Talanta 85 (2011) 2479–2483

Contents lists available at ScienceDirect

Talanta

journal homepage: www.elsevier.com/locate/talanta

The ability of spectrum autocorrelation models to predict the lycopene concentration in foods through visible spectroscopic data

José S. Torrecilla[a,*], Virginia Fernández-Ruiz[b], Montaña Cámara[b], M.Cortes Sánchez Mata[b]

[a] Departamento de Ingeniería Química, Facultad de Ciencias Químicas, Universidad Complutense de Madrid, Avenida Complutense s/n, 28040, Madrid, Spain
[b] Departamento de Nutrición y Bromatología II, Facultad de Farmacia, Universidad Complutense de Madrid, Avenida Complutense s/n, 28040, Madrid, Spain

ARTICLE INFO

Article history:
Received 11 May 2011
Received in revised form 28 July 2011
Accepted 29 July 2011
Available online 6 August 2011

Keywords:
Ketchup
Tomato juice
Tomato sauce
Lag-*k* autocorrelation coefficient
Visible spectroscopy
Chaotic parameter

ABSTRACT

We developed a novel computerized approach based on lag-*k* autocorrelation coefficients (LCCs) and linear models (LMs) to estimate the concentration of lycopene in foods by the spectroscopy. The LCCs were calculated using the data obtained using whole visible scans from 400 to 600 nm (*vide supra*) of lycopene standards and food samples (ketchup, tomato juice and tomato sauce). The chaotic parameter (CP) was then transferred into a LM to estimate the concentration of lycopene compound. The integrated LCC/visible spectroscopy method developed can be considered as a satisfactory analytical technique able to estimate lycopene concentration in food samples in a fast accurate way, with a mean prediction error lower than 5.7% and a mean correlation coefficient higher than 0.957.

1. Introduction

Lycopene presence in the human diet is considered of great interest [1] due to its superior capacity to capture free radicals (more than doubling that of β-carotene). It is an anticarcinogenic and antiaterogenic agent taking part in intercellular communication and modulating the immunological mechanisms. It is more effective than other carotenoids found in both in vitro and in vivo [2].

In our diet lycopene is found in tomatoes, watermelon, papaya, persimmon and pink grapefruit, but fresh tomatoes and their processed products are the main contributors to the total lycopene intake. The concentration of lycopene in processed tomato products depends on the different technological treatment and the origin of the raw material, ranging between 0.85 and 94.0 mg/100 g [3,4].

In spite of the progress in lycopene determination in food products, at present food industries still demand environmentally friendly techniques to minimize the use of chemicals but which are highly versatile, fast, sensitive and selective enough to analyze lycopene concentration in complex matrices for developing new commercial foods [5–7].

Chromatographic methods, such as HPLC, show good precision, accuracy and sensitivity for the quantification of lycopene in fruit and vegetable samples and are more specific than spectrophotometry, where β-carotene and lycopene have common regions. Nevertheless, Olives Barba et al. (2006) [8] optimized and compared the visible spectrophotometric standard method with an HPLC method for the determination of lycopene and β-carotene in vegetables and concluded that the spectrophotometric method can be used when lycopene is the main carotenoid as in the case of tomato products.

The application of mathematical models represents a very useful strategy to improve the accuracy of the analytical data obtained from some technical equipment. In the last decades, the systems analyzed by these mathematical tools have been changed from predictable systems that could be exactly represented by explicit equations to those whose behaviour is often unpredictable in line with analysis using traditional linear equations.

In this context, linear and non linear estimative algorithms have been used to provide an adequate resolution of complex spectrums such as those obtained in food analysis. As an example, our research group has solved the overlapping effect of lycopene and β-carotene compounds in foods using non linear models based on neural networks [9,10]. In addition, the complex behaviour of most chemical systems (as food matrix) suggests that chaos-based methodology may be of use for modelling the system dynamics. And this is even more so in situations where foods must comply with many quality specifications [11].

