ISPRS Journal of Photogrammetry and Remote Sensing 67 (2012) 93–104

Contents lists available at SciVerse ScienceDirect

ISPRS Journal of Photogrammetry and Remote Sensing

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

An assessment of the effectiveness of a random forest classifier for land-cover classification

V.F. Rodriguez-Galiano [a,*], B. Ghimire [b], J. Rogan [b], M. Chica-Olmo [a], J.P. Rigol-Sanchez [c]

[a] Dept. de Geodinámica, Universidad de Granada, Granada 18071, Spain
[b] Graduate School of Geography, Clark University, Worcester, MA, USA
[c] Dept. de Geología, Universidad de Jaén, Jaén 23071, Spain

ARTICLE INFO

Article history:
Received 15 March 2011
Received in revised form 31 October 2011
Accepted 3 November 2011
Available online 1 December 2011

Keywords:
Remote sensing
Machine learning
Classification
Random forest
Land-cover
Landsat Thematic Mapper

ABSTRACT

Land cover monitoring using remotely sensed data requires robust classification methods which allow for the accurate mapping of complex land cover and land use categories. Random forest (RF) is a powerful machine learning classifier that is relatively unknown in land remote sensing and has not been evaluated thoroughly by the remote sensing community compared to more conventional pattern recognition techniques. Key advantages of RF include: their non-parametric nature; high classification accuracy; and capability to determine variable importance. However, the split rules for classification are unknown, therefore RF can be considered to be black box type classifier. RF provides an algorithm for estimating missing values; and flexibility to perform several types of data analysis, including regression, classification, survival analysis, and unsupervised learning.

In this paper, the performance of the RF classifier for land cover classification of a complex area is explored. Evaluation was based on several criteria: mapping accuracy, sensitivity to data set size and noise. Landsat-5 Thematic Mapper data captured in European spring and summer were used with auxiliary variables derived from a digital terrain model to classify 14 different land categories in the south of Spain. Results show that the RF algorithm yields accurate land cover classifications, with 92% overall accuracy and a Kappa index of 0.92. RF is robust to training data reduction and noise because significant differences in kappa values were only observed for data reduction and noise addition values greater than 50 and 20%, respectively. Additionally, variables that RF identified as most important for classifying land cover coincided with expectations. A McNemar test indicates an overall better performance of the random forest model over a single decision tree at the 0.00001 significance level.

1. Introduction

Land-cover mapping and monitoring is one of the major applications of Earth observing satellite sensor data and is essential for the estimation of land cover change. Large scale land cover monitoring is important because human and/or natural land cover modifications affect biophysical and biogeochemical properties of land surfaces (Bala et al., 2007; Betts et al., 2007; Bonan, 2008; Brovkin et al., 2004). Large area monitoring is used to estimate land-cover change and deforestation, perform forest inventory, and determine priority areas for biodiversity conservation (Lambin et al., 2001; Mas et al., 2004; Turner et al., 2007). Additionally, changes in land-cover affect the climate through changes in the composition of carbon dioxide and other greenhouse gases in the atmosphere (Bala et al., 2007; Betts et al., 2007; Bonan, 2008; Brovkin et al., 2004; Fearnside, 2000). Thus, many applications rely on reliable and timely land-cover mapping products over large heterogeneous landscapes.

Increased numbers of satellite sensor images have made it easier to establish land-cover monitoring programs for large area mapping over regular time intervals (Friedl et al., 2002). Operational large area land monitoring programs are well established (Franklin and Wulder, 2002). Unfortunately, there are several limitations related to large area monitoring that need to be resolved. First, large area mapping in complex landscapes is difficult because of abrupt changes in environmental gradients (e.g. moisture, elevation and temperature) and a legacy of past disturbance (Rogan and Miller, 2006). Such heterogeneous landscapes are characterized by land-cover categories that are difficult to separate spectrally due to low inter-class separability and high intra-class variability. Second, large area mapping requires algorithms that can be interpreted

* Corresponding author. Tel.: +34 958 243363; fax: +34 958 248527.
E-mail addresses: vrgaliano@ugr.es (V.F. Rodriguez-Galiano), bghimire@clarku.edu (B. Ghimire), jrogan@clarku.edu (J. Rogan), mchica@ugr.es (M. Chica-Olmo), jprigol@ujaen.es (J.P. Rigol-Sanchez).

0924-2716/$ - see front matter © 2011 International Society for Photogrammetry and Remote Sensing, Inc. (ISPRS) Published by Elsevier B.V. All rights reserved.
doi:10.1016/j.isprsjprs.2011.11.002

readily and automated as well as operated easily due to user-defined parameters that are simple to adjust. Third, the choice of a suitable land-cover classification algorithm for large area mapping depends on the ability of the algorithm to handle noisy observations, a complex measurement space, and a small number of training data relative to the size of the study area (DeFries and Chan, 2000; Rogan et al., 2008).

A variety of classification methods have been used to map land cover using remotely sensed data. Classification methods range from unsupervised algorithms such as ISODATA or *K*-means to parametric supervised algorithms such as maximum likelihood (Jensen, 2005); to machine learning algorithms such as artificial neural networks (Mas and Flores, 2008), decision trees (Breiman, 1984), support vector machines (Mountrakis et al., 2011) and ensembles of classifiers (Breiman, 1996). In the last five years, machine learning algorithms have emerged as more accurate and efficient alternatives to conventional parametric algorithms, when faced with large dimensional and complex data spaces and have been used for large area mapping (Hansen et al., 1996; Huang et al., 2002; Rogan et al., 2003). These algorithms are efficient and effective because they do not rely on data distribution assumptions (e.g. normality) and generally have higher accuracy (Foody, 1995; Friedl and Brodley, 1997). However, some machine learning techniques (e.g. neural networks and support vector machines) are complicated due to the large number of parameters that need to be adjusted and are difficult to automate (Atkinson and Tatnall, 1997; Foody, 2004). Additionally these algorithms have a tendency to over-fit the data (Breiman et al., 1984).

An emerging type of machine learning technique which utilizes ensembles of classifications (e.g. neural network ensembles, random forests, bagging and boosting) is receiving highlighted interest (Friedl et al., 1999; Ghimire et al., 2010; Gislason et al., 2006; Hansen and Salamon, 1990; Krogh and Vedelsby, 1995; Sesnie et al., 2008; Steele, 2000). Ensemble learning algorithms use the same base classifier to produce repeated multiple classifications of the same data (Breiman, 2001; Friedl et al., 1999), or use a combination of different base classifiers to generate multiple classifications of the same data or to target different subsets of the data (Mountrakis et al., 2009). The collection of multiple classifiers of the same data are combined using a rule based approach (such as, maximum voting, product, sum, and Bayesian rule), or based on an iterative error minimization technique by reducing the weights for the correctly classified samples (e.g. boosting) (Friedl et al., 1999; Ghimire et al., 2010; Steele, 2000). Ensemble learning techniques have higher accuracy than other machine learning algorithms because the group of classifiers performs more accurately than any single classifier, and utilizes the strengths of the individual group of classifiers while at the same time the classifier weaknesses are circumvented (Ghimire et al., 2010; Kotsiantis and Pintelas, 2004).

An ensemble learning technique called random forests is increasingly being applied in land-cover classification using multispectral and hyperspectral satellite sensor imagery (Chan and Paelinckx, 2008; Ghimire et al., 2010; Lawrence et al., 2006; Pal, 2005; Sesnie et al., 2008), and lidar and radar data (Guo et al., 2011; Latifi et al., 2010; Martinuzzi et al., 2009; Waske and Braun, 2009). However, most studies that have used random forests have focused on relatively small study areas (Pal, 2005; Waske and Braun, 2009), classified few land-cover classes (Gislason et al., 2006; Lawrence et al., 2006; Prasad et al., 2006), or only used single season imagery for classification (Chapman et al., 2010; Ghimire et al., 2010; Ham et al., 2005). Moreover, most studies have not investigated the behavior of the random forest classifier by assessing the influence of training data quality/noise and variations in training data set size on classifier performance (Chan and Paelinckx, 2008; Gislason et al., 2006; Ham et al., 2005; Lawrence et al., 2006; Pal, 2005; Prasad et al., 2006; Sesnie et al., 2008). The objective of this study was to assess the performance of the random forest classifier in a large heterogeneous landscape with diverse land-cover categories using multi-seasonal Landsat, and auxiliary data. The behavior of random forests is assessed by considering multiple criteria related to variations in classifier parameter values, and sensitivity to noise and training size variations. The performance of the RF is also evaluated in comparison to classification trees.

2. Study area

The Province of Granada (GP) is the study area chosen for this project. It is located in the south of Spain on the Mediterranean coast, encircled by the Penibetica mountain range (Fig. 1). This area occupies 12,635 km^2 and elevation ranges from sea level to the

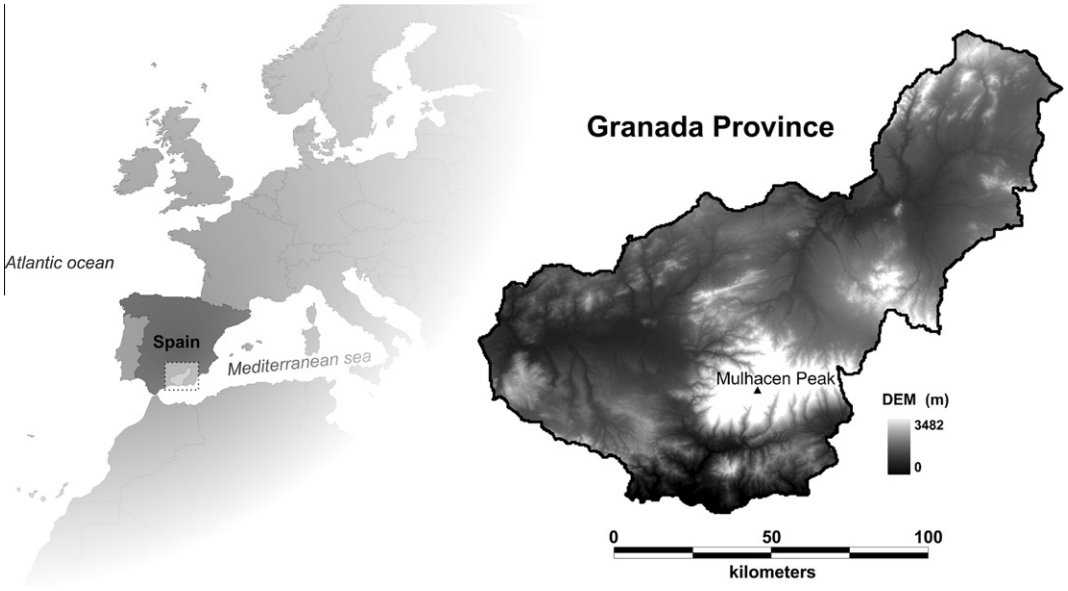

Fig. 1. Location of study area in Spain.

