PARCEL-BASED IMAGE CLASSIFICATION AS A DECISION-MAKING SUPPORTING TOOL FOR THE LAND BANK OF GALICIA (SPAIN)

T. Hermosilla [a]*, J.M. Díaz-Manso [b], L.A. Ruiz [a], J.A. Recio [a], A. Fernández-Sarría [a], P. Ferradáns-Nogueira [b]

[a] Departamento de Ingeniería Cartográfica, Geodesia y Fotogrametría. Universidad Politécnica de Valencia. Camino de Vera s/n, 46022 Valencia, Spain - txohergo@topo.upv.es; (laruiz; jrecio; afernan)@cgf.upv.es

[b] Sociedade para o Desenvolvemento Comarcal de Galicia. Estrada Santiago-Noia, km. 1. A Barcia s/n. Santiago de Compostela. Spain.- jose.marcial.diaz.manso@xunta.es

Commission IV, WG IV/3

KEY WORDS: Object-oriented classification; decision trees; high-resolution imagery; land cover; mapping; land bank.

ABSTRACT:

The results obtained after the application of parcel-oriented classification over two geographic areas of Galicia, in the northwest of Spain, are presented. In this region, forest and shrublands in mountain environments are very heterogeneous, with many private unproductive parcels, some of which are in a high state of abandonment. This situation entails a low economic productivity of the land and a higher vulnerability to wildfires and degradation in the affected areas. In this sense, the local government is promoting new methodologies based on high resolution images in order to classify the territory in basic and generic land uses. This land database will be used to plan specific actions for the sustainable management of degraded parcels, including the creation of a land bank.
The data used were 0.5 m/pixel visible and near infrared aerial imagery, and cadastral cartography employed to define the image objects (parcels) to be analysed. A set of features was computed from the images to quantitative describe different properties of the objects: spectral, texture and structural. In addition, several shape features were extracted from the parcel polygons. The classification was performed by means of decision trees, combined using the boosting technique. For the evaluation, field data samples were collected. An additional test using as a descriptive feature the land use class contained in a previous thematic database was performed. The overall accuracies of the classifications obtained are always above 90%. In one of the two areas tested, where forest and shrublands are especially undefined, the discrimination between these two classes is low. In conclusion, the use of automatic parcel-oriented classification techniques for land use updating tasks, particularly when broad and well defined classes are required, seems to be effective in some of the areas tested.

1. INTRODUCTION

Different scale land use/land cover geospatial databases are key information for territory management and economic monitoring. The accuracy and the reliability of these databases is crucial for territory management and decision-making. The high dynamism of some geographic areas and the need of periodical updating of the information contained in the geospatial databases require a high economic cost that makes difficult to update the information with the appropriate frequency. Image classification can contribute to automate, the processes of land use/land cover geospatial database updating, particularly those that allow the integration of the parcel limits derived from existing cartography for object definition. These methods could substantially reduce costs at production level. Some examples of the use of remote sensing techniques for updating land use land/cover geospatial databases are described in Walter (2000), Marçal et al. (2005), Catani et al. (2005), and Ruiz et al. (2009).

The periodic updating of the information contained in a land use/land cover geospatial database allows for an efficient territory management and avoids the appearance of neglected lands. Abandoned lands generate low economic productivity and a high vulnerability to wildfires and degradation in the affected areas.

This paper presents the results of a preliminary study of the suitability of the employment of parcel-based image classification for land use geospatial database updating in Galicia (Northwest of Spain). In this region, forest and shrublands in mountain environments are very heterogeneous, presenting many private unproductive parcels, some of which are in a high state of abandonment. The local government is promoting new methodologies based in high resolution images in order to classify the territory in basic and generic land uses, with the goal of creating a geospatial database. This land use database will allow for planning specific actions for a sustainable management of neglected parcels, including the creation of a *land bank* of Galicia[1]. Object-oriented classification applied to update agricultural and forest parcels can be focused on different thematic levels. In this case, generic classes would provide operative information for discriminating between productive and unproductive parcels.

