ADDITION OF GEOGRAPHIC ANCILLARY DATA FOR UPDATING GEO-SPATIAL DATABASES

J.A. Recio, T. Hermosilla, L.A. Ruiz, A. Fernández-Sarría

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

Commission IV, WG IV/3

KEY WORDS: Database updating; object-oriented; classification trees; high-resolution imagery.

ABSTRACT:

The geographic characteristics of the territory determine the spatial distribution of land uses and are considered as essential clues by photo-interpreters to determine the land uses. Parcel-based classification of high-resolution images is one of the most reliable alternatives for the automatic updating of land use geospatial databases. Each parcel can be characterized by means of a set of features extracted from the image, its outline, the contextual relationships with its neighbours, etc. Features derived from geographic ancillary data can be considered as descriptive information in order to characterize the objects contained in the database. Several tests have been done in order to evaluate the usefulness of different types of geographic ancillary data to improve the land use/land cover classification. The ancillary data employed are: distance maps to key geographical elements, soil maps and features extracted from digital elevation models. In this study, each database object is described with its spectral feature set extracted from the image, using a per-parcel approach, completing this information with the geographic properties. Afterwards, objects are classified using decision trees combined with boosting techniques. The assigned class is compared with the land use in the database in order to detect changes or errors in any of the compared sources. The classification results demonstrate that a significant increase in overall accuracy can be achieved by combining spectral and textural features with geographic data.

1. INTRODUCTION

Land use-land cover geo-spatial databases are an essential source of information for natural resource management. The updating of this type of databases is expensive and time consuming and requires a high degree of human intervention. Currently, recent advances in quality and quantity of airborne and satellite sensors have entailed an important increase in the availability of high resolution images. At the same time, new methodologies are being developed to analyze these data.

In an object-oriented image analysis the minimal analysis unit is not a pixel but a group of pixels. Image objects can be created grouping pixels by means of automatic segmentation algorithms, or by using available cartographic information, such as cadastral or agricultural cartography. In this approach, the limits of the objects have more geographical meaning than regions created with segmentation algorithms, which produce a space division conditioned by sensor attributes instead of the territorial characteristics. Quantitative description of each object is carried out by means of a set of features which cover different aspects: spectral response, texture, planting pattern, shape of the parcel, etc.

Geographic characteristics of territory, like altitude, slope, aspect, etc. determine the spatial distribution of land uses. Features extracted from geographic data can complete the description of the objects with useful information in addition to the features extracted from images. In a traditional photo-interpretation process, geo-spatial information is considered to assign a class to each object in the database. Therefore, this information must be considered in a detection change semi-automatic process to obtain similar results to the obtained manually.

Integration of ancillary data into the classification has usually been divided in three categories: before, after or during classification. Integration before classification can be done through stratification, where ancillary data are used for dividing zones which have to be analyzed in a different way (Strahler et al., 1978). Some authors have used ancillary data after classification in order to improve or correct the results of the classification. Land use, rainfall (Cohen and Shoshany, 2002) or topographic information (Raclot et al., 2005) has been added in order to better discriminate between classes with a similar spectral response.

Integration of ancillary data during classification can be done in different ways. Many authors (Heipke and Straub, 1999; Olsen et al., 2002; Walter, 2004) employ the land cover/land use contained in agricultural and cartographic geospatial databases to automatically provide training samples for the classifier. The easiest and most employed technique to include ancillary data during the classification process is to use it as an additional descriptive feature. This technique is determined by the data type (continuous or discrete), and also by the classifier employed, because discrete data is not tolerated by statistical or distance-based classifiers. Some authors included land use/land cover information contained in the geospatial database as a descriptive feature (Rogan et al., 2003; Recio, 2009). Ancillary data derived from digital terrain models, such height, slope or aspect, has been included in many studies (Hoffer, 1975; Hutchinson, 1982; Bruzzone et al., 1997; Treltz and Howarth, 2000; Lawrence and Wright, 2001) due to its simplicity of use

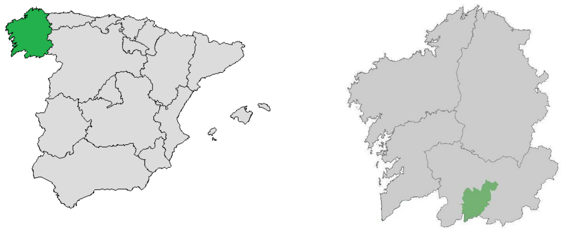
(Pedroni, 2001) and the known improvement in the classification accuracy (Hoffer et al., 1975).

