Peer-reviewed version available at Remote Sens. 2017, 9, 1058; doi:10.3390/rs9101058

Article

Using Different Algorithms and Multi-Seasonal, Textural and Ancillary Information to Increase Classification Accuracy during the Period 2000–2015 in a Mediterranean Semiarid Area

Francisco Gomariz-Castillo [1,2,†,‡], Francisco Alonso-Sarría [2,‡,*] and Fulgencio Cánovas-García [3,4,‡]

1 Instituto Euromediterráneo del Agua, Campus de Espinardo, s/n, 30001 Murcia, Spain; fjgomariz@um.es
2 Instituto Universitario del Agua y Medio Ambiente, Universidad de Murcia. Edificio D, Campus de Espinardo, s/n, 30001 Murcia, Spain; alonsarp@um.es
3 Universidad Politécnica de Cartagena, Unidad Predepartamental de Ingeniería Civil, Paseo Alfonso XIII, 52. 30203 Cartagena, Spain; fulgencio.canovas@upct.es
4 Universidad Técnica Particular de Loja, Departamento de Geología y Minas e Ingeniería Civil, San Cayetano Alto s/n, Loja, Ecuador
* Correspondence: alonsarp@um.es; Tel.: +34-868-88-8695
† Current address: Affiliation 3.
‡ These authors contributed equally to this work.

Abstract: The aim of this study is to evaluate three different strategies to improve classification accuracy in a highly fragmented semiarid area. i) Using different classification algorithms: Maximum Likelihood, Random Forest, Support Vector Machines and Sequential Maximum a Posteriori, with parameter optimisation in the second and third cases; ii) using different feature sets: spectral features, spectral and textural features, and spectral, textural and terrain features; and iii) using different image-sets: winter, spring, summer, autumn, winter+summer, winter+spring+summer; and a four seasons combination. A 3-way ANOVA is used to discern which of these approaches and their interactions significantly increases accuracy. Tukey-Kramer contrast using a heteroscedasticity-consistent estimation of the kappa covariances matrix was used to check for significant differences in accuracy. The experiment was carried out with Landsat TM, ETM, and OLI images corresponding to the period 2000-2015. A combination of four images was the best way to improve accuracy. Maximum Likelihood, Random Forest and Support Vector Machines do not significantly increase accuracy when textural information is added, but do so when terrain features are taken into account. On the other hand, Sequential Maximum a Posteriori increases accuracy when textural features are used, but reduces accuracy substantially when terrain features are included. Random Forest using the three feature subsets and Sequential Maximum a Posteriori with spectral and textural features had the largest kappa values, around 0.9.

Keywords: machine learning; sequential maximum a posteriori; random forest; support vector machines; land use classification; textural information; contextual information

1. Introduction

Several factors hinder the classification of remote sensing imagery in Mediterranean landscapes: the high heterogeneity, due to the presence of small patches dedicated to several different land uses and covers [1]; urban sprawl [2]; the high reflectivity of very dry soils and limestone areas, that mask the presence of vegetation [3]; finally, when rainfed and irrigated areas are mixed, they may be difficult to distinguish by analysing just one image [4].

Several strategies have been developed to overcome these difficulties. Traditional parametric methods, such as Maximum Likelihood (ML) [5], have been substituted by more robust techniques such as Random Forest (RF) [6] or Support Vector Machine (SVM) [7]. Another interesting method, although far less used, is Sequential Maximum a Posteriori (SMAP), a Bayesian multi-scale classification

algorithm [2,8,9]. Several studies have used these methods and reported the comparisons made among them. Li *et al.* [10] used RF, SVM and decision trees to classify forest communities in New York State (USA), obtaining better accuracy with the two first. Sluiter and Pebesma [11] compared seven classification techniques in Mediterranean heterogeneous landscapes, concluding that the more accurate results were obtained both with RF and SVM. Similar results were reported by He *et al.* [12] when mapping Arctic lithology in Canada. Rodríguez-Galiano [13] and Rodriguez-Galiano *et al.* [14] used several different methods to classify land use in a semiarid environment in southern Spain. They reached the same conclusions as the previous authors, stressing that SVM and RF are more robust in the presence of noise in the data. Belgiu and Drăguţ [15], in a recent review of previous research concluded that the RF classifier outperforms decision trees, the Binary Hierarchical Classifier (BHC), Linear Discriminant Analysis (LDA), and Artificial Neural Network classifiers in terms of classification accuracy; RF and SVM classifiers are equally reliable, the accuracy of RF being slightly higher for high dimensional input data such as hyperspectral imagery; however, the SVM classification is more sensitive to the selected features and it is more complicated to use for several parameters have to be set. McCauley and Engel [8], Ehsani [9] and Ehsani and Quiel [16] compared SMAP with ML, finding that the most accurate results were obtained with SMAP. Kumar *et al.* [17] compared SMAP with five machine learning algorithms, including RF and SVM, to classify landsat imagery, the results obtained with SMAP being more accurate than with the machine learning algorithms.

Other research lines have centred on the use of ancillary data (mostly terrain features) to improve accuracy [18]. The conclusion of these studies is that spectral information is not sufficient for an accurate classification, and that terrain information significantly improves accuracy, especially in vegetation types most affected by relief. After an analysis of several studies, Lu and Weng [19], stressed that terrain features improve accuracy in vegetation classification, especially in mountainous regions where vegetation distribution is closely related to topography; however, in urban studies, these features are not commonly used. Sluiter and Pebesma [11] obtained more accurate classifications using ancillary information (elevation, slope, aspect, water stress index and rock types) to complement reflectivity, especially when using RF or SVM. Also, Kumar *et al.* [17] included MDT derived features when comparing classification methods.

Other successful approaches to increase classification accuracy include the use of textural features [3,20] and the use of different images corresponding to different seasons in the same year [14,19]. This last approach has been found especially useful when trying to separate cultivated from natural cover. Classifying multi-seasonal spectral bands using machine learning algorithms such as RF to produce land cover maps helps overcome the difficulty of discriminating between classes which have close spectral characteristics or exhibit a similar phenology [21]. Finally, Gómez *et al.* [22] reviewed and identified methods to incorporate time series information and other novel inputs for annual land cover characterisation.

Since a suitable statistical analysis is needed to test the significance of any improvements in accuracy due to the different strategies tested and also their interactions, the objective of this study is to integrate the above-mentioned four approaches to test their ability to increase accuracy when classifying land cover in a diverse Mediterranean semiarid area. More specifically, we try to increase classification accuracy by using: 1) more sophisticated algorithms and optimising their parameters, 2) several images taken in different seasons of the same year, and 3) different feature sets including textural and contextual (terrain) information. A factorial ANOVA is used to discern which of these approaches, and their interactions, are most relevant for increasing accuracy, and whether the interaction of two such approaches might have a higher impact on accuracy than the sum of individual impacts. Tukey-Kramer contrast using a heteroscedasticity-consistent estimation of the kappa covariances matrix was used to check for significant differences in accuracy when using strategies to improve accuracy.

2. Study area

The research was conducted in a 5079 km^2 semiarid area in South-eastern Spain (Figure 1) in which the Vinalopó and Monnegre river basins (3003 km^2) are included. Precipitation is scarce and irregular; in addition, high temperatures and many hours of sun result in high potential evapotranspiration. Human pressure is quite high, the equivalent population (including both permanent and seasonal) is 1,087,536 [23], including large cities such as Alicante (332,067 hab.) and Elche (228,647 hab.). More than 60 % of the area is dedicated to agricultural activities, mainly irrigated using groundwater and water transfers from the river Tagus. The aridity, pluviometric variability, human pressure and the irrigated agriculture combine to produce a highly diverse area. Such diversity complicates any attempt to classify remote sensing imagery.

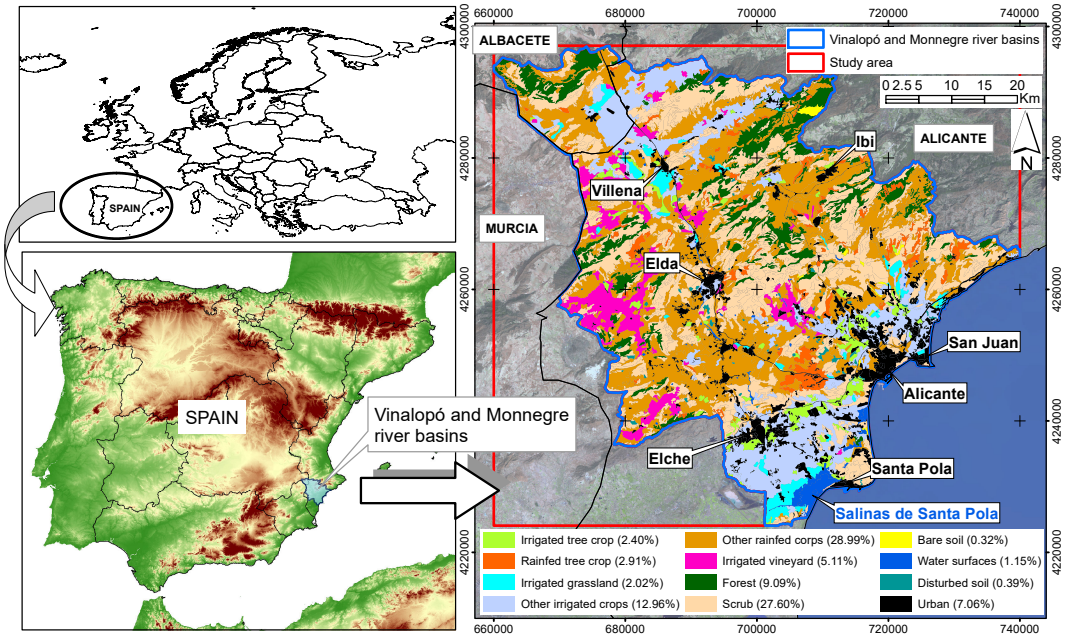

Figure 1. Study area with 2006 Corine Land Cover map is overlayed [24].

3. Material and methods

3.1. Data sources

LandSat-5 Thematic Mapper (TM), LandSat-7 Enhanced Thematic Mapper Plus (ETM+) and LandSat-8 Operational Land Imager (OLI) images along the 2000-2015 time span except 2012 (as we did not find images of sufficient quality to perform the experiment) were classified. When possible, four different images per year (one per season) were taken into account. Table 1 shows the 54 images analysed. After a preliminary quality and cloudiness analysis, 42 images (in bold in Table 1) were finally used. In some years, it was not possible to find good winter images, so the corresponding image from the previous or subsequent year was used.

