



Università degli Studi di Cagliari

**DOTTORATO DI RICERCA
BOTANICA AMBIENTALE ED APPLICATA**

SCUOLA DI DOTTORATO
INGEGNERIA E SCIENZE PER L'AMBIENTE E IL TERRITORIO

XXVI ciclo

BIO/03

**New shape and texture descriptors
for an improved germplasm characterization of the
most representative Mediterranean vascular flora,
by image analysis technology**

Presentata da

Marisol Lo Bianco

Coordinatore Dottorato

Prof. Gianluigi Bacchetta

Tutor

Prof. Gianluigi Bacchetta

Dott. Gianfranco Venora

Co-Tutor

Dott. Oscar Grillo

Esame finale anno accademico 2013 – 2014

To my son

My child, when in a few years you will find this text in the home library, you will read about biodiversity, the need to preserve it and a method, the Image Analysis, as a tool for the correct classification of different seeds.

But, mind you, do not believe you can use the same tool, the image, to discriminate among men; do not stop the appearance but look inside, look at their soul.

*With love
Mom*

Contents

SUMMARY

Biodiversity concepts	11
The Mediterranean: a biodiversity hotspot under threat	18
The ex situ conservation as biodiversity preservation strategy	22
State of the art and innovative aspect of the project	25
Image measurements	27
Project aim and objectives	30
References	34

PART I - THE IMAGE ANALYSIS TECHNIQUE AND STATISTICAL TREATMENT OF DATA	39
--	----

CHAPTER 1 - COMPUTER VISION: FUNDAMENTALS AND IMAGE PROCESSING

Introduction	41
Image acquisition systems	43
Digital images	45
Image processing	49
<i>Noise, contrast and shading correction</i>	49
<i>Color calibration</i>	53
<i>Mathematical morphology</i>	54
References	63

CHAPTER 2 - A NEW SET OF SEED FEATURES: ELLIPTIC
FOURIER DESCRIPTORS AND HARALICK'S PARAMETERS

Introduction	70
Shape measurements: Elliptic Fourier Descriptors (EFDs)	70
Mathematical basis of EFDs method	73
Texture evaluation: Haralick's descriptors	77
References	82

CHAPTER 3 - MATERIALS AND METHODS

Introduction	87
The 'Macro'	89
The germplasm samples	91
Statistics	92
<i>Linear Discriminat Analysis</i>	93
<i>The cross validation</i>	101
References	104

PART II - CASE STUDIES

CHAPTER 1 - INTER AND INTRA-SPECIFIC DIVERSITY IN
CISTUS L. (CISTACEAE) SEEDS, ANALYSED BY COMPUTER
VISION TECHNIQUES

Abstract	110
Introduction	111
Materials and Methods	114
<i>Seed-lot details</i>	114

<i>Image analysis system</i>	117
<i>Statistical analysis</i>	122
Results and Discussion	124
Conclusions	131
Acknowledgements	132
References	133

CHAPTER 2 - SEED IMAGE ANALYSIS PROVIDES EVIDENCE
OF TAXONOMICAL DIFFERENTIATION WITHIN THE
MEDICAGO L. SECT. DENDROTELIS (FABACEAE)

Abstract	141
Introduction	142
Materials and Methods	145
<i>Seed collection and image acquisition</i>	145
<i>Statistical analysis</i>	151
Results	152
Discussion	158
Acknowledgements	161
References	162

CHAPTER 3 – MORPHO-COLORIMETRIC CHARACTERIZATION
OF *MALVA ALLIANCE TAXA* BY SEED IMAGE ANALYSIS

Abstract	170
Introduction	171
Materials and Methods	174
<i>Lavatera and Malva seed lots</i>	174
<i>Image analysis system</i>	179

<i>Statistical analysis</i>	184
Results	186
Discussion	192
Acknowledgements	195
References	196
CONCLUSIONS	204
ACKNOWLEDGMENTS	208

Biodiversity concepts

The variety of life on Earth, its biological diversity and the natural patterns it forms, is commonly referred to *biodiversity*. The number of species of plants, animals, and microorganisms, the enormous diversity of genes in these species, the different ecosystems on the planet, such as deserts, rainforests and coral reefs are all part of a biologically diverse Earth.

The biodiversity we see today is the fruit of billions of years of evolution and natural selection, shaped by natural processes and, increasingly, by the influence of humans. It assures the ecosystems aptitude to adapt to environmental changes, guaranteeing ecological balance and future life. It forms the web of life of which we are an integral part and upon which we so fully depend.

In recent years, biotic and abiotic factors have put into serious difficulties this natural aptitude of macro and micro-ecosystems, undermining the ecological balance throughout the world. As a result, the loss of biological diversity is constantly increasing and the extinction of threatened species is the gravest aspect of this crisis. Particular species can have strong effects on ecosystem processes by directly mediating energy and material fluxes or by altering abiotic conditions that regulate the rates of these processes (Fig. 1). So, the alteration of the species, together with the disturbance regime, and the climate can sensibly affect the ecosystem processes (Chapin *et al.*, 2000).

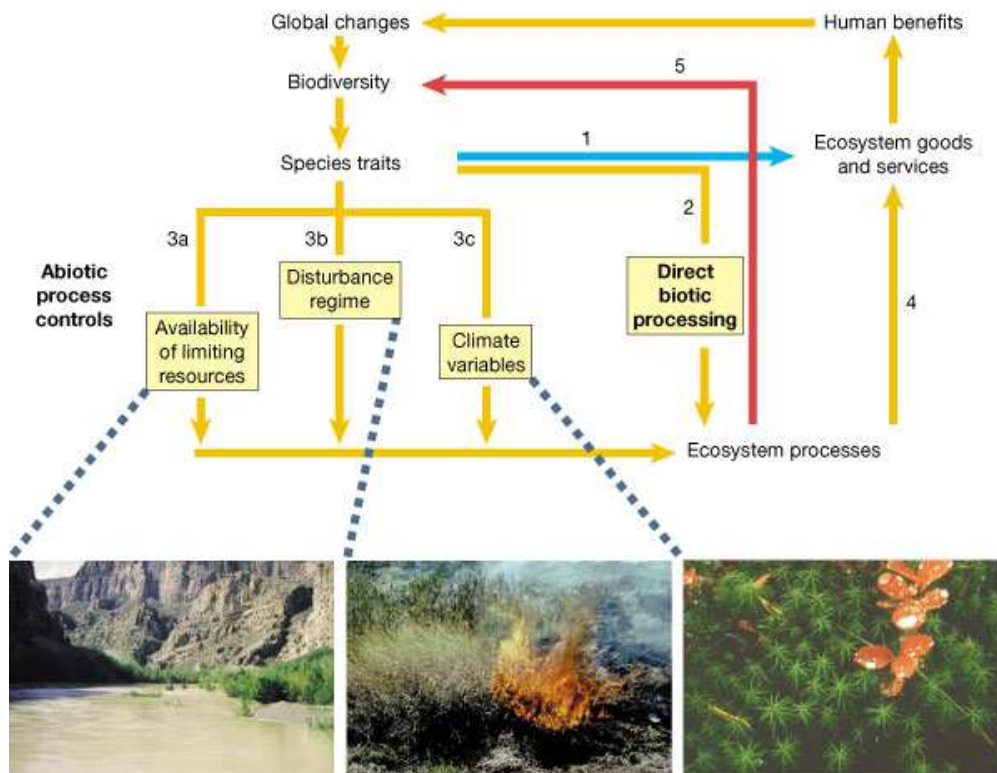


Figure 1. Mechanisms by which species traits affect ecosystem processes.

Changes in biodiversity alter the functional traits of species in an ecosystem in ways that directly influence ecosystem goods and services (1) either positively (for example, increased agricultural or forestry production) or negatively (for example, loss of harvestable species or species with strong aesthetic/cultural value). Changes in species traits affect ecosystem processes directly through changes in biotic controls (2) and indirectly through changes in abiotic controls, such as availability of limiting resources (3a), disturbance regime (3b), or micro- or macroclimate variables (3c). Illustrations of these effects include: reduction in river flow due to invasion of deep-rooted desert trees (3a); increased fire frequency resulting from grass invasion that destroys native trees and shrubs in Hawaii (3b); and insulation of soils by mosses in arctic tundra, contributing to conditions that allow for permafrost (3c). Altered processes can then influence the availability of ecosystem goods and services directly (4) or indirectly by further altering biodiversity (5), resulting in loss of useful species or increases in noxious species.

http://www.nature.com/nature/journal/v405/n6783/fig_tab/405234a0_F4.html

Almost all cultures have in some way or form recognized the importance that nature and its biological diversity has had upon them and the need to maintain it; therefore, appropriate conservation and sustainable development strategies attempt to preserve the declining biodiversity. Biodiversity boosts ecosystem productivity where each species, no matter

how small, all have an important role to play. For example, a larger number of plant species means a greater variety of crops, greater species diversity ensures natural sustainability for all life forms and healthy ecosystems can better withstand and recover from a variety of disasters.

A healthy biodiversity provides a number of natural services for everyone:

- ✓ Ecosystem services, such as
 - Protection of water resources
 - Soils formation and protection
 - Nutrient storage and recycling
 - Pollution breakdown and absorption
 - Contribution to climate stability
 - Maintenance of ecosystems
 - Recovery from unpredictable events
- ✓ Biological resources, such as
 - Food
 - Medicinal resources and pharmaceutical drugs
 - Breeding stocks, population reservoirs
 - Future resources
 - Diversity in genes, species and ecosystems
- ✓ Social benefits, such as
 - Research, education and monitoring
 - Recreation and tourism
 - Cultural values

The cost of replacing these (if possible) would be extremely expensive. It therefore makes economic and development sense to move towards sustainability. A report from *Nature* magazine also explains that genetic diversity helps to prevent the chances of extinction in the wild (Chapin *et al.*, 2000; Tilman, 2000).

To avoid the well known and well documented problems of genetic defects caused by in-breeding, species need a variety of genes to ensure successful survival. Without this, the chances of extinction increases. And as we start destroying, reducing and isolating habitats, the chances for interaction from species with a large gene pool decreases.

While there might be “survival of the fittest” within a given species, each species depends on the services provided by other species to ensure survival. It is a type of cooperation based on mutual survival and is often what a “balanced ecosystem” refers to.

Despite knowing about biodiversity’s importance for a long time, human activity has been causing massive extinctions, and the consequence is a loss of both α diversity (number of species coexisting within a uniform habitat) and β diversity (species turnover rate in function of changing habitats) (Cody, 1986). As the Environment New Service, reported back in August 1999: *“The current extinction rate is now approaching 1,000 times the background rate and may climb to 10,000 times the background rate during the next century, if present trends continue resulting in a loss that would easily equal those of past extinctions”*.

In different parts of the world, species face different levels and types of threats. But overall patterns show a downward trend in most cases. As explained in the United Nations’ 3rd Global Biodiversity Outlook, the rate of biodiversity loss has not been reduced because the 5 principle pressures on biodiversity are persistent, even intensifying:

- I. Habitat loss and degradation
- II. Climate change
- III. Excessive nutrient load and other forms of pollution
- IV. Over-exploitation and unsustainable use
- V. Invasive alien species

Most governments report to the UN Convention on Biological Diversity that these pressures are affecting biodiversity in their country. The International Union for the Conservation of Nature (IUCN) maintains the *Red List* to assess the conservation status of species, subspecies, varieties, and even selected subpopulations on a global scale (Fig. 2).

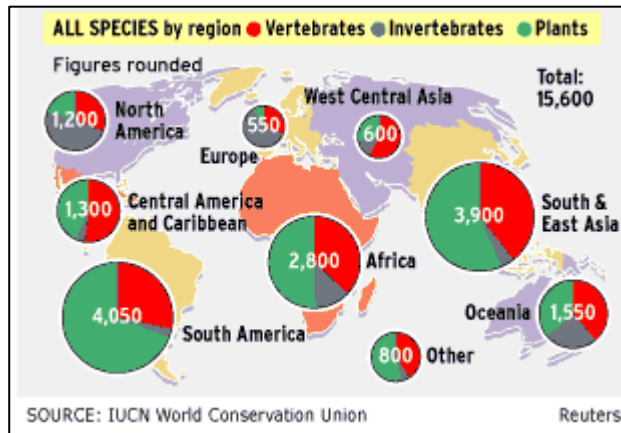


Figure 2. Geographical distribution of the 15,600 species (about 7,270 animal species and 8,330 plant and lichen species) considered at risk of extinction according to the IUCN *Red List* compilation.

A growing attention is given to the conservation of plant biodiversity in outside of the natural environment, both for the species of agronomic interest, and spontaneous flora, in compliance with the obligations under the *Convention on Biological Diversity* (CBD) (Rio de Janeiro, 1992). This was the first global agreement for the conservation and sustainable use of biodiversity on a global level, signed by 192 countries including Italy and the European Union, representing a milestone in international law. For the first time the conservation of biological diversity was recognized as “*common need of mankind*” and an integral part of development.

The Convention has established three main goals: the conservation of biological diversity, the sustainable use of its components, and the fair and

equitable sharing of the benefits from the use of genetic resources. Moreover the article 9 of CBD, “Conservation *ex situ*”, has introduced the *in situ* and *ex situ* conservation concepts, defining the principles to plan strategies and to guarantee the conservation. It indicates a series of measures to be taken to the recovery, restoration and reintroduction of the endangered species, by means of *ex situ* conservation, in addition to the conservation strategies *in situ*.

Besides to preserve existing genetic resources, the conservation allows the study and the development of new cultivars during genetic improvement processes, it provides the populations for reintroduction and repopulation programs of degraded habitats, and then it permits industry, agriculture and scientific research to use essential for progress. Finally, the *ex situ* conservation allows to study the best strategies to apply at the *in situ* conservation of threatened species (Bacchetta, 2011a).

The conservation *in situ* (areas of origin) and the *on farm* (in the areas of cultivation) are, obviously, a priority, but the *ex situ* management is essential in those cases, and there are many, in which the first two, for different reasons, are difficult to achieve. Currently, in fact, the multiple pressures that act on habitat may in some cases threaten the survival of one or more species or the integrity and function of entire ecosystems, making difficult the implementation of the *in situ* conservation strategies. In these cases, only the techniques *ex situ* can guarantee the preservation of genetic variability of germplasm (seeds, pollen, plant parts, spores, etc.) and then the regeneration, reproduction and/or multiplication of the species to be preserved. The conservation *ex situ* also plays an indispensable role in the research and genetic improvement because it promotes the sustainable use of germplasm available.

After the CBD of 1992, several government organizations deal with biodiversity conservation issues. The *Fourth Assessment Report (AR4)* of the

United Nations Intergovernmental Panel on Climate Change (IPCC) (2007) indicated the conservation *ex situ* as one of the main measures of the ecosystems to adapt to climate change in course.

Furthermore, in 2006, 11 Centres of the Consultative Group on International Agricultural Research (CGIAR) and other international collections place their *ex situ* genebank collections under the *International Treaty on Plant Genetic Resources for Food and Agriculture* of the FAO Constitution. Different conventions or agreements were approved; among these, the International Treaty provides in its “Article 15” that the Contracting Parties:

(i) recognize the importance of the *ex situ* collections of plant genetic resources for food and agriculture;

(ii) call upon the International Agricultural Research Centres to sign agreements with the Governing Body of the Treaty with regard to *ex situ* collections.

In 2013, the Commission endorses the updated *Genebank Standards for Plant Genetic Resources for Food and Agriculture*, which provide an overview of the current state of *ex situ* conservation practices, including field genebank management procedures, cryopreservation of germplasm and in vitro practices, as well as the conservation of orthodox seeds.

Finally, the *ex situ* conservation, as an unavoidable system to preserve biodiversity is possible thanks to the activities of structures more and more widespread such as the germplasm banks, gene bank collections, botanical gardens, etc., whose function is not only to preserve threatened species, but also to store, by long-term techniques, seeds, spore, woods, tissues and any other structures that make up the genetic biodiversity of the planet.

The Mediterranean: a biodiversity hotspot under threat

The Conservation International (CI) organization, adopting the Myers' hotspots concept (Myers *et al.*, 2000) as its central strategy, reassessed it introducing quantitative thresholds for the designation of biodiversity hotspots. So, to be qualified as a hotspot, a region must meet two strict criteria: it must contain at least 1500 species of vascular plants (>0.5 percent of the world's total) as endemics, and it has to have lost at least 70% of its original habitat.

The Mediterranean Basin is one of the world's richest places in terms of plant diversity – about 25,000 species are native to the region, and more than half of these are endemic – in other words, they are found nowhere else on earth. This has led to the Mediterranean being recognized as one of the first 34 *Global Biodiversity Hotspots* (Fig. 3).

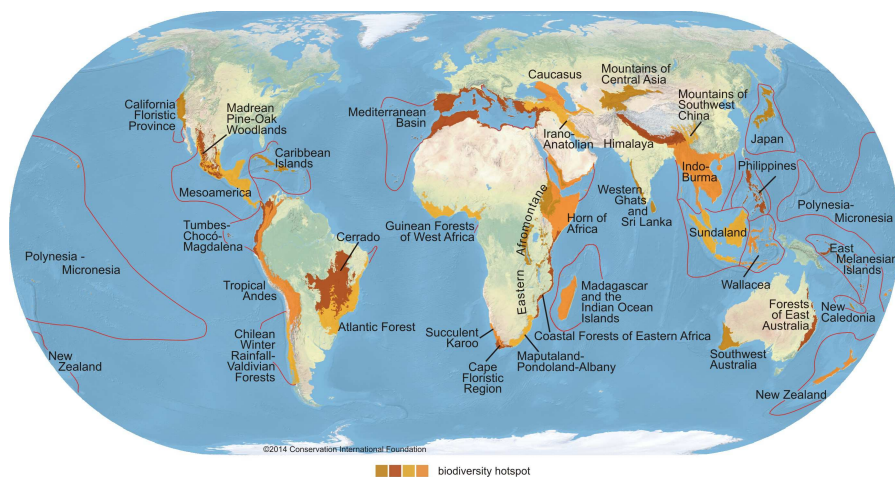


Figure 3. Global biodiversity hotspots map.

The Mediterranean Basin, with its lofty mountains, ancient rivers, deserts, forests, is a mosaic of natural and cultural landscapes, where human civilization and wild nature have coexisted for centuries. The unique conjunction of geography, history, and climate has led to a remarkable

evolutionary radiation that continues to the present day, as animals and plants have adapted to the myriad opportunities for life that the region presents.

The location of basin at the intersection of two major landmasses, Eurasia and Africa, has contributed to its high diversity and spectacular scenery. In particular, in the western basin, plant endemism is very high, due mainly to the age of the geological platform. The northern and southern coasts of Mediterranean basin, present two different situations because of the different human influence (Barbero *et al.*, 1990). In the northern part, the collapse of the agro-sylvo-pastoral system of the past centuries has led to major changes in plant community structure and the extension of woodlands dominated by competitive species. On the other hand, the southern part of the Mediterranean basin (in particular North Africa) has been subjected to the severe effects of constant increases in population and livestock, which have completely destroyed the soils and caused severe erosion and poor regeneration (Médail & Quézel, 1999).

Furthermore, with almost 5,000 islands and islets, the Mediterranean comprises one of the largest groups of islands in the world. Mediterranean islands display extraordinary features, with high rates of endemism, and act as a natural laboratory for evolutionary studies. Their particularities give rise to specific conservation challenges. Thus many of the endemic island plant species are confined to single small locations, they are extremely vulnerable to habitat destruction, overgrazing, and urban expansion (Fig. 4).

The *Top 50 Mediterranean Island Plants* highlights some of the most threatened plant species of the Mediterranean islands, stressing particular situations and conservation needs (Montmollin and Strahm, 2005).

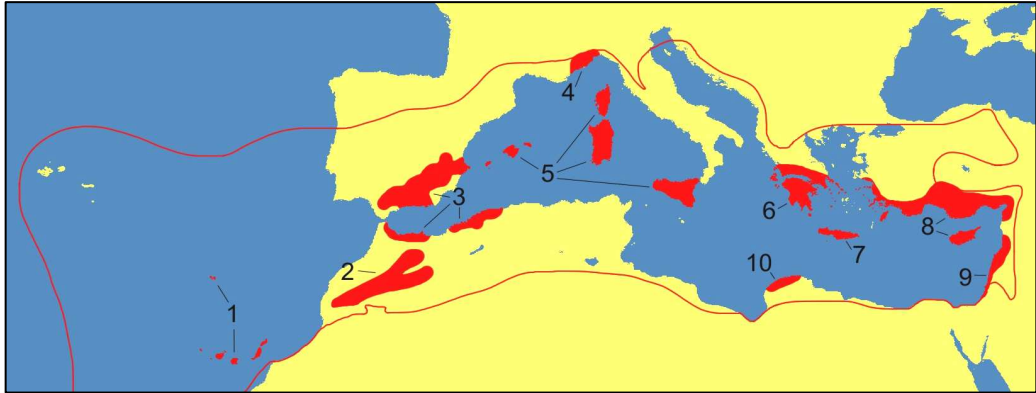


Figure 4. Mediterranean basin hotspots. 1: *Canaries and Madeiran archipelagos*; 2: *High and Middle Atlas Mountains*; 3: *Baetic-Rifan complex*; 4: *Maritime and Ligurian Alps*; 5: *Tyrrhenian islands*; 6: *Southern and Central Greece*; 7: *Crete*. 8: *Anatolia and Cyprus*; 9: *Syria-Lebanon-Israel*; 10: *Mediterranean Cyrenaic*.

Sardinia (Fig. 5) for its orographic condition, geographical location, special chorologic and ecological features, as well as for its low population compared to the extension of the territory, guarded favorable area to the development and maintenance of a large number of endemic species, which appear to be over 10% of the population floristic island.

The endemic species can be grouped in entities related to:

- endemic Sardinian, if it concerns the only Sardinia or restricted area included in it;
- Sardinian-Corsican, extended to Corsica;
- sometimes Sardinia-Corsica-Balearic, also including the Balearic Islands;
- endemism is often extended to the Tuscan Archipelago, Tyrrhenian to the Region or other limited range



Figure 5. Topographic map of Sardinia.

So, Sardinia presents a considerable amount of endemic taxonomic units (Table 1), specially in its mountain massifs, mostly tied to carbonatic substrata, establishing the conditions of ecologic isolation that cause the hot spot effect (Médal & Quézel, 1997; Bacchetta *et al.*, 2005).

Table 1. Vital signs of Sardinia.

	Vascular flora	Endemic flora
Families	135	52
Genera	695	158
Taxa	2 054	347
Endemics as a percentage of total Sardinian vascular flora		16.89
Endemics as a percentage of world total		0.12
Region extent (km ²)		24 090
Human population density (people/ km ²)		69

Furthermore, in Sardinia region, as well as in the whole Mediterranean basin, several of endemic species have a narrow distribution as *Anchusa capellii* Moris, *A. formosa* Selvi, Bigazzi et Bacch., *A. littorea* Moris, *Aquilegia barbaricina* Arrigoni et Nardi, *A. nuragica* Arrigoni et Nardi, *Astragalus maritimus* Moris, *A. verrucosus* Moris, *Borago morisiana* Bigazzi et Ricceri, *Centranthus amazonum* Fridlender et A. Raynal, *Dianthus morisianus* Vals., *Euphrasia genargentea* (Feoli) Diana, *Lamyropsis microcephala* (Moris) Dittrich et Greuter, *Limonium merxmulleri* Erben, *Linum muelleri* Moris, *Nepeta foliosa* Moris, *Polygala sinisica* Arrigoni, *Ribes sardoum* Martelli.

The *ex situ* conservation as biodiversity preservation strategy

Today, the impact of both direct and indirect human activities, such as urbanization, tourism, fires, changes in agricultural practices, introduction of alien and invasive species, and harvesting, as the important and continuing climate and environmental changes, make explicit the need for interventions aimed at the preservation and protection of biological resources now at risk (Montomollin & Strahm, 2005).

As mentioned above, the *ex situ* conservation has demonstrated to be essential in order to ensure biodiversity preservation and, to allow the right practices for collecting and storing seeds, in the last decade, there has been a significant increase in the establishment of centers for the conservation and germplasm banks, with the aim of studying not only the best storage conditions, but even phenology, and ecophysiology of seeds (viability, dormancy, germination range, optimal germination conditions and cardinal temperatures, longevity and soil seed bank), of the stored samples. Furthermore, to characterize the threatened species, it is essential to know their morphology (weight and morphometric traits of seeds) in order to compare these features with the common values.

The germplasm characterization can be carried out through the evaluation of qualitative parameters, related to the shape, size and colour of seeds. The evaluation of seed morphology and the colour definition, in a quantitative way, are complex and not always possible above all because the germplasm of the spontaneous species is characterized by high intraspecific variability (Granitto *et al.*, 2003; Harper *et al.*, 1970). So, these characteristics are difficult to measure and often it is possible only a subjective estimation.

Until a few years ago, sizes are manually measured and colour is roughly determined by comparison with standard colours of specific graphic tables from which it is possible to obtain approximate values of RGB (Red, Green, Blue) and HLS (Hue, Lightness, Saturation) (Fagundez & Izco, 2003). This method is clearly very subjective and not repeatable.

As demonstrated by recent scientific publications, electronics and computer science has provided technologic solutions so that all these limits can be overcome by using image analysis systems able to obtain accurate and precise measurements. Artificial vision is considered a subfield of engineering that is related to informatics, optics, mechanical engineering and industrial automation. Although one of the most common applications of machine vision is the inspection of manufactured goods, more times, this innovative technology proved to be a great help also in biological fields, and in particular it proved to be able to take precise and accurate measures about seeds shape, size and colour (Venora *et al.*, 2007; 2009a; Bacchetta *et al.*, 2008; Mattana *et al.*, 2008; Grillo *et al.*, 2010; Bacchetta *et al.*, 2011; Pinna *et al.*, 2014).

As human inspectors working on visual inspection to judge the quality and the quantity of germplasm features, so machine vision systems use digital

cameras and/or scanners, and image processing software to perform similar inspections.

Even if humans may display finer perception over the short period and greater flexibility in classification and adaptation to new defects and quality assurance policies, many times machine vision systems appear more adequate than human inspectors, specially for visual inspections that require high-speed, high-magnification and/or repeatability of measurements. Frequently these tasks extend roles traditionally occupied by human beings whose degree of failure is classically high through distraction, illness and circumstance. However, computers do not “see” in the same way that human beings are able to. Cameras are not equivalent to human optics and while people can rely on inferences and assumptions, computing devices must “see” by examining individual pixels of images, processing them and attempting to develop conclusions with the assistance of knowledge bases and features such as pattern recognition engines.

Artificial vision concerns to the fundamental part of instrumental acquisition of images; while image analysis, regards their processing and the numerical control, representing so the way to objectify and parameterize measures and evaluations (Symons *et al.*, 2003; Venora *et al.*, 2009b; Grillo, 2009).

Today, the artificial vision has an essential role in the study of vegetal biology, allowing an interaction between knowledges pertaining to disciplines of high technologic and innovative capability, such as electronics and computer science, with competences relating to biological area, in order to bring out multidisciplinary connections between studies and researches many times outwardly very different.

State of the art and innovative aspect of the project

Driven by the excellent results obtained in previous works about the characterization and identification of seeds of cultivated species using the image analysis techniques (Granitto *et al.*, 2003; Shahin & Symons, 2003; Kilic *et al.*, 2007; Venora *et al.*, 2007, 2009a; 2009b), some years ago, within the scientific collaboration between the *Centre for Conservation of Biodiversity (CCB)* of the Department of Botany, University of Cagliari and the *Stazione Sperimentale di Granicoltura per la Sicilia*, a work of germplasm morpho-colorimetric characterization and statistical identification of the most representative families of Mediterranean vascular flora was developed. A database of 33 morpho-colorimetric features (Table 2) of autochthonous germplasm in entry into the Germplasm Bank of Sardinia (BG-SAR) was built and statistical classifiers able to discriminate seeds belonging to different genera and species, were realized, as described in Grillo PhD dissertation (2009).

Such classifiers, based on the *Linear Discriminant Analysis (LDA)*, showed high ability of correct identification and, then, they have been implemented for ten of the most representative families of the Mediterranean vascular flora (Grillo *et al.*, 2010).

Currently, this method is fully accepted, utilized in plant taxonomy studies and contributes to the correct conservation of species in germplasm banks, particularly in the identification of diasporas of wild plant species (Bacchetta *et al.*, 2011b; Grillo *et al.*, 2011, 2013; Pinna *et al.*, 2014; Santo *et al.*, 2015).

Table 2. Thirty-three selected features for seed measurements.

Feature	Description
<i>Mean R</i>	Mean of red channel pixel value, express in grey levels.
<i>StdD R</i>	Standard deviation of red channel pixel value, express in grey levels.
<i>Mean G</i>	Mean of green channel pixel value, express in grey levels.
<i>StdD G</i>	Standard deviation of green channel pixel value, express in grey levels.
<i>Mean B</i>	Mean of blue channel pixel value, express in grey levels.
<i>StdD B</i>	Standard deviation of blue channel pixel value, express in grey levels.
<i>Mean H</i>	Mean of hue channel pixel value, express in grey levels.
<i>StdD H</i>	Standard deviation of hue channel pixel value, express in grey levels.
<i>Mean L</i>	Mean of lightness channel pixel value, express in grey levels.
<i>StdD L</i>	Standard deviation of lightness channel pixel value, express in grey levels.
<i>Mean S</i>	Mean of saturation channel pixel value, express in grey levels.
<i>StdD S</i>	Standard deviation of saturation channel pixel value, express in grey levels.
<i>Mean D</i>	Mean of density pixel value, express in grey levels.
<i>StdD D</i>	Standard deviation of density pixel value, express in grey levels.
<i>Skew</i>	Measure of asymmetry of the density values distribution.
<i>Kurtosis</i>	Measure of concentration or dispersion of the density values.
<i>Energy</i>	Measure of force of the increase in intensity.
<i>Entropy</i>	Measure of force of the dispersion, as chaos of density levels.
<i>Sum D</i>	Sum of density pixel value, express in grey levels.
<i>Sum SQR D</i>	Sum of squares of density pixel value, express in grey levels.
<i>Area</i>	Area of the seed projection, express in mm ² .
<i>Feret min</i>	Minimum diameter of the seed projection, express in mm.
<i>Feret max</i>	Maximum diameter of the seed projection, express in mm.
<i>Feret ratio</i>	Ratio of minimum to maximum diameters.
<i>Perimeter</i>	Perimeter of the seed projection, express in mm.
<i>Convex perimeter</i>	Perimeter of the seed projection, excluding concave zones.
<i>Crofton's perimeter</i>	Perimeter of the seed projection, according to the Crofton's formula.
<i>Perimeter ratio</i>	Ratio of convex to Crofton's perimeters.
<i>F Circle</i>	Shape factor = $(4*\pi*Area)/(Perimeter)^2$.
<i>D Circle</i>	Value of the diameter of the equal area circle.
<i>Ellipse A max</i>	Major axis of the ellipse with same area.
<i>Ellipse A min</i>	Minor axis of the ellipse with same area.
<i>Roundness factor</i>	Roundness factor = $(4*Area)/[\pi*(Feret Max)^2]$

Image measurements

Image analysis attempts to find the descriptive parameters, usually numeric, that briefly represent the information of importance in the image, producing numeric output suitable for statistical analysis or graphical representations. Usually, measurements that can be performed on features in images can be grouped into three classes: size, shape and colour, or more in general, densitometric measurements.

The main basic measures of feature size in digital images are the area, the perimeter and the diameters. For a pixel-based representation, the *Area* (A) simply is the number of pixels within the feature, purely determined by counting. Higher is the resolution of the image and more precise and accurate are the measurements. Of course, it must be remembered that the size of a feature in a two-dimensional image may be related to the size of the corresponding object in three-dimensional space, because commonly the binary images in which the features are measured, are merely projections, or shadows of the objects (Grillo, 2009).

The *Perimeter* (P) of a seed, as well as of any other feature, could be well defined simply by counting the pixels composing the boundary around the seed. Consequently, also in this case, knowing the pixel resolution of the image it is possible to estimate the perimeter length simply by counting of boundary pixels. Even if this pixel counting might be considered a very simple operation, a lot of mathematical calculation and their setting are involved to establish what to measure within the image, such as the *Convex perimeter* and the *Crofton's perimeter*. As showed in Fig. 6, the *Convex perimeter* (P_{conv}) is referred to the perimeter of an object in which all the convexities were filled, and it is measured at the same way in which the net

perimeter is gauged. This parameter is very useful when structures morphologically heterogeneous must be evaluated.

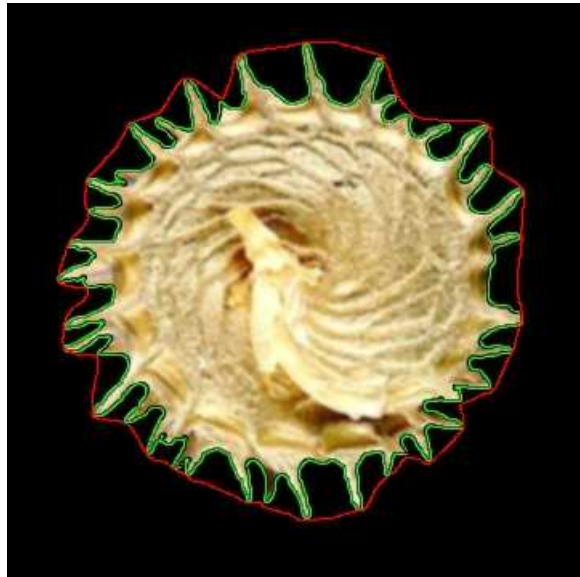


Figure 6. Net (in green) and *Convex perimeter* (in red) in *Medicago polymorpha* legum.

The *Crofton's perimeter* (P_{Crof}), or *Crofton' formula* (Crofton, 1869) allows an approach more mathematical to evaluate the length of the boundary around the seed. It is a classic result of integral geometry relating the length of a curve to the expected number of times a random line intersects it (Santalo, 1953). So, for this study, the ratio between the convex and Crofton's perimeters, was considered too.

The *Calliper dimensions*, or more commonly *Feret's diameters*, represent another measurement of object size. They were used to evaluate length and width of a seed or of other globose objects, but not to measure length and width of a fibre, because it might be twisted. Moreover, it was


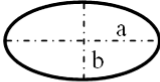
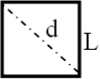
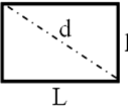
possible to calculate the *minimum* and the *maximum* diameters, by counting of axis pixels.

Shape measurements are dimensionless quantities, independent of their size, and commonly used in image analysis, that numerically describe the shape of an object. For example, the ratio between length and width, or more precisely between the minimum diameters and the maximum orthogonal to it, gives the *aspect ratio*. The *Shape factor* is a value very commonly used to describe the symmetry of an object; it is a function of the perimeter P and the area A , and it is reported as a normalized value. In this case, a factor equal to one represents the perfect circle.

The *Roundness factor* is a parameter less used then the previous. It describes the circularity of an object and it is normalized too; this factor is a function of the maximum diameter D_{max} and the area A . The *Feret ratio* (Fr) is a function of the two diameters D_{min} and D_{max} (Grillo, 2009).

Table 3 shows some examples of values that this three factors can give, relates to four different shapes.

Table 3. Numerical differences between some shape measurements.

		Sf	Rf	Fr
	$r = x$	1	1	1
	$a = 6$ $b = 4$	0,67	0,67	0,67
	$L = 2$ $d = 2,828$	0,785	0,64	0,71
	$L = 5$ $l = 2$ $d = 5,384$	0,16	0,44	0,37

Project aim and objectives

An adequate definition of the seed morpho-colorimetric parameters, represents an important diagnostic factor in the plant taxonomy studies and consequently may be of great help for the improvement of the management and the effective *ex situ* conservation in the germplasm banks (Grillo, 2009).

The discriminant ability of a classification system depends not only on the intra-specific representativeness of *taxa* analyzed, but also, on the quality and quantity of the parameters measured and used to discriminate between groups of belonging. For this reason, it is supposable that an increase in parameters evaluated for each seed, it will be useful to improve the performance of the classifier.

From the recent literature, it appears that the study of surface texture of an object (Fig. 7), whatever its nature, seems to be of great importance for the characterization of the same (Diamond *et al.*, 2004; Gerger & Smolle, 2004; Nanni *et al.*, 2010). There are many texture indicators (Fig. 8), based on Haralick's parameters, able to assess quantitatively how the colour tones may vary within an object by defining, in a particularly detailed way, colour, density and the different chromatic variations. Few results are reported in the literature about this kind of studies on seeds, and so, it might be interesting and original, to including texture parameters (about 20) in the classification system already developed.



Figure 7. Significant variations in texture on flowers surface of *Malva* accessions.

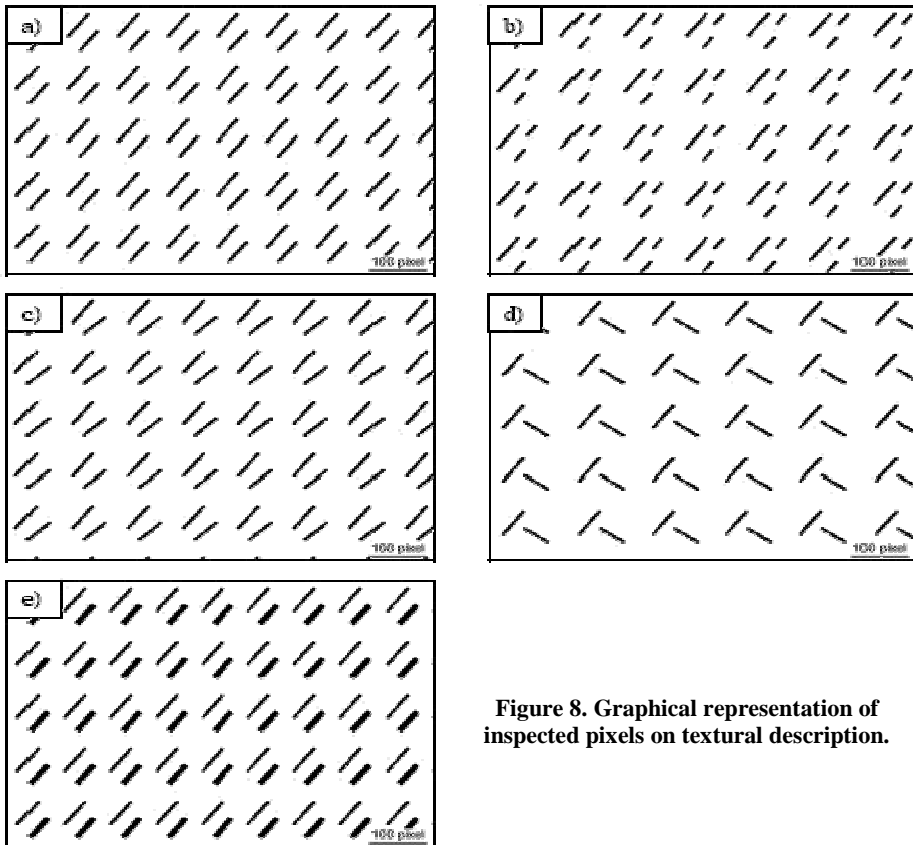


Figure 8. Graphical representation of inspected pixels on textural description.

Also the study of seeds morphology could be improved including new parameters describing the shape: these parameters are the “Elliptic Fourier Descriptors” hereafter EFDs (Iwata *et al.*, 2002, 2004; Kawabata *et al.*, 2009; Yoshioka *et al.*, 2004, Orrù *et al.*, 2013). Based on the profile of a seed projection on the two-dimensional plane, it is possible to generate codes descriptive of the shape. These codes, known as “chain codes”, allow describing, in detail, the outline of a shape and, for comparison with the ellipses geometrically perfect, the Fourier descriptors are obtained.

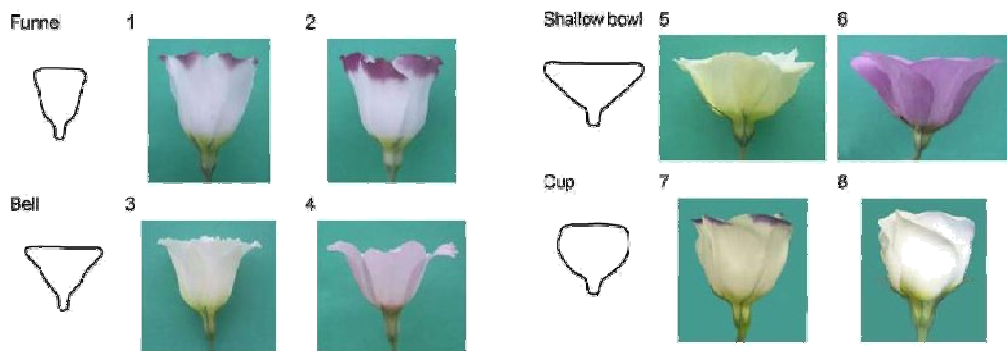


Figure 9. Images of corollas from *Lisianthus* cultivars with four typical corollas shapes.

Then, 78 additional parameters that can be included as common morphological variables in the statistical classification system.

As a consequence, for each seed can be measured and recognized:

- 33 descriptors of morpho-colourimetric features, already commonly used in the taxonomy studies with image analysis tools;
- 20 textural parameters, Haralick’s descriptors;
- 78 EFDs.

Therefore, an overall of about 130 traits, that constitutes a very huge amount of seed traits, never previously achieved, can be determined.

The improvement of this system, which has already proved to be highly effective in *taxa* identification, will be particularly useful for the management of common activities of germplasm banks, in particular for the determination of unknown specie seeds or in confirmation of doubt classifications, for the analysis of soil seed banks, for ecological or archaeobotanical studies, and at infrageneric level, for the determination and revision of critical, or currently under review, taxonomic groups.

Then, the objectives of the research project conducted during the PhD program were:

1. to develop a specific *Macro* introducing Haralick's parameters and EFDs for a more detailed texture and shape characterization of wild seeds;
2. to identify, measure and evaluate, through image analysis, morpho-colourimetric characters of some of the most representative species of the Mediterranean vascular flora stored in the *Sardinian Germplasm Bank* (BG-SAR);
3. to implement statistical classifiers, in order to recognize and discriminate seeds belonging to different families, genera and species. A database, constantly in evolution, should be at the bottom of classifiers, and so it should be necessary to provide for such updating, during the standard management of the germplasm in the seed banks;
4. to improve the *ex situ* conservation of the targeted species and the germplasm banks management trough this innovative classification system.

References

- BACCHETTA G. 2011a. Conservare la natura. In: TAFFETA NI F. (eds.), *Manuale sugli erbari*. Cardini Ed., Firenze.
- BACCHETTA G., IIRITI G., PONTECORVO C. 2005. Contributo alla conoscenza della flora vascolare endemica della Sardegna. *Informatore botanico italiano* 3377, 306-307.
- BACCHETTA G., GARCÍA P.E., GRILLO O., MASCIA F., VENORA G. 2011b. Seed image analysis provides evidence of taxonomical differentiation within the *Lavatera triloba* aggregate (Malvaceae). *Flora* 206, 468-472.
- BACCHETTA G., GRILLO O., MATTANA E., VENORA G. 2008. Morph-colorimetric characterization by image analysis to identify diaspores of wild plant species. *Flora* 203, 669-682.
- BARBERO M., BONIN G., LOISEL R., QUÉZEL P. 1990. Changes and disturbances of forest ecosystems caused by human activities in the western part of the Mediterranean Basin. *Vegetatio* 87, 151-173.
- CHAPIN F.S., ZAVALA E.S., EVINER V.T., NAYLOR R.L., VITOUSEK P.M., REYNOLDS H.L., HOOPER D.U., LAVOREL S., SALA O.E., HOBBIE S.E., MACK M.C., DÍAZ S. 2000. Consequences of changing biodiversity. *Nature* 405, 234-242.
- CODY M.L. 1986. Diversity, rarity, and conservation in Mediterranean-climate regions. In: Soulé ME (ed) *Conservation Biology*, pp 122-152. Sinauer Associates, Sunderland, Massachusetts.
- DIAMOND J., ANDERSON N.H., BARTELS P.H., MONTIRONI R., HAMILTON P.W. 2004. The use of morphological characteristics and texture analysis in the identification of tissue composition in prostatic neoplasia. *Human Pathology* 35, 1121-1131.
- FAGUNDEZ J. & IZCO J. 2003. Seed morphology of *Calluna Salisb.* (Ericaceae). *Acta Botanica Malacitana* 29: 215-220.
- GERGER A. & SMOLLE J. 2003. Diagnostic imaging of melanocytic skin tumors. *Journal of Cutaneous Pathology* 30, 247-252.
- GRANITTO P.M., GARRALDA P.A., VERDES P.F., CECCATO H.A. 2003. Boosting classifiers for weed seeds identification. *Journal of Computer Science and Technology* 3, 34-39.

