UNIVERSITÀ DEGLI STUDI DELLA TUSCIA DI VITERBO

Dipartimento per l'Innovazione nei sistemi Biologici, Agroalimentari e Forestali (DIBAF)

Corso di Dottorato di Ricerca in

SCIENZE AMBIENTALI - XXV Ciclo

A novel classification method to map habitats using multispectral images and ancillary data

BIO/07

Tesi di dottorato di:

Dott. Emiliano Canali

Coordinatore del corso

Prof. Maurizio Petruccioli

Tutore

Dott. Roberto Mancinelli

Co-tutore

Dott. Nicola Lugeri

Giugno 2014

Abstract

Monitoring programmes often use hyper and multispectral images as first data to better understanding the diversity of natural and semi-natural habitats, their spatial distribution, and their conservation status. Despite the large number of studies about remote sensing-based mapping, the vast majority of them focus on the delineation of land cover categories and just a few works aim at mapping habitat types.

This Ph.D. activity presents a novel approach for mapping the spatial distribution of natural habitats using multispectral images and ancillary data. In particular this activity is developed in the frame of the *Carta della Natura* project, envisaged by the Italian Law no. 394 /1991 and carried out by the Institute for Environmental Protection and Research (ISPRA), in order to find more rapid and low-cost tools to produce the *Map of Habitats,* which constitute the basis territorial unit of the whole project.

Partial Least Squares Discriminant Analysis (PLS-DA) combined with GIS and remote sensing procedures was used to map three areas at different scales and with different habitats composition. In order to find the most robust classifier model to be used in the classification PLS-DA technique was tested with a novel recursive algorithm, specifically developed for this activity.

Mapping accuracy was calculated using the official maps produced in the frame of the "*Carta della Natura*" project; moreover, in order to verify the capability of the method with respect to a commercial software of common use, classification results were compared with those obtained using maximum likelihood algorithm available in ESRI ArcGIS.

Results shows a better classification ability of the proposed method than commercial software in all the three areas, with overall accuracies of 55.7% (Monte Vulture volcanic complex), 62.8% (Apulia lagoons) and 72.3% (Campo Pericoli basin). These results, although not very high in absolute terms, can be considered as satisfactory because of the particular context of the study areas, characterized by the complexity and the heterogeneity of their habitats. In particular, the methods shows a good accuracy with the area mapped at the highest scale (Campo Pericoli); these results are very encouraging as habitat such as *Screes* and *Oro-Apennine closed grasslands* are identified in Annex I of the EU Habitats Directive as being of Community interest, and in particular the grasslands are also listed as ‘priority’.

Due to limited data input requirements, the approach developed in this Ph.D. activity is a good alternative to commercial software classifiers and can be used as a starting point for the further steps of photo-interpretation. Moreover it is a flexible method which allows to use, where available, other cartographic base maps, such as land use maps, in order to improve classification

results by excluding habitats of lesser importance for the analysis, which may interfere with the classification.

Riassunto

Le attività di monitoraggio ambientale utilizzano spesso immagini iper- e multi-spettrali come fonte primaria di dati per ottenere una migliore comprensione riguardo la distribuzione e la diversità degli ambienti naturali e semi-naturali. Nonostante il telerilevamento sia un argomento largamente trattato in letteratura, la maggior parte degli studi si focalizza sul riconoscimento e classificazione delle tipologie di copertura ed uso del suolo, e solo pochi lavori si prefiggono come obiettivo la cartografia degli habitat.

La ricerca descritta in questa tesi ha come scopo lo sviluppo di un nuovo metodo di classificazione degli habitat a partire da immagini multi-spettrali e dati ancillari. In particolare, il lavoro si inquadra nelle attività del progetto Carta della Natura, istituito dalla legge quadro sulle aree protette L 394/91 la cui realizzazione è compito dell'ISPRA (Istituto Superiore per la Protezione e la Ricerca Ambientale), allo scopo di sviluppare metodologie rapide ed a basso costo per la produzione della cartografia degli habitat, i quali costituiscono le unità ambientali omogenee di riferimento del progetto.

Il metodo proposto si basa sull'utilizzo della classificazione multivariata supervisionata (Partial Least Square Discriminant Analysis – PLSDA) in combinazione con metodologie di telerilevamento e GIS, per la costruzione di un classificatore e la successiva vettorializzazione e rifinitura finale delle cartografie ottenute. Nella costruzione del classificatore, in particolare, è stato testato un nuovo approccio che propone l'utilizzo ricorsivo della PLS-DA, al fine di individuare il modello più robusto da utilizzare per la discriminazione degli habitat. L'algoritmo è stato sviluppato specificatamente per il presente lavoro di dottorato.

Il metodo è stato applicato su tre aree campione cartografate a diversa scala e contraddistinte da una struttura e composizione degli habitat completamente diversa. L'accuratezza di classificazione è stata verificata utilizzando come riferimento la cartografia ufficiale prodotta nell'ambito del progetto Carta delle Natura; essa risulta essere rispettivamente 62.8% (Lagune pugliesi), 55.7% (Complesso vulcanico del Monte Vulture) and 72.3% (Vallata di Campo Pericoli).

Inoltre, per valutare l'applicabilità del metodo proposto rispetto ai comuni software disponibili in commercio, i risultati sono stati confrontati con quelli ottenuti da cartografie prodotte utilizzando l'algoritmo di massima verosimiglianza disponibile nel software ArcGIS della ESRI. In tutte e tre le aree, il metodo dimostra una migliore capacità di classificazione rispetto a quella ottenuta utilizzando il software commerciale.

I risultati, sebbene non alti in termini assoluti, possono essere considerati soddisfacenti considerando il particolare contesto delle tre aree studio, caratterizzate da elevata complessità

strutturale e da rilevante eterogeneità ambientale. I risultati più incoraggianti si ottengono nella classificazione dell'area contraddistinta dal maggior dettaglio cartografico (Campo Pericoli); in questo contesto il sistema ha riconosciuto habitat di particolare interesse, in quanto inseriti nell'allegato I della Direttiva Habitat, come i *Ghiaioni calcarei e scisto-calcarei montani e alpini* e le *Praterie Compatte Oro-Appenniniche*, che sono inoltre elencate tra gli habitat prioritari della Direttiva.

In conclusione, il metodo sviluppato durante quest'attività di ricerca si pone come valida alternativa all'utilizzo dei classificatori commerciali e rappresenta un punto di partenza nella produzione di carte degli habitat, in quanto le mappe classificate possono essere utilizzate per le successive fasi di foto-interpretazione. La sua flessibilità inoltre permette l'utilizzo di cartografia tematica eventualmente disponibile, come le carte di uso del suolo, allo scopo di migliorare ulteriormente le prestazioni di classificazione, escludendo eventuali habitat che possono risultare di minor interesse, ma che possono interferire con la classificazione.

Acknowledgments

Il lavoro svolto in questa tesi è stato possibile grazie al supporto dell'ISPRA (Istituto Superiore per la Ricerca e la Protezione dell'Ambiente) ed in particolare del Servizio Carta della Natura. Per questo desidero innanzitutto ringraziare Nicola Lugeri per tutta la disponibilità che ha dimostrato, rivedendo tutto il materiale prodotto più e più volte e per i tanti consigli e il sostegno che non ha mai fatto mancare.

Ringrazio, inoltre, tutti i colleghi del servizio Carta della Natura, su tutti Alberto e Robertino, che mi hanno accolto e accompagnato in questi anni e permesso di migliorare il lavoro grazie al loro aiuto, al quotidiano confronto e alle osservazioni e critiche sempre pertinenti e costruttive.

Ringrazio il Prof. Maurizio Petruccioli ed il Dr. Roberto Mancinelli per avermi supportato in tutte le attività.

Ringrazio Domenico Collalti della Regione Abruzzo e l'IptSAT, ed in particolare Fabiano Campo e Stefano De Corso, per la fornitura dei dati necessari alla classificazione.

Grazie a tutti gli ex-colleghi del CRA-ING, primi tra tutti Corrado e Francesca, per tutto quello che mi hanno insegnato e mi hanno lasciato, sia dal punto di vista umano che professionale.

Grazie a Roberta, semplicemente perché c'è…

Infine grazie alla mia famiglia e a tutti quelli che sono stati e continuano ad essere al mio fianco in tutti i momenti davvero importanti ed in particolare ad Alberto, Francesca (che sono gli stessi di prima, e sono prima amici che colleghi), Stefano, Valentina, Silvia, Valerio e Francesco.

Table of contents

List of figures

List of tables

Abbreviations

PCA	Principal Component Analysis
PLS	Partial Least Square Regression
PLS-DA	Partial Least Square Discriminant Analysis
LDA	Linear Discriminant Analysis
NIPALS	Non-Iterative PArtial Least Squares
EU	European Union
IUCN	International Union for Conservation of Nature
EEA	European Environment Agency
ISPRA	Istituto Superiore per la Protezione e la Ricerca Ambientale
ENEA	Agenzia nazionale per le nuove tecnologie, l'energia e lo sviluppo economico sostenibile
AGEA	AGenzia per le Erogazioni in Agricoltura
INSPIRE	INnfrastructure for SPatial Information in Europe
EUNIS	EUropean Nature Information System
Corine	COoRdination of INformation on the Environment
GIS	Geographic Information Systems
EIA	Environmental Impact Assessment
SEA	Strategic Environmental Assessment
SCI	Site of Community Importance
SPA	Special Protection Area
NIR	Near Infrared
SPOT	Satellite Pour l'Observation de la Terre
HRVIR.	Haute Résolution dans le Visible et l'Infra-Rouge
MSI	Multi-Spectral Imager
GSD	Ground Sampling Distance
SAM	Spectral Angle Mapper

ML	Maximum Likelihood algorithm
DN	Digital number DN
LV	Latent Variables
RD	Random selection partition method
KS	Kennard–Stone partition method
SPXY	Sample set Partitioning based on joint x–y distances partition method
DEM	Digital Elevation Model
WGS84	World Geodetic System 1984
NDVI	Normalized Difference Vegetation Index
NDRE	Normalised Difference Red Edge Index

Introduction

1.1 Thesis objective

General aim of this Ph.D. is to develop a novel classification method for multivariate dataset. In particular in the context of habitat mapping it aims to obtain a first set of classified data to be used as starting point in map production, reducing the amount of time spent in the visual photo-interpretation which thus could become the final map-refining step.

For high-dimensional multiclass classification problems a suitable approach lies in the use of multivariate exploratory approaches such as the commonly used Principal Component Analysis (PCA) and the Partial Least Square (PLS) based techniques. In particular, PLS has received much attention in high-dimensional classification problems especially in the fields of computational biology (Chung and Keles, 2010) and chemometrics (Sabatier et al., 2003; Bylesjo et al., 2006; Tominaga, 2006; Menesatti et al., 2008; Antonucci et al., 2012; Costa et al., 2011), despite being criticized for its lack of theoretical justifications. Much work still needs to be done to demonstrate all statistical properties of PLS (Krämer and Sugiyama, 2011). Nevertheless, this computational and exploratory research is extremely popular thanks to its efficiency and predictive ability (Lê Cao et al. 2009).

In this thesis a PLS based multivariate technique (Partial Least Squares Discriminant Analysis, PLS-DA) is used with GIS and remote sensing procedures in order to classify habitats using multispectral images and ancillary data.

1.2 Habitat mapping

In the last decades identification, description, classification and mapping of natural and semi-natural habitats are gaining recognition in the sphere of environmental policy implementation.

Key policy instruments, such as the Habitats Directive and the Bern Convention, implicitly address the need for spatial data necessary to assess the state of environment and to inform long-term and forward planning decision making (EEA, 2014).

Although there's a well established tradition of vegetation maps, usually defined by species composition, after the start of new EU policy orientation (especially with Habitats directive) focused on habitat protection, maps began to be produced to show habitat types or biotopes.

In particular, articles 11 and 17 of the Habitats Directive require member states to report on four parameters of habitat conservation status every 6 years: habitat area, range, indicators of

habitat quality, and future prospects for habitat survival in the member state (European Commission DG Environment 2007; Vanden Borre et al. 2011). In this context habitat maps reporting the distribution of the Annex I habitat types became mandatory.

Habitat maps are also a very useful input to processes of spatial planning, including Environmental Impact Assessments (EIAs) and assessments required under Article 6 of the Directive to protect the Natura 2000 network. They have been used when designing ecological networks from regional to continental scales, as with the Pan-European Ecological Network (PEEN), and will be important in implementing the European Commission's Green Infrastructure Strategy.

Besides that, similar information is also required for compiling the Red Lists of habitats, typically at the national level and currently under development at the European level. The European Red List of habitats will contribute to an IUCN (International Union for Conservation of Nature) led initiative for the Red List of the world's ecosystems. Habitat maps are expected to play an important role in mapping and assessing ecosystem services as ecosystems can be regarded as groupings of habitat types. Typologies based on phytosociological classification are strictly defined by plant communities, whereas habitat types or biotopes take into account geographic, abiotic and biotic features.

Habitat is a widely used term but it has many interpretations developed in different contexts with contrasting meanings (Bunce, et al., 2013). In the ecological sense the use of the term is organism-specific (Miller, 2000; Morrison, 2001; Odum, 1971; Whittaker et al., 1973) and it is explicitly linked to a species or species group that share the same environmental and ecological requirements. In this definition, the behaviour of a given species in terms of foraging, roosting and nesting as well as their competitors and predators all play important roles in determining suitable conditions for a given species or species group.

Another use of the term habitat is landscape-specific (Löfvenhaft et al., 2002) and refers to areas with a defined species composition (both fauna and flora) and associated physical factors (e.g. climate and soil type). It is the common meaning used in Nature Protection Policy by European Union and in the Corine Biotopes, EUNIS (EUropean Nature Information System) and Palaearctic habitats classifications. This definition being the habitat to be characterized by observable and recognizable features and is very convenient to discern from landcover data, aerial photos or satellite images based on geographic information systems (GIS) (Miller and Hobbs, 2007) or even sometimes from field vegetation survey and observation (Maurer et al., 2000).

There is a great variety of large-area habitat-mapping projects in Europe. EEA (2014) reviewed the main projects considering their geographical extent, the types of habitat mapped, their scale, and project duration. According to this report the most important experience are produced at national level in Czech Republic (Czech biotope mapping programme), Spain (Natura 2000 habitat inventory and mapping) and Italy (*Carta della Natura* project, see section 5.2) and all of them are mainly based on visual interpretation of aerial orthophoto and validation field survey. This process is highly time consuming; the classification method proposed in this thesis aims to develop new tools in the frame of the project "Carta della Natura" in order to obtain more rapid habitat maps.

1.3 The mapping scale

The mapping (or cartographic) scale is a fundamental concept in mapping activities. A map's scale is a statement about the relationship between map distance and distance in the real world and represents the amount of reduction that has occurred to get the representation of the world on the map.

The map scale affects the amount of information that can be depicted legibly on the map and the way they can be represented. Potential use of maps depends on their scale: mid-scale maps (e.g. 1:100 000 or 1:50 000) are often used for regional planning, environmental impact assessment (EIA) studies as well as many other similar analysis, finer detailed maps (e.g 1:10000 or 1:5000) can be used for local-focusing analysis.

Also the legend that is to be used in the classification depends on the chosen scale of the map; indeed, the discrimination of homogenous objects depends on the recognition of particular common features which emerge in different ways according to the observation scale. Starting from coarser to finer scales the territory is usually described considering first its main landforms and landscape (mountains, hills, plains etc), going through habitats (in associations or individually), until reaching a local level where single vegetation species are detected and mapped. In this respect, the great advantage arising from hierarchical legend systems, is that the proper hierarchical level can be chosen according to the scale of the study.

In this thesis the proposed classification method was tested on two cartographic scales: a mid-scale (1:50.000 for Monte Vulture volcanic complex and Apulia lagoons study areas) and a local scale (1:10.000, for Campo Pericoli basin study area)

1.4 Tools for mapping

Vegetation analysis and habitat mapping over large areas and at larger scales can benefit greatly from remote sensing data (Langley et al., 2001; Nordberg and Evertson, 2003).

Although visual interpretation of aerial photography is the traditional approach, more advanced technologies, including automated image interpretation and satellite imagery, are now in wide use. Satellite imagery for large-area vegetation mapping has been used since the end of the 1980s, and its use has increased since the end of the 1990s; today there is a wide availability of images (see section 2.3) and considerable research has gone into the possibility of producing automated and semi-automated thematic maps representing landscape patterns.

Monitoring programmes often use hyper and multispectral images as first data to better understanding the diversity of natural and semi-natural habitats, their spatial distribution, and their conservation status (Nagendra, 2001; Turner et al., 2003; Kerr and Ostrovsky, 2003; Nagendra et al., 2013).

Despite the large number of studies about remote sensing mapping, the vast majority of them focus on the delineation of land cover categories (see for example Meliadis and Meliadis, 2011; Rimal, 2011; Ojigi, 2006; Sudeesh and Sudhakar Reddy, 2012; Rahman et al., 2012; Mengistu and Salami, 2007) and just a few works aim at mapping habitat types which is much harder to undertake. (Bock, 2003; Keramitsoglou et al., 2005; Corbane et al 2013).

Mapping in less complex habitat mosaics, focusing on the discrimination of large macro-categories is relatively straightforward (Lucas et al., 2007; Lengyel et al., 2008) but is far more challenging where landscapes are more heterogeneous and fine-grained and variation between habitats is more continuous (Varela et al., 2008; Lucas et al., 2011).

For this reason, improving classification accuracy is always a hot research topic. In particular the classification method proposed in this thesis aims to develop new tools in the frame of the project "*Carta della Natura*".

1.5 Thesis structure

This thesis is organised as follows:

Chapter 2 provides a background on multispectral imaging techniques and describes the multispectral images datasets and the most popular multispectral sensors available in commerce.

Chapter 3 discusses the various current methods for classification of multispectral images. Initially a basic distinction is proposed basing on three criteria, i) which kind of pixel information is used, ii) whether training samples are used or not and iii) whether classifiers models are known a priori or generated from data; then the chapter describes some of the most used classification methods and their algorithms, focusing on the Principal Component Analysis (PCA) dimensionality reduction method.

Chapter 4 provides a theoretical background of the Partial Least Square (PLS) techniques describing the main statistical properties and the possibility to use them as classifiers (Partial Least Square Discriminant Analysis - PLS-DA)

Chapter 5 describes the classification scheme used in this thesis and the references maps used to assess the classification accuracy of the proposed method

Chapter 6 specifically concerns the proposed classification method. It describes all data used in the classification tests, the main steps of the classification procedure and finally the accuracy assessment performed to evaluate the suitability of the method.

Chapter 7 shows the results of the classification tests performed on three study areas: i) Monte Vulture volcanic complex, ii) Apulia lagoons and iii) Campo Pericoli basin, showing classification accuracy results both *per se* and in comparison with a common used classification software available in commerce.

Last chapter presents a summary of the results from this thesis, and makes proposals for future work in the field.

Chapter 2

Remote Sensing and multispectral images

Remote Sensing is essentially the process of acquiring data/information about an object or scene without coming into physical contact with that object or scene. More specifically passive remote sensing uses electromagnetic radiation emitted or reflected by the targets of interest to gather information about their nature and composition. Reflected sunlight is the most common source of radiation measured by passive sensors. The science of remote sensing uses a variety of advanced instruments to measure electromagnetic radiation. These instruments may be ground-based, on airborne platforms, or mounted on satellites in orbit. The data they collect are very diverse and can be used for many different purposes from monitoring programmes to the production of thematic maps. Many researches have established the value of remote sensing for characterizing atmospheric and surface conditions. These instruments prove to be one of the most cost effective means of recording quantitative information about our earth (Kalyankar, 2013).

This chapter provides an introduction to the field of multispectral imaging: sections 2.1 and 2.2 discuss the nature of light and spectral signatures, focusing on their use in image classification, section 2.3 initially describes multispectral images datasets and then focuses on the most common multispectral sensors available in commerce.

2.1 Nature of light

Light is an electromagnetic field consisting of oscillating electric and magnetic disturbances that can propagate as a wave through a vacuum as well as through a medium. Electromagnetic spectrum comprises of radio waves, microwaves, infrared rays, visible light, ultraviolet rays, X-rays, and gamma rays (Fig. 2.1).

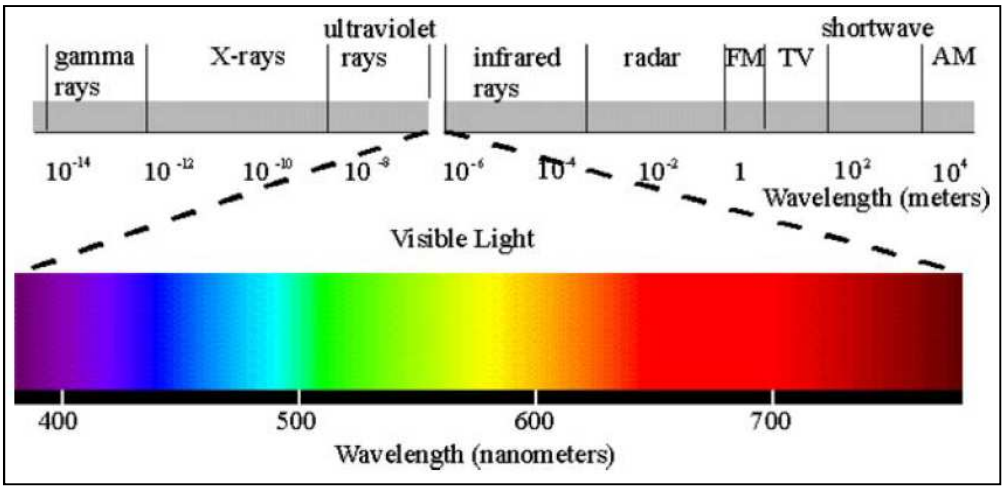


Figure 2.1 - Electromagnetic spectrum (Kaiser and Boynton, 1996)

As indicated, most remote sensing devices make use of electromagnetic energy. However, the electromagnetic spectrum is very broad and not all wavelengths are equally effective for remote sensing purposes. Furthermore, not all have significant interactions with earth surface materials of interest to us. The atmosphere itself causes significant absorption and/or scattering of the shortest wavelengths. In addition, the glass lenses of many sensors also cause significant absorption of shorter wavelengths such as the ultraviolet (UV). As a result, the first significant window (i.e., a region in which energy can significantly pass through the atmosphere) opens up in the visible wavelengths. Even here, the blue wavelengths undergo substantial attenuation by atmospheric scattering, and are thus often left out in remotely sensed images. However, the green, red and near-infrared (IR) wavelengths all provide good opportunities for gauging earth surface interactions without significant interference by the atmosphere. In addition, these regions provide important clues to the nature of many earth surface materials. Chlorophyll, for example, is a very strong absorber of red visible wavelengths, while the near-infrared wavelengths provide important clues to the structures of plant leaves. As a result, the bulk of remotely sensed images used in Geographic information system (GIS) applications are taken in these regions.

Extending into the middle and thermal infrared regions, a variety of good windows can be found. The longer of the middle infrared wavelengths have proven to be useful in a number of geological applications. In this case, the sensors detect thermal radiative properties of the ground, measuring he radiations emitted (emissivity) from the surface of the target, as opposed to optical remote sensing where the measure concerns the reflected portion of the radiation (Prakash, 2000).

The thermal regions have proven to be very useful for monitoring not only the obvious cases of the spatial distribution of heat from industrial activity, but a broad set of applications ranging

from fire monitoring to animal distribution studies to soil moisture conditions. After the thermal IR, the next area of major significance in environmental remote sensing is in the microwave region. A number of important windows exist in this region and are of particular importance for the use of active radar imaging. In this case, sensors transmit a microwave (radio) signal towards a target and detect the backscattered radiation. The strength of the backscattered signal is measured to discriminate between different targets and the time delay between the transmitted and reflected signals (phase) determines the distance (or range) to the target.

The texture of earth surface materials causes significant interactions with several of the microwave wavelength regions. This can thus be used as a supplement to information gained in other wavelengths, and also offers the significant advantage of being usable at night (because as an active system it is independent of solar radiation) and in regions of persistent cloud cover (since radar wavelengths are not significantly affected by clouds).

2.2 Spectral signatures

When electromagnetic energy strikes a material, three types of interaction can follow: reflection, absorption and/or transmission. The reflected portion is usually the portion which is detected by the sensor system. It varies and depends upon the nature of the material and where in the electromagnetic spectrum our measurement is being taken. Remote sensing is definitely based on the ability to discriminate among different ground objects through the analysis of their spectral response, which is usually called spectral signature.

It describes the ratio of the reflected energy to the incident energy (reflectance) as a function of wavelength (expressed as a %). The values of the spectral reflectance are averaged over different, well-defined wavelength intervals (bands); the number and width of them depends on sensor's spectral resolution: the distinctions among the categories of sensors is loose. Usually, a sensor with up to approximately 20 bands is considered as "multispectral"; above that spectral resolution, the sensor is defined "hyperspectral" and can collect data over hundreds to thousands of bands. To obtain the necessary ground truth for the interpretation of hyper and multispectral images, the spectral characteristics of various natural objects have been extensively measured and recorded (fig 2.2).

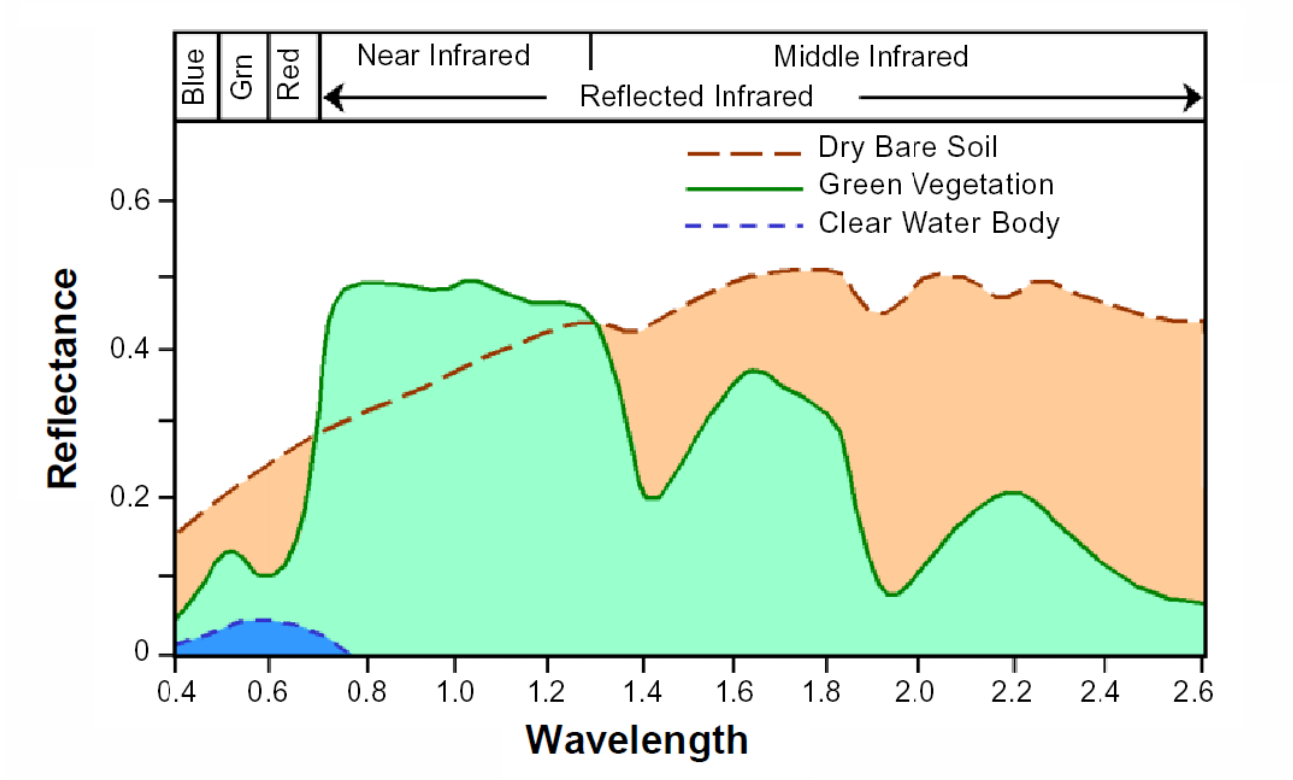


Figure 2.2 - Spectral reflectance of soil, vegetation and water (Penna, 2005)

Spectral signatures that are measured and placed into databases typically represent data collected under very specific conditions of illumination and collection. In a laboratory setting it is relatively easy to determine the reflectance properties of a material because one can easily measure the intensity of the light energy when it hits an object. The light path is controlled, so not much energy is lost between the light source and the target or the target and the detector. Also, the illumination and target orientation are known with high accuracy and precision. In the world of satellite remote sensing, the situation is very different. Remote sensors measure intensity of light when it hits the detector surface (radiance). It depends on the quality and amount of incoming solar radiation, which can vary significantly according to date and time of day (solar zenith angle, path length), and on atmospheric conditions and weather, topography and local orientation of the surface, and intervening features (e.g. forest canopies, shadowing) which may further alter or attenuate it. These additional factors significantly alters the signal before and after interacting with the target affecting the ability to retrieve accurate spectral reflectance values for ground features.

In order to directly compare remote sensing spectra with reference reflectance spectra, the encoded radiance values in the image must be converted to reflectance. Reflectance is generally the standard unit in remote sensing when using information from different sensors, times, or locations,

or when comparing spectral measurements against known reflectance properties of objects (e.g. mineral identification from spectral libraries, geologic exploration) (Peddle et al 2001).

Several strategies are common used to perform comprehensive conversion taking into account the solar source spectrum, lighting effects due to sun angle and topography, atmospheric transmission, and sensor properties.

2.3 Multispectral images

A multispectral image is a multilayer data set formed by a stack of images acquired at different wavelengths (fig 2.3). It is a hypercube where X and Y represent the spatial position and the Z dimension represents the position of the detector's bands in terms of wavelength or frequency of the EM radiation. The data along the Z dimension of the hypercube represents, for each pixel, the spectral information acquired in each band. In the raw data it is expressed as Digital Number (DN), a discrete value which represents the result of the integration of the radiation coming from the selected pixel, over the respective bandwidth. The quantization of this value is another characteristic "resolution" of the sensor, and it is described according to the so-called bit-depth: for example 8-bit images correspond to values ranging from 0 to 255.

The bands are usually equally spaced within the considered region of the electromagnetic radiation. For visual display, each band of the image may be displayed one band at a time as a grey scale image, or in combination of three bands at a time as a colour composite image. The easiest example of classical colour composition are ordinary colour image with only the three bands (red, green and blue) from the visible part of the spectrum; if there are some available additional bands other band-combinations are possible (e.g. false colours).

Interpretation of a multispectral colour composite image will require the knowledge of the spectral reflectance signature of the targets in the scene. In this case, the spectral information content of the image is utilized in the interpretation.

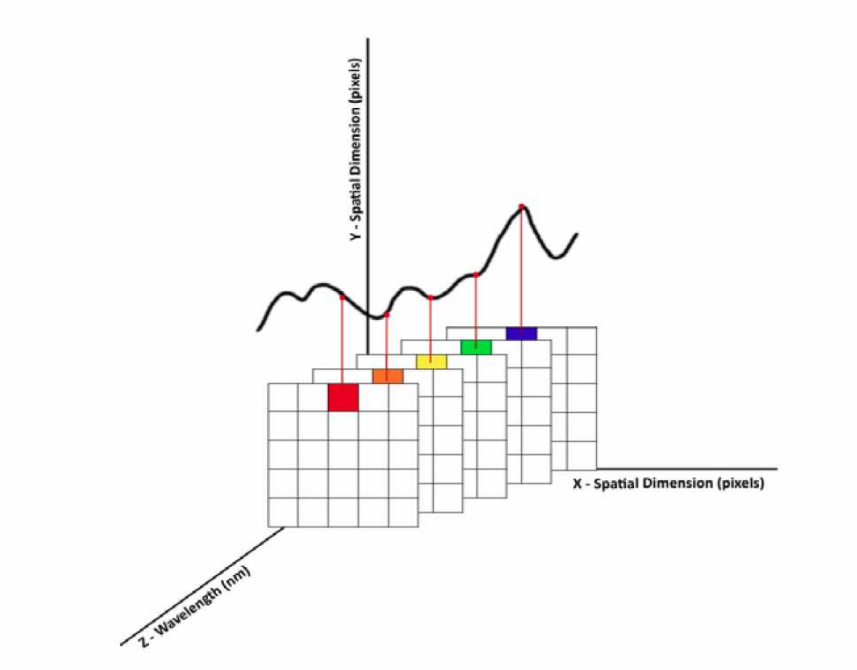


Figure 2.3 - Image data hypercube (Rebollo San Miguel, 2011)

The main elements describing a multispectral image are the number of bands and its spectral and spatial resolution. Spectral resolution refers to the wavelength range that is measured by a imager band. Spatial resolution refers to the X and Y dimensions and represents the size of an image pixel as it is projected on the ground. It is often specified in meters (or kilometres) representing one side of a square area on the ground constituting a pixel.

