MULTIVARIATE DATA ANALYSIS IN ANALYTICAL CHEMISTRY

Majek Pavel[1], Majekova Magdalena[2]

[1]Institute of Analytical Chemistry, Faculty of Chemical and Food Technology, Slovak University of Technology in Bratislava, Slovakia

[2]Center of Experimental Medicine of SAS, Institute of Experimental Pharmacology & Toxicology, Slovak Academy of Sciences, Bratislava, Slovakia

pavel.majek@stuba.sk

Introduction

Multivariate data analysis, MVA is the investigation of many variables, simultaneously, in order to understand the relationships that may exist between them. Analytical chemistry and chemical analysis in the laboratories are mainly associated with the development of chemistry as a scientific discipline and with the progress in techniques and instrumentation because even a simple analyte often represents a multicomponent system. The result of the analysis is a huge data set and its structure is described by a number of variables that contribute to the overall information about the object being investigated. Data handling of this set is performed by chemometrics, where multi-dimensional statistical analysis and graphic visualization represent one of its most important parts.

Aim of study

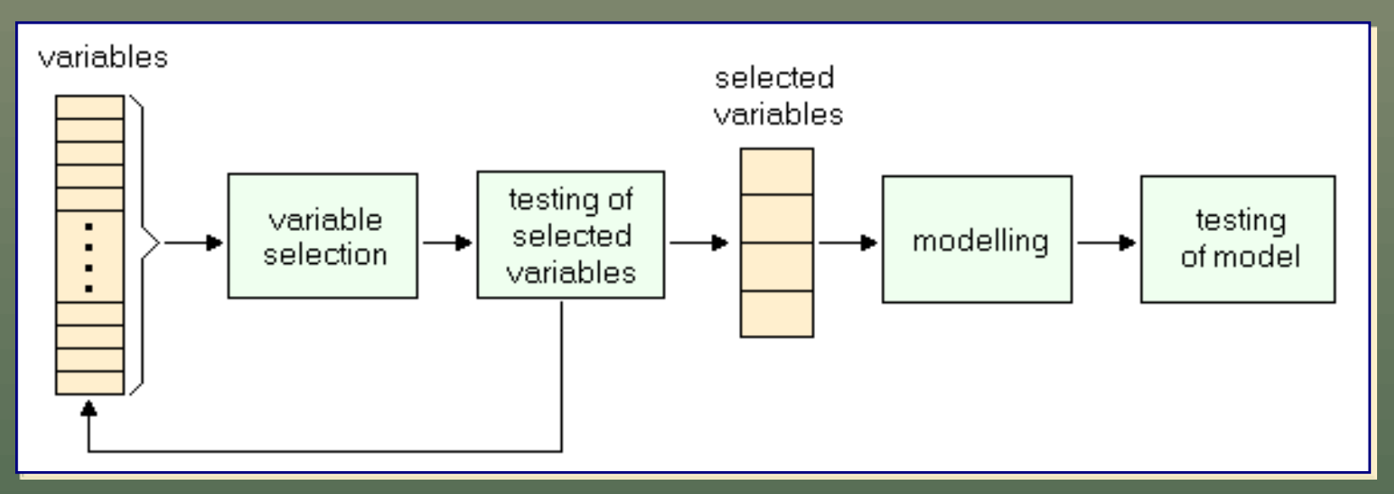
❑ ***Data types:*** a matrix sample data set obtained from measurements, create multi-dimensional or multiway data, Table 1, which is the source for data mining. Number of variables, their structure and the amount of data play crucial role which method of MVA will be used for data evaluation and which sort of information (*qualitative* or *quantitative*) will be discovered.

Table 1 Classification of data sets by array and statistics type.

