Scientific paper

Classification of White Varietal Wines Using Chemical Analysis and Sensorial Evaluations

Katja Šnuderl,[1] Jan Mocak,[2,*] Darinka Brodnjak-Vončina[3] and Bibiana Sedláčková[4]

[1] *Metrology Institute of the Republic of Slovenia, Tkalska 15, SI-3000 Celje, Slovenia*

[2] *University of Ss. Cyril and Methodius, Faculty of Natural Sciences, Nam. J. Herdu 2, SK-91701, Trnava, Slovakia*

[3] *Faculty of Chemistry and Chemical Engineering, Smetanova 17, SI-2000 Maribor, Slovenia*

[4] *Strečnianska 16, 85105 Bratislava, Slovakia*

* *Corresponding author: E-mail: jan.mocak@ucm.sk*

Received: 16-07-2008

Abstract

The ways of application of multivariate data analysis and ANOVA to classification of white varietal wines are here demonstrated. Wine classification was performed using the following classification criteria: wine variety, year of production, wine producer, and wine quality, as found by sensorial testing (bouquet, colour, and taste). Subjective wine evaluation, made by wine experts, is combined with commonly used chemical and physico-chemical properties, measured in analytical laboratory. Importance of the measured variables was determined by principal component analysis and confirmed by analysis of variance. Linear discriminant analysis enabled not only a very successful wine classification but also prediction of the wine category for unknown samples. The wine categories were set up either by three wine varieties, or two vintages, wine producers; two or three wine categories established by wine quality reflected either total points obtained in sensorial evaluation or the points obtained for a particular quality descriptor like colour, taste and bouquet.

Keywords*:* Multivariate data analysis; Principal component analysis; Discriminant analysis; Feature selection; ANOVA; Sensory analysis

1. Introduction

Wine belongs to the commodities, which are very frequent objects of falsification.[1,2] Wine is considered falsified when it has not been made in accordance with a specified method but is presented as a valuable product under an official trademark or when it's declared location of production is not true. Therefore it is necessary to develop procedures which make possible wine classification and authentication, i.e. verification of the selected sample with regard to the wine variety.[1–4] In addition; wine classification according to its producer or locality as well as year of production is also frequently demanded.

Methods of multivariate (multidimensional) data analysis (MVA) use multidimensional statistics for investigation of relations and interactions inside a large table of data.[5,6] They are often employed in analysis of food, natural substances, or environment. For wine classification the MVA methods are especially useful.[2,7–9] Measured or observed wine properties represent variables, which characterize the studied wine sample (generally considered as an object). Each variable can be regarded geometrically as an axis in the multidimensional space defined by all variables. Then each wine sample represents a point in this multidimensional space and its coordinates are given by the corresponding values of the used variables.

Performed research has been focused on several possible ways of classification of white varietal wines, based on the results of chemical analysis. Three typical kinds of Slovak white varietal wines, Welsch Riesling, Grüner Veltliner and Chardonnay, were analyzed during two consecutive years using eighteen selected chemical and physico-chemical descriptors (variables), most frequently used for wine characterization. In addition, sensorial analysis of all examined wine samples was provided. This study concerns an optimal choice of variables with regard to their effective use for wine classification and employs several methods of multivariate data analysis. Significant differences among the studied wine samples render a satisfactory characterization and classification of wines with respect to (a) variety, (b) vintage, and (c) sensorial quality. For this purpose, principal component analysis, discriminant analysis and ANOVA were used as the main chemometrical tools.

2. Experimental

2. 1. Wine Samples

Altogether 46 samples of varietal wines, namely Welsch Riesling (22 samples), Grüner Veltliner (18 samples), and Chardonnay (6 samples) of the vintages 1999 (22 samples) and 2000 (24 samples) were analyzed successively. The wines were produced by two Slovak producers, located in Bratislava and Hlohovec. Sampling was made by the Research Institute of Viticulture and Viniculture, Bratislava, Slovakia. Determination of the concentrations of commonly analyzed wine components was used for the chemical characterization of wine samples. All variables used in this study are listed in Table 1. Table 2 represents the averaged concentrations of physical and chemical parameters for each variety in both vintages of wine samples used for statistics. The codes 0, 1 and 2 stated in the "*Variety*" column denote Veltliner, Riesling and Chardonnay, respectively.

Table 1. Investigated characteristics of wine samples representing variables in chemometrical evaluation and their corresponding codes.

Code	Variable	Code	Variable	Code	Variable
v1	SO_2 free	v7	Tartaric acid	v13	Ethanol
v2	SO_2 total	v8	Lactic acid	v14	Total extract
v3	Total acidity	v9	Reducing sugars	v15	Sugar-free extract
v4	Volatile acidity	v10	Glucose	v16	Ash
v5	Citric acid	v11	Fructose	v17	pH
v6	Malic acid	v12	Density	v18	Polyphenols

2. 2. Sensorial Analysis

Sensorial analysis was made by a group of seven and eight wine experts for the vintages 1999 and 2000, respectively. They assessed the following wine properties: colour, limpidity, bouquet, and taste (where also the overall impression was evaluated) using in total a twenty-point scale, expressing the total sum of the acquired points. This overall evaluation was used as the main wine quality descriptor. The maximal number of the points ascribed to colour as well as limpidity was 2.0, that for bouquet was 4.0 and 12.0 for taste. This way of evaluation was commonly used by the Research Institute in the study of Slovak white wines. Since the wine limpidity was almost equal for the evaluated samples and all evaluators, this particular feature was omitted from the final data table. Consequently, the finally evaluated sensorial criteria were colour, bouquet, taste and total points.

