MODIFICACIÓN DEL ALGORITMO RANDOM FOREST PARA SU EMPLEO EN CLASIFICACIÓN DE IMÁGENES DE TELEDETECCIÓN

Fulgencio Cánovas-García [1,2], Francisco Alonso-Sarría[3], Francisco Gomariz-Castillo[3,4]

[1] Universidad Técnica Particular de Loja, Departamento de Geología y Minas e Ingeniería Civil, San Cayetano Alto s/n, Loja, Ecuador. fulgencio.canovas@um.es
[2] Universidad de Cuenca, Departamento de Ingeniería Civil, Av. 12 de abril. Ciudadela Universitaria, Cuenca, Ecuador.
[3] Universidad de Murcia, Instituto Universitario del Agua y del Medio Ambiente, Edificio D Campus de Espinardo s/n, 30100 Murcia, España. alonsarp@um.es
[4] Fundación Instituto Euromediterráneo del Agua, Campus de Espinardo, s/n, 30100 Murcia, España. fjgomariz@um.es

RESUMEN

Random Forest es uno de los algoritmos de clasificación de imágenes más usados en teledetección. Una de sus ventajas es que aporta una estimación interna de exactitud mediante una forma de validación cruzada. La hipótesis que plantea este trabajo es que esta estimación subestima el error de predicción cuando se clasifican coberturas del suelo con áreas de entrenamiento compuestas por varios píxeles, ya que se viola la necesaria independencia estadística entre casos de entrenamiento y de validación. El objetivo de esta investigación es modificar el algoritmo *Random Forest* para evitar este problema y obtener una estimación de la exactitud más realista. La hipótesis previa y el nuevo algoritmo se contrastan mediante la clasificación de coberturas del suelo de una imagen del satélite Landsat 5 del 24 de julio de 2009. Al aplicar el algoritmo original, este subestima claramente el error de clasificación. Sin embargo, con el algoritmo modificado se obtiene un error global de clasificación muy parecido al obtenido con una validación independiente del algoritmo; además, su capacidad de predicción no disminuye. Al comparar por clases se constata que, con el algoritmo modificado, los errores por clases obtenidos del *out-of-bag* son equivalentes a los de una validación cruzada, mientras que al emplear el algoritmo original son muy inferiores.

Palabras clave: Clasificación de imágenes; *Random Forest*; *Bagging*; independencia estadística.

ABSTRACT

Random Forest is one of the most used classification algorithms in remote sensing. One of its advantages is that it produces an internal accuracy estimation using a form of cross-validation. The hypothesis raised by this work is that this estimation underestimate the prediction error when classifying land cover with training areas composed of several pixels, as the statistical independece among training and validation cases is compromised. The objective of this research is to modify the Random Forest algorithm so that, this independence is not compromised, obtaining a more realistic accuracy estimation. This hypothesis and the new algorithm are tested by classifying land cover with a Landsat 5 satellite image from July 24, 2009. When applying the original algorithm, out-of-bag errors clearly underestimates actual errors. However, when using the modified algorithm, the out-of-bag error estimation is quite similar to the error obtained with a validation separated from the algorithm; additionally, the predictive power of the modified algorithm is not reduced. Carrying out a comparison by class, it has been found that the out-of-bag errors from the modified algorithm are equivalent to those of a cross-validation, whereas when using the original algorithm classification errors were much lower than the cross-validation estimation.

Keywords: Classification of images; Random Forest; bagging; statistical independence.

1. INTRODUCCIÓN

Hasta mediados de la década de los 90 los métodos de clasificación supervisada aplicados al análisis de imágenes de satélite se basaban principalmente en técnicas estadísticas convencionales como la clasificación por máxima verosimilitud o por mínima distancia. Aunque estas técnicas pueden dar buenos resultados, su capacidad para resolver problemas de confusión entre clases es muy limitada. En consecuencia, a raíz de los avances experimentados en el campo de la computación, se han propuesto estrategias alternativas basadas en técnicas de inteligencia artificial y aprendizaje automático, particularmente el uso de redes neuronales artificiales, árboles de decisión, máquinas de vectores soporte o métodos derivados de la teoría de la lógica borrosa (Tso *et al.,* 2009).

Sin embargo, los algoritmos de clasificación empleados en teledetección suelen proceder de la estadística o del aprendizaje automático, no siendo habitual que se creen algoritmos específicos para su uso en teledetección. Asimilar el problema general de la clasificación al de la clasificación de imágenes no es plenamente correcto ya que no es posible equiparar los individuos clasificados en un problema estándar de clasificación y los píxeles clasificación en teledetección. Los primeros son individuos reales, pero a los segundos hemos decidido considerarlos como tales por una cuestión operativa.

Por otro lado, la estadística tradicional se basa en el supuesto de que los individuos analizados son independientes entre sí. Sin embargo este supuesto no se cumple en estadística espacial ya que dos puntos cercanos no son independientes entre sí, y menos si se trata de dos píxeles contiguos como ocurre con los que forman las áreas de entrenamiento y validación que suelen utilizarse para clasificar una imagen de satélite.

