1. INTRODUCTION

1.1 Food authentication and chemometrics..........................10
1.2 The many ways the data are and how to explore them....14
1.3 Objective of the thesis and structure.............................20
References..22

1. INTRODUCTION

1.1 Food authentication and chemometrics

The European Community regulation n° 178/2002 implemented in Italy in 2005 provides the guidelines for the food traceability and define it as the ability to trace and follow a food, a feed, a food-producing animal or a substance intended to be, or expected to be, part of a food or a feed, through all stages of production, processing and distribution. So traceability is in the mean time the activity of leaving a track, for example by means of appropriate labeling, and the capacity to find a trace in the opposite direction [1].

This keyword, regarding the scenario of foodstuff, implies many aspects, new mandatory parameters for health and hygiene, growing demands for food safety by the consumer and the need for companies to innovate their standards to be more competitive. To meet all these demands companies are first required to register and document the information process that follows the product from the beginning to the end of the chain, and after this they should be able to go back, moving from downstream to upstream, at the same information distributed along the chain. In few words, people that operates in the field of food or feed, must know the provenience of the raw material used and how they are produced, moreover such operators should have systems and procedures to identify the other companies to which their products have been supplied [2]. The complete documentation must be provided in case of controls to the competent authorities.

Main aim of this regulation is that of allowing a rapid localization of the purchaser of the final products, thus facilitating the withdrawal of the product from sell, if it is established the presence of risk for human health or environment and in addition to facilitate the identification of the responsible supplier for damage caused by non conform food. Above all the traceability is an assurance for producers and consumers in term of ‘quality’ that the product respects predefined specifics, not only chemical, physical, biological, but also organoleptic and geographical [3].

Nowadays these procedures consist in the organization of information related to the manufacturing process, thorough the use of paper documents; clearly this solution is rather limited and it poses problems in relation to the research of information and, moreover, this procedure can be subjected to falsification of the documents, recurring in cases of alimentary frauds. Even if an accurate labeling and exposition of documents attesting the products provenience is mandatory, the hundreds of seizures realized in Italy only in this last year, are the proof that more powerful instruments of control must be applied. These questions assume even more relevance when considering the high added values products, namely those identified by Protected Designation of Origin (PDO), where further than the compliance with the specifics, the customer want to be sure that the difference of price, is due to the higher standards of quality provided. So the necessity is that of reliable devices, which can assess the quality requisites through objective measurable parameters and can assure that what is declared in the label correspond to truth namely authentication procedures [4].

A scientific approach, involving chemical analysis have been recently demonstrated to be a valid instrument, also to support the too much subjective panel test used up to now for some products. In chemistry there is a growing literature regarding the food authentication issue and the development of methodologies able to identify the origin of alimentary products; just to mention few examples: metabolomics, proteomics and DNA assay, are certainly the most recent approaches [5-9], but also hyphenated techniques, like chromatography coupled with a mass spectrometry detection (GC-MS HPLC-MS) [10-12] are successfully applied for authentication of the most various alimentary products. The qualitative and quantitative determination of components of the analyzed food in the case of authentication task, due to the natural complexity of the food samples, requires an extended characterization of the food commodities. This is nowadays also a natural consequence of the improvement of modern chemical instrumentation, which in a single analysis, can determinate a large number of species. Thus an high number of variables, defining the sample, are obtained, that means the chemist sometimes had to deal with more numbers than laboratory tubes. This is the basic reason for which in food authentication issue, increasingly often chemometrics is used. In fact, chemometrics is the discipline aimed at efficiently extracting information from chemical systems, using multivariate data analysis [13]. The advantage of coupling multivariate data analysis and its powerful tools for data representation and interpretation, with chemical analysis is extremely relevant since it allows as well building quantitative models to asses authenticity and quality, capable to predict constituents content, and/or to classify unknown samples. In particular, for problems like authentication of quality, verification of labeling compliance or traceability of origin, it is important to verify whether a sample belongs to a specific class or not: in statistical terms, this problems all fall in the realm of pattern recognition.

In parallel with this aspect another, must be considered, namely the continuous demand from industry of methods that are compatible with the timing of production process and applicable to monitor the different steps of production, these control methods must be rapid, since the number of samples to be analyzed is large and at the same time economic, namely they must use simple instrumentation, with which even inexperienced operators can work and with minimum waste of materials and solvents.