- GRILLO O. 2009. Germplasm morpho-colorimetric characterization by image analysis and statistical classification of the most representative families of Mediterranean vascular flora (*Doctoral dissertation*).
- GRILLO O., MATTANA E., FENU G., VENORA G., BACCHETTA G. 2013. Geographic isolation affects inter- and intra-specific seed variability in the *Astragalus tragacantha* complex, as assessed by morpho-colorimetric analysis. *Comptes Rendus de Biologies* 336, 102-108.
- GRILLO O., MATTANA E., VENORA G., BACCHETTA G. 2010. Statistical seed classifiers of 10 plant families representative of the Mediterranean vascular flora. *Seed Science and Technology* 38, 455-476.
- GRILLO O., MICELI C., VENORA G. 2011. Computerised image analysis applied to inspection of vetch seeds for varietal identification. *Seed Science and Technology* 39, 490-500.
- HARPER J.L., LOVELL P.H., MOORE K.G. 1970. The shapes and sizes of seeds. *Annual Review Ecology and Systematics* 1, 327-356
- KAWABATA S. YOKOO M., NII K. 2009. Quantitative analysis of corolla shapes and petal contours in single-flower cultivars of *Lisianthus*. *Scientia Horticulturae* 121, 206-212.
- KILIC K., BOYACI I.H., KOKSEL H., KUSMENOGLU U.I. 2007. A classification system for beans using computer vision system and artificial neural networks. *Journal of Food Engineering* 78, 897-904.
- IUCN. 2011. IUCN Red List of Threatened Species (ver. 2011.1). Available at: <http://www.iucnredlist.org>. (Accessed: 16 June 2011).
- IWATA H., NESUMI H., NINOMIYA S., TAKANO Y., UKAI Y. 2002. Diallel analysis of leaf shape variations of *Citrus* varieties based on Elliptic Fourier Descriptors. *Breeding Science* 52, 89-94.
- IWATA H., NIIKURA S., SEIJI M., TAKANO Y., UKAI Y. 2004. Genetic control of root shape at different growth stages in radish (*Raphanus sativus* L.). *Breeding Science* 54, 117-124.
- MATTANA E., GRILLO O., VENORA G., BACCHETTA G. 2008. Germplasm image analysis of *Astragalus maritimus* and *A. verrucosus* of Sardinia (subgen. *Trimeniaeus*, Fabaceae). *Anales Jardin Botanico de Madrid* 65: 149-155.
- MÉDAL F. & QUÉZEL P. 1997. Hot-spots analysis for conservation of plant biodiversity in the Mediterranean basin. *Mediterranean Plant Biodiversity* 84: 112-127.

- MÉDAL F. & QUÉZEL P. 1999. Biodiversity Hotspots in the Mediterranean Basin: Setting Global Conservation Priorities. *Conservation Biology* 13, 1510-1513.
- MONTMOLLIN B. & STRAHM W. (eds). 2005. *The Top 50 Mediterranean Island Plants: Wild plants at the brink of extinction, and what is needed to save them*. IUCN SSC Mediterranean Islands Plant Specialist Group. IUCN, Gland, Switzerland and Cambridge, UK.
- MYERS N., MITTERMEIER R.A., MITTERMEIER C.G., DA FONSECA G.A.B., KENT J. 2000. Biodiversity hotspots for conservation priorities. *Nature* 403, 853-858.
- NANNI L., SHI J.Y., BRAHNAM S., LUMINI A. 2010. Protein classification using texture descriptors extracted from the protein backbone image. *Journal of Theoretical Biology* 264, 1024-1032.
- ORRÙ M., GRILLO O., LOVICU G., VENORA G., BACCHETTA G. 2013. Morphological characterisation of *Vitis vinifera* L. seeds by image analysis and comparison with archaeological remains. *Vegetation History and Archaeobotany* 22, 231-242.
- PINNA M.S., GRILLO O., MATTANA E., CAÑADAS E.M., BACCHETTA G. 2014. Inter- and intraspecific morphometric variability in *Juniperus* L. seeds (Cupressaceae). *Systematics and Biodiversity* 12, 211-223.
- SANTO A., MATTANA E., GRILLO O., BACCHETTA G. 2015. Morpho-colorimetric analysis, germination variability and heteromorphy of *Brassica insularis* Moris (Brassicaceae) seeds. *Plant Biology*, doi:10.1111/plb.12236.
- SHAHIN M.A. & SYMONS S.J. 2003. Lentil type identification using machine vision. *Canadian Biosystems Engineering* 45, 3.5–3.11.
- SYMONS S.J., VAN SCHEPDAEL L., DEXTER J.E. 2003. Measurement of hard vitreous kernels in durum wheat by machine vision. *Cereal Chemistry* 80, 511-517.
- TILMAN D. 2000. Causes, consequences and ethics of biodiversity. *Nature* 405, 208-211
- VENORA G., GRILLO O., RAVALLI C., CREMONINI R. 2009a. Identification of Italian landraces of bean (*Phaseolus vulgaris* L.) using an image analysis system. *Scientia Horticulturae* 121, 410-418.
- VENORA G., GRILLO O., SACCONI R. 2009b. Durum wheat storage centres of Sicily: evaluation of vitreous, starchy and shrunken kernels by image analysis system. *Journal of Cereal Science* 49, 429-440.
- VENORA G., GRILLO O., SHAHIN M.A., SYMONS S.J. 2007. Identification of Sicilian landraces and Canadian cultivars of lentil using image analysis system. *Food Research International* 40, 161-166.

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. *Annals of Botany* 94, 657-664.

The image analysis technique and statistical treatment of data

Morphometry is the science of measuring of quantitative parameters of object morphology, while the colorimetry is the science used to quantify and describe physically the human colour perception. Morpho-colorimetric evaluations are commonly employed as a tool to assess shape, size and colour of objects, in order to relate these physical characters with quality aspects. Compared to conventional measurements, computer-aided morpho-colorimetry is exponentially faster, more accurate, precise and efficient, providing a significantly broader spectrum of measurements of morphological and colorimetric features and, at the same time, replacing subjective estimations with objective quantifications. The first part of this dissertation introduces the fundamentals of image analysis, starting with the essentials of computer vision, the elaboration and processing techniques usually applied to digital images and to achieve binary images used as masks to measure the objects of interest (Chapter 1). Chapter 2 deals with the new shape and texture features of seeds producing by Elliptic Fourier Descriptors (EFDs) and Haralick's parameters introduced in the image analysis system to improve the germplasm characterization. Chapter 3 gives a detailed argumentation concerning the materials and the methods used in this work, dealing about the selected and analysed germplasm, the applied methodologies and the used tools, including an explanation regarding statistical treatment of the raw data achieved by image analysis.

Introduction

Although the human eye is a high precision tool, it has some important restrictions if compared with the use of imaging devices based on computers for technical purposes. Human vision is especially poor at judging color or brightness of features; it is inherently qualitative and comparative rather than quantitative, responding to the relative size, angle, or position of several objects but unable to support numeric measures unless one of the reference objects is a measuring device (Russ, 2007).

Moreover human vision is limited to wavelength ranged between 380 and 700 nm, with a higher sensibility at 550 nm, hence only a little portion of the spectrum. In addition, human vision tends to overestimate and underestimate the information at the borderline between objects with different intensity according to the “Mach Band effect” phenomenon (Fig. 1).

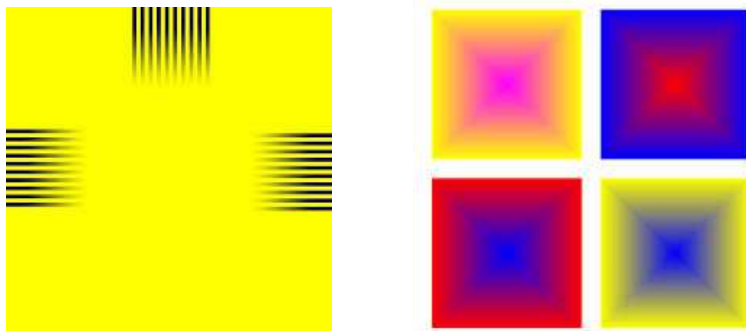


Figure 1. Mach Band effect. *The thin lines or “bands” along the gradient are illusory.*

At last, because of the “Simultaneous Contrast effect”, human vision is particularly influenced by the background light (Fig. 2).

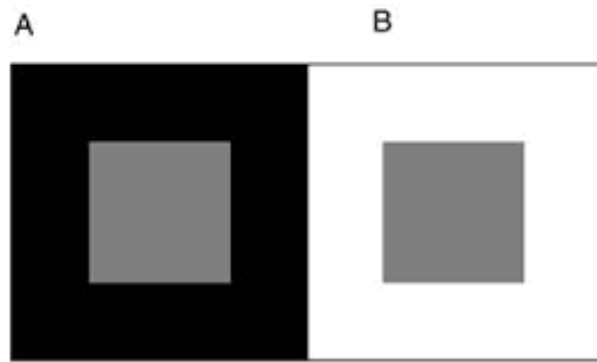


Figure 2. Simultaneous Contrast effect. *It is possible to notice that the square B on the right is more contrasted and visible than the left A.*

These faults of the human vision are the cause of various visual illusions (Fig. 3).

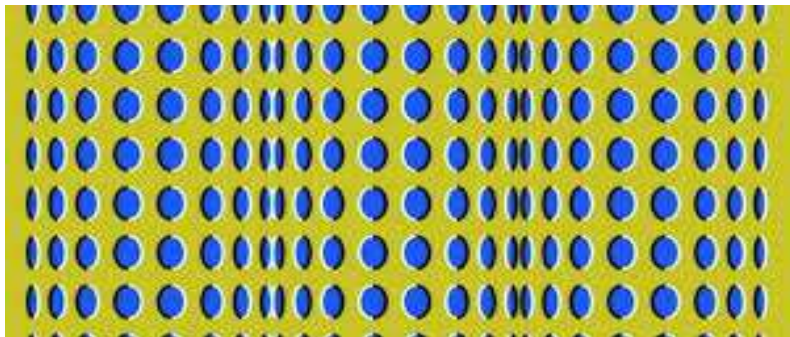


Figure 3. A moving optical illusion.

Traditionally, in any field, quality inspection is performed by trained human inspectors. In addition to being costly, this method is highly variable and decisions are not always consistent between inspectors or from day to day. This is, however, changing with the advent electronic imaging systems

and with the rapid decline in cost of computers, peripherals and other digital devices. Moreover, especially for various quality factors of biological elements, the inspection can be a very repetitive task, but also very subjective. In this type of environment, machine vision systems are ideally suited for routine inspections and qualitative and quantitative assessments. To date, machine vision has extensively been applied to solve various problems, ranging from simple quality evaluation of food products to complicated robot guidance applications (Tao *et al.*, 1995; Pearson, 1996; Abdullah *et al.*, 2000; 2008).

Image acquisition systems

The application of machine vision has increased considerably in recent years. There are many fields in which machine vision is involved: terrestrial and aerial mapping of natural resources (Hirano *et al.*, 2003), crop monitoring (Ling *et al.*, 1996), robotics (Blasco *et al.*, 2002), quality control (Daley & Britton, 2003; Zheng & Sun, 2009), non-destructive inspections (Venora *et al.*, 2009b; 2009c), and many more. In the last 20 years, many authors have successfully applied some of the techniques developed in these fields for botanical study, with the aim to make easy and over all objective, the dimensional evaluation of anatomical elements (Venora & Calcagno, 1991; Venora & Porta-Puglia, 1993; Hu *et al.*, 2006; Xiang Du *et al.*, 2007; Bachetta *et al.*, 2008; Sánchez del Álamo *et al.*, 2008), identification of different physical defects (Blasco *et al.*, 2007; Venora *et al.*, 2009b), and also the classification of germplasm to distinguish different species or agronomical varieties, belonging to the same taxonomic rank (Venora *et al.*, 2007; 2009a; Mattana *et al.*, 2008; Bacchetta *et al.*, 2008, 2011a, 2011b; Zapotoczny *et al.*, 2008, Grillo *et al.*, 2009, 2010, 2013).

The breadth of applications depends, among many other things, on the fact that machine vision systems provide substantial information about the nature and the attributes of the objects present in the scene. Another important feature of such system is that they open the possibility of studying this object in regions of the electromagnetic spectrum, where human vision is unable to operate, as in the ultraviolet or infrared regions (Moltó & Blasco, 2009). In addition to the opportunity to work applying non-destructive and automatic techniques, image analysis provides greater reliability and objectivity than human inspection, because the decision made by operators are easily affected by external factors such as fatigue, acquired habits, competences and frequently also by culture (Studman & Ouyang, 1997; Venora *et al.*, 2009b). For the same reason, machine vision allows to obtain higher repeatability than human inspection, minimizing or standardizing the possible mistakes, and furthermore great speed during the execution of the analysis. Finally, image analysis allows to execute computations, giving the opportunity to take more information by the same object.

Depending on the nature of the sample, on the size and location of it, the acquiring system must suit certain needs, and consequently it take to have some specifications, but, in many cases, a video or photo-camera can be adequate to capture images. Working on the detection of external defects, or on the classification of objects with different morpho-colorimetric characteristics, with a digital camera it is possible to obtain high quality images containing all the need information. In some cases, it can be helpful to apply specific filters to detect specific elements. For example, to evaluate the quality of many food, using different filters able to absorb different wavelength light, it is possible to distinguish a defect from another, or one type of damage from another (Blasco *et al.*, 2007).

Also the flatbed scanner proved to be the ideal acquiring source in a lot of application. Generally, when the size and over all the depth of the samples are small, and when the capture time are not important, the flatbed scanner is probably the best acquiring system, although unlike the camera, the flatbed scanner not allows to regulate the focus to optimize the sharpness of the sample. This kind of acquiring system, are image gathering devices that incorporate a fixed relationship between the illuminating source (lamp) and the solid state sensors of the scanning head, allowing to capture the sample images in a constant manner. Due to their increasing popularity, the cost of these devices is dropping rapidly. These characteristics may turn the flatbed scanners into the image acquisition system of choice (Shahin & Symons, 2001; Grillo, 2009).

Digital images

The hardware configuration of computer-aided machine vision systems is relatively standard. Typically, a vision system consist of:

- an illumination device which illuminates the sample;
- an image acquisition system, such as a camera, a scanner or a image reconstruction apparatus;
- a personal computer or a microprocessor system to provide the image processing, analysis and storage;
- a high resolution colour monitor which allows the visualization of images and the effects of various processing routines.

Figure 4 shows a complete set-up, provided with a professional flatbed scanner equipped with a trans-illuminator cover.

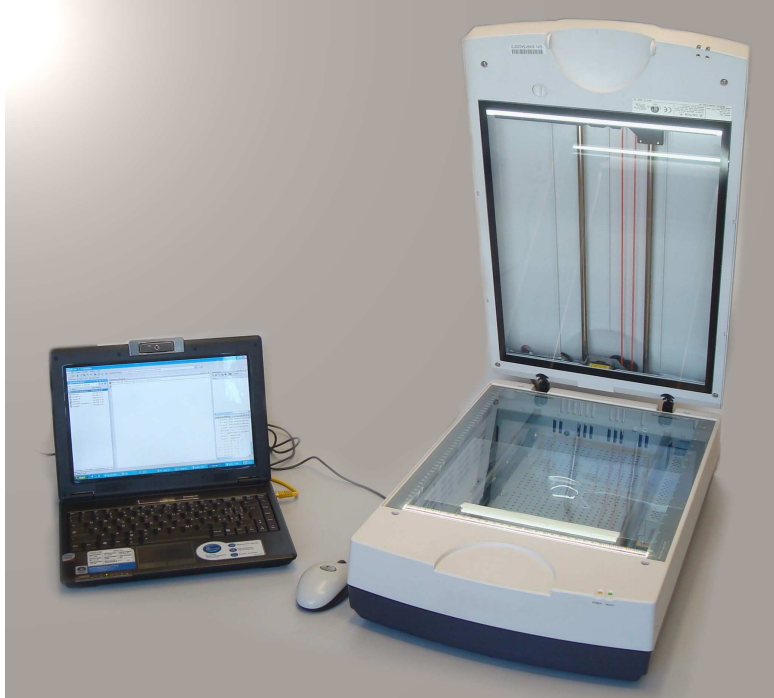


Figure 4. Essential elements of a computer vision system.

A part from the used acquisition system, the first goal to achieve is the image digitalization. This mathematical procedure, executed by the capture device, consists in the conversion of a real image into a digital image, that simply is a matrix of number. A digital image $a[m,n]$ described in a 2D discrete space is derived from an analogical image $a(x,y)$ in a 2D continuous space through a sampling process that is frequently referred to as digitization (Young *et al.* 1995).

The 2D continuous image $a(x,y)$ is divided into N rows and M columns. The intersection of a row and a column is termed a *pixel* (Fig. 5). The value assigned to the integer coordinates $[m,n]$ with $\{m = 0,1,2,\dots,M-1\}$ and $\{n = 0,1,2,\dots,N-1\}$ is $a[m,n]$. In fact, in most cases $a(x,y)$, which we might consider to be the physical signal that impinges on the face of a 2D sensor, is

actually a function of many variables including depth (z), colour (l), and time (t) (Young *et al.* 1995).

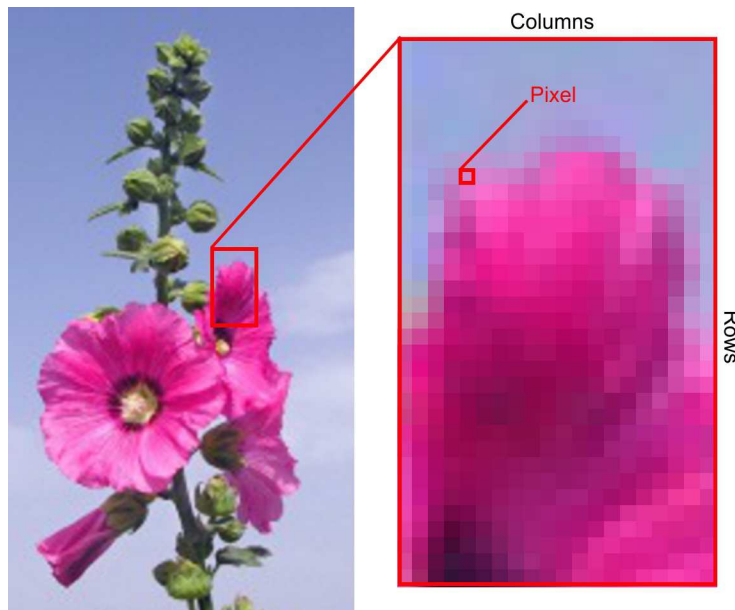


Figure 5. Digitization of a continuous image.

The number of pixel per unit of measurement can be used to define the *resolution* of a digital image, even though a lot of international standards specify that it should not be so used, at least in the digital camera field. In this cases, the convention is to describe the *pixel resolution* with the set of two positive integer numbers (M,N), where the first number is the number of pixel columns (width) and the second is the number of pixel rows (height), for example as 640×480 .

Another popular convention is to cite resolution as total number of pixels in the image, typically given as number of megapixels, which can be calculated by multiplying pixel columns by pixel rows and dividing by one

million. Other conventions include describing pixels per length unit or pixels per area unit, such as pixels per inch (PPI) or per square inch (DPI). But, even if they are widely referred to as such, none of these *pixel resolutions* are true resolutions, because they simply describe the geometric resolution of a digital image (Fig. 6).



Figure 6. Different pixel resolution image.

As stated above, the value assigned to the integer coordinates $[m, n]$ (the pixel), is a function of some variables, including depth (z), colour (l), and/or time (t). Consequently, the quality of a grey scale image q is defined by a tern of integer numbers, $q(m, n, z)$, where m and n are the pixel coordinates, and z are the grey depth.

To define numerically the grey or colour depth, it is need to introduce another important notion, helpful to understand how many information are included in a digital image. The terms *color depth* or *bit depth* describe the number of bits used to represent the colour of a single pixel. *Bit* (binary unit) is the unit of measurement of digital information, and it uses a binary decoding system, only constituted by two integer number, 0 and 1. The number of bits of a digital image defines the grey tones resolution that increase esponetially relating to the bit levels. Figure 7 shows the same image with different bit depth. It is possible to note that the more bit depth increases, the more image definition appears detailed (Grillo, 2009).



Figure 7. Different bit depth image.

Image processing

Owing to the imperfections of image acquisition systems, often the captured image are subject to various defects that could affect the subsequent processing and consequently the image analysis. Therefore, it is preferable to correct the image, after they have been acquired and digitalized it (Zheng and Sun, 2008). Generally this procedure is fast and relatively low-cost; for example, it includes noise removing operations, smoothing filters applications, contrast regulations, image histograms equalizations and much more. All operations of defects correction and image preparation to the analysis are commonly defined *image processing* or *digital picture processing*, as it was often called.

Noise, contrast and shading correction

An acquired image is always subject to different types of noise, such as the *readout noise*, produced by the sensor of camera, or the *electronic noise* caused by electronic circuit of the capture device during the analog/digital conversion, or the *salt and pepper noise* that generates, in the image scattered pixels with very different color or intensity from their

surrounding pixels. All this leads degradation of image quality (Fig. 8). There are various solutions that can be used to adjust images with this kind of defects, and all applies similar mathematical algorithms as smoothing filter (Lee, 1983; Mastin, 1985; Rank *et al.*, 1999; Freeman *et al.*, 2006).



Figure 8. Example of an image corrupted by noise (salt and pepper effect).

Smoothing filters are used to blur an image and reduce noise. There are different types of such filters that can be applied for different kinds of problems. Some of them are linear filters, as the *Mean* and the *Gaussian* filters, while others are non-linear filter, as the *Median* filter. In linear filters, the output pixel value is calculated using the weighted sum of the input pixels. A non-linear filter does not calculate the weighted sum of pixels in the neighbourhood. It assigns a value to the output pixel, which is directly based on the values of the pixels in the neighbourhood.

Figure 9 shows the effects of the *Gaussian* and *Median* filter applied to the noised image showed in figure 8. Images A and B are the result of a *Gaussian* functions application, respectively with a 3 x 3 and 9 x 9 template

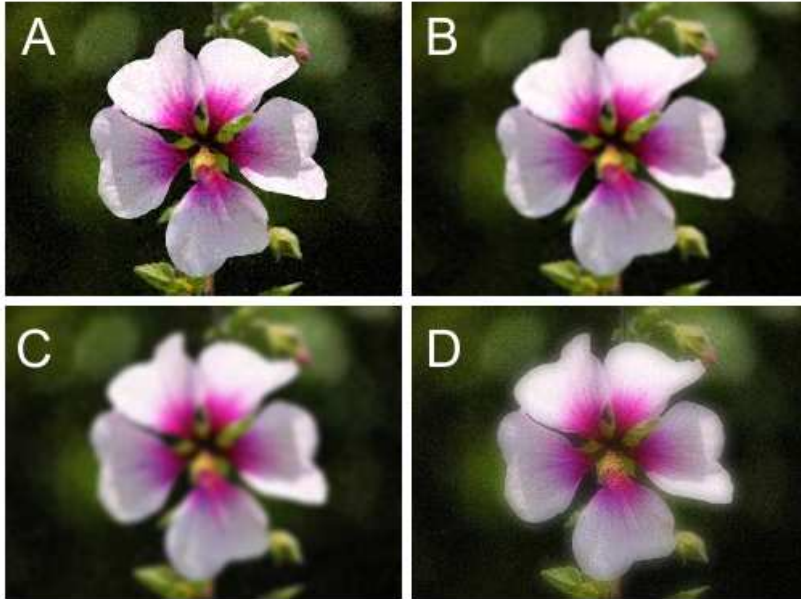


Figure 9. Example of image correction by using smoothing filters.

size. It is possible to note that the salt and pepper effect is only lightly reduced and the blurring of the image is remarkable. Instead, images C and D are the result of a *Median* filter, applying respectively templates with size of 3×3 and 9×9 . In this case it is evident the 3×3 *Median* function (C) provides the best result without blurring or smoothing effects on the output image, while a template of 9×9 (D) appears overloaded.

Today it is possible to acquire images with an optimum quality, because the high technology of the capturing devices allows to do it. However, sometimes captured image are not enough contrasted or, in other words, the intensity values of the image are restricted to a small range of intensity levels, and thus pixels with different intensity values are not well distinguished from each other (Zheng & Sun, 2008).

Most of the contrast enhancing tools use the image histogram (Jain, 1989), a plot of the number of pixels with each possible brightness level. It is a valuable tool for examining the contrast in the image (Russ, 2007). Figure 10, shows an example image in which the histogram covers the full dynamic range, indicating good contrast.

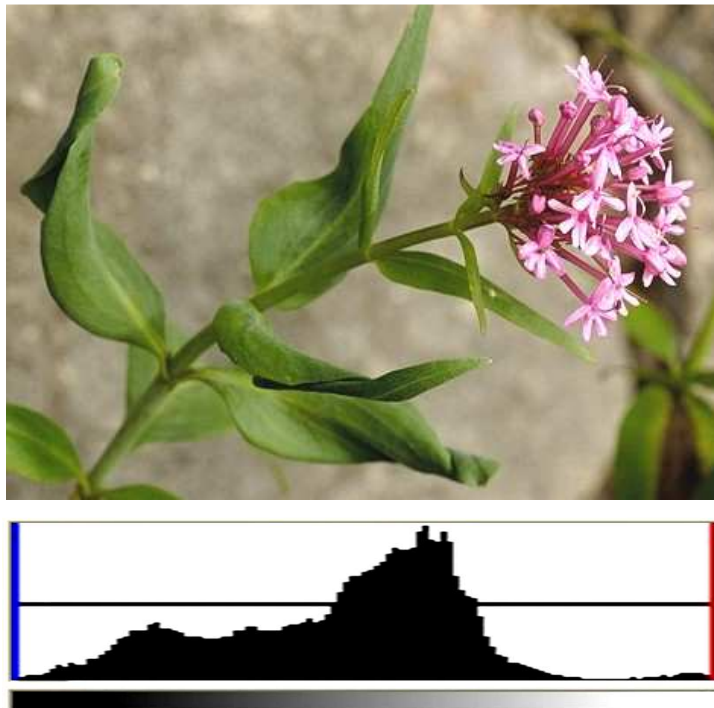


Figure 10. Example of a good exposure adjustment, since the brightness values cover the entire range without clipping at black or white.

Various formulas and various modalities exist to detect and apply the optimal histogram width (Scott, 1979; Hui *et al.*, 1999; Kim *et al.*, 2001), some of these are automatics and other interactively adjustable, but always, the original histogram is transferred from one scale to another, mostly from a smaller scale to larger one. Accordingly, the difference between two neighbouring intensity values is increased (Zheng & Sun, 2008).

Color calibration

Frequently, colour images present significant problems related to the lack of homogeneity of light source. Basically, the solution derives from a subtraction operation between the original not well illuminated image and the background not well illuminated image, but it can be applied only when the imperfection of illumination appears constant (Grillo, 2009).

Furthermore, in electronic imaging, and above all in computer vision, the imaging devices, cameras, scanners, colour monitors, require careful calibration to ensure that they reproduce standardized images (Lee, 2005).

Basically, the aim of colour calibration is to measure or adjust the colour response of a device (input or output) to establish a known relationship to a standard colour space. In image analysis, its importance is get involved both with the possibility to use different kinds of acquisition systems and with the necessity to exclude any light variation due to the wear on the illumination device, granting constant results. The device which has to be calibrated is sometimes known as *calibration source*, while the color space that serves as a standard is known as *calibration target*.

One of the most common process of colour calibration works for image matching, is reported in this research work. The *ad hoc* method developped by Shahin & Symons (2000) was applied to calibrate and standardize the images acquired using a flatbed scanner. This method uses a

Kodak Q60 Target Color Chart (Fig. 11) as reference image, to carry out a *Look-Up Table* (LUT) useful to match, compare and adjust the acquired image.

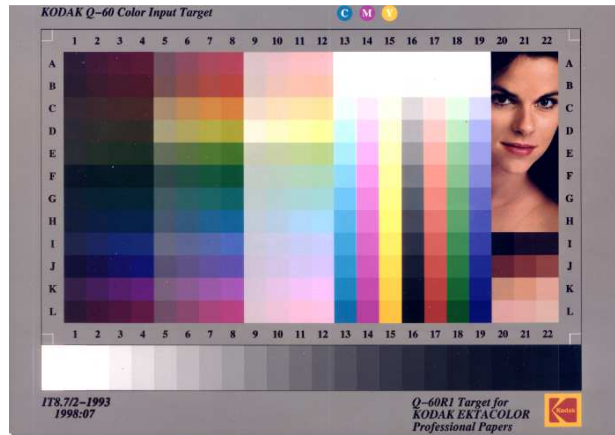


Figure 11. Kodak Q60 Target Color Chart reference image.

Mathematical morphology

The most important process in the switch from the pre-elaboration of an image to its measure, is the *segmentation*. It is a crucial step that allows to reduce images to information, dividing the image into regions and distinguishing the objects of interest. Segmentation is often described by analogy to visual processes as a foreground/background separation, implying that the selection procedure concentrates on a single kind of features and discards the rest. Although this is not quite true for computer systems, which generally deal much better than humans with scenes containing more than one type of features of interest (Russ, 2007), this analogy appear really appropriate to well understanding the concept of segmentation. Indeed, the result of segmentation is usually a binary image, in which the regions of interest (*ROIs*) are white and the background is black.

The aim of image segmentation is the domain-independent partition of the image into a set of region which are visually distinct and uniform with respect to some property, such as grey level, texture or colour (Freixenet *et al.*, 2002). The problem of segmentation has been, and still is, an important research field and many segmentation methods have been proposed in literature (Malik *et al.*, 2001; Frucci & Sanniti di Baja, 2008). Depending on the complexity of the processed image, it is possible to apply different segmentation methods to obtain a binary image in which to measure the regions of interest. Although hundreds of segmentation algorithms have been proposed in the last 30 years, basically two different approaches exist to tackle the segmentation: for *discontinuity* and for *similarity*. The methods based on the discontinuity property of the pixels, also called boundary-based methods, try to detect isolated dots, lines and borders to reconstruct the contours of the regions of interest. To do this, some filters that allow to identify the borderlines of the objects are used and a few operations of mathematical morphology are applied. Instead, the approach for similarity, commonly called region-based method, that is the most widely used, is useful when the regions of interest are segmented imposing an intensity threshold (Freixenet *et al.*, 2002; Riva, 2004).

Selecting features within an image is an important prerequisite for most kind of measurement or understanding of the scene. Traditionally, one simple way thresholding accomplished to define the range of brightness values in the original image, selects the pixels within this range as belonging to the foreground (*ROI*) and rejects all of the other pixels in the background. This is the simplest method of image segmentation, using black and white or other colours to distinguish the regions (Shapiro *et al.*, 2001 Russ, 2007). In a thresholding operation, the input is typically a greyscale or colour image and,

in the simplest implementation, the output is a binary image representing the segmentation and it is determined by a single parameter known as the *intensity threshold* (Fig. 12).

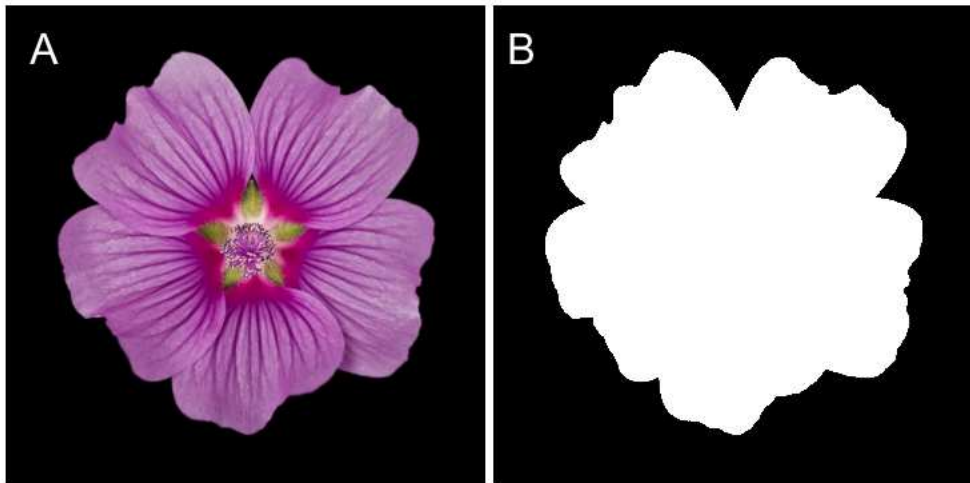


Figure 12. Thresholding a colour image. *Original image (A) and binary image (B).*

Thresholding may be set interactively by user watching the image and using a coloured overlay to preview the result and adjusting the setting. The brightness histogram of the image, or of a region of it, is very useful for making adjustments (Fig. 13).

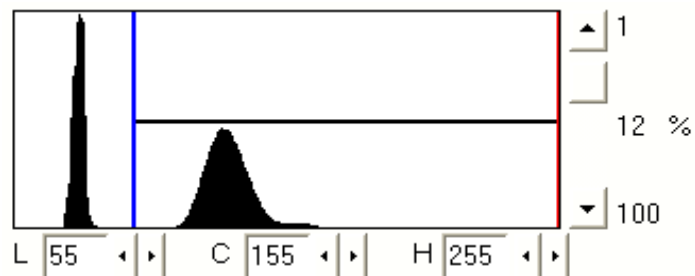


Figure 13. Threshold histogram for the brightness values selection.

There are also many automatic methods to adjust thresholding settings, using either the histogram or the image itself as a guide (Prewitt and Mendelson, 1966; Weszka, 1978; Otsu, 1979; Kittler *et al.*, 1985; Rigaut, 1988; Russ and Russ, 1988; Sahoo *et al.*, 1988; Lee *et al.*, 1990; Russ, 1995; Sezgin and Sankur, 2004).

Segmentation of grey scale images into regions for measurement or recognition is probably the most important area for image analysis. Many other thresholding methods exist and were used extensively in a lot of artificial intelligence applications (Fukunaga, 1990), as well as the *thresholding from texture*, that is a very interesting practice allowing the image segmentation on the bases of different texture orientation and/or spatial frequencies (Haralick *et al.*, 1975), or the *boundary lines* and *contour criteria*, that perform image thresholding on the bases of the boundary information, and many novel techniques too, that are rather *ad hoc* and narrow in their range of applicability, are constantly implemented. Review articles by Fu & Mui (1981) and Haralick & Shapiro (1988) present good guides to the literature, and most standard image analysis textbooks, such as Castleman (1979), Rosenfeld & Kak (1982), Gonzalez & Woods (2007), and Pratt (2007) also contain sections on image segmentation.

As result of segmentation, binary image represents the starting point for a geometric evaluation and it works as a mask for a colour assessment and for following image combinations. But the product of segmentation rarely is perfect. For images of realistic complexity, even the most elaborate segmentation routines misclassify some pixels as foreground or background. Generally, these are pixels belonging to the boundaries of regions or patches of noise within regions (Russ, 2007). The main tools that allow to correct this kind of mistakes, can be organized into two groups of operations: *Boolean*

logical operations for combining images, and *morphological operations* which modify single pixels within images.

These include *erosion* and *dilation*, *opening* and *closing*, *scrapping* and *filling* operations, as well as all the possible combination of them. All are fundamentally neighbour operations that work in spatial domain. Although these operations are discussed in literature in terms of set theory, a much simpler and more empirical approach is taken simply describing these operations in terms of adding and removing pixels from the binary image according to certain rules, which depend on the pattern of neighbouring pixels (Russ, 2007).

Erosion removes pixels from an image or, equivalently, discards any pixel touching other pixels that are part of the background (that is already OFF). This operation removes a layer of pixels from around the periphery of whole region of interest, causing some shrinking of dimensions (Fig. 14). As erosion removes pixels, the complementary operation of *dilation* adds pixels to the perimeter of the region. Figure 15 shows an example of a practical use of this morphological operations.

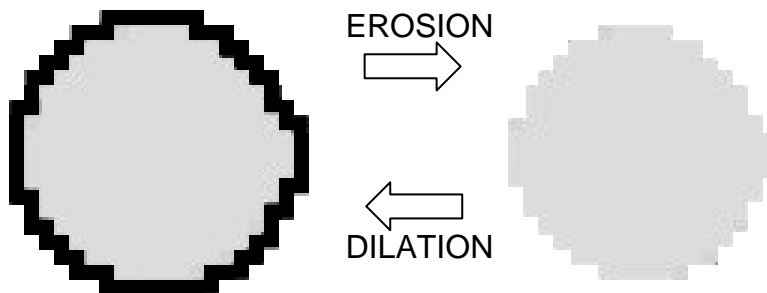


Figure 14. Erosion and dilation morphological operations.

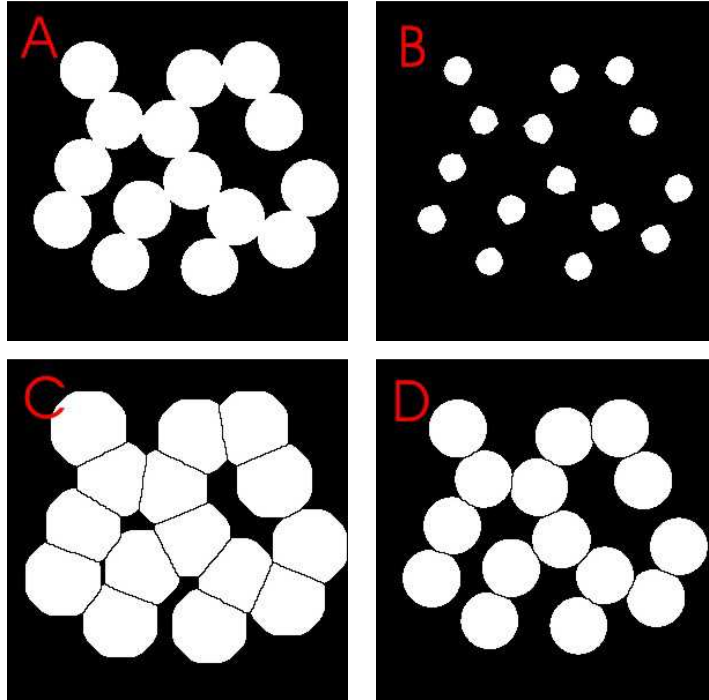


Figure 15. Separation of touching regions using morphological and Boolean operations. (A) Input image; (B) after a few cycles of erosion; (C) some cycles of dilation applied to B image using logic to prevent merging of regions; (D) AND Boolean operator between images A and C.

When the objects in the scene are all similar in size, as it might happen working with seeds, it is possible to apply some cycles of erosion until all of them are separated but not completely erased (image B). Afterwards, a few dilation operations grow the objects making them bigger than their original size. Logical operations are imposed to prevent that the objects will merge again (image C). Employing an AND Boolean operator between the input image and the image in which the objects are separated, a new image with the original objects separated is produced (image D).

Because erosion and dilation cause respectively a reduction and increasing in the size of objects, and for this reason they are sometimes known as etching and plating or shrinking and growing, there are several rules to adjust these operations. Particularly it is possible to adjust the

neighbour pattern, that allows to set the direction (horizontal, vertical or both) of erosion or dilation, and the number of iterations, also called depth of the operation, that roughly corresponds to the distance that boundaries will grow or shrink radially (Grillo, 2009).

Two similar morphological operators that can be used in similar conditions, but in different orders, are *opening* and *closing* operations. Generally, the first is helpful to enlarge pixel holes or coves within the regions of interest, while the closing operator is used to close up breaks in objects (Fig. 16).

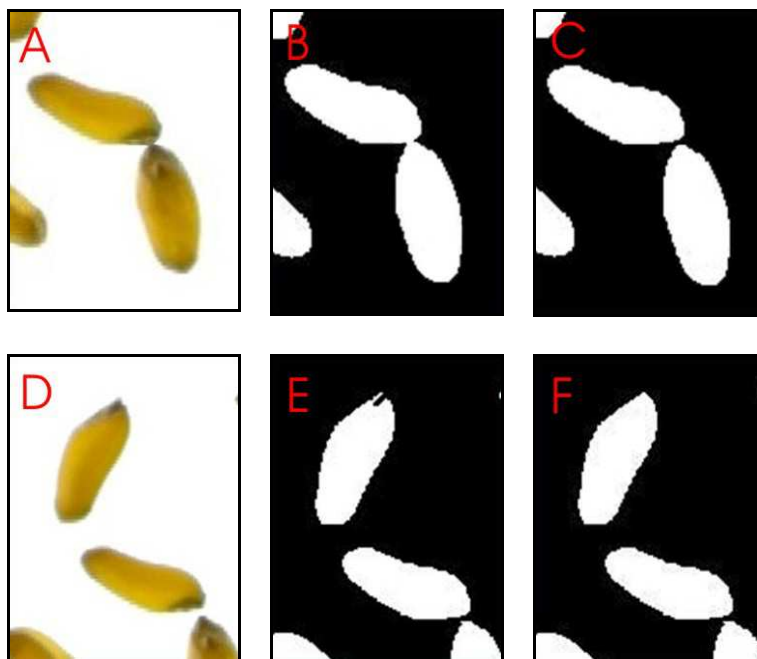


Figure 16. Opening operation to separate touching seeds (images A, B and C); and closing operation to fill a cove in a seed (images D, E and F).

Finally, the morphological operators *fill* and *scrap*, allows exclusively to fill holes in the objects and to erase spots or noise in the binary image, on the bases of dimension settings.

A particular morphological operator is the *skeletonization*. It could be described as a forced erosion of an object, applied up to make it one pixel thick (Pavlidis, 1980; Nevatia & Babu, 1980; Davidson, 1991; Lam *et al.*, 1992; Ritter & Wilson, 2001). This function is often used to obtain the length of an object (Fig. 17).

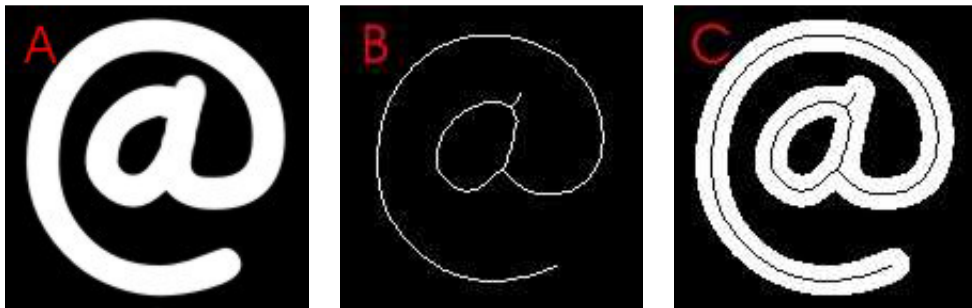


Figure 17. Skeletonization of an object.

Just as the skeleton of objects may be determined in an image, it is also possible to skeletonize the background. Indeed, considering equidistant points from objects boundaries, this operation effectively divides the image into regions around each object (Serra, 1982). This is a very common morphological operation because it is often used, in combination with erosion and dilation operations, to separate objects with different size and shape (Fig. 18). Moreover, modifying the setting parameters, the skeletonization may be helpful to identify object contours (Grillo, 2009).