Mulhacen Peak (3482 m) in Sierra Nevada National Park. The climate of Granada is Mediterranean with a continental influence, characterized by hot and dry summers and wet and cold winters. Average annual temperatures range from 18 °C at the coast to 10 °C in the mountains. Climate ranges from arid to semi-arid (between 300 and 500 mm). The study area is composed of a variety of land-cover types, mainly including agriculture (46%), with tobacco and corn fields, olive trees, tropical crops and substantial greenhouse production. The remainder of the study area is characterized by the presence of upland conifer forest (18%), shrub-grasslands (22%) and oak grove (8%). The natural vegetation zone has the highest biodiversity in the Iberian Peninsula and Europe, with 66 endemic vascular species, which have remained unaltered due to substrate features, the extreme climate and their inaccessibility over time (Mota et al., 2002).

3. Methods

3.1. Satellite and ancillary data

Multi-temporal Landsat images are commonly employed to characterize phenological variation in the state of vegetal cover. Several studies have shown that a combination of multi-seasonal images can increase the separation between spectrally similar covers because it represents the phenological vegetation condition (Lunetta and Balogh, 1999; Oetter et al., 2001; Wolter et al., 1995; Yuan et al., 2005). In this study, spring and summer images have been employed (April and August, respectively) in the land-cover classification since these images contain most of the phenological variations (Brewster et al., 1999; Brisco and Brown, 1995; Pax-Lenney et al., 1996; Pax-Lenney and Woodcock, 1997; Schriever and Congalton, 1995). These two dates represent peaks in productivity of the phenological development of the major vegetation types in the area, which is critical for the accurate classification of land cover types. In summer images, annual crops (e.g. tobacco and corn) can be confused with conifer forests and poplar groves. Highly reflective surfaces, such as urban areas, can be confused with bare soils. The inclusion of spring images allows discrimination between annual crops and evergreen natural vegetation. On the other hand, soils, which remain bare during drought periods (summer), are usually covered by grass in spring time, which can facilitate differentiation from urban areas (Yuan et al., 2005). The definition of summer and spring is related to the Northern Hemisphere.

Two Landsat Thematic Mapper-5 scenes of the same area in southeast Spain were captured. The images were acquired on 18 August and 12 April 2004. Image location corresponds with path 200 row 34 of Landsat worldwide reference system (WRS), with coordinate center 0030822 W 372400 N WGS-84.

Images were corrected independently for geometric offset by using digital orthophotos of 1 m spatial resolution. 150 points for each image provided a third-order polynomial transformation with less than one-half pixel root mean square (RMS) error. Nearest neighbor resampling was chosen to preserve the original values of the pixels. The images were converted to radiance values, and then into reflectance values using a radiative transfer code based on MODTRAN4 (MODerate Resolution TRANsmittance) (Abreu and Anderson, 1996; Berk and Adler-Golden, 2002), FLAASH (Fast Line-of-sight Atmospheric Analysis of Spectral Hypercubes). The reflectance values obtained for both dates were rescaled from 0 to 255 (8 bit reflectance).

The images were enhanced spectrally using the Thomas linear transformation before being used in classification. This transformation produced six features: summer brightness, summer greenness, summer wetness, spring brightness, and both spring greenness and wetness.

Following Franklin (1998), several topographic variables were included as input variables to each RF classification: elevation, slope and aspect. These ancillary variables were derived from a 20 m resolution digital elevation model (DEM) and rescaled to the spatial resolution of the spectral variables (30 m).

3.2. Land-cover categories and reference data

The land cover of the Mediterranean can be very complex and challenging to classify (Berberoglu et al., 2000). Relief complexity and high anthropogenic influence results in a very heterogeneous landscape which makes it possible that 14 different thematic categories can be distinguished in the study area. The classification scheme was based on Andalusian land-cover maps (ALCM) developed in 2003 by the Andalusian Regional Government (Moreira-Madueño, 2006).

Jensen (2005) proposed that the number of training pixels should at least be equal to ten times the number of variables used in the classification model for a parametric classification approach. However, several studies have shown that non parametric machine learning algorithms need a larger number of training data in order to attain optimal results (Foody, 1995; Foody and Arora, 1997; Pal, 2005; Pal and Mather, 2003).

To create an exhaustive database with an optimal size for the training and accuracy assessment it was necessary to resort to auxiliary information due to the retrospective nature of this study. Reference data were obtained from a combination of a set of crop reference sites collected in the summer of 2004 and a stratified random sampling scheme using pre-existing land-cover maps (ALCM). More specifically, the ALCM was reclassified into 14 categories, and 150 sites were sampled randomly from each category. The digital true-color orthophotos (1:10000), corresponding to the sample sites and acquired during 2004, were then interpreted and 2100 sites were obtained. The ground reference dataset was divided randomly into 2/3 and 1/3 for training and testing, respectively. The number of the training sites per class was kept equal (100 training sites and 50 testing sites per land cover category). McCoy (2005) establishes that the minimum dimensions of the sample areas A should be estimated as $A = P(1 + 2L)$, where P is the ground sampling distance and L is the positional accuracy of the geometric registration in terms of pixels. In a multi-temporal context, when at least two image dates are used, the minimum dimension of a sample is equal to the mean value between time periods. In this study, the average between the area of the spring image, $A_{spring} = 30(1 + 2 * 0.51)$ and the area of the summer image $A_{summer} = 30(1 + 2 * 0.47)$ results in a minimum size of 59.4 by 59.4 which is roughly equal to 2 pixels. Thus, plot size was fixed on a 90 by 90 m sample area in a conservative way.

3.3. Random forest classification

Ensemble learning algorithms (e.g. random forest, bagging and boosting) have received increasing interest because they are more accurate and robust to noise than single classifiers, (Breiman, 1996; Dietterich, 2000). The philosophy behind classifier ensembles is based upon the basic premise that a set of classifiers do perform better classifications than an individual classifier does. Breiman suggested a new and promising classifier in (2001) called random forest, which presents many advantages for its application in remote sensing:

- It runs efficiently on large data bases.
- It can handle thousands of input variables without variable deletion.
- It gives estimates of what variables are important in the classification.

- It generates an internal unbiased estimate of the generalization error (oob error).
- It computes proximities between pairs of cases that can be used in locating outliers.
- It is relatively robust to outliers and noise.
- It is computationally lighter than other tree ensemble methods (e.g. Boosting).

A RF consists of a combination of classifiers where each classifier contributes with a single vote for the assignation of the most frequent class to the input vector ($\mathbf{x}$), $\hat{C}_{rf}^{B} = majority\,vote\left\{\hat{C}_b(\mathbf{x})\right\}_1^B$, where $\hat{C}_b(\mathbf{x})$ is the class prediction of the bth random forest tree. The fact that it is a combination of many classifiers confers RF some special characteristics which make it substantially different to a traditional classification trees (CT) and, therefore, it must be understood as a new concept of classifiers. A RF increases the diversity of the trees by making them grow from different training data subsets created through bagging or bootstrap aggregating (Breiman, 1996). Bootstrap aggregating is a technique used for training data creation by resampling randomly the original dataset with replacement (i.e., with no deletion of the data selected from the input sample for generating the next subset). Random forest (RF) is an ensemble classification algorithm which uses trees as base classifiers, $\{h(x,\Theta_k),\quad k=1,\ldots,\}$, where x is the input vector and $\{\Theta_k\}$ are the independent and identically distributed random vectors (Breiman, 2001; Hastie et al., 2009). Hence, some data may be used more than once in the training of classifiers, while others might never be used. Thus, greater classifier stability is achieved, as it makes it more robust when facing slight variations in input data and, at the same time, it increases classification accuracy (Breiman, 2001). Several studies have demonstrated that methods based on bagging such as RF, unlike other methods based on boosting, are not sensitive to noise or overtraining (Briem et al., 2002; Chan and Paelinckx, 2008; Pal and Mather, 2003).

Each subset selected using bagging to make each individual bth-tree grow usually contains 2/3 of the calibration dataset. The samples which are not present in the calibration subset are included as part of another subset called out-of-bag (oob). Note that a different oob subset is formed for every bth-tree, from the non-selected elements by the bootstrapping process. These oob elements, which are not considered for the training of the bth-tree, can be classified by the bth-tree to evaluate performance. At the end of the run, each input element of the oob subset has been classified on average by a third of the overall number of the trees generated in the ensemble (Peters et al., 2007). The proportion between the misclassifications and the total number of oob elements contributes an unbiased estimation of the generalization error (Breiman, 2001). The generalization error converges as the number of trees increases, therefore, the RF does not over fit the data. Furthermore, when the RF makes a tree grow, it uses the best split of a random subset of input features or predictive variables in the division of every node, instead of using the best split variables. Therefore, this can decrease the strength of every single tree, but it reduces the correlation between the trees, which reduces the generalization error (Breiman, 2001). Another characteristic of interest is that the trees of a RF classifier grow with no pruning, which makes it light, from a computational perspective.

