CHEMOMETRIC CHARACTERIZATION AND CLASSIFICATION OF NEW WHEAT GENOTYPES

FILIP KRAIC[1], JÁN MOCÁK[1], TIBOR ROHÁČIK[2],
JANA SOKOLOVIČOVÁ[2]

[1]Department of Chemistry, University of SS. Cyril and Methodius, J. Herdu 2, Trnava, SK-917 01, Slovak Republic (filip.kraic@ucm.sk)
[2]SELEKT, Research and Breeding Institute, Bučany 591, SK-919 28 Slovak Republic

Abstract: The final goal of this work is development of new genotypes of wheat with better properties compared to the standard set. Prediction of optimal descriptors and properties related to the production characteristics is performed using several statistical and chemometrical tools, like correlation analysis, principal component analysis, cluster analysis and linear discriminant analysis. Optimisation of wheat genotypes is directed towards high food quality.

Key words: wheat genotypes, crop, correlation analysis, principal component analysis, cluster analysis.

1. Introduction

In the last two years a large number of data concerning various wheat genotypes, their properties and observations were prepared in a form of spreadsheets. They stemmed the databases, which are periodically complemented by several Slovak breeding stations, state and private research institutions. The main part of the data in this work is from SELEKT, Research and Breeding Institute, Bučany. The evaluated wheat descriptors are of different character; they concern the plant itself (plant height, mass of thousand grains), possible diseases (rust, powdery mildew, fusariosis, septoriosis) as well as some laboratory results (content of gluten, swellingness, sedimentation test, etc.).

Diversity studies have already been performed on wheat using grain quality (HÖGY *et al.*, 2008; EVERY *et al.*, 2008), fysiological factors (REYNOLDS *et al.*, 2002; REYNOLDS *et al.*, 2004; GOGGIN *et al.*, 2004), and resistance against fungal diseases (HAMZEHZARGHANI *et al.*, 2005; SNIJDERS, 2004).

Our previous chemometrical study revealed interesting latent relations among the soil quality (with regard to nutrition elements), content of nutrition elements in the vine leaves, vine crop and quality if the produced wine (KRAIC *et al.*, 2008). Success of that chemometrical data treatment encouraged us to investigate the present agruicultural problem by means of several statistical/chemometrical tools.

2. Material and methods

2.1 Description of the studied data

Seventy wheat samples were investigated; among them 16 samples from Bučany (designated Bu-108 to Bu-124 but without Bu-122); 13 was from Trebišov (Bu-108 to Bu-124 but without Bu-109, Bu-110, Bu-116, Bu-119), 16 from Malý Šariš (Bu-108 to

Bu-124 but without Bu-117) and 17 was from Pstruša (Bu-108 to Bu-124). In addition, the data from all mentioned agricultural locations contained 2 standard wheat strains named Ilona and Bardotka. The samples omitted from individual locations did not contain all measured (or observed) quantities therefore these samples with the missing values were not further statistically processed.

All investigated samples were characterized by the following 6 variables – wet and dry gluten (denominated as *WGluten*, *DGluten*, resp.), Prugar's number (*Prugar*) – which represents a relationship between wet gluten and swell gluten, swell (*Swell*), sedimentation (*Sedimen*) and viscosity (*Viscos*). It should be noted that the names starting with capital letter and stressed by Italics fonts denote the variables selected for further chemometrical processing.

The wheat samples were splitted into two categories according to the value of two categorical variables representing two different classification criteria: (1) Sample quality, by which the samples are distributed to four classes designated as A1 (best), A2, B1 and B2 (worst); this way of categorization was made by the agricultural experts in Bučany Research and Breeding Institute. (2) Sample origin - respecting four locations in Slovakia, namely Bučany, Trebišov, Malý Šariš and Pstruša.

2.2 Multidimensional data analysis

Statistical calculations were performed using following techniques: correlation analysis, principal component analysis (PCA), cluster analysis (CA) and linear discriminant analysis (LDA). In calculations two contemporary software commercial packages were used: STAGRAPHICS Plus 5.1 and SAS JMP 7.0.

