

Application of a hybrid classifier to discriminate Mediterranean crops and forests. Different problems and solutions.

Serra, P.¹; Moré, G.²; Pons, X.^{1,2}

- (1) Departament de Geografia, Edifici B, Universitat Autònoma de Barcelona, 08193 Bellaterra.
(2) Centre de Recerca Ecològica i Aplicacions Forestals (CREAF), Edifici C, Universitat Autònoma de Barcelona, 08193 Bellaterra.

ABSTRACT

Identifying forest covers and crops with a detailed legend and a high accuracy is still an important challenge for remote sensing researchers because of the spatial fragmentation and dynamic phenology of these covers in many areas. To achieve these objectives a high number of remote sensing images is usually needed. Such elevated number of images involves the presence of clouds, shadows, smog or snow, among other, producing a loss of significant classifiable surface. This paper presents an improved hybrid classifier based on two modules: ISOMM that allows classifying more study area through considering NODATA values, and CLSMIX, that assigns spectral categories to thematic classes through spatial correspondence. Results show an increase in the classified surface of 46.1%, in the case of crop discrimination, and of 25.3%, in the case of forest covers, all of them maintaining a similar and high thematic accuracy (85% to 90%).

1. INTRODUCTION

Traditionally, classification methods of remote sensing images have been divided into two broad categories: supervised, involving the selection of areas on the image which statistically characterise the thematic categories of interest, and unsupervised, attempting to identify spectrally clusters within the image that are later assigned to categories. The most commonly applied supervised classification method, the maximum likelihood procedure, is not very effective when covers are not normally distributed which is not uncommon (Richards 1993). On the other hand, under the conventional procedure of the unsupervised classification, spectral classes of pixels are first identified by cluster analysis being ISODATA (Interactive Self Organizing Data Analysis) a non-hierarchical clustering algorithm commonly used in remote sensing; once the clusters are obtained, 'rules of correspondence' between the spectral and the land-cover and land-use (LCLU) categories are established; these rules are normally known through fieldwork or ancillary data. The standard procedure of unsupervised classification is based on the assumption that each spectral class corresponds to one and only one thematic category and vice-versa, but this does not always work because there are different possible patterns of correspondence. Another classification option is to apply a hybrid classifier that combines an unsupervised classification and training areas, those collected as in a conventional supervised classification. In our case, the unsupervised classification is based on ISODATA algorithm (Duda & Hart, 1973) while the assignation of spectral categories to training areas is made using the CLSMIX module of MiraMon (Pons, 2002). Serra *et al* (2003) applied this methodology in Mediterranean crop classification obtaining satisfactory results.

Discriminating Mediterranean crops and forests with a detailed legend is a difficult issue due to the high spatial fragmentation. The medium temporal resolution and pixel size of Landsat images allow analysing the evolution of vegetation phenology with relative low cost (Cohen and Goward, 2004). In our case a Landsat subscription made available a large number of images from 2002 to nowadays (an image every 16 days over three full frames). Nevertheless, an important part of that time series information is usually turned down due to the fact that a considerable quantity of images is not really useful because of the presence of missing values (NODATA values) (figure 1). There are two main reasons explaining the existence of NODATA values in an image. Firstly, the presence of clouds, smog

or snow in the scene causes a high number of images being completely useless (Viñas & Baulies, 1995; Fuller et al., 1995; Wulder, 2002; Homer et al., 1997), or partially useless if a mask is applied, excluding these zones from the classification. On the other hand, images are radiometrically corrected, reducing the number of undesired artefacts that are due to the effects of the atmosphere or the differential illumination which is in turn due to the time of the day, the location in the Earth and the relief (zones that are more illuminated than others, shadows, etc) (Pons & Solé-Sugrañes, 1994). In spite of these main radiometric benefits, in some mountainous regions, radiometric correction often invalidates some zones due to a bad radiometric quality (relief shadows, non lambertian behaviour, etc.).

