Eur Food Res Technol (2011) 233:769–779
DOI 10.1007/s00217-011-1573-y

ORIGINAL PAPER

Discrimination of wheat grain varieties using image analysis: morphological features

Piotr Zapotoczny

Received: 16 June 2011 / Accepted: 19 August 2011 / Published online: 9 September 2011

Abstract This paper presents the results of a study on the discrimination of 11 wheat grain varieties in three successive years of cultivation and at the grain humidity of 12, 14 and 16%. Each grain was described with the use of 54 geometric variables which, after reduction of variables, left 20 for use in the main analysis. Variables calculated from linear dimensions had the greatest share in the group of discriminating variables, with shape-related indexes being of lesser importance. Seven methods of variables selection based on genetic algorithms (the *Class Ranker* and *Class Rankers-Search* methods) were used in the study. The final discriminant analysis was performed with the use of *stepwise progressive analysis* and *Meta MultiClass Classifier*. The proposed statistical model classified varieties with 90–100% accuracy, depending on the experimental group. Grain images were acquired with a flat scanner, and grains were arranged with a specially designed matrix which enabled arranging 552 grains in rows and columns within several minutes; this makes the method usable in the cereal industry.

Keywords Digital image analysis · Size · Image morphological feature · Discrimination · Flatbed scanner · Classification

Introduction

Measurement of geometric features of different types of grain or other crops is of fundamental importance to the processing industry. Knowledge about the basic dimensions can be used in designing machines for sorting, washing, grinding or transporting devices of different kinds (e.g., conveyors). Such knowledge, combined with a knowledge of chemical composition, may help producers or processors to perform quick evaluation of the technological usability of a batch of grain [6]. Before the video systems were developed, shape and dimensions could only be measured with rulers or different kinds of sieves. Currently, owing to the use of video systems, 2D or 3D images can be fed into a computer and even the most complicated geometrical parameters can be determined automatically. The first studies of the subject were conducted as early as in the 1980s [18]. Paliwal et al. [8], Luo et al. [4] used a video system to identify grain damage caused by diseases. Luo et al. [4] determined 68 geometric attributes of shape, which were used to correctly identify diseases or damage with 90–100% accuracy. Shouche et al. [12] described the use of shape indexes and moments of inertia to characterize wheat varieties. Measurement of geometric attributes can be useful in automatic cultivar classification. Various grading systems using different morphological features for the classification of different cereal grains and varieties have been reported in literature [5]. The authors presented the application of image processing techniques in the identification of Australian wheat varieties. They determined 23 geometric features, 10 of which were used in cultivar classification. The classification accuracy ranged from 97 to 100%. Grain harvested in 1994 and only one humidity level were taken into account in the experiment. There is no information on how the model performed in subsequent years of harvest. Pablo et al. [7], Visen et al. [16] classified grain varieties using their color along with geometric features. They developed models with the use of neural networks, which resulted in a classification accuracy

P. Zapotoczny (✉)
Department of Agri-Food Process Engineering,
University of Warmia and Mazury in Olsztyn,
Heweliusza 14, 10-718 Olsztyn, Poland
e-mail: zap@uwm.edu.pl

of 40–96%. Paliwal et al. [9] used a combination of geometric features, grain surface texture and its color to develop a statistical model that used selected variables to identify varieties. They selected 20 features in each group. Depending on the species or the number of variables in the model, the accuracy of classification ranged from 88 to 98%. The experiment was also conducted within 1 year, and there is no information whether it can be used in subsequent years of cultivation. Similar studies using geometric parameters, together with texture and color of grain or seeds of papilionaceous plants, were conducted by [1, 2, 10, 11, 14, 15].

However, the reports do not provide any information about the operation of the statistical model on the data obtained in successive years of cultivation. Therefore, the aim of this study is to develop a statistical model to classify grain of spring and winter wheat harvested in three consecutive years of cultivation, at three humidity levels of 12, 14 and 16%. The system is based on the use of images acquired with a flat scanner, and the method of arranging grains on the measurement scene made it possible to reduce the time of analysis to just a few minutes.

Materials and methods

Grain samples

The experimental material comprised cleaned grain of common spring* and winter wheat of four quality classes (elite wheat: *Torka**, prime quality wheat: *Nawra*, Koksa*, Zyta, Sukces, Tonacja, Fregata,* bread wheat: *Cytra*, Soraja, Nutka*, forage wheat: *Symfonia*). The study covered three cultivation years (2005, 2006, 2007), and 11 varieties (seven winter and four spring varieties) were analyzed each year at three moisture content levels—12, 14 and 16%. Initial moisture content was determined in two replications using the drying method according to Polish standard PN-71A-75101. The samples were ground and placed in a laboratory dryer at a temperature of 100 °C for 4 h. Samples characterized by low initial moisture content values were hydrated. Water was added, grain was stirred for 24 h, and it was placed in tight plastic containers and stored for 48 h at room temperature to ensure equal moisture distribution through the sample. Moisture content was again determined after the applied hydration treatment.

