Treball de Final de Grau/Màster / *Trabajo de Final de Grado/Màster*

TÍTOL / *TÍTULO/ TITLE*

Classifiers ensemble in remote ***sensing: a comparative analysis***

Autor/a / *Autor/a/ Author:*

Hernán Cortés Rodríguez

Director/a / *Director/a/ Supervisor:*

Pr. Reyes Rengel

Tutor/a o supervisor/a / *Tutor/a o supervisor/a/ Co-supervisors:*

Pr. Mario Caetano
Pr. Roberto Henriques

Data de lectura / *Fecha de lectura/ Date of Thesis Defense:*

06/04/2014

Resum / *Resumen/ Abstract:*

Land Cover and Land Use (LCLU) maps are very important tools for understanding the relationships between human activities and the natural environment. Defining accurately all the features over the Earth's surface is essential to assure their management properly. The basic data which are being used to derive those maps are remote sensing imagery (RSI), and concretely, satellite images. Hence, new techniques and methods able to deal with those data and at the same time, do it accurately, have been demanded.

In this work, our goal was to have a brief review over some of the currently approaches in the scientific community to face this challenge, to get higher accuracy in LCLU maps. Although, we will be focus on the study of the classifiers ensembles and the different strategies that those ensembles present in the literature. We have proposed different ensembles strategies based in our data and previous work, in order to increase the accuracy of previous LCLU maps made by using the same data and single classifiers.

Finally, only one of the ensembles proposed have got significantly higher accuracy, in the classification of LCLU map, than the better single classifier performance with the same data. Also, it was proved that diversity did not play an important role in the success of this ensemble.

Paraules clau / *Palabras clave/ Key words:*

Accuracy, Bagging, Boosting, CART, Classifiers Ensemble, Diversity, Feature Selection, Land Cover and Land Use Maps, Linear Discriminant Classifier, Majority Voting, Neural Networks, Random Forest, Regularized Discriminant Classifier, Remote Sensing Imagery, Stacked Description, Single Classifiers, Support Vector Machine

Classifiers ensemble in remote sensing: a comparative analysis

Dissertation Document

Erasmus Mundus Master Program in Geospatial Technologies

Hernán Cortés Rodríguez

Education and Culture

Erasmus Mundus

Erasmus Mundus Master Program in Geospatial Technologies

Dissertation Title:

Classifiers ensemble in remote sensing: a comparative analysis

Dissertation supervised by:

- Dr. Mario Caetano
- Dr. Roberto Henriques
- Dr. Reyes Rengel

Hernán Cortés Rodríguez

Castellón, March 2014

UNIVERSIDADE NOVA ISEGI Estatística e Gestão de Informação

UNIVERSITAT JAUME·I

WESTFÄLISCHE WILHELMS-UNIVERSITÄT MÜNSTER

Acknowledges

I would like to greet all my supervisors: Professor Mario Caetano, Professor Roberto Henriques and Professor Reyes Rengel, who were always available to guide and help me during the process of elaborating this dissertation. I would like to thanks especially to the PhD student and researcher Joel Silva, who was the most inexhaustible support at any time. I would like to thank Professor Marco Painho and the Doctor Alan Glenn for their guideline and ideas during the thesis follow-up meetings.

As always, an special mention to my family, my friends and my flatmates in Lisbon, who had always a good feeling and a smile to share.

Dissertation Title:

Classifiers ensemble in remote sensing: a comparative analysis

Abstract

Land Cover and Land Use (LCLU) maps are very important tools for understanding the relationships between human activities and the natural environment. Defining accurately all the features over the Earth's surface is essential to assure their management properly. The basic data which are being used to derive those maps are remote sensing imagery (RSI), and concretely, satellite images. Hence, new techniques and methods able to deal with those data and at the same time, do it accurately, have been demanded.

In this work, our goal was to have a brief review over some of the currently approaches in the scientific community to face this challenge, to get higher accuracy in LCLU maps. Although, we will be focus on the study of the classifiers ensembles and the different strategies that those ensembles present in the literature. We have proposed different ensembles strategies based in our data and previous work, in order to increase the accuracy of previous LCLU maps made by using the same data and single classifiers.

Finally, only one of the ensembles proposed have got significantly higher accuracy, in the classification of LCLU map, than the better single classifier performance with the same data. Also, it was proved that diversity did not play an important role in the success of this ensemble.

Keywords

Accuracy

Bagging

Boosting

CART

Classifiers Ensemble

Diversity

Feature Selection

Land Cover and Land Use Maps

Linear Discriminant Classifier

Majority Voting

Neural Networks

Random Forest

Regularized Discriminant Classifier

Remote Sensing Imagery

Stacked Description

Single Classifiers

Support Vector Machine

Acronyms

ANN – Artificial Neural Network

AL – Active Learning

CART - Classification And Regression Tree

CLC – CORINE Land Cover Cartography

CORINE - Coordination of Information on the Environment

COS - Carta de Ocupaçao/Uso do Solo de Portugal

DT – Decision Tree

FAO – Food and Agriculture Organization

GIS – Geographic Information Systems

IF - Forestry Inventory of Portugal

KNN – K Nearest Neighbour

LANDAU -

LANDSAT – Land use Satellite

LCLU – Land Cover and Land Use

LDC – Linear Discriminant Classifier

MCS – Multiple Classifier System

ME – Mixture of Expert

ML – Maximum Likelihood

NDVI - Normalized Difference Vegetation Index

NN – Neural Network

PCA – Principal Component Analysis

RDC – Regularized Discriminant Classifier

RSI – Remote Sensing Imagery

SVM – Support Vector Machine

Index of Text

Index of tables

Index of figures

1 Introduction

1.1 Why LCLU maps are important?

Land cover is the physical material on the Earth's surface, as a result of the expression of human activities and changes (Di Gregorio and Jansen, 2000). Land cover comprises trees, asphalt, water, etc. Consequently, land cover is a geographical feature that can form a reference base for applications ranging from forest management and monitoring, agriculture, urban planning, investment, biodiversity, climate change, to desertification control, and so on (Di Gregorio and Jansen, 1997). In the other hand, land use could be explained as the use that humans give to the features on the surface, in this sense, a tree (unique land cover feature) could be a interpreted as forest, urban park, rain-fed agriculture, etc. Figure 1 is an example of land cover map of Portugal done under the CORINE project's umbrella in 2006.

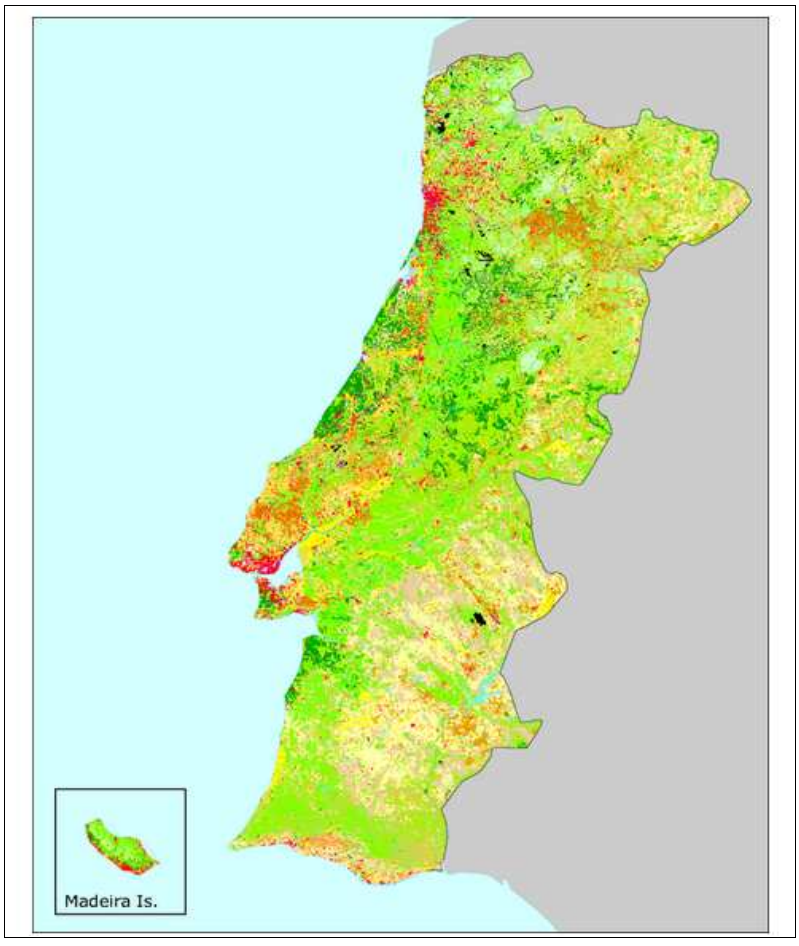

Figure 1. CORINE land cover: Portugal 2006

There is a higher demand for precisely define the Earth's surface due to an increase need for defining and classify accurately the land cover, in order to offer the best tools to the decision-makers. In this sense, the CORINE report (1995) from the European Environmental Commission, also states that if our environment and natural heritage have to be properly managed, decision-makers need to be provided with both an overview of existing knowledge, and information which is as complete and up-to-date as possible on changes in certain features of the biosphere. In this sense, the term LCLU Change (LCLUC), also known as land change, is adopted to point out to any modification of the Earth's surface. From a conceptual perspective, study of land-cover changes permits identification of long-term trends in time and space and the formation of policy in anticipation of the problems that go with the changes in land-cover (Jensen, 1996).

Land-cover of the Earth's land surface has been changing since time immemorial at a range of spatial scales from local to global and at temporal frequencies of days to millennia an is likely to change in the future (Townshend, J., et al., 1991). LCLU features have the particularity to show spatial and structural changes as a reaction to changes in physical, economic and cultural circumstances.

It is a fact that humans have continually reshaped the Earth, but the present magnitude and rate are unprecedented (Di Gregorio and Jansen, 1997). Nowadays, administrations and organizations have realized that it is very important to know how land cover has changed over time, in order to make assessments of the changes one could expect in the (near) future and the impact of these changes will have on peoples' lives (FAO, 1996a). Changes in land cover and land use affect global systems (e.g. atmosphere, climate and sea level) or occur in a localized fashion in enough places to have a significant effect (Meyer and Turner, 1992).

Due to the influence of land-cover change on many of the environmental issues both direct and indirect, such as loss of biodiversity, changes in hydrological, carbon and nitrogen cycles, and climatic change, it is important that these areas under intense change needs to be better understood for adapting suitable management strategies (Schilling et al., 2010). As an example of this importance, could be shown the article of Sala et al. (2000), which calculates that along the coast, land-cover change has been reported to be the prime cause of biodiversity loss, accounting for over 50% of the global biodiversity loss, and thus gained global attention in the last

decade. Also in the context of the management of the natural resources, LCLU has become an important tool, in the sense that an increasing number of socio-economic activities are taking place on those areas.

1.2 LCLU classification

The recent availability of geospatial information technologies with satellite data (in the last four decades) helps us for better understanding of the land-cover change and its effect on human environment. Change detection could be defined as the process of identifying differences in the state of an object or phenomenon by observing it at different time. This process is usually applied to earth surface changes at two or more times (Coppin et al., 2004).

Information describing current land cover is an important input for planning and modelling, but the quality of such data defines the reliability of the simulation outputs (Townshend, 1992). In addition to a high demand for improved land cover data sets, in the middle ninety's there was also a need for standardization and compatibility between data sets and for the possibility to map, evaluate and monitor wide areas in a consistent manner, which technical advances, like the vast amount of remote sensing data that has become available from earth observation satellites, made that increasingly possible (Di Gregorio, 1995).

Remote sensing imagery and, specifically, satellite imagery is the main source for creating LCLU maps. This is due to the following reasons:

- The large amount of data available and produced every day in different spatial and spectral resolution.
- The high temporal resolution which allows a large availability of images from the same area.
- The elevated variability of computational methods to analyse the satellite imagery.
- High availability of images, some of them free of charge, like Landsat images.

Techniques to deal with this increasingly large amount of data had been also improved through these years. We will face this evolution in the following chapter (chapter 2), but as an introduction, we can say that several attempts have been done with different classifiers. More recently different researchers have been trying to combine classifiers in different manners. In this dissertation, we explores a specific possibility, classifiers ensemble and it consists on a machine learning paradigm where multiple algorithms are trained to solve the same problem and combined to use and to get a best solution (Zhou, 2004).

Present work is a modest attempt to prove that most of the time, ensemble methods improve the accuracy of LCLU maps. In this dissertation, we have built different ensemble methods following the evolution they have followed through the last two decades.

1.3 Research Question

Does a Multiple Classifiers System (MCS) or a Classifiers Ensemble (CE) perform better than a single classifier?

In this dissertation, our purpose is to build different ensemble methods to compare and to analyse the results of accuracy obtained (on their classification of satellite images to create LCLU maps) with those results obtained in the LANDAU project, where only single classifiers were used to classify satellite images.

The study area is close to the mouth of the river Tejo and the reason of our choice is that this area was already taken in a previous project (LANDAU), providing results for comparison. In this project different simple classifiers were used to build LCLU maps. Hence, in order to assess the success of our maps, we need to compare the accuracy obtained with the different ensemble methodologies adopted for us and the accuracy achieved in the algorithms used in LANDAU.

Our first step, it will be to study the LANDAU project conclusions and to analyse their outputs, which could be a guide for our first stage. Then, once we are familiar with data and the variables we have to deal with, we should consider if adding more variables (bands, in our case) to the classification procedure, will enlarge our possibilities of success. The more straightforward way to add features (variables) in

this case, is the creation of artificial bands (vegetation index for instance, NDVI), since the majority of the landscape of our study area is natural and include different types of vegetation.