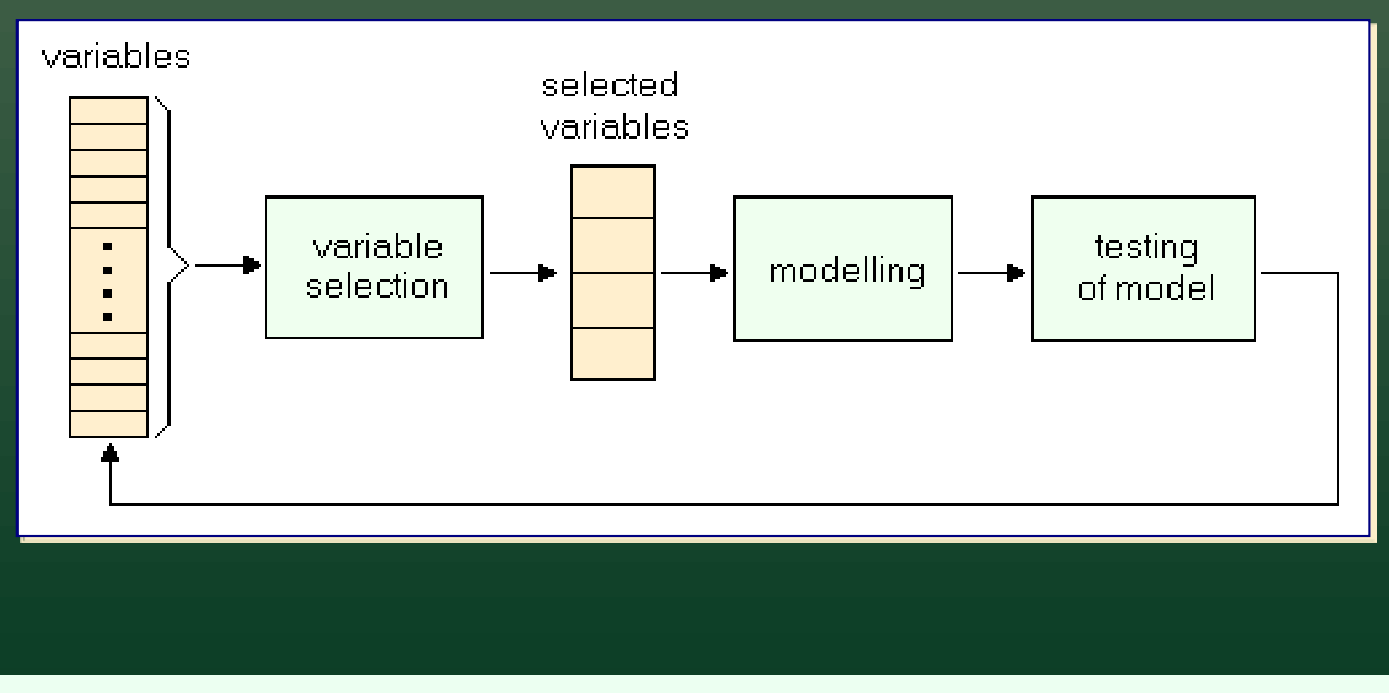
Data order	Tensor order	A sample set	Array	Statistics
p	1	vector	one-way	univariate
p × r	2	matrix	two-way	bivariate
p × r × q	3	3d array	three-way	multivariate
p × r × q × s	4	4d array	four-way	

❑ ***Variable and feature selection:*** is intended to select the *best subset of predictors* – to explain the data in the simplest way (principle of Occam's Razor).

⌖ ***variables selection off-line*** – chemical knowledge, variable ranking, Fisher ratio, correlation



⌖ ***variables selection is included in modeling*** – stepwise regression, discrimination, neural network, genetic algorithm



Example: dataset of 1000 molecules was described by 150 descriptors (experimental, software: ACD/Labs, Dragon 5, Gaussian 09); by variable selection the number of parameters decreased to 50; based on similarity index 6 groups of molecules were obtained. Using stepwise regression and genetic algorithm for the training set in an each group of molecules the number of descriptors reduced from 6 to 8 significant variables.

The given descriptors were apply for prediction of retention time of not measured compounds in dataset by multilinear regression (MLR) and neural networks (ANN) in LC [1].

❑ ***Design of experiment:*** effective realization of the experiment and developing of optimal methods are the basis of a laboratory practice. Helps us to investigate the effects of input ***k*** variables (factors) with ***h*** levels on an output variable (response) at the same time. DOE problem's areas are: (*i*) comparative, (*ii*) screening or characterizing, (*iii*) modeling, (*iv*) optimizing.

