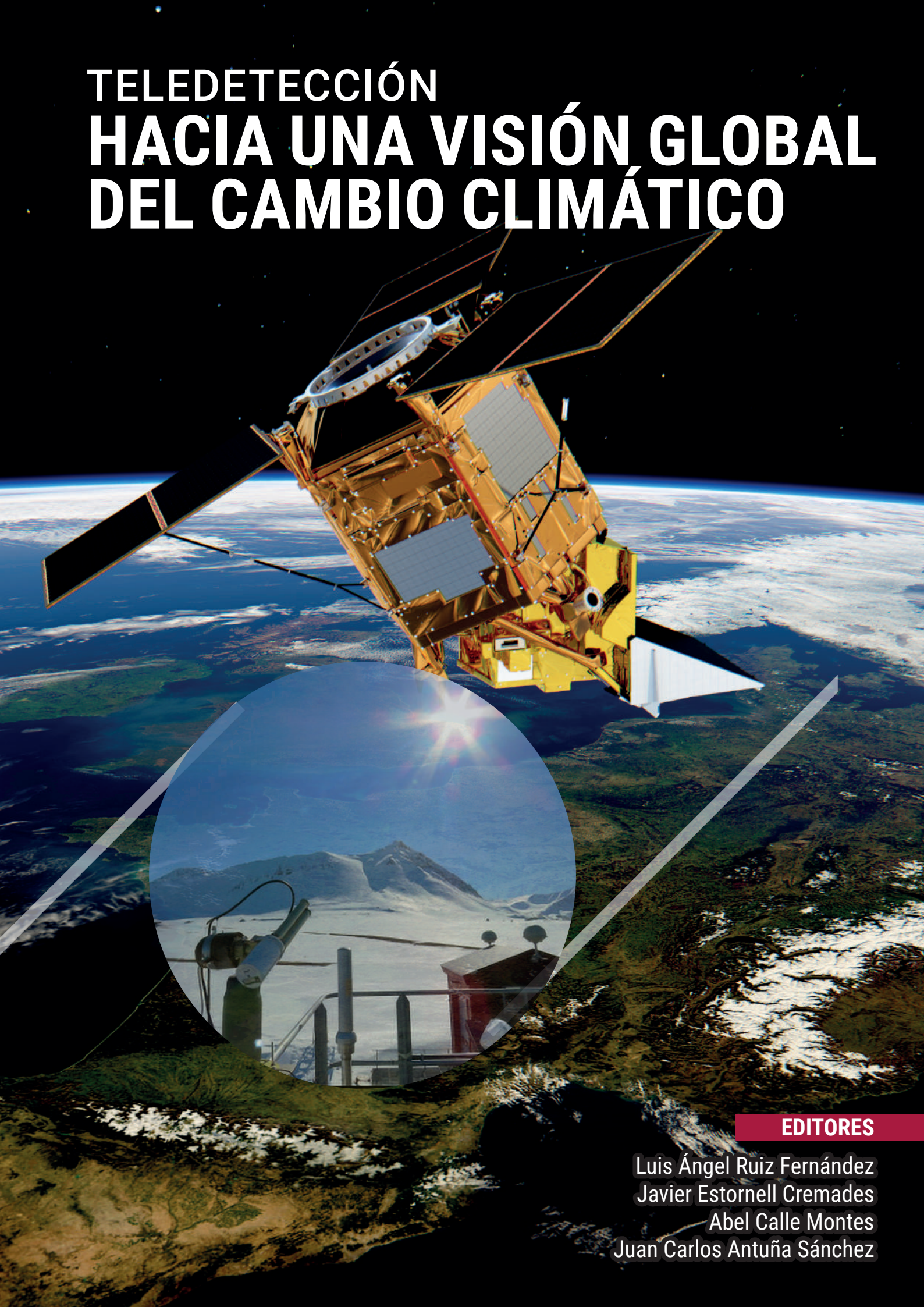
TELEDETECCIÓN

HACIA UNA VISIÓN GLOBAL DEL CAMBIO CLIMÁTICO



EDITORES

Luis Ángel Ruiz Fernández
Javier Estornell Cremades
Abel Calle Montes
Juan Carlos Antuña Sánchez

