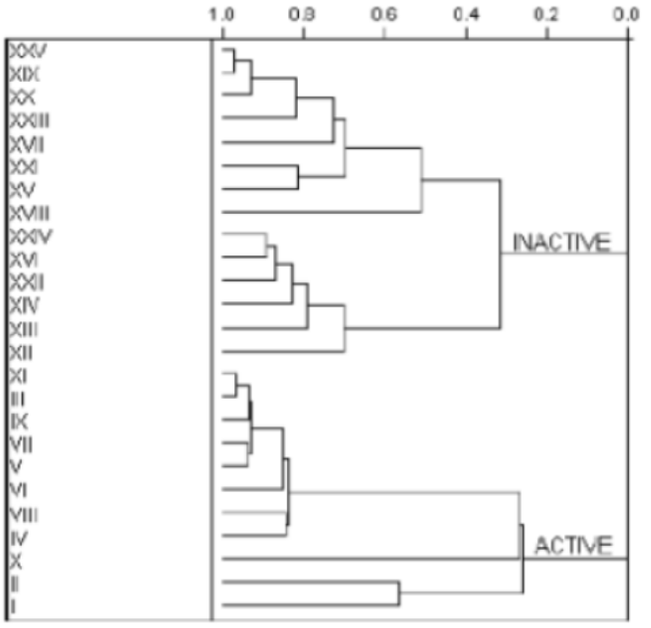

Figure 5. Dendogram for a naphtoquinone series

3.4. Multivariate regression

To construct a QSAR equation (Eq. 1), it is necessary to adopt some kind of multivariate fitting method in order to correlate the descriptors with the BR. The main methods are:

multilinear regression (MLR), principal component regression (PCR) and partial-least squares (PLS).

3.4.1. Multilinear regression (MLR)

The objective of this method is obtaining a relationship among a number of descriptors limited to 1/5 of the number of compounds and the BR, as an equation of the form:

$$BR = \alpha_1\left(\pm\varepsilon_1\right)\cdot D_1 + \alpha_2\left(\pm\varepsilon_1\right)\cdot D_2 + \alpha_3\left(\pm\varepsilon_1\right)\cdot D_3 + \cdots + \varepsilon \tag{10}$$

in which i are the regression coefficients, *Di* are the descriptors, εi, the coefficients confidence interval and ε, the independent term. The model statistical validation is very important, and it requires the consistency in the *Di* descriptors unit, as well as in values magnitude (necessarily). Statistical parameter like the fitting coefficient (*r*), the sample standard deviation (*s*), the cross-validation coefficient (q^2) and the Fischer test (*F*) are used in this task. The MLR is quite sensitive to multicollinearity: variables intercorrelated (tipically, com $r^2 > 0.6$) must not be used together. This is a common problem in multi-descriptor system that may be dealed with other regression methods.

3.4.2. Principal component regression (PCR)

In order to avoid multicollinearity, it is possible to make the regression, not with the descriptors themselves, but with their principal components (PCs) generated in a PCA treatment. The main advantage of this approach is the assurance that every variable are independent and no n-correlated, despite it is necessary to analyze the loading matrix (**L**). In this kind of regression, the variables are defined to maximize the descriptor matrix variance, without force a correlation with the BR

3.4.3. Partial least square (PLS)

Similarly to PCR, the PCs are employed, but in this case, the BR matrix has maximum variability, so that each loading matrix component (L) is a good predictor for each BR matrix component. This is the most used regression method, and it is adequate for dealing with 3D-QSAR problems, in which a set of compounds preciously aligned is put within a grid of interaction points with a molecular probe. Each point energy is a variable in the QSAR equation, which are by their turn corrlated with the BR to achieve a tridimensional profile of the critical sites that favours or disfavours the interaction with a hypothetical biological receptor.

4. Design of experiments

The exploration for new sources of energy such as biodiesel is of great importance today as well as their production processes. The factorial design is an important tool to reduce the search time, waste of reagents and hence operating costs [10]. A factorial design is

performed with the interest to determine the experimental variables and interactions between variables that have significant influence on the different responses of interest [11]. After selecting the significant variables, we must evaluate the experimental methodology and the influence of a particular variable on the yield of the reaction, a statistical experimental design, full factorial type, in which the independent variables are: the nature and concentration of catalyst temperature and the molar ratio between alcohol and oil and the dependent variable is the yield of esters produced. The variables that were not selected must be fixed throughout the experiment [12]. In a subsequent step must be chosen which planning used for estimating the effect (the effect) of the different variables results in a reduced number of conducting experiments. In the screening study the interactions between the variables (main interactions) and second order, usually obtained by full or fractional factorial designs. In the experiments are evaluated best experimental conditions, as well as their simultaneous effects that influence the yield of the reaction are therefore extremely important for understanding the behavior of the system [13]. The values of "p" and greater than or equal to 0.05 indicate that the factors: variable (1), variable (2), variable (3), variable (4) and the interactions of the variables are statistically significant at 95% reliable, since they are greater than 0.05. These parameters were evaluated at a low level (-1) and high (+1) are significant to the process of positive or negative manner. The Figure. 6 shows the profile of the Pareto chart [7]