In order to have a coherent georeferentiation for TM and ETM images, we used 35 control points homogeneously distributed throughout the study area and identifiable in all the images; RSME values were, in all cases, lower than the pixel size. LandSat-8 images are delivered with an accurate enough georeferentiation. Atmospheric correction was carried out using the Chávez [25] method and topographic correction using [26].

Table 1. Landsat images analysed in this study. In bold those finally used for the classification.

Date	Sensor	Date	Sensor	Date	Sensor	Date	Sensor	Date	Sensor
2000		**2001**		**2002**		**2003**		**2004**	
29-01-00	ETM+	**01-12-01**	TM	**06-02-03**	ETM+	**06-02-03**	ETM+	**04-03-04**	TM
21-06-00	ETM+	**21-04-01**	ETM+	**24-04-02**	ETM+	**10-03-03**	TM	13-04-04	ETM+
08-08-00	ETM+	**26-07-01**	ETM+	**19-06-02**	TM	**29-05-03**	TM	19-08-04	ETM+
27-10-00	ETM+	**30-10-01**	ETM+	-	-	26-09-03	TM	**15-11-04**	TM
2005		**2006**		**2007**		**2008**		**2009**	
04-03-04	TM	**24-01-07**	ETM+	**24-01-07**	TM	**14-02-09**	TM	**14-02-09**	TM
18-05-05	ETM+	06-06-06	ETM+	08-05-07	ETM+	**19-06-08**	TM	**05-05-09**	TM
26-06-05	TM	**16-07-06**	TM	**04-08-07**	TM	15-09-08	ETM+	**24-07-09**	TM
12-12-05	ETM+	13-11-06	ETM+	16-11-07	ETM+	01-10-08	ETM	**10-09-09**	TM
2010		**2011**		**2013**		**2014**		**2015**	
16-11-10	TM	**04-02-11**	TM	-	-	**16-03-14**	OLI	**02-02-15**	OLI
24-05-10	TM	**09-04-11**	TM	**14-04-13**	OLI	**04-06-14**	OLI	**07-06-15**	OLI
11-07-10	TM	**28-06-11**	TM	**19-07-13**	OLI	**22-07-14**	OLI	**09-07-15**	OLI
29-09-10	TM	-	-	**14-11-13**	OLI	**26-10-14**	OLI	**30-11-15**	OLI

3.2. Classification methods

To test the effect of different classification methods, we used four classification algorithms, representing different approaches in remote sensing imagery classification: ML, a classical parametric method; RF, an ensemble of decision trees; SVM, a kernel-based algorithm; and SMAP, a multi-scale Bayesian method.

3.2.1. Maximum likelihood

Assuming that features follow a normal multivariate probability distribution, the vectors of means and the variance-covariance matrices of each class can be used to estimate the probability that any given pixel belongs to that class. The pixel is then classified into the most probable class. This probability can also be used as an indicator of the classification certainty, the classification of those pixels with a maximum probability below a given threshold are rejected.

As some classes have a larger presence in the study area, the proportion of each class in the training areas can be used as prior probability, in a bayesian approach that uses the equation:

$$P\left(H|E\right) = \frac{p\left(E|H\right) \cdot p\left(H\right)}{p\left(E\right)} \tag{1}$$

where $P\left(H|E\right)$ is the conditional probability of class H given evidence E (spectral response), $p\left(E\right)$ is the total probability of E, $p\left(H\right)$ is the prior probability of H and $p\left(E|H\right)$ is the conditional probability of E given H.

ML has been widely used in remote sensing; however the basic normality assumption is not always met, especially when including textural or ancillary information, so this assumption should be verified. [27] suggest that ML may be robust enough and not be affected by non-normality; however, it is very sensitive to outliers that may easily appear, especially if ancillary data are used.

3.2.2. Random forest

RF [6] is a non-parametric method based on an ensemble of decision trees. Each tree is trained with a bootstrapped sub-sample of cases, and decisions in each node are made using just a random feature subset. Each tree contributes with one vote to classify each pixel, which, eventually, is attributed to the most voted class. The feature randomisation reduces the correlation among trees, enforcing the ensemble concept. The number of trees (*ntree*) and the number of features used to train each tree (*mtry*)

are parameters whose default values are, respectively, 500 trees and the square root of the number of available features rounded to the closest integer [28,29].

RF produces more accurate results than other classification methods [6,28], even when there are more features than observations or when most of the features are noisy. It does not overfit the model to the data [30] and gives a high generalisation capability [6,31,32]. Since the cases not included in a bootstrapped sample are not used to fit the corresponding tree, they can be used to perform a cross-validation accuracy estimation [33]. However, it has been suggested [34] that this procedure may underestimate errors in Remote Sensing applications. A final advantage is that it is computationally lighter than other meta-classifiers such as boosting [14].

A disadvantage of random forest is that it becomes a black box approach. However, it provides a rank of feature importance that determines which features have had higher weight during the decision process [6,28,29]. It can be used to compare the relative importance among features, so the result is easier to interpret than with other algorithms such as neural networks or SVM.

3.2.3. Support Vector Machines

SVM [35,36] is a very flexible classification algorithm that draws border hyperplanes among classes in the feature space. The distance between such hyperplanes and the cases closest to them, the so-called support vectors, is maximised. These large distances, margins in SVM terminology, give SVM a greater generalisation capacity, because they maximise the probability of correctly classifying new cases located between two different classes. When the classes are not completely separable, a cost parameter C indicates the number of cases allowed on the wrong side of the separating hyperplane. The lower the cost, the more complex the hyperplane needs to be to avoid miss-classifications and the lower the generalisation capacity. SVM uses a default value of $C = 1$.

SVM is included in the category of kernel methods because kernel functions can be used to transform the space of features in order to obtain a linear separation hyperplane even if the borders among classes in the original feature space are not linear. With Radial Basis Function (RBF), the kernel used in this study, γ parameter, controls the width of the function. Generally, low γ values may produce overfitting and very high γ values may produce underfitting.

SVM is increasingly used because of its advantages over traditional methods. Mountrakis *et al.* [37] analysed several studies on SVM used in remote sensing, highlighting its good results due to its generalisation ability and high reliability even with limited training data (both in quality and quantity). Nowadays, It is a fully established method in machine learning [7,38] due to its high generalisation ability compared with other algorithms such as neural networks. It was first applied to classify hyperspectral imagery by Gualtieri and Cromp [39] and Melgani and Bruzzone [40], and recent reviews on the subject can be found in Tso and Mather [41] or Camps-Valls and Bruzzone [36].

Its main disadvantages are that it is a black box and that noisy or co-linear features can affect the results [42].

3.2.4. Sequencial Maximum a Posteriori

SMAP [43,44] is based on contextual classification, a classification of the pixels by region and not individually; in this sense, it can also be considered a segmentation method. It is assumed that the cells that are close in the image are more likely to belong to the same class, so it works by dividing the image in various resolutions. It then uses coarser divisions to obtain a prior density function from which, using a Bayesian approach, an a posteriori distribution in the finer division is obtained [44,45]. The end result is a land use map with larger and more homogeneous polygons avoiding the speckle effect usual in land use maps obtained by image classification.

The only parameter to be defined is the window size, whose purpose is to divide the image to reduce the memory load; however, it can slightly influence the results as the smoothing parameters of the segmentation algorithm are estimated separately for each window. For this reason, using a small

window is recommended. A similar approach is the use of *Random Markov Fields* (MFR) [41], although such algorithms are computationally more intensive than SMAP [44].

3.3. Multi-seasonal approach

To test whether the use of multi-seasonal images improves accuracy classification, seven season combinations were used: spring, summer, autummn, winter, winter+summer, winter+spring+summer, and the combination of the four seasons. After analysing the 2009 images, we concluded that summer images produce the best results when used alone. As a consequence, only the summer image was used to represent one-season classifications for the rest of the years.

3.4. Feature sets

In order to test the effect of textural and ancillary information on classification accuracy, images were classified using three different feature sets: 1) Reflectivity: the six LandSat reflectivity bands; 2) reflectivity and texture estimated by two semivariogram layers; 3) reflectivity, texture and terrain features (height, slope, and aspect sine and cosine) calculated from the Spanish *Instituto Geográfico Nacional* (National Geographic Institute) DEM, a 25 m resolution raster layer derived from LIDAR data taken in 2009.

The two semivariogram layers were calculated from the first principal component of the reflectivity bands (considered as a weighted average of reflectivities) and from the NDVI, respectively. The equation used to calculate the semivariogram is:

$$\gamma = \frac{\sum_{i=1}^{4} (b - b_i)^2}{8} \tag{2}$$

where b is the value in the analysed cell and b_i the value in the four cells surrounding; so, this can be considered a one-pixel lag semivariogram.

The use of parametric classification methods such as ML requires a number of assumptions about the features: mainly that they follow a normal multivariate distribution and the absence of outliers. We analysed these assumptions for the original and transformed variables (logarithmic and inverse) using an exploratory analysis based on box-plots and the Kolmogorov-Smirnov test. A compromise transformation for all classes was selected for each feature that did not comply with those assumptions. Finally, reflectivity and terrain features were not transformed and textural features were logarithmically transformed.

3.5. Training and validation areas

One of the greatest difficulties when historical images are classified is to obtain training and validation areas for the whole period without field work. To identify such areas we used 3 land use maps and 5 orthoimages, in order to cover the time span from 2000 to 2015:

- *Mapa de Cultivos y Aprovechamientos* (crops and land-use map) published by the Spanish *Ministerio de Agricultura, Pesca y Alimentación* (Ministry of Agriculture, Fisheries and Food) with field data collected between 2001 and 2007 at 1:50,000 scale.
- Corine Land Cover maps [24] for 2000 and 2006 at 1:200,000 scale.
- 2002 orthophotography from the *Sistema de Información Geográfica de Parcelas Agrícolas* (Agricultural Plots Geographic Information System) project at 1:5000 scale by the Spanish *Ministerio de Agricultura, Pesca y Alimentación* (Ministry of Agriculture, Fisheries and Food).
- Orthophotography series available in the *Instituto Cartográfico de Valencia* (Cartographic Institute of Valencia) and the *Plan Nacional de Ortofotografía Aérea* (Spanish Orthophotography National Plan, PNOA) for 2005, 2007, and 2012 at 1:10,000 scale by the Spanish *Instituto Geográfico Nacional* (National Geographic Institute).
- Orthophotography from the PNOA for 2009 and 2014 at 1:5000 scale.