Teledetección: hacia una visión global del cambio climático
(Eds. Luis A. Ruiz, Javier Estornell, Abel Calle y Juan Carlos Antuña-Sánchez). pp. 3-6. 2019. ISBN: 978-84-1320-038-5

Comparación de árboles de decisión CART y bajo criterio experto en la clasificación de cultivos de una zona de riego extensa

Guillén, M.[(1)], Rodrigues, M.[(2,3)], Febrer, M.[(3,4)], Casterad, M.A.[(1)]
[(1)] Unidad de Suelos y Riegos (Asociada al CSIC). Centro de Investigación y Tecnologías Agroalimentaria de Aragón, Avda. Montañana 930, 50059-Zaragoza, España. mguillenc@aragon.es
[(2)] Dpto. de Ingeniería Agroforestal, Univ. de Lleida, Pl. de Víctor Siurana, 1, 25001-Lleida, España.
[(3)] Grupo GEOFOREST-IUCA, Dpto. de Geografía y Ordenación del Territorio Univ. de Zaragoza. C/ Pedro Cerbuna, 12, 50009-Zaragoza, España.
[(4)] remOT Technologies S.L., C/ Mariano Esquillor Gómez s/n 50018-Zaragoza, España.

Resumen: En las comunidades de regantes, los gestores demandan estrategias que permitan optimizar el uso del agua. Disponer de información sobre las superficies de cultivos y sus necesidades hídricas es esencial para el manejo eficiente del agua. Las imágenes de satélite son útiles para este fin, gracias a las ventajas que ofrece la información espectral (generación de índices de vegetación) y a la disponibilidad de técnicas como clasificación supervisada, uno de los métodos más utilizados en la clasificación de cultivos. En este trabajo se compara y evalúa la utilización de árboles de decisión CART y árboles de decisión bajo criterio experto (TREE) en la clasificación de tipologías de cultivos en la Zona Regable del Canal de Aragón y Cataluña. Esta clasificación se realiza a partir de información multitemporal de NDVI derivado de imágenes Landsat 8 en 6 fechas distintas a lo largo de la campaña de riegos de 2016. Los dos modelos proporcionan una muy buena clasificación, mejorando CART ($\hat{F}$=0,91) la fiabilidad global de TREE ($\hat{F}$ =0,84).

Palabras clave: CART, Landsat 8, Clasificación digital, NDVI.

Comparing CART and expert criteria decision trees in the classification of crops in a large irrigation area

Abstract: *Water user associations' managers demand strategies for water use optimization. Knowledge on irrigated croplands and their specific hydric requirements is essential to achieve an efficient use of water resources. Remote sensing imagery is particularly useful to this end, due to the benefits offered by spectral information (i.e., vegetation indices) coupled to techniques such as supervised classification, one of the most widely used methods for cropland classification. In this work, we compare the CART algorithm (supervised tree-learning) and expert criteria decision trees (TREE). We compared and evaluated their classification performance for various crop types in the Irrigation Area of Canal de Aragón and Cataluña. Both classification approaches were trained from multi-temporal NDVI data from Landsat 8 imagery measured at six different dates throughout the irrigation season 2016. Both models denoted a very good classification performance, although global accuracy was higher in CART ($\hat{F}$=0.91) compared to TREE ($\hat{F}$=0.84).*

Keywords: *CART, Landsat 8, Digital classification, NDVI.*

1. INTRODUCCIÓN

El conocimiento de la superficie de cultivos es clave en la gestión de zonas regables. Como es sabido, las imágenes de satélite son una herramienta muy útil para la identificación y seguimiento de los cultivos. Además, están adquiriendo gran protagonismo en la mejora e innovación de la gestión del uso del agua con su integración en los nuevos modelos y estrategias de gestión adoptados por Comunidades de Regantes, Confederaciones Hidrográficas, Organismo públicos,…

La clasificación supervisada, uno de los métodos más empleados en la clasificación de cultivos, está dando paso en los últimos años a los métodos basados en algoritmos "*Machine Learning"* que se están aplicando en teledetección para la identificación de cultivos con buenos resultados (Rodríguez-Galiano y Chica-Rivas, 2012; Toro et al., 2015; Larrañaga y Álvarez-Mozos, 2016; Gomáriz-Castillo et al., 2017; Sitonkostantinou et al., 2018).

