Computers and Electronics in Agriculture 144 (2018) 260–268

Contents lists available at ScienceDirect

Computers and Electronics in Agriculture

journal homepage: www.elsevier.com/locate/compag

Original papers

An evaluation of utilizing geometric features for wheat grain classification using X-ray images

Małgorzata Charytanowicz[a,c,*], Piotr Kulczycki[b,c], Piotr A. Kowalski[b,c], Szymon Łukasik[b,c], Róża Czabak-Garbacz[d]

[a] Institute of Mathematics and Computer Science, Faculty of Mathematics, Informatics and Landscape Architecture, The John Paul II Catholic University of Lublin, PL 20–708 Lublin, Poland
[b] Division for Information Technology and Systems Research, Faculty of Physics and Applied Computer Science, AGH University of Science and Technology, PL 30–059 Cracow, Poland
[c] Systems Research Institute, Centre of Information Technology for Data Analysis Methods, Polish Academy of Sciences, Newelska 6, PL 01–447 Warsaw, Poland
[d] Department of Physiopathology, Institute of Rural Health, PL 20-090 Lublin, Poland

ARTICLE INFO

Keywords:
Grain classification
Principal component analysis
Factor analysis
Correlations
Morphological features
Image processing
X-ray imaging
Object recognition

ABSTRACT

Nowadays, with the rapid development of digital image processing, there has been a notable increase in elaborating advanced tools for studying the internal structure of objects. This may be very helpful in characterizing certain morphological traits of grains, as well as in quantifying the differences between them. The current research was carried out to study the structure of the traits and to determine their importance in relation to grain classification and identification. To achieve better performance and deeper understanding of their usefulness, the investigation was done by means of both principal component analysis and multivariate factor analysis. Herein, the percentage of variation explained by the first three factors reached a high 89.97%. Thus, the presented methodology supported reliable discrimination of the wheat varieties as regards their shape descriptors. The conducted study confirmed the practical usefulness and effectiveness of the evolved method when applied to the many practical tasks wherein the image analysis commonly employed in multivariate statistical methods is recommended.

1. Introduction

Recent advances in information technology and in digital image processing have resulted in the development and practical usage of computer-aided techniques in data analysis. The three-dimensional nature of computed tomography scanning allows the same object to be scanned multiple times and provides an opportunity to investigate its particle at any location within a sample. In the last few years, research studies indicate that X-ray computed tomography provides an alternate approach for the nondestructive measuring technique (Papadopoulos et al., 2009; Peth et al., 2010). This is very useful in characterizing the internal structure of objects and in quantifying their geometric features (Charytanowicz, 2014; Charytanowicz and Kulczycki, 2014; Czachor et al., 2015). In our research, the feature extraction method based on the utilization of X-ray images is proposed to measure several grain traits, hence, enabling their classification.

The identification of wheat grain requires some knowledge of their characteristics. High classification accuracy can be obtained by using kernel shape, color, length, and texture (Wiwart et al., 2012; Zapotoczny, 2011). They are considered as major distinctions and can be combined to construct the feature vector, which represents wheat grain. In the previous studies on wheat classification morphology, color, and texture were exploited for wheat variety recognition (Utku, 2000). Indeed, Majumdar and Jayas have suggested several different approaches for classification cereal grains using different types of features and their combinations (Majumdar and Jayas, 2000a, 2000b, 2000c, 2000d).

Various computer-aided systems based on morphological features for the classification have been reported in literature (Guevara-Hernandez and Gomez-Gil, 2011; Niewczas et al., 1995). The majority of different features have involved the identification of grain varieties. The key problem encountered in practice is a very large number of variables. Still, when the dimensionality of the data increases, classification problems become significantly harder. A high number of features can lead to lower classification accuracy. Moreover, the amount of computations required for classification increases exponentially with

* Corresponding author at: Systems Research Institute, Centre of Information Technology for Data Analysis Methods, Polish Academy of Sciences, Newelska 6, PL 01–447 Warsaw, Poland.
E-mail address: mchmat@ibspan.waw.pl (M. Charytanowicz).

https://doi.org/10.1016/j.compag.2017.12.004
Received 9 February 2017; Received in revised form 17 October 2017; Accepted 5 December 2017

the growth of data dimension. On the other hand, a reduction of the space dimensionality leads to a better understanding of a model and simplifies the usage of different visualization techniques (Camastra, 2003). A main problem is identifying a representative set of features from which a classification model for a particular task will be constructed. This addresses the problem of feature selection through a correlation based approach.

A logical attitude for categorizing the traits requires the use of multivariate statistical methods such as principal component analysis and factor analysis. These methods basically reduce a high number of variables to several components without a significant loss in classification accuracy. These approaches also enable the user to detect and explain correlations among variables. In addition, dimensionality reduction aims to reveal meaningful structures and guarantees to show the genuine properties of the original data (Janecek et al., 2008; Tian et al., 2010).

Vahid et al. (2011) has employed factor analysis to research the relationship in totality of some quantitative traits with regard to wheat grain yield under end drought stress via factorial split plot and on the basis of completely randomized block design in three replications. In this work, according to factor analysis, through decomposition to main components, four factors altogether explained 83.51% of all variations. The first factor was called the effective factor to yield. The second factor was that of effective traits to spike characteristics. The third factor was called the effective factor to plant height. Finally, the fourth factor was called the effective factor to plant growth. Leilah and Al-Khateeb (2005) have harnessed factor analysis and principal components analysis to study the relationship between wheat grain yield and its components under the drought conditions. Herein, three main factors accounted for 74.4% of the total variability in the dependent structure. The results showed that biological yield, harvest index, weight of grains/spike, spike length and number of spikes/m^2 had the highest communality, and, consequently, the high relative contribution in wheat grain yield.

Furthermore, the accomplished studies showed that the digital image processing techniques commonly applied in multivariate statistical analysis give reliable results in recognizing wheat varieties. In the previous paper (Charytanowicz et al., 2010), conducted in the earlier stage of the research study, an effective gradient clustering algorithm was proposed for wheat variety classification. This had resource to the basic geometric features, including the kernel area, kernel perimeter, compactness, kernel length, kernel width, asymmetry coefficient and length of kernel groove, common to particular wheat grain varieties. The presented procedure achieved an accuracy of about 92%. The utility of the investigated methodology in the context of discrimination by way of using the geometric features of wheat grain has been also confirmed by the results of discriminant analysis. In the work (Charytanowicz et al., 2016), the selected combination of geometric features, including the kernel perimeter, compactness, asymmetry coefficient, the ratio between the germ length and the kernel length, the ratio between the germ area and the kernel area, and the ratio between the kernel width and the kernel length, had significant contribution to the discrimination, and such features have permitted discriminant analysis to achieve a recognition rate of 89–96%. Moreover, in the paper (Kulczycki and Łukasik, 2014), the problems of reducing data set dimension and size were investigated. After the reduction process, the number of wheat grains assigned to the right class was very high – achieving almost 90%. Finally, the data set of wheat grain was used to verify classification quality results for the probabilistic neural network with simplified structure (Kowalski and Kusy, in press). The obtained results confirmed the practical usefulness of the proposed methodology.