In order to obtain two or three wine categories distinguished by sensorial quality, all wine samples were sorted by the given total points and then the median as well as the lower and upper terciles were calculated. According to the median the wine samples were categorized into two groups of 25 better (denoted as "good") and 21 worse ("bad") wines. The use of the terciles (i.e. the percentiles 0.3333 and 0.6667) resulted in three groups of the evaluated wine samples: 17 "good", 15 "medium" and 14 "bad". Unequal class memberships are due to the assignment of the border values to one of the groups. In fact, "good" and "bad" denote first-class and not fully superior sensorial features of the examined wine samples, respectively, and are used as such only as the labels.

2. 3. Analytical Methods

All analytical methods were made according to Slovak Technical Standards STN 560216, which conform European Union Council Regulations No. 2679/90 of 17 Sept. 1990 determining methods for the analysis of wines. Iodometric titration methods were applied for determinations of free and total sulphur dioxide as well as for determination of reducing sugars. Potentiometric methods using glass and saturated calomel electrodes were employed for determination of total acidity, volatile acidity and pH; volatile acids were separated from the wine by steam distillation and then the distillate was titrated by sodium hydroxide solution in a way similar to the total acid determination. A calibrated pycnometer was used for measuring density as well as for the density measurement of the distillate in ethanol determination. The difference of the total extract, finally obtained also by pycnometry, and the determined content of reducing sugars were used for calculating the value of the sugar-free extract. Ash was determined by ignition of the wine extract at 550 °C followed by a gravimetric endpoint. Citric, malic, tartaric and lactic acids were determined enzymatically with a final spectrophotometric determination. Glucose and fructose were

Table 2. Concentration of wine components and the selected physical-chemical properties representing variables in multivariate data analysis of wines.

