Spanish Journal of Agricultural Research
18 (4), e0208, 12 pages (2020)
eISSN: 2171-9292
https://doi.org/10.5424/sjar/2020184-16904
Instituto Nacional de Investigación y Tecnología Agraria y Alimentaria (INIA)

RESEARCH ARTICLE

OPEN ACCESS

A quantitative multivariate methodology for unsupervised class identification in pistachio (*Pistacia vera* L.) plant leaves size

Francesca Antonucci (Antonucci, F)[1], Rossella Manganiello (Manganiello, R)[2], Corrado Costa (Costa, C)[1], Virgilio Irione (Irione, V)[2], Luciano Ortenzi (Ortenzi, L)[1], and Maria A. Palombi (Palombi, MA)[2,3]

[1] *Consiglio per la Ricerca in Agricoltura e l'Analisi dell'Economia Agraria (CREA), Centro di Ricerca Ingegneria e Trasformazioni Agroalimentari. Via della Pascolare 16, 00015 Monterotondo (Rome), Italy* [2] *CREA, Centro di Ricerca Olivicoltura, Frutticoltura e Agrumicoltura. Via Fioranello 52, 00134 Rome, Italy* [3] *CREA, Centro di Ricerca Viticoltura e Enologia. Via della Cantina sperimentale 1, 00049 Velletri (Rome), Italy*

Abstract

Aim of study: Genetic diversity of pistachio, can be evaluated by using different descriptors, as adopted in international certification systems. Mainly the descriptors are morphological traits as leaf, which represents an important organ for its sensibility to growth conditions during the expansion phase. This study adopted a rapid and quantitative non-hierarchic clustering classification (k-means), to extract size classes basing on the contemporary combination of different morphological traits (i.e., leaf stalk length, terminal leaf length, terminal leaf width and terminal leaf ratio) of a varietal collection composed by 21 pistachio cultivars.

Area of study: Worldwide.

Material and methods: The unsupervised non-hierarchic clustering technique was adopted to the entire samples of pistachio leaves from k=2 to k=15 for both four morphological variables (i.e., leaf stalk length, terminal leaf length, terminal leaf width and terminal leaf ratio) and three morphological variables (i.e., terminal leaf length, terminal leaf width and terminal leaf ratio).

Main results: A classification model only on the three morphological variables (for results of statistical analysis in which the groups resulted to be more separated and different for all the variables), with k= 5 (five groups), was constructed using a non-linear artificial neural network approach. The percentages of bad prediction in both training and testing resulted equal to 0%. The "terminal leaf length" returned the higher impact (44.89%).

Research highlights: The contemporary combination of different morphological leaf traits, allowed to create an automatic classification of size classes of great importance for cultivar identification and comparison.

Additional key words: artificial neural network; morphological analysis; clustering; germplasm collection; k-means.

Abbreviations used: ANN (Artificial Neural Network); ANOVA (Analysis of Variance); AUC (Area Under Curve); CPVO (Community Plant Variety Office); FPR (False Positive Rate); IPGRI (International Plant Genetic Resources Institute); MLP (multi-layer perceptron); PCA (Principal Component Analysis); ROC (Receiver Operating Curve); TPR (True Positive Rate); UPOV (International Union for the Protection of New Varieties of Plants); VIP (Variable Importance in Projection).

Authors' contributions: All the authors equally contributed to the writing of the paper and to its content.

Citation: Antonucci, F; Manganiello, R; Costa, C; Irione, V; Ortenzi, L; Palombi, MA (2020). A quantitative multivariate methodology for unsupervised class identification in pistachio (Pistacia vera L.) plant leaves size. Spanish Journal of Agricultural Research, Volume 18, Issue 4, e0208. https://doi.org/10.5424/sjar/2020184-16904

Received: 11 May 2020. **Accepted:** 12 Nov 2020.

Funding agencies/institutions	**Project / Grant**
Italian Ministry of Agriculture (MiPAAF)	Risorse Genetiche Vegetali-Trattato FAO, V Triennio 2017-2019 (DM 21076/2017)

Competing interests: The authors have declared that no competing interests exist.

Correspondence should be addressed to Francesca Antonucci: francesca.antonucci@crea.gov.it

Introduction

Pistachio (*Pistacia vera* L.) is one of the most popular tree nuts in the world, appreciated for its nutritional value, its health and sensorial attributes and its economic importance (Kashaninejad & Tabil, 2011). It includes about twenty species but only P. vera is edible and worldwide marketable (Fares *et al.*, 2009). The pistachio is native to the Central Asia and was introduced into Mediterranean Europe by the Romans at the beginning of the Christian era (Crane, 1978). Pistachio cultivation extended from its origin to Italy, Spain, and other Mediterranean regions of Southern Europe, North Africa, and the Middle East, as well as to China and, more recently, to the United States and Australia (Hormaza *et al.*, 1998). Nowadays, Iran, the United States, Turkey, and Syria are the main pistachio producers in the world (FAOSTAT, 2017). As reported by Massimo *et al.* (2020), Italy has a large pistachio production, especially in Bronte (Sicily).

However, as reported by Sheikhi *et al.* (2019) its production is affected by the undesired physiological characteristics of alternate bearing, shell indehiscence, blank nuts and susceptibility to abiotic stresses and fungal foliar and root diseases. For these reasons, genetic improvement should be an attempt for future breeding to produce superior pistachio cultivars. Generally, plant genetic resources preserved in *ex situ* gene bank collection could provide material for breeders in the development of cultivars with improved qualities such as increased productivity, adaptability to different agro-climatical and agro-pedological contexts, better resistance to diseases, and higher qualitative, organoleptic and nutritional characteristics (Bacchetta *et al.*, 2015; Gharaghani *et al.*, 2017).

Genetic diversity can be determined by evaluating morphological (Kafkas *et al.*, 2002; Hofer *et al.*, 2014), phenological (Chao *et al.*, 2003; Chatti *et al.*, 2017) and agro-pomological characteristics (Asma & Ozturk, 2005; Scheldeman *et al.*, 2006) as well as determined by the application of DNA markers (Pazouki *et al.*, 2010; Hofer & Peil, 2015). Part of these studies on morphological traits are based on descriptors which have been adopted by the International Union for the Protection of New Varieties of Plants (UPOV, 2020). A descriptors list regarding morphological and carpological traits of pistachio was developed by the International Plant Genetic Resources Institute (IPGRI, 1997). As reported by Antonucci *et al.* (2012), this document provides an international format producing a universally understood "language" for plant genetic resources data collection assisting, with the standardization of descriptor definitions, both the researcher, for the management and maintenance of the collection, and the users of the plant genetic resources.

Usually, in plant species morphological characterization is evaluated by analyzing leaf, flower and fruit descriptors (Hassoon *et al.*, 2018). Leaves represent an extremely important organs for plant (both trees and herbaceous species) because it is very sensitive to growth conditions during the expansion phase (Bayramzadeh *et al.*, 2008). As consequence, leaf characteristics could effectively be used to classify different species (Lin *et al.*, 1984; Kafkas *et al.*, 2002), and to discriminate among varieties (Chatti *et al.*, 2017). Sabzi *et al.* (2020) designed an imaging computer vision system to automatically classify different types of tree leaf images. The analysis of morphological leaf traits supplies deep insight into the taxonomy, genetics, biogeography and evolution (Balduzzi *et al.*, 2017), which are parts of the major classification of scientific areas related to a successful conservation of natural ecosystems. When using leaf to discriminate varieties, various methods to quantitatively evaluate botanical shapes have been suggested. The most common methodology is based on elliptic Fourier descriptors (Costa *et al.*, 2011; Sun *et al.*, 2012; Chitwood & Otani, 2017), which was successfully used on leaves (Jensen *et al.*, 2002; Neto *et al.*, 2006; Wu *et al.*, 2007; Chitwood *et al.*, 2014; Kadir, 2015; López-Santos & Page, 2018), leaflets (Furuta *et al.*, 1995; Olsson *et al.*, 2000), kernels (Iwata *et al.*, 2015), roots (Iwata *et al.*, 1998), flowers (Yoshioka *et al.*, 2004), and fruit (Currie *et al.*, 2000; Goto *et al.*, 2005; Antonucci *et al.*, 2012). This method mathematically describes the entire shape of an object by transforming the contour into Fourier coefficients.

