Universidad de Sevilla

Departamento de Lenguajes y Sistemas Informáticos
Escuela Técnica Superior de Ingeniería Informática

Intelligent techniques on LIDAR for environmental applications

Ph. D. Dissertation

Jorge García Gutiérrez

Sevilla, diciembre de 2011

Intelligent techniques on LIDAR for environmental applications

Jorge García Gutiérrez, 28805506F

jorgarcia@us.es

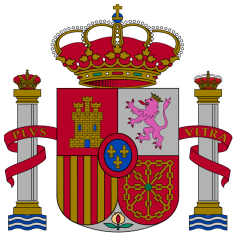
Supervised by Prof. Dr. José Riquelme Santos

Departamento de
Lenguajes y Sistemas Informáticos
Universidad de Sevilla

Thesis project submitted to the Department of Computer Languages and Systems of the University of Sevilla in partial fulfilment of the requirements for the degree of Ph.D. in Computer Engineering.

(Ph. D. Dissertation)

Tesis Doctoral parcialmente subvencionada por la Comisión Interministerial de Ciencia y Tecnología con el proyecto TIN2007-68084-C02-00.



D. José Cristóbal Riquelme Santos, profesor Catedrático de Universidad, adscrito al área de Lenguajes y Sistemas Informáticos,

CERTIFICA QUE:

D. Jorge García Gutiérrez, Ingeniero Informático por la Universidad de Sevilla, ha realizado bajo su supervisión el trabajo de investigación titulado:

INTELLIGENT TECHNIQUES ON LIDAR FOR ENVIRONMENTAL APPLICATIONS

Una vez revisado, autoriza la presentación del mismo como Tesis Doctoral en la Universidad de Sevilla y estima oportuno su presentación al tribunal que habrá de valorarlo. Dicha tesis ha sido realizada dentro del programa de doctorado Tecnología e Ingeniería del Software, con mención de excelencia MEE2011-0129 del Departamento de Lenguajes y Sistemas Informáticos de la Universidad de Sevilla. Igualmente, autoriza su presentación para la obtención de la acreditación de Doctorado Europeo.

Sevilla, diciembre de 2011.

Fdo.: José C. Riquelme Santos

A mi familia, por darme lo que soy,
y a ti Laura, por hacerme mirar más lejos.

Agradecimientos

Desde estas líneas quiero mostrar mi más sincero agradecimiento a todas aquellas personas que, de alguna manera, han contribuido tanto en el plano profesional como en el personal, al desarrollo de esta tesis doctoral.

Especialmente quiero agradecer a Jesús Aguilar el haberme iniciado en este mundo de la investigación y toda la confianza que desde el principio puso en mí. Quiero agradecer a mi director de tesis, José Riquelme, grande como investigador pero más grande como persona, su apoyo incondicional y su cercanía durante estos años que me han hecho sentir que, en el despacho de al lado, tenía más a un amigo que a un catedrático.

También quiero agradecer a todos los miembros del grupo de investigación BIGS y a todos mis compañeros del Departamento de Lenguajes y Sistemas Informáticos su apoyo, tanto moral como técnico.

Muy especialmente tengo que acordarme de los «foráneos» Luis Gonçalves de la Universidad de Porto, Mariano García de la Universidad de Alcalá y Eduardo González, Laura Barreiro y David Miranda de la Universidad de Santiago. Sin su ayuda, el tránsito de este informático por la senda medioambiental no habría sido posible.

Finalmente dedico esta memoria a mis padres, por darme todo lo que soy y lo que tengo, a mi hermano, por su fuerza y su cariño, y, muy especialmente, a Laura, que siempre me animó a dar un paso más y que apostó por mí desde el día que me conoció.

Abstract

This thesis is closely related to the field of intelligent techniques and soft computing. Concretely, the proposed techniques try to improve the results obtained by classical algorithms used in the area of environmental remote sensing, especially using data from sensors on LIDAR (Light Detection and Ranging). Throughout this document, we show novel approaches to explote LIDAR data for tematic mapping and biomass estimation. To this end, our work suggests the use of new or well-known algorithms from data mining on LIDAR isolated or fused with other data sources.

Supervised techniques have traditionally demonstrated high quality for coverage or land use mapping. Thus, we initially show a comparative study of classical classification techniques on LIDAR data source only. We also suggest a set of basic statistics to extract meaningful information from LIDAR to make a per-field classification. Later, we try to improve the results by means of data fusion between LIDAR and aerial images (orthophotography and multispectral images) and we show a comparative study among the most extended supervised techniques in data fusion. Finally, we propose a novel approach called SVMNNS for dealing with LIDAR fused with other data sources such as aerial imagery. It is based on the concepts of ensemble (stacking) and contextual classification. This contextual classifier improves an initial result (achieved by a base classifier) in successive iterations using information from a surrounding area of each instance. SVMNNS uses a support vector machine as an initial classifier and the nearest neighbor technique to refine the support vector's results. Our method had the best results when its performance was compared with a cutting-edge contextual technique.

Regression is another very important technique for LIDAR researchers. This document shows the results obtained by evolutionary feature selection and regression trees in order to generate models for biomass estimation. Their performances are compared with those of stepwise feature selection and multiple linear regressions (MLR) respectively which are the traditional techniques used by environmental engineers. The results show classical techniques can be overcome easily without loss of clarity in the model.

The main contributions shown in this Ph. D. dissertation are directly related to important environmental applications. For example, land use and land cover maps are very important products for land management, flood control or forest monitoring whilst regression models on LIDAR are used to monitor commercial

and natural wood areas. From the foregoing it is clear the importance of innovation in the methods that generate the above products and the need to adapt and apply the most powerful techniques of data mining on LIDAR sensor data.

Resumen

Esta propuesta de tesis está estrechamente relacionada con el campo de las técnicas inteligentes y el soft computing. A lo largo de los nueve capítulos que la conforman, se propone utilizar estas nuevas disciplinas para mejorar los resultados obtenidos por las metodologías y algoritmos clásicos utilizados en el área de la teledetección ambiental, especialmente cuando se aplican a datos provenientes de sensores LIDAR (Light Detection and Ranging). Más concretamente, este documento se centra en la manera de explotar los datos LIDAR para tareas de generación de mapas temáticos y estimación de biomasa y con este fin, nuestro trabajo sugiere el uso de algoritmos nuevos o bien conocidos del mundo de la minería de datos sobre LIDAR, de manera aislada o fusionado con otras fuentes de datos.

Las técnicas de clasificación supervisada tradicionalmente han demostrado ser muy útiles a la hora de generar mapas de coberturas y usos del terreno. Este trabajo muestra inicialmente un estudio comparativo de las técnicas de clasificación clásicas sobre una fuente de datos LIDAR de manera aislada. También, describe un conjunto de estadísticas básicas para extraer información significativa de LIDAR cuando se aplica un enfoque orientado a polígono. Más tarde, se trata de mejorar los resultados a través de la fusión de datos entre LIDAR e imágenes aéreas (ortofotos e imágenes multiespectrales) y se muestra un estudio comparativo entre las técnicas de clasificación más extendidas en la fusión de datos. Por último, se propone una nueva aproximación denominada SVMNNS para tratar datos LIDAR fusionados con otras fuentes como las imágenes aéreas. SVMNNS se basa en los conceptos de stacking (apilamiento) y clasificación contextual. Este clasificador contextual mejora una clasificación inicial (que se logra mediante una técnica supervisada estándar) en sucesivas iteraciones utilizando la información de la vecindad del elemento a clasificar. SVMNNS, en concreto, utiliza una máquina de vector soporte (SVM) como clasificador inicial y la técnica de los vecinos más cercanos en la fase de refinamiento. Los resultados obtenidos en diversas áreas de estudio muestran que SVMNNS tiene mejores prestaciones que otros clasificadores contextuales basados en refinamiento de SVMs.

La regresión es otra técnica muy importante para los investigadores LIDAR. Este documento muestra los resultados obtenidos cuando se aplica una selección de características evolutiva y se aplican árboles de regresión con el fin de generar modelos para la estimación de biomasa. Sus prestaciones se comparan con los

de la selección de características por pasos y la regresión lineal múltiple (MLR) respectivamente, que son las técnicas tradicionalmente utilizadas por los ingenieros ambientales. Los resultados muestran que las técnicas clásicas se pueden mejorar fácilmente sin pérdida de claridad en el modelo.

Las principales contribuciones que se muestran en esta tesis doctoral están directamente relacionadas con importantes tareas en el ámbito medioambiental. Por ejemplo, los mapas de uso del suelo y coberturas del terreno son productos muy importantes para la gestión del territorio, el control de inundaciones o la monitorización de los bosques. Por otro lado, los modelos de regresión a partir de LIDAR son una herramienta clave para monitorizar zonas boscosas naturales o comerciales. De lo anterior, se desprende la gran importancia que tiene la innovación en los métodos clásicos que tradicionalmente generan estos productos y en este sentido, la gran sinergia que puede existir entre las técnicas más potentes de la minería de datos y los datos del sensor LIDAR.

Índice de figuras

Índice de cuadros

Índice general

Capítulo 1

Introducción

In omnibus negotiis prius quam aggrediare,
adhibenda est praeparatio diligens.
(En todos los asuntos, antes de comenzar,
se debe hacer una preparación diligente).
Cicerón.

1.1. Introducción

Hoy en día, en los albores del nuevo milenio, la capacidad de generar información del ser humano llega a límites nunca conocidos y pocas veces imaginados. Algunos autores sostienen que el advenimiento de internet es comparable a la invención de la imprenta de Guttenberg y que al igual que, en su momento, este hecho definió un hito histórico que condujo al cambio de la Edad Media a la Edad Moderna, la revolución que hoy en día vivimos hará lo propio respecto a la Edad Contemporánea (González-Pons, 2000).

Para que este nuevo paso en la historia del hombre se plasme de manera real, la mayor cantidad de información debe ser también acompañada de una mayor capacidad en el tratamiento de dicha información. De esta forma, el objetivo último debería ser la obtención de conocimiento útil a partir de los grandes volúmenes de información disponibles. Esta idea es el motor de una gran comunidad que se dedica a lo que se conoce como minería de datos y que hace uso de otras disciplinas como el aprendizaje automático y el soft computing para extraer este conocimiento a partir de diversas fuentes de información. Durante los últimos años, se han presentado multitud de técnicas en el mundo de la minería de datos y sus aledaños, introduciéndose poco a poco en otras áreas del conocimiento como la biología, la medicina y la industria en general (Hernández Orallo, 2004).

El mundo de la teledetección y, en particular, de los sensores LIDAR (Light Detection and Ranging) ha sido un área industrial muy potente en los últimos tiempos. La necesidad de generar productos útiles a partir de los datos adquiri-

dos de forma remota ha hecho que la actividad investigadora en este campo se dispare. Aún así, todavía no se encuentran verdaderos sistemas que «minen» las grandes cantidades de datos que hoy en día se generan. Esta investigación parte pues, de la premisa de que cualquier base de datos de teledetección y en particular las generadas a partir de los datos LIDAR es susceptible de producir mayor conocimiento cuando las técnicas de minería de datos les son aplicadas. Así, aprovechando las sinergias que en nuestro entorno se dan gracias a entidades como la Consejería de Medio Ambiente de la Junta de Andalucía, planteamos la aplicación de técnicas inteligentes novedosas sobre datos provenientes de sensores LIDAR.

Las técnicas usadas generalmente en el entorno LIDAR suelen ser algoritmos ad-hoc poco relacionados con el aprendizaje automático y algunas técnicas paramétricas como la regresión lineal múltiple. Si se habla de teledetección de manera general, sí que es posible encontrar en la bibliografía multitud de aproximaciones basadas en aprendizaje supervisado, pero existen muy pocas que tengan en cuenta las especiales características del sensor LIDAR y su especial sinergia con otros sensores ópticos. Este nuevo marco de trabajo que se abre sobre los datos LIDAR es el punto de partida de esta investigación, un marco que hace especial énfasis en la necesidad de generar nuevas aproximaciones relacionadas con la minería de datos capaces de obtener mejores resultados tanto sobre datos LIDAR de manera aislada como sobre su fusión con otros sensores.

1.2. Objetivos

En los siguientes puntos, se resumen los objetivos marcados al inicio de esta tesis doctoral y los resultados obtenidos a lo largo del período de investigación:

- **Generación de metodologías**. Nuestro objetivo prioritario en este trabajo de investigación fue generar nuevas metodologías que ayudaran a los expertos en el área de medio ambiente a mejorar los productos derivados de sensores remotos y en concreto, de sensores activos como LIDAR. Como veremos más adelante, la introducción de técnicas inteligentes que traten los datos LIDAR es todavía escasa a pesar de que esta tecnología se ha usado profusamente en los últimos 20 años. Este tipo de sensores viene siendo clave para diversas tareas entre las que destacan el desarrollo de modelos digitales del terreno, la generación de mapas de coberturas y la definición de índices para estimar variables tan importantes como la biomasa en áreas forestales. En este trabajo, se plantean soluciones novedosas basadas en técnicas inteligentes del mundo del aprendizaje automático y el soft computing para el desarrollo de mapas temáticos y modelos de estimación de biomasa. En esta memoría, además de dichas soluciones, se presentan resultados comparativos respecto a otras técnicas.

- **Publicación de resultados**. Una de nuestras prioridades fue en todo momento publicitar nuestro trabajo en foros tanto nacionales como internacionales. Hasta el momento, hemos publicado alrededor de 20 artícu-

los técnicos en conferencias nacionales, congresos internacionales y talleres, así como en dos revistas internacionales de impacto (García-Gutiérrez et al., 2011a, 2012b). Además, se han desarrollado otras aplicaciones que han generado trabajos sometidos o por someter en los próximos tiempos (Parejo et al., 2012a; García-Gutiérrez et al., 2012a,c, To be submitted).

- **Transferencia de los resultados.** Tenemos la intención de integrar parte de los resultados de esta tesis doctoral en una aplicación que genere productos a partir de datos LIDAR y de su fusión con otras fuentes de información. Para ello, planeamos desarrollar una infraestructura completa que mostrará nuestros resultados y todas las técnicas descritas en esta memoria. Los algoritmos necesarios han sido desarrollados en el lenguaje de programación JAVA. Este lenguaje es ya usado en frameworks de desarrollo de herramientas SIG tan importantes como Sextante (Olaya, 2006). Este hecho podría ser clave para transferir nuestros conocimientos a las empresas que participan en proyectos de investigación similares y de esta forma, obtener retroalimentación en nuestro trabajo post-doctoral.

1.3. Período de investigación

1.3.1. Principales aportaciones originales

Artículos publicados en revistas de impacto

Garcia-Gutierrez J, Mateos-Garcia D, Riquelme-Santos JC. Evor-Stack: a label-dependent evolutive stacking on remote sensing data fusion. Neurocomputing, 75 (1), pp. 115-122, 2012. **JCR (2010) IF: 1,429**.

Garcia-Gutierrez J, Gonçalves-Seco L, Riquelme-Santos JC. Automatic environmental quality assessment for mixed-land zones using lidar and intelligent techniques. Expert Systems with Applications, 38 (6), pp. 6805-6813, 2011. **JCR (2010) IF: 1,924**.

Artículos publicados en congresos internacionales

García Gutiérrez J, Mateos García D, Riquelme Santos JC. A non parametric approach for acurate contextual classification of LIDAR and imagery data fusion. HAIS 2012, *In press*, 2012.

Barreiro Fernández L, Buján S, Miranda D, García Gutiérrez J, González Ferreiro E. Location of mature forests using lidar data and aerial orthophotography, Las reservas de la biosfera como estrategia territorial de sostenibilidad, pp. 388-389, 2011.

García Gutiérrez J, González-Ferreiro E, Mateos García D, Riquelme Santos JC, Miranda D. A Comparative Study between Two Regression Methods on LiDAR Data: A Case Study. HAIS 2011, LNAI 6679, pp. 311-318, 2011.

García Gutiérrez J, Martínez Álvarez F, Riquelme Santos JC. Using remote data mining on LIDAR and imagery fusion data to develop land cover maps. IEA-AIE 2010, LNAI 6096, pp. 378-387, 2010.

García Gutiérrez J, Mateos García D, Riquelme Santos JC. A SVM and k-NN Restricted Stacking to Improve Land Use and Land Cover Classification. HAIS 2010, LNCS 6077, pp. 493-500, 2010.

Mateos García D, García Gutiérrez J, Riquelme Santos JC. Label Dependent Evolutionary Feature Weighting for Remote Sensing Data. HAIS 2010, LNCS 6076, pp. 272-279, 2010.

Barreiros Fernández L, García Gutiérrez J et al. Land Cover Classification of Forest Areas Using Lidar and Spectral Data. ForestSat, pp. 159-162, 2010.

García Gutiérrez J, Gonçalves Seco L, Riquelme Santos JC. Decision Trees on Lidar to Classify Land uses and Covers. ISPRS Workshop Laserscanning 2009, pp. 323-328, 2009.

García Gutiérrez J, Martínez Álvarez F, Riquelme Santos JC. Remote Mining: From Clustering to DTM. Silvilaser 2008: 8th International Conference on Lidar Applications in Forest Assessment and Inventory, pp. 389-397, 2008.

Artículos publicados en congresos nacionales

García Gutiérrez J, Martínez Álvarez F, Riquelme Santos JC. Aprendizaje Automático Sobre Datos Lidar para Monitorizar el Avance Urbano en el Medio natural. Actas de la XIII Conferencia de la Asociación Española para la Inteligencia Artificial (CAEPIA 2009), pp. 581-590, 2009.

1.3.2. Proyectos de investigación

Esta investigación ha sido parcialmente subvencionada por diferentes entidades con los siguientes proyectos:

- **Proyecto:** Modelos Avanzados en Minería de Datos: Escalabilidad y Aplicación Biológica.

  Entidad Financiadora: Ministerio de Educación y Ciencia.

  Entidades participantes: Universidad Pablo de Olavide, Universidad de Sevilla.

  Duración: 2007-2011.

  Investigador coordinador: Jesus S. Aguilar Ruiz.

- **Proyecto:** Análisis inteligente de información medioambiental.

Entidad Financiadora: Ministerio de Educación y Ciencia.

Entidades participantes: Universidad de Sevilla.

Duración: 2012-2014.

Investigador coordinador: José C. Riquelme Santos.

- **Proyecto:** Modelos Avanzados para el Análisis Inteligente de Información. Aplicación a Datos Biomédicos y Medioambientales.

  Entidad Financiadora: Junta de Andalucía.

  Entidades participantes: Universidad de Sevilla.

  Duración: 2012-2016.

  Investigador coordinador: Cristina Rubio Escudero.

1.4. Organización

El contenido de esta memoria de investigación se completa con los siguientes capítulos:

- **Capítulo 2: Teledetección y sensores LIDAR.** En este capítulo, se presentan conceptos básicos relacionados con la aplicación de técnicas informáticas a los datos provenientes del sensor LIDAR y de teledetección ambiental en general.
- **Capítulo 3: Aprendizaje automático en teledetección.** Seguidamente, se describen las metodologías más referenciadas relacionadas con el aprendizaje automático en el contexto de la teledetección.
- **Capítulo 4: Datos de estudio.** Posteriormente, se describen los datos cedidos por diversas organizaciones para llevar a cabo la investigación que se recoge en este documento.
- **Capítulo 5: Mapas temáticos a partir de datos LIDAR.** Este capítulo establece la base de este trabajo de investigación. Así, muestra cómo LIDAR puede ser usado para generar mapas de coberturas de manera efectiva mediante técnicas de aprendizaje supervisado clásicas. Además, presenta un estudio comparativo que dio como resultado una publicación en revista de impacto (García-Gutiérrez et al., 2011a).
- **Capítulo 6: Fusión de sensores LIDAR e imágenes.** El capítulo anterior se completa aquí mostrándose las principales técnicas de aprendizaje supervisado aplicadas a la fusión de datos LIDAR. Además, se adjunta una comparativa que recoge sus resultados cuando se aplican sobre la combinación LIDAR e imágenes aéreas con el objetivo de generar mapas de coberturas de alta resolución. Dichos resultados mostrados forman parte de un artículo que se someterá en breve a evaluación (García-Gutiérrez et al., 2012a).

- **Capítulo 7: Uso del contexto sobre fusión de sensores.** En este capítulo, se presenta un clasificador contextual basado en la combinación de técnicas de aprendizaje supervisado, aplicado a fusión de datos LIDAR e imágenes aéreas. Los resultados de este clasificador se comparan con los obtenidos por el clasificador basado en contexto más utilizado en la literatura. Recientemente, el clasificador presentado fue extendido mediante computación evolutiva para el cálculo de pesos en los atributos, dando lugar a una publicación en revista de impacto (García-Gutiérrez et al., 2012b).

- **Capítulo 8: Estimación de variables forestales.** A continuación, se compara la regresión lineal múltiple, técnica usada tradicionalmente cuando se aborda la búsqueda de índices de vegetación para el cálculo de biomasa a partir de datos LIDAR, con otras técnicas como los árboles de regresión. Además, se describe un algoritmo evolutivo para seleccionar los mejores predictores y comparamos sus resultados con los obtenidos por una selección *stepwise*, muy utilizada en el área medioambiental.

- **Cápitulo 9: Conclusions and future work.** Finalmente, se resumen las conclusiones obtenidas durante esta investigación y se proponen futuras líneas de investigación relacionadas con LIDAR en las que los resultados obtenidos nos impulsan a seguir trabajando.

Capítulo 2

Teledetección y sensores LIDAR

A la luz correcta, en el momento correcto, todo es extraordinario.
Aaron Rose.

2.1. Conceptos básicos de la teledetección

La *Teledetección* (del inglés Remote Sensing) (Pinilla, 1996) se puede definir como la técnica que permite obtener información de objetos a distancia, mediante algún tipo de dispositivo, sin que exista un contacto material y que, de manera general, están situados sobre la superficie terrestre. Aunque la definición de teledetección originalmente sólo abarcaba el proceso de adquisición de la información, poco a poco, su significado se ha ido extendiendo (Chuvieco, 2008) hasta abarcar también el conjunto de técnicas que se aplican para extraer conocimiento útil de la información original teledetectada.

La teledetección de cualquier fenómeno requiere el concurso de tres factores: una fuente energética de radiación electro-magnética, la interacción de dicha radiación con la superficie del objeto a estudiar y un sistema de detección que reciba la radiación reflejada. A partir de este punto, las nuevas atribuciones conferidas al concepto teledetección entran en funcionamiento. De esta forma, es posible actuar sobre la información del sensor de manera manual, es decir un experto genera el conocimiento, o de manera automática mediante tratamientos digitales que realicen el proceso de generación del conocimiento útil para el usuario final. El proceso completo puede verse resumido en la Figura 2.1.

Para que la observación en teledetección sea posible, es pues necesario que aunque sin contacto material, exista algún tipo de interacción entre los objetos y el dispositivo sensor. En general, dicha interacción va a ser un flujo de radiación que parte de los objetos y se dirige hacia el sensor. Este flujo puede ser, en cuanto a su origen, de tres tipos:

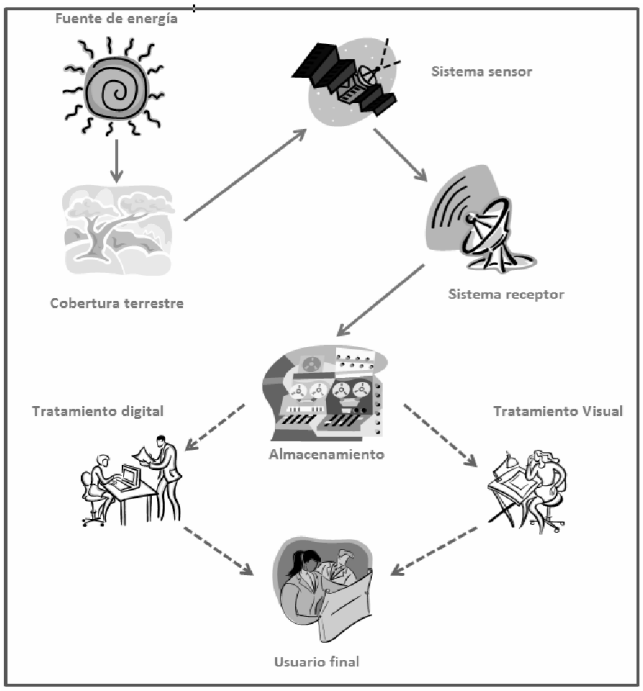


Figura 2.1: Proceso completo de adquisición de conocimiento en teledetección (adaptado de Chuvieco (2008)).

- Radiación solar reflejada por los objetos.
- Radiación terrestre emitida por los objetos.
- Radiación emitida por el sensor y reflejada por los objetos.

En cualquier caso, la interacción entre la cobertura terrestre y el sensor se transmite en forma de radiación electro-magnética. Según la teoría ondulatoria, este tipo de energía se transmite en forma de ondas que vienen definidas por, entre otros elementos, la longitud de onda (λ) y la frecuencia (υ). Además, gracias a la teoría cuántica y la dualidad onda-corpúsculo de la luz, es posible conocer la cantidad de energía (Q) que transmite un fotón (partícula de luz) en función de su frecuencia. Teniendo en cuenta que el producto entre λ y υ es igual a la constante c (velocidad de la luz), la energía que un fotón transmite queda definida por la Ecuación 2.1 donde h es la constante de Planck ($6{,}6 \times 10^{-34}$ J s).

$$Q = h\upsilon = h(\frac{c}{\lambda}) \tag{2.1}$$

De lo anterior, se deduce que para una onda que defina la interacción entre un objeto detectado y un sensor, a menor λ, mayor cantidad de energía transmitida

y viceversa. Este hecho es básico en teledetección ya que la diferencia en la energía registrada por el sensor será la clave para poder diferenciar los elementos teledetectados.

2.1.1. Espectro electro-magnético

Cualquier tipo de energía radiante puede clasificarse en función de su λ o υ. La sucesión de valores de longitud de onda es continua pero se pueden establecer un conjunto de bandas en donde la radiación electro-magnética manifiesta un comportamiento similar. A la organización de longitudes de onda o frecuencias en forma de bandas, se la suele conocer como espectro electro-magnético (Figura 2.2).

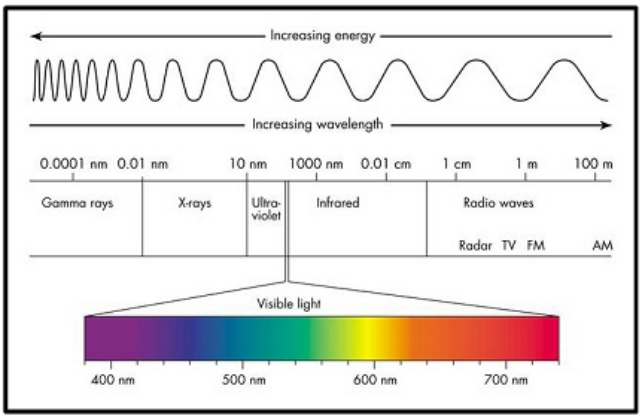


Figura 2.2: Espectro electro-magnético.

Existe un conjunto de frecuencias y longitudes de onda que tiene especial interés para la teledetección debido a que la mayor parte de los sensores trabajan en alguna o varias de dichas frecuencias. Así, la terminología más extendida (Chuvieco, 2008) habla de:

- *Espectro visible.* Es el conjunto de longitudes de onda percibibles por el ojo humano. Dentro de esta región, suele distinguirse entre las bandas azul (400-500 nm), verde (500-600 nm) y roja (600-700 nm), debido a que son los colores primarios que el ojo humano detecta en cada una de ellas.

- *Infrarrojo cercano.* Esta franja (700-1300 nm) es también conocida como infrarrojo próximo, reflejado o fotográfico, puesto que parte de él puede detectarse con sensores ligeramente diferentes a los que se usan para la detección del espectro visible.

- *Infrarrojo medio.* Zona en la que se entremezcla la radiación reflejada denominado en inglés Short Wave Infrarred (SWIR, 1300-2500 nm) y la emitida por los objetos de la superficie terrestre.

- *Infrarrojo lejano o térmico.* En esta zona se concentra el conjunto de emisiones (energía propia de los objetos y no reflejada) de la superficie terrestre. Esta franja del espectro es especialmente interesante para detectar el calor generado por las distintas cubiertas.

- *Micro-ondas.* Este conjunto de longitudes de ondas es especialmente utilizado en sensores que deban atravesar coberturas nubosas ya que su mayor longitud de onda reduce la probabilidad de que impactos contra partículas atmosféricas neutralice la señal completa.

En nuestro caso, nos centraremos en el espectro visible y en el infrarrojo cercano ya que son las bandas registradas principalmente por los distintos tipos de sensores ópticos y LIDAR, materia principal de trabajo en esta propuesta de tesis doctoral.

2.1.2. Coberturas en el espectro visible e infrarrojo cercano

En esta subsección, nos centraremos en tres tipos de coberturas claves para los distintos estudios ambientales: vegetación, suelo y agua. Cada una de estas coberturas tienen una respuesta concreta (reflectividad) cuando se exponen a señales con una frecuencia determinada. Así, para cada cubierta obtendremos lo que se conoce como *firma espectral* que es la respuesta que cualquier superficie del mismo tipo devuelve a un sensor como respuesta a una señal determinada.

Es importante destacar que, aunque las coberturas de un mismo tipo tienen una misma o similar respuesta ante la luz recibida, la señal que recibe el sensor como respuesta no tiene por qué ser la misma. Este hecho se explica porque la señal recibida se ve afectada por otros parámetros entre los que destacan: (i) las condiciones atmosféricas (atenuación de la señal e influencia de la irradiancia vecina), (ii) situación geográfica de la cubierta (efectos de sombras y ángulos de incidencia) y (iii) geometría de la observación (superficies especulares, lambertianas y comportamiento real mixto) (Chuvieco, 2008).

Teniendo en cuenta todo lo anterior, podemos analizar la respuesta de las principales cubiertas de la superficie terrestre en los espectros visibles e infrarrojo cercano.

- Vegetación. La vegetación tiene un comportamiento variable. Cuando los niveles de clorofila son altos y no está sometida a estrés hidrico, la vegetación tiene niveles altos de reflectividad en infrarrojo cercano y verde (de ahí que se observe al ojo humano como ese color) y bajo para el resto de bandas (rojo, azul). Sin embargo, cuando la vegetación sufre por sequía o por cambios estacionales, la vegetación pierde clorofila y el color verde desaparece por una menor reflectividad.

- Suelo. El suelo desnudo tiene un comportamiento mucho más uniforme que el de la vegetación que depende de variables como la relación copa/tronco entre otras. El color del suelo y su reflectividad dependerán del

tipo de composición que tenga pero, como norma general, tendrá mayor reflectividad cuanto menos cantidad de humedad contenga y viceversa.

- Agua. Las superficies acuáticas absorben o transmiten la mayor parte de la radiación óptica que reciben, siendo menor su reflectividad cuanto mayor sea la longitud de onda. Al igual que el suelo, suele ser muy regular en su respuesta y tendrá su mayor reflectivad en agua clara donde el azul (menor λ) será el color predominante y tendrá una respuesta prácticamente nula para la zona del infrarrojo. Si la zona es poco profunda, se puede tener cierto grado de información del fondo. La nieve es la excepción ya que se comporta de manera contraria, teniendo unas tasas de reflectividad muy altas en las bandas visibles (de ahí su color blanco) y menor en el infrarrojo.

2.2. Tipos de sensores y resolución

Existen diversas formas de clasificar los sistemas sensores pero la clasificación inicial más extendida suele ser aquélla que habla de sistemas activos o pasivos según se trate de generadores de la radiación que captan (por ejemplo, radar), o sólo meros receptores de la respuesta del objeto iluminado por una fuente externa que, en general, será el sol (ortofotografía aérea o imágenes de satélite).

Un concepto muy importante en teledetección es la *resolución*. Se denomina resolución de un sistema a su capacidad para discriminar información en un objeto detectado. En teledetección, existen diversos tipos de resolución:

- *Resolución espacial*: Capacidad por la cual, se puede distinguir el objeto más pequeño posible en una imagen.

- *Resolución espectral*: Capacidad del sensor para discriminar la radiancia detectada en distintas longitudes de onda del espectro electromagnético. En este caso, se entiende por radiancia, el total de energía radiada en una determinada dirección por unidad de área y ángulo sólido de medida.

- *Resolución radiométrica*: Capacidad del sensor para ser sensible a distintos niveles o intensidades de radiancia espectral.

- *Resolución temporal*: Capacidad del sensor para discriminar los cambios temporales sufridos por la superficie de estudio.

Existe un cierto antagonismo entre algunos tipos de resolución. Por ejemplo, para obtener mayor resolución temporal, se necesitan órbitas más altas, lo que implica menor resolución espacial. Por otro lado, aumentar la resolución, especialmente la espectral, implica un incremento del tamaño de los datos lo que hace en ocasiones muy costoso el trasiego de dicha información desde el origen hasta el destino y su posterior procesamiento.

Figura 2.3: Ejemplo de imagen obtenida de un sensor remoto.

2.3. Distribución de los datos remotos

Los datos remotos suelen suministrarse en forma de imágenes digitales (ver Figura 2.3) en las que cada píxel contiene la información espectral que el sensor ha capturado para un área concreta. Dicha información se agrupa en forma de canales que aglutinan la información de una o varias bandas del espectro. Como hemos visto, cada banda corresponde a la información recogida en un intervalo de longitud de onda determinado. Así, podemos encontrar canales formados por las bandas visibles roja, verde y azul, y también, infrarrojo cercano, infrarrojo medio, ultravioleta, etc. La cantidad de bandas que se obtengan dependerá del tipo de sensor. Para los sensores pasivos, se puede establecer una clasificación sencilla en función de la cantidad de información que poseen. Así, en la bibliografía se habla de ortofotos cuando se trata de datos georreferenciados en el espectro visible, imágenes multiespectrales cuando dichos datos tienen aproximadamente entre 4 y 7 bandas de distintos espectros (visible, infrarrojo) e imágenes hiperespectrales que reducen el intervalo de longitud de onda detectable recogiendo la información de la escena en varias decenas de bandas.

En cualquier caso, la unidad mínima de información para este tipo de datos es el píxel, que contendrá un valor normalmente entre 0 y 255 para cada banda. A partir de estos datos se puede trabajar con la escena para generar diversos productos.

2.4. Clasificación y medidas de evaluación

Para mejorar la calidad del producto final, la teledetección suele aplicar diversas técnicas de tratamiento de imágenes, entre las que destacan las de corrección de imágenes, el realce de las mismas o las transformaciones globales. Pero en la mayor parte de los casos, todos estos tratamientos van dirigidos a realizar una tarea de clasificación. Este proceso conduce a hacer interpretables los valores numéricos adquiridos por el sensor para cada píxel de la imagen y como consecuencia del mismo, clasificar las diversas zonas de la escena.

En general, la tarea de clasificar se puede resumir en decidir a qué estado de todos los posibles pertenece el píxel que se está estudiando. Así, se debe decidir si se trata de un suelo desnudo, algún cultivo herbáceo, una zona urbana o bosque caducifolio, etc. El resultado final que se obtiene está sometido a diversos errores de asignación. Dichos errores se calculan para cada categoria (k) y se denominan tradicionalmente:

- *Error de omisión o de productor:* se produce cuando, aún perteneciendo a la categoría k, un píxel u objeto no es asignado a ella.
- *Error de comisión o de usuario:* se produce cuando un píxel u objeto es asignado a la categoría k pero pertenece a otra categoría distinta.

El tipo de clasificación que se desee realizar influye de manera directa en el número de errores cometidos. Así, si las categorías son demasiado genéricas, el error de clasificación será muy bajo aunque a costa de extraer una información escasa. Si por el contrario las categorías son muy específicas, el nivel de desagregación será muy alto. Para estimar la calidad de la clasificación realizada, existen dos técnicas principalmente:

- Comparar los resultados con otras fuentes analógicas, como los mapas de usos de suelo, o tabuladas, como las estadísticas agrarias.
- Realizar una campaña sobre una muestra de la escena.

El primero de los métodos implica validar un conocimiento mediante un elemento que ha sido generado a partir de un muestreo previo. A ello, hay que añadir que en ocasiones, estos documentos suelen estar desfasados y no son muy fiables. Por ello, principalmente se debe utilizar la segunda opción y dejar la comparación con otras fuentes sólo para cuando no se pueda realizar un muestreo directamente en campo.

La última etapa de validación es recompilar todos los resultados confeccionando una matriz cuadrada en la que las columnas recogen las clases propuestas por el clasificador y las filas la ocupación real. A esta matriz se la conoce como *matriz de confusión*. Cada elemento de la misma estará ocupado por un número que representa la cantidad de elementos de la muestra analizada que se han clasificado como elementos de la categoría que marca su fila y que la fase de verificación ha demostrado que pertenecen a la categoria asociada a su columna. Así, la diagonal principal contendrá los elementos bien clasificados y

aquellos que estén fuera indicarán errores de asignación. La matriz de confusión es una modalidad de tabla de contingencia, a partir de la cual se podrá extraer información cuantitativa acerca del proceso de verificación.