After that, the main task of the whole process will be to build an ensemble (or a set of them) of classifiers which allow us to get better accuracies in our map, in relation with the accuracies obtained in the LANDAU project. The issue that underlie this approach is the possibility of building a straightforward combination of algorithms to classify a satellite image rather than using a single classifier which should be fine-tuned, in order to get higher accuracies, and probably over-fitted.

We will develop that using as guidance the article from Du et al. (2012), on which a review of the classifiers ensemble is done and some new combinations are proposed. Hence, we will build some of the most used ensemble taking into account this review and then, we will try to build our own hypothesis, by using the same simple classifiers that were compared in the LANDAU project.

Finally, we will try to demonstrate the diversity hypothesis, by testing if the ensembles with better global accuracy results are those which also show higher diversity value. There is a kind of intuition about diversity which say that the more different are the outputs of the classifiers the wider the range of features that could be classified, and this entails an accurate classification. In this sense, when all the algorithms classify on the same way, the diversity is minimum and, when they classify in a completely opposite way, it could be said, that the diversity is maximum. Neither of these extremes is helpful to classify remote sensing imagery. Hence, the issue is to know where the threshold, in which the diversity influence is positive, is.

The remainder of this paper is organized as follows. In Section 2, we introduce some common approaches of RSI classification, summarizing the factors that play a important role in the accuracy of the final map. In Section 3, it is explained the methodology followed in this work, including a brief explanation of the single and classifiers ensemble used in it. Experimental results are presented in Section 4. Section 5 includes the conclusion of this paper.

2 Improving accuracy: factors influencing LCLU mapping accuracy

The main goal of the scientific community when dealing with the classification of remote sensing images to create LCLU maps, is to get as higher accuracy as possible. There are many approaches, which have been followed historically, to face this problem since the first images were obtained. Many of these methods and algorithms are derive from classical statistic, as Linear Discriminant Classifiers or K-Nearest Neighbour, but also those that come from the Machine Learning and Data Mining field are being widely used, as Neural Network, Decision Trees and Support Vector Machine.

Building a LCLU map from remote sensing imagery is basically to assign a class to an object. In remote sensing images those objects are pixel with an intensity value, which represents the measured solar radiance in a given wavelength band reflected from the ground (Liew, 2001). This process to assign a class to any pixel is also a prediction. So, using a group of pixels which class have been already classified and checked in the field, training sample (in the case of supervised classification) we predict the behaviour of the rest of the pixels in the image and provide a class to each of them.

The process of classifying remote sensing images involves many factors, as user's needs, data available, skills of the analyst, the design of the procedure and so on.

The more important steps in the remote sensing classification process are: data selection, classification system and training data selection (in the case of supervised classification, which otherwise used to be the most popular methodology), data preprocessing, feature extraction and selection, suitable classification method choice and accuracy classification assessments (Lu and Weng, 2007). All of these steps are dramatically important to ensure the higher accuracy possible of the final product of the process, which is the LCLU map.

In the following sub-sections, we will try to summarize the main properties and particularities of these steps and how important they are under the accuracy point of view. We will pay special attention to those factors that play an important role in this dissertation.

2.1 Initial steps

Remote sensing images selection is the first step it is followed in the process and it will determine the quality of the final map. The type of image will be depending on the scale and dimensions of the study area, the user's needs and the kind of images available. Economics resources play also an important role, since the price of the images vary dramatically from images free of charge, like Landsat, to images very expensive. Figure 2 is an example of Landsat image of the study area used in this work.

Secondly, defining land cover classification units is also an important task to implement. These units' choice is very related with the spatial resolution of the image and have to be environmentally and ecologically meaningful (Cingolani et al. 2004).

Remote sensing classification is mostly supervised, which entails the existence of a training set. Traditionally, for a wide range of classifiers have been defined a positive relation between the size of the training set and the classification accuracy (Foody, and Mathur, 2004). But the acquisition of a large training set is very costly in terms of financial and time resources. Indeed, some studies like Foody et al. (2006), claim that size is just an attribute of the training set and considerations about the way the classifiers perform can help when selecting the samples of the training set in a small and less costly way.

Other tentative to build smart training sample is Active Learning (AL), an algorithm widely used in Machine Learning, which is being introduced in remote sensing classification lately. Rajan et al. (2008) propose an active learning approach for hyper-spectral images classification, and they get a better accuracy classification than just choosing traditional random samples. Under the philosophy of these methods, lie the acquisition of "smarter" samples which better defines the classes or the border between them (some AL algorithms are built upon the Support Vector Machines -SVM- where the samples are chosen in the margin between different classes). Regarding to Tuia et al. (2011) "Active Learning aims at building efficient training set by improving iteratively the performance of the model". AL models return the pixel with more uncertainty to be classified, which are accurate labelled by the user and reincorporated to the model to reinforce and optimize it.

hen, the preprocessing step includes geometric rectification or image registration,

radiometric calibration, atmospheric and topographic corrections (Lu and Weng, 2007). Atmospheric corrections are not needed when dealing with an single-date image (Song et al. 2001). If ancillary data is used, conversion between different sources or formats will be included in this preprocessing stage. We do not go further in describing this issue, since many articles and books have illustrated it in detail (Jensen 1996, Toutin 2004).

2.2 Feature extraction and selection

Many different features could be used in remote sensing image classification. The most common are spectral bands from one or more dates and vegetation indices; but others kind of data are becoming popular in the pursuing of accuracy improvement, like transformed images, contextual and textural information, multi-sensor images and ancillary data.

Using many variables in the classification procedure could decrease the classification accuracy (Hughes 1968, Price et al. 2002). Thus, selecting those which result more useful to define the classes becomes a must. PCA, discriminant analysis and decision boundary feature extraction are some of the most used methodologies for that.

A special notation should be done about the multi-source imagery analysis, which is still being a trending topic today. Pohl and Genderen (1998), in a review about multi-sensor image fusion, listed the benefits of this “relatively new research field at the leading edge of the available technology”:

- Increase spatial resolution, sharpen images.
- Improve geometric corrections.
- Enhance certain features not visible in either of the single data alone.
- Complement data sets for improved classification.
- Replace defective data and missing information.

Twelve years later, Zhang (2010) stated that “developing effective methods for multi-source fusion and interpretation is still a challenging activity”. The high-speed of new sensor technologies development, new multi-source fusion techniques and

some remains problems in computation effectiveness and efficiency, make the field very dynamic.

Finally, the other strategy to get higher accuracy of the maps at this level, is to use contextual classifiers, where the spatial neighbouring pixel information is used. They were developed to cope with the of intraclass spectral variations (Lu and Weng, 2007). The Markov random field-based contextual classifiers, such as iterated conditional modes, are the most frequently used approaches in contextual classification (Magnussen et al. 2004). By now, computationally is much more expensive than other methods (process become 20% more computing consuming in a 4000 pixel image) and the increase of accuracy is not enough to justify their use, in most of the cases.

2.3 Suitable Classification Method

2.3.1 Single Classifiers Methods

The earlier classical statistics methods for RSI classification have been shifted in to derived methods from Machine Learning and Data Mining fields. Hence, methods like K-Nearest Neighbour (kNN), Linear Discriminant Classifier (LDC) and Maximum Likelihood (ML) were partially, replaced by others like Neural Networks (NN), Decision Trees (DT) and Support Vector Machine (SVM). Although, the classical methods are widely used.

Regarding with this pursuing of accuracy and the availability of higher computing resources, many different classifiers have being used to classify satellite imagery. These methods could be classify (Lu and Weng, 2007), in different ways.

If training set is needed, then the methods may be classified as:

- Supervised, they need some samples known by the user in order to predict the rest of objects (pixels): Maximum Likelihood Classification, k-NN Means, Support Vector Machine, Artificial Neural Networks, etc.
- Unsupervised, there are no need for training samples: Simple One-Pass Clustering, Isodata Classification , Self-Organization.

If covers methods which rely or not on assumptions that the data are following a given distribution, then the methods may be classified as:

- Parametric, data come from a probability distribution and make inferences with parameters, such as mean vector and covariance matrix: Linear Discriminant Classifier, Maximum likelihood.
- Non Parametric, no assumptions about data distribution are given (and needed): k-NN Means, Parzen Windows, Artificial Neural Networks.

If it is taken in account different unit of analysis, then the methods may be classified as:

- Pixel-based: each pixel has assigned a class, most of the classifiers: such as maximum likelihood, minimum distance, artificial neural network, decision tree and support vector machine.
- Sub-Pixel-based: each pixel is a combination of classes: Fuzzy-set classifiers, sub-pixel classifier, spectral mixture analysis.
- Object-oriented: pixel are merged into objects: OBIA (e-Cognition)
- Per-fields: integrating vector and raster data (GIS), the image is divided in parcels: GIS-based classification approaches.

For RSI classification, there is no classifier that could always perform well (Roli et al., 1997). This assumption is known also as the "no free lunch theorem" (Wolpert and Macready, 1997), where it is exposed that every classifier could have a weak performance when facing a classification problem. Based on that, the common strategy followed by many researchers is to compare different methods (Lu and Weng, 2007) and choosing the one which offers better results (accuracy).
One the most used algorithms in the decade of 1980 were the Artificial Neural Networks (ANNs), that include Back-propagation network, fuzzy neural network, Kohonen self-organizing featured map, Hybrid learning vector hierarchical network and so on. An artificial neural network is an interconnected group of nodes, which represents an artificial neuron (the idea of the system come from animal neurons and theirs connections) and an arrow that represents a connection from the output of one

neuron to the input of another (figure 2). The success of the ANNs in RSI classification is based in the size and quality of the training set (Zhang et al., 2013).

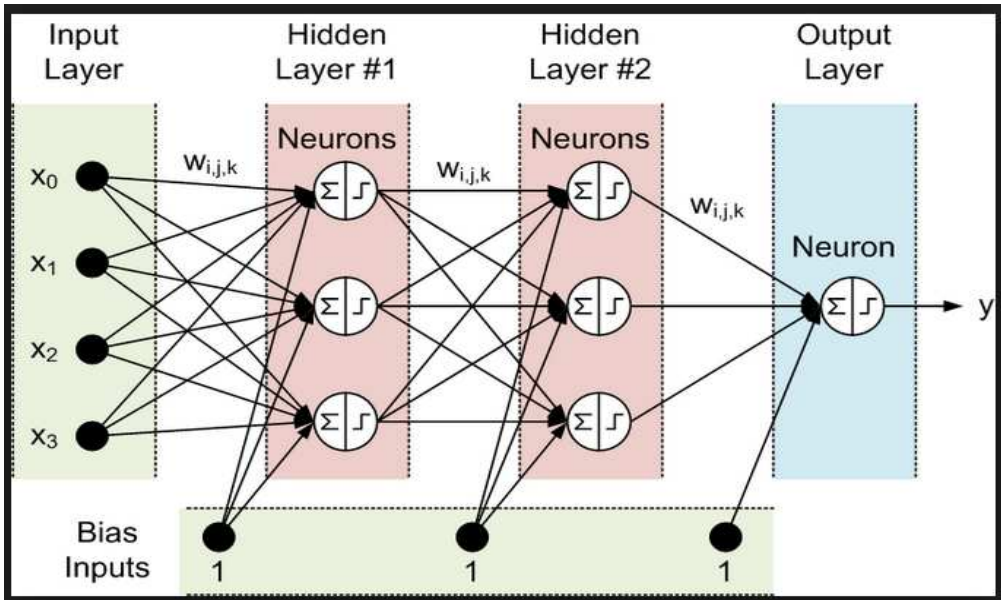

Figure 2. Neural Network classifier structure [1].

In the early ninety's, one the most novel and accurate algorithm used in RSI classification was Support Vector Machine (SVM). This method is based in looking for the optimal separating hyperplane in a multidimensional feature (figure 3).

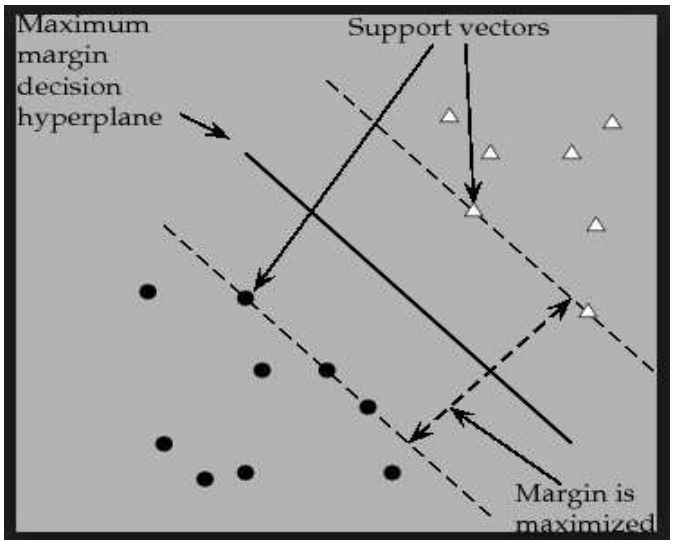

Figure 3. SVM strategy [2].

[1] http://codebase.mql4.com/5738

[2] http://nlp.stanford.edu

It has been used so much, because its stability, convenience and high precision (Zhang et al., 2013). It also does not need a big training set to be effective. It is very used and combined with other methods or techniques.

We will also consider in this section those single classifiers which were used in both projects, LANDAU and this dissertation. These are LDC, QDC (or Maximum Likelihood), Classification and Regression Tree (CART) and SVM (was analysed above).