Thus, this paper demonstrate the utilization of grain geometric traits in wheat variety recognition. The main objective of this work is to determine a basic set of these parameters with respect to wheat grain morphology. Both principal component analysis and multivariate factor analysis are exploited to better comprehend the relations between traits, as well as to identify effective factors in wheat grain classification.

2. Materials and methods

The research was conducted at the Institute of Agrophysics of the Polish Academy of Sciences in Lublin. For this work, we chose combine harvested wheat grain of three varieties: Canadian, Kama, and Rosa. Herein, the relationship between the different grain traits and their effectiveness in grain classification were studied by way of utilizing two multivariate statistical procedures which can be employed for structure detection: principal component analysis and factor analysis (Basilevsky, 1994; Jolliffe, 2002; Morrison, 2005). Furthermore, the Pearson correlation analysis was carried out for all analyzed variables (Draper and Smith, 1981). The significant differences between mean values were tested by analysis of variance and the Tukey's test.

2.1. Data acquisition

In our work, image processing methods were used to acquire the data. In order to evaluate the quantitative traits of wheat grains, a high quality visualization of the internal kernel structure was done through the application of a soft X-ray technique. This is an objective, precise and nondestructive method that is considerably cheaper than other more sophisticated techniques such as magnetic resonance imaging, scanning microscopy or laser technology. For each X-ray exposure, grain kernels were evenly positioned groove down. The images were obtained in the form of photograms at the scale of 5:1.

The photograms were then scanned by way of an Epson Perfection V700 table photo-scanner that was equipped with a transparency adapter, at 600 dpi resolution and 8 bit gray scale levels. This produced bitmap graphics files with a sufficient resolution for reflecting distinct features important for the proper characterization of objects.

Fig. 1 presents the exemplary X-ray images of these kernels for each studied variety: Canadian, Kama, and Rosa.

However, sole visualization of the kernels did not provide quantitative measures of shape parameters and their relations. In order to carry out accurate grain traits measurements, the specialized image processing package Grains (Niewczas and Wozniak, 1999; Strumillo et al., 1999), which incorporates image processing algorithms, was employed for measuring the particular characteristics of any selected grain. Using the commands available in the program menu, automatic boundary detection and diverse measurements relevant to the study were enabled for each individual kernel. In our research, to evaluate the factors affecting the grain differentiation, the following traits were measured: the kernel area, kernel perimeter, compactness, asymmetry coefficient, kernel length, kernel width, length of kernel groove, germ area, germ length, and additionally the ratio between the germ length and the kernel length, the ratio between the germ area and the kernel area, as well the ratio between the kernel width and the kernel length.

Compactness is a shape descriptor computed according to the formula:

$$C = 4\pi \times \frac{A}{P^2} \tag{1}$$

where A denotes the kernel area and P denotes the kernel perimeter. The maximum value of the compactness is equal to one and is taken for a circle. For values close to zero, the shape is increasingly elongated.

The asymmetry coefficient given by the formula

$$AC = \frac{|A_{left} - A_{right}|}{A} \tag{2}$$

is the ratio of two quantities: the absolute value of the difference between areas of the left and right part of a kernel, and the total area of that kernel.

Finally, the obtained data incorporated twelve real-valued

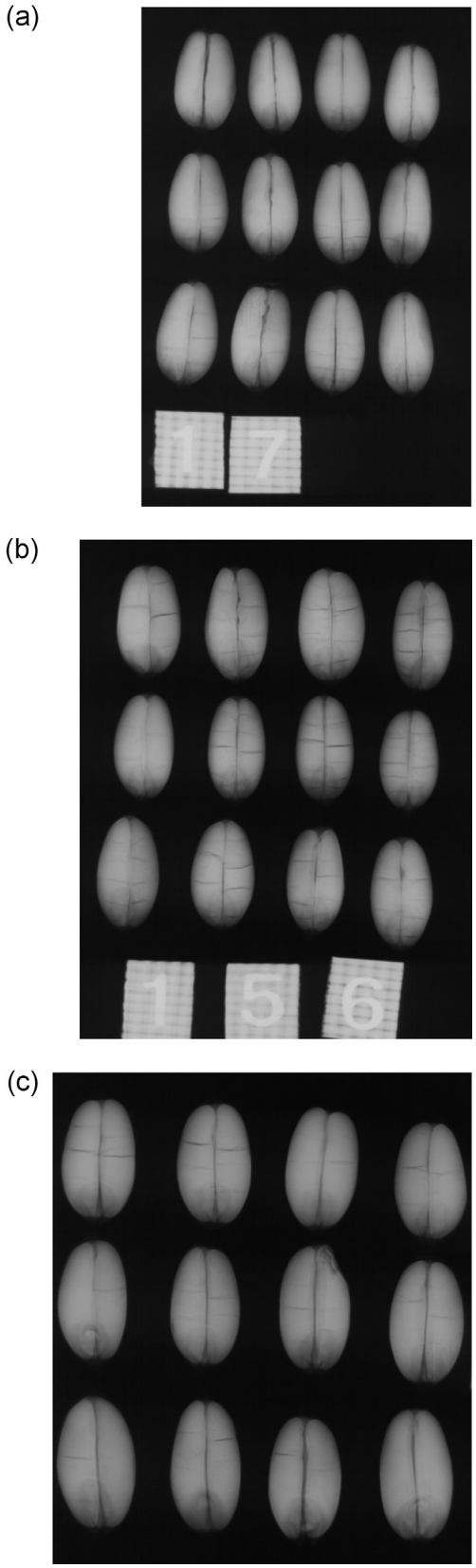

Fig. 1. Exemplary X-ray photograms of Canadian (a), Kama (b), and Rosa (c) wheat kernels.

continuous grain characteristics. The main aim of our research is to analyze their mutual relations and utilization in the classification process.

2.2. The statistical techniques

The combined data, which contained the different parameters extracted from the images, was analyzed by way of applying two statistical methods: principal component analysis and factor analysis. The appropriate data analysis was performed using Statistica 10 software (StatSoft, Poland), this having implemented advanced statistical methods.

2.2.1. Principal component analysis

Principal component analysis is a multivariate technique which can be used to transform a number of correlated variables into a small number of independent variables called principal components, that capture as much of the variability in the original variables as possible.