The criteria to select training and validation areas were: 1) training and validation areas should not be too close in order to guarantee statistical independence; 2) land use should be the same throughout the study period to ensure that no changes occurred; 3) to avoid border effects, areas should be defined inside the real land use polygon, discarding a 50-75 m buffer from their border; 4) infrastructures crossing the areas and other features that could introduce noise should be avoided; 5) the topography inside the areas should be as homogeneous as possible; 6) areas should be as homogeneously distributed as possible; 7) the minimum size of the training set should be between 10 and 30 times the number of features per class (at least in classifications with one or two seasons), as recommended by Mather and Koch [46]; 8) for the validation areas, the area of each land use should be proportional to its area in the Corine land cover map.

Training and validation areas obtained from maps rather than of field work are less reliable, especially when, in a multi-temporal experiment, maps are not available for every year. In order to detect pixels that in certain years might have a different use, an ML-based cross-validation analysis of each pixel and year was made to identify those pixels which, in certain years, have a very low probability of being classified in the class to which they belong according to the maps. These pixels, which were considered noisy and so eliminated, were mostly outliers, so this also functioned as an implicit outlier elimination process.

Finally, 214 areas were obtained. This set was divided into 141 (2/3) training areas and 73 (1/3) validation areas, by means of a stratified random sampling based on the classes studied and elevations in the study area (table 2).

To evaluate the influence of stratified random sampling in the classification, several classifications were made with randomly separated training and validation areas. The resulting changes in validation were not statistically significant.

Table 2. Number and extension (ha) of training and validation areas. Water surfaces include some sea polygons. The percentages refer to the areas. Bare soil was initially included within the Scrub class, but because of its different reflectivity, we have considered it as a new class.

	Training areas			Validation areas		
Use	**N**	**Area**	**%**	**N**	**Area**	**%**
Forest	19	328.21	13.66	10	98.35	9.79
Scrub	22	302.43	12.59	12	213.50	21.26
Rainfed tree crops	13	77.89	3.24	7	51.39	5.12
Irrigated tree crops	14	148.01	6.16	8	39.78	3.96
Rainfed grassland	15	231.85	9.65	8	111.01	11.05
Irrigated grassland	10	293.20	12.20	5	103.74	10.33
Impervious surfaces	16	423.84	17.64	7	112.86	11.24
Water surfaces	11	391.12	16.28	6	207.72	20.69
Bare soil	4	7.56	0.31	2	8.02	0.80
Vineyard	17	198.62	8.27	8	57.81	5.76
Total	**141**	**2402.73**	**100**	**73**	**1004.18**	**100**

3.6. Classification process

The objectives were tackled in three stages. First, year 2009 was classified with all possible combinations of algorithms (4), feature subsets (3) and seasonal imagery (7). making a total of 84 classifications. In the second stage, as the previous results showed that multi-seasonal classifications increased the accuracy, the whole 2000-2015 series (except 2012) was classified using the highest number of seasonal images available (two, three or four) and the 3 feature sets to optimise SVM and RF models. That made 90 optimised and 90 non-optimised classifications. The objective of this stage was to optimise the parameters of the RF and SVM method to test whether such optimisation significantly improves accuracy. Although RF is not sensitive to its parameters [28,35], it can be optimised using

k-fold cross-validation with repetition [47]. Kuhn and Johnson [48] suggested using $k = 10$ when trying to calibrate RF. Following this advice, we made a 10-fold cross-validation with 5 repetitions. The range of values tested was initially $m_{try} \in \{2, \rho\}$ (where ρ is the number of features), but, after some preliminary tests, we changed it to $m_{try} \in \{2, 12\}$.

Both SVM parameters (C and γ) were calibrated using a method similar to that described for RF: a 10-fold cross-validation with five repetitions. In this case, γ was estimated first using Caputo *et al.* [49] methodology, implemented in the R package *kernlab* [50]. Once γ was set, C was optimised for values $\{0.25, 0.5, 1, 2, 4, 8, 16, 32\}$. In a third stage, we classified the whole 2000-2015 period (except 2009, which was classified in the first stage, and 2012 for which there was no appropriate images). Such classifications were carried out using the 3 feature subsets, the 4 classification algorithms and 4 seasonal combinations: summer (the season with more accurate results when used alone in 2009), spring+summer, spring+summer+winter, and the four-images combination. That makes 552 classifications including 2009. As the results of the second stage showed that optimisation does not significantly increase accuracy, RF and SVM classifications were obtained with the default parameters.

That large number of classifications, to which calibration cycles must be added, could not have been carried out on most commercial desktop software, except the most expensive ones. However, this work was successfully carried out using Open Source. GRASS [51,52] was used to store raster data. Processing was carried out both with GRASS and R [53,54]. Bash and R scripting languages were used to write the computing protocols. Some advantages of such an implementation are its low cost, high interoperability between programs, easy automation of processes in high-performance servers, the modularity that allows new processes to be changed or included, and the reproducibility of the work.

3.7. Validation of classifications and evaluation of hypothesis

To asses the quality of the results, both qualitative (visual) and quantitative (confusion matrices) analyses were carried out. Several goodness of fit statistics were estimated from the confusion matrices. Kappa index [55], introduced in remote sensing by Congalton and Mead [56], is an accuracy measurement that evaluates the percentage of improvement over a random classification [57]. The significance of differences between the indices was checked by calculating 95% confidence intervals. Conditional kappa values were estimated to measure the within-class accuracies [58–60]. Conditional kappa involves two values: User's kappa, related with errors of omission, and producer's kappa, related with errors of commission. Kappa values were interpreted following Landis and Koch [61] criteria, that is: 0.00-0.20, Insignificant; 0.21-0.40, Low; 0.41-0.60, Moderate; 0.61-0.80, Good, and 0.81-1.00, Very good.

To evaluate the effects of the different strategies to improve classification accuracy, two factorial ANOVA considering kappa values as dependent variable were carried out:

1. To evaluate how the results improve when RF and SVM parameters are optimised, a factorial ANOVA was conducted to compare the effects of the classification method (*Method*), optimisation (*Optimised*), feature sets (*VarSet*) and the interactions between them. *Method* included two levels (RF; SVM), *Optimised* included two levels (Yes; No) and *VarSet* three levels (Sp: Spectral information; SpTex: Spectral and textural information; SpTexRel: Spectral, textural and contextual information). In this case, classifications were performed using the maximum number of images available per year: four in 2000, 2001, 2009, 2010, 2014, 2015; three in 2002, 2003, 2011, 2013; and two in 2004, 2005, 2006, 2007 and 2008. That makes 180 classifications.
2. To evaluate how classification accuracy improves in the final models, a factorial ANOVA was conducted to compare the main effects of *Method*, *VarSet*, number of seasonal images (*Season*) and the interaction effect between them. In this case, *Method* included four levels (RF; SVM; ML; SMAP) and *Season* four levels (One season; Two seasons; Three seasons; Four seasons). In this case only the years when 4 images were available (2000, 2001, 2009, 2010, 2014 and 2015) were taken into account, making 288 classifications.

The proposed designs include seven null hypotheses to contrast, three related with the simple effects, three related with the interactions among pairs of factors, and a final one for the three factors in conjunction. The orthogonal design for three fixed levels might be represented as:

$$Y_{ijkl} = \mu + \alpha_j + \beta_k + \gamma_l + (\alpha\beta)_{jk} + (\alpha\gamma)_{jl} + (\beta\gamma)_{kl} + (\alpha\beta\gamma)_{jkl} + \epsilon_{ijkl} \quad (3)$$

where μ is the global mean; α_j ,β_k, γ_l are the effects of each level j, k y l; $(\alpha\beta)_{jk}$, $(\alpha\gamma)_{jl}$, $(\beta\gamma)_{kl}$ are the two-way interaction model components; $(\alpha\beta\gamma)_{jkl}$ are the three-way interaction model components; and ϵ_{ijkl} is the error component.

Once the existence of an effect was discovered using ANOVA, a Tukey-Kramer contrast was carried out to identify significant differences among factors; this contrast is based on a Student's t:

$$t = \frac{|X_i - X_j|}{SE_{ij}} \quad (4)$$

where SE_{ij} is a pooled estimation of the standard error of the means obtained from the covariance matrix of the kappa values.

To evaluate the statistical assumptions, we used the Kolmogorov-Smirnov (KS) test to evaluate normality and the Levene test to evaluate homocedasticity in the residuals. In both cases, KS was not significant ($W = 0.08, p = 0.2005$ and $W = 0.077, p = 0.066$ respectively) but the significance of the Levene contrasts ($F(11, 168) = 1.97, p = 0.0346$ and $F(47, 240) = 1.95, p = 0.0007$ respectively) showed heteroscedasticity. In this case, the covariance matrix of the estimated parameters ($Var[\hat{\beta}]$) is not robust enough and the Tukey-Kramer contrast is less reliable.

Several statistical methods have been proposed to correct for heteroscedasticity [62]. In this case we used a heteroscedasticity consistent covariance matrix of the parameters (HC3). With this methodology, heteroscedasticity effects are avoided even when its form is unknown [62]. The basic idea behind an HC estimator is to use residuals ($\hat{e}_i^2$) to estimate the covariance matrix:

$$Var[\hat{\beta}] = HC3 = \left(X'X\right)^{-1} X'\widehat{\Omega}_3 X \left(X'X\right)^{-1} \quad (5)$$

$$\widehat{\Omega}_3 = diag\left\{ \frac{\hat{e}_1^2}{(1-h_{11})^2}, ..., \frac{\hat{e}_i^2}{(1-h_{ii})^2} \right\} \quad (6)$$

where h_{ii} is the ii element in the matrix $\left(X'X\right)^{-1} X'$.

Using this methodology, it is possible to compute a more robust covariance matrix estimator for the Tukey-Kramer contrast, especially when the number of cases is small.

