A PERCEPTRON-BASED FEATURE SELECTION APPROACH FOR DECISION TREE CLASSIFICATION

Carla Jaqueline Casaroti[1] - ORCID: 0000-0002-1250-2839

Jorge Antonio Silva Centeno[2] - ORCID: 0000-0002-2669-7147

Stephan Fuchs[3] - ORCID: 0000-0002-1766-196X

[1]Universidade federal do Paraná – UFPR, Departamento de Geomática, Curitiba - Paraná, Brasil.
E-mail: carlacasaroti@gmail.com

[2]Universidade Federal do Paraná – UFPR, Departamento de Geomática, Curitiba - Paraná, Brasil.
E-mail: centeno@ufpr.br

[3]Karlsruher Institut für Technologie – KIT, Fachbereich Siedlungswasserwirtschaft und Wassergütewirtschaft, Karlsruhe – Baden-Württemberg, Deutschland.
E-mail: stephan.fuchs@kit.edu

Abstract:

The use of OBIA for high spatial resolution image classification can be divided in two main steps, the first being segmentation and the second regarding the labeling of the objects in accordance with a particular set of features and a classifier. Decision trees are often used to represent human knowledge in the latter. The issue falls in how to select a smaller amount of features from a feature space with spatial, spectral and textural variables to describe the classes of interest, which engenders the matter of choosing the best or more convenient feature selection (FS) method. In this work, an approach for FS within a decision tree was introduced using a single perceptron and the Backpropagation algorithm. Three alternatives were compared: single, double and multiple inputs, using a sequential backward search (SBS). Test regions were used to evaluate the efficiency of the proposed methods. Results showed that it is possible to use a single perceptron in each node, with an overall accuracy (OA) between 77.6% and 77.9%. Only SBS reached an OA larger than 88%. Thus, the quality of the proposed solution depends on the number of input features.

Keywords: Feature Selection (FS); Perceptron; Decision tree.

1. Introduction

The increasing development of multi and hyperspectral sensors, as well as the object-based image analysis techniques for classifying high spatial resolution satellite imagery, led to a large amount of data, as illustrated by the substantial set of features available to describe classes of interest in the classification step. As claimed by Haertel and Landgrebe (1999), a high dimensional feature space might cause problems in the estimation of the classes' covariance matrices. When dealing with a parametric classifier, as the feature space dimensionality increases, so does the number of samples required to provide a reliable estimate of the covariance matrix, which is known as the Hughes phenomenon. In his work, Hughes (1968) concluded that the rise in the number of features reflects in the rise of the classifier accuracy, until a maximum accuracy value is reached. Therefore, adding new features to the classification algorithm might reduce the accuracy instead of improving it. Other than that, employing all the features available means a higher computational cost, such as happens with the use of spectral, spatial and textural object image attributes, and methods to reduce the feature space dimensionality have been studied by several authors (Guo et al., 2006; Gasca et al. 2006; Zhang and Chau, 2009; Bartenhagen et al 2010; Geng et al., 2014; Xie et al., 2018 Habermann et al., 2018). Also, decreasing the number of descriptors in the feature space can also reduce the computational cost, for it requires less storage capacity.

Some studies have proposed the use of decision trees as an alternative to the statistical approaches commonly used to conduct image classification and it became popular mainly by its use in the object-based image analysis algorithms (Van Coillie et al., 2007; Mahmoudi et al., 2013; Hamediantar and Shafri 2016; Wang et al., 2016). When it comes to object-based image classification, the system first divides the image into segments according to partitioning parameters and then classifies the segments using spatial, spectral or textural features. The segments classification can be performed with a decision tree proposed by the user. This tree is supposed to represent the human knowledge on recognizing different targets as well as the user's experience, betaking rules and class descriptors derived from the feature space. Once defined the tree, that is, the classes to be represented, the next step lies in proposing the rules to be applied in each tree node. The question addressed in this work is on choosing the features to be analyzed in each node of a given decision tree.

Another critical point to be highlighted is how these classification techniques are part of Machine Learning algorithms. Arthur Samuel introduced the Machine Learning (ML) concept in 1959 based on a simple phrase "How can computers learn to solve problems without being explicitly programmed?". Thus, a comprehension of classification within ML algorithms context, like random forests and decision trees, might be stated along these lines: use of mathematical models, based on sample data in which the algorithm is trained with the training set of samples; then, with the test set, the model accuracy check.

Regarding the focus of this work, a search strategy was drawn to choose the best feature or a combination of two or more features that could best describe and distinguish two classes in each tree node. The proposed method was based on the principles of a single perceptron; one advantage of this technique is that it does not require a specific statistical distribution, unlike the Transformed Divergence, for instance. Another advantage is that it remains proposing a linear combination of two or more features when necessary. Complementing the used method advantages, one of the most important goals in this work is to improve high spatial resolution satellite imagery classification using a simple and yet effective Machine Learning algorithm.

2. Literature Review

Over the last decade, several scientific studies have been using object-based image analysis along with feature selection methods to improve high spatial resolution image (HSRI) classification. Among the examples found in the

literature are Van Coillie et al. (2007), who used genetic algorithms as feature selection methods; Mahmoudi et al. (2013), who applied multi-agent recognition systems; Hamediantar and Shafri (2016), that relied on the C 4.5 algorithm; Wang et al. (2016), who used the support vector machine and random forest classifiers. Meanwhile, even though using spatial information (objects/segments) is highly indexed to classify HSRI, some authors, e.g., Jung and Ehlers (2016), Persello and Bruzzone (2016) and Pu and Bell (2017), still used the pixel-based approach along with feature selection.

Using the object as a processing element opens up a wide array of possibilities in image classification, in which it is possible to combine both spectral and spatial features to describe the classes of interest. Meanwhile, choosing the features that characterize the classes as reliably as possible can be a challenging task due to the large number of combinations that can be made from all of the spatial and spectral descriptors available, thus providing a large dataset.

First of all, techniques aiming to reduce the dimensionality of datasets can be divided in two main parts, the first one regarding feature extraction and the second one, feature selection. In feature extraction, the original variables are combined to create a smaller set of new features that preserve the information of the initial feature space. A typical feature extraction example is to apply the Principal Components transformation in the original feature space and then select the most significant new features based on their percentage of representation in the entire dataset (Weinberger and Saul, 2006). Other well-known approaches to extract information, such as Locally Linear Embedding and Multidimensional Scaling, were also proposed by Zhang and Chau (2009). Over the last years, researches on deep learning techniques are also bringing alternative solutions, as seen in the use of the restricted Boltzmann machine (Sohn et al. 2013). A drawback of such approaches relates to their intentions to represent the original information from the dataset using a lower number of variables without considering what the selection aims for. In this respect, a same set may not be suitable for different classification schemas.

On the other hand, feature selection methods are used to select, from an original set of features, the subset with the most informative and distinctive variables to solve a particular problem according to a search strategy and an evaluation criterion (Tang et al., 2014). Firstly, the search strategy is based on looking for the most suitable subset among all the features available. Then, the evaluation criterion measures the success of each selected feature subset evaluating, for instance, the classification results. More examples of how the search strategy and evaluation criterion work can be found in Aguilar et al. (2012), Guo et al. (2016) and also Xiurui et al. (2014).