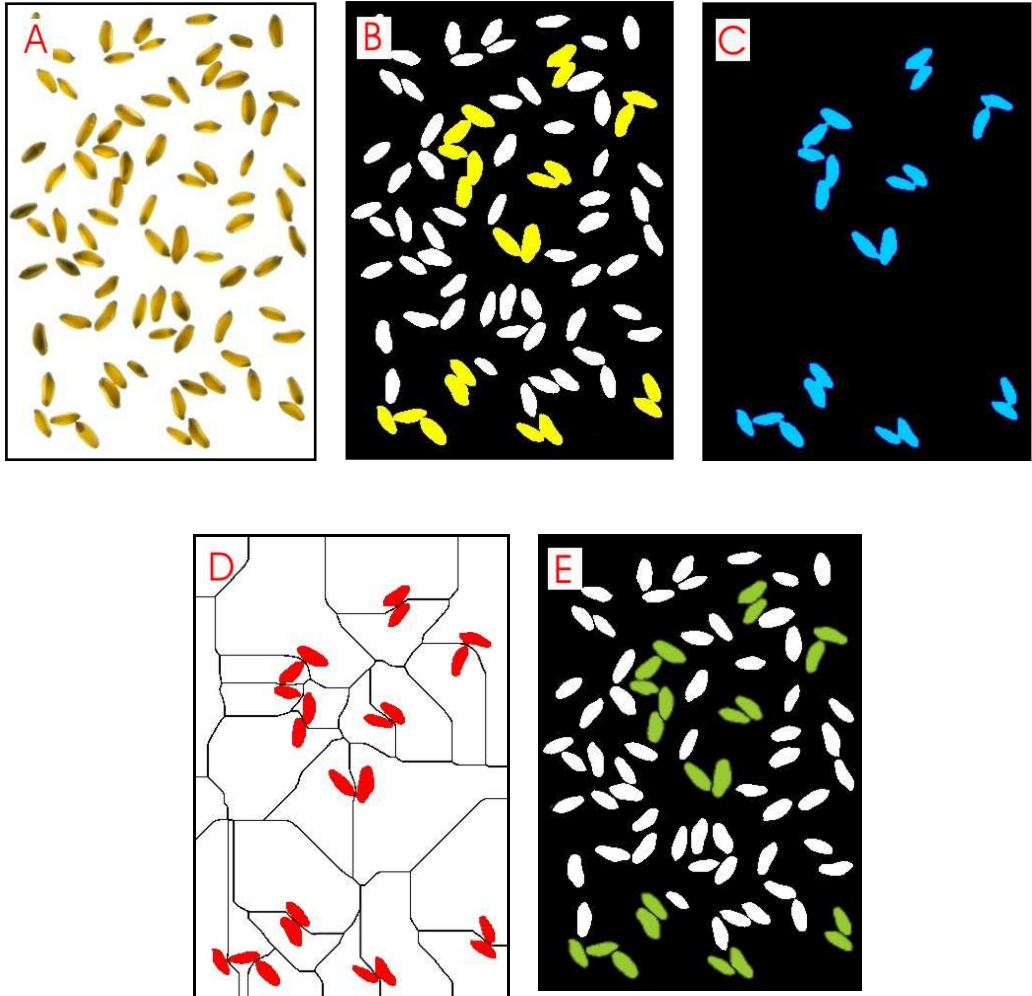


Figure 18. Background skeletonization to separate touching seeds.
(A) Original image; (B) segmentation and selection of touching seeds; (C) isolation of touching seeds; (D) background skeletonization of image C; (E) image with all separated seeds (Grillo, 2009).

References

- ABDULLAH M.Z. 2008. Image acquisition systems, 3-35. In: *Computer Vision Technology for Quality Evaluation*. Edited by Da-Wen Sun, Published by Elsevier Academic Press, San Diego, CA, USA, 583 pages, 2008, ISBN: 978-0-12-373642-0, 0-12-373642-0.
- ABDULLAH M.Z., ABDUL-AZIZ S., DOS-MOHAMED A. M. 2000. Quality inspection of bakery products using color-based machine vision system. *Journal of Food Quality* 23, 39-50.
- BACCHETTA G., FENU G., GRILLO O., MATTANA E., VENORA, G. 2011b. Identification of Sardinian species of *Astragalus* section *Melanocercis* (Fabaceae) by seed image analysis. *Annales Botanici Fennici* 48, 449-454.
- BACCHETTA G., GARCÍA P.E., GRILLO O., MASCIA F., VENORA G. 2011a. Seed image analysis provides evidence of taxonomical differentiation within the *Lavatera triloba* aggregate (Malvaceae). *Flora* 206, 468-472.
- BACCHETTA G., GRILLO O., MATTANA E., VENORA G. 2008. Morpho-colorimetric characterization by image analysis to identify diaspores of wild plant species. *Flora* 203, 669-682.
- BLASCO J., ALEIXOS N., GOMEZ J., MOLTÓ E. 2007. *Citrus* sorting by identification of the most common defects using multispectral computer vision. *Journal of Food Engineering* 83, 384-393.
- BLASCO J., ALEIXOS N., ROGER J.M., RABATEL G., MOLTÓ E. 2002. Automation and Emerging Technologies: Robotic Weed Control using Machine Vision. *Biosystems Engineering* 83, 149-157.
- CASTLEMAN K. R. 1978. *Digital Image Processing*. 1st edition - Published by Prentice Hall. Upper Saddle River, New Jersey, 429 pages.
- DALAY W. & BRITTON D. 2003. Vision-based quality control in poultry processing, 243-258. In: *Machine Vision for the Inspection of Natural Products*. Edited by Mark Graves and Bruce Batchelor, Published by Springer-Verlag London, UK, 471 pages, ISBN: 1-85233-525-4.
- DAVIDSON J. 1991. Thinning and skeletonization: a tutorial and overview. In: *Digital image processing: fundamentals and applications*. Edited by E. Dougherty, Published by Marcel Dekker, New York, USA.
- FREEMAN C.L., SZELISKI W.T., SING BING KANG R. 2006. Noise Estimation from a Single Image. *Computer Vision and Pattern Recognition* 1, 901-908.

- FREIXENET J., MUÑOZ X., RABA D., MARTÍ J., CIUFÍ X. .2002. Yet another survey on image segmentation: region and boundary information integration. *Lecture Notes in Computer Science* 2352, 21-25.
- FRUCCI M. & SANNITI DI BAJA G. 2008. From segmentation to binarization of gray-level Images. *Journal of Pattern Recognition Research* 1, 1-13.
- FU K.S. & MUI J.K. 1981. A survey of image segmentation. *Pattern recognition* 13, 3-16.
- FUKANAGA K. 1990. *Introduction to statistical pattern recognition*. 2nd edition - Published by Academic Press, San Diego, 591pages, ISBN: 0-12-269851-7.
- GONZALEZ R.C. & WOODS R.E. 2007. *Image Digital Processing* (Book review). 3rd edition - Published by Prentice Hall. Upper Saddle River, New Jersey, 954 pages, ISBN: 978-0-13-168728-8, 0-13-168728-X.
- GRILLO O. 2009. Germplasm morpho-colorimetric characterization by image analysis and statistical classification of the most representative families of Mediterranean vascular flora (*Doctoral dissertation*).
- GRILLO O., MATTANA E., FENU G., VENORA G., BACCHETTA, G. 2013. Geographic isolation affects inter- and intra-specific seed variability in the *Astragalus tragacantha* complex, as assessed by morpho-colorimetric analysis. *Comptes Rendus de Biologies* 336, 102-108.
- GRILLO O., MATTANA E., VENORA G., BACCHETTA, G. 2010. Statistical seed classifiers of 10 plant families representative of the Mediterranean vascular flora. *Seed Science and Technology* 38, 455-476.
- HARALICK R.M., SHANMUGAM K., DINSTEN I. 1975. Textural features for image classification. *IEEE Transactions on Systems, Man and Cybernetics* 3, 610-621.
- HARALICK R.M. & SHAPIRO L.G. 1988. Segmentation and its place in machine vision. *Scanning Microscopy* 2, 39-54.
- HIRANO A., MADDEN M., WELCH R. 2003. Hyperspectral image data for mapping wetland vegetation. *Wetlands* 23, 436-448.
- HU Y., RAVALLI C., VENORA G., SCHMIDHALTER U. 2006. Salt tolerance of wheat is associated with the number and size of leaf vessels. *Proceedings of the 50th Italian Society of Agricultural Genetics Annual Congress*. Ischia (Italy), September 10th/14th, 2006.
- HUI Z., CHAN F.H.Y., LAM F.K. 1999. Image contrast enhancement by constrained local histogram equalization. *Computer Vision and Image Understanding*. 73, 281-290.

- JAIN A.K. 1989. *Fundamentals of digital image processing*. Published by Englewood Cliffs: Prentice-Hall, (Prentice Hall Information and system science series) México, 569 pages, 1989.
- KIM J.Y., KIM L.S., HWANG S.H. 2001. An advanced contrast enhancement using partially overlapped sub-block histogram equalization. *Circuits and Systems for Video Technology* 11, 475-484.
- KITTLER J., ILLINGWORTH J., FOGLEIN J. 1985. Threshold selection based on a simple image statistic. *Computer Vision, Graphic and Image Processing* 30, 125-147.
- LAM L., LEE S., SUEN C. 1992. Thinning methodologies - A comprehensive survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 14, 868-885.
- LEE H.C. 2005. Introduction to color imaging science. Published by Cambridge University Press. Cambridge, UK, 695 pages, 2005, ISBN: 978-0-521-84388-X.
- LEE J.S. 1983. Digital image smoothing and the sigma filter. *Computer Vision, Graphics, and Image Processing* 24, 255-269.
- LING P.P., GIACOMELLI G.A., RUSSELL T. 1996. Monitoring of plant development in controlled environment with machine vision. *Advances in Space Research* 18, 101-112.
- MALIK J., BELONGIE S., LEUNG T., SHI J. 2001. Contour and Texture Analysis for Image Segmentation. *International Journal of Computer Vision* 43, 7-27.
- MASTIN G.A. 1985. Adaptive filters for digital image noise smoothing: an evaluation. *Computer Vision, Graphics and Image Processing* 31, 103-12.
- MATTANA E., GRILLO O., VENORA G., BACCHETTA G. 2008. Germplasm image analysis of *Astragalus maritimus* and *A. verrucosus* of Sardinia (subgen. *Trimeniaeus*, *Fabaceae*). *Anales Jardin Botanico de Madrid* 65, 149-155.
- MOLTÓ GARCIA E. & BLASCO J. 2009. Machine vision systems for row material inspection, 84-126. In: *Optical Monitoring of Fresh and Processed Agricultural Crops*. Edited by Manuela Zude, Published by CRC Press. Boca Raton, FL, USA, 537 pages, 2009, ISBN: 978-1-4200-5402-6.
- NEVATIA R. & BABU K. 1980. Linear feature extraction and description. *Computer Vision, Graphics and Image Processing* 13, 257-268.
- OTSU N. 1979. A threshold selection method from gray level histograms. *IEEE Transactions on Systems, Man and Cybernetics* 9, 62-69.
- PAVLIDIS T. 1980. A thinning algorithm for discrete binary images. *Computer Vision, Graphics and Image Processing* 13, 142-157.

- PEARSON T. 1996. Machine vision system for automated detection of stained pistachio nuts. *Lebensmittel Wissenschaft und Technologie* 29, 203-209.
- PRATT W.K. 2007. *Digital Image Processing*. 4th edition - Published by Wiley and Sons, New Jersey, 782 pages, ISBN: 978-0-471-76777-0.
- PREWITT J.M.S. & MENDELSON M.L. 1966. The analysis of cell images. *Annals of the New York Academy of Sciences* 128, 1035-1053.
- RANK K., LENDL M., UNBEHAUEN R. 1999. Estimation of image noise variance. *IEEE Proceedings Vision, Image and Signal Processing*, 146, 80-84.
- RIGAUT J.P. 1988. Automated image segmentation by mathematical morphology and fractal geometry. *Journal of Microscopy* 150, 21-30.
- RITTER X. & WILSON J.N. 2001. *Handbook of computer vision algorithms in image algebra*. 2nd edition - Published by CRC Press. Boca Raton, FL, USA, 431 pages, ISBN: 0-8493-0075-4.
- RIVA M. 2004. *Image analysis*. Published online at http://users.unimi.it/~distam/info/image_0_file/frame.htm.
- ROSENFELD A. & KAK A.C. 1982. *Digital picture processing*, Vol. 1-2. 2nd edition - Published by Academic Press, San Diego, 435 pages, ISBN: 0-125-97301-2, 978-0125973014.
- RUSS J.C. 1995. Thresholding images. *Journal of Computer Assisted Microscopy* 7, 41-164.
- RUSS J.C. 2007. *The Image Processing Handbook*. 5th edition - Published by CRC Press. Boca Raton, FL, USA, 376 pages, 2007, ISBN: 978-0-84-937073-1, 0-84-937073-6.
- SÁNCHEZ DEL ÁLAMO C., SALCEDO GARCÍA-CALVO V., GRILLO O., BOUSO V., SARDINERO S., FERNÁNDEZ-GONZÁLEZ F. 2008. Estudio taxonómico del abedul de los montes de Toledo. *II Congreso de Naturaleza de la Provincia de Toledo*. Toledo (Spain), September 23rd/26th 2008.
- SCOTT D.W. 1979. On optimal and data-based histograms. *Biometria* 66, 605-610.
- SERRA J. 1982. *Image analysis and mathematical morphology*. Published by Academic Press. London, UK, 610 pages, ISBN: 0-12-637240-3.
- SEZGIN M. & SANKUR B. 2004. Survey over image thresholding techniques and quantitative performance evaluation. *Journal of Electronic Imaging* 13, 146-165.
- SHAHIN M.A. & SYMONS S.J. 2000. Color calibration of scanner for scanner-independent grain grading. *Cereal Chemistry* 80, 285-289.
- SHAHIN M.A. & SYMONS S.J. 2001. A machine vision system for grading lentils. *Canadian Biosystems Engineering* 43, 7.7-7.14.

- SHAPIRO L.G. & STOCKMAN G.C. 2001. *Computer Vision*. 1st edition - Published by Prentice Hall. Upper Saddle River, New Jersey, 608 pages, ISBN: 0-130-30796-3, 978-0130307965
- STUDMAN C. & OUYANG L. 1997. Bruise measurement by image analysis. In *Proceedings of V Symposium on Fruit, Nut and Vegetable Production Engineering*. Davis, CA, 1-7.
- TAO Y., HEINEMANN P.H., VARGHESE Z., MORROW C.T., SOMMER III H.J. 1995. Machine vision for color inspection of potatoes and apples. *Transaction of the ASAE* 38, 1555-1561.
- VENORA G. & CALCAGNO F. 1991. Study of stomatal parameters for selection of drought resistant varieties in *Triticum durum* Desf. *Euphytica* 57, 275-283.
- VENORA G., GRILLO O., RAVALLI C., CREMONINI R. 2009a. Identification of Italian landraces of beans (*Phaseolus vulgaris* L.), using an image analysis system. *Scientia Horticulturae* 121, 410-418.
- VENORA G., GRILLO O., SACCONI R. 2009b. Durum wheat storage centers: evaluation of vitreous, starchy and shrunken kernels by image analysis system. *Journal of Cereal Science* 49, 429-440.
- VENORA G., GRILLO O., SHAHIN M.A., SYMONS S.J. 2007. Identification of Sicilian landraces and Canadian cultivars of lentil using image analysis system. *Food Research International* 40, 161-166.
- VENORA G. & PORTA-PUGLIA A. 1993. Observation on outer cell layers of stem in chickpea cultivars susceptible and resistant to *Ascochyta* blight. *Petria* 3, 177-182.
- VENORA G., SACCONI R., GRILLO O., ORLANDO A. 2009c. Stima della resa in semola mediante tecniche di analisi d'immagine. *Tecnica Molitoria* 60, 399-410.
- WESZKA J.S. 1978. A survey of threshold selection techniques. *Computer Graphic Image Processing* 7, 259-265.
- XIANG DU J., WANG X., ZHANG G. 2007. Leaf shape based plant species recognition. *Applied Mathematics and Computation* 185, 883-893.
- YOUNG I.T., GERBRANDS J.J., VAN VLIET L.J. 1995. *Fundamental of Image Processing*. Published by Delft University of Technology, ND, 111 pages, 1995, ISBN: 90-75691-01-7.
- ZAPOTOCZNY P., ZIELINSKA M., NITA Z. 2008. Application of image analysis for the varietal classification of barley: Morphological features. *Journal of Cereal Science* 48, 104-110.

- ZHENG C. & SUN D.W. 2008. Image segmentation techniques, 37-56. In: *Computer Vision Technology for Quality Evaluation*. Edited by Da-Wen Sun, Published by Elsevier Academic Press, San Diego, CA, USA, 583 pages, 2008, ISBN: 978-0-12-373642-0, 0-12-373642-0.
- ZHENG C. & SUN D.W. 2009. Computer vision for quality control, 126-142. In: *Optical Monitoring of Fresh and Processed Agricultural Crops*. Edited by Manuela Zude, Published by CRC Press. Boca Raton, FL, USA, 537 pages, 2009, ISBN: 978-1-4200-5402-6.

A new set of seed features: Elliptic Fourier Descriptors and Haralick's parameters

Introduction

The discriminant ability of the identification system depends not only on the intra-specific representativeness of analyzed *taxa*, but also, on the quality and quantity of the parameters measured and used to differentiate among groups. For this reason, the *Elliptic Fourier Descriptors* (Iwata *et al.*, 2002, 2004; Kawabata *et al.*, 2009; Yoshioka *et al.*, 2004; Orrù *et al.*, 2012; 2013) for a detailed description of the shape, and the *Haralick's parameters*, evaluating the surface texture of seeds (Diamond *et al.*, 2004; Gerger & Smolle, 2004; Nanni *et al.*, 2010), were considered as variables and included in the statistical classification system, in addition to the common seeds morpho-colorimetric traits used in previous similar works (Bacchetta *et al.*, 2008a; 2011a; Grillo *et al.*, 2010, 2013).

Shape measurements: Elliptic Fourier Descriptors (EFDs)

Fourier descriptors (FDs) are very popular shape descriptors and are mainly used in a 2D shape description context. The general idea is to create a mono-dimensional function from a bi-dimensional boundary contour of, for example, the exterior shape of a seed: the shape signature. An example is the centroid distance signature which is a (periodic) function that represents the distance from the boundary to the centroid of the image. This shape signature can be approximated by a Fourier series, where the obtained coefficients are called Fourier descriptors. The larger the set of derived descriptors, the better

the accuracy for shape retrieval will be. The advantages of FDs are (1) their low complexity, (2) each descriptor has a physical meaning, (3) they can easily be normalised and (4) they describe shape features at all scales (Zhang & Lu, 2004, 2005).

Some variations on FDs exist. An example are *Elliptic Fourier descriptors* hereafter EFDs, introduced by Kuhl & Giardina (1982). EFDs describe a closed contour with a series of rotating phasors with elliptical loci; the contour is hence represented as a set of harmonically related ellipses. Elliptical Fourier analysis removes the following three limitations encountered in conventional Fourier analysis: (1) the sampled interval has to be equally divided, (2) the descriptors depend on the chosen coordinate system and (3) the difficulty of dealing with outlines that curve back on themselves (Lestrel, 1989). The drawback of EFDs is that many descriptors have to be used, as each harmonic (ellipse) consists of four descriptors. However, this is not really of concern because the harmonics are computed fairly easy and fast. EFDs have often been used for describing shape variation of biological products, ranging from rice to petals or a stallions sperm heads (Iwata *et al.* 2010; Kawabata *et al.*, 2009; Severa *et al.*, 2010, Rogge *et al.*, 2014)

The main advantage of the EFDs (Crosgriff, 1960; Fritzsche, 1961; Raudseps, 1965; Borel, 1965), is invariance to translation, rotation and scaling of the observed object. Thus the shape description becomes independent of the relative position and size of the object in the input image. In this way, the distance between camera and placement of the object relative to the optical axis of image acquisition system not affects values of the Fourier descriptors.

This method, fundamentally, do not define the shape of the object but allows description of the boundary of the seed projection as an array of complex numbers which correspond to the pixel positions on the seed boundary. So, from the seed apex, defined as the starting point in a Cartesian system, chain codes are generated. A chain code is a lossless compression algorithm for binary images. The basic principle of chain codes is to separately encode each connected component (pixel) in the image. The encoder then moves along the boundary of the image and, at each step, transmits a symbol representing the direction of this movement. This continues until the encoder returns to the starting position (Fig. 1).

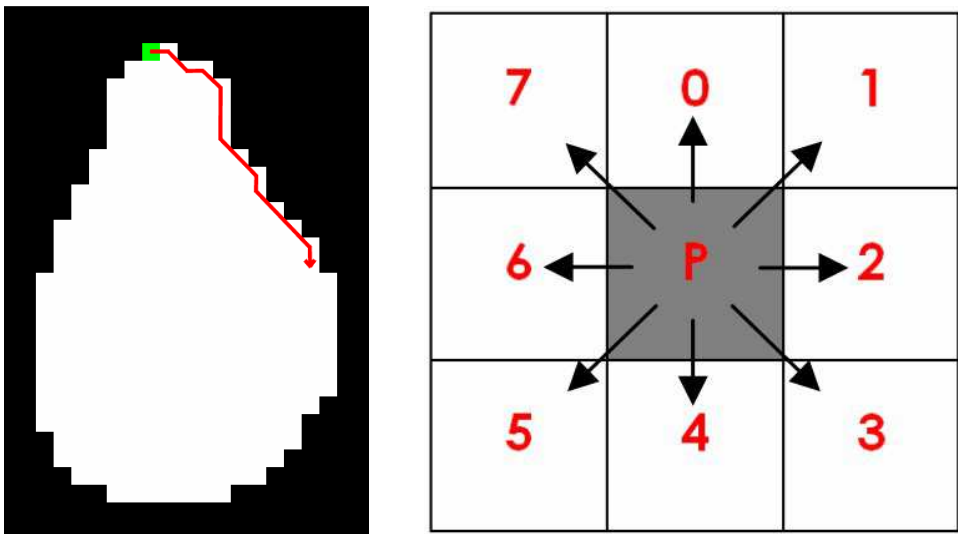


Figure 1. Chain code generation (e.g. 23234443343334...).

Mathematical basis of EFDs method

Extensive description on elliptic Fourier feature analysis is found in Kuhl & Giardina (1982). Briefly, an object's boundary is approximated by Fourier series expansions of the boundary that is first transformed into a time series of a function that arises through chain coding (Freeman, 1974) of the path at constant speed around the object in which individual chain links are at the pixel-to-pixel level, taking on integer values between 0 and 7.

These values represent the direction of movement from one pixel to the next, either being horizontal (0 or 4 for + or - x direction, respectively), vertical (2 or 6 for + and - y direction), or diagonal (1, 3, 5, or 7 for direction intermediate between the corresponding even numbers). The approximations of the x and y positions of the object's boundary are given as truncated series expressions as follows:

$$X_N(t) = A_0 + \sum_{n=1}^N a_n \cos\left(\frac{2n\pi t}{T}\right) + b_n \sin\left(\frac{2n\pi t}{T}\right)$$
$$Y_N(t) = C_0 + \sum_{n=1}^N c_n \cos\left(\frac{2n\pi t}{T}\right) + d_n \sin\left(\frac{2n\pi t}{T}\right)$$

where n is the number of harmonics (N total) and t is the time along the chain path of period T . As $N \rightarrow \infty$, these expressions become $x(t)$ and $y(t)$. Mathematical solution of the expressions for coefficients a_n , b_n , c_n , and d_n are determined by writing the time derivatives of $x(t)$ and $y(t)$ as Fourier series expressions and equating the coefficients with corresponding coefficients from the derivatives of $x(t)$ and $y(t)$.

This yields the following:

$$\begin{aligned}
 a_n &= \frac{T}{2n^2\pi^2} \sum_{p=1}^K \frac{\Delta x_p}{\Delta t_p} \left[\cos\left(\frac{2n\pi t_p}{T}\right) - \cos\left(\frac{2n\pi t_{p-1}}{T}\right) \right] \\
 b_n &= \frac{T}{2n^2\pi^2} \sum_{p=1}^K \frac{\Delta x_p}{\Delta t_p} \left[\sin\left(\frac{2n\pi t_p}{T}\right) - \sin\left(\frac{2n\pi t_{p-1}}{T}\right) \right] \\
 c_n &= \frac{T}{2n^2\pi^2} \sum_{p=1}^K \frac{\Delta y_p}{\Delta t_p} \left[\cos\left(\frac{2n\pi t_p}{T}\right) - \cos\left(\frac{2n\pi t_{p-1}}{T}\right) \right] \\
 d_n &= \frac{T}{2n^2\pi^2} \sum_{p=1}^K \frac{\Delta y_p}{\Delta t_p} \left[\sin\left(\frac{2n\pi t_p}{T}\right) - \sin\left(\frac{2n\pi t_{p-1}}{T}\right) \right]
 \end{aligned}$$

where p is an index that corresponds to each pixel along the boundary contour, Δx_p and Δy_p are the changes in the x and y projections of the chain between the p^{th} and $(p-1)^{\text{th}}$ positions, having possible values of -1, 0, and 1. The time variables t_{p-1} and t_p represent the total time needed to reach the $(p-1)^{\text{th}}$ and p^{th} positions from an arbitrary starting point on the boundary, with Δt_p being the time increment between these points and noting that its value is either 1 or $\sqrt{2}$. Values for the coefficients depend on the starting point of the path of the contour and are therefore difficult to use when comparing objects of different orientations.

At the expense of absolute dimensional information, the series functional expressions may be normalized such that the shapes can be compared among images.

A common procedure is to normalize and align the object with respect to the first harmonic ellipse (Kuhl & Giardina, 1982; Yoshioka *et al.*, 2004; Neto *et al.*, 2006; Mebatsion *et al.*, 2012). Setting aside the translation-determining constant additive terms, A_0 and C_0 , and expressing X_N and Y_N in

phasor notation, the first harmonic phasor is rotated into alignment with the semi-major axis of its locus, whereupon the starting point of the contour is phase shifted to coincide with a maximum value, as determined by setting the derivative of the magnitude of the first harmonic phasor equal to zero. This yields:

$$\theta = \frac{1}{2} \tan^{-1} \left[\frac{2(a_1 b_1 + c_1 d_1)}{a_1^2 + c_1^2 - b_1^2 - d_1^2} \right]$$

so that

$$\begin{bmatrix} a_1^* & c_1^* \\ b_1^* & d_1^* \end{bmatrix} = \begin{bmatrix} \cos \theta & \sin \theta \\ -\sin \theta & \cos \theta \end{bmatrix} \begin{bmatrix} a_1 & c_1 \\ b_1 & d_1 \end{bmatrix}$$

and the spatial rotation ψ is obtained from

$$\psi = \tan^{-1} \frac{c_1^*}{a_1^*}$$

Finally the object is made independent of its size by dividing the coefficients by the magnitude of the semi –major axis, E^*

$$E^* = (a_1^{*2} + c_1^{*2})^{1/2}$$

The standardized coefficients become

$$\begin{bmatrix} a_n^{**} & b_n^{**} \\ c_n^{**} & d_n^{**} \end{bmatrix} = \frac{1}{E^*} \begin{bmatrix} \cos \psi & \sin \psi \\ -\sin \psi & \cos \psi \end{bmatrix} \begin{bmatrix} a_n & b_n \\ c_n & d_n \end{bmatrix} \begin{bmatrix} \cos n\theta & -\sin n\theta \\ \sin n\theta & \cos n\theta \end{bmatrix}$$

Iwata *et al.* (1998) proposed that coefficients a_n^{**} and d_n^{**} define the symmetrical variations, while b_n^{**} and c_n^{**} define the asymmetrical variations. By way of example, the boundary of a kernel is displayed in

Figure 2, in which the contours arising from the series solutions at one and ten harmonics are included. Also included is the ellipse of the equivalent second central moment with its minor and major axes from the morphological properties analysis.

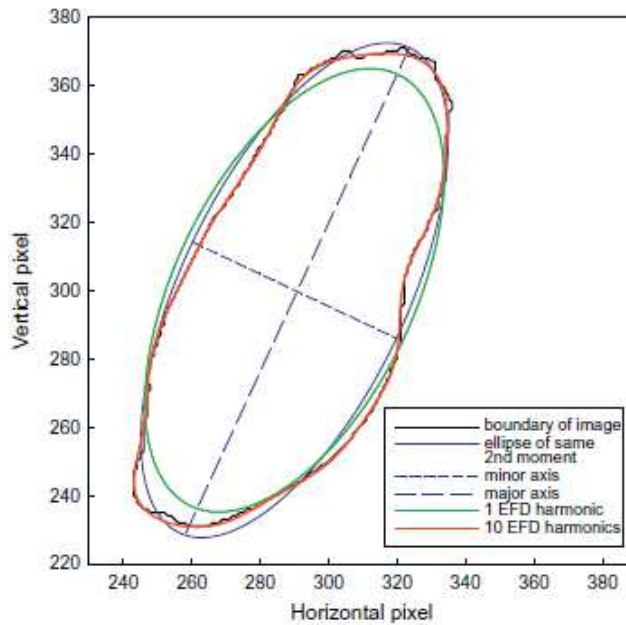


Figure 2. Trace of the boundary of a seed. Also shown are the ellipse, with minor and major axes, of the equivalent second central moment as calculated in morphological property determination and the elliptic Fourier series contour functions at one and ten harmonics.

According to Terral *et al.* (2010), about the use of a number of harmonics for an optimal description of seed outlines, in order to minimize the measurement errors and to optimize the efficiency of shape reconstruction, 20 harmonics were used, in this study, in order to define the seed boundaries, obtaining a further 78 parameters useful to discriminate among the studied seeds (Orrù *et al.* 2012, 2013) (Fig. 3).

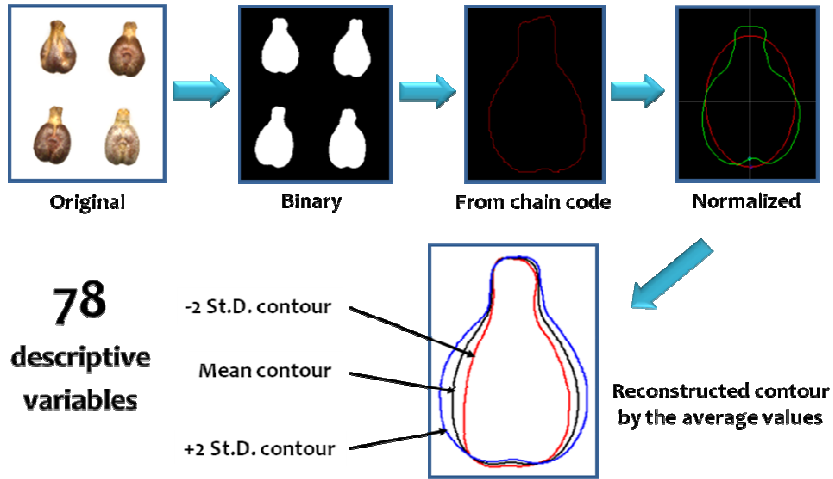


Figure 3. EFDs procedure: from a digitalized image to the chain codes.

Texture evaluation: Haralick's descriptors

The *texture* is series of surface features that gives important information about the densitometry and color distribution of an object.

Although, for the human vision the interpretation of chromaticity changes is easy and natural, *Haralick's descriptors* measured with image analysis system, allow to define mathematically, and so in a quantitative way, the color on a surface, identifying and describing any areas for distribution, intensity and/or homogeneity, characteristic variables of a particular group.

In 1973 Haralick introduced the co-occurrence matrix and texture features for automated classification of rocks into six categories (Haralick & Shanmugam, 1973). Today, these features are widely used representing a popular approach for the analysis and classification of many medical images (Fig. 4), including breast masses and tumors seen in mammograms, diagnosing diseases related to skin (Mittra & Parek, 2001), carotid artery (Hassan *et al.*, 2012), liver (Lee *et al.*, 2007), brain (Dhanalaskshmi &

Rajamani, 2013; Jafarpour *et al.*, 2012; Zulpe *et al.*, 2012), abdomen (Mitrea *et al.*, 2011), and breast (Mclaren *et al.*, 2009) and for microscope images of biological cells too (Harder *et al.*, 2006; Conrad *et al.*, 2004; Sivaramakrishna *et al.*, 2002; Bovis & Singh, 2000; Gupta & Markey, 2005; Lee *et al.*, 2006).

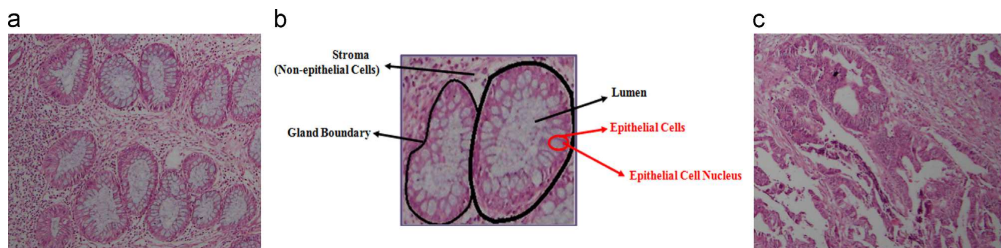


Figure 4. Microscopic images of (a) normal and (c) malignant colon biopsy samples, and (b) regular structure of normal colon tissue.

On the contrary, few results are reported in the literature about this kind of studies on seeds of both agronomical and wild species for tassonomical purposes (Fig. 5) (Diamond *et al.*, 2004; Gerger & Smolle, 2004; Nanni *et al.*, 2010).

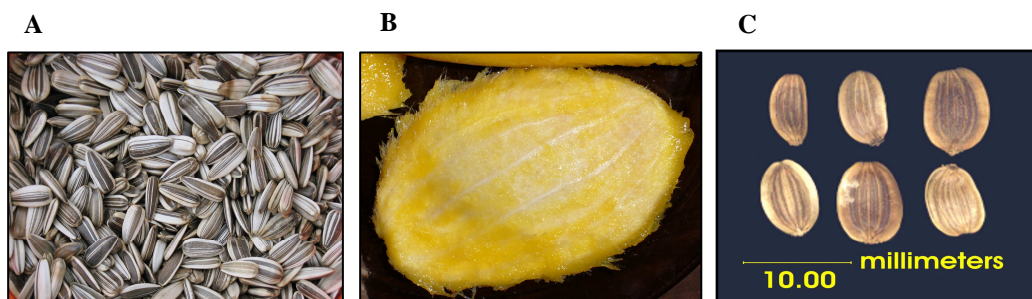


Figure 5. Texture in (A) black striped white seed of sunflower; (B) wet yellow seed of mango; (C) *Ferula arrigoni* seed.

One drawback of the features is the relatively high costs for computation. However, it is possible to speed up the computation using general-purpose graphics processing units (GPUs). Nowadays, GPUs (ordinary computer graphics cards) are more and more used to accelerate graphical as well as non-graphical software by highly parallel execution (Gipp *et al.*, 2009).

Texture analysis, using some or all of the 14 texture features proposed by Haralick & Shanmugam (1973), is based on the spatial gray level dependence (SGLD) matrices, which encapsulates the spatial relationship between pixels of an image. The relationship may be specified in two ways: (1) horizontal and vertical distance of neighbors with the pixel of interest; (2) the spatial relationship between pixel of interest and neighbors lying at various orientations e.g. $\theta = 0^\circ, 45^\circ, 90^\circ, 135^\circ$ (Fig. 6).

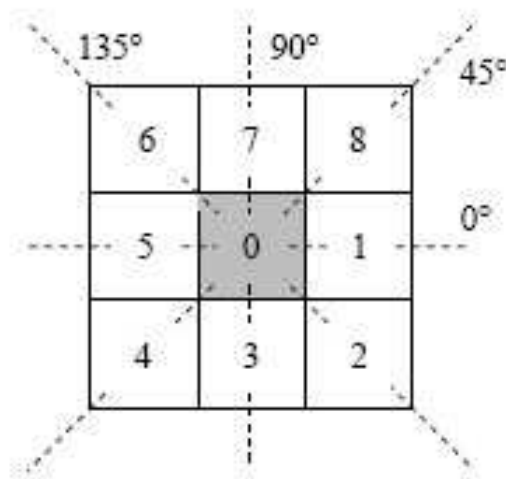


Figure 6. Resolution cells 1 and 5 are 0° (horizontal) nearest neighbors to resolution cell 0; resolution cells 2 and 6 are 135° nearest neighbors; resolution cells 3 and 7 are 90° nearest neighbors; and resolution cells 4 and 8 are 45° nearest neighbors to 0.

In this study, in addition to EFDs, the morpho-colorimetric pattern of feauters recorded for each seed was further improved adding algorithms able to compute 11 Haralick's descriptors and the relative standard deviations for each analyzed seed.

The evaluation of texture, tone and context allows to define the spatial distribution of the image intensities and discrete tonal features. When a small area of the image has little variation of discrete tonal features, the dominant property of that area is grey tone. When a small area has wide variation of discrete tonal features, the dominant property of that area is texture (Haralick and Shapiro, 1991).

According to Haralick *et al.* (1973), the concept of tone is based on varying shades of grey of resolution cells in a photographic image, while texture is concerned with the spatial (statistical) distribution of grey tones. Texture and tone are not independent concepts; rather, they bear an inextricable relationship to one another very much like the relationship between a particle and a wave. Context, texture and tone are always present in the image, although at times one property can dominate the others.

The basis for these features is the gray-level co-occurrence matrix (G in equation 1). This matrix is square with dimension N_g , where N_g is the number of gray levels in the image. Element $[i,j]$ of the matrix is generated by counting the number of times a pixel (p) with value i is adjacent to a pixel with value j and then dividing the entire matrix by the total number of such comparisons made. Each entry is therefore considered to be the probability that a pixel with value i will be found adjacent to a pixel of value j .

$$G = \begin{bmatrix} p(1,1) & p(1,2) & \dots & p(1, N_g) \\ p(2,1) & p(2,2) & \dots & p(2, N_g) \\ \vdots & \vdots & \ddots & \vdots \\ p(N_g, 1) & p(N_g, 2) & \dots & p(N_g, N_g) \end{bmatrix} \quad (1)$$

In Table 1, the 11 Haralick's descriptors measured on each seed to mathematically describe the surface texture, are reported.

Table 1. Haralick's descriptors measured as reported in Haralick *et al.* (1973).

	<i>Feature</i>	<i>Equation</i>
Har 1	Angular second moment	$\sum_i \sum_j p(i, j)^2$
Har 2	Contrast	$\sum_{n=0}^{N_g-1} n^2 \left\{ \sum_{i=1}^{N_g} \sum_{j=1}^{N_g} p(i, j) \right\}, i, j = n$
Har 3	Correlation	$\frac{\sum_i \sum_j (ij) p(i, j) - \mu_x \mu_y}{\sigma_x \sigma_y}$
		where μ_x , μ_y , σ_x and σ_y are the means and the standard deviations of p_x and p_y .
Har 4	Sum of square: variance	$\sum_i \sum_j (i - \mu)^2 p(i, j)$
Har 5	Inverse difference moment	$\sum_i \sum_j \frac{1}{1 + (i - j)^2} p(i, j)$
Har 6	Sum average	$\sum_{n=2}^{2N_g} i p_{x+y}(i)$
		where x and y are the coordinates (row and column) of an entry in the co-occurrence matrix, and $p_{x+y}(i)$ is the probability of co-occurrence matrix coordinates summing to $x+y$.
Har 7	Sum variance	$\sum_{i=2}^{2N_g} (i - f_B)^2 p_{x+y}(i)$
Har 8	Sum entropy	$-\sum_{i=2}^{2N_g} p_{x+y}(i) \log\{p_{x+y}(i)\} = f_B$
Har 9	Entropy	$-\sum_i \sum_j p(i, j) \log[p(i, j)]$
Har 10	Difference variance	$\sum_{n=0}^{N_g-1} i^2 p_{x-y}(i)$
Har 11	Difference entropy	$-\sum_{n=0}^{N_g-1} p_{x-y}(i) \log\{p_{x-y}(i)\}$

References

- BOREL R.J. 1965. *A mathematica pattern recognition technique based on contour shape properties*. Ohio State University Research Foundation, Columbus. Rep. 1801-11, ASTIA AD 476-113.
- BOVIS K. & SINGH S. 2000. Detection of masses in mammograms using texture features. *Proceedings of the 15th International Conference on Pattern Recognition 7*, 267-270.
- CONRAD C., ERFLE H., WARNAT P., DAIGLE N., LÖRCH T., ELLENBERG J., PEPPERKOK R., EILS R. 2004. Automatic identification of subcellular phenotypes on human cell arrays, *Genome Research* 14, 130-1136.
- CROSGRIFF R.L. 1960. *Identification of shape*. Ohio State University Research Foundation, Columbus. Rep. 820-11, ASTIA AD 254-792.
- DHANALAKSHMI K. & RAJAMANI V. 2013. An intelligent mining system for diagnosing medical images using combined texture-histogram features, *International Journal of Imaging Systems and Technology* 23, 194-203.
- DIAMOND J., ANDERSON N.H., BARTELS P.H., MONTIRONI R., HAMILTON P.W. 2004. The use of morphological characteristics and texture analysis in the identification of tissue composition in prostatic neoplasia. *Human Pathology* 35, 1121-1131.
- FREEMAN H., 1974. Computer processing of line drawing images. *Computing Surveys* 6, 57-97.
- FRITZSCHE D.L. 1961. *A systematic method for characte recognition*. Ohio State University Research Foundation, Columbus. Rep. 1222-4, ASTIA AD 268-360.
- GERGER A. & SMOLLE J. 2003. Diagnostic Imaging of Melanocytic Skin Tumors. *Journal of Cutaneous Pathology* 30, 247-252.
- GIPP M., MARCUS G., HARDER N., SURATANEE A., ROHR K., KONIG R., MANNER R. 2012. Haralick's texture features computation accelerated by GPUs for biological applications. *Modeling, Simulation and Optimization of Complex Processes* 127-137.
- GIRILLO O. 2009. Germplasm morpho-colorimetric characterization by image analysis and statistical classification of the most representative families of Mediterranean vascular flora (*Doctoral dissertation*).

- GRILLO O., MATTANA E., FENU G., VENORA G., BACCHETTA G. 2013. Geographic isolation affects inter- and intra-specific seed variability in the *Astragalus tragacantha* complex, as assessed by morpho-colorimetric analysis. *Comptes Rendus de Biologies* 336, 102-108.
- GRILLO O., MATTANA E., VENORA G., BACCHETTA G. 2010. Statistical seed classifiers of 10 plant families representative of the Mediterranean vascular flora. *Seed Science and Technology* 38, 455-476.
- GUPTA S. & MARKEY M.K. 2005. Correspondence in texture features between two mammographic views. *Medical Physics* 36, 1598-1606.
- HARALICK R.M. & SHANMUGAM K. 1973. Computer Classification of Reservoir Sandstones. *IEEE Transactions on Geoscience Electronics* 11, 171-177.
- HARDER N., NEUMANN B., HELD M., LIEBEL U., ERFLE H., ELLENBERG J., EILS R., ROHR K. 2006. Automated recognition of mitotic patterns in fluorescence microscopy images of human cells, *Proc. IEEE Internat. Symposium on Biomedical Imaging: From Nano to Macro (ISBI'06)*, Arlington/VA, USA, pp. 1016-1019
- HASSAN M., CHAUDHRY A., KHAN A., KIM J.Y. 2012. Carotid artery image segmentation using modified spatial fuzzy means and ensemble clustering. *Computer Methods and Programs in Biomedicine* 108, 1261-1276.
- IWATA H., EBANA K., UGA Y., HAYASHI T., JANNINK J.L. 2010. Genome-wide association study of grain shape variation among *Oryza sativa* L. germplasms based on elliptic Fourier analysis. *Molecular Breeding* 25, 203-215.
- IWATA H., NESUMI H., NINOMIYA S., TAKANO Y., UKAI Y. 2002. Diallel analysis of leaf shape variations of *Citrus* varieties based on Elliptic Fourier Descriptors. *Breeding Science* 52, 89-94.
- 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.
- JAFARPOUR S., SEDGHI Z., AMIRANI M.C. 2012. A robust brain MRI classification with GLCM features. *International Journal of Computer Applications* 37, 1-5.
- KAWABATA S. YOKOO, M. NII K. 2009. Quantitative analysis of corolla shapes and petal contours in single-flower cultivars of *Lisianthus*. *Scientia Horticulturae* 121, 206-212.
- KUHL F.P. & GIARDINA C.R. 1982. Elliptic Fourier features of a closed contour. *Computer Graphics* 18, 259-278.

- LEE C.C., CHEN S.H., CHIANG Y.C. 2007. Classification of liver diseases from CT images using a support vector machine. *Journal of Advanced Computational Intelligence and Intelligent Informatics* 11, 396-402.
- LEE G.N., HARA T., FUJITA H. 2006. Classifying masses as benign or malignant based on co-occurrence matrix textures: a comparison study of different gray level quantizations. In: Astley SM, et al Eds. *International Workshop on Digital Mammography*. Manchester, UK, LNCS 4046, pp 332-339.
- LESTREL P.E. 1989. Method for analysing complex two-dimensional forms: elliptical Fourier functions. *American Journal of Human Biology* 1, 149-164.
- MCLAREN C.E., CHEN W.P., NIE K., SU M.Y. 2009. Prediction of malignant breast lesions from MRI features: a comparison of artificial neural network and logistic regression techniques. *Academic Radiology* 16, 842-851.
- MEBATSION H.K., PALIWAL J., JAYAS D.S., 2012. Evaluation of variations in the shape of grain types using principal components analysis of the Elliptic Fourier Descriptors. *Computers and Electronics in Agriculture* 80, 63-70.
- MITREA D., SOCACIU M., BADEA R., GOLEA A. 2011. Texture based characterization and automatic diagnosis of the abdominal tumors from ultrasound images using third order GLCM features. In: *Proceedings of the Fourth International Congress on Image and Signal Processing* pp.1558-1562.
- MITTRA A.K. & PAREKH J. 2001. Automated detection of skin diseases using texture features. *International Journal of Engineering, Science and Technology* 3, 4801-4808.
- NANNI L., SHI J.Y., BRAHNAM S. & LUMINI A. 2010. Protein classification using texture descriptors extracted from the protein backbone image. *Journal of Theoretical Biology* 264, 1024-1032.
- NETO J.C., MEYER G.E., JONES D.D., SAMAL A.K., 2006. Plant species identification using elliptic Fourier leaf shape analysis. *Computers and Electronics in Agriculture* 50, 121-134.
- ORRÙ M., GRILLO O. LOVICU G., VENORA G., BACCHETTA G. 2013. Morphological characterisation of *Vitis vinifera* L. seeds by image analysis and comparison with archaeological remains. *Vegetation History and Archaeobotany* 22, 231-242.
- ORRÙ M., GRILLO O., VENORA G., BACCHETTA G. 2012. Computer vision as a complementary to molecular analysis: grapevines cultivars case study. *Comptes Rendus de Biologies* 335, 602-615.