Suppose that a set of observable variables $X = X_1, X_2, \ldots, X_p$ is given. Principal components are obtained as linear combinations of the observed variables:

$$Z_i = a_{i1}X_1 + a_{i2}X_2 + \ldots + a_{ip}X_p \tag{3}$$

for $i = 1,2,\ldots,p$, where $a_{i1}, a_{i2}, \ldots, a_{ip}$ denote optimal weights of a principal component Z_i for observable variables $X_1, X_2, \ldots, X_p$. Their computing requires finding eigenvectors and eigenvalues from the covariance matrix of the examined variables. Eigenvectors that correspond to the largest eigenvalues are then used to reconstruct a large fraction of the variance of the original data. The first principal component accounts for the maximal amount of total variance in the observed variables. Each subsequent component accounts for a maximal amount of variance in the observed variables that was not accounted for by the preceding components, and is not correlated with the other components.

The sum of the first k eigenvalues divided by the sum of all eigenvalues

$$\frac{\lambda_1 + \lambda_2 + \ldots + \lambda_k}{\lambda_1 + \lambda_2 + \ldots + \lambda_p} \times 100\% \tag{4}$$

represents the proportion of total variation explained by the first k principal components. Similarly, the value defined as

$$\frac{\lambda_i}{\lambda_1 + \lambda_2 + \ldots + \lambda_p} \times 100\% \tag{5}$$

allows to determine the proportion of total variation explained by the i-th principal component.

The number of components to retain are usually determined by Kaiser or scree plotting methods. Finally, the original space has been reduced to the space spanned by a few eigenvectors. In this work, principal component analysis constitutes a variable reduction technique that can be used as an exploratory data analysis tool.

2.2.2. Factor analysis

Factor analysis provides a mathematical model which can be employed to describe a collection of observed variables in terms of a smaller collection of latent variables called factors. The basic assumption is that intercorrelated variables have common factors running through them, and that the aforementioned can be represented more efficiently in terms of these reference uncorrelated factors.

Thus, consider a set of observable variables $X = X_1, X_2, \ldots, X_p$. These can be represented according to the system of linear equations

$$X_i = a_{i1}F_1 + a_{i2}F_2 + \ldots + a_{ir}F_r + e_i \tag{6}$$

for $i = 1,2,\ldots,p$, where $F = F_1, F_2, \ldots, F_r$ are unobservable variables called common factors, whilst a_{ij} denotes the factor loading of variable X_i on factor F_j for $j = 1,2,\ldots,r$ whilst e_i is called a specific factor of X_i with zero mean and finite variance δ_i. Given

$$h_i = a_{i1}^2 + a_{i2}^2 + \ldots + a_{ir}^2 \tag{7}$$

for $i = 1,2,\ldots,p$, otherwise known as the contribution of r common

Table 1
Basic statistics (mean, standard deviation SD, minimum, and maximum) for the estimated traits of wheat.

Variable	Variety*	Mean	SD	Min	Max
Kernel area (V_1)	C	12.07^{c}	1.14	9.42	14.66
	K	14.29^{b}	1.23	11.23	17.08
	R	19.01^{a}	2.18	12.84	23.58
	Total	15.23	3.46	9.42	23.58
Kernel perimeter (V_2)	C	13.33^{c}	0.54	11.87	14.53
	K	14.28^{b}	0.58	12.63	15.46
	R	16.40^{a}	0.92	13.70	18.45
	Total	14.72	1.53	11.87	18.45
Compactness (V_3)	C	0.85^{b}	0.02	0.80	0.91
	K	0.88^{a}	0.02	0.84	0.92
	R	0.89^{a}	0.02	0.85	0.91
	Total	0.87	0.02	0.80	0.92
Kernel length (V_4)	C	5.25^{c}	0.20	4.62	5.76
	K	5.50^{b}	0.23	4.90	6.05
	R	6.25^{a}	0.39	5.32	7.24
	Total	5.69	0.53	4.62	7.24
Kernel width (V_5)	C	2.89^{c}	0.20	2.57	3.47
	K	3.24^{b}	0.18	2.85	3.68
	R	3.75^{a}	0.25	3.03	4.29
	Total	3.30	0.43	2.57	4.29
Asymmetry coefficient (V_6)	C	5.04^{a}	1.64	1.66	9.64
	K	2.85^{c}	1.43	0.77	8.50
	R	3.46^{b}	1.24	1.47	7.77
	Total	3.90	1.71	0.77	9.64
Length of kernel groove (V_7)	C	5.13^{b}	0.22	4.53	5.72
	K	5.09^{b}	0.26	4.52	5.88
	R	6.10^{a}	0.38	5.09	7.15
	Total	5.48	0.57	4.52	7.15
Germ area (V_8)	C	1.77^{b}	0.25	1.06	2.33
	K	1.81^{b}	0.29	1.25	2.60
	R	2.64^{a}	0.57	1.61	4.16
	Total	2.11	0.58	1.06	4.16
Germ length (V_9)	C	1.61^{b}	0.16	1.05	1.97
	K	1.67^{b}	0.15	1.32	2.10
	R	1.85^{a}	0.26	1.33	2.66
	Total	1.71	0.22	1.05	2.66
Ratio of V_8 to V_1 (V_{10})	C	0.15^{a}	0.02	0.08	0.18
	K	0.13^{c}	0.02	0.09	0.16
	R	0.14^{b}	0.02	0.10	0.20
	Total	0.14	0.02	0.08	0.20
Ratio of V_9 to V_4 (V_{11})	C	0.31^{x}	0.03	0.20	0.37
	K	0.30^{y}	0.02	0.25	0.36
	R	0.29^{x}	0.03	0.23	0.38
	Total	0.30	0.03	0.20	0.38
Ratio of V_5 to V_4 (V_{12})	C	0.55^{c}	0.03	0.47	0.66
	K	0.59^{b}	0.03	0.53	0.67
	R	0.60^{a}	0.03	0.52	0.67
	Total	0.58	0.04	0.47	0.67

* C – Canadian, K – Kama, R – Rosa; a, b, c – Means signed by the some letter differ not significantly at alpha = 0.01; x, y – Means signed by the some letter differ significantly at alpha = 0.01.

factors on i-th variable, the value

$$g_j = a_{1j}^2 + a_{2j}^2 + ...+a_{pj}^2 \quad (8)$$

for $j = 1,2,...,r$, is the contribution of factor F_j for X. After extraction of factorial loads of matrix, the matrix is generally followed by a rotation of the factors that were retained to improve the opportunity of achieving meaningful interpretation of each factor. Factor analysis was used to explore and interpret underlying patterns and structure in our data.