Within this context, feature selection techniques require the establishment of a proposed aim as well as both corresponding evaluation criterion and search strategy (Xie et al. 2018). The search strategy can be simple, such as the evaluation of all possible feature combinations, or may consist of optimized search algorithms based, for example, on genetic algorithms (Ribeiro, 2006). A comparison of feature selection methods in remote sensing is available in Serpico et al. (2003). The most popular feature selection methods can be arranged in two groups: sequential backward search (SBS) and sequential forward search (SFS). The SBS method is an iterative process in which the original set of variables is analyzed. In each iteration the less significant variable is discarded until a desired number of variables is reached or when the derived set reaches a satisfactory stopping criterion for the given problem. The SFS method, on its part, starts by selecting the most significant variable. Subsequently, the variables that are more significant between the ones remaining are added iteratively.

Throughout recent decades, studies introduced advances and new methods in feature selection. Gasca et al. (2006) and Ruck et al. (1990) proposed the use of a multilayer perceptron to solve the problem of high dimensionality in the feature space. Such algorithms select a subset of relevant variables by estimating the relative contribution of the input variables from the corresponding classes' problem to the output neurons (Gasca et al. 2006). Another recent example using the perceptron in feature selection was described in Habermann et al. (2018).

The perceptron is an essential learning algorithm in neural networks and machine learning. Although its formulation is relatively simple, it has proved to be very efficient. A Perceptron is also described as a "binary classifier" since it proposes a function, which can decide whether or not an input represented by a feature vector belongs to

a specific class (1 in the positive case, 0 in the negative case). Therefore, its input is a set of values (feature vector) and its output is either 1 or 0. This algorithm is considered supervised as it develops its classification rule based on inputs with known output given by the user. Formally, the classification rule combines the inputs (feature values) to produce output according to equation 1.

$$f(x) = \begin{cases} 1 \rightarrow if\ \vec{w} * \vec{x} + b > 0 \\ 0 \rightarrow otherwise \end{cases} \quad (1.1)$$

When the input vector contains two values, equation 1 can be written as displayed in equation 2. Note that the decision border is a line in a bidimensional feature space.

$$f(x) = w_1 * x_1 + w_2 * x_2 + b \quad (1.2)$$

When more input variables are used, the function becomes the equation of a plane or a hyperplane in a polynomial behavior with n-weights and a-bias for n-input variables. The weights and b-bias computation depend on the given training dataset and it is achieved with the Backpropagation algorithm. This algorithm performs the weights correction according to the difference between the computed output and the desired one

3. Method

Within the framework of object-based image analysis and classification, the first step is image segmentation. The algorithm used in this study was the well-known multi-scale resolution segmentation (FNEA – fractal net evolution approach), that can be stated as a special kind of region growing algorithm. This method depends on homogeneity definitions in combination with local and global optimization techniques and comprises the choice of three main parameters: scale, shape and compactness. Definitions and more information about these parameters and the segmentation approach can be found in Baatz and Schäpe (2000). Besides spectral features, these regions also allow computing spatial and textural features. These features describe the segments and can be used to label them according to previously defined classes of interest, which leads to the next processing step in OBIA: classification.

When a decision tree is proposed, the next problem is to find the best feature (or set of features) that can be used in each node to support the decisions. A decision tree uses an organized tree-like model of decisions and relations between these decisions to obtain the classification result of the branches. The principle is to break up a complex decision (the desired classes of interest) into several more straightforward choices (nodes), hoping that the final product is the desired and most correct classification (Safavian and Landgrebe, 1991). Although the principle is simple, the algorithm requires applying rules in each node, which should be able to perform the binary classification based on a reduced set of variables and thresholds.

In this work, the idea was to use a single perceptron to find the best fitting of variables (or a variable combination) for each classification network node using only one variable, a pair of variables and the SBS. It is relevant to highlight that the decision tree has nodes with binary solutions only. As described in equation 3, the decision is taken based on a polynomial whose size depends on the number of input variables. One solution that the perceptron offers is the computation of the priori unknown parameters of w, given a set of input features and a set of training samples known as the expected classification output. The Backpropagation algorithm is used to estimate the optimal set of parameters (wi) for the linear function used on the classification node, together with the selected features (xi).

$$f(x) = w_1 * x_1 + w_2 * x_2 + ... + w_n * x_n + b \quad (1.3)$$

The evaluation criterion is the degree of partial accuracy achieved with the probable solution. As there are just two possible classes in each node, the accuracy is measured by the percentage of correct classified elements.

Within the methodology here developed, in a first attempt the perceptron was used to select the best feature for each node. For this purpose, all the features (Tables 1 and 2) were evaluated. The previous step was a relatively

simple task for just two parameters had to be computed: the weight w1 and the bias b in equation 4. The feature that best separated the two proposed classes was the solution and the value of the bias was the desired threshold for separating the classes.

$$0 = w_1 * x_1 + w_2 * x_2 + b \quad (1.4)$$

Finally, the perceptron was used to select the less significant feature within a SBS. The search started with all possible variables and computed the set of weights that enabled an optimal solution. The feature associated with the lower weight was then discarded and the procedure was repeated. This iterative process was undertaken until the binary classification accuracy of each node was above a given value. The results were finally evaluated using test samples. For the evaluation, the final classification was considered, not the results in each node. A confusion matrix was also computed for comparison purposes.

4. Experiment

The study was developed using a mosaic of images level 3A from the RapidEye satellite sensor. This product is radiometric, sensor and geometrically corrected and aligned to a cartographic map projection. The mosaic images were acquired in 2014 from the Vossoroca basin having around 136,58 km2, where the Vossoroca dam is placed, more specifically in the city of Tijucas do Sul, located in the state of Paraná, Brazil (figure 1). The images are georeferenced to the WGS84/UTM22 system coordination and have five meters of spatial resolution and five spectral bands; blue (440 – 510 nm), green (520 – 590 nm), red (630 – 685 nm), red edge (690 – 730 nm) and near-infrared (760 – 850 nm). The Vossoroca dam is used for power generation and the basin is predominantly rural. The many land cover classes present in the scene are "forest cover", "agriculture", "bare soil" and small "urban areas". Figure 1 displays the basin in RGB composition as well as its geographic location.
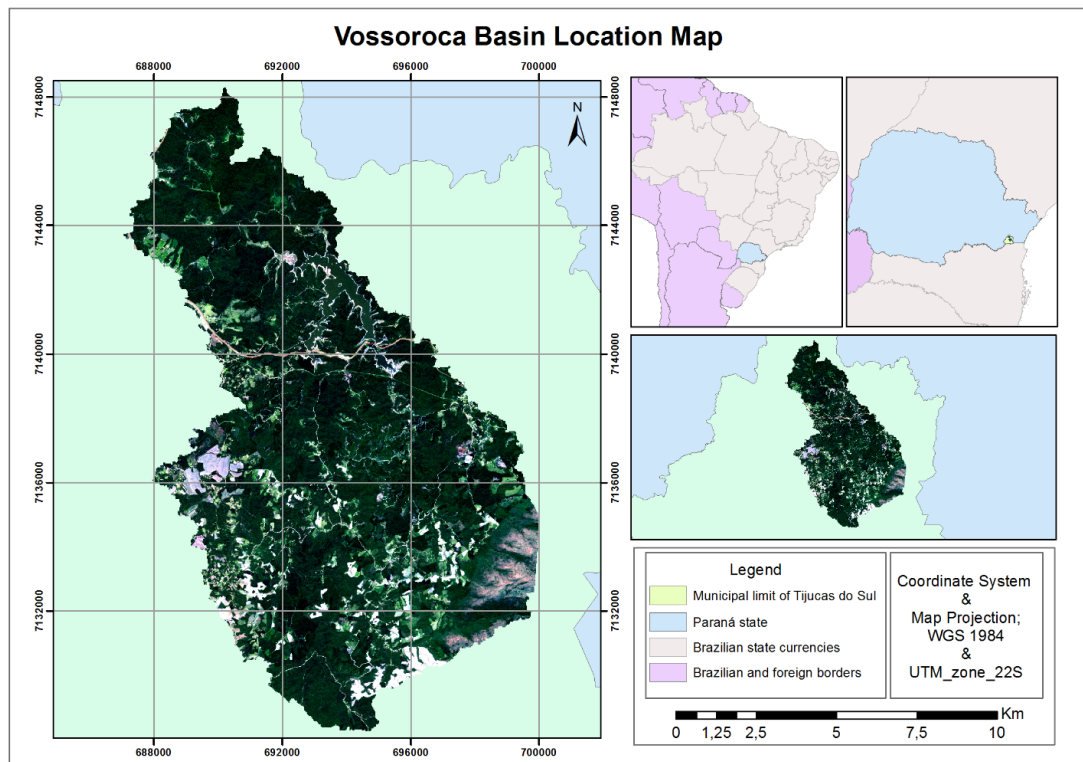
Source: The author