- LDC is a method used to find a linear combination of features which describe or divided two or more classes of objects (pixels). It could be thought as the minimum-error (Bayes) classifier for normal distributed classes with equal covariance matrices, although the results can be surprisingly good even when the classes have no normal distribution (Kuncheva, 2004). LDC is related with Principal Component Analysis and Factor Analysis, in the sense that all of them look for a linear combination of variables which best explains the data (Martínez and Avinash, 2001).
- ML (special case of Quadratic Discriminant Classifier), as in the LDC we assume a normal distribution of the classes but in this cases the covariance matrix of every class is different (Kuncheva, 2004), which entails that allows to separate objects of different classes by a quadric surface. I t could be seen as a generalization of the LDC, justified by the ambition of classifying more complex separating surfaces.
- CART uses a decision tree as a predictive model where the decision process can be traced as a sequence of simple decisions (Kuncheva, 2004). In the tree structure, leaves represent class labels and branches represent combinations of features that guide to these labels. In our case, we will do classification tree analysis, so the predicted outcome is the class to which the data belongs.

2.3.2 Classifiers Ensemble Methods

Multiple Classifier System, Classifier Ensemble or Multiple Combination Methods are a machine learning paradigm where multiple learners are trained to solve the same problem, in others words, are a combination of different single classifiers in order to increase the accuracy of the classification (figure 4).

Ensemble learning methodologies, in contrast to ordinary machine learning approaches which try to learn one hypothesis from training data, try to construct a set of hypotheses and combine them to use (Zhou, 2004).

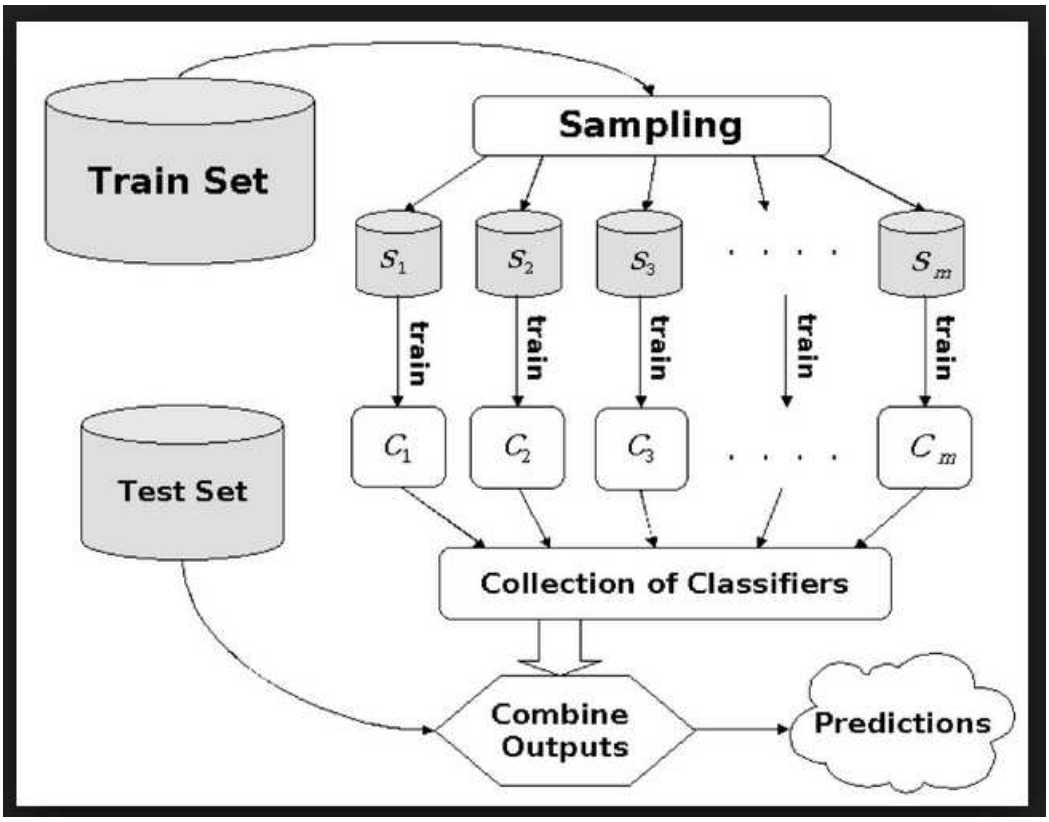

Figure 4. Classifier ensemble notion (Du, 2012).

The evolution of the classifiers ensemble could be described as follow (Polikar, 2006). Firstly, Hansen and Salomon (Hansen and Salomon, 1990) showed that performance of a neural network can be improved by using an ensemble of similarly configured neural networks. But it was the Boosting theory (Schapire 1990) which puts the ensemble systems at the centre of machine learning research, with the idea that a combination of weak classifiers could perform as strong classifier. After a few years, Boosting was improved by creating the AdaBoost algorithm (Freund and Schapire, 1996) which became one of the most popular ensemble learning algorithms. The others "big names" in classifiers ensemble are Bagging (Breiman, 1996) and Stacked generalization (Wolpert, 1992).

According to Polikar (2006), there many approaches and models of building an ensemble learning algorithm, but they usually differ basically in two ways:

- Specific procedure used for generating individual classifiers, which includes Bagging, Boosting, Stacked Generalization and Mixture of Expert.
    - *AdaBoost (Adaptive Boosting)*: It is one of the best known of all ensemble-based algorithms, extends boosting to multi-class and regression problems (Freund and Schapire 1996). AdaBoost is adaptive in the way that classifiers built are modified, by taking into account those instances misclassified by previous classifiers, and is boosting, because is an algorithm for constructing a ”strong” classifier as linear combination of simple “weak” classifiers. Allocates weight to a set of classifiers, as probability of best predicting the label, which will be updated after every training in the data set, the most successful ones gain weight.
    - *Random Forest (Bagging)*: Bagging or bootstrap aggregating is an ensembling method which trains independent and unstable classifiers, by using bootstrap replicate of the training set (bags). Random Forest (Breiman, 2001) operate by constructing a multitude of decision trees at training time and outputting the class that is the mode (most often value) of the classes output by individual trees.
    - *Stacked Description:* an ensemble of classifiers is first trained using bootstrapped samples of the training data, creating Tier 1 or first class classifiers, whose outputs are then used to train a Tier 2 or second class classifier or meta-classifier (Wolpert, 1992).
    - *Mixture of Expert:* it is a similar concept than *Stacked Description,* where a first level of classifiers are trained using bootstrapped samples of the training data, but the combination of the outputs is made by simple combination rules, as random selection or weighted majority. In here, a second level classifier or gating network (usually a neural network) is trained using the raw training data to determine the weight distribution of each classifier. The original Mixture of Experts (ME) model was introduced by Jacobs et al. in 1991.
- The strategy employed for combining the classifiers, in particular, the way in which the output of each classifier are combined. They includes Majority

Voting, Weighted Majority Voting, Naives Bayes Combination and Multinomial Methods. The classification could be divided in two depending if the combination rules apply to class labels or to class-specific continuous outputs.

- Combining class label:
    - *Majority voting*: it is technique (Kuncheva, 2004) where the ensemble choose the class on which, all classifiers agree (unanimous voting), at least, one more than half of the classifiers are agree (simple majority) more classifiers agree (plurality voting).
    - *Weighted majority voting:*it is used when we imagine than some classifier perform better than other, in that case, we weighted heavily those classifiers in order to improve our general performance (Kuncheva, 2004). There are two basic approaches to know which weight should be give to any classifier, by using a validation data set or the training data set (as AdaBoost), and estimate classifiers' future performance.
    - *Behaviour Knowledge Space (BKS),* (Huang and Suen, 1993) developed it firstly, and the procedure consists on keep track of all the labelling combination of the ensemble to finally the class which more times appear on the combination.
    - *Borda count:* each classifier vote each class by rankings. At the end , the most voted class is the chosen in the ensemble decision. It was first developed by Jean Charles Borda in 1770.
- Combining continuous outputs
    - *Algebraic combiners:* the support for a class is obtained by a simple function which includes all the support from all the classifiers. Includes: Mean Rule, Weighted Average, Minimum-Maximum-Median Rule, Product Rule and so on.
    - *Decision templates:* (Kuncheva, 2004) measure the similarity

(Euclidean distance) between every ensemble output and a template, created as averaged of decision profile observed in each class throughout the training.

- *Dempster-Shafer based combination:* the final value is linear combination of values of belief, instead of probability, which is measured in proximity instead of distances. The theory was first introduced by Arthur Dempster and Glenn Shafer (Shafer, 1976).

Ensemble strategies could be also classified (Kuncheva, 2004) on:

- *Classifier selection*, each classifier is trained in a part of the data set, having a good knowledge of it. The combination of the classifiers is then, based on the vicinity of the instance, according to some distance metric, then, the closet classifier obtain the highest credit, in order to be chosen to make the decision.
- *Classifier fusion*, all classifiers are trained over the entire feature space. The combination involves merging the individual classifiers output (which are normally normalized to the [0, 1] interval and entail the support of the classifiers to posterior prediction) to obtain a superior performance. This strategy based the combination in algebraic rules (mean rule, median rule, maximum rule, etc), majority voting or weighted majority voting , fuzzy integral or the Dempster-Shafer based fusion.

Traditionally the ensembles have followed three kind of structures:

- *Parallel:* the classifiers train the data set (modified -bagging- or not) independently and their output are combined.
- *Cascade or sequential:* where the output of a classifier is the input of the subsequent.
- *Mixed or Hierarchical:* is an ensemble which is a mix of both structures above.

The key of the success of ensemble learning is the diversity of the classifiers (Kuncheva and Whitaker, 2003). There is not a strict definition of diversity, but an intuition. The intuition is that if each classifier makes different errors, then a strategic combination of these classifiers can reduce the total error, a concept not too dissimilar to low pass filtering of the noise (Polikar, 2009). When dealing with diversity, the different authors refer to the difference of values obtained in the output of the classifiers that form the ensemble.

Still, the relation between diversity and ensemble accuracy is ambiguous. Nevertheless, many authors have tried to relate them and to generate more diverse ensembles, by:

- Using different training datasets to train individual classifiers. Bootstrapping or bagging is a technique of re-sampling data sets. Real different data sets is very expensive, also in terms of time.
- Uses of weak or more unstable classifiers could allow to get different decision boundaries.
- Use different training parameters for different classifiers, tuning the classifiers in dissimilar way.
- Using completely different types of classifiers.
- Choosing feature selection methods, where each classifier is trained in a separate part of the training set.

Finally, there are many different ways to measure this diversity. Measuring diversity is about measuring distance (Euclidean) between points (Kuncheva and Whitaker, 2003). Hence, when less correlated are the outputs of the classifiers the better the ensemble. In this sense, when the classifiers results are positively correlated the lack of accuracy is slightly reduced, when the correlation is small or negative the accuracy of the ensemble could be better (Turner and Gosh, 1996).
Diversity measures assess the degree of agreement between classifiers (Faria et al. 2013). They can be pairwise, between two classifiers or non-pairwise, the measure takes into account all the classifiers in the ensemble.

Some of the most used measures are (Kuncheva, 2004):

- Pairwise:
    - *Double-Fault Measure:* the ratio of the number of observations on which two classifier classify equally but wrong, to the total number of observations.
    - *Q-Statistic:* measure the ratio of the number of observations where the classifiers perform equally minus when they perform differently to the total number of observations.
    - *Interrater Agreement, k:* Defined by Kuncheva (2004) as the degree of agreement while correcting by chance.
    - *Disagreement Measure:* Defined also by Kuncheva (2004), as the ratio of the number of observations on which two classifiers classify differently to the total number of observations.
    - *Correlation Coefficient:* the diversity of two classifiers is inversely proportional to the correlation between them (Duta, 2009).

- Non-Pairwise:
    - *Entropy Measure, E:* makes the assumption that the diversity is higher if half of the classifiers are correct and the remaining wrong.
    - *Kohavi-Wolpert* Variance: is derived a decomposition formula for the error rate of the classifier.
    - *Measurement of Interrater Agreement, k:* which is similar but not equal to the average pairwise Kappa.
    - *Measure of Difficulty:* related with the difficulty that classifiers meet when trying to define the class of the data set and its distribution, for instance, the same problem with the same data all the classifiers entails low diversity of the ensemble.
    - *Generalized Diversity:* related with the probability of failure of a randomly chosen classifier.

In the figure below (figure 5) it can be seen a summarize of different diversity measures, in which is defined the correlation of the value itself with the measure (low value entail low or high diversity?), if it is pairwise or not and the reference.

Name		↑ / ↓	P	S	Reference
Q-statistic	Q	(↓)	Y	Y	(Yule, 1900)
Correlation coefficient	ρ	(↓)	Y	Y	(Sneath & Sokal, 1973)
Disagreement measure	D	(↑)	Y	Y	(Ho, 1998; Skalak, 1996)
Double-fault measure	DF	(↓)	Y	N	(Giacinto & Roli, 2001)
Kohavi-Wolpert variance	kw	(↑)	N	Y	(Kohavi & Wolpert, 1996)
Interrater agreement	κ	(↓)	N	Y	(Dietterich, 2000b; Fleiss, 1981)
Entropy measure	Ent	(↑)	N	Y	(Cunningham & Carney, 2000)
Measure of difficulty	θ	(↓)	N	N	(Hansen & Salamon, 1990)
Generalised diversity	GD	(↑)	N	N	(Partridge & Krzanowski, 1997)
Coincident failure diversity	CFD	(↑)	N	N	(Partridge & Krzanowski, 1997)

Note: The arrow specifies whether diversity is greater if the measure is lower (↓) or greater (↑). 'P' stands for 'Pairwise' and 'S' stands for 'Symmetrical'.