4. Results and discussion

4.1. Classification of 2009 image

Figure 2 shows mean kappa and 95% confidence intervals for the four algorithms, three feature combinations and seven season combinations in 2009. High kappa values (around 0.8) are reached; given the complexity of the study area and the objectives, the results are quite satisfactory. Figure 3 shows that using four seasons significantly outperforms any other temporal combination in every combination of algorithm and feature subset. In general, the second best option is to use three seasonal images (winter-spring-summer), which in most cases does not significantly differ from using four. When using two seasons (winter and summer) the accuracy does not significantly outperform the accuracy reached using the summer image alone, probably because it is the season in which the highest spectral differences between land uses appear. Images of winter and spring were seen to be the least accurate options.

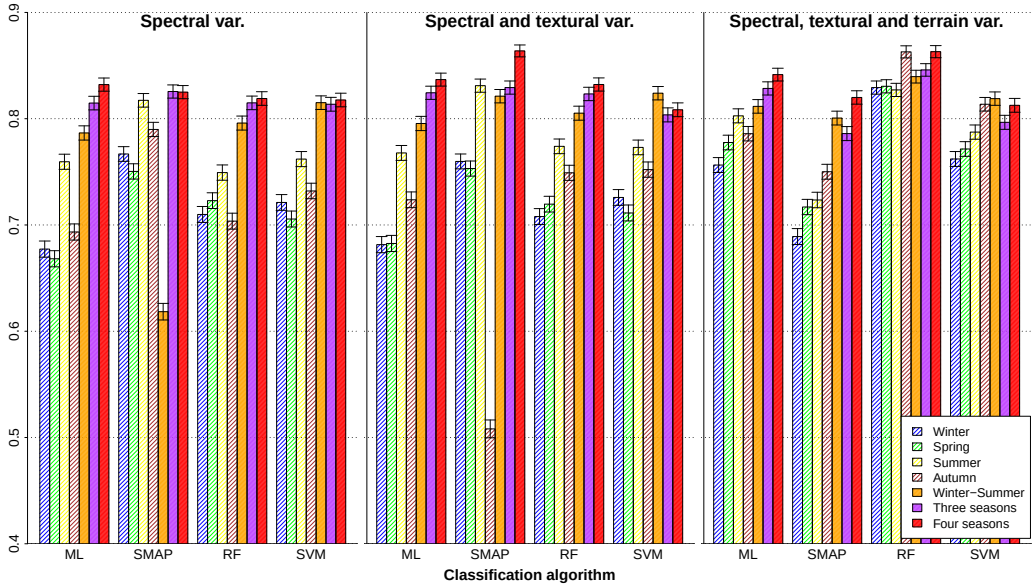

Figure 2. Kappa values and 95% confidence intervals obtained in the 2009 classification. Number of classifications: 84.

There is a general accuracy increase when the number of features increases, especially in the case of SMAP, when spectral and textural features are introduced, and in the case of RF when using all features.

4.2. Parameter optimisation

Having concluded that the best results are obtained with the combination of four seasons, this combination (or, when not available, the combination of three or two seasons) was used, in a second stage to classify the 2000-2015 series using RF and SVM. The objective was to test whether parameter optimisation significantly improves validation accuracy. Figure 3 shows the whole 2000-2015 series of kappa values with 95% confidence intervals. An analysis of variance based on heteroscedasticity-consistent standard errors indicated that the effect *optimised* is not significant ($F\,(1, 168) = 0.0061, p = 0.9378$), so there are no significant differences between optimised and non-optimised models (M = 0.8126, SD = 0.0346 and M=0.8124, SD =0.0343, respectively). By contrast, the effects of *method* ($F\,(1, 168) = 43.4019, p < 0.0001$), *variable set* ($F\,(2, 168) = 60.32, p < 0.0001$), and the interaction between them ($F\,(2, 168) = 24.0107, p < 0.0001$) were significant. Figure 4 summarises the main and simple effects of such significant factors and the homogeneous groups obtained in the post-hoc comparisons. RF (M= 0.8177, SD = 0.04) seems significantly more accurate than SVM (M = 0.8072, SD = 0.0267); SpTexRel (M=0.8374, SD =0.0287, group b) is significantly more accurate than the other two combinations (group a), spTex is not significantly more accurate than classifications using only spectral features, indicating that RF is significantly more accurate when using SpTexRel (M=0.86, SD =0.0133, group c); in general, both methods seem to follow the same pattern, accuracy increasing slightly higher accuracy when features are added to the model; however the differences are not significant.

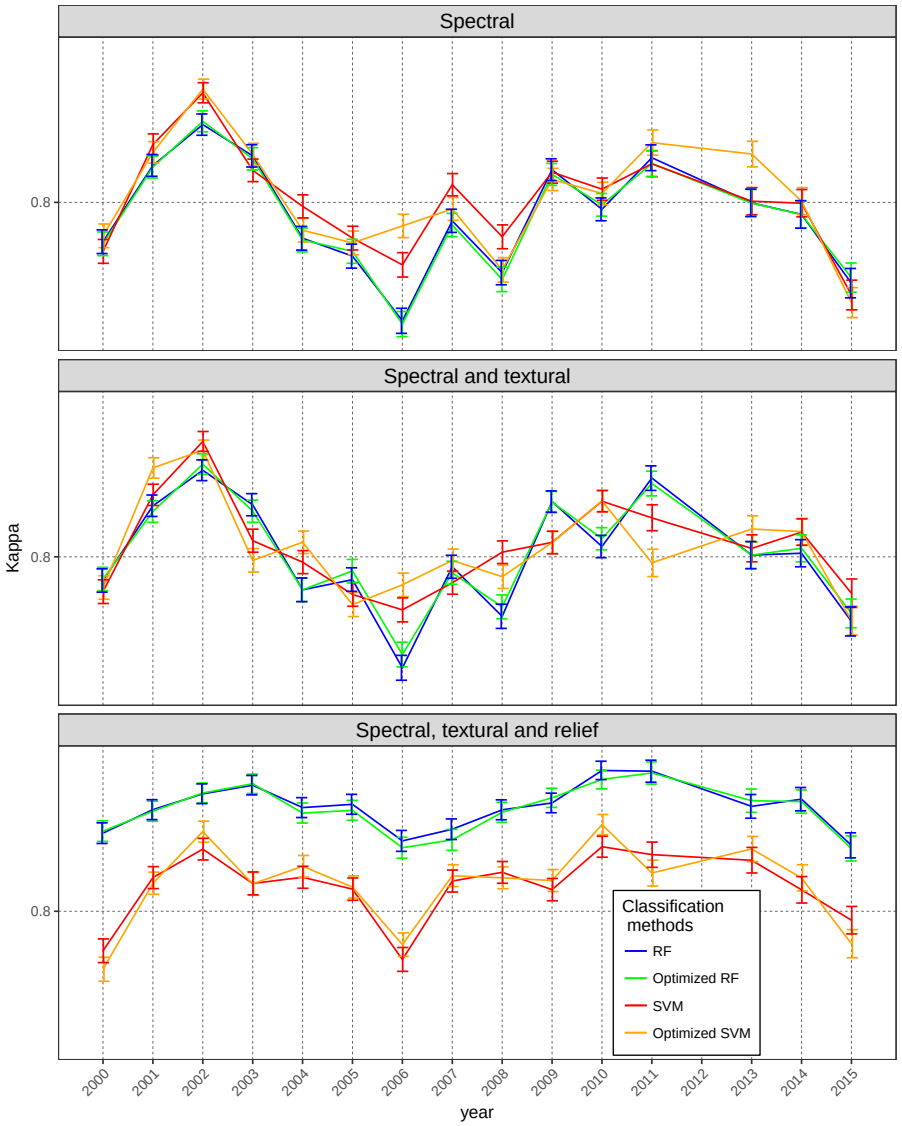

Figure 3. Kappa series with 95% confidence intervals for the RF and SVM algorithms with and without optimisation. Number of classifications: 180.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 September 2017 doi:10.20944/preprints201709.0009.v1

Peer-reviewed version available at Remote Sens. 2017, 9, 1058; doi:10.3390/rs9101058

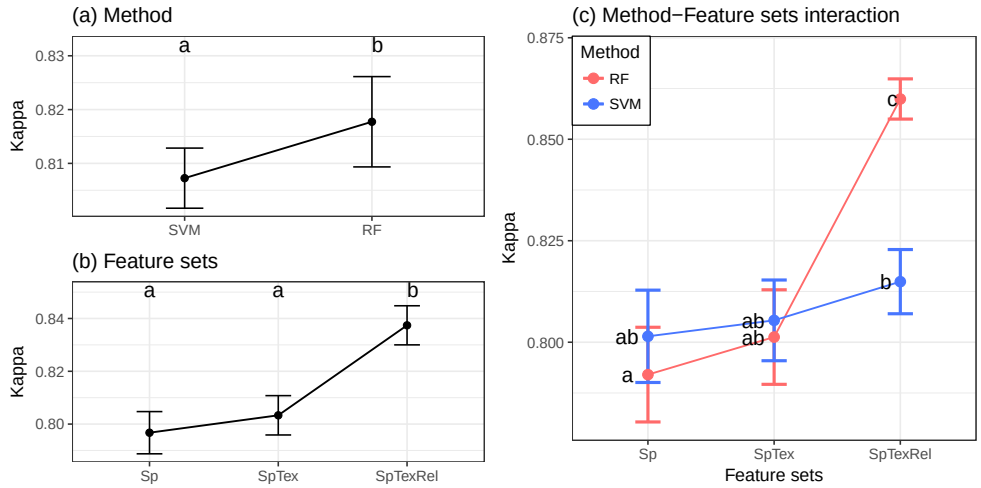

Figure 4. Main effects and simple effects of significant factors (mean and 95% confidence interval). Significantly different groups (Tukey-Kramer contrast using HC3, alpha=0.05) are represented by different letters. Number of classifications: 180.

The main conclusion that can be drawn from these results is that optimisation does not significantly improve accuracy, as only half of the time the optimised model outperforms the non-optimised. A slightly higher accuracy is obtained with SVM when using spectral and spectral plus textural features; however, the accuracy of RF is significantly higher when adding terrain features. These results and the high computing cost of optimisation, especially for RF (0.33 hours without optimisation and 51.2 hours when optimising), led us to omit optimisation for the rest of the season combinations.