Figure 1: Area of interest location map – Vossoroca basin.

This work is a branch of Mudak, a research project of hydrology and non-point pollution sources in the Vossoroca basin. The project is a cooperation between universities and enterprises from Brazil and Germany. Therefore, the land cover classes of interest inserted into the hierarchical semantic network were defined based on the knowledge of both hydrology and remote sensing researchers involved in the project, according to the desired inputs for hydrological models. Thus, the main classes of interest were divided into: "agriculture" to represent the agricultural activity portions of the land cover, "woods" to represent the portions with grass and some single trees around, "forest" regarding the natural forest and reforestation areas, "urban area" to represent the roads and the buildings, "bare land" to represent the agricultural portions without vegetation cover and "water" to represent the reservoir which is the main water body, as well as the other ones.

To easy the classification processing steps, the hierarchical semantic network was used. The first node was composed by "water" and "non-water" classes to distinguish water bodies from the other classes. At that point, from the "non-water" node, the classes "vegetation" and "non-vegetation" were used to split segments with and without vegetation coverage. From the "vegetation" node "low-vegetation" was then named as the final "agriculture" class. The "high vegetation" part was then used to separate "forest" and "woods". The "non-vegetation" (named "bare" in Figure 3) was implemented to distinguish segments without any coverage comprising the classes "bare land" and impervious areas which were designated as "urban". In that way, as the main classes are water, agriculture, woods, forest, bare land and urban, those are the ones which will be used to analyze the results confusion matrix wise.

In respect of image segmentation, only one level was used. This decision was made to test the feature selection algorithm without further significant influence of the user, who could optimize the segmentation using his/her knowledge. The parameters for the segmentation were 100 for the scale parameter and 0.5 for both shape and compactness parameters. All the spectral bands were considered in the segmentation step, attributing the weights of 1 for the red, red edge, green and blue bands, and 2 for the near-infrared (NIR) band. The NIR band was used with more emphasis, one might say, due to the scene land cover, which was mainly composed by vegetated areas and water bodies, and the NIR plays an essential role at describing them and their spectral curves.

The spectral and spatial features were computed and stored in a database for each training segment. The samples were separated into training and test sets. It is important to highlight that the frequency of the classes is not regular in order to make more samples available for some classes. Table 1 displays the list of spectral variables that include the mean value in each spectral band, some spectral indexes and the maximal standard deviation of the region in each band. Table 2 displays the spatial variables computed for each segment.

Table 1: List of features used

Feature	Description
NDWI	Normalized difference water index – definition and equation in Gao (1996)
NDVIre	Normalized difference vegetation index – definition and equation in Tucker (1979), computed with the red edge and NIR bands
NDVIr	Normalized difference vegetation index, computed with the red and NIR bands
Max. diff.	Maximum difference between bands
Max. std. dev.	Maximum standard deviation of the region in each band
Mean (RE)	Mean of the red edge
Mean(R)	Mean of the red band
Mean (NIR)	Mean of the near infrared
Mean(G)	Mean of the green band
Mean(B)	Mean of the blue band
Brightness	Sum of all bands

Figure 2a, 2b, 2c, 2d e 2e displays the spectral response of the classes of interest, regarding the normalized indexes as well as the spectral band used as features and using the training sample data.
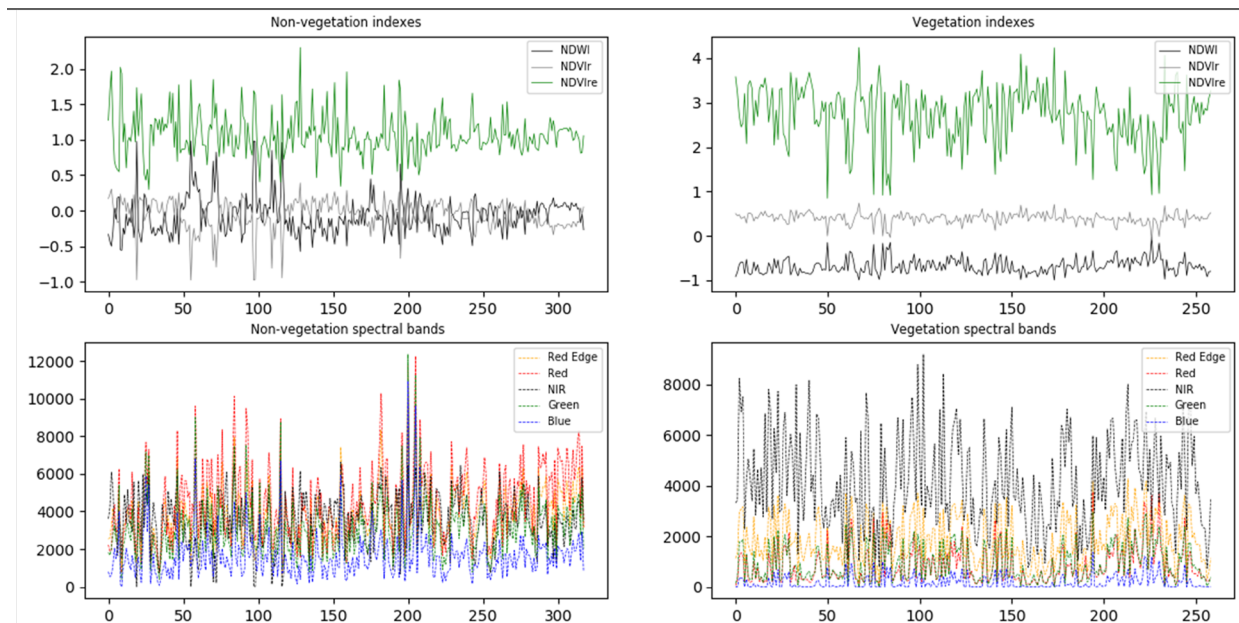
Level 1 – Water and non-water classes
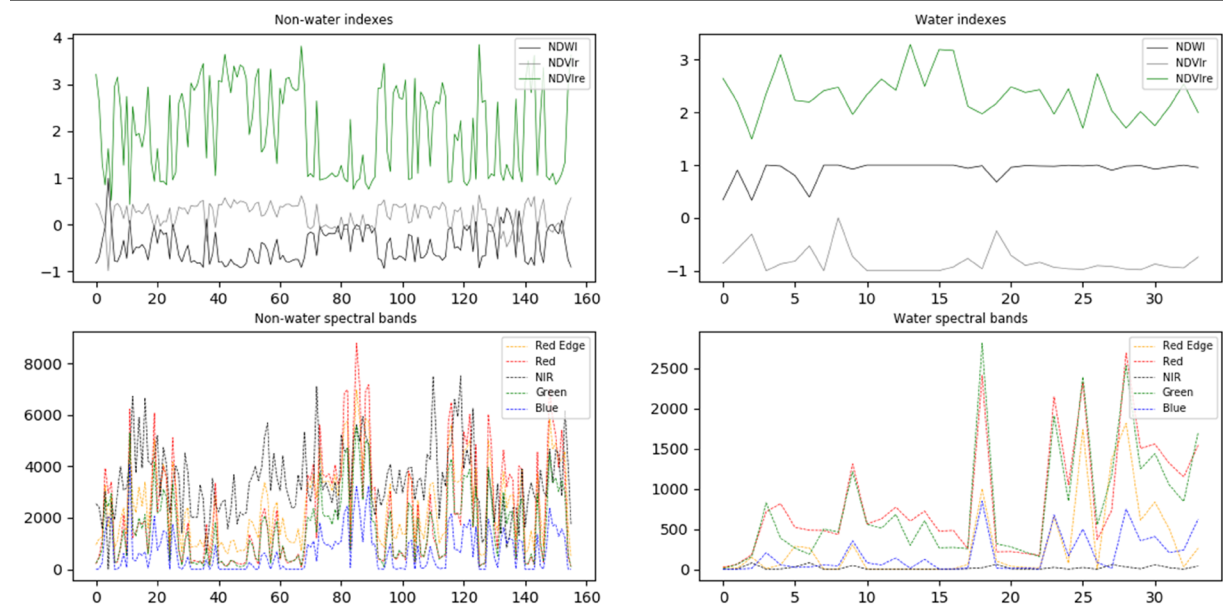
Source: The author