Random Forest es uno de los algoritmos de clasificación de imágenes más usados en teledetección. Una de sus ventajas es que aporta una estimación interna de exactitud mediante una forma de validación cruzada. Sin embargo, esta estimación subestima el error de predicción al clasificar coberturas del suelo con áreas de entrenamiento compuestas por varios píxeles, ya que se viola la necesaria independencia estadística entre casos de entrenamiento y de validación.

1.1 El algoritmo *Random Forest*

Los árboles de decisión (Breiman *et al.*, 1984) están entre los métodos de clasificación supervisada más utilizados. Se trata de un método no paramétrico, robusto y fácil de interpretar. Funcionan haciendo particiones sucesivas en el espacio de variables buscando siempre la variable y el valor umbral de la misma que maximizan la homogeneidad de las particiones resultantes. La homogeneidad de una partición puede medirse de varios modos, uno de los más comunes es el índice de Gini:

$$G = \sum_{k=1}^{K} f_k \cdot (1 - f_k)$$

donde k es cada una de las clases presentes en la partición, K el total de clases presentes en la partición y f_k la proporción de los casos en la partición que pertenecen a la clase k. Para calcular el índice de Gini de un árbol completo, habría que sumar los índices de Gini de todas sus particiones. El proceso de partición continúa hasta que todas las particiones son totalmente homogéneas. En ese momento empieza el proceso de poda (*pruning*) del árbol utilizando un procedimiento de validación cruzada que evita que el árbol se sobreajuste a los datos de entrenamiento. Se trata básicamente de reagrupar las particiones más pequeñas que responden solo al ruido en los datos de entrenamiento. Una vez podado el árbol, a cada partición del espacio de variables le corresponde la clase más frecuente de modo que cualquier nuevo caso se clasifica en función de donde se sitúe en dicho espacio de variables.

El principal problema que tienen los árboles de decisión es que son muy sensibles a pequeñas variaciones en los datos de entrada que pueden encaminar al árbol de decisión por un camino diferente, dando lugar a una clasificación muy diferente. Los clasificadores basados en conjuntos de clasificadores sencillos (*ensemble learning*) han recibido considerable atención como una forma de superar este tipo de problema.

Random Forest (Breiman, 2001) utiliza varios árboles de decisión (entre 500 y 2.000). Cada uno de ellos se entrena con un subconjunto aleatorio de casos (obtenido mediante *bootstrapping*) denominado *in-bag*, el resto de los casos forman el *out-of-bag*. Además, en cada división (nodo) de los árboles se considera solo un subconjunto aleatorio de los predictores. Cada nuevo caso se presenta a cada uno de los árboles (que no han sido previamente podados) y se asigna a la clase más frecuentemente escogida por los árboles. La proporción de árboles que ha votado a cada clase puede también interpretarse como la probabilidad de pertenencia a dicha clase. La aleatoriedad introducida disminuye la correlación entre árboles dando más sentido al uso de un conjunto de clasificadores. Por otra parte, al utilizar varios predictores disminuye el error de generalización (Breiman, 2001; Pal, 2005; Prasad *et al*., 2006) y se obtienen mejores resultados que con otros algoritmos (Breiman, 2001; Liaw y Wiener, 2002). A partir de los casos en el *out-of-bag* se obtiene una estimación del error de clasificación (OOB-CV) válida ya que la respuesta para cada observación se obtiene empleando solo los árboles que no fueron calibrados utilizando esa observación. James *et al.* (2013) afirman que con un número de árboles suficientemente grande la estimación de OOB-CV es prácticamente equivalente a la obtenida con validación cruzada o *leave-one-out cross validation* (LOO-CV).

El algoritmo *Random Forest* utiliza dos parámetros: el número de árboles y el número de predictores a utilizar en cada partición de cada uno de los árboles. Sin embargo, una de las grandes ventajas de este algoritmo es su baja sensibilidad a estos parámetros, por lo que los valores por defecto suelen producir buenos resultados (Liaw y Wiener, 2002; Hastie *et al*., 2009).

El principal problema de *Random Forest*, en comparación con el análisis de un único árbol de clasificación, es que es más difícil de interpretar. Ya no se dispone de un único árbol en el que pueda verse el efecto de cada variable, sino de un gran número de ellos cuyo efecto conjunto no puede visualizarse. Sin embargo, *Random Forest* permite obtener medidas acerca de la importancia que los diferentes predictores han tenido en el modelo, lo que permite en parte interpretar este. La importancia de los predictores se evalúa como el número de veces que han sido utilizados por los diversos árboles y su capacidad para reducir el índice de Gini en ellos.