En este trabajo se compara y evalúa desde un punto de vista metodológico y predictivo la aplicación de árboles de decisión CART y árboles de decisión bajo criterio experto en la clasificación de tipologías de cultivos en la Zona Regable del Canal de Aragón y Cataluña (CAyC). Se elige esta área de estudio por su gran extensión, la diversidad de cultivos y sistemas de riego que presenta, la experiencia previa de monitorización con teledetección de la superficie cultivada en la zona, así como por el interés en mejorar y agilizar la clasificación de cultivos en la zona.

2. MATERIAL Y MÉTODOS

La CAyC, situada en la margen izquierda de la Cuenca del Ebro entre las provincias de Huesca y Lérida, es una de las más importantes de España. Cuenta con unas 105 000 hectáreas, de las cuales, el 60% están en Aragón y el 40% restante en Cataluña. Aproximadamente la mitad de la superficie se riega por aspersión (cobertura fija y pivotes) y la otra mitad por gravedad y goteo, 27 y 23% respectivamente.

Predominan los frutales, tanto de hueso como de pepita, y los cultivos extensivos entre los que destacan

los cereales de invierno, trigo y cebada principalmente; cereales de verano, predominando el maíz; y forrajeras, principalmente alfalfa. Las dobles cosechas han ido ganando terreno en la parte central de la CAyC siendo cebada-maíz la más habitual.

2.1. Obtención de la muestra de cultivos para la clasificación

La muestra de cultivos para la clasificación se obtuvo de la información no sensible sobre cultivos de las declaraciones de la PAC de 2016 (año de estudio) recopilada en el Sistema de Información Geográfica para Parcelas Agrarias (SIGPAC). Esta información fue suministrada por el Departamento de Desarrollo Rural y Sostenibilidad del Gobierno de Aragón al CITA, quién combinó las bases de datos de las declaraciones con el archivo vectorial SIGPAC que incluye el parcelario de 2016 (recintos SIGPAC) actualizado y georreferenciado. Esta capa en formato vectorial fue la utilizada para la selección de parcelas para el modelado.

La experiencia de anteriores clasificaciones en la zona de estudio llevó a establecer las siguientes categorías de cultivos a identificar: 1-Alfalfa, 2-Arroz, 3-Doble cosecha, 4-Extensivos de invierno, 5-Extensivos de verano 6-Leñosos y 7-Sin cultivo. Ello requirió unificar bajo una misma categoría cultivos diferentes.

Las parcelas elegidas y utilizadas en el modelado de clasificación se dividieron en dos grupos: conjunto de parcelas para entrenar el clasificador (Parcelas_training, 70%) y conjunto de parcelas para su validación (Parcelas_test, 30%). El número de parcelas seleccionadas es variable en cada categoría (Tabla 1) pues la representación (superficie) de cada cultivo en el área de estudio no es homogénea. Todas las parcelas elegidas cumplen el requisito impuesto de abarcar al menos una superficie equivalente a 10 píxeles de 30×30 m, tamaño de píxel de las imágenes utilizadas en la clasificación. Para eliminar el efecto borde y disminuir posibles errores asociados en el proceso de clasificación, se aplicó un buffer de 60 m hacia el interior de las parcelas, equivalente a 2 píxeles.