Figure 2a: Indexes and spectral responses for each two classes included in each node – water and non-water classes.

Level 2 – Non-vegetation and vegetation classes

Source: The author

Figure 2b: Indexes and spectral responses for each two classes included in each node – non-vegetation and vegetation classes.

Level 3.1 – High and low vegetation classes

Source: The author

Figure 2c: Indexes and spectral responses for each two classes included in each node – high and low vegetation classes.

Level 3.2 – Bare land and urban classes

Source: The author

Figure 2d: Indexes and spectral responses for each two classes included in each node – bare land and urban classes.
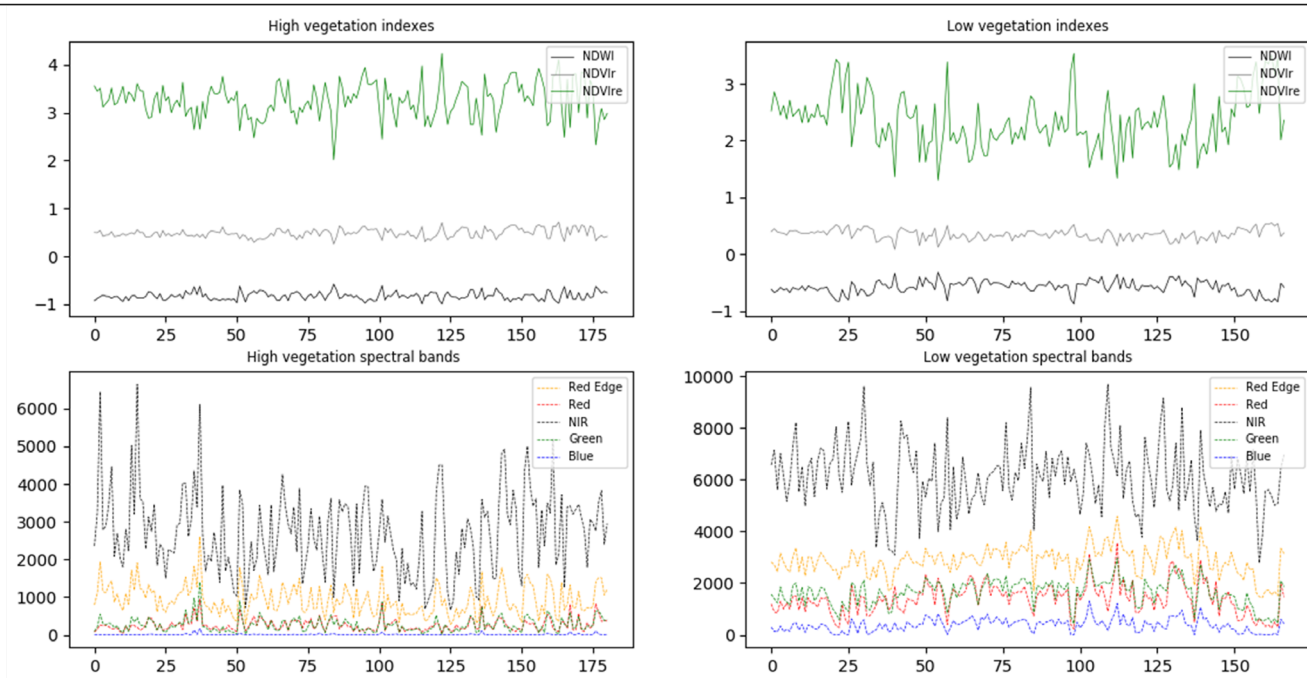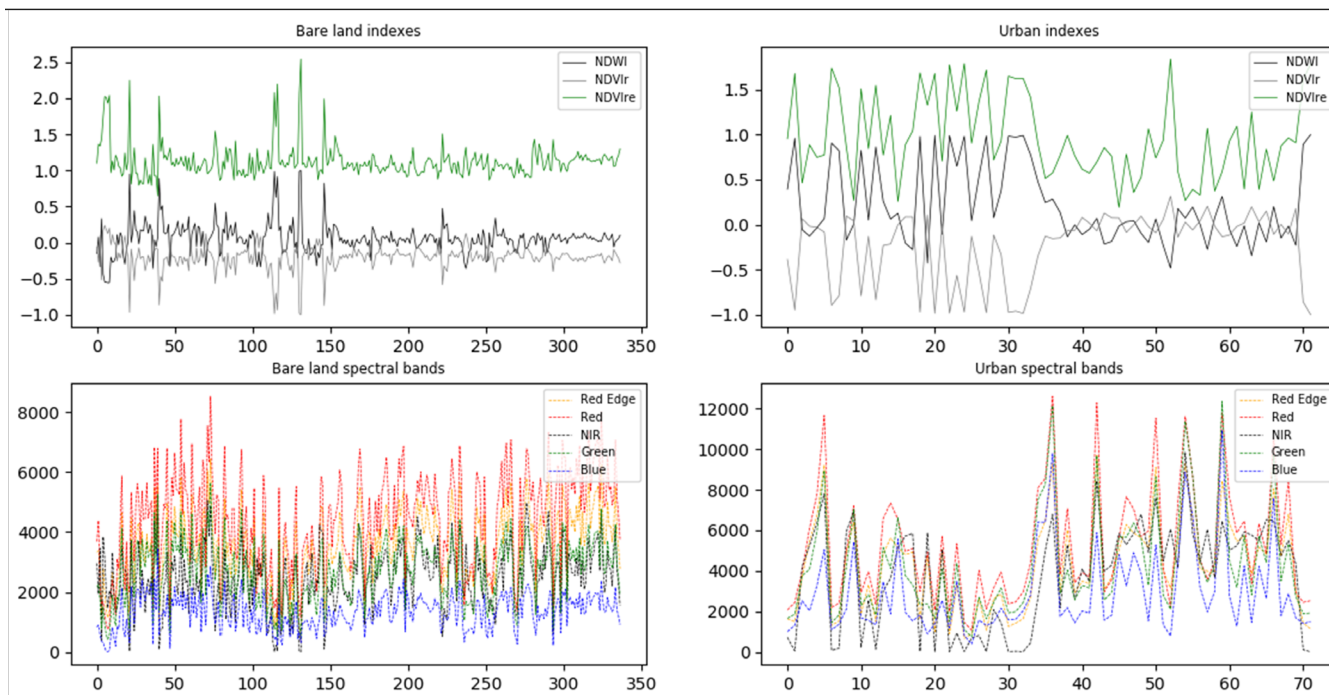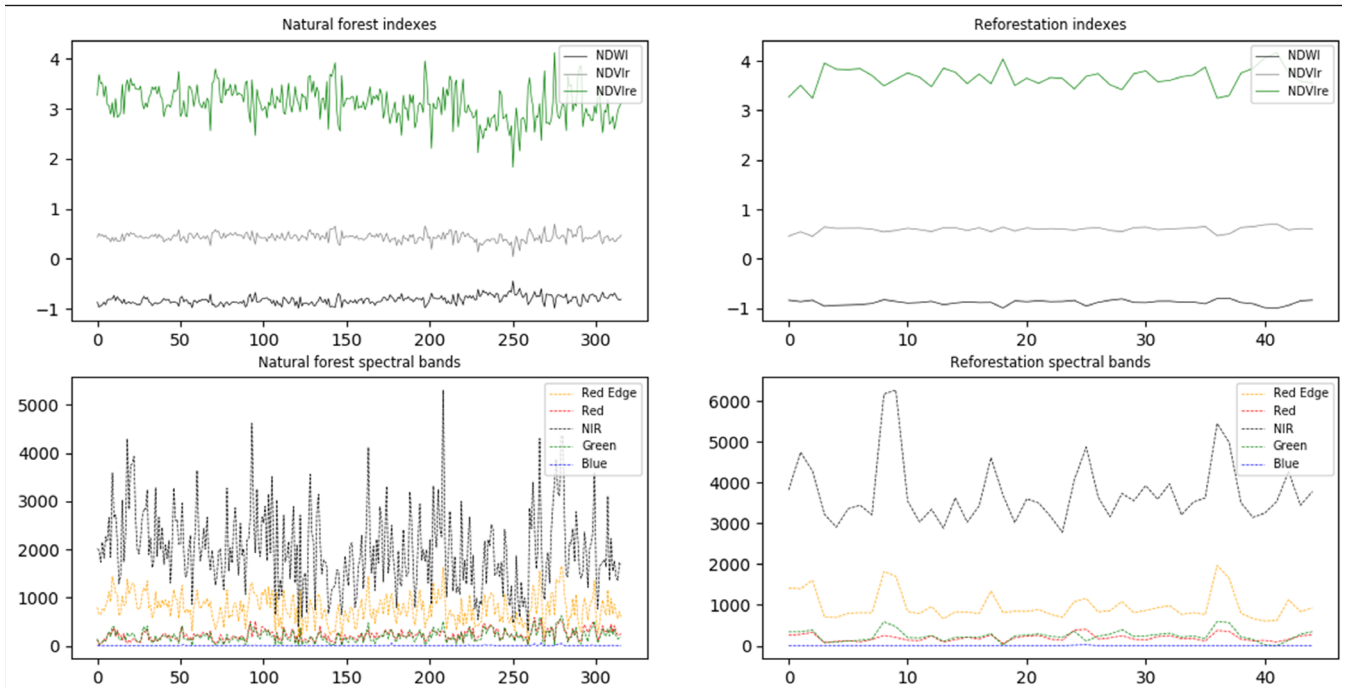
Level 4 – Natural forest and reforestation

Source: The author

Figure 2e: Indexes and spectral responses for each two classes included in each node – natural forest and reforestation.

Table 2 contains examples of the additional spatial features used during the feature selection step along with the spectral features. They were computed using the eCognition software. More detail on how they are calculated, their mathematical formulas and descriptions can be found in Gonzalez and Woods (2002).

Table 2: List of additional spatial features used