Tree design requires choosing a suitable attribute selection measure which maximizes dissimilarity between classes. There are many approximations for selecting attributes which can be used for induction in decision trees. Some of the most frequent ones are gain-ratio (Quinlan, 1993), Gini Index (Breiman et al., 1984) and Chi-square (Mingers, 1989b). A RF usually uses the Gini Index as a measure for the best split selection, which measures the impurity of a given element with respect to the rest of the classes. For a given training dataset T, the Gini Index can be expressed as:

$$\sum\sum_{j\neq i}(f(C_i,T)/|T|)(f(C_j,T)/|T|) \tag{1}$$

Where $f(C_i,T)/|T|$ is the probability that a selected case belongs to class C_i. Thus, by using a given combination of features, a decision tree is made to grow up to its maximum depth (with no pruning). Studies suggest that it is choosing pruning methods, but not the election of attribute selection measures, which has a greater impact on the performance of tree-based methods (Mingers, 1989a; Pal and Mather, 2003). Hence, RF, as it grows without pruning, presents an added advantage.

This RF also provides an assessment of the relative importance of the different features or variables during the classification process. This aspect is useful for multi-source studies, where data dimensionality is very high, and it is important to know how each predictive variable influences the classification model to be able to select the best variables (Ghimire et al., 2010; Gislason et al., 2004; Gislason et al., 2006; Ham et al., 2005; Pal, 2005). To assess the importance of each feature (e.g. satellite image band), the RF switches one of the input random variables while keeping the rest constant, and it measures the decrease in accuracy which has taken place by means of the oob error estimation and of Gini Index decrease (Breiman, 2001).

Finally, a RF can also produce a measure of proximity between each pair of cases. To calculate the proximity between two samples of the same class, the RF counts the number of times the said samples appear at the same terminal node (i.e., how many trees label each possible pair of cases of the same class with the same division rule). Once each tree has been built, and the proximities are computed for each pair of cases, they are normalized by dividing by the number of trees. Proximities are susceptible to being used in replacing missing data and locating outliers (i.e., mislabelled sites into training sets).

3.3.1. Performance of the random forest classifier

To study the performance of the RF algorithm for land-cover classification, different spectral and auxiliary variables were used. On the one hand, spectral variables consisted of the Kauth Thomas multi-seasonal components of the summer and spring images. On the other hand, the auxiliary variables included in the analysis were elevation, slope and aspect, derived from the digital terrain model (Section 3.1).

The RF classifier only needs the definition of two parameters for generating a prediction model: the number of classification trees desired, (k), and the number of prediction variables, (m), used in each node to make the tree grow. In other words, to classify a new dataset a constant number of k random predictive variables is used, and each of the examples of the dataset is classified by a k number of trees defined by the user. This way the final value of the class assigned to each example will be equal to the most frequent value for the total number of k trees generated.

Breiman (1996) suggested that when increasing the number of trees the generalization error always converges and over-training is not a problem due to the "Strong Law of Large Numbers" (Feller, 1968). On the other hand, reducing the number of predictive variables (m) causes each individual tree of the model to be less strong, but also reduces the correlation between trees, which increases the model's accuracy. Taking this into account, it is necessary to optimize the parameters k and m to minimize the generalization error.

There are several commercial and open source implementations for RF model development (Breiman and Cutler, 2004; Liaw and Wiener, 2002; Witten and Frank, 2005). In this study, the randomForest package within the statistical software R 2.10.1 was used. The R implementation of RF (Liaw and Wiener, 2002) contains a function called tuneRF which will automatically select the optimal

value of m with respect to oob correct classification rates. We did not use this function, in part because there is no research as yet to assess the effects of choosing RF parameters such as m to optimize oob error rates on the generalization error rates for RF using remotely sensed data. The performance of RF was compared to a classification tree using the rpart package within the statistical software R2.10.1. The classification tree was pruned based on cost complexity and the tree with the lowest error was selected, to reduce data overfitting. Choice of the classification tree with the lowest error ensured that noise was reduced from the data.

4. Results and discussion

4.1. Effect of the number of trees (k) and predictive variables (m) on the classifier's accuracy

Most conventional statistical classification methodologies (i.e., maximum likelihood) and machine learning methodologies (i.e., neural networks, support vector machines) create complex decision rules from many variables which they use simultaneously to classify a pixel. However, classifiers using trees label pixels considering one or several variables. Taking this into account, classification algorithms can be classified into univariate algorithms, if the decision is made from a single variable, or multivariate, if it is made from a synergy of several variables (Pal and Mather, 2003). A RF is an algorithm based on classification trees; therefore it can be applied both in a univariate and multivariate way. As explained in Section 3.3, a RF brings in an additional parameter which does not appear in traditional classification trees: the m parameter. In each node a subset of m predictive variables, from 1 to the maximum (9 in this case study) has to be specified. This value of m remains constant while the tree is growing, and variable selection is made randomly. Hence the definition of this parameter affects both the correlation and strength of each individual tree and, therefore, it also affects the generalization error and the classifier's accuracy (Breiman, 2001).

To assess the optimal value of m, numerous RF models were created, each of them made up of 1000 trees for the different possible values of m to divide nodes (from 1 to 9). Fig. 2 shows the error depending on the number of trees for m equal to the minimum and maximum (1 and 9, respectively). A distinction is made between two different error measurements: those estimated from the test subset and those estimated from the oob-subset. The oob estimate for the generalization error is the error rate of the oob

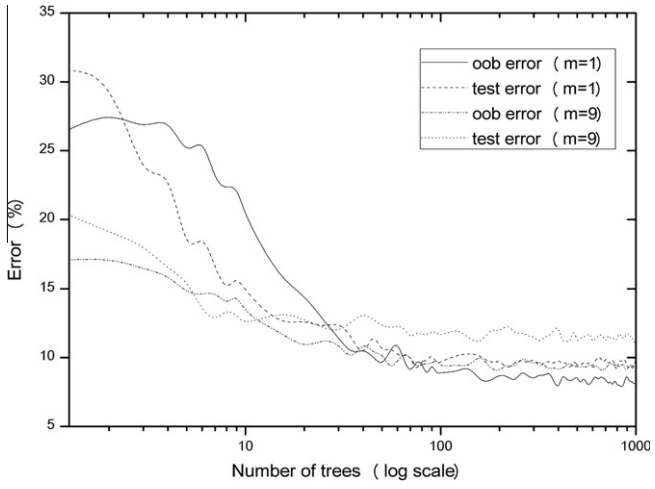

Fig. 2. Effect of number of trees (k) and random split variables (m) on oob and test errors.

classifier of the training set. It is demonstrated in studies carried out by Breiman (1996, 2001) that the oob error is a good estimator of the generalization error depending on the number of trees. It can be seen how from approximately 100 trees, the oob error converges 8% and 9% of the times for m equal to 1 and 9, respectively. The addition of more trees neither increases nor decreases the generalization error. Very small values of k resulted in lower classification performance, larger values of k resulted in more stable classifications and variable importance measures. The results of this analysis show that the differences in stability after 100 trees are very small and the computation time increases for larger ranges of possible values of k. The other two curves show the error in the test set depending on the number of trees and they represent the proportion of test elements predicted incorrectly. Errors in the test set were approximately of 9% and 11% for m equal to 1 and 9, respectively at the end of the construction of the RF models. Similarly as in the case of the oob error, the test set error converges from tree number 100. Since the error rate decreases as the number of combinations increases, the oob estimates will tend to overestimate the current error rate.

RF does not provoke an over-adjustment when adding up more decision trees to the model, so in order to obtain unbiased estimates of the oob error it is possible to run past the point where the test set error converges (Breiman, 2001).

The average of the absolute differences between the estimation of the oob and test errors was 1.8% and 2% for 1 and 9 random variables, respectively. Therefore the oob estimation is as accurate as the estimation based on the test subset (Breiman, 1996, 2001). However, unlike the test estimation, the oob estimation is unbiased. Taking this into account, it is advisable to use the oob estimation as an error internal measurement and, therefore, it is not necessary to use an independent test dataset regarding the training set of the algorithm. A RF, therefore, provides a comparative advantage compared to other algorithms for the classification of areas in which the number of data is reduced.

The average of the absolute differences of the oob estimation between the minimum and maximum of the possible random variables (1 and 9) from tree number 100 is less than 1%, from which it can be inferred that a RF is not sensitive to the value of m once the point has been reached in which the error converges.

4.2. Selection of optimal parameters for random forest model calibration

A RF combines the classifications made by many individual decision trees. For the training of this type of model it is necessary not only to define the space dimensionality of random features, but also the number of trees to be built based on the decision limits generated by said division variables.

