XVI Congreso Nacional de Tecnologías de la Información Geográfica
25, 26 y 27 de Junio de 2014. Alicante.

Clasificación multitemporal de usos del suelo en la Cuenca del Río Vinalopó (Comunidad Valenciana) mediante diferentes algoritmos de clasificación supervisada y variables auxiliares

Francisco Gomariz Castillo[ab]*, Francisco Alonso Sarría[c], Fulgencio Cánovas García[d]

[a]*Instituto Euromediterráneo del Agua*
[b]*Departamento de Ciencias del Mar y Biología Aplicada, Universidad de Alicante*
[c] *Instituto del Agua y del Medio Ambiente (INUAMA), Universidad de Murcia, Edificio D*
[d] *Departamento de Ingeniería Civil, Universidad de Cuenca*

Resumen

La dificultad y coste de obtener información en continuo dificulta la cuantificación de los procesos que intervienen en el ciclo hidrológico. Se requieren por tanto métodos de obtención indirecta como la teledetección. El objetivo de este trabajo es la caracterización espacio-temporal de los usos del suelo en la cuenca del Vinalopó, representativa de paisajes fuertemente antropizados y heterogéneos del sureste español.

Como información de partida se han utilizado imágenes del sensor Landsat TM para la serie temporal 2000-2010. Se evalúan diferentes aspectos como la mejora en la estimación al incluir imágenes de varias estaciones para un mismo año, (hasta cuatro fechas representativas de las cuatro estaciones del año) o variables auxiliares derivadas del relieve (elevaciones, pendientes y orientaciones) y texturales (semivariograma del albedo y el NDVI).

E-mail: francisco.gomariz@ua.es.

Así mismo se evalúan diferentes métodos de clasificación. Un método paramétrico: Máxima Verosimilitud (ML); dos no paramétricos: Random Forest (RF) y Máquinas de Vectores Soporte (SVM) con kernel radial; y el método paramétrico contextual Sequential Maximum a Posteriori (SMAP). Los parámetros de RF y SVM se optimizan mediante validación cruzada y minimización del error de clasificación. Como medida de bondad en la clasificación se ha utilizado el índice kappa, estimado mediante intervalos de confianza.

El proceso de trabajo se desarrolla sobre una plataforma de bajo coste, utilizando programas Open Source (GRASS y R) y como fuentes de información las plataformas liberalizadas de productos Landsat y el Plan Nacional de Teledetección.

Los resultados revelan que el uso de imágenes de varias estaciones y variables auxiliares mejora las clasificaciones en todos los algoritmos. En lo referente a los algoritmos de clasificación, el exhaustivo trabajo realizado sobre los polígonos de entrenamiento y validación mejora los resultados de ML, no siendo significativamente peor al resto, a priori más robustos en estas zonas, caracterizadas por la alta variabilidad y falta de normalidad de las variables. También se aprecia una mejora en los resultados de RF y SVM al optimizar sus parámetros.

Palabras clave: Teledetección; aprendizaje automático; máquinas de vectores soporte; random forest; sequential maximum a posteriori; usos del suelo

1. Introducción

La cuantificación de los procesos que intervienen en el ciclo hidrológico resulta compleja debido al coste y dificultad de obtener información en continuo. En este sentido, la teledetección facilita la estimación espacio-temporal de variables y parámetros hidrológicos, ofreciendo un marco metodológico para comprender mejor la hidrología a gran escala sin las limitaciones de la observación tradicional a escala local (Tang y otros, 2009), siendo la estimación de coberturas de usos la aplicación más frecuente.

No obstante, la elevada heterogeneidad de los paisajes mediterráneos debido a su fuerte fragmentación y la gran variabilidad espacial de la cubierta vegetal (Alrababah y Alhamad, 2006), y la alta reflectividad en zonas calcáreas y suelos muy secos (Berberoglu et al., 2007) dificultan el empleo de la teledetección para estos fines.

Recientemente, diversos algoritmos utilizados en aprendizaje automático se han venido aplicando a la clasificación de imágenes de satélite. Lu y Weng (2007) analizan algunos de los métodos de clasificación más avanzados, su tipología y las técnicas para mejorar su precisión. Mountrakis y otros (2011) analizan decenas de trabajos sobre Support Vector Machines (SVM), destacando su capacidad de generalización, auto adaptación y fiabilidad incluso con información limitada. En lo referente a Random Forest (RF), Gislason, Benediktsson y Sveinson (2006) o Rodríguez-Galiano et al. (2012), aplicándolo en Granada, obtienen buenos resultados, debido a su robustez ante valores anómalos, al no sobre ajustar los modelos. Sequential Maximum a Posteriori (SMAP) ha sido menos probado, aunque autores como Kumar et al. (2012) han obtenido mejores resultados que con RF o SVM.