Although there is no universally accepted definition of chaos, most experts would concur that chaos is the aperiodic, long-term behaviour of a bounded, deterministic system that exhibits sensitive dependence on initial conditions [12]. Given that most chaotic

* Corresponding author. Tel.: +34 91 394 42 44; fax: +34 91 394 42 43.
E-mail address: jstorre@quim.ucm.es (J.S. Torrecilla).

0039-9140/$ – see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.talanta.2011.07.104

regions can represent dynamical systems, and that tools based on chaotic parameters (CPs) can detect slight variations in initial experimental conditions [12,13], models based on CPs could be adequate to determine trace chemicals in food.

In the case of the tomato industry, many companies have implemented low cost equipment for routine analysis and visible spectroscopy is a quick and simple technique available in any quality control laboratory with an affordable cost. In tomato products, where lycopene is the main carotenoid, the spectrometric method could be easily improved using an algorithm based on chaotic parameters to rapidly assess lycopene content. To the best of our knowledge, in the food analysis area, there are few models based on chaotic parameters [14], but the successful results achieved in this and other scientific areas lead us to think that a model based on chaotic parameters would be adequate to estimate the concentration of other bioactive compounds such as lycopene.

The main goal of the present study is to show an innovative approach based on spectroscopic data where the implementation of one CP (an algorithm based on one chaotic parameter) in linear models (LMs) can accurately predict the lycopene concentration in food samples in a fast and reliable way.

2. Materials and methods

2.1. Reagents, standards, and instrumentation

2.1.1. Standard samples

Standard samples of *all-trans* lycopene used in this work were supplied by Sigma–Aldrich–Fluka (St. Louis, MO), with a purity ≥90%. For learning and verification of the model, six individual working standard solutions were prepared by dilution in hexane in the range of 0.4–3.2 μg mL^{-1}. Their purity was checked by calculating the concentration of the standard solution using the extinction coefficient [15,16].

2.1.2. Food samples

A total of 18 individual samples of tomato juices (3), tomato sauces (3) and ketchups (3) which are rich in lycopene were considered for analysis (2 lots of 9 units of three different commercial brands were purchased in local markets). One batch was used for learning and verification and the second for validation purposes. All samples were analyzed in triplicate.

2.1.3. Procedure

Lycopene quantification in tomato products was carried out after extraction by a hexane: acetone: methanol solvent (50:25:25 *v/v/v*) by spectrophotometry at 502 nm of the hexane layer. The validation of the analytical procedure, a study of linearity, accuracy, precision, detection and quantification limits has previously been studied and published in Olives Barba et al., 2006 [8]. In all tomato products, a minimum of three replicate measurements of spectroscopic absorption for each sample were carried out. To check spectrophotometer calibration and for identification and quantification of lycopene in food samples, standard solutions of known amounts of lycopene (Sigma Aldrich) in *n*-hexane (Merk, Darmstadt, Germany) were prepared every day. The response variables were the absorbance values in the visible range (400–600 nm). The acquisition step of the visible spectra was 0.5 nm. A Pharmacia Ultrospec 4000 UV–vis spectrophotometer was employed for absorbance measurements using quartz cells of 1 cm path length. Data acquisition and spectrometric evaluation were performed using PESSW software, version 1.2.

2.2. Chaotic parameter used

To estimate the lycopene concentration, a chaotic parameter or spectrum autocorrelation which consists of some lag-*k* autocorrelation coefficients (LCCs) was calculated using the visible spectra (400–600 nm) of hexane extract of standard samples and lycopene rich food samples. This spectrum autocorrelation was implemented in a simple linear model (*vide infra*).

Lag-*k* autocorrelation coefficient ($R_{\Delta\lambda}$) is also known as *correlogram* or serial correlation function. There is a mathematical relation between the LCC and the presence of chaos [16]. Sugihara and May stated that a sharp decrease of the autocorrelation function could denote the presence of chaos in the database [17]. Later, Kettemann et al. stated that the lag-*k* autocorrelation coefficient is a convenient tool for the characterization of spectral statistic and to describe the chaotic nature of databases [18].