1.2 El problema de la dependencia espacial de OOB-CV y LOO-CV

Es comúnmente aceptado que la OOB-CV de *Random Forest* es un estimador no sesgado de la exactitud de la clasificación general, siendo por tanto innecesario realizar una validación cruzada externa (Efron y Tibshirani, 1997; Breiman, 2001; Svetnik *et al*., 2003). Tan sólo hemos encontrado una referencia argumentando que esta medida interna podría estar sesgada, pero sólo cuando el número de casos es menor que el número de variables (Mitchell, 2011). Sin embargo esto solo es cierto si se asume la independencia entre los casos de entrenamiento y de validación, lo que suele ser un problema cuando se trabaja con datos espaciales. Al clasificar las imágenes de teledetección, los casos se obtienen como áreas de entrenamiento generalmente formadas por varios píxeles contiguos y homogéneos. *Random Forest* asumirá que, aun perteneciendo a la misma parcela, los píxeles son casos independientes y los dividirá entre el *in-bag* y el *out-of-bag*. En este trabajo proponemos la hipótesis de que, en este caso, OOB-CV subestima notablemente el error real de predicción.

El mismo problema puede ocurrir con LOO-CV si, en lugar de utilizar áreas de entrenamiento completas para validar, se utilizan píxeles aislados. Nosotros asumimos por tanto que lo que se debe dejar fuera son todos los píxeles que forman una parcela. De esta forma no se comprometerá la independencia estadística entre los datos de entrenamiento (todas la áreas de validación menos una) y el área de entrenamiento a validar, ya que ninguno de los elementos que forma parte de esta parcela está incluido en el modelo de clasificación

Esta aclaración nos parece importante y creemos que pocas veces ha sido puesta de manifiesto. Además, nos da pié a introducir la nomenclatura que se empleará a lo largo de todo el trabajo. Cuando estemos tratando de validación cruzada diferenciaremos entre validación cruzada dejando-una-fuera (LOO-CV), lo que implica que en cada ciclo de clasificación se deja un píxel fuera para ser evaluado, y por otro lado validación cruzada dejando-una-parcela-fuera (LOPO-CV), en la que en cada ciclo de clasificación se deja fuera a todos los píxeles que pertenezcan a una misma área de entrenamiento (parcela). En clasificación digital de imágenes la LOO-CV no tiene sentido, solo podremos practicar la LOPO-CV.

A continuación se procede a una clarificación de la nomenclatura empleada para referirnos a los distintos tipos de validación.

- VAL: Validación realizada con unos datos diferentes e independientes de los datos de entrenamiento.
- OOB-CV: Validación cruzada interna de *Random Forest* utilizando el *out-of-bag*.
- LOPO-CV: Validación cruzada dejando-una-parcela-fuera.
- LOO-CV: Validación cruzada dejando-un-pixel-fuera.

Si esta nomenclatura lleva una O delante, quiere decir que se ha llevado a cabo empleando el algoritmo original, si lleva una M delante, se ha llevado a cabo empleando el algoritmo modificado. Por ejemplo M-LOPO-CV quiere decir que se ha llevado a cabo una validación cruzada dejando-una-parcela-fuera empleando para clasificar el algoritmo modificado.

1.3 Objetivos

El principal objetivo de este trabajo es modificar el código original del algoritmo *Random Forest* de forma que no se vea comprometida la independencia estadística entre el conjunto de datos de entrenamiento (*in-bag*) y el conjunto de datos que internamente utiliza como validación (*out-of-bag*). Además nos proponemos conseguir varios objetivos específicos:

- Demostrar que la estimación OOB-CV (O-OOB-CV) obtenido con el algoritmo original subestima el error real.
- Demostrar que esta subestimación puede afectar al resultado de un proceso de selección de variables.
- Programar una modificación del algoritmo original que, sin alejarse del planteamiento original del algoritmo, no viole la independencia estadística en el reparto interno entre *in-bag* y *out-of-bag*.
- Demostrar que tras esta modificación la estimación M-OOB-CV se puede considerar una estimación aceptable del error real y que esta es equivalente a la estimación LOPO-CV.
- Demostrar que tras la modificación propuesta no se produce una merma en la capacidad de predicción del algoritmo.

2. ÁMBITO DE ESTUDIO Y DATOS EMPLEADOS

El área de estudio escogida ha sido la cuenca del río Vinalopó, con una superficie aproximada de 3.000 km^2. Ubicada al sureste de la Península Ibérica, al sur de la Provincia de Alicante, es una cuenca litoral característica de las zonas semiáridas del sureste español, con fuerte presión antrópica y grandes mosaicos de cultivos, estando ocupada más del 62% por usos antrópicos (Gomariz-Castillo *et al*., 2014).

Se ha utilizado una imagen del satélite Landsat 5, sensor *Thematic Mapper* (path 199, row 33). La fecha de esta imagen corresponde al 24 de julio de 2009. De las siete bandas disponibles se han utilizado seis, tres del visible, una del infrarrojo cercano y dos del infrarrojo medio. Como parte del preprocesado de la imagen se llevó a cabo una corrección atmosférica y de iluminación basada en los métodos de Chávez (1988) y del lambertiano C (Teillet *et al*., 1982). Además se ha empleado información del relieve mediante el Modelo Digital de Elevaciones del Instituto Geográfico Nacional, escala 1:25.000.