Tabla 1. *Parcelas seleccionadas como entrenamiento (training) y para validación (test).*

	Total Parcelas		Parcelas_ training		Parcelas_test	
Clases	Nº	ha	Nº	ha	Nº	ha
Alfalfa	304	1477,91	213	994,56	91	483,35
Arroz	1	2,25	--	--	--	--
Doble Cosecha	282	1268,84	197	871,78	85	397,06
Ext. Invierno	209	922,9	146	692,66	63	230,24
Ext. Verano	183	1061,35	128	763,66	55	297,69
Leñoso	268	1063,54	188	753,23	80	275,58
Sin Cultivo	26	104,49	18	70,30	8	34,19
TOTAL	1272	5901,28	890	4146,19	382	1752,84

2.2. Imágenes utilizadas en la clasificación

Se usaron 6 imágenes Landsat 8, nivel de procesamiento L1T, descargadas del servidor *United States Geological Survey* (USGS https://earthexplorer.usgs.gov/).

Posteriormente se corrigieron atmosféricamente por el método del objeto oscuro y los valores de reflectividad se obtuvieron según las indicaciones que proporciona USGS. Las imágenes elegidas corresponden a la campaña de riego que se extiende desde marzo hasta primeros de octubre. Se intentó tener una imagen por mes, completando la serie con una más de finales de junio, periodo de transición de cultivos de invierno a cultivos de verano. La presencia de nubosidad limitó la disponibilidad de imágenes en primavera. Las fechas escogidas fueron 30 de marzo, 2 de junio, 25 de junio, 4 de julio, 12 de agosto y 6 de septiembre de 2016.

2.3. Clasificación TREE

Esta clasificación consistió en la definición de árboles de decisión bajo criterio experto que se programaron con la herramienta "*Knowledge Enginner*" de Erdas Imagine. Como parámetros para la clasificación se utilizaron los NDVI de 6 imágenes Landsat 8 anteriormente indicadas.

Los diferentes árboles de decisión se plantearon sobre la experiencia adquirida en anteriores clasificaciones realizadas en la Zona Regable en 2013 y 2014 en la zona de estudio y con el apoyo de los mapas de cultivos y ocupación correspondientes a dicho año. Estos árboles de decisión, 7 en total, se aplicaron secuencialmente empezando con aquellos basados en reglas de decisión que asegurasen alto acierto en la asignación de los píxeles a cada categoría. Las reglas de decisión seguidas fueron multicriterio en cada categoría a discriminar. Cada árbol de decisión se aplicó sobre lo que había quedado sin clasificar al aplicar el precedente (Figura 1). Finalmente se clasificaron las superficies no asignadas con los árboles de decisión. Más detalles de la clasificación en Casterad et al. (2016).

Como una de las utilidades de la clasificación es generar el mapa de tipologías de cultivos, se ensayó la aplicación de un filtro de mayoría 3×3 a dicha clasificación, ya que proporciona un mapa de cultivos más limpio visualmente al disminuir los píxeles aislados de cada categoría.

2.4. Clasificación CART

Se trabajó con este clasificador en el software libre estadístico R 3.5 con los paquetes Rpart 4.1-13 (Therneau et al., 2018). Como parámetros para la clasificación se utilizaron las imágenes ráster de NDVI de las seis fechas seleccionadas y las Parcelas-training en formato vectorial para el entrenamiento del clasificador. A cada una de las parcelas se le asignó el valor de la mediana de sus píxeles en cada uno de los parámetros estudiados (valores de NDVI en distintas fechas). Se eligió la mediana sobre la media debido a una mayor estabilidad frente a los outliers, los cuales se puede deber a anomalías, problemas de suelo en una parcela, etc. (Nitze et al., 2012).

El primer paso fue generar y ajustar el modelo de predicción. Para ello se entrenó al modelo con las Parcelas-training. La variable objetivo fue la categoría de cultivo (código de cultivo) y las variables predictoras fueron los valores de NDVI en las 6 fechas seleccionadas.