2.4.1. Tipos de enfoques en la clasificación

En teledetección, existen diversos enfoques para realizar la clasificación de las imagenes (Lu and Weng, 2007) según el tipo de instancia utilizada para realizar la tarea. El enfoque más tradicional es el que se denomina de orientación a píxel. En él, se extrae de cada píxel la información espectral y se realiza la clasificación en función del conjunto de valores disponibles que, en general, coincidirá con el número de bandas registradas en la imagen. Su principal ventaja es la simplicidad, ya que el proceso de extracción de la información es directo. De esta forma, es posible generar las bases de datos y realizar la clasificación con un procesado intermedio prácticamente nulo. Por otra parte, las propuestas orientadas a píxel adolecen de ciertos problemas como son la incapacidad para tratar píxeles mixtos (combinación de varias clases) que, en muchos casos, dan lugar a un tipo de ruido conocido como «sal y pimienta» reduciendo la calidad del producto final.

Una solución para los problemas del enfoque orientado a píxel es lo que se denomina orientación a campo o polígono (field-oriented)[1]. En este tipo de aproximaciones, el clasificador divide el área de trabajo en cuadrículas y les asigna como información propia los valores medios de los píxeles que engloban. En base a esos estadísticos, se realiza la clasificación posterior. El problema principal de este tipo de aproximaciones es que la información real es generalmente irregular y es complicado adecuar la estructura contenedora para que los píxeles que contenga sean homogéneos, es decir, no haya mezclas de etiquetas.

Para solucionar el problema de los polígonos de píxeles heterogéneos, muchos autores defienden el uso del enfoque que se denomina de orientación o basado en objetos y que se generalizó con el lanzamiento del software propietario *eCognition* eCognition Software (2011). En este tipo de aproximaciones, la unidad mínima de información ya no es el píxel sino el objeto. Un objeto es una aglomeración de píxeles, no forzosamente regular, con características comunes y que se construye mediante un proceso de segmentación de los datos originales. Una vez que el conjunto de objetos se tiene perfectamente definido, se aplica la fase de clasificación usando las características de este nuevo tipo de instancias como datos de entrada. Aunque existen numerosos estudios que muestran la superioridad de esta técnica sobre las técnicas más tradicionales (Whiteside et al., 2011; Myint et al., 2011) también introduce un proceso, la segmentación de imágenes, costoso en muchos aspectos y que tiene líneas de investigación propias, lo que implica que no es una tarea sencilla.

Además de la clasificación orientada a objetos y a polígonos, también existen otras aproximaciones que pueden verse como un punto medio entre los anterio-

[1]En este texto, se usará principalmente el nombre parcela y en menor medida, campo o polígono, debido a que parcela es más descriptivo, a pesar de que no sea el término más exacto para traducir «field».

res y que reciben el nombre de aproximaciones contextuales. Esta aproximación es similar a la orientación a objetos pero tiene como diferencia fundamental el hecho de que no se lleva a cabo una segmentación previa de la escena. En su lugar, la clasificación contextual explota la información de los píxeles vecinos para mejorar los resultados de la clasificación per-píxel. En general, las clasificaciones contextuales se pueden clasificar en *pre-smoothing* y *post-smoothing*. Los clasificadores *pre-smoothing* incorporan información contextual como bandas nuevas y a continuación, llevan a cabo una clasificación utilizando ténicas estándares. Por contra, los clasificadores post-smoothing realizan procesos de refinamiento sobre las imágenes previamente clasificadas.

2.5. Aplicaciones

En lo referente a las aplicaciones de la teledetección, es importante destacar que son cada día más numerosas en diversos campos. Un índice demostrativo del interés suscitado por esta disciplina en los últimos años reside en el incremento del número de ponencias, comunicaciones a congresos y artículos publicados en revistas que muestran su aplicación al campo medioambiental. A grosso modo, es posible agrupar estas aplicaciones según el objeto de estudio. Así, podemos encontrar aplicaciones de la teledetección relacionadas con:

- Estudio de la litosfera. Se trata de estudios realizados sobre el componente sólido e inerte del planeta. Así, se estudian los riesgos de erosión, la geomorfología y la geología del terreno o el nivel de radiación solar a nivel de superficie.
- Estudio de la hidrosfera. La hidrología estudia el agua desde el punto de vista de su interacción con el medio físico terrestre. La evolución del ciclo hidrológico está muy relacionado con las características del terreno y determinados grupos de investigación (Ojeda Zújar et al., 2006), intentan realizar estudios para detectar terrenos inundables o que tiendan a ser fácilmente afectados por la escasez o exceso de lluvias.
- Estudio de la atmósfera. Aunque esta rama de la teledetección tiene muy diversas aplicaciones, la más conocida está en la predicción metereológica y en el estudio del clima de manera general.
- Estudio de la biosfera. Estos estudios se centran, en general, en el estudio de la vegetación y, en particular, en intentar determinar la producción anual de ciertos cultivos.
- Cartografía. Su objetivo es por un lado el análisis territorial (definición y reconocimiento) y por otro la planificación territorial (determinación de aptitudes, control y riesgos de los suelos, etc.).

Toda esta información suele recompilarse en grandes bases de datos que dan lugar a la necesidad de herramientas como los sistemas de información geográfica

(SIG) y que tienen una gran importancia en diversos ámbitos y especialmente en el de las administraciones públicas. Para aumentar el potencial de estas técnicas, no dejan de aparecer nuevas tecnologías de sensores que dan mayor resolución a los datos obtenidos y que mejoran las expectativas de explotación.

2.6. Sistemas de información geográfica

El desarrollo de la teledetección ha propiciado la necesidad de herramientas para los especialistas. Entre ellas, destacan los Sistemas de Información Geográfica (SIG o GIS en inglés) (Bosque Sendra, 1997). Dichos sistemas han sido definidos de distintas formas:

Definición 1. Base de datos computerizada que contiene información espacial (Cebrián, 1988).

Definición 2. Una tecnología informática para gestionar la información espacial (Bosque Sendra, 1997).

Definición 3. Un conjunto de herramientas para reunir, introducir (en el ordenador), almacenar, recuperar, transformar y cartografiar datos espaciales sobre el mundo real para un conjunto particular de objetos (Burrough, 1988).

Definición 4. Un tipo especializado de base de datos, que se caracteriza por su capacidad de manejar datos geográficos, es decir, espacialmente referenciados, los cuales se pueden representar gráficamente como imágenes (Bracken and Wesbter, 1990).

Definición 5. Un sistema hardware, software y procedimientos elaborados para facilitar la obtención, gestión, manipulación, análisis, modelado, representación y salida de datos espacialmente referenciados, para resolver problemas complejos de planificación y gestión (National Center for Geographic Information and Analysis/University of California, 1990).

Pero más allá de estas definiciones, un SIG no es más que un sistema de información que contiene un conjunto de mapas de un mismo área donde un lugar concreto tiene la misma localización en todos los mapas incluidos. La mayor parte de los SIG sólo actúan como visores de datos de teletedección. Sin embargo, nuevos desarrollos como Sextante (Olaya, 2006), plantean ya la introducción de rutinas que permitan mediante técnicas inteligentes generar conocimiento a partir de la información que el sistema posea. De ahí, la necesidad de que la comunidad informática y más concretamente, los expertos en minería de datos trabajen en el área para generar mejores algoritmos adaptados a las necesidades de la teledetección.

2.7. Sensores LIDAR

2.7.1. Definición y caracterísicas principales

La tecnología LIDAR (Light Detection And Ranking) se puede definir, según el Ministerio de Medio Ambiente de España (Ministerio de Medio Ambiente, 2011), como un sistema telemático activo de captura de datos que, procesados de

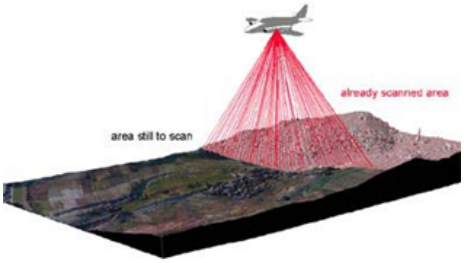


Figura 2.4: LIDAR aerotransportado (adaptado de Chuvieco (2008)).

manera adecuada, pueden proporcionar información muy útil en determinadas áreas de conocimiento. Otros autores (Pinilla, 1996) lo definen como un sistema activo que se caracteriza por al emisión de pulsos de luz polarizada entre el ultravioleta y el infrarrojo próximo mediante un emisor láser.

El instrumental utilizado en LIDAR emite pulsos de luz láser en la banda del espectro electromagnético generalmente en el infrarrojo cercano (ver Figura 2.4) para posteriormente capturar la señal reflejada por la superficie topográfica barrida, midiendo el tiempo empleado por cada una de las señales emitidas en recorrer el espacio que separa al transmisor de la superficie física del terreno. Mediante ese tiempo, el sistema es capaz de localizar la posición del impacto y dicho dato es almacenado en el sistema junto al resto de la información del retorno. Utiliza los mismos principios que la tecnología radar, si bien lo que difiere es la longitud de onda de la señal utilizada. Dicha longitud de onda es mucho menor, por lo que el detalle de la imagen resultante es mayor que la que provee un sistema radar ya que, en general, es posible medir objetos que tengan un tamaño comparable o mayor a la longitud de onda de emisión, por lo que un sistema lidar es en general mucho más sensible en la medición de gases y aerosoles que un sistema radar. La otra gran diferencia es que LIDAR necesita el empleo de técnicas de posicionamiento global (GPS) en modo diferencial y en tiempo real, y sistemas de navegación inercial que permitan caracterizar la posición espacial del instrumental de medición para poder determinar la posición espacial de cada impacto.

Dentro de los sensores LIDAR, suele establecerse una clasificación en función del tamaño del pulso en el suelo. Así, se habla de LIDAR de huella pequeña si el tamaño del pulso en suelo está entre los 20 y 40 cm., y de huella grande a aquellos dispositivos con huellas del orden de las decenas de metros. En función del tipo de preprocesado realizado, se puede hablar también de LIDAR de onda completa (full-wave) o de LIDAR discreto. Originalmente, todos los vuelos LIDAR son de onda completa ya que el sensor registra la actividad de los pulsos de manera continua. Es en la etapa de preprocesado donde los proveedores discretizan la onda buscando los máximos energéticos registrados. Dependiendo del número de máximos detectados, los datos contendrán sólo primer y último impacto por pulso (lo más común) o un mayor número de impactos registrados.

Item	Formato	Tamaño	Requerido
X	long	4 bytes	*
Y	long	4 bytes	*
Z	long	4 bytes	*
Intensidad	unsigned short	2 bytes	
Identificador Retorno	bits 0, 1, 2	3 bits	*
Total Retornos	bits 3, 4, 5	3 bits	*
Dirección Escaneo	bit 6	1 bit	*
Lado de vuelo	bit 7	1 bit	*
Clasificación	unsigned char	1 byte	*
Ángulo	char	1 byte	*
Datos de Usuario	unsigned char	1 byte	
ID Fuente	unsigned short	2 bytes	*
GPS	double	8 bytes	*

Cuadro 2.1: Formato de un punto de la nube LIDAR en el fichero LAS.

El tipo de datos LIDAR más común es el LIDAR discreto de huella pequeña (Chuvieco, 2008). En estos casos, el resultado de un vuelo LIDAR de pulso discreto es una nube de puntos que, en general, se provee en formato LAS (ASP, 2005). A partir de esta nube es posible obtener para cada pulso un conjunto de datos (ver Cuadro 2.1) tales como su posición espacial, su intensidad de retorno, el momento del pulso, etc.

2.7.2. Principales aplicaciones

Los vuelos LIDAR, al igual que otros sensores en teledetección tienen aplicación en multiples campos desde el militar al medioambiental pasando por la ingeniería civil o la arqueología. Para el presente trabajo, nos centraremos en el área medioambiental. Los principales productos derivados del vuelo LIDAR para este campo son:

- Modelo Digital de Terreno o de Elevaciones (Digital Terrain or Elevation Model, DEM), obtenido a partir del último pulso (base de la modelación hidráulica y cartográfica).

- Modelo Digital de Superfice (Digital Surface or Canopy Model, DSM) obtenido a partir del primer pulso, sobre el cual podemos distinguir las alturas de edificios, vegetación, puentes...

- Modelo de Intensidades (Lidar Intensity Image), a partir de la amplitud de la señal que vuelve al avión después de rebotar en la superficie terrestre obtenemos una imagen de intensidades que permite realizar distinciones entre superficies, identificando carreteras, edificios, vegetación, etc.

Las ventajas de la tecnología LIDAR frente a las técnicas tradicionales de trabajo en campo son importantes, destacando un notable menor coste de la cartografía y una mayor precisión en los puntos obtenidos (Ministerio de Medio Ambiente, 2011). Del mismo modo, puesto que se calcula directamente el modelo digital del mismo, no es necesario interpolarlo a partir de la cartografía tradicional, con lo que mejora notablemente la precisión de los estudios y se produce un ahorro de tiempo. Como desventajas, la necesidad en determinados casos de realizar batimetrías y la no obtención de información asociada a la cartografía (toponimia, etc.).

Capítulo 3

Aprendizaje automático en teledetección

Cualquier tecnología suficientemente avanzada
es indistinguible de la magia.
Arthur Clarke.

3.1. Introducción

Los mapas generados mediante clasificación de imágenes (comúnmente llamados mapas temáticos) son probablemente el principal producto a partir de datos remotos. En términos generales, los métodos para generar este tipo de productos se pueden dividir en tres grandes familias (Tuia and Camps-Valls, 2009b): métodos no supervisados, supervisados y semi-supervisados.

El aprendizaje no supervisado se puede definir como el conjunto de técnicas que permiten establecer un modelo sobre un conjunto de datos sin necesidad de tener un conocimiento previo. En general, el concepto de aprendizaje no supervisado está íntimamente relacionado al de *clustering* o agregación. Así, las técnicas no supervisadas tratan de agrupar los píxeles de una imagen según su similitud de acuerdo a un número predefinido de grupos. Una de las principales aplicaciones de este tipo de técnicas es el reconocimiento de cambios en tiempo real (Ghosh et al., 2011).

El aprendizaje supervisado puede definirse como la aplicación de una técnica inteligente para desarrollar un modelo a partir de un conjunto de entrenamiento inicialmente etiquetado por el usuario. Dicho modelo es utilizado posteriormente para etiquetar nuevas instancias de manera automática. En la actualidad, este campo es, probablemente, el más activo en el procesamiento de imágenes de teledetección dando lugar a aproximaciones que aplican las técnicas más avanzadas de aprendizaje automático y *soft computing* (García et al., 2011; Easson and Momm, 2010).

Por último, habría otro grupo formado por las técnicas semi-supervisadas que

realizan una combinación de los datos etiquetados (por lo general, del orden de las centenas) y de la información que se encuentra en los datos no etiquetados.

Además de estos grandes grupos, en los últimos tiempos están apareciendo otras técnicas que no son fácilmente clasificables dentro de las categorías anteriores (Camps-Valls, 2009). Así, tenemos técnicas de aprendizaje múltiple (manifold learning), aprendizaje por transferencia (transfer learning), aprendizaje estructurado (structured learning) o aprendizaje activo (active learning).

En este capítulo, se presentan las principales técnicas usadas en tareas relacionadas con la teledetección ambiental, clasificadas según el tipo de aprendizaje. Así, veremos principalmente las técnicas clásicas basadas en aprendizaje supervisado y no supervisado, haremos referencia a las técnicas de aprendizaje semi-supervisado más importantes y mostraremos algunos resultados del aprendizaje activo dentro del conjunto de técnicas más novedosas.

3.2. Aprendizaje supervisado

3.2.1. Clasificador de máxima verosimilitud

El clasificador de máxima verosimilitud (Maximum Likelihood Classifier, MLC) es uno de los clasificadores más utilizados en teledetección. El MLC es un clasificador paramétrico que se basa en la teoría de probabilidades bayesiana. Para este clasificador, cada instancia (un píxel o un objeto) puede ser visto como un vector de características (x) y análogamente, como un conjunto de valores para una variable aleatoria continua. La etiqueta y_i de una instancia está determinada por la probabilidad condicionada asociada a cada clase $P(x|y_i)$ así como por su probabilidad a priori $P(y_i)$. Para $P(x|y_i)$, si se asume una densidad normal o gaussiana, su función de distribución condicionada $p(x|y_i)$ viene dada por la Ecuación 3.1 que depende del número de atributos o bandas N con el que se trabaja, el vector de medias μ y la matriz de covarianza Σ (Lillesand and Kiefer, 2000).

$$p(x|y_i) = \frac{1}{\sqrt{(2\pi)^N |\Sigma_i|}} e^{\frac{-1}{2}(x-\mu_i)^t |\Sigma_i|^{-1}(x-\mu_i)} \tag{3.1}$$

Es posible obtener una función discriminante para el MLC (Richards and Jia, 1999) tomando logaritmos neperianos como puede verse en la ecuación 3.2, teniendo en cuenta que se suele suponer la igualdad de probabilidades a priori ($P(y_i) = P(y_j), \forall i, j$). De esta forma, el discriminante del MLC se puede utilizar para seleccionar la categoría en lugar de la regla bayesiana original.

$$g_i(x) = -ln|\Sigma_i| - (x-\mu_i)^t \Sigma_i^{-1}(x-\mu_i) \tag{3.2}$$

Muchos estudios han demostrado la idoneidad del MLC cuando se aplica sobre datos multiespectrales con el objetivo de clasificar tipos de vegetación (Hagner and Reese, 2007) o tipos de combustibles (Lasaponara and Lanorte, 2007) aunque se recomienda su aplicación cuando se tenga la certeza de que las condiciones de normalidad se cumplen.

3.2.2. Redes neuronales artificiales

Las redes neuronales artificiales (artificial neural network, ANN) son clasificadores no paramétricos basados en el modo de funcionamiento del cerebro humano (McCulloch and Pitts, 1943) muy utilizado en procesamiento de imágenes (Egmont-Petersen et al., 2002). En Haykin (1998), se puede encontrar una definición más formal que define una ANN como «un procesador distribuido, masivamente paralelo, hecho de unidades de procesamiento simple, que tienen una propensión natural para acumular conocimiento experimental y disponibilizarlo para su uso».

Las ANNs combinan una gran cantidad de elementos de procesado altamente interconectados que se adaptan para resolver problemas concretos mediante un proceso de aprendizaje. Los elementos de procesado suelen denominarse nodos o neuronas y se estructuran en capas (ver Figura 3.1). Esta estructura las convierte en una técnica de modelado de gran flexibilidad y que suele obtener buenos resultados en problemas complejos.

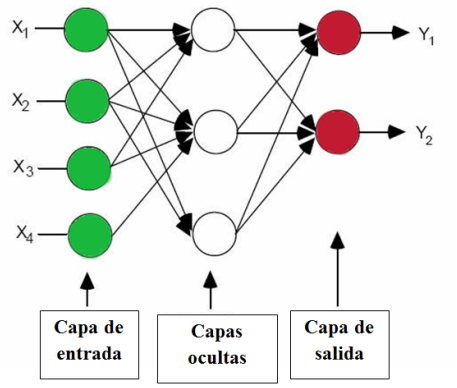


Figura 3.1: Ejemplo de estructura de red neuronal con una capa oculta y transmisión hacia adelante (Multilayer Perceptron, Rosenblatt (1962)).

Cada unidad de procesamiento de una ANN se compone de un conjunto de conexiones de entrada, una función de activación (encargada de computar la entrada total combinada de todas las conexiones), un núcleo central de procesado (encargado de aplicar la función de activación) y una salida o función de transferencia (por donde se transmite el valor de activación a otras unidades). De esta forma, los elementos que caracterizan un nodo de una ANN son:

- **Conjunto de sinapsis o uniones ponderadas.** Correspondería al conjunto de pesos que influirán en el valor estímulo que una instancia produce en el nodo.
- **Función de activación.** Elemento (en general, se tratará de un sumador) encargado de recoger todos los estímulos ponderados en cada nodo y

devolver un valor como respuesta.

- **Función de salida o transferencia.** Se encarga de limitar el valor de salida de manera que la respuesta de la red neuronal esté dentro de unos valores preestablecidos, ya sea un valor numérico o una categoría. Existen diversas posibilidades para implementarla que oscilan entre una función simple como la aplicación de un umbral hasta opciones más avanzadas como la aplicación de funciones no lineales.

Existen distintas formas de transmisión de la información entre los nodos de una ANN, que determinan la naturaleza de la misma. Así, los tipos de transmisión se pueden clasificar en: hacia adelante (*feedforward*), lateral y hacia atrás (*feedback*), en los que la información fluye desde la capa de entrada hacia la de salida, entre los nodos de una misma capa, y desde los nodos de salida hacia los de entrada, respectivamente.

El modelo de ANN ha sido explotado en teledetección de manera profusa en los últimos 20 años. Así, en Benediktsson et al. (1990) se demuestra que las redes neuronales pueden mejorar los resultados de las técnicas bayesianas cuando se aplican a la generación de mapas de usos de suelo a partir de datos multisensor. Más recientemente, se han propuesto redes neuronales granulares que permiten detectar los bordes de las distintos tipos de uso o coberturas (Vasilakos and Stathakis, 2004) para mejorar los valores de precisión global o directamente se han combinado varias redes con distintas características y técnicas de boosting para mejorar sus resultados por separado en el mismo contexto (Canty, 2009).

3.2.3. Árboles de decisión

Un árbol de decisión (decision tree, DT) puede definirse como un modelo que, recurrentemente, realiza particiones de un conjunto de datos en subdivisiones más pequeñas para facilitar la toma de decisiones relacionada con dichos datos. Los DTs se suelen utilizar como modelo tanto para predicción como para clasificación y a pesar de que son bien conocidos desde mediados de los años 80, pueden verse en muchas aplicaciones recientes tales como la extracción de edificios en datos de alta resolución (Tooke et al., 2009).

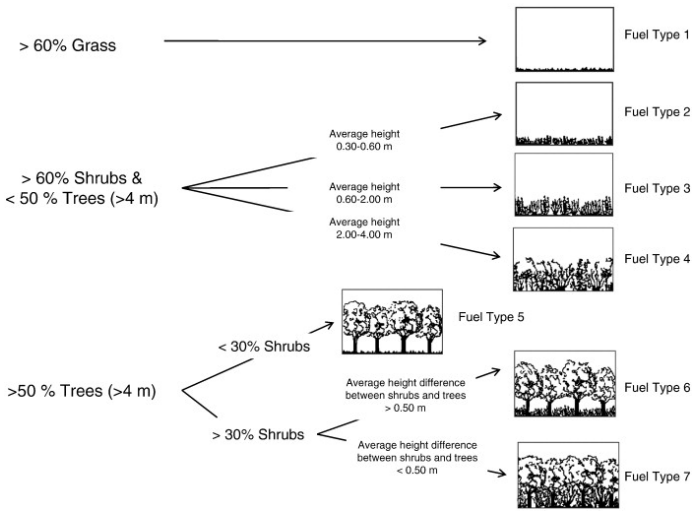


Figura 3.2: Modelo de árbol de decisión (extraído de García et al. (2011)).

Un DT se caracteriza por una estructura ramificada a partir de un punto inicial (raíz) que termina en un conjunto de nodos finales (hojas) representando la salida prevista por el modelo como puede verse en el ejemplo de la Figura 3.2. Uno de los algoritmos de generación automática de DTs más conocido es el algoritmo C4.5 (Quinlan, 1996). C4.5 construye un árbol de decisión utilizando el concepto de razón de ganancia de información. La razón de ganancia se puede expresar como en la Ecuación 3.3 y se basa en el concepto de entropía (Ecuación 3.4). Para cada nodo, C4.5 selecciona el atributo A que maximiza la relación entre la ganancia de un subconjunto de registros T, donde T_i representa cada división de T de acuerdo con un valor de A, C es el número de etiquetas y $p(T, j)$ es la probabilidad de que un conjunto de registros tengan la etiqueta j. El atributo que maximice la razón de ganancia, se selecciona como la variable candidata para ser la raíz del actual nivel. El proceso continúa con las ramas posteriores hasta que no haya más divisiones posibles.

$$Gain(A, T) = E(T) - \sum_{i=1}^{k} \frac{|T_i|}{|T|} E(T_i) \tag{3.3}$$

$$E(T) = -\sum_{j=1}^{C} p(T, j) log_2(p(T, j)) \tag{3.4}$$

Los DTs tienen varias ventajas (Friedl and Brodley, 1997; García-Gutiérrez et al., 2012b) sobre los métodos tradicionales de clasificación supervisada en teledetección. Así, al contrario que MLC, se trata de clasificadores no paramétricos por lo que no requieren que se cumplan supuestos sobre la distribución de los datos de entrada. Además, pueden tratar con relaciones no lineales entre los

atributos y las clases, se permiten los valores *missing*, y son capaces de manejar entradas numéricas y categóricas. Por último, la ventaja más importante del DT respecto a sus competidores es su capacidad para dividir un proceso complejo de toma de decisiones en una colección de decisiones simples y de esta forma, ofrecer una solución que a menudo es más fácil de interpretar.

3.2.4. Vecino más cercano

La técnica de los vecino más cercano (nearest neighbour, NN) es un clásico en la familia de los clasificadores no paramétricos. También se les conoce como clasificadores basados en instancias o perezosos debido a que no se crea un modelo a partir de los datos de entrenamiento sino que los datos en sí mismos, forman el modelo (Dasarathy, 1990; Aha and Kibler, 1991). NN se basa en el concepto de similitud (Ecuación 3.5) o distancia. Así, la etiqueta de la instancia a clasificar será la misma que la de la instancia de entrenamiento que más se ajuste a ella según la función de similitud.

$$Similarity(x, y) = -\sqrt{\sum_{i=0}^{n} f(x_i, y_i)} \tag{3.5}$$

La regla NN puede generalizarse calculando los k vecinos más cercanos y asignando la clase mayoritaria. Tal generalización se denomina k-NN. Este algoritmo necesita la especificación del parámetro k a priori. El cálculo del k óptimo es un problema ampliamente tratado en la bibliografía (Dietterich, 1995). Además, hay que tener en cuenta otras limitaciones de este clasificador relacionadas con su coste computacional. Cada vez que se necesita clasificar una nueva instancia, el algoritmo debe recorrer el conjunto de entrenamiento completo para obtener los k vecinos más cercanos. Esto hace que el algoritmo k-NN original sea costoso tanto en tiempo, ya que necesita recorrer todos los ejemplos en cada predicción, como en espacio (en memoria), por la necesidad de mantener almacenado todo el conjunto de entrenamiento. Para superar estos problemas, es posible encontrar en la bibliografía diferentes aproximaciones buscando una mejora en su eficiencia (Mico et al., 1996; Gomez Ballester et al., 2006).

Pese a los inconvenientes respecto a la eficiencia (coste computacional) y la eficacia (elección de la métrica y el k adecuados), el algoritmo k-NN tiene un buen comportamiento en general, y como los clasificadores clásicos anteriores, NN se puede encontrar aplicado en la actualidad en diferentes áreas relacionadas con la teledetección. Así, encontramos trabajos en los que se utiliza para establecer una relación entre indicadores LIDAR y tipos de especies vegetales (Hudak et al., 2008) o se aplica para realizar inventarios forestales (Magnussen et al., 2009). En los últimos tiempos, se pueden encontrar soluciones para realizar tareas sobre datos remotos que usan NN de manera híbrida con otras técnicas como el análisis discriminante (Thessler et al., 2008).

3.2.5. Máquinas de vector soporte

Las máquinas de vector soporte (Support Vector Machines, SVM) son un método supervisado no paramétrico cuyas bases fueron establecidas por Cortés y Vapnik (Cortes and Vapnik, 1995). El algoritmo SVM tiene como objetivo encontrar el hiperplano óptimo que separa los datos de entrenamiento en un espacio de características multidimensional. Una característica importante de las SVMs es que la búsqueda del hiperplano se basa en la minimización del riesgo estructural intentando maximizar los márgenes entre el hiperplano óptimo de separación y las instancias más cercanas. A estos márgenes se les conoce como vectores soporte (Gunn, 1998).

El caso más simple de una SVM sería una clasificación binaria, donde las muestras puedan ser separadas linealmente. En este caso, el hiperplano óptimo se define por la Ecuación 3.6 donde ω es el vector normal, x es un punto del hiperplano óptimo y b es un valor de sesgo. Esta ecuación plantea un problema de optimización definido por la ecuación 3.7 donde C es un valor de penalización para los casos en el lado equivocado del hiperplano, y ε_i son las variables de anchura que indican la distancia de los puntos clasificados incorrectamente desde el hiperplano óptimo (Foody and Mathur, 2004).

$$f(x) = \omega x \ + \ b \tag{3.6}$$

$$\text{mín}[\frac{\omega^2}{2} + C\sum_{i=1}^{n}\varepsilon_i] \tag{3.7}$$

Si dos clases no pueden ser separadas linealmente, las SVMs también son capaces de representar la no linealidad mediante la proyección de los datos de entrada en un espacio de características de dimensión superior por medio de una función de kernel. La función de decisión final se describe en la Ecuación 3.8 donde $\alpha_i, i = 1, \ldots, n$ son multiplicadores de Lagrange, $y_i = \pm 1$, y k es una función de kernel (gaussiana, de base radial, polinomial, ...):

$$f(x) = sgn(\sum_{i=1}^{n}\alpha_i y_i k(x, x_i) + b) \tag{3.8}$$

A pesar de que fue desarrollado originalmente como un clasificador binario, han aparecido varias aproximaciones («uno contra todos» y «uno contra uno») para el uso de SVMs sobre datos multiclase, en las que las clasificaciones se reducen a un conjunto de problemas binarios (Foody and Mathur, 2004).

Las SVMs se pueden encontrar en un gran número de aplicaciones relacionadas con la teledetección, especialmente en los últimos años (Mountrakis et al., 2011). Así, se han aplicado a la clasificación de tipos de bosques (Huang et al., 2002; Knorn et al., 2009), a la estimación de la degradación de áreas naturales (Cao et al., 2009; Kuemmerle et al., 2009), al desarrollo de mapas de coberturas vegetales para la evaluación de riesgo de incendios (Koetz et al., 2008; García et al., 2011) o para cartografía urbana (Zhu and Blumberg, 2002; Esch et al., 2009), por mencionar sólo algunos ejemplos.

3.3. Aprendizaje no supervisado

3.3.1. Algoritmo Esperanza-Maximización

El algoritmo Esperanza-Maximización (Expectation-Maximization, EM), es un procedimiento iterativo y eficiente (Dempster et al., 1977) que calcula el estimador de maxima verosimilitud en presencia de datos incompletos (Borman, 2004). Una vez conocido este estimador, es posible aplicarlo para conocer la probabilidad de que una instancia pertenezca a un agrupamiento en concreto.

EM se caracteriza por la ejecución de una serie de iteraciones en las que cada una de ellas se divide en dos pasos. En el primero o fase E, dado un conjunto de instancias representando los valores observados y una asignación inicial de dichas instancias al conjunto de agrupamientos (para la primera iteración, la asignación es aleatoria), se construye un estimador de máxima verosimilitud. En la fase M, se recalcula el agrupamiento asignado a cada instancias de manera que la función de similitud se maximice. En este caso, el estimador responde con el aprendizaje adquirido en el paso E. La convergencia de este proceso se garantiza debido a que el algoritmo produce siempre un incremento de la similitud en cada iteración (Jank, 2005).

La bibliografía muestra numerosas aplicaciones de este algoritmo paramétrico en el campo de la teledetección y suele ser uno de los métodos clásicos con los que se comparan las nuevas propuestas (Deng and Clausi, 2004). Recientemente además, se ha propuesto usar una versión estocástica del algoritmo EM (Cariou and Chehdi, 2008) (el algoritmo se modifica para que distintas ejecuciones no converjan en el mismo punto estacionario) con el objetivo de superar el problema del cálculo del número de agrupamientos, que generalmente suele ser un problema no contemplado.

3.3.2. Redes de Kohonen

Los mapas auto-organizados (Self-Organizing Map, SOM) o redes de Kohonen (Kohonen, 1982; Kohonen et al., 2001) son redes neuronales competitivas con aprendizaje no supervisado. Se caracterizan por definir una estructura en forma de rejilla para representar el conocimiento adquirido (mapa neuronal) como puede verse en la Figura 3.3. La vencidad en la rejilla implica similaridad en las características de las neuronas.

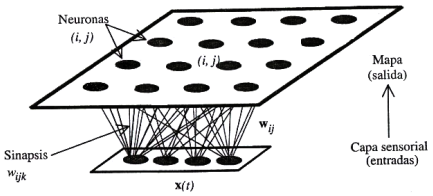


Figura 3.3: Rejilla generada por una red de Kohonen.

El objetivo de la red de Kohonen es descubrir rasgos comunes en los datos de entrada. Para ello, cada neurona compite por activarse en función de los datos de entrada. Esto ocurre cuando dicha neurona es seleccionada como la más parecida al dato de entrada o forma parte de la vecindad de la seleccionada. En el diseño original, las neuronas más próximas representan clusters parecidos. Esto hace que, al final de la fase de entrenamiento de la red, sea posible visualizar los conjuntos de condiciones que mejor se adaptan a los datos de entrada y de ahí, se pueda establecer un clustering eficiente de dichos datos.

Se han generado múltiples soluciones para adaptar las redes de Kohonen a las especiales características de los datos remotos (Villmann et al., 2003). Como ejemplo, se puede citar el trabajo de Ghosh et al. (Ghosh et al., 2009) en el que se implementa un SOM para detectar cambios en imágenes a lo largo del tiempo. Los autores hacen corresponder una neurona por cada posición en la imagen estudiada y actualiza el estado de las neurona de acuerdo a distintas medidas de similaridad. Cuando termina el entrenamiento, se tiene un mapa con el grado de cambios detectados en cada posición de la imagen. Otra propuesta reciente es la de Molinier et al. (Molinier et al., 2007) que sigue el mismo planteamiento pero que analiza no el píxel en sí, sino la información que contiene tras aplicar una fase previa de extracción de conocimiento.

3.3.3. Algoritmo k-medias y derivados

El algoritmo k-medias (MacQueen, 1967) es uno de los algoritmos de clustering más conocidos en el área del aprendizaje automático. Se trata de un algoritmo que intenta particionar el conjunto de datos de manera que se agrupen en k conjuntos. Así, cada cluster o grupo de datos se puede caracterizar mediante su centroide (coordenada formada por los puntos medios del grupo para cada atributo o característica).

El algoritmo k-medias suele implementarse siguiendo lo que se denomina como algoritmo de Lloyd (Lloyd, 1982). Este algoritmo se caracteriza por dos pasos. En el primero, se calculan los centroides (μ) de los k clusters. Cada centroide es una instancia que tiene como valores de sus atributos la media de los correspondientes a las instancias pertenencientes al cluster en cuestión. En el segundo paso, se reasigna cada instancia a un cluster respetando la condición descrita en la Ecuación 3.9, donde x_j representa cualquier instancia de un cluster S_i con centroide μ_i. Estos dos pasos se repiten hasta que el sistema converge.

$$S_i = \left\{ \mathbf{x}_j : \left\| \mathbf{x}_j - \mu_i \right\| \leq \left\| \mathbf{x}_j - \mu_l \right\| \forall l = 1, \ldots, k \right\} \tag{3.9}$$

El algoritmo k-medias clásico ha tenido un papel preponderante en numerosas aplicaciones relacionadas con datos remotos. Así lo atestigua el hecho de que sea uno de los métodos más utilizados como referencia para comparaciones con otras metodologías más avanzadas (Duda and Canty, 2002). Además, la aparición de variantes como ISODATA (del inglés Self-Organizing Data Analysis Techniques), que se diferencia de k-medias en que permite combinar y romper agrupamientos en cada iteración de acuerdo a una serie de parámetros fijados

por el usuario, también han sido exploradas con relativo éxito (Bachmann et al., 2002).

En el entorno LIDAR, algunos autores han usado k-medias para segmentar los datos y conseguir aislar los pulsos a nivel de árbol individual (Morsdorf et al., 2004). Un dato importante es que existen estudios recientes que muestran la capacidad de k-medias para establecer sinergias con otras metodologías como la selección de atributos basadas en componentes principales (Celik, 2009). Esto hace que, a pesar de ser un clásico, el algoritmo k-medias siga siendo materia de trabajo para muchos autores hoy en día.