As a consequence of this requirement, some widespread and economic analytical techniques have been reevaluated and approached with a different mood, also thanks to the fine interpretation of the data, obtained from multivariate analysis. For example spectroscopic analysis with near- or mean-infrared spectroscopy NIR-MIR, or nuclear magnetic resonance NMR if coupled with chemometrics models can supply methods able to single out chemical profiles related to the origin [14-17]. The signal obtained is a sum of the signal profiles provided from each species present in the samples, and, even if it is not possible to resolve the single profiles, hence to determine each species and the relative concentration, the total signal may as well represents a fingerprint of the product, named usually fingerprinting information, by which the different samples can be classified.

Usually, the species on which the studies of traceability and authenticity focuses are called indicators and can be classified in direct or primary and indirect or secondary indicators.

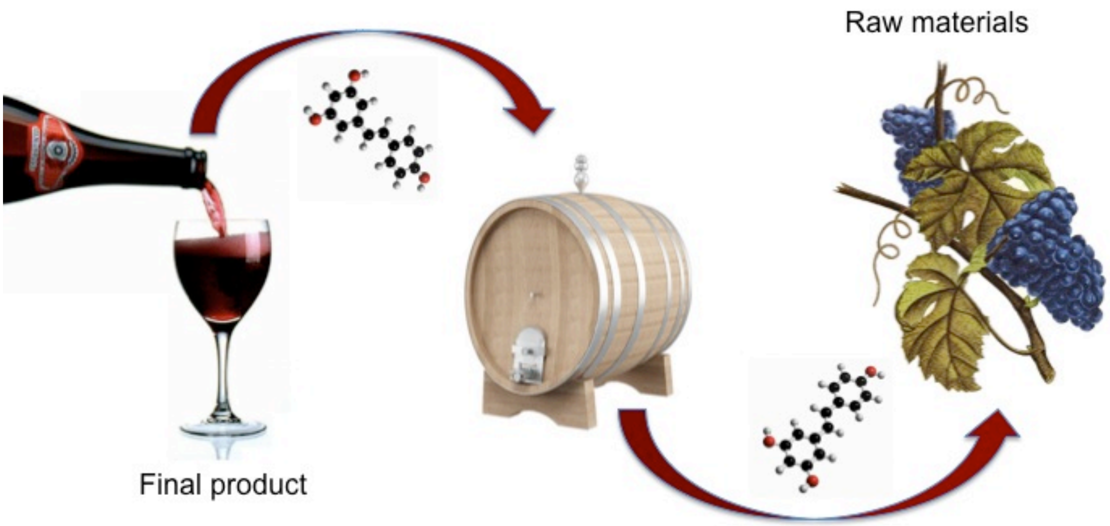

Figure 1: Tracing enological products with indirect indicators.

The firsts can be traced in every step of the production and they can be used to directly correlate the final product with the territorial characteristics of the place in which it originated, for instance metal contents or isotope ratios, this argument is largely described in many publications [18-20]. In this research, we will refer only to the use of secondary indicators, namely those species linked to the food chemical composition and to the production process, that only indirectly, through an extensive characterization of the food matrix may allow grouping products with the same origin and discriminating them from all other similar ones. In other words, extensive characterization corresponds to acquire a characteristic fingerprint of the product [21].

Part of the work of thesis is under a long term project (AGER Agroalimentare e Ricerca, New analytical methodologies for varietal and geographical traceability of oenological products [22]) that focus on creation of analytical methods and models of traceability and authentication of enological products.