3. Results and discussion

The studied material was firstly subjected by image analysis to determine twelve geometric traits. All measurements were made from a total of 288 samples of three wheat varieties: Canadian, Kama, and Rosa, containing 108, 72, and 108 kernels, respectively. Table 1 shows the arithmetic mean, minimum and maximum values, as well as the standard deviation for all estimated wheat traits, both separately for each variety and combined.

The ANOVA results indicated significant differences between Canadian, Kama and Rosa varieties. Average measurements, including the kernel area, kernel perimeter, kernel length and kernel width, as well as the ratio between the kernel width and the kernel length, were significantly higher in Rosa, and significantly lower in the Canadian variety. The Kama measurements fell between the values for Canadian and Rosa varieties. The average length of the kernel groove, germ area and germ length, were significantly higher for Rosa, in comparison with the two other varieties, Canadian and Kama, which were not significantly differentiated. The average asymmetry coefficient and the average ratio between the germ area and the kernel area were significantly higher in Canadian, and significantly lower in the Kama variety. The compactness differentiated only the Canadian variety from the other two varieties with significantly higher values. The average ratio between the germ length and the kernel length significantly differentiated only Canadian and Rosa varieties. No significant differences were observed between Kama and Canadian, as well as Kama and Rosa varieties. The Rosa variety was marked by greater variation than Canadian and Kama varieties, with regard to the majority of studied shape measurements.

For each variety, relatively high variation was observed for the asymmetry coefficient, whereas relatively low variation was observed for compactness (herein, the coefficient of variation was between 1.7% and 2.8%).

The majority of the analyzed variables were significantly correlated. The derived results revealed that all traits characterizing geometric measures, including the kernel area, kernel perimeter, compactness, kernel length, kernel width, length of kernel groove, germ area, and germ length were significantly and positively correlated with each other. The kernel area and kernel perimeter were the most perfectly correlated. However, one is determined by the other and such measures did not add any additional information.

Correlation coefficients of variables with each other are presented in Table 2.

Furthermore, correlation coefficients between the kernel area and other mentioned measures were very high (r values from 0.62 to 0.99). Of course, the same results were noticed for the kernel perimeter (r values from 0.56 to 0.99). However, the asymmetry coefficient had significant negative correlations with these geometric measures (r values from −0.43 to −0.36). Indeed, the last three characteristics were not always significantly correlated with investigated traits. The ratio between the germ length and the kernel length and, consequently, the ratio between the germ area and the kernel area, were not significantly correlated with the kernel area, kernel perimeter, kernel length and length of kernel groove. The correlation coefficients were close to zero (r values from −0.08 to 0.12). The opposite results were observed for the ratio between the kernel width and the kernel length (significant r values from 0.24 to 0.57). The final evaluation based on simple correlation coefficients alone cannot provide complete information on the complex relations of the traits, but it does indicate the existence of distinct structure in the data. Additionally, if the number of features is too high, then it is beneficial to reduce the number of features through a feature extraction technique. Hence, so as to take into consideration the various advantages of multivariate statistical methods, principal component analysis, as well as factor analysis, were used in the current study.

Table 2
A matrix of correlation coefficients for the estimated traits of wheat grains.

	V_1	V_2	V_3	V_4	V_5	V_6	V_7	V_8	V_9	V_{10}	V_{11}
V_1	.	.	.	.	.	.	.	.	.	.	.
V_2	0.99*	.	.	.	.	.	.	.	.	.	.
V_3	0.62*	0.56*	.	.	.	.	.	.	.	.	.
V_4	0.96*	0.98*	0.42*	.	.	.	.	.	.	.	.
V_5	0.97*	0.96*	0.76*	0.89*	.	.	.	.	.	.	.
V_6	−0.40*	−0.39*	−0.43*	−0.36*	−0.41*	.	.	.	.	.	.
V_7	0.90*	0.92*	0.31*	0.96*	0.81*	−0.25*	.	.	.	.	.
V_8	0.83*	0.84*	0.31*	0.84*	0.76*	−0.22*	0.84*	.	.	.	.
V_9	0.65*	0.66*	0.14*	0.72*	0.56*	−0.23*	0.69*	0.86*	.	.	.
V_{10}	−0.08	−0.05	−0.45*	0.03	−0.16*	0.26*	0.12	0.49*	0.53*	.	.
V_{11}	−0.08	−0.07	−0.25*	−0.01	−0.12*	0.05	0.00	0.34*	0.69*	0.73*	.
V_{12}	0.57*	0.51*	0.95*	0.34*	0.73*	−0.32*	0.24*	0.29*	0.08	−0.40*	−0.25*

* Significant at the 0.01 level.

3.1. Principal component analysis

The results presented in the previous section, as well as the significance generated by means of the Bartlett test, confirmed the meaningfulness of principal component analysis. The method was carried out on the basis of the correlation matrix. In Table 3, eigenvalues of principal components and the percentage of the primary variable variance carried by these, are shown.

According to the Kaiser criterion (which keeps any components that has an eigenvalue greater than one), and evaluation of the scree plotting (Cattell, 1996), three components were chosen for further analysis (Fig. 2). These components all together accounted for 89.97% of the total variability: the first component carried 56.56%, the second 24.66% and the third 8.75% of the variation.

Accurate interpretation of the components was possible by means of factor loadings. On the basis of the values presented in Table 4, it may be concluded that the first component was most positively correlated with seven variables which constituted basic geometric measurements. These were: the kernel area, kernel perimeter, kernel length, kernel width, length of kernel groove, germ area, and germ length (loadings greater than 0.7). Additionally, compactness and ratio between the kernel width and the kernel length had loadings greater than 0.5. The asymmetry coefficient had a negative loading equal to −0.438. A weak correlation with the ratio between the germ length and the kernel length, and the ratio between the germ area and the kernel area, was consistent with the results obtained for the second component – which was most negatively correlated with the above variables (loadings less than −0.8). The second component is also mostly correlated with the compactness and germ length, but the strength of the correlation was worse, in comparison with the first component. The third component was most negatively correlated with the compactness, the ratio between the germ length and the kernel length, and the ratio between the kernel width and the kernel length. Herein, the strength of the correlation was moderate (loadings less than −0.4), in comparison with the first and second principal components.

Thus, the first component could be viewed as a measure of shape, whilst the second component carried information related mostly to the characteristics describing ratios between measurements of germ and kernel. The third component had lower absolute values of these three loadings. It is worth noting that the distinguished variables were affected in the same direction on the corresponding components. Table 5 shows the computed communalities which carried the total amount of variance of the original variables, as explained by each component.

Table 3
Eigenvalues from PCA, and related statistics.