Ordinal number	Variety	Vintage	Total points	v1	v2	v3	v4	v5	v6	v7	v8	v9	v10	v11	v12	v13	v14	v15	v16	v17	v18
1	0	1999	17.60	35.0	94	6.00	0.29	0.12	0.30	3.72	1.93	1.00	0.17	0.22	0.99	10.99	19.80	18.80	2.05	3.38	260
2	0	1999	17.85	25.0	117	5.60	0.35	0.09	0.37	3.07	1.81	1.10	0.18	0.21	0.99	10.96	18.00	16.90	1.98	3.38	260
3	0	1999	17.35	31.0	82	5.60	0.40	0.24	0.21	2.95	2.06	1.10	0.32	0.03	0.99	11.23	19.60	18.50	1.92	3.44	280
4	0	1999	17.80	29.0	114	5.40	0.36	0.09	0.21	2.75	1.75	1.10	0.32	0.02	0.99	11.20	17.60	16.50	1.89	3.42	275
5	0	1999	17.70	20.0	102	7.40	0.74	0.14	0.21	4.75	2.01	1.10	0.26	0.05	0.99	11.86	20.90	19.80	1.81	3.33	310
6	0	1999	17.75	45.0	95	5.70	0.43	0.24	0.33	3.06	2.04	1.10	0.33	0.06	0.99	11.88	18.80	17.70	1.84	3.35	240
7	0	1999	17.95	23.0	83	4.00	0.41	0.32	0.11	2.74	1.59	1.10	0.30	0.19	0.99	11.36	17.80	16.70	1.89	3.28	210
8	0	1999	17.80	26.0	68	5.60	0.37	0.08	0.18	3.22	1.88	1.00	0.27	0.34	0.99	11.98	18.00	17.00	1.84	3.32	230
9	0	1999	18.10	21.0	108	6.30	0.32	0.27	2.20	2.86	0.28	1.10	0.21	0.35	0.99	11.20	17.30	16.20	1.94	3.28	195
10	1	1999	17.75	31.0	74	6.40	0.31	0.14	0.33	3.57	1.79	3.90	0.30	0.12	0.99	11.46	19.00	15.10	1.55	3.14	215
11	1	1999	17.95	22.0	53	5.70	0.32	0.20	0.26	3.12	1.78	3.70	0.23	0.38	0.99	11.28	18.80	15.10	1.60	3.35	170
12	1	1999	17.85	45.0	80	5.90	0.54	0.10	0.20	3.14	1.85	3.80	0.25	0.45	0.99	11.53	19.00	15.20	1.61	3.36	185
13	2	2000	18.00	57.0	120	6.00	0.44	0.17	0.84	2.01	2.05	3.40	0.60	3.15	0.99	13.80	25.30	21.90	1.56	3.22	221
14	2	2000	17.60	16.0	222	5.50	0.53	0.22	1.70	2.26	0.83	3.40	0.75	2.98	0.99	13.49	19.80	16.40	1.74	3.24	293
15	2	2000	17.20	13.0	278	6.10	0.49	0.30	2.40	2.26	0.77	5.40	1.42	4.50	0.99	13.31	24.00	18.60	1.94	3.15	293
16	0	2000	17.60	61.0	127	8.90	0.36	0.35	1.66	5.59	0.93	1.30	1.93	2.97	0.99	12.63	22.70	21.40	1.83	2.80	194
17	0	2000	17.55	44.0	273	8.40	0.54	0.39	1.62	4.82	0.70	14.40	1.44	10.58	1.00	11.97	33.60	19.30	1.95	2.69	241
18	0	2000	18.00	24.0	236	8.10	0.51	0.32	1.59	5.01	0.69	14.70	1.51	12.05	1.00	11.66	34.40	19.80	1.60	2.75	226
19	0	2000	17.55	69.0	124	5.10	0.49	0.32	0.33	3.10	1.60	1.30	0.30	0.82	0.99	13.27	18.80	17.50	1.68	3.23	218
20	0	2000	17.75	19.0	153	4.80	0.33	0.40	0.88	2.92	0.86	0.80	0.75	1.41	0.99	12.95	17.10	16.30	1.77	3.20	241
21	0	2000	17.35	111.0	202	5.20	0.45	0.29	0.44	3.01	1.33	4.50	0.92	3.47	0.99	13.11	22.70	18.20	1.96	3.11	188
22	0	2000	17.70	35.0	167	4.80	0.32	0.38	0.75	2.90	0.66	0.70	1.03	1.24	0.99	12.80	17.80	17.20	1.75	3.13	248
23	0	2000	17.90	39.0	159	4.90	0.31	0.38	0.68	2.96	0.71	0.90	2.10	0.67	0.99	13.28	17.10	16.30	1.69	3.12	251
24	1	2000	17.70	63.0	147	6.30	0.46	0.27	1.23	2.40	0.97	31.60	2.63	13.65	1.00	13.08	50.70	19.10	1.73	3.17	307
25	1	2000	18.00	18.0	96	4.00	0.34	0.20	0.93	2.34	0.43	1.10	1.50	1.09	0.99	14.90	17.10	16.00	1.43	3.39	237
26	0	1999	18.05	36.0	99	7.30	0.50	0.24	1.93	4.57	0.31	0.70	0.25	0.03	0.99	11.41	18.30	17.60	1.85	3.20	185
27	0	1999	18.10	35.0	104	8.00	0.36	0.23	2.17	4.62	0.21	1.00	0.28	0.03	0.99	11.93	18.80	17.80	1.81	3.20	182
28	0	1999	18.10	23.0	83	4.90	0.54	0.35	0.19	2.68	1.70	1.00	0.30	0.04	0.99	11.87	18.50	17.50	1.88	3.36	220
29	0	1999	18.30	24.0	122	6.80	0.42	0.31	2.20	2.97	0.33	1.00	0.21	0.34	0.99	11.94	18.50	17.50	1.95	3.31	210
30	1	1999	18.10	40.0	100	7.20	0.32	0.26	2.64	3.61	0.12	1.70	0.27	0.04	0.99	11.74	17.60	15.90	1.74	3.35	280
31	1	1999	17.90	39.0	100	7.10	0.32	0.24	2.53	3.61	0.21	2.10	0.30	0.04	0.99	11.70	17.80	15.70	1.71	3.42	177
32	1	1999	18.00	35.0	97	6.20	0.32	0.27	1.86	3.20	0.22	0.90	0.21	0.37	0.99	11.79	16.50	15.60	1.72	3.42	155
33	1	1999	18.20	45.0	95	6.00	0.43	0.26	1.88	3.12	0.15	1.00	0.22	0.38	0.99	11.17	15.80	14.80	1.72	3.35	157
34	1	1999	18.05	32.0	90	6.20	0.41	0.27	1.74	3.10	0.16	0.80	0.26	0.40	0.99	11.76	16.30	15.50	1.71	3.28	167
35	1	1999	18.10	40.0	97	5.90	0.37	0.26	1.85	3.07	0.17	1.00	0.37	0.36	0.99	11.76	16.30	15.30	1.70	3.32	165
36	2	2000	18.35	9.0	197	5.60	0.57	0.35	1.53	2.24	0.67	2.00	0.77	2.82	0.99	14.10	19.30	17.30	1.61	3.30	279
37	2	2000	18.40	69.0	136	5.40	0.34	0.35	1.93	1.98	0.82	5.00	0.72	3.61	0.99	13.98	22.90	17.90	1.64	3.32	210
38	2	2000	18.10	25.0	228	6.00	0.49	0.41	1.73	2.38	0.72	1.80	1.14	2.08	0.99	13.87	21.10	19.40	1.94	3.23	277
39	0	2000	18.10	25.0	137	4.90	0.29	0.41	0.94	2.99	0.66	0.90	0.79	1.38	0.99	13.17	17.30	16.40	1.66	3.18	214
40	1	2000	18.15	73.0	139	7.60	0.35	0.30	2.59	3.49	0.74	1.00	1.21	1.20	0.99	13.40	23.50	22.60	1.66	2.87	206
41	1	2000	18.05	15.0	143	7.80	0.49	0.22	2.32	4.01	0.48	3.10	1.46	3.40	0.99	13.06	19.60	16.50	1.56	2.88	211
42	1	2000	18.45	25.0	160	7.80	0.62	0.28	2.41	3.53	0.62	3.50	1.46	3.40	0.99	13.01	21.90	18.40	1.63	2.84	194
43	1	2000	18.40	34.0	102	5.60	0.45	0.23	1.42	2.78	0.60	15.70	1.49	3.37	1.00	14.01	34.90	19.20	1.42	3.17	265
44	1	2000	18.25	33.0	139	4.40	0.33	0.30	0.87	2.27	0.54	0.80	1.96	10.27	0.99	14.95	17.60	16.80	1.46	3.20	239
45	1	2000	18.20	55.0	170	4.40	0.34	0.31	0.89	2.26	0.55	0.90	1.80	0.54	0.99	14.93	17.80	16.90	1.61	3.20	242
46	1	2000	18.20	19.0	99	3.90	0.33	0.20	0.94	2.00	0.69	1.00	1.57	0.57	0.99	14.90	16.80	15.90	1.39	3.32	241