The aim of this study is to analyze the influence of considering ancillary qualitative geographically referenced information in objects description and classification, for a particular case of study. Land uses often show a strong correlation with geographic characteristics and, therefore, geographic attributes provided by Digital Terrain Models (DTM) or geological map can be particularly useful when integrated with descriptive features derived from the image.

2. STUDY ZONE AND DATA

2.1 Study zone

The study area is the local region (*comarca*) of *A Limia* (see Figure 1) which is located in the province of Ourense, in Galicia (Spain). The study area presents large areas of agricultural crops, forest and shrublands.



a. Location of Galicia in Spain

b. Location of analyzed region 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 (R), green (G), blue (B) and near infrared (NIR). The images were acquired between May and July of 2007.

Cartographical boundaries have been obtained from the Geographical Identification System for Agricultural Parcels (SIGPAC) existing in Spain since 2005 for the management of agricultural aids. This is a registry of analogous properties to the cadastre.

Ancillary data were extracted from a Digital Terrain Model (DEM) with 10 meter resolution. This was employed to obtain the geographical features of the parcels and to derive the drainage network of the study area. In addition, the geological map of Galicia at a scale of 1:250,000 provided basic lithological information of the working area.

3. METHODOLOGY

Main steps of database updating by means of object oriented image classification are shown in figure 2. Firstly, objects are generated through information pre-processing and integration of different data. In addition, a ground truth database must be collected in order to train the classifier and to evaluate the results of the classification. Besides, a descriptive feature extraction process is developed to describe intensely the objects in the database. When the training sites are fully described, they can be used to train the classifier which would assign a class to each object considering its features. The assessment is done by comparing ground truth database with the classification results. Finally, discrepancies between classification and database to update determine the changes to be reviewed by a photo-interpreter.

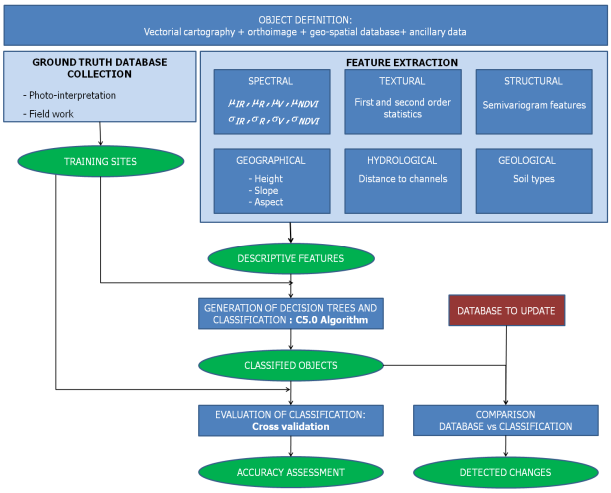
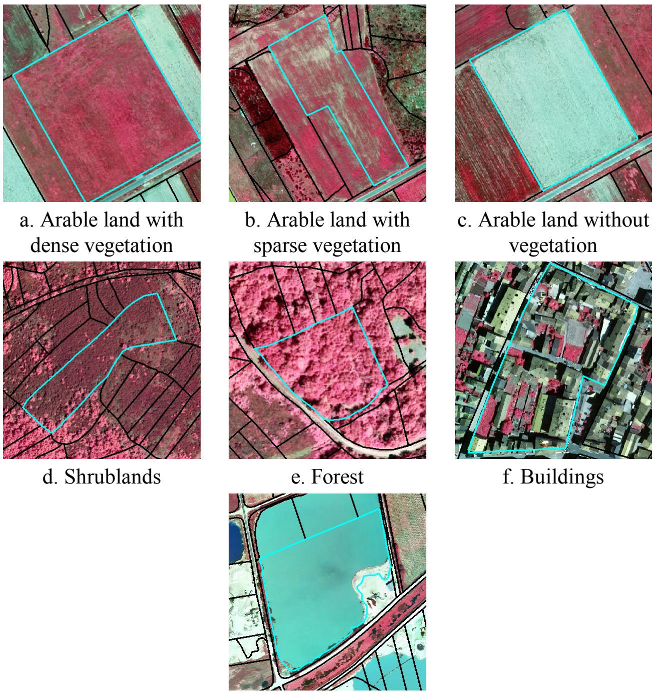