El objetivo de la clasificación ha sido la obtención de un mapa de coberturas con las siguientes clases: Bosque (Bos); Vegetación arbustiva (VArb); Arbóreo poco denso (ArbND); Arbóreo denso (ArbD); Herbáceo de secano (HerS); Herbáceo de regadío (HerR); Superficies impermeables (Imp); Láminas de agua (Agu); Suelo desnudo (SueD); Viñedo (Vid).

3. METODOLOGÍA

3.1 Modificación del algoritmo Random Forest

La modificación que se ha hecho a la función original de Liaw y Wiener (2002) en R consiste simplemente en que el usuario debe indicar un vector que contiene los datos que permiten identificar el área de entrenamiento a la que pertenece cada píxel. Posteriormente en lugar de hacer *bootstrapping* de los píxeles, se hace de las áreas de entrenamiento, de manera que en cada árbol todos los píxeles en una misma área de entrenamiento van o bien al *in-bag* o bien al *out-of-bag*, pero no se reparten entre uno y otro.

3.2 Obtención de las áreas de entrenamiento y validación

Las características generales de las muestras de entrenamiento y validación aparecen en la tabla 1.

Clase	*Bos*	*VArb*	*ArbND*	*ArbD*	*HerS*	*HerR*	*Imp*	*Agu*	*SueD*	*Vid*	*Total*
Entrenamiento											
Parcelas	19	22	13	14	15	10	16	11	4	17	141
Píxeles	5.267	4.841	1.241	2.374	3.715	4.695	6.783	6.262	118	3.177	38.473
Validación											
Parcelas	10	12	7	8	8	5	7	6	2	8	73
Píxeles	1.563	3.410	828	636	1.744	1.653	1.798	3.327	129	928	16.046

Tabla 1. Número de parcelas y de píxeles que componen el conjunto de datos de entrenamiento y validación.

Se ha hecho un muestreo estratificado, intentando que todas las clases estuviesen bien representadas, por lo que el tamaño de los estratos no es proporcional a la superficie ocupada por las clases. Las áreas de validación se obtuvieron mediante un muestreo aleatorio. Por el contrario, las áreas de entrenamiento se eligieron buscando aquellas que representasen adecuadamente las distintas clases.

3.3 Variables utilizadas

La tabla 2 muestra las variables extraídas de los píxeles agrupadas en cinco grandes categorías. Se ha añadido una pequeña descripción cuando se ha considerado apropiado. En total hay 55 variables, 14 espectrales, siete relacionadas con el MDT y 34 texturales (Haralik *et al.,* 1973).

3.4 Ordenación y selección de variables

Disponer de un elevado número de predictores no es necesariamente una ventaja en aprendizaje automático (Hughes, 1968). Es necesario, por tanto, establecer un proceso de selección de variables para eliminar aquellas que resulten redundantes o no aporten información. Un enfoque adecuado es considerar el proceso de selección como un procedimiento heurístico en el que se especifica un subconjunto de variables en cada paso de una búsqueda iterativa (Blum *et al.*, 1997). Tal procedimiento implica 3 pasos:

1. Ordenar las variables de acuerdo con su relevancia para clasificar el conjunto de datos. Se ha utilizado como criterio la importancia de las variables calculada por *Random Forest*.

2. Iterativamente, modificar un modelo de clasificación eliminando variables de acuerdo a su rango.

3. Seleccionar el mejor subconjunto de variables en función de una medida de exactitud en la clasificación.

Bandas originales		*Derivadas del DEM*	
B1 (1)	Azul (0.45-0.52 µm)	SLOPE (1)	Pendiente
B2 (1)	Verde (0.52-0.60 µm)	ASP (1)	Orientación
B3 (1)	Rojo (0.63-0.69 µm)	CURV.perp (1)	Curvatura perpendicular
B4 (1)	Infrarrojo cercano (0.76-0.90 µm)	CURV.tang (1)	Curvatura tangencial

B5 (1)	Infrarrojo de onda corta (1.55-1.75 µm)	ASP.sin (1)	Seno de la orientación
B7 (1)	Infrarrojo de onda corta (2.08-2.35 µm)	ASP.cos (1)	Coseno del a orientación
DEM (1)	Modelo digital de elevaciones		
Capas de textura basadas en el semivariograma experimental		Índices y transformaciones	
VARIO.tc.1 (1)	Semivariograma experimental calculado con la primera capa de la transformación Tasselled Cup	NDVI (1)	Índice de Vegetación de Diferencia Normalizada
		INTENSITY (1)	Intensidad (transformación TC)
VARIO.ndvi (1)	Semivariograma experimental calculado con el NDVI	HUE (1)	Brillo (transformación TC)
		SATURATION (1)	Saturación (transformación TC)
Características de textura (Haralick *et al*., 1973) calculadas a partir de la primera capa obtenida de la trasformación Tasselled Cup			
GLCM.homo (5)	Homogeneidad	GLCM.asm (5)	Segundo Momento Angular
GLCM.cont (5)	Contraste	GLCM.coor (5)	Correlación
GLCM.ent (5)	Entropía	GLCM.var (5)	Varianza