Figure 5. Summary of the measures of diversity (Kuncheva and Whitaker, 2003).

2.4 Accuracy assessment

The quality of LCLU map produced from the classification of a satellite image is estimated by measuring the accuracy between the classification of the land-cover made by any method and the reality. It is obvious that it is impossible to check the behaviour of every method at any surface unit (pixel), hence in our case; we do the validation using the testing sample. The most common way to do this comparison is by using a confusion/error matrix (Foody, 2002). In this method, the classification obtained by the methods at one place is compared with the class defined in the testing observation. If both classify in the same way, we could say that there is concordance, if not there is discordance. At the end, we sum up all the concordances divided by the number of testing observations and we have the Global Accuracy of the method.

There are other measure of validation derived from the confusion/error matrix, i.e. Producer's Accuracy and User's Accuracy. The Producer's Accuracy is defined as probability of finding in the map the same class that it is been checking in the field. The User's Accuracy is the probability to find in the ground the same class that it is been pointed out in the map. As an example to better understand both accuracies, let us use a class, like water bodies, which could have a value of hundred percent user's accuracy and eighty percent producer's accuracy, hence every water body in the map is in the ground, but only eighty percent of the water bodies in the reality are in the map. There is an error of omission (in the map). If the producer's accuracy is higher than the user's the error is of commission.

Methodology

3.1 Study Area and Data Selection

3.1.1 Study Area

he study area is located in the centre of continental Portugal, in the administration area of Alentejo, close to the mouth of the river Tejo (figure 6). The reasons why this area was selected are mainly, because is a flat area which facilitate the preprocessing stage, there are a wide variety of features (in this case, land use types) and it coincides with one of the study areas that were used in the LANDAU project, which is essential to establish comparison between results.

Figure 6. Location of the study area (Google Earth).

3.1.2 Data set

All the data needed is a Landsat 5 image from 2007 (July). This image has an spatial resolution of 30 meters and 7 spectral bands, although the sixth is not used commonly in these kind of analysis. Instead, we have added a synthetic band, a

vegetation index band (NDVI), which is very helpful to identify the vegetation. The idea of taking a summer image is because the atmosphere should be clean of clouds and the differences between the rain-fed and the irrigated agricultural fields should be significantly noticeable.

We have chosen just one kind of satellite imagery, Landsat (figure 7), being aware of our limitations of time, although it will be more recommended to try different spatial resolution images, which allow us to get better conclusions of our work.

The image is available on-line and can be downloaded from the following web address: http://landsatlook.usgs.gov/.

Figure 7. Landsat image from the study area

Other kind of ancillary data were used to determine the training and the testing set to train the algorithms, those data are:

- Aerial imagery (orto-rectified) with a spatial resolution of 0,05 meters and a spectral resolution of four bands, from the following years (1995, 2005 and 2007).
- Forestry Inventory of Portugal, IF (2005).

- CORINE Land Cover Cartography, CLC (2000, 2006).
- Carta de Ocupaçao/Uso do Solo de Portugal, COS (2007).

3.1.3 Nomenclature

The nomenclature of the features that appear in our LCLU maps were proposed in the LANDAU project (Dinis et al. 2012). Only 11 out of 15 categories defined were used in this work, due to the inapplicability of them in this area. Hence, the final categories in this work are keeping the same names and codes than in LANDAU:

- 1.1 - Discontinuous Artificial Areas.
- 2.1 - Irrigated Agriculture.
- 2.2 - Non-Irrigated Agriculture.
- 2.3 - Rice crops.
- 3.1 - Broadleaved Forest.
- 3.2 - Coniferous Forest.
- 3.4 - Grassland.
- 3.5 - Shrubs-land.
- 4 - Bare-land.
- 6 - Wetlands.
- 7 - Water bodies.

3.1.4 Software

In this work, we have used different software. Matlab was the software where the image was treated as a matrix of data and where all the algorithms were implemented and the outputs, presented as values of accuracy or LCLU maps were obtained. PrTool and Libsvm libraries were used in the implementation of the algorithms in Matlab.
We have also used Excel to analyse the results and present some table and graphics. Finally, the maps were displayed in ArcGIS.

3.2 Methodological procedure

Following the theoretical evolution of the classifiers ensemble methodologies (Polikar -2009- and Du -2012-) and their using in remote sensing imagery classification, we have try to implement in Matlab the most used and well known of these methods. Also, taking into consideration the analysis of the data and results of the LANDAU project, we choose some other ensembles that we thought could fit the analysis of the previous work.

The methods we have implemented are: Boosting Trees, Random Forest, Boosting Discriminant, Bagging Discriminant, Regularized Discriminant Classifier and a SVM Ensemble Strategy. Below, we will describe briefly their structure and our motivation to include them in our dissertation (section 3.2.2).

Firstly, it is necessary to explain that all of these methods are supervised, which entails to have a training data set to fit a model (in our case, the ensembles) that can be used to predict the not known values. And, also, it is required to have a testing set to validate the accuracy of our prediction (in our case, classification).

3.2.1 Training and Testing Set and Validation

The training set is made up of 10980 sample points, deterministically extracted from the satellite image by using the CORINE Land Cover Cartography (2006) and ancillary data (Dinis et al, 2012b).

The testing set was recollected directly in the study area. An amount of 550 observations taken in a random way and covering approximately equally all the classes (Dinis et al, 2012b).

The quality of LCLU map produced from the classification of a satellite image is estimated by measuring the accuracy between the classification of the land-cover made by any method and the reality. It is obvious that it is impossible to check the behaviour of every method at any surface unit (pixel), hence in our case; we do the validation using the testing sample. The most common way to do this comparison is by using a confusion/error matrix. In this method, the classification obtained by the methods at one place is compared with the class defined in the testing observation. If both classify in the same way, we could say that there is concordance, if not there is

discordance. At the end, we sum up all the concordances divided by the number of testing observations and we have the Global Accuracy of the method.
There are other measure of validation derived from the confusion/error matrix, i.e. Producer's Accuracy and User's Accuracy. The Producer's Accuracy is defined as probability of finding in the map the same class that it is been checking in the field. The User's Accuracy is the probability to find in the ground the same class that it is been pointed out in the map. As an example to better understand both accuracies, let us use a class, like water bodies, which could have a value of hundred percent user's accuracy and eighty percent producer's accuracy, hence every water body in the map is in the ground, but only eighty percent of the water bodies in the reality are in the map. There is an error of omission (in the map). If the producer's accuracy is higher than the user's the error is of commission.

3.2.2 Ensemble methods description

As we assess in the beginning of chapter 3.2, in this section we are going to describe all the methods that were used in this work. Those are: Boosting Trees, Random Forest, Boosting Discriminant, Bagging Discriminant, Regularized Discriminant Classifier and SVM Ensemble Strategy.

- Boosting Trees: Classification tree analysis provides an effective collection of algorithms for classifying remotely sensed data, but has the limitations of not searching for the optimal tree structure or being adversely affected by outliers, inaccurate training data, and unbalanced data sets (Lawrence et al., 2004). Boosting is a technique developed to increase classification accuracy by forcing the learning algorithm to concentrate on those training observations that are most difficult to classify (Frield et al., 1999). Boosting which is an adaptive and iterative training technique allows the combination of trees to find the best structure and being insensitive to noise (Shapire, 1990). Boosting is one of the most important strategies in constructing ensemble (figure 8). This ensemble construction is also used in Du's work (Du, 2012).

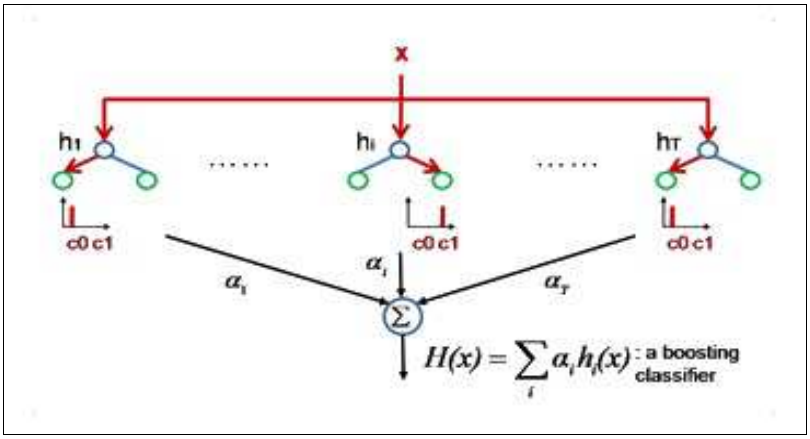

Figure 8. Boosting tree diagram[3].

- Boosting Discriminant: The structure of this ensemble is similar to the previous, but instead of classification Trees, there are a combination of Linear Discriminant Classifiers (LDC), which had the best single classifiers performance in the LANDAU project. The reasons to choose this ensemble are the same that the exposed in the ensemble above. Also, to include the most successful single classifier in the LANDAU project (LDC).
- Random Forest: It is one the most known classifiers ensemble and it was introduced by Breiman in 2001. The idea in this ensemble is the using of bagging to improve the performance of the combination of classification Trees. What bagging offers is a bootstrap replicate of the training set with replacement (kind of "bags"), (I,x) in the figure 9, in which the different Trees are trained, in order to get more diversity in the outputs.

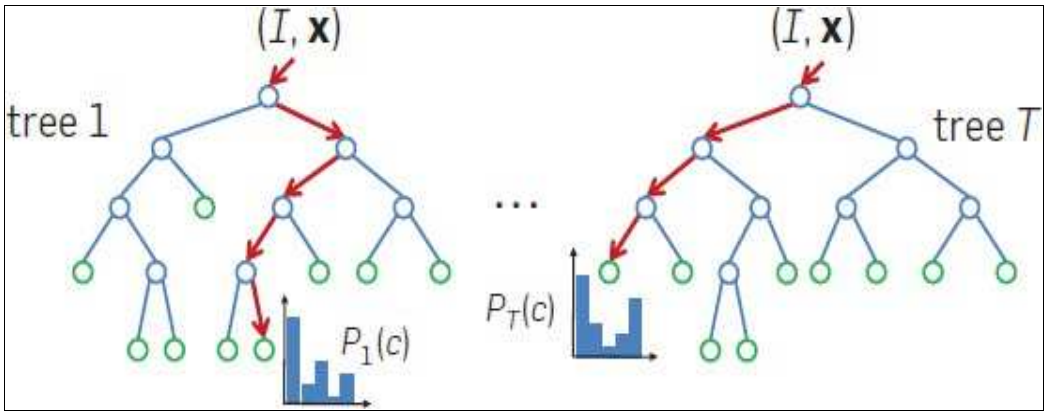

Figure 9. Random Forest diagram[4].

[3] http://www.iis.ee.ic.ac.uk/icvl/iccv09_tutorial.html

[4] http://www.iis.ee.ic.ac.uk/icvl/iccv09_tutorial.html

- Bagging Discriminant: The strategy of the ensemble is to create determined number of bootstrap replicate of the training set and to train them with (in our case) LDC. Bagging is the other great ensemble strategy together with boosting. Again, the most successful classifier in LANDAU (LDC) is used in this ensemble.
- Regularized Discriminant Classifier: This is an ensemble that could use the bagging technique and the algorithms that train the replicate of the training data are a set of classifier constructed as a linear combination of Linear Discriminant Classifier and Quadratic Discriminant Classifier (Maximum Likelihood), being both of them the first and the last of them (figure 10). This methodology is not very popular in the literature, just maybe because is not implemented in the main algorithms toolbox. The idea of using this ensemble came because these two single classifiers (Linear Discriminant Classifier and Quadratic Discriminant Classifier) get the highest accuracy in the LANDAU project. Hence, we thought that an ensemble where every classifier is a combination of them (linear in this case) should throw good results.

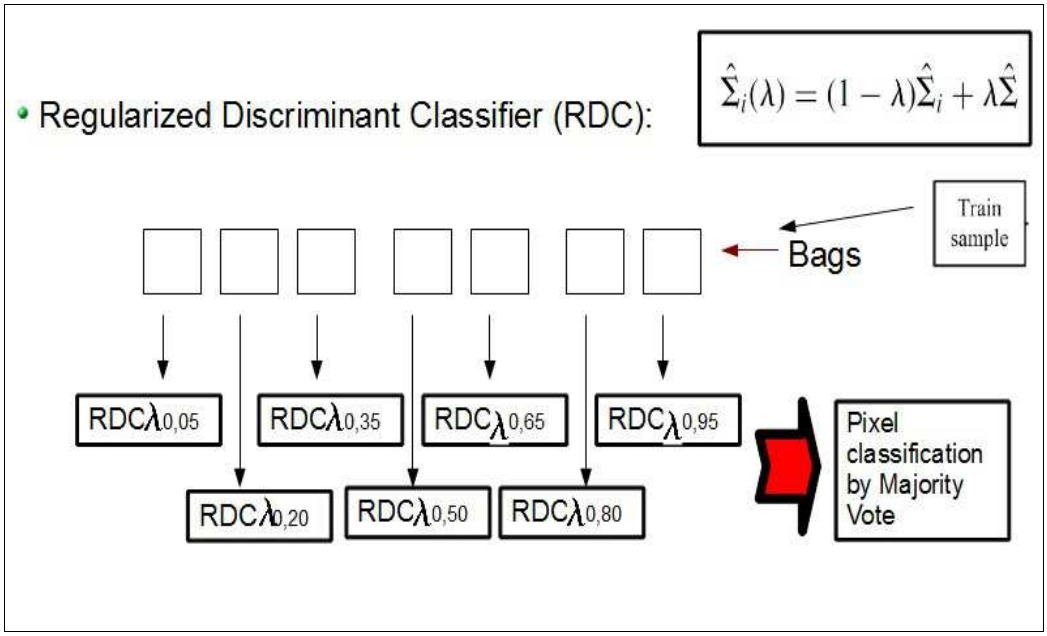

Figure 10. Regularized Discriminant Classifier (RDC) diagram.