OBJECT DEFINITION:
Vectorial cartography + orthoimage + geo-spatial database+ ancillary data
GROUND TRUTH DATABASE COLLECTION
- Photo-interpretation
- Field work
FEATURE EXTRACTION
SPECTRAL
$\mu_{IR}, \mu_R, \mu_V, \mu_{NDVI}$
$\sigma_{IR}, \sigma_R, \sigma_V, \sigma_{NDVI}$
TEXTURAL
First and second order statistics
STRUCTURAL
Semivariogram features
TRAINING SITES
GEOGRAPHICAL
- Height
- Slope
- Aspect
HYDROLOGICAL
Distance to channels
GEOLOGICAL
Soil types
DESCRIPTIVE FEATURES
GENERATION OF DECISION TREES AND CLASSIFICATION : C5.0 Algorithm
DATABASE TO UPDATE
CLASSIFIED OBJECTS
EVALUATION OF CLASSIFICATION: Cross validation
COMPARISON DATABASE vs CLASSIFICATION
ACCURACY ASSESSMENT
DETECTED CHANGES

Figure 2. Workflow diagram for spatial database updating using digital image classification.

3.1 Pre-processing

High resolution images presented a high pre-processing degree: geometric rectification, panchromatic and multispectral fusion, mosaicking, and radiometric adjustments. Additionally, in order to facilitate the descriptive feature extraction process, images were resampled to 0.5 m/pixel using bilinear interpolation. This spatial resolution has been considered optimum for this application. Reference systems of maps (parcel database, geological and DEM) and images were synchronised in order to guarantee the geometric concordance of the data.

Seven classes were defined (see figure 3): *Mass of water, Buildings*, *Forest*, *Shrublands*, and *Arable land* divided in three classes to differenciate the vegetation level of the parcel: without, sparse and dense vegetation. Training sites are necessary to provide the inductive learning algorithm with samples about the classes to be trained and to evaluate the classification results. Around 1300 training samples were selected using mainly field work and also photointerpretation techniques, in order to avoid the underrepresentation of some classes, especially *Water layers*, *Forest* and *Shrub lands*. Objects are described as contiguous pixel groups with similar characteristics to the real world elements that are modelling. The main SIGPAC geographical objects are parcel and plot (Mirón, 2005). We will define a **parcel** as a continuous area of land with a unique alpha-numerical reference, and a **plot** (*recinto*) as the continuous area of land within a parcel for a single agricultural use. In this study, spatial objects are created using the plots limits contained in the SIGPAC database. To avoid the inclusion of pixels not belonging to the plot, due to errors in the delineation of limits or due to positional defects, a morphological erosion filtering was applied to each object with circular structuring element of 5 pixels diameter.



a. Arable land with dense vegetation

b. Arable land with sparse vegetation

c. Arable land without vegetation

d. Shrublands

e. Forest

f. Buildings

g. Mass of water

Figure 3. Examples of the classes defined.

3.2 Feature extraction

The use of valuable features is essential for an accurate classification. Each plot has been 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, texture, structural and ancillary data based.

Spectral features provide information about the spectral response of objects, which depends on land coverage types, state of vegetation, soil composition, construction materials, etc. These features are especially useful in the characterization of spectrally homogeneous objects, 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.