Image analysis

The image acquisition and image analysis workstation consisted of an Epson Perfection 4490 Photo flat scanner connected with a graphic station based on an Intel Pentium D 830 processor. SilverFast Epson v 6.4.3 software was used. Before each series of images was acquired, the scanner was calibrated with an IT8.7/2 template, supplied with the scanner software. Grains were arranged on the measurement scene in 24 rows and 23 columns, so 552 grains could be scanned simultaneously. Grains were arranged with the use of a specially designed matrix, and it took about 5 min to arrange one scene. In total, over 6,500 grains were scanned and analyzed in each cultivar for each year and humidity level. Before a proper analysis of the image was performed, an algorithm of image segmentation was developed. It is an issue of special importance because an incorrectly established segmentation threshold can significantly affect the results. The segmentation algorithm has morphological and non-linear filters implemented in it. The analytical procedure involved a series of the following successive steps: scanner calibration, kernel arrangement in the matrix, matrix removal, scanning and image saving (2,673 × 4,031 resolution, 400 DPI, 24-bit color depth, TIFF format). The next step was image segmentation to generate a mask for the original image. At the final stage, a 1-bit mask of the original image was obtained, and the surface area occupied by pixels identified as belonging to a single kernel was subjected to a geometric analysis.

The methodology developed in this study allowed an unlimited number of images to be analyzed automatically. The computer-aided image analysis was performed by MaZda 4.3 software [13].

Each grain was described by 54 geometric variables that include linear measurements and shape indexes (Table 1).

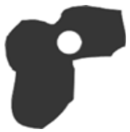
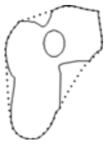
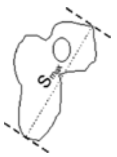
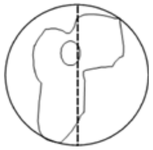
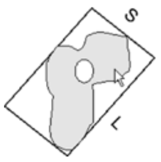
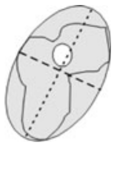
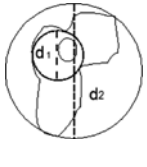
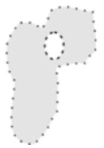
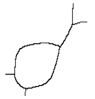
Statistical analysis of results

The analysis of results was performed at several stages. Initially, a histogram distribution for individual variables was checked. At the next stage, in order to reduce the dispersion of results with respect to the mean value, randomly taken values were averaged to give one value. At the next stage, variables were reduced to a set of the 20 best ones. Supervised and unsupervised selection was used. At the last stage, multidimensional analysis was performed in order to discriminate the varieties. To that end, the usability of several methods of discrimination was analyzed and 7 were chosen based on *decision trees*, *Bayes classifiers* and *Lazy classifiers.*

Variables reduction

As each case was described with over 50 geometric variables and the discriminant power of variables could cancel each other out, they were reduced to a set of the 20 best ones. 7 methods of selection were analyzed, based on genetic algorithms, methods based on *Class Ranker* and *Class RankersSearch*. In the first one, the selected

Table 1 Listing of calculated linear dimensions and shape indexes

8 *Geo_F* area, number of the object pixels 	45 *Geo_Uw* convex perimeter 	40 *Geo_S*$_{\max}$ maximal diameter 	9 *Geo_Fd*2 area of circumscribing circle 	36 16 *Geo_S*, *Geo_L* 	17 18 *Geo_L*$_{\max E}$, *Geo_L*$_{\min E}$ 	2 3 *Geo_D*1, *Geo_ D*2 	43 *Geo_Ug* sum of distances between centers of neighboring contour pixels 
41 *Geo_Spol* diameter of the area equivalent circle $Geo\_S_{pol} = 2 * \sqrt{\frac{Ft}{4}}$	46 *Geo_W*1 shape factor $Geo\_W1 = \frac{L_{\max E}}{L_{\min E}}$	53 *Geo_W*2 shape factor $Geo\_W2 = \frac{4\pi*F}{U_w^2}$	54 *Geo_W*3 shape factor $Geo\_W3 = \frac{U_w^2}{F}$	55 *Geo_W*4 shape factor $Geo\_W4 = \frac{U_w}{U_g}$	56 *Geo_W*5 shape factor $Geo\_W5 = \frac{F}{L_{sz}}$	58 *Geo_W*6 shape factor $Geo\_W6 = \frac{U_w}{4\pi*F}$	33 *Geo_R*$_H$, Haralic ratio $Geo\_R_H = \sqrt{\frac{\left(\sum_i d_i\right)^2}{n*\sum_i d_i^2 - 1}}$
59 *Geo_W*7 shape factor $Geo\_W7 = \frac{d_2}{d_1}$	60 *Geo_W*8 shape factor $Geo\_W8 = \frac{L}{S}$	61 *Geo_W*9 Shape factor $Geo\_W9 = \frac{L*S}{F}$	47 *Geo_W*10 shape factor $Geo\_W10 = \frac{M_{\max}}{M_{\min}}$	48 *Geo_W*11 shape factor $Geo\_W11 = S_{\max} - S_{\min}$	49 *Geo_W*12 shape factor $Geo\_W12 = \frac{4\pi*F}{\pi*S_{\max}^2}$	50 *Geo_W*13 shape factor $Geo\_W13 = \frac{S_{\max}}{F}$	51 *Geo_W*14 shape factor $Geo\_W14 = \frac{F}{S_{\max}^3}$
52 *Geo_W*15 shape factor $Geo\_W15 = \frac{4*F}{\pi*S_{\min}*S_{\max}}$	34 *Geo_R*$_M$ Malinowska ratio $Geo\_R_M = \frac{U}{2*\sqrt{\pi*F}} - 1$	26 *Geo_R*$_B$ Blair-Bliss ratio $Geo\_R_B = \frac{F}{\sqrt{2\pi*\sum_i r_i^2}}$	30 *Geo_R*$_D$ Danielsson ratio $Geo\_R_D = \frac{F^3}{\left(\sum_i l_i\right)^2}$	19 *Geo_Lsz* skeleton length 	27 *Geo_Rc* circularity *Rc*1/*Rc*2 $Geo\_R_c = \frac{R_{c1}}{R_{c2}}$	28 *Geo_Rc*1 $Geo\_R_{C1} = 2 * \sqrt{\frac{F}{\pi}}$	29 *Geo_Rc*2 $Geo\_R_{C2} = \frac{U}{\pi}$