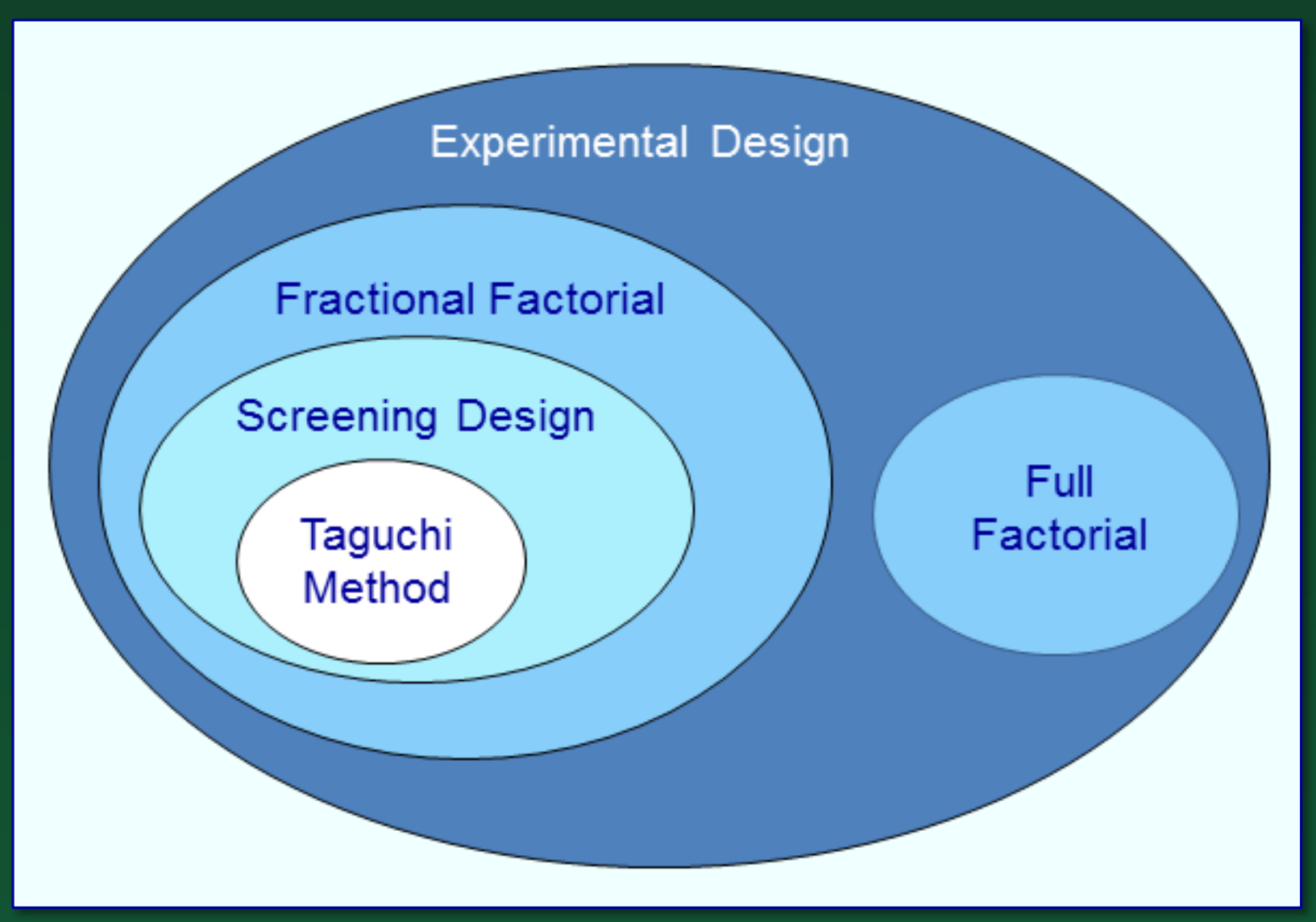


Fig. 1 Types of design of experiments.