The objective of this study is to define and evaluate a productive methodology based on parcel-oriented classification of high-resolution images, for updating a generic land cover database. This database could be eventually used to detect abandoned agricultural and forest parcels.

[1] http://www.bantegal.com/

2. STUDY ZONES AND DATA

2.1 Study zones

The study has been performed on two local administrative areas (*comarcas*) of Galicia: *Baixo Miño* and *A Limia* (see Figure 1). The first one is located in the Atlantic coast of the province of Pontevedra, and is mainly covered by forest, agricultural crops and vineyards. The administrative area of *A Limia* is located in the province of Ourense and presents large areas of agricultural crops, forest and shrublands.

a. Location of Galicia in Spain

b. Location of analyzed regions in Galicia.

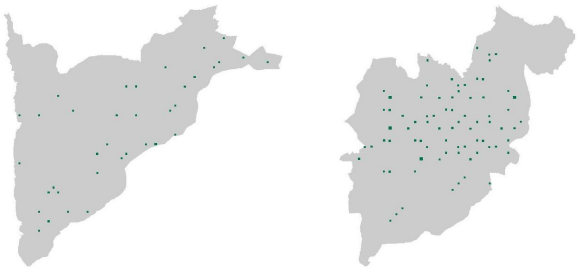
Figure 1. General location maps.

2.2 Data

The images employed were acquired from the *Spanish National Plan of Aerial Orthophotography* (PNOA). These images have a spatial resolution of 0.25 m/pixel and 4 spectral bands: red, green, blue and near infrared. The images of *A Limia* were acquired between May and July of 2007, and those of *Baixo Miño* in the same months of 2008.

Cartographic boundaries to define the final objects (plots) were obtained from the *Land Parcel Information System* (SIGPAC), a geospatial database oriented to agriculture management. The plots represent a continuous area of land within a parcel for a single agricultural use, being the total number of plots 468,721 in *A Limia* and 255,347 in *Baixo Miño*.

There have been available field samples collected in the same date that the images employed for each region. These samples have square shape with side sizes of 350 or 500 meters. Figure 2 shows the distribution of field samples on both administrative areas.



a. *Baixo Miño*

b. *A Limia*

Figure 2. Distribution of field data in the two administrative areas.

3. METHODOLOGY

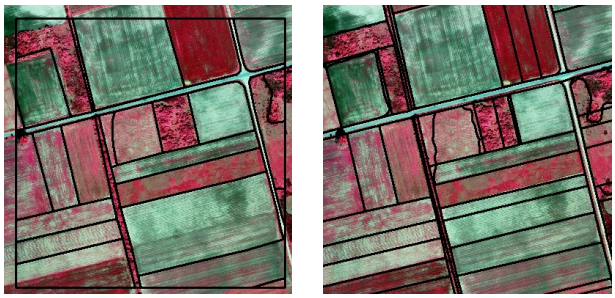
In this section, a general description of the steps followed in the methodological approach is done, with references to documents containing a more exhaustive explanation. Main classification steps include: Image and data pre-processing, selection of training samples, descriptive feature extraction, classification method, post-processing and evaluation.

3.1 Pre-processing

The images used were already orthorectified and georreferenced, panchromatic and multispectral bands fused, mosaicking and radiometric adjustments applied, as a part of the PNOA project (Arozarena et al., 2008). Additionally, in order to facilitate the descriptive feature extraction process, images were resampled to 0.5 m/pixel, using bilinear interpolation. This spatial resolution was considered as optimum for information extraction in our particular conditions and classes.

A number of common classes for both regions were defined: *Buildings*, *Forest*, *Shrublands* and Arable *and crop land*. Due to the definition accuracy presented by the classes regarding to roads and rivers in the SIGPAC, these classes were transferred directly from the geospatial database. *Arable and crop lands* class includes also pastures, being built up by aggregation of three sub-classes, differencing the vegetation level of a plot: no vegetation, medium vegetation and cultivated field. Besides, some additional classes were defined in order to adapt the legend to the reality of each region. Thus, a *Water* layer class was defined in *A Limia* to classify new flooded areas not registered in the SIGPAC database. In *Baixo Miño*, the additional *Vineyards* and *Greenhouse* classes were defined.