The number of the variable is given in the top left-hand corner

4 Geo_El (average distance from contour), *5 Geo_El*2 (average square distance from contour), *6 Geo_Er* (average distance from gravity center), *7 Geo_Er*2 (average square distance from gravity center), *10 Geo_FE* (area of circumscribing ellipsis), *12 Geo_Fmax* (maximal Feret's diameter), *13 Geo_Fmin* (minimal Feret's diameter), *14 Geo_Ft* (profile area), *20 Geo_M2x* (horizontal second order moment of inertia), *21 Geo_M2y* (vertical second order moment of inertia), *22 Geo_M2xy* (second order moment of inertia), *23 Geo_Maver* (Martin's average radius—average distance between gravity center and contour pixels), *24 Geo_Mmax* (Martin's maximal radius—maximum distance between gravity center and contour pixels), *25 Geo_Mmin* (Martin's minimal radius—minimum distance between gravity center and contour pixels), *31 Geo_Rf* (GeoFh/GeoFv), *32 Geo_Rff* (GeoFmax/GeoFmin), *35 Geo_Rs* (GeoU2/4p GeoFt), *37 Geo_S*1 (contour-skeleton maximal thickness), *38 Geo_S*2 (contour-skeleton minimal distance), *39 Geo_SigR* (standard deviation of all radii), *42 Geo_SxGeo_L* (area of circumscribing rectangle), *44 Geo_Ul* (profile specific perimeter), *57 Geo_W5b = Geo_W5/Geo_Lsz* attributes were evaluated by the *InfoGainAttributeEvaluate* method, which involves attributes by measuring their information gain with respect to the class. It discretizes numeric attributes first using the MDL-based discretization method (it can be set to binarize them instead). This method, along with the next three, can treat missing as a separate value or distribute the counts among other values in proportion to their frequency [17]. Another method was based on the *ChiSquared*. *ChiSquaredAttributeEvaluate* statistic evaluates attributes by computing the chi-squared statistic with respect to the class. *GainRatioAttributeEval* evaluates attributes by measuring their gain ratio with respect to the class. *SymmetricalUncertAttributeEval* evaluates an attribute *A* by measuring its symmetric uncertainty with respect to the class *C* [17]. In the *Class RankerSearch* method, the quality of attributes was evaluated by the *CfsSubsetEvaluate* and *ConsistencySubset-Evaluate* methods. Statistical analysis was performed with the use of WEKA v. 3.7 software [3].

Multidimensional analysis

Once the variables had been selected, the multidimensional analysis was started. Cultivar classification was performed with the use of 6 classification methods, i.e., *Bayes*, *Lazy, Meta classifiers, Decision tree* and *Stepwise discriminant analysis. Discriminant stepwise progressive analysis* was performed with the use of the Statistica v 9.0 (StatSoft. Inc) statistical package; the other analyses were performed with the use of WEKA v3.7. The strategy adopted in developing the statistical model involved division of a data set into two subsets: the test set accounted for 30% of the whole and the training set—for 70%. At that stage of the analysis, a method was being sought to ensure the minimal error in the classification of 11 varieties of wheat grain in successive years of cultivation and at specified humidity levels.

Results and discussion

Statistical characteristics of the study material

Selected results of measurements of the geometric features are shown in Table 2. The grain length (*Geo_L*) ranged from 6.30 to 7.58 mm. Grains of the spring varieties were shorter than the winter ones. No permanent tendency to change the grain length depending on the year of cultivation was observed. In the *CYTRA* cultivar, the length increased every year, whereas the reverse tendency was observed in the *ZYTA* cultivar. The grain widths (*Geo_S*) ranged from 3.15 to 3.77 mm. The average width of the spring varieties differed by 0.49 mm from the winter ones. It is noteworthy that the grain width of the spring varieties changed significantly in 2007. Their width was the same as those of winter varieties. The smallest width in the winter varieties was recorded in 2006. It was a year when the weather conditions were adverse and the grain was not as well developed as in 2005 or 2007. The tendency was also observable in the spring varieties. The projection area (*Geo_F*) of spring varieties grains differed by 3.88 mm^2 from the winter ones. As in the case of grain of spring varieties, 2007 was significantly different from 2005 to 2006. The projection area in the winter varieties decreased in the successive years of cultivation. The shape index *Geo_W*6 describes the extent to which an object surface is folded; its value for a circle is 1. On the other hand, the value of the *Geo_Rb* is not sensitive to a change of the object scale and it describes the shape precisely. The perimeter of the projection area of the spring varieties grains was more folded than in the winter varieties; the variability of the index between the years was greater. The values of the indexes for the winter varieties were more stable, and the value of the *Geo_W*6 index showed that the grains of the winter varieties were more oblong and less folded. The CYTRA cultivar had the most irregular and the least stable shape.

In order to analyze the usability of the geometric dimensions in cultivar discrimination, the distribution of histograms for each variable was analyzed. Ideally, the intervals of dispersion of a variable for different varieties should not overlap. Figure 1 shows histograms for the year 2005, for selected varieties and variable geometries. The distribution of the histogram shape was normal, but, unfortunately, their intervals overlapped. This showed that there were grains in each cultivar whose linear dimensions of the shape indexes were not statistically different than those in other experimental groups. This had a negative impact on the discriminant power of individual variables. Table 3 shows the results for selected methods of discriminant analysis for a set of "raw" data. Discrimination of individual varieties was highly unsatisfactory, with a classification error ranging from 47 to 70%. For this reason, the number of cases was reduced by averaging 25 cases to 1. The operation resulted in reducing the set of data to 280 cases for each cultivar and mainly to reducing diversity within one cultivar. Figure 1 shows the histograms after the cases were averaged. This procedure resulted in more distinct "clusters" of cases for each cultivar, and the histograms did not overlap as for the original data.