Response Surface Methods (RSM): are suitable for quantification and interpretation of the relationship between the response and effect of factors. Common models are quadratic or cubic polynomials where response ***y*** is the function of variables or factors ***x*** as follows:

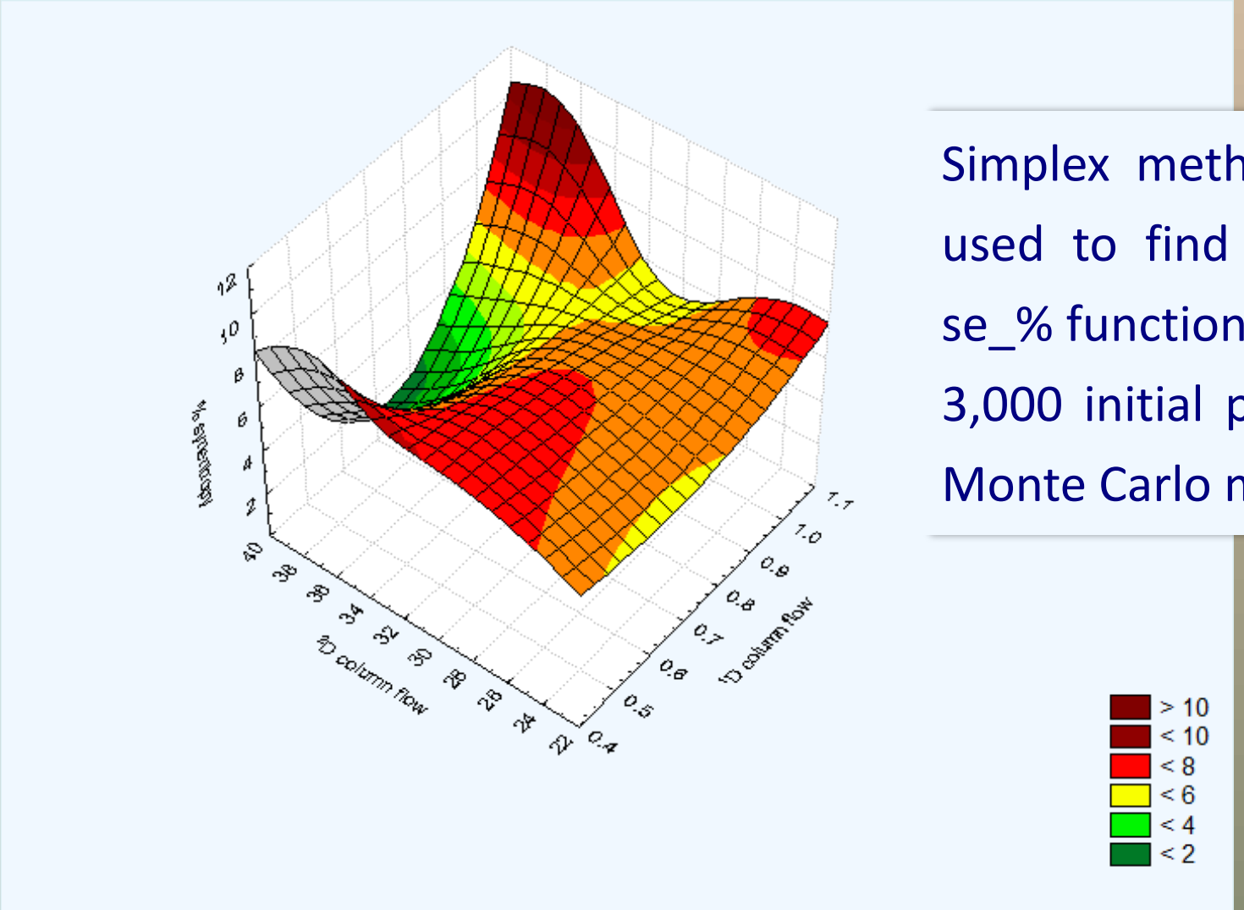
$$y = b_0 + \sum_{i=1}^{k} b_i x_i + \sum_{1 \le i \le j}^{k} b_{ij} x_i x_j + \sum_{i=1}^{k} b_{ii} x_i^2$$

where *k* – number of variable (factors), b_0 – intercept, b_i, b_{ii}, b_{ij} – regression coefficient.