A pesar de que el uso combinado del algoritmo k-medias y los sistemas de lógica borrosa (*fuzzy logic*) en la forma del algoritmo «fuzzy k-means», más conocido como «fuzzy c-means» (FCM, Bezdek (1981); Bezdek et al. (1984)), se conocen desde antiguo, ha sido recientemente cuando ha ido ganando en protagonismo en el área de la teledetección. Así, FCM ha demostrado superar a las técnicas de clustering deterministas o jerárquicas sobre datos remotos (Duda and Canty, 2002). Básicamente, FCM se comporta como el algoritmo k-medias pero la pertenencia a un cluster se controla de manera difusa minimizando la Ecuación 3.10, donde m es una medida de dispersión y debe ser mayor que 1, μ_{ij} es el grado de pertenencia de x_i al cluster c_j y $\|*\|$ es la ecuacion que define la distancia (en general, Euclidea) entre dos elementos. El paso de actualización de centroides viene definido por la Ecuación 3.11 y finalmente, el cálculo de las pertenencias a cada clusters se actualiza en cada iteración mediante la Ecuación 3.12. De esta forma, una instancia tendrá asociada una probabilidad de pertenecer a cada cluster en función generalmente, de la distancia al centroide del cluster. El proceso termina cuando las distancias de los conjuntos de μ entre las iteraciones actual y próxima son menores a un valor ϵ provisto por el usuario.

$$J_m = \sum_{i=1}^{N}\sum_{j=1}^{C} \mu_{ij}^m \|x_i - c_j\|^2 \tag{3.10}$$

$$c_j = \frac{\sum_{i=1}^{N} \mu_{ij}^m \cdot x_i}{\sum_{i=1}^{N} \mu_{ij}^m} \tag{3.11}$$

$$\mu_{ij}^m = \frac{1}{\sum_{k=1}^{C} \left(\frac{\|x_i - c_j\|}{\|x_i - c_k\|} \right)^{\frac{2}{m-1}}} \tag{3.12}$$

FCM se ha utilizado con éxito para realizar tareas de segmentación de imágenes (Fan et al., 2009) y detección de cambios en imágenes hiperspectrales (Ghosh et al., 2011) y hoy en día, es una herramienta que sigue siendo materia de trabajo para muchos investigadores que buscan mejorar sus capacidades. El principal escollo con el que se enfrentan es hacer frente a la ingentes cantidad de datos disponibles. Este problema es propio de todos los métodos presentados en esta subsección (Burrough et al., 2000). Además, es previsible que el incremento del tamaño de las bases de datos sea una constante en los próximos años, debido a que los sensores que van apareciendo tienen mejores capacidades en términos de resolución y registro de información.

3.4. Aprendizaje semi-supervisado

En los últimos tiempos, el uso de sistemas semi-supervisados (Zhu and Goldberg, 2009) ha demostrado ser un tipo de aprendizaje que se adapta perfectamente a las especiales características de los datos remotos. Un sistema semi-supervisado es aquél que realiza el aprendizaje fusionando los enfoques supervisado y no supervisado a la vez. Así, en una primera fase, se utilizan los datos de entrenamiento proporcionados por el usuario para generar el conocimiento que posteriormente es refinado mediante el uso de instancias no etiquetadas. Este tipo de aprendizaje se adapta muy bien al entorno de la teledetección en el que, de manera general, se tienen grandes bases de datos que son costosas de etiquetar y que por tanto, tienen pocas instancias de entrenamiento. Un caso prototípico en teledetección sería un conjunto del orden de cientos de miles de instancias y una base de entrenamiento de sólo el orden de las centenas.

Los sistemas semi-supervisados utilizados en teledetección por tanto, tienden a tratar de superar el problema de la ausencia de datos. En algunos casos, dichos sistemas se aplican para generar mapas de coberturas a partir de otras fuentes de datos, como mapas anteriores (Bruzzone and Fernández Prieto, 2002), sin necesidad de nuevas instancias etiquetadas. En otros, el sistema delega en un tipo de aprendizaje u otro dependiendo de determinados parámetros como el número de diferencias entre dos imágenes (Castellana et al., 2007).

Algunos autores han combinado el aprendizaje semi-supervisado con técnicas como las funciones de kernel y sus resultados demuestran la superioridad de este tipo de sistemas sobre los basados en técnicas supervisadas como las SVMs (Camps-Valls et al., 2007; Tuia and Camps-Valls, 2009a; Ratle et al., 2010).

A pesar de que este tipo de sistemas obtiene buenos resultados, su uso en entornos comerciales y reales está todavía poco extendido. Las técnicas supervisadas en general, y en menor medida, las no supervisadas son mucho más comunes en las principales aplicaciones comerciales orientadas a la explotación de usuarios no expertos (eCognition Software, 2011).

3.5. Aprendizaje activo

El aprendizaje activo es un concepto similar al aprendizaje semi-supervisado (Tuia et al., 2011b) en tanto que se utiliza la mezcla de datos etiquetados y no etiquetados pero la forma en que estos se incorporan al proceso de aprendizaje es ligeramente diferente. En este tipo de aprendizaje, el sistema detecta en un conjunto no etiquetado aquellas instancias que más información pueden aportarle. En una segunda fase, el sistema pide al usuario que etiquete dichas instancias y a partir de aquí, se realiza el resto de las tareas que en general, será algún tipo de clasificación.

En la bibliografía, se pueden encontrar aproximaciones que plantean este tipo de sistemas con relativo éxito. Así, es posible encontrar clasificadores supervisados con apoyo activo previo (Mitra et al., 2004; Di and Crawford, 2011) o propuestas que se centran directamente en la búsqueda de los píxeles más intere-

santes usando técnicas de clustering o estocásticas a partir de un subconjunto inicial formado por datos de trabajo de campo (Tuia et al., 2011a) y que generan nuevas bases etiquetadas que el usuario se encarga de explotar posteriormente.

Al igual que en el caso del aprendizaje semi-supervisado, a pesar de obtener resultados prometedores, este paradigma está poco explotado en los entornos reales, aunque se prevé que en un tiempo relativamente corto, empiecen a verse soluciones software que incluyan técnicas con esta orientación.

Capítulo 4

Datos de estudio

Con ninguna cantidad de experimentos se podrá demostrar que estoy en lo cierto, un solo experimento puede demostrar que estoy equivocado.
Albert Einstein.

4.1. Introducción

La presente propuesta de tesis nació de la necesidad de la Consejería de Medio Ambiente de la Junta de Andalucía de explotar de manera novedosa los nuevos datos LIDAR adquiridos en el período 2004-2010. Así, el grupo TIC-134 del que forma parte el aspirante a doctor, firmó un convenio de colaboración en el año 2007 con la Junta de Andalucía para explorar los datos LIDAR de su propiedad. La colaboración se llevó a cabo mediante la cesión de datos provenientes de tres áreas de Andalucía (ver Figura 4.1) que cubrían alrededor de nueve kilómetros cuadrados, y de las imágenes aéreas de las áreas en cuestión.

De los datos cedidos, dos fueron descartados para este estudio. Así, los datos de la zona de la provincia de Cádiz, cerca de San Fernando (UTM29;741233E, 4046715N), describían un área inundable situada en una zona de antiguas salinas con poca variación de alturas y de tipos de suelo. Por otro lado, también se descartó una zona situada en Córdoba, cercana a Cerro Muriano (UTM30; 344878E, 4207357N), que había sufrido un gran incendio. De nuevo, las limitaciones en cuanto al número de tipos de suelo que se iban a encontrar y la condición de que prácticamente todo el área estaba quemada, aconsejaron descartar este conjunto de datos. De esta forma, el trabajo se inició con los datos de la zona de Huelva que se describe en la siguiente sección.

Posteriormente, se establecieron relaciones con otros grupos de trabajo que han dado como resultado colaboraciones y cesiones de datos. En concreto, el grupo del Laboratorio del Territorio de la Universidad de Santiago (Campus de Lugo) cedió datos de varias zonas de Galicia y el departamento de Geografía de la Universidad de Alcalá cedió datos de una zona de Guadalajara volada por instituciones británicas.

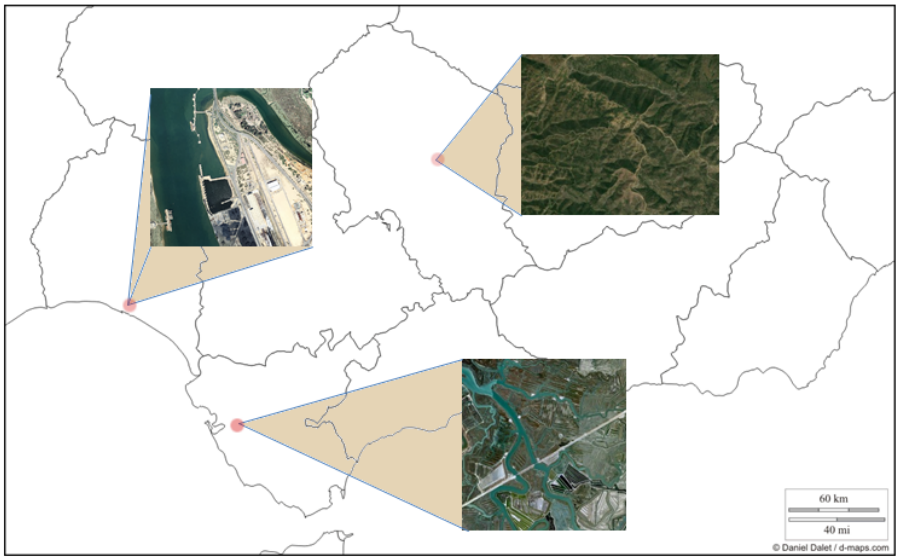


Figura 4.1: Zonas de estudio cedidas por la Junta de Andalucía

El presente capítulo tiene pues como objetivo, presentar las distintas zonas en las que se ha basado este trabajo de tesis, haciendo especial hincapié en las características de los vuelos, los datos auxiliares disponibles y los principales tipos de usos y coberturas predominantes en las distintas áreas.

4.2. Huelva

Los datos de este área fueron proporcionados, como hemos comentado, por la Consejería de Medio Ambiente a través de la REDIAM (Red de Información Ambiental de Andalucía) a la que el grupo TIC-134 pertenece y cuyo principal socio es la Junta de Andalucía. Los datos de Huelva provienen de vuelos realizados en zonas costeras de las provincias de Huelva (ver Figura 4.2) y Cádiz entre los días 23 y 25 de Septiembre de 2007. Los datos cedidos corresponden a una parte del vuelo que cubre un área semi-urbana en la desembocadura de los ríos Tinto y Odiel. El vuelo fue operado a una altura media de 1200 m con ángulos de inclinación bajos ($< 11°$) y densidad nominal de 1 pulso/m^2. Esta densidad baja de pulsos se justifica debido a que su uso estaba planificado originalmente para el desarrollo de modelos digitales del terreno (MDTs) para el que dicha densidad se ha demostrado suficiente. Los pulsos fueron georreferenciados y validados correctamente por el distribuidor de los datos. La precisión reportada indica un error máximo de 0,5 m. en las posiciones $x - y$, y de 0,15 m en la posición z. En este caso, los datos fueron cedidos en formato LAS por lo que se tenía la información completa (intensidad, ángulo, etc.) para cada retorno de

los pulsos.

Junto con el vuelo LIDAR, se tomaron fotografías aéreas de la zona que también fueron georreferenciadas. Las fotografías se tomaron con una resolución de 0,2 m y en este caso tan sólo contenían información de las bandas roja, verde y azul en el espectro visible.

Figura 4.2: Zonas de estudio de la desembocadura del Tinto y el Odiel.

A partir de los datos de esta zona, se generaron dos conjuntos de etiquetas distintos. Las muestras de la base de conocimiento fueron recolectadas en todo momento a partir de la observación de las ortoimágenes disponibles con resolución 0,2 m. Para definir los tipos de coberturas o usos, fue necesario utilizar tanto las ortofotografías de la zona como el mapa de usos del período 1999-2003 en su versión actualizada del 2006, ya que no se disponía de datos de campo en la zona.

4.2.1. Clasificación de tipos de suelo inicial

Para generar el conjunto de etiquetas posibles, hay que tener en cuenta la situación de los datos. Así, en la desembocadura de un río y cerca de la capital, los datos presentan una mezcla de usos y coberturas de origen industrial y natural. En este caso, las clases de coberturas del terreno tratan de controlar el impacto humano en una zona ribereña, por lo que las clases de interés fueron los siguientes siete tipos de usos o coberturas:

- Agua. Dado que se trata de un puerto fluvial, la zona inundada es una clase clave.

- Marisma. Las zonas marismales forman el principal tipo de suelo natural en las zonas no inundadas más cercanas al río.
- Carreteras y otras vías de comunicación. En el área se encuentra una importante carretera junto a algunas vías ferroviarias y otras carreteras secundarias que sirven para acceder a las áreas de servicios como por ejemplo, la zona de descarga de materiales.
- Vegetación baja. Esta etiqueta delimitaría la zona en la que existen zonas de hierbas bajas o directamente suelo desnudo y que no fuera considerada como marisma.
- Vegetación media. Esta clase engloba las zonas de matorrales, muy numerosas cerca de la carretera principal.
- Vegetación alta. En el área, existe una zona con árboles del género *eucalyptus* aunque ocupando una extensión mínima respecto al total del área.
- Zona urbana. Finalmente, se clasificaría como zona urbana cualquier área con estructuras desarrolladas por el hombre no contemplada anteriormente, como vertederos, edificios o zonas de almacenamiento de mercancias.

4.2.2. Clasificación de tipos de suelo ampliada

Posteriormente, la clase zona urbana se dividió en dos (edificios y zonas industriales, y vertederos) para realizar una clasificación más fina de la escena. De esta forma, se pasó de siete a ocho tipos de suelo. La clase edificios incluiría también las zonas de servicios como las de carga-descarga o almacenamiento. Como en el caso anterior, para la selección de muestras se utilizó principalmente la interpretación de ortoimágenes con conocimiento previo de la zona extraído de anteriores mapas de uso del suelo de la Consejería de Medio Ambiente.

4.3. Trabada

El Laboratorio del Territorio de la Universidad de Santiago (Campus de Lugo), cedió para los estudios relacionados con esta propuesta de tesis, dos conjuntos de datos del norte de España. El primero de ellos engloba el área de estudio de Trabada, que se localiza en el noreste de Galicia (punto azul sobre fondo amarillo en Figura 4.3) y se compone básicamente de pequeñas zonas residenciales y bosques cuya especie dominante es el *eucalyptus globulus*. Los datos LIDAR se adquirieron en noviembre de 2004 con un sensor *ALTM Optech 2033* a una altitud de vuelo de 1,500 m. El sensor utiliza una longitud de onda del láser de 1,064 nm, y la divergencia del haz se fijó a 0,3 mrad. La frecuencia de pulsos fue de 33 kHz, la frecuencia del análisis de 50 Hz y la amplitud del ángulo de barrido de $\pm 10^{\circ}$. Se registraron sólo los primeros y últimos retornos de cada pulso y los valores de altura e intensidad. El área de estudio completa

Figura 4.3: Zonas de estudio cedidas por el Laboratorio del Territorio. En azul con fondo amarillo, Trabada. En azul con fondo verde, Guitiriz.

fue dividida en 18 franjas y cada franja fue volada tres veces, lo que dio una densidad media de aproximadamente 4 pulsos/m^2.

Aparte de los datos LIDAR, se tienen imágenes aéreas tomadas de la zona con una resolución de 0,5 m^2 y que, de nuevo, contienen únicamente información del espectro visible.

4.3.1. Clasificación de tipos de suelo

Del conjunto total de datos del vuelo de Trabada, se seleccionó un área de aproximadamente 0.4 km^2 para tareas de clasificación (ver Figura 4.4). En este caso, sólo se consideraron cuatro tipos de usos del terreno:

- Edificios. Formado por las casas de la aldea que aparecen en la escena.
- Carreteras. Esta etiqueta define las zonas con caminos y las calles que comunican el pueblo.
- Cultivos y áreas de pastos. Zona de hierba baja o suelo desnudo de la escena.
- Zonas forestales. Principal clase en el área formada por los individuos de la especie *eucalyptus globulus*.

4.3.2. Estimación de biomasa

Una de las tareas en las que LIDAR ha demostrado una gran utilidad ha sido en la extracción de variables biofísicas para estimación de biomasa (Gonçalves-Seco et al., 2011). Para este tipo de estudios, se suele realizar trabajo de campo que mide las variables biofísicas que definen la cantidad de biomasa en un conjunto de parcelas. Posteriormente mediante técnicas de regresión, se buscan ecuaciones que las relacionen con estadísticas de la nube de puntos. Para esta propuesta de tesis, se usó trabajo de campo realizado en una amplia zona del conjunto volado (véase la Figura 4.5).

Figura 4.4: Zona de estudio de Trabada utilizada para realizar clasificación.

En este caso, se llevó a cabo un inventario forestal formado por 39 parcelas cuadradas de 15 m^2 en plantaciones maduras de *eucalyptus globulus* durante los meses de febrero y marzo de 2005. A partir de ese trabajo de campo, se calcularon la biomasa de copa (W_{cr}), la biomasa del tronco (W_{st}) y la biomasa total (W_{abg}) de las 39 parcelas.

4.4. Guitiriz

El segundo área en Galicia se localiza en la zona de Guitiriz (punto azul sobre fondo verde en Figura 4.3). Esta zona, al igual que Trabada, se encuentra en la provincia de Lugo, pero en su parte occidental. El área registrada cubre aproximadamente 36 km^2 de bosques cuya especie dominante es el *Pinus radiata* y los datos volados se encuentran dentro de un rectángulo de $4,130 \times 8,787$ con las coordenadas UTM: (586315; 4783000), (595102; 4787130). Los bosques de esta zona son representativos de los rodales naturales de *Pinus radiata* atlánticos localizados en Galicia, que se caracterizan por tratamientos silvícolas de baja intensidad y por presencia de arbustos altos.

Los datos LiDAR fueron adquiridos en septiembre de 2007 con un sistema Optech ALTM 3025, que opera a 1064 nm, con una tasa de repetición de 25 kHz, una frecuencia de barrido de 200 Hz, un ángulo de lectura máxima de $\pm 17°$ y una altura de vuelo de 1300 m sobre el nivel del mar. El vuelo se planteó para que hubiera un solapamiento entre las pasadas del 60 % por lo que la densidad teórica es de 8 $pulsos/m^2$.

De cada pulso se registró: el tipo de retorno (primero o último), coordenadas

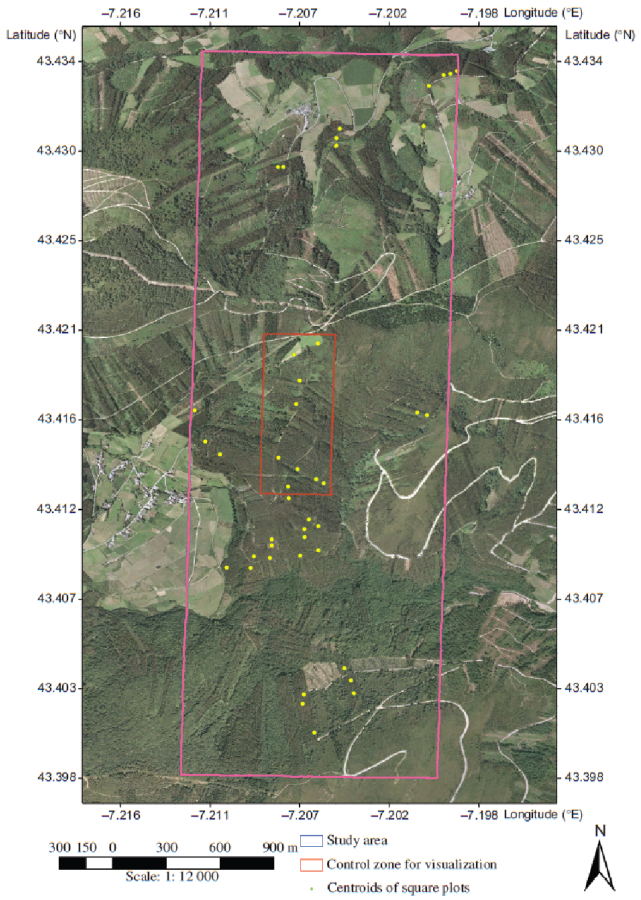


Figura 4.5: Área de Trabada extraída de Gonçalves-Seco et al. (2011). En púrpura, los límites de la zona volada. En rojo, las zonas de control para tareas de visualización. En amarillo, el centro de cada parcela del inventario.

X, Y y Z, y el valor de intensidad. Los retornos únicos se registraron dos veces con la misma información pero con distinto tipo de retorno. Como hemos comentado anteriormente, la densidad nominal teórico del vuelo LiDAR fue de 8 $pulsos/m^2$, lo que implica tener 16 $retornos/m^2$.

Una posterior selección aleatoria de los pulsos LiDAR permitió extraer un dataset con densidad 0,5 pulsos/m^2. Estos datos son utilizados para evaluar la influencia de la densidad en las conclusiones finales. En nuestro caso, trataremos estos datos como nueva información puesto que nuestro objetivo será, como veremos más adelante, la comparación de técnicas de regresión y no la influencia de los parámetros de vuelo en los resultados finales.

4.4.1. Estimación de variables forestales

Para poder aplicar técnicas de regresión, un total de 54 parcelas cuadradas de 225 m^2 fueron geolocalizadas y medidas en las plantaciones de *Pinus radiata* (ver Figura 4.6) entre agosto y diciembre de 2007. Dichas parcelas fueron seleccionadas para representar la variedad existente en la zona en lo que respecta a edades, densidades y tipos de localizaciones.

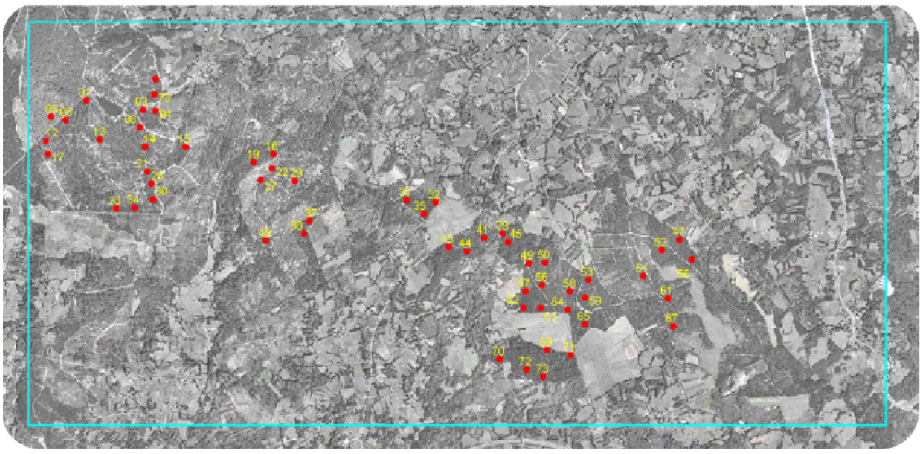


Figura 4.6: Área de Guitiriz: zona volada delimitada por rectángulo azul y parcelas en las que se realizó trabajo de campo junto con su identificador.

Para todos los árboles en cada parcela de muestreo, se realizaron dos mediciones de diámetro a la altura del pecho (1,3m sobre el nivel del suelo) y se calculó la media aritmética de las dos mediciones. Además, se midió la altura total en todos los árboles. Las mediciones de campo (alturas y diámetros), los volúmenes estimados y el peso seco de las fracciones de biomasa fueron utilizadas para estimar distintas variables de cada parcela en base al número de hectáreas: altura media (HM), altura dominante (Hd), área basal (G), volumen (V), biomasa de copa (W_{cr}), biomasa de tronco (W_{st}) y biomasa total (W_{abg}). Estas estadísticas definen las variables dependientes a la hora de establecer las regresiones.

4.5. Alto Tajo

El último conjunto de datos utilizado fue cedido, como hemos comentado, por el Dpto. de Geografía de la Universidad de Alcalá previo permiso de los propietarios reales del vuelo. Así, se cedieron los datos de una zona del Parque Natural del Alto Tajo en Guadalajara (Figura 4.7), en el centro de España (UL: 40° 56’ 49” N; 2° 14’ 49” W; LR: 40° 48’ 25” N; 2° 13’ 21” W). El área tiene

una topografía accidentada, con una altitud media de 1200 m, que oscilan entre los 895 hasta los 1403 m.

El United Kingdom Natural Environment Research Council (NERC) Airborne Research and Survey Facility voló sobre el área de estudio dos veces al final de la primavera de 2006 (16 de mayo y 3 de junio). Ambos vuelos incluyeron LIDAR aerotransportado y fotografía multiespectral (Airborne Thematic Mapper, ATM) con resolución de 2 m.

El sistema LIDAR utilizado fue un Optech-ALTM3033, con un pulso láser de 33 kHz. La altura media de vuelo fue de entre 750 m y 775 m sobre el nivel del suelo para el primer y segundo vuelo respectivamente, con un ángulo de escaneo máximo de $\pm 2°$ y divergencia máxima del haz de 0,2 mrad, creando un diámetro de huella en el nadir de aproximadamente 18 cm. La densidad de puntos de media para cada línea de vuelo fue de aproximadamente de 1,5 a 6 retornos m^{-2}. Los datos facilitados por el NERC incluyeron posición X, Y, Z e intensidad del primer y último retorno.

De los datos ópticos, sólo la imagen correspondiente a la primera fecha fue seleccionada por el poco tiempo de diferencia entre los dos vuelos y la falta de evidencia de cambios fenológicos.

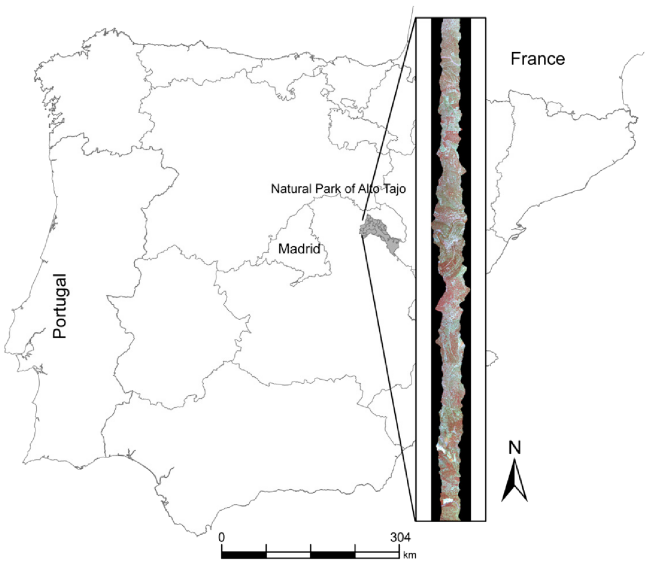


Figura 4.7: Zonas de estudio de Alto Tajo en Guadalajara.

4.5.1. Clasificación de tipos de suelo

Para esta zona de trabajo, se utilizó un subconjunto del vuelo original, con una superficie de $9 \times 0,3\ km^2$. Aunque el número de especies en el área es muy

variado, el número de clases se fijó en seis, debido a que las clases con similares respuestas espectrales se fusionaron. Así, debido a que la clasificación de la cobertura del terreno estaba destinada a la gestión del riesgo de incendios las clases de interés fueron pastos (dos clases espectrales), matorrales (dos clases espectrales), árboles (pinos, encinas y álamos), así como una clase adicional correspondiente a coberturas no pertenecientes al conjunto de combustibles (elementos interesantes para estudios de propagación del fuego). Desde el punto de vista biológico, las etiquetas seleccionadas se correspondieron con:

- Bosque de pinos. Principalmente individuos de la especie *Pinus nigra* Arn.
- Bosques de encinas. Áreas con predominio de *Quercus ilex* L.
- Plantación de álamo. Etiqueta que determinó las zonas con predominio de individuos de la especie *Populus tremula* L.
- Estrato arbustivo. Esta clase está formada a su vez por los individuos jóvenes de pino (*Pinus nigra* Arn.; *Pinus sylvestris* L.; *Pinus pinaster* Ait.), encina y cedro español (*Juniperus oxycedrus* L.) y zonas de aulaga (*Genista scorpius*).
- Pastos. Las áreas despejadas con poca vegatación se clasificaron con esta etiqueta.
- No combustibles. Clase final para las zonas como suelo desnudo o carreteras y construcciones que no forman parte del conjunto de combustibles.

Para la selección de la base de conocimiento usada en las fases de entrenamiento y validación de los diferentes clasificadores, se usaron muestras de la imagen a resolución 0,5 m seleccionadas aprovechando el conocimiento previo de la zona adquirido durante el trabajo de campo. Para discriminar las propiedades verticales de ciertas clases, especialmente entre los arbustos de encina joven (≤ 4 m) y encinas (> 4 m), se utilizaron las alturas normalizadas de LIDAR.

4.5.2. Estimación de biomasa forestales

De nuevo, los datos LIDAR se aprovecharon para realizar estudios sobre las variables forestales. Así, como en los anteriores casos, se realizó trabajo de campo durante los meses de julio y agosto de 2006 y 2008. En concreto, se estudiaron 45 parcelas circulares con un radio de 10 m en varias zonas distribuidas a lo largo de las tres líneas de vuelo. Las parcelas representan las principales especies presentes en el área cubierta por los datos LiDAR (*P. nigra Arn.*, *Juniper thurifera L.*, *Quercus ilex L.*) y su selección se realizó tomando como referencia la imagen ATM adquirida simultáneamente con los datos LiDAR.

En cada parcela, se registraron las especies de árboles y el diámetro a la altura del pecho (DAP) para cada árbol con un DAP ≥ 10 cm. Posteriormente se aplicaron ecuaciones alométricas específicas de la especie para estimar W_{cr}, W_{st} y W_{abg}. Finalmente, para cada parcela los valores para W_{cr}, W_{st} y W_{abg} se obtuvieron sumando los valores de los árboles individuales.

Capítulo 5

Mapas temáticos a partir de datos LIDAR

Un mapa no es el territorio que representa, pero, de ser correcto, tiene una estructura similar al territorio, razón por la cual resulta útil.
Alfred Korzybski.

5.1. Introducción

La influencia humana sobre los entornos naturales es un hecho a nivel global. El ser humano y su actividad es el principal factor que influye en la evolución del medio natural y la presión humana es un factor clave a la hora de planificar políticas de protección de los espacios naturales. Dos factores claves que se suelen estudiar para detectar actividad humana son las vias de comunicación y los núcleos urbanos (Goetz et al., 2009; Svancara et al., 2009). Estos parámetros permiten evaluar el riesgo de daños en las zonas naturales cercanas (Jones et al., 2009) y trazar políticas para corregir los posibles efectos del desarrollo humano. Una de las herramientas más importantes para controlar estos factores es el control y la monitorización de los usos de suelo y las coberturas del terreno mediante mapas temáticos generados a partir de teledetección.

Muchos autores han usado los mapas temáticos generados a partir de datos remotos para tareas tales como monitorizar especies (Stow et al., 2008) o cambios en las ciudades (Gamanya et al., 2009), detectar distintos tipos de terrenos para desarrollar explotaciones madereras (Hughes et al., 2009), caracterizar la morfología de los nucleos urbanos (Gill et al., 2008), estudiar la riqueza de las especies de aves en zonas naturales (Goetz et al., 2007) o la severidad de incendios (Kokaly et al., 2007), etc. La mayor parte de estos autores apuestan por técnicas de teledetección clásicas (aprendizaje supervisado) basadas en imágenes de satélites (Pignatti et al., 2009; Townsend et al., 2009; Fraser et al., 2009).

Aunque la principal aplicación de LIDAR sea generar MDTs, su capacidad para registrar la altura de los objetos provee una nueva fuente de datos con

fuertes sinergias respecto a las imágenes aéreas. Es por ello que los sensores LIDAR están empezando a convertirse en una excelente herramienta para mejorar los resultados de la clasificación en teledetección (Chen, 2007). Sin embargo, son pocos los estudios que evalúen la capacidad de LIDAR para generar mapas temáticos de manera aislada, centrándose la mayor parte de los trabajos directamente en los resultados de su fusión con otras fuentes.

En este capítulo nos proponemos evaluar LIDAR como fuente de datos para generar mapas temáticos. Para ello, es preciso comprobar la capacidad de los clasificadores clásicos sobre este tipo de datos. Es importante subrayar que, hasta donde sabemos, no existe ninguna comparativa entre clasificadores aplicada a este tipo de sensores de manera aislada. En este capítulo en concreto, establecemos la comparación de las distintas técnicas en basa a la precisión alcanzada a la hora de clasificar la zona descrita en la subsección 4.2.1.

5.2. Trabajos relacionados

Muchos investigadores han optado por utilizar LIDAR como tecnología de apoyo a las imágenes. Así, se apuesta por la fusión de sensores con el objetivo de mejorar los resultados obtenidos por separado (Arroyo et al., 2008; Bork and Su, 2007; Chust et al., 2008; Dalponte et al., 2008). Sin embargo, frente a la corriente mayoritaria, es posible encontrar un grupo más reducido de autores que defiende el uso de LIDAR como única fuente de datos con resultados notables (Pascual et al., 2008; Chen, 2007).

Cada estrategia tiene sus propios pros y contras. La fusión proporciona gran cantidad de datos que produce un extra de información para cualquier método de clasificación pero también, necesita un mayor esfuerzo para adaptar datos desde múltiples sensores dando lugar a un incremento en el tiempo de desarrollo y dimensionalidad de los datos. Además, algunos estudios muestran poca mejora en las clasificaciones basadas en LIDAR cuando se usa con otros sensores para realizar determinadas tareas (Jensen et al., 2008; Thessler et al., 2008) mientras que otros trabajos (Townsend et al., 2009) aconsejan ser precavidos a la hora de fusionar datos de manera general, incluso si son de la misma tipología como pueden ser las imágenes de satélite.

En los últimos tiempos, se han comenzado a aplicar técnicas orientadas a objetos sobre LIDAR como única fuente de datos con buenos resultados (Antonarakis et al., 2008; Donoghue et al., 2007). A pesar de que los resultados son muy prometedores, existen problemas aún no resueltos. Fundamentalmente porque la segmentación en datos LIDAR no es un proceso fácilmente automatizable necesitando de la interacción con el usuario para alcanzar buenos resultados. Otro problema adicional es que uno de los parámetros más utilizados para realizar la segmentación en LIDAR es la reflectividad de los retornos. Este dato puede ser afectado por otros factores (angulo de incidencia, distancia al sensor (Hofle and Pfeifer, 2007)) que pueden modificar su valor e introducir ruido en el resultado final.

El grado de automatización del proceso de generación del mapa, también

tiene que tenerse en cuenta. Así, en muchos casos, los mapas se generan a partir de un modelo extraído de manera manual estudiando los datos (Antonarakis et al., 2008). En otros casos, se aplican clasificadores avanzados en la generación de mapas temáticos como SVMs (Koetz et al., 2008), ANNs (Brzank et al., 2008) o clustering (Pascual et al., 2008) sobre los datos LIDAR como única fuente de datos. De cualquier manera, no es sencillo encontrar aportaciones en lo que se refiere a la generaración de mapas temáticos con LIDAR de manera exclusiva, y en los casos que se pueden encontrar, suelen centrarse en un único clasificador sin mostrar ninguna comparativa con otras técnicas de clasificación.

5.3. Metodología

Como se ha comentado anteriormente, los mapas temáticos de usos de suelo y coberturas vegetales (land use and land cover, LULC) son generados en general a partir de técnicas de teledetección (Dorigo et al., 2007). En nuestro caso, para desarrollar estos productos, es preciso clasificar los datos LIDAR. Con este objetivo, proponemos aplicar técnicas inteligentes de aprendizaje supervisado sobre un conjunto de medidas previamente extraídas de la nube de puntos.

Preprocesamiento

Para clasificar los datos LIDAR hemos seleccionado una estrategia orientada a parcelas (polígonos) debido a que es más fácil automatizarla y no depende de tantos parámetros como la orientación a objetos. Este enfoque nos obliga a crear una matriz representando el área de estudio donde cada elemento es una parcela cuadrada. Cada parcela representa una cantidad del terreno en función de la resolución. El valor de resolución debe ser proporcionado por el usuario. En nuestro caso, se ha determinado fijar la resolución en 4 m^2 para que cada parcela trabaje con una media de 4 pulsos LIDAR (el número de retornos medio estará entre 4 y 12). La resolución depende de la densidad de puntos directamente: 1 pulsos/m^2 no aconseja usar una resolución mayor debido a que no se tendrían suficientes pulsos para aplicar los algoritmos de manera fiable y, por otro lado, resoluciones menores producirían ruidos en las clases pequeñas como las vías de comunicación, que no suelen tener anchos mayores de 3 ó 4 metros. Además de la resolución, es necesario suministrar un MDT para extraer las alturas reales (alturas normalizadas) de los retornos. En nuestro caso, lo generamos mediante filtros morfológicos (Gonçalves-Seco et al., 2011).

Es importante tener en cuenta que el vuelo debe ser preprocesado para eliminar distintos tipos de ruido. En este caso, aplicamos dos tipos de técnicas. Con la primera, se realiza una corrección de la intensidad (Hofle and Pfeifer, 2007) de acuerdo a la Ecuación 5.1 donde I es la intensidad original para un retorno, R es la distancia desde la fuente láser al retorno más lejano y R_s es la distancia real desde la fuente al retorno en sí. La segunda técnica de preprocesado aplica métodos estadísticos de eliminación de outliers sobre la componente z de los puntos. En este caso, se aplica un filtro basado en el percentil 95 sobre la componente

altura de manera que, para los datos de Huelva, se eliminan aquellos puntos con alturas superiores a 17 m. Estos retornos se consideran puntos escapados o rebotes en aves y no se tienen en cuenta en el resto del procedimiento.

$$I(R_s) = I * \frac{R^2}{R_s^2} \tag{5.1}$$

Generación del modelo

El aprendizaje supervisado necesita datos previamente clasificados. Para generarlos, se usó el conocimiento extraído de los productos disponibles en la REDIAM (mapas de usos oficiales y fotografías tomadas en el mismo vuelo que los datos LIDAR) debido a que no existía trabajo de campo en el área. De esta forma, se clasificó manualmente alrededor de un 3 % del total de los datos.