- SVM Ensemble Strategy: this methodology was introduced by Waske et al. in 2010, and is based in a combination of Support Vector Machine (SVM)

which are trained with different features (image bands), also known as feature selection (section 2.2). In this study (Waske et al., 2010) the ensemble bring good results and also, we try a sketch implementation of a single SVM and shows good results, hence we decide to include this methodology in our work.

3.2.3 Difference between proportions

Once, we obtain the values of accuracy (global, in this case) it is necessary to know if those values are significant or not. In others words, if the increase of the accuracy values that we are supposed to get with the ensembles proposed, will improve significantly those results obtained with single classifiers in the LANDAU project.
In our case, it only makes sense to compare the values of accuracy of our ensembles with the best single classifier (higher accuracy) in LANDAU, which is LDC.
In Kuncheva (2004) they are proposed some methods, from which the most used are the McNemar-test and the Z-test. Finally, the McNemar test was used instead of the Z-test, due to the hypothesis of independence between proportions is violated (Dietterich, 1998), since the testing set used in both studies, LANDAU project and this dissertation, was the same.

3.2.4 Diversity between Classifiers outputs

Diversity is one the hottest topics at this moment in the classifiers ensemble researching field. This novelty entails great ideas and advances, but also confusion and no consistent basis is built underneath. In this sense, many diversity measures are developed and applied to solve the same problem. Which is the most appropriate approach? It is something that it is still not clear. There is an intuition about what it is the contribution of diversity (Kuncheva and Whitaker, 2003), but not ground truth, which is the most important concern about diversity. How diversity should be measure? Which is the most reliable type of measures? What are the thresholds in within the values of diversity of the classifiers should be considered?
In this environment of uncertainty, we will try to apply one the most applied diversity measures, Double Fault, in order to prove the relation between the diversity of the classifiers and the higher value of accuracy.

The strategy followed will be to analyse only those ensembles with a significant increase of accuracy. Then apply a pairwise measure (Double Fault, in this case) between a relevant amount of the classifiers within the ensemble and analyse the values obtained (mean, maximum value, minimum).

Double-Fault measure, used also by Giacinto and Roli (2001), is based on the concept that is more important to know when simultaneous errors are committed rather than when both classifiers are correct (Kuncheva, 2004).

As an example of how to apply a pairwise measure to a ensemble of various classifiers, we will take the RDC ensemble. It is built by 200 classifiers as linear combination of LDC and ML. We need to compare the classifiers pairwise; hence the number of comparison needed to get the value is $200*^2$, which is 40000 comparisons. We take a sample of them, 1% of the population, 400 comparisons.

These values could be seen as a clue to go further or not in our deliberations and to implement non-pairwise measures for the whole ensemble, which seem to be more appropriate. The reason to start by pairwise measures is a matter of time. They are much easier to implement and analyse than non-pairwise measures.

4 Results and discussions

In this chapter will be presented, firstly the results obtained from the analysis of the data and their processing, the findings discovered in the interpretation of the variables. Difficulties founded in the development of the methodology will be commented as well and, finally, the comparison of accuracy between the ensembles.

4.1 Data analysis and processing

As it was referred before the data set, a Landsat image from a continental area of As it was referred before the data set, a Landsat image from a continental area of Portugal (in the Alentejo region) coincides with one of those which were used in the LANDAU project. In this project a variety of single classifiers were used to classify the satellite image and global accuracies obtained vary from 76 to 82 (table 1 and 2), being the Linear Discriminant Classifier (LDC) the one which presents better performance.

When analysing those results more deeply, including user's and producer's accuracies, it could be observed some phenomena which could be a hint in the further ensemble construction. We can observe in the following tables (table 1 and 2) the performance of the algorithms in the classification of the different classes.

ACCURACY MATRIX (PRODUCER)	ML	LDC	DQDC	KNN	PARZEN	CART	BMP	TOTAL ACCURACY BY CLASS
1.1 Artificial Discontinuous Areas	71	74	33	84	84	82	85	73
2.1 Irrigated agriculture	79	98	91	95	95	93	89	91
2.2 Rain-fed agriculture	71	64	79	51	50	49	67	62
2.3 Rice fields	81	62	83	81	81	79	76	78
3.1 Deciduous Forest	83	69	66	71	73	70	79	73
3.2 Broad-leaves Forest	93	95	79	86	86	88	93	89
3.4 Grassland	84	93	87	86	86	81	82	86
3.5 Shrub	77	84	75	72	72	66	83	76
4 Bare Soil	97	100	100	97	97	94	94	97
5 Wet lands	86	71	53	67	67	79	0	60
6 Water bodies	86	97	92	87	87	87	97	90
TOTAL ACCURACY BY CLASSIFIER	81	82	76	78	78	77	79	79

Table 1. User's accuracy of LCLU maps from the most representative single classifiers (built from LANDAU project data).

ACCURACY MATRIX (USER)	ML	LDC	DQDC	KNN	PARZEN	CART	BMP	TOTAL ACCURACY BY CLASS
1.1 Artificial Discontinuous Areas	86	73	94	63	63	56	70	72
2.1 Irrigated agriculture	80	72	81	84	84	81	75	80
2.2 Rain-fed agriculture	77	82	66	86	86	82	78	80
2.3 Rice fields	92	90	70	79	79	83	69	80
3.1 Deciduous Forest	81	88	73	71	71	70	79	76
3.2 Broad-leaves Forest	89	80	74	71	73	72	78	77
3.4 Grassland	76	77	73	74	75	76	80	76
3.5 Shrub	91	88	83	82	82	78	91	85
4 Bare Soil	63	89	71	79	79	78	79	77
5 Wet lands	72	79	78	82	82	87	0	69
6 Water bodies	94	93	93	93	93	95	83	92
TOTAL ACCURACY BY CLASSIFIER	81	82	76	78	78	77	79	78

Table 2. Producer's accuracy of LCLU maps from the most representative single classifiers (built from LANDAU project data).

In the table 3, it can be seen the meaning of the colours in the tables 1 and 2. Thus, the red areas symbolize the classes where the algorithms have had a great performance. On the contrary, yellow areas mean very performance of the classifier. Orange, white and light brown colours represent the stages in between.

	VERY HIGH ACCURATE CLASSIFICATION (> 90%)
	HIGH ACCURATE CLASSIFICATION (90%-85%)
	MEDIUM ACCURATE CLASSIFICATION (85%-75%)
	LOW ACCURATE CLASSIFICATION (75%-70%)
	VERY LOW ACCURATE CLASSIFICATION (< 70%)

Table 3. Legend of colour in the error matrices.

From table 2, it can be observed that the performance of Maximum Likelihood Classifier (ML) and LDC is better in most of the classes than the rest of algorithms (majority of red and orange cells in their columns), which could be seen as a hint to build an ensemble from them. Artificial Neural Network (ANN), Parzen classifier and Classification and Regression Tree (CART) obtain the highest accuracy values in some classes, which it could be thought as an advantage to build an ensemble using these algorithms.

Table 1 shows that in some classes, like Water bodies, Bare soil and Irrigated agriculture the values of accuracy are high to most of the classifiers, which is a proof of intern invariability within these classes. Further conclusions could be taken from special combination of individual classes and algorithms, like the high performance of DQDC in Artificial Discontinuous Areas or BMP in Shrubs in Table 2. These findings could be used to fine-tune an ensemble (adding this algorithm to the

ensemble) or when the focus of the study is to map one those classes and not all of them.

4.2 Ensemble results and significance

The selection of ensembles used in this work has been carried out by literature selection, choosing those multiple classifier systems more successful in the scientific literature, and also guided by the analysis of our dataset. Hence, in the following lines we will describe the results associated to any of these ensembles.
In the table below (table 4), it could be seen a summary of the best global accuracies reached by all the ensembles and a comparison with results obtained in the LANDAU project, using single classifiers.

Single Classifier		Overall Accuracy	Ensemble Type		Overall Accuracy
LANDAU			DISSERTATION		
ML		81,00	Boosting Discriminant		83,40
LDC		**82,00**	Boosting Trees		83,45
DQDC		76,00	Bagging Discriminant		80,40
K-NN		78,00	Random Forest		83,20
PARZEN		78,00	RDC		**85,60**
CART		77,00	SVM Ensemble		82,50
BMP		79,00			

Table 4. Summarize of the ensembles better results.

In the figure below (figure 11) it can be seen a graphic in which are represented the global accuracies of both groups of classifiers, single classifiers used in the LANDAU project (represented in blue at the bottom of the graphic) and the classifiers ensembles used in this dissertation (represented in red at the top of the graphic). We can observe how generally the group of ensembles get higher accuracy than the group of classifiers used in LANDAU. We can also see how five out of six ensembles get a higher value of accuracy than the best single classifier, Linear Discriminant Classifier (LDC), the light blue line in the bottom of the graphic. The light red line on the top of the graphic makes reference to the higher global accuracy reached in this work, by the RDC ensemble.

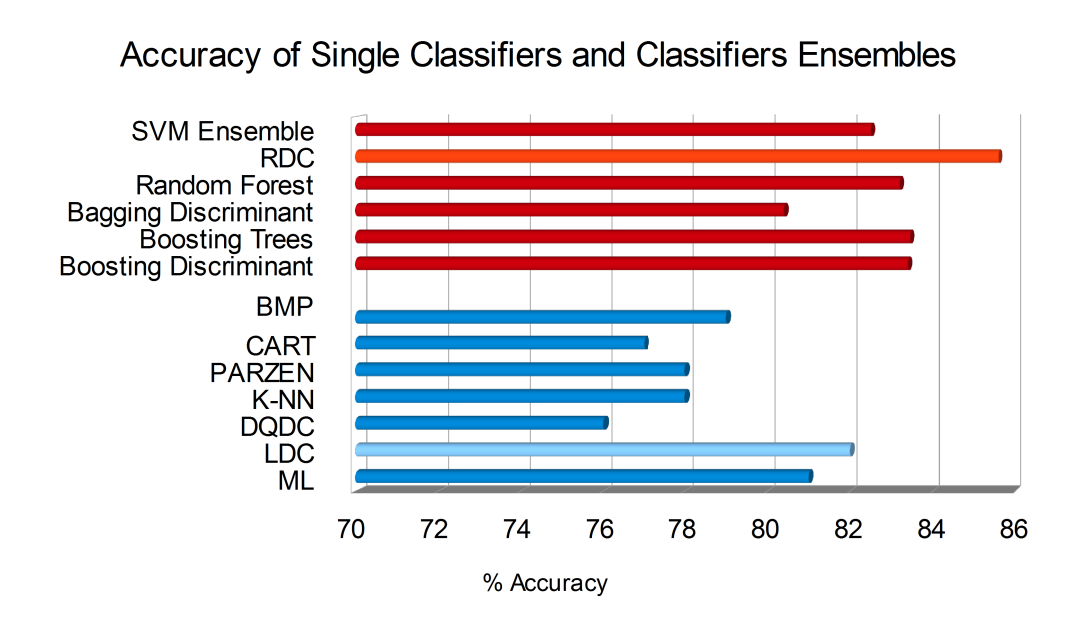

Figure 11. Global Accuracy of Classifiers used in LANDAU and Ensembles used in this dissertation.

In general, all the classifiers ensembles have had a better performance than the single classifiers apart from the Bagging Discriminant ensemble, which get a global accuracy of 80,4 %, being lower than the accuracy obtained by LDC and ML, 82 and 81 percent, respectively. RDC gets the higher value of accuracy, 85,6 %, almost 4 percent higher than the best single classifier, LDC.

The Bagging Discriminant classifier, as we said before, got the lower value, according to Breiman (1996), the cause of this poor performance could be that Bagging is effective on unstable learning algorithms where small changes in the training set result in large changes in predictions, an example of these unstable learning algorithms are neural networks and decision trees. LDC is extremely stable; hence no good results should be expected from this multiple classifier system. However, for data sets where the number of cases is small and the number of features is large, LDC is no longer stable because small changes in the training set might lead to large changes of the classifier. Bagging and “nice bagging” have been found to work for unstable LDC (Kuncheva, 2004).

In this context of bagging, the ensemble Random Forest is a bagging of "Trees", which are unstable classifiers and it was expected to get a higher results than the Bagging Discriminant, then almost 3 more point of accuracy were obtained. Random Forest have been one the most popular ensembles through the last decade.

Boosting Discriminant and Boosting Trees ensembles show good results, both around 83,5 % of accuracy. In the case of Boosting Discriminant, the ensemble was built by LDC algorithms. In the case of the Boosting Trees, the higher results were showed using one hundred classifiers Trees. Those results obtained by applying boosting are better by themselves, but also the boosting strategy avoids creating an over-fitted classifier.

RDC, Regularized Discriminant Classifier, is the ensemble with better results (figure 12).