Note: Variables v1, v2, v18 are measured in mg/L, variables v3 – v11 and v14 – v17 are measured in g/L, variable v12 in g/cm^3, variable v13 in vol. %.

determined also enzymatically with a final spectrophotometric measurement of the reaction products. Total polyphenols were determined by Folin-Ciocalteu assay with a spectrophotometric endpoint.

2. 4. Statistical Analysis

Statistical treatment of the obtained data was performed using program packages SYSTAT 9 (SPSS Inc., Chicago, U.S.A.), STATGRAPHICS Plus 5.0 (Manugistics, Inc., Rockville, U.S.A.), S-PLUS v. 4.0 (Insightful Corp., Seattle, WA, U.S.A.) and Microsoft EXCEL. SYSTAT was used for calculations of linear discriminant analysis and ANOVA. For the latter task, a general linear model (GLM) option was used, enabling various ways of ANOVA, Smirnov-Kolmogorov test of the data normality and Bonferroni test of means for each possible pair of factors corresponding to the selected classification criteria. Calculations of principal component analysis were performed by STATGRAPHICS. The S-PLUS package was exclusively used for bootstrapping, by which 1000 replications were generated for each of the six wine sample groups – given by three wine varieties and two basic classes of wine quality (“good” and “bad”). From the normal distributions, generated for all six groups, seven octiles (0.125, 0.250, 0.375, 0.500, 0.625, 0.750, 0.875 percentiles) were calculated and used as the computer generated wine samples. Altogether 42 test samples were generated in this resampling procedure. MS EXCEL was used for the data preparation, percentile calculation and summarization of the results.

3. Results and discussion

3. 1. Principal Component Analysis

Principal component analysis (PCA) is a basic way used for characterizing multidimensional data, providing a satisfactory representation of the studied objects by projecting the original data set from the high dimensional space onto the lower dimension space. Often two or three most important principal components (PC’s), calculated by linear combination of original variables, sufficiently represent the total variability of the original data. This MVA technique does not need any training set of data (in which the categorization of the objects into the selected classes is known) and represents unsupervised learning.[10]

In our case, the first two PC’s, calculated from all variables, account for 41.7% of the total data variability as shown in Fig. 1, which represents the position of the samples of three wine varieties and two vintages. Even though the data categories are not involved in the PCA calculations, it was practical to mark the wine samples by the category where they belong to so that it might be possible to recognize some natural grouping of the studied wines. This approach was then applied to all considered classification criteria.

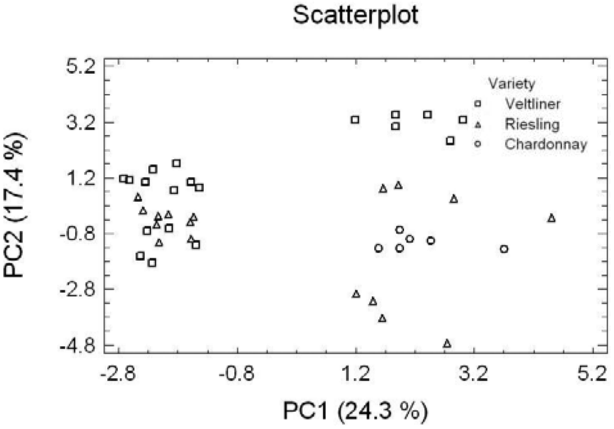

Figure 1. The PCA dependence PC2 vs. PC1. Three wine varieties are marked by different symbols. The left cluster of points belongs to vintage 1999, the right one to vintage 2000.

Inspecting Fig. 1, it is worth to note that the wine clusters observed on the PC2 vs. PC1 graph are not created by the wine varieties but correspond mainly to the vintage categories. The wine samples located on the PC1 axis below –1.0 are from the year 1999; those from the year 2000 are above +1.0. In fact two influencing factors are mixed in this PCA representation, vintage and variety. Since the PC’s are always arranged in a hierarchical order (according to their information content) the plane PC2–PC1 is often satisfactory for representing the position of the objects (wines) in a multidimensional space. Due to a relatively low cumulative percentage (41.7% for the first two PC’s) this is not true in the studied case and further PC’s would be needed for a more precise description.