Variables reduction

The variables for discrimination were selected from the set of data for the years 2005–2007 and from all the three humidity values. At least 10 variables were selected in each

Table 2 The average values of selected linear dimensions and shape indexes for grain with a humidity of 12%

Grain type	Year	*Geo_L* (mm)		*Geo_S* (mm)		*Geo_F* (mm^2)		*Geo_Uw* (mm)		*Geo_Rb* (–)		*Geo_W*6 (–)	
		X_{ave}	±SEM	X_{ave}	±SEM	X_{ave}	±SEM	X_{ave}	±SEM	X_{ave}	±SEM	X_{ave}	±SEM
Cytra[a]	2005	6.30	0.01	3.15	0.00	16.02	0.03	15.45	0.01	3.562	0.005	0.21	0.00
	2006	6.43	0.01	3.00	0.00	15.78	0.04	15.57	0.02	3.453	0.006	0.22	0.00
	2007	6.58	0.01	3.77	0.00	20.08	0.03	16.85	0.01	4.138	0.004	0.18	0.00
Koksa	2005	6.97	0.01	3.19	0.00	17.89	0.03	16.69	0.01	3.667	0.005	0.21	0.00
	2006	7.20	0.01	3.24	0.00	18.89	0.04	17.19	0.01	3.750	0.005	0.22	0.00
	2007	6.84	0.01	3.70	0.00	20.46	0.03	17.19	0.01	4.119	0.004	0.18	0.00
Nawra	2005	7.05	0.01	3.27	0.00	18.24	0.03	16.84	0.01	3.711	0.005	0.20	0.00
	2006	7.24	0.01	2.91	0.00	16.95	0.04	16.85	0.02	3.399	0.005	0.22	0.00
	2007	7.10	0.01	3.59	0.00	20.32	0.04	17.41	0.01	4.020	0.005	0.19	0.00
Torka	2005	6.66	0.01	3.19	0.00	17.01	0.03	16.09	0.01	3.629	0.004	0.21	0.00
	2006	6.77	0.01	3.01	0.00	16.53	0.04	16.14	0.02	3.489	0.005	0.22	0.00
	2007	6.68	0.01	3.64	0.00	19.37	0.03	16.73	0.02	4.014	0.004	0.19	0.00
Fregata[b]	2005	7.17	0.01	3.68	0.00	21.17	0.03	17.63	0.03	4.150	0.004	0.18	0.00
	2006	7.64	0.01	3.66	0.00	22.67	0.07	18.45	0.01	4.187	0.004	0.18	0.00
	2007	6.72	0.01	3.55	0.00	19.04	0.03	16.63	0.01	3.967	0.006	0.19	0.00
Nutka	2005	7.58	0.01	3.68	0.00	22.17	0.03	18.38	0.01	4.159	0.004	0.18	0.00
	2006	7.42	0.01	3.37	0.00	19.95	0.03	17.70	0.01	4.187	0.004	0.19	0.00
	2007	7.05	0.01	3.60	0.00	20.21	0.03	17.33	0.01	3.967	0.004	0.19	0.00
Soraja	2005	7.18	0.01	3.60	0.00	20.80	0.03	17.62	0.01	4.068	0.004	0.19	0.00
	2006	7.12	0.01	3.51	0.00	20.25	0.04	17.42	0.01	3.993	0.004	0.19	0.00
	2007	7.18	0.01	3.76	0.00	21.60	0.04	17.77	0.01	4.200	0.005	0.18	0.00
Sukces	2005	7.40	0.01	3.70	0.00	22.06	0.03	18.16	0.01	4.183	0.004	0.18	0.00
	2006	7.41	0.01	3.48	0.00	20.88	0.03	17.91	0.01	3.997	0.004	0.19	0.00
	2007	6.90	0.01	3.70	0.00	20.14	0.03	17.24	0.01	4.105	0.004	0.18	0.00
Symfonia	2005	6.87	0.01	3.76	0.00	20.12	0.03	17.09	0.01	4.096	0.004	0.18	0.00
	2006	7.00	0.01	3.73	0.00	20.58	0.03	17.34	0.01	4.117	0.004	0.18	0.00
	2007	6.96	0.01	3.68	0.00	20.14	0.03	17.21	0.01	4.056	0.004	0.19	0.00
Tonacja	2005	7.09	0.01	3.76	0.00	21.28	0.03	17.59	0.01	4.191	0.004	0.18	0.00
	2006	7.44	0.01	3.40	0.00	20.18	0.03	17.74	0.01	3.898	0.004	0.19	0.00
	2007	6.87	0.01	3.59	0.00	16.97	0.04	16.96	0.01	4.006	0.004	0.19	0.00
Zyta	2005	6.97	0.01	3.67	0.00	20.59	0.03	17.34	0.01	4.111	0.003	0.18	0.00
	2006	6.97	0.01	3.45	0.00	18.70	0.02	16.64	0.01	3.879	0.003	0.19	0.00
	2007	6.51	0.01	3.55	0.00	18.57	0.03	16.33	0.01	3.939	0.004	0.19	0.00