For efficient management and effective utilization, germplasm collection can be studied through morphological, biochemical and/or genetic methods (Berthaud, 1997; Badenes *et al.*, 1998; Asma & Ozturk, 2005; Scheldeman *et al.*, 2006; Bassil *et al.*, 2009; Hofer *et al.*, 2014; Bacchetta *et al.*, 2015; Gharaghani *et al.*, 2017). In crops species as well as in pistachio, morphological characterization is time-consuming because a lot of traits must to be registered. To reduce time of analysis, in this study the diversity in pistachio germplasm collection maintained at National Fruit Tree Germplasm Collection based on morphological leaf parameters was analyzed. The aim was to adopt a rapid and quantitative non-hierarchic clustering classification (k-means), to extract size classes basing on the contemporary combination of different morphological traits (*i.e.*, leaf stalk length, terminal leaf length, terminal leaf width, and terminal leaf ratio) of a varietal collection composed by 21 cultivars of pistachio.

Material and methods

Data collection

Leaves morphological data were collected from 21 pistachio cultivars (Table 1), provided by the "pistachio germplasm collection" maintained at the National Fruit Tree Germplasm Collection of Consiglio per la Ricerca in Agricoltura e l'Analisi dell'Economia Agraria (CREA) – Centro di Ricerca Olivicoltura, Frutticoltura e Agrumicoltura (Central Italy, lat. 41.8000° N, 12.5690° E, alt. 86 m a.s.l.), according to the IPGRI (1997) protocol.

In particular, leaf stalk length, terminal leaf length, terminal leaf width and terminal leaf ratio (Fig. 1) were measured with a digital caliper, on 10 fully developed leaves, from the middle third of current season shoots, on about three to five samples for each cultivar for two years (2018 and 2019).

Table 1. Leaves data (*i.e.*, cultivars, sex, origin) and mean ± standard deviations of the morphological traits (*i.e.*, stalk length, leaf length, width and ratio) collected from 21 pistachio different cultivars, provided by the "pistachio germplasm collection" maintained at the National Fruit Tree Germplasm Collection of Consiglio per la Ricerca in Agricoltura e l'Analisi dell'Economia Agraria (CREA) – Centro di Ricerca Olivicoltura, Frutticoltura e Agrumicoltura (Central Italy, lat. 41.8000° N, 12.5690° E, alt. 86 m a.s.l.), following the International Plant Genetic Resources Institute (IPGRI, 1997) protocol.

Cultivar	Sex	Origin	Stalk length(mm)	Leaf length (mm)	Leaf width (mm)	Leaf length/width ratio
40 A	Male	Unknown	31.44±8.62	38.19±12.02	16.83±2.90	2.24±0.46
502	Male	Unknown	47.01±11.36	103.50±25.09	71.01±15.63	1.46±0.13
Aegina	Female	Greece	48.78±9.46	105.23±16.16	70.03±11.54	1.51±0.14
ASK	Male	Unknown	41.55±11.05	88.44±20.25	64.73±10.89	1.36±0.21
Baglio	Female	Italy	33.67±8.92	76.81±9.25	57.11±5.39	1.35±0.12
Bianca	Female	Italy	35.63±5.99	83.10±9.71	60.00±7.20	1.39±0.16
Bronte	Female	Italy	39.90±10.39	94.34±8.91	66.40±8.25	1.43±0.14
Cerasuola	Female	Italy	48.44±12.36	106.91±20.32	71.76±8.81	1.49±0.23
Chico	Male	USA	50.49±10.44	62.66±9.62	37.79±5.26	1.68±0.29
Greco	Male	Greece	56.97±9.64	98.00±18.02	62.03±8.44	1.58±0.25
Insolia	Female	Italy	39.34±8.16	93.07±13.75	66.52±8.87	1.42±0.24
Iraq	Female	Iraq	53.77±9.18	90.53±7.88	55.80±6.13	1.63±0.16
Kerman	Female	Iran	55.03±12.97	129.49±13.42	88.68±10.03	1.47±0.15
Larnaka	Female	Cyprus	46.66±11.66	115.23±31.38	74.99±16.40	1.54±0.23
Mateur	Female	Tunisia	36.98±9.60	66.91±32.75	43.44±28.83	1.84±0.54
Napoletana	Female	Italy	36.38±8.31	85.50±10.12	65.64±7.59	1.32±0.21
Naz	Male	Unknown	37.40±10.46	96.16±11.21	61.72±7.58	1.57±0.15
Rashti	Female	Israel	45.11±7.52	79.77±14.14	60.78±8.13	1.32±0.21
Red Aleppo	Female	Syria	46.53±6.77	99.40±9.30	69.86±9.92	1.44±0.16
Sfax	Female	Tunisia	44.50±9.43	107.64±11.66	59.74±11.59	1.83±0.23
Tignusa	Female	Italy	44.62±9.37	94.97±10.08	64.21±6.90	1.49±0.16

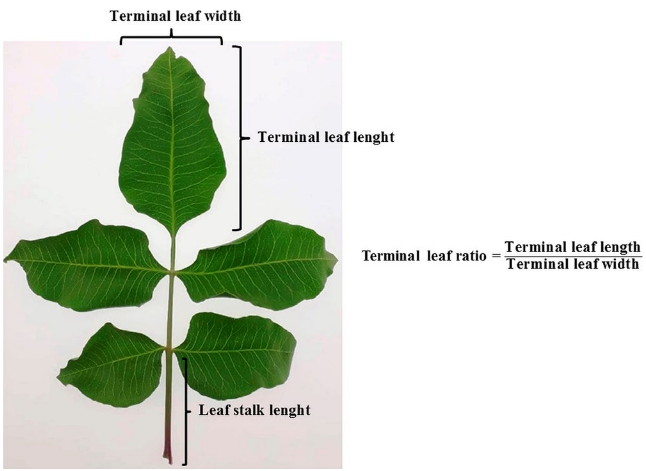

Figure 1. Representation of the measured morphological characteristics of pistachio leaf. Modified from the draft protocol of the International Union for the Protection of New Varieties of Plants (UPOV, 2020).

Cluster analysis

In this study it was adopted the unsupervised non-hierarchic clustering technique (k-means) implemented in the study of Antonucci *et al.* (2012). Generally, in k-means technique the clusters are represented by centers of mass of their members. The algorithm assigns cluster membership for each data vector to the nearest cluster center. Then, it computes the center of each cluster as the centroid of its member data vectors as equivalent to minimize the sum of distances from each object to its cluster centroid, over all clusters (Zha *et al.*, 2002). Moreover, this algorithm moves objects between clusters until the sum cannot be decreased further and the result is a set of clusters that are as compact and well separated as possible. Using the distances of the points from their cluster center it is possible to determine whether the clusters are compact.