In the region of *A Limia*, most of the training samples were selected from the field samples register available from the SIGPAC project, being adequate in number and spatial distribution (Figure 2b). Since the sampling polygons did not coincide with the SIGPAC plots limits (Figure 3), the assignation of samples to each class was manually done. Additional samples were added by photointerpretation in order to avoid the a low representation of some classes, particularly *Water layers*, *Forest* and *Shrublands*. As in the province of *Baixo Miño*, the number of field samples was substantially lower, the training samples were mainly selected using photointerpretation techniques, and using the field registers as ancillary data.



a. Field sampling segment and polygons

b. SIGPAC plots

Figure 3. Example of the geometric differences between field sampling polygons and SIGPAC plots.

Spatial objects were created using the SIGPAC plot limits. In order to confer coherence to the automatic feature extraction process, which requires a minimum parcel surface, and to preserve the classification accuracy, parcels with a surface lower than 60 m^2 were rejected. Besides, SIGPAC parcels with very large dimensions were excluded (representing images with more than 9,000,000), due to the RAM memory limitations for processing the per parcel feature extraction algorithms.

3.2 Feature extraction

The use of efficient features is essential for accurate classification. At this point, every parcel was independently processed to extract descriptive features that characterize the current land use. The features employed in this study can be grouped in four categories: spectral, textural, structural and shape. Besides, the effect of the use of ancillary data from the previous use has been tested.

Spectral features used provide information about the spectral response of objects on the visible and near infrared regions of the spectrum, which depends on land coverage types, state of vegetation, soil composition, construction materials, etc. These features are particularly useful in the characterization of spectrally homogeneous objects, such as herbaceous crops or fallow fields. Mean and standard deviation were computed from the bands NIR, R, G and also from the Normalized Difference Vegetation Index (NDVI).

Texture features inform about the spatial distribution of the intensity values in the image, being useful to quantify properties such as heterogeneity, contrast or uniformity related to each object (Ruiz et al., 2004). These properties are obviously related to the land use/land cover inside an object. For every object, the features proposed by Haralick et al. (1973) based on the grey level co-occurence matrix (GLCM) were computed. This information was completed with the values of kurtosis and skewness of the histogram, and the mean and the standard deviation of the edgeness factor for each parcel (Laws, 1985). The edgeness factor represents the density of edges present in a neighbourhood. These features were derived from the red band.

The semivariogram curve quantifies the spatial associations of the values of a variable, and measures the degree of spatial correlation between different pixels in an image. This is a particularly suitable tool in the characterization of regular patterns. For continuous variables the expression that describes the experimental semivariogram is:

$$(h) = \frac{1}{2N} \sum_{i=1}^{N} [z(x_i) - z(x_i + h)]^2$$

where $z(x_i)$ = value of the variable in position x_i.
N = number of pairs of data considered.
h = separation between elements in a given direction.

The experimental semivariogram representing each object is obtained by computing the mean of the semivariograms calculated in six directions, ranging from 0º to 150º with a step of 30º. Afterwards, each semivariogram curve is filtered using a Gaussian filter with a stencil of 3 positions, in order to smooth its shape and to eliminate experimental fluctuations. Several structural descriptive features were computed considering the singular points of the semivariogram, such as the first maximum, the first minimum, the second maximum, etc., being described in detail in Balaguer et al. (2010).

Shape features inform about the complexity in the shape of the objects. They can contribute to differentiate polygons with specific shapes. Several standard features were extracted for each object: compactness, shape index, fractal dimension, area and perimeter.

Finally, the previous land use, contained in the SIGPAC geospatial database, was included as a qualitative descriptive feature to evaluate its performance.