Tabla 2. Resumen de las características extraídas de las imagen. Las características de tipo textural han sido calculadas en varias direcciones. El número total de características aparece entre paréntesis

Para empezar, se utilizaron todas las variables para entrenar ambos algoritmos (original y modificado); se calculó el índice kappa con la muestra de validación y con el *out-of-bag*. La variable menos importante en cada caso fue eliminada y se repitió el procedimiento de forma recursiva hasta que sólo quedó una de las variables. Se obtuvieron así cuatro vectores de índices kappa: dos aplicando el algoritmo original (uno obtenido del *out-of-bag* y otro de la muestra de validación) y otros dos aplicando el algoritmo modificado (de nuevo *out-of-bag* y validación). Este procedimiento, con algunas modificaciones, se utilizó para seleccionar el subconjunto óptimo de variables para una clasificación (Cánovas-García y Alonso-Sarria, 2015).

Una vez decidido el subconjunto de variables que minimizan el error de clasificación, se analizaron con mayor detalle los resultados mediante unas pirámides de errores de omisión y comisión para estudiar las diferencias en cuanto a exactitud de la clasificación por clase (errores de omisión y comisión) comparando:

- O-OOB-CV y O-LOPO-CV, para conocer el grado de subestimación, si es que existe, del error del OOB con el algoritmo original.
- M-OOB-CV y M-LOPO-CV, para conocer si los resultados del OOB con el algoritmo modificado son similares a los de una validación cruzada dejando-una-parcela-fuera.
- M-LOPO-CV y O-LOPO-CV, para conocer si la modificación del algoritmo ofrece un resultado equivalente al algoritmo original.
- O-VAL y M-VAL, para conocer si la modificación del algoritmo provoca una reducción de su capacidad de generalización.

4. RESULTADOS Y DISCUSIÓN

4.1 Ciclos de clasificación para la selección de variables

La figura 1 muestra la evolución del índice kappa conforme se eliminan variables del modelo empezando por las menos importantes y dejando las más importantes de acuerdo con la ordenación hecha por *Random Forest.* Las curvas que se obtienen con datos de validación son muy similares cuando se utiliza el algoritmo original y cuando se utiliza el modificado; es decir, que la modificación introducida en el algoritmo no altera su comportamiento. Ambas curvas son también muy similares a las que se obtienen, utilizando M-OOB-CV, con el algoritmo modificado. Es decir que la validación con M-OOB-CV y con datos independientes de validación son muy similares si se trabaja con el algoritmo modificado. Por el contrario, los resultados de O-OOB-CV son muy diferentes y están claramente sobreestimados.

Por lo que se refiere a la identificación de un subconjunto óptimo de variables que maximizan la exactitud de la clasificación, los datos de O-OOB-CV no nos permiten identificar el subconjunto de variables óptimo, mientras que los datos del M-OOB-CV sí nos lo permiten. El índice kappa del O-OOB-CV comienza a descender a partir del número de orden 4, mientras que el índice kappa de los datos de validación (algoritmo original) ya comenzaba a descender de manera constante a partir del número de orden 11. Estos gráficos nos permiten seleccionar el subconjunto de variables que maximizan la exactitud de la clasificación dado un conjunto de variables ordenado. De ahora en adelante se seguirán analizando los resultados pormenorizados de los modelos de clasificación generados con las primeras 13 variables (línea vertical azul de la figura 1).