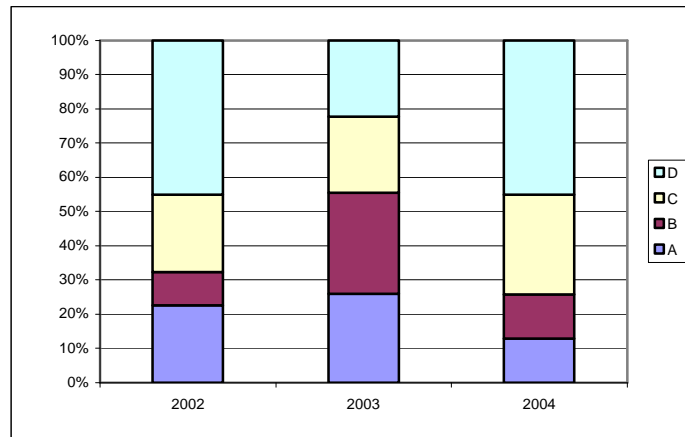


Figure 1: Percentage of images with different cloud cover from 2002 to 2004 for path 197 – row 31 and 198 – row 31 and 32 of Landsat. A = <10%; B = ≥10% and < 50%; C = ≥ 50% and < 80%; D = ≥ 80%

The objective of this paper is to present the benefits of using a modification of the ISODATA algorithm (called ISOMM) for classifying Mediterranean crops and forests by means of a hybrid classifier.

2. STUDY AREA AND MATERIALS

Hybrid classifiers have been applied to two different study areas: one with predominance of crops (study area A) and another one with a high proportion of forest covers (study area B). Location of each zone is represented in figure 1.

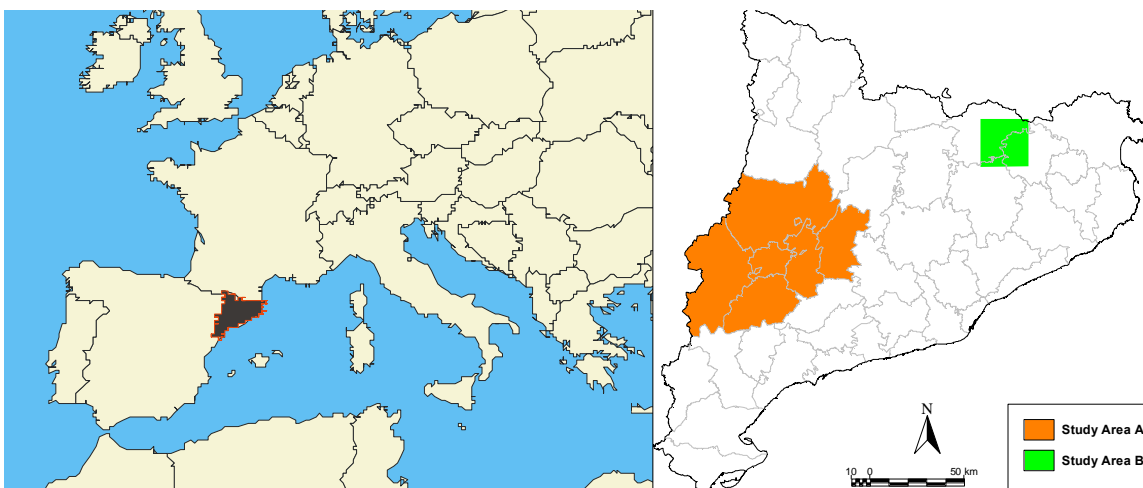


Figure 1: Location of study areas.

The study area A, for crop mapping, belongs to the path 198 and row 31 of Landsat and is located in the centre of Catalonia (North-East of Spain) comprising 348,533 ha. The study area B, for forest mapping, belongs to the path 197 and row 31 of Landsat and is located in the north of Catalonia, covering 50,176 ha.