To establish the optimal value of these parameters, a number of experiments were carried out using a different number of trees and a different number of split variables. The number of trees ranged between 1 and 1000 and the number of split variables was fixed from 1 to 9, using intervals of 1. This gave a total number of 9000 different RF prediction models for the classification of the study area. The resulting models were evaluated using the Kappa index (Congalton and Green, 2009). For the selection of the most accurate classification model we determined the one in which the Kappa index was highest, and the number of split variables lowest. This served as a guarantee for the accuracy of the maps, while avoiding correlation between the trees. Fig. 3 shows the relationship between the number of trees and split variables, which were used for training each RF model, and the classifier accuracy for the test dataset. The relationship between k and m and the Kappa index is directly proportional until a certain number of trees (k) is achieved from which Kappa index remains stable. Breiman

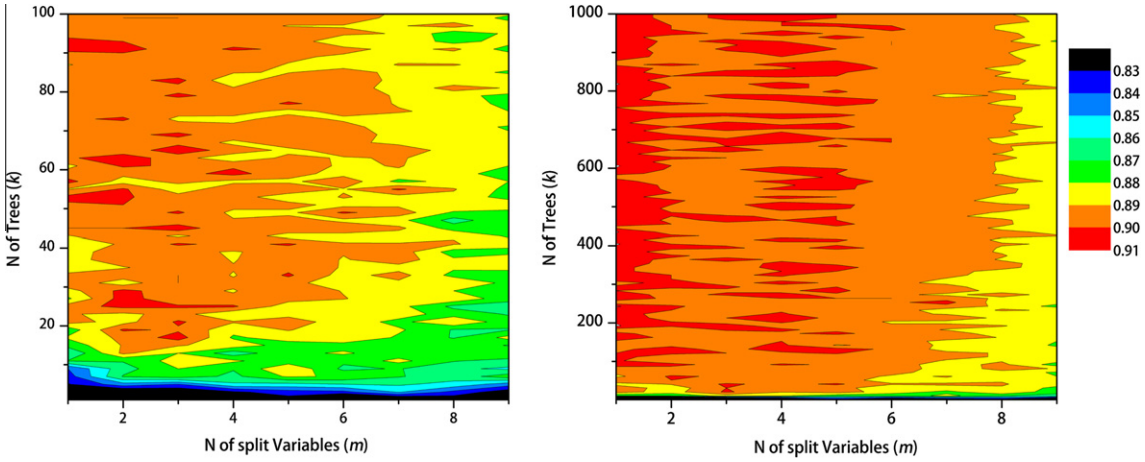

Fig. 3. Results of Kappa Index with relation to the number of trees (k) and the number of random split variables (m).

(2001) demonstrated that RFs do not overfit. Thus, a limiting value of the generalization error is obtained as more trees are added until a threshold value is reached. On the other hand, the value of m, which is constant during forest growth, affects both the correlations between the trees and the strength of the individual trees. Reducing m reduces correlation and strength, increasing m increases both. Taking this into account it is preferable to use a large number of trees (k) and a small number of split variables (m) to reduce the generalization error and the correlation between trees. The strength of the individual trees is not a problem since they are being used in an ensemble way.

The use of a single randomly chosen input variable to split on at each node produced higher levels or accuracy than using several. Therefore, based on the above findings a RF model made up of 479 univariate decision trees was adopted. Hence, even though it has been described in several studies that multivariate decision trees outperform univariate trees when individual trees are being used (Breiman et al., 1984; Pal and Mather, 2003; Utgoff and Brodley, 1990), for this study case, considering a multiple classification tree scenario, it was preferable to combine the classifications made on the basis of univariate decisions of different trees.

4.3. Variable importance by global and per-class context

The layout of the split variables in a decision tree provides information about the importance of the features in the general classification model and in the classification of each category (Pal and Mather, 2003). However, it is almost impossible to carry out this interpretation with classifier ensembles based on multiple decision trees. A RF allows for assessment of the importance of the variables by means of the Gini index and the oob subset (Section 3.3). The importance of the features by means of the oob subset is calculated by switching the m bands used in the prediction of the oob subset. The average of the differences in the accuracy obtained through the modified oob-subset and the original one determines the importance of each variable. The number of votes which each tree contributes as a single unit for the correct class is subtracted from the number of computed votes for the modified subset (Breiman, 2001; Cutler et al., 2007; Chan and Paelinckx, 2008; Gislason et al., 2006). Thus, if variables are too many, as maybe the case for hyperspectral images or multi-source studies, for example, the RF can be applied only to variables which have been identified as most important in the first application of the RF.

Models with 479 trees and one split variable were considered to calculate the importance of the contribution of each variable to the general classification model and to the classification of each category (Breiman, 2001). Fig. 4 shows the contribution of each variable to the classification model generated by considering the KT multi-seasonal bands and topographic variables. According to the Gini Index, the bands with a higher contribution to the RF model are the summer greenness and brightness followed by elevation and spring greenness bands, with values equal to 164 and 156. The contribution of wetness, slope and aspect is smaller than the contribution of the rest of the input variables.

Regarding the oob measure, the contribution of the greenness bands (spring and summer) is the highest one, with values equal to 0.9 in both cases. The elevation band is also very important, with a value of 0.89. In the case of the oob estimation, the aspect variable has a low importance (oob mean decrease in accuracy equal to 0.74).

Table 1 shows the importance of the contribution of each variable to the RF classification of a specific map category. In general terms, the same behavior as for the overall contribution can be observed (Fig. 4), although for some categories the elevation and spring greenness band contribute significantly to increased classification accuracy.

Table 1 shows how the bands derived from the Kauth Thomas transformation have the strongest influence in the class-separability of the classification scheme. It can be seen how when the summer greenness and brightness bands have not been used a decrease in the OOB classification takes place (see oob importance measure).

However, for some specific categories the elevation and spring greenness variables have a significant importance as they enable the classification of those categories, which have a spatial distribution conditioned by relief or temperature. The classes oak grove, ligneous irrigated and tropical crops are located in areas characterized by a given elevation. Hence oak grove and ligneous irrigated crops are distributed over mountainous areas, while tropical crops are distributed along the coast. Moreover, slope has a greater importance in the classification of urban, poplars and herb dry crops categories since they appear on flat areas of low relief.

The spring greenness band has a special interest in the classification of herbaceous dry crops and grasslands, which are only presented during the wet season (spring), having a great vegetation vigor, and, therefore, are not characterized properly in a summer image. In addition, spring greenness helped to differentiate

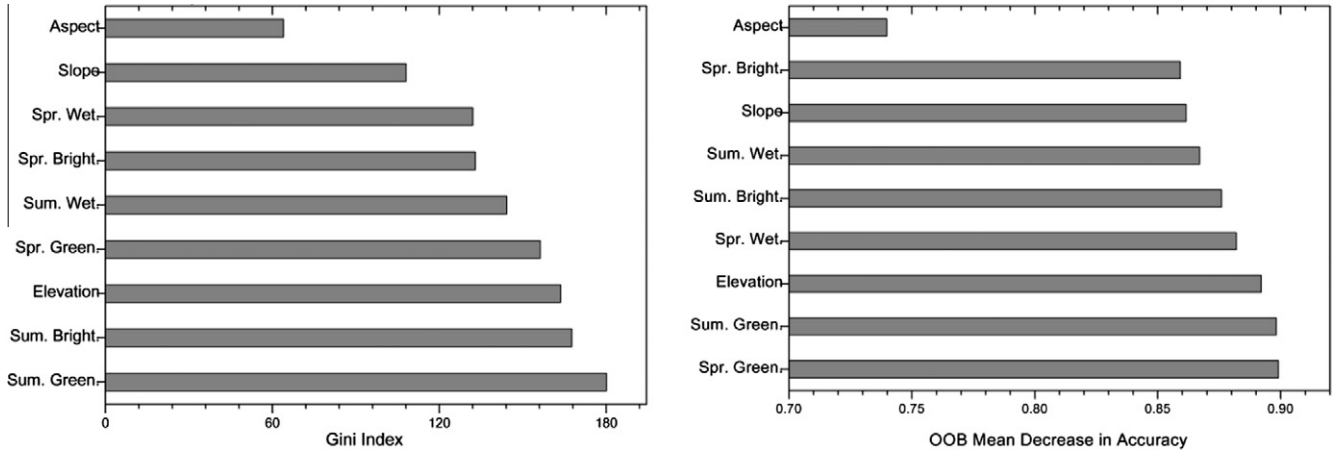

Fig. 4. Variable importance contribution of different bands in terms of Gini Index and oob mean decrease in accuracy.

between urban areas and bare soils, which are covered by some grass in the spring season (Yuan et al., 2005).

4.4. Outliers measure

From the RF point of view an outlier is a case whose proximities to all other cases are small (Section 3.3). Thus, an oulier measure of a case is computed as the number of samples divided by the sum of the squared proximity between that observation and all other observations in the same class, normalized by subtracting the median and dividing by the median absolute deviation within each class. The outlier measure will be large (generally, greater than 10) if the proximity is small, considering it as an outlier.

An outlier measure is computed for each case in the training sample. Fig. 5 shows the outlier measure for all the samples of every class in the training set. As can be observed, almost all the samples are correct, since the proximities between the samples of the same class are high. The samples 347 and 1015 can be considered as exceptions, which belong to the greenhouses and herbal dry classes, respectively.

4.5. Map accuracy assessment

An independent sample of 700 polygons, with 9 pixels for each selected polygon, was selected randomly to assess mapping accuracy. Error matrices were used to assess classification accuracy. Overall accuracy, user's and producer's accuracies, and the Kappa and per category Kappa statistics were then derived from the error matrices (Congalton, 1991; Congalton and Green, 2009) (Table 2). The results of the accuracy assessment of the map obtained through the RF classification model were very positive, taking into account the complexity of the study area and the large number of categories. The overall accuracy and the Kappa statistic were equal to 0.92 and User's and Producer's accuracies of individual classes were also high, 91.33% and 92.08% on average with standard deviations equal to 6.83% and 5.03%, respectively. The average Kappa per category was 0.91 with a standard deviation of 7.68. All the examples corresponding to the categories poplar grove, greenhouses and water were classified correctly. The categories most difficult to classify were those with high intra-class variability, such as shrublands, grasslands and bare soils.

4.6. Comparison of the random forest with a single classification tree

To evaluate the RF technique in a land-cover/land-use classification context, the results achieved by the RF were compared with those produced by a single CT. The aim of this comparison was to determine whether the high accuracy values achieved by the RF are algorithm dependent. The choice for the CT technique allows evaluating the improvement which is implied by the use of classification ensembles as compared to simple classifiers based on decision trees. Furthermore, CT is one of the most widely used tree algorithms in land cover classification (Friedl and Brodley, 1997;

Table 1
Per-class variable importance in terms of oob mean decrease in accuracy. Ranking of importance from top to bottom.