Example: se_% – percent of informational entropy an synentropy were used to optimize flows in the first (f_1) and in the second (f_2) dimension as well as the temperature program rate (*r*) for the flow modulated GC×GC-FID separation of C6-C12 aromatic hydrocarbon in low boiling petrochemical sample. The synentropy hypersurface, obtained from full factorial design ***3³ = 27*** experiments, was describe by a quadratic equation with all interactions between optimized factors:

$$se\_\% = -0.87 + 25.34 f_1 + 1.27 f_2 - 9.26 r + 0.33 f_1^2 - 0.02 f_2^2 + 1.79 r^2 - - 0.71 f_1{*}f_2 - 5.14 f_1{*}r - 0.06 f_2{*}r + 0.14 f_1{*}f_2{*}r$$

Maximum se_% corresponds to the maximum peak distribution of C6-C12 aromatic hydrocarbons measured under optimized values of f_1, f_2 and *r* in GC×GC separation [2].



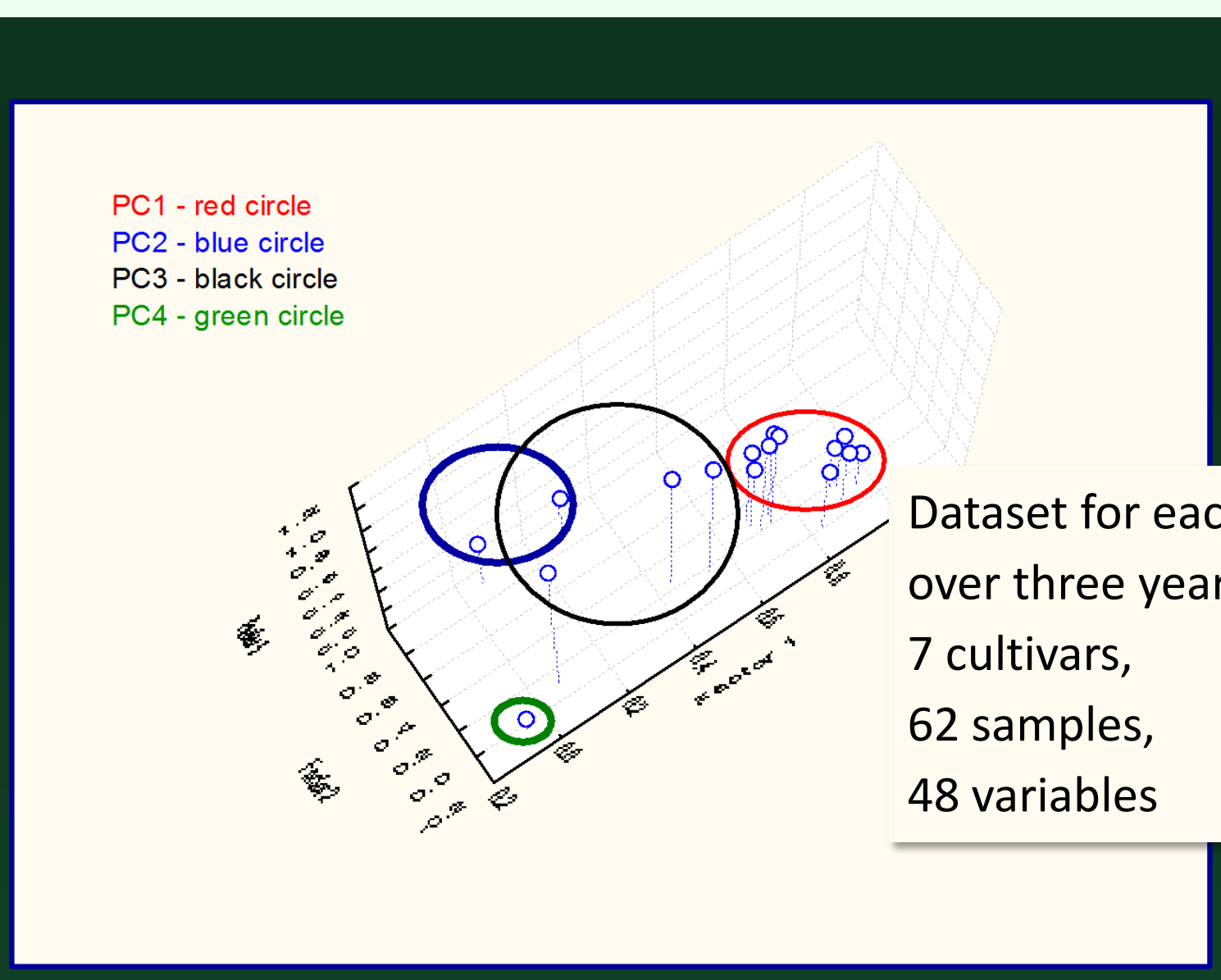
Simplex method with constraints was used to find global maximum of the se_% function with 11 parameters. 3,000 initial points were generated by Monte Carlo method.