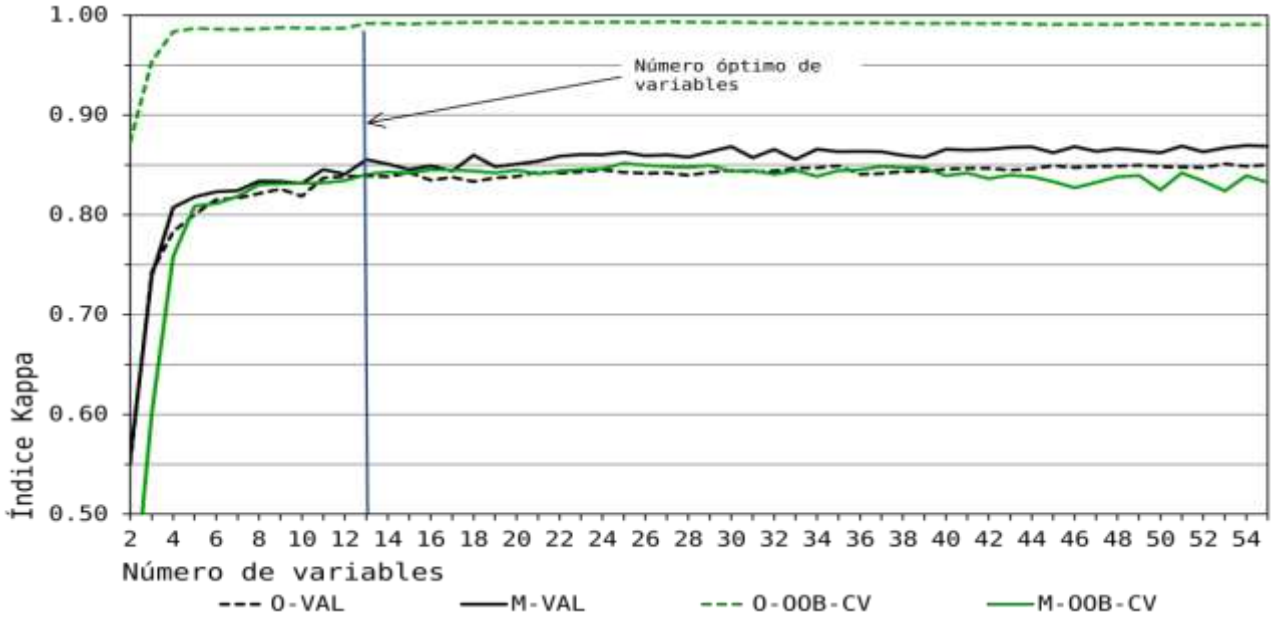

Figura 1. Índices kappa obtenidos con el algoritmo Random Forest original (línea discontinua) y modificado (línea continua) a partir de los datos de validación (en negro) y a partir del *out-of-bag* (en verde).

4.2 Análisis de los errores de omisión y comisión

Una vez decidido el subconjunto de variables que maximiza la exactitud de la clasificación se han ejecutado los correspondientes modelos y se han analizado en profundidad los resultados. Las figuras 2 y 3 muestran 4 pirámides con los errores de omisión y comisión obtenidos con el algoritmo original y el modificado utilizando diferentes formas de validación.

En la figura 2, a la izquierda, se compara O-OOB-CV con O-LOPO-CV. Los errores por clases de O-OOB-CV son muy inferiores a los observados con O-LOPO-CV. Solo se consiguen buenas estimaciones de los errores por clases cuando O-LOPO-CV es cercano a 0. El caso más evidente de subestimación se presenta en la clase *Suelo desnudo,* con O-OOB-CV se obtienen valores cercanos a 0 y con O-LOPO-CV se obtienen valores ligeramente por encima de 0.8. Con los errores de omisión sucede algo parecido. Lo comentado para la clase *Suelo Desnudo* se puede afirmar también para la clase *Arbóreo no denso*. De nuevo, solo se obtienen buenas estimaciones de los errores con O-OOB-CV cuando los errores de O-LOPO-CV son cercanos a 0.

