VALIDATION OF CHEMOMETRIC ANALYSIS OF OSIJEK WHEAT CULTIVARS

ŽELIMIR KURTANJEK[1*], DANIELA HORVAT[2], ZORICA JURKOVIĆ[2], GEORG DREZNER[2], REZICA SUDAR[2], DAMIR MAGDIĆ[3]
[1]*Faculty of Food Technology and Biotechnolgy, University of Zagreb, Zagreb, Croatia*
[2] *Agricultural Institute Osijek, Osijek, Croatia*
[3]*Faculty of Food Technology, University of "J.J. Strossmayer" of Osijek, Osijek, Croatia*

Abstract

Chemometric model based on principal component analysis (PCA) of eleven Croatian wheat cultivars by evaluation of statistics of residuals is assessed. Evaluated are samples of Žitarka, Super Žitarka, Srpanjka, Barbara, Klara, Golubica, Monika, Kata, Ana and Demetra (selected at Agricultural Institute Osijek, Croatia) and cultivar Divana (Jost-Seeds Research, Križevci, Croatia) produced from harvest in 2000. The model is derived from experimental determination of the following 21 variables: High Molecular Weight Glutenin Subunit (HMW) proportions, chemical and physical properties of wheat and dough, and cultivar bread making qualities. Quality of bread crumbs are evaluated by computer image analysis. PCA analysis revealed that the samples can be projected to the subspace of two latent variables with an average residual error of 10 %. Applied is Q statistics for determination of cultivars which are projected outside the model subspace (i.e. which do not confirm to the model of the two latent variables), and T^2 (Hotelling's) for evaluation of cultivars which confirm to the model based on the two latent variables, but have "unusual" properties. Application of the chemometric model for improvement of selection of wheat cultivars and improvement of technology of production of various food products are proposed.

Key words: Chemometrics, wheat cultivars, HMW glutenin, bread making quality

Introduction

Chemometrics is an efficient mathematical and statistical method with wide area of application, from analytical chemistry to various modelling techniques for analysis and process control. It is developed for analysis of functional relationships of multivariate and large data, such as in genomics and proteomics. In contrast to classical statistical modelling method, which is based on determination of pairs of variables with strong correlation, the chemometrics is a holistic approach which accounts for simultaneous variation of all variables. Chemometrics enables

* *Corresponding author address: zkurt@pbf.hr*

extraction of essential and holistic information and removes of stochastic interference from experimental methods or/and human observations.

The scope of this research is to use chemometric methods for selecting lead wheat cultivars, selected at Agricultural Institute Osijek in Croatia, for further development. Experimental data from the harvest 2000 and basic chemometric results are already published in the works of D. Horvat, D. Magdić *et al.* (1-4). The particular aim of this work is to evaluate chemometric model validation by determination of residuals of experimental data from the two component principal component model.

Materials and Methods

Analysed are biochemical, physical and bread making quality (BMQ) properties of wheat cultivars. Grain samples from ten winter wheat cultivars (Prebasic seeds): Žitarka, Super Žitarka, Srpanjka, Barbara, Klara, Golubica, Monika, Kata, Ana and Demetra (selected at Agricultural Institute Osijek, Croatia) and cultivar Divana (Jošt-Sjeme, Križevci, Croatia) as the improver standard, were taken from the harvest of 2000

Table 1. List of measured variables separated into four data sets from S_1 to S_4.

S_1: cultivar	S_2: HMW subunits	S_3: flour properties	S_4: BMQ
Žitarka	Glu-A1x	protein content	volume
Srpanjka	Glu-B1x	sedimentation	average area
Super žitarka	Glu-B1y	wet gluten	total area
Barbara	Glu-D1x	gluten index	radius
Klara	Glu-D1y		min. radius
Golubica	Glu-1x		max. radius
Kata	Glu-1y		roundness
Monika			perimeter
Ana			number of cells
Demetra			
Divana			

For determination of HMW glutenin subunit proportions SDS-PAGE electrophoresis method is applied and standard methods are used for evaluation of physicochemical properties of flour obtained from the pure cultivars. In order to quantify evaluation of bread crumb structure a computer image analysis is applied. Details of the experimental methods and software application are given in (1-4). Table 1. provides overview of measured variables selected into 4 data sets (S_1 – S_4).