Una vez fijada la resolución y preprocesados los datos de entrada, se procede a construir una matriz de datos. Cada celda de la matriz representa una parcela para la que se calcula un conjunto de medidas basadas en la intensidad, altura y distribución de los pulsos. Estas medidas pueden clasificarse como intrapíxel o interpíxel. Las medidas intrapíxel son aquéllas que se calculan con los datos de los pulsos que quedan dentro de un píxel, mientras que las medidas interpíxel se caracterizan por definir una relación entre cada píxel y sus ocho vecinos adyacentes, por lo que se introduce cierto grado de información contextual en la clasificación (en el Capítulo 7, se profundizará en las técnicas contextuales). Con estas medidas, se intentará posteriormente caracterizar el terreno, formalizando las diferencias visuales o morfológicas de las distintas clases.

El Cuadro 5.1 contiene las treinta y tres medidas utilizadas en este estudio. La mayoría de las medidas se han extraído de la bibliografía (Hudak et al., 2008) salvo las siguientes, que fueron contribuciones originales de esta investigación:

- RZDIFF: Sumatorio de las distancias entre los primeros retornos y los posteriores retornos.
- RDIFF: Diferencia de alturas entre píxeles vecinos.
- EMP: Número de píxeles vecinos vacíos.

Cada celda seleccionada como entrenamiento debe ser etiquetada con una clase concreta. Una vez se tiene completa la base de entrenamiento formada por la información de los píxeles etiquetados con sus clases, se eliminan los valores «missing», se escalan los datos al intervalo $[0, 1]$ y se aplica un método de selección de atributos sobre la base de entrenamiento CFS (Hall, 1999). CFS es un selector de atributos clásico (no es objetivo de este estudio profundizar en las técnicas de selección de atributos sino en los métodos de clasificación) que evalúa la valía de cada atributo considerando su capacidad predictiva y el grado de redundancia con el resto de atributos. En concreto, CFS evalúa subconjuntos de atributos y selecciona aquél que esté altamente correlado con la clase y que tenga poca correlación entre los atributos que lo componen. En el

Variable	Description	Type
IMIN	Intensity minimum	Intrapixel
IMAX	Intensity maximum	Intrapíxel
IMEAN	Intensity mean	Intrapixel
IVAR	Intensity variance	Intrapixel
ISTD	Intensity standard deviation	Intrapixel
IAAA	Intensity average absolute deviation	Intrapixel
IRANGE	Intensity range	Intrapixel
HMIN	Height minimum	Intrapixel
HMAX	Height maximum	Intrapixel
HMEAN	Height mean	Intrapixel
HVAR	Height variance	Intrapixel
HSTD	Height standard deviation	Intrapixel
HAAA	Height average absolute deviation	Intrapixel
HRANGE	Height range	Intrapixel
IKURT	Intensity Kurtosis	Intrapixel
ISKEW	Intensity Skewness	Intrapixel
HKURT	Height Kurtosis	Intrapixel
HSKEW	Height Skewness	Intrapixel
ICV	Intensity coefficient of variation	Intrapixel
HCV	Height coefficient of variation	Intrapixel
SLP	Slope	Interpixel
RDIFF	Relative difference among neighbors	Interpixel
RZDIFF	Elevation difference first and last return	Interpixel
PCT1	Percentage 1st returns	Intrapixel
PCT2	Percentage 2nd returns	Intrapixel
PCT3	Percentage 3rd or later returns	Intrapixel
PCT31	Percentage 3rd over 1st returns	Intrapixel
PCT21	Percentage 2nd over 1st returns	Intrapixel
PCT32	Percentage 3rd over 2nd returns	Intrapixel
NOTFIRST	Percentage 2nd or later returns	Intrapixel
EMP	Empty plots surrounding	Interpixel
TPO	Total number of points	Intrapixel
CRR	Canopy relief ratio	Intrapixel

Cuadro 5.1: Treinta y tres variables predictoras. En negrita, las trece seleccionadas tras realizar la fase de selección de atributos.

caso que nos ocupa, el método seleccionó 13 atributos: IMEAN, IMIN, HMEAN, HMIN, HMAX, HCV, SLP, CRR, ISKEW, PCT21, EMP, RDIFF y RZDIFF. Como puede observarse, las 3 nuevas medidas fueron incluidas en el subconjunto, verificándose por tanto, su importancia.

Con las medidas seleccionadas ya generadas para todos los píxeles de la matriz, se pasa a la fase de ejecución del algoritmo de clasificación. Se estudiaron tres tipos de técnicas inteligentes para extraer el modelo: SVMs, ANNs y DTs. Se escogieron las siguientes implementaciones: en el caso de la ANN, un Multilayer Perceptron (Haykin, 1998); para la SVM, se utilizó la implementación LibSVM (El-Manzalawy and Honavar, 2005); para la definición del árbol de decisión, se utilizó el algoritmo C4.5 (Quinlan, 1996). Los tres modelos fueron ejecutados en el entorno WEKA (Hall et al., 2009).

5.4. Resultados

5.4.1. Comparativa entre métodos

Para establecer la comparativa entre los métodos realizamos una validación cruzada con diez carpetas (10-fold cross validation, 10-FCV) sobre los datos. Usamos 10-FCV porque la selección de instancias de test puede provocar que un clasificador obtenga los mejores resultados por sobreajuste accidental, aún cuando no sea el que mejor se adapte a ese tipo de datos de manera general. Mediante un 10-FCV es posible generalizar los resultados ya que los clasificadores deben obtener buenos resultados no en una prueba sino en la media de 10 ejecuciones distintas. Esta técnica ha demostrado ser robusta y nos servirá para posteriormente validar estadísticamente el estudio experimental mediante un test de significación estadística.

Las pruebas realizadas con parámetros estándar para los clasificadores en WEKA muestran que los DTs obtienen los mejores resultados. En los Cuadros 5.2, 5.3 y 5.4 se muestran las precisiones totales, parciales (precisión de usuario y productor) y el coeficiente Kappa obtenido para las tres técnicas como resultado de la 10-FCV. Las tres técnicas obtienen precisiones altas a la hora de clasificar los píxeles, pero la aplicación del árbol de decisión produce una mejora de casi dos puntos porcentuales.

En García et al. (2009), se muestra que se necesita un gran número de conjuntos de datos para garantizar que una comparación estadística sea robusta. Como mínimo, el número de conjuntos de datos debe ser de al menos dos veces el número de clasificadores y cuanto mayor sea el número de diferentes conjuntos de datos, más fiable será el estudio. En nuestro caso, sólo tenemos un conjunto de datos LIDAR por lo que, para generar el número de resultados adecuado, dividimos el conjunto de entrenamiento en 5 subconjuntos disjuntos. A continuación, hicimos un 10-FCV para cada subgrupo. Así, obtuvimos 50 medidas de precisión para cada algoritmo.

Tradicionalmente, se aplican tests paramétricos como ANOVA con el fin de evaluar la significación estadística de las diferencias medidas entre los algorit-

Class \sample	Water	Marsh	Roads & railways	Low Veg.	Mid. Veg.	High Veg.	Urban terrain
Water	87	1	0	0	0	0	0
Marshland	2	624	0	0	3	0	0
Roads and railways	0	1	179	0	2	0	20
Low Veg.	0	0	0	138	1	0	3
Middle Veg.	0	1	2	0	116	1	14
High Veg.	0	0	0	0	1	103	10
Urban terrain	0	19	5	7	11	5	812
Producer's	0.98	0.97	0.96	0.95	0.87	0.94	0.95
User's	0.99	0.99	0.98	0.97	0.87	0.9	0.94
Total	94.97						
KIA	0.93						

Cuadro 5.2: Resumen de las pruebas sobre SVM y matriz de confusión

Class \sample	Water	Marsh	Roads & railways	Low Veg.	Mid. Veg.	High Veg.	Urban terrain
Water	83	5	0	0	0	0	0
Marshland	3	620	0	0	4	0	2
Roads and railways	1	1	185	0	9	0	6
Low Veg.	0	1	0	138	0	0	3
Middle Veg.	0	2	4	0	113	1	14
High Veg.	0	0	0	0	2	105	7
Urban terrain	1	17	7	6	16	12	800
Producer's	0.94	0.95	0.94	0.96	0.78	0.88	0.96
User's	0.94	0.99	0.92	0.97	0.84	0.92	0.93
Total	94.28						
KIA	0.92						

Cuadro 5.3: Resumen de las pruebas sobre ANN y matriz de confusión

Class \sample	Water	Marsh	Roads & railways	Low Veg.	Mid. Veg.	High Veg.	Urban terrain
Water	88	0	0	0	0	0	0
Marshland	0	624	0	0	0	0	5
Roads and railways	0	1	190	0	4	0	7
Low Veg.	0	0	0	134	0	0	8
Middle Veg.	0	1	3	0	120	2	8
High Veg.	0	0	0	0	0	105	9
Urban terrain	0	6	7	1	11	4	830
Producer's	1.0	0.99	0.95	0.99	0.88	0.95	0.96
User's	1.0	0.99	0.94	0.94	0.9	0.92	0.97
Total	96.45						
KIA	0.95						

Cuadro 5.4: Resumen de las pruebas sobre DTs y matriz de confusión

mos. Sin embargo, para que una comparación basada en una prueba paramétrica sea válida, los datos deben cumplir con los criterios de independencia, normalidad y homocedasticidad (Zar, 1999). En este caso, al tratarse de datos provenientes de 10-FCV, no se puede garantizar el principio de independencia por lo que se ha utilizado un procedimiento robusto (García and Herrera, 2008; Demsar, 2006) no paramétrico para comparar clasificadores sobre varios conjuntos de datos. El procedimiento elegido implica el uso del test de Friedman (Friedman, 1937, 1940) y del procedimiento post-hoc de Holm (Holm, 1979).

El primer paso para hacer uso del proceso de comparación es establecer un ranking medio en base a los resultados obtenidos por cada clasificador en cada dataset, donde un valor de 1.0 para un clasificador implicaría que bate al resto en todos los datasets. Seguidamente, se estudia el nivel de significación de las diferencias medidas entre los rankings de dichos clasificadores. Posteriormente, se aplica un procedimiento post-hoc para establecer si las diferencias entre pares de clasificadores son significativas. En nuestro caso, el objetivo fue establecer si la diferencia de precisión del algoritmo C4.5 (algoritmo de control) con la del resto de los algoritmos era significativa.

El test de Friedman es una prueba estadística no paramétrica para evaluar las diferencias entre más de dos medias muestrales relacionadas. Las muestras relacionadas son, en nuestro caso, las precisiones de los clasificadores sobre los distintos conjuntos de datos. Concretamente, estamos trabajando con los resultados de 10-FCV sobre cinco conjuntos de datos artificiales disjuntos seleccionados al azar del conjunto original. La hipótesis nula es pues, que todos los clasificadores realizan la clasificación con la misma calidad y que las diferencias observadas son debidas al azar. El estadístico que se usa en la prueba de Friedman se puede ver en la Ecuación 5.2.

i	Test	Averaged rank	p-value	α/i	Comment
1	ann	2.63	0.00001	0.0253	Reject Ho
2	svm	2.04	0.0017	0.0500	Reject Ho

Cuadro 5.5: Resumen del test de Holm con algoritmo de control C4.5 (ranking medio = 1.33).

$$X_F^2 = \frac{12n}{k(k+1)}(\sum_j {r_j}^2 - \frac{k(k+1)^2}{4}) \tag{5.2}$$

Una vez probado que se rechaza la hipótesis nula de Friedman, se procede al test post-hoc que evalúa las diferencias relativas entre el algoritmo de control y el resto de clasificadores. De esta forma, en el caso que nos ocupa, el procedimiento de Holm comprueba las hipótesis nulas secuencialmente ordenadas por su significancia. Se usa el estadístico z (ver Ecuación 5.3) comparando el i-ésimo clasificador contra el algoritmo de control con un α apropiado en cada paso. Dicho α se reduce en cada paso según la fórmula $\alpha/(k-i)$ para compensar las $k-i$ múltiples comparaciones. En el momento que una de las hipótesis no puede ser rechazada, las siguientes tampoco pueden serlo y termina el procedimiento.

$$z = \frac{(r_i - r_0)}{\sqrt{\frac{k(k+1)}{6N}}} = \frac{(r_i - r_0)}{\sqrt{SE}} \tag{5.3}$$

El análisis de rankings medios reveló que C4.5 obtenía los mejores resultados (ranking medio de 1.33). El resultado del test de Friedman fue de un p-value= $2{,}923E-10$ por lo que se rechazó la hipótesis nula. Habiéndo encontrado que los rankings medios medidos eran significativamente distintos ($\alpha = 0{,}05$), se procedió a aplicar el procedimiento post-hoc de Holm. Holm rechazó todas las hipótesis nulas como puede verse en el Cuadro 5.5 por lo que nuestro análisis confirmó que para un $\alpha = 0{,}05$, la precisión de C4.5 fue superior a la del resto de sus rivales (SVM y ANN).

5.4.2. Precisión del modelo

Como se ha señalado anteriormente, para construir el árbol de decisión se usó el 3 % de los píxeles disponibles. La clasificación posterior del resto de puntos produce la Figura 5.1 que será comentada en la siguiente sección. Para estimar el error de la manera clásica en teledetección (hold-out), se ejecutó un test estratificado con 187 puntos. Los datos de test fueron seleccionados aleatoriamente del conjunto inicial no clasificado y se evaluaron las clases a las que pertenecían a partir de información previa de la REDIAM. En un test estratificado, la proporción entre las clases se mantiene respecto a la proporción original de los datos de entrenamiento. En el Cuadro 5.7, se pueden observar los resultados del test mediante la matriz de confusión, precisiones de usuario y productor, y el estimador kappa del árbol obtenido en la fase de entrenamiento.

Attribute	Tree Level
HMEAN	0
HMAX	1
HMIN	2
RZDIFF	2
CRR	3
HCV	3
IMEAN	3
ISKEW	4
IMIN	5
RDIFF	6

Cuadro 5.6: Atributos finales seleccionados por C4.5 y su distancia a la raíz.

Class \sample	Water	Marsh	Roads & railways	Low Veg.	Mid. Veg.	High Veg.	Urban terrain
Water	32	0	0	0	0	0	0
Marshland	0	30	0	0	0	0	4
Roads and railways	0	0	28	0	0	0	7
Low Veg.	0	0	0	14	0	0	1
Middle Veg.	0	0	0	2	12	0	3
High Veg.	0	0	0	0	0	3	0
Urban terrain	0	0	2	4	5	1	38
Producer's	1.0	1.0	0.93	0.7	0.71	0.75	0.73
User's	1.0	0.88	0.8	0.93	0.71	1.0	0.76
Total	0.85						
KIA	0.81						

Cuadro 5.7: Resumen del conjunto de test y matriz de confusión para el árbol de selección generado.

El algoritmo C4.5 realiza otra selección de atributos aparte de la realizada con el CFS quedándose con 10 medidas finales. En ellas, cabe destacar que aparecen 2 de las nuevas medidas propuestas, RDIFF y RZDIFF. El nivel del nodo en el que se utiliza una medida da la importancia que tiene dicha medida para el algoritmo, siendo mejor cuanto más cercana de la raíz del árbol se encuentre. En el Cuadro 5.6, se puede ver las distancias de cada atributo seleccionado respecto a la raíz del árbol. Nótese que las nuevas medidas RZDIFF y RDIFF se utilizan en un nodo de nivel 2 y en un nodo de nivel 6, respectivamente, aunque como era de esperar, la mayor carga informacional se encuentra en las variables relacionadas con las alturas.

El resultado final de la clasificación es un mapa LULC a una resolución de 4 metros y una precisión media del 85 %. En la Figura 5.1, se muestra la ortofoto original, el conjunto de entrenamiento y el resultado final de la clasificación.

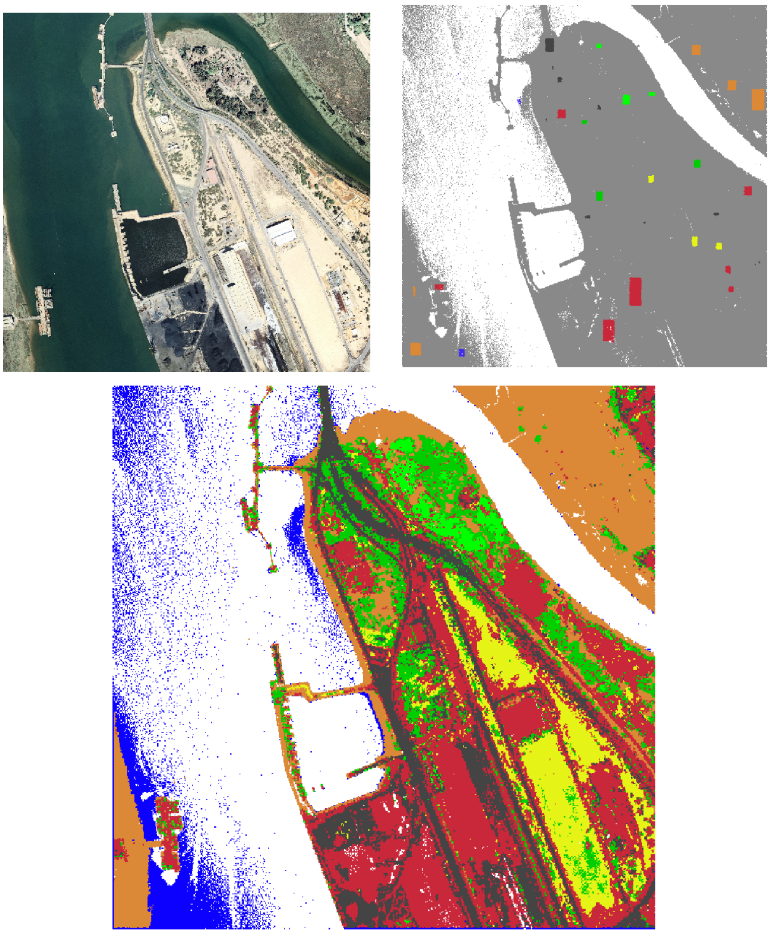


Figura 5.1: Ortofoto original, base de entrenamiento y resultado final : agua en azul, zonas urbanas en rojo, vías de comunicación en gris oscuro, vegetación media en verde, vegetación baja o suelo desnudo en amarillo, vegetación alta en verde claro, marismas en marrón y los datos no clasificados en gris claro.

5.5. Discusión

En lo que se refiere a los resultados del método de clasificación (Cuadro 5.7), hay que tener en cuenta que aunque las zonas ribereñas presentan grandes dificultades para ser clasificadas, los resultados muestran una gran precisión global. Según estudios recientes relacionados con los mapas LULC (Shao and Wu, 2008), las precisiones generales de los mapas realizados para numerosas organizaciones no llegan al 85 % a pesar de que éste sea el estándar para considerarlos útiles. El método propuesto demuestra llegar a este nivel de manera general e incluso para determinadas clases superarlo ampliamente. Analizando más en profundidad los resultados, vemos que se prueba la separabilidad entre clases mediante datos LIDAR y que dichos datos muestran que el método puede ser aplicado sin necesidad de otras fuentes auxiliares.

El hecho de que los DTs hayan obtenido los mejores resultados, no asegura que para un mayor número o complejidad de las clases, y/o mayor cantidad de información espectral, este hecho se vaya a repetir. De hecho, como veremos en el siguiente capítulo, las SVMs se comportan mejor bajo los anteriores supuestos aunque dependen de una mayor parametrización. Esta característica, en determinados ámbitos, puede ser un gran handicap sobre todo cuando usuarios no expertos (en alguna disciplina afín a la minería de datos) son los encargados de explotar la información de los datos LIDAR. En cualquier caso, es importante destacar que todas las técnicas usadas muestran un alta capacidad para clasificar de manera automática los usos de suelo y coberturas vegetales a partir de los datos LIDAR como única fuente de datos.

En lo referente al análisis de cada clase, se puede observar que los mejores resultados coinciden con las zonas inundadas, vías de comunicación y marismas con precisiones que superan el 90 %. Los peores resultados son los que proporciona la clasificación de vegetación media y zonas urbanas con valores alrededor del 72 % de precisión global. Esto último se debe a que, por un lado, la vegetación media es una clase intermedia que puede ser confundida por el algoritmo con baja vegetación y que, por otro lado, la clase suelos urbanos es una clase compleja con muchos subtipos y algunos de ellos pueden compartir características con otras clases, lo que conduce a una mayor probabilidad de fallo. Además, aunque se han obtenido buenos resultados, se puede detectar en el resultado final algunas zonas con un nivel relativamente alto de errores, principalmente en las zonas del puerto. Esto es inherente tanto a las aproximaciones orientadas a píxeles, como a las orientadas a parcelas o polígonos, ya que intentan clasificar unidades informacionales sensibles a ruidos sin tener en cuenta su contexto. El mismo problema aparece en algunas construcciones sin tejado. Este problema volverá a ser tratado en los siguientes capítulos (Capítulo 7).

En el caso de la zona portuaria, los errores son especialmente graves debido a que el tipo de terreno que lo soporta posee una respuesta reflectiva muy parecida a otros tipos de terreno circundantes. Este problema está íntimamente relacionado también con la limitada información espectral que provee LIDAR. Además, algunas zonas muestran serios niveles de ruido que, de nuevo, se deben a las especiales características de estas aproximaciones y su facilidad para ser

afectada por outliers no detectados. Un ejemplo de este comportamiento son las zonas de pasto donde píxeles con igual distribución de alturas y densidades dan lugar a dos clases distintas. Teniendo en cuenta que LIDAR sólo trabaja con alturas, densidades e intensidades, se concluye que el error debe estar en la intensidad. De aquí, que sea necesario seguir trabajando en técnicas más completas de preprocesamiento de la intensidad.

Capítulo 6

Fusión de sensores LIDAR e imágenes

No te establezcas en una forma, adáptala y construye la tuya propia, y déjala crecer, sé como el agua.
Bruce Lee.

6.1. Introducción

Tanto la ortofotografía como las imágenes multiespectrales e hiperespectrales, han sido ampliamente utilizadas para generar mapas LULC a pesar de que tienen sus propias limitaciones como por ejemplo, el efecto que tienen las sombras sobre la clasificación final. Como se mostró en el capítulo anterior, LIDAR también tiene debilidades relacionadas con los problemas asociados a la reflectividad (Yoon et al., 2008). Es por ello que numerosos autores han planteado fusionar la información LIDAR con otras fuentes de datos, como las imágenes multiespectrales, para superar las limitaciones que ambas tecnologías tienen por separado. De esta forma, por un lado se aprovecha la mayor regularidad registrada por las imágenes en lo que se refiere a reflectividad y, por otro, la clasificación se beneficia de la capacidad de LIDAR para registrar las alturas de los objetos.

Como ya comentamos, la fusión entre los sensores aumenta el tamaño de los datos y hace complicado el trabajo de manera manual realizado por operadores aun cuando se utilicen modernas herramientas (eCognition Software, 2011). En este contexto, las técnicas inteligentes son una necesidad si se requiere un proceso completamente automático y abarcable en un proceso industrial serio. Así, algunos autores han demostrado la potencia de las técnicas inteligentes para seleccionar y clasificar a partir de fusión de datos (Nguyen et al., 2005; Mazzoni et al., 2007). Aunque estas aproximaciones abordan técnicas avanzadas de la minería de datos, el método más utilizado sigue basándose en estadística clásica. En concreto, el tradicional análisis de componentes principales (Principal Component Analysis, PCA) y la aplicación del clasificador MLC(Bork and Su,

2007).

Es importante destacar que la fusión de datos LIDAR con imágenes aéreas tiene una especial importancia en nuestro entorno debido a que recientemente, se han planteado por primera vez vuelos LIDAR y ortofotográficos conjuntos en el Plan Nacional De Observación Aérea, PNOA (Nacional, 2009). El principal cometido del sensor LIDAR en estos vuelos es proveer datos para realizar MDTs, pero el alto coste de este tipo de vuelos justifica plenamente su aplicación en nuevas áreas como el desarrollo de mapas temáticos. Así, con el desarrollo de mapas LULC, se puede amortizar las diversas campañas de recogidas de datos generando productos para monitorizar áreas naturales y urbanas e incrementando el retorno de inversión.

Teniendo en cuenta todo lo comentado, este capítulo tiene como objetivo arrojar un poco de luz en lo que se refiere al rendimiento de las principales técnicas de aprendizaje supervisado cuando se aplican sobre fusión de datos LIDAR e imágenes y, concretamente, cuando se usan sobre zonas de la península ibérica con el objetivo de que el estudio sea de aplicación directa a los datos del PNOA. Además, pondremos especial énfasis en comprobar el incremento de calidad (si existe) de los mapas temáticos por fusión de datos respecto a los generados mediante LIDAR de manera aislada.

6.2. Trabajos relacionados

Ya hemos comentado que es posible encontrar un gran número de aproximaciones en la literatura que optan por desarrollar aproximaciones que complementen las capacidades de los sensores pasivos tradicionales fusionándolos con la información LIDAR aprovechando la sinergía existente entre ambas fuentes de datos. Entre las técnicas inteligentes más utilizadas en los últimos tiempos para este tipo de tareas, aparecen las SVMs. Este tipo de clasificador pese a su mayor complejidad respecto a DTs o MLC en lo que se refiere a parametrización, ha obtenido buenos resultados incluso en los tests más extremos como, por ejemplo, la clasificación de especies (Dalponte et al., 2009). También, es fácil encontrar autores que utilizan SVMs sobre fusión de LIDAR e imágenes aéreas para realizar mapas de combustibles (Koetz et al., 2008) o autores que mejoran las anteriores SVMs con reglas de decisión para establecer mapas más complejos (García et al., 2011) pero manteniendo la fortaleza de las SVMs frente al resto de potenciales clasificadores.

En cualquier caso, el problema de discernir qué técnica supervisada es la mejor para trabajar con fusión de datos aún no está claro. Así, podemos encontrar autores que desarrollan mapas de combustibles sobre fusión de LIDAR e imágenes usando el clasificador MLC (Mutlu et al., 2008) mientras que otros defienden el uso de SVMs para la misma tarea (García et al., 2011), o autores que recientemente realizan clasificaciones de especies mediante DTs (Ke et al., 2010) con resultados similares a los obtenidos por las SVMs (Dalponte et al., 2009). Además, podemos encontrar autores que abogan por el uso de ANNs (Grebby et al., 2011) para desarrollar mapas litológicos o por aplicar el algoritmo NN

para realizar clasificaciones más sencillas (Haapanen et al., 2004) donde su rendimiento es comparable a clasificadores más complejos, pero donde su escaso número de parámetros y su sencillez conceptual tienen una mayor importancia.

Como se puede apreciar, con sólo un pequeño subconjunto de trabajos de la extensa bibliografía sobre el tema, prácticamente aparecen todas las familias conocidas de técnicas supervisadas. Sin embargo, son pocos los estudios que planteen qué técnica obtiene mejores resultados a la hora de tratar fusión de datos en general, y en particular, sobre la combinación de LIDAR con otras tecnologías.

Uno de los estudios más recientes en lo que respecta a la comparación de técnicas inteligentes, es la de Szuster et al. (2011) que establece una comparativa entre SVMs, ANNs y MLC y concluye que aunque las SVMs obtienen mejores resultados, MLC no dista mucho en lo que refiere a calidad en resultados. Esta conclusión no deja de ser sorprendente en tanto contradice los resultados del MLC en estudios de autores como Oommen et al. (Oommen et al., 2004) o de Otukei et al. (Otukei and Blaschke, 2010) donde MLC se ve superado por SVM y DT respectivamente. Este dato es especialmente interesante puesto que, como se comentó anteriormente, MLC es la técnica más utilizada en cualquier tarea de generación de mapas temáticos (Lu and Weng, 2007).

Finalmente, es importante recordar que, si bien los autores mayoritariamente coinciden en que la mayor cantidad de información de la fusión mejorará siempre lo obtenido mediante LIDAR, no existen estudios que cuantifiquen esa mejora de manera exacta y muchos autores defienden el uso de LIDAR como única fuente de datos para realizar mapas temáticos (Antonarakis et al., 2008).

6.3. Metodología

Con el fin de ajustarnos a la metodología de trabajo general realizada por los expertos en el área de la fusión de datos, en este capítulo, seguimos las indicaciones de García et al. (2011). Así, la metodología escogida sigue un enfoque orientado a parcela en la que los datos se remuestrean a un tamaño acorde a la resolución LIDAR, se realiza una fase de generación y selección de atributos, y finalmente se realiza una clasificación supervisada. La metodología completa aplicada a las zonas de Huelva (Subsección 4.2.2) y Alto Tajo (Subsección 4.5.1) se describe en las siguientes subsecciones.

6.3.1. Preprocesamiento de los datos de Alto Tajo

Los datos ATM fueron georreferenciados en base a los datos GPS/IMU recogidos durante los vuelos. Además, con el fin de asegurar un adecuado co-registro de los datos ópticos respecto a los datos LIDAR, algunos puntos de control fueron tomados a partir de la imagen de intensidad de los datos LIDAR. El RMSE obtenido fue menor de 1 píxel (2 m).

La imagen ATM se remuestreó de 2 m a un tamaño de píxel de 6 m, teniendo en cuenta el valor medio de todos los píxeles incluidos en cada píxel de 6 metros.

Esta decisión se tomó para asegurar que un número suficiente de retornos LIDAR (más de 54 puntos) se incluyeran dentro de cada píxel. Con el fin de eliminar el efecto de la pendiente del terreno y otros aspectos de la señal registrada por el sensor, la imagen de ATM se corrigió topográficamente utilizando el método propuesto por Soenen et al. (2005).

En cuanto a los datos LIDAR, después de clasificar la nube mediante software propietario (TerraSolid Limited, 2000) en puntos terreno y no terreno, se creó un DEM mediante la interpolación de los retornos del suelo. A continuación, la altura normalizada de cada punto de la vegetación se calculó como la diferencia entre la coordenada Z del punto, y el valor Z del MDE en la posición XY correspondiente.

6.3.2. Generación del modelo para Alto Tajo

Una vez que se terminó de procesar la imagen ATM y los datos LIDAR, se pasó a la fase de extracción de características. Se seleccionaron los valores medios de las bandas disponibles de la imagen (ver Cuadro 6.1) junto con un conjunto de variables extraídas de LIDAR y cuatro índices espectrales: índice de vegetación de diferencia normalizada (Normalized Difference Vegetation Index, NDVI), el índice de vegetación ajustada al terreno (Soil Adjusted Vegetation Index, SAVI), y el índice de diferencia normalizada en infrarrojo (Normalized Difference Infrared Index, NDII) con dos valores infrarrojos distintos.

Después, se derivaron las variables LIDAR a partir de la distribución de alturas de los primeros y últimos retornos dentro de cada celda de la cuadrícula de 6×6 m. De esta forma, es posible caracterizar la estructura de la vegetación de la zona de estudio. Estas variables incluyen la altura máxima, que fue considerada como el percentil 99 para evitar el ruido causado por posibles outliers, la media, la mediana, la desviación estándar, el rango, la asimetría, la curtosis y el coeficiente de variación de las alturas siguiendo el trabajo de otros autores del área (Donoghue et al., 2007; Jensen et al., 2008). Como puede observarse, la intensidad del retorno no es utilizado ya que se supone que esa información ya viene incluída en la banda del infrarrojo cercano de la imagen ATM. Además, el uso de la intensidad, como hemos visto, puede poner en riesgo la bondad de la clasificación en los casos de múltiples retornos (como por ejemplo, las zonas boscosas). Por otro lado, muchos de los atributos vistos en el capítulo anterior o bien, no pueden ser calculados ya que no se tienen todos los retornos para cada pulso (sólo primer y último) o por otro lado, no tienen sentido como el caso de celdas vecinas vacías.

El último paso consiste en fusionar las variables derivadas de LIDAR con la información de las bandas del ATM y los índices espectrales en una instancia para cada celda de la matriz. Partiendo del conjunto de entrenamiento visto en la subsección 4.5.1, es posible extraer las instancias de entrenamiento. Las variables fueron normalizadas para eliminar posibles efectos adversos que pudiera provocar la diversidad de rangos. Posteriormente, al conjunto de variables se le aplicó una selección de atributos. La selección se realizó mediante el software propietario ENVI (Exelis, 2000), basándose en un análisis de separabilidad de

Variable	Description	Variable	Description
B1	420-450 nm	**B2**	450-520 nm
B3	520-600 nm	**B4**	600-620 nm
B5	630-690 nm	**B6**	690-750 nm
B7	750-900 nm	**B8**	910-1050 nm
B9	1550-1750 nm	**B10**	2080-2350 nm
NDVI	$\frac{(B7-B5)}{(B7+B5)}$	**SAVI**	$\frac{(B7-B5)(1,5)}{(B7+B5+1,5)}$
NDII1	$\frac{(B7-B9)}{(B7+B9)}$	NDII2	$\frac{(B7-B10)}{(B7+B10)}$
HMAX	Height maximum	HMEAN	Height mean
HMED	Height median	HKURT	Height kurtosis
HRANGE	Height range	HSKEW	Height skewness
HSTD	Height standard deviation	HCV	Height coefficient of variation

Cuadro 6.1: Atributos y bandas extraídos de los datos LIDAR y de la imagen ATM para Alto Tajo (García et al., 2011). En negrita, atributos finales seleccionados.

los valores medios de cada variable para cada clase (ver Figura 6.1). El conjunto final seleccionado aparece destacado en negrita en el Cuadro 6.1.

6.3.3. Procesamiento de los datos de la desembocadura del Tinto y del Odiel

Como se hizo para el área del Alto Tajo, las ortoimágenes se ajustaron a a los datos LIDAR mediante el uso de puntos de control. En este caso, no hubo corrección radiométrica ya que no se disponía de información para convertir los valores numéricos de las imágenes digitales a la radiación captada por el sensor.

Teniendo en cuenta la densidad de puntos, así como las características de la zona de estudio, las ortoimágenes se remuestrearon a 5 m para que al menos 25 retornos fueran incluidos en cada píxel de la imagen final y poder derivar métricas de los datos LIDAR. En este caso, también se generó un MDE para rectificar las alturas de los retornos y se eliminaron los puntos escapados aplicando un filtro en el percentil 99.

Además, la intensidad media de los pulsos LIDAR se calculó para cada píxel debido a que los datos ópticos no incluyen información sobre la región del infrarrojo cercano. Se utilizó sólo la media para no introducir distorsión en la clasificación. Antes de obtener el valor de la media, se eliminó la dependencia del rango que se tiene en la intensidad LIDAR mediante la normalización de los valores a un rango estándar (Donoghue et al., 2007; García et al., 2010).

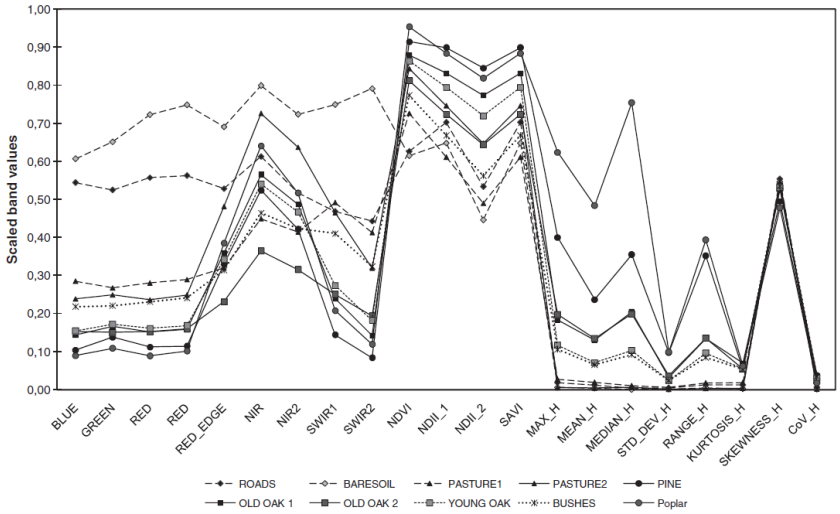


Figura 6.1: Análisis de separabilidad espectral de los datos de Alto Tajo (García et al., 2011).

Variable	Description	Variable	Description
RMEAN	Red Band mean	GMEAN	Green Band mean
BMEAN	Blue Band mean	IMEAN	Intensity mean
HMAX	Height maximum	HMEAN	Height mean
HMED	Height median	HKURT	Height kurtosis
HRANGE	Height range	HSKEW	Height skewness
HSTD	Height standard deviation	HCV	Height coefficient of variation

Cuadro 6.2: Atributos extraídos de los datos LIDAR y de la ortoimagen para Huelva.

6.3.4. Generación del modelo para Huelva

Siguiendo un proceso similar al de Alto Tajo, la altura máxima, la altura media, la altura mediana, su desviación estándar, el rango, la asimetría, la curtosis y el coeficiente de variación se extrajeron de los datos para cada píxel además de la intensidad media. Posteriormente, se calcularon la intensidad media de LIDAR y los valores medios de las bandas R, G y B de la ortoimagen. El conjunto final de atributos y bandas seleccionados puede verse en el Cuadro 6.2.

Como en el caso de Alto Tajo, se aplicaron filtros de eliminación de outliers y de escalado. Por último, al igual que en el caso del Alto Tajo las distintas variables extraídas de LIDAR se unieron con las bandas extraídas de las ortofotos en una sola imagen. En este caso, el análisis de separabilidad entre bandas no eliminó ninguna variable por lo que se trabajó con el conjunto completo.