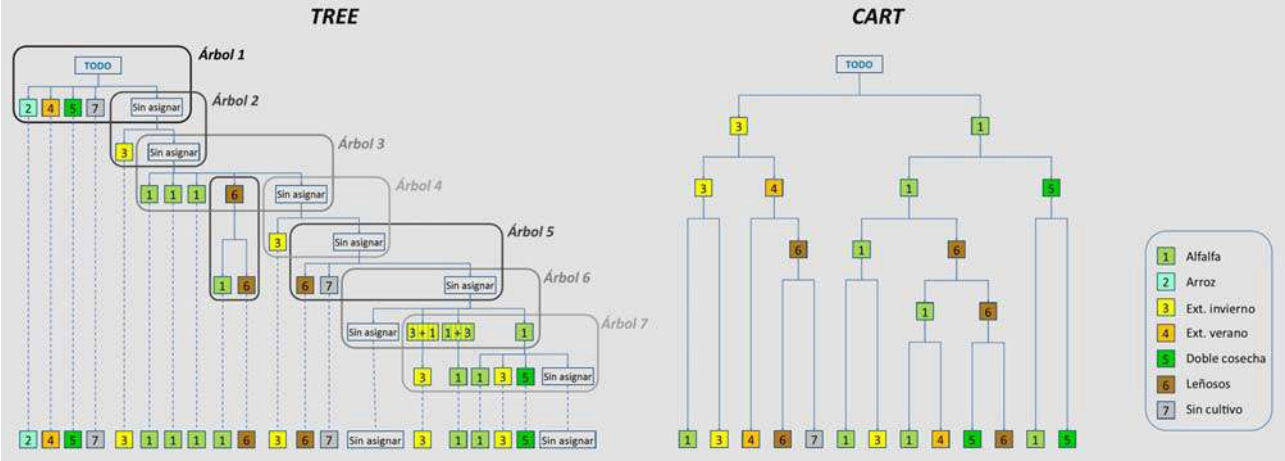


***Figura 1**. Esquema de los modelos adoptados.*

El segundo paso consistió en la aplicación del modelo a los datos ráster para obtener la clasificación de cultivos.

Al igual que en la clasificación TREE se aplicó un filtro de mayoría de 3×3. Las predicciones fueron almacenadas como archivos ráster.

2.5. Verificación de resultados

Para la verificación de las clasificaciones se confrontaron los valores predichos con los reales en las Parcelas_test, comparando la categoría observada con la categoría predicha por el modelo y confirmando su coincidencia o no. La categoría asignada a cada parcela como resultado de la clasificación (valor de predicción) se determinó por el valor de la moda de todos los píxeles de la Parcelas-test.

La validación de los resultados se realizó mediante los estadísticos Fiabilidad global ($\hat{F}$), Fiabilidad del Usuario (*FU*) y del Productor (*FP*), y el índice Kappa (κ).

3. RESULTADOS Y DISCUSIÓN

El modelo CART indica que todas las variables utilizadas en la clasificación, NDVI de las diferentes fechas, son relevantes. Estos resultados entran dentro de lo esperado, pues las fechas utilizadas ya corresponden a una selección previa realizada desde la experiencia en la clasificación de cultivos en la zona de estudio. La variable con mayor contribución es el NDVI de 6 de septiembre y la variable que menor contribución presenta es el NDVI del 2 de junio.

En la Tabla 2 se muestran las matrices de confusión de la clasificación CART y TREE a nivel objeto (Parcelas_test). Para la categoría Arroz, con sólo 0,1% de la superficie total de la Zona Regable, se dispuso únicamente de 1 parcela de muestra, insuficiente para entrenar y validar el modelo en CART. A diferencia de CART, TREE sí permite clasificar categorías con poca representación de verdad terreno, pues las reglas de decisión se establecen de experiencias previas y a partir de umbrales de NDVI sacados del análisis de las curvas de evolución del NDVI de los cultivos en las fechas consideradas.

Por otro lado, CART clasifica todos lo píxeles que se introducen en el modelo y sin embargo TREE deja píxeles sin identificar si no llegan a cumplir todas las condiciones impuestas que se engloban en una categoría denominada. Esto puede ser una ventaja, ya que hay zonas, parcelas, que pueden no ser de ninguna de las categorías a clasificar. Pero también puede ser un inconveniente cuando la superficie que no se identifica es finalmente considerable, como en este caso.