Due to the high number of features extracted from each object, some of them presented a high correlation, being redundant the information provided. The inclusion of these variables in the study could act as noise in the creation of the classification rules. The relations and redundancies existing between features was initially analysed by principal component analysis. Then, linear descriptive discriminant analysis was applied in order to determine the significance of the features, removing from the study those with low significance level.

3.3 Classification through decision trees

Objects were classified by using decision trees. A decision tree is a set of organized conditions in a hierarchical structure, in such a way that the class assigned to an object can be determined following the conditions that are fulfilled from the tree roots (the initial data set) to any of its leaves (the assigned class). The algorithm employed in this study is the C5.0, which is the latest version of the algorithms ID3 and C4.5 developed by Quinlan (1993). This algorithm is the most widely used to deduce decision trees for classifying images (Zhang and Liu, 2005).

The process of building a decision tree begins by dividing the collection of training samples using mutually exclusive conditions. Each of these sample subgroups is iteratively divided until the newly generated subgroups are homogeneous, that is, all the elements in a subgroup belong to the same class. These algorithms are based on searching partitions to obtain purer data subgroups, which are less mixed than the previous group where these come from. For each possible division of the initial data group, the impurity degree of the new subgroups is computed, and the condition which gives the lower impurity degree is chosen. This is iterated until the division of the original data into homogeneous subgroups is carried out by using the gain ratio as splitting criterion. This criterion employs information theory to estimate the size of the sub-trees for each possible attribute and selects the attribute with the largest expected information gain, that is, the attribute that will result in the smallest expected size of the sub-trees.

Objects were classified using 10 decision trees, by means of the boosting multi-classifier method, which allows for increasing the accuracy of the classifier. The methodology followed by the boosting to build the multi-classifier is based on the assignment of weights to training samples. The higher the weight of a sample, the higher its influence in the classifier. After each tree construction, the vector of weights is adjusted to show the model performance. In this way, samples which are erroneously classified increase their weights, whereas the weights of correctly classified samples decrease. Thus, the model obtained in the next iteration will give more relevance to the samples

erroneously classified in the previous step (Hernandez-Orallo et al., 2004). After the construction of the decision tree set, the class to each object is assigned considering the estimated error made in the construction of each tree. The lower the estimated error *e*, the higher the weight given to a tree, according to the formula:

$$-\log\left(\frac{e}{1-e}\right)$$

The sum of the weights of those trees which assign the same class to one object is computed, giving to that object the class with the highest value.

3.4 Evaluation

The evaluation of the classification was done using cross-validation. From confusion matrix, the user's and producer's accuracies per class were computed, that respectively measure the commission and omission errors.

In a typical process of geospatial database updating, the class assigned after the classification is compared to the land use contained in the original database. The differences between them register the potential land use/land cover changes produced in the territory, but also the errors produced in the classification.

In the updating process, correctly classified cases can be divided in two categories: coincidences and detected changes. Coincidences are these cases with equal land use assigned in the classification, reference data and database. A detected change occurs when the classification land use is correctly assigned meanwhile the land use appearing in the database is wrong. The sum of the percentage of coincidences and detected changes is equal to the overall accuracy of the classification. Updating errors can be divided in two respective categories: detectable and undetectable errors. A detectable error is produced when a mistaken land use is assigned in the classification, being the land use contained in the database correct. This error is also given if the classification process assigns an erroneous land use, the land use contained in the database is also incorrect, and both are different. An undetectable error happens when the land use assigned in the classification process and that contained in the database are the same but incorrect. The accumulation of detected changes and detectable errors compose the number of parcels to review in the updating process.

4. RESULTS AND DISCUSSION

4.1 Analysis of the results

4.1.1 ***A Limia***: Table 1 shows the confusion matrix of the classification performed using spectral, texture, structural and shape features. The highest confusion is made between the classes *Arable and crop lands* with *Shrublands*, and *Shrublands* with *Forest*. This confusion is produced due to the high similarities between *Shrublands* and *Forest*.

