

## Clasificador para series de imágenes destinadas a la obtención de mapas detallados de vegetación y cultivos.

X. Pons<sup>(1, 2)</sup>, G. Moré<sup>(2)</sup> y P. Serra<sup>(1)</sup>

<sup>(1)</sup> Departament de Geografia. Edifici B. Universitat Autònoma de Barcelona, E-08193 Bellaterra. Xavier.Pons@uab.es, Pere.Serra@uab.es.

<sup>(2)</sup> Centre de Recerca Ecològica i Aplicacions Forestals (CREAF). Edifici C. Universitat Autònoma de Barcelona, E-08193 Bellaterra, Barcelona. g.more@creaf.uab.es.

### Resumen

La obtención de datos de fenología de la vegetación natural y los cultivos resulta imprescindible para obtener mapas detallados (más de 20 categorías de esta temática) a partir de clasificación de imágenes. Para dicho fin se requiere del uso de múltiples fechas de adquisición. Sin embargo, esta variedad de fechas suele conllevar la presencia de sombras, nubes, nieves, etc, que acarrear la pérdida de una importante superficie clasificable. En esta comunicación se presenta un clasificador no supervisado (IsoMM) que permite un mejor aprovechamiento de las imágenes de una serie temporal gracias a un particular tratamiento de los valores NoData. El nuevo algoritmo ha sido aplicado a la obtención de mapas de zonas forestales y agrícolas de varios miles de ha, con una leyenda de más de 25 categorías en dichas zonas usando imágenes Landsat. En este escenario IsoMM permite clasificar hasta un 40% más de superficie respecto al tratamiento habitual, manteniendo un acierto global del 90%.

### 1. Introducción

El disponer de una serie continua de imágenes Landsat (adquisición de imágenes cada 16 días) permite utilizar la diferente fenología de la vegetación como un potente criterio discriminante para obtener clasificaciones muy detalladas temáticamente. A su vez, y a diferencia de otros sensores en plataformas que ofrecerían una resolución temporal aún mayor pero una pobre resolución espacial, Landsat permite obtener cartografía con un detalle geométrico de unos 30 m abarcando extensiones razonablemente grandes (32400 km<sup>2</sup>). A la hora de escoger el método de clasificación resulta muy difícil mantener con éxito un método supervisado cuando se pretende conseguir una leyenda detallada de vegetación natural y cultivos; ello es debido al hecho que las clases derivadas de áreas de entrenamiento serán, en el escenario descrito, difícilmente asimilables a

clases estadísticas que un clasificador supervisado pueda tratar con el rigor adecuado (clases multimodales por presencia de varias situaciones fenológicas simultáneas en la misma escena, etc). Ello lleva a dejar el proceso en manos de un clasificador no supervisado que se enfrente a la tarea de encontrar las clases estadísticas realmente presentes en la imagen y que puedan ser reasignadas posteriormente a clases temáticas.

Cuando se entregan al clasificador las imágenes disponibles es habitual que una parte muy considerable de ellas no sea realmente utilizable, de modo que, a la práctica, se está desperdiciando una cantidad importante de la información contenida en la serie temporal [1]. Este rechazo es debido a que muchas imágenes contienen abundantes zonas no válidas (NoData). Dichas zonas NoData provienen principalmente de dos fuentes. Por un lado, la presencia de nubes y nieblas y zonas innivadas, provoca que un elevado número de imágenes no sea utilizable, o sólo lo sea si se aplica una máscara para excluir dichas zonas. Por otro lado, las imágenes recibidas se suelen someter a una corrección radiométrica que reduce los artefactos indeseados debidos a los efectos de la atmósfera o a la iluminación diferencial causada por la hora del día, el lugar de la Tierra o el relieve [2]; sin embargo, en regiones montañosas este proceso invalida ciertas zonas por considerarlas de radiometría poco fiable (sombras por ocultamiento del relieve circundante o por autoocultamiento).

Como hemos dicho, dado que la intersección de zonas NoData de las diferentes imágenes potencialmente utilizables (cada 16 días) resulta a menudo en una superficie muy considerable y que los clasificadores implementados en los paquetes *software* convencionales excluyen de la clasificación cualquier pixel que contenga NoData para alguna de las variables, se tiende a desechar las imágenes con valores NoData. Ello evita obtener imágenes con amplias zonas no clasificadas por nubes, nieves o problemas radiométricos en alguna fecha pero conlleva una importante pérdida de potencial de clasificación, dando lugar a resultados

más pobres en el nivel de acierto o a conformarse con leyendas más simples, no tan sensibles a la disponibilidad de menos fechas (menor información fenológica). Sin embargo, el hecho que una zona de la imagen sea declarada NoData no debería impedir aprovechar otras zonas de la misma imagen con una información radiométrica de suficiente calidad.