Chemometric analysis

There are different philosophical approaches to chemometric analysis (5). The most applicable to this research states that "chemometric developments and the accompanying realisation of these developments as computer software provide the means to convert raw data into information, information into knowledge and finally knowledge into intelligence". Mathematical formulation of chememoetric is focused on extraction of information by projection of large dimension space of

multivariate observations into low dimensional set of essential latent variables. The latent variables are not observable by their existence is induced from experimental data, and are expressed as their linear combination. Knowledge is gained by modelling of functional relationships between input and output variables in the space of latent variables.

The chemometric method is focused on mathematical properties of the experimental data collected into a large data matrix. Set of measurement data of *m* samples of *n* variables form a matrix ***X***(*m* x *n*) (usually autoscaled for average and/or standard deviation elimination), which is, by projection from space of dimension *n* into space of dimension *r*, decomposed into a sum of *r* sub-matrices $\boldsymbol{X}_i$ and a residual matrix ***E***:

$$\mathbf{X}(m\,x\,n) = \sum_{i=1}^{i=r} \mathbf{X}_i(m\,x\,n) + \mathbf{E}(m\,x\,n) \qquad (1)$$

First *r* partial matrices $\boldsymbol{X}_i$ correspond to *i*-th latent variables, and they "capture" deterministic components of measured data, while measurement errors (stochastic components) are retained in the error matrix ***E***. Variances of the sub-matrices are ordered, so that the maximum variance (most of the information) is contained in the first matrix $\boldsymbol{X}_1$ corresponding to the first latent variable. True or natural latent variables are approximated from experimental data by principal vectors expressed as linear combinations of measured variables. Principal components ***p*** are eigenvectors of covariance matrix (assumed are auto-scaled data), defined by:

$$(\mathbf{X}^T \cdot \mathbf{X}) \cdot \mathbf{p}_i = \lambda_i \cdot \mathbf{p}_i \qquad (2)$$

The sub-matrices are determined as the outer products of score (target) vectors $\boldsymbol{t}_i$ and corresponding latent variable (represented by its principal component vector $\boldsymbol{p}_i$):

$$\mathbf{t}_i = \mathbf{X} \cdot \mathbf{p}_i \qquad \mathbf{X}_i = \mathbf{t}_i^T \cdot \mathbf{p}_i \qquad (3)$$

For numerical evaluation of chemometric analysis applied are software tools (6-8).

Model development and validation

In the Fig. 1. are presented mathematical models proposed for analysis and further development of Osijek wheat cultivars. The models relate functional dependencies between input and output variables and are organised into three levels, starting from information of individual cultivars and cultivation conditions, to proteomics, further to flour physicochemical properties, and finally to their BMQ properties.

Linear multivariate models based on information contained in the subspace of the latent variables are proposed. Considered is application of the Partial Least Squares (PLS) method for estimation of model parameters β. The model structure is given by:

$$y = \sum_{i=1}^{i=r} \beta_i \cdot t_i = \beta_1 \cdot (\mathbf{x}^T \cdot \mathbf{p}_1) + \beta_2 \cdot (\mathbf{x}^T \cdot \mathbf{p}_2) \Lambda\ \beta_r \cdot (\mathbf{x}^T \cdot \mathbf{p}_r) \qquad (4)$$

In (4) the model is based on *r*-dimensional latent space. From the results presented in (1) is concluded that two dimensional, *r*=2, and/or three dimensional latent space, *r*=3, can explain up to 80 and 95% respectively, of the total variation (information) contained in the space of 21 measured variables. This indicates strong linear interdependence (correlation) among biochemical data, flour properties and BMQ data by image analysis. Proposed model (4) has minimal set of parameters and captures true functional relationship. Due to rejection of correlated variables and parameters, the proposed model is robust on experimental errors and could be used as a basis for further development of cultivars. However, analysis of residuals of chemometric analysis is critical for model validation and successful application. In Fig.2 is presented the idea of a model development in the space of latent variables and determination of residuals of chemometric analysis.