Many different multispectral instruments have been used in both airborne and space applications. The characteristics of the most commonly sensors used in environmental studies are summarized in Table 2.1 and the most popular missions are described below.

Sensor	Platform	launch year	Spectral Characteristics	Spatial Resolution
ASTER	EOS Terra	1999	Multispectral (14 bands)	15 m (VIS and NIR bands) 30 m (SWIR bands) 90 m (TIR bands)
MISR	EOS Terra	1999	Multispectral (4 bands)	Commandable (275 m, 550 m or 1.1 km)
MODIS	EOS Terra and Aqua	1999	Multispectral (36 bands)	250 m (bands 1–2) 500 m (bands 3–7) 1 km (bands 8–36)
ETM+	Landsat-7	1999	Multispectral (7 band) and Panchromatic	30 m (MS bands) 15 m (PAN band)
HRG	SPOT-5	2002	Multispectral (4 bands) and Panchromatic	10 m (VIS and NIR bands) 20m (SWIR band) until 2.5 m (PAN band)
HRS	SPOT-5	2002	Panchromatic	10m
VEGETATION 2	SPOT-5	2002	Multispectral (4 bands)	1 km
QuickBird	Digital Globe Satellites	2001	Multispectral (4 bands) and Panchromatic	2.4 m (VIS and NIR bands) 0.6 m (PAN band)
WorldView-1	Digital Globe Satellites	2007	Panchromatic	0.55 m
WorldView-2	Digital Globe Satellites	2009	Multispectral (8 bands) and Panchromatic	2.4 m at 20° off Nadir up to 1.8 m at Nadir (MS bands) 0.56 m at 20° off Nadir up to 0.52 m at Nadir (PAN band)
GeoEye-1	GeoEye Satellites	2008	Multispectral (4 bands) and Panchromatic	1.65 m (VIS and NIR bands) 0.41 m (PAN band)
IKONOS	GeoEye Satellites	1999	Multispectral (4 bands) and Panchromatic	3.2 m (VIS and NIR bands) 0.82 m (PAN band)
Rapid-Eye	BlackBridge Satellites	2008	Multispectral (5 bands)	6.5 m (Nadir) - 5 m (orthorectified)
Leica ADS40	Aircraft		Multispectral (4 bands) and Panchromatic	various resolutions

Table 2.1 - Main features of image products from the different sensors

Probably NASA's Landsat satellites has been the most widely used in environmental studies (Artiola et al 2004; Nagendra et al 2013), especially in vegetation and habitat mapping. Since the first Landsat satellite was launched in 1972, a series of more sophisticated multispectral imaging sensors, named TM—Thematic Mapper, has been added ranging from Landsat 4 (1982) to 7 (1999) (Enhanced Thematic Mapper Plus, ETM+). Most Landsat images contain 3 visible light and 4 IR bands at a 30m spatial resolution. Landsat 7 data also provides an 8th panchromatic band (a greyscale band covering the visible portion of the electromagnetic spectrum, basically integrating the red, green, and blue bands of the corresponding multispectral sensor). The first satellite of the SPOT (Système Pour l'Observation de la Terre) series was launched in 1986 with the HRV-sensor (High Resolution Visible). The HRV-sensor has a 10 m panchromatic mode and a three-band 20 m

resolution multispectral mode. In 1998 the SPOT-4 was launched. It mounts a HRVIR sensor, with ad additional middle infrared (MIR) band at the same spatial resolution, and has also onboard the first VEGETATION instrument, that collects data at a spatial resolution of 1 km and a temporal resolution of 1 day, permitting observation at a global level. In 2002 the SPOT-5 was presented. It has a HRG sensor with the same spectral band of SPOT-4. It permits a 10 m resolution in multispectral mode and a 20 m resolution for 1,58 - 1,75 µm spectral range (now called Shortwave Infrared (SWIR) band). Moreover the VEGETATION sensor remains unchanged in comparison to the one installed onboard SPOT 4 and ensures the continuity of global data delivery. IKONOS is a commercial sun-synchronous (which passes over the same part of the Earth at roughly the same local time each day) earth observation satellite launched in 1999 and was the first to collect high-resolution imagery at 1 and 4 m resolution, for civilian purposes. It has two imagery sensors, multispectral and panchromatic. Panchromatic sensor collects image at 1 m while the multispectral bands (blue, green, red and near infrared) have a spatial resolution at 4 m. QuickBird offers highly accurate and even higher resolution imagery with panchromatic imagery at 60–70 cm resolution and multispectral imagery at 2.4 and 2.8 m resolutions. It is the only spacecraft sensor able to offer submeter resolution imagery so far. QuickBird images are usually used to study special topics in relatively small areas (or at a local scale) since it is impractical to apply QuickBird imagery for applications in large area due to its high cost and rigid technical parameters (Xie et al., 2008). Rapid eye BlackBridge's satellites are a constellation of 5 identical Earth Observation satellites which permit to acquire images of large-areas with a frequent revisit intervals, an high spatial resolution and multi-spectral capabilities. RapidEye’s Multi-Spectral Imager (MSI) acquires image data in five different spectral bands (RGB, Red Edge Band and NIR Band) each one with a geometric pixel resolution (or ground sampling distance, GSD) of 6.5m (at nadir). Final images are returned with a 5m pixel resolution. In the last years the use of airborne digital sensors are making it possible to produce digital multispectral orthophoto which can be used both as base-maps and informative layer to perform classification studies. In Italy in recent years several projects provide a full image coverage of the territory both at national and regional scale; the most common collections are AGEA and TerraItaly images.

Chapter 3

Classification methods

Classification of multispectral image data is used to assign unique levels to groups of pixels with homogeneous characteristics, with the aim of discriminating multiple objects from each other within the image.

Classification will be executed on the basis of spectral values or spectrally defined features, such as density, texture etc. in the feature space. It can be said that classification divides the feature space into several classes based on a decision rule.

General image classification procedures include (Gong and Howarth, 1990):

i. Design image classification scheme: usually involves the collection of preliminary information about classes to be mapped such as urban, agriculture, forest areas.
ii. Preprocessing of the image, including radiometric, atmospheric, geometric and topographic corrections, image enhancement, and initial image clustering.
iii. Select representative areas on the image and analyze the initial clustering results or generate training signatures.
iv. Image classification
v. Post-processing: complete geometric correction & filtering and classification decorating.
vi. Accuracy assessment: compare classification results with field studies.

There is no single, universally accepted methodology for automatic analysis of multispectral image data and there are a number of different procedure and algorithms available.

Classification methods can be grouped in many ways (pixel based *vs.* object based, supervised *vs.* unsupervised, deterministic approach *vs.* statistical learning etc); in this chapter initially some basic distinctions are proposed basing on three criteria:

- Which kind of pixel information is used - Pixel based *vs.* object based classification (Sections 3.1)
- Whether training samples are used or not - Supervised *vs.* unsupervised classification (Sections 3.2)
- Whether classifiers models are known a priori or generated from data - Deterministic approach *vs* statistical learning (Sections 3.3)

then sections 3.4 to 3.7 examine some of the most used classification methods and their algorithms (Clustering algorithms (3.4), Minimum Distance (3.5) and Maximum Likelihood Classifiers (3.6), and dimensionality reduction methods (3.7)).

The most used dimensionality reduction methods are the Principal Component Analysis (PCA) and the Partial Least Square (PLS) techniques. The first it is introduced on section 3.7.1 while, as the classification method proposed in this thesis is a PLS-based classifier, the seconds will be described in details in the following chapter.

3.1 Pixel based *vs.* object based classification

The pixel based/object based distinction concerns the unit of classification, which is the single pixel in the first case and an object of many pixels grouped together in the latter.

Pixel-based classification uses multi-spectral classification techniques that assign a pixel to a class by considering its similarities with the class or with other classes (Casals-Carrasco et al., 2000). It identifies the class of each pixel in the imagery by comparing the n-dimensional data vector of each pixel with the prototype vector of each class. The data vectors typically consist of a sequence of pixel-level values from multi-spectral channels (Shackelford and Davis, 2003). Although conventional pixel-based classification is widely used to extract thematic information from images, limitations still exist and several studies and approach try to improve classification accuracy.

Object based classification uses textural and contextual information as well as the spectral information in order to recognize and extract homogenous groups of pixels. The idea to classify objects stems from the fact that most image data exhibit characteristic texture features which are neglected in conventional classifications (Blaschke and Strobl, 2001). It involves two main steps: first a image segmentation is performed and the image is divided into homogeneous, continuous, and contiguous regions. These object have geometric attributes, such as shape and length, and topological entities, such as, for example, adjacency and ‘found within’ (Baatz et al., 2004) which are used in the classification process (Whiteside and Ahmad, 2005).

3.2 Supervised *vs.* unsupervised classification

The unsupervised classification is based on the analysis of an image data without the user providing ground sample representing field observation. Resulting classified maps then require

knowledge of the scene area in order to determine what each class (i.e. cluster) may represent in the real world.

Supervised classification requires a previously selection of training samples or training areas (Regions of Interest - ROI) in the image, that are representative of specific classes and can be used as references for the classification of all other areas in the image (Jensen, 2005). Training points are selected basing on the knowledge of the user and should reflect all the classes he want to describe in the area. The user can also set a threshold which specify how similar other pixels must be to group them together. This threshold is often set based on the spectral characteristics of the image to be classified.

Many analysts use a combination of supervised and unsupervised classification processes to develop final output analysis and classified maps (Karem et al., 2012; Markowska-Kaczmar and Switek, 2009; Enderle and Weih, 2005).

3.3 Deterministic approach *vs* statistical learning

In the deterministic approach the value, or DN, measured by the sensor is explained on the basis of mathematical models of the interaction between the electromagnetic radiation and the surface (Rasmussen and Olesen, 1988). These models are then used in algorithms to classify the data. A large amount of preprocessing of the multi and hyperspectral data is required by the user to transform the data and to remove noise. This implies that sensor calibration effects and atmospheric influence must be adequately accounted for, in order to translate the raw sensor's response into values of the relevant physical parameter describing the reflectance properties of the surface. Further, most deterministic model algorithms assume that the spectral signature gathered at each pixel is a linear superposition of the composite material signatures present at the source. While this assumption simplifies the model, it nevertheless ignores the inherently nonlinear characteristics of hyperspectral data gathered in the field (Wang, 2008).

Example of these approach is the traditional Spectral Angle Mapper (SAM) algorithm available in ENVI (Excelisvis, 2013) where classification is achieved by matching samples data to a reference spectrum presented in a spectral library.

Statistical learning approach aims to utilize the data to build the classifier model, thus no expert knowledge is required during classification. Unlike deterministic methods, statistical learning approach do not need to know the model that describe the data but they learn a model basing on the dataset itself. The resulted classifier is dependent on the data and specific to the sensor and/or the

situation. Theoretically, statistical learning methods can therefore provide classification algorithms which better match data to classes.

3.4 Clustering algorithms

A variety of statistical clustering algorithms are used to group data basing upon their inherent variability (Lillesand et al., 2008); three of major methods generally used are:

K-Means: initially the number of clusters is specified by the analyst and an arbitrary set of cluster centers is generated in the multidimensional measurement space. Then cluster centers are iteratively repositioned until optimal spectral separability is achieved based on distance-to-mean.

Fuzzy C-Means (Dunn, 1974; Bezdek, 1981): the method is similar to K-Means but also incorporates fuzzy logic in later processing.

The Iterative Self-Organizing Data Analysis Technique (ISODATA) (Tou et al., 1974): the method is an iterative process where samples are classified using a minimum spectral distance formula. During each iteration all samples are assigned to existing cluster centers and new means are recalculated for every class, then a new classification is performed relative to the new mean locations. This cycle repeats until the number of samples in each class changes by less than an user-specified convergence limit or when a maximum number of iterations is reached. At the end of the process, the user assigns the class or classes to a desired feature.

3.5 Minimum Distance Classifiers

These classifier methods assign each samples to classes which minimize the distance between them and the class in the multi-feature space.

The distance is defined as an index of similarity so that the minimum distance represents the maximum similarity. The following distances are often used in this procedure:

Euclidean distance (3.1) : is used in cases where the variances of the population classes are different to each other.

$$d_k{}^2 = (X - \mu_k)^t \, (X - \mu_k) \qquad (3.1)$$

Normalized Euclidean distance (3.2): the classical Euclidean Distance is normalized to make it independent from size and physical dimensions of the samples.

$$d_k{}^2 = (X - \mu_k)^t \, \sigma_k{}^{-1} \, (X - \mu_k) \qquad (3.2)$$

Mahalanobis Distance (3.3): it is based on the correlation between variables or the variance-covariance matrix. It differs from the Euclidean distance in that it takes into account the correlation of the data set and does not depend on the scale of measurement.

$$d_k{}^2 = (X - \mu_k)^t \, \beta_k{}^{-1} \, (X - \mu_k) \qquad (3.3)$$

Where:

X: vector of image data (n bands)

μ_k: mean of the k-th class

$\sigma_k{}^{-1}$: variance matrix

$\beta_k{}^{-1}$: variance-covariance matrix

3.6 Maximum Likelihood Classifiers

Maximum Likelihood algorithm (ML) is a supervised statistical classification technique that allocates each pixel of an image to the class with which it has the highest likelihood or '*a posteriori*' probability of membership. It is based on a normalized (Gaussian) estimate of the probability density function of each class (Pedroni, 2003).

The maximum likelihood classifier quantitatively evaluates both the variance and covariance of the spectral response patterns of the category when classifying an unknown pixel.

With the assumption that the distribution of a class sample is normal, a class can be characterized by the mean vector and the covariance matrix. Given these two characteristics for each cell value, the statistical probability is computed for each class to determine the membership of the cells to the class.

Finally, the pixel would be assigned to the one with highest probability value or be labeled "unknown" if the probability values are all below a threshold set by the analyst (Lillesand, 2008).

3.7 Dimensionality reduction methods

For the purpose of this thesis, the spectral information at each pixel in the multispectral image is treated as a point in a multi dimensional space. The main purpose of dimensionality reduction techniques is to reduce the number of variables and dimensions with the aim to extract new information (feature extraction), improve interpretability of the data, enhance class separability or remove a certain amount of noise.

In this way classification can be performed on data in the low dimensional space, thus reducing errors in these procedures. Consequently, the application of dimensionality reduction techniques to hyperspectral image classification has received considerable attention (Du and Chang, 2004; Qian, 2004).

In general, dimensionality reduction algorithms are subdivided into linear and non-linear dimensionality reduction. Linear methods assume that data come from some linear subspace, whereas non-linear methods allow presence of non-linear manifolds.

The most used dimensionality reduction methods are the Principal Component Analysis (PCA) and the Partial Least Square Regression method (PLS), which can also be used to classify qualitative classes in discriminant analysis mode (PLS-DA).

The following section describes PCA algorithm while, as the classification method proposed in this thesis is based on PLS-DA classifier it will be described in details in the following chapter.

3.7.1 Principal Component Analysis (PCA)

One of the most widely used methods for dimensionality reduction is Principal Component Analysis (PCA) (Duda et al., 2001). PCA finds a low-dimensional embedding of the data points that best preserves their variance as measured in the high-dimensional input space. In PCA, the eigenvectors and eigenvalues of the data covariance matrix are calculated. Eigenvalues are ordered in decreasing order, according to the respective portion of variance explained. The dimension of the reduced space is determined by the first *v* eigenvectors which corresponding eigenvalues, summed together, keep the level of explained variance above a threshold defined by the user.

Let $Z = \{z_1, z_2, \dots z_N\}$ be a set of *V*-dimensional data samples, so that $Z \in \boldsymbol{R}^{D \times N}$ and $z_i \in \boldsymbol{R}^{D}$. The sample covariance matrix is defined by

$$\Sigma = \frac{1}{N}\sum_{i=1}^{N}(z_i - m)(z_i - m)^T \quad (3.4)$$

where $m = \frac{1}{N}\sum_{i=1}^{N} z_i$ is the sample mean vector. The eigenvectors e_i of the covariance matrix are given by:

$$\sum e_i = \lambda_i e_i \quad (3.5)$$

The low *v*-dimensional projection of the original data is $X = \{x_1, x_2, \ldots x_N\}$ where each x_1 is given by:

$$x_i = A^T(z_i - m) \quad (3.6)$$

where A is a (V x *v*) matrix whose columns are the *v* eigenvectors with the largest *v* eigenvalues λ_i sorted in decrising order.

Chapter 4

Partial Least Square Regression (PLS) and Partial Least Square Discriminant Analysis (PLS-DA) methods

4.1 Theory

Partial Least Square Regression method (PLS) focuses on maximizing the variance of the dependent variables explained by the independent ones instead of reproducing the empirical covariance matrix. Compared to other reduction methods such as PCA, the PLS takes into account both inputs (X-block) and outputs (Y-block) to build its subspace, leading to better performances.

PLS operates by forming so called *latent variables* as linear combinations of the predictor variables in a supervised manner, and then operating a regression of the response on these latent variables (Chung et al., 2010). PLS model consists of a structural part, which reflects the relationships between the latent variables, and a measurement component, which shows how the latent variables and their indicators are related; but it also has a third component, the weight relations, which are used to estimate case values for the latent variables (Chin and Newsted, 1999).

There are several ways to calculate PLS model parameters (Wise et al., 2006). Perhaps the most intuitive method is known as (Non-Iterative PArtial Least Squares - NIPALS) (Wold, 1975). NIPALS calculates scores $\boldsymbol{T}$ (the projection of the observables in the eigenvector space) and loadings $\boldsymbol{P}$ (the projection of the eigenvector in the original observable space), and an additional set of vectors known as weights, $\boldsymbol{W}$ (with the same dimensionality as the loadings $\boldsymbol{P}$). The addition of weights in PLS is required to maintain orthogonal scores.

NIPALS algorithm for PLS also work when there is more than one predicted variable ($\boldsymbol{Y}$). In such cases of multiple Y-variables, scores $\boldsymbol{U}$ and loadings $\boldsymbol{Q}$ matrices are also calculated for the Y-block. A vector of “inner-relationship” coefficients, $\boldsymbol{b}$, which relate the X- and Y-block scores, must also be calculated; then the scores, weights, loadings and inner-coefficients are calculated sequentially.

The PLS decomposition is started by selecting one column of $\boldsymbol{Y}$, y_j, as the starting estimate for u_1. Usually the column of Y with the greatest variance is chosen. Of course, in the case of univariate y, $u_1 = y$.

Starting in the X data block:

$$w_1 = \frac{X^T \mathrm{u}_1}{\|X^T \mathrm{u}_1\|} \quad (4.1)$$

$$t_1 = X w_1 \quad (4.2)$$

In the *y* data block:

$$q_1 = \frac{Y^T \mathrm{t}_1}{\|Y^T \mathrm{t}_1\|} \quad (4.3)$$

$$\mathrm{u}_1 = Y q_1 \quad (4.4)$$

Convergence is then checked by comparing t_1 in equation 4.2 with the value from the previous iteration. If they are equal within rounding error, the algorithm proceeds to Equation 4.5 below. If not, the algorithm returns to Equation 4.1, using the u_1 obtained previously from Equation 4.4. If the Y-block is univariate, Equations 4.3 and 4.4 can be omitted, q_1 can be set to 1, and no iteration is required.

The X data block loadings are then calculated, and the scores and weights are rescaled accordingly:

$$\mathrm{p}_1 = \frac{X^T \mathrm{t}_1}{\|\mathrm{t}_1{}^{\mathrm{T}} \mathrm{t}_1\|} \quad (4.5)$$

$$\mathrm{p}_{1new} = \frac{\mathrm{p}_{1old}}{\|\mathrm{p}_{1old}\|} \quad (4.6)$$

$$\mathrm{t}_{1new} = \mathrm{t}_{1old} \|\mathrm{p}_{1old}\| \quad (4.7)$$

$$w_{1new} = w_{1old} \| p_{1old} \| \quad (4.8)$$

The regression coefficient (b) for the inner relation is then calculated as:

$$b_1 = \frac{u_1{}^T t_1}{\| t_1{}^T t_1 \|} \quad (4.9)$$

After the scores and loadings have been calculated for the first factor (which is commonly called a Latent Variable in PLS), the X- and Y-block residuals are calculated as follows:

$$E_1 = X - t_1 p_1{}^T \quad (4.10)$$

$$F_1 = Y - b_1 t_1 q_1{}^T \quad (4.11)$$

The entire procedure is now repeated for the next latent variable, starting with equation 4.1. However, X and Y in Equations 4.1 through 4.4 are replaced with their residuals E1 and F1, respectively, and all subscripts are incremented by one.

PLS methods have been extensively used in remote-sensing data processing since they are well suited to dealing with collinearity problems, such as those encountered when analysing multidimensional remote-sensing data (e.g. hyperspectral images) (Wolter et al., 2008; Barker and Rayens, 2003) justified the use of PLS for high-dimensional classification problems by establishing its connection with Fisher's linear discriminant analysis (LDA). The resulting methodology, called partial least square discriminant analysis (PLSDA), (Sjöström et al., 1986) searches for the PLS components allowing the best separation of the classes. It basically consists of a PLS Regression where the dependent variables are the indicators of the class membership which are transformed into a dummy matrix, and this dummy matrix provides the response variables for PLS (Afendi et al., 2013).

Let $\boldsymbol{X}$ be a centering matrix ($I \times J$) of predictive variables of the calibration set and $\boldsymbol{g}$ be a vector of I integer values coding the qualitative groups, such as $\boldsymbol{g_i}$ gives the group number associated with the i-th observation and $\boldsymbol{G}$ denotes the total number of qualitative groups.

In a binary classification problem ($\boldsymbol{G} = 2$) the $\boldsymbol{Y}$ variable can be easily defined by setting its values to 1 if the objects are in the class and 0 if not. Then, the model will give a calculated $\boldsymbol{Y}$, in the same way as for a regression approach; the calculated $\boldsymbol{Y}$ will not have either 1 or 0 values perfectly, so a threshold (equal to 0.5, for example) can be defined to decide if an object is assigned to the class (calculated $\boldsymbol{Y}$ greater than 0.5) or not (calculated $\boldsymbol{Y}$ lower than 0.5).

When dealing with multiclass problems, the same approach cannot be used (Ballabio and Todeschini, 2009) and it is necessary to unfold the class vector $\boldsymbol{g}$ by building an indicator matrix $\boldsymbol{Y}$, dimensioned ($I \times G$) such as $\boldsymbol{y_{ig}}$ is equal to 1 if the observation of index i is belonging to group $\boldsymbol{g}$, and 0 otherwise (Sjöström et al., 1986).

Then, the PLS regression model can be applied on $\boldsymbol{X}$ and $\boldsymbol{Y}$ by varying the PLS dimensions. For each object, PLS-DA will return the prediction as a vector of size $\boldsymbol{G}$, with values in-between 0 and 1: a $\boldsymbol{g}$-th value closer to zero indicates that the object does not belong to the $\boldsymbol{g}$-th class, while a value closer to one the opposite. Since predicted vectors will not have the form (0, 0, … , 1, … ,0) but real values in the range between 0 and 1, a classification rule must be applied; the object can be assigned to the class with the maximum value in the predicted vector or, alternatively, a threshold between zero and one can be determined for each class.

4.2 PLS and PLSDA applications

PLS and PLS-DA are both used to relate a set of explanatory variables to a set of observed ones. While PLS model quantitative responses on the base of a set of explanatory variables, PLS-DA is a quantitative method for the modelling of qualitative responses. In other words, it aims to find mathematical relationships between a set of descriptive variables (e.g. ecological features) and a qualitative variable (i.e. the membership to a defined category).

There are many studies in different research fields which utilize PLS-DA as classifier of multivariate datasets (see for example Menesatti et al., 2008; Corbane et al., 2013; Dale et al., 2013; Capoccioni et al., 2011). The common approach in PLS-based techniques involves the following steps (Antonucci et al., 2012): (1) building of a training sample dataset (DS) with the attributes to be used as X-block variables and the corresponding reference or response variable (Y-block); (2) separation of the DS into two subsets, one (CS) for the calibration and one (VS) for the external validation test; (3) application of pre-processing algorithm to the X-block and, where possible, to the Y-block; (4) application of PLS/PLS-DA both to calibration and validation subsets; (5) calculation of efficiency parameter of prediction/classification.

According to this strategy, one of the main aspect to be considered in the procedure regards the need to make the prediction/classification model the more robust as possible. Robustness means that the prediction/classification error of the model remains within acceptable limits when different instrument, location, or physical sample conditions are tested (Swierenga et al., 1998). Actually PLS-based techniques, even if adjusted for loss of degrees of freedom due to the number of predictors in the model, can give a misleading, over optimistic view of accuracy of prediction/classification ability when the model is applied to the external dataset different from the one the model was built with.

In order to assure a better classification ability it is therefore advisable to "validate" the model by testing it on data not used to fit the model itself. Several approaches to validation are available: the most advisable one is to use an external and totally independent dataset, while the most rapid is the cross-validation, where a series of models is fit each time deleting a different observation subset from the calibration set, and using the obtained model to classify it. A third possible validation strategy is the split-sample validation where the model is fit to some portion of the data (for example the second half), and accuracy is measured on the classifications for the other part of the data. In this case, one of the main aspects to be considered is how to divide the original dataset into the calibration and validation subsets (partitioning). The above introduced concepts will be described in detail in the following sections.

4.3 Validation strategies

4.3.1 External validation

This method is validation in the purest sense, and requires the use of a second independent dataset to test the model performance. Typically, the primary dataset is initially collected and a predictive/classifier model is derived from it; then further relevant data are assembled to form the secondary dataset and the performance of the model is evaluated. In order to assure independence, the validation data must in no way enter into the calibration of the first model. A practical difficulty of applying external validation is the lack of sufficient high quality data to build both the calibration and the validation datasets; users generally prefer to use the biggest possible calibration dataset in the hope of including the most representative high-frequency and low-frequency components of variability.

In the classification problems the purpose is usually to assign all the population members to a set of classes basing on the information available from a group of them, so that the external

validation becomes the final test to evaluate the classifier. However in the image classification it is usually the case, and so is in this thesis, that the external dataset is represented by the entire image to be classified, and the validation accuracy is performed by using a reference map of the area.

4.3.2 Cross validation

In this method a series of iterative data extractions is followed to build a series of estimates that can be used to validate a model calibrated on the remaining dataset. For a given data set, the N available objects are divided into *k* removing groups following a predetermined scheme (e.g. contiguous blocks, venetian blind etc.). The model is then computed *k* times, using each time a single removing group as validation subset and the remaining objects as calibration subset. In this way each computation permits to test a model with objects that were not used to build it. A typical cross-validation procedure usually involves more than one calibration experiment, each of which involves the selection of different subsets of samples for model building and model testing. There are several different cross-validation methods, and these vary with respect to how the different sample subsets are selected for these sub-validation experiments. Among these (Eigenvector, 2014):

a) *Venetian Blinds*: each i^{th} test set is determined by selecting every i^{th} object in the dataset starting at objects numbered 1 through *k*

b) *Contiguous Blocks*: each test sets is determined by selecting contiguous blocks of objects in the data set, starting at block number 1 through *k*

c) *Random Subsets*: *k* different test subsets are determined through random selection of N/*k* objects in the data set (N = number of total objects), such that no single object is in more than one test set. This procedure is repeated r times

d) *Leave-One-Out*: each single object in the data set is used as a single removing group (test subset).

4.3.3 Split-sample validation

This method consists in splitting the dataset into two parts and use one to calibrate the model and one to validate it. In this way data splitting provides a data set to measure the in-use prediction accuracy of the model and simulates the complete or partial replication of a study.

Several works have addressed the problem of reproducing the composition variability of the original dataset in the validation subset by selecting a representative subset from a large pool of samples (Snee, 1977; Wu et al., 1996; Tominaga, 1998; Sales et al., 2000; Daszykowski et al., 2002). In this context, two main aspects have to be considered: the ratio of the calibration/validation subsets and the partition method to be used to extract the validation samples from the original dataset. Concerning the first issue several approaches has been proposed, but so far, none of these methods has been clearly recognized as a reference. About the second aspect several partition methods has been described in literature, among these the most used are:

- *Random selection (RD)*: it is a popular technique because of its simplicity and also because a large group of data randomly extracted from larger datasets follows the statistical distribution of the entire set. However, RD does not guarantee the representativity of the set, nor does it prevent extrapolation problems (Rajer-Kanduč et al., 2003). In fact, RD does not ensure that the samples on the boundaries of the set are included in the calibration.

- *Kennard–Stone (KS)* algorithm (Kennard and Stone, 1969): it is aimed at covering the multidimensional space in a uniform manner by maximizing the Euclidean distances between the instrumental response vectors (x) of the selected samples (Wu et al., 1996; Bouveresse et al., 1996). In order to ensure a uniform distribution of such a subset along the *x* data space, KS follows a stepwise procedure in which new selections are taken in regions of the space far from the samples already selected. For this purpose, the algorithm employs the Euclidean distances d_x (p, q) between the x-vectors of each pair (*p, q*) of samples calculated as

$$d_x(p,q) = \sqrt{\sum_{j=1}^{J}\left[x_p(j) - x_q(j)\right]^2} \qquad p,q \in [1, N] \qquad (4.12)$$

For multilayer image data $x_p(j)$ and $x_q(j)$ are the variable responses for samples p and q at the j-th value along J-dimension. J denotes the number of layers in the multilayer matrix. The selection starts by taking the pair (p_1, p_2) of samples for which the distance d_x (p_1, p_2) is the largest. At each subsequent iteration, the algorithm selects the sample that exhibits the largest minimum distance with respect to any sample already selected. Such a procedure is repeated until the number of samples specified by the analyst is achieved.

- *SPXY algorithm* (Galvão et al., 2005): it takes into account the variability in both X and Y spaces. It increases the distance (4.12) with an Euclidean distance d_y calculated in the dependent variable (y) space for the parameter under consideration:

$$d_y(p,q) = \sqrt{\sum (y_p - y_q)^2} \qquad p,q \in [1, N] \qquad (4.13)$$

In order to assign equal importance to the distribution of the samples in the X and Y spaces, distances $d_x(p,q)$ and $d_y(p,q)$ are divided by their maximum values in the data set. In this manner, a normalized xy distance is calculated as:

$$d_{xy}(p,q) = \frac{d_x(p,q)}{\max_{p,q \in [1,N]} d_x(p,q)} + \frac{d_y(p,q)}{\max_{p,q \in [1,N]} d_y(p,q)} \qquad p,q \in [1, N] \quad (4.14)$$

A stepwise selection procedure similar to the KS algorithm can then be applied with $d_{xy}(p,q)$ instead of $d_x(p,q)$ alone.

Chapter 5

Habitat classification and mapping

A fundamental prerequisite for a successful classification is the choice of a suitable classification system and the use of predefined and fixed set of legend units. Such an approach requires thorough preparation, but it is of the utmost importance in order to achieve an acceptable level of comparability of the classification both in space and time.

Following the approach developed by Salafsky et al. (2003) a classification system should be:

a. Hierarchical - Creates a logical way of grouping classes;
b. Comprehensive - Covers all possible objects on the scene by a class label;
c. Consistent - All entries at a given level of the taxonomy are of the same type;
d. Expandable - New classes can be added without changing the full hierarchy;
e. Exclusive - Any given ‘‘object’’ can only be placed in one position within the hierarchy;
f. Geographically invariant - The labeling of a same object is invariant across different locations

For mapping purposes, systems meeting all these criteria are relevant to ensure a full coverage of the landscape and avoid uncertainty in describing objects (Tomaselli et al., 2013).

In Europe habitat mapping has become increasingly important especially since the EU Habitats Directive (1992) was issued. The need for habitat identification has several driving forces: habitat and species conservation, protection legislation, inventories, biodiversity monitoring and reporting, description of a species' habitat requirements etc.

Habitat types may be defined on different spatial and typological scales; their classification attempts to define dividing lines between them, but these divisions are often much less clear-cut and more subjective than those between species. For this reason there is no a clearly agreed ‘taxonomy' and many different systems have been developed, often independently of each other and for different purposes.

In Europe, the first comprehensive habitats classifications were the Corine biotopes (Devillers et al., 1991) and, afterwards the Palaearctic classification (Devillers and Devillers - Terschuren, 1996; Hill et al., 2004), which extended the geographical coverage. Both are now partially superseded by the EEA’s EUNIS system, which was proposed to became the common reference for habitats identification by the EU INSPIRE Directive (1997).

The large number of large-area habitat mapping projects both at continental, national scale are based on one of these classification schemes; in particular, in this thesis we utilize a revised version of the Corine Biotopes scheme, adopted by the Italian official project "*Carta della Natura*".