Urban	Poplar Grove	Conifers	Greenhouses	Shrublands	Olive Grove	Grasslands	Oak Grove	Herb. Irrig.	Lig. Irrig.	Herb. Dry	Bare soils	Trop. Crops	Water
Sum. Green.	Sum. Bright.	Sum. Bright.	Elevation	Sum. Green.	Sum. Green.	Spr. Green.	Elevation	Spr. Green.	Elevation	Spr. Green.	Sum. Green.	Elevation	Sum. Wet.
Spr. Green.	Sum. Wet.	Sum. Green.	Sum. Wet.	Sum. Wet.	Spr. Bright.	Sum. Wet.	Sum. Green.	Sum. Wet.	Spr. Wet.	Sum. Bright.	Sum. Bright.	Sum. Green.	Spr. Green.
Spr. Wet.	Slope	Sum. Wet.	Sum. Bright.	Spr. Green.	Spr. Wet.	Elevation	Sum. Bright.	Elevation	Spr. Bright.	Sum. Wet.	Spr. Green.	Spr. Bright.	Sum. Bright.
Spr. Bright.	Sum. Green.	Elevation	Spr. Wet.	Sum. Bright.	Sum. Bright.	Spr. Wet.	Slope	Spr. Wet.	Sum. Wet.	Spr. Wet.	Spr. Bright.	Sum. Bright.	Spr. Bright.
Sum. Bright.	Elevation	Spr. Bright.	Spr. Bright.	Slope	Spr. Green.	Sum. Green.	Sum. Wet.	Sum. Green.	Sum. Bright.	Spr. Bright.	Elevation	Sum. Wet.	Sum. Green.
Sum. Wet.	Spr. Green.	Spr. Wet.	Sum. Green.	Spr. Wet.	Sum. Wet.	Sum. Bright.	Spr. Green.	Sum. Bright.	Slope	Elevation	Sum. Wet.	Spr. Green.	Spr. Wet.
Slope	Spr. Wet.	Spr. Green.	Spr. Green.	Elevation	Elevation	Spr. Bright.	Spr. Wet.	Spr. Bright.	Spr. Green.	Slope	Spr. Wet.	Spr. Wet.	Elevation
Elevation	Spr. Bright.	Slope	Aspect	Spr. Bright.	Slope	Slope	Aspect	Slope	Sum. Green.	Sum. Green.	Slope	Slope	Aspect
Aspect	Aspect	Aspect	Slope	Aspect	Aspect	Aspect	Spr. Bright.	Aspect	Aspect	Aspect	Aspect	Aspect	Slope

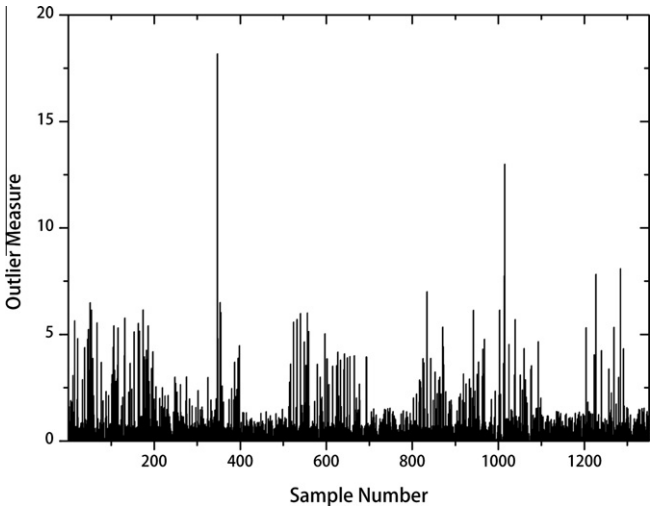

Fig. 5. Outlier analysis for the individual classes.

Hansen et al., 1996; Lippitt et al., 2008; Pal and Mather, 2003; Wessels et al., 2004).

As seen in Section 3.3, RF, as a consequence of its holistic perspective, includes a series of special characteristics which make it different to CT classifiers. On the one hand, RF decision trees are trained from different training data subsets (e.g. two thirds of the total training set). On the other hand, the selection of split variables is random, and those which maximize a function (Gini or gain ratio) are not selected. These procedures, bagging and random selection of split variables, result in that the individual trees of the forest present a low correlation. Hence, learnt patterns diversity and ensemble generalization ability increase (Breiman, 2001).

We used a multivariate CT which uses the Gini Index (Breiman et al., 1984) for data set partition. The original tree was retrospectively pruned to prevent the tree from overfitting (i.e., the tree is allowed to grow up to its maximum depth and it is then when it is pruned). In this way, although computational cost is higher, the tree can explore more partition possibilities (Pal and Mather, 2003).

Table 3 shows the per-categories Kappa index and the Kappa increase pattern for the different RF models (over CT). The RF classifier produces a higher level of classification accuracy than the CT classifier (0.92 over 0.86). The RF increased accuracy greatly in the mapping of certain categories with respect to the CT. Conifers, quercus sp., urban, olive trees, and bare soil present Kappa increases of over 10%, (namely, 30%, 22%, 20%, 18% and 12%, respectively). As already explained in Section 2, the same sample of ground data sites was used in the training and validation of the algorithm. Therefore, these samples are not independent. To compare the differences in accuracies in a rigorous fashion the McNemar test was carried out between the RF and CT classification results. This test considers that a difference in accuracy is statistically significant at the 5% level of significance, thus, for a Z value greater than 1.96 (Foody, 2004, 2009). The Z-value derived from this test, 4.47, shows that the differences in accuracy between the RF and CT are statistically significant and the RF provides a more significant differentiation of the land cover categories.

Figs. 6 and 7 show gains and losses between categories classified by CT and RF. The results expressed in Fig. 7 show the percentage of pixels which change between both classifications taking CT classification as a reference. Classified maps in Fig. 7 represent those transitions derived from the application of the algorithms. From these figures it can be seen how the RF brought a profit in of 35% for all categories. The classification tree overestimated the classification of ligneous irrigated crops and urban areas at 182% and 151%, respectively. On the one hand, RF allows a better differentiation between different types of ligneous vegetation: ligneous irrigated crops, shrublands, tropical crops and oak groves, which were misclassified by CT algorithm. On the other hand, the CT did not make a correct differentiation either between those categories which have greater reflectivity and it classified bare soils as urban areas.

Table 2
Summary of classification accuracies (%) obtained through RF.

Class	Producer's accuracy	User's accuracy	Kappa per category
Urban	0.88	0.90	0.87
Poplar Grove	1	1.00	1.00
Conifers	0.92	0.88	0.91
Greenhouses	1	1.00	1.00
Shrublands	0.8	0.83	0.79
Olive Grove	0.98	0.91	0.98
Grasslands	0.84	1.00	0.83
Oak Grove	0.86	0.84	0.85
Herb. Irrig.	1	0.98	1.00
Lig. Irrig.	0.86	0.98	0.85
Herb. Dry	1	0.89	1.00
Bare soils	0.82	0.84	0.81
Trop. Crops	0.98	0.91	0.98
Water	1	1.00	1.00

Table 3
Per class Kappa values of classification tree and RF classifier.

Class no	CT	RF	Increase in Kappa of RF over CT
1	0.72	0.87	20.40
2	0.98	1.00	2.20
3	0.70	0.91	30.37
4	0.96	1.00	4.49
5	0.85	0.79	−7.14
6	0.83	0.98	18.01
7	0.79	0.83	5.41
8	0.70	0.85	21.88
9	0.91	1.00	9.40
10	0.85	0.85	0.11
11	0.93	1.00	6.98
12	0.72	0.81	12.41
13	0.94	0.98	4.61
14	0.98	1.00	2.20

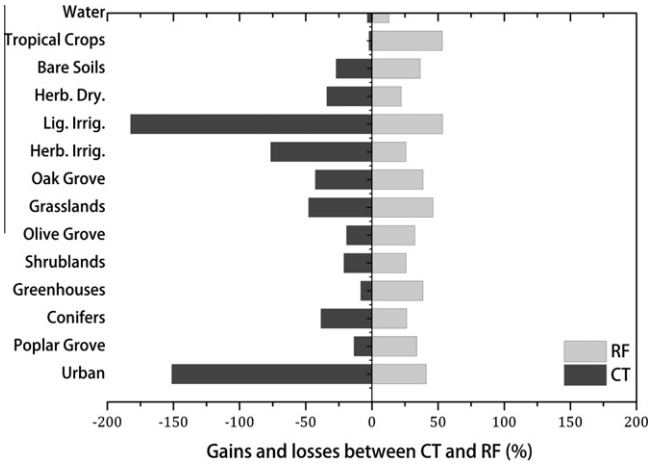

Fig. 6. Comparison of the classifications made by CT and RF in terms of pixel percentage, whose class changed when classified by RF.