The autocorrelation function is a linear measure which quantifies the extent to which X_λ *versus* $X_{\lambda-k}$ (here X represents the absorbance of the sample at a given wavelength λ) is a straight line. This parameter measures linearly how strongly on average each data point depends on the wavelength lag ($\Delta\lambda$). This is the ratio of the autocovariance to the variance of the data. For a given spectrum, $R_{\Delta\lambda}$ is between 1 ($\Delta\lambda = 0$) and 0 (large $\Delta\lambda$) [16]. $R_{\Delta\lambda}$ is defined by Eq. (1) [14,19],

$$R_{\Delta\lambda} = \frac{\sum_{n=1}^{N-k}(X_n - \bar{X})(X_{n-k} - \bar{X})}{\sqrt[z]{\sum_{n=1}^{N-k}(X_n - \bar{X})\sum_{n=1}^{N-k}(X_{n-k} - \bar{X})}} \quad (1)$$

where X, $\bar{X}$ and N represent the UV–vis absorbance set, their average and the total number of absorbance sets for a given spectrum, respectively. Given that the $\Delta\lambda$ value ranges between 5 and 105 nm, with steps of 5 nm, 21 values have been obtained. This $\Delta\lambda$ range was selected to use the whole representative LCC range (LCC = 0.980, $\Delta\lambda = 5$, to LCC < 1×10^{-5}, $\Delta\lambda = 105$ nm). For instance, in the case of $\Delta\lambda = 10$ or $\Delta\lambda = 20$, throughout the work $R_{\Delta\lambda}$ has been referred to as R_{10} or R_{20} (second or fourth lag-*k* autocorrelation coefficient), respectively. An example of these calculations is shown in Fig. 1.

2.3. Learning, verification and validation sample

The method of building the learning, verification and validation samples is the same. It consists of calculating the lag-*k* autocorrelation coefficients following Eq. (1) and using the UV–vis absorbance values of the standard chemical mixtures and food samples (*vide supra*). Every data set of the learning and verification samples is composed of twenty one lag-*k* autocorrelation coefficients (*vide supra*) with their respective lycopene concentrations, in μg mL^{-1}, which came from the standard chemical mixtures and food samples. The formats of these matrixes of databases are 22 columns (21 LCCs and a lycopene concentration) and as many rows as numbers of data sets. The concentration ranges between 0 and 3.2 μg mL^{-1}. As an example, visible scans of ketchup (0.500, 0.620 and 1.130 μg mL^{-1}), tomato juices (0.656, 0.718 and 1.518) and tomato sauce (1.033, 1.138 and 1.157 μg mL^{-1}) are shown in Fig. 1. The negative slopes of LCC profile trends and their graphical superposition are mainly based on the fact that the UV profiles also present similar contours and the nature of the LCC parameter which relates UV absorbance values for different wavelengths, Eq. (1). Nevertheless, these slight differences are used by the chaotic algorithms to estimate adequately the concentration of lycopene (*vide infra*).

The learning and verification samples were composed of a database including 78 data sets (rows of the matrix) including standard, juices, ketchup and tomato sauces (Table 1). The only dif-