- RAUDSEPS J.G. 1965. *Some aspects of tangent-angle vs arc length representation of contours*. Ohio State University Research Foundation, Columbus. Rep. 1801-6, ASTIA AD 462-877.
- ROGGE S., BEYENE S.D., HERREMANS E., HERTOOG M.L., DEFRAEYE T., VERBOVEN P., NICOLAI B.M. 2014. A Geometrical Model Generator for Quasi-Axisymmetric Biological Products. *Food Bioprocessing Technology* 7, 1783-1792.
- SEVERA L., MÁCHAL L., ŠVÁBOVÁ L., MAMICA O. 2010. Evaluation of shape variability of stallion sperm heads by means of image analysis and Fourier descriptors. *Animal Reproduction Science* 119, 50-55.
- SIVARAMAKRISHNA R., POWEL K.A., LIEBER M.L., CHILCOTE W.A., SHEKHAR R. 2002. Texture analysis of lesions in breast ultrasound images. *Computerized Medical Imaging and Graphics* 26, 303-307.
- TERRAL J., TABARD E., BOUBY L., IVORRA S., PASTOR T., FIGUEIRAL I., PICQ S., CHEVANCEJ.B., JUNG C., FABRE L., TARDY C., COMPAN M., BACILIERI R., LACOMBE T., THIS P. 2010. Evolution and history of grapevine (*Vitis vinifera*) under domestication: new morphometric perspectives to understand seed domestication syndrome and reveal origins of ancient European cultivars. *Annals of Botany* 105, 443-455.
- 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. *Annals of Botany* 94, 657-664.
- ZULPE N. & PAWAR V. 2012. GLCM textural features for brain tumor classification. *International Journal of Computer Science* 9, 354-359.

Introduction

The purpose of this study was to implement an image analysis system previously developed by Grillo *et al.* (2010) through the introduction of a new set of seed morpho-colorimetric variables, Elliptic Fourier Descriptors, hereafter EFDs and Haralick's parameters described in Chapter 2.

The targeted species were selected among those present in the *Sardinian Germplasm Bank* (BG-SAR), at the *Centre for Conservation of Biodiversity* (CCB) of the Department of Botany, University of Cagliari (Mattana *et al.*, 2005).

The best moment for the seeds harvest, the methods and the quantity of the material are regulated by ethical and scientific criteria that provide a high quality of the collected material and avoid the pauperization of the *in situ* genetic resources. So, germplasm was collected following internationally recognized protocols to guarantee the greatest representativeness of the genetic diversity of original population (Guarino *et al.*, 1995; Bacchetta *et al.*, 2008a). The lots in admittance at the *BG-SAR* were submitted to a period of post-maturation under controlled conditions (30% of relative humidity), and then cleaned and manually selected by sieves or by the aid of variable air flow gravimetric separators (Agriculex CB-2 Column Seed Cleaner). During the standard treatments of the accessions for their correct conservation, the images of the seeds lots were acquired in digital format, before their entrance into the dehydration room (15°C at the 15% of relative humidity), in order to avoid each possible variation in shape and colour (Bacchetta *et al.*, 2008a).

The quantity of the seeds to analyze depends on the material availability of the *BG-SAR*. For the analysis of every accession, a sample constituted by no less of 100 units was randomly prepared, but when the original accession was lower than 100 units, the analysis was executed on the totality of the whole lot. This should guarantee the representativity of the accession, and at the same time, minimize the intraspecific variability of the seeds morpho-colorimetric characteristics, generally due to the seed position inside the fruit and to the fruit position in the plant (Harper *et al.*, 1970; Rovner & Gyulai, 2008).

Sample images were acquired using a flatbed scanner (Epson Perfection-V600), with a resolution of 400 dpi and a scanning area not superior to 1024 x 1024 pixels. As discussed in previous chapter, the employ of a flatbed scanner to capture digital images, represents a cheap and quick solution to carry out and file image libraries of high quality, exploitable for morphological (McCormac *et al.*, 1990; McDonald *et al.*, 2001) and colorimetric measures (Shahin & Symons, 1999; Shahin *et al.*, 2006). Furthermore, a so simple image acquiring system can be undoubtedly integrated in the daily germplasm bank management.

Seeds were arranged on the scanner glass flat, so that they did not touch each other, they were also covered with a box dressed with opaque paper to avoid interference of environmental light. A couple images were acquired for each seed sample. The first was captured using a cover box dressed with opaque black paper, while the second was acquired covering the seeds with another box dressed with opaque white paper and with a reduced height, in order to avoid that the vividness of the seed shadows on a white background can corrupt the real dimension and colour of seeds. This procedure, that was followed for each accession, allows to apply the same

segmentation method aside from the seeds colour, without manually editing the resulted image to correct few little binarization errors.

The digital images so obtained were stored in TIFF format (*Tagged Image File Format*) and, together with the Kodak Q60 reference image that was monthly acquired in order to calibrate the scanner, they were send by e-mail to the Image Analysis laboratories of the *Stazione Sperimentale di Granicoltura per la Sicilia (SSG)*, where they were calibrated, processed and analyzed using specific macros developed, with a commercial image analysis software (Grillo, 2009).

The ‘Macro’

A macro could be defined as a list of program lines, compiled in the proprietary language of the used software for image processing and analysis, that allows to execute automatically and quickly all the routine functions and operation of acquiring, colour and geometric calibration, elaboration, processing and measurement of images.

To reach the aims of this study, the *KS-400 release 3.0* image analysis software by *Carl Zeiss Vision GmbH (Germany)* was used, together to its library of algorithms and functions. A macro, expressly developed for the image analysis of wild species seeds (Bacchetta *et al.*, 2008b; Grillo *et al.*, 2010), was partially modified, implemented with the introduction of the new set of variables, EFDs and Haralick’s descriptors, and used to achieve size, shape and colour measures of individual seeds in the images.

Hereafter, an illustrative sequence of various steps of this macro, called *germplasm-analysis_1.mcr*, is reported to explain the work method adopted to analyze the seeds.

Even if, as explained above, a separation algorithm could be applied to electronically singulate the seeds after the acquisition, depending on the

hardware speed, it might result a very slow procedure, and for this reason it was chosen to place the seeds on the top of the scanner, in singulated arrangement. Moreover, because of the high variability in seed morphology of the studied *taxa*, the setting of the conditions that would have allowed the electronic separation, would be resulted very hard and tedious.

The two original images of the seeds, that one with black background and that the other with white background, in this case *Medicago arborea* (Fig. 1), are acquired using a flatbed scanner and then standardized to correct the colour of the image, as discussed in previous chapter.

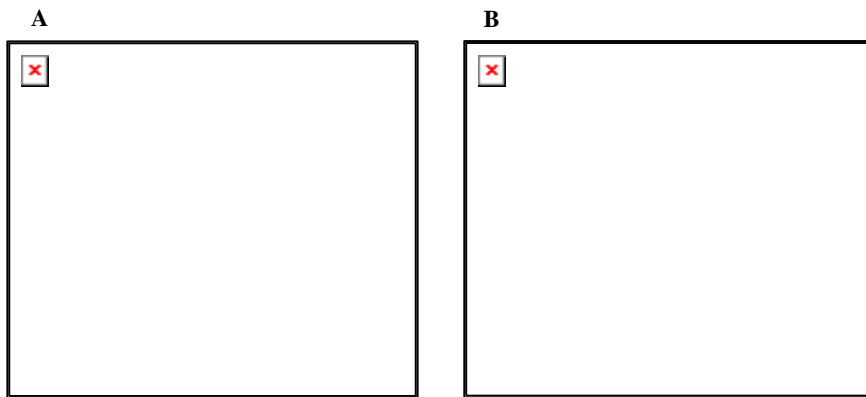


Figure 1. Black background (A) and white background (B) original images.

Using an interactive threshold to segment the contrasted image, the binary image is produced and used as a mask to execute the selected measures. A coloured label mask is superimposed to the binary resulted image, simply to control that all the seeds within the scene are separate. Before to perform the measurements, applying a conversion algorithm the RGB standardized image is transformed in HLS colour space. This image allows to extract information about colour hue, lightness and saturation of each seed.

All the seeds within the image are automatically selected for the measuring process, both in the RGB and in the HLS image, but clearly, maintaining the chance to deselect any seed that appear not well segmented. Finally, all the seeds are numbered with the purpose to have a reference indication related to the obtained data, because these numbers, labelled on the seeds, also indicate the order in which they are analyzed by the system (Fig. 2).

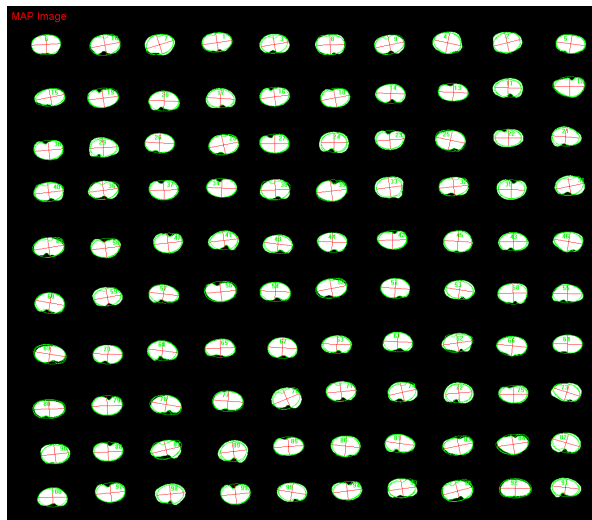


Figure 2. MAP image of selected seeds.

The germplasm samples

In this study, 157 accessions were analysed, belonging to three different genera, *Cistus*, *Medicago*, *Lavatera* and *Malva*, mainly collected in Sardinia, for a total of about 13,000 seeds. Some material, collected in other territories of the Mediterranean basin (e.g. Italian and Iberian peninsulas, Corsica, Balearic islands, Greece, Morocco) was also analysed, as well as seeds obtained by *ex situ* cultivation in the *Botanic Gardens of Cagliari* or provided by other scientific institutions (*Botanical Department - University*

of Catania, Botanical Department - University of Bari, Conservatoire Botanique National Méditerranéen de Porquerolles, Dirección General de Medio Natural de Murcia, Jardí Botànic de València, Institut y Jardí Botànic de Barcelona, Banco de semillas de la Universidad Politécnica de Madrid, Jardín Botánico de Córdoba, Banca del Germoplasma dell'Università degli Studi del Molise, Conservatoire Botanique National de Corse, Seed Conservation Department of the Kew Gardens, Mediterranean Agronomic Institute of Chania MAICh of Crete).

Statistics

In the statistic field, the aptitude to reorganize the raw data in few numbers or significative indicators able to describe the whole quantity of data without modify the overall meaning, is defined *descriptive statistic*. Particularly in the scientific research, the employ of a suitable treatment of data is very important, in order to overcome all the problems due to the *experimental error*, that is the cluster of the variations led by non-controlled factors, whose effects are overlapped to that one of the studied factor.

As discussed in the first chapter, one of the aims of this study consists in the implementation of statistical classifiers able to recognize and discriminate seeds belonging to different botanical ranks. In order to achieve this goal, the more significant morphocolorimetric features measured by image analysis, were used to describe size, shape and colour of each analysed seed, to identify and classify them on the basis of these morphological and colorimetric parameters.

Linear Discriminant Analysis

Sometimes, analyzing the data of various hundreds of groups of objects, in this case families, genera and species of seeds, the principal practical limitation of the pattern recognition is the high-dimensionality of the dataset. In the past several decades, many dimensionality reduction techniques have been proposed. The *Linear Discriminant Analysis (LDA)* (Fukunaga, 1990) is one of the most popular supervised methods for linear dimensionality reduction. It has been proven to be very powerful in many practical applications. The *LDA* is a multivariate statistical analysis and allows to analyze simultaneously measurements of many characters (qualitative and/or quantitative variables) from many samples. Fundamentally, this kind of statistical analysis aims to summarize the cases and simplify their structure to obtain the most correct grouping of them. The *LDA* is a very well-known method for dimensionality reduction and classification that projects high-dimensional data onto a low-dimensional space where the data achieve maximum class separability (Fukunaga, 1990; Duda *et al.*, 2000; Hastie *et al.*, 2001).

The derived features in *LDA*, also called *discriminant functions*, are linear combinations of the original features, where the coefficients are from the transformation matrix. The optimal projection or transformation in classical *LDA* is obtained by minimizing the within-class distance and maximizing the between-class distance simultaneously, thus achieving maximum class discrimination.

Calling J this objective, the original *LDA* formulation, also known as the *Fisher Linear Discriminant Analysis (FLDA)* (Fisher, 1936; 1940), that deals with binary-class classifications, can be described by the following formula:

$$J_F(w) = \frac{w^T S_b w}{w^T S_w w}$$

where w is a linear transformation matrix, S_b is the between-class scatter matrix and S_w is the within-class scatter matrix.

As discussed above, the purpose of the *LDA* is to maximize the between-class scatter, minimizing, at the same time, the within-class scatter. The two scatter matrices, S_b (between-class) and S_w (within-class), are defined as:

$$S_b = \sum_{i=1}^c p_i (m_i - m)(m_i - m)^T$$

$$S_w = \sum_{i=1}^c p_i S_i$$

where c is the number of classes; m_i and p_i are the mean vector and a priori probability of class i , respectively; $m = \sum_{i=1}^c p_i m_i$ is the total mean vector; S_i is the covariance matrix of class i (Grillo, 2009).

Generally, to obtain the discriminant functions and consequently classify objects (seeds in this case) into one of two or more groups, on the base of a set of features that describe the objects (e.g. area, perimeter, red,

green or blue channel, etc.), it is need to assign an object to one of a number of predetermined groups, based on observations made on the object. It is important to note that the groups are known or predetermined a priori.

So, it is possible summarize that the basic tasks of the *LDA* are two:

- to detect set of features that better can determine group membership of the object;
- to identify the classification model (or rule) that better can separate the groups.

The first of these two purposes, the detection of feature set, is a process of variables selection by steps, that allows to define the *LDA* method as *stepwise Linear Discriminant Analysis (sLDA)*. Using this method, only the best features for the identification of the different seed samples were detected, in order to implement a statistical classifier able to discriminate and classify the seeds, on the basis of morpho-colorimetric features. When there are a lot of predictors, the stepwise method can be useful to select automatically the best variables to be used in the classification model. The stepwise method starts with a model that doesn't include any of the predictor. At each step, the predictor with the largest *F to Enter* value that exceeds the entry criteria ($F \geq 3.84$) is added to the model. The variables left out of the analysis at the last step have *F to Enter* values smaller than 3.84, so no more are added. The *Tolerance* value indicates the proportion of a variable's variance not accounted for by other independent variables in the equation. A variable with very low *Tolerance* value proves little information to a model. *F to Remove* value describes the power of each variable in the model and it is useful to describe what happens if a variable is removed from the current model. The process was automatically stopped when no remaining feature increased the discrimination ability (Venora *et al.*, 2009a; 2009b).

The second purpose concerns the model or rule of classification to predict the membership of a new object on the base of the model. This approach is commonly used for the classification/identification of unknown groups characterized by quantitative and qualitative variables (Fisher, 1936; 1940).

This method requires a teaching procedure that use information derived by previous identified sample groups (also called *training set*) allowing to develop and to teach all the classifiers used in the study.

In a *Linear Discriminant Analysis*, the class categories or the groups that represent what it is looking for, are called *dependent variable*; while each measured feature, that describes the analysed object, is statistically defined *independent variable*. Hence, in these cases study, the analysed objects are seeds, the *dependent variable* is the considered taxonomic rank (family, genera or species) of the germplasm accessions, while the descriptive features (area, perimeter, red, green or blue channel, etc.) are the *independent variables*.

Thus, the *LDA* finds a set of *discriminant functions*, whose values are as close as possible within groups and as far apart as possible between groups. A *discriminant function* (f_{ni}), that above was defined as a linear combination of the discriminating variables, has the following mathematical form:

$$f_{ni} = a_0 + a_1x_{1ni} + a_2x_{2ni} + a_3x_{3ni} + \dots + a_kx_{kni}$$

where (f_{ni}) is the value (or score) on the canonical discriminant function for case i in the group n ; x is the value on discriminant variable for

the same case in the same group; and a is the autovalor which produce the desired characteristics in the function.

The coefficients $a_1, a_2, a_3, \dots, a_k$, for the first discriminant function are derived so as to maximize the differences between the group means. At the same way, the coefficients $a_1, a_2, a_3, \dots, a_k$, for the second discriminant function are also derived in order to maximize the difference between the group means, but they are subject to the constraint that the values, on the second discriminant function, are not correlated with the values on the first discriminant function, and so on for the discriminant function that follow. In geometrical terms, the second discriminant function is orthogonal to the first, and the third discriminant function is orthogonal to the second, and so on. The maximum number of unique functions that can be derived is equal to the number of groups minus one or equal to the number of discriminating variables.

Summarizing, the discriminant functions were selected so that:

- f_1 reflects, as much as possible, the differences between the groups;
- f_2 reflects, as much as possible, the differences between the groups, not highlighted by f_1 ;
- f_3 reflects, as much as possible, the differences between the groups, not highlighted by f_1 and f_2 ;
- ... and so on.

Each linear discriminant function explains a certain percentage of the total variance (or variability) of the cases, and all of them explain the 100% of the variability. Generally, it is desirable that the first two discriminant functions explain variability levels higher than 60-70% (Grillo, 2009).

The simplest case of a *LDA* assumes that the groups are linearly separable. It occurs when the groups can be separated by a linear combination of features that describe the objects. If the *independent variables* are only two, the separator between object groups is simply a line; if the features are three, the separator is a plane; while if the number of *independent variables* is more than three, the separator become a hyper-plane. In this last case, the great utility of linear dimensionality reduction of the *LDA* method, is evident, and can be better explained taking advantage of graphical representations. Figure 3 shows a graphical representation of a case in which objects belonging to three different groups (dependent variables) are detected by only two discriminant functions.

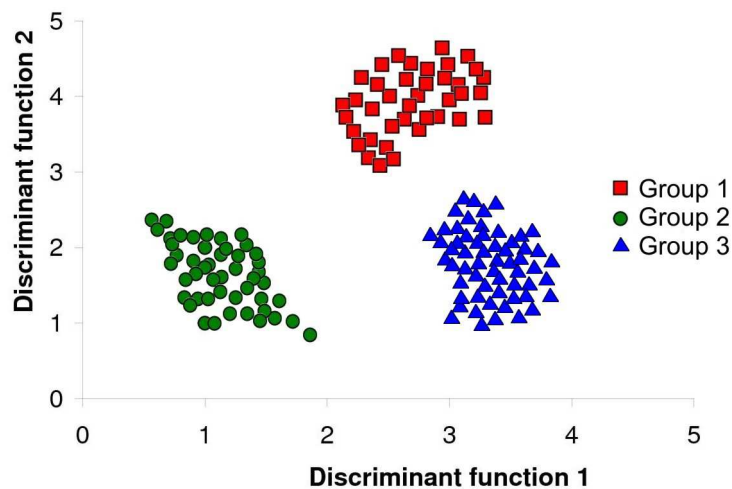


Figure 3. Bi-dimensional plot of a Linear Discriminant Analysis.

In Figure 4, a 3D plot shows the scores of three of the discriminating functions used to distinguish the objects belonging to five different groups.

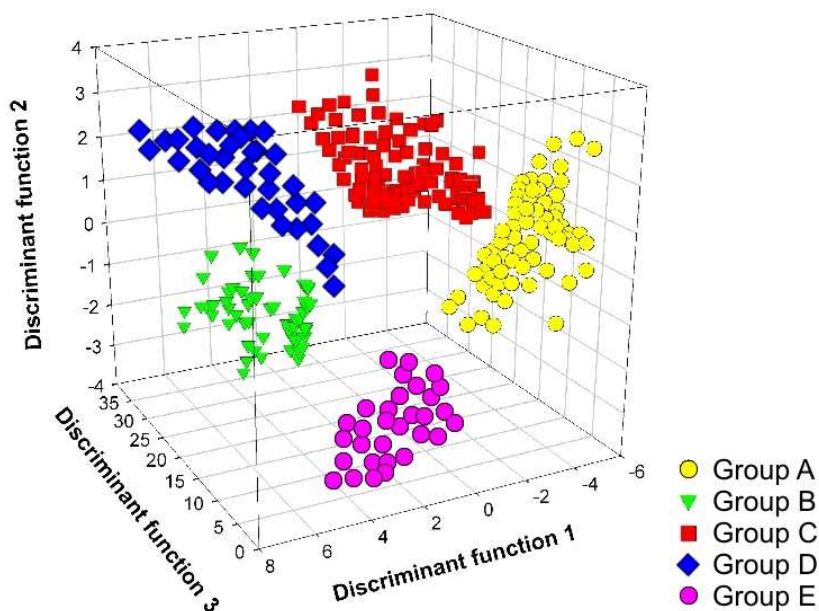


Figure 4. Three-dimensional graphic representation of a LDA.

The principal advantage of the multidimensional plot is in the possibility to represent graphically the discriminant scores in a biggest space than a classical Cartesian plane. In this way, it is simplest to visually appreciate the distances among groups.

When different groups of objects have to be discriminated and only two discriminant functions are available, it is possible to insert a third function that allows to draw a multidimensional plot, the *Mahalanobis distance*. This is a measure introduced by the Indian statistician Prasanta Chandra Mahalanobis (1936) and it is based on correlations among variables by which different patterns can be identified and analysed. It determines similarity of an unknown sample set to a known one. In other words, Mahalanobis distance is a measure of distance between two data points in the space defined by two or more discriminant functions; a high value indicates that a particular case includes extreme values for one or more independent

variables and it can be considered not similar to other cases (Bacchetta *et al.*, 2008b).

Formally, the Mahalanobis distance from a group of values with mean $\mu=(\mu_1,\mu_2,\mu_3,\dots,\mu_n)^T$, and covariance matrix S for a multivariate vector $x=(x_1,x_2,x_3,\dots,x_n)^T$ is defined as (De Maesschalck *et al.*, 2000):

$$D_M(x) = \sqrt{(x - \mu)^T S^{-1} (x - \mu)}$$

In Figure 5, a three dimensional graphic representation is showed, in which objects belonging to different groups are detected by two discriminant functions and the Mahalanobis distance.

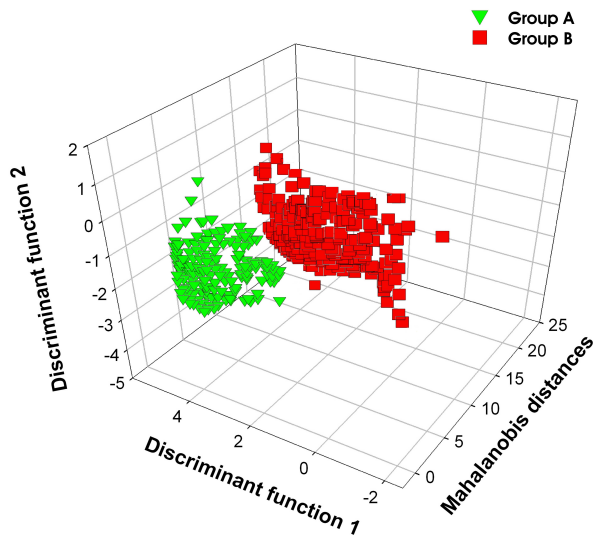


Figure 5. 3D plot of a LDA with two discriminant functions and the Mahalanobis distance.

A major drawback of *LDA* is that it often suffers from the small sample size problem when dealing with the high dimensional data. When there are not enough training samples, S_w (in the 6.1) may become singular and it is difficult to compute the *LDA* vectors. Several approaches have been proposed to address this problem (Liu *et al.*, 1992; Belhumeur *et al.*, 1997; Chen *et al.*, 2000; Yu & Yang, 2001), but a common problem with all these proposed variant *LDA* approaches is that they all lose some discriminative information in the high dimensional space. Anyway, the stepwise *LDA* method has been applied successfully in many different applications (Swets & Weng, 1996; Venora *et al.*, 2007; 2009a; Bacchetta *et al.*, 2008b, Grillo, 2009).

The cross-validation

In this research study, the *cross-validation* procedure, also called *rotation estimation* (Picard & Cook, 1984; Kohavi, 1995), was applied, both to evaluate the performance and to validate any classifier, and to avoid problems and/or mistakes that might arise on account of seed samples not enough numerically representative. Indeed, this procedure is usually applied for small amount of data, in lack of a broad group of new unknown cases (*test set*). It tests the individual cases and classifies them on the basis of all the others (SPSS, 2007).

The most common types of cross-validation are three. The *repeated random sub-sampling validation*, is a method that randomly splits the dataset into training and test (or validation) set. For each such split, the model is fit to the training set of data, and predictive accuracy is assessed using the test set of data. The results are then averaged over the splits. The advantage of this method is that the proportion of the training/test split is not dependent on the number of iterations (as it occurs for the *k-fold cross validation* type). The

disadvantage of this method is that it is very expensive from the point of view of the optimal use of the available dataset.

In *K-fold cross-validation*, another very common type of cross-validation, the original sample is partitioned into K subsamples. One of the K subsamples is put aside as the test dataset to validate the model, and the remaining $K-1$ subsamples are used as training set. The cross-validation process is then repeated K times (the folds), with each of the K subsamples used exactly once as the validation data. Then, the K results from the folds can be averaged (or otherwise combined) to produce a single estimation. The advantage of this method is that all cases are used for both training and validation, and each case is used for validation exactly once, but as hinted above, the ratio between the split training set and the test set, is closely related to the number K of process iterations.

The third common type of cross-validation is the *leave-one-out cross-validation (LOOCV)*. As the name suggests, it involves using a single case from the original sample set as the validation dataset, and the remaining cases as the training set. This is repeated such that each case in the sample set is used once as the test set. This is the same as a *K-fold cross-validation* with K being equal to the number of cases in the original sample. Unfortunately, the *leave-one-out cross-validation* is often computationally expensive because of the large number of times the training process is repeated.

Finally, in order to evaluate the quality of the discriminant functions achieved for each statistical comparison, the Wilks' Lambda, the percentage of explained variance and the canonical correlation between the discriminant functions and the group membership, were computed. The Box's M tests was executed to assess the homogeneity of covariance matrices of the features chosen by the stepwise LDA while the analysis of the standardized residuals

was performed to verify the homoscedasticity of the variance of the dependent variables used to discriminate among the groups' membership (Box, 1949; Haberman, 1973; Morrison, 2004). Kolmogorov-Smirnov's test was performed to compare the empirical distribution of the discriminant functions with the relative cumulative distribution function of the reference probability distribution, while the and Levene's test was executed to assess the equality of variances for the used discriminant functions calculated for groups' membership (Gastwirth *et al.*, 2009; Levene, 1960; Lopes, 2011).

References

- BACCHETTA G., BUENO SANCHEZ A., FENU G., JIMENEZ-ALFARO B., MATTANA E., PIOTTO B., VIREVAIRE M. 2008a. *Conservacion ex situ de plantas silvestres*. Principado de Asturias / La Caixa.
- BACCHETTA G., GRILLO O., MATTANA E., VENORA G. 2008b. Morpho-colorimetric characterization by image analysis to identify diaspores of wild plant species. *Flora* 203, 669-682.
- BELHUMEOUR P., HESPANHA J., KRIEGMAN D. 1997. Eigenfaces vs. Fisherfaces: Recognition using class specific linear projection. *IEEE Transaction Pattern Analysis and Machine Intelligence* 19, 711-720.
- BOX G.E.P. 1949. A general distribution theory for a class of likelihood criteria. *Biometrika* 36, 317-346.
- CHEN L., LIAO H., KO M., LIN J., YU G. 2000. A new LDA-based face recognition system which can solve the small sample size problem. *Pattern Recognition* 33, 1713-1726.
- DE MAESSCHALCK R., JOUAN-RIMBAUD D., MASSART D.L. 2000. The Mahalanobis distance. *Chemometrics and Intelligent Laboratory Systems* 50, 1-18.
- DUDA R., HART P., STORK D. 2000. *Pattern classification*. 2nd edition - Published by Wiley, New York, NY, USA, 860 pages, 2000, ISBN: 0-471-5669-3.
- FISHER R.A. 1936. The use of multiple measurements in taxonomic problems. *Annals of Eugenics* 7, 179-188.
- FISHER R.A. 1940. The precision of discriminant functions. *Annals of Eugenics* 10, 422-429.
- FUKUNAGA K. 1990. *Introduction to statistical pattern recognition*. 2nd edition - Published by Academic Press, San Diego, USA, pages 592, 1990. ISBN: 0-12-269851-7.
- GASTWIRTH J.L., GEL Y.R., MIAO W. 2009. The impact of Levene's test of equality of variances on statistical theory and practice. *Statistical Science* 24, 343-360.
- GRILLO O. 2009. Germplasm morpho-colorimetric characterization by image analysis and statistical classification of the most representative families of Mediterranean vascular flora (*Doctoral dissertation*).
- GRILLO O., MATTANA E., VENORA G., BACCHETTA G. 2010. Statistical seed classifiers of 10 plant families representative of the Mediterranean vascular flora. *Seed Science and Technology* 38, 455-476.

- GUARINO L., RAMANANTHA RAO V., REID R. 1995. *Collecting Plant Genetic Diversity* - Technical guidelines. CABI Wallingford, UK, pages 748.
- HABERMAN S.J. 1973. The analysis of residuals in cross-classified tables. *Biometrics* 29, 205-220.
- HARPER J.L., LOVELL P.H., MOORE K.G. 1970. The shapes and sizes of seeds. *Annual Review Ecology and Systematics* 1, 327-356.
- HASTIE T., TIBSHIRANI R., FRIEDMAN J. 2001. *The elements of statistical learning: Data mining, inference, and prediction*. 2nd edition - Published by Springer, New York, NY, USA, 741 pages, 2001, ISBN: 0-387-95284-5, 978-0387-95284-0.
- KOHAVI R. 1995. A study of cross-validation and bootstrap for accuracy estimation and model selection. *Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence* 2, 1137-1143.
- LEVENE H. 1960. Robust tests for equality of variances. In: Olkin, I., Ghurye, S.G., Hoeffding, W., Madow, W.G. & Mann H.B., Eds., *Contributions to Probability and Statistics: Essays in Honor of Harold Hotelling*. Stanford University Press. pp. 278-292.
- LIU K., CHENG Y., YANG J. 1992. A generalized optimal set of discriminant vectors. *Pattern Recognition* 25, 731-739.
- LOPES R.H.C. 2011. Kolmogorov-Smirnov Test. In: Lovric M., Eds., *International Encyclopedia of Statistical Science*. Springer Berlin Heidelberg. pp. 718-720.
- MAHALANOBIS P.C. 1936. On the generalised distance in statistics. *Proceedings of National Institute of Science of India* 12, 49-55.
- MATTANA E., FENU G., BACCHETTA G. 2005. La Banca del Germoplasma della Sardegna (BG-SAR): uno strumento per la conservazione del germoplasma autoctono sardo. *Informatore Botanico Italiano* 37, 145-146.
- MCCORMAC A.C., KEEFE P.D., DRAPER S.R. 1990. Automated vigour testing of field vegetables using image analysis. *Seed Science Technology* 18, 103-112.
- MCDONALD M.B., EVANS A.F., BENNET M.A. 2001. Using scanners to improve seed and seedling evaluations. *Seed Science Technology* 29, 683-689.
- MORRISON D.F. 2004. *Multivariate Statistical Methods*. 4th edition. Cengage Learning Duxbury Press.
- PICARD R. & COOK D. 1984. Cross-Validation of Regression Models. *Journal of the American Statistical Association* 79, 575-583.

- ROVNER I. & GYULAI F. 2009. Computer-Assisted Morphometry: A New Method for Assessing and Distinguishing Morphological Variation in Wild and Domestic Seed Populations. *Economic Botany* 61, 154-172.
- SHAHIN M.A. & SYMONS S.J. 1999. A machine vision system for color classification of lentils. *ASAE Annual International Meeting* 3, 1-16, Toronto, July 18th/21st, 1999.
- SHAHIN M.A., SYMONS S.J., POYSA V.W. 2006. Determining soya bean seed size uniformity with image analysis. *Biosystems Engineering* 94, 191-198.
- SPSS, 2007. Base 16.0 Application Guide. Prentice Hall, USA, New Jersey.
- SWETS D.L. & WENG J. 1996. Using discriminant eigenfeatures for image retrieval. *IEEE Transaction Pattern Analysis and Machine Intelligence* 18, 831-836.
- VENORA G., GRILLO O., RAVALLI C., CREMONINI R. 2009a. Identification of Italian landraces of beans (*Phaseolus vulgaris* L.), using an image analysis system. *Scientia Horticulturae* 121, 410-418.
- VENORA G., GRILLO O., SACCONI R. 2009b. Durum wheat storage centers: evaluation of vitreous, starchy and shrunken kernels by image analysis system. *Journal of Cereal Science* 49, 429-440.
- VENORA G., GRILLO O., SHAHIN M.A., SYMONS S.J. 2007. Identification of Sicilian landraces and Canadian cultivars of lentil using image analysis system. *Food Research International* 40, 161-166.
- YU H. & YANG J. 2001. A direct LDA algorithm for high-dimensional data with application to face recognition. *Pattern Recognition* 34, 2067-2070.

The second part of this dissertation concerned some applications of the image analysis technologies discussed above.

In Chapter 1, in particular, the morpho-colorimetric characterization of *Cistus* L. (Cistaceae) seeds by image analysis was treated: a database of morphometric and colorimetric data was carried out to statistically discriminate and identify at inter an intra-specific level.

Seed image analysis provided evidence of taxonomical differentiation within the *Medicago* L. sect. *Dendrotelis* (Fabaceae) (Chapter 2).

Finally, the relationships among 79 *taxa* belonging to the *Lavatera* and *Malva* genera were discussed in Chapter 3, in order to contribute to their doubtful systematic treatment.

Inter and intra-specific diversity in Cistus L. (Cistaceae) seeds, analysed by computer vision techniques

Abstract

Seed mean weight and 137 morpho-colorimetric quantitative variables describing shape, size, colour and textural seed traits, were measured using image analysis techniques, with the aim to discriminate among different species and subspecies of the genus *Cistus*. Also, the intra-specific phenotypic differentiation of *C. creticus* through the comparison of three subspecies (*C. creticus* subsp. *creticus*, *C. c.* subsp. *eriocephalus* and *C. c.* subsp. *corsicus*) and the inter-population variability among five *C. creticus* subsp. *eriocephalus* populations were evaluated. Data obtained were analysed applying stepwise Linear Discriminant Analysis method, recording an overall cross-validated classification performance of 80.6% at species level. With regard to *C. creticus* as case study, percentages of correct discrimination of 96.7% and 99.6% were achieved at intraspecific and inter-population levels, respectively. In this classification model, the relevance of the colorimetric and textural descriptive features was highlighted besides to the seed mean weight that was the most discriminant feature at specific and intraspecific level. These achievements proved the ability of image analysis system to be highly diagnostic in the statistically assessment of the morpho-colorimetric traits variability of studied *taxa* seeds.

Introduction

The rockrose (*Cistus* L., Cistaceae) is one of the most representative and widespread genus of the Mediterranean vascular flora, includes about 20 species from the Mediterranean area, reaching the Caucasus mountains to the East and the Canary Islands to the West (Ferrer-Gallego *et al.*, 2013). Its highest diversity is found in the Western Mediterranean region, with 14 species occurring in the Iberian Peninsula and North-Western Africa (Guzmán & Vargas, 2005; Fernández-Mazuecos & Vargas, 2010). A long history of human activities has favoured distribution and abundance of *Cistus* species in the Mediterranean Basin, which plants are formed as early successional stages following woodland disturbances such as fire and soil overturning (Thompson, 2005).

In marked contrast to the detailed knowledge of ecological characteristics, understanding of the evolution of morphological characters and phylogenetic relationships within the genus is extremely limited. Even if *Cistus* is a relatively small genus, it is complex because shows a significant morphological diversification, caused by the polymorphism of a number of species and the hybridization between related species (Simonet & Ansereau, 1939; Pawluczyk *et al.*, 2012). Indeed, hybridization has been reported to be an active process in *Cistus* genus (Ellul *et al.*, 2002), and many hybrid combinations within and among pink or white-flowered species have been recorded (in the field), based on intermediate morphological characters (Paolini *et al.*, 2009).

The taxonomy of *Cistus* has traditionally been based on vegetative (nerve number, shape, and hairiness of leaves) and reproductive characters (sepal number, petal colour, style length, and number of fruit valves), although evolutionary mechanisms responsible for the morphological diversity within the genus remain poorly understood (Guzman & Vargas,

2009; Fernández-Mazuecos & Vargas, 2010). For taxonomic purposes, several investigations on the anatomical and morphological leaf traits of *Cistus* species have been previously reported (e.g. Jeanmonod & Gamisans 2007; Tattini *et al.*, 2007; Catoni *et al.*, 2012). On the other hand, few attempts of taxonomic classification of *Cistus* based on the morpho-colorimetric description of seeds, have been reported in literature (Cerabolini *et al.*, 2003; Delgado *et al.*, 2008; Moreira *et al.*, 2012). Although it is not wholly determined by morphological characters, many authors testified the importance of seed size and shape also to predict the seed persistence in the soil (Funes *et al.*, 1999; Saatkamp *et al.*, 2009), at least in some climatic conditions. This aspect seems also to be indirectly related to a certain inter-specific variability, proved on many plant genera (Hernández-Martínez *et al.*, 2011; Grillo *et al.*, 2013), but not in others (Pinna *et al.*, 2014). In contrast, because the seed morphological aspects are genetically fixed characters, undoubtedly the seed shows diagnostic features much more statistically significant in comparison to other plant characters.

In the last two decades, image analysis has achieved several goals in morpho-colorimetric evaluation of seeds (Wiesnerová & Wiesner, 2008; Granitto *et al.*, 2003; Venora *et al.*, 2009a) for identification of both wild plant (Rovner & Gyulai, 2007; Bacchetta *et al.*, 2008a; Grillo *et al.*, 2012) and agronomical important species (Shahin & Symons, 2003a; Venora *et al.*, 2007, 2009b; Firatligil-Durmus *et al.*, 2010; Grillo *et al.*, 2011; Smykalova *et al.*, 2011; 2013), proving to be a performance analytical tool for taxonomic studies.

The discriminant ability of the identification system depends not only on the intra-specific representativeness of analyzed *taxa*, but also, on the quality and quantity of the parameters measured and used to differentiate among groups. For this reason, as reported in recent literature, Haralick's

parameters, evaluating the surface texture of seeds (Diamond *et al.*, 2004; Gerger & Smolle, 2003; Nanni *et al.*, 2010), and the Elliptic Fourier Descriptors hereafter EFDs (Iwata *et al.*, 2002, 2004; Yoshioka *et al.*, 2004; Kawabata *et al.*, 2009; Orrù *et al.*, 2012; 2013) for a detailed description of the shape, were considered as variables and included in the statistical classification system, in addition to the common seeds morpho-colorimetric traits used in previous similar works (Bacchetta *et al.*, 2008a; 2011a; Grillo *et al.*, 2010, 2013).

Applying computer vision techniques, the family of Cistaceae, and six *taxa*, belonging to genus *Cistus*, were involved in a previous work (Bacchetta *et al.*, 2008a). Incrementing the number of features and the amount of analysed seed lots, Grillo *et al.* (2010) improved the classification performance, for each of the ten studied families, also included the Cistaceae and the relative *taxa*. Compared to previous studies, we increase here the number of *taxa* and parameters, including Haralick's parameters and EFDs, assessing two taxonomic levels (species and subspecies), as well as the population variability of one case study.

Specifically, following the systematic treatment proposed by *The Plant List* (2013), the aims of the present work were to: (1) characterize the genus *Cistus* at species level on the basis of seed mean weight, shape, size, colour and textural measurements by computer vision; (2) evaluate both the intra-specific phenotypic differentiation of *C. creticus* through the comparison of three subspecies (*C. creticus* subsp. *creticus*, *C. creticus* subsp. *eriocephalus* and *C. creticus* subsp. *corsicus*) and the inter-population variability among five *C. creticus* subsp. *eriocephalus* populations.

Material and methods

Seed-lot details

Seeds of 14 *taxa* of the genus *Cistus*, belonging to 49 populations, were collected during a period of 21 years (from 1993 to 2013), in eight Mediterranean regions (Corse, France, Greece, Italy, Morocco, Sardinia, Sicily and Spain) for an overall of 65 accessions and 6,475 seeds (Table 1). Considering the high statistical significance that seed mean weight has been proved to have in previous methodologically similar works (Grillo *et al.*, 2010; Bacchetta *et al.*, 2011a; 2011b; Smykalova *et al.*, 2011; 2013), this feature was considered attempting to discriminate among the studied *Cistus taxa* and, then, recorded before acquiring the seed lots image. Afterwards, the seeds were ultra-dried out down to 3% R.H., guaranteeing homogeneity and regularity in seed size and weight (Pérez-García *et al.*, 2007). All accessions were stored at -25°C in the Sardinian Germplasm Bank (BG-SAR), according to the protocols reported in Bacchetta *et al.* (2008b).

Table 1. Geographical regions, sampling years and seed amount of the studied *Cistus taxa* populations.