El objetivo de este trabajo es evaluar algunas de las técnicas de aprendizaje automático más empleadas y flexibles. Otro objetivo es comprobar si la inclusión de variables auxiliares y escenas correspondientes a varias estaciones del año mejora la clasificación.

2. Metodología, materiales, datos y herramientas

2.1. Área de estudio y fuentes de información

El área de estudio escogida ha sido la Cuenca del Río Vinalopó (Fig. 1), con 3.000km^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 zonas de mosaicos de cultivos, estando ocupada más del 62% por usos antrópicos.

Como información espectral se ha utilizado Landsat 5 y 7, con resolución espacial y espectral suficiente para clasificar con precisión gran variedad de paisajes, como paisajes mediterráneos heterogéneos (Rozenstein & Karnieli, 2011). Tras analizar su calidad se han seleccionado 40 imágenes del Plan Nacional de Teledetección (PNT), el United States Geological Survey y las bibliotecas de imágenes del Instituto del Agua y Medio Ambiente e Instituto Euromediterráneo del Agua. La serie temporal ha sido 2000-2010, utilizando una escena por estación del año, o en su defecto la integración invierno-verano.

Como variables auxiliares en la clasificación se ha incluido información del relieve (Modelo Digital de Elevaciones –MDE- del Instituto Geográfico Nacional, escala 1:25.000 y capas derivadas de pendientes, seno y coseno de las orientaciones). También se incluyen dos capas de información textural, que suele dar buenos resultados en ambientes heterogéneos (Berberoglu et al., 2007): albedo (primer componente de un Análisis de Componentes Principales) y el índice de vegetación de diferencia normalizada (NDVI).

2.2. Implementación

Para el almacenamiento de la información raster y su procesamiento se ha utilizado de forma integrada el SIG GRASS (Neteler & Mitasova, 2008) y el programa de análisis R (Venables, Smith & R Core Team, 2012). El proceso de trabajo se ha programado bajo los lenguajes ScriptBash y R, bajo un servidor Linux con 16 núcleos y 92 Gb. de memoria RAM.

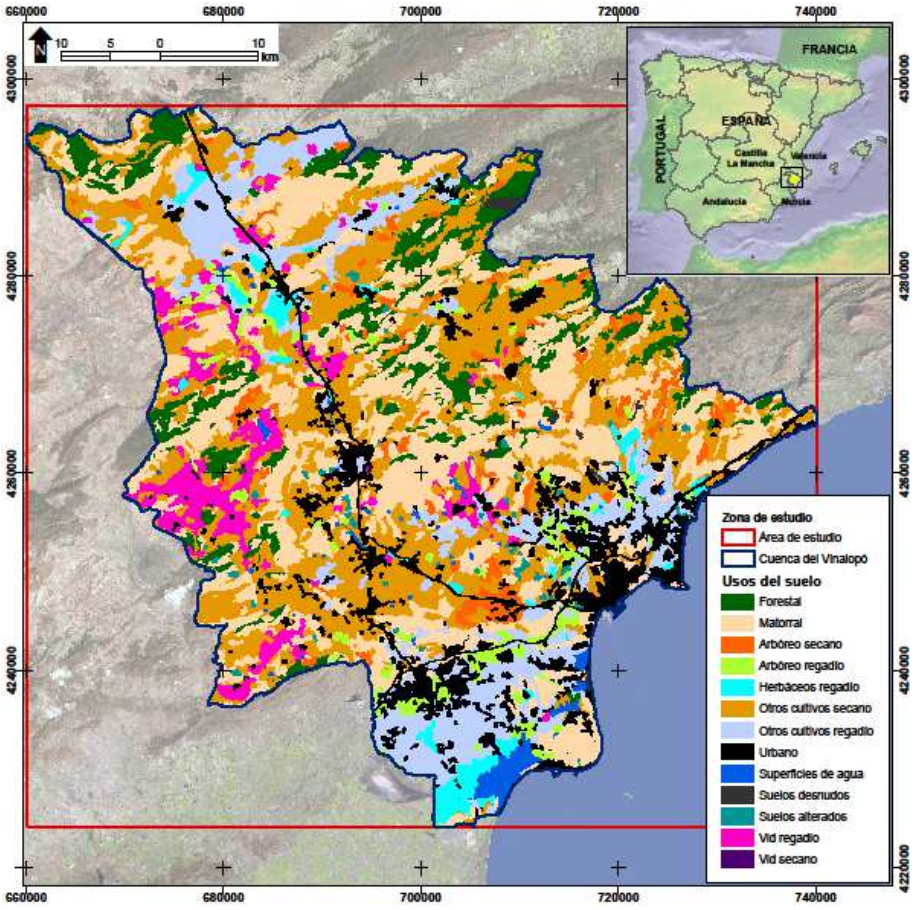

Fig. 1. Área de estudio y usos del suelo. Fuente: A partir de CORINE LandCover 2000.