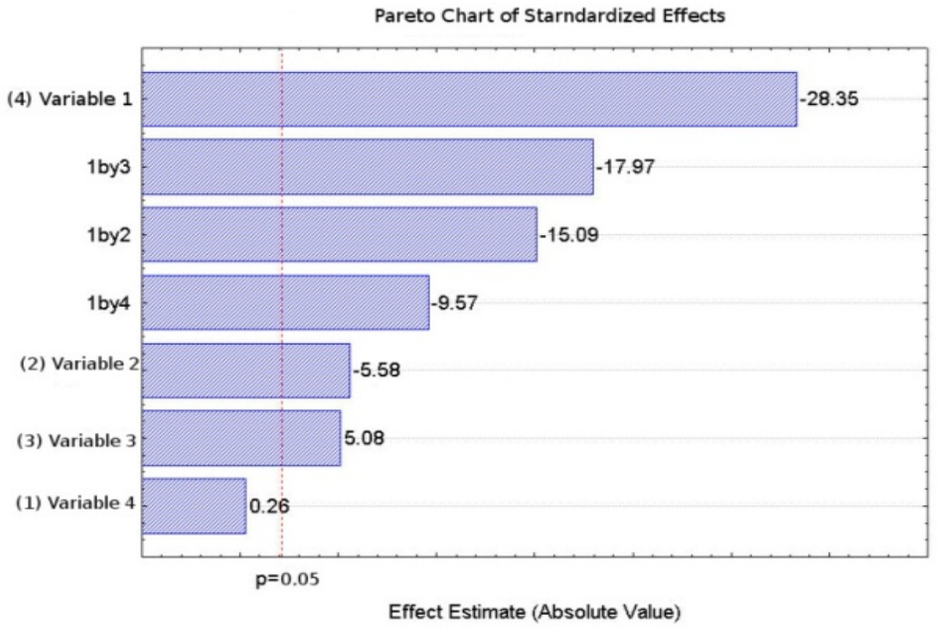

Figure 6. Pareto chart of the resulting fractional factorial design to evaluate the effects of each variable and their interactions in the reaction yield.

The analysis parameters obtained by means of multivariate optimization consists in choosing the conditions for preliminary assessment of experimental variables (fractional factorial design) followed by a response surface methodology (central composite design) made from the screening of the variables that may affect the synthesis of biodiesel. Generated model and the set of significant effects can evaluate through the study of

response surface methodology, as shown in Figure 7 and 8, and their interference in the response, ie the yield of the reaction, in which the dark area demonstrates the conditions that process has higher yield.

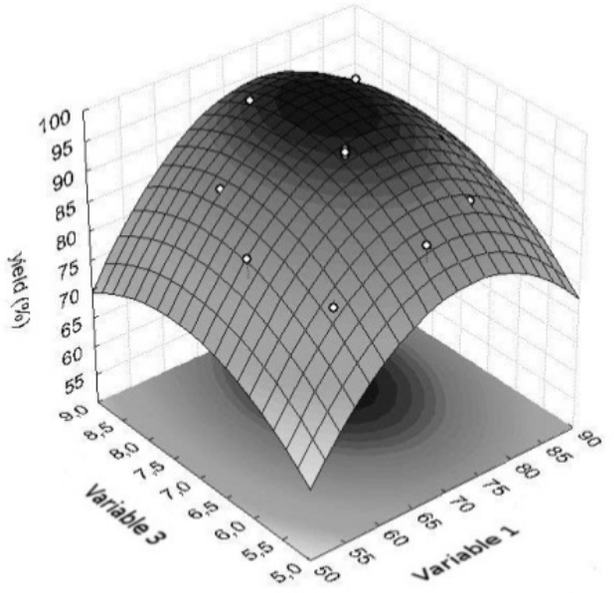

Figure 7. (a) Response surface generated by the central composite design for optimization of variables 1 and 3