Taxon	Population	Geographical region	Collecting year	Seed amount
<i>C. albidus</i>	Prox Chechaouen	Morocco	2001	100
<i>C. albidus</i>	Mt. Corrasi (Oliena)	Sardinia	2006	100
<i>C. albidus</i>	Miniera Sos Enattos (Lula)	Spain	2006	100
<i>C. albidus</i>	Jaen (JA)	Spain	2008	100
<i>C. albidus</i>	Sierra Elvira (GR)	Spain	2009	100
<i>C. albidus</i>	Sierra Elvira (GR)	Spain	2009	100
<i>C. albidus</i>	Rio Cabril (Cuenca)	Spain	2013	100
<i>C. clusii</i>	CIEF Valencia	Spain	2000	100
<i>C. clusii</i>	Lesina (FG)	Italy	2006	100
<i>C. clusii</i>	La Resinera (JA)	Spain	2009	100
<i>C. clusii</i>	Sierra de Lujar (GR)	Spain	2009	100
<i>C. clusii</i>	Ragusa	Italy	2009	100
<i>C. clusii</i>	Pineta di Vittoria (RG)	Italy	2013	97
<i>C. creticus</i> subsp. <i>corsicus</i>	Santo Pietro di Tenda	Corse	1993	100
<i>C. creticus</i> subsp. <i>creticus</i>	Karfas (Chios)	Greece	2006	100
<i>C. creticus</i> subsp. <i>creticus</i>	Leonforte (EN)	Sicily	2010	100
<i>C. creticus</i> subsp. <i>creticus</i>	Akrotiri (Chania)	Greece	2012	100
<i>C. creticus</i> subsp. <i>eriocephalus</i>	Agruxiau (CI)	Sardinia	2006	100
<i>C. creticus</i> subsp. <i>eriocephalus</i>	Agruxiau (CI)	Sardinia	2007	100
<i>C. creticus</i> subsp. <i>eriocephalus</i>	Pineta della Foce del Garigliano (CE)	Italy	2009	100
<i>C. creticus</i> subsp. <i>eriocephalus</i>	Casargiu (CI)	Sardinia	2010	100
<i>C. creticus</i> subsp. <i>eriocephalus</i>	Agruxiau (CI)	Sardinia	2010	100
<i>C. creticus</i> subsp. <i>eriocephalus</i>	Portixeddu (Buggerru)	Sardinia	2011	100

Table 1. Continue				
<i>C. creticus</i> subsp. <i>eriocephalus</i>	Piscinamamma (Pula)	Sardinia	2012	100
<i>C. crispus</i>	Hyerer (Porquerolles)	France	2005	100
<i>C. crispus</i>	Colle S. Rizzo (ME)	Sicily	2006	86
<i>C. crispus</i>	Hinojos (JA)	Spain	2010	100
<i>C. crispus</i>	Ctra a Bab Berret	Morocco	2011	100
<i>C. heterophyllus</i> subsp. <i>cartaginensis</i>	CIEF Valencia	Spain	2007	100
<i>C. heterophyllus</i> subsp. <i>cartaginensis</i>	DGMN Murcia	Spain	2007	100
<i>C. ladanifer</i>	Hyerer (Porquerolles)	France	2005	100
<i>C. ladanifer</i>	Andujar (JA)	Spain	2010	100
<i>C. ladanifer</i>	Prox Chechaouen	Morocco	2011	97
<i>C. laurifolius</i>	Sierra de Baza (GR)	Spain	2001	100
<i>C. laurifolius</i>	CIEF Valencia	Spain	2007	100
<i>C. laurifolius</i>	Sierra de Lujar (GR)	Spain	2009	100
<i>C. laurifolius</i>	Ketama a Jbel Tidighine	Morocco	2011	100
<i>C. monspelliensis</i>	Agruxiau (CI)	Sardinia	2006	100
<i>C. monspelliensis</i>	Pantano Quebrajano (JA)	Spain	2008	100
<i>C. monspelliensis</i>	Diga Sterili (S. Giorgio - CI)	Sardinia	2010	100
<i>C. monspelliensis</i>	Montevecchio (Guspini - CI)	Sardinia	2010	100
<i>C. monspelliensis</i>	Prox Chechaouen	Morocco	2011	100
<i>C. monspelliensis</i>	Khamis M'Diq a Bab Berret	Morocco	2011	100
<i>C. monspelliensis</i>	Calaverde (Pula)	Sardinia	2012	100
<i>C. parviflorus</i>	Karave - Isola di Gavdos (Creta)	Greece	2012	100
<i>C. parviflorus</i>	Giaudos	Greece	2013	99
<i>C. populifolius</i>	Serra Penas Altas	Spain	2004	99
<i>C. populifolius</i>	Pantano Quebrajano (JA)	Spain	2008	100
<i>C. salviifolius</i>	Huelva (HU)	Spain	1999	100
<i>C. salviifolius</i>	Agruxiau (CI)	Sardinia	2006	100

Table 1. Continue				
<i>C. salviifolius</i>	Porto Campana (Chia)	Sardinia	2007	100
<i>C. salviifolius</i>	Quartu S. Elena (CA)	Sardinia	2007	100
<i>C. salviifolius</i>	Agruxiau (CI)	Sardinia	2007	100
<i>C. salviifolius</i>	Porto Campana (Chia)	Sardinia	2007	100
<i>C. salviifolius</i>	Monte Altesina - Nicosia (EN)	Sicily	2010	100
<i>C. salviifolius</i>	Cungiau (CI)	Sardinia	2010	100
<i>C. salviifolius</i>	Mt. Vecchio - Guspini (CI)	Sardinia	2010	100
<i>C. salviifolius</i>	Is Arenas (Arbus)	Sardinia	2010	100
<i>C. salviifolius</i>	Simius (Villasimius)	Sardinia	2010	100
<i>C. salviifolius</i>	Porto Campana (Chia)	Sardinia	2010	100
<i>C. salviifolius</i>	Portixeddu (Buggerru)	Sardinia	2011	100
<i>C. salviifolius</i>	Ctra a Bab Berret	Morocco	2011	100
<i>C. salviifolius</i>	Calaverde (Pula)	Sardinia	2012	100
<i>C. salviifolius</i>	Taranto	Italy	2013	100
<i>C. salviifolius</i>	Collesano (PA)	Sicily	2013	97

Image analysis system

Samples digital images, consisting of 100 seeds randomly disposed on tray, were acquired using a flatbed scanner (Epson GT-15000) with a digital resolution of 400 dpi and a scanning area not exceeding 1024×1024 pixel. For accessions of fewer than 100 seeds, the analysis was executed on the whole batch. A total of 6,475 seeds were analyzed. Before the image acquisition was performed, the scanner was calibrated for colour matching following the protocol of Shahin and Symons (2003b) before seed samples image acquisition, as suggested by Venora *et al.* (2009b).

Digital images of seeds were processed and analyzed using the software package KS-400 V. 3.0 (Carl Zeiss, Vision, Oberkochen, Germany). A macro specifically developed for the characterization of seeds (Venora *et*

al., 2009b), was modified to perform automatically all the analysis procedures, reducing the execution time and contextually mistakes in the process.

In order to improve the discrimination power, this macro was further enhanced adding algorithms able to compute the Elliptic Fourier Descriptors (EFDs) for each analyzed seed. This method allows description of the boundary of the seed projection as an array of complex numbers which correspond to the pixel positions on the seed boundary. So, from the seed apex, defined as the starting point in a Cartesian system, chain codes are generated. A chain code is a lossless compression algorithm for binary images. The basic principle of chain codes is to separately encode each connected component (pixel) in the image. The encoder then moves along the boundary of the image and, at each step, transmits a symbol representing the direction of this movement. This continues until the encoder returns to the starting position. This method is based on separate Fourier decompositions of the incremental changes of the X and Y coordinates as a function of the cumulative length along the boundary (Kuhl and Giardina 1982). Each harmonic (n) corresponds to four coefficients (an , bn , cn and dn) defining the ellipse in the XY plane. The coefficients of the first harmonic, describing the best fitting ellipse of outlines, are used to standardize size (surface area) and to orientate seeds (Terral *et al.* 2010). According to Terral *et al.* (2010), about the use of a number of harmonics for an optimal description of seed outlines, in order to minimize the measurement errors and to optimize the efficiency of shape reconstruction, 20 harmonics were used to define the seed boundaries, obtaining a further 78 parameters useful to discriminate among the studied seeds (Orrù *et al.* 2012).

Moreover, the macro was further improved adding algorithms able to compute 11 Haralick's descriptors and the relative standard deviations for

each analyzed seed. These parameters are generally used when the objects in the images cannot be separated due to indefinite grey values variations. In these cases, the evaluation of texture, tone and context allows to define the spatial distribution of the image intensities and discrete tonal features. When a small area of the image has little variation of discrete tonal features, the dominant property of that area is grey tone. When a small area has wide variation of discrete tonal features, the dominant property of that area is texture (Haralick & Shapiro, 1991). According to Haralick *et al.* (1973), the concept of tone is based on varying shades of grey of resolution cells in a photographic image, while texture is concerned with the spatial statistical distribution of grey tones. Texture and tone are not independent concepts; rather, they bear an inextricable relationship to one another very much like the relationship between a particle and a wave. Context, texture and tone are always present in the image, although at times one property can dominate the others.

The basis for these features is the grey-level co-occurrence matrix (G in equation 1). This matrix is square with dimension N_g , where N_g is the number of grey levels in the image. Element $[i,j]$ of the matrix is generated by counting the number of times a pixel with value i is adjacent to a pixel with value j and then dividing the entire matrix by the total number of such comparisons made. Each entry is therefore considered to be the probability that a pixel with value i will be found adjacent to a pixel of value j .

$$G = \begin{bmatrix} p(1,1) & p(1,2) & \dots & p(1, N_g) \\ p(2,1) & p(2,2) & \dots & p(2, N_g) \\ \vdots & \vdots & \ddots & \vdots \\ p(N_g, 1) & p(N_g, 2) & \dots & p(N_g, N_g) \end{bmatrix} \quad (1)$$

In Table 2, the 11 Haralick's descriptors measured on each seed to mathematically describe the surface texture, are reported.

Table 2. Haralick's descriptors measured as reported in Haralick et al. (1973).

	<i>Feature</i>	<i>Equation</i>
Har 1	Angular second moment	$\sum_i \sum_j p(i, j)^2$
Har 2	Contrast	$\sum_{n=0}^{N_g-1} n^2 \left\{ \sum_{i=1}^{N_g} \sum_{j=1}^{N_g} p(i, j) \right\}, i, j = n$
Har 3	Correlation	$\frac{\sum_i \sum_j (ij) p(i, j) - \mu_x \mu_y}{\sigma_x \sigma_y}$
		where μ_x , μ_y , σ_x and σ_y are the means and the standard deviations of p_x and p_y .
Har 4	Sum of square: variance	$\sum_i \sum_j (i - \mu)^2 p(i, j)$
Har 5	Inverse difference moment	$\sum_i \sum_j \frac{1}{1 + (i - j)^2} p(i, j)$
Har 6	Sum average	$\sum_{n=2}^{2N_g} i p_{x+y}(i)$
		where x and y are the coordinates (row and column) of an entry in the co-occurrence matrix, and $p_{x+y}(i)$ is the probability of co-occurrence matrix coordinates summing to $x+y$.
Har 7	Sum variance	$\sum_{i=2}^{2N_g} (i - f_B)^2 p_{x+y}(i)$
Har 8	Sum entropy	$-\sum_{i=2}^{2N_g} p_{x+y}(i) \log\{p_{x+y}(i)\} = f_B$
Har 9	Entropy	$-\sum_i \sum_j p(i, j) \log[p(i, j)]$
Har 10	Difference variance	$\sum_{n=0}^{N_g-1} i^2 p_{x-y}(i)$
Har 11	Difference entropy	$-\sum_{n=0}^{N_g-1} p_{x-y}(i) \log\{p_{x-y}(i)\}$

Seed mean weight and 137 morphometric, colorimetric and textural characters were measured on each seed (Table 3).

Table 3. List of morpho-colorimetric features measured on seeds, excluding the 78 Elliptic Fourier Descriptors calculated according to Hâruta (2011).

Feature	Description
<i>A</i>	Area Seed area (mm ²)
<i>P</i>	Perimeter Seed perimeter (mm)
<i>P_{conv}</i>	Convex Perimeter Convex perimeter of the seed (mm)
<i>P_{Crof}</i>	Crofton's Perimeter Perimeter of the seed calculated using the Crofton's formula (mm)
<i>P_{conv}/P_{Crof}</i>	Perimeter ratio Ratio between convex and Crofton's perimeters
<i>D_{max}</i>	Max diameter Maximum diameter of the seed (mm)
<i>D_{min}</i>	Min diameter Minimum diameter of the seed (mm)
<i>D_{min}/D_{max}</i>	Feret ratio Ratio between minimum and maximum diameters
<i>Sf</i>	Shape Factor Seed shape descriptor = (4 x π x area)/perimeter ² (normalized value)
<i>Rf</i>	Roundness Factor Seed roundness descriptor = (4 x area)/(π x max diameter ²) (normalized value)
<i>Ecd</i>	Eq. circular diameter Diameter of a circle with an area equivalent to that of the seed (mm)
<i>EA_{max}</i>	Maximum ellipse axis Maximum axis of an ellipse with equivalent area (mm)
<i>EA_{min}</i>	Minimum ellipse axis Minimum axis of an ellipse with equivalent area (mm)
<i>R_{mean}</i>	Mean red channel Red channel mean value of seed pixels (grey levels)
<i>R_{sd}</i>	Red std. deviation Red channel standard deviation of seed pixels
<i>G_{mean}</i>	Mean green channel Green channel mean value of seed pixels (grey levels)
<i>G_{sd}</i>	Green std. deviation Green channel standard deviation of seed pixels
<i>B_{mean}</i>	Mean blue channel Blue channel mean value of seed pixels (grey levels)
<i>B_{sd}</i>	Blue std. deviation Blue channel standard deviation of seed pixels
<i>H_{mean}</i>	Mean hue channel Hue channel mean value of seed pixels (grey levels)
<i>H_{sd}</i>	Hue std. deviation Hue channel standard deviation of seed pixels
<i>L_{mean}</i>	Mean lightness channel Lightness channel mean value of seed pixels (grey levels)
<i>L_{sd}</i>	Lightness std. deviation Lightness channel standard deviation of seed pixels
<i>S_{mean}</i>	Mean saturation channel Saturation channel mean value of seed pixels (grey levels)
<i>S_{sd}</i>	Saturation std. deviation Saturation channel standard deviation of seed pixels
<i>D_{mean}</i>	Mean density Density channel mean value of seed pixels (grey levels)
<i>D_{sd}</i>	Density std. deviation Density channel standard deviation of seed pixels
<i>S</i>	Skewness Asymmetry degree of intensity values distribution (grey levels)
<i>K</i>	Kurtosis Peakness degree of intensity values distribution (densitometric units)
<i>H</i>	Energy Measure of the increasing intensity power (densitometric units)
<i>E</i>	Entropy Dispersion power (bit)
<i>D_{sum}</i>	Density sum Sum of density values of the seed pixels (grey levels)
<i>SqD_{sum}</i>	Square density sum Sum of the squares of density values (grey levels)

Statistical analysis

The achieved data were used to build a database including seed mean weight, morpho-colorimetric, EFDs and Haralick's descriptors. Statistical elaborations were executed using SPSS software package release 16.0 (SPSS Inc. for Windows, Chicago, Illinois, USA), and the stepwise Linear Discriminant Analysis method (LDA) was applied to identify and discriminate among the investigated *Cistus* accessions.

This approach is commonly used to classify/identify unknown groups characterized by quantitative and qualitative variables (Fisher, 1936; 1940; Sugiyama, 2007), finding the combination of predictor variables with the aim of minimizing the within-class distance and maximizing the between-class distance simultaneously, thus achieving maximum class discrimination (Hastie *et al.*, 2001; Holden *et al.*, 2011; Alvin & William, 2012; Kuhn & Johnson, 2013). The stepwise method identifies and selects the most statistically significant features among the 138 measured on each seed, using three statistical variables: Tolerance, *F*-to-enter and *F*-to-remove. The Tolerance value indicates the proportion of a variable variance not accounted for by other independent variables in the equation. *F*-to-enter and *F*-to-remove values define the power of each variable in the model and are useful to describe what happens if a variable is inserted and removed, respectively, from the current model. This method starts with a model that does not include any of the variables. At each step, the variable with the largest *F*-to-enter value that exceeds the entry criterion chosen ($F \geq 3.84$) is added to the model. The variables left out of the analysis at the last step have *F*-to-enter values smaller than 3.84, and therefore no more are added stopping the process (Venora *et al.*, 2009b; Grillo *et al.*, 2012). Finally, a cross-validation procedure was applied to verify the performance of the identification system,

testing individual unknown cases and classifying them on the basis of all others (SPSS, 2007).

All the raw data were standardized before starting any statistical elaboration. Moreover, in order to evaluate the quality of the discriminant functions achieved for each statistical comparison, the Wilks' Lambda, the percentage of explained variance and the canonical correlation between the discriminant functions and the group membership, were computed. The Box's M tests was executed to assess the homogeneity of covariance matrices of the features chosen by the stepwise LDA; while the analysis of the standardized residuals was performed to verify the homoscedasticity of the variance of the dependent variables used to discriminate among the groups' membership (Box, 1949; Haberman, 1973; Morrison, 2004). Kolmogorov-Smirnov's test was performed to compare the empirical distribution of the discriminant functions with the relative cumulative distribution function of the reference probability distribution, while the and Levene's test was executed to assess the equality of variances for the used discriminant functions calculated for groups' membership (Levene, 1960; Gastwirth *et al.*, 2009; Lopes, 2011).

To graphically highlight the differences among groups, multidimensional plots were drawn using the first three discriminant functions or, alternatively, when the number of discriminant groups n did not allow to obtain at least three discriminant functions ($n-1$), the two available discriminant functions and the Mahalanobis' square distance (Mahalanobis, 1936) were used. This measure of distance is defined by two or more discriminant functions and ranges from 0 to infinite. Samples are increasingly similar at values closer to zero. Higher values indicate that a particular case includes extreme values for one or more independent variables, and can be

considered significantly different to other cases of the same group (Bacchetta *et al.*, 2008a).

Results and Discussion

The discriminant analysis at species level showed an overall cross-validated classification performance of 80.6% (Table 4). *C. parviflorus* and *C. heterophyllus* showed the highest percentage of correct discrimination, recording values of 95% or higher. In contrast, *C. laurifolius* was correctly discriminated with a percentage of 60%, misclassified as *C. salviifolius* in the 13.8% of cases and as *C. creticus* and *C. albidus* for 11.3% and 8.5% respectively, in addition to less important erroneous attributions with other species.

Table 4. Percentage of correct identification at species level. In parenthesis, the number of analysed seeds.

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	Total
<i>C. albidus</i> (1)	77.6 (543)	-	0.7 (5)	-	2.7 (19)	-	2.9 (20)	4.3 (30)	-	11.6 (81)	0.3 (2)	100.0 (700)
<i>C. clusii</i> (2)	-	93.0 (555)	1.2 (7)	4.4 (26)	0.0	1.5 (9)	-	-	-	-	-	100.0 (597)
<i>C. creticus</i> (3)	0.3 (3)	0.5 (6)	92.2 (1014)	4.7 (52)	0.1 (1)	0.4 (4)	0.9 (10)	-	-	0.3 (6)	0.5 (4)	100.0 (1100)
<i>C. crispus</i> (4)	-	-	12.7 (49)	87.0 (336)	-	1.7 (5)	-	-	-	-	0.3 (1)	100.0 (386)
<i>C. heterophyllus</i> (5)	2.5 (5)	-	-	-	95.5 (191)	-	-	-	1.5 (3)	0.5 (1)	-	100.0 (200)
<i>C. ladanifer</i> (6)	-	4.4 (13)	0.7 (2)	4.0 (12)	-	90.9 (270)	-	-	-	-	-	100.0 (297)
<i>C. laurifolius</i> (7)	8.5 (34)	-	11.3 (45)	1.0 (4)	0.8 (3)	-	61.3 (245)	0.3 (1)	3.0 (12)	13.8 (55)	0.3 (1)	100.0 (400)
<i>C. monspelliensis</i> (8)	1.0 (7)	-	0.3 (2)	-	-	-	0.3 (2)	73.6 (515)	-	24.9 (174)	-	100.0 (700)
<i>C. populifolius</i> (9)	0.5 (1)	-	0.0	-	29.6 (59)	-	-	1.5 (3)	67.8 (135)	0.5 (1)	-	100.0 (199)
<i>C. salviifolius</i> (10)	2.0 (34)	0.9 (15)	5.5 (93)	2.7 (46)	0.1 (2)	-	2.7 (46)	13.3 (225)	0.4 (7)	72.2 (1226)	0.2 (3)	100.0 (1697)
<i>C. parviflorus</i> (11)	0.5 (1)	-	4.5 (9)	-	-	-	-	-	-	-	95.0 (189)	100.0 (199)
Overall												80.6 (6575)

The first discriminating ten variables at species level, chosen by the LDA discrimination process, are reported in Table 5. The seed mean weight represented the most powerful parameter, showing a significantly high value of *F*-to-remove. According to the achievements previously reached by Grillo *et al.* (2010), the other best 35 variables, selected over the available 138, were principally colorimetric features (RGB and HLS colour channels) and densitometric descriptors, in addition to a few of other dimensional parameters, explaining the wide within-species variability of seed size (Delgado *et al.*, 2008; Tavşanoğlu & Çatav, 2012). In the present analysis the EFDs not entered into the classification system used to discriminate *Cistus* species; this result is probably due to seed shape homogeneity that characterized all the investigated *taxa*.

Table 5. Tolerance, F-to-remove and Wilks' lambda values of the best ten key parameters chosen by the LDA to discriminate the 11 studied *Cistus* species.

	Tolerance	<i>F</i> to remove	Wilks λ
<i>SW</i>	0,647	1023,746	$8,11 \cdot 10^{-03}$
<i>G_{sd}</i>	0,016	233,773	$4,26 \cdot 10^{-03}$
<i>D_{sd}</i>	0,011	143,175	$3,82 \cdot 10^{-03}$
<i>D_{sum}</i>	0,014	140,214	$3,81 \cdot 10^{-03}$
<i>SqD_{sum}</i>	0,013	134,938	$3,78 \cdot 10^{-03}$
<i>A</i>	0,007	133,390	$3,78 \cdot 10^{-03}$
<i>D_{mean}</i>	0,002	131,830	$3,77 \cdot 10^{-03}$
<i>B_{sd}</i>	0,028	113,834	$3,68 \cdot 10^{-03}$
<i>L_{mean}</i>	0,001	98,198	$3,60 \cdot 10^{-03}$
<i>L_{sd}</i>	0,006	94,354	$3,59 \cdot 10^{-03}$

The analysis of *C. creticus* at infra-specific level showed an overall performance of 96.7% (Fig 1a), with misattributions mostly related to *C. creticus* subsp. *corsicus*, misclassified as *C. creticus* subsp. *eriocephalus* in 12% of cases (data not shown). The histogram of the standardized residuals

(Fig. 1*b*), the normal probability plot (Fig. 1*c*) and the dispersion plot of the standardized residuals (Fig. 1*d*) were also included to better understand the normal distribution of the data.

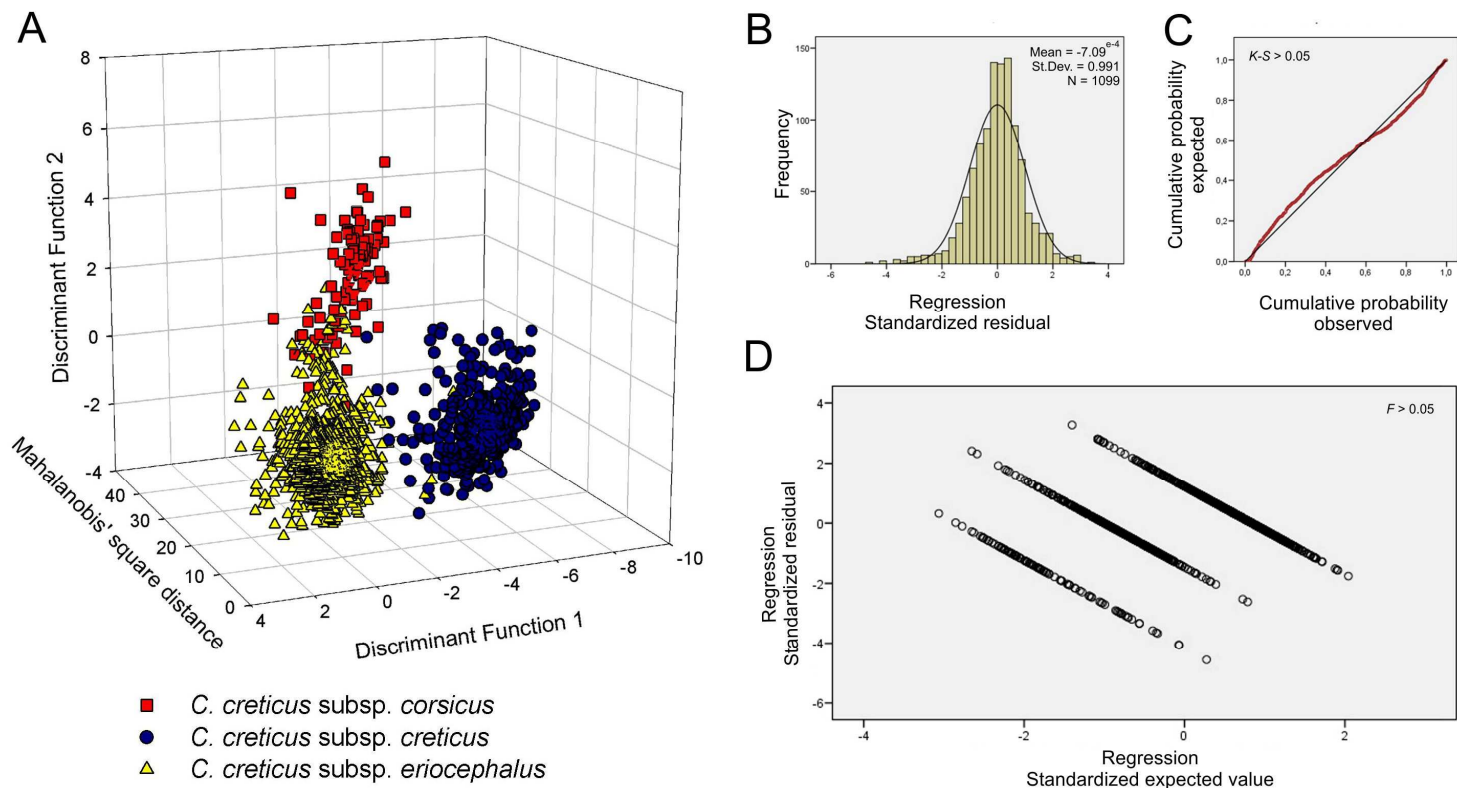


Figure 1. A) Discriminant scores 3D graphical representation of the three studied *Cistus creticus* subspecies; B) histogram of the standardized residuals; C) Normal Probability Plot (P-P) tested with the Kolmogorov-Smirnov's test (K-S); D) dispersion plot of the standardized residuals tested with the Levene's test (F).

According to the previous classification model, the seed mean weight was the most discriminant feature, even though the relative value of *F*-to-remove recorded minor statistical relevance (Table 6). As expected, this feature is comparable among *taxa* belonging to the same species, nevertheless statistically differences exist. Also in this case, the most discriminating parameters were descriptive of colorimetric and textural traits of the seed surface, but only 24 steps have been enough to identify these three *taxa* by means of LDA (Table 6). These results well fit with those obtained by Grillo *et al.* (2010), that similarly found seed mean weight as the most discriminant feature, followed by mean and standard deviation values of RGB colour channels.

Table 6. Tolerance, F-to-remove and Wilks' lambda values of the best ten key parameters chosen by the LDA to discriminate the three studied *Cistus creticus* subspecies.

	Tolerance	<i>F</i> to remove	Wilks λ
<i>SW</i>	0,242	97,060	$8,26 \cdot 10^{-02}$
<i>G_{sd}</i>	0,020	65,452	$7,84 \cdot 10^{-02}$
<i>B_{sd}</i>	0,018	64,436	$7,83 \cdot 10^{-02}$
<i>A</i>	0,029	58,085	$7,75 \cdot 10^{-02}$
<i>D_{sum}</i>	0,002	53,227	$7,68 \cdot 10^{-02}$
<i>SqD_{sum}</i>	0,003	40,636	$7,52 \cdot 10^{-02}$
<i>S</i>	0,504	38,748	$7,50 \cdot 10^{-02}$
<i>S_{sd}</i>	0,111	27,020	$7,34 \cdot 10^{-02}$
<i>B_{mean}</i>	0,005	24,603	$7,31 \cdot 10^{-02}$
<i>R_{mean}</i>	0,004	22,760	$7,29 \cdot 10^{-02}$

Regarding the variability among the *C. creticus* subsp. *eriocephalus* populations, the analysis achieved an overall cross validated discrimination percentage of 99.6% (Fig. 2a), ranged from 98% of correct discrimination for the Italian population of Pineta della Foce del Garigliano and 100% for three of the Sardinian populations (Agriuxiau, Portixeddu and Piscinamanna). Also for this discrimination model, the histogram of the standardized residuals (Fig. 2b), the normal probability plot (Fig. 2c) and the dispersion plot of the standardized residuals (Fig. 2d) were computed. This high differentiation among populations was achieved by a statistical model involving, among the 51 chosen variables, 25 Haralick's and EFDs descriptors, but unlike previous analyses, the seed mean weight was not the most important feature in the classification system (data not shown).

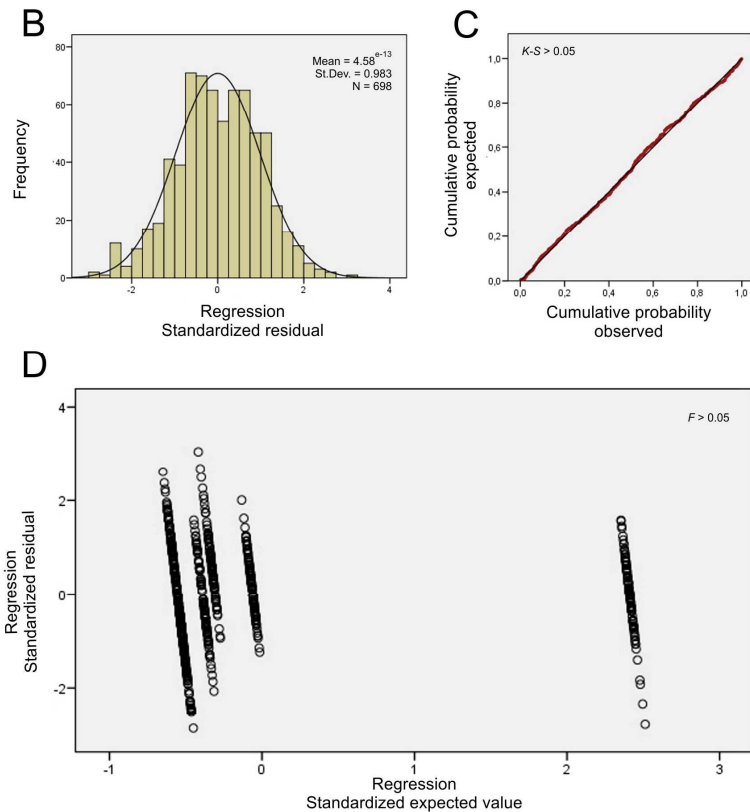
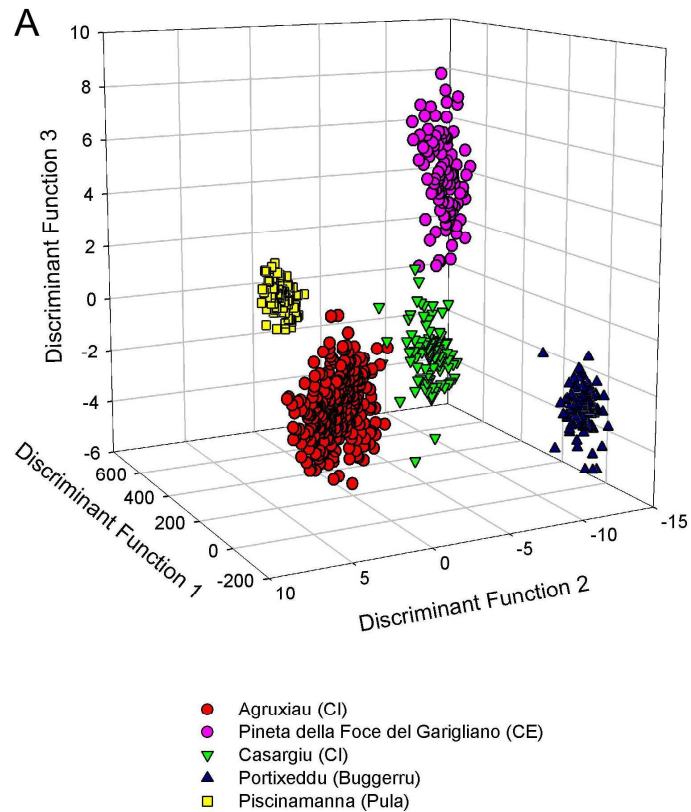


Figure 2. A) Discriminant scores 3D graphical representation of the five studied *Cistus eriocephalus* populations; B) histogram of the standardized residuals; C) Normal Probability Plot (P-P) tested with the Kolmogorov-Smirnov's test (K-S); D) dispersion plot of the standardized residuals tested with the Levene's test (F).

Definitively, as reported in Grillo *et al.* (2010), the seed mean weight proved to be the most discriminating feature both at specific and infra-specific level, but not at population level. Furthermore, the obtained data explain that parameters related to the surface, seed colour, density and distribution proved to be more discriminant than the size and shape ones. Effectively, the genus *Cistus* is not the only one showing these phenotypic characteristics. Grillo *et al.* (2010) found that the families of Cistaceae and Scrophulariaceae seem exclusively to show non-morphological features in the best key parameters able to discriminate at specific or infra-specific level. Similarly, according to Bacchetta *et al.* (2011a), the seeds of *Astragalus* sect. *Melanocercis*, showed chromatic features as the most discriminant.

Conclusions

The satisfactory discrimination performances reached by the statistical comparisons among *Cistus* species, subspecies and populations, on the basis of seed morpho-colorimetric data, agree with the results reported in the previous papers on the same *taxa* (Bacchetta *et al.*, 2008a; Grillo *et al.*, 2010). Comparing with the present study, the number of cases was raised, making more difficult the discrimination among *taxa* because of increased infra-specific variability. On the other hand, the improvement of the image analysis system adopted, in which an overall of 138 seed features was evaluated, allowed to reinforce the discrimination power. In addition, except the mean seed weight that resulted to be the most discriminant character in the comparisons conducted at species and infra-specific level, colorimetric and textural parameters resulted key variables in the statistical elaborations. The resulting effect of these developments led to a slight reduction of the overall performance but also to the ability of the system to discriminate among a larger amount of different *taxa*. Considering the high similarity in

seed morpho-colorimetric traits, it could be considered a good compromise for the development of an identification system based on seed characters, anyway affected by biotic and abiotic factors.

Finally, confirming the current taxonomic treatment accepted by *The Plant List* (2013) at the inter- and infra-specific levels, these achievements, prove that this method is effective also when the morphometric variability within each group should be extremely reduced such as in inter-population groups.

Acknowledgement

I would thank to Prof. Gianluigi Bacchetta, Dr. Oscar Grillo, Dr. Gianfranco Venora and Dr. Eva Canadas for their great contribution to the writing of this chapter. Thanks also to the researchers of all the institutions that kindly provided seed material: Dr. Christine Fournaraki of MAICH (Crete), Dr. Esteban Bermejo and Dr. Paqui Herrera Molina of Jardin Botanico de Cordoba, Dr. Caroline Favier of Conservatoire Botanique National de Corse, Prof. Salvatore Brullo of Dipartimento di Botanica, Università degli studi di Catania and Prof. Luigi Forte of Orto Botanico, Università degli studi di Bari. This research was supported by the “Provincia di Cagliari” and “Ente Foreste della Sardegna”.

References

- ALVIN C.R. & WILLIAM F.C. 2012. *Methods of Multivariate Analysis*. 3rd edition. John Wiley & Sons.
- BACCHETTA G., BUENO SANCHEZ A., FENU G., JIMENEZ-ALFARO B., MATTANA E., PIOTTO B., VIREVAIRE M. 2008b. *Conservacion ex situ de plantas silvestres*. Principado de Asturias / La Caixa.
- BACCHETTA G., FENU G., GRILLO O., MATTANA E., VENORA, G. 2011b. Identification of Sardinian species of *Astragalus* section *Melanocercis* (Fabaceae) by seed image analysis. *Annales Botanici Fennici* 48, 449-454.
- BACCHETTA G., GARCÍA P.E., GRILLO O., MASCIA F., VENORA G. 2011a. Seed image analysis provides evidence of taxonomical differentiation within the *Lavatera triloba* aggregate (Malvaceae). *Flora* 206, 468-472.
- BACCHETTA G., GRILLO O., MATTANA E., VENORA, G. 2008a. Morpho-colorimetric characterization by image analysis to identify diaspores of wild plant species. *Flora* 203, 669-682.
- BOX G.E.P. 1949. A general distribution theory for a class of likelihood criteria. *Biometrika* 36, 317-346.
- CATONI R., GRATANI L., VARONE L. 2012. Physiological, morphological and anatomical trait variations between winter and summer leaves of *Cistus* species. *Flora* 207, 442-449.
- CERABOLINI B., CERIANI R.M., CACCIANIGA M., DE ANDREIS R., RAIMONDI B. 2003. Seed size, shape and persistence in soil: a test on Italian flora from Alps to Mediterranean coasts. *Seed Science Research* 13, 75-85.
- DELGADO J.A., SERRANO J.M., LÓPEZ F., ACOSTA, F.J. 2008. Seed size and seed germination in the Mediterranean fire-prone shrub *Cistus ladanifer*. *Plant Ecology* 197, 269-276.
- DIAMOND J., ANDERSON N.H., BARTELS P.H., MONTIRONI R., HAMILTON P.W. 2004. The use of morphological characteristics and texture analysis in the identification of tissue composition in prostatic neoplasia. *Human Pathology* 35, 1121-1131.
- ELLUL P., BOSCAIU M., VICENTE O., MORENO V., ROSELLÓ, J.A. 2002. Intra- and interspecific variation in DNA content in *Cistus* (Cistaceae). *Annals of Botany* 90, 345-351.

- FERNÁNDEZ-MAZUECOS M. & VARGAS P. 2010. Ecological rather than geographical isolation dominates Quaternary formation of Mediterranean *Cistus* species. *Molecular Ecology* 19, 1381-1395.
- FERRER-GALLEGO P.P., LAGUNA E., CRESPO M.B. 2013. Typification of Linnaean names in *Cistus*. *Taxon* 62, 1046-1049.
- FIRATLIGIL-DURMUŞ E., ŠÁRKA E., BUBNÍK Z., SCHEJBAL M., KADLEC P. 2010. Size properties of legume seeds of different varieties using image analysis. *Journal of Food Engineering* 99, 445-451.
- FISHER R.A. 1936. The use of multiple measurements in taxonomic problems. *Annals of Eugenics* 7, 179-188.
- FISHER R.A. 1940. The precision of discriminant functions. *Annals of Eugenics* 10, 422-429.
- FUNES G., BASCONCELO S., DÍAZ S., CABIDO M. 1999. Seed size and shape are good predictors of seed persistence in soil in temperate mountain grasslands of Argentina. *Seed Science Research* 9, 341-345.
- GASTWIRTH J.L., GEL Y.R., MIAO W. 2009. The impact of Levene's test of equality of variances on statistical theory and practice. *Statistical Science* 24, 343-360.
- GERGER A. & SMOLLE J. 2003. Diagnostic imaging of melanocytic skin tumors. *Journal of Cutaneous Pathology* 30, 247-252.
- GRANITTO P.M., GARRALDA P.A., VERDES P.F., CECCATO H.A. 2003. Boosting classifiers for weed seeds identification. *Journal of Computer Science and Technology* 3, 34-39.
- GRILLO O., DRAPER D., VENORA G., MARTÍNEZ-LABORDE J.B. 2012. Seed image analysis and taxonomy of *Diplotaxis* DC. (Brassicaceae, Brassicaceae). *Systematic and Biodiversity* 10, 57-70.
- GRILLO O., MATTANA E., FENU G., VENORA G., BACCHETTA G. 2013. Geographic isolation affects inter- and intra-specific seed variability in the *Astragalus tragacantha* complex, as assessed by morpho-colorimetric analysis. *Comptes Rendus de Biologies* 336, 102-108.
- GRILLO O., MATTANA E., VENORA G., BACCHETTA, G. 2010. Statistical seed classifiers of 10 plant families representative of the Mediterranean vascular flora. *Seed Science and Technology* 38, 455-476.

- GRILLO O., MICELI C., VENORA G. 2011. Computerised image analysis applied to inspection of vetch seeds for varietal identification. *Seed Science and Technology* 39, 490-500.
- GUZMÁN B. & VARGAS P. 2005. Systematics, character evolution, and biogeography of *Cistus* L. (Cistaceae) based on ITS, trnL-trnF, and matK sequences. *Molecular Phylogenetics and Evolution* 37, 644-660.
- GUZMÁN B. & VARGAS P. 2009. Historical biogeography and character evolution of Cistaceae (Malvales) based on analysis of plastid rbcL and trnL-trnF sequences. *Organisms Diversity and Evolution* 9, 83-99.
- HABERMAN S.J. 1973. The analysis of residuals in cross-classified tables. *Biometrics* 29, 205-220.
- HARALICK R.M. & SHAPIRO L.G. 1991. Glossary of computer vision terms. *Pattern Recognition* 24, 69-93.
- HARALICK R.M., SHANMUGAM K., DINSTEN, I. 1973. Textural features for image classification. *IEEE Transactions on Systems, Man and Cybernetics* 3, 610-621.
- HÂRUTA O. 2011. Elliptic Fourier analysis of crown shapes in *Quercus petraea* trees. *Annals of Forest Research* 54, 99-117.
- HASTIE T., TIBSHIRANI R., FRIEDMAN J. 2001. *The elements of statistical learning: Data mining, inference, and prediction*. New York: Springer.
- HERNÁNDEZ-MARTÍNEZ M.Á., NÚÑEZ-COLÍN C.A., GUZMÁN-MALDONADO S.H., ESPINOSA-TRUJILLO E., HERRERA-HERNÁNDEZ M.G. 2011. Morphological variability by means of seed traits of populations of *Amelanchier denticulata* (Kunth) Koch, from Guanajuato, Mexico. *Chapingo Serie Horticultura* 17, 161-172.
- HOLDEN J.E., FINCH W.H., KELLY K. 2011. A Comparison of two-group classification methods. *Educational and Psychological Measurement* 715, 870-901.
- IWATA H., NESUMI H., NINOMIYA S., TAKANO Y., UKAI Y. 2002. Diallel analysis of leaf shape variations of *Citrus* varieties based on Elliptic Fourier Descriptors. *Breeding Science* 52, 89-94.
- IWATA H., NIIKURA S., SEIJI M., TAKANO Y., UKAI Y. 2004. Genetic control of root shape at different growth stages in Radish (*Raphanus sativus* L.). *Breeding Science* 54, 117-124.

- JEANMONOD D. & GAMISANS J. 2007. *Flora Corsica*. Edisud, Aix-en-Provence. pp. 581-583.
- KAWABATA S. YOKOO M., NII K. 2009. Quantitative analysis of corolla shapes and petal contours in single-flower cultivars of *Lisianthus*. *Scientia Horticulturae* 121, 206-212.
- KUHL F.P. & GIARDINA C.R. 1982. Elliptic Fourier features of a closed contour. *Computer Graphics* 18, 259-278.
- KUHN M. & JOHNSON K. 2013. Discriminant Analysis and Other Linear Classification Models. In: *Applied Predictive Modeling* pp. 275-328. Springer New York. ISBN: 978-1-4614-6848-6
- LEVENE H. 1960. Robust tests for equality of variances. In: Olkin, I., Ghurye, S.G., Hoeffding, W., Madow, W.G. & Mann H.B., Eds., *Contributions to Probability and Statistics: Essays in Honor of Harold Hotelling*. Stanford University Press. pp. 278-292.
- LOPES R.H.C. 2011. Kolmogorov-Smirnov Test. In: Lovric M., Eds., *International Encyclopedia of Statistical Science*. Springer Berlin Heidelberg. pp. 718-720.
- MAHALANOBIS P.C. 1936. On the generalized distance in statistics. *Proceedings of the National Institute of Science of India* 12, 49-55.
- MOREIRA B., TAVSANOGLU Ç., PAUSAS J.G. 2012: Local versus regional intraspecific variability in regeneration traits. *Oecologia* 168, 671-677.
- MORRISON, D.F. 2004. *Multivariate Statistical Methods*. 4th edition. Cengage Learning Duxbury Press.
- NANNI, L., SHI, J.Y., BRAHNAM, S., LUMINI A. 2010. Protein classification using texture descriptors extracted from the protein backbone image. *Journal of Theoretical Biology* 264, 1024-1032.
- ORRÙ M., GRILLO O., LOVICU G., VENORA G., BACCHETTA G. 2013. Morphological characterisation of *Vitis vinifera* L. seeds by image analysis and comparison with archaeological remains. *Vegetation History and Archaeobotany* 22, 231-242.
- ORRÙ, M., GRILLO, O., VENORA, G., BACCHETTA G. 2012. Computer vision as a complementary to molecular analysis: grapevines cultivars case study. *Comptes Rendus de Biologies* 335, 602-615.
- PAOLINI J., FALCHI A., QUILICHINI Y., DESJOBERT J.M., DE CIAN M.C., VARESI L., COSTA J. 2009. Morphological, chemical and genetic differentiation of two

- subspecies of *Cistus creticus* L. (*C. creticus* subsp. *eriocephalus* and *C. creticus* subsp. *corsicus*). *Phytochemistry* 70, 1146-1160.
- PAWLUCZYK M., WEISS J., VICENTE-COLOMER M.J., EGEEA-CORTINES M. 2012. Two alleles of *rpoB* and *rpoC1* distinguish an endemic European population from *Cistus heterophyllus* and its putative hybrid (*C. × clausonis*) with *C. albidus*. *Plant Systematics and Evolution* 298, 409-419.
- PÉREZ-GARCÍA F., GONZÁLEZ-BENITO M.E., GÓMEZ-CAMPO C., 2007. High viability recorded in ultradrying seeds of 37 species of Brassicaceae after almost 40 years of storage. *Seed Science and Technology* 35, 143-153.
- PINNA M.S., GRILLO O., MATTANA E., CAÑADAS E.M., BACCHETTA G. 2014. Inter- and intraspecific morphometric variability in *Juniperus* L. seeds (Cupressaceae). *Systematics and Biodiversity* 12, 211-223.
- ROVNER I. & GYULAI F., 2007. Computer-assisted morphometry: a new method for assessing and distinguishing morphological variation in wild and domestic seed populations. *Economic Botany* 61, 154-172.
- SAATKAMP A., AFFRE L., DUTOIT T., POSCHLOD P. 2009. The seed bank longevity index revisited: limited reliability evident from a burial experiment and database analyses. *Annals of Botany* 104, 715-724.
- SHAHIN M.A. & SYMONS S.J., 2003a. Lentil type identification using machine vision. *Canadian Biosystems Engineering* 45, 3.5-3.11.
- SHAHIN M.A. & SYMONS S.J., 2003b. Colour calibration of scanners for scanner independent grain grading. *Cereal Chemistry* 80, 285-289.
- SIMONET M. & ANSERAU P. 1939 The meiosis of two *Cistus* hybrids: *C. x hybridus* Pouri and *C. x Rodier Verg var. antipolintensis*. Dans. pp 1526-1527.
- SMYKALOVA I., GRILLO O., BJELKOVA M., HYBL M., VENORA G. 2011. Morpho-colorimetric traits of *Pisum* seeds measured by an image analysis system. *Seed Science and Technology* 39, 612-626.
- SMYKALOVA I., GRILLO O., BJELKOVA M., PAVELEK M., VENORA, G. 2013. Phenotypic evaluation of flax seeds by image analysis. *Industrial Crops and Products* 47, 232-238.
- SPSS. 2007. Base 16.0 Application Guide. Prentice Hall, USA, New Jersey.