[a] spring cultivars

[b] winter cultivars

experimental setting. This produced a group of 63 sets of variables from all the experimental groups (3 years, 3 humidity levels and 7 methods of selection). In the next step, the number of sets was reduced from 63 to 21, by combining sets of variables obtained at a specific humidity level and all the methods of selection. The sets were combined by selecting variables that were first on the list. Table 4 shows in a synthetic way the numbers of variables which were chosen for further discriminant analysis. The variables were shown according to the frequency of their occurrence (methods of reduction). The most frequently chosen included the following: *Geo_Er*2 (average square distance from gravity center), *Geo_Fd*2 (area of circumscribing circle), *Geo_Fmax* (maximal Feret's diameter) and *Geo_L* (length). The majority of these are variables determined from linear dimensions. The most frequently chosen shape index was the *Danielsson's index* (*Geo_R$_D$*) and *Rc*2. Of all the 54 analyzed geometric variables, 17 parameters were chosen more than 4 times, including most of those which are not shape indexes.

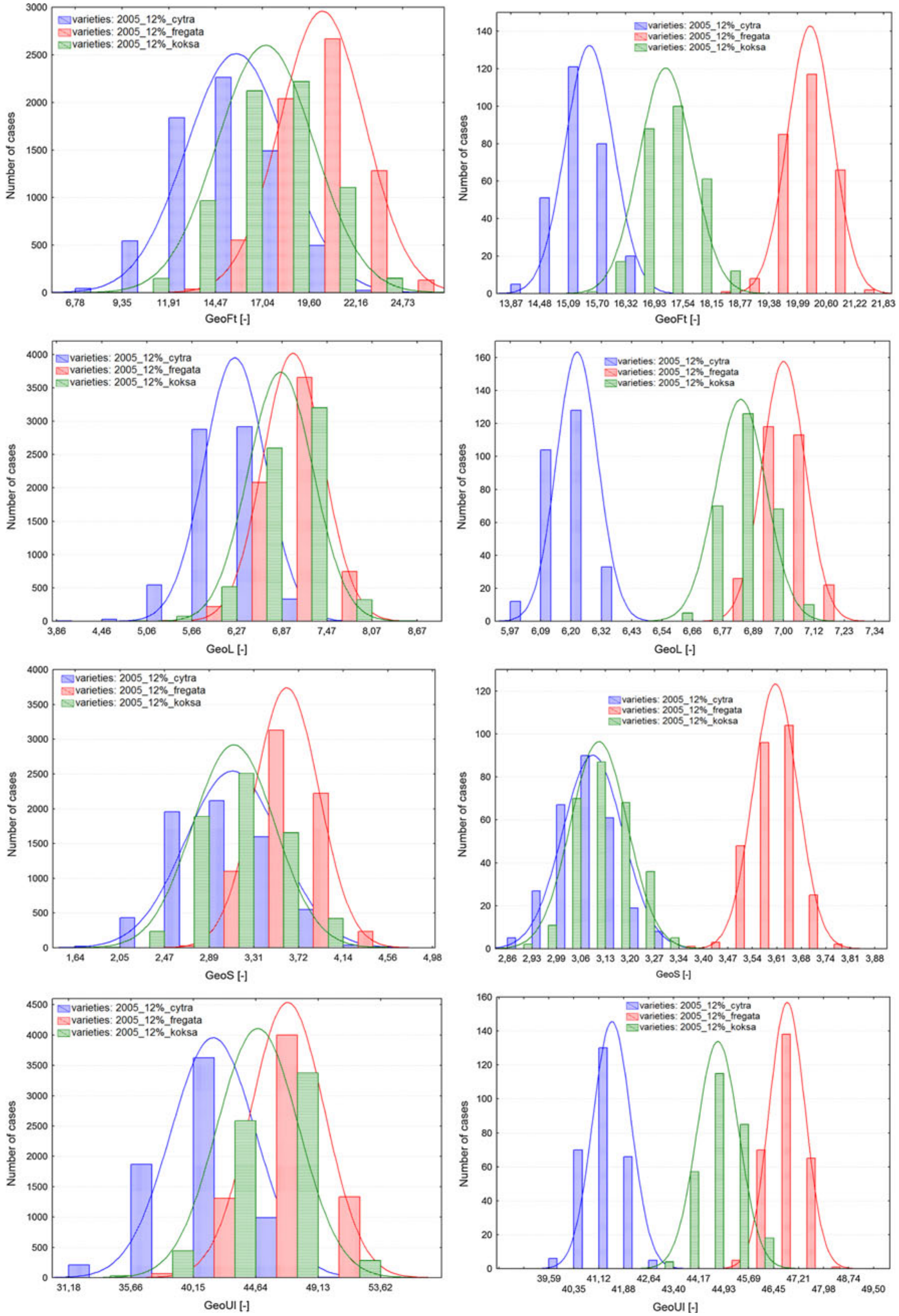


Fig. 1 Dispersion of histograms of selected variables before (*left*) and after (*right*) the cases was averaged

Table 3 The results of discriminant analysis for the raw data Training set 1840, test set 785, method of selection Ranker + ChiSquaredAttributEval

	Method of dydiscrimination	Total (%)	CYTRA[a]	TORKA[a]	KOKSA[a]	NAWRA[a]	NUTKA	SORAJA	SUKCES	SYMFONIA	TONACJA	FREGATA	ZYTA
2005	Meta MultiClass Classifier	43	80	47	40	36	45	45	24	45	55	17	42
	Discriminant analysis	44	79	47	41	38	44	46	25	45	56	14	47
2006	Meta MultiClass Classifier	40	64	46	24	52	23	27	53	61	35	31	0
	Discriminant analysis	39	66	44	26	57	24	24	55	63	34	33	2
2007	Meta MultiClass Classifier	30	36	20	48	23	27	32	23	1	2	38	41
	Discriminant analysis	30	38	21	49	22	29	34	25	3	5	35	43