The following sections describe initially the Corine Biotopes classification scheme (section 5.1) and then the Italian project "*Carta della Natura*" (section 5.2), whose maps are also utilized as reference to assess the accuracy of the proposed classification method.

5.1 The Corine Biotopes

The Corine (COoRdination of INformation on the Environment) Biotopes classification was published in 1991 as part of the Corine project which aimed to identify and describe the habitats of major importance for the conservation within the European Community (at that time comprising only 12 Member States). It is a hierarchical classification system intended to cover all habitat types but with a focus on natural and semi-natural habitats and a limited coverage of marine habitat types. Although it is clearly based on phytosociological classifications, it also includes other factors like geomorphology, climate and soil, and covers several habitat types giving them a geomorphological connotation (e.g. glaciers and lava tubes). The original version of the Annex I of the EU Habitats Directive, as published in 1992, is a selection of the Corine biotopes classification (Evans, 2010).

The highest hierarchical level, defining the broadest division, includes seven categories: coastal, wetland, grassland and scrub, woodland, marsh and bog, rocky and agricultural habitats. The second digit defines the most important subdivisions of each of these categories. The first two digits together denote the 'generic habitat type', of which there are 44 in all. A decimal point separates these two digits from up to five further alphanumeric digits which are used to define individual habitat types or phytosociological associations with increasing precision. Any code which has at least two decimal digits is referred to as a 'detailed habitat code'.

Figure 5.1 shows an example of the Corine hierarchical coding system.

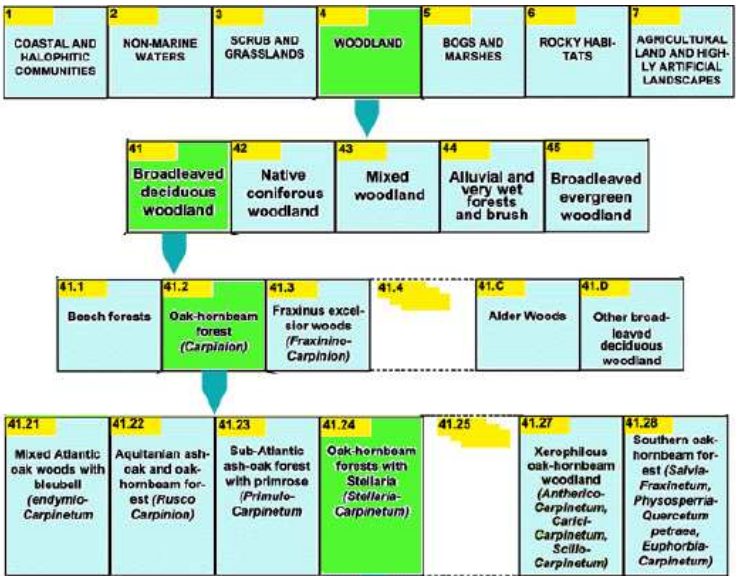


Figure 5.4 - Illustration of the CORINE habitat coding system (Commission of the European Communities, 1991)

The advantages of this system are that all habitat types and associations which are known in the Community can be included, and that the hierarchical nature makes it possible to retrieve information at the required level of detail and to add new categories to the coding system without disturbing any of the existing codes. The system can easily be expanded to accommodate highly detailed sub-divisions of the more important habitats, for example for use in national or regional inventories, while retaining upward compatibility with the Community-wide system.

In 1996 the Corine biotopes classification system was extended in the Palaearctic Habitats Classification to cover all of Europe. Although the new system extended the geographical coverage, treatment of marine habitats remained poor and no criteria to distinguish these habitat types were given.

5.2 The "*Carta della Natura*" project

The Italian *Carta della Natura* system is one of the most important projects of habitat mapping at a national level (EEA, 2014). The project began with the Italian Law no. 394 /1991 and aimed to identify the status of the natural environment in Italy and to assess the quality and fragility of Italian habitats.

Now the project is in charge of the Italian National Institute for Environmental Protection and Research (ISPRA) which coordinates the activities and built and maintains an informative system structured at different scale of analysis with different map outputs: a map of the Italian Landscape Units at a scale of 1:250 000 completed in 2001 (ISPRA, 2003) and habitats maps at regional (1:50 000) and local (1:10 000) scales.

In particular, the regional and local scales are organized in two main products: the map of habitats, which identifies and classifies homogeneous patches with respect to their characteristics, and the assessment system which, with the use of indicators and indexes, assigns to each of those patches values representing its Value and risk of degradation in terms of sensitivity and fragility (ISPRA, 2009a).

The final indexes used to represent the state of the habitat are:

- *Ecological value* representing the quality of a biotope from an environmental point of view, defined by the law as "natural value"..
- *Ecological Sensitivity* representing the intrinsic predisposition of the biotope to be degraded by an external disturbance
- *Anthropogenic Pressure* representing the external disturbance (only limited to human activities)
- *Environmental Fragility* which measures the state of vulnerability of the environmental unit from the natural and environmental point of view, as a result of the combination of the previous two indexes.

The main application of the *Carta della Natura* maps has been thought as the main “knowledge tool” in order to set out the main guidelines to territorial planning at a national or regional scale in the country. The maps and the associated databases can be used especially to identify areas with high environmental value (to be put under different possible levels of protection) with more “natural” boundaries, because often protected areas follow administrative boundaries and cut ecosystem units which should be entirely protected.

In addition to this first applications, in this years, the *Carta della Natura* data have been widely used in the SEA (Strategic environmental assessment) and EIA (Environmental Impact Assessment) procedures or in the ecological network and biodiversity studies and, more generally, they can be (and have been) used for any kind of environmental study or planning procedure relying on a deep knowledge of the territory, and in particular on its natural systems.

5.2.1 Habitat legend

The legend is defined according to the Corine Biotope classification following the Palaearctic revision. Starting from Corine Biotopes categories, a subset of them was selected at different levels of the hierarchy, in order to take into account both the local situations and the need to have a general standard for the whole national territory at the 1:50 000 scale.

At the present time, the *Carta della Natura* legend contains 231 types of habitat (ISPRA, 2009b), 19 of which are not mapped in Italy yet.

Moreover, a legend database has been organized in order to highlight the correspondences between Corine Biotopes legend and other European classification systems (EUNIS, Natura 2000 etc).

5.2.2 Habitat mapping activity

Habitat detection, identification and mapping are carried out by integrating information from satellite images, field surveys and other spatial data (e.g. land use or forest type maps). The experimental phase of habitat mapping only used satellite images (Landsat 7 ETM +) processed through a supervised classification. However, in the last years the remote sensing phase showed several limitations and became only a preliminary activity to the visual interpretation of aerial photography which, associated with field campaigns, took over in importance in mapping activities.

The new strategy allows to obtain a more detailed and accurate maps, but it is more time-consuming and more prone to subject interpretations.

This thesis aims to develop a novel semi-automated classification method which could be useful in the first steps of mapping, reducing the amount of photo-interpretation.

Chapter 6

Use of PLS-DA classifier to map habitat in Italy

6.1 Introduction

This chapter describes the main steps of the proposed classification method. It combines multivariate statistical techniques (in particular PLS-DA classification algorithm) with GIS and remote sensing procedures in order to build a classification procedure to be used to map habitats distribution.

In summary, the main contributions of this methods are:

- improvement of the amount of information available for the classification by using ancillary data with the classical multispectral image dataset
- optimizing the classification ability using a stepwise “two-level” approach
- proposal of a new recursive algorithm in order to improve the classification ability of the PLS-DA on external datasets
- use of GIS and remote sensing procedures to overlap the classified images and to refine the final map.

All of the tools used in the proposed procedure were implemented in Matlab (for the statistical analysis) and ESRI ArcGIS (for the GIS and remote sensing analysis) softwares.

In order to verify the suitability of the classification under various conditions, the method was tested on three study areas (Monte Vulture volcanic complex, Apulia lagoons and Campo Pericoli basin) with different habitat composition and on different cartographic scales.

Finally, the classification ability was determined and compared with the one obtained using a commercial software (ESRI ArcGIS).

The chapter is organized as follows: section 6.2 describes the three image datasets used as a basis to classify the test areas. Section 6.3 describes in detail each steps of the method and finally section 6.4 describes the accuracy assessment performed on the classified maps.

6.2 Datasets for classification

Each image dataset used in the classification tests is a multilayer matrix composed by a multi-spectral image (Rapid-eye images/4 bands orthophoto) and by other layers representing the ancillary data to be considered in the classification model. It is a hypercube matrix where the X and

Y dimensions specify the spatial position of a pixel and the Z dimension is composed by the N layers representing multispectral image bands and ancillary data. In the following sections the multispectral images (section 6.2.1) and the ancillary data (section 6.2.2) used to classify the three test areas will be described, while table 6.1 shows a short description.

	Lesina lagoon	***Vulture Mount***	***Campo Pericoli***
Mapping scale	50.000	50.000	10.000
Minimum mapping unit	1 hectare	1 hectare	0.04 hectare
Multispectral image	Rapid-eye SAT (5 bands)	Rapid-eye SAT (5 bands)	Vexcel UltraCam (4 bands)
Ancillary data	altitude	altitude	altitude
	slope	slope	slope
	exposure	exposure	exposure
	solar radiation	solar radiation	solar radiation
	NDVI	NDVI	NDVI
	NDRE	NDRE	

Table 6.1 - Study areas characteristics and dataset used for classification

6.2.1 Multispectral image dataset

6.2.1.1 Rapid-eye images

The Rapid Eye constellation (owned by Blackbridge company) comprises five satellites equipped with identical sensors (multi-spectral push broom imager) and located in the same orbital plane (630 km, sun-synchronous). Its particular configuration allows to acquire images of very large areas (up to 4 million square kilometres *per* day) at 6.5 meter nominal ground resolution (5 meters pixel size when orthorectified) with a short revisit time.

Spectral acquisition comprises five spectral bands: Blue (440 – 510 nm), Green(520 – 590 nm), Red (630 – 685 nm), Red Edge (690 – 730 nm) and NIR (760-850 nm). Images are made

available in TIFF format with 16 bit depth digital number (DN) representing reflectance values. Three levels of product with different post-processing are available:

- level 1B is the basic level with just radiometric and sensor corrections to the data without correction for any geometric distortions inherent in the imaging process. Resulting images are not mapped in a cartographic projection
- in level 3A, radiometric, sensor and geometric corrections are applied to the data. The product accuracy depends on the quality of the ground control and DEMs used. Product is processed as an individual 25 km by 25 km tile.
- in level 3B multiple images are adjusted together with radiometric and geometric corrections to cover larger areas with fewer files

In Italy the local official distributor for Rapid-Eye images is IptSAT company.

6.2.1.2 Orthophoto

In the Campo Pericoli classification test an image with an higher ground resolution was used. Orthophotos were collected in 2009 within a collaboration between the Italian *Agenzia per le Erogazioni in Agricoltura* (AGEA) and *Regione Abruzzo*. Images were acquired using the Vexcel UltraCam with a four band spectral acquisition, (RGB+NIR). The resulting products were made available in TIFF format with 8 bit depth digital number representing reflectance values and with a 0.2 m ground spatial resolution.

In order to use orthophoto images with other ancillary data with different spatial resolution (2 meters), they have been resampled to 2 meters pixel-size.

6.2.2 Ancillary data

In this study, remote sensing information was integrated with other ancillary data which describe the physical conditions that can influence the spatial distribution of habitats present in the study area. Use of supplementary information can be useful to help distinguish between spectrally inseparable vegetation (Yu, et al 2006) or habitat classes and lead to more effective classification. Environmental or topographic factors, such as elevation, slope or soil moisture are widely used as ancillary data (Gould, 2000; Dobos et Al., 2000, Dymond and Johnson, 2002; Gastellu-Etchegorry et al., 1993, McIver and Friedl, 2002) and their contribution can be particular important when working with remote sensing imagery at very high spatial resolution due to the complexity of class description and the limited spectral resolution (few spectral bands) in order to minimize uncertainty

in mapping (Lechner et al. 2012). According this approach several layers are used as additional information sources:

6.2.2.1 Elevation

Elevation data are mapped using a Digital Elevation Model (DEM) raster layer where each pixel represents the mean altitude value of that cell. In each study area a different resolution elevation model was used depending on mapping scale area and data availability:

- Monte Vulture volcanic complex: a 20 meters resolution DEM, made available by the Italian Geoportal was used. In order to overlay these data with Rapid-eye images layer were resampled at 5 meters pixel-size using nearest neighbour algorithm.
- Apulia lagoons: a 8 meters resolution DEM, made available by Regione Puglia was used. Siilar to the previous case also these data were resampled at 5 meters pixel-size.
- Campo Pericoli basin: a 2 meters resolution DEM of the area was produced during the Ph.D. activity starting from Regional Technical Map (CTR) at 1:5000 scale. Both isodistance curves and reference elevation points were used to build the elevation raster by interpolating their values using the ANUDEM program (Hutchinson, 1996, Hutchinson, 2011) available in ArcGIS (ESRI, rel 10.1)

6.2.2.2 Slope

Slope layer represents the inclination of terrain in degrees. It is calculated from the DEM layer using ESRI ArcGIS tool. For each pixel the tool calculates the maximum rate of change in value from that cell to its neighbours. Basically, the maximum change in elevation over the distance between the cell and its eight neighbours identifies the steepest altitude change from or to the cell.

6.2.2.3 Exposure

Exposure layer identifies the geographical (compass) direction that the surface faces at that location. It is built from the DEM layer using ArcGIS Aspect tool, the resulting raster is then reclassified considering the calculated clockwise values in degrees, from 0 (due north) to 360 (again due north).

6.2.2.4 Insolation

Insolation layer represents the yearly amount of incoming solar radiation for a particular location, expressed as (MW h/m2). Insolation map is calculated from the DEM layer, using the *Hemispherical viewshed* algorithm developed by Rich et al. (1994), modified by Rich (Rich and Fu, 2000) and Fu (Fu and Rich, 2002), and available in ArcGIS.

6.2.2.5 Normalized Difference Vegetation Index (NDVI)

NDVI layer was used to express vegetation productivity. NDVI is a common vegetation index used as a proxy to provide information on the amount of live green vegetation in an area. It is calculated considering the strong energy absorption by the chlorophyll in the red portion of the electromagnetic spectrum (RED), and the energy scattered by the internal structure of leaves in the near-infrared (NIR). The relative proportions of red and near-infra red reflections according to the formula (Rouse et al. 1974) define the index:

$$NDVI = \frac{NIR - RED}{NIR + RED} \qquad (6.1)$$

Moreover, on Rapid-Eye images a modification of the NDVI index, called Normalised Difference Red Edge Index (NDRE) was calculated in order to provide additional information to be used in the classification procedures (Recio et al 2011; Schuster et al 2012). NDRE differs from NDVI in using bands along the red edge, instead of the main absorption and reflectance peaks.

$$NDRE = \frac{RedEdge - RED}{RedEdge + RED} \qquad (6.2)$$

6.3 The Classification method

The PLS-DA method used in this study belongs to the family of supervised classification algorithms. In the supervised classification framework, the classifier needs to be trained using a sample of homogeneous areas that can be identified either directly on the image or using thematic

products (e.g. existing maps), derived from field visits, or through a combination of both approaches.

The classification was carried out through five main steps:

- Data collection and definition of the classification levels (section *6.3.1*)
- Training datasets preparation (section *6.3.2*)
- Application of the recursive PLS-DA algorithm and selection of the most robust model for each level of classification (section *6.3.3*)
- Classification of the entire image on each classification level (section *6.3.4*)
- Image final reconstruction and vectorization (section *6.3.5*)

6.3.1 Data collection and definition of the classification levels

For each study area an exhaustive number of ground samples, representing the habitat existing in the zone, were collected using available spatial datasets (ISPRA, 2011a) or with dedicated field campaigns. For each sample its geographical position (WGS84 coordinate systems) and the corresponding habitat attribution according to the Corine Biotopes legend were recorded.

Several works demonstrated that PLS-DA classifier shows a better prediction ability with a limited number of classes (Eriksson, 2006) so, given the large number of habitat types to be identified, it was decided to perform the classification through a stepwise procedure which initially aims to classify the area according to habitats groups (macro-categories) and then to identify the final habitat class by discriminating them within each macro-category.

Following this approach all the samples were labelled with a *level-2* and a *level-1* class using two criteria. The lower level of classification (level-2) is based on the classes of the original training set (obtained from the Corine Biotopes legend) while the upper, more general, level of macro-classification (level-1) was identified by grouping the original classes in super-classes based on habitat ecological features and spectral affinity.

6.3.2 Training datasets preparation

The training set data to be used in PLS-DA classification is composed by an X-block including all the multivariate variables which are involved in the classification and an Y-block with the corresponding class attribution.

The classification method proposed in this thesis involves a stepwise procedure, so for each study area more classification tests were performed, each one using a different training dataset (X and Y blocks) depending on the macro-groups or habitat classes to be identified.

Initially, for each study area a raw dataset was built by extracting from the training image the pixel values corresponding to the position of ground samples (X-block) and by assigning them both labels corresponding to level-1 and level-2 classes (Y-block). In this way to each sample is assigned a vector *X* (x_1; x_2; x_3;… x_N) with the values of the N variables which contribute to the classification and a vector *Y* (y_1; y_2) with a numeric attribution representing respectively the habitat class (level-2) and macro-category (level-1).

In order to build the final training sets to be used in level-1 classification and the level-2 divisions within each macro-area, the raw dataset was then used considering different class attributions (fig 6.1).

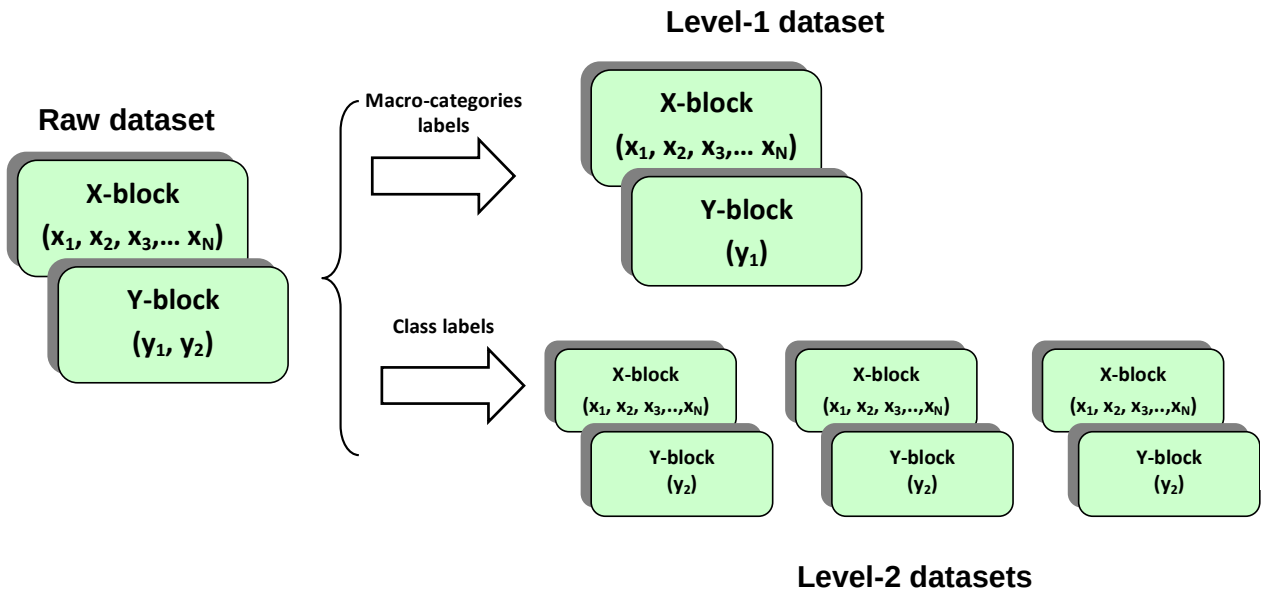


Figure 6.1 – Training datasets and subsets composition

The level-1 dataset was built by taking all sample data (with their X-block variables) and assigning the level-1 categories as Y-block. The level-2 subsets were extracted taking all control points belonging to the same macro-category and using the corresponding level-2 classes as Y-block.

6.3.3 Application of the recursive PLS-DA algorithm and selection of the most robust model for each level of classification

Some works demonstrate the possibility to use PLS and PLS-DA (Baker and Rayens, 2003) to classify habitat and vegetation (Lieckfeld et al., 2006; Corbane et al., 2013; Roelofsen et al., 2014).

Generally the PLS-DA modelling procedure involves a 'selective' approach (see section 4.2) in which the user develops the model in sequential steps, choosing among different optimal strategies in terms of:

- model components (pre-processing and Latent Variables (LV))
- model validation (size and composition of the validation subset)

in order to obtain the optimum predictive performance.

Essentially, this is a "trial and error" approach to find the most appropriate data pre-processing and data partition parameters for optimum classification performance in the chosen multivariate technique. This process relies on the visual assessment of the statistical and graphical output for each of the classifiers developed. Classifiers with poor performance are discarded and the cycle continues until favourable outcomes are reached. The strength of these models is then tested with independent data.

In this way PLS-DA classifiers are not routinely tested to determine their robustness and the final choice about which model could have the better classification ability is left to user's ability and experience. Furthermore, there may be many model results that are acceptable for final use. Overall, there is no standard protocol for developing PLS-DA classification models.

In this Ph.D. activity a new recursive algorithm for the PLS-DA was developed and tested to identify the most robust model which could be developed for the selected area. It permits to test different validation subsets (obtained using different partition methods and partitioning ratios) and multiple parameters (in terms of pre-processing and LV) to be used in the classification modelling procedure without choosing *a priori* any components but stressing all the models in order to find the most robust one to be used on external area.

In the recursive algorithm PLS-DA models were developed using the NIPALS algorithm in MATLAB (rel. R2010b - The Mathworks, Natick, MA, USA) and cross-validated using the Venetian blind approach (Matlab rel. R2010b, PLSToolbox Eigenvector rel. 7.3.1).

A range of partition methods, partition ratios and data pre-processing transformations were used. Two partition systems were employed to split the dataset into calibration and validation data: Random Selection (RD), and Kennard Stone algorithm (KS; Kennard and Stone, 1969). Nine

partition ratios divided the data into calibration and validation samples on the basis of increasing calibration size; 50:50, 55:45, 60:40, 65:35, 70:30, 75:25, 80:20, 85:15, and 90:10. Twenty-four pre-processing transformations (described in table 6.2 at the end of the chapter) were applied to the data.

Each cross-validation was run with all possible latent variables and tested with independent validation data. Finally for each model the accuracy of classification (in terms of percentage of correctly classified samples in the validation subset) was calculated.

Figure 6.2 shows a flow chart describing the recursive algorithm procedure. In order to find the most robust model the following steps were carried out:

Step 1: Independent classifiers were generated for each possible combination of partition method, partition ratio, X block data pre-processing transformation and LV.

Step 2: Classification results obtained in step 1 (% of correct classification in the calibration and validation subsets) with common pre-processing and number of LV's were averaged and results were ranked by percentage of correct classification in the test subset. This determined the optimal parameters for the classifier.

Step 3: a new classifier model was built applying model parameters obtained in step 2 to all the dataset.

Steps 1 to 3 were repeated both for level-1 and level-2 datasets in order to obtain all models to be used in the classification levels.

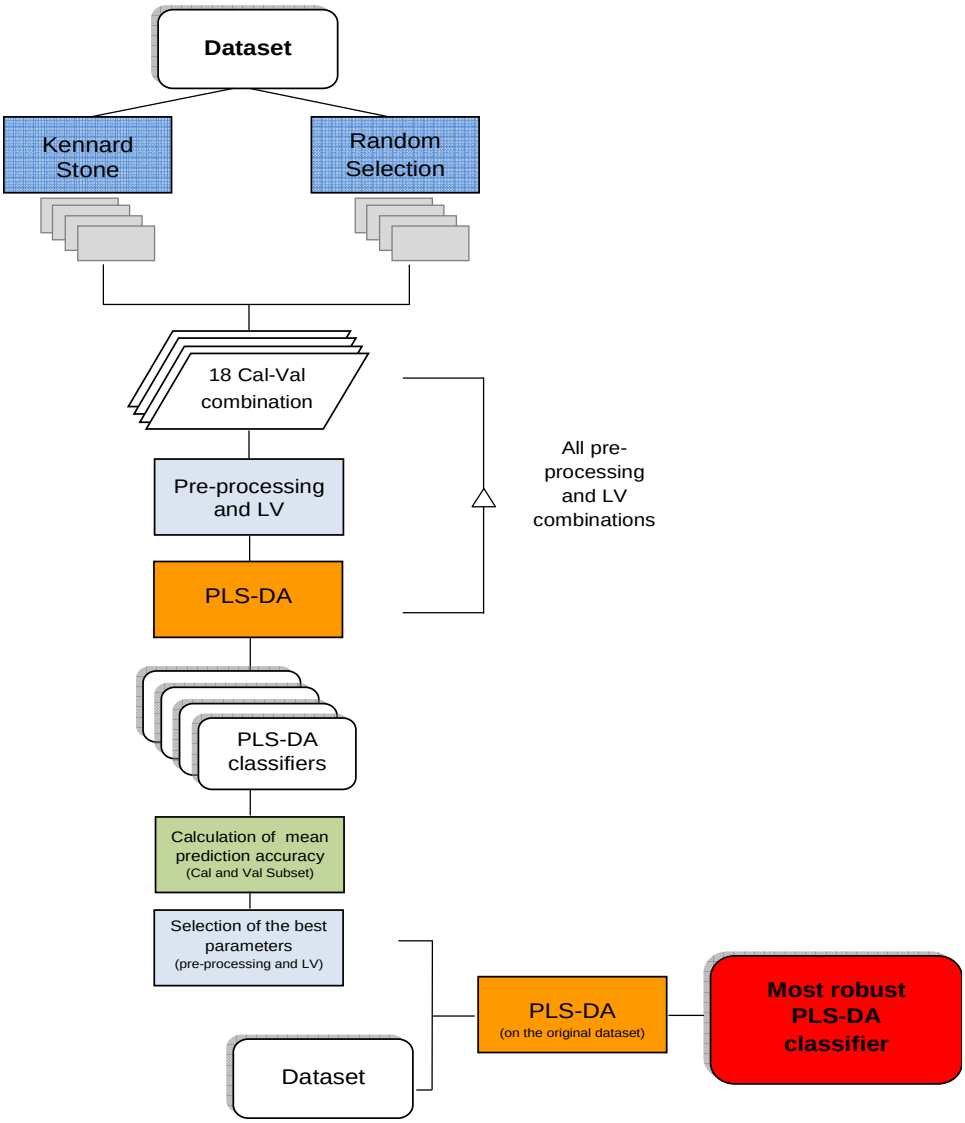


Figure 6.2 - Recursive algorithm procedure

6.3.4 Classification of the entire image on each classification level

The classifier models built in the previous step were used to classify the entire image dataset according to level-1 labels and, within each macro-category, according to level-2 labels.

By applying a model to the entire image a multilayer matrix (with the number of layers equal to the N classes to be represented) was obtained. In the matrix the n^{th} layer represents the probability of that pixel to belong to the n^{th} class.

Each classified image was then composed assigning each pixel to the class with the highest probability and labelling as “not classified” those pixels where no probability layers were greater than a threshold of 0.5.

In this way for each study area we obtained:

- one classified image representing the macrocategories (level-1 labels)
- several classified images, each one representing habitat classes (level-2 labels) belonging to a macro-category

All image extraction were performed using the Multivariate Image Analysis toolbox (MIA toolbox Eigenvector rel. 2.8.5).

6.3.5 Image final reconstruction and vectorization

The classified map was then composed by the overlay of all the previous layers.

Initially, the level-1 image was used as a mask to identify the macro-areas division. Masking is a basic process in image analysis which is generally used to hide or highlight regions in the area. The mask is a binary image consisting of values of 0 and 1, and it must have the same spatial extent and projection as the input image.

For each macro-category in the level-1 image, a single mask was obtained by reclassifying the corresponding areas as 1 and setting the others as 0. The habitat classes within each macro-category were then identified by multiplying the level-2 images for their corresponding masks.

Finally the raw classified map representing habitat distribution was built by overlaying all the level-2 images.

Moreover, a map with “not classified areas” was produced. This map represents areas where classification models show a lower accuracy; these areas have eventually been classified using GIS and remote sensing procedures, but it is advisable to control them in the further phase of visual validation of the map.

The resulting image was then processed with two filters in order to force the not classified areas and to remove the "salt and pepper" effect (Lillesand et al 2008). The first (nibble filter) replaces the non-classified cells with the values of the nearest neighbour (ESRI, 2014) while the second (majority filter) identify isolated pixels and replaces their value using a 3x3 moving window. Each pixel value is replaced if there are at least three (of its eight) contiguous cells which have the same value; otherwise cell retains its value.

In conclusion the filtered image was vectorized and smoothed, and patches smaller than the minimum mapping unit were eliminated in order to obtain the final classified image.

Figure 6.3 shows a scheme of the images overlaying procedure.

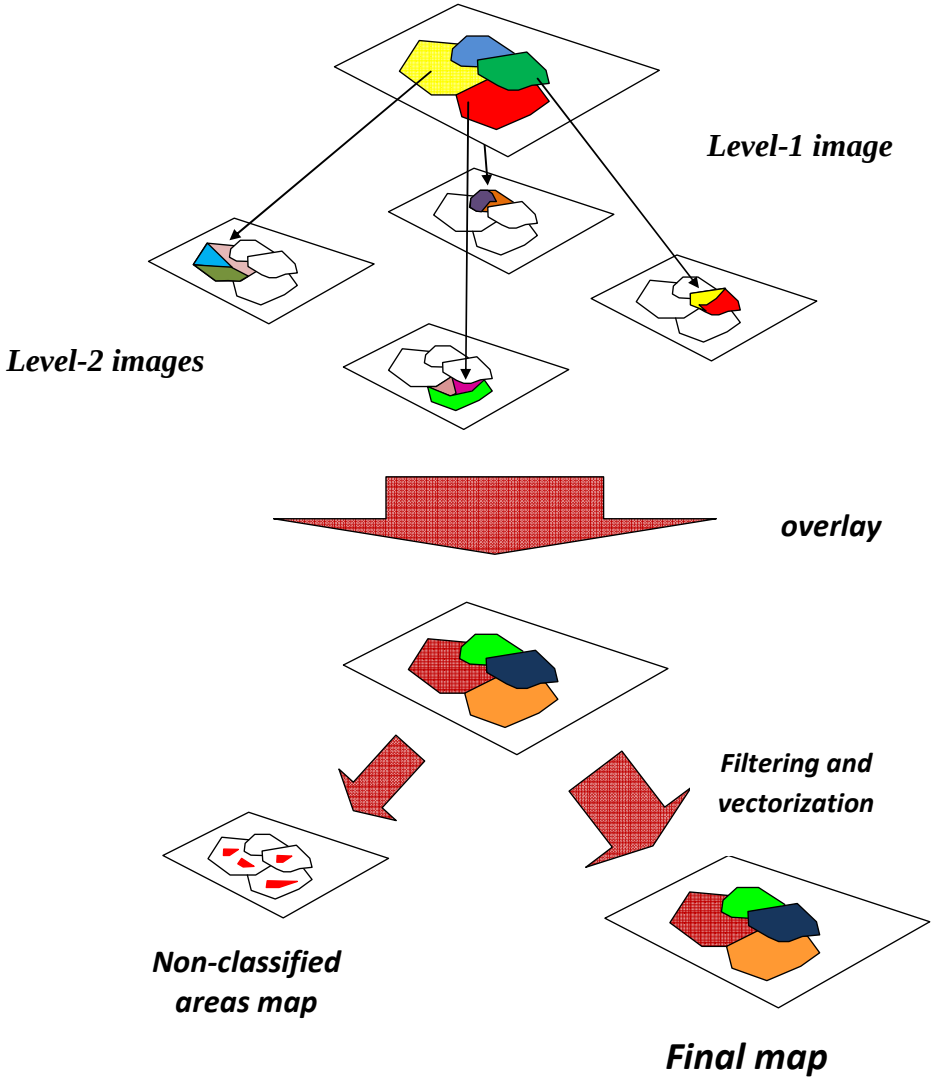


Figure 6.3 - Overlaying procedure scheme

6.4 Accuracy assessment

Accuracy assessment is a necessary and an integral part of mapping. Contemporary scientific and planning processes demand increasingly quantitative methods to evaluate spatial products reliability. The accuracy and precision of thematic maps can be quantified and accounted for in a number of ways; the scientific literature abounds with discussions on this topic and several methods are proposed both in general (e.g. Landis and Koch 1977, Story and Congalton 1986, Congalton 1991, Fitzgerald and Lees 1994, Gopal and Woodcock 1994, Green and Strawderman 1994, Hammond and Verbyla 1996, van Deusen 1996, Stehman 1997, Milliken et al. 1998, Stehman and Czaplewski 1998, Stehman 2005) and specifically in vegetation mapping for the interpretation and use of vegetation maps (e.g. Regan et al. 2002, Elith et al. 2002).