Naturally formed clusters, reflecting the sensorial quality of the wine samples, are shown in Fig. 2. As demonstrated, the wines with larger total points (assessed as “good” wines) are predominantly located at the lower PC2

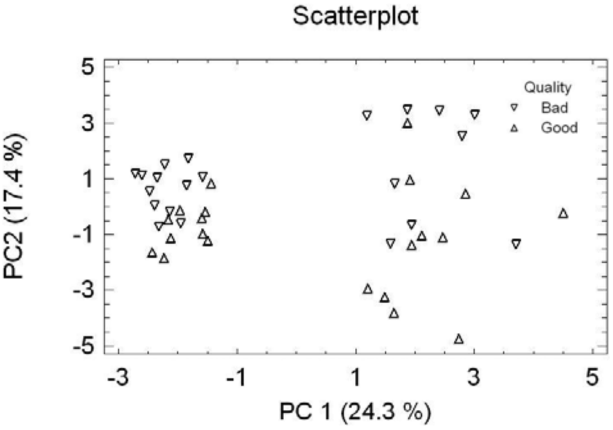

Figure 2. The PCA dependence PC2 vs. PC1 for two groups of wine samples created by sensorial assessment (total points) and designated by different symbols. The “good” wines are located mostly at the lower PC2 values. The left cluster belongs to vintage 1999, the right one to vintage 2000.

values, the samples of “bad“ wine are located at the higher PC2 values. Based on the PCA results it might be anticipated that the discrimination and prediction of the wine categories should be easiest when using the vintage criterion; the categorization of wine samples according to sensorial quality and variety of wines appears more difficult.

It should be again noted that the pictures depicted in the PC2–PC1 plane are simplifications of the multidimensional reality. In the studied case also further PC’s are important since the percentage of total variance is 13.8%, 12.5%, 7.1%, 5.7%, and 5.0% from PC3 to PC7, respectively.

3. 2. Discriminant Analysis

Linear discriminant analysis (LDA) is a supervised learning method, in which a classification model is constructed using the data of the objects pre-categorized into known categories (the training data set) and the calculation algorithm is trained to discriminate the objects (e.g. wine samples) into the given categories (classes).[6,10,11] The developed model is then used to classify the samples, which create the test set of data. In the investigated wine problem, several ways of classification were used since the categories were made according to the wine *variety*, *year of vintage*, *total sensorial quality* and even by several partial sensorial characteristics like *colour*, *taste* and *bouquet* of wine. The classification criteria used in all performed linear discriminant analyses are summarized in the first two columns of Table 3.

Success of classification (the number of the correctly classified samples to the total number), evaluated according to the *variety*, *vintage* as well as *producer* criteria for the training data set containing 46 wine samples was always 100%. Fig. 3 exemplifies such a very successful classification performed according to *variety*.

Classification success calculated by *sensorial quality* of wine was over 91% for the two-category classifications regardless the classification criterion and concerns *total points* or some partial sensorial component (*colour*, *bouquet* or *taste*). Since the three-classes classifications are more difficult, the corresponding classification results were less successful. The least successful categorization was that for three *colour* classes (*cf.* Table 3, columns 3 and 4).

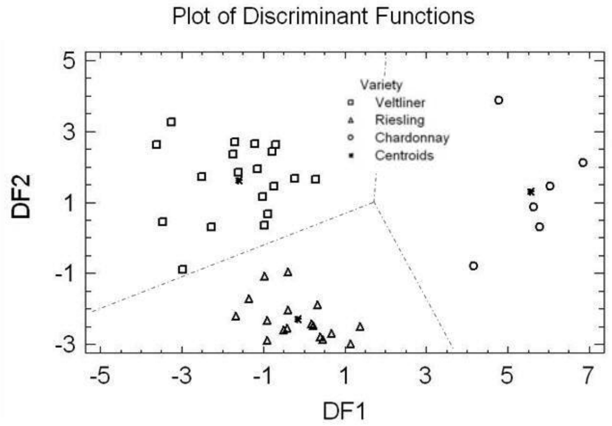

Figure 3. The plot of the discriminant functions DF2 vs. DF1 exhibiting the classification by the wine *variety* indicated by the used symbols. A 100% classification success of 46 wine samples was found.

Backward selection technique is one of the frequently applied feature reduction techniques.[6,12] The stepwise discriminate analysis was applied to the complete set of variables in order to select, in a consecutive way, those variables, which are most important with regard to the chosen classification criterion. In this process, the success of classification for every group of variables was evaluated and the best combinations are surveyed in the last two columns of Table 3.

It can be seen in Fig. 4 that only two best variables are sufficient for a 100% classification of wines by *vintage*. Similarly, only five best variables are sufficient for a 93% wine classification by *sensorial quality* as shown in Fig. 5.

Table 3. Criteria for wine classification and success in the LDA classification when all or the best variables were used.