In particular, the intra-cluster distance is the distance between a point and its cluster center meanwhile, the inter-cluster distance, or the distance between clusters, should be as big as possible and it is calculated as the distance between cluster centers and takes the minimum of this value. Only the minimum of this value was taken and since both of these measures determine a good clustering, the ratio between these two measures was calculated and indicated as "validity":

$$\text{Validity} = \text{mean intra} / \text{mean inter} \qquad (1)$$

The clustering which gives a minimum value for the validity measure is the ideal value of k in the k-means procedure. This measurement was proposed by Ray & Turi (1999) and modified in the work of Antonucci *et al.* (2012). This procedure, which introduces an iteration (1,000 cycles) to smooth the stochastic attitude of the k-means procedure, was adopted to the entire samples of pistachio leaves from k=2 to k=15 for both four morphological variables (*i.e.*, leaf stalk length, terminal leaf length, terminal leaf width and terminal leaf ratio) and three morphological variables (*i.e.*, terminal leaf length, terminal leaf width and terminal leaf ratio). The coefficients of morphological variables for each sample were clustered (k-means) with the unsupervised k-means clustering technique, attributing the group which follows the mode value for each sample.

Statistical analysis

To visualize the groups (extracted from the cluster analysis) distribution considering both four morphological variables (*i.e.*, leaf stalk length, terminal leaf length, terminal leaf width and terminal leaf ratio) and three morphological variables (*i.e.*, terminal leaf length, terminal leaf width and terminal leaf ratio), principal component analysis (PCA) were carried out. In addition, box plots and Analysis of Variance (ANOVA) were performed to evaluate statistical significance differences among the same groups. All these analyses were performed with the software PAST (v. 2.17; Hammer *et al.*, 2001).

Classification analysis

Basing on the k-means grouping suggestion a classification model has been constructed using an artificial intelligence approach. k-means clustering suggested two different kinds of classification based on four morphological variables (*i.e.*, leaf stalk length, terminal leaf length, terminal leaf width and terminal leaf ratio) or on three morphological variables (*i.e.*, terminal leaf length, terminal leaf width and terminal leaf ratio). Once chosen the best number of variables (basing on the results of the statistical analysis of PCA, box plot and ANOVA) a classification model will be constructed using a non-linear artificial neural network (ANN) approach.

The ANNs were developed basing on an input layer (x-block) to estimate the dummy output layer (groups attribution; y-block) developing a multi-layer perceptron (MLP) structure. The MLP is a feedforward neural network, which is widely used in classification and pattern recognition problems (for the procedure see Mossalam & Arafa, 2017; Proto *et al.*, 2020). From the 400 observations, to avoid overfitting only 320 samples (80%) were used to construct the models. The remaining 80 samples (20%) were then used to test the performance of the models (internal test). The partitioning of the two datasets was optimally chosen with Euclidean distances, based on the algorithm developed by Kennard & Stone (1969), which selects objects without a priori knowledge of a regression model. The percentage of bad prediction on train and test sub-sets were reported. To better visualize the results, some numerical statistical validation measures for the test (*i.e.*, accuracy, sensitivity and specificity) and some powerful classification performances as the receiver operating curve (ROC), the area under the curve (AUC) and the f1-score were extracted. The ROC represents a technique for visualizing, organizing and selecting classifiers based on their performance (Fawcett, 2006), while the AUC represents the degree or measure of separability describing how much the model is capable of distinguishing between classes. The ROC curve was obtained by plotting the true positive rate (TPR) as a function of the false positive rate (FPR), and then the area under curve (AUC) was calculated (Guan *et al.*, 2020). In addition, the confusion matrices of both training and the test of the MLP model were reported.

Moreover, the variable impact neural network analysis was performed to assess the relative importance of each variable (Abdou *et al.*, 2012). This index is similar to the linear regression variable importance in the projection (VIP) scores (Febbi *et al.*, 2015). The ANN analysis has been performed using Palisade Neural Tools 7.6.

The variable impact Δ^k for the k-th independent variable, was calculated in the following way (Palisade Knowledge Base, 2020). The training set was considered made of m samples. Each sample in the training set is a row vector $\mathbf{x}$ with n columns. The samples were stored in a matrix $\mathbf{X}$ having m rows and n columns. As a result, a generic element of that matrix is X_m^n The output $\mathbf{y}$ is a column vector with m rows obtained by applying the operator $\boldsymbol{N}$ to the matrix $\mathbf{X}$:

$$\mathbf{y}=\boldsymbol{N}\mathbf{X} \qquad (2)$$

In particular, $\mathbf{y}=\boldsymbol{N}\mathbf{X}_k$ is a row number representing the class of the k-th element of the training set and X_k is the row vector representing the k-th row of the **X** matrix. In this study $\boldsymbol{N}$ is a nonlinear operator and can be expressed as the tensor product of several linear and nonlinear operators. It represents indeed the MLP. The first row X_1 of the matrix **X** was considered and the operator $\boldsymbol{N}$ was applied *n* times to X_1 choosing each time a different value of X_1^1 among the values of X^1, the latter being the first column of the matrix **X**. The 1-case dependent variable impact of the first variable Δ_1^1 was then defined as:

$$\Delta_1^1 = \max_{X_1^1 \in X^1} \boldsymbol{N}X_1 - \min_{X_1^1 \in X^1} \boldsymbol{N}X_1 \quad (3)$$

This procedure was repeated for all the *n* variables over all the training set. The *i*-case dependent variable impact of the *k*-variable Δ_i^k was then defined as:

$$\Delta_i^k = \max_{X_i^k \in X^k} (\boldsymbol{N}X_i) - \min_{X_i^k \in X^k} (\boldsymbol{N}X_i) \quad (4)$$

Finally, the variable impact of the *k*-th dependent variable was then obtained by averaging Δ_i^k over all the m training cases:

$$\Delta^k = \frac{\sum_{i=1}^{m} \Delta_i^k}{m} \quad (5)$$

Results

Cluster analysis

The results of the procedure of validation to find the best number of k-clusters are reported in Fig. 2. The best number of clusters to be used on this dataset was chosen above the second negative peak for both A) four morphological variables (*i.e.*, leaf stalk length, terminal leaf length, terminal leaf width and terminal leaf ratio) and for B) three morphological variables (*i.e.*, terminal leaf length, terminal leaf width and terminal leaf ratio). Considering four variables, the analysis extracted a value of k=6 (six groups), while using three variables k is equal to 5 (five groups) (Fig. 2A and B respectively).

Statistical analysis

Four morphological variables (6 groups)

Figure 3 reports the PCA performed on four morphological variables (*i.e.*, leaf stalk length, terminal leaf length, terminal leaf width and terminal leaf ratio) for the six groups extracted from the cluster analysis. The interactions among the morphological variables and the six groups were highlighted by the convex hulls. The first component (PC1), reported an explained variance equal to 54.8%, and was related with the leaf stalk length, terminal leaf length and terminal leaf width. The second component (PC2) (presenting an explained variance equal to 24.3%) was mainly related with the terminal leaf ratio. Five groups (1, 3, 4, 5 and 6) resulted partially overlapped and related to high values of leaf stalk length, terminal leaf length and terminal leaf width. Meanwhile, the group 2 positioned on the negative side of PC1 associating with high value of terminal leaf ratio and low values of leaf stalk length, terminal leaf length and terminal leaf width.

Figure 4 reports the box plots performed on the four morphological variables [*i.e.*, leaf stalk length (A), terminal leaf length (B), terminal leaf width (C) and terminal leaf ratio (D)]. It was possible to observe as only for the terminal leaf ratio the six groups resulted to be all similar.