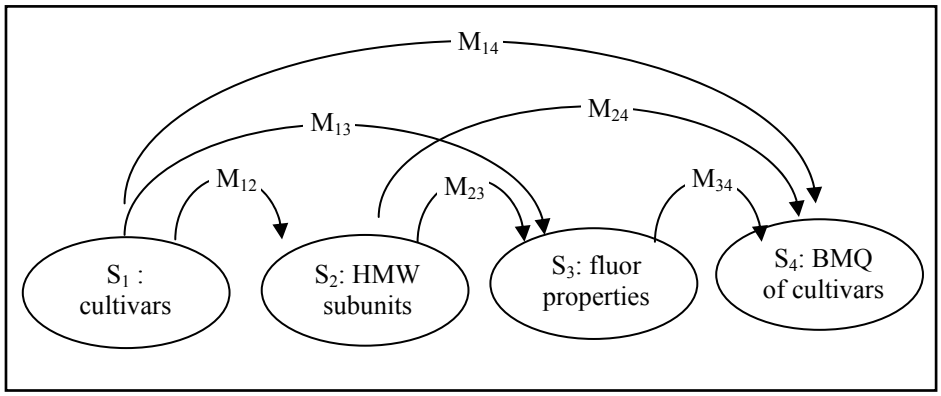

Figure 1. Graphical presentation of the proposed three level hierarchy of partial least squares models for predictions between the following data sets: S_1 - cultivar and cultivation conditions, S_2 - HMW subunits proportions, S_3 - flour properties from pure cultivars, S_4 -bread making quality BMQ produced from pure cultivars.

Experimental data are given as points in n-dimensional space of measured variables (x_1, x_2,..x_n). However, due to their inner dependence (linear correlation) they occupy a plane in a subspace of latent variables (l_1, l2). The latent space is extracted from experimental data by principal component analysis given by (1-3). There are possibly various reasons why experimental data deviate from the latent space, such as: measurement or observation errors; unmeasured variable responsible for expression of the latent variables, dynamic effects in unsteady state experiments, nonlinear dependencies between the latent and observed variables, etc. Discrepancy between experimental data and the model (projections of the experiment data into the latent subspace) can be quantified by two types of residuals, *Q* and Hottelling's T^2 , Fig. 2. They are determined by the following relations:

$$Q_i = \mathbf{e}_i \cdot \mathbf{e}_i^T = \mathbf{x}_i \cdot \left(\mathbf{I} - \mathbf{P}_k \cdot \mathbf{P}_k^T\right) \cdot \mathbf{x}_i^T \tag{5}$$

$$T_i^2 = \mathbf{t}_i \cdot \lambda^{-1} \cdot \mathbf{t}_i^T = \mathbf{x}_i \cdot \mathbf{P} \cdot \lambda^{-1} \cdot \mathbf{P}^T \cdot \mathbf{x}_i^T \tag{6}$$

Q residual is a measure of deviation of an experiment from a model defined by the latent space, i.e. it is a deviation which can not be accounted by the latent variables. Such data are outlayers with respect to the latent space; however their projections into the latent space do not

deviate from the rest of experiments. Hotelling's T^2 residual is a measure of deviation of an experiment which complies with the latent space, but is projected outside the region of the majority of data.

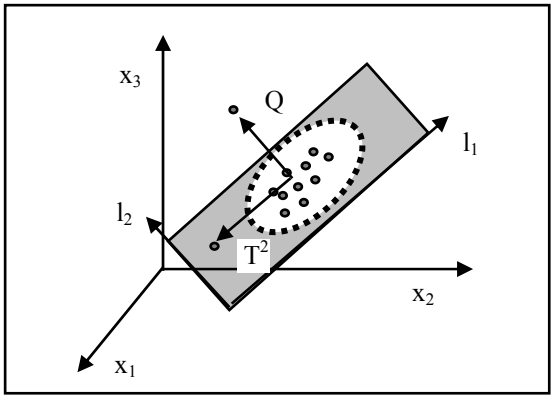

Figure 2. Presentation of the subspace defined by latent variables. Measured variables are denoted as x_1, x_2, x_3, while the latent subspace has coordinates l_1 and l_2. Experimental data are denoted as full circles. Presented is a residual Q of experimental data which do not belong to the latent subspace and T^2 data which is in the subspace but is an outlier.

