UNIVERSIDAD DE CANTABRIA

ÁREA DE INGENIERIA CARTOGRÁFICA, GEODESIA Y FOTOGRAMETRÍA

Evaluación y propuesta de metodologías de clasificación a partir del procesado combinado de datos LiDAR e imágenes aéreas georreferenciadas

TESIS DOCTORAL

Presentada por:
Mª Pilar Martínez Blanco

Dirigida por:
Dr. JAVIER Mª SÁNCHEZ ESPESO
Dr. AITOR BASTARRIKA IZAGIRRE

Santander, enero 2016

Dedicada a Rafa:

por fin ha visto la luz.

AGRADECIMIENTOS

Con estas líneas quisiera agradecer a todos aquellos que directa o indirectamente han sufrido, de distintas maneras, el largo proceso de aprendizaje que aquí presento.

Especialmente debo expresar **mi más sincera gratitud** a mis directores de tesis y amigos **Dr. Javier Sánchez Espeso**, Profesor Titular de la Universidad de Cantabria y **Dr. Aitor Bastarrika Izagirre**, Profesor Agregado de la Universidad del País Vasco. Gracias a vosotros porque me habéis acompañado con empeño y afecto a lo largo de todo el proceso, no sólo en los momentos buenos, sino si cabe, con más impetu en las situaciones más desoladoras. Con vuestras contribuciones y rectificaciones habéis permitido mejorar significativamente el fruto de esta investigación. Sin vuestra ayuda el resultado hubiera sido otro.

Al Gobierno Vasco, y en especial a Juan Carlos Barroso y Agustín Fernández, pertenecientes al Servicio de Información Territorial de la Dirección de Planificación Territorial y Urbanismo, además de por el servicio de descarga FTP que permite acceder a la información cartográfica, por la atención personalizada, cuando ha sido necesaria, para obtener información no disponible a través de esta fuente.

A la UPV/EHU por las licencias parciales de tesis concebidas en los dos últimos cuatrimestres, liberándome de parte de mis obligaciones docentes en la Escuela Universitaria de Ingeniería de Vitoria-Gasteiz.

Mi más franco agradecimiento a los profesores de la Escuela Universitaria de Ingeniería de Vitoria-Gasteiz: Aitor, Amaia, Leyre, Karmele y Ruper, profesores de la sección de Topografía; y a Ortzi y Fran, profesores del departamento de Expresión Gráfica y Proyectos de Ingeniería. A todos vosotros gracias por la comprensión, ánimo, humor y apoyo que me habeis ido brindando a lo largo de este arduo trabajo.

A toda mi familia, que han constituido un gran soporte. En especial a Sabin y Etxe que no me han dejado sola en ningún momento; a mis padres y hermana, que han podido notar mi falta en momentos importantes; a mis tíos y primos de Bergara que se han preocupado y me han ayudado con mi situación.

A mis amigos, que en los últimos tiempos no me han sentido cerca.

A todos vosotros, GRACIAS. Espero haber estado a la altura de vuestras espectativas.

Vitoria-Gasteiz, enero de 2016

ÍNDICE DE CONTENIDO

ÍNDICE DE FIGURAS

ÍNDICE DE TABLAS

1. PRESENTACIÓN

En las siguientes líneas se ha tratado de describir uno de los últimos avances tecnológicos que ha revolucionado no sólo el mundo cartográfico sino también las ciencias afines de la Tierra, permitiendo su descripción tridimensional de una manera fácil e inteligible.

En ese ámbito, tras la realización de una breve introducción sobre la técnica, se han marcado los objetivos que se persiguen en esta investigación explicando cómo se van a exponer en las siguientes secciones.

1.1. INTRODUCCIÓN

En los últimos años la irrupción de la tecnología LiDAR (*Light Distance And Ranging*) ha revolucionado la captura de datos en el ámbito cartográfico, pasando de usar metodologías eminentemente discretas, observándose las posiciones estrictas y necesarias de los objetos del mundo real que se quieren capturar, a una nueva totalmente masiva. La representación de la superficie terrestre por medio de nube de puntos ha permitido llegar a cotas insospechadas, de forma que resulta común el uso de esta materialización para el dibujo de carreteras, túneles, edificios, etc. a través de sistemas de escaneado móviles (*Mobil Laser Scanning*, MLS) (Ussyshkin 2009).