Este trabajo expone un clasificador no supervisado que presenta, entre otras ventajas, la capacidad de sortear los anteriores problemas y clasificar píxeles aunque no dispongan de valores útiles en todas las variables. El clasificador no supervisado aquí presentado, denominado IsoMM, se entiende como parte de un clasificador mixto o híbrido: Posteriormente a la etapa no supervisada, las clases estadísticas serán reasignadas a las clases temáticas definidas por el usuario mediante áreas de entrenamiento, fase que aquí no se describe en detalle y que corresponde al algoritmo ClsMix [3].

## 2. Consideraciones sobre IsoMM

El algoritmo base utilizado se inspira en IsoData [4], procedimiento iterativo que, mediante un criterio de mínima distancia, agrupa (formando *clusters*) los píxeles de una escena en diferentes clases; dichas clases han sido inicialmente ubicadas según algún criterio adecuado (centros en posición aleatoria, diagonal multivariante, etc). En cada iteración las características de las clases son de nuevo recalculadas teniendo en consideración los valores de todos los píxeles que se han incorporado a cada clase durante esa iteración. El proceso finaliza cuando las asignaciones de los píxeles a las clases llegan a una situación de estabilidad satisfactoria, o bien cuando se alcanza un número máximo de iteraciones, siendo ambos criterios fijados *a priori* por el usuario. La implementación aquí presentada ha ido modificando las fases del algoritmo original para dar solución a los problemas habitualmente presentados en las clasificaciones detalladas de vegetación natural y cultivos.

A continuación se comentará brevemente los rasgos básicos del algoritmo y, con mayor detalle, el particular tratamiento de valores NoData durante la clasificación.

### 2.1. Características básicas del clasificador

- Varios centenares de variables admitidas, con lo que se permite utilizar series temporales de imágenes de satélite con alta resolución temporal así como otras variables que ayuden a discriminar las categorías de interés (variables de carácter climático, relieve, etc.).

- Se admiten formatos de tipo entero (1 y 2 *bytes*) y de tipo real (4 *bytes*).
- Un máximo de 32767 *clusters* admitidos.
- Criterios de distancia euclidiana y Manhattan.
- Eliminación de *clusters* estadísticamente cercanos según un umbral configurado.
- Fin de iteraciones por umbral de convergencia o por número máximo de iteraciones definidos por el usuario.
- Tres estrategias distintas para la distribución de las semillas iniciales: 1/ A lo largo de la diagonal multivariante de los datos, repartidas de forma equidistante; 2/ Aleatoriamente en todo el espacio multivariante; 3/ En base a un muestreo sobre la imagen.

### 2.2. Tratamientos de valores NoData

A menudo los algoritmos de clasificación de imágenes dan poca o ninguna importancia al tratamiento de los valores NoData (también citados como valores no válidos, o *missing values*). En los *software* menos cuidadosos los valores NoData se tratan como valores ordinarios, dando lugar a inaceptables resultados. En los casos en que sí se considera su especial naturaleza, la solución más utilizada es la de considerar un píxel como NoData cuando éste contiene por lo menos un valor NoData en su conjunto de variables utilizado.

Sin embargo, la utilización masiva de imágenes de satélite (alta resolución temporal) implica una notable presencia de valores NoData por lo expuesto en la Introducción. Estos valores NoData se distribuyen de forma diferente en función de la fecha en la que fue tomada la imagen desde el satélite. Por ejemplo, en una clasificación donde intervienen 34 variables podemos tener un píxel que contenga valores NoData en solamente cinco de las variables introducidas. Sin embargo, es posible que, por este determinado píxel, las 29 variables restantes con información útil sean suficientes para su correcta clasificación.

El uso de núcleos clasificadores débilmente paramétricos (distancia euclidiana o Manhattan), que no utilizan la información sobre dispersión estadística multivariante, facilita que los cálculos de distancia entre píxel y centro de *cluster* utilicen conjuntos de variables diferentes en función de las variables disponibles para cada píxel.

En la implementación de IsoMM el usuario puede determinar cuántas variables con NoData (sin información) tolera en el momento de clasificar un

píxel. El rango de opciones va desde 0 (equivalente al antes comentado tratamiento estadístico habitual) hasta un valor equivalente al número total de variables menos uno (la opción menos restrictiva, capaz de clasificar en base a una sola variable).