Table 2 reports the results of ANOVA performed on the four morphological variables (*i.e.*, leaf stalk length, terminal leaf length, terminal leaf width and terminal leaf ratio) for the six groups extracted from cluster analysis.

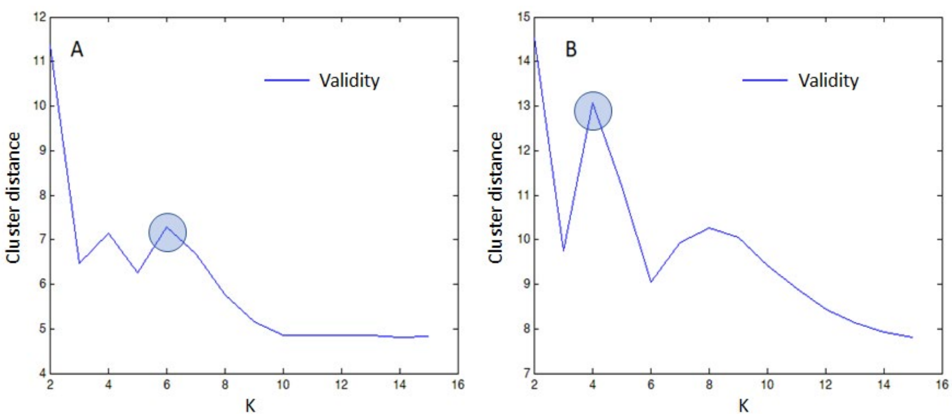

Figure 2. Results of the validation procedure to find the best number of k-clusters (highlighted by a circle) for A) four morphological variables (*i.e.*, leaf stalk length, terminal leaf length, terminal leaf width and terminal leaf ratio) and for B) three morphological variables (*i.e.*, terminal leaf length, terminal leaf width and terminal leaf ratio).

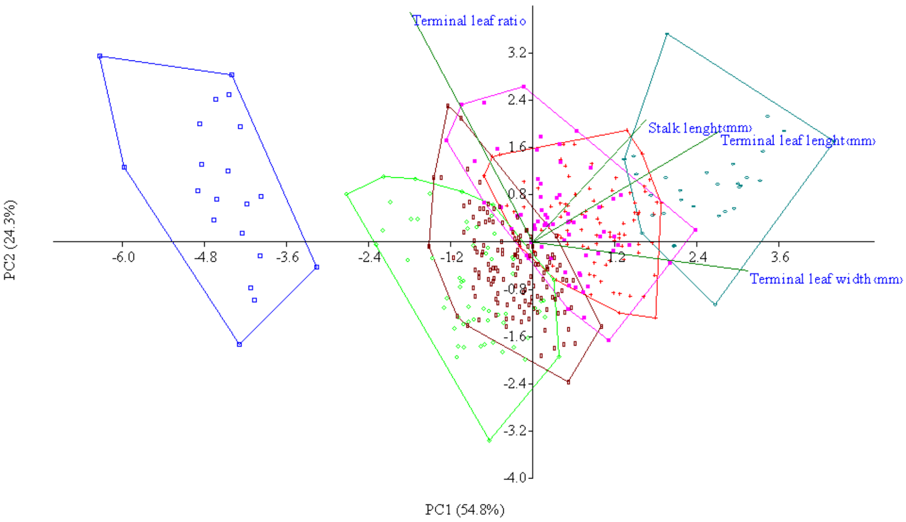

Figure 3. Principal component analysis (PCA) performed on four morphological variables (*i.e.,* leaf stalk length, terminal leaf length, terminal leaf width and terminal leaf ratio) for the 6 groups [group 1 (red), group 2 (blue), group 3 (purple), group 4 (green), group 5 (brown) and group 6 (light blue)] extracted from the cluster analysis.

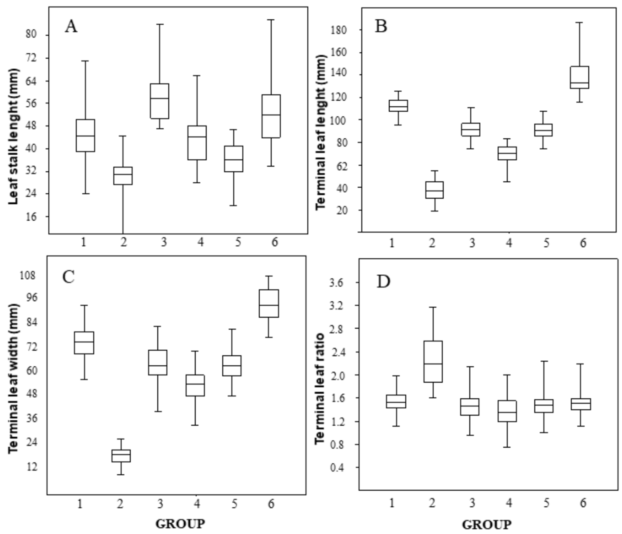

Figure 4. Box plots performed on the four morphological variables for the six groups extracted from cluster analysis.

Generally, all the six groups statistically differentiated for all the four variables except for the group 1 and 4 for the variable "leaf stalk length", for the groups 3 and 5 for the "terminal leaf length" and the "terminal leaf width". Meanwhile, resulted few statistically significant differences between groups for the variable "terminal leaf ratio".

Three morphological variables (5 groups)

Figure 5 shows the PCA performed on three morphological for the five groups (1, 2, 3, 4 and 5) extracted from the cluster analysis. The interactions among the morphological variables and the five groups were highlighted by the convex hulls. In this case, the five groups resulted well separated on the first axes PC1 (explained variance equal to 67.8%) and in particular with the trend (from low to high values of terminal leaf length and terminal leaf

Table 2. Results of analysis of variance (ANOVA; p<0.05) performed on the four morphological variables (*i.e.*, leaf stalk length, terminal leaf length, terminal leaf width and terminal leaf ratio) for the six groups extracted from cluster analysis.

	Four morphological variables (6 groups)					
	1	**2**	**3**	**4**	**5**	**6**
Leaf stalk length	45.00±9.02^{a}	30.20±6.72b	58.06±8.05^{c}	43.36±9.00^{a}	36.05±6.11^{d}	53.73±11.63^{c}
Terminal leaf length	111.81±7.44^{a}	37.05±9.81^{a}	91.21±8.80^{b}	69.81±8.85^{c}	90.82±6.63^{b}	139.58±17.86^{d}
Terminal leaf width	73.49±7.58^{a}	16.71±4.14^{a}	62.68±8.17^{b}	51.62±8.44^{c}	62.16±6.94^{b}	91.85±9.05^{d}
Terminal leaf ratio	1.54±0.18^{b}	2.25±0.43^{c}	1.48±0.25^{ab}	1.39±0.27^{a}	1.48±0.20^{ab}	1.53±0.19^{b}

Different letters in the same row denote significant differences among samples means (p < 0.05).