Value number	Eigenvalue	% of total variation	Cumulative eigenvalue	Cumulative %
1	6.788	56.565	6.788	56.565
2	2.960	24.663	9.747	81.228
3	1.050	8.746	10.797	89.974
4	0.802	6.685	11.599	96.659
5	0.290	2.418	11.889	99.076
6	0.056	0.469	11.945	99.545
7	0.037	0.307	11.982	99.851
8	0.010	0.081	11.992	99.933
9	0.005	0.045	11.997	99.978
10	0.001	0.012	11.999	99.990
11	0.001	0.007	12.000	99.997
12	0.000	0.003	12.000	100.000

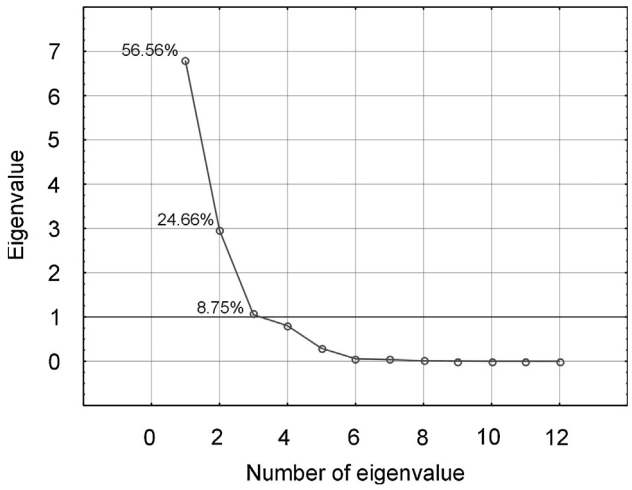

Fig. 2. Scree plot showing eigenvalues in response to number of components for the estimated variables.

Table 4
Factor loadings of the principal components for the estimated variables of wheat.

Variable	*PC*1	*PC*2	*PC*3
Kernel area	0.989*	0.069	0.114
Kernel perimeter	0.982*	0.025	0.166
Compactness	0.637	0.601	−0.413
Kernel length	0.949*	−0.105	0.263
Kernel width	0.971*	0.207	−0.023
Asymmetry coefficient	−0.438	−0.264	0.341
Length of kernel groove	0.893*	−0.188	0.360
Germ area	0.868*	−0.436	−0.009
Germ length	0.717*	−0.633	−0.187
Ratio of V_8 to V_1	0.008	−0.911*	−0.138
Ratio of V_9 to V_4	0.037	−0.813*	−0.532
Ratio of V_5 to V_4	0.581	0.593	−0.430

* Loadings in absolute value greater than 0.7.

Table 5
Cumulative % of variation explained by way of the principal components for the estimated variables of wheat.

Variable	*PC*1	*PC*2	*PC*3
Kernel area	0.978	0.983	0.996
Kernel perimeter	0.965	0.966	0.993
Compactness	0.406	0.767	0.938
Kernel length	0.901	0.912	0.982
Kernel width	0.942	0.985	0.985
Asymmetry coefficient	0.192	0.262	0.378
Length of kernel groove	0.798	0.833	0.962
Germ area	0.753	0.943	0.943
Germ length	0.514	0.916	0.951
Ratio of V_8 to V_1	0.000	0.830	0.849
Ratio of V_9 to V_4	0.001	0.662	0.946
Ratio of V_5 to V_4	0.337	0.689	0.874

These values, besides the values corresponding to the asymmetry coefficient, were very high.

Component scores were computed as linear combinations of the observed variables weighted by eigenvectors. The first three components took the form:

$$PC_1 = 0.380V_1 + 0.377V_2 + 0.244V_3 + 0.364V_4 + 0.373V_5 - 0.168V_6 + 0.343V_7 + 0.333V_8 + 0.275V_9 + 0.003V_{10} + 0.014V_{11} + 0.223V_{12}, \quad (9)$$

$$PC_2 = 0.040V_1 + 0.015V_2 + 0.350V_3 - 0.061V_4 + 0.120V_5 - 0.153V_6 - 0.109V_7 - 0.253V_8 - 0.368V_9 + 0.530V_{10} - 0.437V_{11} - 0.345V_{12}, \quad (10)$$

$$PC_3 = 0.111V_1 + 0.162V_2 - 0.403V_3 + 0.257V_4 - 0.023V_5 - 0.333V_6 + 0.351V_7 - 0.009V_8 - 0.183V_9 - 0.135V_{10} - 0.519V_{11} - 0.420V_{12} \quad (11)$$

The wheat grain data projection on the axes of the first three principal components is presented in Fig. 3.

The first component was found to differentiate mostly between Rosa, and Canadian and Kama combined. The second and third components differentiated Canadian and Kama varieties.

3.2. Factor analysis

Factor analysis was performed by means of the principal components method. Table 6 presents the factor loadings of matrix, as well as an estimation of the number of factors.

In pursuance of the assumed criteria, a three–factors model was fitted to the data. The factors rotation was done by way of varimax rotation. Varimax rotation decreased the amount of variation that is explained by the first two factors and increased the amount of variation that is explained by the third factor.

A summary of high factor loadings for the estimated variables of wheat is presented in Table 7. According to the obtained results, twelve traits, subsequently divided into three comprehensive factors, explained 89.97 percent of the total variability in the dependent structure. The share of each factor is 47.9%, 19.8% and 22.3% respectively.

The first factor, which accounted the greatest bulk of data variation, had positive and very high loads for six traits: the kernel area, kernel perimeter, kernel length, kernel width, length of kernel groove, and germ area (loadings greater than 0.8). These could be regarded mainly as the factor relating to basic shape measurements of a kernel. Additionally, a moderate correlation was observed with the germ length (the positive load equal to 0.668) and a weak correlation with the ratio between the germ area and the kernel area and ratio between the germ length and the kernel (the absolute values of loadings less than 0.1). These were consistent with the results obtained for the second factor, which included higher loads with these traits, equaled to 0.707, 0.968 and 0.813, correspondingly. This could be regarded primarily as the factor relating to the ratio between germ and kernel. The third factor included compactness and the ratio of kernel width to kernel length, and showed both high and positive loads, equaled to 0.902 and