Results and Discussion

Principal component analyses of Osijek wheat cultivars, harvest 2000, are presented in literature (1, 4). The analysis reveals the 21 measured variables for 11 cultivars. Table 1. can be explained by the two and/or three principal components (approximation of the latent variables) up to 85 and 95 % of variance respectively. The score plots of cultivars and variables are given in (1). The score plots reveal cluster of cultivars with similar properties, and their relations with HMW glutenin proportions indicate that Glu-1x and Glu-1y are possible the first two latent variables.

Here are presented the results for Q and T^2 residuals based on the two dimensional principal components. In Fig. 3 are depicted Q and T^2 residuals for the principal component model derived from the S_2 and S_3 data sets.

Residuals for the model derived form BMQ data evaluated by computer image analysis are given in Fig. 4. In Fig. 3-4 are depicted 95 % confidence intervals by which a statistical confidence of the proposed models are asserted. All of the cultivars from harvest 2000 have residuals inside the confidence intervals, i.e. Osijek cultivars and Divana (improver standard selected by Jost-Seed Research, Križevci, Croatia) comply with two dimensional principal component model deduced either from HMW glutenin proportions, flour physicochemical properties or BMQ data. Divana, as a specific cultivar, complies very closely with the model derived from HMW and a physicochemical property, as indicated by very small Q residual in Fig.3, but has the largest T^2 residual . In other words, Divana properties can be explained from the same model of latent variables as Osijek cultivars, but has distinct properties. Srpanjka cultivar the least complies with the model, has the largest Q residual in Fig. 3, and cultivar Kata has approximately the same magnitude of T^2 residual, but from principal component analysis (1) it is revealed that is projected opposite from Divana. The residuals of BMQ data do not point to a cultivar with specific properties (Fig.4). However, principal component analyses (1) indicate that cultivars Ana, Srpanjka and Demetra approach the BMQ properties of improver Divana.

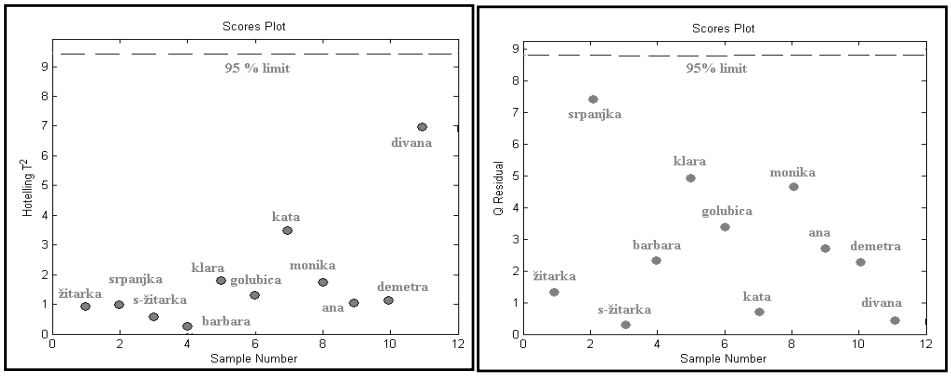

Figure 3. Confidence limit 95 % of T^2 and Q residuals for scores from data sets S_2 (HMW subunit proportions) and S_3 (flour properties) for Osijek cultivars and Divana determined by projection into two dimensional latent subspace. The first two principal components account for 43.21 and 26.43 % of the total variance.