4.3. Global validation

Figure 5 summarises the kappa valuesfor the period 2000-2015 and Figure 6 shows the complete series. Figure 5 does not include the period 2004-2008, because, for those years, only two images per season were used in the classification. According to these results, ML, RF and SVM slightly increase their accuracy when textural features are added, but the increase is higher when terrain features are taken into account, especially with RF. In contrast, SMAP improves accuracy when textural features are used but reduces it considerably when terrain features are added, producing several moderate (lower than 0.65) kappa values. Using just two images (2004-2008) has a clear effect on classification accuracy (Figure 6); whereas for the four-image classification (2001, 2002, 2009, 2010, 2014 and 2015) kappa values are around 0.8, which is the threshold value for a very good classification, the two-image classifications (2004-2008) drops to 0.7, except when SMAP is used to classify or when the three feature sets are taken into account.

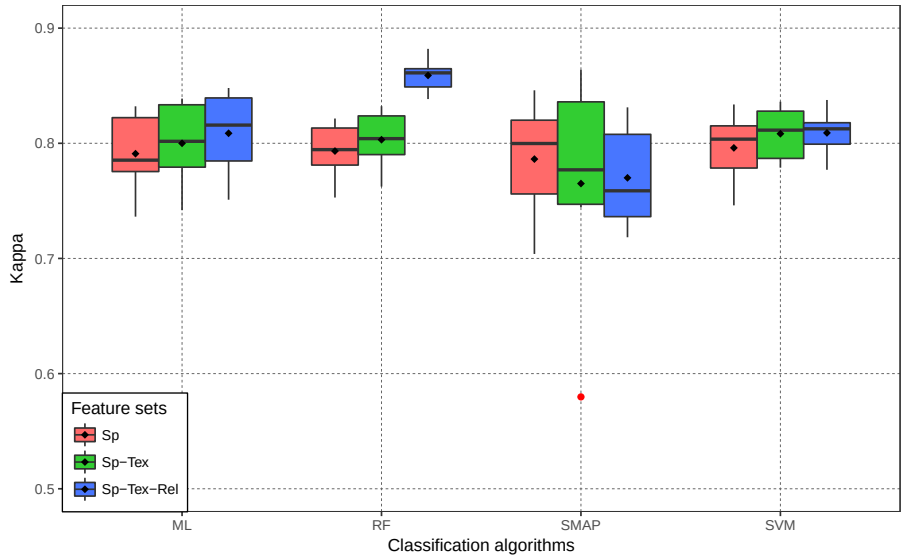

Figure 5. Kappa values obtained with the four algorithms and the three feature combinations using four season imagery. Number of classifications: 72. Sp: Spectral features; Sp-Tex: Spectral and textural features; Sp-Tex-Ter: Spectral, textural and terrain features.

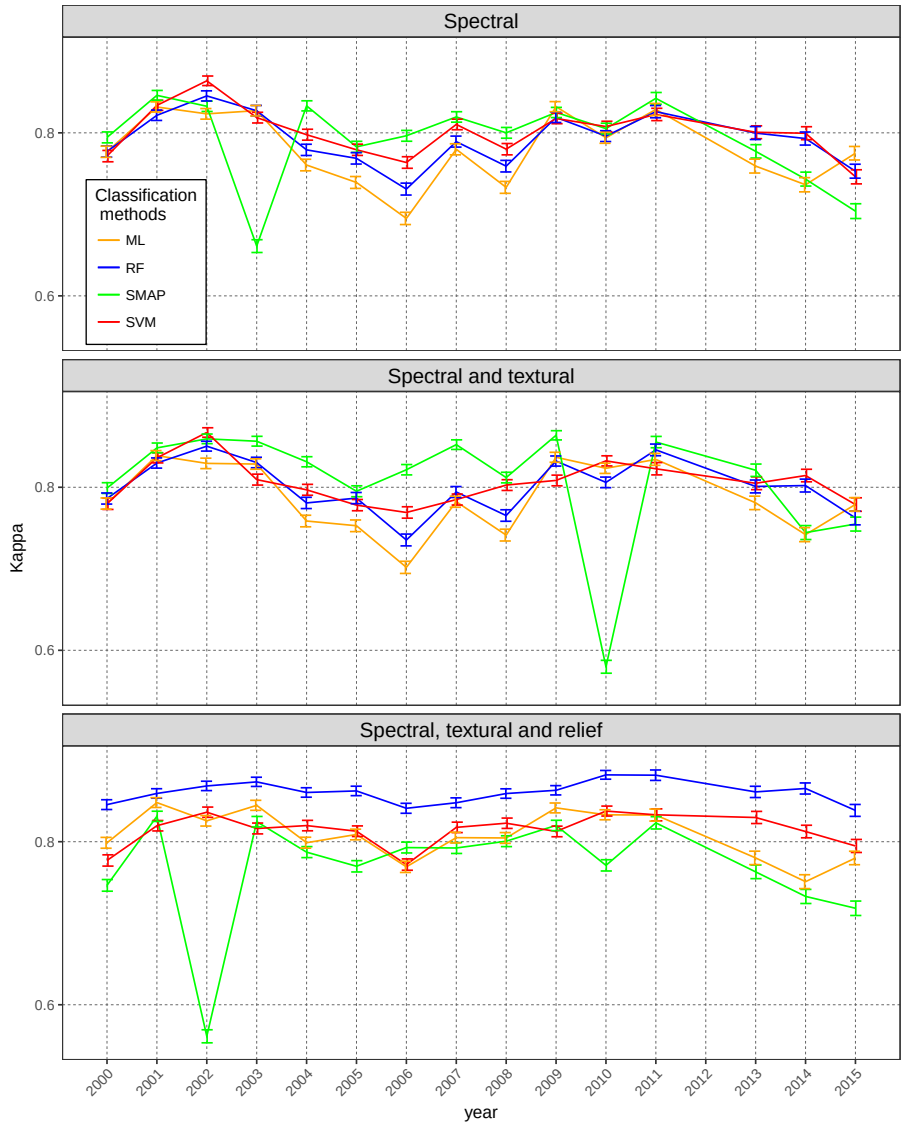

Figure 6. Validation kappa series with 95% confidence intervals for the four algorithms and the three feature combinations, using four, trhee or two seasons depending on their availability. Number of classifications: 180.

The analysis of variance based on heteroscedasticity-consistent standard errors (conducted for the 6 years when four images were available) indicated that the effects on kappa values of *method* ($F\,(3, 240) = 14.1289$, $p < 0.0001$), *seasons* ($F\,(3, 240) = 20.7537$, $p < 0.0001$), *variable set* ($F\,(2, 240) = 51.1315$, $p < 0.0001$), and the interaction between *method* and *seasons* ($F\,(6, 240) = 9.6975$, $p < 0.0001$) were significant.

In relation with the main effects (Figure 7 (a)), SMAP ($M = 0.7561, SD = 0.084$, group a) is the least accurate method, whereas accuracy differences in the other methods are not statistically significant (group b), although RF seems slightly better. In relation to the *feature sets* effect (Figure 7 (b)), the value of the SpTexRel factor ($M = 0.7987, SD = 0.0512$) is significantly more accurate than the others (group b), indicating that this feature combination substantially improves accuracy. Finally, the *multi-seasonal* effect (Figure 7 (c), not evaluated when optimising, indicates that including different images for different seasons improves accuracy as phenological differences among classes are better identified. Although the best results are obtained with four seasons ($M = 0.7991, SD = 0.0467$, group b) , there were not significant differences with the two seasons and three seasons options. Only the

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 September 2017 doi:10.20944/preprints201709.0009.v1

Peer-reviewed version available at Remote Sens. 2017, 9, 1058; doi:10.3390/rs9101058

one-season classification ($M = 0.7573; SD = 0.0523$, group a) is significantly less accurate than the others.

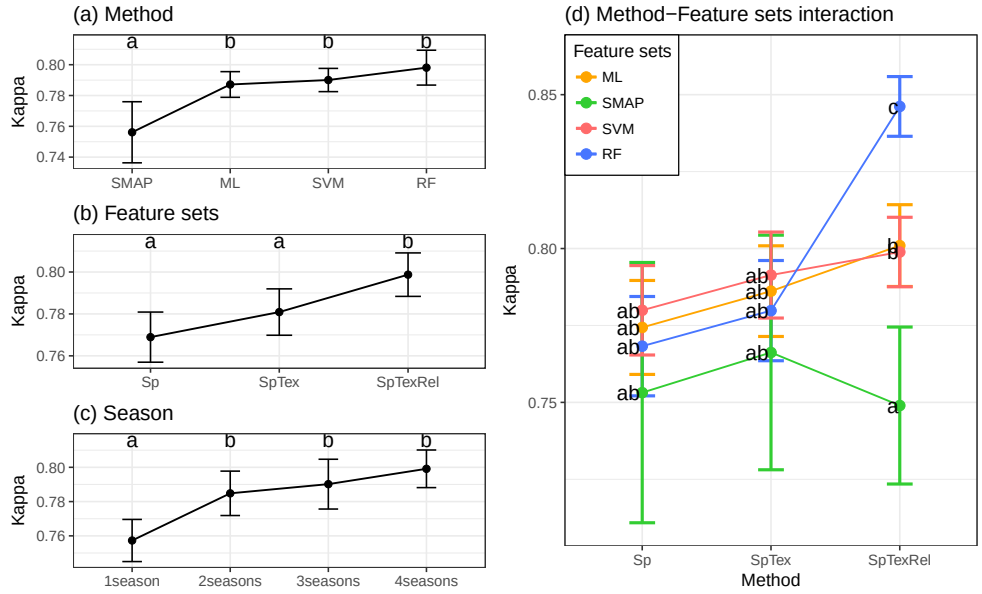

Figure 7. Main effects and simple effects in significant factors (mean ±). Means with different letters are significantly different (Tukey-Kramer contrast using HC3, alpha=0.05). Number of classifications: 288.

ML classification is usually reported to be less accurate than machine learning methods; however, in this study, its kappa values are quite high with all feature combinations. When considering only the spectral variables, this algorithm is significantly better than RF and SVM in some years, such as 2001. In the remaining years it provides results that are almost as good as RF and SVM. We think these results are due to the precautions taken when the training and validation areas were selected and the resulting reduction in noise and outliers in the dataset. Such results indicate that machine learning methods outperform classical statistical methods when data is noisy indicating their greater robustness, although this is not necessarily the case when noise is removed.

Finally, although LM, RF and SVM were not significantly different, the significance of the interaction *Method-Feature sets* (Figure 7 (d)) indicates that the accuracy of RF is significantly higher than in other methods when using SpTexRel data-set ($M = 0.8462, SD = 0.023$, group c). On the other hand, SMAP provides the lowest kappa values ($M = 0.749, SD = 0.0604$, group a) while its wider confidence intervals are also of note. The four methods follow a similar pattern with a slight increase in accuracy when new features are added; however, the difference is not significant.

