

Improvements on Classification by Tolerating NoData Values

Application to a Hybrid Classifier to Discriminate Mediterranean Vegetation with a Detailed Legend Using Multitemporal Series of Images

Gerard Moré¹, Xavier Pons^{2,1}

¹Center for Ecological Research and Forestry Applications,
CREAF
Bellaterra, Spain
g.more@creaf.uab.es, x.pons@uab.es

Pere Serra²

²Department of Geography
UAB, Autonomous University of Barcelona
Bellaterra, Spain
p.serra@uab.es

Abstract—Natural and crop vegetation phenologic data become indispensable when creating thematically and geographically detailed maps through satellite images classification. Several date acquisition is necessary to achieve this cartography. However, the presence of clouds, shadows, snow, etc, is usual when many different dates are used and that fact implies an important loss in classifiable surface. This work presents a hybrid classifier designed to deal with the common problems appeared in the classification of Mediterranean vegetation. Specifically, IsoMM, the first phase of the hybrid methodology, is an unsupervised classifier that allows a better use of temporal series thanks to a particular treatment of NoData values (or missing values) in the images. This methodology has been applied to a Mediterranean forestry zone with a legend of eleven categories and has been compared to a Maximum Likelihood classifier. The presented improvements allow classifying more surface than a common NoData treatment strategy (wheter unsupervised, Maximum Likelihood classification or the extraction of a problematic date) and achieving high accuracy level.

Keywords: IsoData; Landsat; NoData; missing values

I. INTRODUCTION

Classification methods of remote sensing images have been usually divided into two broad categories: supervised, involving the selection of areas on the image which statistically characterize the thematic categories of interest, and unsupervised, attempting to identify clusters within the image that are later assigned to categories [1]. Supervised methods usually assume that each thematic class belongs to a unique statistic class. In other words, the classes we want to classify fit with a unique and characteristic probability function. However, this assumption is not always fulfilled in Mediterranean vegetation classifications due to the species richness, intra-species variability and age and habitat fragmentation [2]. On the other hand, the number of clusters we want to classify is an important decision when using an unsupervised method. When a low number of clusters are required, the resulting classification becomes an excessively simplified version of the scene and not all the covers of interest are always represented. Nonetheless, assignation of spectral classes to thematic classes

becomes hard to do manually if a high number of identified clusters are required (e.g., more than 50 clusters).

Our proposal is a hybrid classifier consisting of two phases: an unsupervised classifier, and an automatic assignation of statistical classes to thematic classes that uses training areas. The aim of this classifier is to take advantage of both traditional classification methods in remote sensing and to obtain better results in Mediterranean vegetation. In our case, thematic classes don't have to fit to a particular probability distribution. That permits the existence of reasonably heterogeneous or multimodal classes due to, for example, the presence of different phenologic stages or different biophysical properties in the same scene [3]. A high number of clusters is usually required to accurately represent all the statistical variability in the scene. The statistical class to thematic class procedure permits doing an automatic and quite objective assignation through the correspondence of the unsupervised classification and training areas while a certain control is left to the user.

Discriminating Mediterranean crops and forests with a detailed legend using remote sensing is a difficult issue that requires spectral information as well as temporal information [4]. The medium temporal resolution and pixel size of Landsat images allow analyzing the evolution of vegetation phenology with relative low cost [5]. 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). That permits using phenology as an important factor in discriminating among vegetation types or even vegetation species. Moreover, and unlike other sensors in platforms that can offer a better temporal resolution but a poor spatial resolution, Landsat permits obtaining cartography with a 30 meters geographic detail and over reasonably large regions (32400 km²).

Usually, when a high number of images are available, an important number of them are not really usable, so, in practice, we are rejecting a part of the information contained in the temporal series. This rejection is due to the fact that some images contain some zones with no valid values (NoData or

missing values). These NoData zones mainly come from two sources. For a side, the presence of clouds, fog and snow makes a high number of images be useless or, at most, be partially useful after an appropriate masking procedure. For another side, a radiometric correction is applied to the images to reduce the undesired artifacts that are due to the effects of the atmosphere or to the differential illumination which is in turn due to the time of the day, the location on Earth and the relief (zones that are more illuminated than others, shadows, etc) [6]. Some mountainous zones can become invalidated by this procedure by considering that the radiometry in these zones is not reliably enough (*i.e.*, relief shadows, non lambertian behaviour, etc.).