Classifier	Huelva	Alto Tajo
ANN	$hl = 10; m = 0{,}2; lr = 0{,}3$	$hl = 13; m = 0{,}2; lr = 0{,}3$
SVM	$C = 90{,}5097; \gamma = 0{,}353553$	$C = 207{,}94; \gamma = 3{,}25$
k-NN	$k = 5$	$k = 3$
DT	None	None
MLC	None	None

Cuadro 6.3: Parámetros seleccionados para cada clasificador y área.

6.3.5. Clasificación de la imagen

Los experimentos se llevaron a cabo mediante la reutilización de WEKA (Hall et al., 2009). Concretamente, se utilizó J48 para la implementación de DT, SMO para SVM, IBK para k-NN, un perceptrón multicapa para ANN y un algoritmo ad-hoc basado en la aplicación del clasificador Naive Bayes de WEKA para MLC.

En SVM, k-NN y ANN, la selección de parámetros adecuados puede afectar significativamente a los resultados. En este caso, se utilizó un procedimiento de búsqueda controlada para seleccionar estos parámetros. Para el caso de SVM, se usó un kernel gaussiano. Dicho kernel tiene dos parámetros que determinan la exactitud de la clasificación, en concreto C y γ. El procedimiento de búsqueda descrito en el framework LIBSVM de Chang y Li (Chang and Lin, 2011), usa una validación cruzada con cinco carpetas. En esta búsqueda, se prueban pares de (C, γ) en los datos de entrenamiento y el par que obtiene la mejor precisión se selecciona. La búsqueda incluye dos pasos (Oommen et al., 2004). En primer lugar, se utiliza una secuencia de crecimiento exponencial ($C = 2^{-5}, 2^{-3}, \ldots, 2^{15}$ y $\gamma = 2^{-15}, 2^{-13}, \ldots, 2^{3}, 2^{5}$), seguido por una búsqueda más fina que aumenta ligeramente la precisión. Para el caso de los parámetros asociados al algoritmo k-NN y ANN se usaron procedimientos análogos. En el caso del algoritmo k-NN: $k = 1, \ldots, 10$; para ANN: número de capas ocultas (conjuntos adicionales de neuronas internas; $hl = 0, 1, \ldots,$ cantidad de atributos + número de etiquetas), momentum (factor corrección de backtracking, $m = 0{,}1; 0{,}2; \ldots; 1{,}0$) y tasas de aprendizaje (coeficientes para controlar los cambios en los pesos de las neuronas; $lr = 0{,}1; 0{,}2; \ldots; 1{,}0$). Los valores seleccionados pueden verse en el Cuadro 6.3.

El MLC utilizado es una aplicación ad-hoc basadas en Naive Bayes. MLC se basa en el cálculo del estimador de máxima verosimilitud (ver Ecuación 6.1).

$$Y_{mle} = \arg\max_{Y} \hat{\ell}(Y | X_1, \ldots, X_n) \tag{6.1}$$

Naive Bayes utiliza la Ecuación 6.2 para seleccionar la etiqueta a asignar, donde Y es el valor de la etiqueta y X es el conjunto de valores para cada atributo o banda.

$$Y = \arg\max_{Y} P(X|Y)P(Y) \tag{6.2}$$

Para calcular las probabilidades condicionales, Naive Bayes utiliza el estimador de máxima verosimilitud para una distribución gaussiana, pero su cálculo es ligeramente diferente, ya que lo simplifica aplicando la hipótesis de independencia (Ecuacion 6.3) para superar los problemas relacionados con la estimación de las probabilidades condicionadas.

$$p(Y, X_1, \ldots, X_n) \propto p(Y) \ \cdots \propto p(Y) \prod_{i=1}^{n} p(X_i|Y) \tag{6.3}$$

Además, Naive Bayes se basa en probabilidades a priori ($P(Y)$). Así pues, si queremos un MLC puro es necesario modificar los valores de $P(Y_i)$ tal que $P(Y_i) = P(Y_j) = 1{,}0 \ \forall i, j$. De esta manera, es posible obtener una aproximación del clasificador de máxima verosimilitud mediante la reutilización del código del clasificador Naive Bayes de WEKA.

6.4. Resultados

La comparación de los clasificadores se basa en dos estrategias bien conocidas: un hold-out (una técnica comúnmente utilizada en teledetección), y una 10-FCV. En el caso del hold-out, la precisión de la clasificación se evaluó mediante una matriz de confusión y el coeficiente kappa de Cohen (Congalton and Green, 2008), con alrededor del 25 % de la base de conocimiento como datos de test y el resto para entrenamiento.

6.4.1. Precisión de los modelos

Cada clasificador fue probado en los datos de referencia descrito en el apartado anterior. El Cuadro 6.4 muestra los resultados obtenidos para cada clasificador en la zona del Alto Tajo, y el Cuadro 6.5 muestra los resultados para el área de la desembocadura del Tinto y del Odiel.

El análisis de las matrices de confusión mostró que las SVMs dan los mejores resultados en promedio. Sin embargo, este hecho no es suficiente en el estudio de las técnicas de minería de datos en general, y del aprendizaje automático en particular, como hemos visto en el capítulo anterior. Así, pasamos a examinar los resultados de la 10-FCV y a realizar un análisis estadístico del rendimiento de cada clasificador en las dos áreas de estudio.

6.4.2. Significación estadística de resultados

Debido a que, de nuevo, sólo se tenían dos conjuntos de datos, se decidió realizar una división aleatoria estratificada en cada conjunto etiquetado por los expertos. En concreto, se crearon cinco divisiones aleatorias para cada zona. Después, se ejecutó una 10-FCV para cada división, y los 10 resultados parciales de guardaron, resultando en 50 medidas de precisión para cada clasificador y

Alto Tajo Classification Method										
	MLC		k-NN		DT		ANN		SVM	
label\ accuracy	User's	Producer's	User's	Producer's	User's	Producer's	User's	Producer's	User's	Producer's
non fuel	100.00	97.09	98.00	98.00	100.00	97.09	100.00	96.15	100.00	96.15
bushes	91.74	94.34	90.83	97.06	89.91	94.23	93.58	93.58	92.66	98.06
poplar	81.67	69.01	80.00	82.76	93.33	82.35	86.67	85.25	88.33	80.30
oak	89.63	89.63	88.15	92.25	91.85	94.66	94.81	93.43	94.07	94.07
pasture	77.14	84.38	89.52	85.45	89.52	93.07	89.52	94.95	89.52	94.95
pine	81.58	83.78	89.47	79.07	88.16	85.90	89.47	90.67	88.16	85.90
overall	87.69		89.91		92.14		92.99		92.65	
Kappa	0.85		0.88		0.9		0.91		0.91	

Cuadro 6.4: Resumen de las pruebas hold-out 75 %-25 % para el Alto Tajo.

Tinto and Odiel Classification Method										
	MLC		k-NN		DT		ANN		SVM	
label\ accuracy	User's	Producer's	User's	Producer's	User's	Producer's	User's	Producer's	User's	Producer's
water	100,00	92.33	97.18	95.34	98.81	98.23	97.77	95.37	100.00	95.87
marshland	77.69	58.85	86.88	75.57	85.56	77.80	87.40	76.73	86.61	95.87
roads	32.09	81.13	82.46	82.77	90.67	87.41	92.16	79.94	91.79	75.51
low veg.	84.74	88.46	84.74	96.99	84.74	97.58	77.89	96.73	80.53	83.11
mid veg.	62.50	51.28	62.50	72.73	46.09	73.75	60.16	59.69	64.84	96.23
high veg.	93.67	61.16	84.81	75.28	70.89	44.44	86.08	71.58	87.34	69.75
urban	72.34	96.68	81.21	80.63	78.72	79.86	74.82	93.36	79.79	73.40
landfills	33.93	38.78	21.43	70.59	37.50	61.76	19.64	52.38	17.86	95.34
overall	77.45		85.18		85.23		85.23		86.98	
Kappa	0.72		0.82		0.82		0.82		0.84	

Cuadro 6.5: Resumen de las pruebas hold-out 75 %-25 % para el Tinto y el Odiel.

Algorithm	Ranking
SVM	2.14
ANN	2.85
k-NN	3.06
DT	3.08
MLC	3.88

Cuadro 6.6: Rankings medios después de los tests 10-FCV.

	algorithm	$p - value$	Holm
4	MLC	0.0000	0.0125
3	DT	0.0000	0.0167
2	k-NN	0.0000	0.0250
1	ANN	0.0015	0.0500

Cuadro 6.7: P-values para el procedimiento de Holm.

área de estudio. Después de la obtención de estas medidas, ya es posible el establecimiento de un análisis estadístico de la precisión con los 100 (50×2 zonas) resultados para cada clasificador.

Hemos utilizado un 10-FCV para generar los resultados de la comparación por lo que el uso de una prueba paramétrica no está recomendado. De nuevo, usamos un proceso no paramétrico para realizar la comparación (García and Herrera, 2008) basándonos en el test de Friedman y el procedimiento post-hoc de Holm.

El primer paso del proceso no paramétrico es establecer un rango medio entre los algoritmos basados en la precisión alcanzada por los clasificadores en cada dataset. Así, se asignó un valor entre 1 y 5 a cada clasificador para todo los datasets de la prueba 10-FCV, siendo 1 el mejor y 5 el peor de los resultados obtenidos. A partir de estos datos, se pudo generar el ranking medio para los clasificadores. De acuerdo con los resultados de la 10-FCV, SVM obtiene la clasificación más alta, como muestra el Cuadro 6.6.

Recordemos que la hipótesis nula de la prueba de Friedman consiste en que los valores de los rankings de los clasificadores no son lo suficientemente diferentes del ranking medio r (en nuestro caso 3.0). Después de ejecutar la prueba, el p-value obtenido fue menor de $3{,}65E^{-11}$, de modo que la hipótesis nula fue rechazada. Después de demostrar que los clasificadores se comportan de manera diferente en general, se aplica el procedimiento post-hoc de Holm.

El procedimiento post-hoc de Holm detecta rigurosamente las diferencias significativas entre pares cuando se compara un algoritmo de control con el resto. Por lo tanto, para nuestro estudio, la hipótesis nula es que no existen diferencias significativas entre el clasificador SVM (nuestro método de control) y cada uno del resto de clasificadores. El p-value obtenido en cada comparación fue, en todos los casos, inferior al valor requerido por el test de Holm (columna Holm en el Cuadro 6.7) de tal manera que las hipótesis nulas fueron rechazadas.

Approach	Huelva	Alto Tajo
Fusion	77 %	93 %
Only Image	58 %	86 %
Only LIDAR	69 %	65 %
Only LIDAR (Chapter 5)	85 %	84 %

Cuadro 6.8: Estudio de la precisión del modelo por sensor y conjunto de atributos cuando se aplica el clasificador SVM.

Finalmente, dado que las diferencias entre los métodos fueron significativamente diferentes para $\alpha = 0{,}05$, el análisis de rankings concluyó que la precisión del método SVM fue significativamente mejor que la de sus competidores para los datos de ambas áreas de estudio.

6.4.3. Importancia relativa de cada sensor en la clasificación

Una cuestión importante no tratada hasta el momento es el grado de mejora obtenido cuando se aplica fusión respecto al uso de los sensores de manera aislada. Es de esperar que el uso conjunto de varios sensores mejore la clasificación individual, pero saber si dicha mejora es lo suficientemente alta para justificar su uso es un dato muy importante, dado que la fusión implica una inversión extra en el vuelo y en post-procesado de los datos que genere.

El Cuadro 6.8 recoge la experimentación realizada con los datos de este estudio. En ella, se muestran los resultados obtenidos para el mismo clasificador (SVM con los parámetros optimizados para cada zona) pero usando exclusivamente el conjunto de atributos generado por cada sensor/enfoque. En todos los casos, se respetaron las metodologías seguidas en este y en el anterior capítulo (según su caso). El porcentaje que se presenta es la precisión global aproximada del clasificador sobre la base de entrenamiento para una 10-FCV en el que todas las metodologías trabajaron con las mismas carpetas.

6.5. Discusión

En lo que respecta al rendimiento de los clasificadores, hay que decir que MLC obtuvo menor precisión e índice kappa en ambas áreas de estudio, mientras que SVM y ANN alcanzaron los valores más altos. Las diferencias en la precisión total fueron sólo del 5.3 % para el Alto Tajo y del 9.53 % para el Tinto y el Odiel mientras que las diferencias en el índice kappa fueron del 0,06 y 0,12 para el Alto Tajo y el Tinto y el Odiel, respectivamente. Las resultados obtenidos están de acuerdo con las conclusiones de Oommen et al. (2004) y Waske and Benediktsson (2007) en tanto las diferencias entre los métodos no paramétricos son escasas. En cualquier caso, el análisis estadístico posterior muestra que aunque estas diferencias son pequeñas, también son significativas. Por el contrario,

las diferencias entre el MLC y el resto de clasificadores reflejan que los supuestos en los que se basan los clasificadores paramétricos rara vez se cumplen cuando se aplican a datos remotos de múltiples fuentes.

En Alto Tajo, la clase que muestra la menor precisión fue álamo. Este hecho puede explciarse en parte por lo reducido del área de la plantación respecto al resto de clases. En minería de datos, este fenómeno se conoce como desbalanceo de clases (Japkowicz and Stephen, 2002) y se caracteriza por provocar un alto nivel de errores para las clases minoritarias. Para la zona de estudio del Tinto y el Odiel, la clase que muestra la menor precisión es vertedero, y el desbalanceo de clases pueden ser, de nuevo, la causa principal del error.

Las Figuras 6.2 y 6.3 muestran los resultados finales. Se puede observar cierto grado de confusión entre el agua y las áreas urbanas en el área del Tinto y el Odiel (Figura 6.3). Algunas áreas de escasa vegetación también fueron clasificadas como clase urbana. La presencia de objetos con una altura similar a algunos de los objetos en la zona de edificios y la escasa información espectral disponible (sólo RGB e intensidad) podrían explicar estos problemas de clasificación.

Las precisiones obtenidas en la clasificación para el Alto Tajo son ligeramente superiores a las obtenidas para la zona de Huelva. La causa de la mejora radica en la mayor información espectral de los datos ATM, mucho más rico que las ortoimágenes RGB para el sitio de estudio, y además, por el menor número de clases en la zona Alto Tajo, que disminuye la dificultad de la tarea de clasificación.

Los resultados de Huelva, dada la escasa información espectral, indican la clasificación se benefició claramente de la inclusión de la información proporcionada por el sensor LIDAR. Este hecho se ve subrayado si atendemos a los resultados del Cuadro 6.8. Un dato muy importante que se extrae de dicho cuadro es que los resultados del subconjunto de atributos LIDAR utilizado en este capítulo no mejoran los obtenidos si hubiéramos aplicado el conjunto de estadísticos del capítulo anterior (69 % frente al 85 % para el caso de Huelva, y 65 % frente al 84 % para el caso de Alto Tajo). Esto implica que LIDAR tiene más información oculta que los atributos más comunes en fusión no explotan al máximo. Es por ello, que en el próximo capítulo exploraremos vías para mejorar la fusión partiendo del conjunto de atributos visto en el anterior capítulo.

Más allá de la aportación de LIDAR según se use un subconjunto de atributos u otro, hay que subrayar que LIDAR aporta más información que las bandas RGB pero menos que ATM que lo supera en el caso de Alto Tajo en un 21 % (86 % vs. 65 %) y en un 2 % (86 % vs. 84 %) según se use la aproximación presentada en este capítulo o la usada en el anterior, respectivamente. Este dato es lógico si tenemos en cuenta que ATM trabaja sobre una cantidad de información espectral mucho mayor que LIDAR. En todo caso, la fusión de ambas tecnologías obtiene los mejores resultados para las dos áreas de estudio ya que, aunque el enfoque del capítulo anterior consiga mejores resultados para el área de Huelva con LIDAR exclusivamente, es lógico pensar que la mejora también se produciría si se tomase toda la información posible de LIDAR.

Finalmente, aunque el número de etiquetas es distinto (ver Subsección 4.2.1), se puede establecer una comparativa visual entre las Figuras 6.3 y 5.1 que mues-

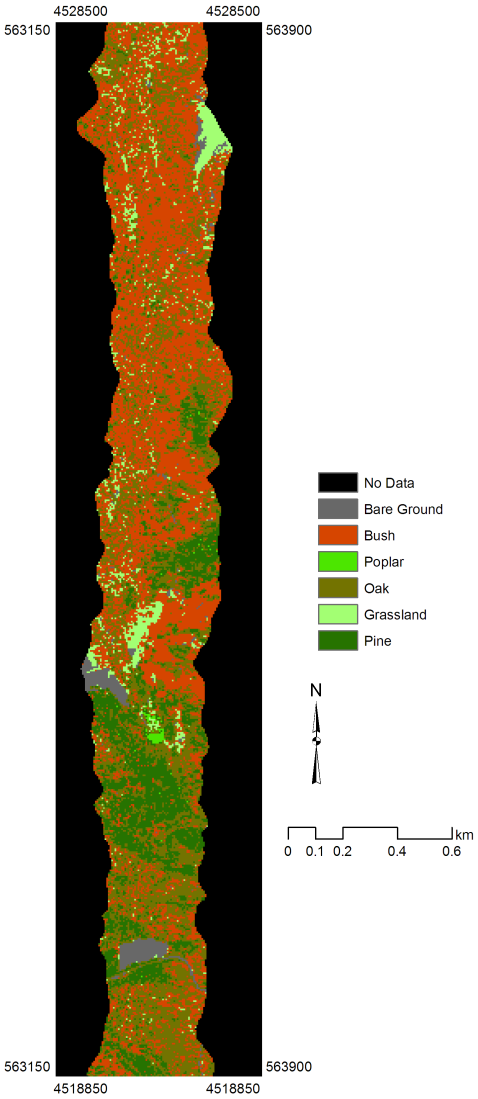


Figura 6.2: Resultados para SVM (mejor clasificador) sobre la zona de Alto Tajo.

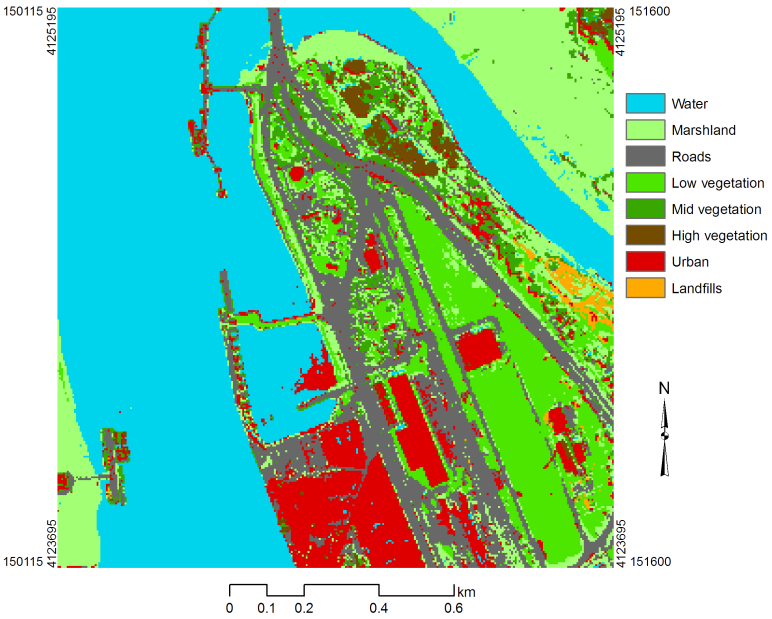


Figura 6.3: Resultados para SVM (mejor clasificador) sobre la zona de Huelva.

tran los resultados sobre los datos de Huelva de los mejores clasificadores en cada capítulo, esto es SVM y DT, espectivamente. Así, es posible observar que el resultado final es bastante similar con la ventaja de que la inclusión de la nueva etiqueta *vertedero* permite discernir mejor las diferencias entre estos y otras zonas urbanas. También desaparecen errores de clasificación entre zona urbana y zona de vegetación baja. Por contra, aparecen, como hemos dicho, clasificaciones erróneas en algunas zonas inundadas por la influencia del color del agua en determinadas áreas. Un dato muy importante es que aunque existe menor grado de ruido de «sal y pimienta» en la Figura 6.3, éste todavía enturbia la imagen final. En el siguiente capítulo, abordaremos el uso del contexto como herramienta para eliminar este tipo de problemas.

Capítulo 7

Uso del contexto sobre fusión de sensores

Todos estamos en el contexto y el contexto está en nosotros. Messi representa el paradigma sistémico y contextual.
Juan Manuel Lillo Díez.

7.1. Introducción

Durante los capítulos anteriores, se pudo observar que el uso de los datos LIDAR junto a la aplicación de técnicas de minería de datos puede mejorar en gran manera los resultados obtenidos en lo que respecta a la generación de mapas temáticos (por ejemplo, mapas de usos y coberturas vegetales). También vimos que, aunque los resultados obtenidos mediante fusión de LIDAR con otros sensores son de calidad, las técnicas de fusión clásicas no explotan completamente la potencia del sensor LIDAR, especialmente cuando las imágenes no contienen una cantidad de información espectral suficiente, como es el caso de la ortofotografía aérea básica.

Teniendo en cuenta el desarrollo de planes como el PNOA, en este capítulo nos centraremos en la fusión de ortofotografía básica (bandas RGB) y LIDAR. En este contexto, es necesario establecer una metodología que cubra dos aspectos básicos para la generación de mapas temáticos. Por un lado, cómo explotar la mayor cantidad posible de información, entendiendo información, en este caso, como la capacidad para discernir entre distintos tipos de coberturas o usos al mayor detalle posible. Así pues, es básico obtener un conjunto amplio de descriptores (sobre todo de LIDAR) para posteriormente filtrar aquellos que mejor se adapten a los requisitos de separabilidad de las áreas de estudio. Por otro lado, hay que prestar especial atención a los ruídos generados en el proceso de clasificación. Ya hemos visto el tipo de ruido clásico de «sal y pimienta» que se genera en la orientación a píxeles y a polígonos. Entre las posibles soluciones a este problema se encuentran adoptar una clasificación orientada a objetos o refinar

la clasificación mediante el contexto. En este capítulo, aplicaremos una solución basada en contexto para evitar los problemas del trabajo con segmentación ya comentados previamente.

Así pues, en este capítulo, exploramos el uso de clasificadores contextuales para mejorar los resultados obtenidos en fusión de datos con el objetivo de desarrollar mapas LULC de alta resolución a partir de ortofotografía RGB y datos LIDAR. En concreto, mostraremos un nuevo clasificador denominado SVMNNS que combina el uso de SVMs y de NNs para introducir la información contextual. Posteriormente, compararemos su rendimiento con el de una SVM clásica y con los resultados de una SVMMRF, combinación de SVM y regularización mediante campos aleatorios de Markov, que como clasificador contextual ha mostrado recientemente buenos resultados para la clasificación de otros sensores (hiperespectrales) de alta resolución (Tarabalka et al., 2010).

7.2. Trabajos relacionados

Las campos aleatorios de Markov (Markov Random Field, MRF) son una técnica estocástica usada de manera profusa en el mundo de la visión por computador (Geman and Geman, 1984). El objetivo principal de una MRF es calcular un estimador sencillo de la máxima probabilidad a posteriori (MAP). De manera informal, los MRF aplicados en la restauración de imágenes se caracterizan por las siguientes premisas:

- Cualquier cambio local en la imagen depende exclusivamente del conjunto de valores actuales de los píxeles y de los valores de sus vecinos. Este cambio es generado por muestreo de una distribución de probabilidad condicional local.

- Las distribuciones condicionales locales dependen de un parámetro de control global comunmente denominado «temperatura». A bajas temperaturas las distribuciones condicionales locales se concentran en los estados que incrementan una función objetivo, mientras que a altas temperaturas, la distribución de la secuencia de cambios es prácticamente uniforme.

- Las restauraciones de imágenes basadas en MRF evitan los máximos locales usando altas temperaturas al principio, donde muchos de los cambios estocásticos disminuyen el ajuste de la imagen a su entorno. A medida que el algoritmo avanza, la temperatura va disminuyendo gradualmente y el proceso se va comportando como un proceso de mejora iterativo convergiendo progresivamente a un estado final.

El uso de la regularización MRF es la principal vía de explotación del contexto en la bibliografía reciente en lo que a clasificación contextual se refiere, a pesar de que sus virtudes sobre datos provenientes de múltiples fuentes se conocen desde los años 90 (Solberg, 1999). De esta forma, podemos encontrar MRFs para clasificar en función de un resultado parcial previo que sólo abarque

los casos más sencillos (Luo and Mountrakis, 2011), o también múltiples MRFs que se combinan para realizar el proceso completo (Levada et al., 2010).

Uno de los ejemplos más recientes de un clasificador con post-procesamiento basado en MRF puede verse en Tarabalka et al. (2010). Los autores proponen una técnica llamada SVMMRF para una clasificación hiperespectral que consta de dos pasos. En el primer paso, se aplica una SVM para realizar una clasificación orientada a píxel sobre imágenes hiperespectrales. En el segundo paso, la información contextual-espacial se utiliza para refinar los resultados de la clasificación con una regularización basada en MRF. Los resultados muestran que SVMMRF supera a otras técnicas propuestas recientemente. En la misma línea, el trabajo de Fauvel et al. (2012) muestra un metodo novedoso que combina información contextual y espectral, y que posteriormente, se proporciona a una SVM para generar la clasificación final. En la comparativa que los autores presentan, se muestra que aunque el método presentado es competitivo, no se puede concluir que mejore los resultados de SVMMRF.

Para desarrollar los modelos de regularización MRF existen diversas metodologías. Una de las más extendidas es el método de modas condicionales iteradas (Iterated Conditional Modes, ICM) de Besag (1986). El método ICM es un método determinista que parte de una clasificación inicial, generalmente basada en el criterio de máxima verosimilitud. Dicha clasificación se va refinando en cada iteración buscando la etiqueta para cada píxel que maximice la probabilidad a posteriori según la Ecuación 7.1:

$$\hat{x}_i = \arg\max_{x_i} P(x_i|y, \hat{x}_{S-\{i\}}) \propto f(y_i|x_i)p_i(x_i|\hat{x}_{\delta_i}) \tag{7.1}$$

En dicha Ecuación, i representa un píxel cualquiera, S es el conjunto de píxeles de la escena y δ_i es el conjunto de píxeles vecinos de i. Además, y_i representa al conjunto de valores de los atributos calculados para el píxel i, $\hat{x}_{S-\{i\}}$ es el conjunto de etiquetas asignadadas a S excluyendo al píxel i, $f(y_i|x_i)$ es la función de densidad de la observación y_i dada la etiqueta x_i, y $p_i(x_i|\hat{x}_{\delta_i})$ es la probabilidad de que se dé la etiqueta x_i dada una estimación $\hat{x}$ calculada a partir del entorno δ_i del píxel i. Nótese que esta probabilidad depende de cada píxel, de ahí el subíndice.

Es importante destacar que a la hora de aplicar ICM, en lugar de maximizar la Ecuación 7.1 se suele minimizar una función de energia U. La función de energía vendrá dada por una parte, por la carga espectral del píxel y por otra, por el conjunto de vecinos, que aportarán la carga espacial de la clasificación como puede verse en la Ecuación 7.2. La espeficificación de cada una de las componentes dependerá del tipo de MRF aplicada. Así, por ejemplo, en el caso de una SVMMRF, la función U viene dada por la suma de las Ecuaciones 7.3 y 7.4 donde $k(x_i, x_j)$ es el operador de kronecker que devuelve un valor 1 si ambas etiquetas son iguales y 0 en caso contrario, y β es un parámetro que controla la influencia del espacio en la clasificación final de cada instancia.

$$U(x_i) = U(x_i)_{espectral} + U(x_i)_{espacial} \tag{7.2}$$

$$U(x_i)_{espectral} = -ln(P(y_i|x_i)) \tag{7.3}$$

$$U(x_i)_{espacial} = \sum_{j \in \delta_i} \beta(1 - k(x_i, x_j)) \tag{7.4}$$

Es importante destacar que, en la mayoría de los casos, las MRFs se basan en supuestos gaussianos (Gaussian Markov Random Field) para facilitar el cálculo de aspectos como las probabilidades condicionadas y que en la versión original de ICM, la carga espectral se calculaba según un MLC que también proporcionaba las estimaciones iniciales.

Por otra parte, hay que hacer una mención especial respecto al parámetro β. Dicho parámetro se suele establecer con un valor 1,5 (Besag, 1986) de manera general, aunque también se puede actualizar a medida que el algoritmo avanza controlando la influencia del espacio en la clasificación final (variable temperatura del MRF). El valor de β condiciona la ejecución de ICM. Así, si $\beta = 0$, se estaría usando directamente el clasificador original y si $\beta = \infty$, entonces se calcularía el valor de la nueva etiqueta como el resultado de una votación entre los vecinos y la carga espectral sólo sería tenida en cuenta en caso de empate.

7.3. Metodología

7.3.1. Procesamiento de los datos

La Figura 7.1 muestra cada uno de los pasos del método que se presenta como principal aportación en este capítulo y que se ha denominado SVMNNS. El primer paso del método SVMNNS es el procesamiento de los datos de entrada (en nuestro caso, LIDAR e imágenes RGB). Así, las alturas de los datos LIDAR se normalizan a través de un MDE generado de los datos LIDAR correspondientes mediante filtros morfológicos (Gonçalves-Seco et al., 2006). También es importante normalizar las intensidades LIDAR para reducir su dependencia del rango de alturas (Hofle and Pfeifer, 2007) ya que no vamos a tener el soporte infrarrojo que poseen otras fuentes de datos como las imágenes multiespectrales.

Para realizar el proceso de clasificación, primero se debe fijar un tamaño de parcela que contenga el suficiente número de retornos LIDAR para poder generar información estadística. En este capítulo, usaremos los datasets de Huelva y Trabada (Subsecciones 4.2.2 y 4.3.1) para evaluar la calidad del algoritmo SVMNNS. El tamaño de parcela fue de 3 m^2 para Huelva y de 1 m^2 para Trabada, respectivamente. De esta forma, dividimos las áreas de estudio en parcelas de tamaño fijo. Para cada parcela, como hicimos en los anteriores capítulos extrajimos un conjunto de medidas estadísticas a partir de las bandas disponibles.

En lo que respecta a las estadísticas usadas, hay que tener en cuenta los resultados de los dos anteriores capítulos que mostraban la necesidad de extraer el máximo de información LIDAR para superar las limitaciones espectrales que las imágenes RGB tenían. Así pues, se calcularon 71 atributos distintos extraídos

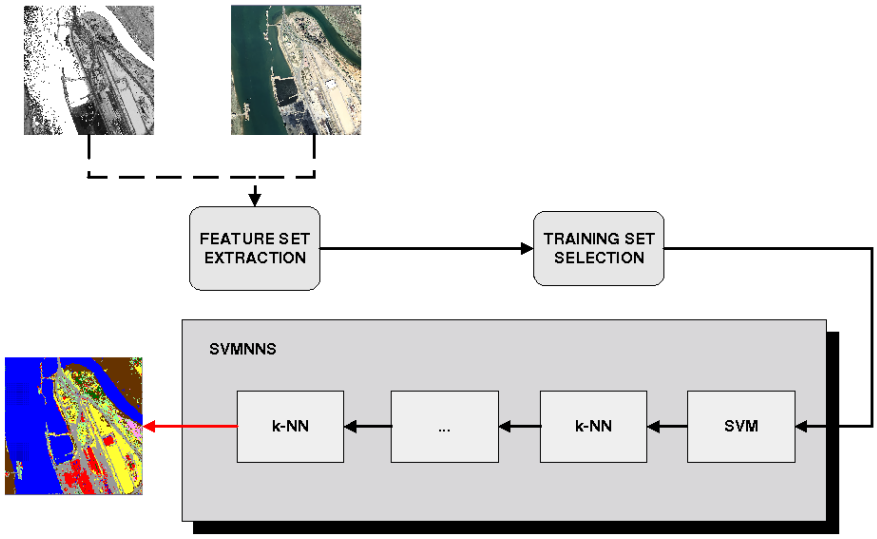


Figura 7.1: Método SVMNNS.

principalmente de la bibliografía relacionada a los anteriores capítulos (García-Gutiérrez et al., 2012b; García et al., 2011) para cada parcela. En el Cuadro 7.1, se muestran 54 atributos, esto es, 9 medidas para cada una de las 6 bandas disponibles: RGB de las imágenes, intensidad LIDAR, altura normalizada LIDAR y SNDVI (Simulated Normalized Difference Vegetation Index). La banda SNDVI simula la banda NDVI (Normalized Difference Vegetation Index), pero sustituyendo el valor infrarrojo, por un valor de intensidad de retorno LIDAR. Además, otros 17 atributos fueron definidos a partir de la distribución de los impactos LIDAR (Cuadro 7.2).

Symbol	Description	Symbol	Description
MAX	Maximum	MIN	Minimum
RANG	Range	STD	Standard Deviation
VAR	Variance	MEAN	Mean
KURT	Kurtosis	CV	Coefficient
SKEW	Skewness		of Variation

Cuadro 7.1: Conjunto de atributos calculados para cada banda: R (rojo), G (verde), B (azul), H (altura normalizada), I (intensidad LIDAR) y SNDVI (NDVI simulado).

Finalmente, las medidas se unen para dar lugar a una instancia para cada parcela con la información fusionada. El conjunto de todas las parcelas, forma el conjunto de datos con el que se trabajará para generar la clasificación final.

Symbol	Description
PCTR1	Percentage of first return
NOTFIRST	Number of non-first impacts
PCTR2	Percentage of second return
NEMP	Number of empty neighbours
PCTR3	Percentage of third or later return
TOTALR	Total number of impacts
PCTR21	PCTR2 over PCTR1
PEC	Penetration coefficient
PCTR31	PCTR3 over PCTR1
HPD	Height difference between first and last impacts in the pixel
PCTR32	PCTR3 over PCTR2
PCTN1	Percentage of single impact
EXTRASLP	Slope among every neighbor
PCTN2	Percentage of double impact
INTRASLP	Slope in the pixel
PCTN3	Percentage of triple or more impact
CRR	Canopy Relief Ratio

Cuadro 7.2: Conjunto de atributos calculados a partir de la distribución LIDAR.

7.3.2. Selección de datos de entrenamiento

Con el conjunto de datos generado, el siguiente paso es la extracción de la base de entrenamiento. Para ello, se clasificó un conjunto de parcelas a través de inspección visual de las imágenes aéreas y de otras fuentes de datos (como mapas LULC anteriores) asignándoles una etiqueta de acuerdo con su tipo de cobertura o de uso. Se etiquetaron 618 parcelas para el área de Huelva y 304, para Trabada. Estos datos pasaron a formar la base de entrenamiento para cada zona. En ambos casos, la cantidad de instancias de entrenamiento fue de aproximadamente el 1 % del total.

Antes de abordar la clasificación, se debe aplicar un pre-procesado a los datos (Zhang et al., 2003). De nuevo, se utilizaron dos tipos de filtros. En primer lugar, los valores missing se sustituyeron por el valor medio del correspondiente atributo. Posteriormente, se aplicó un método de selección de atributos (en nuestro caso, CFS). Los dos filtros se ejecutan utilizando el framework Weka (Hall et al., 2009) con los parámetros por defecto. Una vez que se obtuvo la base de datos filtrada, la siguiente fase fue la ejecución del algoritmo SVMNNS.

El conjunto de atributos seleccionado para las dos áreas de estudio seleccionadas puede verse en el Cuadro 7.3 y la importancia del hecho de que 9 de los atributos coincidan se discutirá más adelante en este capítulo. Aparte de este hecho importante, si analizamos la distribución de atributos seleccionados podemos observar que LIDAR aparece como la principal fuente de información definiendo o influyendo en 10/17 atributos en el caso de Huelva, y 15/20 en

Huelva	Trabada
SNDVIMIN	**SNDVIMIN**
SNDVIMAX	**SNDVIMAX**
SNDVIMEAN	**SNDVIMEAN**
IMEAN	**IMEAN**
IMIN	**IMIN**
HMAX	**HMAX**
IKURT	**IKURT**
RVAR	**RVAR**
GMEAN	**GMEAN**
INTRASLP	SNDVISTD
ISKEW	SNDVIRANGE
HCV	IMAX
RMIN	HMEAN
GKURT	HMIN
GCV	HRANGE
BMEAN	EXTRASLP
BVAR	HPD
	RKURT
	RCV
	BCV

Cuadro 7.3: Conjunto de atributos seleccionados para las áreas de Trabada y Huelva a partir de los 71 originales.

el caso de Trabada. Por otro lado, sólo 1 atributo fue seleccionado del Cuadro 7.2 en el caso de Huelva (INTRASLP) y 2 en el caso de Trabada (EXTRASLP, HPD), por lo que se puede concluir que la mayor parte de la información LIDAR viene definida por las estadísticas básicas descritas en el Cuadro 7.1.

7.3.3. SVMNNS

SVMNNS es el resultado de combinar una SVM y un refinamiento basado en k-NNs de forma secuencial. En nuestro caso, cada clasificador posterior utiliza la información contextual para refinar la clasificación del paso previo. Es importante destacar que SVMNNS usa como información contextual no sólo las etiquetas de sus vecinos sino su carga espectral. Ésta es la principal diferencia respecto a SVMMRF que por contra, tiene en cuenta exclusivamente las etiquetas de los vecinos y deja la carga espectral en manos del clasificador original.