The expectations on the classification accuracy of a map can vary according to its potential use and its cartographic scale. For example, mid-scale maps (e.g. 1:100 000 or 1:50 000) are often used for regional planning, environmental impact assessment (EIA) studies as well as many other similar analysis. Such products must portray the spatial extent of vegetation types accurately at the sub-regional scale but would not be expected to be accurate at 100% of any given point. Such mapping may be considered acceptable when accuracy is between 60 and 80% (Sivertsen, D. 2009). On the other hand, fine-scale maps (e.g. 1:25 000 or finer) may be used for local activities and they are expected by the user to be more accurate. In practical terms this means that, as the demands on the mapping increase, additional expense and effort will be required to test, establish and improve the accuracy of that mapping.

Moreover, errors in the classification should have different weights depending on their magnitude. Indeed certain kinds of mismatches can be considered greater in error than others. For example, assigning a map label of water to an area of coniferous trees might be considered a larger error than confusing types of different grasslands. Furthermore, assigning an area of 100 percent coniferous trees to broad-leaved forest might be a more serious error than assigning the same label to an area that is 60 percent coniferous trees and 40 percent broad-leaved trees. In this context Gopal and Woodcock (1994) proposed a linguistic scale of correctness based on five degrees (absolutely wrong, understandable but wrong, reasonable or acceptable, good, absolutely right) which could be used to distinguish between the levels of accuracy required for the maps. Obviously, the expectations for a very detailed map will be different from those of a broader scale one.

In an accuracy assessment, classified areas are compared to a set of reference localities (validation data) that are regarded as ‘true’; the extent to which these two classifications agree is defined as map accuracy.

A common traditional method to quantify the accuracy of a thematic map is the confusion (or error) matrix (Card, 1982; Congalton, 1991; Hoffer and Fleming, 1978; Rosenfield and Fitzpatrick-Lins, 1986). It is a square matrix (see fig 6.4 for an example) whose columns usually represent the ground data (which are assumed to be corrected) and rows indicate the mapped data. Each element in the matrix gives the number of map units labelled according to reference data, which are assigned to a particular category; the elements of the principal diagonal of the matrix represent the correct matches and the remaining elements, the mismatches. Obviously, the ideal situation is represented by a diagonal matrix where only principal diagonal elements have non-zero values; all areas on the map have been correctly classified (Van Genderen and Lock, 1977; Mead and Szajgin, 1981; Congalton et al.,1983). This situation is rarely the case.

	C1	C2	..	Ck
as C1	n_{11}	n_{12}	…	n_{1k}
as C2	n_{21}	n_{22}	…	n_{2k}
…	…	…	…	…
as Ck	n_{k1}	n_{k2}	…	n_{kk}

Figure 6.4 – Example of confusion matrix

In this thesis the confusion matrices were used to evaluate the proposed classification method. The three test areas were compared using the official maps produced in the frame of the "Carta della Natura" project (see chapter 5). The classification accuracy was evaluated both considering the area units (A%) and ground point units (P%), using an equidistant grid of samples.

Four separate statistics are used to provide indications of the level of accuracy in the data:

- an overall accuracy (6.3), which is the proportion of all correctly classified units within the whole matrix for all habitat units examined. It is calculated by adding all cell values in the matrix diagonal and dividing by the total map units.
- a producer's accuracy (6.4), which is the probability that any data point within a unit has been correctly classified on a map. It is calculated for each reference class by dividing the number correctly classified by the column sum for that class and it usually expressed as a percentage
- a user's accuracy (6.5), which is the probability that a classified data point within a unit actually represents that unit in the field. Is is calculated for each map class by dividing the number correctly classified by the row sum for that class and it usually expressed as a percentage.
- the kappa coefficient (Cohen, 1960; Rosenfield and Fitzpatrick-Lins, 1986) (6.6) which provides a measure of the overall classification accuracy while correcting for matching that occurs by chance.

$$Overall\ Accuracy = \frac{\sum_{i=1}^{k} n_{ii}}{n} \tag{6.3}$$

$$Producer's\ Accuracy = \frac{n_{ik}}{\sum n_{ir}} \tag{6.4}$$

$$(class\ k)$$

$$User's\ Accuracy = \frac{n_{ki}}{\sum n_{ri}} \tag{6.5}$$

$$(class\ k)$$

$$kappa = \frac{N \sum_{i=1}^{k} n_{ii} - \sum_{r=1}^{k} (\sum n_{ir} \times \sum n_{ri})}{N^2 - \sum_{r=1}^{k} (\sum n_{ir} \times \sum n_{ri})} \tag{6.6}$$

Among these, the overall accuracy and the user's accuracy can be considered better indicators because they provide, respectively, a general indication and a user's point of view about classification performance. The use of the kappa coefficient, although reported in the results, has often been questioned because it can lead to an underestimation of map accuracy (Foody, 1992 Stehman, 1997; Turk, 2002; Jung, 2003)

Moreover, in order to evaluate the potential benefits of the proposed method in the classification analysis, a comparison test was performed by evaluating the resulting classified images with respect to those obtained by using a commercial classification software (ESRI ArcGIS, rel 10.1), with the same input data. In each study area two comparison tests were performed, in order to evaluate the relative importance of the two main components (PLS-DA recursive algorithm and 2-level stepwise classification) of the classification method:

- The first test, which will be referred to as "tutorial", was performed by comparing the overall accuracy of the classified maps to that of the classification produced using the Image Classification Toolbar available in ArcGIS, following the procedures of the tutorial provided by the software.

- The second test aims to evaluate the classification ability of PLS-DA algorithm vs the Maximum Likelihoood classification algorithm used by ArcGIS, removing the influence of the stepwise classification in the final result. Therefore this test was performed using a comparison map obtained by performing a two-step classification but using the maximum likelihood algorithm embedded in ArcGIS.

Finally, some quality issues associated with the use of a reference map for validation need to be considered:

a) there is a time gap between the reference maps and image acquisition dates that could be resulting in possible habitat changes. Indeed, the classical methods to produce thematic maps, and in particular vegetation and habitat maps, are very time consuming and their updates may require years to be released and validated, so it is difficult to have a perfect matching between the two datasets.

b) there is a certain degree of subjectivity of the photointerpreter. Indeed, in the delineation of homogeneous patches different analysts tend to give different interpretations according to their ability and experience. Moreover they could prefer to map habitats of greatest conservation importance

For this reasons, the results of the accuracy assessment presented in this thesis need to be interpreted with caution.

Type	Label	Description
Normalization	Normalize	Normalization of the rows dividing each variable by the sum of the absolute value of all variables in the sample.
	MSC mean	Multiplicative scatter correction: performs a regression of the measured spectrum against the mean spectrum and corrects the measured spectrum using the slope of this fit.
	MSC median	Multiplicative scatter correction: performs a regression of the measured spectrum against the median spectrum and corrects the measured spectrum using the slope of this fit.
	SNV	Standard Normal Variate: Divides each variable by standard deviation of all variables in the sample.
Variable centring	Mean centre	Calculates the mean of each column and substracts this from the column.
	Median centre	Calculates the median of each column and substracts this from the column.
	Logdecay	Scales each measure by a continuously decreasing log function of the form where si is the scaling for variable i, n is the total number of variables and τ is set at the default value of 0.3.
	Class centroid	Centers data to the centroid of all classes by calculating initially all class means and then the "class centroid" (as mean of these class means) and remove that from all samples. Samples belonging to class 0 (unknown class) are not used in calculating the centroid or pooled variance
	Class centroid and scale	Centers data to the centroid of all classes and scales to intra-class variance
Variable scaling	Autoscale	Centres columns to zero mean and scales to unit variance. It permits to correct different variable scaling and units if the predominant source of variance in each variable is signal rather than noise.
	Variance (std) scaling	Scales each variable by its standard deviation without mean-centering
	Groupscale	Performs scaling based on standard deviations by splitting the variables into a predefined number of equally-sized blocks and scaling each block by the grand mean of their standard deviations.
	Sqmnsc	Scale each variable by the square root of its mean.
	Pareto scaling	Scales each variable by the square root of its standard deviation

Noise, Offset & Baseline	Baseline	Removes a signal which is assumed to be interference ("the baseline"). It allows the user to fit a polynomial of a specific order to points which are known to be baseline (no-signal) points. This method is typically used in spectroscopic applications where the signal in some variables is due only to baseline (background). These variables serve as good references for how much background should be removed from nearby variables.
	Baseline (Automatic Weighted Least Squares)	Removes a signal which is assumed to be interference ("the baseline") using an automatic approach to determine which points are most likely due to baseline alone. It iteratively fits a baseline to each spectrum and determining which variables are clearly above the baseline (i.e., signal) and which are below the baseline. The net effect is an automatic removal of background while avoiding the creation of highly negative peaks.
	Detrend	Remove a constant, linear, or curved offset by substracting a polynomial of a given order to the entire sample.
	Smooth	Savitsky-Golay smoothing (Savitsky and Golay, 1964): is a low-pass filter used for removing high-frequency noise from samples. Smoothing assumes that variables which are near to each other in the data matrix (i.e., adjacent columns) are related to each other and contain similar information which can be averaged together to reduce noise without significant loss of the signal of interest. The algorithm essentially fits individual polynomials to windows around each point in the dataset. These polynomials are then used to smooth the data. The algorithm requires selection of both the size of the window (filter width) and the order of the polynomial. The larger the window and lower the polynomial order, the more smoothing that occurs
	Derivatives	Savitsky-Golay smoothing and derivatives. Derivatives are a form of high-pass filter and frequency-dependent scaling each variable (point) in a sample is subtracted from its immediate neighbouring variable (point). This subtraction removes the signal which is the same between the two variables and leaves only the part of the signal which is different. Because derivatives de-emphasize lower frequencies and emphasize higher frequencies, they tend to accentuate noise (high frequency signal). For this reason, the Savitzky-Golay algorithm is often used to simultaneously smooth the data as it takes the derivative, greatly improving the utility of derivatized data

Multivariate filtering	GLS weighting	Calculates a filter matrix based on the differences between pairs/groups of samples which should otherwise be similar. These differences are considered interferences and the filter attempts to shrink those interferences.
	OSC	Orthogonal Signal Correction: Removes variance in the X-block which is orthogonal to the Y-block in order to obtain data contained only in those covariance patterns which are useful or interesting in the context of the model.
	EPO	Performs a "hard" orthogonalization to the clutter. A PCA model is calculated for the clutter and the given number of PCs are extracted. The filter then orthogonalizes (removes) all the variance which matches these PCs. If the selected number of PCs is large, more variance will be removed and the filter may remove variance which is not clutter, but part of the signal of interest.
Multiway	Centering	Performs centering, when handling multi-way data, across one of more models of a multiway array. The result is a matrix which has a mean of zero of the given modes.
	Scaling	Performs scaling, when handling multi-way data, across one/more modes of a multiway array.

Table 6.2 – Pre-processing transformation applied to classification data

Chapter 7

Results

The proposed classification method was tested on three study areas: i) Monte Vulture volcanic complex, ii) Apulia lagoons and iii) Campo Pericoli basin. This chapter presents classification accuracy results for each site and it is organized as follows: first sections provide a short description of the site and of the training datasets used to perform the classification both for the macro-categories and the final habitat classes. Last sections initially present the results of the classification models built with the recursive PLS-DA algorithm and then the overall classification accuracy of the method, calculated using the reference maps produced in the frame of the Carta Natura project. The latter is assessed both *per se* and in comparison with a common used classification software available in commerce.

7.1 Monte Vulture volcanic complex

7.1.1 Description of study area

The study area of Monte Vulture volcanic complex (hereafter, Monte Vulture) (fig 7.1) is located between 15°31’26’’ to 15°58’50’’ E longitude and 40°44’56’’ to 41°1’48’’ N latitude in Basilicata administrative region and covers an area of approximately 490 km^2. Its specific shape depends on the presence of clouds in the Rapid Eye image available for the classification, which forced to cut several parts from the original rectangular study area.

In the study area there are five protected areas (*Grotticelle*, *Agromonte Spacciaboschi*, *Coste Castello* and *I Pisconi* nature reserves and *Grotticelle di Monticchio* site of community importance (SCI)).

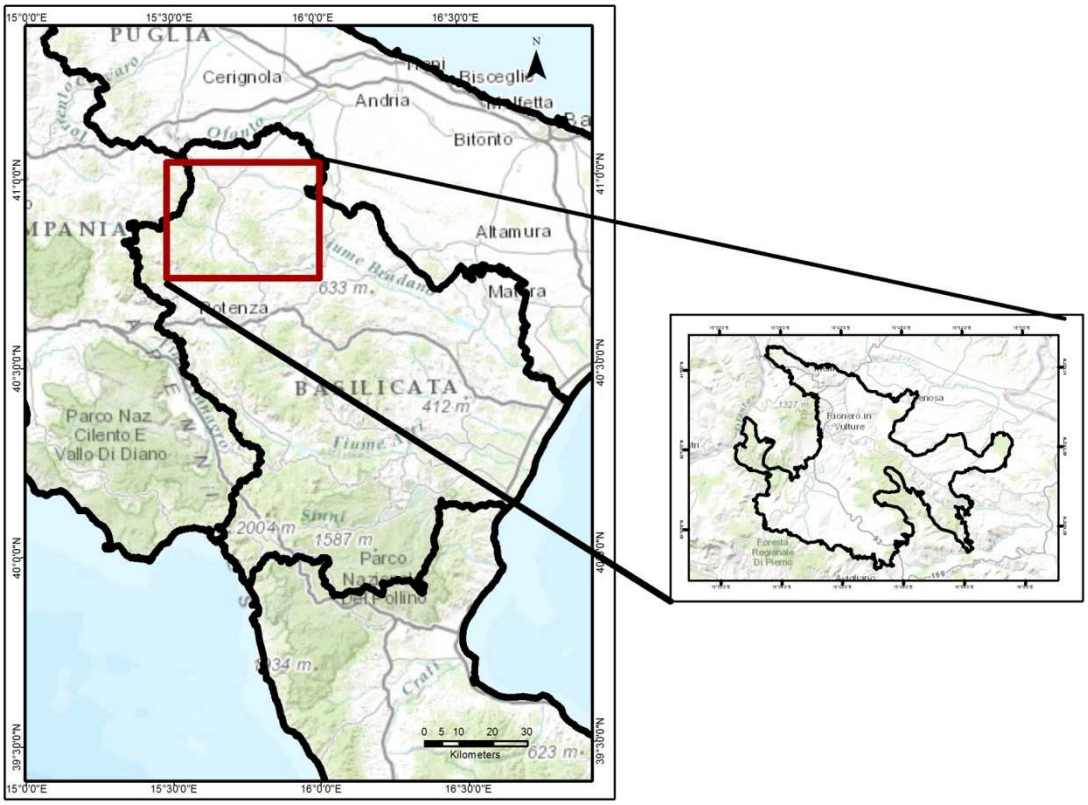


Figure 7.5 - Monte Vulture volcanic complex study area

Landform is mainly characterized by hills; elevation ranges between 250 m (Arcidiaconata torrent, in the north of the area of investigation) and 1238 m (Monte Caruso). The Monte Vulture (1326 m), an ancient volcano arising from the Fossa Bradanica depression, is located slightly out of the area. The area is included in two basins (fig 7.2a): the Ofanto river basin, which comprises the most of the area, and the Basento river basin, in the South East of the map.

In the area there are several ditches developing with a radial pattern from the Monte Vulture. They finally flow into two torrents named Antella and Arcidiaconata located respectively to the South and the East of the Monte Volture.

From the Italian Lithological Map 1:500,000 (derived from the Geological Map of Italy scale 1:500,000; Compagnoni et al.,1983) the geological framework of the study area is composed by three main units (fig 7.2b):

1. Lucano Apennine: commonly recognizable by terrigenous complexes with a flysch-like structure (from middle-late Cretaceous to Miocene).

2. *Fossa Bradanica* depression: a tectonic depression which was filled by predominantly clay and clay-sandy Plio-Pleistocene soils following the formation of a terraced series of normal faults that lowered the marginal portion of the Murgia platform (Tadolini and Bruno, 1984).

3. Monte Vulture volcanic complex

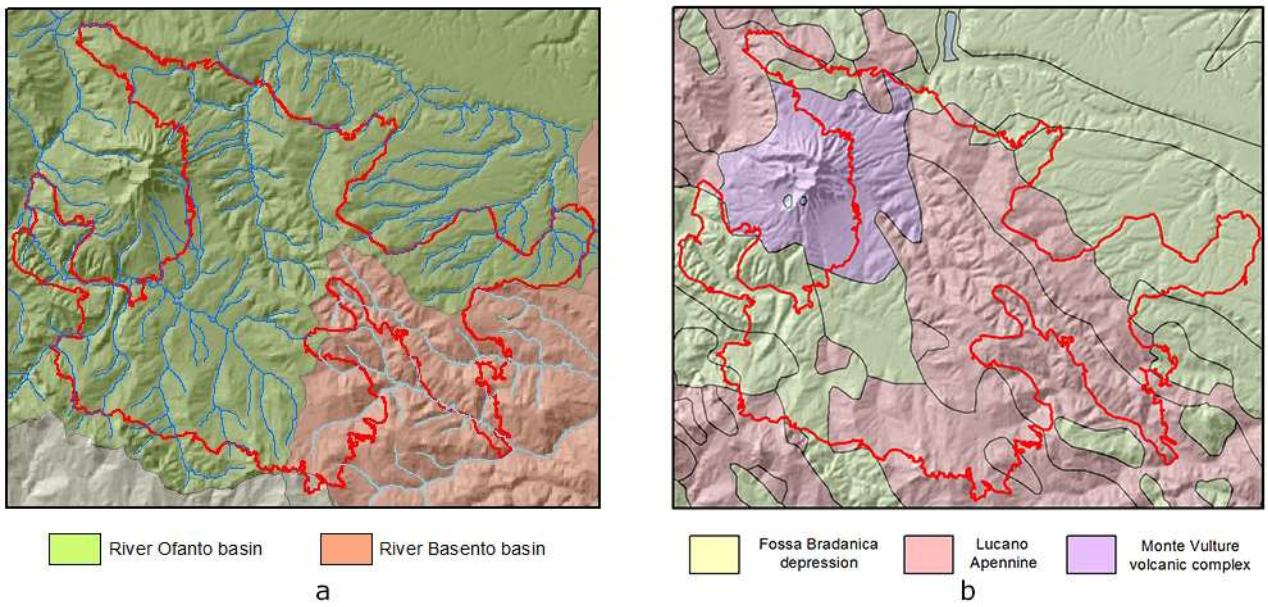


Figure 7.6 –Basins (a) and geological framework (b)

The main landscape units which can be identified from the physiographic units map (ISPRA, 2013a) at 1:250,000 are shown in table 7.1:

Landscape type	%
Volcano	0,05
Clay hills	4,21
Open plain	7,14
Terrigenous hilly landscape with plateaus	10,37
Volcanic hilly landscape with plateaus	10,86
Terrigenous reliefs with rocky ridges	67,37

Table 7.2 - Landscape units in Monte Vulture volcanic complex study area

The study area is characterized by a mediterranean climate. The annual average rainfall is about 875 mm, ranging from a minimum of 14 mm, recorded in June, to a maximum of 248 mm in January. Average temperatures (recorded in Rionero in Vulture station - 40°56' N, 15°40’ E) range from about 4.5°C in January to 23°C in August (table 7.2; ENEA 2002).

Month	Min		Max		Mean
	Mean	Extreme	Mean	Extreme	
1	1.1	-6.9	7.7	14.9	4.4
2	1.9	-6.3	9.5	18.6	5.7
3	3.3	-3.1	12.1	20.0	7.7
4	6.4	1.1	17.1	25.3	11.8
5	9.0	4.6	21.3	30.1	15.2
6	14.5	6.8	26.9	34.9	20.7
7	15.7	10.9	29.7	36.7	22.7
8	15.7	10.4	30.3	38.0	23.0
9	13.1	8.3	25.5	33.3	19.3
10	9.2	4.8	18.4	26.1	13.8
11	5.9	-1.2	13.3	20.6	9.6
12	3.6	-3.1	10.5	16.2	7.0

Table 7.3 - Mean monthly temperatures in Monte Vulture volcanic complex area

Vegetation in the area is distributed in three bioclimatic zones (Nimis and Martellos, 2008):

1. Sub-Mediterranean dry zone: characterized by deciduous xerophytes oak woods, mainly with pubescent oak (*Quercus pubescens)*

2. Sub-Mediterranean wet zone: characterized by mesophile oak woods mainly with Turkey oak (*Quercus cerris*) and Italian oak (*Quercus frainetto*), sometimes substitued by chestnuts (*Castanea sativa*)

3. Mountain zone: mainly characterized by beech woods (*Fagus sylvatica*).

From the *Carta della Natura* map of Regione Basilicata (ISPRA, 2012) the main habitats present at this site are as follows:

31.8 Western Palaearctic temperate thickets: Pre- and postforest formations, mostly deciduous, characteristic of the western Palaearctic deciduous forest zone

34.32 Sub-Atlantic semidry calcareous grasslands: More or less mesophile, closed formations dominated by perennial, tuft-forming grasses (with *Bromus erectus* and *Brachypodium rupestre*), colonizing relatively deep, mostly calcareous soils, in the sub-Mediterranean mountains of the Italian peninsula.

38.A Mediterranean subnitrophilous grass communities and Mesophile pastures: Graminoid formations which may cover vast expanses of post-cultural or extensive pasture lands or on fertilised and well-drained soils. Groups 34.81 and 38.1 Corine habitats.

41.7 Thermophilous and supra-Mediterranean oak woods: Forests or woods of submediterranean climate regions and supramediterranean altitudinal levels dominated by deciduous or semideciduous thermophilous *Quercus* species or by other southern trees such as *Carpinus orientalis* or *Ostrya carpinifolia*. Thermophilous deciduous trees may, under local microclimatic or edaphic conditions, replace the evergreen oak forests in mesomediterranean or thermomediterranean areas, and occur locally to the north in central and western Europe.

44.61 Mediterranean riparian poplar forests: Mediterranean multi-layered riverine forests of base-rich soils submitted to seasonal prolonged inundation with slow drainage, with *Populus alba*, *Populus nigra*, *Fraxinus angustifolia*, *Ulmus minor*, *Salix alba*, *Salix* spp., *Alnus* spp., lianas and often species of the *Quercetalia ilicis*, distributed in the mediterranean regions of the Iberian peninsula, southern France, the Italic peninsula, the large Tyrrhenian islands, the Hellenic peninsula, the southern Balkan peninsula, North Africa, and their zones of transition to adjacent climatic zones. Formations physiognomically dominated by tall *Populus alba* and/or *Populus nigra* are listed here. The poplars may, however, be absent or sparse in some associations which are then dominated by *Fraxinus angustifolia*, *Ulmus minor* and/or *Salix* spp. The poplar forests are usually the tall ligneous vegetation belt closest to the water in riverside catenas.

83.11 Olive groves: Mediterranean formations of *Olea europaea* var. *europaea*. Group both Ancient olive groves, often made of very old trees shading herbaceous layer and extensive cultivations. Sometimes substrate is maintained as semi-arid pasture lands leading to a confusion with abandoned crops. For this reason during the level-1 classification they was grouped into the *Grassland and scrubs* macrocategory.

82.A Field crops and Extensive cultivation: groups both intensive and traditional extensively cultivated crops (82.1 and 82.3 Corine habitats)

7.1.2 Training datasets

The image dataset used in this classification test is composed by a multilayer image with eleven bands (5 bands of the Rapid-eye image and 6 representing ancillary data) with a spatial resolution of 5 meters.

The training dataset containing the reference data on habitat distribution is composed by 357 ground samples collected between 2010 and 2012 and available in the ISPRA's "habitat check dataset – regione Basilicata" (ISPRA, 2011). Ground check data represent 16 habitats, classified according to Corine Biotopes legend; for the purpose of the classification only eight classes (level-2) were used grouping Corine Classes which are separable only by visual interpretation. Finally these eight classes were grouped into three macro-categories (level-1) representing: *a) Anthropogenic habitats, b) Grasslands and scrubs* and *c) Deciduous forests.*

Table 7.3 shows the classification legend used for this area, the corresponding level-1 and Corine Biotopes codes and the number of check samples.

Corine Biotope codes	*Level 2 classes*	*Level 1 classes*	*Training samples*
82.1, 82.3 Field crops and Extensive cultivation	82.A - Field crops and Extensive cultivation	Anthropogenic habitats	49
86.1 – Towns and Active industrial sites	86.A - Towns and Active industrial sites	Anthropogenic habitats	40
31.8A, 31.81, 31.844 - Tyrrhenian sub-Mediterranean deciduous thickets; Medio-European rich-soil thickets and Tyrrhenian broom fields	31.8 - Western Palaearctic temperate thickets	Grasslands and scrubs	35
34.323, 34.326 - Middle European [*Brachypodium*] semidry grasslands and Sub-Mediterranean [*Mesobromion*]	34.32 - Sub-Atlantic semidry calcareous grasslands	Grasslands and scrubs	43
34.81, 38,1 - Mediterranean subnitrophilous grass communities and Mesophile pastures	38.A - Mediterranean subnitrophilous grass communities and Mesophile pastures	Grasslands and scrubs	70
83.11 - Olive groves	83.11- Olive groves	Grasslands and scrubs	35
44.61 - Mediterranean riparian poplar forests	44.61 - Mediterranean riparian poplar forests	Deciduous forests	37
41.732, 41.737B, 41.7511 and 41.7512 - Italo-Sicilian [*Quercus pubescens*] woods; Eastern sub-Mediterranean white oak woods of Southern Italy; Southern Italic [*Quercus cerris*] woods and Southern Italic [*Quercus frainetto*] woods	41.7 - Thermophilous and supra-Mediterranean oak woods	Deciduous forests	48

Table 7.4 - Classification legend for Vulture Mount vocanic complex area

7.1.3 Classifier models performance

Independently validated model classifiers were generated in each level of classification for every possible combination of partition method (n = 2), partition ratio (n = 9), X pre-processing transformation (n =24) and number of LV (n = 11) and produced a total of 4752 suitable classification results.

In order to determine the optimum parameters for the best model classifiers, the validated classification results were averaged over common pre-processing and number of LV to produce a total of 264 parameters combinations results.

7.1.3.1 Macro-categories

In *level-1* classification test a total of 1340 classifier models show a prediction ability greater than 70%. Results obtained with Random Selection partition method are generally higher than those obtained with the Kennard-Stone (fig 7.3a), while no real predominance can be observed considering the different partition ratios used to split the calibration and validation subsets (fig 7.3b).

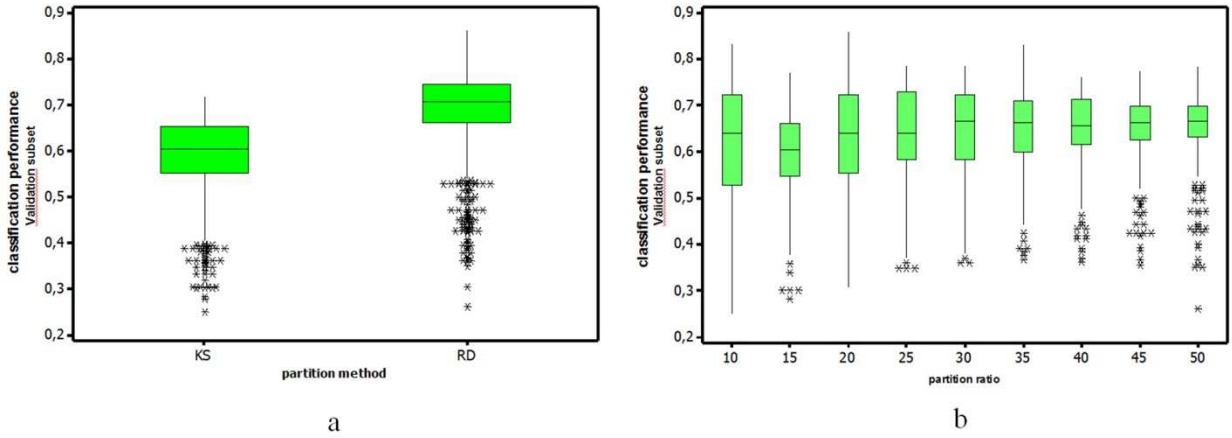


Figure 7.7 - Distribution of classification performance results considering the different partition methods (a) and partition ratios (b)

The best classification performance over all the partition methods and partition ratios was obtained using a GLS Weighting pre-processing with three latent variables. These parameters are then used to build the model to be used to classify the entire image.

Model parameters and results, both for calibration and validation phases, are presented in Table 7.4.

LV	3
Pre-processing X-Block	GLS Weighting
Mean of % classification ability (validation subset)	72.92
Mean of % classification ability (calibration subset)	77.90

Table 7.5 - Model parameters and results for level-1 classification

Figure 7.4 shows the importance of each variable involved in building the PLS-DA model. The heat map is obtained by calculating the VIP score (Chong and Jun, 2005) which relates the variable with each class estimating its contribution in discriminating pixels belonging to that class.

The most important variables are hence identifiable by high positive (dark red) values while values closer to zero (green) indicate less important variables which might be good candidates for exclusion from the model.

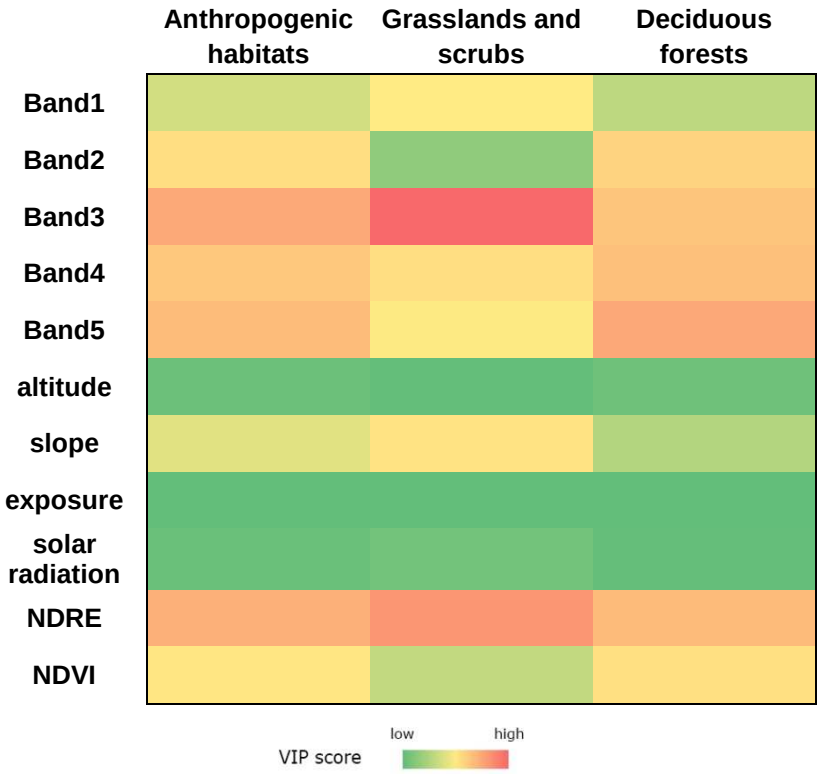


Figure 7.8 - Heat map of variable importance in building level-1 PLS-DA model

From figure 7.4 it's possible to see that Elevation, Aspect and Solar radiation data have a poor contribute in building the model while Rapid eye band 3 and Normalised Difference Red Edge Index (NDRE) are the most sensitive, especially in discriminating *Grassland and scrubs* macro-category.

7.1.3.2 Habitat classes

In order to produce the final habitats map a new classification test was performed within each macro-category identified in the level-1. Classifiers built for *Anthropogenic habitats* and *Deciduous forests* groups show generally a good classification ability with respectively 1379 and

1936 models of 4752 having a result higher than 70%. On the contrary, for the *Grassland and scrubs* all the models obtained show very poor results.