Conventional image classification software tends to reject those pixels with at least a NoData value within the entire vector of variables introduced in the classification process. That provokes that images with NoData zones (due to the presence of clouds, snow or radiometric problems) are usually rejected to avoid obtaining an important surface of unclassified pixels in the final classified image. But, at the same time, a loss in phenologic information is implied in this rejection resulting in a lower classification power. In other words, having fewer images, poorer classification accuracies will be obtained, or the use of a more simplistic legend (describing a lower number of categories being less sensitive to phenologic information) will be necessary to maintain high accuracy levels in the results. Nonetheless, labeling a part of a continuous image as NoData is not a reason to reject the entire image if there is valid radiometric information in other zones of the same image.

The aim of this work is to expose an unsupervised classifier that presents, among other features, the ability to overcome the problems mentioned before and to classify a pixel despite not having valid values in all the variables used in the classification. The exposed unsupervised classifier, called IsoMM, has to be understood in the context of a hybrid classifier: After the unsupervised phase, statistical classes will be assigned to thematic classes defined by the user through training areas. That second phase corresponds to the ClsMix algorithm and will only be described briefly in this work.

II. HYBRID CLASSIFIER MAIN CHARACTERISTICS

A. *Unsupervised Phase: IsoMM*

IsoMM is a clustering algorithm based on IsoData [7] that divides spectral or statistical space in homogeneous groups through the use of an iterative procedure in which every pixel of the image is assigned to the more statistically similar cluster using some statistical distance measure until the process is stabilized. IsoMM has three options to throw away initial seeds (*i.e.*, reference groups in the first iteration): i) along the multivariate diagonal calculated from all the input variables, ii) random distribution in all multivariate space, iii) a distribution based on an equidistant sample over the image (*e.g.*, a seed every 50 pixels). IsoMM admits a high number of input variables (*i.e.*, hundreds), regardless of their data format (*i.e.*, byte, integer or real), in order to work with temporal satellite series and other topographic and climatic variables. An elevated number of statistical categories (a maximum of 32767 clusters) can be obtained. The user can define the statistical

distance criteria used (Euclidian or Manhattan), the number of clusters desired, the maximum statistical distance threshold to fuse to clusters being very similar, the minimum number of pixels forming a cluster to consider it valid, the maximum numbers of iterations, a convergence threshold value for terminating the algorithm (*i.e.*, minimum acceptable proportion of pixels that do not change from a cluster to another between two iterations), and, finally, the maximum number of variables with NoData accepted when classifying a pixel (the "NoData tolerance" threshold).

Classification algorithms usually pay little attention to the treatment of NoData values (missing values). In the worst cases, NoData values are considered as ordinary values giving rise to unacceptable results. When these pixels are considered as a special value, the most common treatment consists in considering that a pixel can not be classified if it contains at least one NoData value within the entire variables vector in the classification.

A remarkable presence of NoData values appears when using a massive number of images (high temporal resolution), as mentioned previously. For instance, only about 20% of the 2004 images for path 197-row 31 and path 198-row 31 and 32 of Landsat have a cloud cover lower than 50% of the scene, and less than 15% of the images had a cloud cover lower than 10%. These NoData values tend to be distributed in different zones of the scene depending on the date of the image. For instance, in a classification using 6 satellite images taken in different dates, it's probable to find clouds in one of these images. The pixels affected for the cloud cover will have valid information in 5 variables. The common solutions are to remove the date containing information or to ignore those pixels under the cloud cover. However, this date can contain valuable phenologic information and can be particularly useful to discriminate some categories as well as those pixels under the cloud cover can be correctly classified using the information contained in their 5 valid variables.