2.3. Preclasificación

El tratamiento previo de las escenas sigue la metodología desarrollada en Alonso Sarría et al. (2010), mediante una corrección geométrica con 35 puntos de control localizados en toda la serie y una corrección atmosférica y de iluminación basada en los métodos de Chávez (1988) y del lambertiano C (Teillet et al., 1982).

Debido a la naturaleza multitemporal del trabajo, se han seleccionado como verdad terreno un conjunto de polígonos representativo de todas las fechas mediante los mapas de cultivos y aprovechamientos de España (MCA), Corine LandCover y ortofotografía del Sistema de Información Geográfica de Parcelas Agrícolas (SIGPAC), el Institut Cartogràfic Valencià (ICV) y el Plan Nacional de Ortofotografía Aérea (PNOA).

Para corregir posibles errores e incertidumbres (mala interpretación en la asignación del operador, errores puntuales dentro de los polígonos por otros usos discriminables) se ha programado un algoritmo consistente en una validación cruzada uno a uno (leave-one-out-cross-validation o LOOCV) de la predicción de cada

polígono respecto a la clase asignada; la predicción se obtiene a partir de una clasificación por ML calibrada con el resto de los polígonos, identificando mediante las tablas de contingencia los píxeles dentro del polígono clasificados en diferentes categorías e identificando su localización dentro éste. Tras este proceso, es posible corregir malas asignaciones o eliminar píxeles dentro del polígono si realmente la asignación realizada se debe a errores y no a la variabilidad propia de las observaciones.

Tras esta depuración, se han dividido los polígonos en dos subconjuntos (entrenamiento y validación) mediante muestreo aleatorio estratificado por la altitud, obteniendo 141 de entrenamiento (2.402 ha) y 73 de validación (1.004 ha).

Para evaluar los supuestos estadísticos de los métodos empleados se ha efectuado un análisis estadístico de los polígonos y las variables y sus transformadas mediante análisis exploratorio y contrastes de normalidad de Kolmogorov-Smirnof, tras lo cual se ha analizado la separabilidad entre clases mediante clusters para cada escena y las diferentes integraciones, utilizando como estimación de distancias entre clases el índice de divergencia transformada (Swain & Davis, 1978), de gran utilidad para comprobar la mejora en la separabilidad en función de las escenas integradas y las variables auxiliares.

2.4. Clasificación supervisada y validación

Las clasificaciones supervisadas se han obtenido mediante los siguientes algoritmos, desarrollados en Tso y Mather (2009) y Kuhn y Johnson (2013):

- *Máxima Verosimilitud* (ML) (Maselli y otros, 1992): método paramétrico más utilizado por su sencillez y rapidez de cómputo, basado en el uso del vector varianzas-covarianzas para estimar la probabilidad de pertenencia y asignación del píxel a una clase.
- *Sequential Maximum a Posteriori* (SMAP) (Bouman & Shapiro, 1994), método bayesiano basado en una clasificación multiescalar de campos aleatorios.
- *Random Forest* (RF) (Breiman, 2001), algoritmo no paramétrico basado en árboles de decisión.
- *Máquinas de Vectores Soporte* (SVM) (Cortes & Vapnik, 1995) con kernel radial, basado en la búsqueda de hiperplanos que minimizan la separabilidad entre clases.

RF y SVM utilizan algunos parámetros que pueden calibrarse mediante los datos disponibles. Siguiendo la metodología propuesta por Kuhn y Johnson (2013), se han calibrado mediante validación cruzada de k $k=10$ grupos o iteraciones (K-fold-cross-validation), repitiendo el proceso cinco veces; este método se aconseja por sus buenas propiedades de varianza aceptable, bajo sesgo y no requerir un excesivo coste computacional (Molinaro et al., 2005; Kuhn & Johnson, 2013).

Para la validación de la clasificación se ha utilizado el índice kappa estimado por intervalos de confianza al 95% como medida global del acuerdo; estas estimaciones permiten comparar si las clasificaciones son significativamente diferentes entre sí. Como medida de acuerdo por clase se ha estimado kappa condicional del usuario y del productor. En Foody (2009) se desarrollan estas medidas.