Figure 7.5 shows the distributions of classification performance values considering respectively the different partition methods and the partition ratios used in all the three classification tests to build the validation subsets

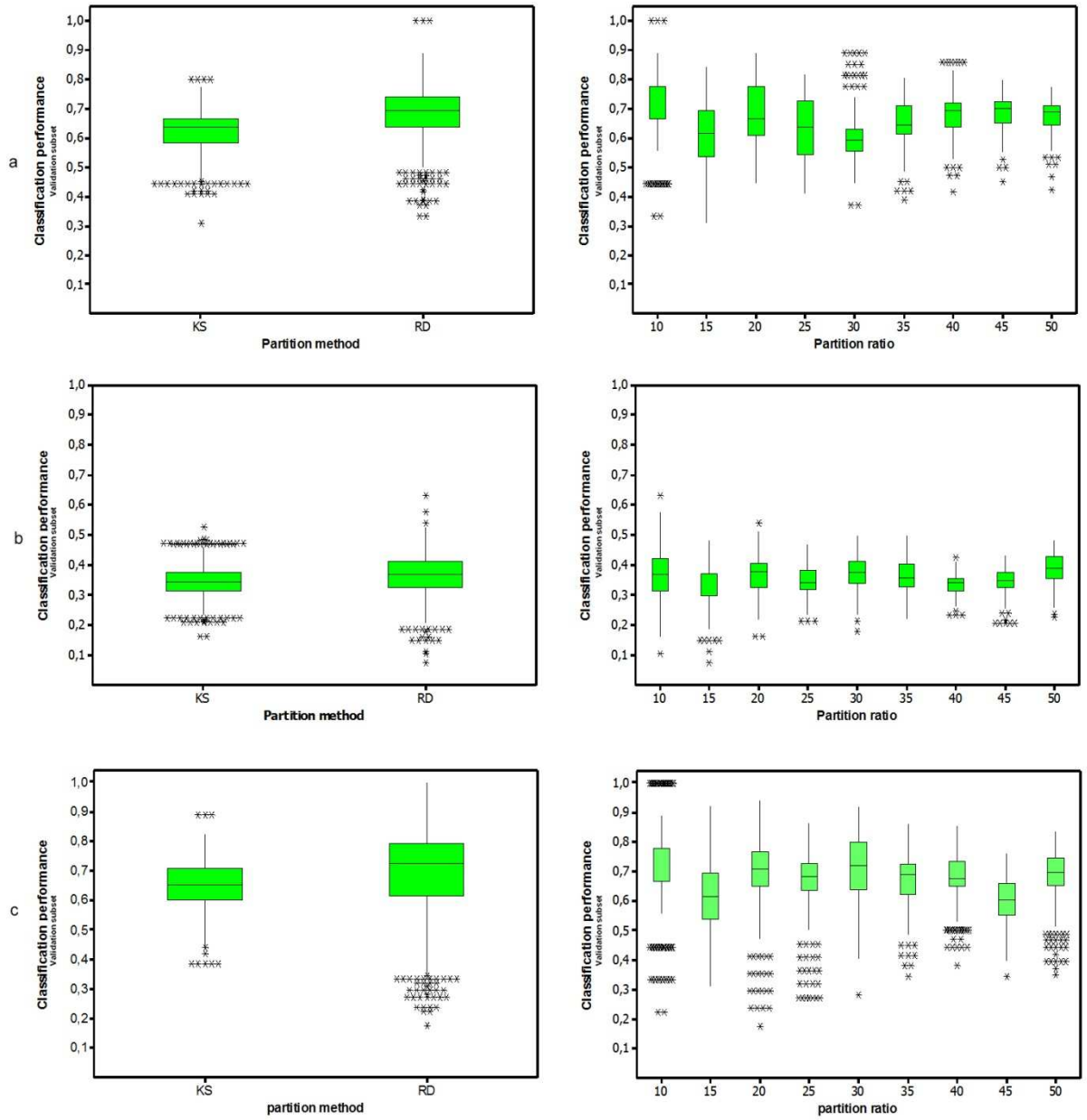


Figure 7.9 - Distributions of classification performance results considering the different partition methods and partition ratios for Anthropogenic habitats (a), Grassland and scrubs (b) and Deciduous forests (c) groups

Considering the partition methods, RD produces better performing results than KS both in *Anthropogenic habitats* and *Deciduous forests* groups, while no predominance can be observed in

grassland and scrubs macro-category. No predominance can be observed considering the different partition ratios as well.

Also averaged results shows a good classification performance for *Anthropogenic habitats* (71.66 in validation and 74.55 in calibration phase) and *Deciduous forests* (76.07 in validation and 76.85 in calibration phase) groups, while the low result of *Grassland and scrubs* indicate a general poor classification ability within that macro-category.

Model results and parameters for each macro-category are shown in table 7.5.

	Anthropogenic habitats	**Grasslands and scrubs**	**Deciduous forests**
LV	1	5	11
Pre-processing X-Block	Derivative	Normalize	Autoscale
Mean of % classification ability (validation subset)	71.66	40.99	76.07
Mean of % classification ability (calibration subset)	74.55	41.16	76.85

Table 7.6 - Model parameters and results for level-2 classifications

Heat maps for all PLS-DA models utilized in level-2 classifications were computed (fig 7.6). NIR band and elevation data are the most important variables in classifying agricultural and urban habitats (a); NIR band indicates that the amount of vegetation is a crucial factor to discriminate the two classes.

Classification model within *Grassland and scrubs* macro-category (b) is based mainly on Rapid-eye image bands, this is an evidence that other ancillary data should be used to increase the information available to improve the discrimination ability of PLS-DA classifier.

Finally in *Deciduous forests* group (c) slope and elevation data are the most significant factors useful to discriminate *Thermophilous and supra-Mediterranean oak woods* from riparian forests, which can be expected to occur predominantly on plains of low slope such as valleys bottom.

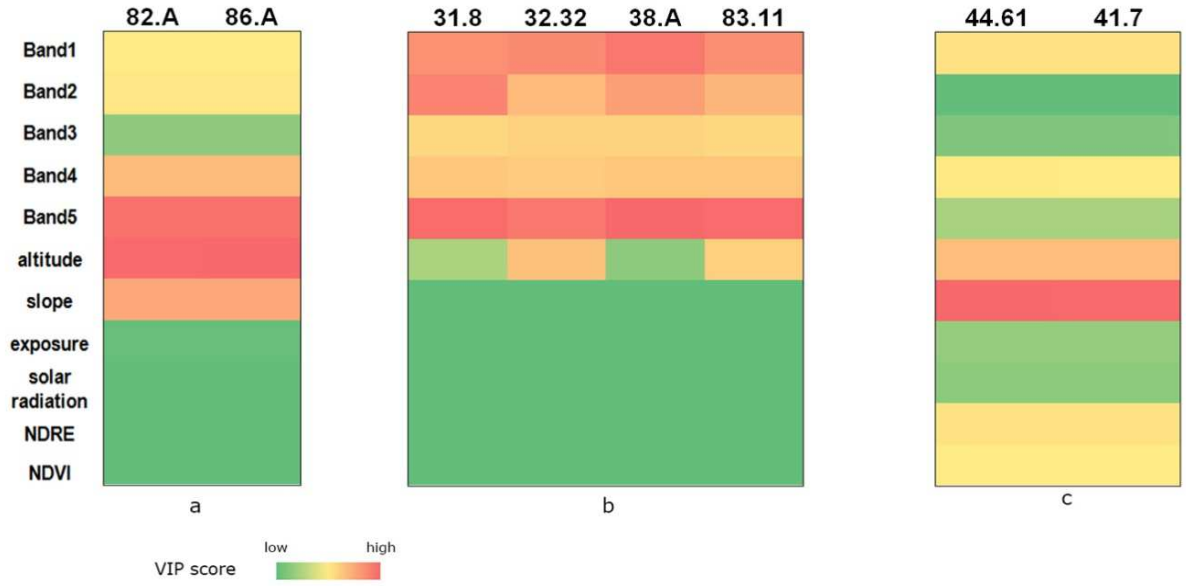


Figure 7.10 - Heat map of variable importance in building PLS-DA model for Anthropogenic habitats (a), Grassland and scrubs (b) and Deciduous forests (c) groups

7.1.4 Accuracy assessment

Model parameters were applied to the entire training datasets, to build the final classifiers to be used on the image dataset. The classified raster where then overlapped and processed (see section 6.3.5) to obtain the final classified map (at 1:50.000 scale, minimum mapping unit = 1 hectare) and the map of Non-Classified zones (annex I of the thesis).

In order to evaluate the classification accuracy classified map was compared using the official "*Carta della Natura della Regione Basilicata*" map, produced in the frame of the "*Carta della Natura*" project (ISPRA,2012). Figures 7.7 and 7.8 show the confusion matrices which compare the classified map *vs Carta della Natura* map. Accuracy values are calculated considering respectively the total mapped areas (A%) and a series of reference ground samples (P%) obtained using an equidistant grid of points. In the confusion matrix, the C0 class represents some small habitats (in total 2.6% of the study area) which were mapped in the reference map but not included in the check points collected in the field survey, and this is the reason why they were not represented in the training dataset. In particular, in the study area there are some vineyards producing an Italian famous wine named *Aglianico del Vulture*. However they are not represented in the classification legend because in the area the cultivations are fragmented into small patches, so in the map they have been underestimated because they have been included either in the more extensive olive groves or in extensive cultivation.

	C0	82.A	86.A	31.8	34.32	38.A	83.11	44.61	41.7	user's accuracy
as 82.A	773,051	114,569,793	651,126	843,395	2,853,096	4,070,970	809,211	1,028,166	1,723,759	89.98%
as 86.A	671,540	32,599,350	9,029,873	382,812	182,731	2,078,900	606,307	181,592	809,836	19.40%
as 31.8	649,383	2,490,775	394,932	662,368	1,429,576	1,252,622	1,299,935	260,939	7,059,973	4.27%
as 34.32	541,364	20,161,568	555,905	2,005,708	10,235,036	4,903,492	104,308	311,008	3,292,904	24.30%
as 38.A	559,772	21,398,024	1,124,555	821,531	3,205,655	4,579,519	1,582,259	188,614	806,452	13.36%
as 83.11	5,139,636	23,412,639	3,432,837	1,828,171	1,828,678	11,549,813	30,125,174	2,502,499	5,351,830	35.37%
as 44.61	2,603,311	1,193,996	119,077	127,810	219,884	834,358	2,365,127	2,405,172	13,275,673	10.39%
as 41.7	1,509,296	3,845,332	439,621	1,753,745	1,473,349	1,451,910	825,901	1,541,338	102,189,821	88.84%
producer's accuracy		52.16%	57.34%	7.86%	47.76%	14.91%	79.87%	28.57%	75.97%	

Overall accuracy 55.98%
kappa coefficent 0.46

Figure 7.11 - Confusion matrix (m^2 correctly mapped) for Monte Vulture volcanic complex area and classification accuracy indexes

	C0	82.A	86.A	31.8	34.32	38.A	83.11	44.61	41.7	user's accuracy
as 82.A		255	3	2	4	7	4	2	4	90.75%
as 86.A	2	68	18			4		1	2	18.95%
as 31.8	2	3	1	1	5	6	2		16	2.78%
as 34.32	2	48		3	14	9		1	8	16.47%
as 38.A	1	46	2	1	6	8	1		6	11.27%
as 83.11	11	41	8	6	5	23	60	5	15	34.48%
as 44.61	4	4				3	4	8	23	17.39%
as 41.7	1	8	2	3	4	2	3	4	204	88.31%
producer's accuracy		53.91%	52.94%	6.25%	36.84%	12.90%	81.08%	38.10%	73.38%	

Overall accuracy 55.74%
kappa coefficent 0.45

Figure 7.12 - Confusion matrix (number of reference ground points correctly mapped) for Monte Vulture volcanic complex area and classification accuracy indexes

The overall accuracy is 55.98% considering all mapped areas and 55.74% using the validation grid.

As shown in the model results, lower classification rates are given by *grassland and scrubs* classes. In particular 34.32 and 38.A habitats, as well as olive trees (83.11), are confused with cultivations; on the other hand 82.A, together with 41.7 class, presents an high classification marks. This can be due to the heterogeneous structure of the agricultural landscape in the area which groups many different cultivation types, so, if on the one hand the system succeeds in detecting the mapped cultivation patches, on the other hand it produces many false positives.

Table 7.6 shows the user's classification accuracy calculated within each macro-category. Results are still high for the *Anthropogenic habitats* and *Deciduous forests* groups and low for *grassland and scrubs*; this confirms that in this class the misclassification events are due also to errors in the first step when performing the macro-categories division, as habitats were confused mainly with 82.A class.

	Anthropogenic habitats	**Grasslands and scrubs**	**Deciduous forests**
User's accuracy (A%)	90,21	43,72	86,42
User's accuracy (P%)	91,49	40,98	86,28

Table 7.7 - User's classification accuracy calculated within each macro-category

In order to verify the capability of the method with respect to a commercial software of common use, a new classification test was performed using the maximum likelihood algorithm available in ArcGIS software (rel. 10.1). Two comparison tests were performed using the same training datasets: the first by classifying the areas according to the tutorial software and the second using the classifier with a two-level approach (see section 6.4).

The overall classification accuracy indexes for the proposed classification methods and for comparison tests are presented in table 7.7. Although classification accuracy of PLS-DA method is not very high in absolute terms, it is higher than those obtained with the commercial software in both experimental trials.

	PLS-DA	**Maximum likelihood** stepwise classification	**Maximum likelihood**
A%	**55.98**	**51.45**	**51.50**
P%	**55.74**	**52.21**	**51.91**

Table 7.8 - Classification accuracy indexes comparison

7.2 Apulia lagoons

7.2.1 Description of study area

The study area of Apulia lagoons (fig 7.9) is located between 15°30’7.7’’ to 15°42’58’’ E longitude and 41°50’2’’ to 41°55’50’’ N latitude in Puglia administrative region and covers an area of approximately 166 km^2. In the area there are five protected areas (*Gargano* National Park, *Isola Varano* and *Lago di Lesina* nature reserves, *Duna e Lago di Lesina - Foce del Fortore* (SCI), *Isola e Lago di varano* (SCI) and *Laghi di Lesina e Varano* (Special protected area - SPA)).

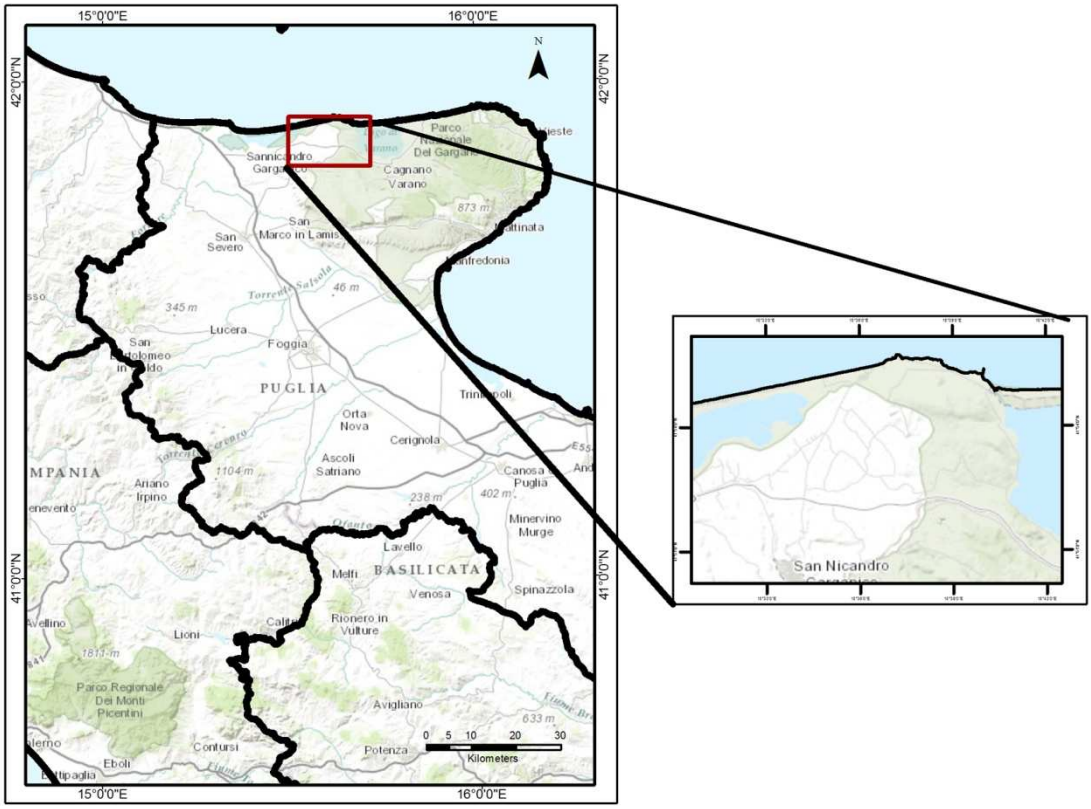


Figure 7.13 – Apulia lagoons study area

Landscape is characterized by two coastal plains (Lesina lagoon plain and Varano lagoon plain); elevation ranges from the coastal line to 718 m m.s.l. (Monte Rosella) in the south-east of the area.

The hydrography of the area is characterized by two lagoons:

- The brackish *coastal lagoon of Lesina* is connected to the sea through two inlets and has several freshwaters (FW) inputs, such as two perennial tributaries and several domestic and agricultural canals (Nonnis Marzano et al. 2003). Seasonal salinity fluctuations are not

severe due to the FW inputs that mitigate the effect of evaporation in summer months. On the other hand, an east–west spatial salinity gradient exists all over the year between different areas of the lagoon due to the reduced water circulation and the considerable contribution of FW inputs concentrated on the south-eastern side of the lagoon (Manzo 2010). Water temperature ranges from 10.3 to 27.6 °C (mean value: 18.4 °C; Roselli et al. 2009).

- *Varano lagoon* is a shallow lagoon with a mean depth of 5 m. A coastal barrier, 10 km long and 1 km wide, separates the lagoon from the Adriatic sea. The lagoon is connected to the sea through two artificial channels, that allow to exchange waters and sediments following the tidal cycle. Varano lagoon receives freshwater inputs characterized by a high organic content originating from urban and agricultural runoff, fish-farming and zootecnical activities (Villani et al., 2000; Spagnoli et al., 2002). The freshwater inputs to the lagoon originate from groundwater springs in the south-western basin of the lagoon, while in the south-eastern zone, urban wastewaters and drainage watercourses discharge into the lagoon through the two effluents Antonino and S.Francesco (Capoccioni 2013, Capoccioni et al 2014).

From the Lithological Map 1:500,000 the geological framework of the study area is composed by three main units (fig 7.10b):

1. Gargano massif: essentially formed of limestone and dolomite rocks (from the Mesozoic) with frequent inclusions of flint (nodules, slabs), covered with thin layers of calcarenites (from the Tertiary), and in some stretches (lake and coastal zones) include marine and watercourse deposits (from the Quaternary; Lopez, 2003).
2. Alluvial plain
3. Lakes complex

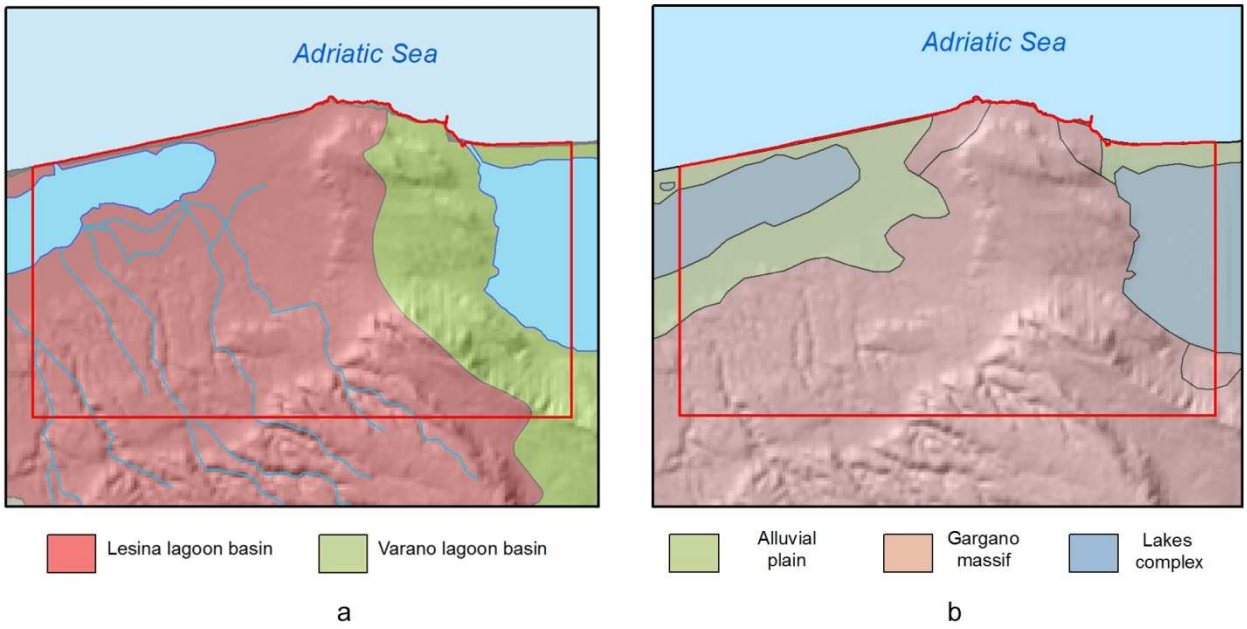


Figure 7.14 - Basins (a) and geological framework (b)

The main landscape units which can be identified from the physiographic units map (ISPRA, 2013a) at 1:250.000 scale are:

Landscape type	%
Carbonatic mountains	16.4
Coastal plain	68.0
Lake	15.6

Figure 7.15 - Landscape units in Apulia lagoons study area

The study area is characterized by a Mediterranean climate. Annual mean rainfall is about 594 mm with more than 70% occur during the period from, October to March. The maximum in recorded in December (88.2 mm) and the minimum in July (16.3). (ENEA, 2002). Average temperatures (recorded in Lesina station - 41°52' N, 15°21' E) range from about 7.4°C in January to 25.2°C in July (table 7.8).

Month	Min		Max		Mean
	Mean	Extreme	Mean	Extreme	
1	3.8	-2.1	11.1	17.6	7.4
2	3.8	-1.6	12.0	19.1	7.9
3	5.6	1.1	14.7	22.4	10.1
4	8.1	4.1	19.6	26.7	13.8
5	11.4	6.9	24.0	31.9	17.7
6	15.4	10.1	28.9	36.1	22.2
7	18.8	14.3	31.7	36.4	25.2
8	18.7	14.5	31.3	37.3	25.0
9	15.8	12.1	27.7	33.9	21.7
10	12.0	7.4	22.2	28.1	17.1
11	8.2	3.7	16.7	23.1	12.4
12	5.5	0.2	13.3	19.2	9.4

Table 7.9 - Mean monthly temperatures in Apulia lagoons area

Vegetation in the area is distributed in two bioclimatic zones (Nimis and Martellos, 2008):

1. Mediterranean dry zone: characterized by sclerophyllous woodlands and shrub formations.

2. Sub-mediterranean wet zone: characterized by mesophile oak woods mainly with Turkey oak (*Quercus cerris*)

From *Carta della Natura* map of Regione Puglia (ISPRA, 2009a) the main habitats present at this site are as follows:

21 Coastal lagoon: Saline or hypersaline lake connected with sea

82.A Field crops and Extensive cultivation: groups both intensive and traditional extensively cultivated crops (82.1 and 82.3 Corine habitats)

86.A Towns and Active industrial sites: built-up areas or site with current industrial or commercial use where buildings, roads and other impermeable surfaces occupy at least 30% of the land.

16.1 Sand beaches: gently sloping sand-covered shorelines fashioned by wave action along the coasts of the oceans, their connected seas and associated coastal lagoons

31.8A Tyrrhenian sub-Mediterranean deciduous thickets: mostly deciduous shrubs and hedges, often tall, luxuriant and rich in lianas, of submediterranean areas and moist stations in Mediterranean areas of peninsular Italy, Sicily, Sardinia and Corsica

32.6 Supra-Mediterranean garrigues: low shrub formations with pronounced Mediterranean affinities formed as a degradation stage of thermophilous deciduous woodland or sometimes of evergreen *Quercus* woodland in the supra-Mediterranean belt of the Mediterranean region. Included here are only those formations that are characteristic of the supra-Mediterranean level

53.1 Reed beds: water-fringing stands of tall vegetation by lakes (including brackish lakes), rivers and brooks, usually species-poor and often dominated by one species growing in stagnant or slowly flowing water of fluctuating depths, and sometimes on waterlogged ground.

83.11 Olive groves: Mediterranean formations of *Olea europaea* var. *europaea*. Group both Ancient olive groves, often made of very old trees shading herbaceous layer and estensive cultivations. Sometimes substrate is maintained as semi-arid pasture lands leading to a confusion with abandoned crops

45.A Sclerophyllous woodlands: thermo-Mediterranean woodland or arborescent matorrals with *Olea europaea* var. *sylvestris*, *Pistacia lentiscus*, *Ceratonia siliqua*, or. Groups both degradation or colonisation stages (32.12) and forest (45.1) and Southern Italian holm-oak forests (45.31A)

42.84 Aleppo pine forests: woods of Pinus halepensis, a frequent colonist of thermo- and calcicolous meso-mediterranean scrubs. The distinction between spontaneous forests and long-established formations of artificial origin is often difficult. The latter are thus included here, while recent, obviously artificial groves are not

41.A Broad-leaved deciduous forests: woodlands and forests dominated by *Quercus cerris* and *Quercus frainetto*. On the highest slopes it is substituded by *Ostrya carpinifolia* in low organic matter soil or by *Acer* sp. and *Fraxinus* sp. on the wet zones. Groups 41.7511, 41.41 and 41.81 codes.

7.2.2 Training datasets

Rapid-eye images at 5 meters spatial resolution were utilized also in this test area to build the multilayer image to be used as base for the classification.

The training dataset containing the reference data on habitat distribution is composed by 1159 ground samples collected between 2008 and 2009 and available in the ISPRA's "habitat check dataset – regione Puglia" (ISPRA, 2009b). Ground check data represent 16 Corine Biotopes habitats which were labelled in 11 habitat classes, grouping that habitats which are separable just by visual interpretation. Level-1 macro-categories were used to group 82.A, 86.A and 16.1 classes as "*Anthropogenic habitats*", 31.8A, 32.6 and 53.1 classes as "*Scrubs and reeds"* and finally 83.11 and 45.A classes as "*Non deciduous forests*". Two habitats (*Coastal lagoons* and *Broad-leaved deciduous forests*) were not grouped in any macro-category.

Table 7.9 shows classification legend used for this area, the corresponding level-1 and Corine Biotopes codes and the number of check samples.

Corine Biotope codes	*Level 2 classes*	*Level 1 classes*	*Training samples*
21 - Coastal lagoons	21 - Coastal lagoons	Coastal lagoons	147
82.1, 82.3 Field crops and Extensive cultivation	82.A - Field crops and Extensive cultivation	Anthropogenic habitats	158
86.1, 86.3 – Towns, Active industrial sites	86.A - Towns and Active industrial sites	Anthropogenic habitats	129
16.1 - Sand beaches	16.1 - Sand beaches	Anthropogenic habitats	47
31.8A - Tyrrhenian sub-Mediterranean deciduous thickets	31.8A - Tyrrhenian sub-Mediterranean deciduous thickets	Scrubs and reeds	15
32.6 - Supra-Mediterranean garrigues	32.6 - Supra-Mediterranean garrigues	Scrubs and reeds	62
53.1 - Reed beds	53.1 - Reed beds	Scrubs and reeds	150
83.11 - Olive groves	83.11- Olive groves	Non deciduous forests	148
32.211, 45.1 , 45.31A - Oleo-lentisc brush and Olive-carob forests and Southern Italian holm-oak forests	45.A – Sclerophyllous woodlands	Non deciduous forests	150
42.84 - Aleppo pine forests	42.84 - Aleppo pine forests	Non deciduous forests	6
41.7511, 41.41, 41.81 - Southern Italic [*Quercus cerris*] woods, Medio-European ravine forests, Hop-hornbeam woods	41.A - Broad-leaved deciduous forests	Deciduous forests	147

Table 7.10 - Classification legend for Apulia lagoons area

7.2.3 Classifier models performance

Independently validated model classifiers were generated in each level of classification for every possible combination of partition method (n = 2), partition ratio (n = 9), X pre-processing transformation (n=24) and number of LV (n = 11) and produced a total of 4158 suitable classification results.

In order to determine the optimum parameters for the best model classifiers, the validated classification results were averaged over common pre-processing and number of LV to produce a total of 231 parameters combinations results.

7.2.3.1 Macro-categories

In *level-1* classification test a total of 2307 out of 4158 classifier models returned a prediction ability greater than 70%, showing a general good capacity of PLS-DA classifier to discriminate the three macro-categories in this area. Also in this test, the Random Selection partition method produced generally higher results than Kennard-Stone method (fig 7.12a) while no real predominance can be observed considering the different partition ratios used to split the calibration and validation subsets (fig 7.12b).

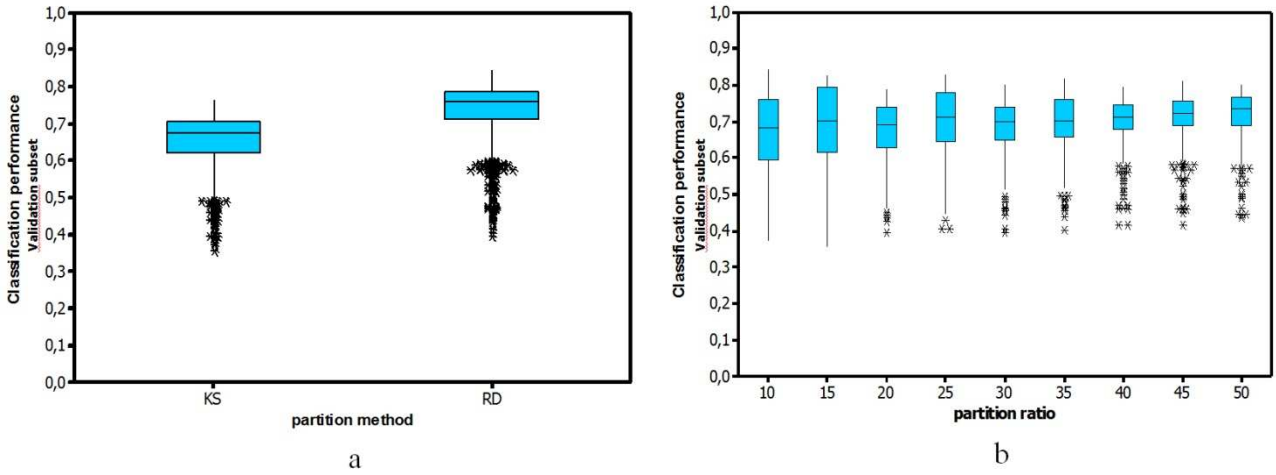


Figure 7.16 - Distribution of classification performance results considering the different partition methods (a) and partition ratios (b)

The best classification performance over all the partition methods and partition ratios was obtained using Autoscale pre-processing with seven latent variables. These parameters were then used to build the model to be used to classify the entire image.

Model parameters and results, both for calibration and validation phase, are presented in Table 7.10.

LV	7
Pre-processing X-Block	Autoscale
Mean of % classification ability (validation subset)	77.3
Mean of % classification ability (calibration subset)	80.9

Table 7.11 - Model parameters and results for level-1 classification

As *Coastal lagoon* (21) and *Broad-leaved deciduous forests* (41.A) habitats were not grouped into macro-categories, their detection depends on the level-1 classification. Considering the lagoon class, the most discriminant variables are NIR and Red-edge bands, indicating that the difference in the amount of biomass allows to separate this class from others (fig 7.13). For 41.A class the incoming solar radiation and the slope seem to be less important for the discrimination of this habitat.