Images were acquired through a Thematic Mapper (TM) Landsat subscription (initially Landsat-7 ETM+). The subscription implies that all the images were purchased from January to December whatever the percentage of cloud cover. Images were geometrically corrected using the procedure developed by Palà and Pons (1995). Georeferencing was done using a mean of 26 Ground Control Points (GCPs) per image. The accuracy of the georeferencing was assessed through the root mean square (RMS) of the location of independent test GCPs (a mean of 14 GCPs per image). The images had a mean RMS error of about 0.7 pixels. On the other hand, in the radiometric correction digital numbers were converted into reflectance values using the sensor calibration parameters and other factors such as atmospheric effects, solar incident angle accounting for the relief, etc. (Pons and Solé-Sugrañes, 1994). The resultant corrected images presented a coherent range of reflectance values.

Table 1 shows the classification variables for each study area. In the case of crops (study area A), classification used all the original ETM+ or TM bands, six per image not including the thermal band due to its pixel size (80 or 120 m), plus the Normalized Difference Vegetation Index (NDVI) and the Wetness Index extracted from the Tasseled Cap transformation (Crist and Cicone, 1984). In the case of forest covers (study area B) the six original bands were also introduced, plus the NDVI of each date, plus climatic and relief variables. All these data were standardized to avoid undesired artefacts due to the fact that they use different physical units (reflectance, centigrade degrees, etc). The legend used can also be seen in table 1.

Study Area	Dates	Radiometric Variables	Other Variables	Legend
A	May 16, June 1 and 17, July 19, August 4, October 23 and November 8, all from 2004.	1, 2, 3, 4, 5 & 7 Landsat TM bands.	NDVI, Wetness Index.	Winter cereals, maize, alfalfa, rice, fruit trees, sunflower, fallow land, olive trees, vineyards and other herbaceous crops.
B	June 13 2002, March 12 2003, August 16 2003, April 29 2004.	1, 2, 3, 4, 5 & 7 Landsat TM (after May 31 2003) or ETM+ bands (before May 31 2003)	NDVI, minimum and maximum annual temperature, average annual solar radiation, average annual precipitation, slope.	<i>Quercus ilex</i> , <i>Fagus sylvatica</i> , <i>Freaxinus sp.</i> , high mountain shrub, Mediterranean shrub, grass, <i>Pinus uncinata</i> , <i>Pinus sylvestris</i> , <i>Pinus pinaster</i> , <i>Quercus canariensis</i> , <i>Quercus humilis</i> .

Table 1: Variables and legend used in ISOMM for study areas A and B.

Once all the images were corrected and included, and in order to avoid classification confusions, the next step was to mask crops and forests using the *Mapa de Cobertes del sòl de Catalunya* (Land Cover Map of Catalonia; MCSC, 1999). This map is the result of photo interpretation of colour orthophotos 1:25 000 from 1993.

3. METHODOLOGY

In the ISOMM module, clusters are formed by iterative assignments of n-dimensional pixels. These assignments are based on the minimum Euclidean distance of a pixel from all current cluster centroids. The initial set of centroids, the seeds, is obtained prior to the clustering run. Cluster centroids, after each iteration, are updated to the centroid of all currently assigned pixels.

3.1. ISOMM characteristics

One of the main characteristics of ISOMM is that admit a high number of input variables (hundreds). The main utility of this property is to allow the use of high temporal resolution satellite series and other topographic and climatic

variables. It also accepts different data formats as byte, integer or real. The module permit obtaining an elevated number of statistical categories (32767) that may be eliminated or modified by the user using two different parameters: the minimum Euclidean distance between two valid clusters and the minimum number of pixels per cluster in order to consider the cluster valid. In the former case, clusters are fused in a single category when the Euclidean distance is lower than a minimum value define by the user, while in the latter case a cluster is eliminated if its total area is lower that a threshold established by the user. Finally, the module requires the introduction of the following parameters: the numbers of desired clusters, the maximum number of iterations before terminates and a threshold value for terminating the algorithm (minimum acceptable proportion of pixels that do not change from a cluster to another between two iterations).