Structural features describe spatial distribution and spatial relations between the elements contained in the objects (regularity patterns, distances between elements, etc.). They are important to describe tree crops with regular planting pattern. The structural features used in this work are based on the semivariogram. The semivariogram 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. The parameters computed consider the singular points of the semivariogram, such as, the first maximum, the first minimum, the second maximum, etc., and are fully described in Balaguer et al. (2010).

The ancillary data-based features are extracted from diverse data sources and contribute to the object description adding spatial and contextual characteristics. The features employed in this study are: mean and standard deviation of the elevation, slope and aspect, average distance to the rivers and most frequent lithology.

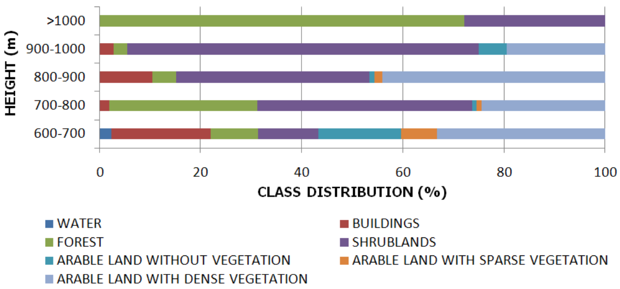


Figure 4. Distribution of classes according to terrain elevation.

Elevation is a determining factor for distribution of spontaneous vegetation and crops. Figure 4 shows the class distribution of the training samples depending on the terrain height. Land uses located in areas with elevation higher than 1,000 m above sea level are limited to *Forest* and *Shrublands*, meanwhile agricultural classes and *Mass of water* trend to be placed at lowest levels. Landforms also condition the land use. As it is shown in figure 5, plains are kept for agricultural uses while spontaneous vegetation predominates in steepest terrains. Information about landforms can be derived from the DEM, by calculating the local slope along the steepest direction. Aspect of hillsides is the major determining factor in soil water content, that influences the vegetation distribution. This feature is related with land use in mountainous areas, but has a reduced significance in plain areas.

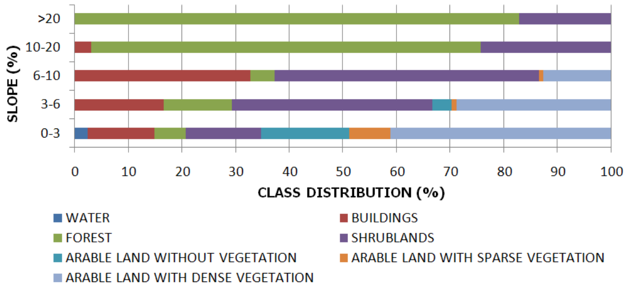


Figure 5. Distribution of classes according to terrain slope.

In mountainous terrains, crop areas are located along the valleys close to the rivers. The feature *"Distance to the rivers"* can be useful to describe the position of the parcels with respect to the channels, providing contextual information. This feature has been computed from the drainage network deduced from the DEM.

Lithology data is useful to model species distribution, even when detailled knowledge on soil vegetal species preferences is not available for many species (Gastón et al., 2009). In addition, lithologic properties are related with landforms and in some manner with the land use. A lithology class is assigned to each plot as the most frequent type given. This information was obtained from the geological map of Galicia. Geologic materials in *A Limia* are grouped in three categories: about 30% of the surface is covered with detritus deposits, 52% with granites and 18% with slates. This feature is defined as a discrete data type, requiring classifiers able to manage thematic features.

3.3 Classification through decision trees

Objects have been classified in several tests by using decision trees with different descriptive features. 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 C5.0 algorithm can manage several data types, such as continuous or discrete, which highly increases the possibility of adding descriptive features coming from diverse data sources to perform the classification.

Objects were classified using 10 decision trees, by means of the boosting multi-classifier method, which allows for increasing the accuracy of the classifier (Freund, 1995). The methodology followed by the boosting to build the multi-classifier is based on the assignment of weights to training samples (Freund and Shapire, 1997). After each tree construction, the weights vector is adjusted to show the model performance. In this way, samples erroneously classified increase their weights, whereas the weights of correctly classified samples are decreased. Thus, the model obtained in the next iteration will give more relevance to the previously wrongly classified samples (Hernandez-Orallo et al., 2004). After the decision tree set is constructed, the class assigned to an object will be done considering the estimated error made in the construction of each tree. The sum of the weights of those trees which assign the same class to one object is computed, giving that object the class with higher value.