1.2 The many ways the data are and how to explore them

To describe the different data structure existing is like making a sort of historical excursus of the development and improvement of chemical instrumentation: data has changed during the year due to more complex apparatus used for the analysis and along with this the data elaboration techniques have evolved. The simplest data we can obtain from an analytical determination is a number; imagine you have to calculate the concentration of a compound through UV-spectrophotometry, the correct procedure is to set the instrument on a specific wavelength, normally the maximum of its absorbance spectrum is used, then a value of absorbance is collected and converted in concentration according e.g. to Lambert-Beer law. The number obtained can be considered a 0-way data set and it can be seen as a matrix of dimension 1x1, or better a matrix element or single entry. If the same instruments is able to record, e.g. a UV spectrum of the compound, this means that a values of absorbance is calculated for each of M wavelengths in a specific range, the 1-way data set is a vector, namely a matrix line of dimension 1xM; other examples can be chromatograms, data obtained from the most common spectroscopies and titrimetric curves. When a diode array detector is coupled to liquid chromatography for detection, the system records continuously the absorbance for a range of wavelengths, this means that for each time point a spectrum is collected, and as a result for a chromatographic run a 2D-landscape is the response of the detector. This kind of data has the mathematical shape of a matrix, in which column vectors are chromatograms recorded at different wavelengths and row vectors are spectra and represents the 2-way data set. Clearly also other kind of data can assume this form, among the hyphenated techniques chromatography coupled with mass spectrometry provides them, but also other spectroscopic techniques like emission-excitation matrix fluorescence, or furthermore environmental data where the two dimensions can be sampling sites vs. contaminants. If more samples are analyzed with one of these techniques, a further dimension can be added to the data, namely the samples dimension, the data obtained can be represented mathematically with a cube of data, the structure is called a 3-way data set. In Figure 2 this evolution is graphically schematized: every square represent a number, namely an element of the matrix. There is also the possibility of supplying data of higher dimensionality adding new dimension of variables, or using some modern chromatographic apparatus where two different chromatography are performed on samples and more than one detection are used these are n-way data sets, obviously cannot be easily sketched, but are not so uncommon.

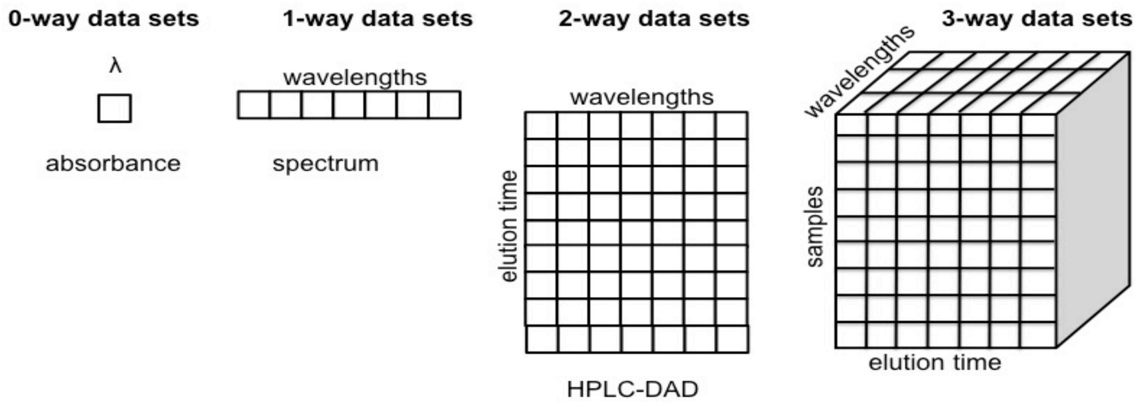

Figure 2: Representation of data sets from zero to three dimensions, referring to spectroscopy and chromatography.

The 3-way and in general n-way structures maintain the natural dimensionality of data extending the landscapes of hyphenated chromatographic techniques, this is one of the way in which data can be arranged, namely positioning matrices of samples one on the top of the other, the final array obtained is $\underline{\mathbf{X}}_{I,J,K}$ where I, J and K are the rank of data cube in each dimension.

A widespread alternative, instead of considering and analyzing the 3-way data preserving their 3-way data structure, is that of unfolding them, this means that every single matrix related to an experiment (2-way data) is converted in a vector in which lines of the chromatograms are placed one next to the other; final data shape is a matrix and the procedure is called row-wise unfolding, it is illustrated in Figure 3. The data set is indicated with $\underline{\mathbf{X}}_{I,JK}$ in analogy to the previous indicization.

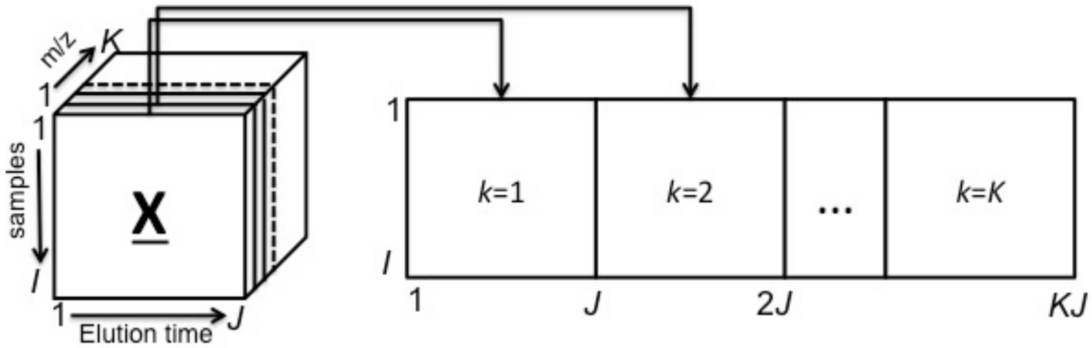

Figure 3: Row-wise unfolding from the 3-way dataset.