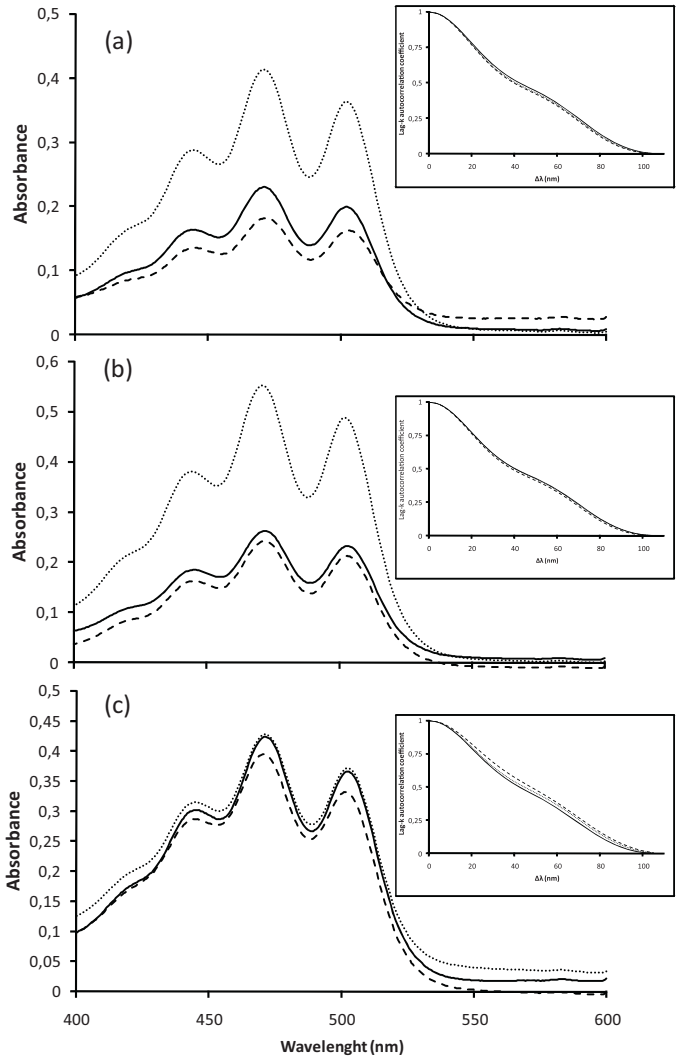

Fig. 1. Absorbance profiles for ketchups (a) (**....** 1.130 μg mL^{-1}; **___** 0.620 μg mL^{-1}; **- - -** 0.500 μg mL^{-1}), tomato juices (b) (**....** 1.518 μg mL^{-1}; **___** 0.718 μg mL^{-1}; **- - -** 0.656 μg mL^{-1}) and tomato sauces (c) (**....** 1.157 μg mL^{-1}; **___** 1.138 μg mL^{-1}; **- - -** 1.033 μg mL^{-1}).

ference between the verification and learning samples is that the latter is composed of 90% (68 data sets) of data and the former of the remaining 10%. Taking into account that every datum of the verification sample should be interpolated within the learning range, the data sets were randomly distributed between both samples [20].

On the other hand, with relation to the external validation process, the above mentioned spectrum autocorrelation has been calculated using different visible scans from ketchup and tomato juices and tomato sauces (9 different samples, Fig. 1). The concentrations of lycopene in the external validation sample are within the range of the learning sample [20].

Table 1
Main composition of learning, verification and external validation samples used.

	Databases	Number of samples	Trademarks
Standard	Learning and verification samples	51	Lycopene (Sigma Aldrich) Tomato juice: Hero, Granini and Kasfruit.
Commercial	External validation sample	27	Ketchups: Calve, Prima and Heinz. Tomato sauces: Solis, Orlando and Apis.

2.4. Linear models

The regression analysis is a simple method for investigating functional relationships among variables. These relationships are expressed in the form of an equation or a model connecting the response or dependent variable and one or more explanatory variables (a.k.a. independent variables, covariates, regressors, factors, carriers, *etc.*) [21].

In this case, a linear equation containing all those variables can be constructed, Eq. (2):

$$[\text{Lycopene}] = \beta_0 + \sum_{i=1}^{n} \beta_i \text{LCC}_i + \varepsilon \quad (2)$$

where, [Lycopene], n, β_i (β_0, β_1,..., β_n), LCC_i ($i = 1, 2, \ldots, n$) and ε represent a response variable (here lycopene concentration), number of observations, regression parameters or coefficients of the model, independent variables (here lag-k autocorrelation coefficients), and random error, respectively. The error term is an unobservable random variable that represents the residual variation and will be assumed to have zero mean, constant variance and a normal distribution [21].

These multiple regression models present different numbers of independent variable combinations which range from one to twenty-one independent variables. The regression models selection analysis ranks the best subsets of the explanatory variables based on statistical criterion which consists of their correlation coefficient (real *versus* estimated values) and mean prediction error. In this way, the best group of multiple regression models can be chosen. After the final regression models have been selected and thoroughly checked, these models should be validated using the relationship between the dependent variable and the final set of independent variables.