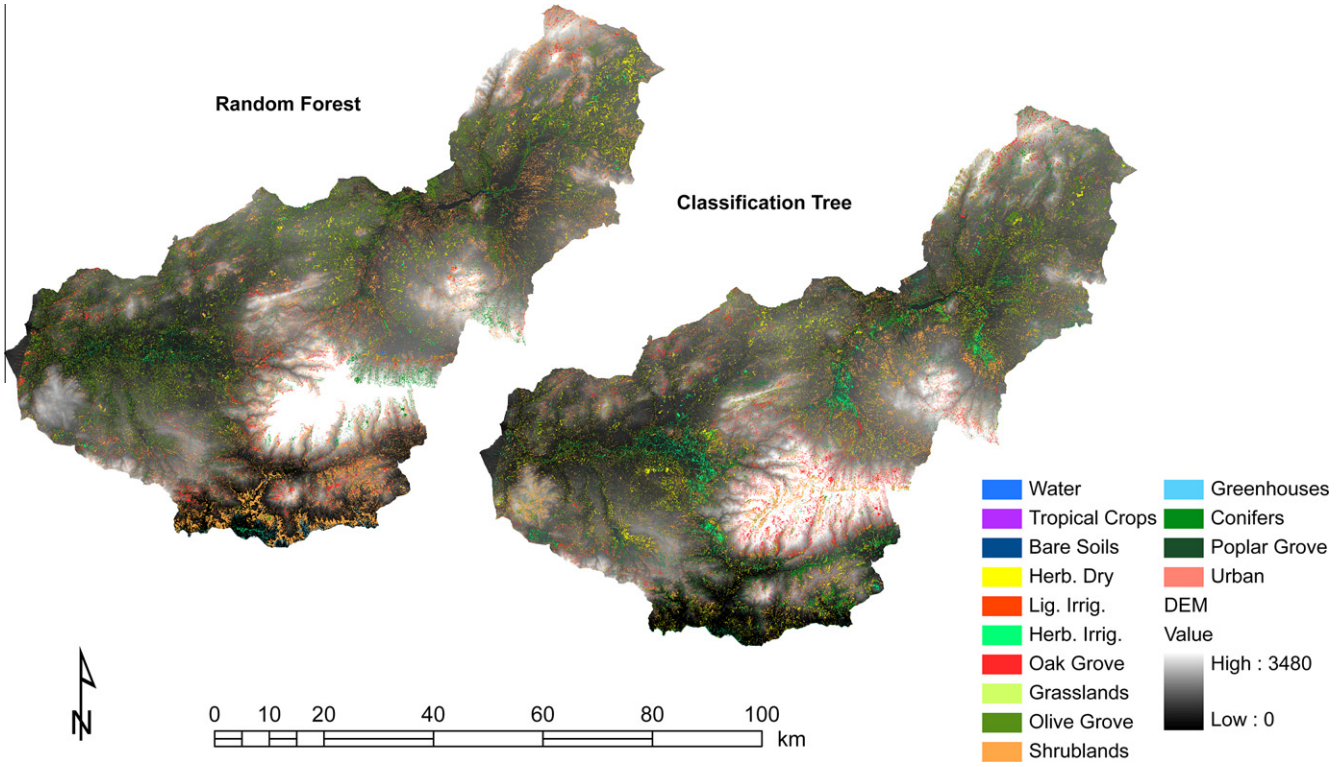

Fig. 7. Land-cover change maps between CT and RF for Granada Province: (a) RF and (b) classification tree.

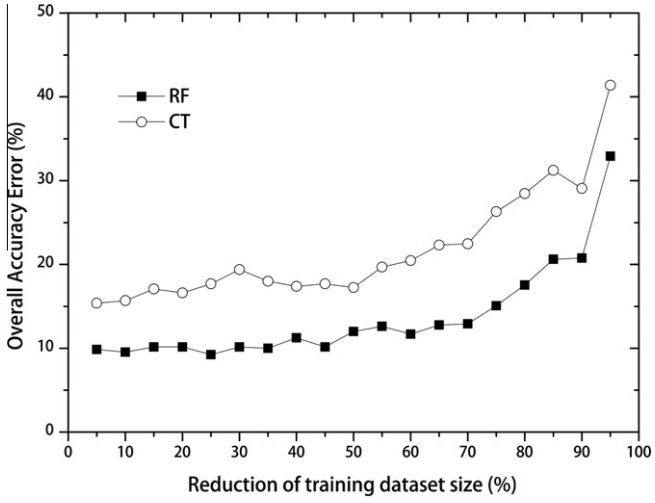

Fig. 8. Effects of training set size reduction in the classification accuracy.

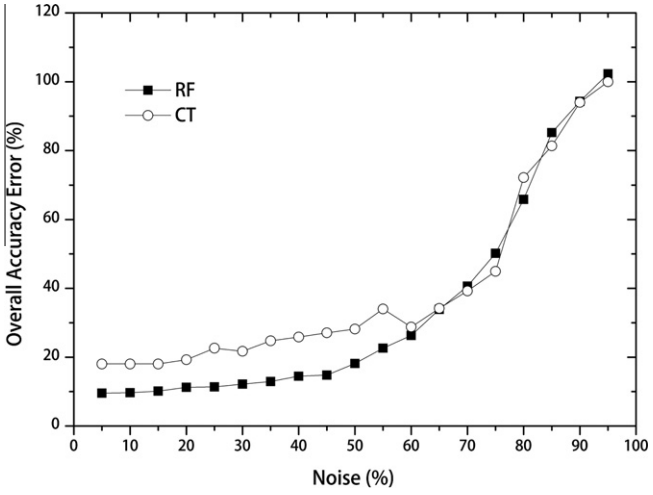

Fig. 9. Effects of noise in the classification accuracy.

4.7. Effect of reduction in the training set

In the classification of complex areas with a large number of categories, some of them with high intra-class variability, the acquisition of ground reference information (training sites) to train the classifier is a time-consuming task, not to mention its high economic cost. On the one hand, it is necessary that the number of training areas be as large as possible to represent all the variability present in a category (Pal and Mather, 2003). On the other hand, it is essential to design a sampling scheme which can be operative both in economic and time terms, and with which an acceptable mapping accuracy level can be achieved (Lippitt et al., 2008; Rogan et al., 2003, 2008).

For homogeneous classes, training data size is not problematic, but it is necessary to use a larger number of training sites for classes with high variability. The effect of the training set size and

Table 4
Pairwise *Z* values to test significant differences between Kappa coefficients of classifications produced from the original and the models built from reduced training set sites.

Reduction (%)	5	10	15	20	25	30	35	40	45	50	55	60	65	70	75	80	85	90	95
RF	0.00	1.39	0.00	0.45	0.77	0.58	0.69	1.46	1.62	2.40	3.09	1.86	2.87	3.16	4.44	4.64	5.66	6.48	10.90
CT	0.22	0.56	1.95	1.17	2.10	2.71	1.86	1.48	1.69	1.35	3.11	3.21	3.94	4.31	5.96	6.98	7.54	7.03	10.71

Table 5
Pairwise Z values to test significant differences between Kappa coefficients of classifications produced from the original and the models trained with different percentages of noise.

Noise (%)	5	10	15	20	25	30	35	40	45	50	55	60	65	70	75	80	85	90	95
RF	0.33	0.28	1.15	1.98	2.69	2.78	3.54	3.68	4.30	5.67	8.64	9.10	11.46	13.38	16.04	18.39	21.83	23.37	24.58
CT	2.36	2.29	2.32	3.20	4.90	4.24	5.61	6.49	6.73	6.86	8.99	7.49	9.35	10.52	11.99	18.45	20.09	22.07	27.00

noise on RF performance was evaluated using the errors in the classification of the training sites as a metric and altering the training sets in increments of 5%, ranging from 5% to 95%.Fig. 8 and Table 4 show how the RF has low sensitivity to the training set size reduction. The difference between the test error of the calibrated model and the total number of data is less than 5% up to reaching a 70% reduction in the number of training data. From the 70% threshold, accuracy decreases more abruptly to attain a Kappa equal to 0.3, considering a 95% reduction. For the CT classifier, this threshold (5%) was reached for reductions equal to and greater than 30%. However, from Table 4 (McNemar test; see Section 4.6) it can be observed how for reduction values lower than before, the accuracy differences were already significant, that is reduction values equal to 50% and 15% for RF and CT. RF had lower overall accuracy error with increased reduction of training data compared to the CT. However, the relative increase in overall error with training dataset size reduction was similar for both RF and classification tree.

4.8. Effect of noise in the training set

The information necessary to train classification algorithms can be obtained through fieldwork or by interpreting aerial photographs. In either case, even though this information is assumed to be correct, errors occur in the labeling of the types of covers, above all in those circumstances in which categories are very heterogeneous and the landscape is complex (Lippitt et al., 2008; Rogan et al., 2008). The effect of mislabeling some examples of training information contributes to increase intra-class variability and, therefore, it has a direct effect on the accuracy of those maps classified from this information. Several studies show the negative effect noise has in the classification of remote sensing data using machine learning algorithms (DeFries and Chan, 2000; Miller and Franklin, 2002; Simard et al., 2000), although none of them state the significance of their results. In this sense it is worth mentioning that the RF can detect those samples associated to noise from the proximity and outlyingness measures (Section 4.4).

The strength level of RF against noise has been assessed by mislabeling training instances following the same pattern as in the previous Section (4.7). Fig. 9 shows how RF is relatively little noise sensitive (error test less than 10%), as long as the noise threshold of 20% is not exceeded. Once the threshold is exceeded the error increases exponentially. The CT classifier always yielded to classification errors greater than 18%. In the study carried out by Rogan et al. (2008) the addition of 30% of noise reduced the overall map accuracy of a CT 47% in the context of change maps. RF for a noise threshold equal to 30% only reduced classification accuracy 12%. Table 5 shows that the differences between RF classifiers were significant only for noise thresholds greater than 20%, that is, for error values higher than 10%. In relation to the CT classifier, the addition of just a 5% of misclassified instances yielded a significant difference in accuracy, which confirms that CTs tend to overfit the data (Breiman, 1984). The greater resistance of RF to noise with respect to a traditional classification tree stems in that RF only uses two thirds of the calibration instances in the classification made by each of the individual CTs which make it up. Hence, the probability that the trees learn the noise (i.e., overfit) is lower than that of the algorithms that use the whole amount of data. However, the relative increase in overall accuracy error with increase in training dataset noise was similar for both RF and classification tree.

5. Conclusions

This study aimed to evaluate the performance of the RF classifier for land-cover classification of a heterogeneous area: Granada Province. By incorporating a suite of multitemporal Landsat data and digital terrain model variables, the RF performed well in the context of classifications with 14 categories. The specific objectives were to study the behavior of the RF with a number of trees and random split variables, reduction in training data and noise addition in terms of oob and test accuracy. The results of this research provide new insights into the performance of MLAs in the context of mapping land-cover over complex and heterogeneous areas.