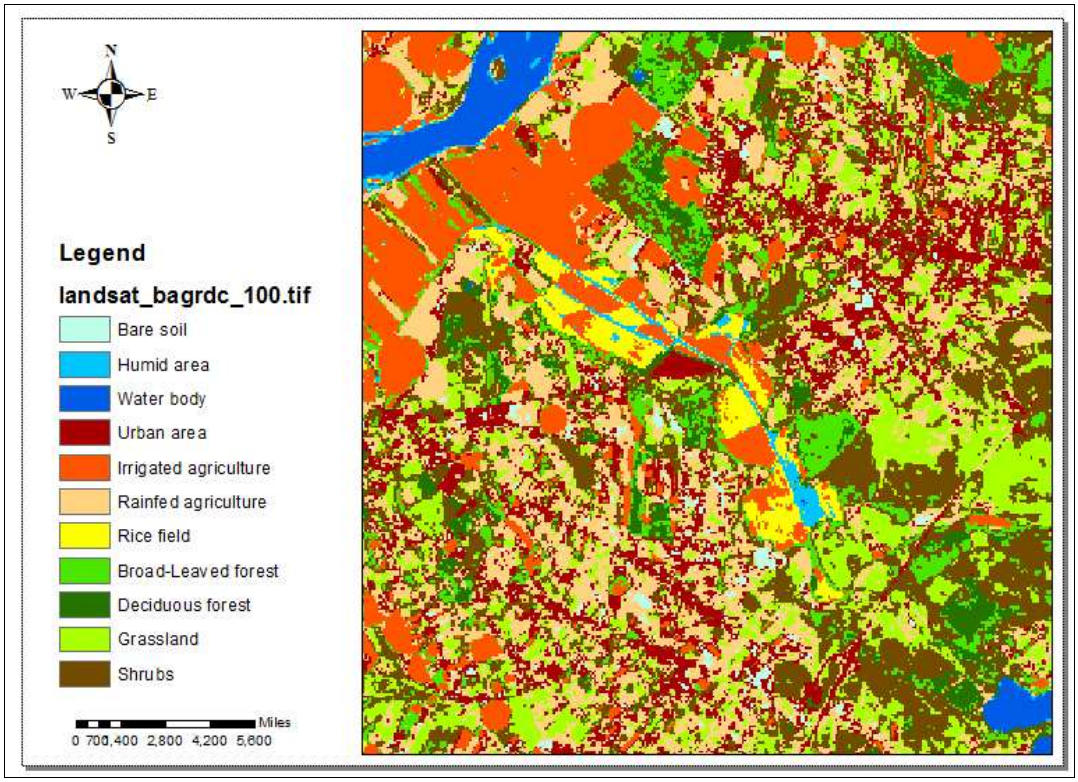

Figure 12. LCLU map, from Landsat image classified by RDC ensemble.

Initially, the idea was to build a RDC ensemble based in the bagging strategy. After the results obtained in the Bagging Discriminant Classifier, we understood that this strategy was not appropriate to this algorithm. Hence, the last version of the ensemble was a two hundred classifiers ensemble, all of them trained in the same dataset and with feature (bands) selection of five out of seven. The main reason to choose this ensemble was that the performance of the single classifiers LDC and ML (type of Quadratic Discriminant Classifier), in the LANDAU project was the best. Hence, intuitively we decided to use an ensemble in which every classifier is a linear combination of both.

Finally, we decide to apply a feature selection over an ensemble of Support Vector Machine (SVM) algorithms. This structure based Waske's work (Waske et al., 2010) try to gain diversity from the feature selection process. SVM is an algorithm which has been widely used in satellite image classification (Mountrakis et al., 2011) for general land cover and land use tasks. Firstly, we tried to implement a single SVM algorithm, and we got an accuracy of 83,45%, which take us to think than an ensemble of them could perform even better .But it did not work in this way and an ensemble of SVM with feature selection, only got an accuracy of 82,36 %. We faced some problem in the programming step of the development of the ensemble and we are aware that probably the structure of it is not the most suitable to get better results. We propose for further studies to go deeply in this model, by better implementing the code.

Once we get the percentages of accuracy of all the methods used, we tested which of them add a significant increase of accuracy in the classification of the image. The difference of proportion methods are used to accomplish this task. As we explain in a previous section (section 3.2.3) we used the McNemar test. In the table below (table 5), it is show the proportion between the accuracy of the best the single classifiers from the LANDAU project , LDC with a global accuracy of 82 %, and all the ensembles built in this work.

COMPARISON OF CLASSIFIERS ENSEMBLES: McNEMAR TEST							
	LDC (Best LANDAU)	Boosting Discr.	Boosting Trees	Bagging Discr.	Random Forest	RDC	SVM Ensemble
G. Accuracy	82,00%	83,40%	83,45%	80,40%	83,20%	85,60%	82,36%
	THE VALUE IN THE TEST HAS TO BE HIGHER THAN **3,841.**						
Comparison		Boosting Discr.	Boosting Trees	Bagging Discr.	Random Forest	RDC	SVM Ensemble
LDC	0	**0,7**	**0,75**	**(-) 1,33**	**0,35**	**9,5**	**0,1**

Table 5. McNemar test between all the ensembles and the best single classifier.

McNemar Test assess that the value of the test between two classifiers (in our case a single classifier and classifiers ensemble) have to be higher than 3,841, to consider significant the difference of accuracy between them.

In our case, only the RDC ensemble obtains a higher value than 3,841, which is 9,5, so only the difference of accuracy showed by this ensemble is significant in relation to the best single classifier, LDC. It means that we have only increased significantly

the value of accuracy for image classification of this dataset with the RDC ensemble. In this case, was completely unworthy to spend time in building any classifiers ensemble, apart from the RDC ensemble. It is something to have into account, because not all the classifiers ensemble get significantly better results than single classifiers, and sometimes is more recommended not to spend time and resources in building an ensemble, if the results are not considerably better.

4.3 Diversity measures

As it was said before, diversity is still a field which need to be explored more scientifically, and it will be a good purpose for further research to analyse all these ensembles and theirs diversity using different measure, to try to find some conclusions in one direction or the opposite.

As we explain in section 3.2.4, we only tried one measure, Double-Fault measure. We have applied the measure just to one ensemble, the only one with a significantly higher value of accuracy than the best single classifier used in the LANDAU project (LDC). This ensemble is RDC and by applying this diversity measure, we wanted to check if its success was because its high value of diversity.

Also, it was exposed in the section 3.2.4 that the RDC ensemble was built by 200 classifiers as linear combination of LDC and ML, which entails 40000 pairwise comparisons between them. We take a sample of those comparison and we get the results showed in the table below (table 6).

DIVERSITY MEASURES				
Double Fault	It is a **Pairwise Measure** RDC with 200 CLASSIFIERS entails 40000 Pairwise Measures (Comparison), So we took a Sample (1%, **400 MEASURES**).			
	MIN VALUE	MAX VALUE	MEAN VALUE	**CONCLUSIONS:**
	14,50%	46,00%	22,91%	LOW VALUES OF DIVERSITY BETWEEN CLASSIFIERS
Generalized Diversity???	It is **Non-Pairwise Measure** **No** need for using it. The values of Double Fault may indicate its **uselessness.**			

Table 6. Double-Fault diversity measures results.

In this test, the double-Fault measure of diversity between classifiers, we observed that the diversity between the classifiers within the ensemble, is not too high, being

the highest score recorded 46 out of 100 (high values of diversity are considered from 80). Hence, we can conclude that diversity does not explain the success of the RDC ensemble when classifying a Landsat image.

At least, this diversity measure does not show good results, as it used to happen in many studies (Kuncheva, 2003; Dutta, 2009; Faria et al., 2013). It is probable that the utilization of a non-pairwise diversity measure, as the Generalized Diversity measure, to quantify the diversity of the RDC ensemble, could be more suitable for this study, but the approach and the coding process was too complex to go into it in this work. Going further in those diversity measures is a researching line which could be very interesting in the future.

The approach of the Double-Fault measure was done with the intention to test randomly the diversity between classifiers in the RDC ensemble. It has been prove to be effective and very straightforward to implement. Since not stimulating diversity values were obtained, we refused to go further in our research and not to try another diversity measures.

5 Conclusions

The initial objective of this work was to prove that a classifiers ensemble can perform better than strong single classifiers in the task of classification of remote sensing images. In most of the cases we have analysed the results were better than the results from the best single classifier (LDC) used in the LANDAU project. Of course, we are referring in terms of accuracy, which not always means in terms of map quality. We will analyse this later on in this section.

One of the statement that have to be more in our minds is the "No Free-Lunch Theorem", which basically set up that there is no optimal solution for every circumstances, there is no optimal classifier, or in this case, classifiers ensemble, which fit for every data set. Hence, no further conclusions could be taken from this work, apart from that this data set (a Landsat image) is best classified by the RDC ensemble. But we do not know the behaviour of this ensemble in other circumstances.

Another statement, which made the classifiers ensemble to be in the centre of the machine learning research, is the possibility of building a strong classifier from a combination of weak classifier. This is the base of the Boosting theory developed by Schapire.

In our case, the results from the Boosting Discriminant and Boosting Trees ensembles were better than those from the single classifiers but they were not spectacular, as could be thought at the beginning of the process. Boosting was thought for weak classifiers, those that show an accuracy of 50% or less (Schapire, 1996), and “our” single classifiers were not that weak, since they have around 80% of accuracy. This reason could explain better but no spectacular results.

We can assess that our results for those ensemble which follow the Bagging strategy, give consistency to the statement of Breiman which establishes that the bagging strategy for ensemble works better for unstable classifiers (Breiman, 1996), those where small changes in the training set result in large changes in predictions, as our Trees (CART). In fact, the performance of the bagging LDC was worse than the performance of the LDC, as a single classifier. Like Kuncheva also relate in his book (Kuncheva, 2004), no good results could be expected from a multiple linear discriminant system using bagging.

One of the deceptions of this work was the performance of the SVM ensemble with feature selection. When we run the SVM as a single classifier, we got a high accuracy value, sometimes higher than the one get with LDC (82%), which guides us to think in better results when applying a group of them in the same system. Unfortunately the output was even worse than the one obtained with the simple algorithm. We are sure that we missed something in the process of coding the algorithm, because the ensemble seems to be very powerful (Waske, 2011). We were, also trying to fine-tune the ensemble but the structure of the SVMs is very hard to understand (like the different kind of kernel's parameter, sort margin parameter and alpha value). Hence, finally we did not have time to better develop a SVM ensemble. However, further work could be done in this direction, since the strength of this methodology could be very high.

Once, we obtained the results of accuracy of our ensembles, the next step was to confirm how good were our scores and testing the significance of them in comparison with the best of the accuracy mark in the single classifiers. We applied one of the most used differences of proportion methods, the McNemar Test, to check the importance of the output of the classifiers ensemble trained in this work. As we argued before, McNemar Test was selected because of the character of the validation dataset. The McNemar Test showed up that the only important value of accuracy from all the ensembles that were tested in this work is the one achieve by the RDC ensemble, 85,6 %.

Then the following question to answer is why the RDC ensemble is the one which get a higher value of global accuracy rather than the other multiple classifier systems.

We slightly faced this question under the point of view of the diversity of the ensemble. As we see before, we applied the Double-Fault measure to the RDC ensemble to figure out if the diversity between their two hundreds classifiers was high enough to explain the value of global accuracy. The result was negative, the values of diversity were far away from top values. Even, as we thought that the Double-Fault measure was not the best approach to calculate the diversity, because is a pairwise measure, not actually design for ensembles, it calculate the diversity very straightforward and could be taken as an indicator. In fact, after getting these extremely low values of diversity we gave up to apply another kind of measure. It could be an interesting topic for further investigation, because we still have

problems in understanding the concept of diversity. It sounds intuitive (Kuncheva, 2003) that having a group of different classifiers which produce different outputs should better than having a group of similar classifier which produce the same outputs, in the last case it does not make sense to build an ensemble, since one of them could show up the same results.

If we have in consideration that many ensembles work better than a single classifier, we should think that the variation that different algorithms add to the whole system must be positive. Under our perspective, two problems appear, when we deal with the concept diversity, the thresholds and the measures. Obviously, it seems clear that certain degree of variance between the elements of an ensemble will give to it more power to predict, but which are these thresholds within the diversity, where the ensemble gain strength. As an example, two completely different algorithms that produce extremely different outputs do not create a powerful ensemble, but the contrary. Hence, how much different should be the algorithm between them within the ensemble is still a mystery.

There are many measures approved by the scientific community, most of them come from Kuncheva and Whitaker (2003), to calculate the diversity between classifiers or within the ensemble, and many more are being developed at present, but we consider that there are still not science over there. A lot of mysteries, a lot of incongruences, a lot of eager to measure something that even is not completely explained. How can you find in the literature for the last ten years, more than ten different measure to calculate the same thing? It seems to us that many efforts are being done in this direction to clarify the term and the way to used and measure, but it is still insufficient.

Then, to finally answer the question about why the RDC ensemble produces the best outputs, we were analysing how the algorithms within the ensemble work. RDC or RDA, as it was defined by Friedman (1989), is a regularized version of the discriminant analysis. The ensemble of RDC takes many linear combinations from LDC to QDC as defined. So the objects, in our case pixels, could be classified by using a common covariance to all of them as happen in the LDC, a unique covariance to each of them like occur in the QDC or everything which is between of both classifiers. So, we can argue that the success of this ensemble resides in the versatility to build covariance matrices à la carte, in order to explain the behaviour of the variables and also do better prediction.

Finally, having a look over the maps that were produced by using the all the classifiers ensembles, including the map generated from the best single classifier (LDC) and a google map image of the study area (figure 13), we can see how the differences on classification are independent of the degree of accuracy.