When the previous land use contained in the SIGPAC geospatial database is added as descriptive feature in the classification, the producer's and user's accuracies of the three classes with higher confusion degree increase. The higher number of errors is still produced between *Arable and crop lands* and *Shrublands*, and *Shrublands* and *Forest*. Adding the previous land use as a descriptive feature increases the overall accuracy, as proved by Recio et al. (2009). In this case, the increase produced is approximately of 3%.

		Reference					
		Water	Buildings	Forest	Shrublands	Arable lands	User's accuracy
Classification	Water	17				2	89.5
	Buildings	1	175	3	5	4	93.1
	Forest		2	143	18		87.7
	Shrublands		3	14	251	28	84.8
	Arable lands	1	3	2	23	574	95.2
Producer's accuracy		89.5	95.6	88.3	84.5	94.4	**91.4**

Table 1. Confusion matrix of the classification performed using spectral, texture, structural and shape features in *A Limia*.

		Reference					
		Water	Buildings	Forest	Shrublands	Arable lands	User's accuracy
Classification	Water	16				1	94.1
	Buildings	2	177		1	5	95.7
	Forest		2	145	5		95.4
	Shrublands		1	15	279	22	88.0
	Arable lands	1	3	2	12	580	97.0
Producer's accuracy		84.2	96.7	89.5	93.9	95.4	**94.3**

Table 2. Confusion matrix of the classification performed using spectral, texture, structural, shape and **previous land use** features in *A Limia*.

In order to analyse the effect of the previous land use in the updating process, the classification results are compared with the information contained in the database. Even when the land uses defined in the database are different to the employed in the classification, they were grouped to produce a direct correspondence between land uses, in order to detect the changes produced. Figure 4 shows the distribution of errors in the different cases defined. Assuming a thematic updating of land use contained in the SIGPAC, the number of changes detected without employing the previous land use is 5%, meanwhile employing this information, detected changes are reduced to 3.3%. Detectable errors without using ancillary data represent a 6.2% and a 1.4% using this information. This means that when previous land use is employed the number of parcels to be revised (possible detected changes) is reduced to a 4.7%, against 11.2% without using it. However, the proportion of detected changes is notably reduced.

The undetectable errors in a later classification are practically doubled (from 2.4% to 4.3%) when the ancillary information is introduced in the classification. This means that even when the

use of the previous land use as descriptive feature improves the classification accuracy, in an updating process the undetectable errors would be increased, which are the proportion of parcels where the land use has changed but has been misclassified with the previous land use.

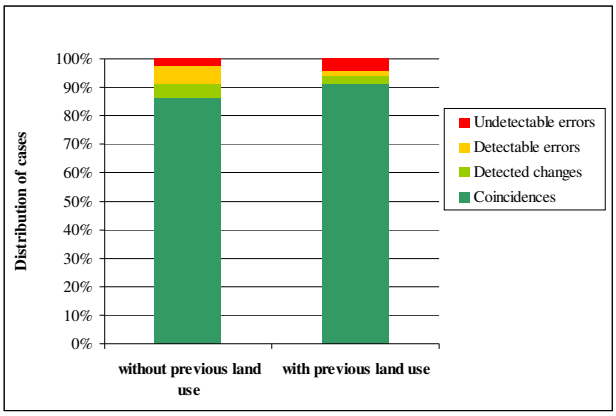


Figure 4. Distribution of coincidences, changes and errors including or not the previous land use contained in the SIGPAC geospatial database as descriptive feature (*A Limia*).

4.1.2. ***Baixo Miño*:** The confusion matrix of the classification using spectral, texture, structural and shape features (Table 3) shows high confusion between *Greenhouse* with *Buildings* and *Arable and crop lands* classes. In addition, some errors are produced between *Shrublands* with *Forest*, *Vineyards* and *Arable and crop land* due to the heterogeneity presented by the *Shrublands* class. The overall accuracy of the classification reaches 89.2%. Introducing the previous land use contained in the database as descriptive feature (see Table 4), the overall accuracy increases up to 92.9%, but the errors are mostly produced between the same classes.