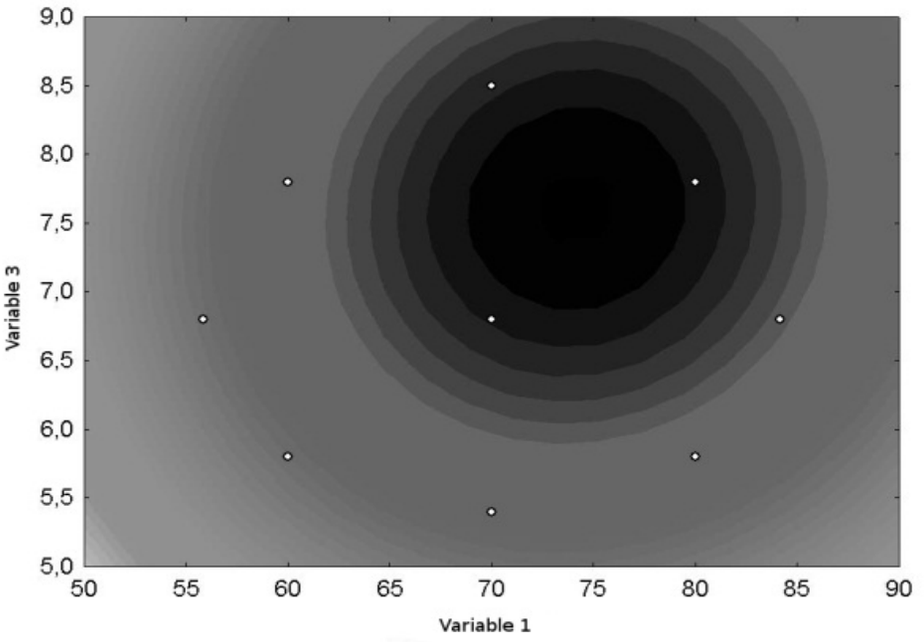

Figure 8. Zoom applied to the surface region of response.

Thus, the statistical analysis shown to be an important tool to evaluate, select and propose new technological routes, either through raw materials and / or process evaluation of the parameters that most influence the transesterification reaction to obtain for biofuels.

5. Conclusion of chapter

This chapter had as aim to show the versatility tools chemometrics in several areas. Was showed application chemometrics theory in drug design, natural products chemistry but it is not limited in theses area. Well, we hope to have expanded the range of chemometrics

Author details

Hilton Túlio Lima dos Santos and Wagner Freitas
University of São Paulo (USP), Brazil

André Maurício de Oliveira
Federal Center of Technology – Minas Gerais (CEFET - MG), Brazil

Patrícia Gontijo de Melo
Federal University of Uberlândia (UFU), Brazil

Ana Paula Rodrigues de Freitas
State University of São Paulo (UNESP), Brazil

6. References

[1] Otto M. Chemometrics- Statistic and Computer Application in Analytical Chemistry. Ed. Wiley-VCH. 1999
[2] Beebe K., Pell R., Seasholtz M., Chemometrics – A Practical Guide. Ed.Wiley Interscience Publication. 1998.
[3] Hammett, Louis P. (1937). J. Am. Chem. Soc. 59: 96.
[4] Hansch, C. (1969) A Quantitative Approach to Biochemical Structure-Activity Relationships. Acct. Chem. Res. 2: 232-239.
[5] Hall, Lowell H.; Kier, Lemont B. (1976). Molecular connectivity in chemistry and drug research. Boston: Academic Press.
[6] Wiener, H. (1947). "Structural determination of paraffin boiling points". J. Am. Chem. Soc. 1 (69): 17–20.
[7] R. W. Taft, Linear free energy relationships from rates of esterification and hydrolysis of aliphatic and ortho-substituted benzoate esters. J. Am. Chem. Soc. 1952, 74, 2729-2732.
[8] Hansch, C.; Sammes, P. G.; Taylor, J. B.; Comprehensive medicinal chemistry: the rational design, mechanistic study & therapeutic application of chemical compounds, Pergamon Press: Oxford, 1990, vol. 4.
[9] Ferreira, M.M.C. J. Braz. Chem. Soc., Vol. 13, No. 6, 742-753, 2002
[10] Charoenchaitrakool, M., & Thienmethangkoon, J. (2011). Statistical optimization for biodiesel production from waste frying oil through two-step catalyzed process. Fuel Processing Technology, 92(1), 112-118.
[11] Berrios, M., Gutiérrez, M. C., Martín, M. A., & Martín, A. (2009). Application of the factorial design of experiments to biodiesel production from lard. Fuel Processing Technology, 90(12), 1447-1451.

[12] Melo, P. G. (2012). Production and characterization of obtained from Macaúba (Acrocomia aculeata). Master degree thesis. Univesity of Federal of Uberlandia – Brazil.
[13] Atadashi, I. M., Aroua, M. K., & Aziz, A. A. (2010). High quality biodiesel and its diesel engine application: A review. Renewable and Sustainable Energy Reviews, 14(7), 1999-2008.
[14] Mingoti, S. A., (2007). Data analysis through methods of multivariete statistical approach applied. Federal University of Minas Gerais

IntechOpen