Finally, data can be also organized in multiset structures; these can be column wise augmented matrices, row wise augmented matrices, but also column and row wise augmented matrices (Figure 4), formed by a number of submatrices equal to the number of experiments. In building the structure one direction is maintained common to all the submatrices, namely one dimension must represent the same variables and have the same meaning, while the other is completely free and can have different size and different meaning. An example is represented by several experiment monitored with the same technique or the same process monitored with different techniques, but also several experiment monitored with several techniques.

Figure 4: Multiset structures, column wise, row wise and column and row wise augmented, the fixed direction is highlighted by a red and blu line.

Usually for chromatographic experiments collected in the same conditions, are sort in a column wise augmented matrix where the spectra dimension is fixed, this arrangement can be also see like a column wise unfolding from the 3-way data illustrated in Figure 5.

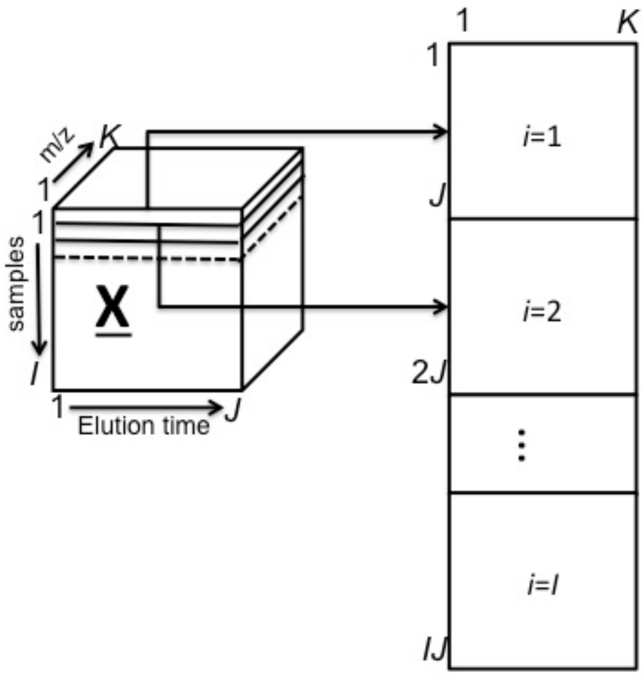

Figure 5: Column wise unfolding representing, from which a multiset is obtained.

It easy to understand that only the 3-way array maintains the original structure of the data and consider distinct the sources of variance that each mode represent, while with unfolding, the transformation from three-dimensional to bi-dimensional data, the risk is to lose the information due to interactions among variables and to overwhelm distinct sources of variance in one of the two dimension considered. The multiset arrangement also implies a bilinear data structure but is based on a different model with respect to principal component analysis (PCA) and PCA like methods, that do not seek direction of maximum variance, hence the data structure can be recovered, e.g. if for instance HPLC-DAD maps are acquired for a series of samples and considering the spectral dimension as common, the samples / elution times dimensions are compressed in one but still the chromatographic profile can be recovered. Moreover, if data follow a trilinear structure this can be imposed as a constraint in the model and taken into account.

Summarizing, there is now a widespread use of instrumentation that generates multi-way data in food analysis, processing and control.

Thus, multi-way data analysis is gaining more and more interest, and it has been demonstrated that data analysis can be more effective when using multi-way methods on multi-way data compared to unfolding procedures. Unfolding methods may result in more complex models with an associated risk of poor predictive ability. Unfolding may also be less efficient in terms of capturing and interpreting the underlying structure in a data set. On the contrary, the use of multi-way methods potentially simplifies the interpretation of the results and provides more adequate and robust models using relatively few parameters.

The multi-way classification methods that recently have been developed and applied in this work are NPLS [23] (multi-way Partial Least Squares) and NSIMCA [24] (multi-way Soft Independent Model of Class Analogies), the first is a discriminant methods and the second is a class modeling method. Class modeling is considered a more suitable approach to food authentication, since unlike discriminant analysis where the space is divided is as many region as the number of classes and the object is necessarily assigned to one of these, here the attribution is more flexible: a class space is defined and the object can belong to a class if it is projected into the class boundaries, it can be outside all the class boundaries so it does not belong to any class, but can also belong at the main time to more than one class and in this case it is defined a confuse object.