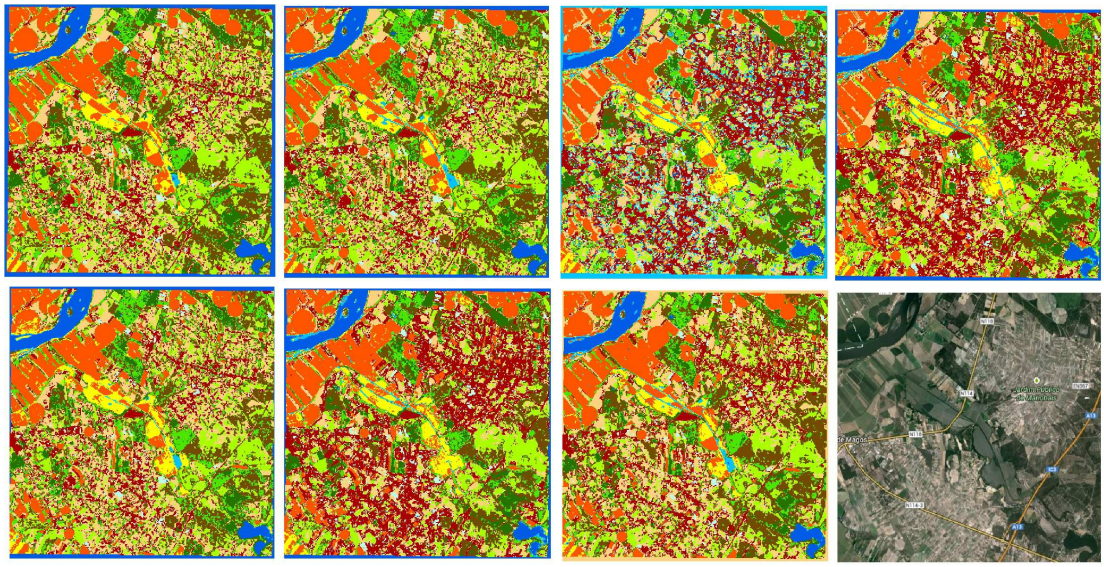

Figure 13. LCLU maps, from Landsat image classified by all ensembles analysed and google map image. From left to right and from up to the bottom, following the scores of global accuracy from lower to higher these maps belong to: Bagging Discriminant (1), LDC (2), SVM ensemble (3), Random Forest (4), Boosted Discriminant (5), Boosted Trees(6) and RDC ensemble (7), respectively.

Looking at the maps, it is hard to say which is the map with the best appearance. At a first glance, it could be said that the second, the fourth and the seventh maps (following the order in the figure 13) show a more compact structure. Their global accuracies values are 82%, 83,2% and 85,6%, respectively.

For further research it would be taken in consideration a deep study about diversity and the different diversity measures, and also a better implementation of the SVM ensemble by using feature selection, which seems to offer great results.

Bibliographic References:

- Breiman, Leo. "Bagging predictors." *Machine learning* 24.2 (1996): 123-140.
- Breiman, Leo. "Random forests." *Machine learning* 45.1 (2001): 5-32.
- Campbell, J. B. “Landuse and Land Cover Inventory.” *Manual of Phothographic Interpretation (2nd edition).* American Society of Photogrammetry and Remote Sensing (2001):335-364
- Cingolani, Ana M., et al. "Mapping vegetation in a heterogeneous mountain rangeland using Landsat data: an alternative method to define and classify land-cover units." *Remote sensing of environment* 92.1 (2004): 84-97.
- CORINE Land Cover Report. Commission of the European Communities. 1995
- Di Gregorio, A. "FAO land use statistics: A case study for three countries using remote sensing and GIS technology." *Consultancy Report for FAO Statistics Division, Rome* (1995).
- Di Gregorio, A., and Jansen, L.J.M. 1996a. Part I - Technical document on the Africover Land Cover Classification Scheme. FAO. Africover Land Cover Classification (1997): 4-33; 63-76.
- Di Gregorio, A., and Jansen, L.J.M. Land Cover Classification System (LCCS): Classification concepts and user manual. FAO (2000).
- Dietterich, T.G., "Approximate statistical tests for comparing supervised classification learning algorithms." *Neural computation* 10.7 (1998): 1895-1923.
- Dinis, J., Gonçalves, M., Nicolau, R. and Reis, R. Proposta de uma Nomenclatura de Ocupaçâo do Solo. Relatório de execuçâo do projecto LANDAU, Action 3.1 – Task 3. Lisboa, Instituto Geográfico Portuguès (2012).
- Dinis, J., Rodrigues, P. and Nicolau, R.. Clasificaçâo uni-temporal da ocupaçâo do solo através de dados de elevada reloluçâo espacial - LANDSAT. Relatório de execuçâo do projecto LANDAU, Action 3.2 – Task 3. Lisboa, Instituto Geográfico Portuguès (2012).
- Du, Peijun, et al. "Multiple classifier system for remote sensing image

classification: a review." *Sensors (Basel, Switzerland)* 12.4 (2012): 4764.

- Dutta, Haimonti. "Measuring Diversity in Regression Ensembles." *IICAI*. Vol. 9. 2009.
- Faria, Fabio A., et al. "Classifier Selection based on the Correlation of Diversity Measures: When Fewer is More." *Graphics, Patterns and Images (SIBGRAPI), 2013 26th SIBGRAPI-Conference on*. IEEE, 2013.
- Foody, G.M., “Status of land cover classification accuracy assessment”. Remote Sensing of Environment, 80 (2002): 185–201.
- Foody, G. M., et al. "Training set size requirements for the classification of a specific class." *Remote Sensing of Environment* 104.1 (2006): 1-14.
- Foody, G. M., and Mathur, A.. “A relative evaluation of multiclass image classification by support vector machines”. IEEE Transactions on Geoscience and Remote Sensing, 42 (2004):1335−1343.
- Friedl, Mark A., Carla E. Brodley, and Alan H. Strahler. "Maximizing land cover classification accuracies produced by decision trees at continental to global scales." *Geoscience and Remote Sensing, IEEE Transactions on* 37.2 (1999): 969-977.
- Freund, Yoav, and Robert E. Schapire. "Experiments with a new boosting algorithm." *ICML*. Vol. 96. 1996.
- Giacinto G. and Roli F.. “Design of effective neural network ensembles for image classification processes". Image Vision and Computing Journal, 19 (2001): 699–707.
- Hansen, Lars Kai, and Peter Salamon. "Neural network ensembles." *Pattern Analysis and Machine Intelligence, IEEE Transactions on* 12.10 (1990): 993-1001.
- Huang, Y. S., and C. Y. Suen. "The behavior-knowledge space method for combination of multiple classifiers." *Computer Vision and Pattern Recognition, 1993. Proceedings CVPR'93., 1993 IEEE Computer Society Conference on*. IEEE, 1993.
- Hughes, G. "On the mean accuracy of statistical pattern recognizers". *Information Theory, IEEE Transactions on* 14.1 (1968): 55-63.
- Jacobs, Robert A., et al. "Adaptive mixtures of local experts." *Neural computation* 3.1 (1991): 79-87.

- Jensen, John R. *Introductory digital image processing: a remote sensing perspective*. No. Ed. 2. Prentice-Hall Inc., 1996.
- Kuncheva, Ludmila I. *Combining Pattern Classifiers: Methods and Algorithms*. Wiley. com, 2004.
- Kuncheva, Ludmila I., and Christopher J. Whitaker. "Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy."*Machine learning* 51.2 (2003): 181-207.
- Lawrence, R., Bunn, A., Powell, S., and Zambon, M.. Classification of remotely sensed imagery using stochastic gradient boosting as a refinement of classification tree analysis. *Remote sensing of environment, 90* (2004): 331-336.
- Liew, S.C.. "Principles of Remote Sensing". *Centre for Remote Imaging Sensing and Processing. National University of Singapore. 2001*
- Lu, D., and Q. Weng. "A survey of image classification methods and techniques for improving classification performance." *International journal of Remote sensing* 28.5 (2007): 823-870.
- Magnussen, S., P. Boudewyn, and M. Wulder. "Contextual classification of Landsat TM images to forest inventory cover types." *International Journal of Remote Sensing* 25.12 (2004): 2421-2440.
- Martínez, A. M., and Avinash C. K. "Pca versus Lda". *Pattern Analysis and Machine Intelligence, IEEE Transactions on* 23.2 (2001): 228-233.
- Meyer, William B., and Billie L. Turner. "Human population growth and global land-use/cover change." *Annual review of ecology and systematics* 23 (1992): 39-61.
- Mountrakis, G., Im, J., and Ogole, C.. Support vector machines in remote sensing: A review. *ISPRS Journal of Photogrammetry and Remote Sensing*,*66* (2011): 247-259.
- Pohl, Cle, and J. L. Van Genderen. "Review article multisensor image fusion in remote sensing: concepts, methods and applications." *International journal of remote sensing* 19.5 (1998): 823-854.
- Polikar R. (2009) Ensemble learning. Scholarpedia, 4(1):2776., revision #91224
- Price, K.P., Guo, X. and Stiles, J.M., Optimal Landsat TM band

combinations and vegetation indices for discrimination of six grassland types in eastern Kansas. International Journal of Remote Sensing, 23 (2002): 5031–5042.

- Rajan, Suju, Joydeep Ghosh, and Melba M. Crawford. "An active learning approach to hyperspectral data classification." *Geoscience and Remote Sensing, IEEE Transactions on* 46.4 (2008): 1231-1242.
- Roli, Fabio, Giorgio Giacinto, and Gianni Vernazza. "Comparison and combination of statistical and neural network algorithms for remote-sensing image classification." *Neurocomputation in Remote Sensing Data Analysis*. Springer Berlin Heidelberg, (1997): 117-124.
- Schapire, Robert E. "The strength of weak learnability." *Machine learning* 5.2 (1990): 197-227.
- Schilling, Keith E., et al. "Quantifying the effect of land use land cover change on increasing discharge in the Upper Mississippi River." *Journal of Hydrology*387.3 (2010): 343-345.
- Shafer, Glenn. *A mathematical theory of evidence*. Vol. 1. Princeton: Princeton university press, 1976.
- Song, C., Woodcock, C.E., Seto, K.C., LenneY, M.P. and Macomber, S.A., Classification and change detection using Landsat TM data: when and how to correct atmospheric effect. Remote Sensing of Environment, 75 (2001): 230–244.
- Waske, B., van der Linden, S., Benediktsson, J. A., Rabe, A., and Hostert, P. "Sensitivity of support vector machines to random feature selection in classification of hyperspectral data." *Geoscience and Remote Sensing, IEEE Transactions on* 48.7 (2010): 2880-2889
- Wolpert, David H. "Stacked generalization." *Neural networks* 5.2 (1992): 241-259.
- Wolpert, David H., and William G. Macready. "No free lunch theorems for optimization." *Evolutionary Computation, IEEE Transactions on* 1.1 (1997): 67-82.
- Toutin, T., Geometric processing of remote sensing images: models, algorithms and methods. International Journal of Remote Sensing, 25, (2004): 1893–1924.

- Townshend, JR G. "Land cover." *International Journal of Remote Sensing* 13.6-7 (1992): 1319-1328.
- Townshend, John, et al. "Global land cover classification by remote sensing: present capabilities and future possibilities." *Remote Sensing of Environment* 35.2 (1991): 243-255.
- Tuia, Devis, et al. "A survey of active learning algorithms for supervised remote sensing image classification." *Selected Topics in Signal Processing, IEEE Journal of* 5.3 (2011): 606-617.
- Tumer, Kagan, and Joydeep Ghosh. "Analysis of decision boundaries in linearly combined neural classifiers." *Pattern Recognition* 29.2 (1996): 341-348.
- Zhang, Jixian. "Multi-source remote sensing data fusion: status and trends."*International Journal of Image and Data Fusion* 1.1 (2010): 5-24.
- Zhang, Y., Baoguo W., and Dong W.. "Research Dynamics of the Classification Methods of Remote Sensing Images." Asian Agricultural Research 5.03, 2013.
- Zhou, Z.H. “Ensemble Learning”, National Key Laboratory for Novel Software Technology, Nanjing University (2004).

Annex I – Confusion Matrix of the Classifiers Ensembles

ACCURACY MATRIX	ML	LDC	DQDC	KNN	PARZEN	CART	BMP	TOTAL ACCURACY BY CLASS
1.1 Artificial Discontinuous Areas	78	73	63	73	74	69	78	73
2.1 Irrigated agriculture	80	85	86	90	89	87	82	86
2.2 Rain-fed agriculture	74	73	73	66	68	65	73	70
2.3 Rice fields	87	76	77	80	80	77	72	78
3.1 Deciduous Forest	82	78	79	71	72	70	79	76
3.2 Broad-leaves Forest	91	88	77	79	80	80	86	83
3.4 Grassland	80	85	80	80	81	79	81	81
3.5 Shrub	84	86	79	77	77	72	87	80
4 Bare Soil	80	95	86	88	90	86	87	87
5 Wet lands	79	75	67	74	75	83	0	65
6 Water bodies	90	95	93	90	90	91	90	91
TOTAL ACCURACY BY CLASSIFIER	81	82	76	78	78	77	79	79

Table 7. Global accuracy of LCLU maps of more representative single classifiers (built from LANDAU project data).