Features (part I)	Features (part II)
Border length	Compactness (polygon)
Number of edges (polygon)	Average branch length
Width	Std. dev. of length of edges (polygon)
Asymmetry	Volume
Length/Width (only main line)	Radius of the largest enclosed ellipse
Rel. Border to Image Border	Perimeter (polygon)
Avrg. area represented by segments	Length/Thickness
Elliptic Fit	Std. dev. of the area represented by the segments
Std. dev. Curvature (only main line)	Main direction
Length of longest edge (polygon)	Length of main line (no cycles)
Density	Shape index
Average length of edges (polygon)	Degree of skeleton branching
Number of pixels	Thickness
Polygon self-intersection (polygon)	Number of segments
Radius of smallest enclosing ellipse	Maximum branch length
Area (excluding inner polygons)	Compactness

Continue...

Table 2: Continuation.

Features (part I)	Features (part II)
Rectangular Fit	Curvature/length (only main line)
Length of main line (regarding cycles)	Roundness
Length	Area
Area (including inner polygons)	Border index
Number of inner objects (polygon)	Width (only mainline)
Length/Width	

The resulting decision tree for the classification is displayed in figure 3 and follows the hydrological relation between the groups of classes.
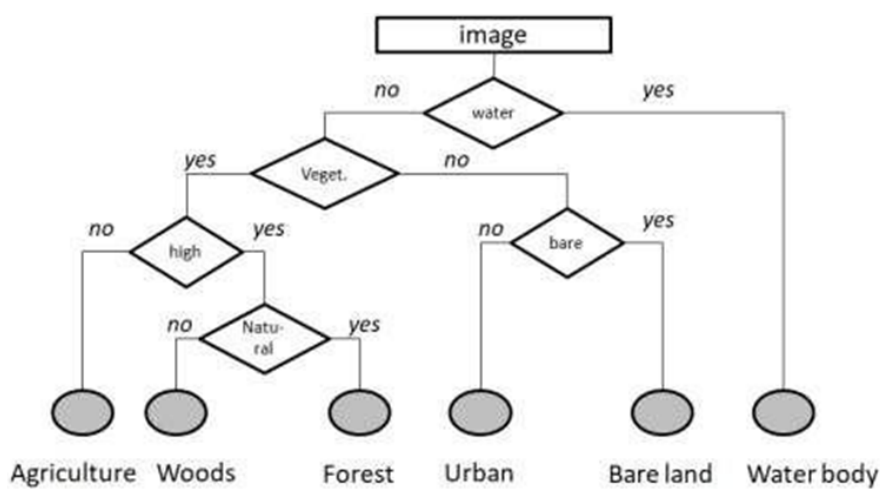
Source: The author

Figure 3: Proposed decision tree for the land cover classification.