These techniques are compared with the classical two-way approaches performed on the unfolded data so PLS-DA [25] (Partial Least Square - Discriminant Analysis) and two-way SIMCA [26], whose original version was developed by Wold in the late 70s [27].

Multiset structure are typical of Multivariate Curve Resolution methods, this method allows to solve chromatographic peaks, recorded with a diode array or mass spectrometry detector ‘a posteriori’ using the second order information of the spectra recorded, the MCR-ALS algorithm, which make use of the alternating least square algorithm, was implemented by Tauler [28]. Recently also a discriminant approach using MCR-ALS was developed and it is one of the salient points of the discriminant studies in this Thesis.

1.3 Objective of the thesis and structure

This thesis focus on two principal topics regarding the use of chemometric in food authentication science: a multivariate approach to the chromatography, in the optic of optimization of the analysis and a comparative study of the more modern chemometric tools to build authenticity models. These two arguments will be carefully described trough out the script here only some preliminary information is given.

To improve the hyphenated chromatographic methods to assess the quality of oenological products is one hot topic, since chromatography is a widespread analytical technique. Main advantages are the ability of analyzing a large number of compounds at the same time, the use of different detector systems, and the possibility of interfacing with mass spectrometer. One of the aim of this work is the discussion of how chromatographic resolution and quantification of compounds can be optimized and improved using techniques of ‘a posteriori resolution’ instead of recurring to more sophisticated and costly instrumentation / procedures. Multivariate Curve Resolution technique is used for this scope and an example of its application in a real case of wine variety characterization is described.

The second part of the work regards the development of authentication models: among all the chemometric classification tools available, fundamentally the main distinction is between discriminant approaches and class modeling ones.
The use of class modeling methods, which on one hand have the disadvantage of being more complex and requiring the availability of a greater number of samples, on the other hand allows a more flexible assignation of the samples, a characteristics that is very useful in traceability studies. Indeed, while discriminant methods classify any unknown sample into one and only one of the categories for which training samples were

available, class-modeling tools can identify a sample as belonging to one category univocally, to more than one (confused object) or to none of the investigated classes. Moreover, by building independent models for each category, modeling tools allow to investigate the asymmetric cases when only a class is of interest (e.g. "is my wine sample, labeled as "Chianti", compliant to that denomination?").

So the different approaches are compared when applied to fingerprinting techniques, also considering another important aspect namely the dimensionality of the data: multi-way methodologies in food authentication are described and applied in authentication and detection of alimentary provenience; moreover new approaches are proposed for the classification.

The manuscript is organized in two main parts: Chapter 2 describes the used chemometric methods distinguishing them on the base of data organization, and a section is dedicated to the different data preprocessing; Chapter 3 and 4 reports respectively information on operative conditions for the data sets, acquired during the thesis' work or taken from bibliography and the results obtained, namely some applications on data sets. In Chapter 5 conclusions are reported. Finally I would like to highlight the content of the appendices since one of them wants to be a manual of instruction for recently developed methods not yet so familiar (NSIMCA).

This work has permitted to publish two papers and a chapter in a book the list of publications is reported in Appendix IV.

1. INTRODUCTION

References

[1] CE n° 178/2002.

[2] E. H: Golan, B. Krissoff, F. Kuchler, K. E. Nelson, G. K. Price, L. Calvin; Traceability for Food Safety and Quality Assurance: Mandatory Systems Miss the Mark. Current Agriculture, Food and Resource Issues (2003) 4.

[3] CE n° 510/2006.

[4] D. M. A. M. Luykx, S. M. van Ruth; An overview of analytical methods for determining the geographical origin of food products. Food Chemistry (2008) 107 897–911.

[5] D. S. Wishart; *Metabolomics: applications to food science and nutrition research*. Trends in Food Science and Technology (2008), 19(9) 482-493.

[6] S. Vaidyanathan, R. Goodacre; *Proteome and metabolome analysis for food authentication*. Food authentication and traceability (2003), ed. M. Lees, 71-100.

[7] Ö. Aslan, R. M. Hamill, T. Sweeney, W. Reardon, A. M. Mullen; *Integrity of nuclear genomic deoxyribonucleic acid in cooked meat: Implications for food traceability*. Journal of Animal Science (2009), 87(1), 57-61.