Error Matrix		Bare Soil	Wetlands	Water body	Urban Area	Irrigated agri.	Rainfed agri.	Rice field	Broad-leav. F	Deciduous F	Grassland	Shrub	User's ACC
	0	4	6	7	12	21	22	23	31	32	34	35	
Bare Soil	4	35	0	0	0	0	2	0	0	0	0	0	0,946
Wetlands	6	0	15	0	0	0	0	2	0	0	0	0	0,882
Water body	7	0	6	65	0	0	0	0	0	0	0	0	0,915
Urban Area	12	0	0	0	38	0	11	0	0	0	0	0	0,776
Irrigated agri.	21	0	1	0	0	42	0	9	3	0	0	0	0,764
Rainfed agri.	22	0	0	0	9	0	50	0	0	0	2	0	0,820
Rice field	23	0	12	0	0	0	0	27	0	1	0	0	0,675
Broad-leav. F	31	0	0	0	1	1	0	0	27	4	1	0	0,794
Deciduous F	32	0	0	3	0	0	0	2	9	37	0	1	0,712
Grassland	34	0	0	0	0	0	7	1	1	0	53	9	0,746
Shrub	35	0	0	0	1	1	1	0	2	1	4	53	0,841
Prod's ACC		1,000	0,441	0,956	0,776	0,955	0,704	0,659	0,643	0,860	0,883	0,841	**0,804**
													Overall ACC

Table 8. Error matrix of LCLU map get from Bagging Discriminant ensemble.

Error Matrix		Bare Soil	Wetlands	Water body	Urban Area	Irrigated agri.	Rainfed agri.	Rice field	Broad-leav. F	Deciduous F	Grassland	Shrub	User's ACC
	0	4	6	7	12	21	22	23	31	32	34	35	
Bare Soil	4	35	0	0	0	0	2	0	0	0	0	0	0,946
Wetlands	6	0	15	0	0	0	0	2	0	0	0	0	0,882
Water body	7	0	6	65	0	0	0	0	0	0	0	0	0,915
Urban Area	12	0	0	0	38	0	11	0	0	0	0	0	0,776
Irrigated agri.	21	0	1	0	0	42	0	9	3	0	0	0	0,764
Rainfed agri.	22	0	0	0	9	0	50	0	0	0	2	0	0,820
Rice field	23	0	12	0	0	0	0	27	0	1	0	0	0,675
Broad-leav. F	31	0	0	0	1	1	0	0	27	4	1	0	0,794
Deciduous F	32	0	0	3	0	0	0	2	9	37	0	1	0,712
Grassland	34	0	0	0	0	0	7	1	1	0	53	9	0,746
Shrub	35	0	0	0	1	1	1	0	2	1	4	53	0,841
Prod's ACC		1,000	0,441	0,956	0,776	0,955	0,704	0,659	0,643	0,860	0,883	0,841	**0,804**
													Overall ACC

Table 9. Error matrix of LCLU map get from Boosting Discriminant ensemble.

Error Matrix		Bare Soil	Wetlands	Water body	Urban Area	Irrigated agri.	Rainfed agri.	Rice field	Broad-leav. F	Deciduous F	Grassland	Shrub	User's ACC
	0	4	6	7	12	21	22	23	31	32	34	35	
Bare Soil	4	34	0	0	0	0	2	0	0	0	0	0	0,944
Wetlands	6	0	36	5	0	0	0	0	0	0	0	0	0,878
Water body	7	1	3	52	0	0	0	0	0	0	0	0	0,929
Urban Area	12	0	0	0	45	0	20	0	0	0	2	0	0,672
Irrigated agri.	21	0	0	0	0	39	0	4	3	0	0	0	0,848
Rainfed agri.	22	0	0	1	3	0	47	0	0	0	0	0	0,922
Rice field	23	0	3	1	0	1	0	36	0	0	0	0	0,878
Broad-leav. F	31	0	0	0	0	1	0	0	31	2	1	0	0,886
Deciduous F	32	0	0	1	0	0	0	3	6	39	0	2	0,765
Grassland	34	0	0	0	0	0	3	1	1	0	54	13	0,750
Shrub	35	0	0	0	1	0	0	0	2	2	3	46	0,852
Prod's ACC		0,971	0,857	0,867	0,918	0,951	0,653	0,818	0,721	0,907	0,900	0,754	**0,835**
													Overall ACC

Table 10. Error matrix of LCLU map get from Boosting Trees ensemble.

Error Matrix		Bare Soil	Wetlands	Water body	Urban Area	Irrigated agri.	Rainfed agri.	Rice field	Broad-leav. F	Deciduous F	Grassland	Shrub	User's ACC
	0	4	6	7	12	21	22	23	31	32	34	35	
Bare Soil	4	32	0	0	0	0	3	0	0	0	0	1	0,889
Wetlands	6	0	36	6	0	0	0	1	0	0	0	0	0,837
Water body	7	1	3	52	0	0	0	0	0	0	0	0	0,929
Urban Area	12	2	0	0	46	0	19	0	0	0	1	0	0,676
Irrigated agri.	21	0	0	0	1	41	0	5	2	0	0	0	0,837
Rainfed agri.	22	0	0	1	1	0	46	0	0	0	0	0	0,958
Rice field	23	0	3	0	0	0	0	35	1	0	0	0	0,897
Broad-leav. F	31	0	0	0	0	0	0	0	31	1	1	3	0,861
Deciduous F	32	0	0	1	0	0	0	2	6	40	0	1	0,800
Grassland	34	0	0	0	0	0	4	1	1	0	55	12	0,753
Shrub	35	0	0	0	1	0	0	0	2	2	4	43	0,827
Prod's ACC		0,914	0,857	0,867	0,939	1,000	0,639	0,795	0,721	0,930	0,902	0,717	**0,831**
													Overall ACC

Table 11. Error matrix of LCLU map get from Random Forest ensemble.

Error Matrix		Bare Soil	Wetlands	Water body	Urban Area	Irrigated agri.	Rainfed agri.	Rice field	Broad-leav. F	Deciduous F	Grassland	Shrub	User's ACC
	0	4	6	7	12	21	22	23	31	32	34	35	
Bare Soil	4	34	0	0	0	0	2	0	0	0	0	0	0,944
Wetlands	6	0	29	0	0	0	0	5	0	0	0	0	0,853
Water body	7	0	4	63	0	0	0	0	0	0	0	0	0,940
Urban Area	12	0	0	0	44	0	7	0	0	0	0	0	0,863
Irrigated agri.	21	0	1	0	0	42	0	7	3	0	0	0	0,792
Rainfed agri.	22	0	0	1	5	0	57	0	0	0	4	0	0,851
Rice field	23	0	2	0	0	0	0	26	0	0	0	0	0,929
Broad-leav. F	31	0	1	0	0	0	0	0	34	2	1	1	0,872
Deciduous F	32	0	0	1	0	0	0	3	3	38	0	1	0,826
Grassland	34	0	0	0	0	0	5	1	1	0	51	8	0,773
Shrub	35	0	0	0	1	1	0	0	1	3	4	53	0,841
Prod's ACC		1,000	0,784	0,969	0,880	0,977	0,803	0,619	0,810	0,884	0,850	0,841	**0,856**
													Overall ACC

Table 12. Error matrix of LCLU map get from RDC ensemble.

Annex II – Ensembles implementation Code, an example.

RDC ensemble code, implemented in Matlab.

1. Algorithm definition:

```
function outclass = rdc(sample,training,group,type,prior,alpha)
% Input 'type' is not to be used
type = [];
% grp2idx sorts a numeric grouping var ascending, and a string
grouping
% var by order of first occurrence
[gindex,groups,glevels] = grp2idx(group);
nans = find(isnan(gindex));
if ~isempty(nans)
    training(nans,:) = [];
    gindex(nans) = [];
end
ngroups = length(groups);
gsize = hist(gindex,1:ngroups);
nonemptygroups = find(gsize>0);
nusedgroups = length(nonemptygroups);
if ngroups > nusedgroups
    warning(message('stats:classify:EmptyGroups'));
end
[n,d] = size(training);
if size(gindex,1) ~= n
    error(message('stats:classify:TrGrpSizeMismatch'));
elseif isempty(sample)
    sample = zeros(0,d,class(sample));  % accept any empty array but
force correct size
elseif size(sample,2) ~= d
    error(message('stats:classify:SampleTrColSizeMismatch'));
end
m = size(sample,1);

% if nargin < 4 || isempty(type)
%     type = 'linear';
% elseif ischar(type)
%     types =
{'linear','quadratic','diaglinear','diagquadratic','mahalanobis'};
%     type = internal.stats.getParamVal(type,types,'TYPE');
% else
%     error(message('stats:classify:BadType'));
% end

% Default to a uniform prior
if nargin < 5 || isempty(prior)
    prior = ones(1, ngroups) / nusedgroups;
    prior(gsize==0) = 0;
    % Estimate prior from relative group sizes
elseif ischar(prior) && strncmpi(prior,'empirical',length(prior))
    %~isempty(strmatch(lower(prior), 'empirical'))
    prior = gsize(:)' / sum(gsize);
```

```
    % Explicit prior
elseif isnumeric(prior)
    if min(size(prior)) ~= 1 || max(size(prior)) ~= ngroups
        error(message('stats:classify:GrpPriorSizeMismatch'));
    elseif any(prior < 0)
        error(message('stats:classify:BadPrior'));
    end
    %drop empty groups
    prior(gsize==0)=0;
    prior = prior(:)' / sum(prior); % force a normalized row vector
elseif isstruct(prior)
    [pgindex,pgroups] = grp2idx(prior.group);

    ord = NaN(1,ngroups);
    for i = 1:ngroups
      j = find(strcmp(groups(i), pgroups(pgindex)));
        if ~isempty(j)
            ord(i) = j;
        end
    end
    if any(isnan(ord))
        error(message('stats:classify:PriorBadGrpup'));
    end
    prior = prior.prob(ord);
    if any(prior < 0)
        error(message('stats:classify:PriorBadProb'));
    end
    prior(gsize==0)=0;
    prior = prior(:)' / sum(prior); % force a normalized row vector
else
    error(message('stats:classify:BadPriorType'));
end
% Add training data to sample for error rate estimation
if nargout > 1
    sample = [sample; training];
    mm = m+n;
else
    mm = m;
end

gmeans = NaN(ngroups, d);
for k = nonemptygroups
    gmeans(k,:) = mean(training(gindex==k,:),1);
end

% Linear
% computed without unpermuting.  Instead use SVD to find rank of R.
[Q,R] = qr(training - gmeans(gindex,:), 0);
R = R / sqrt(n - nusedgroups); % SigmaHat = R'*R

% Quadratic
D = NaN(mm, ngroups);
logDetSigma = zeros(n,1);
for k = nonemptygroups
    [Q,Rk] = qr(bsxfun(@minus,training(gindex==k,:),gmeans(k,:)),
0);
    Rk = Rk / sqrt(gsize(k) - 1); % SigmaHat = R'*R
    % the average between Rk and R
    Rk = alpha*Rk + (1-alpha)*R;
```

```
    s = svd(Rk);
    if any(s <= max(gsize(k),d) * eps(max(s)))
        error(message('stats:classify:BadQuadVar'));
    end
    logDetSigma(k) = 2*sum(log(s)); % avoid over/underflow
    A = bsxfun(@minus, sample, gmeans(k,:))/Rk;
    D(:,k) = log(prior(k)) - .5*(sum(A .* A, 2) + logDetSigma(k));
end
% find nearest group to each observation in sample data
[maxD,outclass] = max(D,[],2);
%Convert outclass back to original grouping variable type
outclass = glevels(outclass,:);
end
```

2. Accuracy Assessment:

```
load('../data/mat/landsat_a.mat');

%% The inputs
nparts = 200;

pfeat = 0.8;
ptrain = 1.0;

freplace = 0; % feature replacement
treplace = 1; % training unit replacement

%% Processing

ntrain = size(datatr,1);
k = floor(pfeat*size(datatr,2)); % should be <= n. of features, i.e.
nbands
m = floor(ptrain*size(datatr,1)); % n. of training samples

% alpha ranging from 0 to 1 in 0.01
alpha = [];
factor = 1/nparts;
for i = 0:nparts
   alpha = [alpha; factor * i];
end

outclasses = [];
for i = 1:length(alpha)
    fidx = randsample(nbands,k,freplace);
    tidx = randsample(ntrain,m,treplace);
    outclass = rdc(datats(:,fidx),datatr(tidx,fidx),labeltr(tidx),
[],[],alpha(i));
    outclasses = [outclasses outclass];
end

outclass = mode(outclasses,2);

%% Validate the model
labelf = [];
for i = 1:length(labelts)
```

```
    if outclass(i)==labelts(i,1) || outclass(i)==labelts(i,2)
        labelf = [labelf; outclass(i)];
    else
        labelf = [labelf; labelts(i,1)];
    end
end

[C codes] = confusionmat(outclass,labelf);

OA = sum(diag(C))/sum(C(:));
PA = diag(C)'./sum(C,1);
UA = diag(C)./sum(C,2);

EM = C;
EM = [EM UA];
EM = [EM; PA OA];
EM = [[codes; 0] EM];
EM = [0 codes' 0; EM];
```

3. LCLU map creation:

```
%% Input data

matname = '..\data\mat\Landsat_A.mat';

load(matname);

%% Process stuff

% This is my random forest function
map = rdc(image,datatr,labeltr,[] ,5 ,0.005);

% reshape this into a map format
map = reshape(map,nrows,ncols);

%% Output it

imwrite(uint8(map),'..\data\single_maps\landsat_RDC1.tif','tif');
worldfilewrite(refmat, '../data/single_maps/landsat_RDC1.tfw');
```