4.4. Per class validation

Figure 8 summarises user's and producer's kappa values for each feature combination and algorithm used in the time series. The classes that correspond to natural vegetation (Forest and Scrub) are classified more accurately than other classes in almost all cases (kappa around 0.9 or higher) and show less dispersion. These two natural vegetation classes are frequently confused. In both cases, the best classifier seems to be SMAP with spectral and textural features, although the differences are very small.

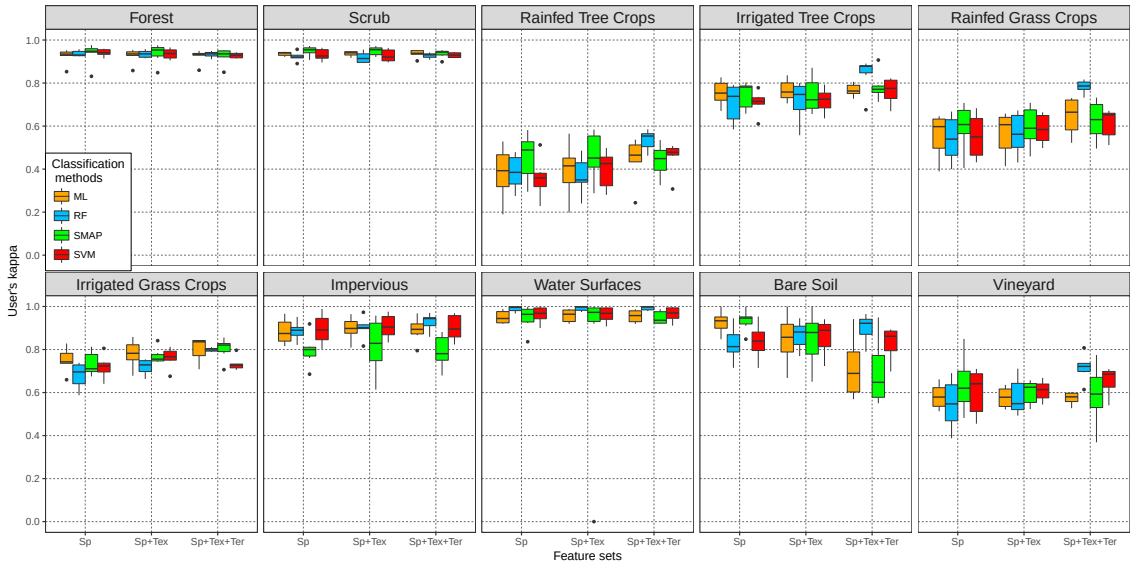

Figure 8. Per class user's average conditional kappa values (user's and producer's) using four season imagery. Number of classifications: 72. Sp: Spectral features; Tx: Textural features; Tr: Terrain features.

Figure 9 shows the confusion matrix for 2015. In the diagonal, the number of correctly classified pixels and the user's (left) and producer's (right) kappa values are represented. The remaining cells show missclassified pixels, the percentage of pixels of one class incorrectly assigned to other classes, and the percentage of pixels incorrectly assigned to the class analysed. the colour gradient reflects the average of both percentages.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 September 2017 doi:10.20944/preprints201709.0009.v1

Peer-reviewed version available at Remote Sens. 2017, 9, 1058; doi:10.3390/rs9101058

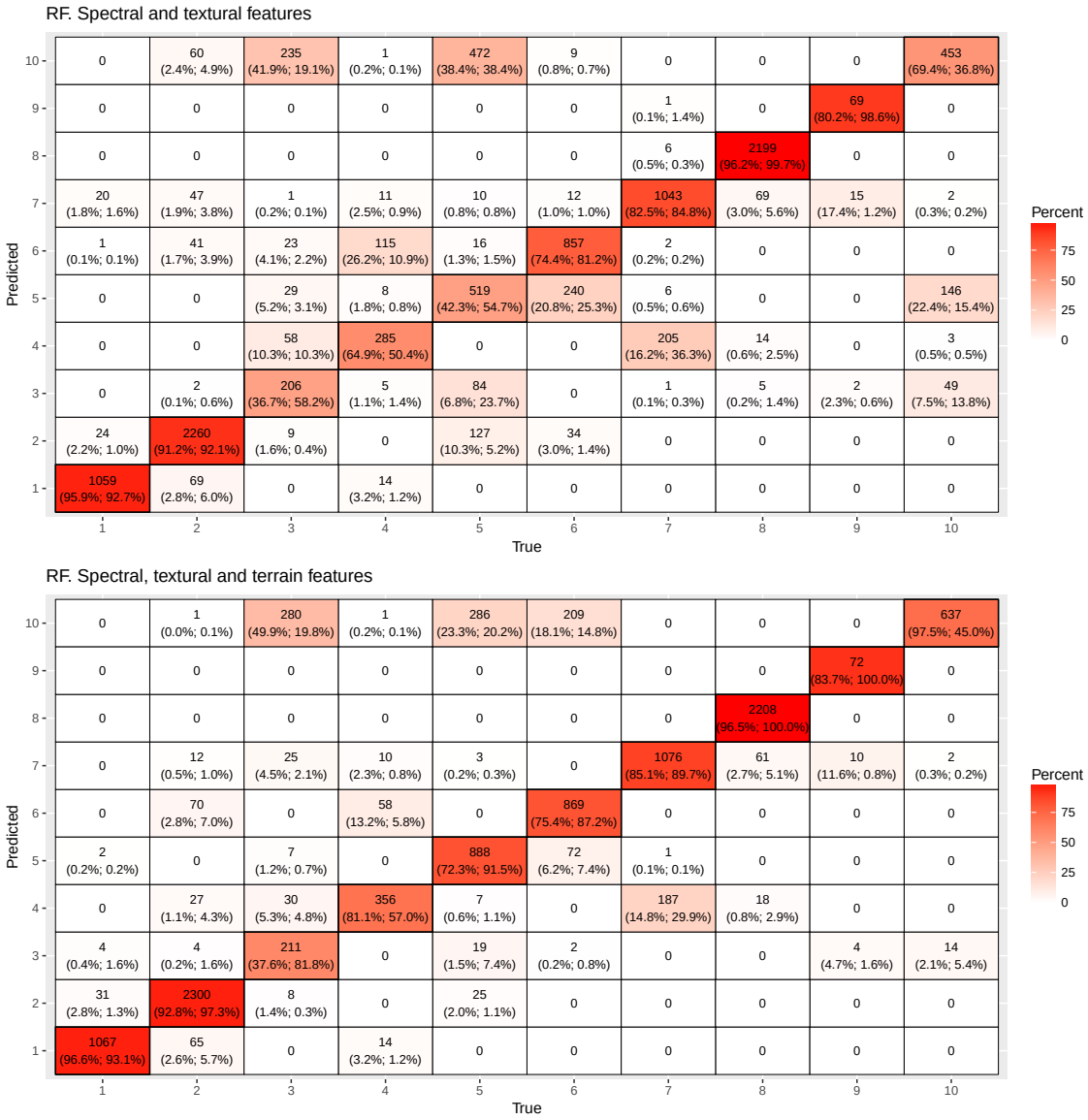

Figure 9. Four season confusion matrix in 2015 using spectral and textural features (top) and spectral, textural and terrain features (bottom) with Random Forest algorithm. In the diagonal appear the number of correctly classified pixels and the user's (left) and producer's (right) kappa values. Outside the diagonal appear the number of confusions, the percentage of the column class incorrectly classified as the row class (left) and the percentage of the row class that truly belongs to column class (right).

Natural uses (Forest and Scrub) and water are the most accurately classified land uses with a reduction in error when relief features are included. The greatest confusion appears among classes that are similar, both in terms of reflectivity and in terms of agronomic properties. For instance, when using the three feature sets, RF classifies almost 50% of the Rainfed tree crops as Vineyard; the former is the class least accurately classified in most of the classifications. This sort of confusion appears especially in the surrounding of Elda and Villena (Figura 10 bottom) and produces an overestimation of vineyards. The inclusion of new features does not contribute to the accuracy. The best option to classify this class is SMAP with three feature sets, followed by SMAP with spectral and textural features.

Irrigated tree crops have higher user's and producer's kappa values than the previous class. They also present less dispersion, especially the producer's kappa, whose values are between 0.6

and 0.8. When analysing user's kappa, it is clear that SMAP and LM obtain more accurate values (kappa larger than 0.8), whereas RF and SVM kappa values are between 0.6 and 0.8, except when using terrain features, when the values are around 0.8. RF also benefits from the inclusion of terrain features (SVM only in user's kappa) and ML producer's kappa is reduced when terrain features are included. Confusion occurs with Irrigated grass crops and, to a lesser extent, with Rainfed tree crops and Rainfed grass crops. As a whole, RF with the three feature sets produces the most accurate classification followed by LM with spectral and textural features.

Another frequent confusion is the misclassification of Impervious surfaces such as Irrigated tree crops. The mixing of both uses is frequent in orchard landscapes near the Mediterranean coast. This fact and the high spectral variability of urban land use explain this confusion.

Rainfed grass crops have, in general, low producer's kappa values (between 0.2 and 0.6) and higher user's kappa (between 0.5 and 0.9). Confusion tends to occur with Irrigated grass crops and, to a lesser extent, with Rainfed tree crops and Vineyards. All methods increase user's accuracy when adding new feature sets, but only LM and RF do the same with producer's kappa, while SMAP clearly reduces accuracy. In general, the best option is LM with the three feature sets followed by SMAP with spectral and textural features.

Irrigated grass crops (including orchard areas) is a heterogeneous class that reaches quite high accuracy values (producer's kappa between 0.5 and 0.9, and user's kappa between 0.6 and 1). All methods improves both kappas when new features are included. Confusion occurs with class Rainfed grass crops and, to a lesser extent, with Irrigated tree crops and Rainfed tree crops. RF with the three feature sets is, in general, the best classification option followed by SMAP or LM with spectral and textural features.

Impervious surfaces (buildings and infrastructures) are classified with great accuracy. However, of interest are the SMAP lower user's kappa values (0.6-0.9) while all the other methods have values between 0.7 and 1. Producer's kappa is higher, with values above 0.9 except for SVM. RF benefits in both kappas from the inclusion of new feature sets. Confusion occurs with classes like Bare soil, Irrigated grass crops and Rainfed tree crops. The best classification option is RF with the three feature sets. No significant differences appear among algorithms when spectral and textural features are used to classify.