Training areas were selected for each class using the image objects (segments) as samples, respecting the number of segments bigger than 30 as usually carried out in statistics (Shewhart, 1939). The training samples collected by the analyst were based on the objects that could best describe the classes of interest. The quantities of each class are presented below (table 3):

Table 3: Set of training samples

Classes	Training samples (in objects)	Training samples (in pixels)
Water	44	35918
Non-water	156	42226
Vegetation	259	65600
Non-vegetation	318	36390
Agriculture	167	26584
High vegetation	181	76519
Wood	112	17472
Forest	316	122304
Bare land	252	41378
Urban area	72	8484

The samples were collected in accordance with the logic of selecting well-distributed objects around the image scene based on image interpretation and analyst knowledge of the area. From the total area, less than 8% of the pixels were used to the set of training samples. The test samples under the same rule of well-distributed objects, but they were collected apart, respecting the classes' proportions in the scene after classification as well as a minimum number of 30 samples. The confusion matrices using the test samples are presented in tables 5, 6 e 7.

5. Results and Discussion

In the first experiment, the perceptron concept was applied in each node to select the best features to distinguish the classes; the set of features for each node is displayed in the second column of Table 3. The third column of Table 3 shows the selected variables for each node and the weights combining two features each time. In some cases, in this part of the experiment, the perceptron converged very quickly, demonstrating that most of the time only one feature would be enough to separate two classes. This event happened when one of the feature weights was too small compared to the other one. Other than that, the solution was a balanced combination of two variables.

Figure 4a displays an example in which only one variable is necessary. In this case, the spatial feature "border length" was used, from which one can readily understand that the decision border is a horizontal line. On the other hand, in some nodes the decision border is only achieved when using a linear combination of two variables, as shown in figure 3b. In those, it is clear that a horizontal line was not enough to distinguish the two classes, hence the need of two variables to separate low from high vegetation.
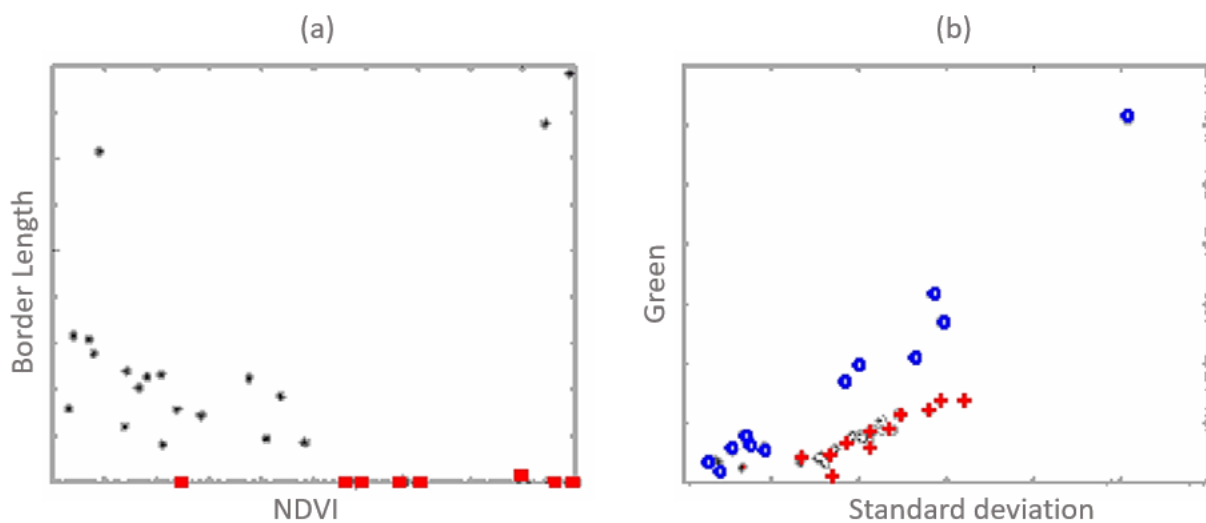
Source: The author

**Figure 4:** Examples of the computed solutions. (a) The classes can be separated using only one feature (class one represented by the black color and class 2, by the red color); (b) A linear combination of two features is necessary to separate classes 1 (blue) and 2 (red).

Finally, the variables selected with the SBS appear in the last column of Table 3. The solutions accuracies were found between 77 and 88% (table 8).

Table 4: Set of features and weight chosen for each node, according to the methods tested: one variable, pair of variables and SBS

Node	One variable	Pairwise selection Variables and weight	(SBS) Variables and weight
Water/no water	Border length	Brightness 54% max.Std.dev. 46%	Mean(R)
Vegetation/non-vegetation	Max. diff.	NDVIr	NDVIr 41% NDVIre 37% NDWI 22%
Low/high vegetation	Mean(G)	NDWI 52% Mean(G) 48%	NDVIre 59% NDVIr 25% NDWI 16%
Wood/forest	NDVIre	NDVIre	NDWI 40% Max. diff. 30% NDVIre 15% NDVIr 15%
Bare land/urban	Brightness	Mean(G) 66% Std. dev. 34%	max.Std.dev. 33% Mean(G) 28% Mean(R) 17%

Although it was expected that the water index NDWI would be the best option to separate water from other objects, the methods proposed other features. The SBS approach selected various optimal solutions and the red channel was chosen as the best option. As the expected feature NDWI was not accepted, a plot (Figure 5) using this feature from water and non-water samples was performed. As a result, an overlap of areas with water and the other classes was evident; this justifies the SBS algorithm choice.
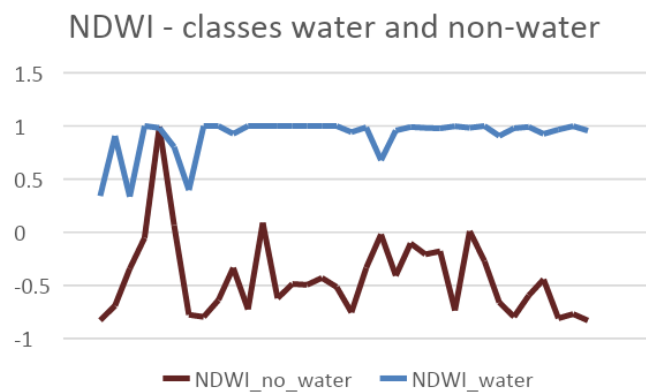
Source: The author

Figure 5: Chart using the NDWI feature for the classes: water and non-water.

Normalized difference vegetation indices are considered to be a reasonable solution for separating areas covered by vegetation. Still, it did not happen for most of the nodes with vegetation classes. In their case, when only one variable was used as input, the solution proposed by the pairwise method, as well as those recommended by the SBS, seemed more reasonable since they select a vegetation index. For example, the separation of vegetation/non-vegetation using the SBS proposes to combine the NDVIr with the NDVIre, which are very similar and reflect the contrast between near-infrared and red. The inclusion of the water index has a relatively lower weight.