Criterion	**Number of classes**	**Classification success in %**		**Codes of the selected best variables**
		All *	**Best ***	**(as given in Table 1)**
Variety	3	100.0	95.7 (3)	v7, v3, v16
Vintage	2	100.0	100.0 (2)	v14, v12
Quality	2	93.5	93.5 (5)	v8, v10, v13, v2, v3
Quality	3	87.0 †	78.3 (9)	v8, v18, v1, v7, v14, v9 v2, v5, v16
Colour	2	91.3	89.1 (7)	v8, v7, v9, v3, v2, v11, v16
Colour	3	76.1	78.3 (6)	v6, v7, v5, v10, v18, v1
Bouquet	2	97.8	93.5 (8)	v8, v7, v9, v14, v2, v4, v18, v1
Bouquet	3	93.5	84.8 (10)	v8, v2, v4, v7, v9, v12, v13, v5, v10, v17
Taste	2	91.3	91.3 (10)	v6, v7, v2, v4, v9, v17, v8, v14, v10, v11
Taste	3	87.0	91.3 (7)	v6, v7, v13, v2, v4, v1, v18
Producer	2	100.0	100.0 (4)	v7, v6, v16, v8

* "All" refers to 18 originally used variables. "Best" refers to the optimally selected variables with their number in brackets and the codes in the next column. † All variables except v6 and v15.

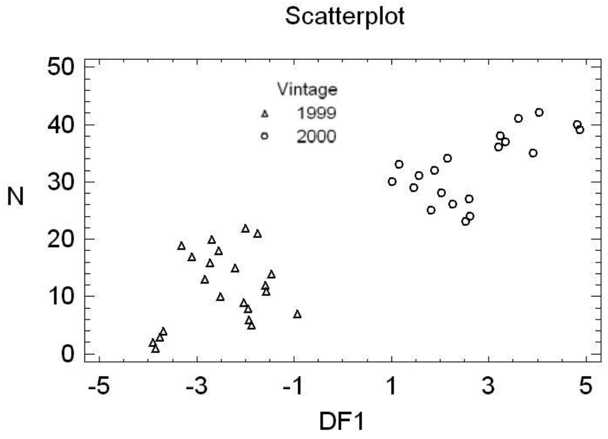

Figure 4. The ordinal sample number *N* vs. DF1 dependence shows the values of the first (and only) discriminant function DF1 for various *N*. The samples were here renumbered in order to better recognize two classes of wines differing by the *year of production*, 1999 vs. 2000. A 100% classification success for 46 wine samples was achieved by using only two best variables: ethanol (v13) and pH (v17).

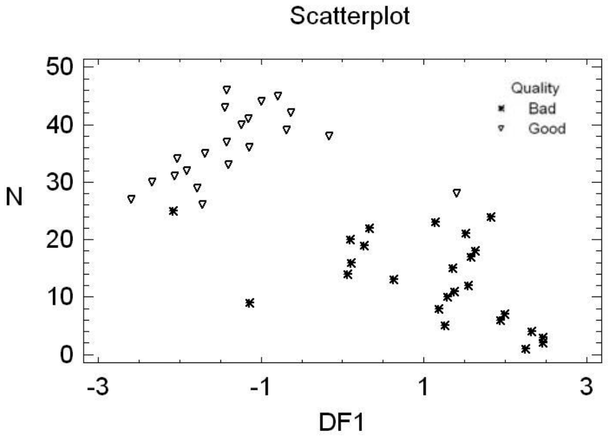

Figure 5. The sample number *N* vs. DF1 dependence exhibiting the values of the first (and only) discriminant function DF1 for various *N*. The samples were renumbered in order to better recognize two classes of wines differing by *sensorial quality* (total points). A 93.5% classification success for 46 wine samples was achieved using only five best variables: lactic acid (v8), glucose (v10), ethanol (v13), SO_2 total (v2), and total acids (v3).

Cross-validation of the calculated discriminant model was made in two different ways: (1) Using the leave-one-out (jack-knife) procedure, which can be performed also by SYSTAT software.[11–13] (2) Using 42 test samples generated by the resampling procedure.[11–14] However, due to an extensive way of calculations, this way of validation was made only for two classification criteria – *variety* and *sensorial quality* (using total points).

For 42 test samples, prepared by the bootstrap procedure and categorized by *variety*, a 100% classification success was found not only for all originally used variables but also for three best variables (v7, v3, and v16) determined by the feature selection technique. When total points were used for classification of wine by *sensorial quality* ("good" and "bad" wines), a 90.5% success was observed with all variables used and, even better, a 100% success was reached with 5 best variables (v8, v10, v13, v2, v3).

For the same classification criteria but using the leave-one-out validation technique the following results were obtained: 87.0% and 91.0% for all and seven best variables, respectively, when the wines were categorized by *variety*. The same way of validation and wine categorization into two classes by *sensorial quality* (total points) exhibited 78.0% and 87.0% success for all and five best variables, respectively.

3. 3. Analysis of Variance

Analysis of variance, ANOVA, makes it possible to evaluate independently the importance of any used variable with respect to the given classification criterion, which is here called a factor.[5,10] Modern software packages enable to calculate the *p*- values, indicating the probability of the null hypothesis stating that the effect of the factor on the particular variable does not exists; the most important factors have the lowest p-values. Therefore in Table 4 all *p*- values smaller than 0.05 are marked by bold faces to stress the importance of the corresponding factor for the given variable. It is understandable that the same factor may be important for some variable and unimportant for others. In the investigated problem, the received ANOVA results are in a reasonable agreement with the outcomes of the LDA calculations.