***Tabla 2**. Matrices de confusión de las clasificaciones.*

	NºParcelas	VERDAD TERRENO ALF	DC	EI	EV	LEÑ	SC	TOTAL
CART	Alfalfa	82	5	0	0	4	0	91
	Doble Cosecha	2	79	0	2	0	0	83
	Ext. Invierno	1	1	62	0	0	1	65
	Ext. Verano	2	0	0	52	5	1	60
	Leñoso	4	0	1	1	68	1	75
	Sin Cultivo	0	0	0	0	3	5	8
	TOTAL	91	85	63	55	80	8	382

	NºParcelas	VERDAD TERRENO ALF	DC	EI	EV	LEÑ	SC	TOTAL
TREE	Alfalfa	65	1	0	7	3	1	77
	Doble Cosecha	1	80	0	0	0	1	82
	Ext. Invierno	0	1	61	0	1	1	64
	Ext. Verano	1	0	0	48	0	0	49
	Leñoso	1	0	0	0	61	0	62
	Sin Cultivo	0	0	0	0	6	4	10
	Sin Identificar	23	3	2	0	9	1	38
	TOTAL	91	85	63	55	80	8	382

La $\hat{F}$ = 84% para TREE y $\hat{F}$ = 91% y κ= 0,89 en CART indican que los dos modelos proporcionan una muy buena clasificación. κ no pudo ser calculado para el modelo TREE al no tener una matriz cuadrada, al incluir la categoría *Sin identificar*.

La precisión obtenida es ligeramente superior a lo reportado por Sitonkostantinou et al. (2018), Larrañaga y Álvarez-Mozos (2016), y del Toro et al. (2015), que estudiaron cultivos similares en España. En estos estudios los valores de κ estuvieron entre 0,73 y 0,89 y los de $\hat{F}$ entre 0,79 y 0,86. En cambio son similares a los que obtuvo Rodríguez-Galiano y Chicas-Riva (2012) que lograron con árboles de decisión κ de 0,86 y $\hat{F}$ de 0,85 al clasificar otras ocupaciones del suelo, más generalistas como herbáceos de regadío, herbáceos de secano, suelo desnudo, pastizal olivar, etc.

Todas las categorías, a excepción de *Sin cultivo*, se discriminan con gran exactitud por ambos clasificadores con FU y FP superiores en general al 85% y en la mayoría de los casos al 90%. CART proporciona mejores FP que TREE para 5 de las 6 categorías identificadas. Sin

embargo, la FU es en CART superior en tres categorías a la de TREE mientras que para las otras 3 es superior TREE que CART.

Alfalfa es la categoría que, con diferencia respecto a las otras, más nodos terminales precisa para su discriminación en los dos modelos aplicados. Para esta categoría, la FP obtenida con el clasificador TREE es claramente inferior a la obtenida con CART, donde el 25% de las Parcelas-test de esta categoría quedan sin identificar como tal. En cuanto a FU es, junto con *Sin cultivo*, la única categoría que no supera el 90% con el clasificador TREE confundiéndose principalmente con *Extensivos de Verano.*

Sin cultivo es la que menos nodos requiere y peores fiabilidades obtiene, entre 40 y 63%. Estas fiabilidades pueden estar influenciadas por la baja representación de esta categoría en la zona de estudio, con pocas Parcelas-training y Parcelas-test. Esta categoría se confunde con *Leñosos*, debido en gran medida a que dependiendo de la edad de los frutales el suelo tiene gran peso en la variable independiente que se utiliza en la clasificación (NDVI). Las plantaciones jóvenes son las que mayor confusión provocan. *Leñosos* es, tras la *Alfalfa,* la categoría con más Parcelas-test incluidas en *Sin identificar* (11%) en TREE.

Leñosos es la que más superficie ocupa en la zona de estudio con valores de 37,6% para CART y 23,7% para TREE, seguida de *Doble Cosecha* con valores de 16,2% y 18,8% respectivamente. Las zonas *Sin Cultivo* son las que menos representan con valores de entre 5, 7%.