A significant difference concerning the chosen features was found in node three, which was used to separate

low from high vegetation. For this node, the SBS solution was based on the normalized difference indexes, whereas the pairwise selection includes the green channel and the NDWI. As the NDWI formula undertaken was based on the difference between the Near Infrared and the Green bands, the healthy and dense vegetation would present higher values, which is reasonable, since this index can also be called GNDVI (Green NDVI) (Gitelson at al., 1996). The single input perceptron selected the green channel, but with lower performance.

When considering woods and forests, the SBS proposes the use of the normalized difference indexes and the NDWI, composed by the NIR band, mainly because of the leaf water content. The other methods chose the NDVIre calculated with the Red Edge band instead of the Red one, commonly used to separate vegetation species based on the leaf structure (Sims and Gamon, 2002; Centeno, 2009). This is a quite appropriate choice in terms of spectral properties of the leaf.

In the last node, urban areas were separated from regions where the soil had no cover (bare land). The SBS selected the maximum difference between the bands, the green and the blue channels, while the pairwise selection decided by the difference between bands and the standard deviation. The choice of bands in the visible spectrum is suitable since urban areas are better identified in these regions of the spectrum. The single input approach did not achieve a satisfactory solution in the last node (bare land and urban classes) and the brightness was selected with very low classification accuracy as a result.

Once the variables for each node were selected, the classification applying the decision tree with the three computed decision rules for each node was performed. Methods' accuracies were evaluated by using the decision tree for other samples of the same area, the test samples. The number of samples was not equal for all the classes because the classes were not homogeneously distributed in the image. Tables 5, 6 and 7 show the confusion matrix of each experiment.

Table 5: Confusion matrix for the single input experiment

	Agriculture	Wood	Forest	Urban	Bare land	Water	Producer's accuracy %
Agriculture	27	0	2	0	14	0	63.3
Woods	0	42	3	0	0	0	93.3
Forest	3	4	32	0	0	0	81.0
Urban	0	0	0	14	28	0	33.6
Bare land	1	0	0	1	57	0	96.6
Water	0	0	0	1	3	39	90.7
User's accuracy %	86.5	90.6	87.7	86.8	56.3	100.0	

Table 6: Confusion matrix for the pairwise selection experiment

	Agriculture	Wood	Forest	Urban	Bare land	Water	Producer's accuracy %
Agriculture	30	0	1	0	11	0	71.6
Woods	0	42	3	0	0	0	93.3
Forest	11	5	21	0	0	3	52.5
Urban	1	0	0	24	17	0	57.6
Bare land	4	0	0	2	54	0	90.7
Water	0	0	0	1	3	39	90.7
User's accuracy %	66.2	90.3	84.7	89.7	63.5	91.3	

Table 7: Confusion matrix for the SBS experiment

	Agriculture	Woods	Forest	Urban	Bare land	Water	Producer's accuracy %
Agriculture	41	1	0	1	0	0	97,0
Woods	0	42	3	0	0	0	93,3
Forest	3	4	33	0	0	0	84,2
Urban	2	0	0	0	0	4	0,0
Bare land	1	0	0	1	13	1	83,2
Water	0	0	0	1	3	39	90,7
User's accuracy %	89,2	91,1	91,1	0,0	81,3	90,1	

For the comparison of the three methods, table 8 shows their overall accuracy and Kappa index values for each classification method performed.

Table 8: Statistics of the experiments

Method	Overall accuracy	Kappa index
single input	77.86	0.73
pair wise selection	77.5	0.72
SBS	88.20	0.80

The single input method and the pairwise selection performed relatively equally, with an overall accuracy around 78% and a Kappa index around 0.73, which is not considered a suitable result, according to Anderson et al. (1976). On the other hand, the overall accuracy of the thematic image applying the rules selected by the SBS method was of 88.20%, a value that characterizes a good result according to the same authors. There was significant confusion between forest and woods, a fact that affects the producer's accuracy of these classes. The Kappa value computed from the confusion matrix was 0.806, which can also be considered adequate, as stated by Jensen (2005), for whom Kappa coefficients above 0.80 show strong consistency, Kappa values between 0.40 and 0.80 show moderate consistency and values below 0.40 are weak. Considering the previous statistic values, it is reasonable to state that the SBS approach successfully performed the classification.

The second method, the pairwise selection, had lower performance. It was possible to notice a considerable confusion between urban areas and bare land. As a consequence, there was a decrease in both user and producer's accuracy of "urban areas" and in the user's accuracy of "bare land". The pairwise selection overall accuracy reached only 77.50% (moderate agreement) and the Kappa index was 0.72, values below those obtained in the SBS experiment. Although the values were lower, it must be taken into account that, in the second experiment, the selection of a maximum number of two features while using the SBS was imposed; the limit was performance-wise. Given the considerations made in the earlier paragraphs, the pairwise selection results were not worthless.

Similar results were obtained using a single feature as input. The overall accuracy value was of 77.86% (moderate agreement) and the kappa index was 0.73. The results were similar to those acquired using two variables as input. Although the producer's accuracy values were lower, the user's accuracy values were better. The classes "urban" and "bare land" showed higher confusion. This result was already expected since the accuracy of selecting the features was already low.

6. Conclusion

In this work, three feature selection approaches based on the perceptron concept were compared. The first advantage regarding the use of perceptron for the methods here developed is that this technique does not require any prior statistical assumption on data distribution, such as the premise of normality within the classes, for instance. Not making any statistical assumptions is a crucial point mainly because, when using the OBIA approach for image classification purposes, pixels are grouped in the segmentation step, reducing the number of available samples.

The pairwise selection demands high computational effort for it consists of an exhaustive search within a bidimensional space. As the size of the feature space increases substantially, including the spatial features, the number of possible combinations also has a significant increase. The advantage of these methods is that they can find solutions to the classes' separation problem by combining more than one feature in a linear discriminant function. Moreover, the use of the perceptron with a single input proved to achieve results comparable to those using a pair of variables as input, but with lower effort.

The SBS algorithm produced the best results and enabled the identification of the most significant variables within a set of spatial, spectral and textural features. Whereas the single perceptron allows computing new variables through a linear combination of the original features that enable the classification in a decision tree.

The use of the perceptron to find the best variables for a decision tree proved to be efficient and can be an alternative for feature selection. Not only does it avoid a human visual comparison of features but also presents the advantage of proposing linear variable combinations as a solution. More sophisticated systems, such as a multilayer perceptron, could as well be used to estimate more complex discriminant functions.

AUTHOR´S CONTRIBUTION

All authors contribute equally.

REFERENCES

Aguilar, M., Vicente, R., Aguilar, F., Fernández, A. and Saldaña, M., 2012. Optimizing Object-Based Classification in Urban Environments Using Very High Resolution GEOEYE-1 Imagery. *ISPRS Annals of Photogrammetry, Remote Sensing and Spatial Information Sciences*, I-7, pp.99-104.

Anderson, J., Hardy, E., Roach, J. and Witmer, R., 1976. A Land Use and Land Cover Classification System for Use with Remote Sensor Data. 1st ed. Washington: U.S. Govt. Print. Off. [online] Available at: http://www.pbcgis.com/data_basics/anderson.pdf [Accessed 9 May. 2019].

Baatz, M. and Schäpe, A., 2000. Multiresolution Segmentation: An Optimization Approach for High Quality Multi-Scale Image Segmentation. In: *Angewandte Geographische Informationsverarbeitung XII: Beiträgezum AGIT-Symposium*. Heidelberg: Hebert Wichmann Verlag, pp.12-23.