[a] Spring cultivars

Subsequently, the usability of different sets for cultivar discrimination was tested on the set of data from the year 2005 and the humidity level of 12%. Table 5 shows the collective results of the multidimensional analysis conducted by 6 discriminant methods. The multidimensional analysis proper was conducted on the entire set with a set of variables obtained by the selection method: *Ranker + ChiSquared AttributEval* and *Best First.*

Multidimensional analysis

Preliminary multidimensional analysis

Due to a complicated experimental setting (number of years, levels of humidity, varieties) and the methods of selection of variables and multidimensional analyses, a preliminary evaluation of the usability of the classification methods was carried out. The results are shown in Table 5. The analysis was performed on the set of data from the years 2005–2007 and the humidity level of 12%. The cumulative error of classification, depending on the method applied, ranged from 56 to 99%. The worst results were achieved for the *Bayes Net* and *Naive Bayes* methods, whereas the best results were achieved for the methods: *Meta MultiClass Classifier* and *Discriminant analysis.* For this reason, those two methods of discrimination were chosen for further analysis. Regardless of the applied method of selection of variables and of classification, the lowest error of varieties discrimination was achieved in 2005, followed by 2006 and the worst was in 2007.

Main multidimensional analysis

In the final stage of the analysis, the discrimination of 11 grain varieties from three successive years of cultivation and at three level of humidity was conducted. As described previously, the varieties discrimination was conducted by the *Ranker + ChiSquaredAttributEval* and *Best First* methods, whereas classification was conducted by the *Meta MultiClass Classifier* and *Discriminant analysis* methods.

Discrimination by the Meta MultiClass Classifier Ranker method The results of the classification analysis are shown in Table 6, where the training set comprised 1,840 cases, whereas the test set comprised 785 cases. The cumulative classification ranged from 90 to 99%, depending on the year of cultivation. A decreasing tendency in classification quality depending on the year of cultivation was observed. The best results were achieved for the year 2005 and the worst for 2007, with the differences not being significant and not higher than 8%. The worst results with respect to varieties were achieved for *Nutka* and *Tonacja.* In the case of the latter, the maximum accuracy of

Table 4 The results of discriminant analysis for the raw data setting of selected variables

Multiplicity	1	2	3	4	5	6	7
Number variable	2, 15, 25, 26, 31, 32, 47, 53, 59;	13, 14, 18, 19, 23, 27, 33, 34, 42, 45;	5, 10, 11,21, 43, 44;	15, 39, 40, 48, 51;	3, 6, 8, 29, 30;	17, 22, 24;	7, 9, 12, 16;

Table 5 The results of discriminant analysis for the raw data method of selection

Method discrimination	Year	Test options	Genetic search	Best first	Linear forward selection	Ranker + InfoGain AttributeEval	Ranker + ChiSquared AttributEval	RankSearch + CfsSubsetEval	RankSearch + consistency subset eval
Naive bayes	2005	% split	77	86	80	75	81	79	79
	2006		86	87	87	80	85	85	85
	2007		61	65	65	57	62	59	59
Bayes net	2005	% split	77	84	79	73	79	77	77
	2006		85	85	87	79	85	84	84
	2007		61	62	62	56	61	59	59
Lazy.IB1	2005	% split	82	90	85	89	90	89	89
	2006		89	90	88	90	91	90	90
	2007		60	67	70	77	74	75	75
Meta MultiClass Classifier	2005	% split	95	96	97	97	98	98	98
	2006		98	99	98	97	98	97	97
	2007		88	93	92	93	93	92	92
Trees.J48	2005	% split	84	92	86	86	89	88	88
	2006		90	90	89	89	90	91	91
	2007		64	66	69	74	72	70	70
Discriminant analysis	2005	% split	94	97	96	95	98	97	97
	2006		97	97	96	97	97	96	96
	2007		92	95	93	95	96	95	95

classification achieved was 98%. For the majority of varieties, the accuracy of classification ranged from 96 to 100%. No negative effect of humidity on discrimination was observed in any of the varieties.

Discrimination by the stepwise progressive method This method of discrimination employed stepwise progressive analysis, assuming that the analysis will be conducted until all the variables are introduced in the model or until the value of Wilks' lambda statistics of min. 0.00001 is achieved.

Figure 2 shows diagrams of dispersion of canonical variables for the varieties under analysis. As in the method discussed earlier, the accurate classification index ranged from 92 to 97%. The worst results were again achieved for the *Nutka* and *Tonacja* varieties. The decreasing tendency in discrimination quality was observed depending on the cultivation year; 2005 was the best, and 2007 was the worst.

The discrimination analyses conducted made it possible to distinguish between spring and winter varieties. The *Cytra, Torka, Koksa* and *Nawra* varieties occupied a distinct area in the dispersion diagram (Fig. 2). A winter cultivar—*Zyta*—was also included in the same space, but only in 2005. In the other years, the winter varieties were separate from the spring ones.