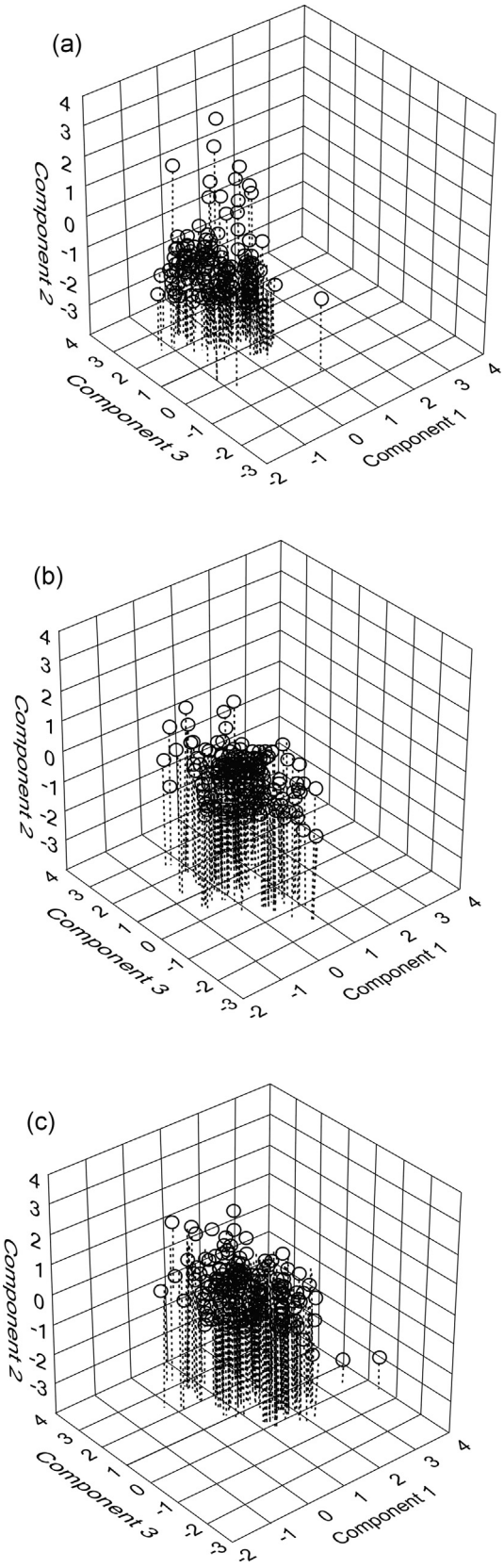

Fig. 3. 3D scatterplots of component scores for three wheat varieties: Canadian (a), Kama (b), and Rosa (c).

Table 6
Rotated factor loadings and communalities C for the estimated variables of wheat.

Variable	$F1$	$F2$	$F3$	C
Kernel area	0.926*	0.001	0.372	0.999
Kernel perimeter	0.948*	0.005	0.307	0.998
Compactness	0.312	−0.164	0.902*	0.963
Kernel length	0.978*	0.048	0.150	0.999
Kernel width	0.832*	−0.032	0.540	0.999
Asymmetry coefficient	−0.215	−0.041	−0.575	0.411
Length of kernel groove	0.980*	0.052	0.011	0.944
Germ area	0.845*	0.463	0.118	0.994
Germ length	0.668	0.707*	0.067	0.998
Ratio of V_8 to V_1	0.093	0.813*	−0.423	0.983
Ratio of V_9 to V_4	−0.057	0.968*	−0.077	0.996
Ratio of V_5 to V_4	0.256	−0.154	0.886*	0.997
Total	5.747	2.370	2.679	10.797
Proportion	0.479	0.198	0.223	0.899

* Loadings greater than 0.7.

Table 7
Summary of factor loadings for the estimated variables of wheat.

Variable	Loading	% of the total variability
Factor 1	5.747	47.9%
Kernel area	0.926	
Kernel perimeter	0.948	
Kernel length	0.978	
Kernel width	0.832	
Length of kernel groove	0.980	
Germ area	0.845	
Factor 2	2.370	19.8%
Germ length	0.707	
Ratio of V_8 to V_1	0.813	
Ratio of V_9 to V_4	0.968	
Factor 3	2.679	22.3%
Asymmetry coefficient	−0.575	
Compactness	0.902	
Ratio of V_5 to V_4	0.886	

0.886 correspondingly. This also revealed a negative correlation with the asymmetry coefficient (the load equaled to −0.575). The aforementioned final factor could be regarded as the factor relating to the additional measures based on the basic geometric traits of any kernel.

The obtained results demonstrated the importance of all examined shape descriptors in wheat variety classification. These traits besides the asymmetry coefficient, had very high communalities. Similarly, all loadings besides the asymmetry coefficient, were very high. Subsequently, based on factor score coefficients, factor scores were calculated and analyzed in relation to wheat classification. These factors took the form:

$$F_1 = 0.172V_1 + 0.193V_2 - 0.107V_3 + 0.233V_4 + 0.109V_5 + 0.088V_6 + 0.268V_7 + 0.134V_8 + 0.056V_9 - 0.005V_{10} - 0.158V_{11} - 0.121V_{12}, \quad (12)$$

$$F_2 = -0.064V_1 - 0.081V_2 + 0.078V_3 - 0.100V_4 - 0.026V_5 - 0.126V_6 - 0.132V_7 + 0.139V_8 + 0.289V_9 + 0.324V_{10} + 0.517V_{11} + 0.088V_{12}, \quad (13)$$

$$F_3 = -0.002V_1 - 0.046V_2 + 0.433V_3 - 0.139V_4 + 0.115V_5 - 0.307V_6 - 0.223V_7 - 0.025V_8 + 0.047V_9 - 0.084V_{10} + 0.201V_{11} + 0.439V_{12}. \quad (14)$$

Fig. 4 presents 3D scatterplots of factor scores for each wheat variety. The first factor distinctly differentiated Rosa, from Canadian and Kama combined. The third factor was found to differentiate mostly between the Canadian and Kama varieties. Herein, the Rosa variety was better differentiated, whilst Kama and Canadian varieties were less successfully distinguished.