In this work, the linear models and statistical analyses were carried out by SPSS software version 17.

3. Results and discussion

As expected, there are similar trends between absorbance and LCCs profiles of all the tomato samples analyzed (with lycopene as the major carotenoid, >90%) and lycopene standard, with a characteristic maximum at 502 nm. No interference of other minor carotenoids was detected (Fig. 1). The only difference between the profiles shown in Fig. 1 is based on the chemical matrix where the lycopene can be found.

The data base used to design and verify the model consists of the composition of the standard samples composed of the concentration of lycopene and lag-k autocorrelation coefficients (21 values). The spectrum autocorrelation has been calculated using whole visible scans from 400 to 600 nm (*vide supra*) and then, the description and design of a linear model has been given.

As a first step in order to guarantee the reliability of the estimations calculated by these models, the applicability domain of the experimental measurements has been evaluated by searching the spectrum set with cross-validated standardized residuals greater than three standard deviations. In this evaluation, no response outlier was determined [22,23].

Once the domain of the whole data from visible scans has been tested, the lag-k autocorrelation coefficients were calculated using Eq. (1). The lag-k autocorrelating coefficients range between 0.98 and 1×10^{-5}. With this information, the learning, verification, and external validation samples were created (*vide supra*).

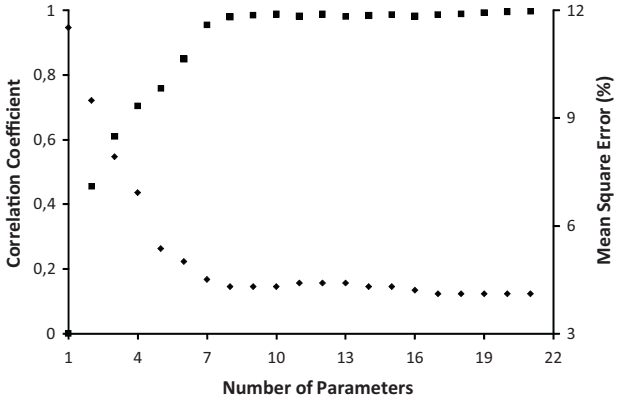

Fig. 2. Statistical results of the linear models used to estimate lycopene concentrations (■ correlation coefficient; ♦, mean prediction error).

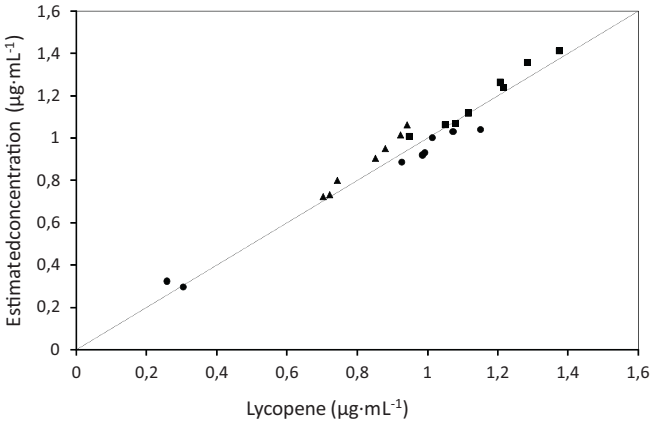

Fig. 3. Estimation of the lycopene concentrations in ketchups (●)Tomato juices (▲),and Tomato sauces (■).