Este tratamiento permite utilizar escenas que rechazaríamos con el tratamiento habitual (normalmente imágenes con un moderado recubrimiento de nubes y/o sombras) al mismo tiempo que ampliamos el área clasificada (podemos llegar a clasificar zonas en las que no se dispone de datos válidos por alguna de las fechas utilizadas).

Para garantizar la consistencia estadística del proceso, los cálculos estadísticos de los centros de clase se realizan únicamente a partir de aquellos píxeles que contengan todo el vector de variables con valores útiles (es decir, totalmente sin NoData).

### 3. Área de estudio y material

Los resultados expuestos en este trabajo se han obtenido en una zona de estudio con unas 50 200 ha de superficie con vegetación natural, situada en la Alta Garrotxa (noreste de Cataluña).

En el proceso de clasificación no supervisada se han introducido un total de 34 variables, donde se incluyen variables radiométricas (todas las bandas excepto el canal térmico de las imágenes Landsat 7 ETM+ con fecha 29-04-2003, 12-03-2003, 13-06-2002, 16-08-2002), el índice NDVI para cada una de las fechas utilizadas, y otras variables de carácter topoclimático (pendiente del terreno, precipitación media anual, radiación solar media anual, temperatura media de las mínimas en los meses de enero y abril, y temperatura media de las máximas del mes de julio). Todas las variables han sido normalizadas (media 0 y desviación estándar 1) antes de ser introducidas en IsoMM.

Las áreas de entrenamiento se han obtenido a partir del procesamiento del Mapa de Hábitats de Cataluña tal y como se describe en [5].

### 4. Resultados y conclusiones

En la figura 1 se resume gráficamente el efecto del parámetro de tolerancia de variables con NoData influye en los resultados de la clasificación.

Observamos que, a medida que toleramos un mayor número de variables, IsoMM consigue clasificar una mayor superficie de la zona de estudio (gráfica IS). La gran mayoría de píxeles que contienen algún valor NoData tiene estos valores en siete o menos variables. Esto explica el gran salto y la posterior estabilización que se produce al pasar del tratamiento habitual (0 variables con NoData toleradas), donde se clasifica el 70.85% de la

escena, al caso de tolerar 7 variables con NoData, con un 97.62% de escena clasificada.

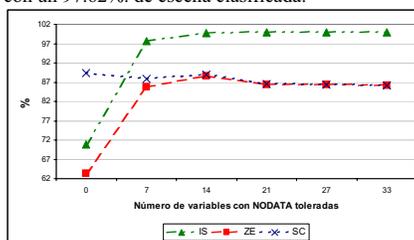


Figura 1: Efecto de la tolerancia a variables con NoData en una clasificación. (IS= Superficie clasificada por IsoMM respecto al área de la escena; ZE= Sup. clasificada por ClsMix respecto al área de la escena; SC= Sup. clasificada final respecto a la sup. clasificable).

Las mejoras también son evidentes en la segunda fase del proceso (reasignación de clases espectrales a temáticas con ClsMix). En efecto, la superficie clasificada por ClsMix respecto a la zona de estudio (gráfica ZE) sigue una tendencia paralela a la anterior, donde con el tratamiento típico de valores NoData sólo conseguimos clasificar 63.4% de la escena mientras que con una tolerancia de 7 variables se consigue clasificar alrededor del 85.7% (prácticamente el 90% en el mejor de los casos, con 14 variables toleradas).

Otro éxito del proceso es que el acierto global de la clasificación (88.7, 90.3, 90.1, 89.8, 89.9, 90.0%, respectivamente) y la superficie que es capaz de clasificar ClsMix respecto el área clasificada por IsoMM (gráfica SC) no quedan alterados negativamente por el incremento de superficie clasificada al tolerar píxeles con el vector de variables incompleto.

A modo de ejemplo, presentamos la clasificación correspondiente a la clasificación con una tolerancia de 14 variables con NODATA (figura 2). El acierto global conseguido es del 90.1% sin considerar los píxeles no clasificados y del 85.9% contando los píxeles sin clasificar. En la tabla 2 se muestra las categorías de la leyenda y sus respectivos aciertos del usuario (columna Usu) y del productor (columna Prod), así como la superficie que ocupan dentro de la clasificación (columna Sup).