Estos sistemas MLS se basan en la tecnología desarrollada por los escáner láser aerotransportados (*Airbone Laser Scanning*, ALS) también denominados *Airbone Laser Terrain Mapper* (ALTM). Dentro de este conjunto de instrumentos destaca el denominado LiDAR, el cual permite la adquisición de datos altimétricos de la superficie terrestre, constituyendo las llamadas nubes de puntos tridimensionales (figura 1.1).

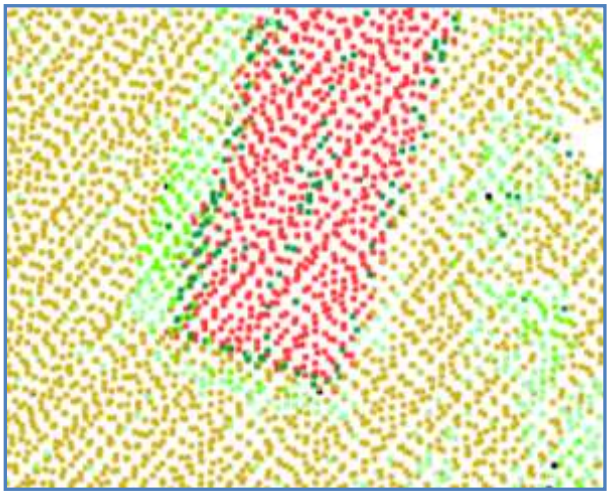

Figura 1.1. Nube de puntos tridimensionales

El LiDAR adquiere datos espaciales de todos elementos de la superficie de trabajo, siguiendo patrones regulares sin hacer diferencia alguna en los distintos objetos para los que se determina su posición, capturándose conjuntamente puntos sobre la superficie en una carretera o en el tejado de un edificio, tomando junto a ellos información sobre objetos no relevantes desde un punto de vista cartográfico como pueden ser, a modo de ejemplo, la sombrilla en una playa o un vehículo, en nubes de puntos de millones de datos.

Se trata de una tecnología (figura 1.2) que usa un escaneado láser en vuelo desde una posición y dirección conocida, gracias al uso de sistemas integrados GPS (*Global Positioning Systems*) o GNSS (*Global Navigation Satellite* Systems) e INS (*Inertial Navigation System*), midiendo la distancia relativa existente desde el punto de emisión al punto de recepción, así como la intensidad del pulso láser recibido; consiguiendo, de esta forma, un gran volumen de puntos (entorno a 150.000 pulsos por segundo) para la representación de los objetos escaneados.

El sensor utilizado para este fin suele ser un sensor de barrido que va acompañado de un espejo rotatorio, éste hace que los rayos láser se dirijan a ambos lados del avión en función del ángulo de apertura especificado, obteniendo así un barrido de una franja del terreno, que será mayor o menor en función de la altura de vuelo, a medida que avanza el avión o helicóptero. Para cubrir una gran zona con puntos láser es necesario establecer líneas de vuelo paralelas guiadas con GPS, de manera que se pueda asegurar la cobertura de toda la zona utilizando corredores estrechos, que a su vez permitan el solapamiento de datos a lo largo de los ejes, evitando áreas sin datos (Harding 2000).