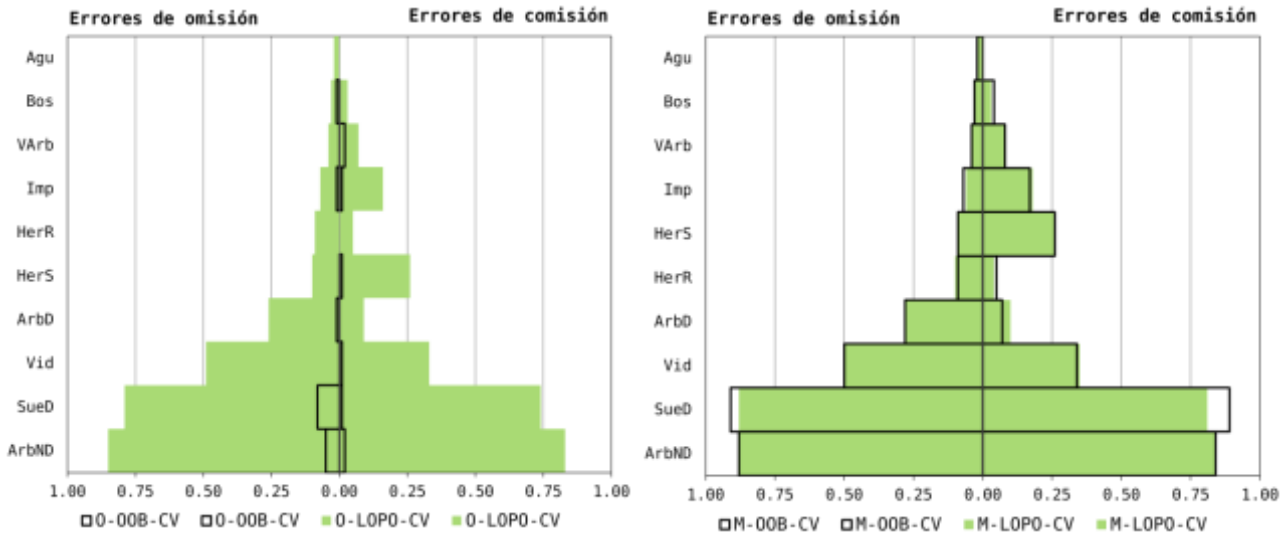

Figura 2. A la izquierda, comparación de errores de omisión y comisión estimados con el *out-of-bag* y con validación cruzada dejando una parcela fuera. En ambos casos se trata del algoritmo original. A la derecha, comparación de errores de omisión y comisión estimados con el *out-of-bag* y con validación cruzada dejando-una-parcela-fuera; en ambos casos se trata del algoritmo modificado

En esta investigación se propone la hipótesis de que al modificar el algoritmo *Random Forest* según lo indicado en el apartado 3.1, el resultado de la validación interna del algoritmo (M-OOB-CV) es equiparable a una validación cruzada dejando-una-parcela-fuera (M-LOPO-CV). Para contrastar esta hipótesis, la figura 2 (derecha) muestra que tanto los errores de omisión como los de comisión son prácticamente iguales; solo existen pequeñas diferencias en suelo desnudo (omisión y comisión). Pero estas diferencias son muy pequeñas, debidas posiblemente a la propia variabilidad en la generación de los modelos de clasificación.

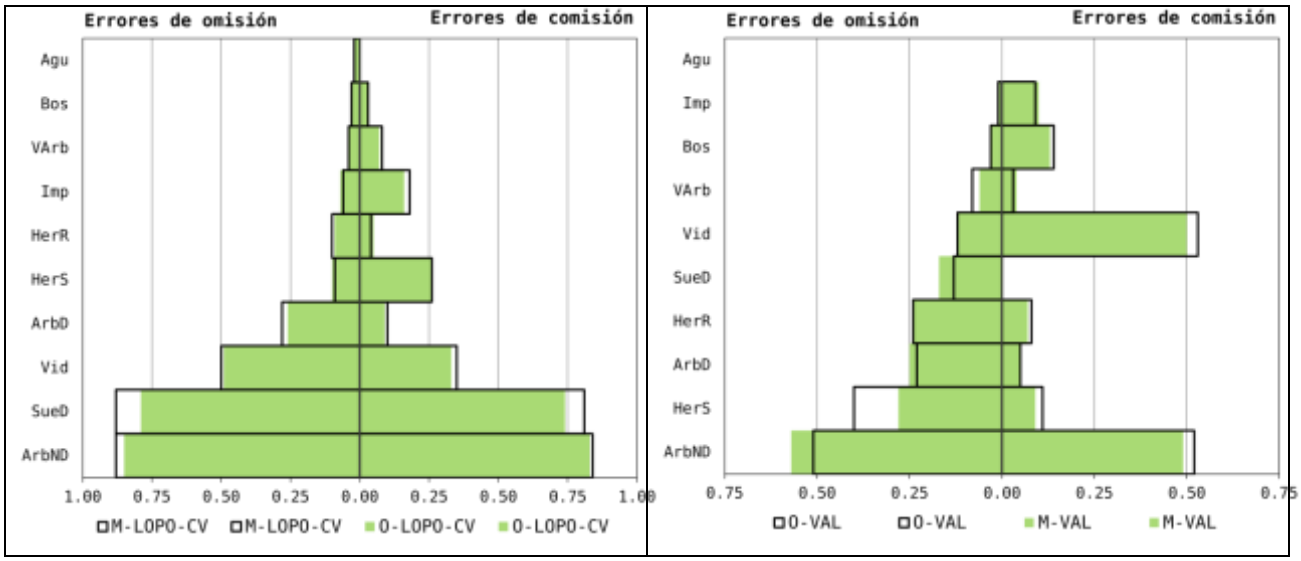

Figura 3. A la izquierda, comparación de errores de omisión y comisión estimados con validación cruzada dejando-una-parcela-fuera; utilizando tanto el algoritmo original como el modificado. A la derecha, comparación de errores de omisión y comisión estimados en áreas de validación independientes tanto con el algoritmo original como con el modificado.

Por último, se lleva a cabo la comparación de los resultados de *leave-one-out-CV* empleando el algoritmo original y el modificado (figura 3 a la izquierda). La mayor parte de las clases obtienen resultados equiparables. Solo hay una clase que obtiene valores algo diferentes. La clase *Suelo desnudo*, en la que se obtienen errores de omisión y comisión mayores con el algoritmo modificado. Aun así, las diferencias no son importantes.

La figura 3 (derecha) muestra los resultados obtenidos con una muestra independiente de validación, ambos algoritmos arrojan resultados muy similares, solo en el caso de los cultivos herbáceos de secano se observa una cierta diferencia.