The module presents three options for obtaining the initial seeds: i) along the multivariate diagonal calculated from all the input variables, ii) a random distribution in all the multivariate space, iii) a distribution based on a equidistant sample over the image (for example a seed every 50 pixels, etc.).

As previously discussed, one of the main problems when a set of images is used, is the presence of NODATA values whatever the origin (clouds, etc., figure 1). In most cases this fact requires using other alternative images or subimages to avoid NODATA. The strategy of most software is to remove a pixel when it appears as NODATA in at least one of the input variables. This approach is statistically correct but in some cases it means an excessive decrease of the study area being classified. NODATA values are usually distributed in a different pattern when changing the date (clouds and shadows). For example, in a classification with 56 variables, we can find a pixel with NODATA in 8 of the variables. The classic strategy will remove this pixel although it could be correctly classified using the rest of 48 valid variables. In the present implementation, the user determines the parameter “NODATA tolerance” or, in other words, how many variables with NODATA will be acceptable in the classification stage. The possibilities can vary from 0 (that value implies the use of the traditional strategy, so removing any pixel with at least one NODATA value) to a number equal to the total number of variables less one (the less restrictive option because it is able to classify a pixel with only one valid variable). Figure 3 shows an example of this situation: Image A corresponds to an original RGB combination (4,5,3 of July) without NODATA. Image B corresponds to another original RGB combination (4,5,3, of August) with NODATA due to clouds and shadows (in dark colour). Image C corresponds to ISOMM classification in the less restrictive option. Although image B contains NODATA values, ISOMM is able to classify all the study area coherently, specially parcel 1 where a part is NODATA. Note that with the traditional strategy all the pixels with NODATA of image B would not be classified, losing an important part of the study area.

To ensure the statistical consistence of the process, cluster centroids are only calculated from those pixels without any NODATA, assign the rest to them according to the statistical similarity. For this reason, it is suitable to avoid including images with excessive NODATA.

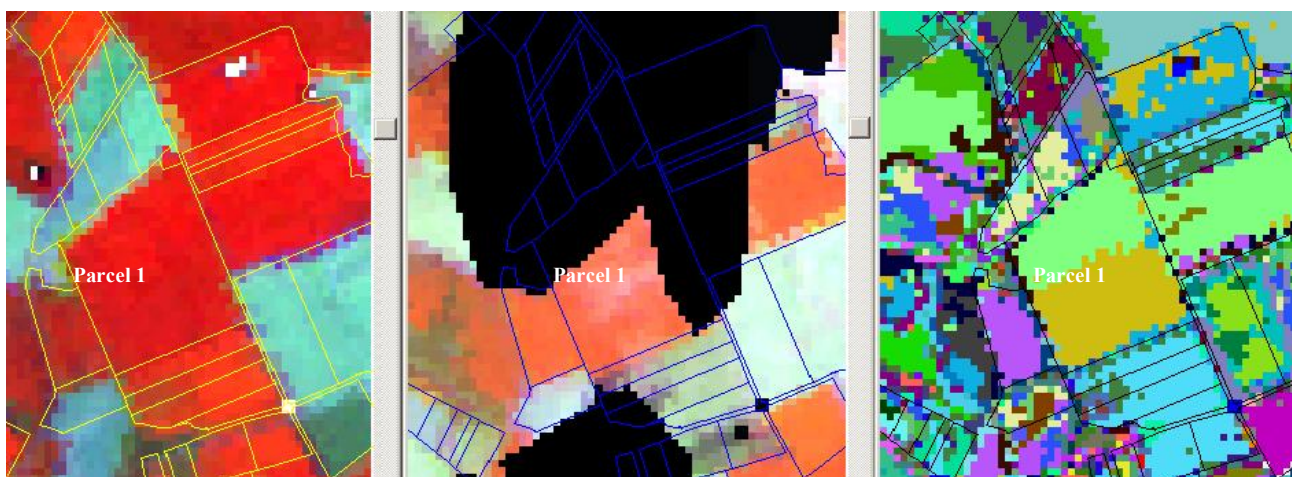


Figure 3: Image A (left) corresponds to an original RGB combination (4-5-3 of July 19) without any NODATA pixel. Image B (centre) corresponds to another original RGB combination (4-5-3 of August 4) with some NODATA pixels due to clouds and shadows (in dark colour). Image C (right) corresponds to ISOMM classification in the less restrictive option. Rural cadastre has been overlaid in all of them.