4. Conclusions

White varietal wines were classified in several ways, primarily by variety, vintage and the wine producer. An additional and hitherto rare way of classification was also performed, based on sensorial assessment of the wine quality, either using total points assigned to the wine sample or using partial sensory features like colour, bouquet and taste. All these properties were subjectively evaluated by a group of wine experts. Qualitative preview of natural grouping of wine samples in the space of the used variables was simplified by calculating first two principal components in the principal component analysis. ANOVA technique provided an additional assessment of variables which are most sensitive to the changes of the chosen individual factor, i.e. the wine characterizing criterion.

A quantitative discrimination among the wine samples according to all selected classification criteria was obtained by linear discriminant analysis. This technique allows assessing, which of the variables, representing basic chemical and physico-chemical characteristics of wines, are important for the chosen way of classification. Using stepwise variable selection it was possible to evaluate clas-

Table 4. The probability *p*- values in the ANOVA classification of wines using different classification factors and eighteen variables (v1 to v18) described in Table 1.

Variable Code	**Factor**							
	Variety	**Vintage**	**Producer**	**Quality(2)**	**Quality(3)**	**Taste**	**Bouquet**	**Colour**
	p							
v1	0.314	0.712	0.642	0.879	0.212	0.576	0.487	0.995
v2	**0.002**	**0.000**	0.569	0.892	0.886	0.989	0.783	0.705
v3	0.950	0.331	**0.000**	0.531	0.808	0.781	0.460	0.375
v4	0.156	0.404	0.182	0.834	0.448	0.480	0.685	0.142
v5	0.539	**0.000**	0.896	**0.027**	**0.013**	0.089	0.109	0.252
v6	**0.010**	**0.017**	**0.021**	**0.000**	**0.002**	**0.001**	**0.012**	**0.017**
v7	**0.000**	**0.047**	**0.000**	0.572	0.427	0.099	**0.047**	**0.013**
v8	**0.018**	0.645	0.147	**0.000**	**0.000**	**0.000**	**0.009**	**0.005**
v9	0.091	**0.012**	0.242	0.179	0.463	0.418	0.245	0.279
v10	0.181	**0.000**	0.578	0.780	0.799	0.960	0.488	0.933
v11	**0.023**	**0.000**	0.446	0.680	0.791	0.831	0.637	0.823
v12	0.464	0.612	0.083	**0.048**	0.077	0.057	**0.028**	**0.010**
v13	**0.002**	0.000	0.195	0.094	0.130	0.112	**0.018**	0.076
v14	0.604	0.003	0.161	0.164	0.323	0.313	0.389	0.099
v15	**0.019**	**0.004**	0.055	0.352	0.452	0.803	0.810	0.310
v16	**0.000**	**0.005**	0.827	**0.045**	**0.010**	0.131	**0.035**	**0.030**
v17	0.855	**0.000**	**0.000**	0.762	0.874	0.759	0.827	0.298
v18	**0.023**	**0.027**	0.402	**0.025**	**0.022**	0.146	0.055	0.130

* All *p*- values are rounded to three decimal places. The smaller *p*- value, the larger effect has the factor to the respected variable. Quality(2) and Quality(3) denote the two- and three- class criteria for evaluation of the wine samples by sensorial quality, respectively.

sification success in percents for a given set of the selected variables and find the optimum set of the employed variables. The classification model, created by discriminant analysis, was cross-validated either by the leave-one-out (jack-knife, holdout) procedure or by evaluating the classification success for the samples belonging to a special test data set. Due to the sample lack, the test samples were generated by the bootstrapping procedure. The established and validated discriminant model was finally used for the category prediction of the unclassified wine samples.

The obtained results of principal component analysis, ANOVA and discriminant analysis enable to state most important variables among the studied chemical and physical properties of wine samples even though the outputs of these methods are not equivalent; an imperfect method equivalency is due to a different way and weight by which the particular variable is influenced by other variables. In case of high correlation between variables only one of them is shown as important in discriminant analysis, the influence of the second one is hidden. Despite these facts, it is possible to give a significant overview on the role of individual wine properties in white wine classifications performed by various aspects, as shown below.

The variables most important with respect to all classification criteria were malic acid (v6), tartaric acid (v7) and lactic acid (v8). All effects mentioned here are considered regardless whether they influence the given criterion, e.g. the wine quality in a positive or in a negative way.

The levels of ethanol (v13), ash (v16) and polyphenols (v18) were pronounced as very important for classifications by all investigated principles except the producer.

The content of total sulphur dioxide (v2) was found important for classifications of white wines by variety and vintage; sugar-free extract (v15) and pH (v17) were found important for classifications by vintage and producer.

White wine classification by vintage is significantly influenced also by the following variables: citric acid (v5), reducing sugars (v9), glucose (v10), fructose (v11) and total extract (v14). Total acid concentration (v3) helps to classify wines only by the producer.

Without any significant influence were found sulphur dioxide free (v1) and volatile acids (v4).