Water surfaces (artificial or natural) show very high accuracy values, especially RF, which reaches 100% with any combination of feature sets. SMAP with spectral and textural features also produces very high kappa values. Confusion happens mainly with the Impervious class.

Bare soil is represented exclusively by a quarry in the south of the study area, the area of which is small compared with the other classes. The user's kappa values are high and similar for the different algorithms, increasing when new features are added. The producer's kappa is in most cases in the range 0.6-0.8 in the case of RF and SVM, and between 0.8 and 1 in the case of SMAP and ML. Confusion appears with class Impervious and to a lesser extent with Rainfed tree crops. In these cases the best results are reached with LM and SMAP, using just spectral features.

Finally, Vineyard includes both irrigated and rainfed systems because their separability is very low. This class is easily confused with similar crops, like Rainfed tree crops. A high dispersion is observed in kappa values. Producer's kappa is in the range 0.6-0.8 for LM, around 0.9 for RF and SVM with the three feature sets, and around 0.5 for SMAP with the three feature sets. Confusion appears with classes Rainfed tree crops, Rainfed grass crops and Irrigated grass crops. The best results are obtained, as a rule, with RF and SVM, whose values improved when terrain features were added. SMAP with spectral and textural features also gives a high accuracy.

4.5. Visual validation

Figure 10 show the land use maps obtained with RF, using spectral and textural features (top) and spectral, textural and terrain features (bottom) in 2015. Although this is not the year in which the

best results are obtained (see Figure 6), the kappa value is still quite high when using spectral, textural and terrain features ($k = 0.838$). In addition , it is the most recent year of the series.

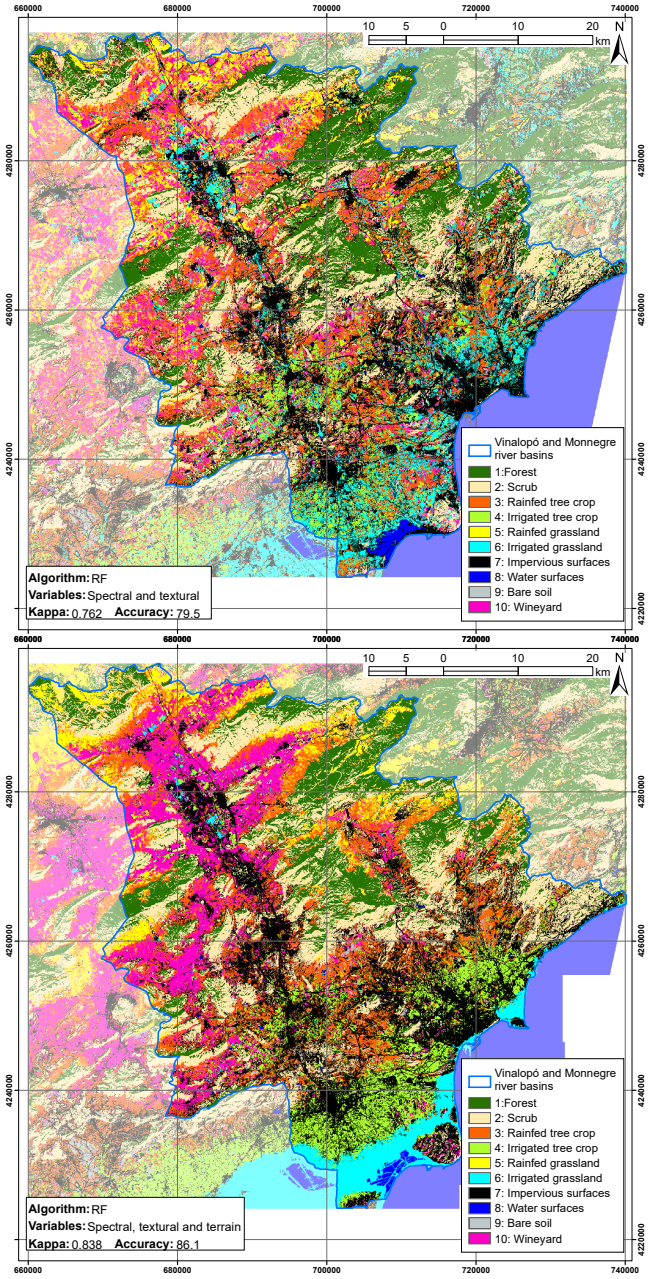

Figure 10. Four seasons imagery clasification in 2015 using spectral and textural features (top) and spectral, textural and terrain features (bottom) using Random Forest algorithm.

SMAP is the classifier that produces the visually poorest result because of a strong overestimation of class 7 (Impervious surfaces), although in some years, such as 2009, it produces a good classification both visually and quantitatively ($k = 0.864$).

Some serious errors appear when ML or SMAP are used as the Santa Pola coastal marshes are classified as urban; both RF andd SVM classify this area correctly, although some overestimation of Impervious surfaces is observed when including the three feature sets. In the case of RF, we think the reason is that the biased distribution of urban areas in the study area, concentrated in low height and low slope sites, might have led RF to missclassify low height cells as urban. In addition, the importance

of the variables obtained with RF give much higher importance to relief features than to the spectral or textural.

In other cases, mistakes were observed in coastal areas even in the most accurate classifications (Figure 10, bottom), classifying as class 6 (irrigated grassland) some coastal urban areas (San Juan or Santa Pola) and, at the same time, expanding the urban areas beyond its real limit.

In summary, it is necessary to visually check the classification results to avoid the sort of problems we have mentioned when using very flexible classifiers with a large number of features.

5. Conclusions

Using images from several seasons significantly improves accuracy, as the differences in the phenological calendars of different land covers are taken into account; however, there are no significant differences between using two or more images. When classifying only one image per year, summer is the best option, probably due to strong differences in the water content between irrigated and non-irrigated covers.

Parameter optimisation in RF and SVM does not improve accuracy significantly but greatly increased the computing time. However, it may be interesting to try other more expensive methods of calibration, such as a simultaneous calibration parameters in grid, or use other resampling methods.

Adding textural features to the spectral features does not increase accuracy significantly, but when terrain features are also added, there is a significant increase in accuracy.

SMAP is significantly less accurate than ML, RF or SVM, these latter algorithms not significantly differing in terms of accuracy. However, the interaction between feature sets and algorithm produces a significant increase in RF accuracy over SVM and ML when terrain features are added. We think that the good accuracy of LM is partly due to the restrictions taken into account when the training and validation areas were identified and to outlier elimination.

Although some algorithms and feature subsets perform better overall, the results are not so clear when classes are analysed. Some classes benefit from the inclusion of additional feature sets, but others do not. This may indicate that the combined use of different algorithms depending on classes can be a good line to follow to get better results. Similarly, it may be interesting to use different combination strategies depending on the class of interest. Thus, the variables derived from the MDE should not be used in classes such as urban, while it may be a good strategy in natural classes as forest or scrub.

Water surfaces, Forest and Scrub are the most accurately classified land use. The problem arises with crops, especially with Sparse tree crops, mostly in rainfed land. In this case, the greatest confusion occurred with Vineyards due to the similar characteristics of both uses. Confusion errors were also detected in Rainfed grass crops, especially errors of commission, almost always resulting in confusion with other crops, especially Irrigated grass crops.

Supplementary Materials: The R scripts and GRASS python scripts (folder grassr-clasif) and a test database (folder database) are available as supplementary material in the compressed file grassr-clasif.zip.

Acknowledgments: The work was developed within the project *Modelización Hidrológica en Zonas Semiáridas, Subproyecto: Modelización Numérica de Procesos Hidrológicos y Sistemas de Recursos Hídricos*, leaded by F-IEA and INUAMA, and funded by *D.G. de Investigación y Política Científica de la Consejería de Educación, Ciencia e Investigación de la Región de Murcia*. This work partly results from a post-doctoral contract included in the *Programa Saavedra Fajardo* (20023/SF/16) funded by *Consejería de Educación y Universidades de la Comunidad Autónoma de la Región de Murcia* throught *Fundación Séneca-Agencia de Ciencia y Tecnología de la Región de Murcia*.

Author Contributions: The three authors contributed equally to this work.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Alrababah, M.; Alhamad, M. Land use/cover classification of arid and semi-arid Mediterranean landscapes using Landsat ETM. *International Journal of Remote Sensing* **2006**, *27*, 2703–2718.