		Reference						User's accuracy
		Buildings	Forest	Greenhouse	Shrublands	Arable lands	Vineyards	
Classification	Buildings	241		19	2	7	1	89.3
	Forest		387	2	21		6	93.0
	Greenhouse	10	1	64	6	4		75.3
	Shrublands	2	5	3	187	10	17	83.5
	Arable lands	6	1	16	30	498	4	89.7
	Vineyards		2	3	15	2	231	91.3
Producer's accuracy		93.1	97.7	59.8	71.6	95.6	89.2	**89.2**

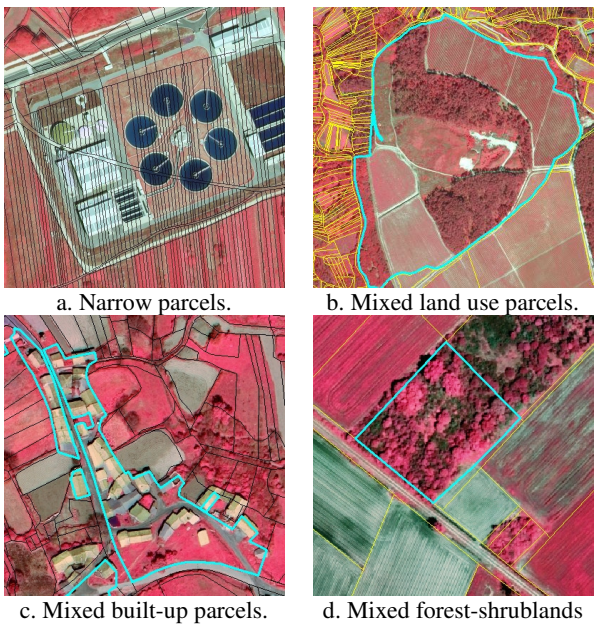
Table 3. Confusion matrix of the classification performed using spectral, texture, structural and shape features in *Baixo Miño*.

		Reference						User's accuracy
		Buildings	Forest	Greenhouse	Shrublands	Arable lands	Vineyards	
Classification	Buildings	252		13	4	2		93
	Forest	1	387		13			96.5
	Greenhouse	2	1	73	5	5		84.9
	Shrublands	4	8	4	212	9	7	86.9
	Arable lands			15	18	500	1	93.6
	Vineyards			2	9	5	251	94
Producer's accuracy		97.3	97.7	68.2	81.2	96.0	96.9	**92.9**

Table 4. Confusion matrix of the classification performed using spectral, texture, structural, shape and **previous land use** features in *Baixo Miño*.

4.2 Discussion and problems

One of the most important group of problems that introduce errors in the classification process in both areas of study is related with the geometry and shape of the plots. Some polygons, particularly in *Baixo Miño*, presented extremely long and narrow shapes (Figure 5,a). In other cases, the very small area of the plots makes the feature extraction process more difficult. On the other hand, the parcels with large dimensions normally present mixed land uses (Figure 5.b). This problem can be solved using automatic segmentation algorithms and classifying the generated sub-objects. Similarly, since the analyzed regions are basically rural areas, some built-up zones are contained in parcels mixed with vegetation (Figure 5.c). In this sense, the introduction of a post-processing step to filter or control improbable changes could reduce this type of errors and improve the classification accuracy for the classes involved.



a. Narrow parcels. b. Mixed land use parcels.

c. Mixed built-up parcels. d. Mixed forest-shrublands

Figure 5. Examples of cases that difficult a correct classification.

In the training sample selection process, partially done through photointerpretation, the visual discrimination between forest and shrubland was sometimes difficult. The SIGPAC database plots present mixed classes very frequently (Figure 5.d). In addition, the class *Shrublands* presents a significant internal heterogeneity in the area of *Baixo Miño*. Finally, since the regions of study were significantly large, those land cover classes having a very low representation were not considered in the classification legend.