3. Resultados

Se han obtenido 132 clasificaciones anuales (combinación de cuatro escenas y entre 2004-2008 combinación de 2 por la calidad en las imágenes) más 20 adicionales en 2009 (una por estación e invierno-verano).

Para evaluar la mejora en la clasificación al incluir diversas escenas, se ha estimado kappa por intervalos de confianza al 95% para la clasificación de una escena representativa de cada estación en 2009, la integración de las escenas de invierno-verano y la integración de cuatro escenas (Fig. 2). Se observa para todos los algoritmos (incluyendo o no variables texturales y de relieve) cómo los resultados para cuatro estaciones son significativamente mayores al resto, seguida de la combinación de dos estaciones y las clasificaciones de verano; el mejor acuerdo ($\kappa = 0.87$) se obtiene en SMAP considerando variables texturales.

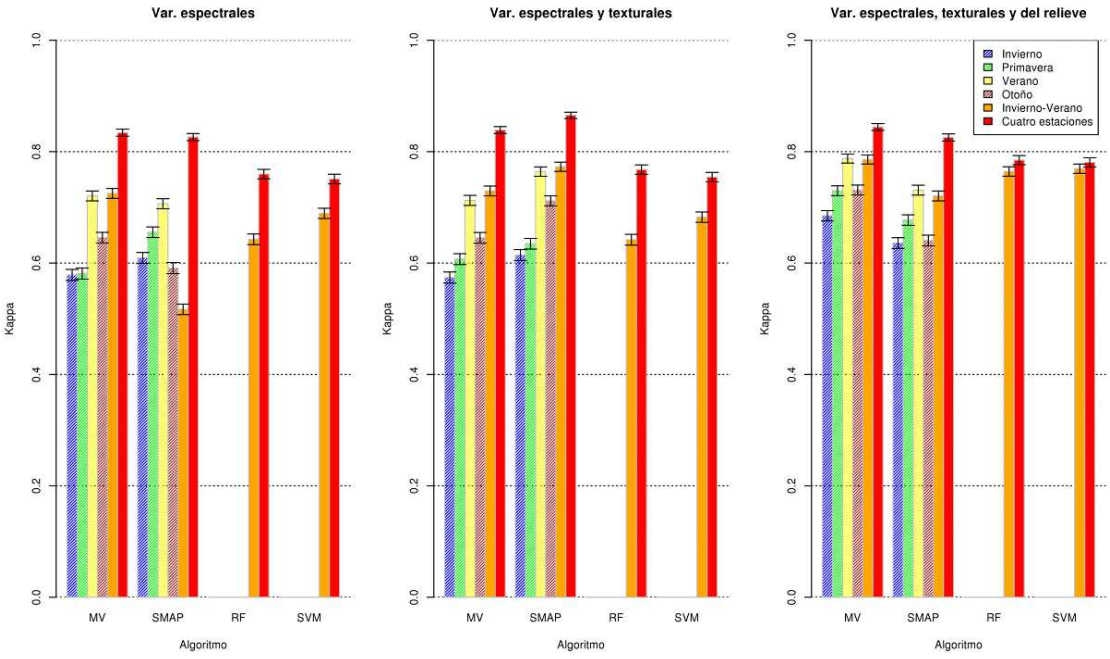

Fig. 2. Validación global de las clasificaciones en 2009 (kappa e intervalos de confianza al 95 %) de los algoritmos en función del número de variables, para cada estación del año, la combinación invierno-verano y la combinación de cuatro estaciones.

Al analizar la evolución de kappa anual de las clasificaciones integrando escenas para varias estaciones y la agrupación de los resultados por tipo de algoritmo y combinación de variables (Fig. 3 y 4), se observa cómo en la mayoría de los años y algoritmos utilizados se encuentra en el intervalo $\kappa \in \{0{,}6;0{,}8\}$ (buena concordancia), superando en muchos casos $\kappa = 0.8$ (concordancia muy buena). En general, kappa mejora en las clasificaciones con solo dos estaciones (2004-2008) y conforme se integran variables adicionales, al tiempo que disminuye la dispersión en los resultados. No obstante, al utilizar las variables relativas al relieve se observa una sobreestimación en muchos usos; en este caso, el análisis de importancia de las variables de RF desvela que éste da más importancia al relieve que al resto, generando errores en la predicción. SMAP