3.1. Prediction of lycopene concentrations

Twenty one LCC coefficients were calculated with $k = 5$ nm wavelength ($\Delta\lambda$, Eq. (1)). These LCCs have been combined and optimized in multiple regression models (*vide supra*) which are polynomial functions from the first to the twenty first degree. These have been created by the combination of 21 independent variables, Eq. (2). In all cases, correlation coefficient (R^2) and mean prediction error values (MPE, Eq. (3)) of estimated and real values have been calculated, Fig. 2. As can be seen, using linear models with 0–21 LCCs, R^2 values increase from 0 to 0.99 and the MPE values decrease from 11.50 to 4.20%. Taking into account the statistical results shown in Fig. 2, in an attempt to reduce the complexity of the model, a stepwise variable selection was used to reduce the number of independent variables required. The best model consists of eight independent variables and is adequate as R^2 is higher than 0.984, and MPE (4.30%) is close to the minimum MPE value. For that reason this linear model was selected and then validated externally (Eq. (4)).

$$\mathrm{MPE} = \frac{1}{N}\sum_{i}\frac{\left|[\mathrm{Lycopene}]_i - [\mathrm{Lycopene}]_i^{\mathrm{est}}\right|}{[\mathrm{Lycopene}]_i}\cdot 100 \quad (3)$$

$$\begin{aligned}[\mathrm{Lycopene}] &= -1789.27 + 3092.89\cdot R_{10} - 5907.74\cdot R_{25} \\ &+6265.21\cdot R_{30} - 6897.43\cdot R_{45} \\ &+8525.91\cdot R_{50} - 3487.32\cdot R_{55} + 233.589\cdot \\ &R_{75}(R^2 > 0.984;\ \mathrm{MPE} < 4.30\%)\end{aligned} \quad (4)$$

In Eq. (3), N, $[\mathrm{Lycopene}]_i$, and $[\mathrm{Lycopene}]_i^{\mathrm{est}}$ are the number of observations, lycopene concentration value and their estimation, respectively. It is worthwhile to highlight that 8 parameter models can be optimized and validated using short experimental databases in comparison with the required databases to design models based on non linear models.

In Table 2, the statistical results of three comparable methods to estimate the lycopene concentrations are shown. Torrecilla et al. 2008 compared linear and NN models in the simultaneous estimation of β-carotene and lycopene and proposed a method to solve the overlapping effect in the simultaneous determination of both compounds in binary systems. The MPE and correlation coefficient of a linear model which uses UV absorbance values were unacceptable. In the case of the neural network (NN) model, although it also uses UV absorbance values, it consists of more complex algorithms and therefore the statistical results have notably improved. Although the third model (LCC model) is not applied to the same database, the three models are qualitatively comparable. The linear/CPs approach required less parameters and shorter computing time than the Neural Network model but as it is also based on non linear algorithms, its statistical results (estimated *versus* experimental values) are better than in the case of linear models, Table 2. More importantly, the lycopene concentration can be estimated using the LCC/visible approach with the least MPE.

Table 2
Statistical results of linear, non linear and spectrum autocorrelation models calculated using the external validation sample (*[10]).

		Linear model*	NN model*	LCC model
Lycopene	R^2	0.970	>0.999	0.957
	MPE (%)	2.028	4.57	5.70

NN = Neural network; LCC = lag-k autocorrelation coefficients; MPE = mean prediction error.

3.2. External validation

To externally validate the model described by Eq. (4), a new database has been used (*vide supra*). The estimations and real values of the external validation sample are plotted in Fig. 3. The correlation coefficient and MPE values of estimated *versus* real values are 0.957 and 5.7%, respectively. In the light of these results, the 8 parameter linear model used to estimate the lycopene concentration in the range studied is adequate. These statistical results lead us to think that linear models based on the LCCs can be used to determine the lycopene concentration in food samples.

In conclusion, the proposed LCC/visible approach is a reliable tool to determine the concentration of lycopene in food samples. The most important advantages of the proposed method are (i) its capacity to estimate the lycopene concentration by calculating autocorrelation coefficients, (ii) decreases the use of organic solvents, (iii) the easy, short sample preparation time required. These autocorrelation coefficients can be calculated easily (Eq. (1)), and the essential information can be extracted from huge databases such as visible scans. This model is suitable to estimate the lycopene concentration both off-line and on-line and therefore could be appropriate for quality and process control in the food industry.