Fig. 2 Experimental 3D plot of synentropy hypersurface described by above equation for the section plane at the temperature gradient *r* = 3.5.

❑ ***Multivariate data analysis***: enable us to find out patterns and relationships which determine the essential content of the total information in the large data set. Here there will be provide an overview of some MVA methods that represent the most suitable procedures of datamining and its application in analytical chemistry: principal component analysis (PCA), cluster analysis (CA), linear discriminant analysis (LDA), multilinear regression (MLR), partial least square regression (PLS) and artificial neural network (ANN).

Principal Component Analysis: is a tool to reduce multidimensional data to lower dimensions while retaining most of the information. The data matrix is decomposed to small matrices – the ***score matrix*** (*k* – PCs, the map of samples in PC coordinates) and ***loading matrix*** (influence of variables in responsible PC). PC_1 corresponds to the direction with the maximum amount of variation in the dataset.

Example: PCA extraction was applied for selection and determination of 16 significant anthocyanins, which characterize 92 % of total variance of their content in South African red wines [3].



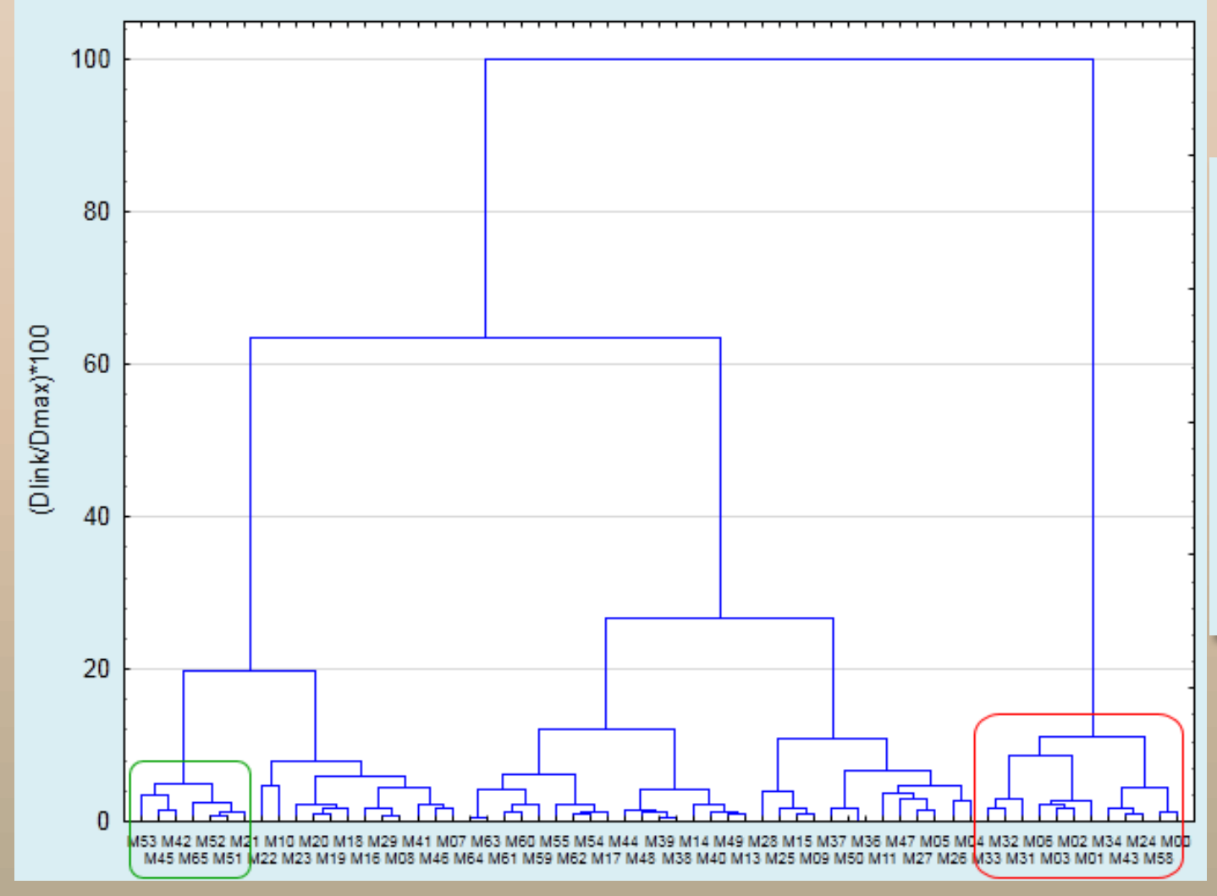
Dataset for each year over three years:
7 cultivars,
62 samples,
48 variables