5. CONCLUSIONS

A methodology for land use/land cover geospatial database creation and updating, based on generic classes and a parcel-based classification approach from high resolution multispectral images has been presented and analysed. The classification is based on the combination of several descriptive features derived from images, parcel shape and ancillary data. The analysis has been focused on detecting neglected agricultural and forest parcels, since this information is required by the Land Bank of Galicia to eventually intermediate between owners and potential users for a productive reutilization of the land.

The results obtained in this study show a high capability of the proposed techniques as a supporting tool for updating and managing this land cover / land use information. Main classification errors are produced in the discrimination between forest and shrublands areas, because of the complexity of the landscape. Several problems were found differentiating these classes using photointerpretation techniques. The effect of the use, as a descriptive feature, of the previous land use contained in the geospatial database has also been tested. The results show that even when the addition of this feature improves the classification accuracy reducing the errors, it produces a significant increase of the undetectable errors, difficulting the process of geospatial database updating.

ACKNOWLEDGEMENTS

The authors appreciate the support provided by the *Xunta de Galicia* and *Sociedade para o Desenvolvemento Comarcal de Galícia*.

REFERENCES

Arozarena, A., García, L., Villa, G., 2008. *Plan Nacional de Observación del Territorio en España.* Congreso Internacional de Ingeniería Geomática y Topográfica, TOP-CART, 18-21 February, Valencia, Spain.

Balaguer, A., Ruiz, L.A., Hermosilla, T., Recio, J.A., 2010. *Definition of a comprehensive set of texture semivariogram features and their evaluation for object-oriented image classification.* Computers & Geosciences, 36(2), pp. 231-240.

Catani F., Ermini, L., Kukavicic, M.., Moretti, S., Righini, G., 2005. *Detecting land cover changes through remote sensing and GIS techniques.* 31st International Symposium on Remote Sensing of Environment, 20-24 June 2005, St. Petersburg, Russia.

Haralick, R.M., Shanmugam, K., Dinstein, I., 1973. *Texture features for image classification.* IEEE Transactions on Systems, Man and Cybernetics 3(6), pp. 610-622.

Hernández Orallo, J., Ramírez Quintana, M.J., Ferri Ramírez, C., 2004. *Introducción a la minería de datos.* Pearson Educación S.A., Madrid.

Laws, K.I., 1985. *Goal-directed texture image segmentation.* Applications of Artificial Intelligence II, SPIE 548, pp.19-26.

Marçal, A.R.S., Borges, J.S., Gomes, J.A., Pinto Da Costa, J.F., 2005. *Land cover update by supervised classification of segmented ASTER images.* International Journal of Remote Sensing, 26(7), pp. 1347-1362.

Quinlan, J.R., 1993. *C4.5: Programs for machine learning*. Morgan Kaufmann Publishing, San Francisco.

Recio, J.A, Hermosilla, T., Ruiz, L.A., Fernández-Sarria, A., 2009. *Analysis of the addition of qualitative ancillary data on parcel-based classification.* International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, 38(1-4-7/W5) 6p.

Ruiz, L.A, Fernández-Sarria, A., Recio, J.A., 2004. *Texture feature extraction for classification of remote sensing data using wavelet decomposition: A comparative study.* International Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences. 35(B4), pp. 1109-1115.

Ruiz, L.A., Recio, J.A, Hermosilla, T., Fernández-Sarria, A., 2009. *Identification of Agricultural and Land Cover Database Changes Using Object-oriented Classification Techniques.* 33rd International Symposium on Remote Sensing of Environment, May 4 - 8, Stresa, Italy.

Walter, V., 2000. *Automatic change detection in GIS databases based on classification of multispectral data.* International Archives of Photogrammetry and Remote Sensing, 34(B4), pp. 1138-1145.

Zhang, S., Liu, X., 2005. *Realization of Data Mining Model for Expert Classification Using Multi-Scale Spatial Data.* International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, 26(4/W6), pp. 107-111.