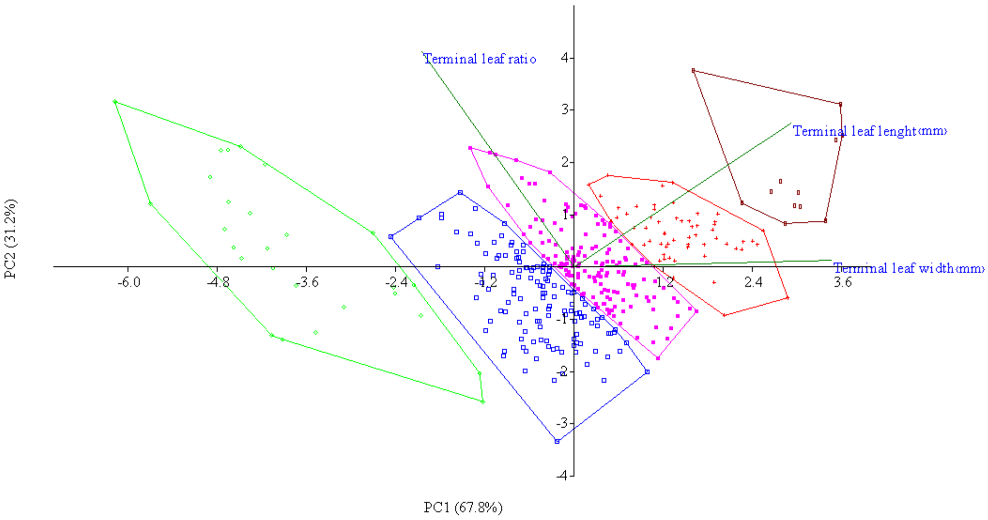

Figure 5. Principal component analysis (PCA) performed on three morphological variables (*i.e.*, terminal leaf length, terminal leaf width and terminal leaf ratio) for the 5 groups [group 1 (red), group 2 (blue), group 3 (purple), group 4 (green) and group 5 (brown)] extracted from the cluster analysis.

width and from low to high value of terminal leaf ratio) of groups 4, 2, 3, 1 and 5.

Figure 6 reports the box plots performed on the three morphological variables. Also here, as in Fig. 4, it was observed that the five groups resulted to be all similar only for the terminal leaf ratio. Table 3 shows the results of ANOVA performed on the three morphological variables for the five groups extracted from cluster analysis. Also in this case, all the five groups presented statistically significant differences for all the variables except for the "terminal leaf ratio".

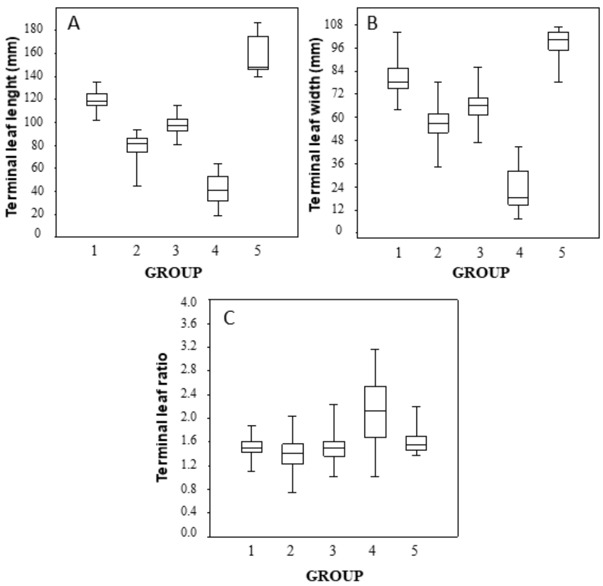

Figure 6. Box plots performed on the three morphological variables for the five groups extracted from cluster analysis.

Classification

It has been chosen to classify with the MLP model only the three morphological variables (*i.e.*, terminal leaf length, terminal leaf width and terminal leaf ratio) since from the results of the statistical analysis (*i.e.*, PCA), the groups extracted from clustering were more separated than in the four morphological variables one with a better statistically significant differences for all the variables (*i.e.*, ANOVA).

Table 4 reports these results about the performances of the MLP model (training and test) to predict the classification of the five groups. The best model resulted to be constructed with 6 nodes. The percentages of bad prediction in both training and testing resulted equal to 0% (0% incorrect). In addition, the training time was of 00:19:59.

Table 5 reports the confusion matrices for both the training and the test set of the MLP model. The accuracy (number of TPR and FPR divided by total number of cases) resulted equal to 1. To better analyze the model, 10,500 random trials were run. The training set size/total number of samples ratio varied between 0.75 and 0.85 (step of 0.5) and each time 500 trials were run, the dataset was reshuffled. For each trial the ROC curve was calculated. The curves obtained by averaging the results are reported in Figure 7 for each class. The mean AUC value for each class was also calculated and reported in Table 6 together with the relative standard deviations. Generally, the 10,500 AUC values resulted strictly < 1. However, being their distribution highly asymmetric, the sum of the standard deviation and the mean value were > 1 for four classes out of five.

Table 3. Results of analysis of variance (ANOVA; p<0.05) performed on the three morphological variables (*i.e.*, terminal leaf length, terminal leaf width and terminal leaf ratio) for the five groups extracted from cluster analysis.

	Three morphological variables (5 groups)				
	1	**2**	**3**	**4**	**5**
Terminal leaf length	119.89±7.38^{a}	79.03±8.53^{b}	97.99±7.35^{c}	41.91±12.20^{d}	156.98±16.03^{e}
Terminal leaf width	80.00±7.86^{a}	56.61±7.46^{b}	65.82±7.18^{c}	22.06±10.21^{d}	98.18±7.87^{e}
Terminal leaf ratio	1.51±0.15^{ab}	1.42±0.24^{a}	1.51±0.22^{ab}	2.07±0.51^{c}	1.61±0.22^{b}

Different letters in the same row denote significant differences among samples means ($p < 0.05$).

Table 4. Characteristics and principal results of the multi-layer perceptron (MLP) (training and internal test) to classify the five groups extracted from the cluster analysis considering the three morphological variables (*i.e*., terminal leaf length, terminal leaf width and terminal leaf ratio)

Training (80%)	
Number of cases	320
Training time	00:19:59
Number of trials	1,534,909
Bad predictions	0%
Testing (20%)	
Number of cases	80
Bad predictions	0%

Table 5. Confusion matrices for both training and internal test of the results of the multi-layer perceptron (MLP) to classify the five groups extracted from the cluster analysis considering the three morphological variables (*i.e*., terminal leaf length, terminal leaf width and terminal leaf ratio).

	1	**2**	**3**	**4**	**5**
	Confusion matrix (training)				
1	50	0	0	0	0
2	0	108	0	0	0
3	0	0	130	0	0
4	0	0	0	22	0
5	0	0	0	0	10
	Confusion matrix (testing)				
1	14	0	0	0	0
2	0	23	0	0	0
3	0	0	36	0	0
4	0	0	0	5	0
5	0	0	0	0	2

Figure 8 shows the variable impact on the MLP model underlining that the "terminal leaf length" returned the higher impact (44.89%) for the five groups classification. This variable was followed by "terminal leaf width" and the "terminal leaf ratio" (38.97% and 16.14% respectively) which also have a high impact.