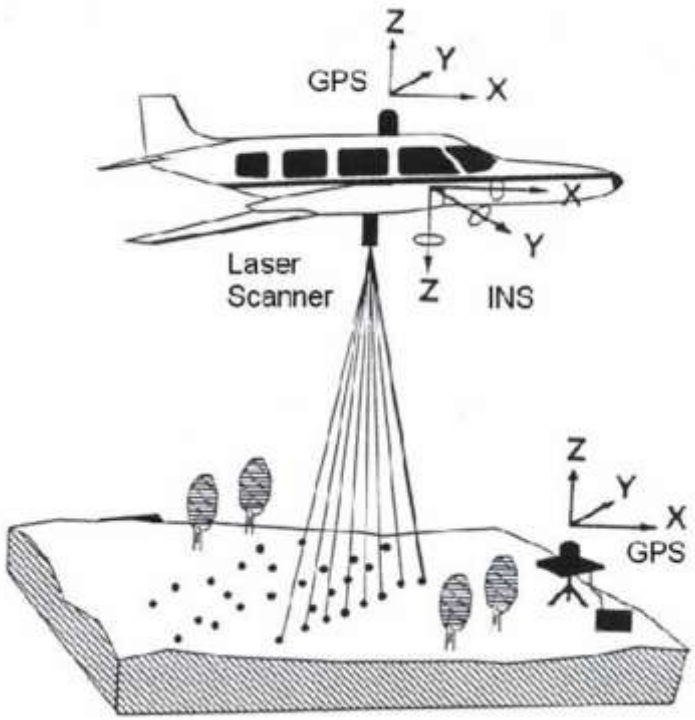

Figura 1.2. Esquema de captura de datos LiDAR (OSGeoLive)

El uso de GPS/INS permite georreferenciar en el sistema de referencia y la proyección indicada cada uno de esos puntos reflejados. Siendo habitual el uso de sistemas cartesianos que admiten su representación mediante los valores *X*, *Y*, *Z* (Shan and Toth 2008). El GPS ofrece la posición de la plataforma de acuerdo a un post-procesamiento que se ha venido llamando procesamiento de bases cortas, en el que la distancia entre el GPS del avión y el de referencia de Tierra no puede ser mayor a 30 km. El sistema inercial (INS) dispone la actitud del instrumento indicando los ángulos de orientación (*roll* → balanceo, *pitch* → inclinación y *heading* → cabeceo), así como la dirección del pulso láser al medir la orientación del espejo del escáner en todo momento.