- SUGIYAMA M. 2007. Dimensionality reduction of multimodal labeled data by local Fisher discriminant analysis. *The Journal of Machine Learning Research* 8, 1027-1061.
- TATTINI M., MATTEINI P., SARACINI E., TRAVERSI M.L., GIORDANO C., AGATI G. 2007. Morphology and biochemistry of non-glandular trichomes in *Cistus salviifolius* L. leaves growing in extreme habitats of the Mediterranean basin. *Plant Biology* 9, 411-419.
- TAVŞANOĞLU C. & SERTER ÇATAV S. 2012. Seed size explains within-population variability in post-fire germination of *Cistus salviifolius*. *Annales Botanici Fennici* 49, 331-340.
- TERRAL J., TABARD E., BOUBY L., IVORRA S., PASTOR T., FIGUEIRAL I., PICQ S., CHEVANCEJ.B., JUNG C., FABRE L., TARDY C., COMPAN M., BACILIERI R., LACOMBE T., THIS P. 2010. Evolution and history of grapevine (*Vitis vinifera*) under domestication: new morphometric perspectives to understand seed domestication syndrome and reveal origins of ancient European cultivars. *Annals of Botany* 105, 443-455.
- THE PLANT LIST 2013. Version 1.1. Published on the Internet; <http://www.theplantlist.org/> (accessed 19 August 2014).
- THOMPSON J.D. 2005. Plant Evolution in the Mediterranean. Oxford University Press, Oxford 288 pp.
- VENORA G., GRILLO O., RAVALLI C., CREMONINI R., 2009b. Identification of Italian landraces of bean (*Phaseolus vulgaris* L.) using an image analysis system. *Scientia Horticulturae* 121, 410-418.
- VENORA G., GRILLO O., SACCONI R. 2009a. Quality assessment of durum wheat storage centres in Sicily: Evaluation of vitreous, starchy and shrunken kernels using an image analysis system. *Journal Cereal Science* 49, 429-440.
- VENORA G., GRILLO O., SHAHIN M.A., SYMONS S.J., 2007. Identification of Sicilian landraces and Canadian cultivars of lentil using image analysis system. *Food Research International* 40, 161-166.
- WIESNEROVA D. & WIESNER L. 2008. Computer image analysis of seed shape and seed color for flax cultivar description. *Computers and Electronics in Agriculture* 61, 126-135.

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. *Annals of Botany* 94, 657-664.

**Seed image analysis provides evidence of taxonomical differentiation
within the *Medicago* L. sect. *Dendrotelis* (Fabaceae)**

Abstract

Morpho-colorimetric quantitative variables describing seed size, shape, colour and texture were analyzed using image analysis techniques, in order to evaluate the variability among *Medicago taxa* sect. *Dendrotelis* and verify the current taxonomical treatment which divide this section into three species: *M. arborea* L., *M. citrina* (Font Quer) Greuter and *M. strasseri* Greuter, Matthäs & H. Risse. Further comparisons were conducted to discriminate among populations and regions of provenance. Data obtained were statistically analysed applying stepwise Linear Discriminant Analysis method (LDA), recording an overall cross-validated classification performance of 100.0% at species level. With regard to inter-population comparisons, percentages of correct discrimination above 98% were achieved and high performance was recorded in the discrimination among *M. arborea taxa* distinguished by region of provenance. For each of these statistical comparisons, the best discriminant variables chosen by the stepwise LDA were related to colour and textural information. Finally, the obtained results confirmed the validity of the proposed method to be highly diagnostic in the statistically assessment of the morpho-colorimetric traits variability of *Medicago taxa* seeds, both for the taxonomic differentiation at specific levels and regional and populational groups.

Introduction

Medicago L. is a large genus of the legume family (Trifolieae, Fabaceae) that includes a big amount of agriculturally and economically important species (e.g. *M. sativa* L.). This genus comprises about 83-85 species, grouped in 12 sections (Bena, 2001; Lock, 2005; Small & Jomphe, 1989). The natural distribution area of the genus covers broad regions of Eurasia (Mediterranean Region and W to C Asia) and Northern Africa (Heyn, 1963; Lesins & Lesins, 1979; Mehregan *et al.*, 2002).

In contrast with the remaining 11 sections, formed by annual or perennial herbs, the section *Dendrotelis* (Vassilcz.) Lassen comprises woody shrubs showing physiological adaptations to water and salt-stressed environments (Chebbi *et al.*, 1994; Koning *et al.*, 2000; Sibole *et al.*, 2003, 2005). The unique features that characterize the section are the presence of perennial stems, showing annual rings of wood and bark produced by cambia (Small & Jomphe, 1989). One to three species, depending on systematic criteria (Bolòs & Vigo, 1974; Lesins & Lesins, 1979; Small & Jomphe, 1989; Sobrino *et al.*, 2000), have been recognized within this section: *M. arborea* L., *M. citrina* (Font Quer) Greuter and *M. strasseri* Greuter, Matthäs & H. Risse. All of them are restricted to rocky and cliff faces in coastal places of the Mediterranean Basin.

All the species of the section *Dendrotelis* are polyploid. *Medicago arborea* and *M. strasseri* are tetraploids ($2n=32$; Cluster *et al.*, 1996; Falistocco, 1987; González-Andrés *et al.*, 1999; Rosato *et al.*, 2008), whereas *M. citrina* is hexaploid ($2n=48$ chromosomes; Boscaiu *et al.*, 1997; Rosato *et al.*, 2008). Recently, molecular cytogenetic studies have supported the close relationships between *M. arborea* and *M. strasseri* that showed a single 45S rDNA locus and two 5S rDNA loci. By contrast, the hexaploid *M. citrina* could be distinguished from the tetraploid species by the presence of four 45S

rDNA and five 5S rDNA (Rosato *et al.*, 2008). These findings suggest that fine cytogenetic approaches could be relevant to assess the origins of the polyploidy in the section *Dendrotelis*, and to gain insights on the level of karyological distinctiveness, and hence evolutionary divergence, among the three species. The cytogenetic data indicate a clear evolutionary split in woody medics (tetraploid vs. hexaploid species), reflecting divergent patterns of karyological evolution.

Medicago arborea has been taxonomically recognized by all authors dealing with the genus. It has been widely cultivated as a forage plant in the Mediterranean region (Olives, 1969), and introduced as ornamental in other areas of Europe, North Africa and Asia, blurring the boundaries of its natural distribution. Some authors have suggested that this species was originally endemic to some small islets of the Aegean Sea, being later introduced throughout most places of its current Mediterranean range (Greuter, 1986).

Medicago strasseri is endemic of Crete, being known from only two limestone gorges in the central part of the island (Greuter *et al.*, 1982). It is closely related to *M. arborea* and some authors have included it within that species at the subspecific level (Sobrino *et al.*, 2000).

Medicago citrina is a Spanish endemic restricted to a few small islets surrounding the Balearic Islands (Alomar *et al.*, 1997), and the Valencian Community, where the species is represented at the volcanic Columbretes archipelago (Bolòs & Vigo, 1984), and a small islet by the East coast of the Iberian Peninsula (Serra *et al.*, 2001). It was first described as a pale, yellow-flowered variety of *M. arborea* (Font Quer, 1924), and concerns about its specific status have been reported by several authors (Bolòs & Vigo, 1974; Lesins & Lesins, 1979; Pérez-Bañón *et al.*, 2003; Small & Jomphe, 1989). Taken together, all the karyological data unequivocally support the recognition of *M. citrina* as a distinct species (Rosato & Rosselló, 2009).

Although no clear evidences have been published and further studies will be needed, we have noticed (personnel observations, E. Laguna and P. Ferrer-Gallego) that for some of the most apparent external, morphometric characters (size of leaves and flowers, size of flower pieces, legume shape), several differences can be easily appreciated, both *in situ* and *ex situ* collections of adult plants, for *M. citrina*. Juan (2002) demonstrated that the population of *M. citrina* from a small islet -‘Illot de la Mona’ or ‘Escull del Cap’- very close to the continent in NE of Alicante, shows clear morphometric differences with respect to the rest of the known native populations - Columbretes and Balearic islands - for the flower pieces. In addition, further studies both using AFLPs (Juan *et al.*, 2004) and mixed genetic-morphological analyses (Crespo *et al.*, 2008) have found genetic differences among the four main populations known for this species: Alicante, Columbretes, Cabrera and Ibiza islands. Besides the data and evidences for *M. citrina*, the cultivated plants of *M. arborea* also shows apparent differences in fruit shape and leaf colour (E. Laguna, pers. obs.).

The potential of biometric indices is well known and many authors exploited it for various studies related to seeds, particularly regarding morpho-colorimetric evaluations (e.g. Granitto *et al.*, 2003; Grillo *et al.*, 2011; Kiliç *et al.*, 2007; Shahin & Symons, 2003; Smykalova *et al.*, 2013; Venora *et al.*, 2007; Venora *et al.*, 2009a; Wiesnerova & Wiesner, 2008). Bacchetta *et al.* (2008) characterized seeds of wild vascular plants of the Mediterranean Basin, using digital images and implementing statistical classifiers able to discriminate seeds belonging to different genera and species. Then, Grillo *et al.* (2010) improved that classification system, developing 10 specific statistical classifiers at the family level for Angiosperms and testing the system on the genus *Juniperus* L. (Cupressaceae), demonstrating that the method is also reliable for

Gymnosperms. Recently, Pinna *et al.* (2014) focalized this topic on Mediterranean *taxa* of *Juniperus* at interspecific, specific and intraspecific levels, and shortly before Orrù *et al.* (2012) confirmed the effectiveness of this identification method, studying seeds of *Vitis vinifera* L. varieties. Afterwards, many authors have successfully used Elliptic Fourier Descriptors in seed studies (e.g. Mebatsion *et al.*, 2012; Orrù *et al.*, 2013; Terral *et al.*, 2010).

Applying the same technical approach, the aim of this study is to investigate the *Medicago* sect. *Dendrotelis* *taxa*, in order to evaluate the seed morpho-colorimetric variability, finding additional evidences to reinforce the most recent taxonomical treatment which divide this section into three species, as found by Juan *et al.* (2003) and Rosato *et al.* (2008) using molecular techniques.

Material and Methods

Seed collection and image acquisition

A total of 1295 seeds of 13 accessions belonging to three *Medicago* species of the sect. *Dendrotelis* and three accessions of the related *M. marina* L. from section *Medicago*, used as out-group, were collected, during a period of ten years, in Mediterranean Basin countries (Table 1) and stored in the Sardinian Germplasm Bank (BG-SAR), according to Bacchetta *et al.* (2008). Digital images of seeds were acquired using the same equipment and following the same procedure reported in Chapter I, and processed using the software package KS-400 V. 3.0 (Carl Zeiss, Vision, Oberkochen, Germany). A macro specifically developed for the characterization of seeds (Venora *et al.*, 2009b), was modified to perform automatically all the analysis procedures, reducing the execution time and contextually mistakes in the analysis process.

Table 1. Collecting years, localities, geographic coordinates and seed amount of the three studied *Medicago* species of the sect. *Dendrotelis*.

Taxon	Collecting year	Locality	Biogeographic Province according to Rivas-Martínez (2004)	Geographic coordinates		Seed number
				N	E	
<i>M. citrina</i>	2012	Columbretes (Spain)	Balearic	39°53'58.49''	0°48'0.88''	100
	2012	Illot de la Mona (Spain)	Valencian-Catalonian	38°48'0.88''	0°41'2.18''	100
		Illot de Ses Bledes (Balearic Islands)	Balearic	39° 8' 18.67''	2° 57' 41.93''	100
	2013					
<i>M. arborea</i>	2005	Castel Boccale (Italy)	Padanian	45°8'9.34''	10°54'54.90''	100
		Castel Boccale (Italy)	Padanian			100
	2006	Castel Boccale (Italy)	Padanian			100
		Parque Natural El Montgó (Spain)	Valencian-Catalonian	38°49'4.96''	0°5'43.74''	100
	2012	Capoterra (Sardinia)	Sardinian	39°10'14.79''	8°57'24.16''	100
		Nebida (Sardinia)	Sardinian	39°19'2.02''	8°26'13.24''	98
	2013	Chios (Greece)	Western Anatolian	38°21'35.85''	26°7'22.07''	100
	2006	Villefranche-sur-Mer (France)	Occitanian-Provençal	43°41'45.89''	7°18'15.89''	100
		Porquerolles (France)	Occitanian-Provençal	43°0'0.12''	6°12'13.12''	97
	2010					
	2005					
<i>M. strasseri</i>	2014	Crete (Greece)	Cretan	35°30'14.84''	24° 2'29.69''	100
<i>M. marina</i>	2009	Playa Mesqueda - Mallorca (Balearic Islands)	Balearic	39°44'55.87''	3°24'38.53''	100
		Is Arenas (Sardinia)	Sardinian	39°31'31.73''	8°26'24.61''	100
	2009	Castel Porziano (Italy)	Coastal West Italian	41°41'50.91''	12°23'20.67''	100

Morpho-colorimetric analysis: shape and texture descriptors

In order to improve the discrimination power, the macro was further enhanced adding algorithms able to compute the Elliptic Fourier Descriptors, hereafter EFDs (Iwata *et al.*, 2002, 2004; Kawabata *et al.*, 2009; Orrù *et al.*, 2012, 2013; Yoshioka *et al.*, 2004) for each analyzed seed. This method allows description of the boundary of the seed projection as an array of complex numbers which correspond to the pixel positions on the seed boundary. So, from the seed apex, defined as the starting point in a Cartesian system, chain codes are generated. A chain code is a lossless compression algorithm for binary images. The basic principle of chain codes is to separately encode each connected component (pixel) in the image. The encoder then moves along the boundary of the image and, at each step, transmits a symbol representing the direction of this movement.

This continues until the encoder returns to the starting position. This method is based on separate Fourier decompositions of the incremental changes of the X and Y coordinates as a function of the cumulative length along the boundary (Kuhl & Giardina, 1982). Each harmonic (n) corresponds to four coefficients (a_n , b_n , c_n and d_n) defining the ellipse in the XY plane. The coefficients of the first harmonic, describing the best fitting ellipse of outlines, are used to standardize size (surface area) and to orientate seeds (Terral *et al.*, 2010). According to Terral *et al.* (2010), about the use of a number of harmonics for an optimal description of seed outlines, in order to minimize the measurement errors and to optimize the efficiency of shape reconstruction, 20 harmonics were used to define the seed boundaries, obtaining a further 78 parameters useful to discriminate among the studied seeds (Orrù *et al.*, 2013).

Moreover, the macro was further improved adding algorithms able to compute 11 Haralick's descriptors and the relative standard deviations for

each analyzed seed (Lo Bianco *et al.*, submitted). These parameters are generally used when the objects in the images cannot be separated due to indefinite grey values variations. In these cases, the evaluation of texture, tone and context allows to define the spatial distribution of the image intensities and discrete tonal features. When a small area of the image has little variation of discrete tonal features, the dominant property of that area is grey tone. When a small area has wide variation of discrete tonal features, the dominant property of that area is texture (Haralick & Shapiro, 1991). According to Haralick *et al.* (1973), the concept of tone is based on varying shades of grey of resolution cells in a photographic image, while texture is concerned with the spatial (statistical) distribution of grey tones. Texture and tone are not independent concepts; rather, they bear an inextricable relationship to one another very much like the relationship between a particle and a wave. Context, texture and tone are always present in the image, although at times one property can dominate the others.

The basis for these features is the gray-level co-occurrence matrix (G in equation 1). This matrix is square with dimension N_g , where N_g is the number of gray levels in the image. Element [i, j] of the matrix is generated by counting the number of times a pixel with value i is adjacent to a pixel with value j and then dividing the entire matrix by the total number of such comparisons made. Each entry is therefore considered to be the probability that a pixel with value i will be found adjacent to a pixel of value j.

$$G = \begin{bmatrix} p(1,1) & p(1,2) & \cdots & p(1, N_g) \\ p(2,1) & p(2,2) & \cdots & p(2, N_g) \\ \vdots & \vdots & \ddots & \vdots \\ p(N_g, 1) & p(N_g, 2) & \cdots & p(N_g, N_g) \end{bmatrix} \quad (1)$$

In table 2, the 11 Haralick's descriptors measured on each seed to mathematically describe the surface texture, are reported.

Table 2. Haralick's descriptors measured as reported in Haralick et al. (1973).

	<i>Feature</i>	<i>Equation</i>
Har 1	Angular second moment	$\sum_i \sum_j p(i, j)^2$
Har 2	Contrast	$\sum_{n=0}^{N_g-1} n^2 \left\{ \sum_{i=1}^{N_g} \sum_{j=1}^{N_g} p(i, j) \right\}, i, j = n$
Har 3	Correlation	$\frac{\sum_i \sum_j (ij) p(i, j) - \mu_x \mu_y}{\sigma_x \sigma_y}$
		where μ_x , μ_y , σ_x and σ_y are the means and the standard deviations of p_x and p_y .
Har 4	Sum of square: variance	$\sum_i \sum_j (i - \mu)^2 p(i, j)$
Har 5	Inverse difference moment	$\sum_i \sum_j \frac{1}{1 + (i - j)^2} p(i, j)$
Har 6	Sum average	$\sum_{n=2}^{2N_g} i p_{x+y}(i)$
		where x and y are the coordinates (row and column) of an entry in the co-occurrence matrix, and $p_{x+y}(i)$ is the probability of co-occurrence matrix coordinates summing to $x+y$.
Har 7	Sum variance	$\sum_{i=2}^{2N_g} (i - f_B)^2 p_{x+y}(i)$
Har 8	Sum entropy	$-\sum_{i=2}^{2N_g} p_{x+y}(i) \log\{p_{x+y}(i)\} = f_B$
Har 9	Entropy	$-\sum_i \sum_j p(i, j) \log[p(i, j)]$
Har 10	Difference variance	$\sum_{n=0}^{N_g-1} i^2 p_{x-y}(i)$
Har 11	Difference entropy	$-\sum_{n=0}^{N_g-1} p_{x-y}(i) \log\{p_{x-y}(i)\}$

Mean seed weight and 137 morphometric, colorimetric and textural characters were measured on each seed (Table 3).

Table 3. List of morpho-colorimetric features measured on seeds, excluding the 78 Elliptic Fourier Descriptors calculated according to Háruta (2011).

Feature	Description
<i>A</i>	Area Seed area (mm ²)
<i>P</i>	Perimeter Seed perimeter (mm)
<i>P_{conv}</i>	Convex Perimeter Convex perimeter of the seed (mm)
<i>P_{Crof}</i>	Crofton's Perimeter Perimeter of the seed calculated using the Crofton's formula (mm)
<i>P_{conv}/P_{Crof}</i>	Perimeter ratio Ratio between convex and Crofton's perimeters
<i>D_{max}</i>	Max diameter Maximum diameter of the seed (mm)
<i>D_{min}</i>	Min diameter Minimum diameter of the seed (mm)
<i>D_{min}/D_{max}</i>	Feret ratio Ratio between minimum and maximum diameters
<i>Sf</i>	Shape Factor Seed shape descriptor = (4 x π x area)/perimeter ² (normalized value)
<i>Rf</i>	Roundness Factor Seed roundness descriptor = (4 x area)/(π x max diameter ²) (normalized value)
<i>Ecd</i>	Eq. circular diameter Diameter of a circle with an area equivalent to that of the seed (mm)
<i>EA_{max}</i>	Maximum ellipse axis Maximum axis of an ellipse with equivalent area (mm)
<i>EA_{min}</i>	Minimum ellipse axis Minimum axis of an ellipse with equivalent area (mm)
<i>R_{mean}</i>	Mean red channel Red channel mean value of seed pixels (grey levels)
<i>R_{sd}</i>	Red std. deviation Red channel standard deviation of seed pixels
<i>G_{mean}</i>	Mean green channel Green channel mean value of seed pixels (grey levels)
<i>G_{sd}</i>	Green std. deviation Green channel standard deviation of seed pixels
<i>B_{mean}</i>	Mean blue channel Blue channel mean value of seed pixels (grey levels)
<i>B_{sd}</i>	Blue std. deviation Blue channel standard deviation of seed pixels
<i>H_{mean}</i>	Mean hue channel Hue channel mean value of seed pixels (grey levels)
<i>H_{sd}</i>	Hue std. deviation Hue channel standard deviation of seed pixels
<i>L_{mean}</i>	Mean lightness channel Lightness channel mean value of seed pixels (grey levels)
<i>L_{sd}</i>	Lightness std. deviation Lightness channel standard deviation of seed pixels
<i>S_{mean}</i>	Mean saturation channel Saturation channel mean value of seed pixels (grey levels)
<i>S_{sd}</i>	Saturation std. deviation Saturation channel standard deviation of seed pixels
<i>D_{mean}</i>	Mean density Density channel mean value of seed pixels (grey levels)
<i>D_{sd}</i>	Density std. deviation Density channel standard deviation of seed pixels
<i>S</i>	Skewness Asymmetry degree of intensity values distribution (grey levels)
<i>K</i>	Kurtosis Peakness degree of intensity values distribution (densitometric units)
<i>H</i>	Energy Measure of the increasing intensity power (densitometric units)
<i>E</i>	Entropy Dispersion power (bit)
<i>D_{sum}</i>	Density sum Sum of density values of the seed pixels (grey levels)
<i>SqD_{sum}</i>	Square density sum Sum of the squares of density values (grey levels)

Statistical analysis

The achieved results were used to build a database including morpho-colorimetric, EFDs and Haralick's descriptors. Statistical elaborations were executed using SPSS software package release 15 (SPSS, 2007), and the stepwise Linear Discriminant Analysis (LDA) method was applied to identify and discriminate among the investigated *Medicago* accessions.

This approach is commonly used to classify/identify unknown groups characterized by quantitative and qualitative variables (Duda *et al.*, 2000; Fisher, 1936, 1940; Fukunaga, 1990), finding the combination of predictor variables with the aim of minimizing the within-class distance and maximizing the between-class distance simultaneously, thus achieving maximum class discrimination (Hastie *et al.*, 2001; Holden *et al.*, 2011; Kuhn & Johnson, 2013; Rencher & Christensen, 2012). Then, the stepwise procedure identifies and selects the most statistically significant features among the 137 measured on each seed (Grillo *et al.*, 2012; Venora *et al.*, 2009a). Finally, a cross-validation procedure was applied to verify the performance of the identification system, testing individual unknown cases and classifying them on the basis of all others (SPSS, 2007).

To graphically highlight the differences among groups, multidimensional plots were drawn using the first three discriminant functions or, alternatively, when the number of discriminant groups n did not allow to obtain at least three discriminant functions ($n-1$), bidimensional plots were drawn. To represent the morpho-colorimetric variability among taxonomical groups, box plots were drawn using the Mahalanobis' square distance values (Mahalanobis, 1936). This measure of distance is defined by two or more discriminant functions and ranges from 0 to infinite. Samples are increasingly similar at values closer to zero. Higher values indicate that a particular case includes extreme values for one or more independent

variables, and can be considered significantly different to other cases of the same group (Bacchetta *et al.*, 2008).

Results

Data obtained by measuring mean seed weight and 137 morpho-colorimetric quantitative variables describing seed size, shape and colour, were analysed by stepwise LDA, and statistical classifiers were developed in order to distinguish the three studied *taxa*.

The three *taxa* belonging to the sect. *Dendrotelis* were perfectly identified and classified (data not shown). Figure 1 report a graphical distribution of the three taxonomical groups on the basis of the two available discriminant function. In order to validate the comparison among the three studied *taxa*, *M. marina* was included in this analysis as out-group, resulting perfectly distinguished from the other species of the section *Dendrotelis*.

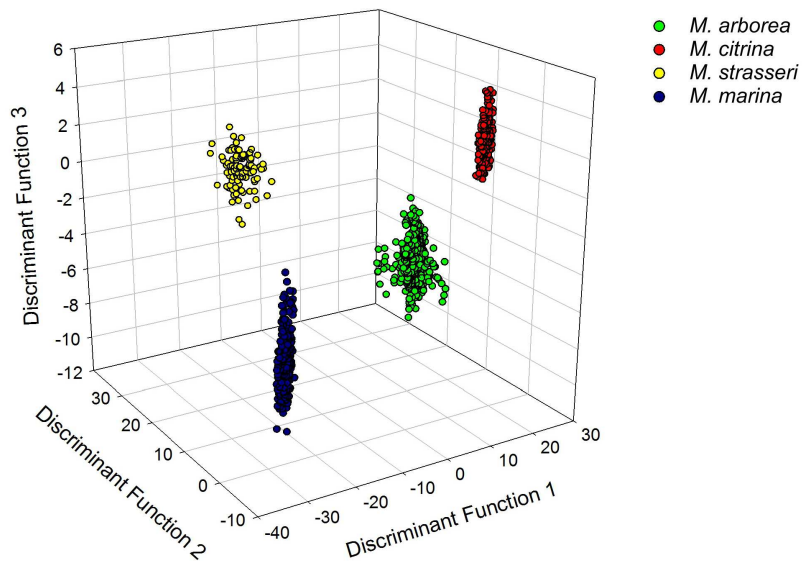


Figure 1. Graphical representation of the discriminant analysis for *Medicago* species.

A further comparison was implemented between *M. arborea* and *M. citrina*, in order to assess their taxonomic identity or their similarity level without the influence of *M. strasseri*, because it has a distribution area non-overlapping with the other two *taxa*. Also in this case, a perfect discrimination was achieved (data not shown). To evaluate and compare the specific morpho-colorimetric variability of these two *taxa*, the Mahalanobis' square distance values were used highlighting the spatial dispersion between *M. arborea* and *M. citrina* species (Fig. 2).

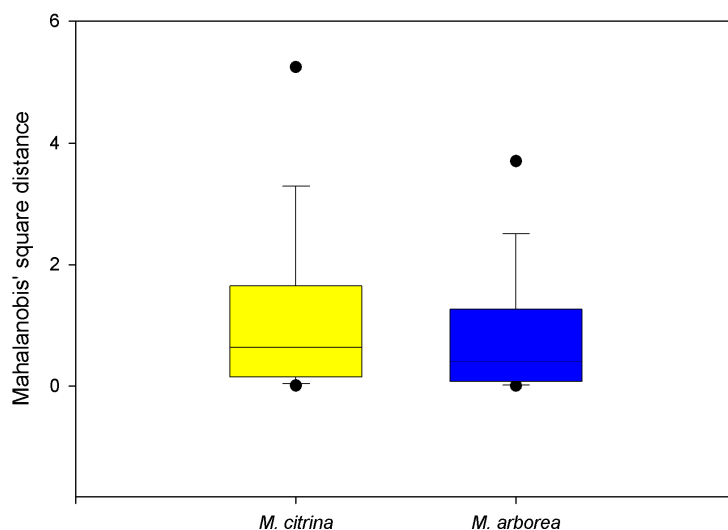


Figure 2. Graphic representation of the Mahalanobis' square distance values of *M. arborea* and *M. citrina* species.

Analyzing in more detail the relationship between *M. arborea* and *M. citrina*, the seeds of the available Spanish populations - mainland Spain and close archipelagos, excluding Balearic Islands - were compared. The *M. citrina* populations of Columbretes and Illot de la Mona, and that of *M. arborea* from Parque Natural El Montgó, resulted perfectly distinguishable

with an overall cross-validated percentage of correct identification of 100.0% (Table 4). Applying the same statistical model, the seeds of the two populations of *M. citrina* were grouped, incrementing the within variability of this species group, and compared with the Spanish population of *M. arborea*. One more time, 100.0% of correct identification was reached at species level (data not shown).

Table 4. Percentage of correct identification among *M. citrina* and *M. arborea* Spanish populations. In parenthesis, the number of seeds analyzed.

	<i>M. citrina</i> Columbretes (Castellon)	<i>M. citrina</i> Illot de la Mona (Alicante)	<i>M. arborea</i> Parque Natural El Montgó (Alicante)	Total
<i>M. citrina</i> Columbretes (Castellon)	100.0 (100)	-	-	100.0 (100)
<i>M. citrina</i> Illot de la Mona (Alicante)	-	100.0 (100)	-	100.0 (100)
<i>M. arborea</i> Parque Natural el Montgó (Alicante)	-	-	100.0 (100)	100.0 (100)
Overall				100.0 (300)

A separate comparison was carried out among the three studied populations of *M. citrina* to evaluate the inter-population variability (Fig. 3).

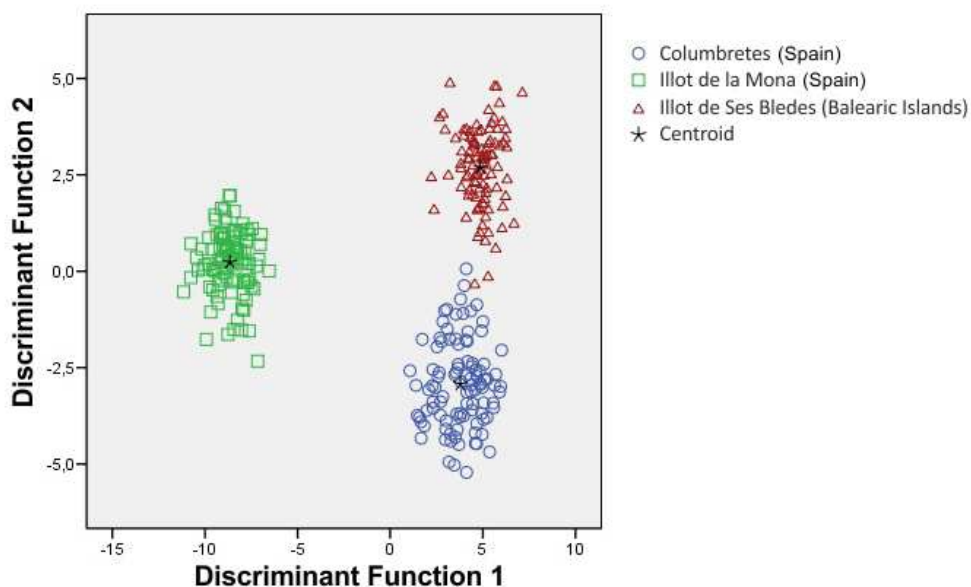


Figure 3. Graphic representation of the discriminant function scores for the three population of *M. citrina*.

Also in this case, high performances were obtained, achieving a percentage of correct classification of 99.0%. The population of Illot de la Mona (Spain) was perfectly discriminated, while slight misclassifications were recorded between *M. citrina* seeds from Columbretes (Spain) and from Illot de Ses Bledes (Balearic Islands), correctly distinguished in 99.0 and 98.0% of the cases, respectively (data not shown).

Furthermore, *M. arborea taxa* were compared, distinguishing them by region of provenance and reaching 94.2% of overall accuracy (Table 5, Fig. 4). In particular, all the seeds from Spain were correctly assigned, and no seed from other regions was mistakenly assigned to this group. Seeds from Sardinia were correctly discriminated with 93.4% accuracy, erroneously attributed to Italy and France in 5.6% and 1.0% of the cases, respectively. Seeds from Italy were correctly discriminated with 94.0% accuracy, equally distributing the misclassified cases with Sardinian and French seed groups. A satisfactory result was obtained for seeds from Greece, which percentage of correct classification was 98.0%, while only 2.0% was wrongly classified as seed from Italy. Contrastingly, seeds from France showed the lowest percentage of correct assignment (90.4%), due to the assignment of 6.6% of them to Italy seed group, 2.5% to Sardinia group and 0.5% to Greece one.

Table 5. Percentage of correct identification among *M. arborea taxa*, carried out for regions of provenance. In parenthesis, the number of seeds analyzed.

	Sardinia	Italy	Spain	Greece	France	Total
Sardinia	93.4 (185)	5.6 (11)	-	-	1.0 (2)	100.0 (198)
Italy	3.0 (9)	94.0 (282)	-	-	3.0 (9)	100.0 (300)
Spain	-	-	100.0 (100)	-	-	100.0 (100)
Greece	-	2.0 (2)	-	98.0 (98)	-	100.0 (100)
France	2.5 (5)	6.6 (13)	-	0.5 (1)	90.4 (178)	100.0 (197)
Overall						94.2 (895)

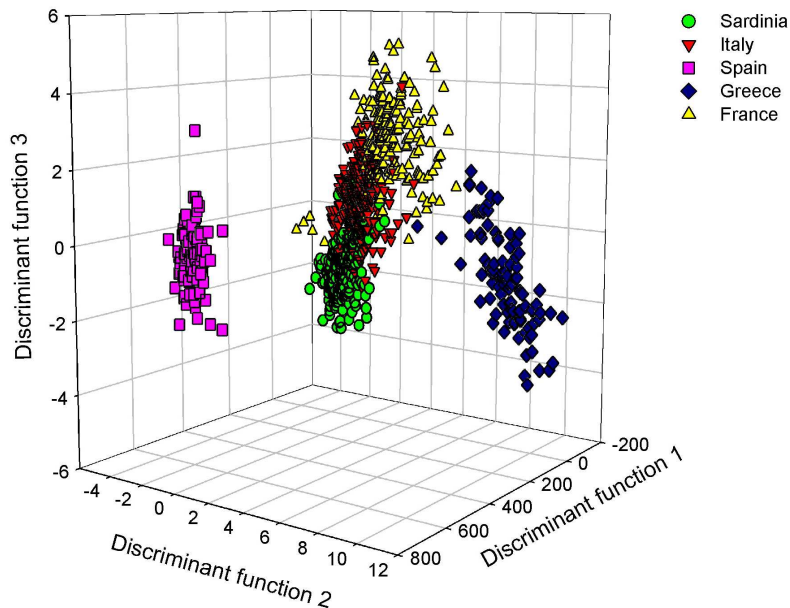


Figure 4. Graphic representation of the discriminant function scores for the five regions of provenance of *M. arborea*.

For each of these statistical comparisons, the best five discriminant variables chosen by the stepwise method are shown in Table 6. In the evaluation of the parameters that more than other influenced the discrimination process of the *Medicago taxa*, the most important variables were related to colour and textural information. Only for the first two classifiers at species level, the mean seed weight represented the most powerful parameter used in the discriminant functions, showing a significantly high value of F-to-remove (Table 6). For the other processing analyses, this parameter was not present in the discriminant function or appeared after the first ten most important ones, as occurred in the discrimination of *M. arborea taxa* by geographical region. In this last, as well as in the comparison carried out to correlate the Spanish populations of *M. arborea* and *M. citrina*, the Haralick's descriptors were found to be particularly powerful among the best five key parameters (Table 6).

Table 6. Number of groups, discriminant steps and performance of identification. Ranking of the best five discriminant parameters and the percentage of correct classification are reported for each of the implemented statistical comparisons (C1 = comparison among *M. arborea*, *M. citrina* and *M. strasseri*; C2 = comparison between *M. arborea* and *M. citrina*; C3 = comparison among the Spanish population of *M. arborea* and *M. citrina*; C4 = comparison among the *M. arborea* taxa distinguished by region of provenance; C5 = comparison among the *M. citrina* populations).

	C1	C2	C3	C4	C5
N of groups	3	2	3	5	3
N of steps	28	23	29	47	27
1st discriminant parameter	<i>Weight</i> (0.29; 10585.55; 0.13)	<i>Weight</i> (0.36; 3910.32; 9.17 ^{e-6})	<i>Har₆</i> (0.03; 8861.38; 9.91 ^{e-6})	<i>S_{sd}</i> (0.06; 169.42; 0.01)	<i>Har₆</i> (0.03; 5060.28; 5.51 ^{e-6})
2nd discriminant parameter	<i>G_{mean}</i> (0.03; 166.56; 0.02)	<i>E</i> (0.01; 138.56; 1.12 ^{e-6})	<i>Har₁₀</i> (0.01; 437.41; 6.30 ^{e-6})	<i>S_{mean}</i> (0.02; 105.16; 0.01)	<i>Har₁₀</i> (0.01; 591.27; 8.39 ^{e-6})
3rd discriminant parameter	<i>R_{sd}</i> (0.01; 152.17; 0.02)	<i>Har₇</i> (0.02; 5.37; 9.17 ^{e-6})	<i>S_{mean}</i> (0.04; 280.76; 4.58 ^{e-6})	<i>S</i> (0.43; 66.92; 0.01)	<i>Har₉</i> (0.03; 387.09; 6.25 ^{e-6})
4th discriminant parameter	<i>L_{mean}</i> (0.01; 120.88; 0.01)	<i>SqD_{sum}</i> (0.01; 109.86; 1.06 ^{e-6})	<i>Har₉</i> (0.03; 272.48; 4.48 ^{e-6})	<i>Har₁₀</i> (0.17; 41.77; 0.01)	<i>Har_{sd10}</i> (0.08; 108.67; 3.34 ^{e-6})
5th discriminant parameter	<i>L_{sd}</i> (0.01; 152.17; 0.02)	<i>G_{sd}</i> (0.10; 103.57; 1.05 ^{e-6})	<i>S_{sd}</i> (0.08; 239.08; 4.12 ^{e-6})	<i>E</i> (0.01; 35.75; 0.01)	<i>Har_{sd4}</i> (0.16; 106.10; 3.31 ^{e-6})
Percentage of correct identification between groups	100.0%	100.0%	100.0%	94.2%	99.0%

For each parameter, the tolerance, F-to-remove and Wilks' lambda values are reported in brackets.

Discussion

With the aim of confirm the most recent taxonomical treatment that includes into the section *Dendrotelis* three species of *Medicago*, seed morpho-colorimetric variability of *M. arborea*, *M. citrina* and *M. strasseri* was investigated applying image analysis techniques to extract accurate measurements and the LDA to statistically compare them.

According to the first preliminary comparison, the three studied species were perfectly distinguishable (Fig. 1). This satisfactory discrimination agrees with the results reported by Juan *et al.* (2003) and Rosato *et al.* (2008) on the basis of molecular techniques, confirming the current taxonomic treatment at the section level.

Considering the non-overlapping geographical distribution of *M. strasseri* respect to that the couple *M. arborea* and *M. citrina*, a comparison was conducted between these last two species, highlighting the marked morpho-colorimetric differentiation between them and confirming the clear taxonomic distance between these species. Molecular cytogenetic studies have been relevant to assess the recognition of *M. citrina* (hexaploid) as a distinct species relating to *M. arborea* and *M. strasseri* too (tetraploids) (Boscaiu *et al.*, 1997; Cluster *et al.*, 1996; Falistocco, 1987; González-Andrés *et al.*, 1999; Rosato *et al.*, 2008). Moreover, analyzing the spatial dispersion of the statistical cases by the Mahalanobis' square distance values, it is possible to deduce that the inter-specific morpho-colorimetric variability of the *M. citrina* is sensibly higher than *M. arborea*, although *M. arborea* seed sample is numerically more conspicuous than *M. citrina* (Fig. 2).

To better understand the relationship between *M. arborea* and *M. citrina* a comparison among the three Spanish populations of these two species was implemented, giving a perfect correct identification also in this case. This achievement suggests that seed colour and texture descriptors are able to discriminate among different populations from the same geographical area (Tab. 4). These results support as reported by Juan *et al.* (2003) on flower pieces, demonstrating that the population of *M. citrina* from Illot de la Mona shows clear morphometric differences with respect to the Columbretes and Balearic Islands populations. Furthermore, the same authors (Juan *et al.*, 2004) and later Crespo *et al.* (2008) found genetic differences between the four main populations known for this species by means of mixed genetic and morphological analyses. Similar results were reached comparing the three available populations of *M. citrina*, supporting these findings (Fig. 3).

In order to analyze more in detail the *M. arborea* variability, a comparison among the region of provenience was conducted (Table 5, Fig. 4). The perfect identification obtained for the seeds from Spain should indicate that these populations are independent from the morphological, genetic and evolutive points of view, respect to the populations from the other studied Mediterranean regions. On the other hand, the misidentifications highlighted among the seeds from Sardinia, Italy and France, although in low percentages, could suggest a certain connection among these regions and a consequent genetic flow mirrored in some morpho-colorimetric seed characters. The little misattribution of 2% of the seeds from Greece as Italy should support this idea, allowing to presume a plausible link with Italy.

By evaluating the contribution of the variables, using the discrimination algorithm (LDA), it was possible to identify the features that, more than others, were relevant for the discrimination of the *Medicago taxa* studied. At specific and intraspecific level, parameters related to the seed colour and texture proved to be more discriminant than the size or shape-descriptive ones.

One important difference among the three species of section *Dendrotelis* is the weight, necessarily related to size of seeds, larger in *M. citrina*, intermediate in *M. arborea* and smaller in *M. strasseri* (Greuter *et al.*, 1982; Robledo *et al.*, 1993). Furthermore, as reported by González-Andrés (1999) the seeds of *M. citrina* seems to be more rounded, the seeds of *M. strasseri* have a longer hilum compared with the total length of the seed, and the hilum angle is higher in *M. citrina*, lower in *M. strasseri* and intermediate in *M. arborea*. Nevertheless, features describing of the seed shape don't appear among the first five discriminant parameters (Table 6).

On the other hand, the recent literature proves that the study of surface texture of an object, whatever its nature, seems to be of great importance for its characterization (Diamond *et al.*, 2004; Gerger & Smolle, 2004; Nanni *et al.*, 2010). The results reported in table 6 confirm this assumption. Except the mean seed weight that resulted to be the most discriminant character in the two comparisons conducted at species level, only colorimetric and textural parameters appear among the best five for each executed statistical comparison. This achievement highlights the importance of the introduction of these descriptors, improving the image analysis system previously developed by Grillo *et al.* (2010) in which morphometric features were the first discriminant parameters. Also in Bacchetta *et al.* (2011), regarding the *Lavatera triloba* aggregate, the first three parameters with the highest discriminatory power were of morphological type, although colour evaluation was very important in this work for correct seed identification. The present results confirmed the validity of the proposed method for the taxonomic differentiation of *Medicago* at specific levels, and its identification capability of regional and populational groups.

Acknowledgement

I would thank to Prof. Gianluigi Bacchetta, Dr. Oscar Grillo, Dr. Gianfranco Venora for their great contribution to the writing of this chapter. Thanks also to Pablo Ferrer-Gallego and Emilio Laguna of the *Centro para la Investigación y Experimentación Forestal, Servicio de Vida Silvestre, (Spain)* for their suggestions and kind collaboration at this study.

References

- ALOMAR G., MUS M., ROSSELLÓ J.A. 1997. *Flora endèmica de les Balears*. Palma de Mallorca: Consell Insular de Mallorca.
- BACCHETTA G., GARCÌA P.E. GRILLO O., MASCIA F., VENORA G. 2011. Seed image analysis provides evidence of taxonomical differentiation within the *Lavatera triloba* aggregate (Malvaceae). *Flora* 206, 468-472.
- BACCHETTA G., GRILLO O., MATTANA E., VENORA G. 2008. Morpho-colorimetric characterization by image analysis to identify diaspores of wild plant species. *Flora* 203, 669-682.
- BENA G. 2001. Molecular phylogeny supports the morphologically based taxonomic transfer of the “medicagoid” *Trigonella* species to the genus *Medicago*. *Plant Systematic and Evolution* 229, 217-236.
- BOLÒS O. & VIGO J. 1974. Notes sobre taxonomia i nomenclatura de les plantes. I. *Butlletí de la Institució Catalana d’Història Natural* 38, 61-69.
- BOLÒS O. & VIGO J. 1984. *Flora dels Països Catalans*. Barcelona: Editorial Barcino.
- BOSCAIU M., RIERA J., ESTRELLES E., GÜEMES J. 1997. Números cromosòmics de plantes occidentals, 751-776. *Anales Jardín Botánico Madrid* 55, 430-431.
- CHEBBI H., RÍOS S., PASCUAL-VILLALOBOS M.J., CORREAL E. 1994. El grupo *Medicago arborea* en la cuenca Mediterrànea. II. Comportamiento frente a la sequía. *Pastos* 24, 177-188.
- CLUSTER P.D., CALDERINI O., PUPILLI F., CREA F., DAMIANI F., ARCIONI S. 1996. The fate of ribosomal genes in three interspecific somatic hybrids of *Medicago sativa*: three different outcomes including the rapid amplification of new spacer-length variants. *Theoretical and Applied Genetics* 93, 801-808.
- CRESPO M.B., JUAN A., ALONSO M.A., MARTÍNEZ-FLORES F., MARTINEZ-AZORÍN M. 2008. Biodiversidad vegetal del Parque Nacional de Cabrera: Biología de la conservación y diseño de estrategias de gestión de endemismos vasculares insulares. In *Red de Parques Nacionales: Proyectos de investigación en parques nacionales: 2003-2006*, (pp. 129-148). Madrid: Ministerio de Medio Ambiente y Medio Rural y Marino.
- DIAMOND J., ANDERSON N.H., BARTELS P.H., MONTIRONI R., HAMILTON P.W. 2004. The use of morphological characteristics and texture analysis in the identification of tissue composition in prostatic neoplasia. *Human Pathology* 35,1121-1131.