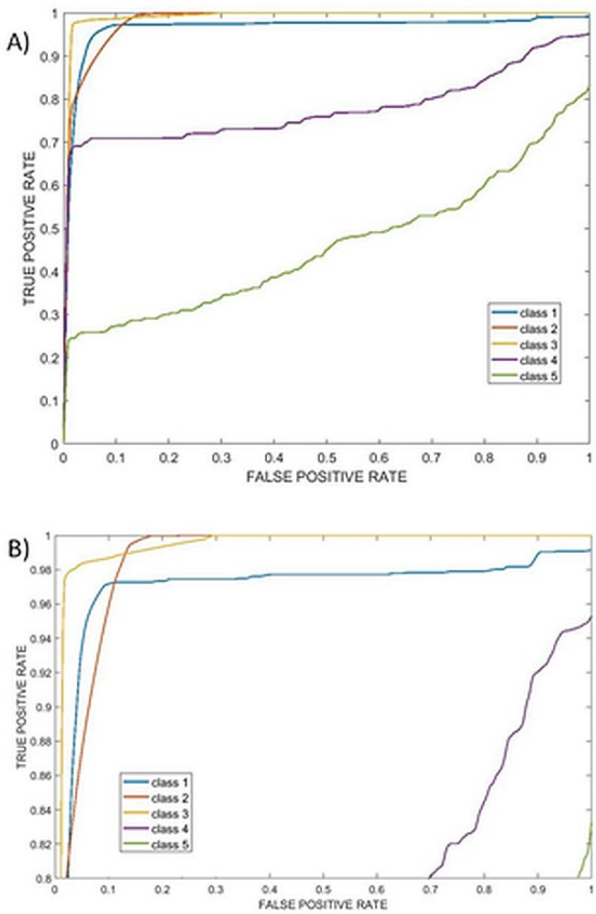

Figure 7. A) Receiver operating curves (ROC) [obtained by plotting the true positive rate (TPR) against the false positive rate (FPR)], averaged over 10500 random trials, for the five classes extracted from cluster analysis considering only the test set of the three morphological variables, after binarization (one class versus the rest of classes). B) Zoom of the ROC curves in the range of TPR [0.8:1].

Table 6. Mean area under the curve (AUC) values, and relative standard deviations, for the five classes extracted from cluster analysis considering the test set of the three morphological variables

	Class 1	Class 2	Class 3	Class 4	Class 5
Mean AUC value	0.9768	0.9812	0.9895	0.8083	0.4562
St. dev.	0.0920	0.0300	0.0285	0.3264	0.3992

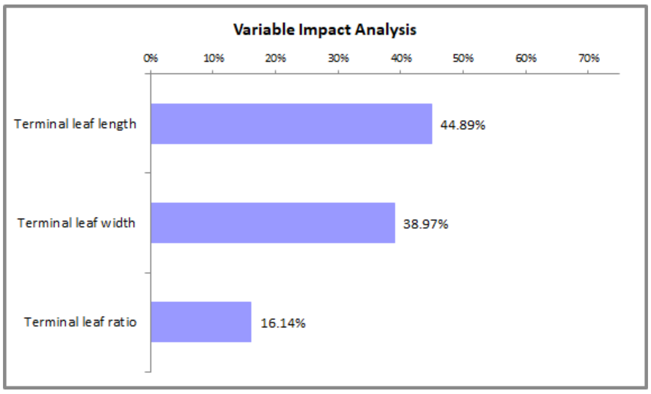

Figure 8. Variable impact analysis in the multi-layer perceptron (MLP) (training and internal test) to classify the five groups extracted from the cluster analysis considering the three morphological variables.

Discussion

Generally, the automatic classification of size classes basing on the contemporary combination of different morphological traits (*i.e.*, leaf stalk length, terminal leaf length, terminal leaf width and terminal leaf ratio) could be a valid and efficient instrument for the germplasm collection researches. As reported by Sun *et al.* (2012) shape, in terms of length, width, volume or ratios is a crucial aspect to identify species, and cultivar authenticity. The IPGRI located particular emphasis to collect, conserve and promote the utilization of the world's plant germplasm (Ayad, 1986). As reported by Chatti *et al.* (2017) for pistachio, being leaf an important growing organ, it could be used to classify different species and to discriminate among varieties. In particular leaf stalk length, terminal leaf length, width and ratio are highly discriminating quantitative characters which are continuously variable; so, they could be measured in a group of plants and recorded in scales, to assess phenotypes and discriminate among varieties.

Certification systems based on the technical protocols of the UPOV (at international level) and Community Plant Variety Office (CPVO, at European one), identify appropriate characteristics for variety description set out in visual charts (*e.g.*, specific technical protocols or technical protocols for different crops specie) and objective measurable observations against a calibrated linear scale, made on a large varietal collection.

The pistachio morphological characterization follows the IPGRI protocols, which are time consuming because a lot of traits must be separately considered. This required the collection of a high number of descriptors resulting time consuming. This aspect could be by-passed automatically establishing size classes in crops specie. As also reported in the study of Lootens *et al.* (2013), where chicory roots morphological traits were studied using elliptic Fourier descriptors, finding that it is possible to objectively use the root shape also to characterize varieties.

The importance of automatically establishing size classes of a pistachio germplasm collection basing on different morphological leaf variables (*i.e.*, leaf stalk length, terminal leaf length, terminal leaf width and terminal leaf ratio) simultaneously and not only separately, was one of the fundamental aspects of this study.

In our study we adopted a similar approach based on different morphological leaf variables (*i.e.*, leaf stalk length, terminal leaf length, terminal leaf width and terminal leaf ratio) simultaneously and not only separately. The aim was to adopt a rapid and quantitative non-hierarchic clustering classification (k-means), already experimented for the extraction of almond shape classes in the study of Antonucci *et al.* (2012), in combination with the use of an algorithm of the artificial neural network (MLP). Moreover, an automatic leaves image selective classification system was presented by Arribas *et al.* (2011) to discriminate among sunflower crops using neural networks.

In particular, when the four morphological variables (*i.e.*, leaf stalk length, terminal leaf length, terminal leaf width and terminal leaf ratio) were considered, the cluster analysis extracted six weakly differentiated groups; meanwhile considering the three morphological variables (*i.e.*, terminal leaf length, terminal leaf width and terminal leaf ratio) the analysis extracted five groups more statistically differentiated. These results indicated that these three morphological variables can be used to classify groups of leaf size classes with the MLP.

This method could be generalized for germplasm collections of different fruits and morphological traits, just taking into consideration a very important variable, *i.e.* the occurrence of leaf variations under different ecological conditions and environmental factors which could induce structural variations (Belhadj *et al.*, 2007). Quantifying the shape in its multivariate and multidimensional complexity is a very important aspect for a quality grading in the industry (Sun *et al.*, 2012; Lootens *et al.*, 2013). In addition, the selection of suitable pistachio phenotypes, basing on specific morphological traits which reflect different environmental and soil conditions and diseases, are important for increasing yield efficiency and the property of this important crop (Karimi *et al.*, 2009).

In this study, the contemporary combination of different morphological leaf traits (*i.e.*, leaf stalk length, terminal leaf length, terminal leaf width and terminal leaf ratio) recorded on the pistachio germplasm collection, allowed to create an automatic classification of size classes of great importance for cultivar identification and comparison. The proposed quantitative methodology is rapid, non-hierarchic and provides k-clusters successfully used for pistachio leaf size classification, representing a practical efficient instrument in many different research fields, such as genetics and agronomy. At k=5 (three morphological variables) the system performed better than k=6 (four morphological variables), because the five groups extracted from cluster analysis resulted well separated and presented statistically significant differences for all the variables except for the "terminal leaf ratio".