3.2. CLSMIX

In the second part of the classification process, CLSMIX assigns every spectral class to a thematic class using two different parameters: fidelity and representativity. In one hand, the threshold proportion at which to accept a spectral class as being a part of a thematic class in terms of the proportion of the spectral class that is inside the thematic class. For example, 0.7 will mean that if 70% or more of the spectral class inside the training areas is under a given category of these areas, then this spectral class will be assigned to this category. On the other hand, the threshold proportion at which to accept a spectral class as being a part of an LCLU category in terms of the proportion of the LCLU category that is formed by a given spectral class. For example, 0.01 will mean that if 1% or more of the LCLU category is formed by a given spectral class, this spectral class will be assigned to the LCLU category. Table 2 shows a very clear example of the first situation: in that case spectral class #40 will be assigned to winter cereals because 98.63% of the pixels belong to that LCLU category. Simultaneously, the proportion of winter cereals that are formed by spectral class #40 is 1.95%.

Spectral class #40	Number of pixels	%
Fruit trees	4	0.04
Other herbaceous crops	93	0.99
Winter cereals	9279	98.63
Vineyards	7	0.07
Fallow land	25	0.27

Table 2: Example of CLSMIX assignation: Proportion of the spectral class that is inside the thematic class.

Note that when classifying a given pixel, the module chooses the category that has the more ‘reasonable’ assignation: i) The spatial correspondence between the spectral class and the training areas of that LCLU category (the spectral class is inside the training area), ii) The spectral class is mainly inside this LCLU category (an important proportion of the spectral class belongs to the category) and iii) The spectral class is a not insignificant part of the LCLU category.

Conversely, a pixel will remain unclassified if no training area covers pixels in the same spectral class or if, given the input thresholds, no spectral class is adequate for it: either the pixel belongs to a class that is split too much between two or more LCLU categories (no clear LCLU tendency of the spectral class) or the pixel belongs to a class that is poorly representative of the total area of any LCLU category (perhaps the spectral class is noisy).

4. RESULTS

Classification results for the two study areas were obtained. Final crop maps show that with the traditional strategy (tolerance NODATA = 0) ISOMM considers only 48.7% of whole study area due to clouds and shadows in some dates, while CLSMIX classifies 40.1 % of total study area with a thematic accuracy of 86.5%. With the option less restrictive (tolerance NODATA = 55) ISOMM considers 100% of the total study area while CLSMIX classifies 87.1% with a total accuracy of 86.2% (figure 4).