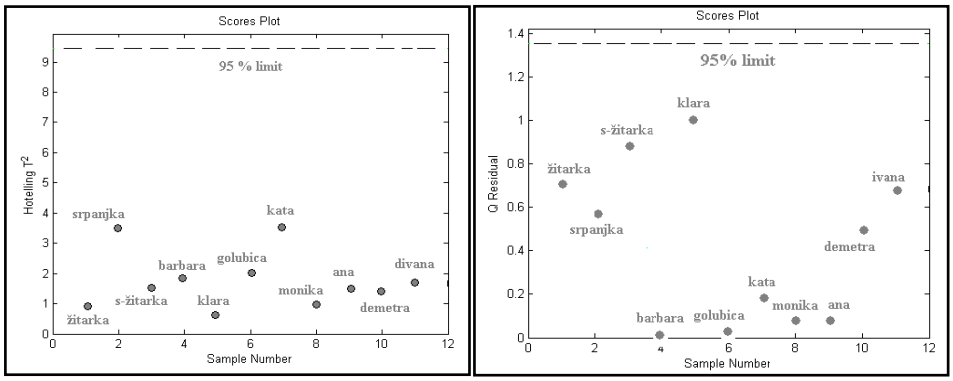

Figure 4. Confidence limit 95 % of T^2 and Q residuals for Osijek cultivars and Divana determined from S_4 data sets of BMQ by computer image analysis projected into two dimensional latent spaces. The first two principal components account for 57.14 and 16.99 % of the total variance.

Conclusions

Determined are Q and T^2 residuals for the models with two latent variables. Proposed are models based on HMW glutenin proportion, flour physicochemical properties, and BMQ properties determined by computer image analysis. All Osijek cultivars and Divana confirm with the proposed model, and their scores are inside the 95 % confidence intervals.

Based on T^2 residuals from HMW proportions is evident that cultivar Divana has distinguished properties, however, it does confirm with the same model as for Osijek cultivars.

Proposed is a hierarchy of partial least square models for determination of functional relationship between HMW glutenin proportions, flour physicochemical properties and BMQ data

from computer image analysis. These are linear models defined in the two dimensional latent space, include only two parameters, and are robust to experimental and observation errors.

The proposed mathematical models enable various applications, such as:

1) application linear programming and simplex optimisation of cultivar composition for production of flour for specific products (baker's products, industrial use, etc.)
2) improvement in experimental determination of wheat cultivar properties by use of the models for prediction of properties based on previous experimental data
3) the models can be applied in the experiment design for selection and improvement of existing wheat cultivars.

References

1. Magdić D., Horvat D., Jurković Z., Sudar R., Kurtanjek Ž. Chemometric analysis of high molecular mass glutenin fractions and image data of bread crumb structure from Croatian wheat cultivars. Food Technol. Biotechnol. 2002; 40(4):331-341.
2. Horvat D. The High Molecular Weight Glutenin Subunits (HMW) of OS Wheat Cultivars and Their Relationship With Bread-making Quality (in Croatian), M.Sc. Thesis, University of Zagreb, Faculty of Food Technology and Biotechnology, Zagreb, 2001.
3. Magdić D., Digital Image Analysis Algorithm of Bread Medium Part (in Croatian), M.Sc. Thesis, University of Zagreb, Faculty of Food Technology, Zagreb, 1999.
4. Horvat D., Magdić D., Jurković Z., Šimić G., Drezner G., Kurtanjek Ž. Chemometric And Image Analysis Of Croatian Wheat Cultivars And Their Bread Making Quality, Proceedings of ITI 2002, Ed. V. Glavinić, V. Hljuz-Dobrić, D. Šimić; Dubrovnik, June, 24-27, p.91-95
5. Workman J., Chemometrics in a network economy, NAmICS Newsletter, 2002; 22; 3-7.
6. STATISTICA v. 6.0, StatSoft, Inc., Tulsa, OK, USA, 2002.
7. Wise B.M., Gallagher, N.B.,"PLS_ Tool Box 2.0", Eigenvector Research, Inc., Manson, WA, USA; 1998.
8. MATLAB, v. 6.5, The MathWorks Inc., Natick, MA, USA; 2002.