Fig. 3 PCA extraction of anthocyanins derivatives in SA red wines.

Cluster Analysis: objects are aggregated stepwise according to the similarity of their feature. As a result, hierarchically or non hierarchically ordered clusters are formed. Measure of similarity is different for qualitative and quantitative variables and proximity measure depends on problem which is solved by CA: e.g. Minkowski distance, $\boldsymbol{d_{ij}}$ or similarity, $\boldsymbol{S_{ij}}$ measures for ***n*** variables (***i***, ***j*** indices for objects, ***r*** dimensionality) are defined:

$$d_{ij} = \left(\sum_{k=1}^{n} | x_{ik} - x_{jk} |^{r} \right)^{1/r} \qquad S_{ij} = 1 - \frac{d_{ij}}{d_{ij}(\max)}$$

Example: in the study of hexapyridoindole compounds (HPI) with neuroprotective effects: 61 derivatives of HPI were tested for the inhibition of oxidative impairment of creatine kinase. The stepwise MLR selection of descriptors (experimental and software: Dragon, Gaussian) with various sequence of F_{enter} and F_{remove} was used in the CA and Soergel distance D_{AB} as a measure of dissimilarity was adopted [4].



Subclusters:

red box – the lowest IC_{50} value,

green box – the upper IC_{50} value.

Fig. 4 Dendrogram of creatine kinase experiment, linkage rule: Ward's method, distances: matrix dissimilarities.

Discriminant Analysis: builds a predictive model for known group membership – analyze data when the dependent variable is categorical and the independent variable is interval in nature. Linear discriminant functions (DFs) are linear combinations of the original measured variables and then can be applied to new cases that have measurements for the predictor variables but have unknown group membership.

Example: stepwise LDA was performed for volatile compounds of white wine where two DFs provide 97.4 % correct predictions for all analyzed white wines [5].
Dataset: 3 cultivars, 38 samples, 11 variables in each year over three years.