5. CONCLUSIONES

1. Al clasificar imágenes de satélite existe una clara diferencia entre la estimación del error de clasificación que se obtiene a partir del OOB-CV con el algoritmo original de Random Forest y el que se obtiene con una validación independiente de la del algoritmo. En el primer caso se sobreestima claramente el índice kappa.

2. La modificación propuesta del algoritmo Random Forest permite obtener unas estimaciones de error con OOB-CV mucho más próximas a las que se obtienen con una muestra de validación independiente de la de entrenamiento.

3. La modificación propuesta no resta capacidad de predicción al algoritmo. Los resultados obtenidos con ambos algoritmos son similares.

4. Al analizar los errores de omisión y comisión por clases los resultados son similares. Existe una gran diferencia, cuando se trabaja con el algoritmo original, entre los resultados obtenidos con OOB-CV y con validación cruzada externa. Por el contrario, los errores obtenidos con OOB-CV con el algoritmo modificado son muy similares a los que se obtienen con una validación externa. Finalmente, el algoritmo modificado no pierde capacidad de predicción ya que los resultados de validación externa de ambos algoritmos son muy similares.

6. BIBLIOGRAFÍA

Blum,A.L. y Langley,P. (1997) "Selection of relevant features and examples in machine learning", *Artificial Intelligence,* 97, pp. 245-271.

Breiman, L., Friedman. J.H., Olshen, R.A., Stone, C.J. (1984) *Classification and Regression Trees,* New York, Chapman and Hall/CRC.

Breiman, L. (2001) "Random Forests", *Machine Learning*, 45, 5-32.

Cánovas-García, F., Alonso-Sarría,F. (2015) "Optimal Combination of Classification Algorithms and Feature Ranking Methods for Object-Based Classification of Submeter Resolution Z/I-Imaging DMC Imagery", *Remote Sensing,* 7, 4, pp. 4651-4677.

Chávez, P.S. (1988) "An improved dark-object substraction technique for atmospheric scattering correction of multispectral data", *Remote Sensing of Environment,* 24, pp. 459-479.

Efron, B. Y Tibshirani, R. (1997) "Improvements on Cross-Validation: The .632+ Bootstrap Method*", Journal of the American Statistical Association*, 92, 438, 548-560.

Gomariz Castillo, F., Alonso Sarría, F. y Cánovas García, F. (2014) "Clasificación multitemporal de usos del suelo en la Cuenca del Río Vinalopó (Comunidad Valenciana) mediante diferentes algoritmos de clasificación supervisada y variables auxiliares" *XVI Congreso Nacional de Tecnologías de la Información Geográfica* 25, 26 y 27 de Junio de 2014. Alicante.

Haralick, R.M., Shanmugan, K. y Dinstein, I. (1973) "Textural features for image classification", *Systems, man and cybernetics*, SMC-3, 6, pp. 610-621.

Hastie,T., Tibshirani,R. y Friedman,J. (2009) *The Elements of Statistical Learning: Data Mining, Inference, and Prediction,* New York, Springer.

Hughes, G.F. (1968) "On the mean accuracy of statistical patterns recognizers", *IEEE Transactions Information Theory,* 14, 1, 55-63.

James, G., Witten, D., Hastie, T. y Tibshirani, R. (2013) *An Introduction to Statistical Learning: with Applications in R*, New York, Springer.

Liaw, A., Wiener, M. (2002) "Classication and Regression by randomForest", *R News*, 2, 3, 18-22.

Mitchell, M. (2011) "Bias of Random Forest Out-of-Bag (OOB) Error for Certain Input Parameters", *Open Journal of Statistics*, 1, 205-211.

Pal, M. (2005) "Random forest classifier for remote sensing classification", *International Journal of Remote Sensing,* 26, 217-222.

Prasad, A.M., Iverson, L.R., Liaw, A. (2006). "Newer classification and regression tree techniques: bagging and random forests for ecological prediction", *Ecosystems,* 9, 181-199.

Puissant, A., Rougier, S. y Stumpf, A. (2014) "Object-oriented mapping of urban trees using random forest classifiers", *International Journal of Applied Earth Observation and Geoinformation*, 26, pp. 235-245.

Svetnik, V., Law, A., Tong, C., Culberson, J.C: Sheridan, R.P. y Feuston, B.P. (2003) "Random Forest: A Classification and Regression Tool for Compound Classification and QSAR Modeling", *Journal of Chemical Information and Computer Sciences*, 43, pp. 1947-1958.

Teillet, P.M., Guindon, B. y Goodenough, D.G. (1982) "On the slope-aspect correction of multispectral scanner data", *Canadian Journal of Remote Sensing*, 58, pp. 84-106.

Tso,B. y Mather,P.M. (2009) *Classification Methods for Remotely Sensed Data*, Londres, Taylor & Francis.

AGRADECIMIENTOS

Este trabajo ha sido financiado por el Proyecto Prometeo de la Secretaría de Educación Superior, Ciencia, Tecnología e Innovación del Gobierno de Ecuador.