Así pues, el método SVMNNS genera a partir de los datos de entrenamiento una SVM inicial usando el algoritmo de optimización SMO (Keerthi et al., 2001). La SVM realiza una clasificación de primer nivel de la escena y asigna a cada parcela una etiqueta.

El siguiente paso es la aplicación de un k-NN para reclasificar las parcelas. Esto es, el k-NN se entrena con los vecinos 8-adyacentes de cada píxel y reclasifica la instancia del píxel actual de acuerdo con la información espectral que contienen sus vecinos. Esta etapa de post-procesamiento es equivalente a la aplicación de un k-NN en stacking (Wolpert, 1992) respecto al anterior clasificador, en el que la posición espacial tuviera un peso mucho mayor que el resto de características.

Es importante destacar que el valor de k es un parámetro que el usuario tiene que configurar al inicio de la ejecución. Para las zonas que se utilizaron en este estudio, se demostró mediante un análisis experimental que los mejores resultados se obtuvieron con un valor de $k = 3$, es decir, cada NN se basa en los 3 vecinos más cercanos de los 8 elementos adyacentes a cada parcela. En el caso extremo de que $k = \#vecinos$, el clasificador se comportaría como un post-procesamiento en el que cada instancias se clasificaría por votación.

El número de refinamientos mediante k-NN, que puede ser visto como el número de iteraciones del método, es también un parámetro configurable n. El valor de n es también definido por el usuario al inicio de la ejecución del método SVMNNS. Con respecto a las áreas de este estudio y para establecer una comparación justa respecto a SVMMRF se estableció un número de iteraciones $n = 6$ que es el número en el que el algoritmo ICM suele estabilizarse (Besag, 1986).

Finalmente, tras el último refinamiento, el algoritmo genera un mapa temático en el que se asocia a cada píxel la etiqueta de la parcela que le corresponde geográficamente.

7.4. Resultados

Para confirmar la calidad del método SVMNNS, se estableció una comparativa con dos otros clasificadores: SVM y SVMMRF. El método SVMMRF y en particular el método seleccionado para implementarlo (ICM) es sensible a la clasificación inicial (Cortijo and Blanca, 1998; Cao et al., 2011). Por esta razón y para hacer una comparación justa de los resultados, los algoritmos contextuales (SVMNNS y SVMMRF) utilizaron la clasificación inicial obtenida por SVM. Dicha SVM fue generada mediante el algoritmo SMO optimizando sus parámetros C (índice de complejidad) y γ (parámetro básico para el kernel de base radial) mediante computación evolutiva (Huang and Wang, 2006). En este caso, SMO utilizó los valores $C = 7{,}93$ y $\gamma = 4{,}15$ para los datos de Trabada, y $C = 8{,}12$ y $\gamma = 6{,}17$ para Huelva. Además, ambos clasificadores contextuales usaron el mismo número de iteraciones de refinamiento en los resultados ($n = 6$) que es el valor al que ICM alcanza la convergencia de manera general (Cortijo and Blanca, 1998). Además, se pudo comprobar empíricamente que en ambos casos el clasificador SVMNNS obtuvo los mejores resultados para $k = 7$ por lo que dicho valor fue seleccionado para la comparativa.

En este capítulo, la comparación se basó exclusivamente en un hold-out. Esto se debe a que la introducción del contexto no permite dividir las bases de datos al estilo de lo realizado en los dos capítulos anteriores ya que, de hacerse, la vecindad podría escogerse fuera del contexto real de cada píxel e introducir un importante grado de ruído en los resultados finales.

Para realizar la prueba hold-out, se eligieron y etiquetaron 7501 instancias del área de estudio de Huelva y 2320 del área de estudio Trabada. Se utilizó aproximadamente el 10 % del número total de casos de cada conjunto de datos para entrenar los diferentes clasificadores. Los Cuadros 7.4 y 7.6 muestran los resultados obtenido por SVM, SVMMRF con ICM, y SVMNNS en Huelva y Trabada, respectivamente. Los cuadros muestran el número de casos de prueba para cada clase, los valores de los errores de comisión y omisión, la precisión global, y el índice Kappa (KIA). En los Cuadros 7.5 y 7.7, se muestran los valores de precisión parcial y el valor de *F-measure* por etiqueta. La *F-measure* es una medida estadística que evalúa la calidad de una clasificación mediante la ponderación de los errores de tipo I (errores de usuario) y tipo II (errores de productor).

Como hemos comentado antes, debido a que el contexto no puede ser definido propiamente en subconjuntos aleatorios, no podemos aplicar el mismo estudio de significación que en anteriores capítulos. Así pues, en este caso, nos centramos en establecer la significación de los resultados sobre los dos conjuntos de test exclusivamente. Para ello, en lugar de estudiar los resultados globales, iteramos sobre los valores asociados a cada instancia de test y comparamos las etiquetas asignadas por cada clasificador con la etiqueta verdad de referencia. Para cada caso, asignamos un valor 1 si el clasificador acierta y un valor 0 si falla. De esta forma, a partir de las clasificaciones obtenemos una distribución binaria para cada clasificador. Posteriormente, es posible aplicar una prueba Q de Cochran (Conover, 1999) para determinar si las diferencias en las distribuciones dicotómicas recogidas son estadísticamente significativas y un análisis

		SVM		SVMMRF		SVMNNS	
Class	#	Comm. error	Omiss. error	Comm. error	Omiss. error	Comm. error	Omiss. error
Water	2151	3,00	4,10	0,00	3,50	0,00	0,30
Marsh	1266	14,60	27,50	3,00	25,50	3,50	18,40
Roads	1083	23,30	10,70	6,40	4,00	4,50	4,30
Low Veg.	686	24,10	18,10	20,80	9,30	14,40	11,10
Mid. Veg.	464	49,80	48,50	29,50	13,50	30,20	11,20
High Veg.	329	37,10	42,50	19,50	21,40	4,30	12,50
Buildings	1314	20,20	22,90	11,70	4,80	4,10	3,40
Landfills	209	66,50	27,10	81,80	0,00	63,60	0,00
KIA		0,77		0,87		0,91	
Accuracy		81,02		89,66		92,90	

Cuadro 7.4: Resultados porcentuales del hold-out para la zona de Huelva.

	SVM		SVMMRF		SVMNNS	
Class	Partial Accur.	F-Mea.	Partial Accur.	F-Mea.	Partial Accur.	F-Mea.
Water	0,97	0,964	1	0,982	1	0,999
Marsh	0,854	0,784	0,97	0,843	0,965	0,884
Roads	0,767	0,825	0,936	0,948	0,955	0,956
Low. Veg.	0,759	0,788	0,792	0,845	0,856	0,872
Middle Veg.	0,502	0,509	0,705	0,777	0,698	0,782
High Veg.	0,629	0,601	0,805	0,796	0,957	0,914
Buildings	0,798	0,784	0,883	0,916	0,959	0,962
Landfills	0,335	0,459	0,182	0,308	0,364	0,533
Minimun	0,335	0,459	0,182	0,308	0,364	0,533
Mean	0,661	0,686	0,717	0,747	0,791	0,826

Cuadro 7.5: Precisiones parciales por clase para el área de Huelva.

		SVM		SVMMRF		SVMNNS	
Class	#	Comm. error	Omiss. error	Comm. error	Omiss. error	Comm. error	Omiss. error
Roads	640	38,00	23,40	35,50	17,90	18,00	12,20
Low. Veg.	549	12,60	32,40	4,20	31,00	4,60	18,00
High Veg.	765	14,10	21,10	5,50	16,70	7,30	10,70
Buildings	366	40,20	15,40	49,50	1,10	21,90	1,00
KIA		0,67		0,72		0,84	
Accuracy		75,56		79,61		88,10	

Cuadro 7.6: Resultados porcentuales del hold-out para la zona de Trabada.

	SVM		SVMMRF		SVMNNS	
Class	Partial Accur.	F-Mea.	Partial Accur.	F-Mea.	Partial Accur.	F-Mea.
Roads	0,62	0,686	0,645	0,723	0,82	0,848
Low. Veg.	0,874	0,763	0,958	0,802	0,954	0,882
High Veg.	0,859	0,822	0,945	0,885	0,927	0,91
Buildings	0,598	0,701	0,505	0,669	0,781	0,873
Minimun	0,598	0,686	0,505	0,669	0,781	0,848
Mean	0,738	0,743	0,763	0,770	0,870	0,878

Cuadro 7.7: Precisiones parciales por clase para el área de Trabada.

	algorithm	χ^2	p-value	Holm's
2	SVMMRF	300.62	$1{,}32^{-10}$	0.025
1	SVM	908.40	$3{,}17^{-10}$	0.05

Cuadro 7.8: Holm's adjusted p-values for the significance McNemar's tests.

post-hoc basado en su homólogo para comparar entre pares, el test de McNemar (McNemar, 1947), una prueba no paramétrica que tradicionalmente se usa en teledetección (Tarabalka et al., 2010) para determinar si dos matrices de confusión son estadísticamente diferentes.

La prueba Q de Cochran es una generalización para más de dos conjuntos de datos del test de McNemar. El estadístico de Cochran se calcula según la Ecuación 7.5 donde k es el número de clasificadores, b es el número de instancias de test, $X_{i.}$ es la suma de los valores asociados a cada clasificador para una instancia de test en concreto, $X_{.j}$ es la suma de los valores para un clasificador en concreto y N es la suma total de valores de los k clasificadores. El estadístico de Cochran sigue una distribución χ^2 con $k-1$ grados de libertad.

$$T = (k)(k-1)\frac{\sum_{j=1}^{k}(X_{.j} - \frac{N}{k})^2}{\sum_{i=1}^{b} X_{i.}(k - X_{i.})} \tag{7.5}$$

En nuestro caso, la prueba Q de Cochran demostró que existen diferencias significativas entre los resultados de los tres métodos estudiados (p- value$<$ 10^{-9}), $\alpha = 0{,}05$). El posterior análisis post-hoc por pares mediante pruebas de McNemar corregidas con el procedimiento de Holm reveló que SVMNNS obtiene precisiones significativamente diferentes respecto a sus competidores, como puede verse en el Cuadro 7.8.

Teniendo en cuenta que en ambos casos, SVMNNS obtiene mejores resultados y que dichos resultados son estadísticamente significativos, podemos concluir que para ambas áreas SVMNNS tiene un comportamiento mejor que el resto de competidores.

La Figura 7.2 y la Figura 7.3 muestran las imágenes aérea, la imagen de intensidad LIDAR, las áreas de entrenamiento y test, y los mapas temáticos

resultantes para Huelva y Trabada respectivamente.

7.5. Discusión

Uno de los datos más llamativos de este capítulo es el hecho de que 9 de los atributos seleccionados para cada área coincidan. De hecho, es interesante ver que dos de ellos, IMEAN y HMAX, forman parte de los conjuntos seleccionados tanto en el Capítulo 5 como en el Capítulo 6. Lo que hace pensar que ambos atributos son clave a la hora de valorar la cantidad de información aportada por LIDAR al sistema.

Otro hallazgo muy importante mostrado en este capítulo es que es posible desarrollar índices de vegetación simulados aunque no se posea la información exacta de una banda en cuestión. Así, el conjunto de estadísticos relativos a SNDVI tiene una gran importancia en la clasificación final ya que se han seleccionado 3 (5 en el caso de Trabada) de los 7 estadísticos generados relativos a ese índice.

Para analizar los datos relativos a la precisión de los clasificadores y teniendo en cuenta que los datos de estudio son muy diferentes, es interesante establecer el análisis por separado para cada uno de los resultados de manera que, posteriormente, se puedan obtener conclusiones generales.

Así pues, centrándonos en los resultados de Huelva, encontramos que el nivel de precisión de la SVM es relativamente alto, por encima del 81 % (Cuadro 7.4). A pesar de ello, hay que destacar que existe una enorme diferencia entre la SVM y los resultados para los algoritmos contextuales (diferencia de más de 8 puntos). Este hecho subraya la importancia de utilizar algoritmos contextuales cuando se trabaja con orientaciones a píxel o parcela en datos de alta resolución.

Otro dato importante es que SVMNNS obtiene los mejores resultados para la precisión global y local a excepción de 2 etiquetas (marismas y vegetación media). Sin embargo, si analizamos los resultados de la medida F para tener en cuenta el error de omisión, vemos que comparando ambos clasificadores contextuales, SVMMRF obtiene los mejores resultados en todas las clases (véase el Cuadro 7.5). Así pues, se puede concluir que SVMNNS se comporta mejor para el conjunto de prueba en el área de Huelva ya que de manera global, bate a SVM y SVMMRF, y además, sus precisiones locales en el peor de los casos son similares a las obtenidas por sus competidores cuando no las supera ampliamente.

Si comparamos el resultado obtenido en este capítulo con el obtenido en el anterior, nos damos cuenta de que el resultado del procesado contextual ha generado una mejor clasificación final (precisión del 90 % vs. 86 % del capítulo anterior) a pesar de que la clasificación en este capítulo se ha llevado a cabo a una resolución superior. De lo anterior y en consonancia con lo demostrado por otros autores (Tarabalka et al., 2010; Fauvel et al., 2012), se debe aplicar un procedimiento contextual siempre que se aplique una clasificación por píxel o parcela para resolver los problemas asociados a estos paradigmas, ya comentados en capítulos anteriores.

Desde el punto de vista visual (Figura 7.2), se puede confirmar que el tratamiento contextual mejora los resultados generales de la SVM. La sensación cuando se compara un mapa con tratamiento contextual con uno sin dicho tratamiento es que el mapa es más «limpio». Esto es debido a que los tratamiento contextuales tienden a homogeneizar el resultado final. De cualquier manera, el análisis de los resultados debe hacerse principalmente de manera cuantitativa a partir de los resultados de test y como hemos visto, dichos datos confirman la sensación visual.

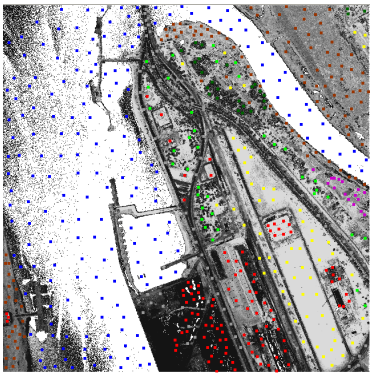
La zona de Trabada sólo tenía cuatro clases, y aunque se podría suponer que los resultados serían mejores, las precisiones son del mismo orden que en Huelva o menores. En concreto, SVM no alcanza el 76 % (Cuadro 7.6). Este hecho se puede atribuir a un nivel más bajo de separabilidad entre clases a partir de la información espectral disponible aunque observando los resultados obtenidos por SVMNNS, es más lógico pensar que también existe un mayor número de instancias atípicas que empobrecen la clasificación original. Es importante destacar que el uso del contexto sigue mejorando notablemente los resultados de la SVM aunque principalmente en el caso de SVMNNS (79 % para SVMMRF y 88 % para SVMNNS) que de nuevo obtiene el mejor resultado. La diferencia con respecto a SVMMRF es de casi 10 puntos. En este caso, al analizar la precisión por clase (Cuadro 7.7), se puede observar que SVMMRF obtiene bajas precisiones (medida F inferior al 85 %) para todas las etiquetas excepto vegetación alta. Por contra, SVMNNS matiene una precisión alta prácticamente para todas las clases.

En lo que refiere al aspecto visual (Figura 7.3), en esta ocasión puede verse perfectamente los problemas de SVMMRF comparándolo con SVMNNS. De nuevo, el tratamiento contextual mejora la clasificación inicial, pero en este caso, SVMMRF introduce zonas de eucaliptos en medio del área urbana, rodeando a los principales edificios debido a la excesiva carga espectral que introduce en la asignación de etiquetas. Por contra, SVMNNS compacta la zona de edificios eliminando los anteriores problemas.

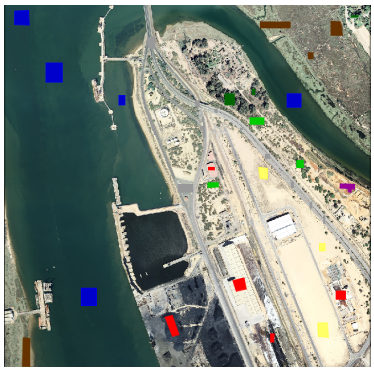
Valorando los resultados de manera conjunta, se debe concluir que, en primer lugar, se confirma la gran sinergia entre LIDAR y la ortofotografía área para separar las clases de ambas áreas de estudio por lo que su uso con datos del PNOA parece perfectamente viable. Es importante tener en cuenta los altos niveles de precisión que se han alcanzado (>85 %) para SVMNNS en ambas bases de datos. En segundo lugar, hay que destacar que el método propuesto, SVMNNS, mejora tanto la SVM clásica como SVMMRF no sólo globalmente, sino también para cada etiqueta de manera general, de acuerdo con las precisiones parciales y la medida F. Esta mejora con respecto a SVMMRF fue mayor para los datos de Trabada que para los datos de Huelva. Es decir, se obtuvo un mejor rendimiento con datos de mayor resolución (1m en Trabada vs. 3m en Huelva). La estrategia de la combinación de clasificadores SVM y NN resultó ser más apropiada en este contexto que la regularización MRF que, para conjuntos de datos con resoluciones bajas, funciona mejor pero que cuando se enfrenta a datos más complejos tiende a eliminar los conjuntos de etiquetas aislados en beneficio de las etiquetas mayoritarias en el área. Este problema no se tiene en SVMNNS

pero sin embargo, al no tener en cuenta la probabilidad original generada por la SVM, la clasificación pasa a depender demasiado de la relación espectral entre vecinos. Así, dos valores erroneos con respuestas espectrales similares al píxel actual, pueden hacer que la etiqueta actual bien asignada, se cambie e introduzca un error en la clasificación final. Este razonamiento explicaría el por qué a mayores valores de k se obtuvo un rendimiento visual mejor. En cualquier caso, este problema deberá ser tratado en el futuro trabajo post-doctoral.

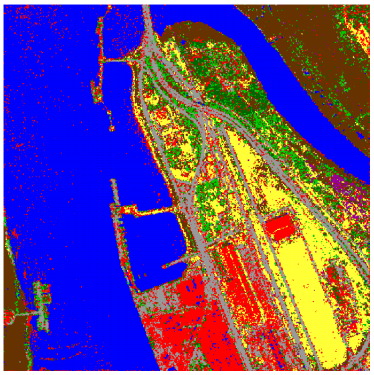
Finalmente, un dato importante no tratado es el rendimiento sobre datos no balanceados (Chawla et al., 2004). En lo que respecta a las dos bases de datos tratadas, sólo Huelva tiene un nivel de desbalanceo grave. Así, la clase «vertederos» tiene muy poca presencia tanto en la base de entrenamiento como en el conjunto global. Los niveles de precisión para esa clase son claramente insuficientes aunque el efecto perjudicial del desbalanceo parece ser menor en el caso de SVMNNS. El desbalanceo en los datos es una situación típica en teledetección (Guerrero-Curieses et al., 2004) y también deberá ser abordadas en futuros trabajos.



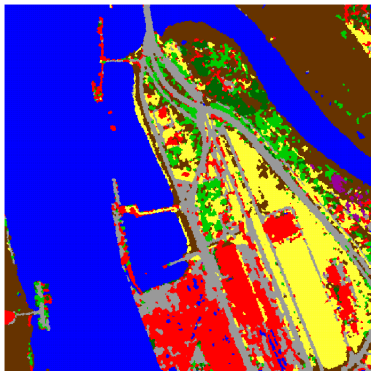
(a) Imagen de intensidad LIDAR en Huelva y puntos de entrenamiento



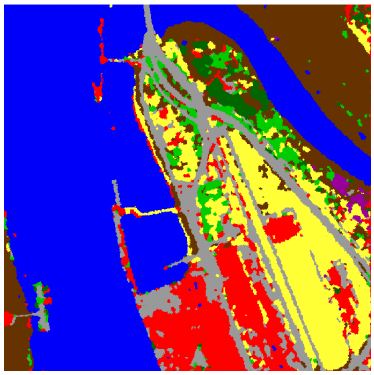
(b) Imágen aérea de Huelva y áreas de test



(c) Mapa temático SVM

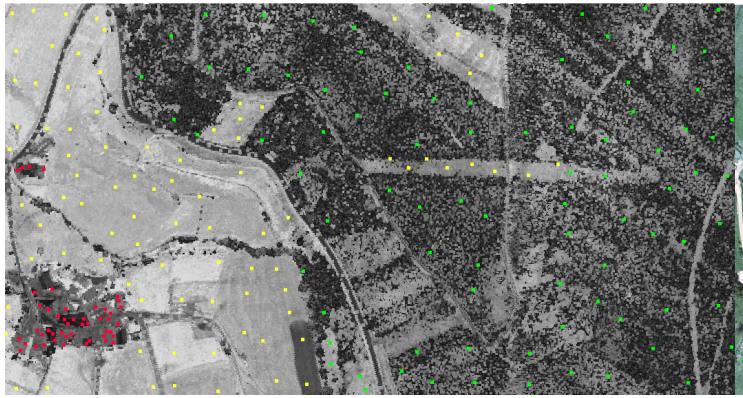


(d) Mapa temático SVMMRF

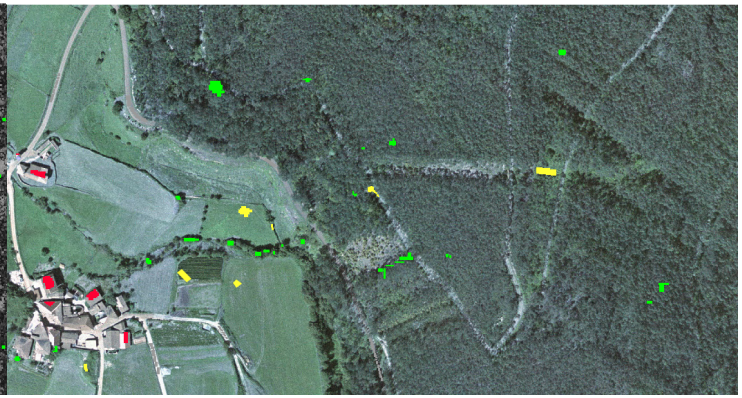


(e) Mapa temático SVMNNS

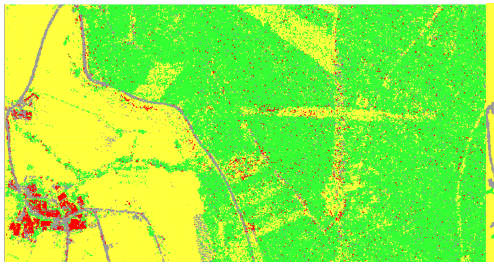
Figura 7.2: Clasificación final obtenida en Huelva. Agua en color azul, pantanos en marrón, carreteras y vías férreas en gris, vegetación baja y suelo desnudo en amarillo, vegetación media en color verde claro, eucaliptos en verde oscuro, edificios en rojo y vertederos en color morado.



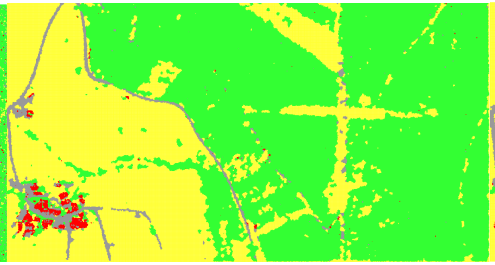
(a) Imagen de intensidad LIDAR en Trabada y puntos de entrenamiento



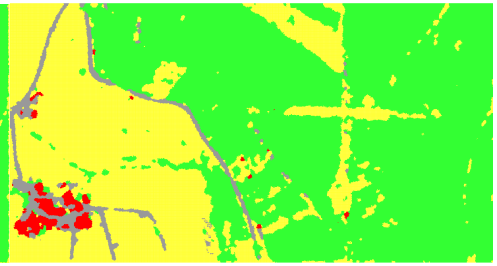
(b) Imágen aérea de Trabada y áreas de test



(c) Mapa temático SVM



(d) Mapa temático SVMMRF



(e) Mapa temático SVMNNS

Figura 7.3: Clasificación final obtenida en Trabada. Carreteras en gris, vegetación baja y suelo desnudo en amarillo, eucaliptos en verde y edificios en rojo.

Capítulo 8

Estimación de variables forestales

Dios tomó al hombre y lo puso en el jardín del Edén
para que lo cultivara y lo cuidara.
Génesis 2:15.

8.1. Introducción

Hasta el momento, nos hemos centrado en LIDAR como sensor para el desarrollo de mapas temáticos en lo que sería una aplicación clásica de teledetección. Así, hemos visto distintas formas de generar mapas de manera automática mediante técnicas de aprendizaje supervisado y posteriormente, nos hemos centrado en cómo introducir información del contexto espacial en la propia clasificación.

En este capítulo, nos centraremos en la que puede ser la aplicación más importante del sensor LIDAR junto al desarrollo de MDTs: la estimación de variables forestales y más concretamente de variables relacionadas con la biomasa forestal. En este contexto, es posible encontrar una bibliografía extensa en la que LIDAR aparece como un factor clave, debido principalmente a su capacidad para registrar la información de las alturas. Sin embargo, existen pocas aportaciones que se centren en la mejora del proceso de extracción del conocimiento a partir de la nube de puntos. En este capítulo concretamente, exploraremos la aplicación de técnicas de minería de datos y de *soft computing* para mejorar la estimación de biomasa basada en el sensor LIDAR.

La principal motivación para estimar biomasa viene de la necesidad de garantizar la sostenibilidad de los bosques. La Unión Europea ha desarrollado en los últimos tiempos políticas activas orientadas a la gestión, control y protección de los recursos naturales (Comisión al Consejo y al Parlamento Europeo, 2001). En el caso de los bosques, es de vital importancia tener en cuenta sus función tanto económica como ecológica, por lo que se hace necesario cuantificar los re-

cursos existentes para realizar una planificación estratégica, táctica y operativa en lo que respecta a los distintos tratamientos silvícolas y forestales.

La biomasa forestal es un indicador que proporciona una estimación del material celulósico como fuente potencial de energía renovable (Popescu, 2007), da una medida de la producción forestal en general, y también, da información del secuestro de carbono en los árboles. El secuestro de carbono es la capacidad de un árbol para almacenar carbono en su estructura. La capacidad de los bosques para secuestrar carbono es un hecho que interesa en extremo a los investigadores, debido a la posible relación entre la acumulación de CO_2 en la atmosfera y el cambio climático (Torres et al., 2010).

La mayoría de los sistemas de gestión de áreas forestales se basan en el uso de variables biofísicas medidas en campo (Naesset et al., 2004). Sin embargo, estas variables son muy caras de extraer debido principalmente al alto coste en tiempo y recursos que conlleva (Hall et al., 2005). Además, existen limitaciones para el conocimiento obtenido por esta vía. Así, por ejemplo, pueden existir zonas con difícil acceso que limiten el establecimiento de suficientes parcelas de muestra para capturar la variabilidad real (Lovell et al., 2005). También hay que tener en cuenta que la estimación clásica de biomasa a menudo implica un muestreo destructivo (Popescu, 2007) con el consiguiente deterioro de la materia prima.

Con todo lo anterior en mente, el uso de LIDAR aparece como una tecnología no invasiva que permite reducir los costes de los procesos relacionados con la detección, gestión y medición de los cambios en los bosques. A partir de LIDAR, es posible extraer estimaciones de variables biofísicas que posteriormente se usarán para generar información (García et al., 2010) relacionada con las áreas boscosas como los inventarios forestales (Anderson et al., 2008) o los modelos de combustible (Mutlu et al., 2008).

Si LIDAR es la fuente de datos para estimar las variables en las que se basan los distintos modelos, por lo general, dichas estimaciones se realizan mediante regresión lineal múltiple (Multiple Linear Regression, MLR) entre un conjunto de mediciones en campo y un conjunto de indicadores LIDAR. La principal ventaja de utilizar MLR es su simplicidad. Por contra, este método también tiene algunas desventajas. Por ejemplo, en la mayoría de los estudios, MLR emplea «un conjunto de estadísticos relativamente amplio calculados a partir de la altura LIDAR y de los datos de su intensidad, que son sistemáticamente eliminados del modelo mediante un proceso gradual y que se traduce en un conjunto de predictores con poca justificación física» (Muss et al., 2011). Así pues, las metodologías para establecer las regresiones entre los datos LIDAR y las variables clave para la caracterización de los bosques están viéndose revisadas (Salas et al., 2010). Además, se ha comenzado a explorar el uso de nuevas técnicas de regresión no paramétricas y de metaheurísticas aplicadas a la selección de predictores (Gong et al., 2011; Latifi et al., 2010b) mejorando los resultados pero también, perdiendo parte de la simplicidad y la claridad del modelo de regresión lineal.

Buscando un punto medio entre las opciones antes comentadas, este capítulo plantea el uso de árboles de regresión con selección genética para generar las

estimaciones de distintas variables relacionadas con la biomasa a partir de datos LIDAR. En concreto, estudiaremos qué tipo de selección es más apropiada (*stepwise* o basada en algoritmos genéticos) y por otro lado, compararemos el rendimiento de un modelo basado en árboles de regresión M5P (Quinlan, 1992) contra otras técnicas recogidas en la bibliografía. Así, este capítulo pretende:

- Mostrar la mayor calidad de la computación evolutiva para guiar el proceso de selección de predictores.
- Establecer un respaldo solido para futuros trabajos que exploren nuevas mejoras en los árboles de regresión aplicados a la generación de productos derivados de LIDAR.

8.2. Trabajos relacionados

En esta sección, mostramos las principales características de los estudios aplicados a LIDAR y a estimación de variables forestales desde el punto de vista metodológico con especial énfasis en las técnicas utilizadas para establecer las regresiones, las medidas de calidad para dichas regresiones y los procesos de selección de predictores.

8.2.1. Técnicas de regresión

En el área medioambiental, las técnicas parámetricas (concretamente, MLR) han sido la herramienta básica para generar conocimiento a partir de las nubes de datos LIDAR en lo que a estimación de variables forestales se refiere. Este conjunto de técnicas parten de un conjunto de asunciones que no siempre son tenidas en cuenta y que pocas veces son comprobadas (Osborne and Waters, 2002):

- Normalidad. La regresión múltiple asume que las variables siguen una distribución normal.
- Linealidad. La regresión múltiple estándar sólo puede estimar con exactitud la relación entre las variables dependientes e independientes si las relaciones son de naturaleza lineal. Si la relación entre las variables no es lineal, los resultados del análisis de regresión pueden subestimar la verdadera relación. En cualquier caso, es posible encontrar en la literatura especializada, transformaciones en los datos para incrementar la linealidad. Así, para describir relaciones exponenciales, muchos autores toman logaritmos neperianos de los datos para las variables dependientes e independientes y linealizan el modelo. Al resultado final una vez que se deshace la transformación se le conoce como ecuación alométrica ($Y = a \times X^b$).
- Fiabilidad. A lo largo de la recogida de datos, existen múltiples situaciones que pueden generar errores. Estos errores deben ser eliminados ya que en otro caso, la relación descrita por MLR puede verse afectada de manera grave.

- Homocedasticidad. Esta propiedad asegura que las varianzas de los datos son homogéneas. Si se tiene una situación de heterocedasticidad grave, se puede producir una grave distorsión en los resultados y debilitar seriamente el análisis.

Para las situaciones en las que no es posible aplicar una regresión paramétrica, existen técnicas que se ajustan a los datos sin hacer ningún tipo de presunción sobre ellos (regresión no paramétrica). Este tipo de técnicas están generando un gran interés en el mundo LIDAR y existen ya diversos estudios que comparan los resultados de las técnicas clásicas (MLR) con técnicas tales como los procesos gaussianos (Zhao et al., 2011) o los vecinos más cercanos (Tian et al., 2012).

El potencial del aprendizaje automático en el área medioambiental puede ser mejor evaluado si se tiene en cuenta que existen estudios explorando el uso de las técnicas más novedosas en el área como es el caso de los ensembles para regresión. Así, podemos encontrar aplicaciones basadas en *Random Forests* (Stojanova et al., 2010; Shataeea et al., 2011) para realizar tareas de regresión con el objetivo de modelar la altura o la extensión de las zonas de vegetación, usando dichos ensembles como herramienta básica para combinar árboles de regresión más simples.

8.2.2. Medidas de diagnóstico

Las técnicas de regresión y en concreto MLR, se utilizan para encontrar el modelo lineal que mejor predice una variable dependiente a partir de un conjunto de variables independientes. Así, es fácil demostrar que en el caso de MLR, para p variables independientes, se tienen 2^p modelos posibles formado por sus posibles subconjuntos. Para cada modelo, se suele calcular una medida de diagnóstico con el objetivo de ayudar a determinar qué modelo es el "mejor". Estas medidas de diagnóstico incluyen el error cuadrático medio ($RMSE$) y el coeficiente de determinación (R^2) calculadas mediante las Ecuaciones 8.2 y 8.4. Un modelo lineal de buena calidad tiene un $RMSE$ bajo y R^2 alto, cercano a 1. Sin embargo, estas medidas por sí solas no son suficientes para determinar el mejor modelo ya que existen otros factores que no se tienen en cuenta como el grado de multicolinearidad y la simplicidad del modelo (principio de simplicidad) que son claves a la hora de desarrollar modelos medioambientales a partir de LIDAR.

$$SSE = \sum_{i}^{N} (x_i - \hat{x})^2 \tag{8.1}$$

$$RMSE = \sqrt{\frac{SSE}{N - p}} \tag{8.2}$$

$$SST = \sum_{i}^{N} (x_i - \bar{x})^2 \tag{8.3}$$

$$R^2 = 1 - \frac{SSE}{SST} \tag{8.4}$$

Para introducir la simplicidad en el modelo de regresión han ido apareciendo distintas medidas que potencian la disminución de predictores en el modelo final. Entre las más importantes están los índices Criterio de Información de Akaike (Akaike Information Criterion, AIC) y Criterio de Información Bayesiano (Bayesian Information Criterion, BIC).

AIC (Akaike, 1974) es una medida de la bondad relativa de ajuste de un modelo estadístico. Se basa en el concepto de entropía de la información y ofrece una medida relativa de la pérdida de información cuando un determinado modelo se utiliza para describir la realidad. AIC viene dado por la Ecuación 8.5 donde k es el número de parámetros del modelo y L es el valor maximizado de la función de similitud obtenido por dicho modelo.

$$AIC = 2k - 2ln(L) \tag{8.5}$$

BIC (Schwarz, 1978) está íntimamente relacionado con AIC pero penaliza de manera más grave el número de parámetros. En la Ecuación 8.6, k de nuevo representa el número de parámetros y N el número de instancias con las que se genera el modelo.

$$BIC = -2ln(L) + k \cdot ln(N) \tag{8.6}$$

A pesar de que el uso de BIC como selector de modelos está muy extendido, existen autores Weakliem (1999) que critican su uso por dos razones sustanciales: primero, las conclusiones alcanzadas mediante BIC dependen de que los valores sigan una misma distribución que no todas las combinaciones de predictores tienen por qué respetar; en segundo lugar, BIC tiende a favorecer modelos excesivamente simples en la práctica.

En lo que se refiere a la multicolinealidad, dos índices suele ser usados para controlarla: el número de condición (Condition Number, CN) y el factor de inflación de varianza (Variance Inflation Factor, VIF). CN (Montgomery and Peck, 1992) se calcula a partir de la matriz formada por los valores asociados a los predictores seleccionados. Posteriormente, se calculan los autovalores de dicha matriz y finalmente se calcula el valor de CN como el cociente entre el autovalor máximo y el mínimo (Ecuación 8.7).

$$CN = \left| \frac{\lambda_{max}}{\lambda_{min}} \right| \tag{8.7}$$

Aunque no existe un valor para el número de condición que sirva de umbral de multicolinearidad, existe una opinión generalizada (Belsley, 1991; Kleinbaum et al., 1988) que sostiene que un número de condición superior a 10 implica multicolinearidad y un valor de 20 o superior implica un problema de multicolinearidad de moderado a fuerte.

Por su parte, VIF es un estadístico clásico para detectar problemas de multicolinearidad que se calcula para cada predictor del modelo. Así, se define un

modelo que use el predictor a estudiar como variable dependiente y al resto de predictores como variables independientes. Una vez que se establece el modelo, el valor VIF se calcula mediante la Ecuación 8.8. El valor de VIF límite que suele establecerse es de 10 (García et al., 2010), pero cuanto menor sea el valor de VIF para cada predictor más seguros estaremos de que no existe multicolinealidad en el modelo respecto a dicho predictor.

$$VIF = \frac{1}{1 - R^2} \tag{8.8}$$

8.2.3. Selección de atributos

La forma en la que se busca el conjunto de predictores óptimo también es un factor importante a la hora de establecer las ecuaciones que describirán las relaciones entre LIDAR y el trabajo de campo. Entre los procesos de selección más utilizados están:

- **Selección hacia adelante**. La construcción de un modelo con este tipo de proceso comienza con sólo un predictor en el modelo (en general, el que mayor R^2 tenga en una regresión lineal simple (García et al., 2010)). Para cada una de las variables independientes bajo el supuesto de normalidad, se realiza un F-test para determinar si la contribución de cada variable en el modelo es significativa. La variable con el menor pvalue por debajo de un valor de corte especificado ($\alpha = 0,05$) es la elegida para ser introducida. Una vez que la variable está en el modelo, se vuelve a ejecutar el proceso con en el resto ($p - 1$ variables independientes). Este proceso continúa hasta que las variables restantes no obtienen pvalues por debajo de α.