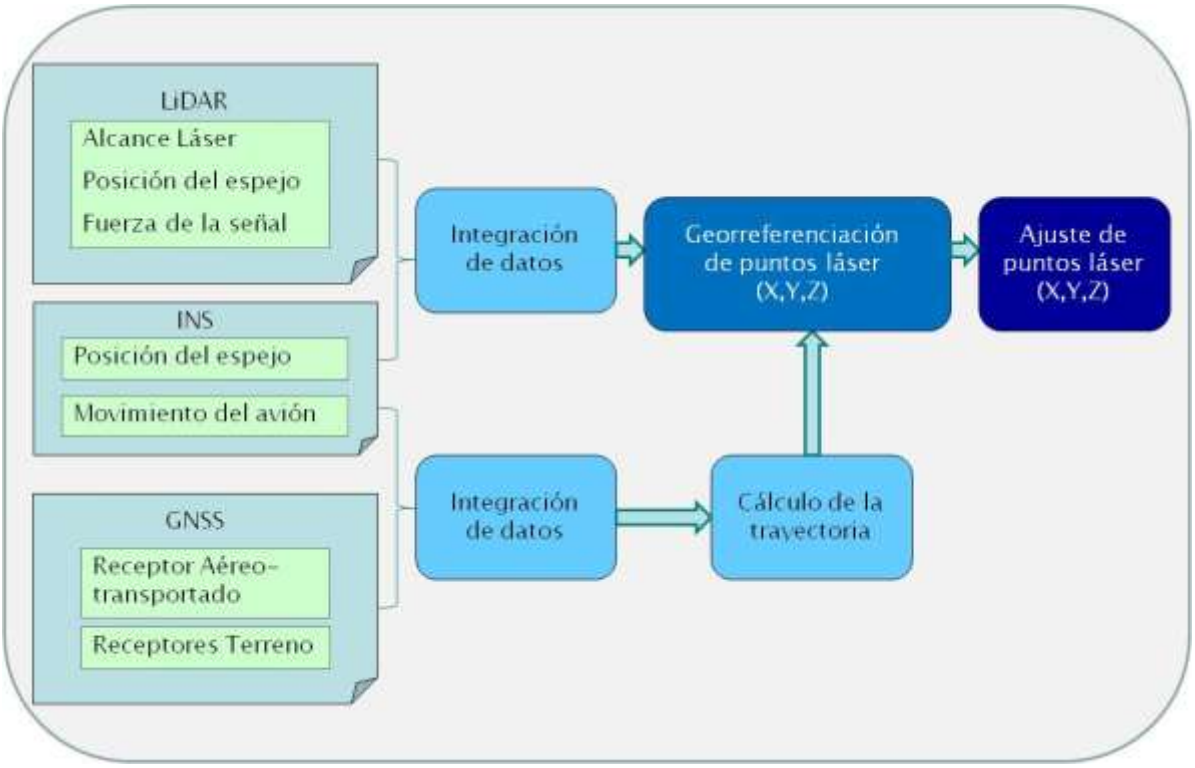

Figura 1.3. Esquema de georreferenciación de puntos láser

Calculada la trayectoria GPS/INS mediante la integración de ambos tipos de datos utilizando para ello datos filtrados y considerando los desplazamientos relativos entre los diversos sistemas, se decodifican los datos tomados por el láser y se ajustan en altura según las zonas de test, resultando finalmente determinados una vez consideradas las diferencias en alturas entre pasada y con respecto a las zonas de control, consiguiendo de esta forma que todos los retornos láser queden perfectamente georreferenciados (figura 1.3).

Así, tras ese proceso en el que se calcula la trayectoria GPS/INS y se determinan y ajustan los puntos láser, según ICC 2005 también hay que verificar, mediante el análisis visual de los datos, que la zona cubierta se corresponde con el área que se pretendía volar, para posteriormente editar los puntos láser y generar los productos correspondientes (figura 1.4).

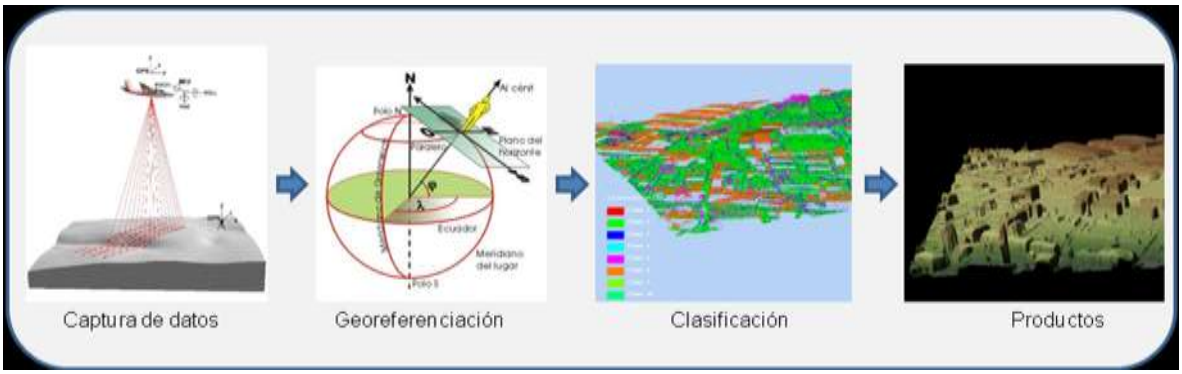

Figura 1.4. Esquema de las fases a desarrollar en el procesado de datos LiDAR
(Captura de datos: Usda; Georeferenciación: Puente)

La edición de puntos láser es lo que se conoce como clasificación y suele estar compuesta por dos etapas: en la primera, se realiza una clasificación automática; en la segunda, se realiza la edición manual reclasificando aquellos puntos que han resultado erróneos en la primera fase, también se suelen introducir las líneas de rotura o puntos de cota en aquellas zonas que se considere necesario, aunque a veces esto se realiza en una fase posterior.

Los parámetros que hacen referencia a las características de la captura de los puntos y su posterior procesado tras el vuelo constituyen los atributos que suelen venir debidamente recogidos en su base de datos correspondiente, donde además de las coordenadas (*X*, *Y*, *Z*) y el valor de clasificación se recogen otros valores tales como el ángulo y dirección de escaneo, el eje de la línea de vuelo y el tiempo GPS, entre otros (apartado 3.1). Si se trata de sistemas LiDAR de múltiples retornos, para un mismo punto (igual *X*, *Y*) pueden existir diferentes valores de altitud (*Z*), tantos como retornos haya para ese haz (figura 1.5).