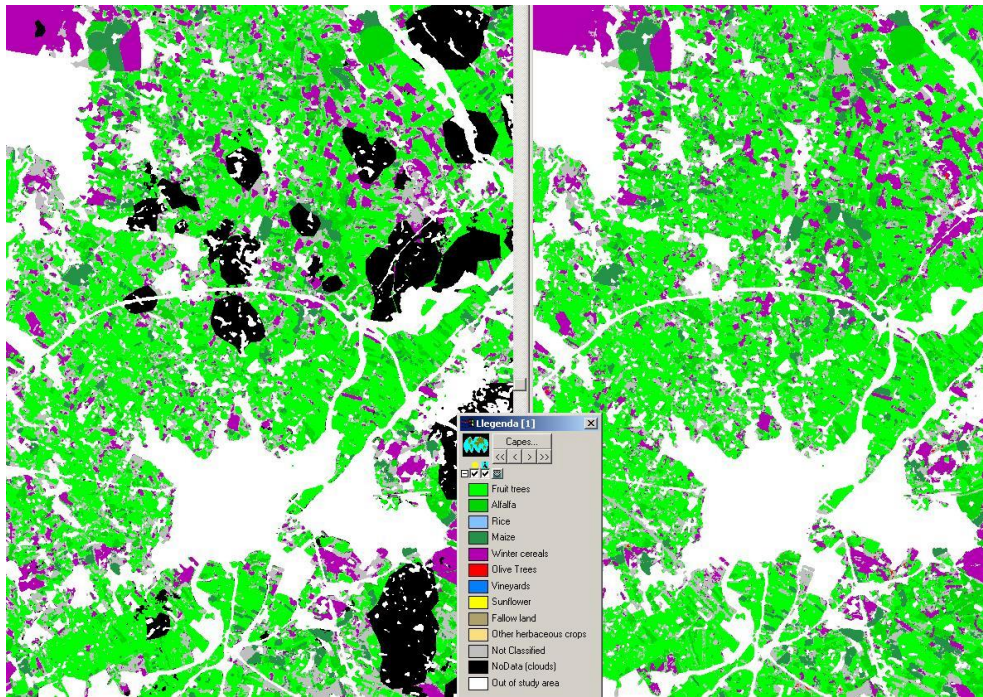


Figure 4: Two examples of crop maps: with tolerance NODATA = 0 (left) and with tolerance NODATA = 55 (right).

In figure 5 results from final forest maps are shown, being these results very similar to crop maps. When more variables with NODATA are accepted, much surface is classified (IS line). With the traditional strategy (tolerance NODATA = 0) ISOMM classifies 70.8% of the study area, while CLSMIX considers 63.4% (CL line) with a thematic accuracy of 88.7% (TA line). With the less restrictive option, ISOMM considers 100%, while CLSMIX discriminates 86.1% of the study area with a total accuracy of 90.0%

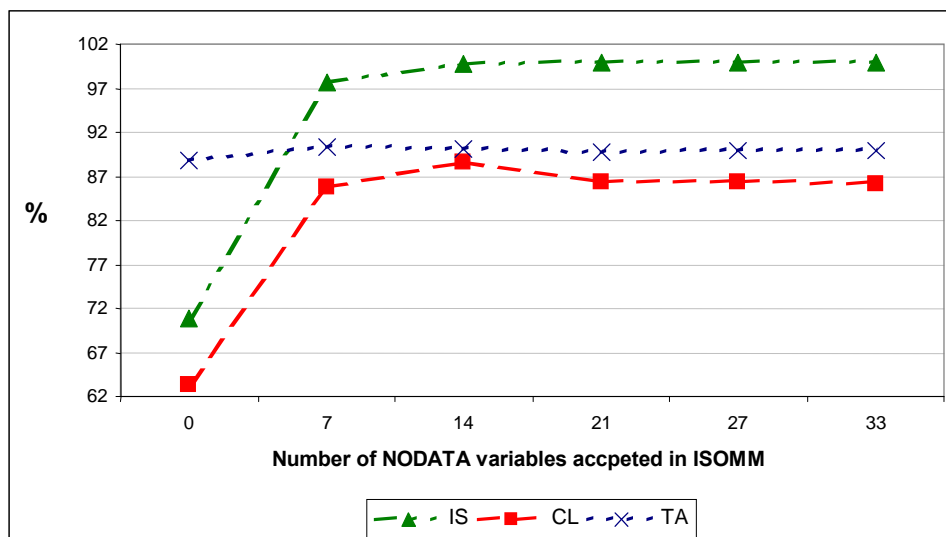


Figure 5: Effect of the “NODATA tolerance” parameter in a classification. (IS = Surface classified by ISOMM in relation to the entire study area; CL = Surface classified by CLSMIX in relation to the entire study area; TA = Thematic accuracy).