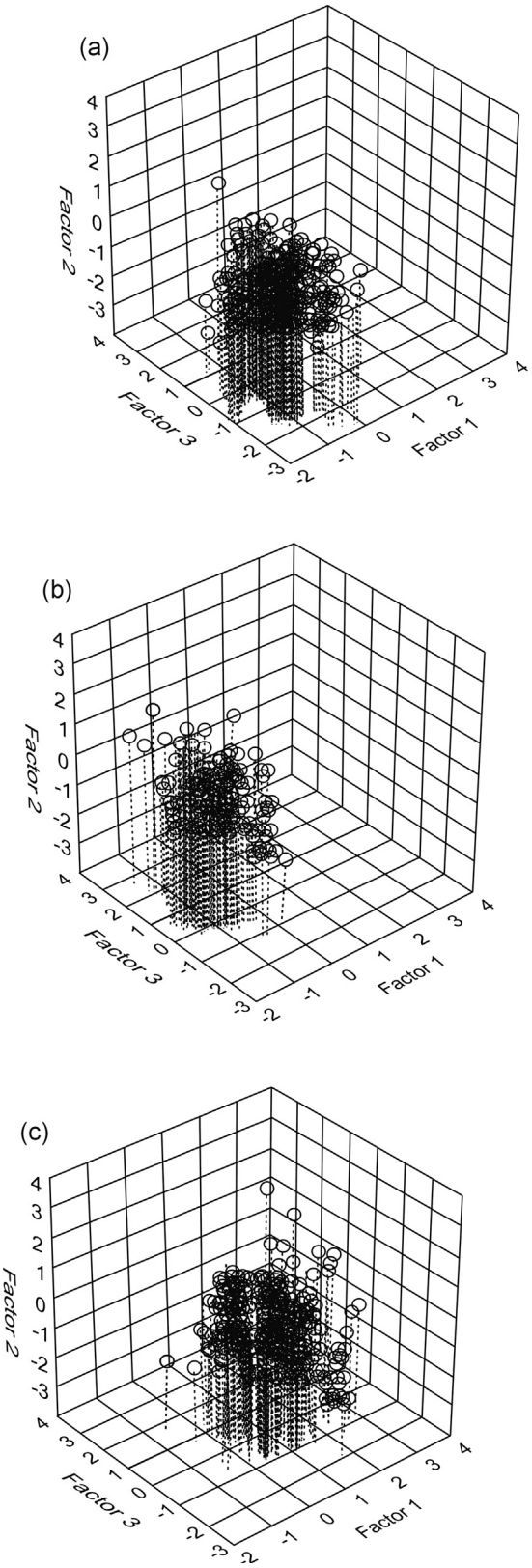

Fig. 4. 3D scatterplots of factor scores for three wheat varieties: Canadian (a), Kama (b), and Rosa (c).

4. Final remarks and summary

Computer vision-based systems, along with data analysis procedures based on classical statistical methods or computation intelligence can

be successfully exploited for classification tasks (Forczmański and Markiewicz, 2013, 2016; Hu et al., 1998; Sabanci et al., 2017). The conducted research has shown the usefulness of the main geometric features of three wheat varieties: Kama, Rosa and Canadian in such problems. Thus, in works (Charytanowicz et al., 2010; Kulczycki et al., 2012), the complete gradient clustering algorithm using nonparametric kernel estimation (Kulczycki, 2008; Silverman, 1986; Wand and Jones, 1994) was proposed. The main idea of this algorithm assumed that each cluster was identified by the local maxima of the kernel density estimator of the data distribution. The whole procedure did not need strict assumptions regarding the desired number or shape of clusters. This allowed the number obtained to be better suited to a real data structure. The number of correctly classified grains was, in order, 96%, 84%, 96% for Kama, Rosa and Canadian varieties (respectively), giving almost 92% of the total properly classified objects. The above results were comparable to that of the classic *K*-means method, although in this case, it did require additional correct information regarding the number of clusters. Furthermore, the study carried out in work (Charytanowicz et al., 2016) on an enlarged set of wheat grain geometric features, confirmed the positive properties of the proposed methodology. This gave a classification rate of 89–96%. The kernel perimeter, and, subsequently, the ratio between the germ length and kernel length, as well as the ratio between the germ area and the kernel area, were established as being most important in discrimination. The paper (Kulczycki and Łukasik, 2014) deals with the algorithm of reducing the dimension and size of a data set for the domain's fundamental tasks of exploratory data analysis. Among these tasks are clustering, classification and detection of atypical elements. In numerical experiments, the classification procedure, realized through applying the nearest neighbor algorithm, reached a 90% rate of classification for the initial space of wheat grains. Satisfactory results of classification of this data set were also obtained when the classifier based on probabilistic neural networking was used. In the article (Kowalski and Kusy, in press), the classification correctness of the probabilistic neural network with both full and reduced structure were calculated. These gave test quality values between 0.90 and 0.93.

In our research, the structure of various wheat grain morphological traits was studied to determine the factors which best explain the variability in the dataset. The result of such work is that the data, which initially incorporated twelve geometrical variables, has been reduced to a three-dimension model which justified 89.97% of the data variation as a whole. The first factor can be regarded mainly as the factor relating to the kernel's basic shape measurements. This factor accounts the greatest bulk of data variation and has positive and very high loads for six traits: the kernel area, kernel perimeter, kernel length, kernel width, length of kernel groove, and germ area The second factor is considered primarily as the factor relating to the ratio between germ and kernel. The third factor makes reference to additional measurements based on a kernel's basic geometric traits. All traits, besides the asymmetry coefficient, have very high communalities.

The study also discussed the importance of grain size and shape in the wheat varieties classification process. The results indicated significant differences between Canadian, Kama and Rosa varieties. Herein, average kernel measurements of the Rosa variety were significantly higher, in comparison to the Canadian and Kama varieties. Of note: the Rosa variety was better recognized, whilst the Canadian and Kama varieties were less successfully differentiated.

Summarizing, the multivariate statistical methods used in this work revealed the importance of the considered traits in the reliable analysis of wheat grains. The proposed basic geometric features constitutes a crucial set of parameters with respect to wheat grain morphology which best differentiate wheat varieties. The presented methodology combining image analysis and statistical methods supported reliable discrimination of the wheat varieties as regards their shape descriptors, and it allowed nondestructive and automatic feature detection. The conducted study confirmed the practical usefulness and effectiveness of the evolved method in classification practices. It should be underlined, however, that the study was conducted as a practical trial so as to clarify the relationship between the traits.

Acknowledgments

Our heartfelt thanks go to our colleague Professor Jerzy Niewczas, with whom we commenced the research presented in the previous papers. We are also very grateful for his valuable advice and inspiration.

References

Basilevsky, A.T. Statistical Factor Analysis and Related Methods: Theory and Applications. John Wiley and Sons, Inc. 1994.

Camastra, F., 2003. Data dimensionality estimation methods: a survey. Pattern Recogn. 36 (12), 2945–2954.

Cattell, R.B., 1996. The scree test for the number of factors. Multivar. Behav. Res. 1 (2), 245–276.

Charytanowicz, M., 2014. An Algorithm for the Pore Size Determination using Digital Image Analysis. Information Technologies in Biomedicine. In: Pietka, E,. Kawa, J., Wieclawek, W. (Eds.), Advances in Intelligent Systems and Soft Computing, Springer, vol. 283(3), pp. 223–234.

Charytanowicz, M., Kulczycki, P., 2014. An image analysis algorithm for soil structure identification. In: Filev, D., Jablkowski, J., Kacprzyk, J., Popchev, I., Rutkowski, L., Sgurev, V., Sotirova, E., Szynkarczyk, P., Zadrozny, S. (Eds.), Intelligent Systems '2014, Advances in Intelligent Systems and Computing, Springer, vol. 2, pp. 681–692.