The effect of the inclusion of ancillary data-based features was evaluated by comparing the results of several classifications. Table 1 shows the combinations of descriptive features used in the 42 classifications performed (6 groups of image-based features and seven alternatives of inclusion of the ancillary data-based features).

The performance of a classifier on the training samples from which it was constructed gives a poor estimate of its accuracy on new cases. The accuracy of the classifier can be estimated by using a separate sample set; either way, the classifier is evaluated on cases that were not used to build it. However, this estimate can be unreliable unless the numbers of cases used to build and evaluate the classifier are both large. One way to get a more reliable estimate of predictive accuracy is by *f*-fold cross-validation. In our work, the training sample set was divided into 10 blocks of roughly the same size and class distribution. For each block in turn, a classifier is constructed from the cases in the remaining blocks and tested on the cases in the hold-out block. In this way, each case is used just once as a test case. The error rate of a classifier produced from all the cases is estimated as the ratio of the total number of errors on the hold-out cases to the total number of cases.

4. RESULTS AND DISCUSSION

First row of table 1 shows the overall accuracies of classifications without considering the ancillary features. Results show that spectral features are, in this case, the image-based features with the highest discriminant power. As expected, the combination of spectral information with texture and structural features produced moderate increments at the overall accuracies.

		Image-based features					
		Spectral	Textural	Structural	Spectral+Textural	Spectral+Structural	Spectr+Text+Struct
Ancillary data-based features	None	80.8	69.7	66.5	84.1	82.0	84.4
	Height	82.6	74.8	72.0	85.3	84.9	86.9
	Slope	81.2	73.4	71.9	84.9	84.8	85.5
	Aspect	79.2	70.3	68.5	82.8	83.5	84.2
	Lithology	81.5	72.0	68.0	85.4	84.2	85.4
	River distance	84.3	74.7	72.4	86.7	86.7	86.6
	All	86.6	80.2	77.4	86.7	86.0	86.8

Table 1. Overall accuracies of the classifications with different input data

The addition of ancillary data-based features produced overall accuracies increments with the exception of considering the feature *Aspect* which, in some cases, entailed a slight accuracy decrease. This is due to the fact that this feature does not give additional information in mainly plain terrains and should be considered jointly with the *Slope*. The overall accuracy increments are more significant as worst is the description of the objects with the image-based features. The ancillary feature that presents a higher discriminative power is *Distance to the rivers*. The addition of this feature produced increments on the overall accuracies ranging from 2.2% to 5.9%, meanwhile the *Height* produced slightly lower accuracy increments.

The analysis of confusion matrixes reveals that the *Distance to the rivers* increases the separability of the classes *Forest* and *Shrubland*. Using this feature, average increments of the producer's accuracy of these classes was 11.6% and 4.5%, respectively. Moreover, the average user´s accuracies incremented 7.5% for *Forest* and 6.5% for *Shrubland. However,* this is the most subjective feature because its value depends on the criterion employed to define the channels from the DEM.

Lithological properties of plots had a reduced effect in the classifications. In all cases, overall accuracies obtained were slightly higher when this feature was considered, not having a negative effect. In our study area, no correlation was observed between land uses and lithology classes, furthermore, the low level of detail of the geologic cartography employed made difficult to properly describe the geologic properties of the plots.

In *A Limia*, crops are limited to slopes lower than 10%, whereas forest and shrublands predominate in higher slopes. This feature had a positive effect in classification but it was less significant than other features.

The jointly addition of ancillary data-based features involved increments of the overall accuracies ranging from of 2.4% when the plots were described with spectral, textural and structural features, to a 10.9% when only the structural features were employed.

5. CONCLUSIONS

This study evaluates the contribution of geographic ancillary information into object-based classification of high resolution images for database updating in a particular area of study. Numerous alternatives to describe the objects contained in spatial databases have been compared in order to deduce the object class.