3. Results and discussion

3.1 Correlation analysis

The output of correlation analysis is the correlation table, which contains pair (Pearson) correlation coefficients expressing the strength of correlation between all possible pairs of variables. The entries of this table are symmetrical according to diagonal. The correlation table comprising mutual dependence of all pairs of six chemical or physico-chemical variables relevant to the wheat quality is summarized in Table 1.

The following conclusions may be drawn from the correlation table: (a) The highest correlation is between *WGluten* and *DGluten*. (b) Very high correlations were found between *WGluten* and *Prugar* as well as *DGluten* and *Prugar*. (c) Very significant correlations ($r_{crit} \geq 0.306$ at $p \leq 0.01$) are between the following pairs of variables: *Prugar* and *Sedimen*, *Swell* and *Viscos*, *WGluten* and *Sedimen*, *DGluten* and *Sedimen*, *Viscos* and *Prugar*, *WGluten* and *Swell* (inverse dependence), and *Sedimen* and *Viscos*. (d) A significant correlation (at the 95 % or higher probability level, $p \leq 0.05$) is between *DGluten* and *Swell* (ireverse dependence), and *Prugar* and *Swell*. All hitherto mentioned correlation coefficients are larger or equal than the critical value of the correlation coefficient, $r_{crit} = 0.184$, and are marked by bold faces. (e) No significant correlation was proved in all other pairs of variables.

Table 1. Pearson correlation coefficients exhibiting the strength of correlation between individual pairs of variables for 70 studied wheat samples.

	WGluten	*DGluten*	*Prugar*	*Swell*	*Sedimen*	*Viscos*
WGluten	1					
DGluten	**0.988**	1				
Prugar	**0.762**	**0.837**	1			
Swell	**-0.376**	**-0.256**	**0.293**	1		
Sedimen	**0.452**	**0.468**	**0.497**	0.073	1	
Viscos	0.024	0.069	**0.389**	**0.474**	**0.305**	1

Critical values of the correlation coefficient (absolute values) for n = 70 are: r_{crit} = 0.198 (p=0.10), r_{crit} = 0.235 (p=0.05), r_{crit} = 0.306 (p=0.01); significant correlations are marked bold.

3.2 Cluster analysis

Generally, the agglomeration process in cluster analysis may be performed either with the studied objects or variables (KHATTREE and NAIK, 2000). In this work, the clustering was made for the studied six chemical/physico-chemical variables. The result of the performed cluster analysis is a dendrogram depicted in Fig. 1. The basis for the performed calculations were the data of 6 selected variables characterizing 70 samples of the wheat strains at 4 quality levels and 4 research locations. Ward's method of clustering and squared Euclidean distance was used in all calculations.

Two main clusters of the progressively agglomerated variables can be seen in Fig. 1. The first cluster, connected at the lowest Distance level, is formed by *WGluten*, *DGluten*, which are most similar, gradually connected to *Prugar* and further to *Sedimen*. The second main cluster is formed by *Swell* and *Viscos*. The variables forming the same cluster are most similar; the measure of mutual similarity is given by the distance on the vertical axis of dendrogram. The results of cluster analysis are in agreement with the outputs of correlation analysis.

3.3 Principal component analysis

In principal component analysis, PCA, some natural grouping of the objects (the wheat samples in this work) and the studied variables (the sample characteristics) may be seen. The principal components, PCs, are calculated as the linear combinations of original variables (SHARMA, 1996; KHATTREE and NAIK, 2000). According to the computed eigenvalues only three out of six principal components (PCs) were found important as their value was larger than 1, which is usually considered as the criterion of significancy. Three kinds of graphical outputs are used in the PCA, namely scatterplot showing the objects, the loadings plot showing the variables, and the biplot where both, the objects and variables are depicted together. The advantage of first two graphical representations is a possibility to obtain the 2D graph as well as the 3D illustration where usually the first two or three most important PCs are used as the axes. On the other hand, the biplot, even though plotted in two dimensions, provides some additional information about the studied problem.