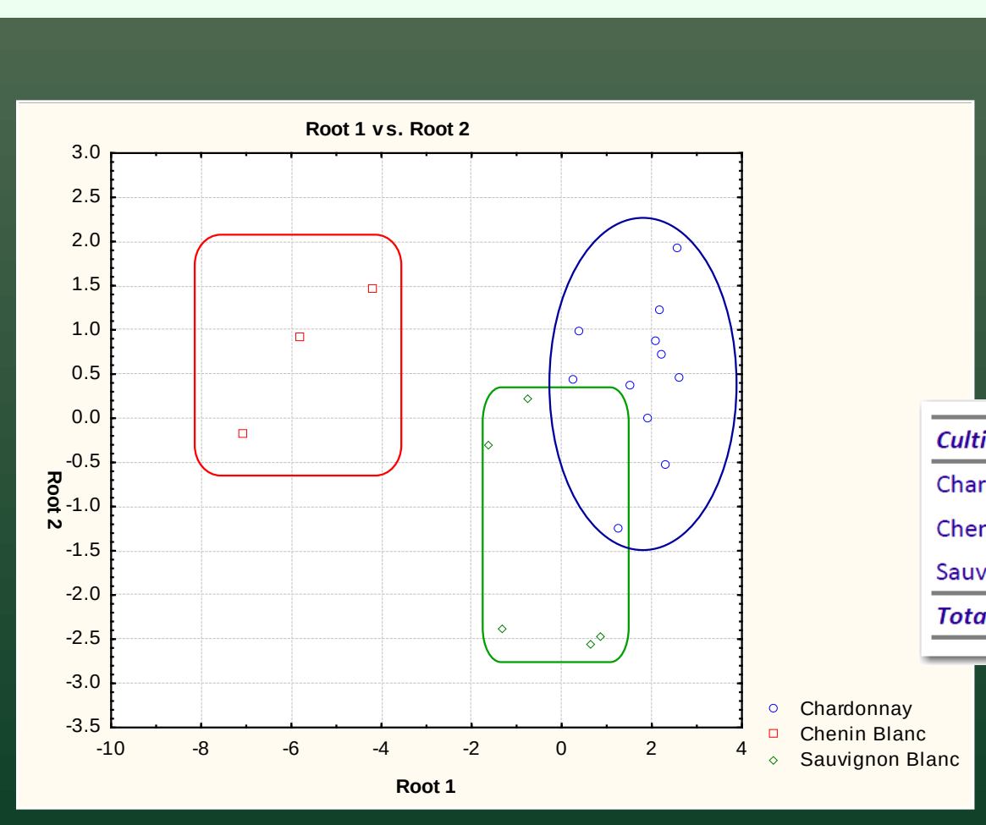


Table 1 Classification matrix for white wine, stepwise LDA, rows: observed classifications, columns: predicted classifications.

Cultivar	Percent	Chardonnay	Chenin Blanc	Sauvignon Blanc
Chardonnay	100	15	0	0
Chenin Blanc	90	0	9	1
Sauvignon Blanc	100	0	0	13
Total	*97.37*	*15*	*9*	*14*

Fig. 5 Scatter plot of canonical score DF1 vs. DF2 for SA white wines.

Multivariate calibration, MVC: is suitable for determining multiple components in mixture simultaneously. MVC are inverse calibration methods where concentrations are treated as functions of responses. The most useful MVC methods are: multilinear regression (MLR), principal component regression (PCR), partial least square regression, (PLS), neural network (NN).

Example: PLS calibration model of fluoxetine enantiomers in the presence of *β*-CD as chiral selector, obtained from UV/VIS spectral data, is alternative method to routine chiral analysis and control of enantiomeric composition for enantiomers of fluoxetine in racemic mixture [6].

References:

1. Majek, P., Szucz, R.: Pfizer Analytical Research Center, Ghent University, 2010, *unpublished results*.
2. Majek, P., Krupcik, J., Gorovenko, R. et al.: *Computerized optimization of flows and temperature gradient in flow modulated comprehensive two-dimensional gas chromatography*. J. Chromatogr. A, 1349, 135-138 (2014).
3. de Villiers, A., Vanhoenacker, G., Majek, P. et al.: *Determination of anthocyanins in wine by direct injection liquid chromatography-diode array detection-mass spectrometry and classification of wines using discriminant analysis*. J. Chromatogr. A 1054 (1-2), 195-204 (2004).
4. Majekova, M., Kovacikova, L., Stolc, S. et al.: *The use of advanced statistical methods in the study of pyridoindole compounds with neuro-protective effects*. 19th EuroQSAR, Abstract Book, University of Vienna, p. 195 (2012).
5. Tredoux, A., de Villiers, A., Majek, P. at. al.: *Stir bar sorptive extraction combined with GC-MS analysis and chemometric methods for the classification of south African wines according to the volatile composition*. J. Agric. Food Chem. 56 (12), 4286-4296 (2008).
6. Polacek, R., Majek, P.: *UV/VIS spectrometry with multivariate calibration as a new approach for the chiral analysis of fluoxetine*. IJDR 5(8), 5240-5244 (2015).

Acknowledgements: This work was supported by the Scientific Grant Agency of the Ministry of Education of Slovak Republic and the Slovak Academy of Sciences VEGA 2/0127/18 and the Slovak Research and Development Agency under the contract No. APVV-15-0455.