Features regarding to topographic properties and spatial arrangement provide useful information to describe objects and complement spectral features in a similar way to the textural and structural features but with fewer variables.

ACKNOWLEDGEMENTS

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

REFERENCES

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.

Bruzzone, L., Conese, C. Maselli, F., Roli, F., 1997. *Multisource Classification of Complex Rural Areas by Statistical and Neural-Network Approaches.* Photogrammetric Engineering & Remote Sensing, 63(5): 523-533.

Cohen, Y., Shoshany, M., 2002. *Integration of remote sensing, GIS and expert knowledge in national knowledge-based crop recognition in Mediterranean environment.* International Journal of Applied Earth Observation and GeoInformation, 4: 75-78.

Freund, Y., 1995. *Boosting a weak learning algorithm for majority.* Information and Computation, 121(2): 256-285.

Freund, Y., Shapire, R.E., 1997. *A decision-theoretic generalization of on-line learning and an application to boosting.* Journal of Computer and System Sciences, 55(1): 119-139.

Gastón, A. Soriano, C., Gómez-Miguel,V. 2009. *Lithologic data improve plant species distribution models based on coarse-grained occurrence data.* Investigación Agraria: Sistemas y Recursos Forestales, 18(1):42-49.

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.

Heipke, C., Straub, B.M., 1999. *Towards the automatic GIS update of vegetation areas from satellite imagery using digital lanscape model as prior information.* IAPRS, 32 - Part 3-2W5:, 8-11 September 1999, Munchen, Germany, pp. 167-174

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

Hoffer, R.M., 1975. *Natural resource mapping in mountainous terrain by computer analysis of ERTS-1 Satellite Data.* LARS Information Note 061575, Purdue University, Indiana, 124 p.

Hutchinson, C.F., 1982. *Techniques for combining Landsat and ancillary data for digital classification improvement.* Photogrammetric Engineering & Remote Sensing, 48(1): 123-130.

Lawrence, R.L., Wright, A., 2001. *Rule-based classification systems using classification and regression tree (CART) analysis.* Photogrammetric Engineering & Remote Sensing, 67(10): 1137-1142.

Mirón, J., 2005. *Cadastre and the reform of European union's common agricultural policy. Implementation of the SIGPAC(1).* Catastro, 54:161-172.

Olsen, B.P., Knudsen, T., Frederiksen, P., 2002. *Digital Change detection for map database update.* International Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences, 34(2): 357-363.

Pedroni, L., 2001. *Discriminación de diferentes tipos de bosque tropical mediante imágenes de satélite y datos auxiliares.* Revista Forestal Centroamericana, 34: 12-18.

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

Raclot, D., Colin,F., Puech, C., 2005. *Updating land cover classification using a rule-based decision system.* International Journal of Remote Sensing, 26(7): 1309-1321.

Recio, J.A., 2009. *Técnicas de extracción de características y clasificación de imágenes orientada a objetos aplicadas a la* *actualización de bases de datos de ocupación del suelo.* PhD Thesis. Universidad Politécnica de Valencia http://hdl.handle.net/10251/6848

Rogan, J., Miller,J., Stow,D. Frankling, J., Levien, L., Fischer, C., 2003. *Land cover change mapping in California using classification trees with multitemporal Landsat TM and ancillary data.* Photogrammetric Engineering and Remote Sensing, 69(7): 793-804.

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.

Strahler, A.H., Logan, T.L., Bryant, A., 1978. *Improving forest cover classification accuracy from Landsat by incorporating topographic information.* Proceedings of the 12th International Symposium on Remote Sensing of Environment, Ann Arbor, Michigan (Environmental Research Institute of Michigan), pp. 927-942.

Treltz, P., Howarth, P., 2000. *Integrating Spectral, Spatial, and Terrain Variables for Forest Ecosystem Classification.* Photogrammetric Engineering and Remote Sensing, 66(3): 305-317.

Walter, V., 2004. *Object-based classification of remote sensing data for change detection.* ISPRS Journal of Photogrammetry & Remote Sensing 58(3-4): 225– 238.

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.