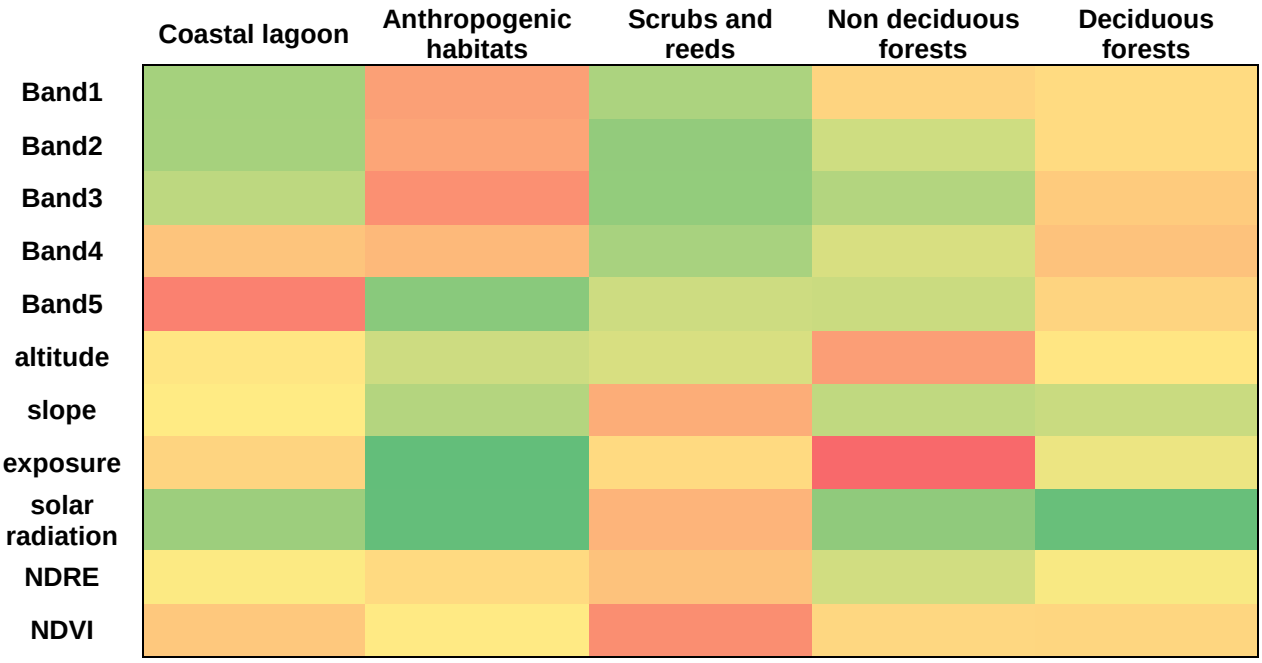


Figure 7.17 - Heat map of variable importance in building level-1 PLS-DA model

7.2.3.2 Habitat classes

Level-2 classifications were performed within *Anthropogenic habitats*, *Scrubs and reeds* and *Non deciduous forests* macro-categories. Classifiers built in level-2 tests show generally a good classification ability in all the three habitat groups. The parameters with the best classification results over all the partition method and partition ratios used to build the validation subsets are shown in table 7.11.

	Anthropogenic habitats	**Scrubs and reeds**	**Non deciduous forests**
LV	8	11	7
Pre-processing X-Block	msc	Median centre	Autoscale
Mean of % classification ability (validation subset)	81.5	95.4	77.5
Mean of % classification ability (calibration subset)	81.9	94.5	76.1

Table 7.12 - Model parameters and results for level-2 classifications

Fig 7.14 shows the distributions of classification performance values grouped by the partition methods and partition ratios. Although results are generally high, all distributions show a large number of outliers; this situation helps explain the use of the recursive algorithm in choosing the model to be used in the classification; indeed, in this case, the use of the best performing models instead of the most robust ones, could be an erroneous choice leading to poorer results in the final map classification.

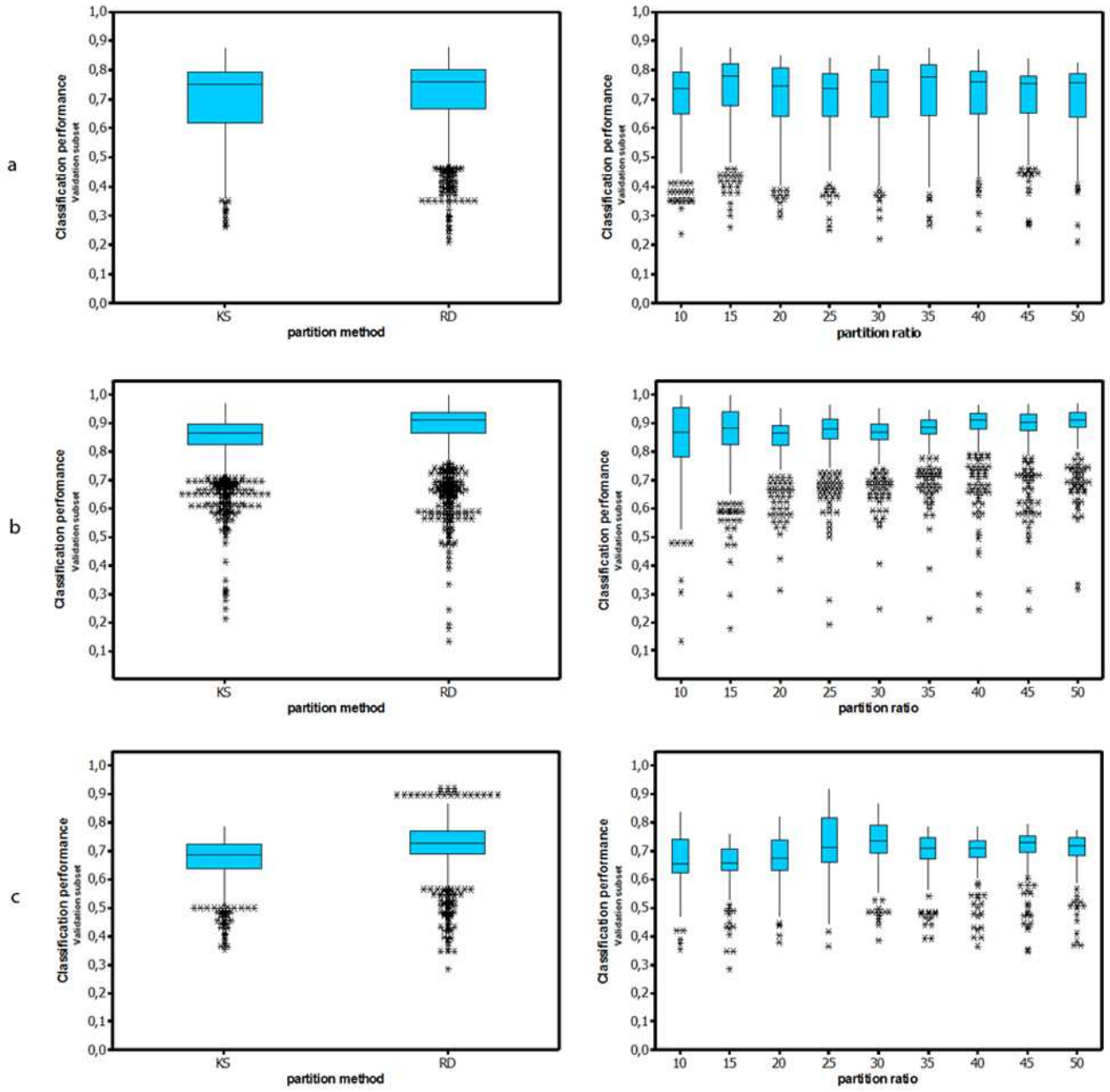


Figure 7.18 - Distributions of classification performance results considering the different partition methods and partition ratios used in the classification for Anthropogenic habitats (a), Scrubs and reeds (b) and Non deciduous forests (c) groups

Heat maps for all PLS-DA models utilized in level-2 classifications were computed (fig 7.15). In the first group (a) the variables contributing most to the separation of habitats are essentially the spectral bands of the Rapid Eye images while ancillary data (topographic variables and vegetation indexes) seem to be less important for the discrimination of these classes.

In the *scrubs and reeds* group (b) red-edge band is important in the detection of 53.1 habitat while the two vegetation indexes are scarcely used by PLS-DA model and seem to be less important for the discrimination of the habitat in the macrocategory.

Finally in the third group (c) red-edge and NIR bands, as well as the exposure and the incoming solar radiation, seems not to be useful in class discrimination.

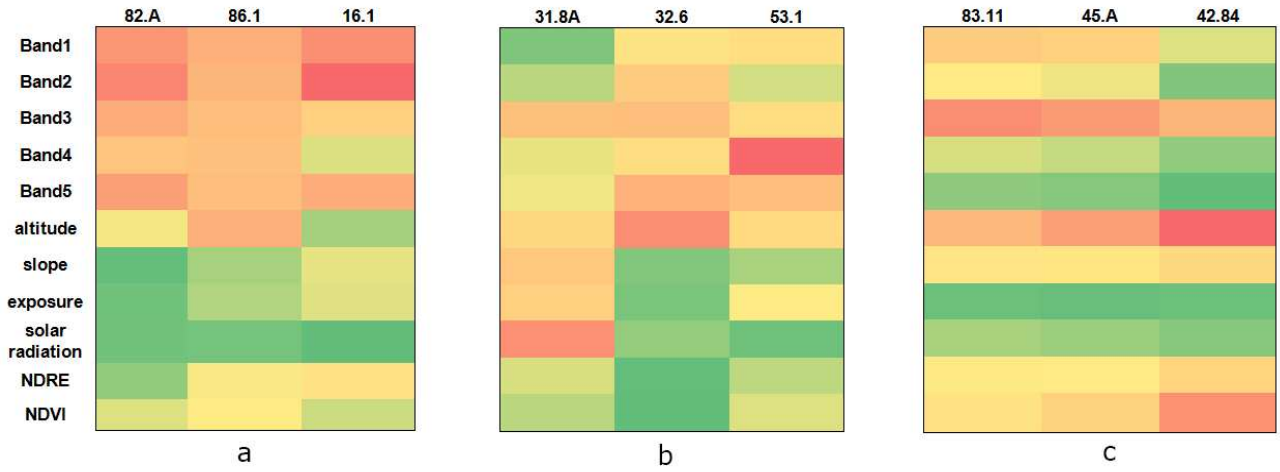


Figure 7.19 - Heat map of variable importance in building PLS-DA model for Anthropogenic habitats (a), Scrubs and reeds (b) and Non deciduous forests (c) groups

7.2.4 Accuracy assessment

Model parameters were applied to the entire training datasets and the resulted classifiers were used to obtain the final classified map (at 1:50.000 scale, minimum mapping unit = 1 hectare) and the map of Non-Classified zones (annex II of the thesis).

Accuracy assessment was performed using the official "*Carta della Natura della Regione Puglia*" map produced in the frame of the "*Carta della Natura*" project (ISPRA,2009a). Confusion matrixes which comparing the classified map *vs Carta della Natura* map are shown in figures 7.16 and 7.17.

Again, C0 class represents some small habitats (in total 1.44% of the study area) which were cartographed in the reference map but not represented in the check points collected in the field surveys.

	C0	21	82.A	86.1	16.1	31.8A	32.6	53.1	83.11	45.A	42.84	41.A	user's accuracy
as 21	3,581	20,740,039	59,456	47,230	1,577	833		286,517	54,236	11,505			97.81%
as 82.A	282,003	20,442	25,650,980	1,474,189	27,490	75,707	147,864	76,049	2,573,669	1,200,417			81.36%
as 86.1	141,741	4,736	3,021,512	2,745,309	46,595	13,473	29,471	4,265	534,042	1,780,585		47,424	32.80%
as 16.1	116,899	9,522	79,828	138,509	614,553	20,738			26,022	72,990			56.95%
as 31.8A	3,778		45,122			16,850			24,024	530,725			2.72%
as 32.6	53,152		2,306,636	13,706		26,580	1,036,942		591,134	4,124,555	41,535	149,617	12.43%
as 53.1	1,293,892	483,411	2,000,645	259,013	187		101,736	3,673,634	101,187	230,884			45.11%
as 83.11	388,626	1,217	7,883,646	103,592	71	148,217	484,259		11,392,436	6,991,907	6,314	16,099	41.55%
as 45.A	104,555	2,240	2,294,213	296,961	8,080	1,427,446	581,777		1,884,182	33,937,987	14,128	175,228	83.33%
as 42.84			15,730						10,436	258,332	10,058		3.41%
as 41.A			277,167	13,515		5,029	2,692,336		88,592	6,570,493		8,554,967	47.00%
producer's accuracy		97.55%	58.79%	53.91%	87.98%	0.97%	20.43%	90.92%	65.93%	60.92%	13.96%	95,66%	

Overall accuracy 65.31%

kappa coefficent 0.58

Figure 7.20 - Confusion matrix (m^2 correctly mapped) for Apulia lagoons area and classification accuracy indexes

	C0	21	82.A	86.1	16.1	31.8A	32.6	53.1	83.11	45.A	42.84	41.A	user's accuracy
as 21		122						1					99.19%
as 82.A	2		150	11		2	2	1	16	13			76.14%
as 86.1			23	15					2	14		1	27.27%
as 16.1	2				9				1				75.00%
as 31.8A										3			0.00%
as 32.6			15				8		5	31		1	13.33%
as 53.1	4	2	14	2			1	25		1			51.02%
as 83.11	2		49				3		77	43			44.25%
as 45.A	3		18	4		8	4		11	193	1	2	79.10%
as 42.84										1			0.00%
as 41.A			1				18			47		51	43.59%
producer's accuracy		98.39%	55.56%	46.88%	100.00%	0.00%	22.22%	92.59%	68.75%	55.78%	0.00%	92.73%	

Overall accuracy 62.80%

kappa coefficent 0.55

Figure 7.21 - Confusion matrix (number of reference ground points correctly mapped) for Apulia lagoons area and classification accuracy indexes

The overall accuracy is 65.31% considering all mapped areas and 58.21% using the validation grid. In particular, in this area the proposed procedure shows a good classification ability (>80%) with *lagoon*, *field crops and extensive cultivations* and *Sclerophyllous woodlands*. Reeds habitats and beaches show high values of producer's accuracy but lower results considering user's point of view because of the high number of false positive classifications. 31.8A and 42.84 classes show very poor results, probably because such two habitats are scarcely represented in the area (both with only one polygon mapped). Considering classification accuracy within each

macrocategory (table 7.12), in the *non-deciduous forests* group there are some misclassification cases between 45.A category, which is composed principally by olive-carob forests, and olive groves (83.11). This is probably due to the irregular structure of olive plantations in the area: scattered olive trees can be scarcely differentiated from uncultivated populations. Besides that, the olive groves are in part confused with other cultivations (82.A) probably due to the heterogeneity of the agricultural landscape in the area.

	Coastal lagoon	**Anthropogenic habitats**	**Scrubs and reeds**	**Non deciduous forests**	**Deciduous forests**
User's accuracy (A%)	97.81	82.48	28.38	79.64	47.00
User's accuracy (P%)	99.19	78.79	30.36	77.80	43.59

Table 7.13 - User's classification accuracy calculated within each macro-category

Table 7.13 shows the classification accuracies of the proposed method in comparison with commercial software's performances. The accuracy is higher than that obtained with both the ArcGIS tests. In particular, the classification performed using the software tutorial is particularly unsuccessful (5.39% and 5.12%) suggesting that the stepwise procedure could be useful to improve classification performance.

	PLS-DA	**Maximum likelihood** stepwise classification	**Maximum likelihood**
A%	**65.31**	**58.21**	**5.39**
P%	**62.80**	**58.07**	**5.12**

Table 7.14 - Classification accuracy indexes comparison

7.3 Campo Pericoli basin

7.3.1 Description of study area

Campo Pericoli is an inner basin of the Gran Sasso Massive. The study site covers an area of approximately 3 km^2 in the Gran Sasso and Monti della Laga National Park (fig 7.18). It is included

both a Site of Community Importance (SCI) and in a Special Protection Area (SPA) of the Natura 2000 network.

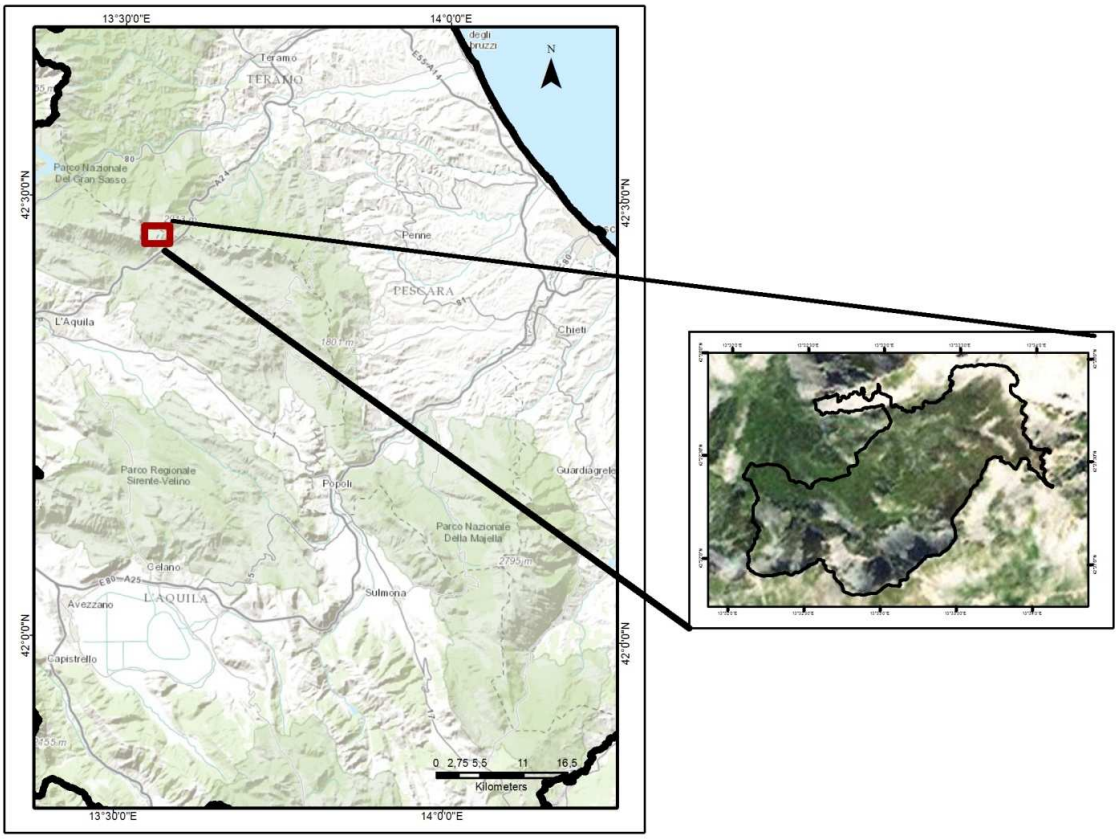


Figure 7.22 – Campo Pericoli basin study area

Altitude ranges from 2000 to 2600 m a.s.l.. The area is surrounded by ridges with a unique mouth towards *Val Maone* valley. It is closed by *Corno Grande* southern slope and *Primo Scrimone* ridge (at North), by the *Sella di Corno Grande*, *Monte Aquila* and *Sella di Monte Aquila* crest (at East), by *Monte Portell*a ridge (at South) and finally at West by the ridges developing from *La Portella* pass to *Pizzo di Intermesoli* mount.

Ridges structure is composed by limestone and calcareous marls (Mesozoic-Cenozoic); in particular in the study area there are both bedded and massive carbonatic rocks (Giurassic-Cretacic), partially covered by glacial deposits (Pleistocene) in the lower areas with smaller acclivity, and by largely active scree slopes in the steep slopes.

The geomorphologic framework is characterized by inactive landforms due to pleistocenic glacialism (moraine, cirque, overdeepened hollow) and by largely active cryogenic, running waters, gravity and karst landforms (protalus rampart, striated soils, solifluction, turf banked terrace, turf

hummocks, degradational scarps, slope debris, scree slope, field of small dolines, karren) (D'alessandro et al., 2003).

The hydrography of the area is poor and is characterized by ephemeral or intermittent streams due to the seasonal snowmelt.

The elevation and the orographic features in the area determine typical alpine climatic conditions; winters are relatively long and the beginning of spring is delayed by the persistence of snow on the ground (6-9 month/year). Both climate conditions and habitats structure are influenced by the marked action of the wind, mainly in summit crests.

Climate data were recorded in *Campo Imperatore* station (42° 27' 0" N, 13° 42' 0" E) and cover a range from 1951 to 1981 (Biondi, 1999). Annual mean rainfall is about 1143 mm with about the than 70% occur during the period from September to April. The maximum in recorded in November (119 mm) with two minima in march (90 mm) and July (55 mm). The snow covers the area generally from October to April with the highest depth in February and March. Average temperatures ranges from -3.2°C in February to 12.1°C in August, with values below zero from December to March. In the observation period the recorded minimum temperature was -23.9°C.

With the respect to *Campo Imperatore* station, the study area is located to an higher altitude and has a greater exposure to north with lower incoming solar radiation, so it can expect to have a more "alpine" climatic conditions.

Vegetation in the area is characterized by formations typical of alpine and sub-nival levels: grassland, brushes and screes. From *Carta della Natura* map (ISPRA, 2013b) almost all mapped habitats are listed in the Annex I of Habitats Directive and three of them are labelled as "priority"

The main habitats, classified according to Palearctic system are as follows:

31.431 Mountain *Juniperus nana* scrub: Thermophile *Juniperus nana*-dominated heaths of the upper levels, mostly of the subalpine or equivalent levels, of the Alps. In the study area can be found at the edge of the moraines and of the calcareous outcrops within the grassland matrix, mostly characterized by 36.436 and 36.414.

36.A Oro-Apennine and Pyreneo-Alpine grasslands: is a composed class dominated by mesophile, closed, short turfs of the subalpine and alpine levels of the southern and central Apennines, developed locally above treeline, on both calcareous and siliceous substrates (36.38). In the study area they form a mosaic with subalpine and alpine hygro-mesophile (36.313), sinkholes "dolina" habitats (36.3A) and apennine tall herb communities (37.816)

36.3464 Alpine [*Juncus trifidus*] swards: *Juncus trifidus*-dominated swards of the siliceous inner Alps and of lime-free anomalous stations of the calcareous outer Alps. In the study area they can be found between 2050 and 2450 m in the slope zones with a northern exposition.

36.4A Apennine naked-rush swards and Violet fescue: groups closed grasslands of the subalpine and lower alpine levels of the Alps, the Pyrenees and the Apennines dominated by *Festuca violacea* or *Festuca nigrescens* and small formations of *Elyna myosuroides*. In the study area they can be found in the wind slopes and crests which delimitate the basin.

36.436 Apennine stripped grasslands: Open, xerophile, stripped, stepped, scraped and garland grasslands of alpine and subalpine slopes and summits of the central and southern Apennines, dominated by *Sesleria apennina, Sesleria nitida, Sesleria italica, Festuca dimorpha, Carex kitaibeliana.*

61.A Screes: calcareous and calcschist screes of high altitudes and cool sites in mountain ranges of the nemoral zone, including the Alps, Pyrenees and Caucasus. Usually sparse vegetation cover, unstabile, on steep slope. This class also groups Alpide [*Salix retusa-reticulata*] snowbed communities which can be found on the slopes and are difficulty discriminated from non-vegetated zones.

62.15 Alpine and sub-mediterranean cinquefoil cliffs: Calcareous cliff and rock communities of the Alps and the Carpathians, of lesser satellite ranges and of sub-Mediterranean areas of the northern Tyrrhenian periphery. Dominant species include ferns *Asplenium ruta-muraria, Asplenium trichomanes, Asplenium viride, Cystopteris fragilis, Gymnocarpium robertianum* vascular plants (e.g.*Saxifraga paniculata*) and mosses. In the map representation they have a long narrow shape.

7.3.2 Training datasets

As Campo Pericoli basin is mapped at an higher cartographic scale (1:10.000) than the other two test areas, the base map used to classify this zone is a digital color-infrared (4 bands) aerial image with an higher spatial resolution (0.2 meters). The orthophoto was then resampled to obtain pixels of 2x2 meters in order to build the final image dataset composed by the four aerial image bands and five additional layers representing ancillary data (table 6.1).

The training dataset containing the reference data on habitat distribution is composed by 525 ground samples collected with dedicated field campaigns organized by ISPRA between 2011 and 2013; ground check data represents 14 Corine Biotopes habitats which were labelled in 7 habitat

classes, grouping that habitats which are separable just by visual interpretation. Level-2 habitats were grouped into three macro-categories identifying respectively *Wet grasslands and scrubs* (grouping 31.431, 36.A and 36.3464 classes), *Dry grasslands* (grouping 36.42 and 36.436 classes) and *Outcrops* (grouping 61.A and 62.15 classes).

Table 7.14 shows classification legend used for this area, the corresponding level-1 and Corine Biotopes codes and the number of check samples.

Corine Biotope codes	***Level 2 classes***	***Level 1 classes***	***Training samples***
31.431 - Mountain [*Juniperus nana*] scrub	31.431 - Mountain [*Juniperus nana*] scrub	Wet grasslands and scrubs	29
36.38, 36.313, 36.3A, 37.816 - Oro-Apennine closed grasslands, Pyreneo-Alpine hygrophile foxtail swards, sinkholes, Apennine tall herb communities	36.A - Oro-Apennine and Pyreneo-Alpine grasslands	Wet grasslands and scrubs	131
36.3464 - Alpine [*Juncus trifidus*] swards	36.3464 - Alpine [*Juncus trifidus*] swards	Wet grasslands and scrubs	61
36.414, 36.424 - Violet fescue swards and related communities, Apennine naked-rush swards	36.4A - Apennine naked-rush swards and Violet fescue	Dry grasslands	62
36.436 - Apennine stripped grasslands	36.436 - Apennine stripped grasslands	Dry grasslands	58
61.22, 61.3B1, 62.31, 36.12211 - Alpine pennycress screes, Central Mediterranean calcareous screes, Pavements, rock slabs, rock domes, Alpide [*Salix retusa-reticulata*] snowbed communities	61.A - Screes	Outcrops	139
62.15 - Alpine and sub-mediterranean cinquefoil cliffs	62.15 - Alpine and sub-mediterranean cinquefoil cliffs	Outcrops	45

Table 7.15 - Classification legend for Campo Pericoli area

7.3.3 Classifier models performance

Independently validated model classifiers were generated in each level of classification for every possible combination of partition method (n = 2), partition ratio (n = 9), X pre-processing transformation (n = 24) and number of LV (n = 9) and produced a total of 3888 suitable classification results.

In order to determine the optimum parameters for the best model classifiers, the validated classification results were averaged over common pre-processing and number of LV to produce a total of 216 parameters combinations results.

7.3.3.1 Macro-categories

In *level-1* classification test about the 75% (2905 of 3888) of the total classifier models show a prediction ability greater than 70%. Figure 7.19 shows the distribution of classification performance values for each of the tested partition methods (a) and partition ratios (b).

Results obtained with Random Selection partition method are generally higher than results obtained with Kennard-Stone (fig 7.19a) with a total of 1673 *vs.* 1232 models with a classification performance higher than 70%.

Considering different test ratios, models which were validated using smaller test subsets (< 30%) have generally lower results, although the smallest test ratio (10%) produces a lower number of outliers.

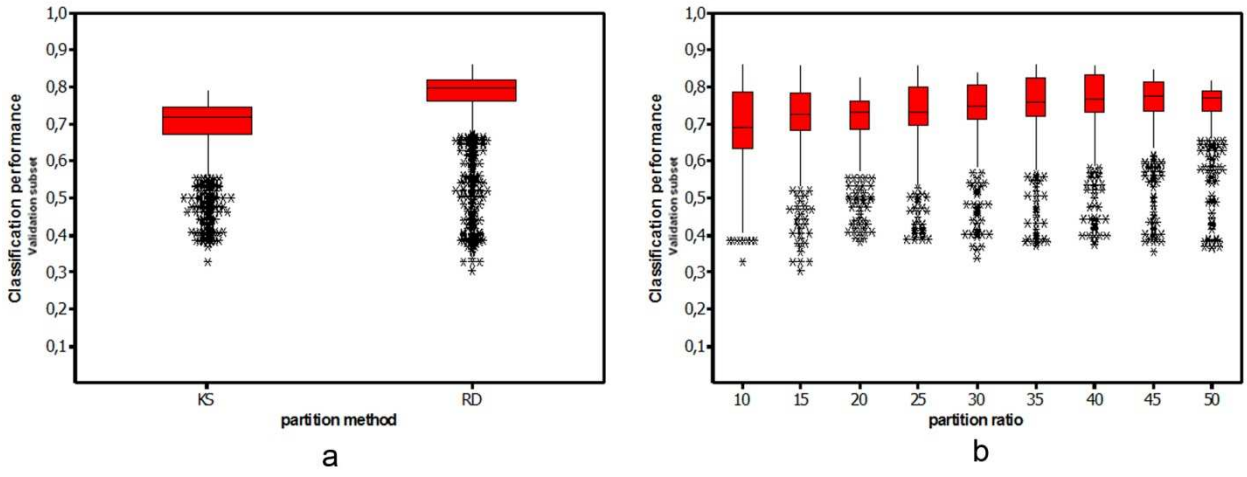


Figure 7.23 - Distribution of classification performance results considering the different partition methods (a) and partition ratios (b) used in the classification

The best classification performance over all the partition methods and partition ratios was obtained using a *Detrend* pre-processing with six latent variables. These parameters are then used to build the model to be used to classify the entire image.

Model parameters and results, both for calibration and validation phases, are presented in Table 7.15.

LV	6
Pre-processing X-Block	Detrend
Mean of % classification ability (validation subset)	79.40
Mean of % classification ability (calibration subset)	81.50

Table 7.16 - Model parameters and results for level-1 classification

Heat map represented in figure 7.20 shows the importance of each variable involved in building the PLS-DA model. Elevation is the most discriminating factor for all the three macro-categories; this can be due to the habitat structure of the basin with wet grassland generally arising below the dry swards and with outcrops which group habitats predominantly occurring at higher altitude areas. Band 4 (NIR) allows the discrimination of *wet grasslands and scrubs* macrocategory; indeed the NIR band is sensitive to organic matter, and it is thus realistic to expect the selection of this variable for the discrimination of wet grasslands.

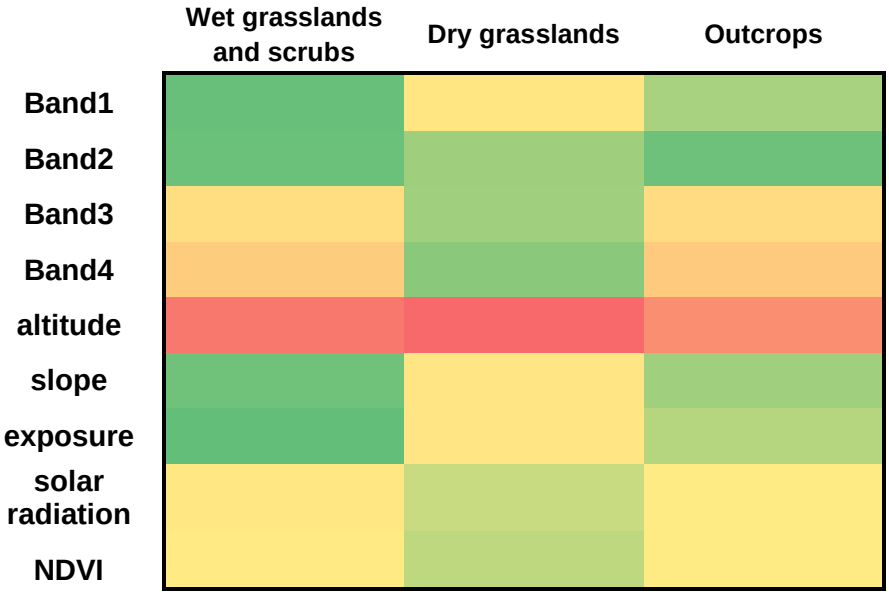


Figure 7.24 - Heat map of variable importance in building level-1 PLS-DA model

7.3.3.2 Habitat classes

Level-2 classifications were performed within each macro-category identified in the level-1. The models built within each class group generally show a good classification ability in all the three habitat groups. The parameters with the best classification results over all the partition method and partition ratios used to build the validation subsets are shown in fig 7.16.

	Wet grasslands and scrubs	**Dry grasslands**	**Outcrops**
LV	2	1	6
Pre-processing X-Block	GLS Weighting	smooth	msc (median)
Mean of % classification ability (validation subset)	81.90	78.05	89.74
Mean of % classification ability (calibration subset)	84.64	84.36	89.02

Table 7.17 - Model parameters and results for level-2 classifications

Figure 7.21 shows the distributions of classification performance values considering respectively the different partition methods and the partition ratios used in all the three classification tests to build the validation subsets.

Considering the partition methods RD produces better performing results than KS both in *Wet grasslands and scrubs* and *Dry grasslands* groups, while no predominance can be observed in *Outcrops* macro-category. Considering the different partition size no predominance can be observed for *Wet grassland and scrubs* and *Outcrops* groups while in the *Dry grasslands* macro-category, models obtained with smaller validation subsets (< 25% of the total dataset) show lower classification performance, although their distribution has a lower number of outliers.