References

Abdou HA, Kuzmic A, Pointon J, Lister RJ, 2012. Determinants of capital structure in the UK retail industry: A comparison of multiple regression and generalized regression neural network. Intell Syst Account Finance Manag 19 (3): 151-169. https://doi.org/10.1002/isaf.1330

Antonucci F, Costa C, Pallottino F, Paglia G, Rimatori V, De Giorgio D, Menesatti P, 2012. Quantitative method for shape description of almond cultivars (*Prunus amygdalus* Batsch). Food Bioprocess Technol 5: 768-785. https://doi.org/10.1007/s11947-010-0389-2

Arribas JI, Sánchez-Ferrero GV, Ruiz-Ruiz G, Gómez-Gil J, 2011. Leaf classification in sunflower crops by computer vision and neural networks. Comput Electron Agric 78 (1): 9-18. https://doi.org/10.1016/j.compag.2011.05.007

Asma BM, Ozturk K, 2005. Analysis of morphological, pomological and yield characteristics of some apricot germplasm in Turkey. Genet Resour Crop Evol 52: 305-313. https://doi.org/10.1007/s10722-003-1384-5

Ayad WG, 1986. Conservation of crop germplasm: an overview of the FAO/IBPGR regional programme for South West Asia. Proc R Soc Edinb B 89: 265-271. https://doi.org/10.1017/S0269727000009088

Bacchetta L, Rovira M, Tronci C, Aramini M, Drogoudi P, Silva AP, Solar A, Avanzato D, Botta R, Valentini N, Boccacci P, 2015. A multidisciplinary approach to enhance the conservation and use of hazelnut *Corylus avellana* L. genetic resources. Genet Resour Crop Evol 62: 649-663. https://doi.org/10.1007/s10722-014-0173-7

Badenes ML, Martinez-Calvo J, Llacer G, 1998. Analysis of apricot germplasm from the European ecogeographical group. Euphytica 102: 93-99. https://doi.org/10.1023/A:1018332312570

Balduzzi M, Binder BM, Bucksch A, Chang C, Hong L, Iyer-Pascuzzi AS, Pradal C, Sparks EE, 2017. Reshaping plant biology: qualitative and quantitative descriptors for plant morphology. Front Plant Sci 8: 117. https://doi.org/10.3389/fpls.2017.00117

Bassil N, Hummer KE, Postman JD, Fazio G, Baldo A, Armas I, Williams R, 2009. Nomenclature and genetic relationships of apples and pears from Terceira Island. Genet Resour Crop Evol 56: 339-352. https://doi.org/10.1007/s10722-008-9369-z

Bayramzadeh V, Funada R, Kubo T, 2008. Relationships between vessel element anatomy and physiological as well as morphological traits of leaves in *Fagus crenata* seedlings originating from different provenances. Trees 22: 217-224. https://doi.org/10.1007/s00468-007-0178-3

Belhadj S, Derridj A, Aigouy T, Gers C, Gauquelin T, Mevy JP, 2007. Comparative morphology of leaf epidermis in eight populations of Atlas pistachio (*Pistacia atlantica* Desf., Anacardiaceae). Microsc Res Tech 70 (10): 837-846. https://doi.org/10.1002/jemt.20483

Berthaud J, 1997. Strategies for conservation of genetic resources in relation with their utilization. Euphytica 96: 1-12. https://doi.org/10.1023/A:1002922220521

Chao CT, Parfitt DE, Ferguson L, Kallsen C, Maranto J, 2003. Genetic analyses of phenological traits of pistachio (P*istacia vera* L.). Euphytica 129: 345-349. https://doi.org/10.1023/A:1022206911350

Chatti K, Choulak S, Guenni K, Salhi-Hannachi A, 2017. Genetic diversity analysis using morphological parameters in Tunisian Pistachio (*Pistacia vera* L.). Int Res J Biol Sci 2: 29-34.

Chitwood DH, Ranjan A, Martinez CC, Headland LR, Thiem T, Kumar R, *et al.*, 2014. A modern ampelogray: a genetic basis for leaf shape and venation patterning in grape. Plant Physiol 164 (1): 259-272. https://doi.org/10.1104/pp.113.229708

Chitwood DH, Otani WG, 2017. Morphometric analysis of Passiflora leaves: the relationships between landmarks of the vasculature and elliptic Fourier descriptors of the blade. GigaScience 6: 1-13. https://doi.org/10.1093/gigascience/giw008

Costa C, Antonucci F, Pallottino F, Aguzzi J, Sun DW, Menesatti P, 2011. Shape analysis of agricultural products: a review of recent research advances and potential application to computer vision. Food Bioprocess Technol 4: 673-692. https://doi.org/10.1007/s11947-011-0556-0

Crane JC, 1978. Pistachio tree nuts. Avipublishing Co., Westport, California.

Currie AJ, Ganeshanandam S, Noiton DA, Garrick D, Shelbourne CJA, Oraguzie N, 2000. Quantitative evaluation of apple (*Malus x domestica* Borkh.) fruit shape by principal component analysis and Fourier descriptors. Euphytica 111: 219-227. https://doi.org/10.1023/A:1003862525814

FAOSTAT, 2017. FAO web page. http://www.fao.org/faostat [30 March 2020].

Fares K, Guasmi F, Touil L, Triki T, Ferchichi A, 2009. Genetic diversity of pistachio tree using inter-simple sequence repeat markers ISSR supported by morphological and chemical markers. Biotechnol 8: 24-34. https://doi.org/10.3923/biotech.2009.24.34

Fawcett T, 2006. An introduction to ROC analysis. Pattern Recognition Letters 27 (8): 861-874. https://doi.org/10.1016/j.patrec.2005.10.010

Febbi P, Menesatti P, Costa C, Pari L, Cecchini M, 2015. Automated determination of poplar chip size distribution based on combined image and multivariate analyses. Biomass Bioenergy 73: 1-10. https://doi.org/10.1016/j.biombioe.2014.12.001

Furuta N, Ninomiya S, Takahashi S, Ohmor H, Ukai Y, 1995. Quantitative evaluation of soybean (*Glycine max* L., Merr.) leaflet shape by principal component scores based on elliptic Fourier descriptor. Breed Sci 45 (3): 315-320. https://doi.org/10.1270/jsbbs1951.45.315

Gharaghani A, Solhjoo S, Oraguzie N, 2017. A review of genetic resources of almonds and stone fruits (*Prunus* spp.) in Iran. Genet Resour Crop Evol 64: 611-640. https://doi.org/10.1007/s10722-016-0485-x

Goto S, Iwata H, Shibano S, Ohya K, Suzuki A, Ogawa H, 2005. Fruit shape variation in *Fraxinus mandshurica* var. *japonica* characterized using elliptic Fourier descriptors and the effect on flight duration. Ecol Res 20: 733-738. https://doi.org/10.1007/s11284-005-0090-5

Guan X, Zhang L, Huang S, Peng Z, 2020. Infrared small target detection via non-convex tensor rank surrogate joint local contrast energy. Remote Sens 12(9): 1520. https://doi.org/10.3390/rs12091520

Hammer Ø, Harper DAT, Ryan PD, 2001. Past: paleontological statistics software package for education and data analysis. Palaeontol Electron 4: 1-9.

Hassoon IM, Kassir SA, Altaie SM, 2018. A review of plant species identification techniques. Int Jour Sci Res 7 (8): 2319-7064.

Höfer M, Eldin Ali MAMS, Sellmann J, Peil A, 2014. Phenotypic evaluation and characterization of a collection of Malus species. Genet Resour Crop Evol 61: 943-964. https://doi.org/10.1007/s10722-014-0088-3

Höfer M, Peil A, 2015. Phenotypic and genotypic characterization in the collection of sour and duke cherries (*Prunus cerasus* and x *P.* x *gondouini*) of the Fruit Genebank in Dresden-Pillnitz, Germany. Genet Resour Crop Evol 62 (4): 551-566. https://doi.org/10.1007/s10722-014-0180-8

Hormaza JI, Pinney K, Polito VS, 1998. Genetic diversity of pistachio (*Pistacia vera*, Anacardiaceae) germplasm based on randomly amplified polymorphic DNA (RAPD) markers. Econ Bot 52: 78-87. https://doi.org/10.1007/BF02861298

IPGRI, 1997. Descriptors for pistachio (*Pistacia vera* L.). International Plant Genetic Resources Institute. Rome, Italy.