tiene una concordancia superior al resto al considerar variables espectrales y texturales y obtiene buenos rendimientos aunque se utilicen solo dos escenas por año; no obstante, en algunos casos se omite alguna clase en la predicción final (clasificaciones con kappa anormalmente bajo de la Fig. 3). Se observa también rendimientos ligeramente superiores de ML sobre RF y SVM, debido a la depuración de los polígonos de entrenamiento y validación, que disminuye la incertidumbre; en estos casos, métodos a priori más flexibles ante dichas incertidumbres, empeoran al sobre ajustar los datos (la fiabilidad en el entrenamiento ha sido siempre superior a 0,97). Al analizar kappa condicional se detecta mayor confusión entre los cultivos con menor separabilidad (vid con arbóreos dispersos y herbáceos entre sí). Las clases con mejores resultados son aquellas de vegetación natural y superficies de agua.

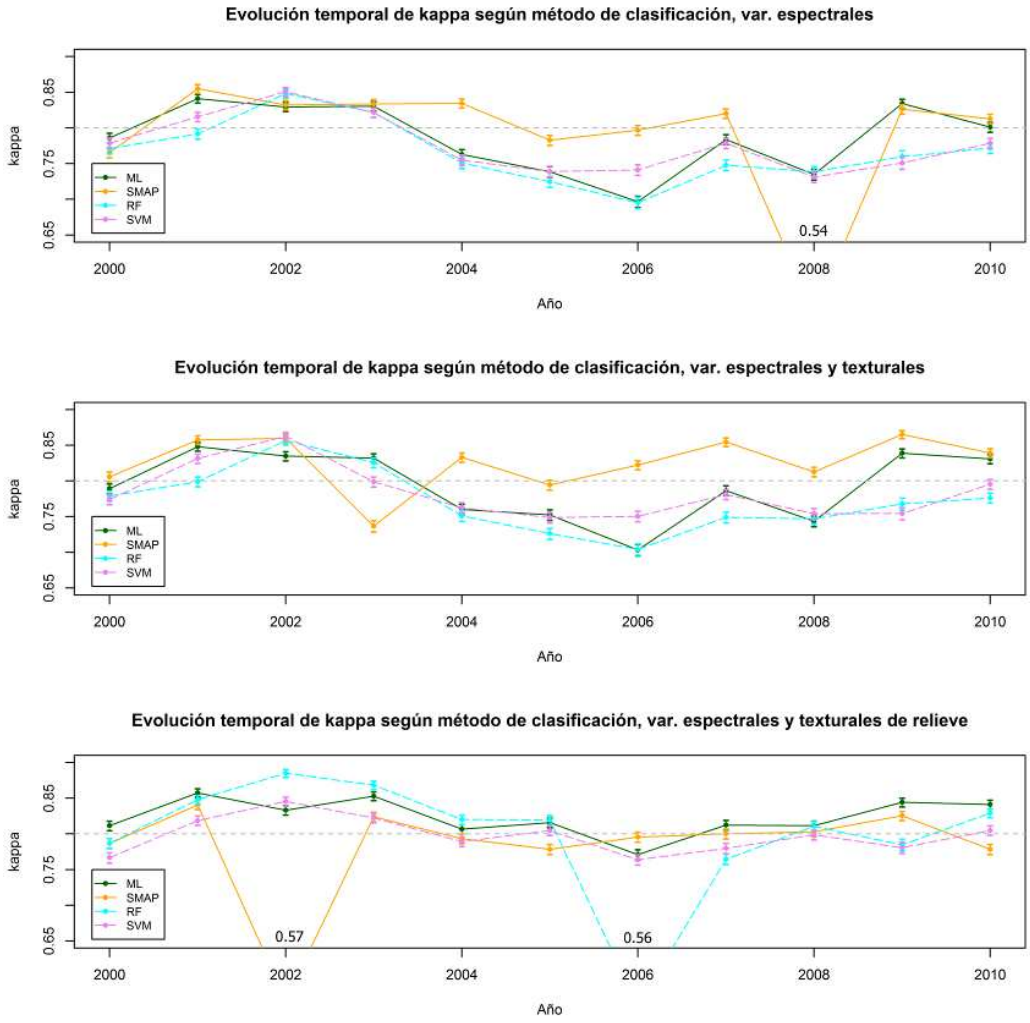

Fig. 3. Evolución de kappa por intervalos de confianza (95%) según la combinación de variables.

Como ejemplo de las clasificaciones resultantes, en la Fig.5 se presentan los mapas para 32 variables y los cuatro algoritmos. Resalta la sobre estimación de RF sobre urbanos o forestal en SVM y RF o las confusiones entre los diferentes usos de parecida respuesta espectral. También se puede ver una mayor adecuación visual en SMAP al clasificar a diversas escalas, corrigiendo efectos de sal y pimienta propios del resto clasificadores píxel a píxel.