The RF algorithm generates an internal unbiased estimation of the generalization error (oob error) so it is not necessary to use a test subset independent from the training or resort to cross validation. The RF does not overfit because of the Law of Large Numbers, and it requires two user-defined parameters to be set only: the number of trees and the number of random split variables. The number of trees is directly proportional to the classifier's accuracy until reaching a state (100 trees) in which the generalization error converges adopting values lower than 10%. Once the error has converged, the number of random variables (m) only alters the classifier's accuracy slightly, so the RF can be applied with almost no guidance.

Furthermore, the algorithm can estimate the importance of variables (bands) for the general classification of the land-cover categories and for the classification of each category by means of the Gini Index and oob estimation. Such estimation is of importance for variable selection in the classification of complex areas where it is mandatory to use large multisource data sets with a large number of variables. These important measures show how the summer Kauth Thomas bands have the strongest influence in class-separability in the study area, although elevation and slope bands have a high importance in the classification of some categories which have a spatial distribution conditioned by relief or temperature. The RF algorithm can also measure 'outlyingness' using proximity, which can be very useful to detect mislabeled training areas. The main disadvantage of RF was that it can be difficult to understand the rules used to generate the final classification because of the multiple classification trees generated from resampling the same dataset.

Results show clearly that the RF is superior to standard classification approaches such as a simple decision tree, as it allowed an increased differentiation between the different categories of the study area. The overall Kappa accuracies for the RF and CT were equal to 0.92 and 0.86, respectively. The RF attained an average increase of Kappa per categories of 0.94 and achieved a more reliable classification of the most heterogeneous categories, which are the most difficult to classify, for example, shrublands (Franklin et al., 2000; Rogan et al., 2008), whose accuracy increased by 30% using the RF classifier. On the other hand, the RF, unlike simple decision trees, runs efficiently on high-dimensional data sets (Breiman, 2001; Pal and Mather, 2003).

The RF is relatively robust to the reduction of the training set size and noise and outperforms the CT because the RF had lower overall accuracy error with increased reduction of training data

and noise compared to the CT. However, the relative increase in overall error with increase in training dataset size reduction and noise was similar for both the RF and CT. The reduction of training dataset size and the addition of noise did not have a significant effect on the classifier's accuracy, until reaching a 50% and 20% data reduction and noise addition threshold, respectively. In the case of reduction it could have been due, partly, to the existence of redundancy in some cases of the training set. However, as regards noise, the RF's strength can be attributed to the fact that as only two thirds of the data were used in the training (which is a consequence of the internal bagging process, Section 3.3), the algorithm's overtraining probability is lower.

Acknowledgements

The first author is a FPU Grant holder from the Ministry of Education and Science of Spain. We are grateful for the financial support given by the Spanish MICINN (Project CGL2010-17629) and Junta de Andalucía (Group RNM122). We would like to thank the reviewers for their constructive criticism.

References

Abreu, L.W., Anderson, G.P., 1996. The MODTRAN 2/3 Report and LOWTRAN 7 MODEL.

Atkinson, P., Tatnall, A., 1997. Introduction neural networks in remote sensing. International Journal of Remote Sensing 18 (4), 699–709.

Bala, G., Caldeira, K., Wickett, M., Phillips, T., Lobell, D., Delire, C., Mirin, A., 2007. Combined climate and carbon-cycle effects of large-scale deforestation. Proceedings of the National Academy of Sciences 104 (16), 6550–6555.

Berberoglu, S., Lloyd, C.D., Atkinson, P.M., Curran, P.J., 2000. The integration of spectral and textural information using neural networks for land cover mapping in the Mediterranean. Computers and Geosciences 26 (4), 385–396.

Berk, A., Adler-Golden, S.M., 2002. Exploiting MODTRAN radiation transport for atmospheric correction: the FLAASH algorithm. In: Fifth International Conference on Information Fusion, Annapolis. pp. 798–803.

Betts, R., Falloon, P., Goldewijk, K., Ramankutty, N., 2007. Biogeophysical effects of land use on climate: model simulations of radiative forcing and large-scale temperature change. Agricultural and Forest Meteorology 142 (2–4), 216–233.

Bonan, G., 2008. Forests and climate change: forcings, feedbacks, and the climate benefits of forests. Science 320 (5882), 1444–1449.

Breiman, L., 1984. Classification and Regression Trees. Chapman & Hall/CRC.

Breiman, L., 1996. Bagging predictors. Machine Learning 24 (2), 123–140.

Breiman, L., 2001. Random forests. Machine Learning 45 (1), 5–32.

Breiman, L., Cutler, A., 2004. Random Forest. <http://www.stat.berkeley.edu/breiman/RandomForests/cc_home.htm> (accessed 10.04.10).

Breiman, L., Friedman, J., Stone, C.J., Olshen, R.A., 1984. Classification and Regression Trees, first ed. Chapman and Hall/CRC, Belmont, CA.

Brewster, C.C., Allen, J.C., Kropp, D.D., 1999. IPM from space: using satellite imagery to construct regional crop maps for studying crop insect interaction. American Entomologist 45 (2), 105–117.

Briem, G.J., Benediktsson, J.A., Sveinsson, J.R., 2002. Multiple classifiers applied to multisource remote sensing data. IEEE Transactions on Geoscience and Remote Sensing 40 (10), 2291–2299.

Brisco, B., Brown, R.J., 1995. Multidate SAR/TM synergism for crop classification in western Canada. Photogrammetric Engineering and Remote Sensing 61 (8), 1009–1014.

Brovkin, V., Sitch, S., Von Bloh, W., Claussen, M., Bauer, E., Cramer, W., 2004. Role of land cover changes for atmospheric CO_2 increase and climate change during the last 150 years. Global Change Biology 10 (8), 1253–1266.

Congalton, R., 1991. A review of assessing the accuracy of classifications of remotely sensed data. Remote Sensing of Environment 37 (1), 35–46.

Congalton, R.G., Green, K., 2009. Assessing the Accuracy of Remotely Sensed Data: Principles and Practices, second ed. CRC Press, Boca Raton, Florida.

Cutler, D.R., Edwards, T.C., Beard, K.H., Cutler, A., Hess, K.T., 2007. Random forests for classification in ecology. Ecology 88, 2783–2792.

Chan, J.C.-W., Paelinckx, D., 2008. Evaluation of Random Forest and Adaboost tree-based ensemble classification and spectral band selection for ecotope mapping using airborne hyperspectral imagery. Remote Sensing of Environment 112 (6), 2999–3011.

Chapman, D.S., Bonn, A., Kunin, W.E., Cornell, S.J., 2010. Random Forest characterization of upland vegetation and management burning from aerial imagery. Journal of Biogeography 37 (1), 37–46.

DeFries, R.S., Chan, J.C.-W., 2000. Multiple criteria for evaluating machine learning algorithms for land cover classification from satellite data. Remote Sensing of Environment 74 (3), 503–515.

Dietterich, T.G., 2000. An experimental comparison of three methods for constructing ensembles of decision trees: bagging, boosting, and randomization. Machine Learning 40 (2), 139–157.

Fearnside, P.M., 2000. Global warming and tropical land-use change: greenhouse gas emissions from biomass burning, decomposition and soils in forest conversion, shifting cultivation and secondary vegetation. Climatic Change 46 (1), 115–158.

Feller, W., 1968. An Introduction to Probability Theory and its Application, third ed. Willey, New York, USA.

Foody, G.M., 1995. Land cover classification by an artificial neural network with ancillary information. International Journal of Geographical Information Systems 9 (5), 527–542.

Foody, G.M., 2004. Thematic map comparison: evaluating the statistical significance of differences in classification accuracy. Photogrammetric Engineering and Remote Sensing 70 (5), 627–633.

Foody, G.M., 2009. Sample size determination for image classification accuracy assessment and comparison. International Journal of Remote Sensing 30 (20), 5273–5291.

Foody, G.M., Arora, M.K., 1997. An evaluation of some factors affecting the accuracy of classification by an artificial neural network. International Journal of Remote Sensing 18 (4), 799–810.

Franklin, J., 1998. Predicting the distribution of shrub species in southern California from climate and terrain-derived variables. Journal of Vegetation Science 9 (5), 733–748.

Franklin, S., Wulder, M., 2002. Remote sensing methods in medium spatial resolution satellite data land cover classification of large areas. Progress in Physical Geography 26 (2), 173–205.

Franklin, S.E., Hall, R.J., Moskal, L.M., Maudie, A.J., Lavigne, M.B., 2000. Incorporating texture into classification of forest species composition from airborne multispectral images. International Journal of Remote Sensing 21 (1), 61–79.

Friedl, M., Brodley, C., Strahler, A., 2002. Maximizing land cover classification accuracies produced by decision trees at continental to global scales. IEEE Transactions on Geoscience and Remote Sensing 37 (2), 969–977.

Friedl, M.A., Brodley, C.E., 1997. Decision tree classification of land cover from remotely sensed data. Remote Sensing of Environment 61 (3), 399–409.

Friedl, M.A., Brodley, C.E., Strahler, A.H., 1999. Maximizing land cover classification accuracies produced by decision trees at continental to global scales. IEEE Transactions on Geoscience and Remote Sensing 37 (2), 969–977.

Ghimire, B., Rogan, J., Miller, J., 2010. Contextual land-cover classification: incorporating spatial dependence in land-cover classification models using random forests and the Getis statistic. Remote Sensing Letters 1, 45–54.

Gislason, P.O., Benediktsson, J.A., Sveinsson, J.R., 2004. Random forest classification of multisource remote sensing and geographic data, Igarss 2004. In: IEEE International Geoscience and Remote Sensing Symposium, Anchorage, AK, 20–24 September. pp. 1049–1052.

Gislason, P.O., Benediktsson, J.A., Sveinsson, J.R., 2006. Random Forests for land cover classification. Pattern Recognition Letters 27 (4), 294–300.

Guo, L., Chehata, N., Mallet, C., Boukir, S., 2011. Relevance of airborne lidar and multispectral image data for urban scene classification using Random Forests. ISPRS Journal of Photogrammetry and Remote Sensing 66 (1), 56–66.