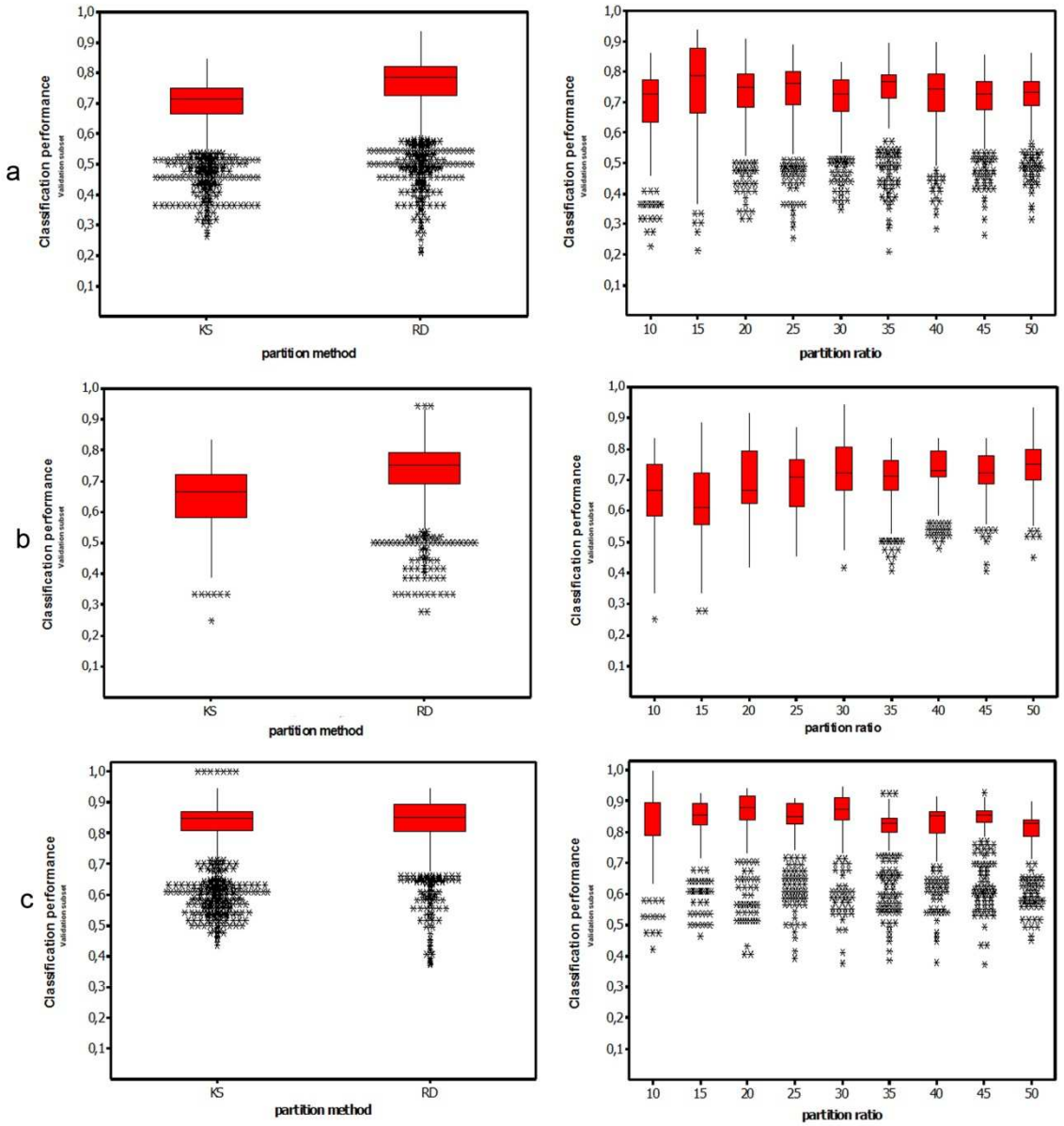


Figure 7.25 - Distributions of classification performance results considering the different partition methods and partition ratios for Wet grasslands and scrubs (a), Dry grasslands (b) and Outcrops (c) groups

Heat maps for all PLS-DA models utilized in level-2 classifications were computed (fig 7.22). In the first habitat group (a) 'band 1' appear to be the less important variable to build the model; on the contrary, slope is crucial to discriminate 36.A and 36.3464 classes. Besides that, topographic variables 'slope' and 'elevation' are also important in the discrimination of habitats belonging to *Dry grasslands* (b) and *Outcrops* (c) groups.

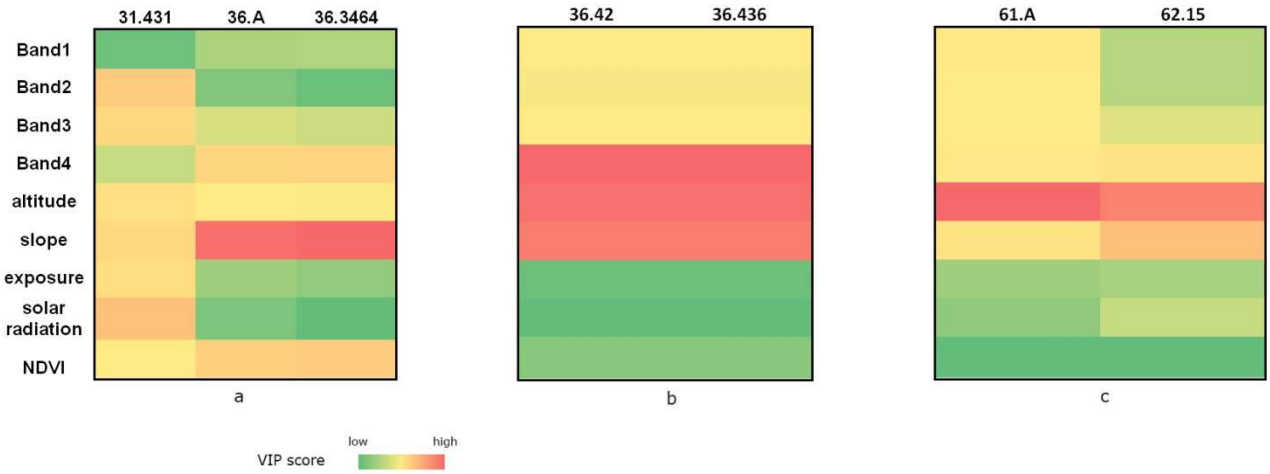


Figure 7.26 - Heat map of variable importance in building PLS-DA model for Wet grasslands and scrubs (a), Dry grasslands (b) and Outcrops (c) groups

7.3.4 Accuracy assessment

Model parameters were applied to the entire training datasets and the resulted classifiers were used to obtain the final classified map (at 1:10.000 scale, minimum mapping unit = 400 m^2) and the map of Non-Classified zones (annex III of the thesis).

Accuracy assessment was performed using a map produced during the Ph.D. activity in the frame of the "*Carta della Natura*" project (ISPRA,2013b). Table 7.17 shows classification accuracy within each macrocategory. Results show an high classification ability in the first step of the classification procedure and are comparable with those obtained by model's results.

	Wet grasslands and scrubs	**Dry grasslands**	**Outcrops**
User's accuracy (A%)	76,66	70,78	86,02
User's accuracy (P%)	81,08	73,80	89,43

Table 7.18 - User's classification accuracy calculated within each macro-category

Confusion matrices comparing the classified map *vs Carta della Natura* map are shown in figures 7.23 and 7.24.

	31.431	36.A	36.3464	36.4A	36.436	61.A	62.15	user's accuracy
as 31.431	29,446	35,676	19,668	27,323	21,921	28,306	7,259	**17.36%**
as 36.A	5,985	599,651	22,188	57,338	38,219	14,768	669	**81.16%**
as 36.3464	8,355	31,440	148,895	17,583	13,279	38,707	9,060	**55.70%**
as 36.4A		20,586	14,609	168,658	37,532	16,151	69	**65.47%**
as 36.436	7,259	98,541	15,579	45,616	197,135	9,754	2,801	**52.33%**
as 61.A	10,548	10,491	16,741	32,543	44,516	712,241	42,485	**81.91%**
as 62.15	5,910	2,529	12,854	2,799	2,404	29,903	85,090	**60.14%**
producer's accuracy	**43.62%**	**75.06%**	**59.43%**	**47.93%**	**55.53%**	**83.81%**	**57.71%**	

Overall accuracy	**68.81%**
kappa coefficent	**0.61**

Figure 7.27 - Confusion matrix (m^2 correctly mapped) for Campo Pericoli area and classification accuracy indexes

	31.431	36.A	36.3464	36.4A	36.436	61.A	62.15	user's accuracy
as 31.431	11	11	9	9	7	11	3	**18.03%**
as 36.A	1	219	8	15	9	3		**85.88%**
as 36.3464	3	12	56	4	4	11	1	**61.54%**
as 36.4A		7	6	63	14	8		**64.29%**
as 36.436	2	30	4	18	74	2	1	**56.49%**
as 61.A	4	1	3	10	10	250	11	**86.51%**
as 62.15			6		1	9	26	**61.90%**
producer's accuracy	**52.38%**	**78.21%**	**60.87%**	**52.94%**	**62.18%**	**85.03%**	**61,90%**	

Overall accuracy	**72.29%**
kappa coefficent	**0.65**

Figure 7.28 - Confusion matrix (number of reference ground points correctly mapped) for Campo Pericoli area and classification accuracy indexes

Table 7.18 shows the overall classification accuracy indexes for the proposed classification methods and for comparison tests.

	PLS-DA	**Maximum likelihood** stepwise classification	**Maximum likelihood**
A%	**68.81**	**63.32**	**61.42**
P%	**72.29**	**64.84**	**64.53**

Table 7.19 - classification accuracy indexes comparison

Overall classification accuracy is 68.81% considering all mapped areas and 72.29% using the validation grid. In this case, classification gives good results with screes habitats (61.A) and with Oro-Apennine closed grasslands (36.A). These results are very encouraging as habitat belonging to these classes were identified in Annex 1 of the EU Habitats Directive as being of Community interest, and in particular 36.38 habitat, which is predominant in 36.A class, is listed as ‘priority’.

The lowest value in classification accuracy is obtained with habitat codes 31.431; this can be due to habitat’s scrub-structure which makes it difficultly recognised by an automated system.

Chapter 8

Conclusion

Identification, description, classification and mapping of natural and semi-natural habitats are gaining recognition in the sphere of environmental policy implementation and frequently find applications in land planning and management as well as nature protection measures. Although visual interpretation of optical aerial photography with field surveys remains the more accurate approach to produce detailed habitat and vegetations maps, remote sensing data sets from space or air-borne multi/hyperspectral sensors are increasingly being considered by EU Member States in order to fulfil their reporting obligations under the Habitats Directive (Lengyel et al., 2008). However, although the potential of remote sensing techniques has been more clearly demonstrated for mapping the land cover categories, its use for accurate, detailed and complete conservation status assessment and monitoring of natural and semi-natural habitats, as required under the Habitats Directive, is still rare (Vanden Borre et al., 2011).

In this study, a new pixel-based classification method was tested by combining PLS-DA classifier with GIS and remote sensing procedures in order to classify natural habitats using multispectral images and ancillary data.

The method was tested under different conditions in order to verify its suitability to be used both at middle-range (eg 1:50.000) and at a more detailed mapping scale (e.g. 1:10.000); three study areas with different habitat composition were used for the classification: i) Monte Vulture volcanic complex (1:50.000), ii) Apulia lagoons (1:50.000) and iii) Campo Pericoli basin (1:10.000) and the respective classification accuracy was calculated using official maps produced in the frame of the italian *Carta della Natura* project.

Moreover, in order to evaluate the potential use of the proposed method in the common classification analysis, a comparison test was performed by evaluating the resulting classified images with respect to those obtained by using a commercial classification software (ESRI ArcGIS, rel 10.1), using the same input data. In order to stress the relative importance of the two main components of the classification method here presented (PLS-DA recursive algorithm and 2-level stepwise classification), the comparison with the performances of the commercial classifier has been carried out over two classification schemes implemented within ArcGIS: firstly, the “raw” classification as available in the software tutorial; secondly, a stepwise classification on the same two levels as in this method, but using the embedded maximum likelihood algorithm in the software.

Results show a better prediction ability of the proposed method in all the three areas considering both comparison tests. Overall accuracies were: 55.7% for Monte Vulture volcanic complex, 62.8% for Apulia Lagoons and 72.3% for Campo Pericoli basin. These results, although not very high in absolute terms, can be considered as satisfactory because of the particular context of the study areas, characterized by the complexity and the heterogeneity of their habitats. In particular the methods shows a good accuracy with the area mapped at the highest scale (Campo Pericoli); these results are very encouraging as habitat belonging to these classes were identified in Annex 1 of the EU Habitats Directive as being of Community interest, and in particular 36.38 habitat, which is predominant in 36.A class, is listed as 'priority'.

Several benefits brought about by this approach can be highlighted:

- ✓ PLS-DA multivariate technique is a well suited method to analyze and classify remote sensing data in general and for discriminate habitat classes in particular. It permits to reduce data noise and to avoid data collinearity when it is necessary to use ancillary data in addition to the classical information available from image bands, in order to better discriminate habitat patches. This is particularly important in the case of detailed habitat classification legends such as the *Corine Biotopes*, which often includes "interdisciplinary" descriptive information in habitats descriptions.

  Moreover, the proposed recursive algorithm is a non-selective approach which results in a rapid and automated data-mining strategy for exploratory analysis, which allows to identify the most robust classifier model to be used in order to obtain the best possible classification accuracy.
- ✓ the stepwise approach of classification permits to obtain a more detailed result by reducing the number of classes to be examined. Moreover, it permits to better identify possible weak points in the classification procedure, by isolating that classes or macrocategories that show a worse classification ability.
- ✓ data used in the classification (Ortophoto/RapidEye images and Digital Elevation Models) are available at reasonable cost and with limited technical constraints, demonstrating the ease of implementation of the method.

Although manual interpretation of aerial photography remains the more accurate approach this method can be used as a starting point for the further steps of photo-interpretation, thus allowing to reduce the amount of time spent in the visual interpretation and improving consistency of the final products when large areas need to be mapped by more authors

8.1 Recommendations for further work

In order to achieve better results in terms of discrimination of habitat classes, several strategies are possible. Corbane et al. (2013) evaluate the possibility to use PLS-DA classification in an object-oriented approach, using the multivariate technique as a following step afterobjects' recognition (segmentation); in this way it is possible to insert in the classification training data also features of different sources such as spectral values, texture, shape, context relationships, along with the thematic or continuous information supplied by objects obtained from the first segmentation phase.

Other data sources could also be tested and included in the PLS-DA analysis. Above all, the use of existing data such as urban or agricultural maps could be used to exclude such habitats from the classification, in order to have a better classification ability.

Multitemporal remote-sensing data can also be very useful for discriminating habitats basing on their seasonal features; spectral responses from images taken on dates where two species are at different phenological stages allow species to be distinguished and bring about additional information in habitat classification (Lucas et al., 2007 and 2011).

References

http://wiki.eigenvector.com/index.php?title=Using_Cross-Validation – visited March 20th 2014.

Afendi F.M., Ono N., Nakamura Y., Nakamura K., Darusman L.K., Kibinge N., Morita AH, Tanaka K., Horai H., Altaf-Ul-Amin Md., Kanaya S., 2013. Data Mining Methods for Omics and Knowledge of Crude Medicinal Plants toward Big Data Biology. *Computational and Structural Biotechnology Journal* 4.

Antonucci F., Menesatti P., Holden N. M., Canali E., Giorgi S., Maienza A., Stazi S.R., 2012. Hyperspectral Visible and Near-Infrared Determination of Copper Concentration in Agricultural Polluted Soils. *Communications in Soil Science and Plant Analysis* 43(10), pp. 1401-1411.

Artiola J.F., Pepper I.L., Brusseau M.L. (Eds.). 2004. Environmental monitoring and characterization. Academic Press.

Baatz M., Benz U., Dehghani S., Heynen M., Höltje A., Hofmann P., Lingenfelder I., Mimler M., Sohlbach M., Weber M., Willhauck G., 2004. eCognition Professional: User guide 4. Munich: Definiens-Imaging.

Barker M., Rayens W., 2003. Partial least squares for discrimination. *Journal of chemometrics* 17(3), pp. 166-173.

Ballabio D., Todeschini R., 2009. Multivariate classification for qualitative analysis. *Infrared spectroscopy for food quality analysis and control*. Ed. D. Sun (Academic Press: Burlington, MA) pp. 83-104.

Bezdek J.C., 1981. Pattern Recognition with Fuzzy Objective Function Algorithms, Plenum, NY.

Biondi E., 1999. Ricerche di Geobotanica ed Ecologia vegetale di Campo Imperatore (Gran Sasso d'Italia).

Blaschke T., Strobl J., 2001. What's wrong with pixels? Some recent developments interfacing remote sensing and GIS. *GeoBIT/GIS* 6(01), pp. 12-17.

Bock M., 2003. Remote Sensing and GIS-Based Techniques for the Classification and Monitoring of Biotopes: Case Examples for a Wet Grass- and Moor Land Area in Northern Germany. *Journal for Nature Conservation* 11, pp. 145–155.

Bouveresse E., Hartmann C., Massart D.L., Last I.R., Prebble K.A., 1996. Standardization of near-infrared spectrometric instruments. *Analytical Chemistry* 68(6), pp. 982-990.

Bunce R.G.H., Bogers M.M.B., Evans D., Halada L., Jongman R.H.G., Mucher C.A., Bauch B., De Blust G., Parr T.W., Olsvig-Whittaker L., 2013. The significance of habitats as indicators of biodiversity and their links to species. *Ecological Indicators* 33, pp. 19-25.

Bylesjo M, Rantalainen M, Cloarec O, Nicholson JK, Holmes E & Trygg J (2006) OPLS discriminant analysis: combining the strengths of PLS-DA and SIMCA classification. *Journal of Chemometrics* 20, pp. 341-351.

Capoccioni F., 2013. Biological features of eel (*Anguilla anguilla*, L. 1758) local stocks in the mediterranean area, as a function of different ecological conditions. PhD thesis. Università degli Studi di Tor Vergata.

Capoccioni F., Costa C., Aguzzi J., Menesatti P., Lombarte A., Ciccotti E., 2011. Ontogenetic and environmental effects on otolith shape variability in three Mediterranean European eel (*Anguilla anguilla*, L.) local stocks. *Journal of experimental marine biology and ecology*, 397(1), pp. 1-7.

Capoccioni F., Lin D., Iizuka Y., Tzeng W.N., Ciccotti E., 2014. Phenotypic plasticity in habitat use and growth of the European eel (*Anguilla anguilla*) in transitional waters in the Mediterranean area. *Ecology of Freshwater Fish* 23, pp. 65–76.

Casals-Carrasco P., Kubo S., Babu Madhavan B., 2000. Application of spectral mixture analysis for terrain evaluation studies. *International Journal of Remote Sensing* 21(16), pp.3039-3055.

Chin W.W., Newsted P.R., 1999. Structural equation modelling analysis with small samples using partial least squares. In: R.H. Hoyle (Ed.), *Statistical strategies for small sample research*, 1999. pp.307–341. Thousand Oaks, CA: Sage.

Chong I.G., Jun C.H., 2005. Performance of some variable selection methods when multicollinearity is present. *Chemometrics and Intelligent Laboratory Systems*, 78(1), pp. 103-112.

Chung D., Keles S., 2010. Sparse Partial Least Squares Classification for High Dimensional Data. *Statistical Applications in Genetics and Molecular Biology* 9, pp. 1–30.

Cohen J., 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement* 20 (1), pp. 37–46.

Commission of the European Communities, 1991. CORINE biotopes, The design, compilation and use of an inventory of sites of major importance for nature conservation in the European Community.

Compagnoni B., Damiani A.V., Valletta M., Finetti I., Cirese E., Pannuti S., Sorrentino F., Rigano C., 1976-1984. Carta Geologica d'Italia alla scala 1:500.000 (5 fogli). Servizio Geologico d'Italia, Roma.

Congalton R.G., 1991. A review of assessing the accuracy of classifications of remotely sensed data. *Remote sensing of environment* 37(1), pp. 35-46.

Congalton R.G., Oderwald R.G., Mead R.A., 1983. Assessing Landsat classification accuracy using discrete multivariate analysis statistical techniques. *Photogrammetric Engineering and Remote Sensing*.

Corbane C., Alleaume S., Deshayes M., 2013. Mapping natural habitats using remote sensing and Sparse partial least square discriminant analysis. *International Journal of Remote Sensing* 34(21), pp. 7625 – 7647.

Costa C, Antonucci F, Pallottino F, Aguzzi J, Sun D.W., Menesatti P., 2011. Shape analysis of agricultural products: a review of recent research advances and potential application to computer vision. *Food and Bioprocess Technology* 4, pp. 673-692.

D'alessandro L., De Sisti G., D'Orefice M., Pecci M., Ventura R., 2003. Geomorphology of the summit area of the Gran Sasso d'Italia (Abruzzo, Italy). *Geografia Fisica e Dinamica Quaternaria* 26, pp. 125-141.

Dale L.M., Thewis A., Boudry C., Rotar I., Păcurar F.S., Abbas O., Dardenne P., Baeten V., Pfister J., Fernández Pierna J.A., 2013. Discrimination of grassland species and their classification in botanical families by laboratory scale NIR hyperspectral imaging: Preliminary results. *Talanta*, 116, pp. 149-154.

Daszykowski M., Walczak B., Massart D.L., 2002. Representative subset selection. *Analytica Chimica Acta* 468(1), pp. 91-103.

Devillers P., Devillers - Terschuren J., 1996, A classification of Palaearctic habitats, *Nature and environment*, 78, pp. 194.

Devillers P., Devillers - Terschuren J., Ledant J.P., 1991, Corine biotopes manual. Vol. 2. Habitats of the European Community, Office for Official Publications of the European Communities, Luxembourg.

Dobos E, Micheli E, Baumgardner M.F., 2000. Use of combined digital elevation model and satellite radiometric data for regional soil mapping. *Geoderma* 97, pp. 367- 391.

Du Q., Chang C.I., 2004. Linear mixture analysis-based compression for hyperspectral image analysis. *Geoscience and Remote Sensing, IEEE Transactions on* 42(4), pp. 875-891.

Duda R.O., Hart P.E., Stork D.G., 2001. Pattern classification. John Wiley, Section, 10, l, NY.

Dunn J.C., 1974. A Fuzzy Relative of the ISODATA Process and its Use in Detecting Compact, Well Separated Clusters.

Dymond C., Johnson E., 2002. Mapping vegetation spatial patterns from modeled water, temperature and solar radiation gradients. *ISPRS Journal of Photogrammetry and Remote Sensing* 57(1–2), pp. 69–85.

Elith J, Burgman M.A., Regan H.M., 2002. Mapping epistemic uncertainties and vague concepts in predictions of species distributions. *Ecological Modelling* 157, pp. 313–329.

Enderle D.I., Weih Jr, R.C., 2005. Integrating supervised and unsupervised classification methods to develop a more accurate land cover classification. *Journal of the Arkansas Academy of Science* 59, pp. 65-73.

ENEA, 2002. ENEA climate archive. Available at http://clisun.casaccia.enea.it/Pagine/Index.htm (visited 29th of April 2014)

Eriksson L., 2006. Multi-and megavariate data analysis. MKS Umetrics AB.

European Commission DG Environment, 2007. Interpretation Manual of European Union Habitats – EUR 27. Brussels: European Commission, DG Environment

European Commission, 2009. Composite Report on the Conservation Status of Habitat Types and Species as required under Article 17 of the Habitats Directive.

European Union, 1992. Council Directive 92/43/EEC of 21 May 1992 on the conservation of natural habitats and of wild fauna and flora. Brussels, Belgium.

European Union, 1997. Directive 2007/2/EC of the European Parliament and of the Council of 14 March 2007 establishing an Infrastructure for Spatial Information in the European Community (INSPIRE).

Evans D., 2010. Interpreting the habitats of Annex I: past, present and future. *Acta Botanica Gallica* 157(4), pp. 677-686.

Excelisvis, 2013. ENVI classic tutorial: classification methods – available at http://www.exelisvis.com/portals/0/pdfs/envi/Classification_Methods.pdf

Fitzgerald R.W., Lees B.G., 1994. Assessing the classification accuracy of multi-source remote sensing data. *Remote sensing of environment* 47, pp. 362–368.

Fu P., Rich P.M., 2002. A Geometric Solar Radiation Model with Applications in Agriculture and forestry. *Computers and Electronics in Agriculture* 37, pp. 25–35.

Galvão R.K.H., Araujo M.C.U., José G.E., Pontes M.J.C., Silva E.C., Saldanha T.C.B., 2005. A method for calibration and validation subset partitioning. *Talanta* 67(4), pp. 736-740.

Gastellu-Etchegorry J.P., Estregull C., Mougin E.A., 1993. GIS based methodology for small scale monitoring of tropical forests - a case study in Sumatra. *International Journal of Remote Sensing* 14(12), pp. 2349-2368.

Gong P., Howarth P., 1990. An assessment of some factors influencing multispectral land-cover classification. *Photogrammetric Engineering and Remote Sensing* 56(5), pp. 597-603.

Gopal S, Woodcock C., 1994. Theory and methods for accuracy assessment of thematic maps using fuzzy sets. *Photogrammetric Engineering and Remote Sensing* 60(2), pp. 181–188.

Gould W., 2000. Remote sensing of vegetation, plant species richness, and regional biodiversity hotspots. *Ecological Applications* 10(6), pp. 1861–1870.

Green E.J., Strawderman W.E. 1994. Determining accuracy of thematic maps. *The Statistician* 43(1), pp. 77–85.

Hammond T.O., Verbyla D.L., 1996. Optimistic bias in classification accuracy assessment. *International Journal of Remote Sensing* 17(6), pp. 1261–1266.

Hill M.O., Moss D., Davies C.E., 2004. Revision of habitat descriptions originating from Devillers et al. (2001). European Topic Centre on Nature Protection and Biodiversity, Paris.

Hodgson M.E., 1988. Reducing the computational requirements of the minimum-distance classifier. *Remote Sensing of Environment* 24.

Hutchinson M.F., 1996. A locally adaptive approach to the interpolation of digital elevation models. *Third International Conference/Workshop on Integrating GIS and Environmental Modeling*. NCGIA, University of California, Santa Barbara.

Hutchinson M.F., Xu T., Stein J.A. 2011. Recent Progress in the ANUDEM Elevation Gridding Procedure. *Geomorphometry 2011*, pp. 19-22.

ISPRA, 2003. Il Progetto Carta della Natura alla scala 1: 250.000. Metodologia di realizzazione. *APAT, Manuali e linee guida* 17/2003, Rome.

ISPRA, 2009c. Dati del sistema informativo di Carta della Natura – scala 1:50.000 - Regione Puglia. Available at http://www.geoviewer.isprambiente.it/

ISPRA, 2009a. Il progetto Carta della Natura. Linee guida per la cartografia e la valutazione degli habitat alla scala 1:50.000, *ISPRA Manuali e Linee Guida* 48/2009, Rome.

ISPRA, 2009d. Archivio dei rilievi degli habitat del sistema informativo di Carta della Natura – scala 1:50.000 - Regione Puglia.

ISPRA, 2009b. Gli habitat in Carta della Natura. *ISPRA Manuali e Linee Guida* n. 49/2009, Rome.

ISPRA, 2012b. Archivio dei rilievi degli habitat del sistema informativo di Carta della Natura – scala 1:50.000 - Regione Basilicata.

ISPRA, 2012a. Dati del sistema informativo di Carta della Natura – scala 1:50.000 - Regione Basilicata. Available at http://www.geoviewer.isprambiente.it/

ISPRA, 2013a. Carta delle Unità Fisiografiche di Paesaggio d'Italia (scala 1:250.000). Available at http://www.geoviewer.isprambiente.it/

ISPRA, 2013b. Dati del sistema informativo di Carta della Natura – scala 1:10.000 – Area di Campo Pericoli (Abruzzo)

Jensen J.R., 2005. Introductory digital image processing: a remote sensing perspective - 3rd ed. Prentice Hall. United States of America.

Jung H.W., 2003. Evaluating interrater agreement in SPICE-based assessments. *Computer Standards & Interfaces* 25(5), pp. 477-499.

Kaiser P.K., Boynton R.M., 1996. Human color vision.

Kalyankar N.V., 2013. Major limitations of satellite images. *Journal of Global Research in Computer Science* 4(5), pp. 51-59.

Karem F., Dhibi M., Martin A., 2012. Combination of supervised and unsupervised classification using the theory of belief functions. In: *Belief Functions: Theory and Applications*, pp. 85-92. Springer Berlin Heidelberg.

Kennard R.W., Stone L.A., 1969. Computer aided design of experiments. *Technometrics* 11(1), pp. 137-148

Keramitsoglou I., Kontoes C., Sifakis N, Mitchley J., Xofis P., 2005. Kernel Based Re-Classification of Earth Observation Data for Fine Scale Habitat Mapping. *Journal for Nature Conservation* 13, pp. 91–99.

Kerr J.T., Ostrovsky M., 2003. From Space to Species: Ecological Applications for Remote Sensing. *Trends in Ecology & Evolution* 18, pp. 299–305.

Krämer N., Sugiyama M., 2011. The degrees of freedom of partial least squares regression. *Journal of the American Statistical Association* 106 (494).

Landis J.R., Koch G.G., 1977. The measurement of observer agreement for categorical data. *Biometrics* 33, pp. 159–174.

Langley S.K., Cheshire H.M., Humes K.S., 2001. A comparison of single date and multitemporal satellite image classifications in a semi-arid grassland. *Journal of Arid Environments* 49(2), pp. 401-411.

Lê Cao K.A., Martin P. G. P., Robert-Granié C., Besse P., 2009. Sparse Canonical Methods for Biological Data Integration: Application to a Cross-Platform Study. *BMC Bioinformatics* 10(1) pp. 34.

Lechner A.M., Langford W.T., Bekessy S.A., Jones S.D., 2012. Are landscape ecologists addressing uncertainty in their remote sensing data? *Landscape ecology* 27(9), pp. 1249-1261.

Lengyel S., Déri E., Varga Z., Horváth R., Tóthmérész B., Henry P.Y., Kobler A., Kutnar L., Babij V., Seliskar A., Christia C., Papastergiadou E., Gruber B., Henle K., 2008. Habitat monitoring in Europe: a description of current practices. *Biodiversity and Conservation* 17(14), pp. 3327-3339.

Lieckfeld L., Oldeland J., Weber B., Schultz C., Müller A., Schmidt M., 2006. Use of hyperspectral data to assess the effects of different land use strategies on vegetation types in savannahs in Central Namibia. *ESA Workshop Proceedings of the fourth CHRIS/Proba Workshop, September 19-21, 2006, Esrin Frascati, Italy.*

Lillesand T., Kiefer R.W., Chipman J., 2008. Remote Sensing and Image Interpretation. 6th Edition, John Wiley & Sons, Inc., New York.

Löfvenhaft K., Björn C., Ihse M., 2002. Biotope patterns in urban areas: A conceptual model integrating biodiversity issues in spatial planning. *Landscape and Urban Planning* 58, pp. 223-240.

Lopez N., 2003. Caratterizzazione geologica e geomorfologia del Gargano. In: *Caratterizzazione agroecologica del Gargano*, a cura di Flagella Z. & Tarantino E., Università degli Studi di Foggia, Claudio Grenzi Editore

Lucas R., Medcalf K., Brown A., Bunting P., Breyer J., Clewley D., Keyworth S., Blackmore P., 2011. Updating the Phase 1 Habitat Map of Wales, UK, using Satellite Sensor Data. *ISPRS Journal of Photogrammetry and Remote Sensing* 66. pp 81–102.

Lucas R., Rowlands A., Brown A., Keyworth S., Bunting P., 2007. Rule-based classification of multitemporal satellite imagery for habitat and agricultural land cover mapping. *ISPRS Journal of photogrammetry and remote sensing* 62(3), pp. 165-185.

Manzo C., 2010. Fish assemblages in three Mediterranean coastal lagoons: structure, functioning and spatio-temporal dynamics. PhD thesis. Università degli Studi di Tor Vergata. Available at: http://hdl.handle.net/2108/1274.

Markowska-Kaczmar U., Switek T., 2009. Combined unsupervised-supervised classification method. In: *Knowledge-Based and Intelligent Information and Engineering Systems*, pp. 861-868. Springer Berlin Heidelberg.

Maurer U., Peschel T., Schmitz S., 2000. The flora of selected urban land-use types in Berlin and Potsdam with regard to nature conservation in cities. *Landscape and Urban Planning* 46, pp. 209-215.

McIver D.K., Friedl M.A., 2002. Using prior probabilities in decision-tree classification of remotely sensed data. *Remote Sensing of Environment* 81(2–3), pp. 253–261.

Mead R.A., Szajgin J., 1981. Landsat Classification Accuracy Assessment Procedures: An Account of a National Working Conference.

Meliadis I., Meliadis M., 2011. Multi-temporal Landsat image classification and change analysis of land cover/use in the Prefecture of Thessaloiniki, Greece. *Proceedings of the International Academy of Ecology and Environmental Sciences* 1(1), pp. 15-25.

Menesatti P., Costa C., Paglia G., Pallottino F., D'Andrea S., Rimatori V., Aguzzi J., 2008. Shape-based methodology for multivariate discrimination among Italian hazelnut cultivars. *Biosystem Engineering* 101(4), pp. 417-424.