The use of Euclidian or Manhattan distance in the classifier core of IsoMM (two measures that don't need statistical multivariate dispersion matrices and not being heavily parametric) facilitates the use of different sets of images depending on the availability of each pixel in the calculation of the statistical distance between a pixel and a cluster. It is on the user's decision to determine how many variables with NoData (without information) are acceptable to classify a pixel. The "NoData tolerance" parameter can take values from 0 (corresponding to the traditional strategy of removing those pixels not containing the entire variables vector with valid information) to a value equal to the number of input variables less one (the most permissive option where a pixel can be classified using a unique valid variable). Pixels containing a number of variables with NoData higher than the "NoData tolerance" threshold will remain unclassified, whereas those pixels with a number of variables with NoData lower than the "NoData tolerance" threshold will be classified using those variables with valid information.

To ensure statistical consistence of the process, cluster centroids are only calculated from those pixels without any variable with NoData, and the remaining pixels are assigned to

those clusters according to their statistical similarity. That implies a limitation in adding images with excessive NoData surface since the probability to define representative clusters for the entire scene can be dramatically reduced.

B. Statistical Class to Thematic Class Assignment: ClsMix

The second phase of the classifier is carried by ClsMix, an algorithm that assigns every spectral class to a thematic class through training areas defined by the user. ClsMix uses two different parameters to deal with the main problems in the assignment procedure [8]: fidelity and representativity. Fidelity is 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; and representativity is the threshold proportion at which to accept a spectral class as being a part of a Land-Cover/Land-Use (LCLU) category in terms of the proportion of the LCLU category that is formed by a given spectral class.

A pixel will remain unclassified if no training area covers pixels in their 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). The procedure is used and explained in more detail in [9].

III. STUDY AREA AND MATERIAL

The classifier has been applied in a natural vegetation zone with a surface of 50200 hectares, situated in the north-east of Catalonia (Spain). The legend consists of eleven categories: *Quercus ilex*, *Fagus sylvatica*, *Freaxinus sp.*, High Mountain Shrub, Mediterranean Shrub, Grass, *Pinus uncinata*, *Pinus sylvestris*, *Pinus pinaster*, *Quercus canariensis*, *Quercus humilis*.

A total amount of 34 variables have been introduced in the classification. These include radiometric bands (all Landsat 7 ETM+ or Landsat 5 TM radiometric bands except thermal and panchromatic channels for 12-03-2003, 29-04-2003, 13-06-2002, 16-08-2002), NDVI indexes for every used date, and topographic and climatic variables (slope, annual average precipitation, annual average solar radiation, minimum average temperature for January and April, maximum average temperature for July). The climatic variables have been obtained from the Digital Climatic Atlas of Catalonia [10]. All variables have been normalized to have mean 0 and standard deviation 1.

A set of IsoMM classification using different values for the "NoData tolerance" parameter have been carried out, as well as a Maximum Likelihood classification to compare the results of both classification methods. Both classification methods have been repeated after the extraction of the March image, which is containing a large NoData zone.

The realization of this work wouldn't be possible without the grant received *Ministerio de Ciencia y Tecnologia* and the funds of *FEDER* through the research project TIC2003-08604-C04.

IV. RESULTS

The classification results in reference to classification surface are presented in Fig. 1. The Unsupervised Classified Surface (UCS) and the Final Map Classified Surface (FMCS) are differentiated because of the characteristics of the hybrid methodology. UCS corresponds to the results in the first phase, that is to say the classified pixels by IsoMM (resulting in an unsupervised map formed by statistical classes). Obviously, UCS has no sense in the Maximum Likelihood classification. For the hybrid methodology, FMCS corresponds to the second and final phase, the classified pixels by ClsMix (giving as a result a thematic class map). FMCS can be represented for the Maximum Likelihood classification. As no probability thresholds have been set, all available pixels will be labeled to a thematic class (and that surface can be compared to the ClsMix classified surface).