[8] I. Mafra, I. M. Ferreira, M. B. Oliveira; *Food authentication by PCR-based methods*. European Food Research and Technology (2008), 227(3) 649-665.

[9] A. Trantakis, T. K. Christopoulos, Stelios Spaniolas, P. Kalaitzis, P. C. Ioannou, G. A. Tucker; *Quantitative Bioluminometric Method for DNA-Based Species/Varietal Identification in Food Authenticity Assessment*. Journal of Agricultural and Food Chemistry (2012) 60(4), 912-916.

[10] V. Di Stefano, G. Avellone, D. Bongiorno, V. Cunsolo, V. Muccilli, S. Sforza, A. Dossena, L. Drahos, K. Vékey; *Applications of liquid chromatography-mass spectrometry for food analysis*. Journal of Chromatography A (2012) 1259 74-85.

[11] A. Taamalli, D. Arráez Román, M. Zarrouk, A. Segura-Carretero, A. Fernández-Gutiérrez; *Classification of 'Chemlali' accessions according to the*

geographical area using chemometric methods of phenolic profiles analysed by HPLC-ESI-TOF-MS. Food Chemistry (2012) 132 561-566.

[12] A. Damascelli, F. Palmisano; *Sesquiterpene Fingerprinting by Headspace SPME-GC-MS: Preliminary Study for a Simple and Powerful Analytical Tool for Traceability of Olive Oils*. Food Analytical Methods (2012).

[13] S. Wold, M. Sjöström; *Chemometrics, present and future success*. Chemometrics and Intelligent Laboratory Systems (1998) 44 3-14.

[14] A. Rohmanab, Y. B. C. Man; *Application of Fourier Transform Infrared Spectroscopy for Authentication of Functional Food Oils*. Applied Spectroscopy Reviews (2012), 47(1), 1-13.

[15] F. Vogt, R. D. Luttrell, J. Rodgers; *New Approaches for Field Analyses of Cotton Quality by Means of Near-IR Spectroscopy Supported by Chemometrics*. Analytical Letters (2011) 44 (14) 2466-2477.

[16] G. Downey, R. Briandet, R. H. Wilson, E. K. Kemsley; *Near- and Mid-Infrared Spectroscopies in Food Authentication: Coffee Varietal Identification*. Journal of Agricultural and Food Chemistry (1997), 45 (11), 4357-4361.

[17] R. Consonni, L. R. Cagliani, C. Cogliati; *NMR based geographical characterization of roasted coffee*. Talanta (2011) 88, 420-426.

[18] L. Rock; *The use of stable isotope techniques in egg authentication schemes: A review*. Trends in Food Science & Technology (2012), in press.

[19] B.L. Guo , Y.M. Wei, J.R. Pan, Y. Li; *Stable C and N isotope ratio analysis for regional geographical traceability of cattle in China*. Food Chemistry (2012) 118 (4) 915-920.

[20] G. Vinci, R. Preti, A. Tieri, S. Vieri; *Authenticity and quality of animal origin food investigated by stable-isotope ratio analysis*. Journal of the Science of Food and Agriculture (2012).

[21] R. A. Fischer; *The use of multiple measurements in taxonomic problems*. Annals of Eugenics (1936) 7 179-188.

[22] AGER 2009, *Agroalimentare e Ricerca: "New analytical methodologies for varietal and geographical traceability of oenological products*", contract n. 2011 - 0285.

[23] R. Bro; Multiway calibration - *Multilinear PLS*. Journal of Chemometrics (1996) 10 47-61.

[24] C. Durante, R. Bro, M. Cocchi; *A classification tool for N-way array based on SIMCA methodology*. Chemometrics and Intelligent Laboratory Systems (2011) 106 (1), 73-85.

[25] M. Barker, W. Rayens; *Partial Least Squares for Discrimination*. Journal of Chemometrics (2003), 17 166-173.

[26] S. Wold; *Pattern recognition by means of disjoint principal component models*. Pattern Recognition (1997) 8 127-139.

[27] S. Wold, M. Soöström; *SIMCA: A Method for Analyzing Chemical Data in Terms of Similarity and Analogy*. Chemometrics: Theory and Application (1977) 12 243-282.

[28] R. Tauler; *Multivariate curve resolution applied to second order data*. Chemometrics and Intelligent Laboratory Systems (1995), 30 (1) 133-146.