Mengistu D.A., Salami A.T., 2007. Application of remote sensing and GIS inland use/land cover mapping and change detection in a part of south western Nigeria. *African Journal of Environmental Science and Technology* 1(5), pp. 99-109.

Miller C.J., 2000. Vegetation and habitat are not synonyms. *Ecological Management and Restoration* 1, pp. 102-104.

Miller J.R., Hobbs R.J., 2007. Habitat restoration: Do we know what we're doing? *Restoration Ecology* 15, pp. 382-390.

Milliken J., Beardsley D., Gill S., 1998. Accuracy assessment of a vegetation map of north-eastern California using permanent plots and fuzzy sets. United States Department of Agriculture Forest Service.

Morrison M.L., 2001. Introduction: Concepts of wildlife and wildlife habitat for ecological restoration. *Ecology* 9, pp. 251-252.

Nagendra H., 2001. Using Remote Sensing to Assess Biodiversity. *International Journal of Remote Sensing* 22. Pp. 2377–2400.

Nagendra H., Lucas R., Honrado J.P., Jongman R.H.G., Tarantino C., Adamo M., Mairota P., 2013. Remote Sensing for Conservation Monitoring: Assessing Protected Areas, Habitat Extent, Habitat Condition, Species Diversity, and Threats. *Ecological Indicators* 33, pp. 45–59.

Nimis P.L., Martellos S., 2008. *ITALIC* - The Information System on Italian Lichens. Version 4.0. University of Trieste, Dept. of Biology, IN4.0/1 (http://dbiodbs.univ.trieste.it/)

Nonnis Marzano C., Scalera C., Liaci L., Fianchini A., Gravina F., Mercurio M., Corriero G., 2003. Distribution, persistence and change in the macrobenthos of the lagoon of Lesina (Apulia, southern Adriatic Sea). *Acta Adriatica* 26, pp. 57–66.

Nordberg M.L., Evertson J., 2003. Vegetation index differencing and linear regression for change detection in a Swedish mountain range using Landsat TM and ETM+ imagery. *Land Degradation & Development* 16, pp. 139–149.

Odum E.P., 1971. Fundamentals of Ecology, 3rd ed. W.B. Saunders Company.

Ojigi L.M., 2006. Analysis of spatial variations of Abuja land use and land cover from image classification algorithms. In: *Symposium Remote Sensing: From Pixel to Processes, Enschede, Netherlands*, p. 6.

Peddle D.R., Peter White H., Soffer R.J., Miller J.R., LeDrew E.F., 2001. Reflectance processing of remote sensing spectroradiometer data. *Computers & Geosciences* 27(2), pp. 203-213.

Pedroni L., 2003. Improved classification of Landsat Thematic Mapper data using modified prior probabilities in large and complex landscapes. *International Journal of Remote Sensing* 24(1), pp. 91-113.

Penna B., 2005. Design and implementation of prediction and transform-based techniques for progressive lossy and lossless hyperspectral image compression. PhD thesis, Politecnico Di Torino.

Prakash A., 2000. Thermal remote sensing: concepts, issues and applications. *International Archives of Photogrammetry and Remote Sensing*, 33 (B1; Part 1), pp. 239-243.

Qian S.E., 2004. Hyperspectral data compression using a fast vector quantization algorithm. *Geoscience and Remote Sensing, IEEE Transactions on*, 42(8), pp. 1791-1798.

Rahman A., Kumar S., Fazal S., Siddiqui M.A., 2012. Assessment of land use/land cover change in the North-West District of Delhi using remote sensing and GIS techniques. *Journal of the Indian Society of Remote Sensing* 40(4), pp. 689-697.

Rajer-Kanduč K., Zupan J., Majcen N., 2003. Separation of data on the training and test set for modelling: a case study for modelling of five colour properties of a white pigment. *Chemometrics and intelligent laboratory systems* 65(2), pp. 221-229.

Rasmussen K., Olesen H.H., 1988. Applications of multivariate statistical analysis in remote sensing of agriculture. *Geografisk Tidsskrift-Danish Journal of Geography* 88(1), pp. 100-107.

Rebollo San Miguel E.P., 2011. Applications of imaging spectroscopy to the chemistry of cultural heritage field. PhD thesis, Università degli Studi di Padova,.

Recio J.A., Helmholz P., Müller S., 2011. Potential evaluation of different types of images and their combination for the classification of GIS objects cropland and grassland. *International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences* 38(4).

Regan H.M., Colyvan M., Burgman M.A., 2002. A taxonomy and treatment of uncertainty for ecology and conservation biology. *Ecological Applications* 12(2), pp. 618–628.

Repaka S.R., Truax D.D., Kolstad E., O'Hara C.G., 2004. Comparing spectral and object based approaches for classification and transportation feature extraction from high resolution multispectral imagery. In: *ASPRS Annual Conference Proceedings*, pp. 23-28.

Rich P.M., Dubayah R., Hetrick W.A., Saving S.C., 1994. Using Viewshed models to calculate intercepted solar radiation: applications in ecology. *American Society for Photogrammetry and Remote Sensing Technical Papers*, pp. 524-529.

Rich P.M., Fu P., 2000. Topoclimatic habitat models. *Proceedings of the Fourth International Conference on Integrating Geographic Information Systems (GIS) and Environmental Modeling, Banff, Alberta, Canada.*

Rimal B., 2011. Application of remote sensing and gis, land use/land cover change in Kathmandu metropolitan city, Nepal. *Journal of Theoretical & Applied Information Technology* 23(2).

Roelofsen H.D., Kooistra L., Van Bodegom P.M., Verrelst J., Krol J., Witte J.P.M., 2014. Mapping a priori defined plant associations using remotely sensed vegetation characteristics. *Remote Sensing of Environment* 140, pp. 639-651.

Roselli L., Fabbrocini A., Manzo C., D'Adamo R., 2009. Hydrological heterogeneity, nutrient dynamics and water quality of a non-tidal lentic ecosystem (Lesina Lagoon, Italy). *Estuarine, Coastal and Shelf Science* 84(4), pp. 539-552.

Rosenfield G.H., Fitzpatricklins-Lins K., 1986. A coefficient of agreement as a measure of thematic classification accuracy. *Photogrammetric Engineering and Remote Sensing* 52(2), pp. 223-227.

Rouse J.W., Haas R.H., Schell J.A., Deering D.W., 1974. Monitoring vegetation systems in the Great Plains with ERTS. In: Fraden S.C., Marcanti E.P., Becker M.A. (eds.), *Third ERTS-1 Symposium, 10–14 Dec. 1973*, NASA SP-351, Washington D.C., pp. 309–317.

Sabatier R., Vivein M., Amenta P., 2003. Two approaches for Discriminant Partial Least Square. In: Schader M, Gaul W & Vichi M (eds) Between data science and applied data analysis. Springer-Verlag, Berlin, Germany.

Salafsky N., Salzer D., Ervin J., Boucher T., Ostlie W., 2003. Conventions for defining, naming, measuring, combining, and mapping threats in conservation: an initial proposal for a standard system. *Conservation Measures Partnership*, Washington, DC.

Sales F., Rius A., Callao M.P., Rius F.X., 2000. Standardization of a multivariate calibration model applied to the determination of chromium in tanning sewage. *Talanta* 52(2), pp. 329-336.

Savitzky A., Golay M.J.E., 1964. *Smoothing and Differentiation of Data by Simplified Least Squares Procedures*. *Analytica Chemica Acta*, 36, pp. 1627-1639.

Schuster C., Förster M., Kleinschmit B., 2012. Testing the red edge channel for improving land-use classifications based on high-resolution multi-spectral satellite data. *International Journal of Remote Sensing* 33(17), pp. 5583-5599.

Shackelford A. K. Davis A.H., 2003. A combined fuzzy pixel-based and object-based approach for classification of high-resolution multispectral data over urban areas. *IEEE Transactions on Geoscience and Remote Sensing* 41, pp. 2354–2364.

Sivertsen D., 2009. Native vegetation interim type standard. Department of Environment, Climate Change and Water NSW, Sydney

Sjöström M., Wold S., Söderström B., 1986. PLS discriminant plots. In: *Pattern recognition in practice II*, pp. 461-470. Elsevier, Amsterdam.

Snee R.D. 1977. Validation of regression models: methods and examples. *Technometrics* 19(4), pp. 415-428.

Spagnoli F., Specchiulli A., Scirocco T., Carapella Spagnoli F., Specchiulli A., Scirocco T., Carapella G., Villani P., Casolino G., Schiavone P., Franchi M., 2002. The lago di Varano: Hydrologic characteristics and sediment composition. *Marine Ecology* 23(1), pp. 384-394.

Stehman S.V. 2005. Comparing estimators of gross change derived from complete coverage mapping versus statistical sampling of remotely sensed data. *Remote Sensing of Environment* 96, pp. 466–474.

Stehman S.V., 1997. Selecting and interpreting measures of thematic classification accuracy. *Remote Sensing of Environment* (62), pp. 77–80.

Stehman S.V., Czaplewski R.L., 1998. Design and analysis for thematic map; accuracy assessment: fundamental principles. *Remote Sensing of Environment* 64, pp. 331–344.

Story M., Congalton R.G., 1986. Accuracy assessment: a user's perspective. *Photogrammetric Engineering and Remote Sensing* 52(3), pp. 397–399.

Sudeesh S., Sudhakar Reddy C., 2012. Vegetation and Land Cover Mapping of Nagarjunasagar-Srisailam Tiger Reserve, Andhra Pradesh, India using Remote Sensing and GIS. *International Journal of Geomatics & Geosciences* 2(4).

Swierenga H., De Groot P.J., De Weijer A.P., Derksen M.W.J., Buydens L.M.C., 1998. Improvement of PLS model transferability by robust wavelength selection. *Chemometrics and Intelligent Laboratory Systems* 41(2), pp. 237-248.

Tadolini T., Bruno G., 1984. The influence of geostructural setting upon water thermo-mineralization in certain areas of Apulia (southern Italy). *I Hydrogeological Processes in Karst Terranes*, 75.

Tomaselli V., Dimopoulos P., Marangi C., Kallimanis A.S., Adamo M., Tarantino C.,, Panitsa M., Terzi M., Veronico G., Lovergine F., Nagendra H., Lucas R., Mairota P., Mucher CA Blonda P., 2013. Translating land cover/land use classifications to habitat taxonomies for landscape monitoring: a Mediterranean assessment. *Landscape Ecology* 28(5), pp. 905-930.

Tominaga Y., 1998. Representative subset selection using genetic algorithms. *Chemometrics and intelligent laboratory systems* 43(1), pp. 157-163.

Tominaga Y., 2006. Comparative study of class data analysis with PCA-LDA, SIMCA, PLS, ANNs, and k-NN. *Chemometrics and Intelligent Laboratory Systems* 49(1), pp. 105-115.

Tou J.T., Gonzalez R.C., 1974. Pattern Recognition Principles.

Turk G., 2002. Map evaluation and "chance correction". *Photogrammetric Engineering and Remote Sensing*, 68(2).

Turner W., Spector S., Gardiner N., Fladeland M., Sterling E., Steininger M., 2003. Remote Sensing for Biodiversity Science and Conservation. *Trends in Ecology & Evolution* 18 pp. 306–314.

Van Deusen P.C., 1996. Unbiased estimates of class proportions from thematic maps. *Photogrammetric Engineering and Remote Sensing* 62(4), pp. 409–412.

Van Genderen J.L., Lock B.F., 1977. Testing land-use map accuracy. *Photogrammetric Engineering and Remote Sensing* 43(9).

Vanden Borre J., Haest B., Lang S., Spanhove T., Forster M., Sifakis N.I., 2011. Towards a Wider Uptake of Remote Sensing in Natura 2000 Monitoring: Streamlining Remote Sensing Products with Users' Needs and Expectations. In *Proceedings, 2nd International Conference on Space Technology (ICST), Athens*, pp. 1–4.

Varela R.A.D., Rego P.R., Iglesias S.C., Sobrino C.M., 2008. Automatic habitat classification methods based on satellite images: a practical assessment in the NW Iberia coastal mountains. *Environmental monitoring and assessment* 144(1-3), pp. 229-250.

Villani P., Carapella G., Scirocco T., Specchiulli A., Maselli M., Schiavone R., Spagnoli F., Marolla V., Casolino G., Franchi M., Schiavone P., Deolo A., 2000. *Progetto Integrato di Recupero*

e Riqualificazione della Zona Umida della Laguna di Varano. Technical Report Consorzio ELTCON, Roma, Italy.

Wang X.R., 2009. Learning and Classification of Hyperspectral Images. Ph.D. thesis, University of Sydney.

Whiteside T., Ahmad W., 2005. A comparison of object-oriented and pixel-based classification methods for mapping land cover in northern Australia. In: *Proceedings of SSC2005 Spatial intelligence, innovation and praxis*: *The national biennial Conference of the Spatial Sciences Institute*, pp. 1225-1231.

Whittaker R.H., Levin S.A., Root R.B., 1973. Niche, habitat and ecotope. *American Naturalist* 107, pp. 321-338.

Wise B.M., Gallagher N.B., Bro R., Shaver J.M., Windig W., Koch R.S., 2006. Chemometrics tutorial for PLS_Toolbox and Solo. Eigenvector Research, Wenatchee.

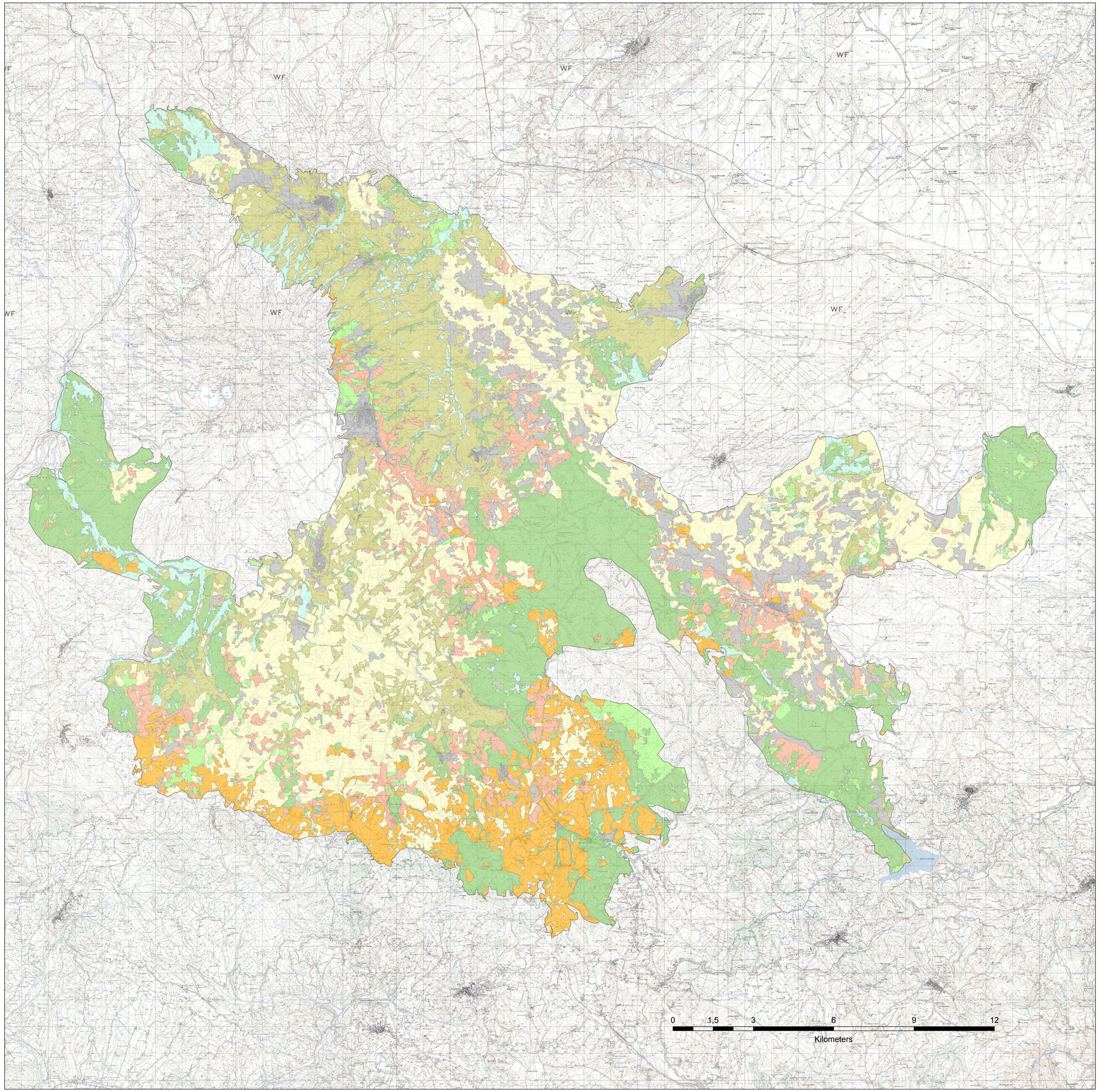
Wold H., 1975. Soft modeling by latent variables: the nonlinear iterative partial least squares approach. *Perspectives in probability and statistics, papers in honour of MS Bartlett*, pp. 520-540.

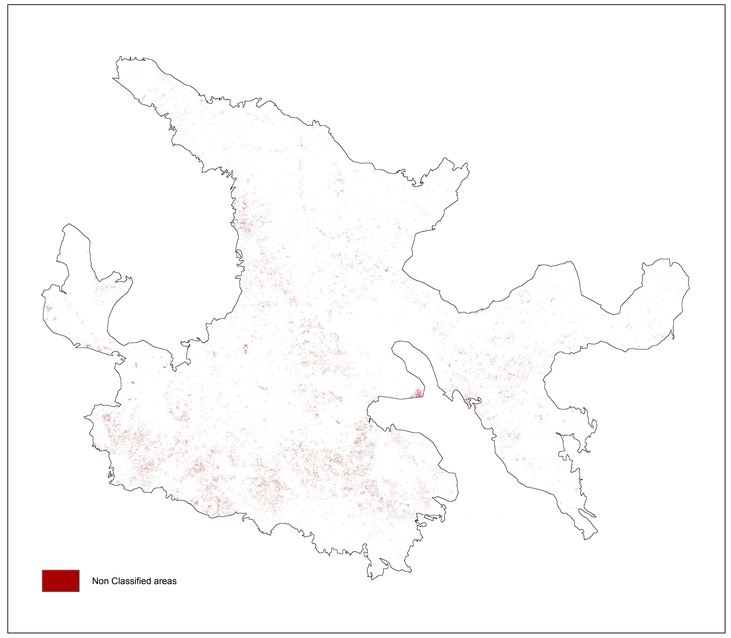
Wolter P.T., Townsend P.A., Sturtevant B.R., Kingdon C.C., 2008. Remote sensing of the distribution and abundance of host species for spruce budworm in Northern Minnesota and Ontario. *Remote Sensing of Environment* 112(10), pp. 3971-3982.

Wu W., Walczak B., Massart D.L., Heuerding S., Erni F., Last I.R., Prebble, K.A., 1996. Artificial neural networks in classification of NIR spectral data: design of the training set. *Chemometrics and intelligent laboratory systems*, 33(1), pp. 35-46

Xie Y., Sha Z., Yu M., 2008. Remote sensing imagery in vegetation mapping: a review. *Journal of Plant Ecology*, 1(1), pp. 9-23.

Yu Q., Gong P., Clinton N., Biging G., Kelly M., Schirokauer D., 2006. Object-based detailed vegetation classification with airborne high spatial resolution remote sensing imagery. *Photogrammetric engineering and remote sensing* 72(7), pp. 799.





Non Classified areas map

82.A 86.A 31.8 34.32 38.A 83.11 44.61 41.7

82.A	Field crops and Extensive cultivation
86.A	Towns and Active industrial sites
31.8	Western Palaearctic temperate thickets
34.32	Sub-Atlantic semidry calcareous grasslands
38.A	Mediterranean subnitrophilous grass communities and Mesophile pastures
83.11	Olive groves
44.61	Mediterranean riparian poplar forests
41.7	Thermophilous and supra-Mediterranean oak woods

N

Ph.D. course

Environmental Science - XXV Cycle

UNIVERSITÀ della Tuscia

ISPRA Istituto Superiore per la Protezione e la Ricerca Ambientale

Annex I:
Monte Vulture volcanic complex classified map

Ph.D. thesis:
A novel classification method to map habitat using multispectral images and ancillary data

Author:

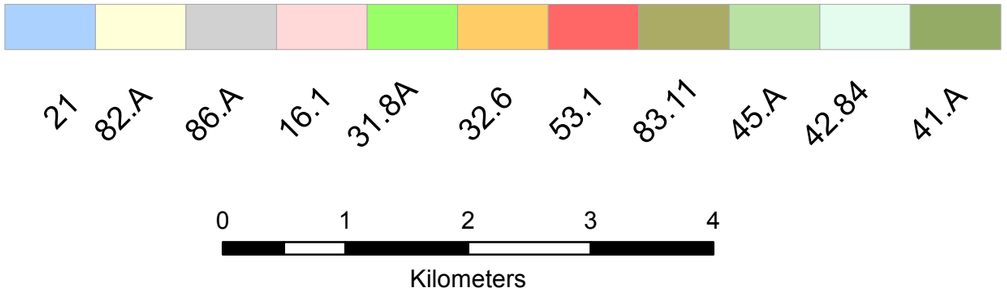
Emiliano Canali

Scale 1:50.000

Topographic map: IGM serie 25

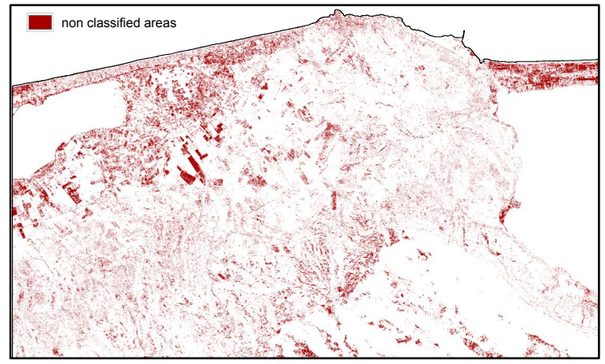
Datum: WGS84 – UTM33N

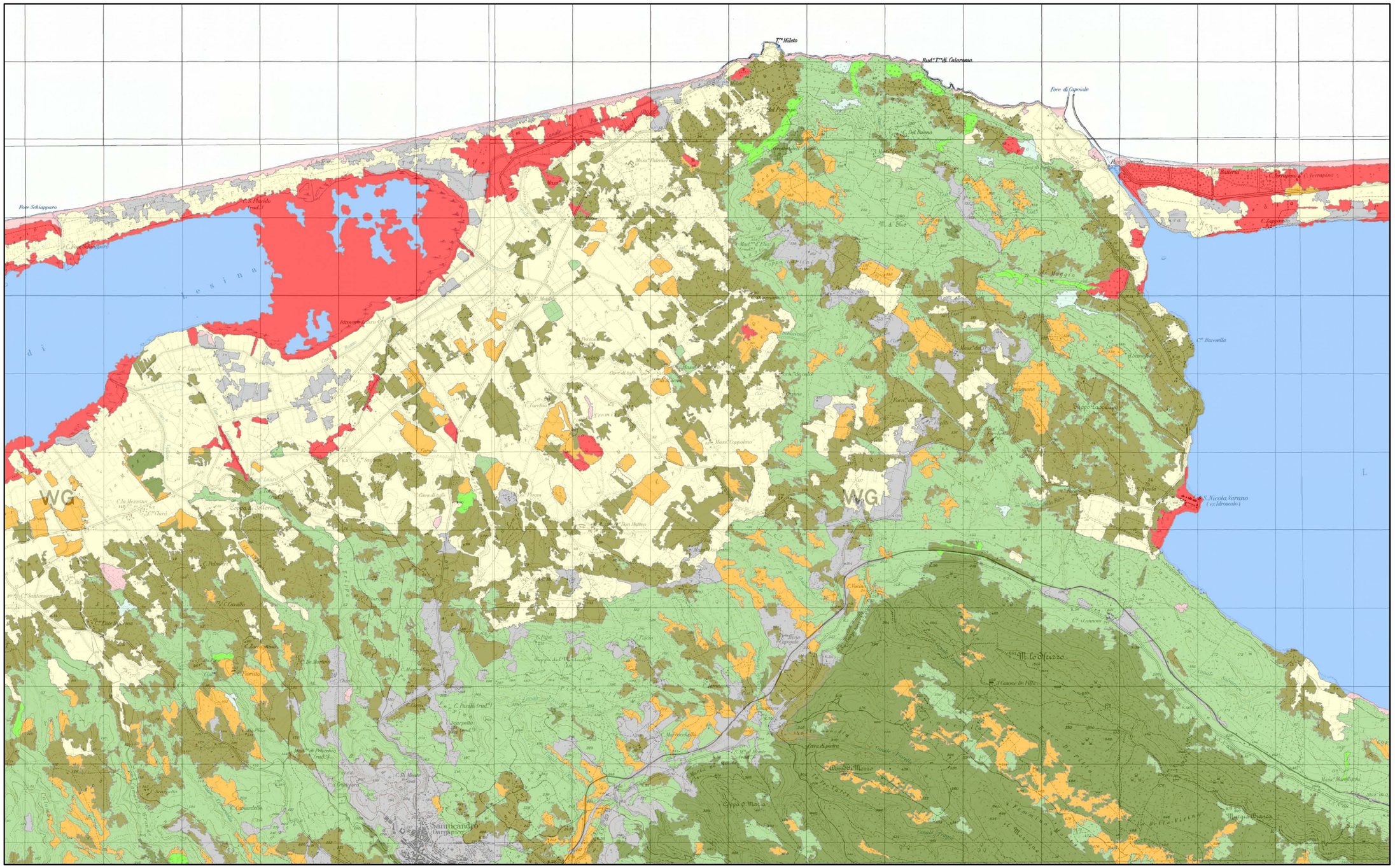
June 2014



21	Coastal lagoons
82.A	Field crops and Extensive cultivation
86.A	Towns and Active industrial sites
16.1	Sand beaches
31.8A	Tyrrhenian sub-Mediterranean deciduous thickets
32.6	Supra-Mediterranean garrigues
53.1	Reed beds
83.11	Olive groves
45.A	Sclerophyllous woodlands
42.84	Aleppo pine forests
41.A	Broad-leaved deciduous forests

Non Classified areas map





N

Ph.D. course

Environmental Science - XXV Cycle

Annex II:
Lesina lagoon classified map

Ph.D. thesis:
A novel classification method to map habitats using multispectral images and ancillary data

Author:

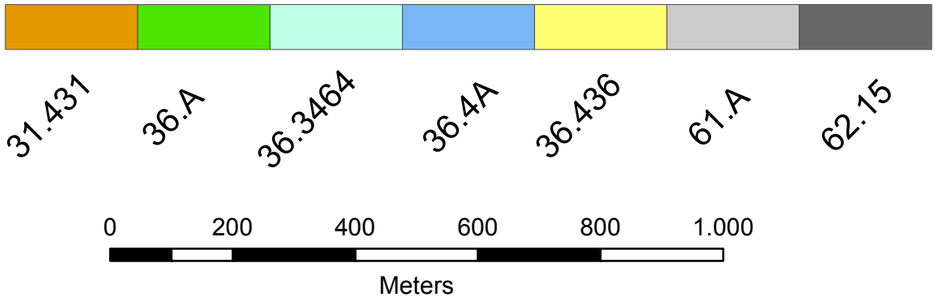
Emiliano Canali

Scale 1:50.000

Topographic map: IGM serie 25

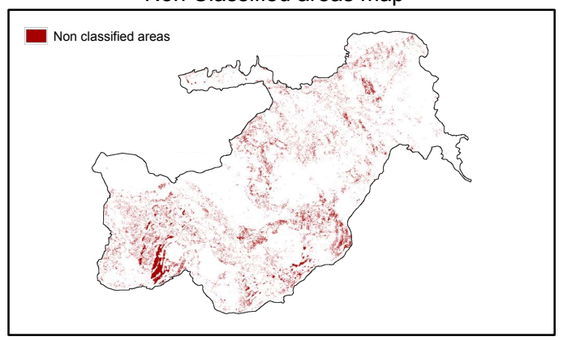
Datum: WGS84 – UTM33N

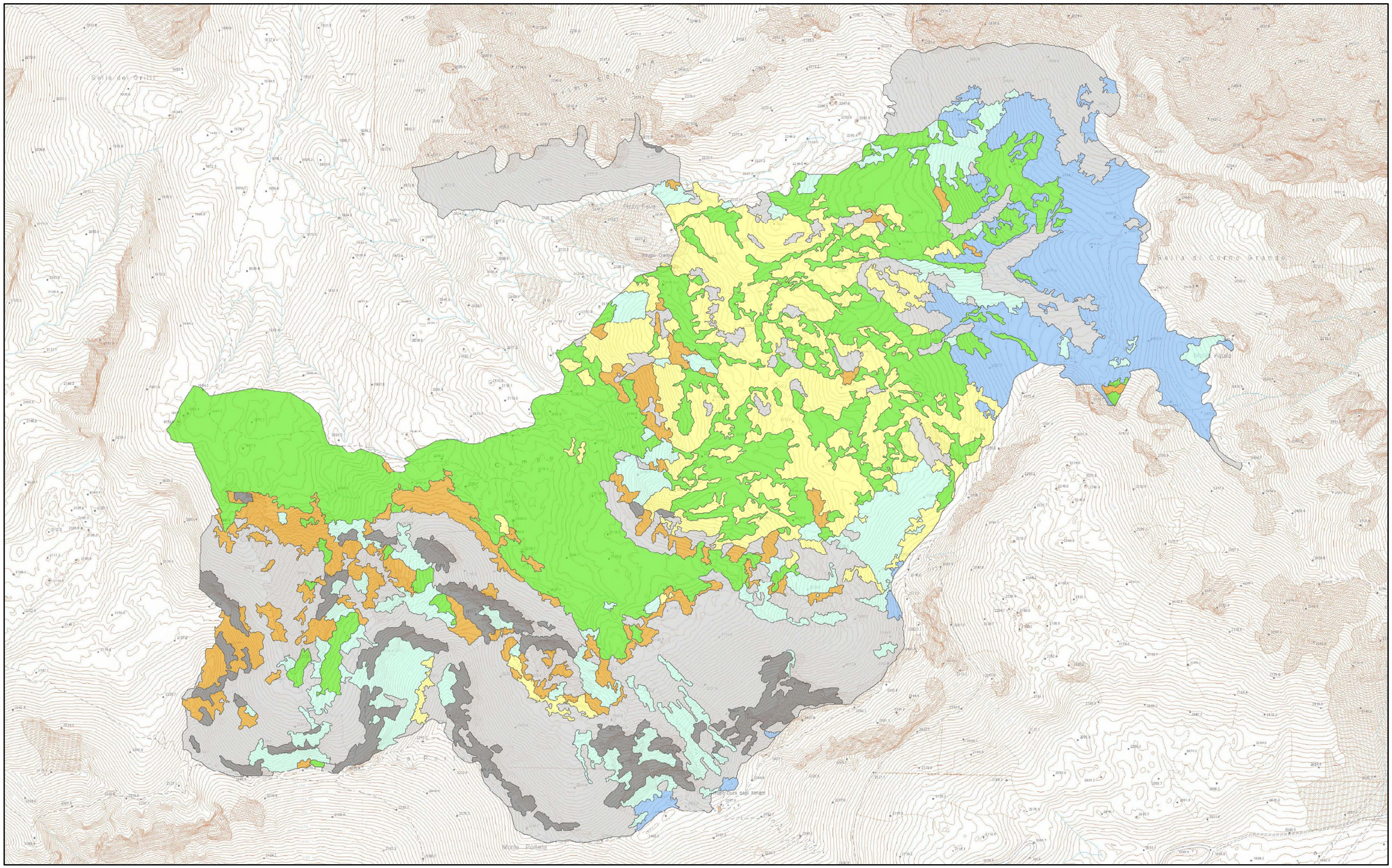
June 2014



31.431	Mountain [*Juniperus nana*] scrub
36.A	Oro-Apennine and Pyreneo-Alpine grasslands
36.3464	Alpine [*Juncus trifidus*] swards
36.4A	Apennine naked-rush swards and Violet fescue
36.436	Apennine stripped grasslands
61.A	Screes
62.15	Alpine and sub-mediterranean cinquefoil cliffs

Non Classified areas map





N

Ph.D. course

Environmental Science - XXV Cycle

UNIVERSITÀ Tuscia

ISPRA Istituto Superiore per la Protezione e la Ricerca Ambientale

Annex III:
Campo Pericoli basin classified map

Ph.D. thesis:
A novel classification method to map habitats using multispectral images and ancillary data

Author:

Emiliano Canali

Scale 1:10.000

Topographic map: Technical map Regione Abruzzo

Datum: WGS84 – UTM33N

June 2014