- DUDA R., HART P., STORK D. 2000. *Pattern classification* (2nd ed.). Hoboken: Wiley.
- FALISTOCCO E. 1987. Cytogenetic investigations and karyological relationships of two *Medicago*: *M. sativa* L. (alfalfa) and *M. arborea* L. *Caryologia* 40, 339-346.
- FISHER R.A. 1936. The use of multiple measurements in taxonomic problems. *Annals of Eugenics* 7, 179-188.
- FISHER R.A. 1940. The precision of discriminant functions. *Annals of Eugenics* 10, 422-429.
- FONT QUER P. 1924. Formes noves de plantes. Memòries del Museu de Ciències Naturals de Barcelona. *Sèrie Botànica* 1, 7-14, pl. I-V.
- FUKUNAGA K. 1990. *Introduction to statistical pattern classification*. San Diego: Academic Press.
- GERGER A., & SMOLLE J. 2003. Diagnostic imaging of melanocytic skin tumors. *Journal of Cutaneous Pathology* 30, 247-252.
- GONZÁLEZ-ANDRÉS F., CHÁVEZ J., MONTAÑEZ G., CERESUELA J.L. 1999. Characterization of woody *Medicago* (sect. *Dendrotelis*) species, on the basis of seed and seedling morphometry. *Genetic Resources and Crop Evolution* 46, 505-519.
- GRANITTO P.M., GARRALDA P.A., VERDES P.F., CECCATO H.A. 2003. Boosting classifiers for weed seeds identification. *Journal of Computer Science and Technology* 3, 34-39.
- GREUTER W. 1986. *Medicago citrina* (Font Quer) Greuter. *Willdenowia* 16, 112.
- GREUTER W., MATTHÄS U., RISSE H. 1982. Notes on Caradaegan plants. 3. *Medicago strasserii*, a new leguminous shrub from Kriti. *Willdenowia* 12, 201-206.
- GRILLO O., DRAPER D., VERONA G., MARTÍNEZ-LABORDE J.B. 2012. Seed image analysis and taxonomy of *Diplotaxis* DC. (Brassicaceae, Brassicaceae). *Systematics and Biodiversity* 10, 57-70.
- GRILLO O., MATTANA E., VENORA G., BACCHETTA G. 2010. Statistical seed classifiers of 10 plant families representative of the Mediterranean vascular flora. *Seed Science and Technology* 38, 455-476.
- GRILLO O., MICELI C., VENORA G. 2011. Computerised image analysis applied to inspection of vetch seeds for varietal identification. *Seed Science and Technology* 39, 490-500.

- HARALICK R.M. & SHAPIRO L.G. 1991. Glossary of computer vision terms. *Pattern Recognition* 24, 69-93.
- HARALICK R.M., SHANMUGAM K., & DINSTEN I. 1973. Textural features for image classification. *IEEE Transactions on Systems, Man and Cybernetics* 3, 610-621.
- HÂRUTA O. 2011. Elliptic Fourier analysis of crown shapes in *Quercus petraea* trees. *Annals of Forest Research* 54, 99-117.
- HASTIE T., TIBSHIRANI R., FRIEDMAN J. 2001. *The elements of statistical learning: Data mining, inference, and prediction*. New York: Springer.
- HEYN C.C. 1963. *The annual species of Medicago. Scripta Hierosol* 12. Jerusalem: Hebrew University.
- HOLDEN J.E., FINCH W.H., KELLY K. 2011. A Comparison of two-group classification methods. *Educational and Psychological Measurement* 715, 870-901.
- IWATA H., NESUMI H., NINOMIYA S., TAKANO Y., UKAI Y. 2002. Diallel analysis of leaf shape variations of *Citrus* varieties based on Elliptic Fourier Descriptors. *Breeding Science* 52, 89-94.
- IWATA H., NIIKURA S., SEIJI M., TAKANO Y., UKAI Y. 2004. Genetic control of root shape at different growth stages in radish (*Raphanus sativus* L.). *Breeding Science* 54, 117-124.
- JUAN A. 2002. *Estudio sobre la morfología, variabilidad molecular y biología reproductiva de Medicago citrina (Font Quer) Greuter (Leguminosae). Bases para su conservación*. (Unpublished doctoral dissertation). Universidad de Alicante, Alicante, Spain.
- JUAN A., CRESPO M.B., RÍOS S. 2003. Remarks on *Medicago citrina* (sect. *Dendrotelis*, Leguminosae). *Flora Mediterranea* 13, 303-316.
- JUAN A., CRESPO M.B., COWAN R.S., LEXER C. FAY M.F. 2004. Patterns of variability and gene flow in *Medicago citrina*, an endangered endemic of islands in the western Mediterranean, as revealed by amplified fragment length polymorphism AFLP. *Molecular Ecology* 13, 2679-2690.
- KAWABATA S., YOKOO M. NII K. 2009. Quantitative analysis of corolla shapes and petal contours in single-flower cultivars of *Lisianthus*. *Scientia Horticulturae* 121, 206-212.

- KILIÇ K., BOYACI I.H., KOKSEL H. KUSMENOGLU U.I. 2007. A classification system for beans using computer vision system and artificial neural networks. *Journal of Food Engineering* 78, 897-904.
- KONING C.T., HUCHES S., LACHLAN D.M. DUNCAN A.J. 2000. *Medicago arborea* - a leguminous fodder shrub for low rainfall farming systems. Proceeding of the 10th meeting of FAOCIHEAM on Pastures and Fodder Crops, Cahiers Options Méditerranéennes 435-438.
- KUHL F.P. & GIARDINA C.R. 1982. Elliptic Fourier features of a closed contour. *Computer Graphics* 18, 259-278.
- KUHN M. & JOHNSON K. 2013. Discriminant analysis and other linear classification models. In: *Applied Predictive Modeling* (pp. 275-328). New York: Springer.
- LESINS K.A. & LESINS I. 1979. Genus *Medicago* (Leguminosae): a taxogenetic study. The Hague: Dr. W. Junk Publishers.
- LO BIANCO M., GRILLO O., CREMONINI R. VENORA G. Seed phenotypic identification of Italian bean landraces (*Phaseolus vulgaris* L.) by biometric and texture descriptors. *Australian Journal of Crop Science* (Manuscript submitted for publication).
- LOCK J.M. 2005. Trifolieae. In G. Lewis, B. Schrire, B. Mackinder, & J.M. Lock (Eds.), *Legumes of the World* (pp. 499-503). Kew: The Royal Botanic Gardens.
- MAHALANOBIS P.C. 1936. On the generalised distance in statistics. *Proceedings of the National Institute of Sciences of India* 12, 49-55.
- MEBATSION H.K., PALIWAL J. JAYAS D.S. 2012. Evaluation of variations in the shape of grain types using principal components analysis of the Elliptic Fourier Descriptors. *Computers and Electronics in Agriculture* 80, 63-70.
- MEHREGAN I., RAHIMINEJAD M.R. AZIZIAN D. 2002. A Taxonomic revision of the genus *Medicago* L. (Fabaceae) in Iran. *Iranian Journal Botany* 9, 207-221.
- NANNI L., SHI J.Y., BRAHNAM S. LUMINI A. 2010. Protein classification using texture descriptors extracted from the protein backbone image. *Journal of Theoretical Biology* 264,1024-1032.
- OLIVES G. 1969. *La alfalfa arbórea*. Madrid: Ministerio de Agricultura, Pesca y Alimentación.

- ORRÙ M., GRILLO O., LOVICU G., VENORA G. BACCHETTA G. 2013. Morphological identification of archaeological remains of *Vitis* L. by image analysis. *Vegetation History and Archaeobotany* 22, 231-242.
- ORRÙ M., GRILLO O., VENORA G. BACCHETTA G. 2012. Computer vision as a complementary to molecular analysis: grapevines cultivars case study. *Comptes Rendus de Biologies* 335, 602-615.
- PÉREZ-BAÑÓN C., JUAN A., PETANIDOU T., MARCOS-GARCÍA M.A. CRESPO M.B. 2003. Pollinator limitation in isolated environments: The reproductive ecology of *Medicago citrine* (Font Quer) Greuter (Leguminosae), a bee-dependent plant from bee-deprived Mediterranean islands. *Plant Systematics and Evolution* 241, 29-46.
- PINNA S., GRILLO O., MATTANA E., CAÑADAS E. BACCHETTA G. 2014. Inter- and intraspecific morphometric variability in *Juniperus* L. seeds (Cupressaceae). *Systematics and Biodiversity* 12, 211-223.
- RENCHER A.C. & CHRISTENSEN W.F. 2012. Methods of multivariate analysis (3rd edition). Hoboken: Wiley.
- RIVAS-MARTÍNEZ S., PENAS A. DÍAZ T.E. 2004. Bioclimatic and biogeographic maps of Europe. www.globalbioclimatics.org/form/maps
- ROBLEDO A., RÍOS S. CORREAL E. 1993. El grupo *Medicago arborea* en la cuenca Mediterránea: Origen, distribución y morfología. *Pastos* 23, 325-336.
- ROSATO M., & ROSSELLÓ J.A. 2009. Karyological observations in *Medicago* Section *Dendrotelis* (Fabaceae). *Folia Geobotanica* 44, 423-433.
- ROSATO M., CASTRO M. ROSSELLÓ J.A. 2008. Relationships of the woody *Medicago* species (section *Dendrotelis*) assessed by molecular cytogenetic analyses. *Annals of Botany* 102, 15-22.
- SERRA L., PÉREZ J. IZQUIERDO J. 2001. *Medicago citrina* (Font Quer) Greuter, en la Península Ibérica. *Anales Jardín Botánico Madrid* 59, 158-159.
- SHAHIN M.A. & SYMONS S.J. 2003. Lentil type identification using machine vision. *Canadian Biosystem Engineering* 45, 3.5-3.11.
- SIBOLE J.V., CABOT C., MICHALKE W., POSCHENRIEDER C. BARCELÓ J. 2005. Relationship between expression of the PM H⁺-ATPase growth and ion partitioning in the leaves of salt-treated *Medicago* species. *Planta* 221 557-566.

- SIBOLE J.V., CABOT C., POSCHENRIEDER C. BARCELÓ J. 2003. Ion allocation in two different salt-tolerant Mediterranean *Medicago* species. *Journal of Plant Physiology* 160, 1361-1365.
- SMALL E. & JOMPHE M. 1989. A synopsis of the genus *Medicago* (Leguminosae). *Canadian Journal of Botany* 67 3260-3294.
- SMYKALOVA I., GRILLO O., BJELKOVA M., PAVELEK M. VENORA G. 2013. Phenotypic evaluation of flax seeds by image analysis. *Industrial Crops and Products* 47, 232-238.
- SOBRINO E., HERVELLA A., CERESUELA J.L., BARBADO A., VIVIANI A., DE ANDRÉS E.F. TENORIO J.L. 2000. Morfología y taxonomía de la Sección *Dendrotelis* del género *Medicago* (Fabaceae). *Portugaliae Acta Biologica* 19, 225-237.
- SPSS. 2007. Application Guide, SPSS version 16.0. Chicago: SPSS Inc.
- TERRAL J., TABARD E., BOUBY L., IVORRA S., PASTOR T., FIGUEIRAL I., PICQ S., CHEVANCEJ.B., JUNG C., FABRE L., TARDY C., COMPAN M., BACILIERI R., LACOMBE T., THIS P. 2010. Evolution and history of grapevine (*Vitis vinifera*) under domestication: new morphometric perspectives to understand seed domestication syndrome and reveal origins of ancient European cultivars. *Annals of Botany* 105, 443-455.
- VENORA G., GRILLO O. SACCONI R. 2009a. Quality assessment of durum wheat storage centres in Sicily: Evaluation of vitreous, starchy and shrunken kernels using an image analysis system. *Journal Cereal Science* 49, 429-440.
- VENORA G., GRILLO O., RAVALLI C. CREMONINI R. 2007. Tuscany beans landraces, on-line identifications from seeds inspection by image analysis and Linear Discriminant Analysis. *Agrochimica* LI 4-5, 254-268.
- VENORA G., GRILLO O., RAVALLI C. CREMONINI R. 2009b. Identification of Italian landraces of bean (*Phaseolus vulgaris* L.) using an image analysis system. *Scientia Horticulturae* 121, 410-418.
- WIESNEROVA D. & WIESNER I. 2008. Computer image analysis of seed shape and seed color for flax cultivar description. *Computers and Electronics in Agriculture* 61, 126-135.

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. *Annals of Botany* 94, 657-664.

**Morpho-colorimetric characterization of *Malva* alliance taxa by seed
image analysis**

Abstract

Seed morphometric and colorimetric features, describing shape, size, and textural seed traits, of 28 *taxa* belonging to the genus *Lavatera* and *Malva* were measured using an image analysis system. The data were statistically analyzed to contribute to the taxonomical treatment of the *Malvae* alliance and to evaluate some doubtful systematic position. A clear differentiation between the *taxa* traditionally attributed respectively to the genus *Lavatera* and *Malva*, was highlighted. Furthermore, the image analysis system here proposed, was able to discriminate among the *Lavatera* sections, confirming the taxonomic organization for this genus. Similarly, the results obtained for *Malva*, both at species level and among sections, supported this analytical tool as diagnostic for systematic purposes.

Introduction

The family of Malvaceae A.L. de Jussieu, is represented by dialipetals and pentamerous plants, with hermaphrodite and actinomorphic flowers or with a weak tendency to zygomorphism (Klitgård, 2013).

Malvaceae includes more than 100 genera and 2000 species, grouped in five tribes, with cosmopolite chorology, some of them invasive, mostly spread at tropics, especially in Southern America (Tutin, 1964). The *Malva* alliance (*Malva*, *Lavatera* and annexed genera; Bates, 1968) led, more than once, to different opinions among those who investigated these *taxa* from a phylogenetic and morphological point of view through traditional classification systems. These problems arise from the high level of homoplasy in morphological characters that distinguishes the entire group (Escobar *et al.*, 2009).

Linneaus (1753) emphasized the characters of the epicalyx as a discriminating factor of *Lavatera* and *Malva* genera. According to this classification system, followed by many others (e.g., de Candolle, 1824; Baker, 1890; Fernandes, 1968a,b) and still the most frequently used in modern floras (e.g., Flora Europaea, Flora USSR, Flora Iberica), the c. 20 species of *Lavatera* (Mediterranean herbs and shrubs with highest diversity in the western Mediterranean, a few shrubby species in California and Mexico, Ethiopia and Western Australia; Tournefort, 1706; Fernandes, 1968b) have three fused epicalyx bracts; while the c. 12 perennial and annual species of *Malva* (native to Eurasia with the center in the western Mediterranean, introduced elsewhere: Morton, 1937; Dalby, 1968), are characterized by also three (sometimes two) but free epicalyx bracts (Escobar *et al.*, 2009).

By molecular analysis, Ray (1995) reassessing the phylogeny of the two genera highlighting the existence of two groups of species, morphologically characterized by the peculiarities of the fruits:

- Clade of Lavateroids: monophyletic group including 16 Euro-Mediterranean species belonging to the genera *Lavatera* and *Malva*, characterized by fruits with melted mericarps that open when ripe releasing the seeds, while the walls remain attached to a carpophore more or less developed in so as to form small patches of hyaline. This type of result is not attributable to a true schizocarp but rather to an intermediate form between a schizocarp and a capsule.
- Clade of Malvoids: monophyletic group of species belonging to both genera *Lavatera* and *Malva*, mainly with cosmopolitan chorology, also including those
- of the genus *Lavatera* distributed in Australia and the New World. These species possess schizocarps, with mericarps with thick walls and angular, not releasing the seed but that detach from the fruiting bodies separately or entirely (as in the case of *Malva nicaeensis* All.).

These morphological differences were strongly supported by analysis of interstitial telomeric sequences (ITS) that allowed the distinction of the two groups of species. According to Ray (1995), the clades are perfectly distinguishable by the analysis of the characteristics of the fruit. These results derived by molecular analysis have been recognized by Bayer & Kubitzki (2003) as a starting point for the division of the two genera and currently the

only genus *Lavatera* would include the species with “Lavateroid” fruit as defined by Ray (1995).

Nevertheless, molecular studies on the *Malva* alliance carried out by Ray and other authors (Fuertes Aguilar *et al.*, 2002, Tate *et al.*, 2005), are currently considered as partial, because based on a not completed samples of *taxa* (Escobar *et al.*, 2009). Recently, challenging the already abandoned Linnean classification scheme based on the characters of epicalyx, but without considering the characteristics of schizocarp, and following the principle of priority established by International Code of Botanical Nomenclature (ICBN), Banfi *et al.* (2005) proposed to consider a single genus *Malva* for all entities, except the Macaronesian endemite *Navaea phoenicea* (Vent.) Webb & Berthel.

Being the taxonomic treatment of *Lavatera* and *Malva* controversial, the potential of biometric indices of the seeds, as tool for the systematic approach and differentiation between the two genera, was investigated. Previously, Bacchetta *et al.* (2011a), successfully demonstrated the evidence on taxonomical differentiation inside the genus *Lavatera* L. sect. *Glandulosae* by seed phenetic characterization using a seed image analysis system. Several authors deal with morpho-colorimetric evaluations of seeds for similar purposes (Granitto *et al.*, 2003; Shahin & Symons, 2003a; Kilic *et al.*, 2007; Venora *et al.*, 2007, 2009a; Grillo *et al.*, 2011; Smykalova *et al.*, 2013). In particular, seed image analysis has gained relevance in morphometrics and colour evaluation (Granitto *et al.*, 2003; Wiesnerova & Wiesner, 2008; Venora *et al.*, 2009a) for its utility in the identification of diaspores of wild plant species (Rovner & Gyulai, 2007; Bacchetta *et al.*, 2008a, 2011b; Grillo *et al.*, 2010, 2013; Pinna *et al.*, 2014; Santo *et al.*, 2015), proving to be a useful tool for taxonomic studies, where very close *taxa* need to be characterized and discriminated.

Afterwards, many authors have successfully used Elliptic Fourier Descriptors, hereafter EFDs in seed studies (e.g. Yoshioka *et al.*, 2004; Terral *et al.*, 2010; Mebatsion *et al.*, 2012; Orrù *et al.*, 2013; Uccesu *et al.*, 2015; Sabato *et al.*, submitted) as well as, “Haralick” parameters, evaluating the surface texture of seeds (Diamond *et al.*, 2004; Gerger & Smolle, 2004; Nanni *et al.*, 2010; Lo Bianco *et al.*, submitted).

The aim of the present work is to contribute to the assessment of the taxonomic position of *taxa* belonging to the *Lavatera* and *Malva* genera, by investigating the morphometric and colorimetric features of the germplasm features of their seeds.

Material and methods

Lavatera and Malva seed lots

Seeds of 20 *taxa* of *Lavatera* belonging to 55 populations and eight *taxa* of *Malva* from eight populations were collected, in ten different geographical regions during a period of ten years. An overall of 79 accessions (Table 1) were studied and, in order to allow an effective long time storage, they were ultra-dried out down to 2-3% R.H., guaranteeing homogeneity and regularity in seed size and weight (Pérez-García *et al.*, 2007). Finally, they were hermetically sealed in glass tubes with capsules of micro-granular silica gel and stored at -25°C in a cold room, in the Sardinian Germplasm Bank (BG-SAR), according to the protocols reported in Bacchetta *et al.* (2008b).

Table 1. Populations, sampling years and seed amount of the studied *Lavatera* and *Malva* taxa.

Taxon	Accepted name	Seccio	Population	Geographical region	Collecting year	Seed amount
<i>Lavatera acerifolia</i>	<i>Malva acerifolia</i> (Cav.) Alef.	<i>Axolopha</i>	Icod de los Vinos (Santa Cruz de Tenerife)	Spain	2005	100
<i>Lavatera agrigentina</i>	<i>Malva agrigentina</i> (Tineo) Soldano & al.	<i>Glandulosae</i>	Agira (Enna)	Sicily	2010	100
<i>Lavatera agrigentina</i>	<i>Malva agrigentina</i> (Tineo) Soldano & al.	<i>Glandulosae</i>	Assoro (Enna)	Sicily	2010	100
<i>Lavatera agrigentina</i>	<i>Malva agrigentina</i> (Tineo) Soldano & al.	<i>Glandulosae</i>	Ponte Capotarso (Caltanissetta)	Sicily	2008	99
<i>Lavatera agrigentina</i>	<i>Malva agrigentina</i> (Tineo) Soldano & al.	<i>Glandulosae</i>	Piana Grande, Ribera (Agrigento)	Sicily	2008	100
<i>Lavatera arborea</i>	<i>Malva arborea</i> (L.) Webb & Berthel.	<i>Anthema</i>	Saline Sant' Antioco (Carbonia-Iglesias)	Sardinia	2008	100
<i>Lavatera arborea</i>	<i>Malva arborea</i> (L.) Webb & Berthel.	<i>Anthema</i>	Mora de Santa Quiteria, Tobarra (Albacete)	Spain	2003	33
<i>Lavatera arborea</i>	<i>Malva arborea</i> (L.) Webb & Berthel.	<i>Anthema</i>	Porto Campana, Domus de Maria (Cagliari)	Sardinia	2007	100
<i>Lavatera arborea</i>	<i>Malva arborea</i> (L.) Webb & Berthel.	<i>Anthema</i>	Porto Campana, Domus de Maria (Cagliari)	Sardinia	2007	37
<i>Lavatera arborea</i>	<i>Malva arborea</i> (L.) Webb & Berthel.	<i>Anthema</i>	Porto Campana, Domus de Maria (Cagliari)	Sardinia	2007	100
<i>Lavatera arborea</i>	<i>Malva arborea</i> (L.) Webb & Berthel.	<i>Anthema</i>	Capo Testa, Santa Teresa di Gallura (Olbia-Tempio)	Sardinia	2012	100
<i>Lavatera assurgentiflora</i>	<i>Malva assurgentiflora</i> (Kellogg) M.F.Ray	<i>Anthema</i>	Strybing Arboretum (California)	USA	-	54
<i>Lavatera bryonifolia</i>	<i>Malva unguiculata</i> (Desf.) Alef.	<i>Olbia</i>	Rethymno (Crete)	Greece	-	11
Table 1. Continue						
<i>Lavatera cretica</i>	<i>Malva multiflora</i> (Cav.) Soldano & al.	<i>Anthema</i>	Calpe, Peñón de Ifach (Alicante)	Spain	2004	90
<i>Lavatera flava</i>	<i>Malva flava</i> (Desf.) Alef.	<i>Glandulosae</i>	Tazaghine (Rif)	Morocco	2009	200
<i>Lavatera maritima</i>	<i>Malva subovata</i> (DC.) Molero & J. M. Monts.	<i>Axolopha</i>	Nebida	Sardinia	2008	100

Table 1. Continue						
<i>Lavatera maritima</i>	<i>Malva subovata</i> (DC.) Molero & J. M. Monts.	<i>Axolopha</i>	Calpe, Peñón de Ifach (Alicante)	Spain	2004	96
<i>Lavatera maritima</i>	<i>Malva subovata</i> (DC.) Molero & J. M. Monts.	<i>Axolopha</i>	BG-SAR	Sardinia	2008	100
<i>Lavatera maritima</i>	<i>Malva subovata</i> (DC.) Molero & J. M. Monts.	<i>Axolopha</i>	BG-SAR	Sardinia	2007	100
<i>Lavatera maritima</i>	<i>Malva subovata</i> (DC.) Molero & J. M. Monts.	<i>Axolopha</i>	BG-SAR	Sardinia	2006	100
<i>Lavatera maritima</i>	<i>Malva subovata</i> (DC.) Molero & J. M. Monts.	<i>Axolopha</i>	BG-SAR	Sardinia	2010	100
<i>Lavatera maroccana</i>	<i>Malva maroccana</i> (Batt. & Trab.) Soldano & al.	<i>Olbia</i>	Cabezas de San Juan, Laguna de la Cigarrera (Sevilla)	Spain	2003	90
<i>Lavatera mauritanica</i>	<i>Malva davaei</i> (Cout.) Valdés	<i>Anthema</i>	Cabo de San Vicente (Algarve)	Portugal	2003	100
<i>Lavatera moschata</i>	<i>Malva moschata</i> L.	<i>Bismalva</i>	Sankt Wolfgang (Salzburg)	Austria	2005	100
<i>Lavatera oblongifolia</i>	<i>Malva oblongifolia</i> (Boiss.) Soldano & al.	<i>Olbia</i>	Ugijar (Granada)	Spain	2010	100
<i>Lavatera oblongifolia</i>	<i>Malva oblongifolia</i> (Boiss.) Soldano & al.	<i>Olbia</i>	Alpujarra (Granada)	Spain	-	100
<i>Lavatera olbia</i>	<i>Malva olbia</i> (L.) Alef.	<i>Olbia</i>	Is molas, Pula (Cagliari)	Sardinia	2012	100
<i>Lavatera olbia</i>	<i>Malva olbia</i> (L.) Alef.	<i>Olbia</i>	Monte Agruxau (Carbonia-Iglesias)	Sardinia	2010	100
<i>Lavatera plazzae</i>	<i>Malva stenopetala</i> (Batt.) Soldano & al.	<i>Olbia</i>	Giave (Sassari)	Sardinia	2006	99
<i>Lavatera plazzae</i>	<i>Malva stenopetala</i> (Batt.) Soldano & al.	<i>Olbia</i>	Giave (Sassari)	Sardinia	2006	100
<i>Lavatera punctata</i>	<i>Malva punctata</i> (All.) Alef.	<i>Olbia</i>	Aydin	Turkey	-	16
<i>Lavatera thuringiaca</i>	<i>Malva thuringiaca</i> (L.) Vis.	<i>Olbia</i>	Wien	Austria	2008	81
<i>Lavatera triloba</i> subsp. <i>pallescens</i>	<i>Malva lusitanica</i> subsp. <i>pallescens</i> (Moris) Valdés	<i>Glandulosae</i>	Buggerru (Carbonia-Iglesias)	Sardinia	2011	100
<i>Lavatera triloba</i> subsp. <i>pallescens</i>	<i>Malva lusitanica</i> subsp. <i>pallescens</i> (Moris) Valdés	<i>Glandulosae</i>	Buggerru (Carbonia-Iglesias)	Sardinia	2011	100
<i>Lavatera triloba</i> subsp. <i>pallescens</i>	<i>Malva lusitanica</i> subsp. <i>pallescens</i> (Moris) Valdés	<i>Glandulosae</i>	Buggerru (Carbonia-Iglesias)	Sardinia	2011	100
<i>Lavatera triloba</i> subsp. <i>pallescens</i>	<i>Malva lusitanica</i> subsp. <i>pallescens</i> (Moris) Valdés	<i>Glandulosae</i>	Buggerru (Carbonia-Iglesias)	Sardinia	2009	100
<i>Lavatera triloba</i> subsp. <i>pallescens</i>	<i>Malva lusitanica</i> subsp. <i>pallescens</i> (Moris) Valdés	<i>Glandulosae</i>	Buggerru (Carbonia-Iglesias)	Sardinia	2009	100

Table 1. Continue						
<i>Lavatera triloba</i> subsp. <i>pallescens</i>	<i>Malva lusitanica</i> subsp. <i>pallescens</i> (Moris) Valdés	<i>Glandulosae</i>	Buggerru (Carbonia- Iglesias)	Sardinia	2010	100
<i>Lavatera triloba</i> subsp. <i>pallescens</i>	<i>Malva lusitanica</i> subsp. <i>pallescens</i> (Moris) Valdés	<i>Glandulosae</i>	Buggerru (Carbonia- Iglesias)	Sardinia	2010	100
<i>Lavatera triloba</i> subsp. <i>pallescens</i>	<i>Malva lusitanica</i> subsp. <i>pallescens</i> (Moris) Valdés	<i>Glandulosae</i>	Chia, Domus de Maria (Cagliari)	Sardinia	2005	90
<i>Lavatera triloba</i> subsp. <i>pallescens</i>	<i>Malva lusitanica</i> subsp. <i>pallescens</i> (Moris) Valdés	<i>Glandulosae</i>	Chia, Domus de Maria (Cagliari)	Sardinia	2005	100
<i>Lavatera triloba</i> subsp. <i>minoricensis</i>	<i>Malva lusitanica</i> subsp. <i>pallescens</i> (Moris) Valdés	<i>Glandulosae</i>	Illa de l'Aire (Menorca)	Balearic Islands	2008	100
<i>Lavatera triloba</i> subsp. <i>minoricensis</i>	<i>Malva lusitanica</i> subsp. <i>pallescens</i> (Moris) Valdés	<i>Glandulosae</i>	Illa de l'Aire (Menorca)	Balearic Islands	2008	80
<i>Lavatera triloba</i> subsp. <i>minoricensis</i>	<i>Malva lusitanica</i> subsp. <i>pallescens</i> (Moris) Valdés	<i>Glandulosae</i>	S'Escilas	Balearic Islands	2008	80
<i>Lavatera triloba</i> subsp. <i>minoricensis</i>	<i>Malva lusitanica</i> subsp. <i>pallescens</i> (Moris) Valdés	<i>Glandulosae</i>	Punta Nati (Menorca)	Balearic Islands	2008	30
<i>Lavatera triloba</i> subsp. <i>triloba</i>	<i>Malva lusitanica</i> (L.) Valdés subsp. <i>lusitanica</i>	<i>Glandulosae</i>	Diebres (Guadajajara)	Spain	2008	25
<i>Lavatera triloba</i> subsp. <i>triloba</i>	<i>Malva lusitanica</i> (L.) Valdés subsp. <i>lusitanica</i>	<i>Glandulosae</i>	Toledo	Spain	2008	100
<i>Lavatera triloba</i> subsp. <i>triloba</i>	<i>Malva lusitanica</i> (L.) Valdés subsp. <i>lusitanica</i>	<i>Glandulosae</i>	Senda Galiana, Sierra de Cascojo (Toledo)	Spain	2008	100
<i>Lavatera triloba</i> subsp. <i>triloba</i>	<i>Malva lusitanica</i> (L.) Valdés subsp. <i>lusitanica</i>	<i>Glandulosae</i>	Toledo	Spain	2008	100
<i>Lavatera triloba</i> subsp. <i>triloba</i>	<i>Malva lusitanica</i> (L.) Valdés subsp. <i>lusitanica</i>	<i>Glandulosae</i>	Almedina, Ciudad Real	Spain	2008	100
<i>Lavatera triloba</i> subsp. <i>triloba</i>	<i>Malva lusitanica</i> (L.) Valdés subsp. <i>lusitanica</i>	<i>Glandulosae</i>	El Bonillo (Albacete)	Spain	2008	100
<i>Lavatera triloba</i> subsp. <i>triloba</i>	<i>Malva lusitanica</i> (L.) Valdés subsp. <i>lusitanica</i>	<i>Glandulosae</i>	Caspe (Zaragoza)	Spain	2008	100
<i>Lavatera triloba</i> subsp. <i>triloba</i>	<i>Malva lusitanica</i> (L.) Valdés subsp. <i>lusitanica</i>	<i>Glandulosae</i>	La Parra (Badajoz)	Spain	2008	100
<i>Lavatera triloba</i> subsp. <i>triloba</i>	<i>Malva lusitanica</i> (L.) Valdés subsp. <i>lusitanica</i>	<i>Glandulosae</i>	Los Santos de Maimona (Badajoz)	Spain	2008	100
<i>Lavatera triloba</i> subsp. <i>triloba</i>	<i>Malva lusitanica</i> (L.) Valdés subsp. <i>lusitanica</i>	<i>Glandulosae</i>	Zafra (Badajoz)	Spain	2008	100
<i>Lavatera triloba</i> subsp. <i>triloba</i>	<i>Malva lusitanica</i> (L.) Valdés subsp. <i>lusitanica</i>	<i>Glandulosae</i>	Matanegra (Badajoz)	Spain	2008	100
<i>Lavatera triloba</i> subsp. <i>triloba</i>	<i>Malva lusitanica</i> (L.) Valdés subsp. <i>lusitanica</i>	<i>Glandulosae</i>	Correinas, Elmas (Cagliari)	Sardinia	2010	100
<i>Lavatera triloba</i> subsp. <i>triloba</i>	<i>Malva lusitanica</i> (L.) Valdés subsp. <i>lusitanica</i>	<i>Glandulosae</i>	Sa Tierra, Assemimi (Cagliari)	Sardinia	2008	100
<i>Lavatera triloba</i> subsp. <i>triloba</i>	<i>Malva lusitanica</i> (L.) Valdés subsp. <i>lusitanica</i>	<i>Glandulosae</i>	Santa Maria, Assemimi (Cagliari)	Sardinia	2008	100

Table 1. Continue							
<i>Lavatera triloba</i> subsp. <i>triloba</i>	<i>Malva lusitanica</i> (L.) Valdés subsp. <i>lusitanica</i>	<i>Glandulosae</i>	Tierra, Domus de Maria (Cagliari)	Sardinia	2010	100	
<i>Lavatera triloba</i> subsp. <i>triloba</i>	<i>Malva lusitanica</i> (L.) Valdés subsp. <i>lusitanica</i>	<i>Glandulosae</i>	Canali Saliu, Pula (Cagliari)	Sardinia	2008	100	
<i>Lavatera triloba</i> subsp. <i>triloba</i>	<i>Malva lusitanica</i> (L.) Valdés subsp. <i>lusitanica</i>	<i>Glandulosae</i>	Canali Saliu, Pula (Cagliari)	Sardinia	2010	100	
<i>Lavatera triloba</i> subsp. <i>triloba</i>	<i>Malva lusitanica</i> (L.) Valdés subsp. <i>lusitanica</i>	<i>Glandulosae</i>	Correinas, Elmas (Cagliari)	Sardinia	2008	100	
<i>Lavatera triloba</i> subsp. <i>triloba</i>	<i>Malva lusitanica</i> (L.) Valdés subsp. <i>lusitanica</i>	<i>Glandulosae</i>	Stani de Serdiana (Cagliari)	Sardinia	2008	100	
<i>Lavatera triloba</i> subsp. <i>triloba</i>	<i>Malva lusitanica</i> (L.) Valdés subsp. <i>lusitanica</i>	<i>Glandulosae</i>	Campu su Gureu, Sestu (Cagliari)	Sardinia	2008	100	
<i>Lavatera triloba</i> subsp. <i>triloba</i>	<i>Malva lusitanica</i> (L.) Valdés subsp. <i>lusitanica</i>	<i>Glandulosae</i>	Riu Saliu, Selargius (Cagliari)	Sardinia	2008	100	
<i>Lavatera triloba</i> subsp. <i>triloba</i>	<i>Malva lusitanica</i> (L.) Valdés subsp. <i>lusitanica</i>	<i>Glandulosae</i>	Chia, Domus de Maria (Cagliari)	Sardinia	2011	100	
<i>Lavatera triloba</i> subsp. <i>triloba</i>	<i>Malva lusitanica</i> (L.) Valdés subsp. <i>lusitanica</i>	<i>Glandulosae</i>	Terrasili, Assemini (Cagliari)	Sardinia	2011	100	
<i>Lavatera triloba</i> subsp. <i>triloba</i>	<i>Malva lusitanica</i> (L.) Valdés subsp. <i>lusitanica</i>	<i>Glandulosae</i>	Pula (Cagliari)	Sardinia	2011	100	
<i>Lavatera</i> <i>trimestris</i>	<i>Malva trimestris</i> (L.) Salisb.	<i>Olbia</i>	Chefchaouen (Rif)	Morocco	2004	107	
<i>Malva alcea</i>	<i>Malva alcea</i> L.	<i>Bismalva</i>	Mijares (Avila)	Spain	2003	77	
<i>Malva hispanica</i>	<i>Malva hispanica</i> L.	<i>Bismalva</i>	Guadajira, La Orden (Badajoz)	Spain	2004	100	
<i>Malva multiflora</i>	<i>Malva multiflora</i> (Cav.) Soldano & al.	<i>Anthema</i>	Poggio dei Pini, Capoterra (Cagliari)	Sardegna	2012	98	
<i>Malva nicaeensis</i>	<i>Malva nicaeensis</i> All.	<i>Malva</i>	Guadajira, La Orden (Badajoz)	Spain	2004	90	
<i>Malva parviflora</i>	<i>Malva parviflora</i> L.	<i>Malva</i>	Guadajira, La Orden (Badajoz)	Spain	2004	100	
<i>Malva sylvestris</i>	<i>Malva sylvestris</i> L.	<i>Malva</i>	Arganda del Rey (Madrid)	Spain	2004	100	
<i>Malva</i> <i>tournefortiana</i>	<i>Malva</i> <i>tournefortiana</i> L.	<i>Bismalva</i>	Talarrubias, Sierra de Puerto Peña (Badajoz)	Spain	2004	16	
<i>Malva</i> <i>verticillata</i>	<i>Malva</i> <i>verticillata</i> L.	<i>Malva</i>	Wien	Austria	-	102	

Image analysis system

Samples digital images, consisting of 100 seeds randomly disposed on tray, were acquired using a flatbed scanner (Epson GT-15000) with a digital resolution of 400 dpi and a scanning area not exceeding 1024×1024 pixel. For accessions of fewer than 100 seeds, the analysis was executed on the whole batch. A total of 7,178 seeds were analyzed.

Image acquisition was performed before drying the seeds at 15°C to 15% of R.H. to avoid spurious variation in dimension, shape and colour. The scanner was calibrated for colour matching following the protocol of Shahin and Symons (2003b) before seed samples image acquisition, as suggested by Venora *et al.* (2009b).

Digital images of seeds were processed and analyzed using the software package KS-400 V. 3.0 (Carl Zeiss, Vision, Oberkochen, Germany). A macro specifically developed for the characterization of seeds (Venora *et al.*, 2009b), was modified to perform automatically all the analysis procedures, reducing the execution time and contextually mistakes in the analysis process.

In order to improve the discrimination power, this macro was further enhanced adding algorithms able to compute the EFDs for each analyzed seed. This method allows description of the boundary of the seed projection as an array of complex numbers which correspond to the pixel positions on the seed boundary. So, from the seed apex, defined as the starting point in a Cartesian system, chain codes are generated. A chain code is a lossless compression algorithm for binary images. The basic principle of chain codes is to separately encode each connected component (pixel) in the image. The encoder then moves along the boundary of the image and, at each step, transmits a symbol representing the direction of this movement. This continues until the encoder returns to the starting position. This method is

based on separate Fourier decompositions of the incremental changes of the X and Y coordinates as a function of the cumulative length along the boundary (Kuhl & Giardina 1982). Each harmonic (n) corresponds to four coefficients (an , bn , cn and dn) defining the ellipse in the XY plane. The coefficients of the first harmonic, describing the best fitting ellipse of outlines, are used to standardize size (surface area) and to orientate seeds (Terral *et al.* 2010). According to Terral *et al.* (2010), about the use of a number of harmonics for an optimal description of seed outlines, in order to minimize the measurement errors and to optimize the efficiency of shape reconstruction, 20 harmonics were used to define the seed boundaries, obtaining a further 78 parameters useful to discriminate among the studied seeds (Orrù *et al.* 2012).

Moreover, the macro was further improved adding algorithms able to compute 11 Haralick's descriptors and the relative standard deviations for each analyzed seed. These parameters are generally used when the objects in the images cannot be separated due to indefinite grey values variations. In these cases, the evaluation of texture, tone and context allows to define the spatial distribution of the image intensities and discrete tonal features (Haralick *et al.*, 1973). When a small area of the image has little variation of discrete tonal features, the dominant property of that area is grey tone. When a small area has wide variation of discrete tonal features, the dominant property of that area is texture (Haralick & Shapiro, 1991). According to Haralick *et al.* (1973), the concept of tone is based on varying shades of grey of resolution cells in a photographic image, while texture is concerned with the spatial (statistical) distribution of grey tones. Texture and tone are not independent concepts; rather, they bear an inextricable relationship to one another very much like the relationship between a particle and a wave

(Haralick, 1979). Context, texture and tone are always present in the image, although at times one property can dominate the others.

The basis for these features is the gray-level co-occurrence matrix (G in equation 1). This matrix is square with dimension N_g , where N_g is the number of gray levels in the image. Element $[i,j]$ of the matrix is generated by counting the number of times a pixel with value i is adjacent to a pixel with value j and then dividing the entire matrix by the total number of such comparisons made. Each entry is therefore considered to be the probability that a pixel with value i will be found adjacent to a pixel of value j .

$$G = \begin{bmatrix} p(1,1) & p(1,2) & \dots & p(1,N_g) \\ p(2,1) & p(2,2) & \dots & p(2,N_g) \\ \vdots & \vdots & \ddots & \vdots \\ p(N_g,1) & p(N_g,2) & \dots & p(N_g,N_g) \end{bmatrix} \quad (1)$$

In Table 2, the 11 Haralick's descriptors measured on each seed to mathematically describe the surface texture, are reported.

A total of 137 morphometric, colorimetric and textural characters were measured on each seed (Table 3).

Table 2. Haralick's descriptors measured as reported in Haralick *et al.* (1973).

	<i>Feature</i>	<i>Equation</i>
Har 1	Angular second moment	$\sum_i \sum_j p(i, j)^2$
Har 2	Contrast	$\sum_{n=0}^{N_g-1} n^2 \left\{ \sum_{i=1}^{N_g} \sum_{j=1}^{N_g} p(i, j) \right\}, i, j = n$
Har 3	Correlation	$\frac{\sum_i \sum_j (ij) p(i, j) - \mu_x \mu_y}{\sigma_x \sigma_y}$
		where μ_x , μ_y , σ_x and σ_y are the means and the standard deviations of p_x and p_y .
Har 4	Sum of square: variance	$\sum_i \sum_j (i - \mu)^2 p(i, j)$
Har 5	Inverse difference moment	$\sum_i \sum_j \frac{1}{1 + (i - j)^2} p(i, j)$
Har 6	Sum average	$\sum_{n=2}^{2N_g} i p_{x+y}(i)$
		where x and y are the coordinates (row and column) of an entry in the co-occurrence matrix, and $p_{x+y}(i)$ is the probability of co-occurrence matrix coordinates summing to $x+y$.
Har 7	Sum variance	$\sum_{i=2}^{2N_g} (i - f_{\oplus})^2 p_{x+y}(i)$
Har 8	Sum entropy	$- \sum_{i=2}^{2N_g} p_{x+y}(i) \log\{p_{x+y}(i)\} = f_{\oplus}$
Har 9	Entropy	$- \sum_i \sum_j p(i, j) \log[p(i, j)]$
Har 10	Difference variance	$\sum_{n=0}^{N_g-1} i^2 p_{x-y}(i)$
Har 11	Difference entropy	$- \sum_{n=0}^{N_g-1} p_{x-y}(i) \log\{p_{x-y}(i)\}$

Table 3. List of morpho-colorimetric features measured on seeds, excluding the 78 Elliptic Fourier Descriptors calculated according to Hâruta (2011).