Conclusions

The experiment and the proposed methodology has resulted in a statistical model that can perform classification of 11 wheat varieties with an accuracy of 90–100%, depending on the method applied, year of cultivation, humidity and cultivar. Cultivar discrimination was based on a model in which 20 geometry variables were implemented, most of which were calculated from linear dimensions, with shape indexes not being as important. The proposed model also distinguished winter varieties from spring varieties. No effect of grain humidity on the discrimination quality was observed. Of the varieties analyzed, *Nutka* and *Tonacja* lowered the quality of

Table 6 The results of multidimensional analysis for 11 varieties of grain, the year of cultivation: 2005, 2006, 2007 and the grain humidity of 12, 14 and 16%

Method of discrimination/Method Selection			Total (%)	CYTRA[a]	TORKA[a]	KOKSA[a]	NAWRA[a]	NUTKA[b]	SORAJA[b]	SUKCES[b]	SYMFONIA[b]	TONACJA[b]	FREGATA[b]	ZYTA[b]
Meta MultiClass Classifier	2005	12%	98	99	97	99	97	100	97	98	99	96	99	100
		14%	96	100	100	100	96	98	93	90	100	92	96	93
		16%	98	100	99	96	100	95	98	97	100	98	98	96
	2006	12%	98	100	100	100	100	96	94	100	100	95	100	99
		14%	97	99	100	95	100	95	97	94	100	94	96	99
		16%	96	99	94	97	100	90	100	96	99	90	97	94
Ranker + ChiSquaredAttributEval	2007	12%	93	99	90	92	87	93	99	92	89	94	90	95
		14%	93	91	98	94	99	81	100	94	88	83	94	100
		16%	93	95	97	95	97	85	96	91	93	86	94	97
Meta MultiClass Classifier	2005	12%	96	98	96	97	96	99	93	94	98	88	97	100
		14%	96	100	96	99	98	96	88	96	98	90	97	93
		16%	96	100	90	95	98	98	92	97	99	98	98	97
	2006	12%	99	100	99	100	100	99	94	100	100	98	98	100
		14%	97	100	100	95	99	92	99	97	98	87	96	99
		16%	93	99	97	96	100	72	95	92	99	84	91	99
Best first	2007	12%	93	99	97	96	100	72	95	72	99	84	91	99
		14%	92	97	100	91	96	80	94	96	91	73	96	99
		16%	90	90	97	78	97	85	96	86	92	80	93	97
Discriminant analysis- krokowa postępującas	2005	12%	98	98	96	99	98	99	98	96	98	98	96	100
		14%	98	100	65	100	99	99	97	100	100	96	99	94
		16%	98	100	100	97	99	100	96	98	100	96	97	97
	2006	12%	97	99	100	98	100	90	89	99	100	95	100	97
		14%	97	100	98	98	100	82	100	100	100	97	98	100
		16%	96	100	100	95	100	83	98	97	99	93	100	95
Ranker + ChiSquaredAttributEval	2007	12%	96	96	99	94	98	97	97	96	91	91	99	97
		14%	94	95	97	91	96	96	91	91	92	90	96	100
		16%	95	95	100	99	97	87	95	96	81	99	95	100
Discriminant analysis- krokowa postępującas	2005	12%	97	97	95	96	99	100	97	96	98	96	96	100
		14%	97	100	94	96	98	98	97	98	100	94	99	97
		16%	97	100	98	97	100	100	95	98	100	91	96	96
	2006	12%	97	99	100	99	100	88	91	100	100	93	100	96
		14%	97	100	97	98	100	81	100	99	100	95	96	100
		16%	95	100	100	98	100	76	98	97	99	84	97	97
Best first	2007	12%	95	96	97	92	96	95	95	94	94	89	99	100
		14%	96	99	99	91	95	96	91	97	95	90	97	100
		16%	92	99	98	79	92	90	97	90	91	96	94	99