Iwata H, Niikura S, Matsuura S, Takano Y, Ukai Y, 1998. Evaluation of variation of root shape of Japanese radish (*Raphanus sativus* L.) based on image analysis using elliptic Fourier descriptors. Euphytica 102: 143-149. https://doi.org/10.1023/A:1018392531226

Iwata H, Ebana K, Uga Y, Hayashi T, 2015. Genomic prediction of biological shape: elliptic Fourier analysis and kernel partial least squares (PLS) regression applied to grain shape prediction in rice (*Oryza sativa* L.). PLoS One 10 (3): e0120610. https://doi.org/10.1371/journal.pone.0120610

Jensen RJ, Ciofani KM, Miramontes LC, 2002. Lines, outlines, and landmarks: morphometric analyses of leaves of *Acer rubrum, Acer saccharinum* (Aceraceae) and their hybrid. Taxon 51: 475-492. https://doi.org/10.2307/1555066

Kadir A, 2015. Leaf identification using Fourier descriptors and other shape features. Gate to Computer Vision and Pattern Recognition 1: 3-7. https://doi.org/10.15579/gtcvpr.0101.003007

Kafkas S, Kafkas E, Perl-Treves R, 2002. Morphological diversity and a germplasm survey of three wild Pistacia species in Turkey. Genet Resour Crop Evol 49 (3): 261-270. https://doi.org/10.1023/A:1015563412096

Karimi HR, Zamani Z, Ebadi A, Fatahi MR, 2009. Morphological diversity of Pistacia species in Iran. Genet Resour Crop Evol 56 (4): 561-571. https://doi.org/10.1007/s10722-008-9386-y

Kashaninejad M, Tabil LG, 2011. Pistachio (*Pistacia vera* L.). In: Postharvest biology and technology of tropical and subtropical fruits, vol 4; Yahia EM (ed). Woodhead Publ Ltd, Cambridge, UK (pp. 218-246). https://doi.org/10.1533/9780857092618.218

Kennard RW, Stone LA, 1969. Computer aided design of experiments. Technometrics 11: 137-148. https://doi.org/10.1080/00401706.1969.10490666

Lin TS, Crane JC, Ryugo K, Polito VS, Dejong TM, 1984. Comparative study of leaf morphology, photosynthesis and leaf conductance in selected *Pistacia* species. J Am Soc Hortic Sci 109: 325-330.

Lootens L, Chaves B, Bert J, Pannecoucque J, Van Waes J, Roldan-Ruiz I, 2013. Comparison of image analysis and direct measurement of UPOV taxonomic characteristics for variety discrimination as determined over five growing season, using industrial chicory as a model top. Euphytica 189: 329-341. https://doi.org/10.1007/s10681-012-0750-9

López-Santos A, Page T, 2018. Elliptic Fourier descriptors of leaf outlines: a tool to discriminate among *Aquilaria* species (Thymelaeaceae). Silvae Genetica 67: 89-92. https://doi.org/10.2478/sg-2018-0012

Massimo L, Laura D, Ginevra LB, 2020. Phytosterols and phytosterol oxides in Bronte's pistachio (*Pistacia vera* L.) and in processed pistachio products. Eur Food Res Technol 246 (2): 307-314. https://doi.org/10.1007/s00217-019-03343-8

Mossalam A, Arafa M, 2017. Using artificial neural networks (ANN) in projects monitoring dashboards' formulation. HBRC J 14 (3): 385-392. https://doi.org/10.1016/j.hbrcj.2017.11.002

Neto JC, Meyer GE, Jones DD, Samal AK, 2006. Plant species identification using Elliptic Fourier leaf shape analysis. Comput Electron Agr 50 (2): 121-134. https://doi.org/10.1016/j.compag.2005.09.004

Olsson Å, Nybom H, Prentice HC, 2000. Relationships between Nordic dogroses (Rosa L. sect. Caninae, Rosaceae) assessed by RAPDs and elliptic Fourier analysis of leaflet shape. Syst Bot 25 (3): 511-521. https://doi.org/10.2307/2666693

Palisade Knowledge Base, 2020. 15.36. Calculation and use of variable impacts. https://kb.palisade.com/index.php?pg=kb.page&id=225 [29 June 2020].

Pazouki L, Mardi M, Shanjani PS, Hagidimitriou M, Pirseyedi SM, Naghavi MR, Avanzato D, Vendramin E, Kafkas S, Ghareyazie B, *et al.*, 2010. Genetic diversity and relationships among *Pistacia* species and cultivars. Conserv Genet 11: 311-318. https://doi.org/10.1007/s10592-009-9812-5

Proto AR, Sperandio G, Costa C, Maesano M, Antonucci F, Macrì G, Scarascia Mugnozza G, Zimbalatti G, 2020. A three-step neural network artificial intelligence modelling approach for time, productivity and costs prediction: a case study in Italian forestry. Croat J For Eng 41: 35-47. https://doi.org/10.5552/crojfe.2020.611

Ray S, Turi RH, 1999. Determination of number of clusters in k-means clustering and application in colour image segmentation. Proc 4th Int Conf on Advances in Pattern Recognition and Digital Techniques (ICAPRDT'99), Calcutta, India, pp: 137-143.

Sabzi S, Pourdarbani R, Arribas JI, 2020. A computer vision system for the automatic classification of five varieties of tree leaf images. Computers 9 (1): 6. https://doi.org/10.3390/computers9010006

Scheldeman X, Van Damme P, Romero Motoche J, Urena Alvarez JV, 2006. Germplasm collection and fruit characterization of cherimoya (*Annona* cherimola) in Loja province, Ecuador, an important centre of biodiversity. Belg J Bot 139 (1): 27-38.

Sheikhi A, Arab MM, Brown PJ, Ferguson L, Akbari M, 2019. Pistachio (*Pistacia* spp.) breeding. In: Advances in plant breeding strategies: nut and beverage crops; Al-Khayri J, Jain S, Johnson D (eds). Springer, Cham, pp. 353-400. https://doi.org/10.1007/978-3-030-23112-5_10

Sun DW, Costa C, Menesatti P, 2012. Advantages of using quantitative shape descriptors in protocols for plant cultivar and postharvest product quality assessment. Food Bioprocess Technol 5: 1-2. https://doi.org/10.1007/s11947-011-0715-3

UPOV, 2020. International Union for the Protection of New Varieties of Plants. https://www.upov.int/edocs/mdocs/upov/en/twf_50/tg_pista_proj_3.pdf [30 March 2020].

Wu SG, Bao FS, Xu EY, Wang YX, Chang YF, Xiang QL, 2007. A leaf recognition algorithm for plant classification using probabilistic neural network. 2007 IEEE Int Symp on Signal Processing and Information Technology, pp: 11-16. https://doi.org/10.1109/ISSPIT.2007.4458016

Yoshioka Y, Iwata H, Ohsawa R, Ninomiya S, 2004. Analysis of petal shape variation of *Primula sieboldii* by elliptic Fourier descriptors and principal component analysis. Ann Bot 94 (5): 657-664. https://doi.org/10.1093/aob/mch190

Zha H, He X, Ding C, Gu M, Simon HD, 2002. Spectral relaxation for k-means clustering. Neural Inform Process Syst 14: 1057-1064.