Charytanowicz, M., Niewczas, J., Kulczycki, P., Kowalski, P.A., Łukasik, S., Żak, S., 2010. Complete Gradient Clustering Algorithm for Features Analysis of X-ray Images. In: Pietka, E., Kawa, J. (Eds.), Information Technologies in Biomedicine. Springer-Verlag, Berlin-Heidelberg, vol. 69(2), pp. 15–24.

Charytanowicz, M., Niewczas, J., Kulczycki, P., Kowalski, P.A., Łukasik, S., 2016. Discrimination of Wheat Grain Varieties Using X-ray Images. In: Pietka, E., Badura, P., Kawa, J., Więcławek, W. (Eds.), Information Technologies in Biomedicine, Advances in Intelligent Systems and Soft Computing, Springer, vol. 471(1), pp. 39–50.

Czachor, H., Charytanowicz, M., Gonet, S., Niewczas, J., Józefaciuk, G., Lichner, L., 2015. Impact of long term mineral and organic fertilization on water stability, wettability and porosity of aggregates of two silt loamy soils. Eur. J. Soil Sci. 66 (3), 577–588.

Draper, N.R., Smith, H., 1981. Applied Regression Analysis. John Wiley and Sons, New York.

Forczmański, P., Markiewicz, A., 2013. Low-level image features for stamps detection and classification. In: 8th International Conference on Computer Recognition Systems CORES 2013, Advances in Intelligent Systems and Computing, vol. 226, pp. 383–392.

Forczmański, P., Markiewicz, A., 2016. Two-stage approach to extracting visual objects from paper documents. Mach. Vision Appl. 27, 1243–1257.

Guevara-Hernandez, F., Gomez-Gil, J., 2011. A machine vision system for classification of wheat and barley grain kernels. Spanish J. Agricult. Res. 9 (3), 672–680.

Hu, B.-G., Gosine, R.G., Cao, L.X., de Silva, C.W., 1998. Application of fuzzy classification technique in computer grading of fish product. IEEE Trans. Fuzzy Syst. 6 (1), 144–152.

Janecek, A., Gansterer, W.N., Demel, M., Ecker, G.F., 2008. On the relationship between feature selection and classification accuracy. JMLR: Workshop and Conference Proceedings 2008, vol. 4(1), pp. 90–105.

Jolliffe, I.T., 2002. Principal Component Analysis. Springer-Verlag, New York.

Kowalski, P.A., Kusy, M., 2017. Sensitivity Analysis for Probabilistic Neural Network Structure Reduction. IEEE Transactions on Neural Networks and Learning Systems, in press, DOI: 10.1109/TNNLS.2017.2688482.

Kulczycki, P., 2008. Kernel estimators in industrial applications. In: Prasad, B. (Ed.), Soft Computing Applications in Industry. Springer Verlag, Berlin, pp. 69–91.

Kulczycki, P., Łukasik, S., 2014. An algorithm for reducing the dimension and size of a sample for data exploration procedures. Int. J. Appl. Math. Comput. Sci. 24 (1), 133–149.

Kulczycki, P., Charytanowicz, M., Kowalski, P.A., Łukasik, S., 2012. The Complete Gradient Clustering Algorithm: properties in practical applications. J. Appl. Stat. 39 (6), 1211–1224.

Leilah, A.A., Al-Khateeb, S.A., 2005. Statistical analysis of wheat yield under drought conditions. J. Arid Environ. 61 (3), 483–496.

Majumdar, S., Jayas, D., 2000a. Classification of cereal grains using machine vision: I. Morphology models. Transact. ASAE 43 (6), 1669–1675.

Majumdar, S., Jayas, D., 2000b. Classification of cereal grains using machine vision: II. Color models. Transact. ASAE 43 (6), 1677–1680.

Majumdar, S., Jayas, D., 2000c. Classification of cereal grains using machine vision: III. Texture models. Transact. ASAE 43 (6), 1681–1687.

Majumdar, S., Jayas, D., 2000d. Classification of cereal grains using machine vision: IV. Combined morphology, color, and texture models. Transact. ASAE 43 (6), 1689–1694.

Morrison, D.F., 2005. Multivariate Statistical Methods. Brooks/Cole Thomson Learning, Belmont, California.

Niewczas, J., Wozniak, W., 1999. Application of GRAINS program for characterization of X-ray images of wheat grains at different moisture content. In: Xth Seminar "Properties of Water in Foods", Warsaw Agricultural University, Department of Food Engineering.

Niewczas, J., Woźniak, W., Guc, A., 1995. Attempt to application of image processing to evaluation of changes in internal structure of wheat grain. Int. Agrophys. 9, 343–347.
Papadopoulos, A., Bird, N.R., Whitmore, A.P., Mooney, S.J., 2009. Investigating the effects of organic and conventional management on soil aggregate stability using X-ray computed tomography. Eur. J. Soil Sci. 60, 360–388.
Peth, S., Nellesen, J., Fischer, G., Horn, R., 2010. Non-invasive 3D analysis of local soil deformation under mechanical and hydraulic stresses by uCT and digital image correlation. Soil Tillage Res. 111, 3–18.
Sabanci, K., Kayabasi, A., Toktas, A., 2017. Computer vision-based method for classification of wheat grains using artificial neural network. J. Sci. Food Agric. 97 (8), 2588–2593.
Silverman, B.W., 1986. Density Estimation for Statistics and Data Analysis. Chapman and Hall, London.
Strumillo, A., Niewczas, J., Szczypinski, P., Makowski, P., Wozniak, W., 1999. Computer system for analysis of X-ray images of wheat grains. Int. Agrophys. 13 (1), 133–140.
Tian, T., Wilcox, R., James, G., 2010. Data reduction in classification: a simulated annealing based projection method. Stat. Anal. Data Min. 3 (5), 319–331.
Utku, H., 2000. Application of the feature selection method to discriminate digitized wheat varieties. J. Food Eng. 46, 211–216.
Vahid, M., Shahreyari, R., Imani, A.A., Khayanezad, M., 2011. Factor analysis of wheat quantitative traits on yield under terminal drought. Am.-Eurasian J. Agricult. Environ. Sci. 10 (2), 157–159.
Wand, M.P., Jones, M.C., 1994. Kernel Smoothing. Chapman and Hall, London.
Wiwart, M., Suchowilska, E., Lajszner, W., Graban, Ł., 2012. Identification of hybrids of spelt and wheat and their parental forms using shape and color descriptors. Comput. Electron. Agric. 83, 68–76.
Zapotoczny, P., 2011. Discrimination of wheat grain varieties using image analysis and neural networks. Part I. Single kernel texture. J. Cereal Sci. 54, 60–68.