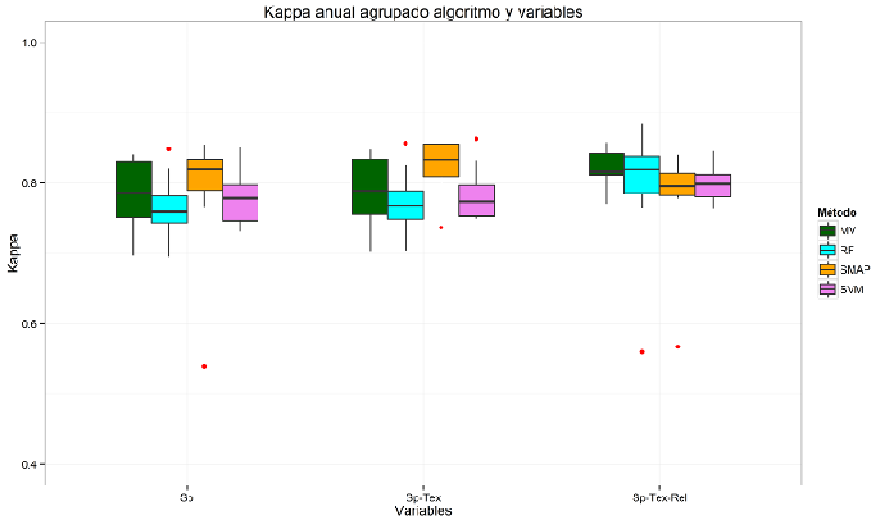

Fig. 4. Gráfico de cajas de kappa anual, agrupados por tipo de algoritmo y combinación de variables.

4. Conclusiones

Dada la complejidad de la zona de estudio y los objetivos marcados, los resultados obtenidos mediante la metodología implementada son satisfactorios, con $\kappa \geq 0.8$ en muchos casos.

La depuración de los polígonos mediante validación cruzada y su división para calibración y validación mediante muestreo aleatorio estratificado mejora los resultados en LM y SMAP frente a métodos más flexibles al eliminar valores anómalos y reducir las incertidumbres, obteniendo los mejores resultados con SMAP. Éste obtiene además un mejor ajuste en este tipo de paisajes fragmentados al ser un método que utiliza el contexto espacial para clasificar. RF y SVM obtienen una menor concordancia debido probablemente a un sobre-ajuste en el proceso de calibración.

Al combinar escenas de diferentes estaciones del año se observa una mejora de los resultados. Se recomienda por tanto el uso de cuatro escenas por año, o al menos dos que reflejen las diferencias fenológicas y déficit hídrico de la vegetación natural y los cultivos dominantes en este tipo de zonas semiáridas.

Al utilizar variables de apoyo, al incluirlas aumenta la separabilidad entre clases y la concordancia. Este hecho es especialmente interesante al utilizar variables texturales debido a la naturaleza heterogénea y fragmentada de los paisajes mediterráneos. Las variables relativas al relieve parecen mejorar kappa, sobre todo en zonas de vegetación natural y usos en los que condiciona su distribución, aunque en la comprobación visual y kappa condicional se observan errores en la clasificación por sobreestimación de usos antrópicos

propios de zonas llanas, menos influenciados por el relieve. Este hecho pone de manifiesto que kappa por si solo para evaluar las clasificaciones puede resultar engañoso y debe complementarse con comprobaciones visuales.

Por último se debe hacer referencia al coste computacional, uno de los factores más importantes para seleccionar un algoritmo concreto. El incremento al aumentar las variables no es muy alto, pero las diferencias entre los algoritmos son notables, desde menos de 10 minutos y 500Mb en la clasificación más costosa de ML o 40 minutos para SMAP hasta más de 20 horas y 20Gb en el caso de RF calibrado (aunque se reduce a 6 horas al paralelizar el proceso). En este contexto, tantas clasificaciones no pueden realizarse con un software comercial sin un considerable gasto económico, por lo que la implementación propuesta basada en software OpenSource, alta interoperabilidad, reproducibilidad del proceso y escalabilidad es una gran ventaja en este tipo de trabajos.