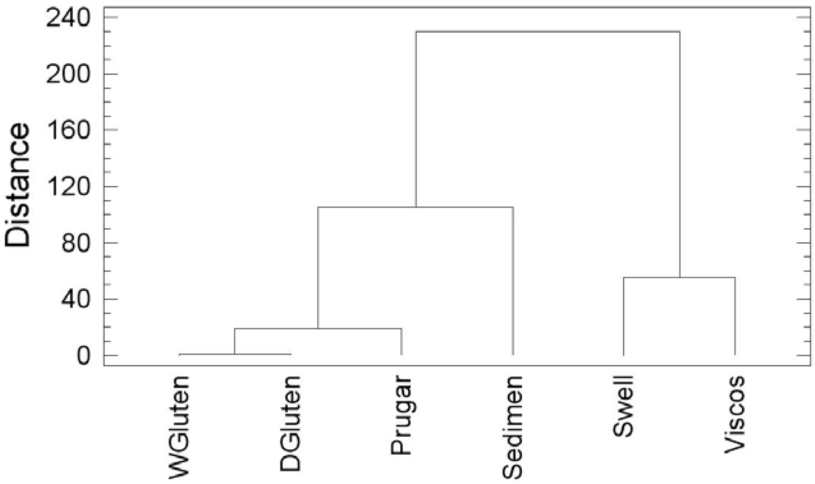

Fig. 1. Cluster analysis of 6 variables characterizing quality of 70 wheat samples obtained from all four research locations. Software Statgraphics Plus 5.1.

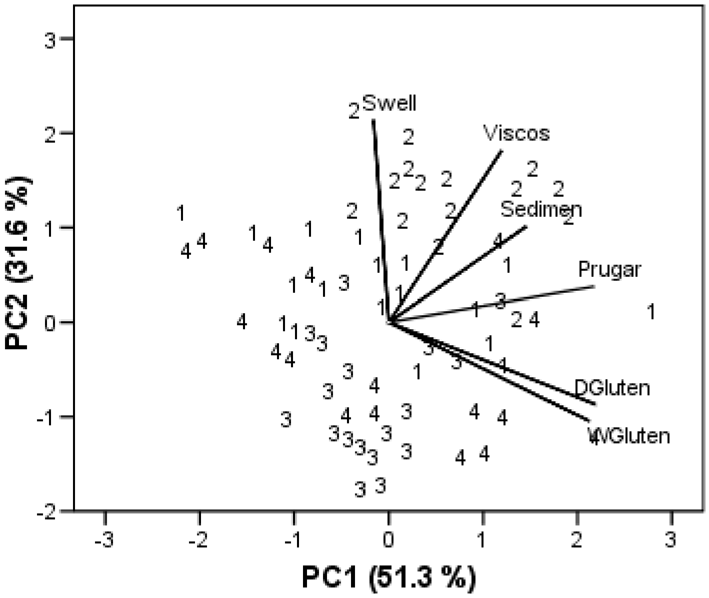

Fig. 2. Biplot PC2 vs. PC1 for 6 measured chemical descriptors and 70 samples from 4 localities (1 – Bučany, 2 - Trebišov, 3 – Malý Šariš, 4 – Pstruša). Software SPSS 15.

Fig. 2 exhibits the biplot, which simultaneously represents the wheat samples (depicted here by the numbers) and six original quality descriptors, depicted by the rays starting from the origin and ending at the point determining the position of the corresponding variable. The samples are here categorized by four locations, namely Bučany, Trebišov, Malý Šariš and Pstruša. Even though PCA method is not used for classification, high values of the second principal component, PC2, are observable for the samples from Trebišov (2), which at the same time means that they are characterized by the highest swelling and viscosity factors. The factors mostly influencing the first principal component, PC1, are wet and dry gluten and Prugar number (with relatively smaller contribution of sedimentation).