5. CONCLUSIONS

The hybrid classifier has shown useful results for discriminating Mediterranean crops and forest covers with Landsat images and detailed legends. The modification of ISODATA algorithm allows solving the annoying problem of

the presence of clouds, shadows, smog or snow in some of the used remote sensing images. In this case, the traditional strategy would be to refuse such images excluding them from the classification procedure. The new module, ISOMM, classifies all the pixels with the same thematic accuracy (above 85%). To ensure the statistical consistence of the process, cluster centroids are only calculated from those pixels without any NODATA, assigning the rest to them according to their statistical similarity.

New research will be developed to analyse the consequences in the classification (thematic accuracy) when excluding cloudy images, probably losing radiometric and phenologic significant information.

BIBLIOGRAPHY

- Cohen, W.B., Goward, S.N. (2004) "Landsat's role in ecological applications of remote sensing". *BioScience*, 54, 6, 535-545.
- Crist, E.P., Cicone, R.C. (1984) "Application of the tasseled cap concept to simulated Thematic Mapper data". *Photogrammetric Engineering and Remote Sensing*, 50, 343-352.
- Duda, R.D. & Hart, P.E. (1973) *Pattern Classification and Scene Analysis*, John Wiley & Sons, New York.
- Fuller, R.M., Groom, G.B. & Jones, A.R. (1994) "The land cover map of Great Britain: an automated classification of Landsat Thematic Mapper data", *Photogrammetric Engineering & Remote Sensing*, 60, 553-562.
- Homer, C.G., Ramsey, D.R., Edwards, T.C. Jr. & Falconer, A. (1997) "Landscape cover-type modeling using a multi-scene Thematic Mapper mosaic", *Photogrammetric Engineering & Remote Sensing*, 63(1): 59-67.
- MCSC (1999) Mapa de Cobertes del Sòl de Catalunya. *The Internet*. (<http://www.creaf.uab.es/mcsc/index.htm>).
- Palà, V. & Pons, X. (1995) "Incorporation of relief into geometric corrections based on polynomials", *Photogrammetric Engineering & Remote Sensing*, 61(7): 935-944.
- Pons, X. (2002) "MiraMon. Geographic information system and remote sensing software", Centre de Recerca Ecològica i Aplicacions Forestals, CREA. Bellaterra. ISBN: 84-931323-5-7
- Pons, X. & Solé-Sugrañes, L. (1994) "A simple radiometric correction model to improve automatic mapping of vegetation from multispectral satellite data". *Remote Sensing of Environment*, 48, 191-204.
- Richards, J.A. (1993) "Remote sensing digital image analysis, an introduction". Berlin, Springer-Verlag.
- Serra, P., Pons, X. & Saurí, D. (2003) "Post-classification change detection with data from different sensors: some accuracy considerations", *International Journal of Remote Sensing*, 24: 3311-3340.
- Viñas, O. & Baulies, X. (1995) "1:250 000 Land-use map of Catalonia (32 000 km²) using multitemporal Landsat-TM data", *International Journal of Remote Sensing*, 16(1): 129-146.
- Wulder, M. (2002) "Mapping the land cover of the forested area of Canada with Landsat data", *Proceedings of 2002 International Geoscience and Remote Sensing Symposium*, June 24-28, Toronto, Canada.

BIOGRAPHY OF PERE SERRA

Pere Serra Ruiz received M.S. degrees in Remote Sensing Applications and GIS from the Institute for Space Studies of Catalonia at Barcelona in 1997. He received the Ph.D. degree in 2002 with the dissertation entitled "Agrarian Landscape Dynamics in the Alt Empordà (1977-1997). An Analysis from Remote Sensing and Geographical Information Systems"

He has been concerned in teaching research in spatial statistical models, concretely in lineal and logistic regressions, supervised and fuzzy classifiers, clustering, principal components analysis and factor analysis. He has been also involved in mathematical morphology, post classification comparisons and error matrices.

He is currently a Junior Lecturer and a Research Assistant at the Department of Geography from the Autonomous University of Barcelona. He is working in monitoring agricultural water needs in Catalonia for the Water Catalan Agency and applying remote sensing images in Mediterranean crops discrimination for the Catalan Ministry of Agriculture, Livestock and Fishing.