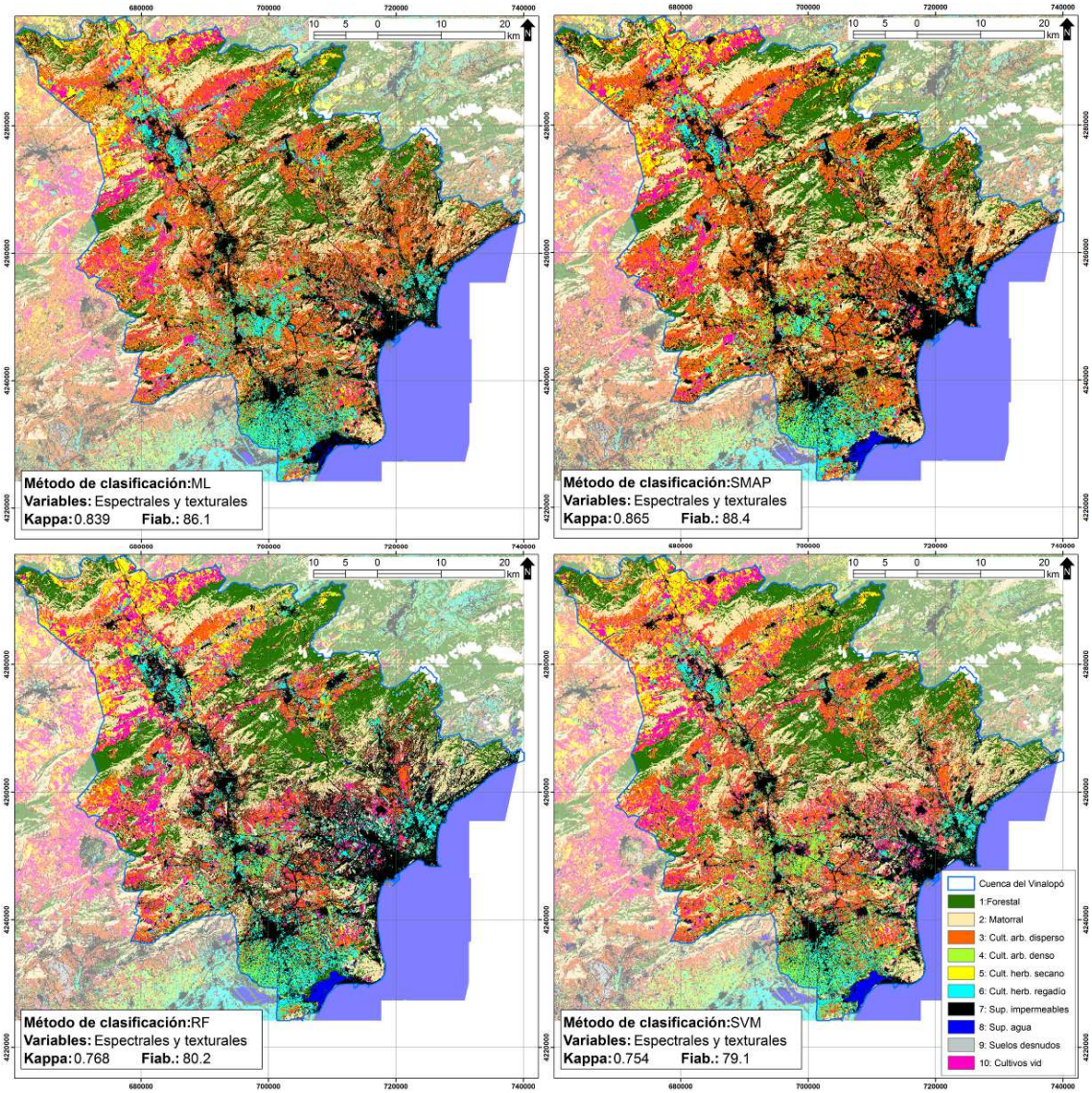

Fig. 5. Clasificaciones en 2009 para los cuatro algoritmos y 32 variables (espectrales de cuatro estaciones y texturales).

Agradecimientos

Esta investigación se ha desarrollado bajo el proyecto *Modelización Hidrológica en Zonas Semiáridas*, realizado por la Fundación Instituto Euromediterráneo del Agua, financiado por la Consejería de Educación, Ciencia e Investigación de la Región de Murcia.

De igual forma, ha sido parcialmente financiado por el Proyecto *Prometeo* de la Secretaría de Educación Superior, Ciencia, Tecnología e Innovación del Gobierno de Ecuador.

Referencias

Alonso Sarría, F., Gomariz Castillo, F. & Cánovas García, F. (2010). Análisis temporal de los cambios de usos del suelo en la Cuenca del Segura mediante teledetección. Implicaciones sobre la degradación. Cuaternario y Geomorfología, 24(3-4), 71-86.