Acknowledgements

The authors are grateful to the Spanish "Comunidad Autónoma de Madrid" for financial support of project S2009/PPQ-1545 and the project OTRI 26/2008, UCM-Agrucon 2008–2009.

References

[1] M. Cámara, M.C. Sánchez Mata, in: V. Rao (Ed.), Tomatoes Lycopene and Human Health, Caledonian Press, Barcelona, 2006, pp. 9–62.
[2] M.J. Periago, I. Martínez-Valverde, G. Ros, C. Martínez, G. López, Anales de Veterinaria (Murcia) 17 (2001) 51–66.
[3] M. Cámara, M.C. Matallana, M.C. Sánchez-Mata, R. Lillo, E. Labra, Acta Hort. 613 (2003) 365–371.
[4] G. Maiani, M.J. Periago Castón, G. Catasta, E. Toti, I. Goñi Cambrodón, A. Bysted, F. Granado-Lorencio, B. Olmedilla-Alonso, P. Knuthsen, M. Valoti, V. Böhm, E. Mayer-Miebach, D. Behsnilian, U. Schlemmer, Mol. Nutr. Food Res. 53 (2009) S194–S218.
[5] D. Bicanic, M. Anese, S. Luterotti, D. Dadarlat, J. Gibkes, M. Lubbers, Rev. Sci. Instrum. 74 (2003) 687–689.
[6] J.M. Roldán-Gutiérrez, M. Dolores Luque de Castro, Trends Anal. Chem. 26 (2007) 163–170.
[7] B. Schoefs, Trends Food Sci. Technol. 13 (2002) 361–371.
[8] A.I. Olives Barba, M. Camara, M.C. Sanchez Mata, V. Fernandez Ruiz, M. Lopez Saenz de Tejada, Food Chem. 95 (2006) 328–336.
[9] M. Cámara, J.S. Torrecilla, J.O. Caceres, M.C. Sánchez-Mata, V. Fernández-Ruiz, J. Agric. Food Chem. 58 (2010) 72–75.
[10] J.S. Torrecilla, M. Cámara, V. Fernández-Ruiz, G. Piera, J.O. Caceres, J. Agric. Food Chem. 56 (2008) 6261–6266.
[11] J.S. Torrecilla, E. Rojo, J.C. Domínguez, F. Rodríguez, Talanta 83 (2010) 404–409.
[12] J.C. Sprott, Chaos and Time-Series Analysis, Oxford University Press Inc., New York, 2003.
[13] J.S. Torrecilla, E. Rojo, J.C. Domínguez, F. Rodríguez, Talanta 79 (2009) 665–668.
[14] J.S. Torrecilla, E. Rojo, J.C. Domínguez, F. Rodríguez, J. Agric. Food Chem. 58 (2010) 1679–1684.
[15] AOAC (Association of Official Analytical Chemists), Official Methods of Analysis, AOAC, Arlington, VA, March Supplement, 1996.
[16] L. Zechmeister, A.L. LeRosen, W.A. Schroeder, A. Polgar, L. Pauling, J. Am. Chem. Soc. 65 (1943) 1940–1950.
[17] G. Sugihara, R. May, Nature 344 (1990) 735–741.
[18] S. Kettemann, D. Klakow, U. Smilansky, J. Phys. A: Math. Gen. 30 (1997) 3643–3662.
[19] H. Kant, T. Schreiber, Nonlinear Time Series analysis, Cambridge University Press, Cambridge, 2005.
[20] Guidance Document on the Validation of (Quantitative) Structure Activity Relationship [(Q)SAR] Models, No. 69, OECD, Series on Testing and Assessment, Organisation of Economic Cooperation and Development, Paris, France, 2007.
[21] S. Chattefuee, A.S. Hadi, Regression Analysis by Example, 4th ed., Wiley Interscience, A John Wiley & Sons, Inc., New Jersey, 2006.
[22] P. Gramatica, E. Giani, E. Papa, J. Mol. Graphics 25 (2007) 755–766.
[23] P. Gramatica, QSAR Comb. Sci. 26 (2007) 670–694.