The training set 1840, the test set—785

[a] spring cultivars

[b] winter cultivars

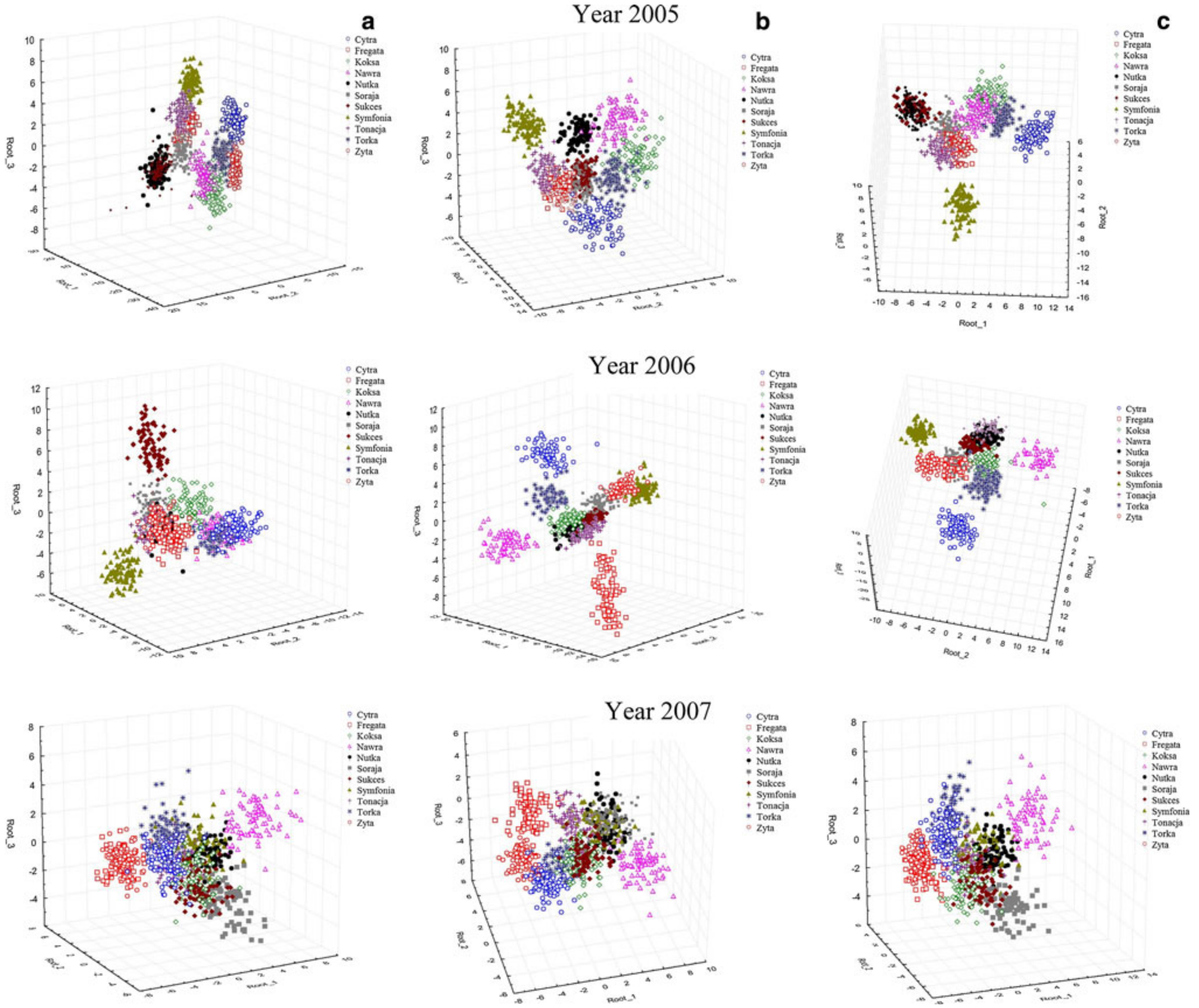


Fig. 2 Diagrams of dispersion of canonical variables for grain humidity of 12, 14 and 16% (from *left* to *right*) and the years of cultivation 2005, 2006 and 2007. Stepwise progressive analysis

classification. After they were removed from the model, the cumulative accuracy of classification ranged from 99 to 100%. Further studies should result in developing such a universal statistical model for successive years of cultivation. Most publications dealing with the issue propose models verified on data from the year in which they were developed.

Acknowledgments The author is grateful for the financial support provided by the Ministry of Scientific Research within the framework of grant no. 1089/P06/2005/29.

References

1. Brosnan T, Da-Wen S (2002) Inspection and grading of agricultural and food products by computer vision systems–a review. Comput Electron Agric 36:193–213
2. Granito PM, Garralda PA, Verdes PF, Ceccato HA (2003) Boosting classifiers for weed seeds identification. J Cereal Sci Technol 3:34–39
3. Hall M, Frank E, Holmes G, Pfahringer B, Reutemann P, Witten IH (2009) The WEKA data mining software: an update. SIGKDD Explor 11(1):10–18
4. Luo X, Jayas DS, Symons SJ (1999) Identifications of damaged kernels in wheat using a colour machine vision system. J Cereal Sci 30:49–59
5. Majumdar S, Jayas DS (2000) Classification of cereal grains using machine vision: I. Morphology models. Am Soc Agric Eng 43:1669–1675

6. Nielsen JP (2003) Evaluation of malting barley quality using exploratory data analysis. II. The use of kernel hardness and image analysis as screening methods. J Cereal Sci 38:247–255
7. Pablo MG, Pablo F, Verdes H, Ceccatto A (2005) Large-scale investigation of weed seed identification by machine vision. Comput Electron Agric 47:15–24
8. Paliwal J, Visen NS, Jayas DS, White NDG (2003) Cereal grain and dockage identification using machine vision. Biosystems Eng 85:51–57
9. Paliwal J, Visen NS, Jayas DS, White NDG (2003) Comparison of a neural network and non-parametric classifier for grain kernel identification. Biosystems Eng 85:405–413
10. Paliwal J, Visen NS, Jayas DS (2001) Evaluation of neural network architectures for cereal classification using morphological features. J Agric Eng Res 79:361–370
11. Shahin MA, Symons SJ (2001) Lentil seed size distribution with machine vision. In: St. Joseph (eds) ASAE annual international meeting, Sacramento, 30 July to 1 Aug, 2001, paper no. 01-3058
12. Shouche SP, Rastogi R, Bhagwat SG, Sainis JK (2001) Shape analysis of grains of Indiana wheat varieties. Comput Electron Agric 33:55–76
13. Szczypiński PM, Strzelecki M, Materka A, Klepaczko A (2009) MaZda-A software package for image texture analysis. Comput Methods Programs Biomed 94:66–76
14. Venora G, Grillo O, Ravalli C, Cremonini R (2009) Identification of Italian landraces of bean (*Phaseolus vulgaris* L.) using an image analysis system. Sci Hortic 121:410–418
15. Venora G, Grillo O, Shahin MA, Symons SJ (2007) Identification of Sicilian landraces and Canadian cultivars of lentil using image analysis system. Food Res Int 40:161–166
16. Visen NS, Paliwal J, Jayas DS, White NDG (2001) Specialist neural networks for cereal grain classification. Biosystems Eng 82:151–159
17. Witten IH, Frank E (2005) Data mining. Practical machine learning tools and techniques, 2nd edn. Elsevier, San Francisco, ISBN 0-12-088407-0
18. Zayas IZ, Steele JL (1990) Image analysis applications for grain science. SPIE—the international society of optical engineering. Opt Agric For 1379:151–161