The higher the number of NoData variables are tolerated by IsoMM to classify a pixel, the more surface is classified in the unsupervised phase (UCS line for the columns labeled with IsoMM* in Fig. 1). Most of the pixels with NoData values have these non valid values in seven or less variables (there are seven variables for each date). Specifically, the March image contains a considerable part of the image as NoData. That explains the important increase in classified surface and the posterior plateau behavior when tolerating more than seven variables. IsoMM is able to classify 70.9% of the total scene surface when the common NoData treatment is used (no tolerance to NoData) and 97.6% when seven variables with NoData are accepted when classifying a pixel. In the most permissive option (tolerating 33 variables of NoData) IsoMM classifies the entire scene.

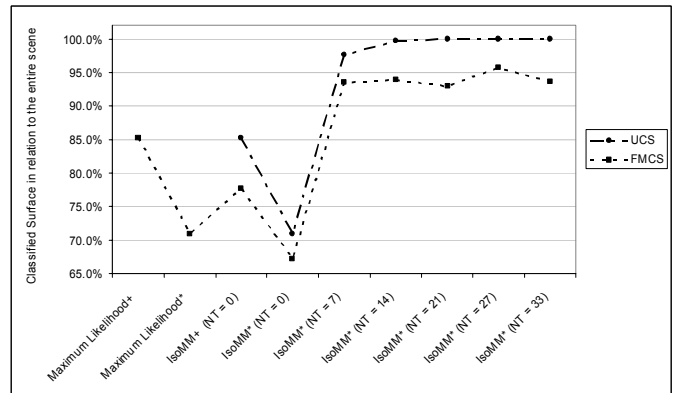


Figure 1. Results of the classification on the classified surface. (NT: Number of variables with NoData tolerated; UCS: Unsupervised Classified Surface; FMCS: Final Map Classified Surface; the executions marked with * have used all the available images and in those marked with + the March image has been extracted.)

These achievements can be applied to the statistical class to thematic class stage (represented in the Fig. 1 by the FMCS line) where the relation between classified surface and number of NoData variables tolerated is parallel to the behavior previously mentioned. In this second phase, it's a common fact that some statistical classes remain unclassified due to confusions in the assignment decision or because these statistical classes don't match with any training area. That is why the Final Map Classified Surface is always lower than the

Unsupervised Classified Surface. Anyway, the classifier with a common NoData treatment in the unsupervised phase will be able to label 67.2% of the total scene surface, while tolerating seven NoData variables is able to classify 95.6%.

The Maximum Likelihood classification is heavily conditioned by the presence of NoData. The available pixels will be those having the entire vector of variables with valid information. That is the reason why FMCS for Maximum Likelihood* (using all the dates) equals the UCS value for IsoMM* when NoData tolerance is 0 (Fig. 1). If the March image is extracted, there will be less pixels affected by NoData values and that will permit an increase in UCS for a common unsupervised classification (IsoMM⁺ with NoData tolerance 0 in Fig. 1) as well as an increase in FMCS for Maximum Likelihood classification (Maximum Likelihood⁺ in Fig. 1). Anyway, the use of IsoMM with all the dates and tolerating NoData values permits classifying more surface than a common strategy whether unsupervised method, Maximum Likelihood classification or extracting the more problematic image.

Another success of the procedure is on the thematic accuracy (Fig. 2). The amount of unclassified pixels provokes a low thematic accuracy when using Maximum Likelihood (62.9%) and IsoMM (58.6%) with all dates (Maximum Likelihood* and IsoMM* with NoData tolerance 0 respectively in Fig.2). Extracting the March image produces an increase in the thematic accuracy (from 58.6% to 62.4%) in IsoMM classification because of the increase in the amount of classifiable pixels. For the Maximum Likelihood, the great increase in the classified surface produced by the extraction of March image (from 70.9% to 85.2%) translates in a slight increase in thematic accuracy (from 62.9% to 67.3%). Anyway, the use of IsoMM with all dates and tolerating NoData values maintains the best accuracy level (between 87.8% and 88.6%).

V. CONCLUSIONS

A hybrid classifier has been presented in order to obtain high accuracy vegetation maps using temporal series of satellite images. A particular treatment of NoData values classifies a pixel despite of the missing of some of the variables in the classification. That permits classifying a larger surface of the scene in spite of the cloud, snow or radiometric problems presence in some of the dates used and, at the same time, it avoids turning down some valuable images (phenologic information) having part of the image occupied with NoData.