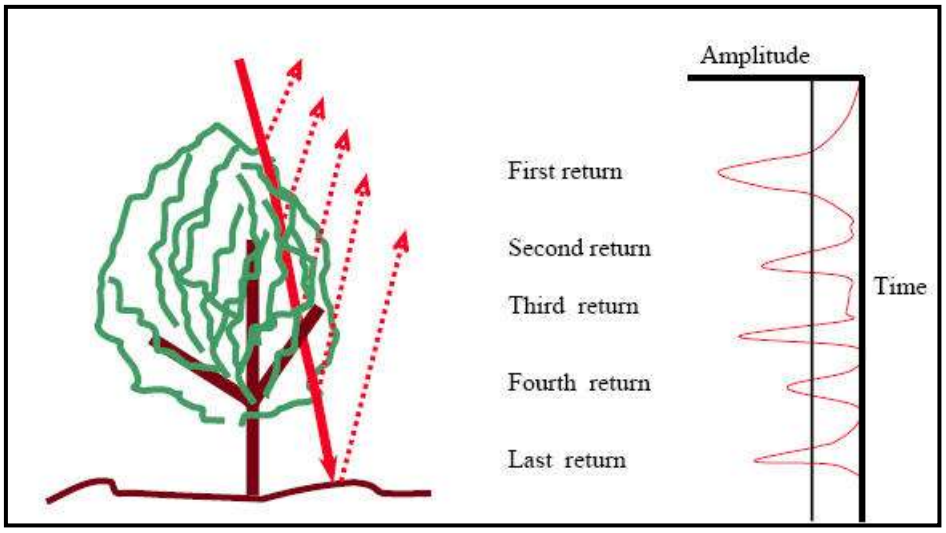

Figura 1.5. Retornos de un mismo punto (Lohani)

Como este sensor capta para cada punto el valor de la cantidad de energía electromagnética reflejada, se almacena también el valor de la intensidad (*i*) que le corresponde. Esta información permite generar imágenes de la superficie en tonalidades grises, denominadas imágenes de intensidad (figura 1.6).

Dada la precisión altimétrica y el volumen de datos que se consigue con esta tecnología, está tendiendo a sustituir a los métodos fotogramétricos clásicos cuando el objetivo es la obtención de modelos de superficies, en los que resulta necesaria la realización de un vuelo para obtener las imágenes aéreas a partir de las cuales, tras un proceso previo, conseguir los datos cartográficos buscados; en este caso, las altitudes que luego permitan generar los productos demandados.

Figura 1.6. Nube de puntos simbolizado según su valor de intensidad

En el ámbito cartográfico las reacciones ante este nuevo producto cartográfico están generando sistemáticamente dos nuevas situaciones:

- Expectación ante un nuevo modelo de datos con una densidad superficial inimaginable hasta hace poco tiempo, apareciendo continuamente nuevos ámbitos en los que esta tecnología puede ser utilizada.

- Y a la vez, de forma pareja, la imperiosa necesidad de herramientas software y de equipos hardware cada día más potentes capaces de aprovechar toda la potencialidad de la información contenida en inmensas nubes de puntos.

De hecho, hoy por hoy, la mayoría de los softwares de gestión de datos espaciales y de tratamiento de imágenes son capaces de entender este tipo de formación, e incorporarlo como otra fuente de datos en su flujo de trabajo. Pero la única posibilidad operativa de gestionar esta información es que los datos LiDAR dispongan de información añadida que permitan su reorganización, selección y/o estructuración, resultando absolutamente imprescindible que cada punto de la nube LiDAR cuente con un atributo que le asigne a una cierta categoría, en concordancia con la finalidad perseguida. La clasificación más básica distingue dos grandes categorías: puntos pertenecientes al suelo y puntos no pertenecientes a él, que permitiría la creación de Modelos Digitales del Terreno (MDT), un producto cartográfico con numerosas aplicaciones en Ingeniería y otras ciencias de la Tierra.

Con esta nueva técnica, una vez realizado el vuelo y tras el proceso de ajuste de los datos láser, al disponer de altitudes se pueden generar los Modelos Digitales de Elevaciones (MDE) , si se

dispone de la clasificación mencionada. Para la creación de estos modelos se deben seguir unos pasos, que tal y como se indica en Meng, et al. 2010 vienen a ser:

1. Eliminación de errores: se trata de borrar los puntos que hagan referencia a datos anómalos, debidos a pájaros, aviones, rebotes del propio sensor, etc.
2. Interpolación, reorganización o remuestreo: en el caso de que se vaya a generar la malla o *grid*. Por lo general, el tamaño de celda suele ser de 1 ó 2 m.
3. Filtrado de puntos: separación de los puntos en terreno y no terreno.
4. Generación del MDE: se genera la superficie a partir de los puntos clasificados en el paso anterior, considerando los puntos terreno si se quiere generar el Modelo Digital del Terreno (MDT) o los puntos no terreno para obtener el Modelo Digital de Superficies (MDS).