2. Di Palma, F.; Amato, F.; Nolè, G.; Martellozzo, F.; Murgante, B. A SMAP Supervised Classification of Landsat Images for Urban Sprawl Evaluation. *ISPRS International Journal of Geo-Information* **2016**, *5*, 109.
3. Berberoglu, S.; Curran, P.; Lloyd, C.; Atkinson, P. Texture classification of Mediterranean land cover. *International Journal of Applied Earth Observation and Geoinformation* **2007**, *9*, 322–334.
4. Senf, C.; Leitão, P.J.; Pflugmacher, D.; van der Linden, S.; Hostert, P. Mapping land cover in complex Mediterranean landscapes using Landsat: Improved classification accuracies from integrating multi-seasonal and synthetic imagery. *Remote Sensing of Environment* **2015**, *156*, 527 – 536.
5. Maselli, F.; Conese, C.; Petkov, L.; Resti, R. Inclusion of prior probabilities derived from a nonparametric process into the maximum likelihood classifier. *Photogrammetric Engineering and Remote Sensing* **1992**, *58*, 201–207.
6. Breiman, L. Random Forests. *Machine Learning* **2001**, *45*, 5–32.
7. Cortes, C.; Vapnik, V. Support-vector network. *Machine Learning* **1995**, *20*, 1–5.
8. McCauley, J.; Engel, B. Comparison of scene segmentations: SMAP, ECHO, and maximum likelihood. *Geoscience and Remote Sensing, IEEE Transactions on* **1995**, *33*, 1313–1316.
9. Ehsani, A. Evaluation of Sequential Maximum a Posteriori (SMAP) Method for Land Cover Classification. Geomatics 90 (National Conference & Exhibition), 2011.
10. Li, M.; Im, J.; Beier, C. Machine learning approaches for forest classification and change analysis using multi-temporal Landsat TM images over Huntington Wildlife Forest. *GIScience & Remote Sensing* **2013**, *50*, 361–384.
11. Sluiter, R.; Pebesma, E.J. Comparing techniques for vegetation classification using multi- and hyperspectral images and ancillary environmental data. *International Journal of Remote Sensing* **2010**, *31*, 6143–6161.
12. He, J.; Harris, J.; Sawada, M.; Behnia, P. A comparison of classification algorithms using Landsat-7 and Landsat-8 data for mapping lithology in Canada's Arctic. *International Journal of Remote Sensing* **2015**, *36*, 2252–2276.
13. Rodríguez-Galiano, V. Metodología basada en teledetección, SIG y geoestadística para cartografía y análisis de cambios de cubiertas del suelo de la Provincia de Granada. PhD thesis, Departamento de Geodinámica, Universidad de Granada, 2011.
14. Rodriguez-Galiano, V.; Ghimire, B.; Rogan, J.; Chica-Olmo, M.; Rigol-Sanchez, J. An assessment of the effectiveness of a random forest classifier for land-cover classification. *{ISPRS} Journal of Photogrammetry and Remote Sensing* **2012**, *67*, 93 – 104.
15. Belgiu, M.; Drăguţ, L. Random forest in remote sensing: A review of applications and future directions. *ISPRS Journal of Photogrammetry and Remote Sensing* **2016**, *114*, 24 – 31.
16. Ehsani, A.; Quiel, F. Efficiency of Landsat ETM+ Thermal Band for Land Cover Classification of the Biosphere Reserve "Eastern Carpathians" (Central Europe) Using SMAP and ML Algorithms. *International Journal of Environmental Research* **2010**, *4*, 741–750.
17. Kumar, U.; Dasgupta, A.; Mukhopadhyay, C.; Ramachandra, T. Advanced Machine Learning Algorithms based Free and Open Source Packages for Landsat ETM+ Data Classification. Proceedings of the OSGEO-India: FOSS4G 2012- First National Conference: OPEN SOURCE GEOSPATIAL RESOURCES TO SPEARHEAD DEVELOPMENT AND GROWTH. 25-27th October 2012, 2012, pp. 1–7.
18. Elumnoh, A.; Shrestha, R. Application of DEM data to Landsat image classification: evaluation in a tropical wet-dry landscape of Thailand. *Photogrammetric Engineering & Remote Sensing* **2000**, *66*, 297–304.
19. Lu, D.; Weng, Q. A survey of image classification methods and techniques for improving classification performance. *International Journal of Remote Sensing* **2007**, *28*, 823–870.
20. Zhou, Q.; Robson, M. Contextual information is ultimately necessary if one is to obtain accurate image classifications. *International Journal of Remote Sensing* **2001**, *22*, 612–625.
21. Eisavi, V.; Homayouni, S.; Yazdi, A.M.; Alimohammadi, A. Land cover mapping based on random forest classification of multitemporal spectral and thermal images. *Environmental Monitoring and Assessment* **2015**, *187*, 187–291.
22. Gómez, C.; White, J.C.; Wulder, M.A. Optical remotely sensed time series data for land cover classification: A review. *{ISPRS} Journal of Photogrammetry and Remote Sensing* **2016**, *116*, 55 – 72.
23. CHJ. Plan Hidrológico de la Demarcación Hidrográfica del Júcar. Technical report, Cemarcación Hidrográfica del Júcar, Ministerio de Medio Ambiente, 2015.

24. Bossard, M.; Feranec, J.; Otahel, J. *CORINE land cover technical guide - Addendum 2000*; Technical report No 40, European Environment Agency: Kongens Nytorv 6, DK-1050 Copenhagen K, Denmark, 2000.
25. Chávez, P. An improved dark-object substraction technique for atmospheric scattering correction of multispectral data. *Remote Sensing of Environment* **1988**, *24*, 459–479.
26. Teillet, P.; Guindon, B.; Goodenough, D. On the slope-aspect correction of multispectral scanner data. *Canadian Journal of Remote Sensing* **1982**, *8*, 84–106.
27. Swain, P.; Davis, S.E. *Remote Sensing: The Quantitative Approach*; McGraw-Hill: New York, USA, 1976; p. 396.
28. Liaw, A.; Wiener, M. Classification and Regression by randomForest. *R News* **2002**, *2*, 18–22.
29. Gislason, P.; Benediktsson, J.; Sveinsson, J. Random Forests for land cover classification. *Pattern Recognition Letters* **2006**, *27*, 294–300. Pattern Recognition in Remote Sensing (PRRS 2004).
30. Ghimire, B.; Rogan, J.; Miller, J. Contextual land-cover classification: incorporating spatial dependence in land-cover classification models using random forests and the Getis statistic. *Remote Sensing Letters* **2010**, *1*, 45–54.
31. Pal, M. Random forest classifier for remote sensing classification. *International Journal of Remote Sensing* **2005**, *26*, 217–222.
32. Prasad, A.; Iverson, L.; Liaw, A. Newer Classification and Regression Tree Techniques: Bagging and Random Forests for Ecological Prediction. *Ecosystems* **2006**, *9*, 181–199.
33. Cutler, D.; Edwards Jr., T.; Beard, K.; Cutler, A.; Hess, K.; Gibson, J.; Lawler, J. Random forest for classification in ecology. *Ecology* **2007**, *88*, 2783–2792.
34. Cánovas-García, F.; Alonso-Sarría, F.; Gomariz-Castillo, F.; Oñate-Valdivieso, F. Modification of the random forest algorithm to avoid statistical dependence problems when classifying remote sensing imagery. *Computers & Geosciences* **2017**, *103*, 1–11.
35. Hastie, T.; Tibshirani, R.; Friedman, J. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, second ed.; Springer, 2009.
36. Camps-Valls, G.; Bruzzone, L., Eds. *Kernel Methods for Remote Sensing Data Analysis*, first ed.; John Wiley & Sons, Ltd: UK, 2009.
37. Mountrakis, G.; Im, J.; Ogole, C. Support vector machines in remote sensing: A review. *{ISPRS} Journal of Photogrammetry and Remote Sensing* **2011**, *66*, 247–259.
38. Vapnik, V. *Statistical Learning Theory*, 1 ed.; Wiley Interscience, 1998; p. 736.
39. Gualtieri, J.; Cromp, R. Support Vector Machines for Hyperspectral Remote Sensing Classification. 27th AIPR Workshop: Advances in Computer Assisted Recognition., 1998.
40. Melgani, F.; Bruzzone, L. Classification of hyperspectral remote sensing images with support vector machines. *IEEE Transactions on Geoscience and Remote Sensing* **2004**, *42*, 1778–1790.
41. Tso, B.; Mather, P. *Classification Methods for Remotely Sensed Data, Second Edition*, second ed.; Taylor & Francis, 2009; p. 352.
42. Auria, L.; Moro, R. Support Vector Machines (SVM) as a Technique for Solvency Analysis. Discussion Papers of DIW Berlin 811, DIW Berlin, German Institute for Economic Research, 2008.
43. Bouman, C.; Shapiro, M. Multispectral Image Segmentation using a Multiscale Image Model. Proceedings of IEEE International Conference on Acoustic, Speech and Signal Processing; , 1992; pp. III 565–III 568.
44. Bouman, C.; Shapiro, M. A Multiscale Random Field Model for Bayesian Image Segmentation. *IEEE Transactions on Image Processing* **1994**, *3*, 162–177.
45. Cheng, H.; Bouman, C.A. Multiscale Bayesian Segmentation Using a Trainable Context Model. *IEEE Trans. on Image Processing* **2001**, *10*, 511–525.
46. Mather, P.; Koch, M. *Computer Processing of Remotely-Sensed Images: An Introduction*, 4 ed.; Computer Processing of Remotely Sensed Images: An Introduction, Wiley, 2010.
47. Molinaro, A.; Simon, R.; Pfeiffer, R. Prediction error estimation: a comparison of resampling methods. *Bioinformatics* **2005**, *21*, 3301–3307.
48. Kuhn, M.; Johnson, K. Over-Fitting and Model Tuning. In *Applied Predictive Modeling*; Springer New York, 2013; pp. 61–92.
49. Caputo, B.; Sim, K.; Furesjo, F.; Smola, A. Appearance-based object recognition using SVMS: which kernel should i use? Proceedings of NIPS Workshop on Statitsical Methods for Computational Experiments in Visual Processing and Computer Vision, 2002.

50. Karatzoglou, A.; Smola, A.; Hornik, K.; Zeileis, A. kernlab – An S4 Package for Kernel Methods in R. *Journal of Statistical Software* **2004**, *11*, 1–20.
51. Neteler, M.; Mitasova, H. *Open source GIS. A GRASS GIS approach*, 3 ed.; Vol. 773, *The International Series in Engineering and Computer Science*, Springer, New York, 2008; p. 486.
52. Neteler, M.; Bowman, M.; Landa, M.; Metz, M. GRASS GIS: A multi-purpose open source GIS. *Environmental Modelling & Software* **2012**, *31*, 124–130.
53. Venables, W.; Smith, D.; the R Development Core Team. *An Introduction to R*, 2012.
54. R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2012.
55. Cohen, J. A coefficient of agreement for nominal scales. *Educational & Psychological Measurement* **1960**, *20*, 37–46.
56. Congalton, R.; Mead, R. A Quantitative Method to Test for Consistency and Correctness in Photointerpretation. *Photogrammetric Engineering and Remote Sensing* **1983**, *49*, 69–74.
57. Chuvieco, E. *Fundamentals of Satellite Remote Sensing. An Environmental Approach*; CRC Press, 2016.
58. Coleman, J. *Measuring concordance in attitudes*; Unpublished manuscript. Department of Social Relations, Johns Hopkins University, 1966.
59. Rosenfield, G.; Fitzpatrick-Lins, K. A coefficient of agreement for nominal scales. *Photogrammetric Engineering & Remote Sensing* **1986**, *52*, 223–227.
60. Hudson, W.; Ramm, C. Correct formulation of the Kappa coefficient of agreement (in remote sensing). *Engineering & Remote Sensing* **1987**, *53*, 421–422.
61. Landis, J.; Koch, G. The Measurement of Observer Agreement for Categorical Data. *Biometrics* **1977**, *33*, 159–174.
62. Long, J.; Ervin, L. Using Heteroscedasticity Consistent Standard Errors in the Linear Regression Model. *The American Statistician* **2000**, *54*, 217–224.