4. CONCLUSIONES

- Los dos modelos comparados en este trabajo proporcionan una muy buena clasificación, si bien CART ($\hat{F}$=0,91) mejora la fiabilidad global de TREE ($\hat{F}$=0,84).
- Todas las variables elegidas, 6 fechas diferentes de valor NDVI obtenidos de imágenes Landsat 8, son importantes para los modelos, lo que entra dentro de lo esperado ya que estas fechas corresponden a una selección previa hecha de la experiencia.
- Los dos clasificadores coinciden en que la categoría *Alfalfa* es la que a través de más nodos terminales se discrimina, mostrando la complejidad de su clasificación. La categoría *Sin Cultivo* es la que con menos nodos terminales se discrimina y peores fiabilidades consigue.

La comparación de un modelo y otro en relación a su aplicación tal y como se presenta en este estudio revela que:

- El modelo CART es más sencillo de interpretar que TREE y permite identificar de forma más rápida y eficiente las variables más importantes. Además, toma las reglas de decisión de manera objetiva, no así en TREE que depende del criterio experto.
- La metodología TREE permite clasificar aquellas categorías con poca representación de verdad terreno. No necesita de verdad terreno, ya que permite diferenciar clases o categorías espectrales similares a partir de la comparación visual de las distintas firmas. A diferencia de CART, que clasifica todos los píxeles que se introducen el modelo, TREE deja píxeles sin identificar si no llega a cumplir todas las condiciones impuestas.

5. AGRADECIMIENTOS

A la Comunidad General de Regantes del Canal de Aragón y Cataluña por apoyar estos trabajos con dos contratos de colaboración con el CITA (2016 y 2016-2018).

6. BIBLIOGRAFÍA

Casterad, M.A., Portero, C., Gómez, R. 2016. Aplicación de la Teledetección por satélite en la gestión del agua en el Canal de Aragón y Cataluña en 2016. Memoria. http://hdl.handle.net/10532/3980.

Gomariz-Castillo, F., Alonso-Sarría, F., Cánovas- García, F. 2017. Improving classification accuracy of multi-temporal Landsat images by assessing the use of different algorithms, textural and ancillary information for a Mediterranean semiarid area from 2000 to 2015. Remote Sensing 9 (10), 1058.

Larrañaga, A., Álvarez-Mozos, J. 2016. On the added value of quad-pol data in a multi-temporal crop classification framework based on RADARSAT-2 Imagery." Remote Sensing 8 (4), 1–19.

Nitze, I, Schulthess, U., Asche, H. 2012. "Comparison of Machine Learning algorithms Random Forest, Artificial Neural Network and Support Vector Machine to Maximum Likelihood for supervised crop type classification. Proceedings of the 4th GEOBIA, May 7- 9, 2012 - Rio de Janeiro - Brazil. pp. 35-40.

Rodríguez-Galiano, V., Chica-Rivas, M. (2012). Clasificación de imágenes de satélite mediante software libre: nuevas tendencias en algoritmos de inteligencia artificial. XV Congreso Nacional de Tecnologías de la Información Geográfica, Madrid, AGE-CSIC, 19-21 de septiembre de 2012, pp. 19–21.

Sitokonstantinou, V., Papoutsis, I., Kontoes, C., Lafarga, A., Armesto A. P. y Garraza, J. A. 2018. Scalable parcel-based crop identification scheme using Sentinel-2 Data Time-Series for the monitoring of the Common Agricultural Policy. Remote Sensing 10 (6), 911.

Therneau, T. M., Atkinson, B., Brian Ripley. 2018. Package 'rpart' 34.

Toro, N., Gomariz-Castillo, F., Cánovas-García, F., Alonso-Sarría, F. 2015. Comparación de métodos de clasificación de imágenes de satélite en la Cuenca Del Río Argos (Región de Murcia). Boletín de La Asociación de Geógrafos Españoles 67, 327-347.