The results presented in this work permits classifying a surface up to 40% greater than using a common methodology (unsupervised classifier with no tolerance to NoData, a Maximum Likelihood classifier, or the rejection of a problematic date) and achieving a clearly greater accuracy (88.6% for the best hybrid classifier result versus 67.2% for the best Maximum Likelihood result).

ACKNOWLEDGMENT

The realization of this work wouldn't be possible without the grant received *Ministerio de Ciencia y Tecnología* and the funds of *FEDER* through the research project TIC2003-08604-C04. We thank *Agència Catalana de l'Aigua* and *Departament de Medi Ambient i Habitatge de la Generalitat de Catalunya* their investment policy and Remote Sensing data availability. We want to thank to our colleagues from Geography Department of UAB and CREAM who have cooperated in the image processing, as well as INTA for their considerate in the imagery subscription service.

REFERENCES

- [1] J.A. Richards, *Remote sensing digital image analysis, an introduction*, Berlin: Springer-Verlag, 1993.
- [2] M. Soshany, "Satellite remote sensing of natural Mediterranean vegetation: a review within an ecological context", *Prog. Phys. Geogr.*, vol. 24, 2000, pp. 153-178.
- [3] O.C. Jensen and A.G. Sánchez-Azofeifa "Satellite-derived ecosystems classification: image segmentation by ecological region for improved classification accuracy, a boreal study case", *Int. J. Remote Sens.*, vol. 27, pp. 233-251, 2006.
- [4] O. Viñas and X. Baulies, "1:250 000 Land-use map of Catalonia (32 000 km²) using multitemporal Landsat-TM data", *Int. J. Remote Sens.*, vol. 16, pp. 129-146, 1995.
- [5] W.B. Cohen and S.N. Goward, "Landsat's role in ecological applications of remote sensing", *Bioscience*, vol. 54, pp. 535-545.
- [6] X. Pons and L. Solé-Sugrañes, "A simple radiometric correction model to improve automatic mapping of vegetation from multispectral satellite data", *Remote Sens. Environ.*, vol. 48, pp. 191-204, 1994.
- [7] R.D. Duda and P.E. Hart, *Pattern Classification and Scene Analysis*, New York: John Wiley & Sons, 1984.
- [8] R.M. Lark, "A reappraisal of unsupervised classification. I: correspondence between spectral and conceptual classes", *Int. J. Remote Sens.*, vol. 16, pp. 1425-1443, 1995.
- [9] P. Serra, X. Pons and D. Saurí, "Post-classification change detection with data from different sensors: some accuracy considerations", *Int. J. Remote Sens.*, vol. 24, pp. 3311-3340, 2003.
- [10] M. Ninyerola, X. Pons, J.M. Roure, "A methodological approach of climatological modelling of air temperature and precipitation through GIS techniques", *Int. J. Climatol.*, vol. 20, pp. 1823-1841, 2000.

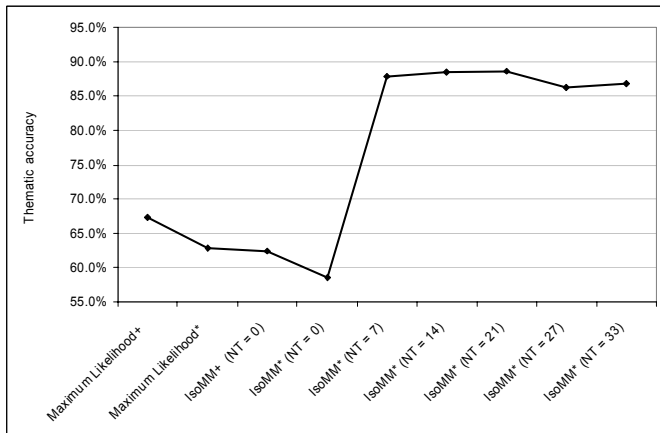


Figure 2. Results of the classification on the global thematic accuracy. (NT: Number of variables with NoData tolerated; the executions marked with ⁺ have used all the available images and in those marked with ^{*} the March image has been extracted.)