- **Eliminación hacia atrás**. Este proceso comienza con todas las variables en el modelo. Así, la aplicación del F-test se realiza para cada variable y aquella con mayor $p - value$ superior al valor especificado por el valor α se retira del modelo. Este proceso continúa hasta que todas las variables dentro del modelo superan el test.

- **Selección paso a paso (stepwise)**. Esta metodología es una modificación de la técnica de selección hacia adelante en la que las variables que ya están en el modelo no necesariamente deben estar al final del proceso. Así, al igual que en la técnica de selección hacia adelante, las variables se agregan de una en una, pero después de que una variable se añada, la técnica paso a paso evalúa todas las variables ya incluidas en el modelo y elimina cualquier variable que no supere el F-test. El proceso paso a paso termina cuando ninguna de las variables excluidas del modelo supera el test, mientras que las que están sí lo hacen.

Estos procesos clásicos en general, usan las diferencias en R^2 antes y después de introducir o eliminar un predictor para establecer el correspondiente test estadístico. Posteriores propuestas (González-Ferreiro et al., 2011) han usado otros estadísticos como el C_p de Mallows (Mallows, 1973). Dicho estadístico se

calcula según la Ecuación 8.9 donde $\sigma^2 \approx SSE$ (SSE para el modelo con todos los predictores), N es el número de instancias de entrenamiento y P el número de predictores para el modelo concreto.

$$C_p = \frac{SSE_p}{\sigma^2} - N + 2P \tag{8.9}$$

Aunque estos estadísticos mejoran los modelos finales desde el punto de vista de la simplicidad, siguen teniendo el mismo problema que los anteriores puesto que en algunos casos, sólo se evalúa un subconjunto de los modelos posibles dependiendo de la secuencia de entrada en el sistema, y además, el uso de test estadísticos paramétricos sobre datos que no tienen por qué cumplir las condiciones paramétricas pueden dar como resultados problemas graves a la hora de seleccionar los mejores atributos. Por todo ello, la aplicación de técnicas evolutivas se han planteado recientemente como una forma de conseguir un subóptimo de mayor calidad (Latifi et al., 2010a).

8.3. Generación de los atributos LIDAR

Esta sección describe los diferentes conjuntos de datos utilizados en este capítulo con especial énfasis en el tipo de procesamiento sobre los datos LIDAR y el conjunto final de estadísticos sobre los que se aplican las distintas técnicas de regresión.

8.3.1. Trabada

El primer conjunto de datasets utilizado se extrajo de la zona de estudio de Trabada (Subsección 4.3.2) para estimar las variables biomasa de copa (W_{cr}), biomasa de tronco (W_{st}) y biomasa total (W_{abg}) medidas en campo.

La nube de puntos LIDAR del área de Trabada se preprocesó siguiendo la metodología descrita en Gonçalves-Seco et al. (2011). Así, se extrajo un conjunto de estadísticos comunes en la bibliografía (Gonçalves-Seco et al., 2011; Antonarakis et al., 2008; Hudak et al., 2008) a partir de las intensidades y de las alturas (normalizadas mediante un MDE). Dicho MDE es generado a partir de los propios datos LIDAR usando un filtro morfológico adaptativo (Gonçalves-Seco et al., 2006).

El conjunto de posibles predictores definido para esta zona se recoge en el Cuadro 8.1. Las métricas generadas se utilizan como variables independientes en los modelos de regresión, mientras que en este caso, W_{cr}, W_{st} y W_{abg} actúan como las variables dependientes. De esta forma, el conjunto total de datasets para este área es de tres (uno para cada variable a regresar).

8.3.2. Guitiriz

El segundo conjunto de datasets se extrajo de la zona de Guitiriz (descrita en la Subsección 4.4.1). Este conjunto de datasets es el más numeroso puesto

Variable	Description	Variable	Description
returns	Number of returns	hmin	Height minimum
hmax	Height maximum	hmean	Height mean
hmedian	Height median	hmode	Height mode
hSD	Height standard deviation	hV	Height variance
hID	Height interquartile range	hSkw	Height skewness
hKurt	Height kurtosis	hAAD	Height average absolute deviation
hP25	25^{th} height pct	hP50	50^{th} height pct
hP75	75^{th} height pct	hP05	5^{th} height pct
hP10	10^{th} height pct	hP20	20^{th} height pct
hP30	30^{th} height pct	hP40	40^{th} height pct
hP60	60^{th} height pct	hP70	70^{th} height pct
hP80	80^{th} height pct	hP90	90^{th} height pct
hP95	95^{th} height pct	imin	Intensity minimum
imax	Intensity maximum	imean	Intensity mean
imedian	Intensity median	imode	Intensity mode
iSD	Intensity standard deviation	iV	Intensity variance
iID	Intensity interquartile range	iSkw	Intensity skewness
iKurt	Intensity kurtosis	hAAD	Intensity average absolute deviation
iP25	25^{th} intensity pct	iP50	50^{th} intensity pct
iP75	75^{th} intensity pct	iP05	5^{th} intensity pct
iP10	10^{th} intensity pct	iP20	20^{th} intensity pct
iP30	30^{th} intensity pct	iP40	40^{th} intensity pct
iP60	60^{th} intensity pct	iP70	70^{th} intensity pct
iP80	80^{th} intensity pct	iP90	90^{th} intensity pct
iP95	95^{th} intensity pct	FP_cover2	Ratio of laser hits above 2-metre height

Cuadro 8.1: Conjunto de estadísticas calculadas para actuar como posibles predictores en la estimación de biomasa a partir de la nube de puntos LIDAR en la zona de Trabada.

que se utilizaron tanto los datos con resolución 0,5 $pulsos/m^2$ como los datos con resolución a 8 $pulsos/m^2$.

Ambos conjuntos de datos, fueron filtrados (separación entre pulsos terreno y no terreno) mediante el software Fusion (McGaughey, 2009), aplicándose una posterior interpolación para generar un MDT y un MDS (Modelo Digital de Superficie). A partir de ambos modelos, se procedió a calcular las mismas variables recogidas en el Cuadro 8.1.

Las estadísticas anteriores basadas en las distribuciones de alturas e intensidades fueron generadas para cada una de las 54 parcelas de campo y de esta forma, se pudo generar un dataset distinto para cada una de las variables medidas en el terreno: Hm (altura media, m), Hd (altura dominante, m), G (area basimétrica, m^2/ha), V (volumen, m^3/ha), W_{cr} (kg/ha), W_{st} (kg/ha), y W_{abg} (kg/ha). El total pues para esta zona de estudio es de 16 datasets distintos (8 variables $\times$ 2 resoluciones).

8.3.3. Alto Tajo

La zona de Alto Tajo también ha sido usada para este estudio (Subsección 4.5.2). Para ello, los puntos LiDAR se clasificaron en terreno y no terreno (vegetación) mediante un filtro morfológico. A partir de este preprocesado, se desarrolló un MDT mediante interpolación de los retornos terreno para normalizar las alturas. Posteriormente, se normalizaron las intensidades según la Ecuación 5.1 vista previamente. No se aplicaron otras regularizaciones puesto que no se tenía disponibles ni ángulos de incidencia ni otros datos que sirvieran para realizar un calibrado de la intensidad. Una vez que se completó el procesado de los datos, se pasó a la generación de los estadísticos del Cuadro 8.2 para cada parcela de estudio. Con dichos datos, se establecerían las regresiones sobre las variables W_{cr}, W_{st} y W_{abg} medidas en campo.

Un dato importante que hay que resaltar es que para hacer que la distribución de las distintas variables se ajustaran a una normal, los valores fueron transformados con distintas técnicas. Estas fueron de dos tipos: inversiones y aplicación de logaritmos neperianos. En lo sucesivo, si la variable usada está transformada mediante inversión, se le antepondrá al nombre el prefijo INV y en el caso de transformaciones logaritmicas, se hará lo propio con el prefijo LN.

8.4. Selección genética de predictores

Uno de los objetivos de este capítulo es demostrar que la selección genética de los predictores LIDAR es una mejor opción que la clásica selección paso a paso. Para ello, se definió un algoritmo genético que realizara esta tarea y cuyas características básicas (población inicial, función de ajuste, cruce y mutación) se detallan en las siguientes subsecciones.

Variable	Description	Variable	Description
P_25i	25th percentile [intensity]	P_25h	25th percentile [height]
P_50i	50th percentile [intensity]	P_50h	50th percentile [height]
P_75i	75th percentile [intensity]	P_75h	75th percentile [height]
P_90i	90th percentile [intensity]	P_90h	90th percentile [height]
P_99i	99th percentile [intensity]	P_99h	99th percentile [height]
Mean_i	Mean [intensity]	Mean_h	Mean [height]
Std_i	Standard deviation [intensity]	Std_h	Standard deviation [height]
CV_i	Coefficient of variation intenstity	CV_h	Coefficient of variation [height]
Range_i	Range of intenstities	CL_h	Canopy length [height]
Skew_i	Skewness [intensity]	Skew_h	Skewness [height]
Kurt_i	Kurtosis intenstity	Kurt_h	Kurtosis [height]
CanHits_i	Canopy Hits proportion [intensity]	CanHits_h	Canopy Hits proportion [height]
%Int_P25	% of [intensity] of the 25th percentile	Dif_99_25_h	P_99h - P25h
%Int_P50	% of [intensity] of the 50th percentile	Dif_99_50_h	P_99h - P50h
%Int_P75	% of [intensity] of the 75th percentile	Dif_90_25_h	P_90h - P25h
%Int_P90	% of [intensity] of the 90th percentile	Dif_90_50_h	P_90h - P50h
%Int_P99	% of [intensity] of the 99th percentile		
CRS	Canopy reflection sum		
DWCRS	Density weighted canopy reflection sum		

Cuadro 8.2: Conjunto de estadísticas calculadas para actuar como posibles predictores en la estimación de biomasa a partir de la nube de puntos LIDAR en la zona de Alto Tajo.

8.4.1. Población inicial

En este epígrafe, describimos la representación de un individuo del espacio de soluciones. En este caso, un individuo de la población es un vector en el que cada celda (genes) representa si un posible predictor en el conjunto de entrenamiento es seleccionado o no. Cada gen del individuo se inicializa con un valor de 1 ó 0 (codificación entera). Así, si la posición $i-$ésima tiene asignada un valor 1, el predictor $i-$ésimo es seleccionado y en el caso contrario, es ignorado.

Es muy importante tener en cuenta que, en la generación de la población inicial, el algoritmo se asegura de introducir todas las rectas de regresión simples posibles, para comenzar a partir del modelo mínimo (un solo predictor seleccionado).

8.4.2. Cruce y mutación

En el diseño de un algoritmo genético siempre es importante establecer un criterio de búsqueda coherente en el espacio de posibles soluciones. Esto sólo puede lograrse con una adecuada selección de los operadores de cruce y mutación.

En este caso, se aplicó la operación de cruce uniforme para dos individuos (padres) seleccionados por el método de la ruleta. El cruce selecciona aleatoriamente un gen para cada posible predictor a partir de los dos valores posibles (los genes de los padres para el predictor correspondiente). De esta forma, el conjunto de genes final es asignado al nuevo individuo.

El operador de mutación se definió para cambiar el valor de un gen con una probabilidad dada. En este caso, la mutación consiste en cambiar un valor al azar del conjunto de valores asociados a los predictores por su complementario (1 a 0 y viceversa).

8.4.3. Función de ajuste

Nuestro objetivo es medirnos con los resultados obtenidos en las tres zonas de estudio en terminos del valor BIC. Así pues, se definió una función de fitness adecuada. Así, se buscó la optimización del índice BIC que será mejor cuanto menor sea. En nuestro caso, los valores por debajo de 0 podían dar un problema ya que el framework utilizado (Dyer, 2010) no soportaba este tipo de valores en la evolución. Para evitar estos problemas, la función de fitness es la suma del BIC más una constante soporte (en nuestro caso, 1000,0) que dependerá del tipo de datos de entrada y que debe ser configurado al inicio de la ejecución.

Además, se introdujeron como funciones de filtro el CN y el VIF, respectivamente. Así, independientemente del valor de fitness que se tenga, si el valor asociado al individuo para la función de filtro supera un determinado umbral, el fitness se actualiza al valor mayor posible para que tenga una mala adaptación al entorno y por consiguiente, no pase a la siguiente generación. De esta forma, controlamos tanto la posibilidad de multicolinearidad como la complejidad del modelo final.

Parámetro	Valor
Población	100
Nº de genes	#atributos
Nº de generaciones	100
Tipo de selección	Ruleta
Cruce	Uniforme
Mutación	$P_M(S_i) = 0{,}8$
VIF umbral	10.0
CN umbral	20.0

Cuadro 8.3: Parámetros evolutivos.

8.4.4. Parametrización

Un aspecto clave en la computación evolutiva es la fase de parametrización. En el Cuadro 8.3, se resumen los parámetros seleccionados para las áreas de prueba.

8.5. Comparativa de métodos de regresión no paramétricos

El segundo objetivo de este capítulo es evaluar la capacidad de las técnicas de regresión no paramétricas para trabajar sobre conjuntos de datos reducidos como los que forman las bases de datos antes descritas. Estas técnicas además de haber demostrado su validez en otras áreas, tienen como principal ventaja la independencia respecto a la distribución concreta de los datos. Así, no realizan asunciones sobre estos y se ajustan a cada instancia de manera local. De esta forma, si se tiene un conjunto de datos lo suficientemente amplio pueden llegar a inferir modelos generales de alta calidad. Por otra parte, los métodos no paramétricos tienden a sobreajustar sus resultados cuando hay un número insuficiente de instancias.

En este caso, hemos seleccionado un conjunto de técnicas no paramétricas que ya han sido utilizadas en la bibliografía para tareas relacionadas con la estimación de variables forestales. Así, comparamos el rendimiento de los árboles de regresión en la forma del algoritmo M5P (García-Gutiérrez et al., 2011b), el algoritmo k-NN (Hudak et al., 2008), los procesos gaussianos (Zhao et al., 2011), en adelante GP, y las ANNs (Niska et al., 2010) en la forma del perceptrón multicapa.

Prácticamente todos los métodos que vamos a utilizar en este estudio se caracterizan por tener un conjunto de parámetros que deben ser ajustados por el usuario para obtener los mejores resultados. Así pues, de nuevo, utilizamos un algoritmo evolutivo para ajustarlos a la vez que se seleccionan los mejores predictores. Este último punto es especialmente importante para el algoritmo k-NN puesto que en otro caso, se corre el riesgo de obtener resultados erróneos por los

Método de regresión	Parámetro	Valor
GP	γ	[0.01, 1.0]
GP	C	[0.01, 10.0]
ANN	Momentum	[0.01, 1.0]
ANN	Learning rate	[0.01, 1.0]
ANN	Hidden layers	[1, 15]
k-NN	k	[1, 10]

Cuadro 8.4: Valores límite para los parámetros de los métodos de regresión utilizados.

problemas asociados a la alta dimensionalidad (Hughes, 1968). Las principales características del algoritmo evolutivo utilizado se presentan en los siguientes epígrafes.

8.5.1. Población inicial

Un individuo de la población está formado por dos subindividuos. El primero contiene un conjunto de valores reales representando los parámetros del método de regresión. El segundo está formado por un individuo como el descrito en la Subsección 8.4.1. Los valores iniciales del primero son valores reales que se generan aleatoriamente dentro de unos límites preestablecidos. En el Cuadro 8.4, se muestran los valores límites de cada parámetro.

8.5.2. Cruce y mutación

En este caso, también utilizamos la operación de cruce uniforme entre dos individuos padres. De esta forma, cada invididuo hijo tiene una probalidad del 50 % de heredar de un padre, el coeficiente asociado a un parámetro del método o a un posible predictor.

Un punto importante del proceso es la especificación del operador de mutación. En este caso, si un indiduo ha sido seleccionado para mutar, antes de hacerlo, se selecciona al azar el subindividuo que mutará. Así, si se elige que mute el vector de posibles predictores, el operador de mutación se comportará de la manera descrita en la Subsección 8.4.2 y si se elige el nuevo subindividuo, el operador cambiará el valor de uno de los parámetros de manera aleatoria a un nuevo valor dentro del intervalo determinado.

8.5.3. Función de ajuste y parametrización

A priori, la función de ajuste podría ser la misma que la descrita en la Subsección 8.4.3, pero el índice BIC hace uso del número de parámetros como un dato imprescindible para su cálculo. El problema es que el concepto de parámetro en BIC hace referencia al número de predictores y no al conjunto de parámetros visto en el Cuadro 8.4. Por ello, en este caso, se decidió maximizar el valor de

R^2 por ser éste un índice que no depende del número de parámetros, ni de su significado concreto.

Para evitar los problemas relacionados con la multicolinealidad y los modelos con demasiados predictores, que en general, provocarían un mayor riesgo de sobreajuste, usamos el valor VIF como filtro. Así, los individuos con valores de ajuste que tengan asociado un valor VIF por encima de 10, ven su valor de ajuste modificado para que se adapten lo peor posible a su entorno (se sustituye su valor R^2 por el valor 0).

En este caso, no se pudo utilizar el CN puesto que en su definición se exige conocer el término independiente, lo que implica que no es posible generalizarlo a otro tipo de técnicas que no sean las regresiones lineales.

Finalmente, respecto a la parametrización, se debe comentar que se ejecutaron utilizando los datos del Cuadro 8.3, puesto que habían funcionado de manera correcta a la hora de seleccionar los mejores predictores como veremos en el siguiente epígrafe.

8.6. Resultados

8.6.1. Comparativa entre selección genética y selección paso a paso de predictores

Comprobar si la selección paso a paso (*stepwise*) de los posibles predictores puede ser mejorada por una selección de características basada en computación evolutiva es uno de los principales objetivos de este capítulo. Debido a la naturaleza aleatoria de los algoritmos genéticos, y con el fin de establecer la comparación, la ejecución de los algoritmos genéticos se repitió diez veces para cada dataset y se seleccionaron los predictores del individuo que obtuvieron el valor mediano de las diez ejecuciones. En el Cuadro 8.5, se pueden ver los predictores seleccionados por cada método junto al coeficiente de determinación R^2 y BIC alcanzados por la técnica MLR con ambos tipos de selección. También se recoge las medidas de multicolinearidad explicadas anteriormente. En el caso de la selección genética, dichos datos se omiten puesto que la función de ajuste limita dichos valores a un rango en el que el grado de multicolinearidad es prácticamente nulo.

Atendiendo a la medida BIC obtenida en cada caso (R^2 mejora siempre a medida que más predictores son añadidos), se puede observar que la selección genética obtiene mejores resultados. Sólo en cuatro casos, la selección stepwise consigue batir a la selección genética. Si analizamos estas cuatro bases de datos y los modelos generados a partir de ellas, se puede observar que en todos los casos se tuvo un valor CN superior a 20 lo que indica multicolinearidad moderada. Teniendo en cuenta que el algoritmo genético se definió para no permitir este tipo de situaciones, se debe valorar aún más la potencia de esta selección frente a su competidora.

Para establecer la significancia estadística de este estudio, aplicamos un test de comparación por pares sobre los resultados BIC de cada tipo de selección. En

Area	Variable	Stepwise VIF	Stepwise CN	Stepwise BIC	Stepwise R^2	Stepwise predictors	Genetic BIC	Genetic R^2	Genetic predictors
Alto Tajo	W_{cr}	1,03	17,59	-50,32	0,67	INVP_50h, %Int_P25	**-52,69**	0,633	INVP50_h, P75_i
Alto Tajo	W_{st}	1,47	233,09	**-61,40**	0,6	P99_i, %Int_P50	-58,03	0,521	INVP90_h, DWCRS, CanHits_i
Alto Tajo	W_{abg}	1,03	17,59	-69,81	0,673	INVP_50h, %Int_P25	-69,81	0,673	INVP_50h, %Int_P25
Trabada	W_{cr}	0,00	13,31	650,51	0,708	hP90	650,51	0,708	hP90
Trabada	W_{st}	1,09	14,29	800,03	0,801	hSkew,hP75	**799,93**	0,776	hP95
Trabada	W_{abg}	0,00	12,95	809,99	0,771	hP95	809,99	0,771	hP95
Guitiriz (0,5)	G	1,46	17,90	254,29	0,620	returns,hSD, hSkw,hP05	**253,00**	0,553	returns,hP40
Guitiriz (0,5)	W_{cr}	1,46	17,90	983,40	0,616	returns,hSD, hSkw,hP05	**981,40**	0,554	returns,hP40
Guitiriz (0,5)	W_{st}	9,07	29,16	1152,60	0,707	returns,hP20, hP40	**1151,71**	0,652	hmedian
Guitiriz (0,5)	W_{abg}	1,15	13,85	1171,62	0,664	returns, hmedian	**1171,49**	0,632	hmedian
Guitiriz (0,5)	Hd	0,00	8,31	82,19	0,846	hP95	82,19	0,846	hP95
Guitiriz (0,5)	Hm	1,01	14,55	105,61	0,784	hP50,iID	**82,19**	0,846	hP95
Guitiriz (0,5)	V	9,07	29,16	504,29	0,706	hP20,hP40, returns	**503,68**	0,650	hP50
Guitiriz (8)	G	1,46	29,77	**246,53**	0,671	returns,imax, hSkw,iAAD	246,92	0,669	hID,hKurt, imin,iSkw
Guitiriz (8)	W_{cr}	1,45	21,10	981,43	0,594	hID,hKurt ,iKurt	**976,32**	0,630	returns,hKurt ,hAAD
Guitiriz (8)	W_{st}	1,23	25,94	**1148,39**	0,729	returns,imax ,hmedian	1151,57	0,738	hKurt,hAAD, ,returns,iSkw
Guitiriz (8)	W_{abg}	1,23	25,94	**1168,19**	0,713	returns,imax ,hmedian	1169,63	0,732	hKurt,hAAD ,returns,iSkw
Guitiriz (8)	Hd	0,00	8,57	81,06	0,849	hP95	81,06	0,849	hP95
Guitiriz (8)	Hm	0,00	6,97	78,34	0,759	hP30	78,34	0,759	hP30
Guitiriz (8)	V	1,23	25,94	**1148,39**	0,729	returns,imax ,hmedian	1151,57	0,738	hKurt,hAAD ,returns,iSkw

Cuadro 8.5: Capacidad de predicción (BIC y R^2) para MLR, cuando se aplican selección paso a paso y genética, respectivamente, a los datos de test. En negrita, mejores BIC para cada base de datos.

este caso, los tests de normalidad de Shapiro-Wilk (Shapiro and Wilk, 1965) y Lilliefors (Lilliefors, 1967) confirmaron que ambas distribuciones no cumplían la condición de normalidad ($pvalue_{shapiro-wilk} < 0{,}047, pvalue_{lilliefors} < 0{,}01$ para ambas distribuciones). Una de las alternativas no paramétricas para este tipo de comparación es el test de signo de Wilcoxon (Wilcoxon, 1945). Así pues, el test de Wilcoxon establece como hipótesis nula que ambas distribuciones forman parte de la misma población de resultados. En este caso, el test de Wilcoxon proporcionó un $pvalue < 6{,}104E-5$ por lo que dicha hipótesis nula es rechazada y se puede concluir que la selección genética obtiene mejores resultados que la selección paso a paso de manera significativa.

8.6.2. Comparativa de técnicas de regresión no paramétricas

Para evaluar el rendimiento de las distintas técnicas de regresión no paramétricas, se realizó un proceso de 5CV sobre cada dataset usando cada una de ellas. Para eliminar los efectos aleatorios de la computación evolutiva, cada 5CV se repitió 10 veces. En cada iteración, se mantuvo el mismo conjunto de carpetas para todos los métodos de regresión. El cuadro 8.6 recoge los resultados medios obtenidos al final de la experimentación llevada a cabo.

Para evaluar la significancia estadística de este estudio, se aplicó la metodología ya utilizada en los anteriores capítulos. En primer lugar, un test omnibus de Friedman y seguidamente un análisis post-hoc de Holms. En este caso, la variante no paramétrica fue utilizado puesto que se pudo comprobar que la distribución asociada a los resultados de GP no seguían una distribución normal ($pvalue$ de 0,029 para el test de Shapiro-Wilk con $\alpha = 0{,}05$). Así pues se estableció el ranking correspondiente que puede verse en el Cuadro 8.7. Como puede observarse MLR obtiene el mejor ranking y GP el peor. A partir de estos datos, el test de Friedman dio como resultado un $pvalue = 1{,}738E-10$ por lo que los métodos tienen un comportamiento estadísticamente diferente de manera global para un $\alpha = 0{,}05$.

El posterior análisis post-hoc mediante el procedimiento de Holms demuesta que existen diferencias significativas entre MLR y el resto de competidores excepto para el caso del árbol de regresión M5P como puede verse en el Cuadro 8.8.

8.7. Discusión

A pesar de que los algoritmos genéticos son técnicas metaheurísticas que proporcionan soluciones cercanas al óptimo pero no por fuerza óptimas, sus resultados son mucho mejores que los de la selección paso a paso si obviamos aquellos resultados con un CN que indique un nivel de colinearidad notable (superior a 20). Este hecho es especialmente interesante puesto que se supone que la selección stepwise cubre el conjunto de combinaciones para extraer el conjunto óptimo. Así pues, la pregunta clave es por qué la búsqueda genética

Area	Variable	MLR	M5P	GP	MLP	NN
Alto Tajo	W_{cr}	**0,568**	0,483	0,071	0,545	0,549
Alto Tajo	W_{st}	**0,457**	0,444	0,213	0,442	0,424
Alto Tajo	W_{abg}	**0,353**	0,343	0,019	0,193	0,323
Trabada	W_{cr}	0,307	0,313	0,220	**0,380**	0,515
Trabada	W_{st}	**0,508**	0,449	0,083	0,442	0,426
Trabada	W_{abg}	0,620	**0,648**	0,239	0,509	0,575
Guitiriz (0,5)	G	0,331	**0,385**	0,093	0,293	0,295
Guitiriz (0,5)	W_{cr}	0,382	0,379	0,103	0,201	**0,395**
Guitiriz (0,5)	W_{st}	0,445	0,524	0,000	0,484	**0,525**
Guitiriz (0,5)	W_{abg}	0,543	0,584	0,001	**0,616**	0,442
Guitiriz (0,5)	Hd	**0,774**	0,759	0,180	0,725	0,654
Guitiriz (0,5)	Hm	0,745	**0,753**	0,077	0,613	0,646
Guitiriz (0,5)	V	0,512	0,481	0,035	**0,536**	0,376
Guitiriz (8)	G	**0,435**	0,363	0,043	0,393	0,194
Guitiriz (8)	W_{cr}	0,420	0,381	0,033	0,269	0,179
Guitiriz (8)	W_{st}	0,614	**0,619**	0,139	0,576	0,575
Guitiriz (8)	W_{abg}	**0,626**	0,615	0,096	0,569	0,468
Guitiriz (8)	Hd	0,803	**0,808**	0,227	0,661	0,416
Guitiriz (8)	Hm	**0,706**	0,689	0,026	0,595	0,525
Guitiriz (8)	V	**0,674**	0,665	0,046	0,450	0,613

Cuadro 8.6: Capacidad de predicción (R^2) para MLR y el resto de métodos no paramétricos con selección genética. En negrita, mejores resultados.

Distribution	Rank
MLR	1,75
M5P	2,05
GP	5,00
MLP	3,00
NN	3,20

Cuadro 8.7: Rankings medios para cada método de regresión.

	algorithm	p-value	Holm's
1	GP	0,0000	0,0125
2	NN	0,0037	0,0167
3	MLP	0,0124	0,025
4	M5P	0,5485	0,050

Cuadro 8.8: P-values ajustados de Holm para los tests de significación entre MLR y el resto de métodos de regresión no paramétricos.

Area	Variable	Kolmogorov-Smirnov	Shapiro-Wilk	Lilliefors
Alto Tajo	W_{cr}	1,0	0,162	0,001
Alto Tajo	W_{st}	1,0	0,107	0,001
Alto Tajo	W_{abg}	1,0	0,260	0,001
Trabada	W_{cr}	1,0	0,633	0,001
Trabada	W_{st}	1,0	0,521	0,001
Trabada	W_{abg}	1,0	0,178	0,001
Guitiriz (0,5)	G	0,05	0,001	0,001
Guitiriz (0,5)	W_{cr}	0,1	0,001	0,001
Guitiriz (0,5)	W_{st}	0,05	0,001	0,001
Guitiriz (0,5)	W_{abg}	0,05	0,001	0,001
Guitiriz (0,5)	Hd	0,05	0,001	0,001
Guitiriz (0,5)	Hm	0,1	0,001	0,001
Guitiriz (0,5)	V	0,2	0,001	0,001
Guitiriz (8)	G	0,1	0,001	0,001
Guitiriz (8)	W_{cr}	0,2	0,001	0,001
Guitiriz (8)	W_{st}	0,01	0,001	0,001
Guitiriz (8)	W_{abg}	0,01	0,001	0,001
Guitiriz (8)	Hd	0,01	0,001	0,001
Guitiriz (8)	Hm	1,0	0,007	0,001
Guitiriz (8)	V	0,01	0,001	0,001

Cuadro 8.9: P-values para los tests de normalidad (Kolmogorov, Shapiro, Lilliefors) sobre los residuos de las regresiones generadas mediante selección paso a paso.

es mejor. Para responder esta cuestión, se debe analizar la forma en que la selección paso a paso se ejecuta. Como hemos visto, basa sus decisiones en test estadísticos que asumen normalidad en los datos. Sin embargo, esta premisa pocas veces es tenida en cuenta. El resultado, cuando evaluamos los modelos finales obtenidos en cada caso utilizando una plataforma ad-hoc (Parejo et al., 2012b), muestran que para un $\alpha = 0{,}05$ y en el caso del test de Lilliefors, la normalidad en los residuos cuadrados obtenidos en las regresiones es rechazada en todos los casos (ver Cuadro 8.9).

La selección genética al no introducir ningún tipo de asunción es una técnica que se ajusta mejor a la «realidad» de los datos. En cualquier caso, MLR no debería ser aplicado si se tiene la certidumbre de que las premisas necesarias para aplicar la regresión paramétrica no van a cumplirse. Así pues en este capítulo, hemos planteado el uso de técnicas de regresión no paramétricas en su lugar, sobre los datos de las diferentes áreas de estudio.

En lo que respecta a la aplicación de métodos no paramétricos, hay que destacar que en general, sus resultados fueron pobres en comparación con los obtenidos por MLR. Este hecho se explica por el tamaño excesivamente pequeño de las bases de datos que oscila entre las 35 y 60 instancias. Los árboles de

regresión debido a su mayor sencillez conceptual y su cercanía a MLR, obtienen resultados comparables. Por otro lado, a excepción de los GPs, todos los métodos no paramétricos consiguen mejorar al resto en algún momento por lo que no se debe descartar su uso en determinados momentos. Las condiciones de uso deberán ser estudiadas más en profundidad en trabajos futuros.

Un dato muy importante, como hemos comentado, es que, en contra de lo que cabía esperar (Zhao et al., 2011), los GPs obtienen los peores resultados. Algunos autores han obtenido muy buenos resultados con dichos métodos pero con la necesidad de aplicar un procesado para evitar el sobreajuste. Para poder usar esta técnica en este estudio, se debería seleccionar y realizar un proceso similar para cada uno de sus competidores.

Para concluir, debemos subrayar de nuevo la importancia de comprobar las asunciones en las que se basa la regresión paramétrica. En el caso de que no se cumpliesen y aún así se quisiera realizar un modelo basado en MLR por su simplicidad, se debería siempre utilizar una selección de predictores genética en lugar de un proceso stepwise o análogo. Si la claridad en el modelo es clave y el uso de MLR no es imprescindible, recomendamos el uso de árboles de regresión (M5P) que no establecen ningún tipo de asunción previa, tienen niveles de claridad en el modelo similares a los de MLR y además, consiguen resultados comparables al de MLR incluso en un entorno con bases de datos que le perjudican debido a que, como hemos visto, las técnicas más avanzadas sufren de sobreajuste por el reducido tamaño de las bases de datos.

Capítulo 9

Conclusions and future work

The beginning and end are common on the circumference of a circle.
Heraclitus of Ephesus.

In recent years, LIDAR has become a very important technology for environmental sciences. There already exist applications based on LIDAR that have been used for the development of digital models and thematic maps, or estimation of biomass in natural areas. These applications are benefiting from lower prices and higher quality LIDAR data collected. This higher quality must be accompanied by an improvement in the techniques that exploit them. Thus, the use of the most advanced techniques on LIDAR data is fully justified.

With the above in mind, professionals involved in computer science should make an effort to address the application of cutting-edge techniques from the field of data mining on LIDAR. Thus, the research conducted in this Ph.D. dissertation explored the synergy of LIDAR sensors and advanced intelligent techniques from data mining and soft computing which have provided the results summarized in this chapter.

- Thematic maps are one of the main products generated from remote sensing data. Although they have traditionally been generated by image classification, other sensors such as LIDAR have begun to generate a great interest because they can register object heights. According to our experiments, we confirmed that LIDAR sensors are perfectly viable for the development of thematic maps. Within classical classification algorithms used in the literature, we recommend the use of DTs in the case of maps from only LIDAR data. DTs have important advantages such as providing a white box model, which can be essential if the user is not an expert in machine learning and an easily understandable set of classification rules is required.

- Although the power of LIDAR as a data source on its own for the development of thematic maps has been demonstrated, most approaches for thematic mapping related to LIDAR are based on data fusion where LIDAR is mainly combined with different types of images (multispectral images or orthophotography). Experiments conducted in this study showed that better results could be obtained with fused LIDAR than with isolated LIDAR. In addition, although DTs obtained acceptable results they were clearly outperformed by SVMs. The influence of the cost of parameterization should be evaluated in each case, but it can be concluded that SVMs is the most suitable classifier for data fusion.

- Independently of the classifier selected, pixel-oriented approaches cause problems such as «salt and pepper» noise due to instances with features similar for two or more different classes. To solve this problem, other approaches such as object orientation have been proposed. This type of approaches involve early phases of segmentation that usually complicate the development of thematic maps by introducing a large number of parameters in the system. Contextual classifiers appear as simpler methods that get similar accuracies but at a lower cost in parameterization time and complexity. In this study, we showed a novel contextual classifier called SVNNNS. SVNNNS consists of an initial classification with an SVM and a later regularization based on non-parametric classifiers (NN). The results of experiments carried out showed that SVMNNS outperformed SVM with MRF regularization for LIDAR and image data fusion even when SVMMRF is a cutting-edge classifier in literature.

- Management of forested areas is one of the most important applications of LIDAR. The general procedure involves the extraction of statistical data from the LIDAR cloud and a later phase which establishes a relationship between these statistics (independent variables) and biophysical variables extracted from field work (dependent variables). MLR after a stepwise feature selection is the usual technique to establish this relationship. Our experiments showed that the common stepwise feature selection should be always substituted by a genetic feature selection because stepwise selection is easily affected by non-normality which is very common in real data. In addition, this Ph. dissertation has shown that other regression techniques, concretely regression trees, can provide non-parametric models with similar levels of simplicity and power.

With regard to future work, there are several possible improvements that should be explored to advance in the studies included in this report and several avenues for future work than can complement our work nowadays:

- An important task not covered in this document is DEM generation. Clustering has proven to be capable of filtering terrain returns from LIDAR data cloud (Filin, 2004) although this task is always hard and difficult. A way to improve the results could be the application of a feature weighting.

The use of evolutionary computation could adjust weighting so that the influence of each feature in the clustering would be modified according to the area being filtered in order to ease the final classification.

- An inherent weakness of image classification is the fact that it generally works on data with categorical labels. When working with real data, the categories may be too restrictive. This situation can be controlled by introducing a certain degree of fuzziness. Another solution is to work on the data using regression techniques rather than classification. In that case we could see a label as a numerical value between 1 and the number of classes studied. From here, it is logical that the transitions between types of classes will be smoother, «salt and pepper» noise will be reduced, and more importantly, the process of recording information will be more likely the way visual information is caught by humans. That is, working with continuous values in lieu of discreet ones.

- Closely related to the previous point, there is a need to explore methods for regression, especially those techniques that are devoted to the combination of regression techniques in a similar way to ensembles of classifiers. This research could improve the results of studies such as LIDAR-based biomass estimation, but also could help identify new ways to generate thematic maps as discussed previously.

- As we saw in the introduction, the full-wave LIDAR data are becoming more and more important in tasks related not only to the environmental field but to other disciplines such as archeology. The use of full-wave data is not yet widespread, mainly because there are no specific tools to deal with this type of data source. In this context, data mining techniques can be helpful and concretely, time series classification could be adapted to this type of signals and thus obtain results from this richer source.