	Feature	Description
<i>A</i>	Area	Seed area (mm ²)
<i>P</i>	Perimeter	Seed perimeter (mm)
<i>P_{conv}</i>	Convex Perimeter	Convex perimeter of the seed (mm)
<i>P_{Crof}</i>	Crofton's Perimeter	Perimeter of the seed calculated using the Crofton's formula (mm)
<i>P_{conv}/P_{Crof}</i>	Perimeter ratio	Ratio between convex and Crofton's perimeters
<i>D_{max}</i>	Max diameter	Maximum diameter of the seed (mm)
<i>D_{min}</i>	Min diameter	Minimum diameter of the seed (mm)
<i>D_{min}/D_{max}</i>	Feret ratio	Ratio between minimum and maximum diameters
<i>Sf</i>	Shape Factor	Seed shape descriptor = (4 x π x area)/perimeter ² (normalized value)
<i>Rf</i>	Roundness Factor	Seed roundness descriptor = (4 x area)/(π x max diameter ²) (normalized value)
<i>Ecd</i>	Eq. circular diameter	Diameter of a circle with an area equivalent to that of the seed (mm)
<i>EA_{max}</i>	Maximum ellipse axis	Maximum axis of an ellipse with equivalent area (mm)
<i>EA_{min}</i>	Minimum ellipse axis	Minimum axis of an ellipse with equivalent area (mm)
<i>R_{mean}</i>	Mean red channel	Red channel mean value of seed pixels (grey levels)
<i>R_{sd}</i>	Red std. deviation	Red channel standard deviation of seed pixels
<i>G_{mean}</i>	Mean green channel	Green channel mean value of seed pixels (grey levels)
<i>G_{sd}</i>	Green std. deviation	Green channel standard deviation of seed pixels
<i>B_{mean}</i>	Mean blue channel	Blue channel mean value of seed pixels (grey levels)
<i>B_{sd}</i>	Blue std. deviation	Blue channel standard deviation of seed pixels
<i>H_{mean}</i>	Mean hue channel	Hue channel mean value of seed pixels (grey levels)
<i>H_{sd}</i>	Hue std. deviation	Hue channel standard deviation of seed pixels
<i>L_{mean}</i>	Mean lightness channel	Lightness channel mean value of seed pixels (grey levels)
<i>L_{sd}</i>	Lightness std. deviation	Lightness channel standard deviation of seed pixels
<i>S_{mean}</i>	Mean saturation channel	Saturation channel mean value of seed pixels (grey levels)
<i>S_{sd}</i>	Saturation std. deviation	Saturation channel standard deviation of seed pixels
<i>D_{mean}</i>	Mean density	Density channel mean value of seed pixels (grey levels)
<i>D_{sd}</i>	Density std. deviation	Density channel standard deviation of seed pixels
<i>S</i>	Skewness	Asymmetry degree of intensity values distribution (grey levels)
<i>K</i>	Kurtosis	Peakness degree of intensity values distribution (densitometric units)
<i>H</i>	Energy	Measure of the increasing intensity power (densitometric units)
<i>E</i>	Entropy	Dispersion power (bit)
<i>D_{sum}</i>	Density sum	Sum of density values of the seed pixels (grey levels)
<i>SqD_{sum}</i>	Square density sum	Sum of the squares of density values (grey levels)

Statistical analysis

The achieved results were used to build a database including morpho-colorimetric, EFDs and Haralick's descriptors. Statistical elaborations were executed using SPSS software package release 16.0 (SPSS Inc. for Windows, Chicago, Illinois, USA), and the stepwise Linear Discriminant Analysis method hereafter LDA was applied to identify and discriminate among the investigated *Lavatera* and *Malva* accessions.

This approach is commonly used to classify/identify unknown groups characterized by quantitative and qualitative variables (Fisher, 1936; 1940; Sugiyama, 2007), finding the combination of predictor variables with the aim of minimizing the within-class distance and maximizing the between-class distance simultaneously, thus achieving maximum class discrimination (Hastie *et al.*, 2001; Holden *et al.*, 2011; Alvin & William, 2012; Kuhn & Johnson, 2013). The stepwise method identifies and selects the most statistically significant features among the 137 measured on each seed, using three statistical variables: Tolerance, *F*-to-enter and *F*-to-remove. The Tolerance value indicates the proportion of a variable variance not accounted by other independent variables in the equation. *F*-to-enter and *F*-to-remove values define the power of each variable in the model and are useful to describe what happens if a variable is inserted and removed, respectively, from the current model. This method starts with a model that does not include any of the variables. At each step, the variable with the largest *F*-to-enter value that exceeds the entry criterion chosen ($F \geq 3.84$) is added to the model. The variables left out of the analysis at the last step have *F*-to-enter values smaller than 3.84, and therefore no more are added stopping the process (Venora *et al.*, 2009b; Grillo *et al.*, 2012). Finally, a cross-validation procedure was applied to verify the performance of the identification system,

testing individual unknown cases and classifying them on the basis of all others (SPSS, 2007).

All the raw data were standardized before starting any statistical elaboration. Moreover, in order to evaluate the quality of the discriminant functions achieved for each statistical comparison, the Wilks' Lambda, the percentage of explained variance and the canonical correlation between the discriminant functions and the group membership, were computed. The Box's M tests was executed to assess the homogeneity of covariance matrices of the features chosen by the stepwise LDA while the analysis of the standardized residuals was performed to verify the homoscedasticity of the variance of the dependent variables used to discriminate among the groups' membership (Box, 1949; Haberman, 1973; Morrison, 2004). Kolmogorov-Smirnov's test was performed to compare the empirical distribution of the discriminant functions with the relative cumulative distribution function of the reference probability distribution, while the and Levene's test was executed to assess the equality of variances for the used discriminant functions calculated for groups' membership (Gastwirth *et al.*, 2009; Levene, 1960; Lopes, 2011).

To graphically highlight the differences among groups, multidimensional plots were drawn using the first three discriminant functions or, alternatively, when the number of discriminant groups n did not allow to obtain at least three discriminant functions ($n-1$), the two available discriminant functions and the Mahalanobis' square distance (Mahalanobis, 1936) were used. This measure of distance is defined by two or more discriminant functions and ranges from 0 to infinite. Samples are increasingly similar at values closer to zero. Higher values indicate that a particular case includes extreme values for one or more independent variables, and can be

considered significantly different to other cases of the same group (Bacchetta *et al.*, 2008a).

Results

A preliminary investigation was carried out to discriminate between the two groups object of the study. Using this model, 97.6% of the cross-validated samples of the all studied *Lavatera* and *Malva taxa* were correctly classified (data not shown).

In order to identify the *taxa* of *Lavatera* that are closer to genus *Malva* and viceversa, all the *taxa* were singularly compared among them, highlighting some mutual misattributions. In particular, *L. maroccana* was misidentified to *Malva* group in the 13% of the cases, wrongly classified mainly as *M. sylvestris* and *M. tournefortiana*, and *L. moschata* was confused with the same *Malva* species for 18% of the cases (data not shown).

Successively, the 20 *taxa* belonging to the genus *Lavatera* were compared among them, without the influence of the genus *Malva*. An overall percentage of correct identification of 87.5% was reached (Table 4). In this group, a correct classification range included between 63.6% (*L. brynonifolia*, misclassified as *L. maroccana* in the 27.3% of cases, and as *L. moschata* in the 9.1%) and 98.0% for *L. olbia* was recorded.

In order to perform a statistical comparison, all the *Lavatera* accessions were grouped into their sections of belonging. The system was able to discriminate among the five groups studied in the 87.8% of the cases (Table 5). As shown, the *Glandulosae* and *Bismalva* sections were correctly distinguished for the 94% of their seeds; on the other hand, the *Olbia* section was discriminate for the 73.7%, misattributed as *Glandulosae* for 15.5%, *Bismalva* for 6.2% and *Anthema* for 4.6% of its seeds.

Table 4. Percentage of correct identification for *Lavatera* accessions at species level. In parenthesis, the number of analysed seeds.

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)	(17)	(18)	(19)	(20)	Total
<i>L. acerifolia</i> (1)	95.0 (95)	-	1.0 (1)	-	-	-	-	-	-	-	-	1.0 (1)	-	-	-	1.0 (1)	1.0 (1)	-	1.0 (1)	-	100.0 (100)
<i>L. agrigentina</i> (2)	-	93.0 (370)	2.5 (10)	-	-	-	-	0.3 (1)	-	-	-	0.3 (1)	-	-	-	-	1.5 (6)	0.8 (3)	1.8 (7)	-	100.0 (398)
<i>L. arborea</i> (3)	4.3 (20)	1.5 (7)	67.2 (316)	0.2 (1)	-	-	0.2 (1)	1.5 (7)	-	0.4 (2)	0.2 (1)	-	-	10.6 (50)	-	0.9 (4)	1.5 (7)	1.5 (7)	10.0 (47)	-	100.0 (470)
<i>L. assurgentifolia</i> (4)	3.8 (2)	-	-	96.2 (51)	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	100.0 (53)
<i>L. bryonifolia</i> (5)	-	-	-	-	63.6 (7)	-	-	-	27.3 (3)	-	9.1 (1)	-	-	-	-	-	-	-	-	-	100.0 (11)
<i>L. cretica</i> (6)	-	-	-	-	-	82.2 (74)	-	-	-	17.8 (16)	-	-	-	-	-	-	-	-	-	-	100.0 (90)
<i>L. flava</i> (7)	-	-	-	-	-	-	86.7 (157)	-	-	-	-	-	-	-	-	-	1.7 (3)	0.6 (1)	11.0 (20)	-	100.0 (181)
<i>L. maritima</i> (8)	1.7 (10)	0.5 (3)	6.5 (39)	0.3 (2)	-	-	0.5 (3)	77.5 (462)	-	-	0.2 (1)	-	-	-	-	1.2 (7)	0.2 (1)	0.2 (1)	11.2 (67)	-	100.0 (596)
<i>L. maroccana</i> (9)	-	-	-	-	2.2 (2)	1.1 (1)	-	-	91.1 (82)	1.1 (1)	-	-	-	-	2.2 (2)	2.2 (2)	-	-	-	-	100.0 (90)
<i>L. mauritanica</i> (10)	-	-	-	-	-	7.0 (7)	-	-	3.0 (3)	86.0 (86)	-	-	-	-	2.0 (2)	2.0 (2)	-	-	-	-	100.0 (100)
<i>L. moschata</i> (11)	-	-	-	-	8.0 (8)	-	-	-	4.0 (4)	-	87.0 (87)	-	-	-	-	1.0 (1)	-	-	-	-	100.0 (100)
<i>L. oblongifolia</i> (12)	4.5 (9)	-	-	-	-	-	-	-	-	-	0.5 (1)	79.5 (159)	0.5 (1)	-	-	15.0 (30)	-	-	-	-	100.0 (200)
<i>L. olbia</i> (13)	-	-	-	-	-	-	-	-	-	-	-	-	98.0 (196)	-	-	-	-	-	0.5 (1)	1.5 (3)	100.0 (200)
<i>L. plazzae</i> (14)	-	1.5 (3)	1.5 (3)	-	-	-	-	-	-	-	-	-	-	80.4 (160)	-	-	2.0 (4)	0.5 (1)	14.1 (28)	-	100.0 (199)
<i>L. punctata</i> (15)	-	-	-	-	-	-	-	-	12.5 (2)	-	-	-	-	-	87.5 (14)	-	-	-	-	-	100.0 (16)
<i>L. thuringiaca</i> (16)	-	-	-	-	-	1.2 (1)	-	-	-	1.2 (1)	1.2 (1)	-	-	-	-	96.3 (78)	-	-	-	-	100.0 (81)
<i>L. triloba</i> subsp. <i>pallescens</i> (17)	-	0.2 (2)	0.1 (1)	-	-	-	2.4 (21)	-	-	-	-	-	-	-	0.2 (2)	-	88.0 (782)	5.2 (46)	3.9 (35)	-	100.0 (889)
<i>L. triloba</i> subsp. <i>minoricensis</i> (18)	-	-	0.3 (1)	-	-	-	-	-	-	0.3 (1)	-	-	-	-	0.7 (2)	-	1.4 (4)	87.2 (252)	10.0 (29)	-	100.0 (289)
<i>L. triloba</i> subsp. <i>triloba</i> (19)	0.0 (1)	0.0 (1)	0.4 (9)	0.0 (1)	-	-	1.9 (44)	0.5 (11)	-	-	-	0.1 (2)	0.0 (1)	2.7 (2.7)	-	-	1.0 (23)	0.9 (21)	92.4 (2148)	0.0 (1)	100.0 (2325)
<i>L. trimestris</i> (20)	-	-	-	-	-	-	-	-	-	0.9 (1)	-	-	-	-	-	-	-	-	-	99.1 (106)	100.0 (107)
Overall																					87.5 (6495)

Table 5. Percentage of correct identification among *Lavatera* sections. In parenthesis, the number of analyzed seeds.

	(1)	(2)	(3)	(4)	(5)	Total
<i>Axolopha</i> (1)	74.9 (521)	9.1 (63)	13.6 (95)	2.0 (14)	0.4 (3)	100.0 (696)
<i>Glandulosae</i> (2)	1.1 (45)	94.9 (3873)	2.0 (81)	2.0 (82)	0.0 (1)	100.0 (4082)
<i>Anthema</i> (3)	7.2 (51)	10.5 (75)	76.7 (547)	5.3 (38)	0.3 (2)	100.0 (713)
<i>Olbia</i> (4)	-	15.5 (140)	4.6 (42)	73.7 (666)	6.2 (56)	100 (904)
<i>Bismalva</i> (5)	-	-	-	6.0 (6)	94.0 (94)	100 (100)
Overall						87.8 (6495)

With the aim to verify the inter-specific seed morpho-colorimetric variability within each group, the five *Lavatera* sections were analysed separately. In Table 6, the percentage of correct identification among *Lavatera* accessions section *Axolopha*, is shown. An overall of cross-validated classification of 99.3% was recorded.

Table 6. Percentage of correct identification among *Lavatera* accessions section *Axolopha*. In parenthesis, the number of analyzed seeds.

	<i>L. acerifolia</i>	<i>L. maritima</i>	Total
<i>L. acerifolia</i>	99.0 (99)	1.0 (1)	100.0 (100)
<i>L. maritima</i>	0.7 (4)	99.3 (92)	100.0 (596)
Overall			99.3 (696)

The *Glandulosae* sectio, involving five species, *L. agrigentina*, *L. flava*, *L. triloba* in its three subspecies *L. triloba* subsp. *pallescens*, *L. triloba* subsp. *minoricensis* and *L. triloba* subsp. *triloba*, was correctly classified for 92.5% (Table 7), ranged from 95.4% for *L. triloba* subsp. *triloba* and 84% for *L. flava*, whose seeds were mainly misclassified among those of *L. triloba* subsp. *triloba*.

Table 7. Percentage of correct identification among *Lavatera* accessions sectio *Glandulosae*. In parenthesis, the number of analyzed seeds.

	(1)	(2)	(3)	(4)	(5)	Total
<i>L. agrigentina</i> (1)	92.7 (369)	-	2.8 (11)	0.8 (3)	3.8 (15)	100.0 (398)
<i>L. flava</i> (2)	-	84.0 (152)	5.0 (9)	0.6 (1)	10.5 (19)	100.0 (181)
<i>L. triloba</i> subsp. <i>pallescens</i> (3)	-	1.6 (14)	88.8 (789)	4.9 (44)	4.7 (42)	100.0 (889)
<i>L. triloba</i> subsp. <i>minoricensis</i> (4)	-	0.7 (2)	1.0 (3)	85.5 (247)	12.8 (37)	100.0 (289)
<i>L. triloba</i> subsp. <i>triloba</i> (5)	0.6 (14)	1.5 (35)	1.4 (33)	1.1 (25)	95.4 (2218)	100.0 (2325)
Overall						92.5 (4082)

To compare these two species, Spanish populations of *L. triloba* subsp. *triloba* and Moroccan seed accessions of *L. flava* were also analysed and compared, achieving an identification performance of 95.4% (Table 8). Seeds of *L. triloba* subsp. *triloba* were clearly distinguishable and only 4.3% were mistaken and classified within *L. flava*. Furthermore, only 6.1% of *L. flava* seeds were mistaken as *L. triloba* subsp. *triloba*.

Finally, to evaluate the inter-population variability of *L. triloba* subsp. *triloba*, Spanish and Sardinian seed accessions were compared, achieving an identification performance of 99.6%. Seeds were clearly distinguishable for geographical region of provenance with slightly misattributions of 0.4% (data not shown).

Table 8. Percentage of correct identification between Moroccan *Lavatera flava* and Spanish *L. triloba* subsp. *triloba* populations. In parenthesis, the number of analyzed seeds.

	<i>L. flava</i>	<i>L. triloba</i> ssp <i>triloba</i>	Total
<i>L. flava</i>	93.9 (170)	6.1 (11)	100.0 (181)
<i>L. triloba</i> ssp <i>triloba</i>	4.3 (44)	95.7 (981)	100.0 (1025)
Overall			95.4 (1206)

The four *Lavatera* species of *Anthema* sectio were clearly distinguished by means of morpho-colorimetric seed traits as reported in the LDA graphical representation (Fig. 1) (data not shown).

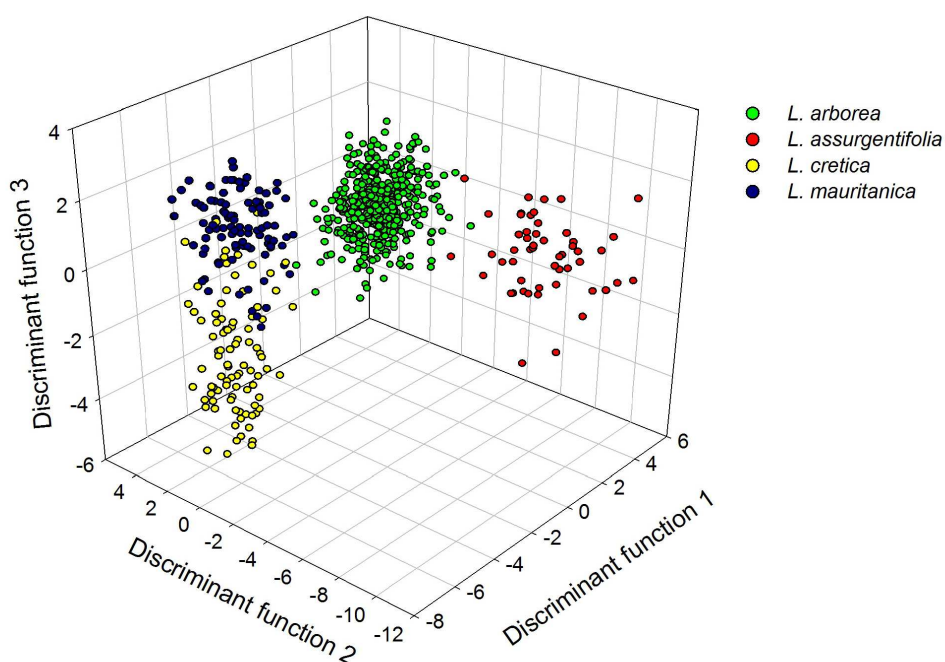


Figure 1. Graphical representation of the discriminant analysis for *Lavatera* sectio *Anthema*.

The *Olbia* sectio, involving eight *Lavatera* species, was also investigated and a correct classification of 97.3% was recorded. All the species were distinguished with percentage above 90% except for *L. maroccana* (83.3%), misclassified mainly as *L. punctata* (6.7%) and *L. bryonifolia* (5.6%) (Table 9). No comparison was carried out for *Bismalva* sectio being *L. moschata* the only available species for this group.

Table 9. Percentage of correct identification among *Lavatera* accessions sectio *Olbia*. In parenthesis, the number of analyzed seeds.

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	Total
<i>L. bryonifolia</i> (1)	90.9 (10)	9.1 (1)	-	-	-	-	-	-	100.0 (11)
<i>L. maroccana</i> (2)	5.6 (5)	83.3 (75)	-	-	-	6.7 (6)	3.3 (3)	1.1 (1)	100.0 (90)
<i>L. oblongifolia</i> (3)	-	-	97.0 (194)	-	-	-	3.0 (6)	-	100.0 (200)
<i>L. olbia</i> (4)	-	-	-	100.0 (200)	-	-	-	-	100.0 (200)
<i>L. plazzae</i> (5)	-	-	-	-	100.0 (199)	-	-	-	100.0 (199)
<i>L. punctata</i> (6)	-	6.3 (1)	-	-	-	93.8 (15)	-	-	100.0 (16)
<i>L. thuringiaca</i> (7)	-	1.2 (1)	-	-	-	-	98.8 (80)	-	100.0 (81)
<i>L. trimestris</i> (8)	-	-	-	-	-	-	-	100.0 (107)	100.0 (107)
Overall									97.3 (904)

As well as the genus *Lavatera*, also the eight *Malva* accessions were compared by means of LDA statistical elaboration. Data processing led to an overall cross-validated discrimination performance of 95.8% (Fig. 2), recording values of 100% for *M. hispanica* and *M. multiflora*. All the other *Malva* accessions were correctly classified with percentages above 89%, with the exception of *M. tournefortiana* discriminated in the 81.3% of the cases, misclassified as *M. parviflora*, *M. sylvestris* and *M. verticillata* (data not shown).

With the aim to record the inter-specific seed morpho-colorimetric variability, *Malva* species were compared within each section studied. Figures 3 and 4 report the graphical representations of the discriminant analysis for *Malva* sectio *Bismalva* and *Malva* sectio *Malva*, respectively. In both the cases, 99.5% of correct classification was obtained (data not shown). Only *M. multiflora* was present in the *Anthema* sectio, so no comparison was achievable.

Discussion

In order to contribute to the taxonomical treatment affiliation of *Lavatera* and *Malva* groups, their seed morpho-colorimetric features were investigated applying image analysis techniques in order to obtain accurate measurements to be subjected to LDA statistical elaborations.

As shown in Table 4, the two groups were perfectly distinguishable confirming the taxonomic treatment at the genus level. Although some *taxa* of *Lavatera* were closer to the genus *Malva* with respect to seed traits, the percentages of misattributions never exceed more than 20%.

The image analysis system here developed, was clearly able to discriminate among the *Lavatera* sections taken into account, confirming the actual and accepted taxonomic organization for this genus. The high

performance of the statistical classifiers was principally due to the textural and EFDs descriptors, that showed significantly high value of *F*-to-remove, in addition to colorimetric (RGB and HLS colour channels) and densitometric features, selected among the available 138. The textural variables and EFDs introduced in this study, were involved in a large extent in the LDA process, representing the 65% and 51% of the whole variable pool of the discriminant function referred to *Lavatera* sectio *Axolopha* and *Anthema* respectively (data not shown).

As reported in Bacchetta *et al.* (2011a), an overall cross-validated percentage of correct identification above 92.5% was again obtained in the discrimination of the five species included in the *Lavatera* sectio *Glandulosae* (Table 5), but the improvement of the image analysis system previously developed by Grillo *et al.* (2010) in which morphometric features were the first discriminant parameters, allowed to reinforce the classifier performance relating to *L. flava*.

In fact, while keeping the same number of seeds compared to previous work, the percentage of correct classification increased from 60.2 % to 84.0%, and a similar trend of misattributions toward *L. triloba* subspecies was observed. Also in this case, the most discriminating parameters were descriptive of colorimetric and textural traits of the seed surface, highlighting the importance of the introduction of the new set of descriptors.

Taking into account *L. triloba* group, *L. triloba* subsp. *pallescens* was classified with a lower efficiency (88.8%) compared to the above mentioned study of Bacchetta *et al.* (100%) because the within variability of the group significantly increased at the performance expense.

Regarding to *L. triloba* subsp. *minoricensis*, the obtained data unambiguously confirmed this *taxa* as an independent subspecies (Escobar Garcia *et al.*,

2010), on the basis of marked seed morpho-colorimetric traits differentiation with respect to the other *L. triloba* subspecies.

In this sense, the taxonomic relationships between *Lavatera triloba* s.l. and *L. flava*, whose similarity with *L. triloba* is absolutely comparable to some of the entities now voted as *L. triloba* subspecies, have yet to be clarified in detail: this morpho-colorimetric investigation, definitely confirms the close relationship between these entities, often morphologically well differentiated.

Within *L. triloba*, a quite perfect differentiation between the Spanish and Sardinian populations was found. Isolation and genetic divergences of the species in Sardinia may explain this outcome (Bacchetta *et al.*, 2011a).

Similarly, in the classification among the *Malva* sections, the best variables were the seed colorimetric and densitometric features, in addition to a few of other dimensional parameters. The new texture and shape descriptors occurred with total percentage of over 30% in the discriminant function, chosen by stepwise LDA (data not shown).

In conclusion, by analyzing morpho-colorimetric seed traits, a clear differentiation between the entities of *Malvae* alliance, traditionally attributed respectively to the genus *Lavatera* and to the genus *Malva*, was highlighted. Regarding to the *Lavatera* genus entities closer to the genus *Malva*, a degree of similarity generally never exceeding 20% was recognized. These results confirm in part the differentiation of Ray (1995) of the two *Malva* alliance clades, supporting once again the validity of the two “historical” genera of *Lavatera* and *Malva*, and confirming that the assimilation of all entities to the single *Malva* genus, proposed by Banfi *et al.* (2005) does not match the morphology of germplasm.

Furthermore, the results obtained, both at species and section level, supported the image analysis tool as diagnostic for systematic purposes and the introduction of Haralick's and EFDs variables proved useful for the system implementation.

Acknowledgement

I would thank to Prof. Gianluigi Bacchetta, Dr. Oscar Grillo, Dr. Gianfranco Venora for their great contribution to the writing of this chapter. Thanks also to Dr. Pedro Escobar Garcia of *Department of Biogeography and Botanical Garden, Faculty Centre Biodiversity - University of Vienna* and Francesco Mascia for their suggestions, revision and kind collaboration at this study.

References

- ALVIN C.R. & WILLIAM F.C. 2012. *Methods of Multivariate Analysis*. 3rd edition. John Wiley & Sons.
- BACCHETTA G., BUENO SANCHEZ A., FENU G., JIMENEZ-ALFARO B., MATTANA E., PIOTTO B. VIREVAIRE M. 2008b. *Conservacion ex situ de plantas silvestres*. Principado de Asturias / La Caixa.
- BACCHETTA G., FENU G., GRILLO O., MATTANA E. VENORA G. 2011b. Identification of Sardinian species of *Astragalus* section *Melanocercis* (Fabaceae) by seed image analysis. *Annales Botanici Fennici* 48, 449-454.
- BACCHETTA G., GARCÍA P.E. GRILLO O., MASCIA F. VENORA G. 2011a. Seed image analysis provides evidence of taxonomical differentiation within the *Lavatera triloba* aggregate (Malvaceae). *Flora* 206, 468-472.
- BACCHETTA G., GRILLO O., MATTANA E., VENORA G. 2008a. Morpho-colorimetric characterization by image analysis to identify diaspores of wild plant species. *Flora* 203, 669-682.
- BAKER W.R. 1890. Synopsis of genera and species of Malveae. *Journal of Botany* 28, 140-145, 207-213, 239-243, 339-343, 367-371.
- BANFI E., GALASSO G., SOLDANO A. 2005. Notes on systematics and taxonomy for the Italian vascular flora. I. *Atti Società italiana di Scienze naturali*. Museo civico di Storia naturale di Milano, Milano, 146, 219-244.
- BAYER C. & KUBITZKI K. 2003. Malvaceae, In: Kubitzki K. (Ed.). *The Families and Genera of Vascular Plants* 5, 225-311.
- BOX G.E.P. 1949. A general distribution theory for a class of likelihood criteria. *Biometrika* 36, 317-346.
- CERABOLINI B., CERIANI R.M., CACCIANIGA M., DE ANDREIS R., RAIMONDI B. 2003. Seed size, shape and persistence in soil: a test on Italian flora from Alps to Mediterranean coasts. *Seed Science Research* 13, 75-85.
- DALBY D.H. 1968. *Malva*. In: Tutin, T.G. et al. (Eds.). *Flora Europaea*, vol. 2. Cambridge University Press, Cambridge, pp. 249-251.
- DE CANDOLLE A.P. 1824. *Prodromus Systematis Naturalis Regni Vegetabilis*. Treuttel and Wurtz, Paris.
- DIAMOND J., ANDERSON N.H., BARTELS P.H., MONTIRONI R. HAMILTON P.W. 2004. The use of morphological characteristics and texture analysis in the

- identification of tissue composition in prostatic neoplasia. *Human Pathology* 35, 1121-1131.
- DRAPER, S.R. & KEEFE, P.D. 1989. Machine vision for the characterization and identification of cultivars. *Plant Varieties and Seeds* 2, 53-62.
- ELLUL P., BOSCAIU M., VICENTE O., MORENO V. ROSELLÓ, J.A. 2002. Intra- and interspecific variation in DNA content in *Cistus* (Cistaceae). *Annals of Botany* 90, 345-351.
- ESCOBAR GARCIA P., MASCIA F., BACCHETTA G. 2010. Typification of the name *Lavatera triloba* subsp. *pallescens* (Moris) Nyman and reassessment of *L. minoricensis* Cambess. (*L. triloba* subsp. *minoricensis* comb. nova). *Anales del Jardín Botánico de Madrid* 67, 79-86.
- ESCOBAR GARCÍA P., SCHÖNSWETTER P., FUERTES AGUILAR J., NIETO FELINER G., SCHNEEWEISS G.M. 2009. Five molecular markers reveal extensive morphological homoplasy and reticulate evolution in the *Malva* alliance (Malvaceae). *Molecular Phylogenetics and Evolution* 50, 226-239.
- FERNANDES R.B. 1968a. Contribuições para o conhecimento do género *Lavatera* L. I. Notas sobre algumas espécies. *Collect. Bot. (Barcelona)* 7, 393-448.
- FERNANDES R.B. 1968b. Contribuições para o conhecimento do género *Lavatera* II: taxonomia. *Bol. Soc. Port. Ci. Nat. Ser. 2a* 12, 67-103.
- FISHER R.A. 1936. The use of multiple measurements in taxonomic problems. *Annals of Eugenics* 7, 179-188.
- FISHER R.A. 1940. The precision of discriminant functions. *Annals of Eugenics* 10, 422-429.
- FUERTES AGUILAR J., RAY M.F., FRANCISCO-ORTEGA J., SANTOS-GUERRA A., JANSEN R.K. 2002. Molecular evidence from chloroplast and nuclear markers for multiple colonizations of *Lavatera* (Malvaceae) in the Canary Islands. *Systematic Botany* 27, 74-83.
- GASTWIRTH J.L., GEL Y.R., MIAO W. 2009. The impact of Levene's test of equality of variances on statistical theory and practice. *Statistical Science* 24, 343-360.
- GERGER, A. & SMOLLE, J. 2003. Diagnostic imaging of melanocytic skin tumors. *Journal of Cutaneous Pathology* 30, 247-252.

- GRANITTO P.M., GARRALDA P.A., VERDES P.F., CECCATO H.A. 2003. Boosting classifiers for weed seeds identification. *Journal of Computer Science and Technology* 3, 34-39.
- GRILLO O., DRAPER D., VENORA G., MARTÍNEZ-LABORDE J.B. 2012. Seed image analysis and taxonomy of *Diploaxis* DC. (Brassicaceae Brassicaceae). *Systematic and Biodiversity* 10, 57-70.
- GRILLO O., MATTANA E., FENU G., VENORA G., BACCHETTA G. 2013. Geographic isolation affects inter- and intra-specific seed variability in the *Astragalus tragacantha* complex, as assessed by morpho-colorimetric analysis. *Comptes Rendus de Biologies* 336, 102-108.
- GRILLO O., MATTANA E., VENORA G., BACCHETTA G. 2010. Statistical seed classifiers of 10 plant families representative of the Mediterranean vascular flora. *Seed Science and Technology* 38, 455-476.
- GRILLO O., MICELI C., VENORA, G. 2011. Computerised image analysis applied to inspection of vetch seeds for varietal identification. *Seed Science and Technology* 39, 490-500.
- HABERMAN S.J. 1973. The analysis of residuals in cross-classified tables. *Biometrics* 29, 205-220.
- HARALICK R.M. & SHAPIRO L.G. 1991. Glossary of computer vision terms. *Pattern Recognition* 24, 69-93.
- HARALICK R.M. 1979. Statistical and structural approaches to texture. *Proceedings of the IEEE* 67, 786-804.
- HARALICK R.M., SHANMUGAM K., DINSTEN I. 1973. Textural features for image classification. *IEEE Transactions on Systems. Man and Cybernetics* 6, 610-621.
- HASTIE T., TIBSHIRANI R., FRIEDMAN J. 2001. *The elements of statistical learning: Data mining, inference, and prediction*. New York: Springer.
- HOLDEN J.E., FINCH W.H., KELLY K. 2011. A Comparison of two-group classification methods. *Educational and Psychological Measurement* 715, 870-901.
- KAWABATA S. YOKOO M., NII K. 2009. Quantitative analysis of corolla shapes and petal contours in single-flower cultivars of *Lisianthus*. *Scientia Horticulturae* 121, 206-212.

- KILIC K., BOYACI I.H., KOKSEL H., KUSMENOGLU U.I. 2007. A classification system for beans using computer vision system and artificial neural networks. *Journal of Food Engineering* 78, 897-904.
- KLITGÅRD B.B. 2013. Neotropical Malvaceae (Bombacoideae). In: Milliken, W., Klitgård, B. & Baracat, A. (2009 onwards), *Neotropikey - Interactive key and information resources for flowering plants of the Neotropics*.
- KUHL F.P. & GIARDINA C.R. 1982. Elliptic Fourier features of a closed contour. *Computer Graphics* 18, 259-278.
- KUHN M. & JOHNSON K. 2013. Discriminant analysis and other linear classification models. In: *Applied Predictive Modeling* pp. 275-328. Springer New York. ISBN: 978-1-4614-6848-6
- LEVENE H. 1960. Robust tests for equality of variances. In: Olkin, I., Ghurye, S.G., Hoeffding, W., Madow, W.G. & Mann H.B., Eds., *Contributions to Probability and Statistics: Essays in Honor of Harold Hotelling*. Stanford University Press. pp. 278-292.
- LINNAEUS C. 1753. *Species Plantarum*. A 1957 facsimile of the first edition of 1753 with an introduction by W.T. Stearn and an appendix by J.L. Heller and W.T. Stearn. Ray Society, London.
- LO BIANCO M., FERRER-GALLEGO P., GRILLO O., LAGUNA E., VENORA G. & BACCHETTA G. Seed image analysis provides evidence of taxonomical differentiation within the *Medicago* L. sect. *Dendrotelis* (Fabaceae). *Systematics and Biodiversity* (Manuscript submitted for publication).
- LOPES R.H.C. 2011. Kolmogorov-Smirnov Test. In: Lovric M., Eds., *International Encyclopedia of Statistical Science*. Springer Berlin Heidelberg. pp. 718-720.
- MAHALANOBIS P.C. 1936. On the generalized distance in statistics. *Proceedings of the National Institute of Science of India* 12, 49-55.
- MEBATSION H.K., PALIWAL J., JAYAS D.S. 2012. Evaluation of variations in the shape of grain types using principal components analysis of the elliptic Fourier descriptors. *Computers and Electronics in Agriculture* 80, 63-70.
- MORRISON D.F. 2004. *Multivariate Statistical Methods*. 4th edition. Cengage Learning Duxbury Press.
- MORTON C.V. 1937. The correct names of the small-flowered mallows. *Rhodora* 39, 98-99.

- NANNI L., SHI J.Y., BRAHNAM S., LUMINI A. 2010. Protein classification using texture descriptors extracted from the protein backbone image. *Journal of Theoretical Biology* 264, 1024-1032.
- ORRÙ M., GRILLO O., LOVICU G., VENORA G., BACCHETTA G. 2013. Morphological characterisation of *Vitis vinifera* L. seeds by image analysis and comparison with archaeological remains. *Vegetation History and Archaeobotany* 22, 231-242.
- ORRÙ M., GRILLO O., VENORA G., BACCHETTA G. 2012. Computer vision as a complementary to molecular analysis: grapevines cultivars case study. *Comptes Rendus de Biologies* 335, 602-615.
- PÉREZ-GARCÍA F., GONZÁLEZ-BENITO M.E., GÓMEZ-CAMPO C. 2007. High viability recorded in ultradrying seeds of 37 species of Brassicaceae after almost 40 years of storage. *Seed Science and Technology* 35, 143-153.
- PINNA M.S., GRILLO O., MATTANA E., CAÑADAS E.M. BACCHETTA G. 2014. Inter- and intraspecific morphometric variability in *Juniperus* L. seeds (Cupressaceae). *Systematics and Biodiversity* 12, 211-223.
- RAY M.F. 1995. Systematics of *Lavatera* and *Malva* (Malvaceae, Malveae) a new perspective. *Plant Systematic and Evolution* 198, 29-53.
- ROVNER I. & GYULAI F. 2007. Computer-assisted morphometry: a new method for assessing and distinguishing morphological variation in wild and domestic seed populations. *Economic Botany* 61, 154-172.
- SABATO D., ESTERAS C., GRILLO O., PICÓ B. & BACCHETTA G. 2015. Seeds morpho-colorimetric analysis as complementary method to molecular characterization of melon diversity. *Scientia Horticulturae* (Manuscript submitted for publication).
- SANTO A., MATTANA E., GRILLO O. & BACCHETTA G. 2015. Morpho-colorimetric analysis, germination variability and heteromorphy of *Brassica insularis* Moris (Brassicaceae) seeds. *Plant Biology*, doi:10.1111/plb.12236.
- SHAHIN M.A. & SYMONS S.J. 2003a. Lentil type identification using machine vision. *Canadian Biosystems Engineering* 45, 3.5-3.11.
- SHAHIN M.A. & SYMONS S.J. 2003b. Colour calibration of scanners for scanner independent grain grading. *Cereal Chemistry* 80, 285-289.

- SMYKALOVA I., GRILLO O., BJELKOVA M., PAVELEK M., VENORA G. 2013. Phenotypic evaluation of flax seeds by image analysis. *Industrial Crops and Products* 47, 232-238.
- SPSS 2007. Base 16.0 Application Guide. Prentice Hall, USA, New Jersey.
- SUGIYAMA M. 2007. Dimensionality reduction of multimodal labeled data by local Fisher discriminant analysis. *The Journal of Machine Learning Research* 8, 1027-1061.
- TATE J.A., FUERTES AGUILAR J., WAGSTAFF S.J., LA DUKE J.C., BODO SLOTTA T.A., SIMPSON B.B. 2005. Phylogenetic relationships within the tribe *Malveae* (Malvaceae, subfamily Malvoideae) as inferred from ITS sequence data. *American Journal of Botany* 92, 584-602.
- TERRAL J., TABARD E., BOUBY L., IVORRA S., PASTOR T., FIGUEIRAL I., PICQ S., CHEVANCEJ.B., JUNG C., FABRE L., TARDY C., COMPAN M., BACILIERI R., LACOMBE T., THIS P. 2010. Evolution and history of grapevine (*Vitis vinifera*) under domestication: new morphometric perspectives to understand seed domestication syndrome and reveal origins of ancient European cultivars. *Annals of Botany* 105, 443-455.
- TOURNEFORT J.P. 1706. Suite de l'établissement de quelque nouveaux genres de plants: *Lavatera*. *Histoire de L'Academic Royale des Sciences*. 1731, 83-87.
- TUTIN T.G., HEYWOOD V.H., BURGESS N.A., VALENTINE D.H, WALTERS S.M., WEBB D.A. (EDS.) (1964-1980). *Flora Europaea*, vol. 2-5. Cambridge University Press.
- UCCHESU M., ORRÙ M., GRILLO O., VENORA G., USAI A., SERRELI P.F., BACCHETTA G. 2015. Earliest evidence of a primitive cultivar of *Vitis vinifera* L. during the Bronze Age in Sardinia (Italy). *Vegetation History and Archaeobotany*, 10.1007/s00334-014-0512-9.
- VENORA G., GRILLO O., RAVALLI C., CREMONINI R. 2009b. Identification of Italian landraces of bean (*Phaseolus vulgaris* L.) using an image analysis system. *Scientia Horticulturae* 121, 410-418.
- VENORA G., GRILLO O., SACCONI R. 2009a. Quality assessment of durum wheat storage centres in Sicily: Evaluation of vitreous, starchy and shrunken kernels using an image analysis system. *Journal Cereal Science* 49, 429-440.

- VENORA G., GRILLO O., SHAHIN M.A., SYMONS, S.J. 2007. Identification of Sicilian landraces and Canadian cultivars of lentil using image analysis system. *Food Research International* 40, 161-166.
- WIESNEROV D. & WIESNER L. 2008. Computer image analysis of seed shape and seed color for flax cultivar description. *Computers and Electronics in Agriculture* 61, 126-135.
- 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. *Annals of Botany* 94, 657-664.

Conclusions

The focus of this doctoral thesis concerned the application of Image Analysis technique for an adequate definition of the seed morpho-colorimetric parameters, that represents an important diagnostic factor in the plant taxonomy studies and consequently may be of great help for the improvement of the management and the effective *ex situ* conservation in the germplasm banks.

The first part of this study was related to the computer vision fundamentals and statistical treatment of data. Based on the clear concept that the discriminant ability of a classification system depends not only on the intra-specific representativeness of *taxa* analyzed, but also on the quality and quantity of the parameters measured and used to discriminate between groups of belonging, a new set of recordable shape and texture variables was introduced in a yet consolidated image analysis system, with the aim to improve the performance of the classifiers. In the second part of this dissertation some case studies are reported as applications of this innovative technology, with the aim to prove its great usefulness for systematic purposes. The results here presented confirmed that an extensive database of morpho-colorimetric traits may be applied for taxonomy screening of species groups, comparing with the current systematic, as well as with groupings more recently revealed by genetic studies.

The morpho-colorimetric characterization of *Cistus* L. (Cistaceae) seeds by image analysis was treated: a database of morphometric and colorimetric data was carried out to statistically discriminate and identify both at inter and intra-specific and populational level. The satisfactory

discrimination performances reached agree with the results reported in the previous papers on the same *taxa*, though, the improvement of the image analysis system adopted, in which an overall of 138 seed features was evaluated, allowed to reinforce the discrimination power also when the morphometric variability within each group, such as in inter-population groups, is extremely reduced. The taxonomical differentiation within the *Medicago* L. sect. *Dendrotelis* (Fabaceae) was also described. The obtained results confirmed the validity of the proposed method for the taxonomic differentiation of *Medicago* at specific levels, and its identification capability of regional and populational groups. Furthermore, the relationships among 79 *taxa* belonging to the *Lavatera* and *Malva* genera were treated, in order to contribute to their doubtful systematic treatment.

Finally, this innovative kind of identification system, which method was specifically developed to identify wild seeds, and that requires only a few seconds for scanning and measurement operations, proved to be a quick, repeatable, reliable and non destructive method. It does not require any chemical reagents, expensive analytical consumables or high priced physical preparation of samples, hence it is a very cheap method. This precise and accurate identification system it was only possible thanks to efficient and useful cooperation between taxonomists and image analysis specialists. The expert, practical experience in such different fields allowed the development of a system so complex in its structure and so simple in the use. Indeed, having a broad database of morpho-colorimetric seed features for an adequate amount of families, genera and species would enable the identification of *taxa* already present in the database. In this way, this innovative tool would open new perspectives in plant taxonomy, but also offer the opportunity for germplasm banks to make identifications in a standard, speedy way.

In addition, the availability of morpho-colorimetric data should be helpful for ecological and/or archeobotanical studies such as the prediction of seed persistence in the soil.

Acknowledgements

A heartfelt thank you to my tutor, Prof. Gianluigi Bacchetta, coordinator of the PhD course in *Botanica Ambientale ed Applicata*, for allowing me to achieve this title with excellent and patient assistance. I wish express my warm thanks to Dr. Gianfranco Venora, director of the Stazione Sperimentale di Granicoltura per la Sicilia, for being kindly helpful in transmitting his expertise and for giving me so much time with great elasticity, despite all my other professional (and not) commitments.

In particular, I extend a special ‘thank you’ to Dr. Oscar Grillo, co-tutor of my doctoral thesis, precious supervisor of this manuscript, but above all a sincere friend who supported me in difficult times.

Many thanks to all the staff of the *Centro Conservazione Biodiversità* of the *Banca del Germoplasma della Sardegna - Università degli Studi di Cagliari*, particularly to Martino Orrù, Andrea Santo and Alba Cuenca Lombraña for their friendly help in providing me useful data for my research. Thanks also to all my PhD colleagues and to Diego Sabato who readily kept me informed, despite the distance. I must extend my thanks to Eva Cañadas and Francesco Mascia for their scientific support.

I would like to express my appreciation to Dr. Pedro Escobar Garcia of *Department of Biogeography and Botanical Garden, Faculty Centre Biodiversity - University of Vienna*, and Dr. Emilio Laguna and Dr. Pablo Ferrer-Gallego of *Servicio de Vida Silvestre, Centro para la Investigación y Experimentación Forestal (CIEF) - Valencia*, for their critical and meaningful review of the case studies chapters.

Thanks also to the researchers of all the institutions that kindly provided seed material: Dr. Christine Fournaraki of *MAICh - Crete*, Dr.

Esteban Bermejo and Dr. Paqui Herrera Molina of *Jardin Botanico de Cordoba*, Dr. Caroline Favier of *Conservatoire Botanique National de Corse*, Prof. Salvatore Brullo of *Dipartimento di Botanica - Università degli studi di Catania* and Prof. Luigi Forte of *Orto Botanico - Università degli studi di Bari*.

And finally, I am enormously grateful to my family, Salvo and Umberto, for encouraging me to achieve my professional goals, and to my parents for their tireless presence in support as daughter and as mom.