Ham, J., Yangchi, C., Crawford, M.M., Ghosh, J., 2005. Investigation of the random forest framework for classification of hyperspectral data. IEEE Transactions on Geoscience and Remote Sensing 43 (3), 492–501.

Hansen, L.K., Salamon, P., 1990. Neural network ensembles. IEEE Transactions on Pattern Analysis and Machine Intelligence 12 (10), 993–1001.

Hansen, M., Dubayah, R., Defries, R., 1996. Classification trees: an alternative to traditional land cover classifiers. International Journal of Remote Sensing 17 (5), 1075–1081.

Hastie, T., Tibshirani, R., Friedman, J., 2009. Random Forests, The Elements of Statistical Learning. Springer, New York. pp. 587–604.

Huang, C., Davis, L.S., Townshend, J.R.G., 2002. An assessment of support vector machines for land cover classification. International Journal of Remote Sensing 23 (4), 725–749.

Jensen, J.R., 2005. Introductory Digital Image Processing, third ed. Prentice Hall, Upper Saddle River, NJ.

Kotsiantis, S., Pintelas, P., 2004. Combining bagging and boosting. International Journal of Computational Intelligence 1 (4), 324–333.

Krogh, A., Vedelsby, J., 1995. Neural network ensembles, cross validation, and active learning. Advances in Neural Information Processing Systems, 231–238.

Lambin, E., Turner, B., Geist, H., Agbola, S., Angelsen, A., Bruce, J., Coomes, O., Dirzo, R., Fischer, G., Folke, C., 2001. The causes of land-use and land-cover change: moving beyond the myths. Global Environmental Change 11 (4), 261–269.

Latifi, H., Nothdurft, A., Koch, B., 2010. Non-parametric prediction and mapping of standing timber volume and biomass in a temperate forest: application of multiple optical/LiDAR-derived predictors. Forestry 83 (4), 395–407.

Lawrence, R., Wood, S., Sheley, R., 2006. Mapping invasive plants using hyperspectral imagery and Breiman Cutler classifications (RandomForest). Remote Sensing of Environment 100 (3), 356–362.

Liaw, A., Wiener, M., 2002. Classification and regression by randomForest. R News 2 (3), 18–22.

Lippitt, C.D., Li, Z., Eastman, J.R., Jones, T.G., 2008. Mapping selective logging in mixed deciduous forest: a comparison of machine learning algorithms. Photogrammetric Engineering and Remote Sensing 74 (10), 1201–1211.

Lunetta, R.S., Balogh, M., 1999. Application of muti-temporal Lansat 5 TM imagery for wetland identification. Photogrammetric Engineering and Remote Sensing 65 (11), 1303–1310.

Martinuzzi, S., Vierling, L., Gould, W., Falkowski, M., Evans, J., Hudak, A., Vierling, K., 2009. Mapping snags and understory shrubs for a LiDAR-based assessment of wildlife habitat suitability. Remote Sensing of Environment 113 (12), 2533–2546.

Mas, J., Velázquez, A., Díaz-Gallegos, J., Mayorga-Saucedo, R., Alcántara, C., Bocco, G., Castro, R., Fernández, T., Pérez-Vega, A., 2004. Assessing land use/cover changes: a nationwide multidate spatial database for Mexico. International Journal of Applied Earth Observation and Geoinformation 5 (4), 249–261.

Mas, J.F., Flores, J.J., 2008. The application of artificial neural networks to the analysis of remotely sensed data. International Journal of Remote Sensing 29 (3), 617–663.

McCoy, R.M., 2005. Field Methods in Remote Sensing, first ed. The Guilford Press, New York.

Miller, J., Franklin, J., 2002. Modeling the distribution of four vegetation alliances using generalized linear models and classification trees with spatial dependence. Ecological Modelling 157 (2–3), 227–247.

Mingers, J., 1989a. An empirical comparison of pruning methods for decision tree induction. Machine Learning 4 (2), 227–243.

Mingers, J., 1989b. An empirical comparison of selection measures for decision-tree induction. Machine Learning 3 (4), 319–342.

Moreira-Madueño, J.M., 2006. El sistema de información geográfica-ambiental de Andalucía. Del SINAMBA a la Red de Información Ambiental de Andalucía. Geofocus 6, 4–10.

Mota, J.F., Pérez-García, F.J., Jiménez, M.L., Amate, J.J., Peñas, J., 2002. Phytogeographical relationships among high mountain areas in the Baetic Ranges (South Spain). Global Ecology and Biogeography 11 (6), 497–504.

Mountrakis, G., Im, J., Ogole, C., 2011. Support vector machines in remote sensing: a review. ISPRS Journal of Photogrammetry and Remote Sensing 66 (3), 247–259.

Mountrakis, G., Watts, R., Luo, L., Wang, J., 2009. Developing collaborative classifiers using an expert-based model. Photogrammetric Engineering and Remote Sensing 75 (7), 831–844.

Oetter, D.R., Cohen, W.B., Berterretche, M., Maiersperger, T.K., Kennedy, R.E., 2001. Land cover mapping in an agricultural setting using multiseasonal Thematic Mapper data. Remote Sensing of Environment 76 (2), 139–155.

Pal, M., 2005. Random forest classifier for remote sensing classification. International Journal of Remote Sensing 26 (1), 217–222.

Pal, M., Mather, P.M., 2003. An assessment of the effectiveness of decision tree methods for land cover classification. Remote Sensing of Environment 86 (4), 554–565.

Pax-Lenney, M., Woodcock, C.E., Collins, J.B., Hamdi, H., 1996. The status of agricultural lands in Egypt: the use of multitemporal NDVI features derived from landsat TM. Remote Sensing of Environment 56 (1), 8–20.

Pax-Lenney, M., Woodcock, C.E., 1997. Monitoring agricultural lands in Egypt with multitemporal Landsat TM imagery: how many images are needed? Remote Sensing of Environment 59 (3), 522–529.

Peters, J., De Baets, B., Verhoest, N.E.C., Samson, R., Degroeve, S., De Becker, P., Huybrechts, W., 2007. Random forests as a tool for ecohydrological distribution modelling. Ecological Modelling 207 (2–4), 304–318.

Prasad, A.M., Iverson, L.R., Liaw, A., 2006. Newer classification and regression tree techniques: bagging and random forests for ecological prediction. Ecosystems 9 (2), 191–199.

Quinlan, J.R., 1993. C4.5 Programs for Machine Learning, first ed. Morgan Kaurmann, San Mateo, CA.

Rogan, J., Franklin, J., Stow, D., Miller, J., Woodcock, C., Roberts, D., 2008. Mapping land-cover modifications over large areas: a comparison of machine learning algorithms. Remote Sensing of Environment 112 (5), 2272–2283.

Rogan, J., Miller, J., 2006. Integrating GIS and remotely sensed data for mapping forest disturbance and change. In: Franklin, M.W.A.S. (Ed.), Understanding Forest Disturbance and Spatial Pattern: Remote Sensing and GIS Approaches. CRC Press, Boca Raton, FL, pp. 133–172.

Rogan, J., Miller, J., Stow, D., Franklin, J., Levien, L., Fischer, C., 2003. Land-cover change monitoring with classification trees using landsat tm and ancillary data. Photogrammetric Engineering and Remote Sensing 69 (7), 784–793.

Schriever, J.R., Congalton, R.G., 1995. Evaluating seasonal variability as an aid to cover-type mapping from Landsat Thematic Mapper data in the Northeast. Photogrammetric Engineering and Remote Sensing 61 (3), 321–327.

Sesnie, S., Gessler, P., Finegan, B., Thessler, S., 2008. Integrating Landsat TM and SRTM-DEM derived variables with decision trees for habitat classification and change detection in complex neotropical environments. Remote Sensing of Environment 112 (5), 2145–2159.

Simard, M., Saatchi, S.S., De Grandi, G., 2000. The use of decision tree and multiscale texture for classification of JERS-1 SAR data over tropical forest. IEEE Transactions on Geoscience and Remote Sensing 38 (5), 2310–2321.

Steele, B.M., 2000. Combining multiple classifiers: an application using spatial and remotely sensed information for land cover type mapping. Remote Sensing of Environment 74 (3), 545–556.

Turner, B., Lambin, E., Reenberg, A., 2007. The emergence of land change science for global environmental change and sustainability. Proceedings of the National Academy of Sciences 104 (52), 20666.

Utgoff, P.E., Brodley, C.E., 1990. An incremental method for finding multivariate splits for decision trees. In: Proceedings of the Seventh International Conference on Machine Learning. Morgan Kaufmann, Austin, TX. pp. 58–65.

Waske, B., Braun, M., 2009. Classifier ensembles for land cover mapping using multitemporal SAR imagery. ISPRS Journal of Photogrammetry and Remote Sensing 64 (5), 450–457.

Wessels, K.J., De Fries, R.S., Dempewolf, J., Anderson, L.O., Hansen, A.J., Powell, S.L., Moran, E.F., 2004. Mapping regional land cover with MODIS data for biological conservation: examples from the Greater Yellowstone Ecosystem, USA and Pará State, Brazil. Remote Sensing of Environment 92 (1), 67–83.

Witten, I.H., Frank, E., 2005. Data Mining: Practical Machine Learning Tools and Techniques, second ed. Morgan Kaufmann, San Francisco, CA.

Wolter, P.T., Mladenoff, D.J., Host, G.E., 1995. Improved forest classification in the Northern Lake States using multi-temporal Landsat imagery. Photogrammetric Engineering and Remote Sensing 61 (9), 1129–1143.

Yuan, F., Bauer, M.E., Heinert, N.J., Holden, G., 2005. Multi-level land cover mapping of the Twin Cities (Minnesota) metropolitan area with multi-seasonal Landsat TM/ETM + data. Geocarto International 20 (2), 5–14.