Tabla 2: Categorías de la leyenda de clasificación, exactitud y superficie

Categorías		Exactitud (%)		Sup. (%)
Código	Descripción	Usu.	Prod	
FSYL	<i>Fagus sylvatica</i>	89.9	89.3	22.8
QHUM	<i>Quercus humilis</i>	87.4	74.9	21.7
PSYL	<i>Pinus sylvestris</i>	94.0	93.7	14.1
QILE	<i>Quercus ilex</i>	96.6	95.9	13.0
MAM	Matorrales alta montaña	83.3	84.8	6.7
PBS	Prados basófilos	65.0	62.3	4.6
PAM	Prados alta montaña	63.7	84.2	2.6
FEXC	<i>Fraxinus excelsior</i>	70.7	28.7	2.1
PAS	Prados acidófilos	81.5	40.1	1.3
MM	Matorrales mediterráneos	60.1	65.8	0.6
PUNC	<i>Pinus uncinata</i>	86.1	68.4	0.5
QCAN	<i>Quercus canariensis</i>	24.4	7.3	0.2
PPIN	<i>Pinus pinaster</i>	100	1.4	0.1

Podemos observar como aquellas categorías más representativas en el área de estudio tienen unos aciertos de usuario y de productor satisfactorios. Del análisis de la matriz de confusión se deduce que las categorías con menores aciertos (y también más minoritarias) acumulan predominantemente las confusiones con categorías similares (confusiones entre categorías de prados, entre pinos y entre caducifolios respectivamente).

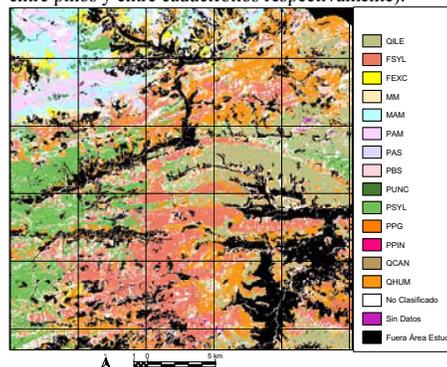


Figura 2: Alta Garrotxa. Imagen clasificada tolerando 14 variables con NoData (consultar tabla 2 para descripción del código de leyenda).

## 5. Conclusiones

IsoMM es un clasificador no supervisado inspirado en IsoData y pensado para la clasificación de vegetación utilizando series temporales de imágenes de satélite.

Un particular tratamiento de los valores NoData permite clasificar un determinado píxel a pesar de

contener algunas de las variables como no válidas, evitando así la pérdida de variables que serían rechazadas y permitiendo clasificar una mayor superficie. Las primeras pruebas dan resultados muy positivos: Se clasifica un 40% más de superficie respecto el tratamiento habitual manteniendo unos aciertos cercanos al 90% y con leyendas de más de 10 categorías.

## 6. Agradecimientos

La realización de este trabajo no hubiera sido posible sin la financiación aportada por el Ministerio de Ciencia y Tecnología y fondos FEDER mediante el proyecto de investigación: "Compresión wavelet de imágenes para aplicaciones de Teledetección y SIG. 4 implicaciones en el proceso digital de imágenes" (TIC2003-08604-C04-03). Expresamos nuestro agradecimiento a la Agència Catalana de l'Aigua y al Departament de Medi Ambient i Habitatge de la Generalitat de Catalunya por su política de inversión y disponibilidad de datos de Teledetección, la cual ha permitido abordar este estudio con plenas garantías. Deseamos expresar también nuestro agradecimiento a los compañeros del Departament de Geografia y del CREAL que han colaborado en el tratamiento de las imágenes y al INTA por su atento servicio de suscripción de imágenes.

## 7. Referencias

- [1] Cihlar, J., "Land cover mapping of large areas from satellites: status and research priorities", *Int J Remote Sens*, 21, 1093-1114, 2000.
- [2] Pons, X. y Solé-Sugrañes, L., "A Simple Radiometric Correction Model to Improve Automatic Mapping of Vegetation from Multispectral Satellite Data", *Remote Sens Environ*, 47:1-14, 1994.
- [3] Serra, P., Pons, X. y Sauri, D., "Post-classification change detection with data from different sensors: some accuracy considerations", *Int J Remote Sens*, 24, 3311-3340, 2003.
- [4] Duda, R.D. and Hart, P.E., *Pattern Classification and Scene Analysis*, John Wiley & Sons, New York, 1973.
- [5] Moré, G., Burriel, J.A., Castells, R., Ibáñez, J., Roijals, X., "Tratamiento estadístico de variables radiométricas, orográficas y climáticas para la obtención de un mapa detallado de vegetación", *Proceedings of MCSIGT 2004, Murcia, 20-23 September 2004*.