Finally, an aspect not addressed in this document is how to transfer the results to the industrial world. This is a key factor in any process of knowledge generation if it is desired to be available for the general public. Taking into account that this document is a Computer Science Ph.D. dissertation, our best way to transfer knowledge is to provide a simple tool that could handle LIDAR data in the fashion seen in this document. Thus, we plan to make the methodologies presented in this study available via web or integrating the methodologies presented in this document into consolidated GIS developments.

Bibliografía

D. Aha and D. Kibler. Instance-based learning algorithms. *Machine Learning*, 6:37–66, 1991.

H. Akaike. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6):716–723, 1974.

J. Anderson, L. Plourde, M. Martin, B. Braswell, M. Smith, R. Dubayah, M. Hofton, and B. Blair. Integrating waveform lidar with hyperspectral imagery for inventory of a northern temperate forest. *Remote Sensing of Environment*, 112(4):1856–1870, 2008.

A. Antonarakis, K. Richards, and J. Brasington. Object-based land covers classification using airborne LiDAR. *Remote Sensing of Environment*, 112:2988–2998, 2008.

L. A. Arroyo, C. Pascual, and J. A. Manzanera. Fire models and methods to map fuel types: The role of remote sensing. *Forest Ecology and Management*, 256:1239 – 1252, 2008.

LAS Specification. ASPRS, 1.1 edition, March 2005.

C. Bachmann, T. Donato, G. Lamela, W. Rhea, M. Bettenhausen, R. Fusina, K. Du Bois, J. Porter, and B. Truitt. Automatic classification of land cover on Smith Island, VA, using HyMAP imagery. *Geoscience and Remote Sensing, IEEE Transactions on*, 40(10):2313 – 2330, 2002.

D. Belsley. *Conditioning Diagnostics: Collinearity and Weak Data in Regression*. Wiley-Interscience, 1991.

J. A. Benediktsson, P. H. Swain, and O. K. Ersoy. Neural network approaches versus statistical methods in classification of multisource remote sensing data. *IEEE Transactions on Geoscience and Remote Sensing*, 28(4):540–552, 1990.

J. Besag. On the statistical analysis of dirty pictures. *Journal of the Royal Statistical Society*, 48:259–302, 1986.

J. C. Bezdek. *Pattern Recognition with Fuzzy Objective Function Algoritms*. Plenum Press, New York, 1981.

J. C. Bezdek, R. Ehrlich, and W. Full. Fcm: The fuzzy c-means clustering algorithm. *Computers & Geosciences*, 10(2-3):191–203, 1984.

E. W. Bork and J. G. Su. Integrating LiDAR data and multispectral imagery for enhanced classification of rangeland vegetation: A meta analysis. *Remote Sensing of Environment*, 111:11–24, 2007.

S. Borman. The Expectation Maximization algorithm – a short tutorial. http://www.seanborman.com/publications/EM/algorithm.pdf, 2004.

J. Bosque Sendra. *Sistemas de información geográfica*. RIALP, 1997.

I. Bracken and C. Wesbter. *Information Technology in Geography and Planning: Including Principles of Gis*. Routledge, 1990.

L. Bruzzone and D. Fernández Prieto. A partially unsupervised cascade classifier for the analysis of multitemporal remote-sensing images. *Pattern Recognition Letters*, 23(9):1063 – 1071, 2002.

A. Brzank, C. Heipke, J. Goepfert, and U. Soergel. Aspects of generating precise digital terrain models in the Wadden Sea from lidar water classification and structure line extraction. *ISPRS Journal of Photogrammetry and Remote Sensing*, 63:510–528, 2008.

P. Burrough. *Principles of Geographical Information Systems for land resources assessment*. Oxford University Press, 1988.

P. Burrough, P. Van Gaans, and R. MacMillan. High-resolution landform classification using fuzzy k -means. *Fuzzy Sets and Systems*, 113(1):37–52, 2000.

G. Camps-Valls. Machine learning in remote sensing data processing. In *Machine Learning for Signal Processing, 2009. MLSP 2009. IEEE International Workshop on*, pages 1–6, 2009.

G. Camps-Valls, T. Bandos Marsheva, and D. Zhou. Semi-supervised graph-based hyperspectral image classification. *IEEE Transactions on Geoscience and Remote Sensing*, 45(10):3044–3054, 2007.

M. J. Canty. Boosting a fast neural network for supervised land cover classification. *Computers & Geoscience*, 35(6):1280–1295, 2009.

X. Cao, J. Chen, B. Matsushita, H. Imura, and L. Wang. An automatic method for burn scar mapping using support vector machines. *International Journal of Remote Sensing*, 30(3):577–594, 2009.

Y. Cao, Y. Luo, and S. Yang. Image denoising based on hierarchical Markov random field. *Pattern Recognition Letters*, 32(2):368–374, 2011.

C. Cariou and K. Chehdi. Unsupervised texture segmentation /classification using 2-D autoregressive modeling and the stochastic expectation-maximization algorithm. *Pattern Recognition Letters*, 29(7):905 – 917, 2008.

L. Castellana, A. D'Addabbo, and G. Pasquariello. A composed supervised/unsupervised approach to improve change detection from remote sensing. *Pattern Recognition Letters*, 28(4):405 – 413, 2007.

J. Cebrián. *Sistemas de Información Geográfica. Aplicaciones de la informática a la Geografía y Ciencias sociales*, pages 125–140. Síntesis, 1988.

T. Celik. Unsupervised change detection in satellite images using principal component analysis and k-means clustering. *Geoscience and Remote Sensing Letters, IEEE*, 6(4):772–776, 2009.

C.-C. Chang and C.-J. Lin. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2:27:1–27:27, 2011. `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.

N. Chawla, N. Japkowicz, and A. Kolcz. Editorial: Special issue on learning from imbalanced data sets. *ACM SIGKDD*, 6(1):1–6, 2004.

Q. Chen. Airborne lidar data processing and information extraction. *Photogammetry engineering Remote Sensing*, 2(73):109–112, 2007.

G. Chust, I. Galparsoro, A. B., J. Franco, and A. Uriarte. Coastal and estuarine habitat mapping, using LIDAR height and intensity and multi-spectral imagery. *Estuarine, Coastal and Shelf Science*, 78(4):633 – 643, 2008.

E. Chuvieco. *Teledetección Ambiental. La observación de la Tierra desde el Espacio*. Ariel, S.A., 2008.

Comisión al Consejo y al Parlamento Europeo. Plan de acción sobre biodiversidad para la conservación de los recursos naturales. online, 2001.

R. G. Congalton and K. Green. *Assessing the accuracy of remotely sensed data: Principles and practices*. CRC Press, Taylor & Francis group, 2nd edition, 2008.

W. J. Conover. *Practical Nonparametric Statistics*. Wiley, New York, NY USA, 1999.

C. Cortes and V. Vapnik. Support-vector networks. *Machine Learning*, 20: 273–297, 1995.

F. J. Cortijo and N. P. D. L. Blanca. Improving classical contextual classifications. *International Journal of Remote Sensing*, 19(8), 1998.

M. Dalponte, L. Bruzzone, and D. Gianelle. Fusion of hyperspectral and LIDAR remote sensing data for classification of complex forest areas. *IEEE Transactions On Geoscience And Remote Sensing*, 46(5):1416–1427, 2008.

M. Dalponte, L. Bruzzone, L. Vescovo, and G. Damiano. The role of spectral resolution and classifier complexity in the analysis of hyperspectral images of forest areas. *Remote Sensing of Environment*, 113:2345–2355, 2009.

B. V. Dasarathy. *Nearest neighbor (NN) norms: NN pattern classification techniques*. Los Alamitos: IEEE Computer Society Press, 1990, 1990.

A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society: Series B*, 39(1):1–38, November 1977.

J. Demsar. Statistical comparisons of classifiers over multiple data sets. *Journal of Machine Learning Research*, 7:1–30, 2006.

H. Deng and D. A. Clausi. Unsupervised image segmentation using a simple MRF model with a new implementation scheme. *Pattern Recognition*, 37(12): 2323 – 2335, 2004.

W. Di and M. Crawford. Active learning via multi-view and local proximity co-regularization for hyperspectral image classification. *IEEE Journal on Selected Topics in Signal Processing*, 5(3):618–628, 2011.

T. Dietterich. An experimental comparison of nearest neighbor and nearest hyperrectangle algorithms. *Machine Learning*, 19(1):5–28, 1995.

D. N. Donoghue, P. J.Watt, N. J. Cox, and J. Wilson. Remote sensing of species mixtures in conifer plantations using LiDAR height and intensity data. *Remote Sensing of Environment*, 110:110–509, 2007.

W. A. Dorigo, R. Zurita-Milla, A. J. W. de Wit, J. Brazile, R. Singh, and M. E. Schaepman. A review on reflective remote sensing and data assimilation techniques for enhanced agroecosystem modeling. *International journal of Applied Earth Observation and Geoinformation*, 9:165–193, 2007.

T. Duda and M. Canty. Unsupervised classification of satellite imagery: Choosing a good algorithm. *International Journal of Remote Sensing*, 23(11): 2193–2212, 2002.

D. W. Dyer. Watchmaker framework. online, 2010.

G. Easson and H. Momm. Evolutionary computation for remote sensing applications. *Geography Compass*, 4(3):172–192, 2010.

eCognition Software. http://www.ecognition.com/. online, 2011.

M. Egmont-Petersen, D. de Ridder, and H. Handels. Image processing with neural networks–a review. *Pattern Recognition*, 35(10):2279 – 2301, 2002.

Y. El-Manzalawy and V. Honavar. *WLSVM: Integrating LibSVM into Weka Environment*, 2005. `http://www.cs.iastate.edu/~yasser/wlsvm`.

T. Esch, V. Himmler, G. Schorcht, M. Thiel, T. Wehrmann, F. Bachofer, C. Conrad, M. Schmidt, and S. Dech. Large-area assessment of impervious surface based on integrated analysis of single-date Landsat-7 images and geospatial vector data. *Remote Sensing of Environment*, 113(8):1678–1690, 2009.

Exelis. http://www.exelisvis.com/productsservices/envi/capabilities.aspx. online, 2000.

J. Fan, M. Han, and J. Wang. Single point iterative weighted fuzzy C-means clustering algorithm for remote sensing image segmentation. *Pattern Recognition*, 42(11):2527 – 2540, 2009.

M. Fauvel, J. Chanussot, and J. Benediktsson. A spatial-spectral kernel-based approach for the classification of remote-sensing images. *Pattern Recognition*, 45(1):381 – 392, 2012.

S. Filin. Surface classification from airborne laser scanning data. *Computers & Geosciences*, 2004.

G. Foody and A. Mathur. Toward intelligent training of supervised image classifications: Directing training data acquisition for SVM classification. *Remote Sensing of Environment*, 93(1-2):107–117, 2004.

R. Fraser, I. Olthof, and D. Pouliot. Monitoring land cover change and ecological integrity in Canada's national parks. *Remote Sensing of Environment*, 113: 1397–1409, 2009.

M. A. Friedl and C. E. Brodley. Decision tree classification of land cover from remotely sensed data. *Remote Sensing of Environment*, 61:399–409, 1997.

M. Friedman. The use of ranks to avoid the assumption of normality implicit in the analysis of variance. *Journal of the American Statistical Association*, 200(32):675–701, 1937.

M. Friedman. A comparison of alternative tests of significance for the problem of m rankings. *The Annals of Mathematical Statistics*, 1(11):86–92, 1940.

R. Gamanya, P. D. Maeyer, and M. D. Dapper. Object-oriented change detection for the city of Harare, Zimbabwe. *Expert Systems with Applications*, 36:571–588, 2009.

M. García, D. Riaño, E. Chuvieco, and F. Danson. Estimating biomass carbon stocks for a mediterranean forest in central Spain using LIDAR height and intensity data. *Remote Sensing of Environment*, 114(4):816–830, 2010.

M. García, D. Riaño, E. Chuvieco, J. Salas, and F. M. Danson. Multispectral and LiDAR data fusion for fuel type mapping using support vector machine and decision rules. *Remote Sensing of Environment*, 115(6):1369–1379, 2011.

S. García and F. Herrera. An extension on statistical comparisons of classifiers over multiple data sets for all pairwise comparisons. *Journal of Machine Learning Research*, 9:2677–2694, 2008.

S. García, A. Fernandez, J. Luengo, and F. Herrera. A study of statistical techniques and performance measures for genetics-based machine learning: Accuracy and interpretability. *Soft Computing*, 13(10):959–977, 2009.

J. García-Gutiérrez, L. Gonçalves-Seco, and J. C. Riquelme-Santos. Automatic environmental quality assessment for mixed-land zones using lidar and intelligent techniques. *Expert Systems with Applications*, 38:6805–6813, 2011a.

J. García-Gutiérrez, E. González-Ferreiro, D. Mateos-García, J. C. R. Santos, and D. Miranda. A comparative study between two regression methods on lidar data: A case study. In *HAIS (2)'11*, pages 311–318, 2011b.

J. García-Gutiérrez, M. García, and J. C. Riquelme-Santos. A comparative study of intelligent techniques applied to LiDAR and imagery data fusion for thematic maps. *International Journal of Remote Sensing*, To be submitted, 2012a.

J. García-Gutiérrez, D. Mateos-García, and J. C. Riquelme-Santos. EVOR-STACK: a label-dependent evolutive stacking on remote sensing data fusion. *Neurocomputing*, 75(1):115–122, 2012b.

J. García-Gutiérrez, D. Mateos-García, and J. C. Riquelme-Santos. A comparison between contextual classifiers for high-resolution thematic mapping. *IEEE Geoscience and Remote Sensing Letters*, To be submitted, 2012c.

J. García-Gutiérrez, E. González-Ferreiro, D. Mateos-García, J. C. R. Santos, and D. Miranda. A comparative study among machine learning regression methods on LiDAR data. *Forest Ecology and Management*, To be submitted.

S. Geman and D. Geman. Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, PAMI-6(6):721 –741, 1984.

A. Ghosh, N. S. Mishra, and S. Ghosh. Fuzzy clustering algorithms for unsupervised change detection in remote sensing images. *Information Sciences*, 181(4):699 – 715, 2011.

S. Ghosh, S. Patra, and A. Ghosh. An unsupervised context-sensitive change detection technique based on modified self-organizing feature map neural network. *International Journal of Approximate Reasoning*, 50(1):37 – 50, 2009.

S. Gill, J. F. Handley, A. R. Ennos, S. Pauleit, N. Theuray, and S. Lindley. Characterising the urban environment of UK cities and towns: A template for landscape planning. *Landscape and Urban Planning*, 87:210–222, 2008.

S. Goetz, D. Steinberg, R. Dubayah, and B. Blair. Laser remote sensing of canopy habitat heterogeneity as a predictor of bird species richness in an eastern temperature forest, USA. *Remote Sensing of Environment*, 108:254–263, 2007.

S. Goetz, P. Jantz, and C. A. Jantz. Connectivity of core habitat in the Northeastern United States: Parks and protected areas in a landscape context. *Remote Sensing of Environment*, 113:1421–1429, 2009.

E. Gomez Ballester, L. Mico, and J. Oncina. Some approaches to improve tree-based nearest neighbour search algorithms. *Pattern Recognition*, 39(2): 171–179, 2006.

L. Gonçalves-Seco, D. Miranda, R. Crecente, and J. Farto. Digital terrain model generation using airborne LIDAR in forested area of Galicia, Spain. In *Accuracy 2006*, pages 169–180, 2006.

L. Gonçalves-Seco, E. González-Ferreiro, U. Diéguez-Aranda, B. Fraga-Bugallo, R. Crecente, and D. Miranda. Assessing the attributes of high-density eucalyptus globulus stands using airborne laser scanner data. *International Journal of Remote Sensing*, 32(24):9821–9841, 2011.

B. Gong, J. Im, and G. Mountrakis. An artificial immune network approach to multi-sensor land use/land cover classification. *Remote Sensing of Environment*, 115(2):600–614, 2011.

E. González-Ferreiro, U. Diéguez-Aranda, , and D. Miranda. Estimation of stand variables in pinus radiata d. don plantations using different lidar pulse densities. *Forestry: An International Journal of Forest Research*, In press, 2011.

E. González-Pons. Internet, el final de la edad contemporánea. *Politica Exterior*, 14:155–161, 2000.

S. Grebby, J. Naden, D. Cunningham, and K. Tansey. Integrating airborne multispectral imagery and airborne LiDAR data for enhanced lithological mapping in vegetated terrain. *Remote Sensing of Environment*, 115(1):214 – 226, 2011.

A. Guerrero-Curieses, R. Alaiz-Rodriguez, and J. Cid-Sueiro. A fixed-point algorithm to minimax learning with neural networks. *Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on*, 34 (4):383–392, 2004.

S. R. Gunn. SVM for classification and regression. Technical report, Image Speech and intelligent system groups, 1998.

R. Haapanen, A. R. Ek, M. E. Bauer, and A. O. Finley. Delineation of forest/nonforest land use classes using nearest neighbor methods. *Remote Sensing of Environment*, 89(3):265 – 271, 2004.

O. Hagner and H. Reese. A method for calibrated maximum likelihood classification of forest types. *Remote Sensing of Environment*, 110:438–444, 2007.

M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten. The WEKA data mining software: An update. *SIGKDD Explorations*, 11(1), 2009.

M. A. Hall. Correlation-based feature selection for machine learning. Technical report, Hamilton, New Zealand, 1999.

S. Hall, I. Burke, D. Box, M. Kaufmann, and J. Stoker. Estimating stand structure using discrete-return lidar: an example from low density, fire prone ponderosa pine forests. *Forest. Ecol. Manag.*, 208:189–209, 2005.

S. Haykin. *Neural Networks: A Comprehensive Foundation*. Prentice Hall PTR, Upper Saddle River, NJ, USA, 2nd edition, 1998.

J. Hernández Orallo. *Introducción a la minería de datos*. Prentice Hall, 2004.

B. Hofle and N. Pfeifer. Correction of laser scanning intensity data: Data and model-driven approaches. *ISPRS Journal of Photogrammetry and Remote Sensing*, 62(6):415 – 433, 2007.

S. Holm. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6:65–70, 1979.

C. Huang, L. S. Davis, and J. R. G. Townshend. An assessment of support vector machines for land cover classification. *International Journal of Remote Sensing*, 23(4):725 – 749, 2002.

C.-L. Huang and C.-J. Wang. A ga-based feature selection and parameters optimizationfor support vector machines. *Expert Systems with Applications*, 31(2):231 – 240, 2006.

A. T. Hudak, N. L. Crookston, J. S. Evans, D. E. Halls, and M. J. Falkowski. Nearest neighbor imputation of species-level, plot-scale forest structure attributes from LiDAR data. *Remote Sensing of Environment*, 112:2232–2245, 2008.

G. Hughes. On the mean accuracy of statistical pattern recognizers. *IEEE Transactions on Information Theory*, 14:55–63, 1968.

M. Hughes, J. Schmidt, and P. C. Almond. Automatic landform stratification and environmental correlation for modelling loess landscapes in North Otago, South Island, New Zealand. *Geoderma*, 149:92–100, 2009.

W. Jank. Stochastic variants of EM: Monte Carlo, quasi-Monte Carlo and more. In *Joint Statistical Meetings*, 2005.

N. Japkowicz and S. Stephen. The class imbalance problem: A systematic study. *Intell. Data Anal.*, 6:429–449, 2002.

J. L. R. Jensen, K. S. Humes, L. A. Vierling, and A. T. Hudak. Discrete return lidar–based prediction of leaf area index in two conifer forests. *Remote Sensing of Environment*, 112:2988–2998, 2008.

D. A. Jones, A. J. Hansen, K. Bly, K. Doherty, J. P. Verschuyl, J. I. Paugh, R. Carle, and S. J. Story. Monitoring land use and cover around parks: A conceptual approach. *Remote Sensing of Environment*, 113:1346–1356, 2009.

Y. Ke, L. J. Quackenbush, and J. Im. Synergistic use of QuickBird multispectral imagery and LIDAR data for object-based forest species classification. *Remote Sensing of Environment*, 114(6):1141 – 1154, 2010.

S. Keerthi, S. Shevade, C. Bhattacharyya, and K. Murthy. Improvements to Platt's SMO algorithm for SVM classifier design. *Neural Computation*, 13(3): 637–649, 2001.

D. Kleinbaum, L. Kupper, and K. Muller. *Applied Regression Analysis and Other Multivariables Methods*. PWS-KENT Publishing Company, 1988.

J. Knorn, A. Rabe, V. Radeloff, T. Kuemmerle, J. Kozak, and P. Hostert. Land cover mapping of large areas using chain classification of neighboring Landsat satellite images. *Remote Sensing of Environment*, 113(5):957–964, 2009.

B. Koetz, F. Morsdorf, S. van der Linden, T. Curt, and B. Allgower. Multi-source land cover classification for forest fire management based on imaging spectrometry and LiDAR data. *Forest Ecology and Management*, 256:263–271, 2008.

T. Kohonen. Self-organized formation of topologically correct feature maps. *Biological Cybernetics*, 43:59–69, 1982.

T. Kohonen, M. R. Schroeder, and T. S. Huang. *Self-Organizing Maps*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 3rd edition, 2001.

R. F. Kokaly, B. W. Rockwell, S. L. Haire, and T. V. V. King. Characterization of post-fire surface cover, soils, and burn severity at the Cerro Grande Fire, New Mexico, using hyperspectral and multispectral remote sensing. *Remote Sensing of Environment*, 106:305–325, 2007.

T. Kuemmerle, T. Chaskovskyy, J. Knorn, V. Radeloff, I. Kruhlov, W. Keeton, and P. Hostert. Forest cover change and illegal logging in the Ukrainian Carpathians in the transition period from 1988 to 2007. *Remote Sensing of Environment*, 113(6):1194–1207, 2009.

R. Lasaponara and A. Lanorte. Remotely sensed characterization of forest fuel types by using satellite ASTER data. *International Journal of Applied Earth Observation and Geoinformation*, 9(3):225 – 234, 2007.

H. Latifi, A. Nothdurft, and B. Koch. Non-parametric prediction and mapping of standing timber volume and biomass in a temperate forest: application of multiple optical/lidar-derived predictors. *Forestry*, 83(4):395–407, 2010a.

H. Latifi, A. Nothdurft, and B. Koch. Non-parametric prediction and mapping of standing timber volume and biomass in a temperate forest: Application of multiple optical/LiDAR-derived predictors. *Forestry*, 83(4):395–407, 2010b.

A. L. Levada, N. D. Mascarenhas, and A. Tannas. A novel MAP-MRF approach for multispectral image contextual classification using combination of suboptimal iterative algorithms. *Pattern Recognition Letters*, 31(13):1795 – 1808, 2010.

T. Lillesand and R. Kiefer. *Remote sensing and image interpretation.* New York: John Wiley and Sons, 2000.

H. Lilliefors. On the kolmogorov smirnov test for normality with mean and variance unknown. *Journal of the American Statistical Association*, 62:399–402, 1967.

S. P. Lloyd. Least squares quantization in PCM. *IEEE Trans. Information Theory*, 28:129–137, 1982.

J. Lovell, D. Jupp, G. Newnham, N. Coops, and D. Culvenor. Simulation study for finding optimal lidar acquisition parameters for forest height retrieval. *Forest Ecol. Manag.*, 214:398 – 412, 2005.

D. Lu and Q. Weng. A survey of image classification methods and techniques for improving classification performance. *International Journal of Remote Sensing*, 28(5):823–870, 2007.

L. Luo and G. Mountrakis. Converting local spectral and spatial information from a priori classifiers into contextual knowledge for impervious surface classification. *ISPRS Journal of Photogrammetry and Remote Sensing*, 66(5):579 – 587, 2011.

J. B. MacQueen. Some methods for classification and analysis of multivariate observations. In *Proceedings of the Fifth Symposium on Math, Statistics,and Probability*, pages 281–297, 1967.

S. Magnussen, R. E. McRoberts, and E. O. Tomppo. Model-based mean square error estimators for k-nearest neighbour predictions and applications using remotely sensed data for forest inventories. *Remote Sensing of Environment*, 113:476–488, 2009.

C. L. Mallows. Some comments on cp. *Technometrics*, 15(4):661–675, 1973.

D. Mazzoni, M. J. Garay, R. Davies, and D. Nelson. An operational MISR pixel classifier using support vector machines. *Remote Sensing of Environment*, 107:149–158, 2007.

W. McCulloch and W. Pitts. A logical calculus of the ideas immanent in nervous activity. *Bulletin of Mathematical Biophysics*, 5:115–133, 1943.

R. McGaughey. Fusion v1.48, lidar data viewer and analysis tool. online, 2009.

Q. McNemar. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, 12:153–157, 1947.

L. Mico, J. Oncina, and R. Carrasco. A fast branch-and-bound nearest-neighbor classifier in metric-spaces. *Pattern Recognition Letters*, 17(7):731–739, 1996.

E. Ministerio de Medio Ambiente. *Nuevos Desarrollos. Obtención de la cartografía mediante tecnología LIDAR y ortofotografía*, 2011.

P. Mitra, B. Uma Shankar, and S. Pal. Segmentation of multispectral remote sensing images using active support vector machines. *Pattern Recognition Letters*, 25(9):1067–1074, 2004.

M. Molinier, J. Laaksonen, and T. Hame. Detecting man-made structures and changes in satellite imagery with a content-based information retrieval system built on self-organizing maps. *Geoscience and Remote Sensing, IEEE Transactions on*, 45(4):861–874, 2007.

D. Montgomery and E. Peck. *Introduction to Linear Regression Analysis*. John Willy and Sons, New York, 1992.

F. Morsdorf, E. Meier, B. Kötz, and K. Itten. Lidar based geometric reconstruction of boreal type forest stands at single tree level for forest and wildland fire management. *Remote Sensing of Environment*, 92:353–362, 2004.

G. Mountrakis, J. Im, and C. Ogole. Support vector machines in remote sensing: A review. *ISPRS Journal of Photogrammetry and Remote Sensing*, 66(3):247–259, 2011.

J. D. Muss, D. J. Mladenoff, and P. A. Townsend. A pseudo-waveform technique to assess forest structure using discrete lidar data. *Remote Sensing of Environment*, 115(4):824–835, 2011.

M. Mutlu, S. C. Popescu, C. Stripling, and T. Spencer. Mapping surface fuel models using lidar and multispectral data fusion for fire behavior. *Remote Sensing of Environment*, 112(1):274–285, 2008.

S. W. Myint, P. Gober, A. Brazel, S. Grossman-Clarke, and Q. Weng. Per-pixel vs. object-based classification of urban land cover extraction using high spatial resolution imagery. *Remote Sensing of Environment*, 115(5):1145 – 1161, 2011.

I. G. Nacional. Memoria de actividades. online, 2009.

E. Naesset, T. Gobakken, J. Holmgren, H. Hyyppa, J. Hyyppa, M. M. M., Nilsson, H. Olsson, A. Persson, and U. Soderman. Laser scanning of forest resources: the Nordic experience. *Scand. J. Forest. Res.*, 19:482–499, 2004.

National Center for Geographic Information and Analysis/University of California. Core curriculum. Introduction to GIS, 1990.

M. Q. Nguyen, P. M. Atkinson, and H. G. Lewis. Superresolution mapping using a hopfield neural network with LIDAR data. *IEEE Geoscience and Remote Sensing Letters*, 2(3):366–370, 2005.

H. Niska, J.-P. Skon, P. Packalen, T. Tokola, M. Maltamo, and M. Kolehmainen. Neural networks for the prediction of species-specific plot volumes using airborne laser scanning and aerial photographs. *Geoscience and Remote Sensing, IEEE Transactions on*, 48(3):1076–1085, 2010.

J. Ojeda Zújar, J. Márquez Pérez, and A. Gómez. Restitución analítica, estereocorrelación y lídar para la generación de modelos digitales de terreno en marismas mareales. *El acceso a la información espacial y nuevas tecnologías geográficas*, XII Congreso Nacional de Tecnologías de la Información Geográfica:1121–1134, 2006.

V. Olaya. *Fundamentos de análisis geográfico con SEXTANTE*. Online, 2006.

T. Oommen, D. Misra, N. K. C. Twarakavi, A. Prakash, B. Sahoo, and S. Bandopadhyay. An objective analysis of support vector machine based classification for remote sensing. *Mathematical Geosciences*, 40:409–424, 2004.

J. Osborne and E. Waters. Four assumptions of multiple regression that researchers should always test. *Practical Assessment, Research & Evaluation*, 8(2), 2002.

J. Otukei and T. Blaschke. Land cover change assessment using decision trees, support vector machines and maximum likelihood classification algorithms. *International Journal of Applied Earth Observation and Geoinformation*, 12, Supplement 1(0):S27 – S31, 2010.

J. A. Parejo, J. García-Gutiérrez, A. Ruiz, and J. C. Riquelme-Santos. Statservices: Providing non-parametric statistical software as web services. *Journal of Statistical Software*, To be submitted, 2012a.

J. A. Parejo, J. García, A. Ruiz-Cortés, and J. C. Riquelme. Statservice: Herramienta de análisis estadístico como soporte para la investigación con metaheurísticas. In *Actas del VIII Congreso Expañol sobre Metaheurísticas, Algoritmos Evolutivos y Bio-inspirados*, 2012b.

C. Pascual, A. García-Abril, L. García-Montero, S. Martín-Fernández, and W. Cohen. Object-based semi-automatic approach for forest structure characterization using lidar data in heterogeneus Pinus Sylvestris stands. *Forest Ecology and Management*, 255:3677–3685, 2008.

S. Pignatti, R. M. Cavalli, V. Cuomo, L. Fusilli, S. Pascucci, M. Poscolieri, and F. Santini. Evaluating Hyperion capability for land cover mapping in a fragmented ecosystem: Pollino National Park, Italy. *Remote Sensing of Environment*, 113:622–634, 2009.

C. Pinilla. *Elementos de teledetección*. RA-MA, 1996.

S. C. Popescu. Estimating biomass of individual pine trees using airborne lidar. *Biomass and Bioenergy*, 31:646 – 655, 2007.

R. J. Quinlan. Learning with continuous classes. In *5th Australian Joint Conference on Artificial Intelligence*, pages 343–348, 1992.

R. J. Quinlan. Improved use of continuous attributes in C4.5. *Journal of Artificial Intelligence Research*, 4:77–90, 1996.

F. Ratle, G. Camps-Valls, and J. Weston. Semisupervised neural networks for efficient hyperspectral image classification. *Geoscience and Remote Sensing, IEEE Transactions on*, 48(5):2271 –2282, 2010.

J. A. Richards and X. Jia. *Remote Sensing Digital Image Analysis: An Introduction*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 3rd edition, 1999.

F. Rosenblatt. *Principles of Neurodynamics*. Spartan, New York, 1962.

C. Salas, L. Ene, T. G. Gregoire, E. Næsset, and T. Gobakken. Modelling tree diameter from airborne laser scanning derived variables: A comparison of spatial statistical models. *Remote Sensing of Environment*, 114(6):1277 – 1285, 2010.

G. E. Schwarz. Estimating the dimension of a model. *Annals of Statistics*, 6(2): 461–464, 1978.

G. Shao and J. Wu. On the accuracy of landscape pattern analysis using remote sensing data. *Landscape Ecology*, 23:505–511, 2008.

S. S. Shapiro and M. B. Wilk. An analysis of variance test for normality (complete samples). *Biometrika*, 52(3-4):591–611, 1965.

S. Shataeea, H. Weinaker, and M. Babanejad. Plot-level forest volume estimation using airborne laser scanner and TM data, comparison of boosting and random forest tree regression algorithms. *Procedia Environmental Sciences*, 7(0):68 – 73, 2011.

S. A. Soenen, D. R. Peddle, and C. A. Coburn. SCS+C: A modified sun-canopy-sensor topographic correction in forested terrain. *IEEE Transactions on Geoscience and Remote Sensing*, 43:2148–2159, 2005.

A. Solberg. Contextual data fusion applied to forest map revision. *Geoscience and Remote Sensing, IEEE Transactions on*, 37(3):1234–1243, 1999.

D. Stojanova, P. Panov, V. Gjorgjioski, A. Kobler, and S. Dzeroski. Estimating vegetation height and canopy cover from remotely sensed data with machine learning. *Ecological Informatics*, 5(4):256 – 266, 2010.

D. Stow, Y. Hamada, L. Coulter, and Z. Anguelova. Monitoring shrubland habitat changes through object-based change identification with airborne multispectral imagery. *Remote Sensing of Environment*, 112:1051–1061, 2008.

L. K. Svancara, J. M. Scott, T. R. Loveland, and A. B. Pidgorna. Assessing the landscape context and conversion risk of protected areas using satellite data products. *Remote Sensing of Environment*, 113:1357–1369, 2009.

B. W. Szuster, Q. Chen, and M. Borger. A comparison of classification techniques to support land cover and land use analysis in tropical coastal zones. *Applied Geography*, 31(2):525 – 532, 2011.

Y. Tarabalka, M. Fauvel, J. Chanussot, and J. Benediktsson. SVM- and MRF-based method for accurate classification of hyperspectral images. *IEEE Geoscience and Remote Sensing Letters*, 7(4):736–740, 2010.

TerraSolid Limited. Terrascan for microstation, user's guide, 2000.

S. Thessler, S. Sesnie, Z. S. R. Bendana, K. Ruokolainen, E. Tomppo, and B. Finegan. Using k-nn and discriminant analyses to classify rain forest types in a landsat tm image over northern costa rica. *Remote Sensing of Environment*, 112:2485–2494, 2008.

X. Tian, Z. Su, E. Chen, Z. Li, C. van der Tol, J. Guo, and Q. He. Estimation of forest above-ground biomass using multi-parameter remote sensing data over a cold and arid area. *International Journal of Applied Earth Observation and Geoinformation*, 14(1):160 – 168, 2012.

T. R. Tooke, N. C. Coops, N. Goodwin, and J. A. Voogt. Extracting urban vegetation characteristics using spectral mixture analysis and decision tree classifications. *Remote Sensing of Environment*, 113:398–407, 2009.

A. B. Torres, R. Marchant, J. C. Lovett, J. C. Smart, and R. Tipper. Analysis of the carbon sequestration costs of afforestation and reforestation agroforestry practices and the use of cost curves to evaluate their potential for implementation of climate change mitigation. *Ecological Economics*, 69(3):469 – 477, 2010.

P. A. Townsend, T. R. Lookingbill, C. C. Kingdon, and R. H. Gardner. Spatial pattern analysis for monitoring protected areas. *Remote Sensing of Environment*, 113:1410–1420, 2009.

D. Tuia and G. Camps-Valls. Semisupervised remote sensing image classification with cluster kernels. *IEEE Geoscience and Remote Sensing Letters*, 6(2):224–228, 2009a.

D. Tuia and G. Camps-Valls. Recent advances in remote sensing image processing. In *Image Processing (ICIP), 2009 16th IEEE International Conference on*, pages 3705–3708, 2009b.

D. Tuia, E. Pasolli, and W. Emery. Using active learning to adapt remote sensing image classifiers. *Remote Sensing of Environment*, 115(9):2232 – 2242, 2011a.

D. Tuia, M. Volpi, L. Copa, M. Kanevski, and J. Muñoz-Mari. A survey of active learning algorithms for supervised remote sensing image classification. *Selected Topics in Signal Processing, IEEE Journal of*, 5(3):606 –617, 2011b.

A. Vasilakos and D. Stathakis. Granular neural networks for land use classification. *Soft Computing*, 9:332–340, 2004.

T. Villmann, E. Merényi, and B. Hammer. Neural maps in remote sensing image analysis. *Neural Networks*, 16(3-4):389 – 403, 2003.

B. Waske and J. Benediktsson. Fusion of support vector machines for classification of multisensor data. *IEEE Transactions on Geoscience and Remote Sensing*, 45(12):3858–3866, 2007.

D. L. Weakliem. A critique of the bayesian information criterion for model selection. *Sociological Methods Research*, 27(3):359–397, 1999.

T. G. Whiteside, G. S. Boggs, and S. W. Maier. Comparing object-based and pixel-based classifications for mapping savannas. *International Journal of Applied Earth Observation and Geoinformation*, 13(6):884 – 893, 2011.

F. Wilcoxon. Individual comparisons by ranking methods. *Biometrics Bulletin*, 1(6):80–83, 1945.

D. H. Wolpert. Stacked generalization. *Neural Networks*, 5:241–259, 1992.

J.-S. Yoon, J.-I. Shin, and K.-S. Lee. Land cover characteristics of airborne LiDAR intensity data: A case study. *Geoscience and Remote Sensing Letters, IEEE*, 5(4):801 –805, 2008.

J. Zar. *Biostatistical Analysis*. Prentice Hall, 1999.

S. Zhang, C. Zhang, and Q. Yang. Data preparation for data mining. *Applied Artificial Intelligence*, 17(5/6):375–381, 2003.

K. Zhao, S. Popescu, X. Meng, Y. Pang, and M. Agca. Characterizing forest canopy structure with lidar composite metrics and machine learning. *Remote Sensing of Environment*, 115(8):1978 – 1996, 2011.

G. Zhu and D. Blumberg. Classification using ASTER data and SVM algorithms. The case study of Beer Sheva, Israel. *Remote Sensing of Environment*, 80(2):233–240, 2002.

X. Zhu and A. B. Goldberg. *Introduction to semi-supervised learning*. Morgan & Claypool Publishers, 2009.