Bartenhagen, C., Klein, H., Ruckert, C., Jiang, X. and Dugas, M., 2010. Comparative study of unsupervised dimension reduction techniques for the visualization of microarray gene expression data. *BMC Bioinformatics*, 11(1).

Centeno, J., A., S. 2009. Sensoriamento Remoto e Processos de Imagens Digitais. Editora UFPR

Gasca, E., Sánchez, J. and Alonso, R., 2006. Eliminating redundancy and irrelevance using a new MLP-based feature selection method. *Pattern Recognition*, 39(2), pp.313-315.

Gao, B., 1996. NDWI—A normalized difference water index for remote sensing of vegetation liquid water from space. *Remote Sensing of Environment*, 58(3), pp.257-266.

Geng, X., Sun, K. and Ji, L., 2014. Band selection for target detection in hyperspectral imagery using sparse CEM. *Remote Sensing Letters*, 5(12), pp.1022-1031.

Gitelson, A. A.; Kaufman, Y. J.; Merzlyak, M. N. 1996. Use of a green channel in remote sensing of global vegetation from EOS-MODIS. *Remote Sensing of Environment*, 58, 289-298.

Gonzalez, R. and Woods, R. 2002. Digital image processing. 2nd Edition, Prentice Hall, Upper Saddle River.

Guo, B.; Gunn, S. R.; Damper, R. I.; Nelson, J. D. B. 2006. Band Selection for Hyperspectral Image Classification Using Mutual Information. *IEEE Geosci. Remote Sens. Lett.* 2006(3): 522–526.

Habermann, M., Frémont, V., Shiguemori, E. 2018. Unsupervised Hyperspectral Band Selection Using Clustering and Single-layer Neural Network. *Revue Française de Photogrammétrie et de Télédétection, Société Française de Photogrammétrie et de Télédétection*, In press, pp.33-42

Haertel, V., Langrebe, D. A. 1999. On the classification of classes with nearly equal spectral response in remote sensing hyperspectral image data. *IEEE Transactions on Geoscience and Remote Sensing*, 37(5), 2374–2386.

Hughes, G. 1968. On the mean accuracy of statistical pattern recognizers. *IEEE Transactions on Information Theory*, 14(1), 55–63.

Hamediantar, A. and Shafri, H. Z. M. 2016. Integrated Approach Using Data Mining-Based Decision Tree and Object-Based Image Analysis for High Resolution Urban Mapping of WorldView-2 Satellite Sensor Data. *Jour. of Applied Rem. Sens.* 10(2), 025001-1 - 025001-21. DOI: 10.1117/1.JRS.10.025001.

Jensen, J., 2005. Introductory Digital Image Processing: A Remote Sensing Perspective. 3rd ed. Upper Saddle River, N.J.: Prentice Hall.

Jung, R. and Ehlers, M. 2016. Comparison of Two Feature Selection Methods for the Separability Analysis of Intertidal Sediments with Spectrometric Datasets in the German Wadden Sea. *Intern. Journal of Applied Earth Observ. and Geoinf.*, 52, 175-19. DOI: 10.1016/j.jag.2016.06.009.

Mahmoudi, F., Samadzadegan, F. and Reinartz, P., 2013. Object oriented image analysis based on multi-agent recognition system. *Computers & Geosciences*, 54, pp.219-230.

Persello, C. and Bruzzone, L., 2016. Kernel-Based Domain-Invariant Feature Selection in Hyperspectral Images for Transfer Learning. *IEEE Transactions on Geoscience and Remote Sensing*, 54(5), pp.2615-2626.

Pu, R. and Bell, S. 2017. Mapping Seagrass Coverage and Spatial Patterns with High Spatial Resolution IKONOS Imagery. *Intern. Journal of Applied Earth Observ. and Geoinf.*, 54, . 145-158. DOI: 10.1109/IGARSS.2016.7730998

Ruck, D. W., Rogers, S. K., Kabrisky, M. 1990. Feature selection using a multilayer perceptron. *Neural Network Computing* 2: 40–48

Safavian, S. R.; Landgrebe, D. 1991. A survey of decision tree classifier methodology. *IEEE Transactions on Systems, Man and Cybernetics*. 21. 660 - 674. 10.1109/21.97458.

Serpico, S. B., D'Inca, M., Melgani, F., Moser, G. 2003. Comparison of feature reduction techniques for classification of hyperspectral remote-sensing data, *Proc. SPIE 4885, Image and Signal Processing for Remote Sensing VIII*, (13 March 2003).

Shewhart, W. A. 1939. Statistical Method from the Viewpoint of Quality Control. 2nd Edition, Lancaster Press, Washington.

Sims, D. A.; Gamon, J. A. 2002. Relationships between leaf pigment content and spectral reflectance across a wide range of species, leaf structures and developmental stages. *Remote Sensing of Environment*, 81, 337-354.

Sohn, K.; Zhou, G.; Lee, C.; Lee, H. 2013. Learning and Selecting Features Jointly with Point-wise Gated Boltzmann

Machines. *Proceedings of the 30th International Conference on Machine Learning, Atlanta, Georgia*, USA, 2013. JMLR: W&CP vol. 28.

Tang, J.; Alelyani, S.; Liu, H. 2014. Feature selection for classification: a review. *Data Classification: Algorithms and Applications*, 37-64.

Tucker, C., 1979. Red and photographic infrared linear combinations for monitoring vegetation. Remote Sensing of Environment, 8(2), pp.127-150.

Van Coillie, F. M. B.; Verbeke, L. P. C., De Wulf, R. R. 2007. Feature Selection by Genetic Algorithms in Object-Based Classification of IKONOS Imagery for Forest Mapping in Flanders, Belgium. *Rem. Sens. of Envir.*, 110, 476-487. DOI: 10.1016/j.rse.2007.03.020.

Wang, X. Y., Guo, Y. G., He, J., Du, L. T. 2016. Fusion of HJ1B and ALOS PALSAR Data for Land Cover Classification Using Machine Learning Methods. *Intern. Journal of Applied Earth Observ. and Geoinf.*, 52, 193-203. DOI: 10.1016/j.jag.2016.06.014.

Weinberger, K. Q. and Saul, L. K. 2006. An introduction to nonlinear dimensionality reduction by maximum variance unfolding. *In: 21st National Conference on Artificial Intelligence. [online] AAAI Press*, pp.1683–1686. Available at: <https://www.aaai.org/Papers/AAAI/2006/AAAI06-280.pdf> [Accessed 1 September 2020].

Xie F, Li, F., Lei C., Ke, L. 2018. Representative Band Selection for Hyperspectral Image Classification*. International Journal of Geo-Information*, 7(338):1-19.

Xiurui, G., Vitasi, S., Luyan J., Yongchao, Z. 2014. A Fast Volume-Gradient-Based Band Selection Method for Hyperspectral Image. *Geoscience and Remote Sensing, IEEE Transactions on.* 52. 7111-7119. 10.1109/TGRS.2014.2307880.

Zhang S. and Chau KW. 2009. Dimension reduction using semi-supervised locally linear embedding for plant leaf classification. In: Huang DS., Jo KH., Lee HH., Kang HJ., Bevilacqua V. (eds) Emerging Intelligent Computing Technology and Applications. ICIC 2009. Lecture Notes in Computer Science, vol 5754. Springer, Berlin, Heidelberg.