3.4 Linear discriminant analysis

Linear discriminant analysis (LDA) is a multivariate technique focused on separating distinct sets of objects into two or more populations and then allocating the objects, not readily known where they belong to, into one of the considered populations (SHARMA, 1996). Its main goal is to ensure a maximal discrimination among the classes of objects. Taking this into account, the original variables are linearly combined into discriminant functions, the number of which is equal to the number of the classes minus one (LANKMAYR *et al.*, 2004).

Classification performance depends on the selected categorical target variable. If *Quality* was used as the target variable (the categorization of wheat samples is described in part 2.1) then 94.3% of the originally grouped objects were correctly classified when the discrimination model was calculated and 92.9% of the objects were correctly classified using leave-one-out cross-validation method. It means that 5 objects out of 70 ones were categorized into a different category then supposed). Classification of the wheat samples by locality was also successful when considering that four localities were used for categorization. Leave-one-out cross-validation showed 82.9 % correct classifications. It is worth noting that in this case 25 % success corresponds to random categorization so that the classification success is clearly pronounced.

4. Conclusions

Principal component analysis and cluster analysis allow display a natural grouping of the wheat samples. Both methods revealed that six utilized variables can be divided into two groups: (1) wet and dry gluten, Prugar number and sediment, (2) swelling and viscosity factors. The obtained results demonstrate a good applicability of the used multivariate statistical methods for graphical representation of the wheat samples in two dimensional visual display. Wheat classification may be conveniently illustrated by linear discriminate analysis, which was proved as an appropriate multidimensional classification technique. Further study on the investigated wheat samples using different measurement techniques and observations is under progress.

References

EVERY, D., MOTOI, L., RAO, P.S., SHOTNER, S.C., SIMMONS, L.D.: Predicting wheat quality – consequences of the ascorbic acid improver effect. J. Cereal Sci., 48, 2008, 339-348.

GOGGIN, D.E., COLMER, T.D.: Wheat genotypes show contrasting abilities to recover from anoxia in spite of similar anoxic carbohydrate metabolism. J. Plant Physiol., 164, 2007, 1605-1611.

HAMZEHZARGHANI, H., KUSHALAPPA A.C., DION, Y., RIOUX, S., COMEAU, A., YAYLAYAN, V., MARSHALL, W.D., MATHEE, D.E.: Metabolic profiling and factor analysis to discriminate quantitative resistance in wheat cultivars against fusarium head blight. Physiol. Mol. Plant Pathol., 66, 2005, 119-133.

HÖGY, P., FANGMEIER, A.: Effects of elevated atmospheric CO_2 on grain quality of wheat. J. Cereal Sci., 48, 2008, 580-591.

KHATTREE, R., NAIK, D.N.: Multivariate data reduction and discrimination. SAS Institute, Cary, North Carolina, USA, 2000.

KRAIC, F., MOCAK, J., ARGAY, M.: Influence of nutrients in soil and vine leaves and meteorological factors upon vine crop and must. Nova Biotechnol., 8, 2008, 71-77.

LANKMAYR, E., MOCAK, J., SERDT, K., BALLA, B., WENZL, T., BANDONIENE, D., GFRERER, M., WAGNER, S.: Chemometrical classification of pumpkin seed oils using UV–Vis, NIR and FTIR spectra. J. Biochem. Biophys. Meth., 61**,** 2004, 95-106.

REYNOLDS, M.P., TRETHOWAN, R., CROSSA, J., VARGAS, M., SAYRE, K.D.: Erratum to "Physiological factors associated with genotype by environment interaction in wheat". Field Crop Res., 85, 2004, 253-274.

REYNOLDS, M.P., TRETHOWAN, R., CROSSA, J., VARGAS, M., SAYRE, K.D.: Physiological factors associated with genotype by environment interaction in wheat. Field Crop Res., 75, 2002, 139-160.

SHARMA, S.: Applied Multivariate Techniques. Wiley, New York, 1996.

SNIJDERS, C.H.A.: Resistance in wheat to fusarium infection and trichothecene formation. Toxicol. Lett., 153, 2004, 37-46.