A wider validity of the above mentioned results can be supposed taking into account differences in white wines caused by the variety, vintage and producer (differences in winemaking process as well as in geographic locality). The chemometrical approach, exemplified in this article, might be also useful when applied to the wine classifications based on entirely different wine characteristics, like amino acids profile,[15,16] polyphenol profile,[17,18] volatile aromatic compounds,[1,2] trace elements profile,[19] or protein fractions,[17,20] suggested in the past.

5. Acknowledgment

The support of the Slovak Grant Agency of Sciences to the projects VEGA 1/1005/09, 1/0066/09 and project APVV-0057-06 is gratefully acknowledged.

6. References

1. B. Medina, Food authentication, In: P. R. Ashurst, M. Dennis (Eds.): Wine authenticity, Chapman & Hall, London, **1996**, pp. 60–107.
2. J. Petka, J. Mocak, P. Farkas, B. Balla, M. Kovac, *J. Sci. Food Agr.* **2001**, *81*, 1533–1539.
3. M. Penza, G. Cassano, *Food Chem.* **2004**, *86*, 283–296.
4. A. P. Ferreira, J. A. Lopes, J. C. Menezes, *Anal. Chim. Acta* **2007**, *595*, 120–127.
5. D. L. Massart, B. G. M. Vandengiste, S. N. Deming, Y. Michotte, L. Kaufman, Handbook of Chemometrics and Qualimetrics: Part A, Elsevier, Amsterdam, **1997**, pp. 121–150, 519– 556.
6. S. Sharma, Applied Multivariate Techniques, Wiley, New York, **1996**, pp. 1–15, 59–66, 264–267, 287–293.
7. M. D. Huerta, M. R. Salinas, T. Masoud, G. L. Alonso, *J. Food Compos. Anal.* **1998***, 11*, 363–374.
8. S. Perez-Magarino, M. Ortega-Heras, M. L. Gonzales San-Jose, *Anal. Chim. Acta* **2002**, *458*, 187–190.
9. M. Urbano, M. D. Luque De Castro, P. M. Perez, J. Garcia-Olmo, M. A. Gomez-Nieto, *Food Chem.* **2006**, *97*, 166–175.
10. M. Otto, Chemometrics, Statistics and Computer Application in Analytical Chemistry, Wiley-VCH, Weinheim, **1999**, pp. 41–47, 124–168.
11. B. G. M. Vandengiste, D. L. Massart, L. M. C. Buydens, S. De Jong, P. J. Lewi, J. Smyers-Verbeke, Handbook of Chemometrics and Qualimetrics: Part B, Elsevier, Amsterdam, **1998**, Chapter 33, pp. 207–242.
12. J. W. Einax, H. W. Zwanziger, S. Geiss, Chemometrics in Environmental Analysis, Wiley-VCH, Weinheim, **1997**, Chapter 5, pp.139–203.
13. O. Parr Rud, Data Mining Cookbook, Wiley, New York, **2003**, Chapter 6, pp. 105–129.
14. B. Efron, R. Tibshirani. An Introduction to the Bootstrap, Wiley, New York, **1993**, pp. 168–177, 338–357.
15. A. M. P. Vasconcelos, H. J. Chaves das Neves, *J. Agric. Food Chem.* **1989,** *37*, 931–937.
16. P. Lehtonen, *Am. J. Enol. Vitic.* **1996**, *47*,127–133.
17. L. Almela, S. Javaloy, J. A. Fernandez-Lopez, J. M. Lopez-Roca, *J. Sci. Food Agric.* **1996**, *70*, 173–180.
18. E. Csomos, K. Héberger, L. Simon-Sarkadi *J. Agric. Food Chem.* **2002**, *50*, 3768–3774.
19. K. Pyrzynska, *Crit. Revs. Anal. Chem.* **2004**, *34*, 69–83.
20. E. Pueyo, M. Dizy, M. C. Polo, *Am. J. Enol. Vitic.* **1993**, *44*, 255–260.

Povzetek

Prikazali smo načine uporabe analize večdimenzionalnih podatkov in ANOVE pri klasifikaciji vin belih vrst. Klasifikacijo smo izvedli z uporabo več klasifikacijskih kriterijev, ki so vrsta vina, letnik, proizvajalec in kakovost vina kot je bila ocenjena s senzoričnim preskušanjem (cvetica, barva in okus). Uporabili smo kombinacijo subjektivnih ocen vzorcev vin s strani enologov in fizikalno kemijskih lastnosti vzorcev vin izmerjenih v analiznem laboratoriju. Vplivnost izmerjenih spremenljivk smo določili z metodo glavnih osi in rezultate potrdili z analizo variance. Z linearno diskriminantno analizo smo vzorce vin uspešno ustrezno uvrstili v razrede, kakor tudi neznane vzorce vin uvrstili v ustrezno kategorijo. Kategorije vina smo določili glede na tri vrste vina, na dva letnika in na dva proizvajalca. Dve ali tri kategorije smo določili glede na kakovost vina, ki je bila odraz skupne ocene senzorične analize vzorcev vin ali ocene vzorcev vin po posameznih deskriptorjih opravljene senzorične analize (barva, okus in cvetica).