Alrababah, M.A. & Alhamad, M.N. (2006). Land use/cover classification of arid and semiarid Mediterranean landscapes using Landsat ETM. International Journal of Remote Sensing, 27(13), 2703-2718. doi:10.1080/01431160500522700

Berberoglu, S., Curran, P.J., Lloyd, C.D. & Atkinson, P.M. (2007). Texture classification of Mediterranean land cover. International Journal of Applied Earth Observation and Geoinformation, 9(3), 322-334. doi:10.1016/j.jag.2006.11.004

Bouman, C. & Shapiro, M. (1994). A Multiscale Random Field Model for Bayesian Image Segmentation. IEEE Transactions on Image Processing, 3(2), 162-177. doi: 10.1109/83.277898

Breiman, L. (2001). Random Forests. Machine Learning, 45, 5-32. doi:10.1023/A:1010933404324

Cortes, C. & Vapnik, V. (1995). Support-vector network. Machine Learning, 20, 1-5. doi:10.1023/A:1022627411411

Chávez, P.S. (1988). An improved dark-object substraction technique for atmospheric scattering correction of multispectral data. Remote Sensing of Environment, 24, 459-479. doi:10.1016/0034-4257(88)90019-3

Foody, G.M. (2009). Classification accuracy comparison: Hypothesis tests and the use of confidence intervals in evaluations of difference, equivalence and non-inferiority. Remote Sensing of Environment, 113, 1658-1663. doi: 10.1016/j.rse.2009.03.014

Gislason, P.O., Benediktsson, J.A. & Sveinsson, J.R. (2006). Random Forests for land cover classification. Pattern Recognition Letters, 27(4), 294-300. doi:10.1016/j.patrec.2005.08.011

Kuhn, M. & Johnson, K. (2013). Applied Predictive Modeling. New York, NY: Springer.

Kumar, U., Dasgupta, A., Mukhopadhyay, C. & Ramachandra, T.V. (2012, 25-27 Octubre). Advanced Machine Learning Algorithms based Free and Open Source Packages for Landsat ETM+ Data Classification. Proceedings of the OSGEO-India: FOSS4G 2012-First National Conference Open Source Geospatial Resources To Spearhead Development And Growth.

Lu, D. & Weng, Q. (2007). A survey of image classification methods and techniques for improving classification performance. International Journal of Remote Sensing 28(5), 823-870. doi: 10.1080/01431160600746456

Maselli, F., Conese, C., Petkov, L. & Resti, R. (1992). Inclusion of prior probabilities derived from a nonparametric process into the maximum likelihood classifier. Geomatics 90 (National Conference & Exhibition)

Molinaro, A.M., Simon, R. & Pfeiffer, R.M. (2005). Prediction error estimation: a comparison of resampling methods. Bioinformatics, 21(15), 3301-3307. doi: 10.1093/bioinformatics/bti499

Mountrakis, G., Im, J. & Caesar, O. (2011). Support vector machines in remote sensing: A review. ISPRS Journal of Photogrammetry and Remote Sensing, 66(3), 247-259. doi: 10.1.1.187.2781

Neteler, M. & Mitasova, H. (2008). Open source GIS. A GRASS GIS approach (3ª Edición). New York, NY: Springer.

Rodriguez-Galiano, V.F., Ghimire, B., Rogan, J., Chica-Olmo, M. & Rigol-Sanchez, J.P. (2012). An assessment of the effectiveness of a random forest classifier for land-cover classification. ISPRS Journal of Photogrammetry and Remote Sensing, 67(0), 93-104. doi: 10.1016/j.isprsjprs.2011.11.002

Rozenstein, O. & Karnieli, A. (2011). Comparison of methods for land-use classification incorporating remote sensing and GIS inputs. Applied Geography, 31(2), 533-544. doi:10.1016/j.apgeog.2010.11.006.

Swain, P.H. & Davis, S.M. (1978). Remote Sensing: The Quantitative Approach. New York, NY: McGraw-Hill.

Tang, Q., Gao, H., Lu, H. & Lettenmaier, D.P. (2009). Remote sensing: Hydrology. Progress in Physical Geography, 33(4), 490-509. doi:10.1177/0309133309346650

Teillet, P.M., Guindon, B. & Goodenough, D.G. (1982). On the slope-aspect correction of multispectral scanner data. Canadian Journal of Remote Sensing, 8(2), 84-106.

Tso, B. y Mather, P. (2009). Classification Methods for Remotely Sensed Data (2ª Edición). New York, NY: Taylor & Francis.

Venables, W.N., Smith, D.M. & R Core Team (2012). An Introduction to R.