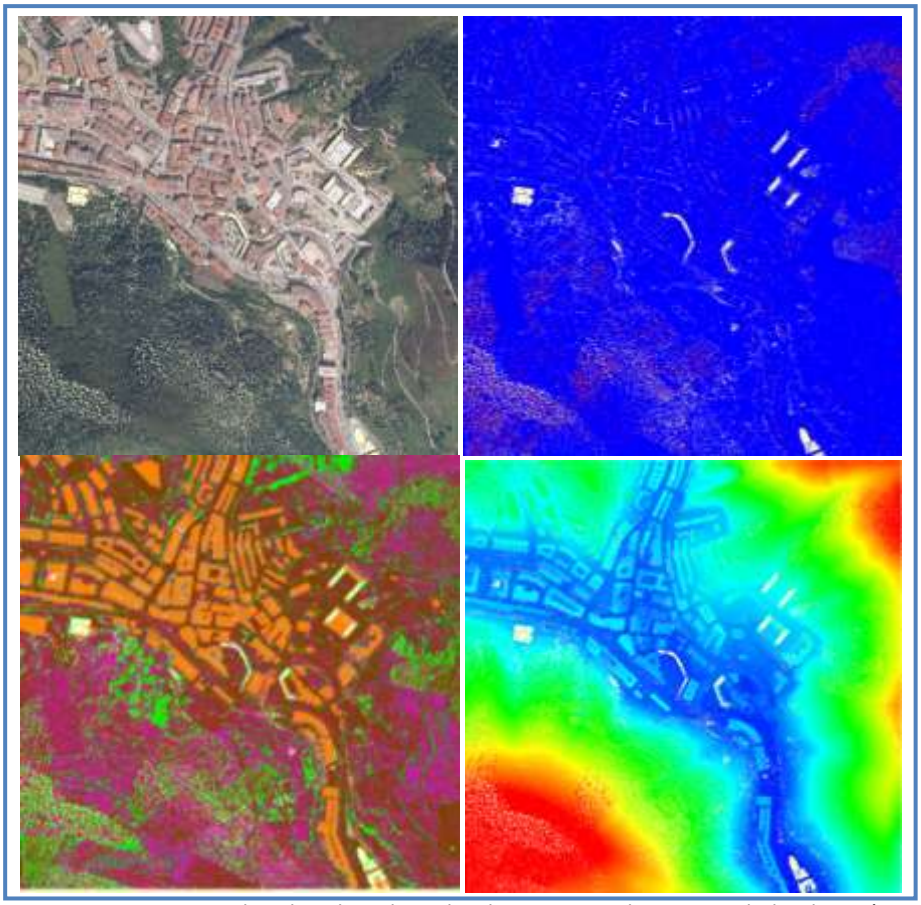

Figura 1.7. De izquierda a derecha y de arriba abajo: imagen de LiDAR simbolizada según Ortofotografía RGB, número de retornos, clasificación y altitud

Este proceso comparado con el que se debe seguir con los métodos fotogramétricos presenta la ventaja de que directamente ya se obtiene el dato buscado - la altitud - por lo que se produce una disminución considerable no sólo en los costes de ejecución sino también en las tareas a desarrollar. Además, frente a los métodos fotogramétricos (discretos) brindan mayor volumen de datos, ya que se trata de sistemas cuya tecnología aporta una toma masiva de

puntos, no sólo de la información requerida sino de toda aquella que se ubica en sus alrededores, permitiendo captar información del terreno en zonas cubiertas como puede ser el caso de las áreas boscosas. También tiene menor dependencia de las condiciones climáticas por usar un sensor activo que no necesita de la luz solar, abriendo las posibilidades a volar de noche o con nubes, ofreciendo la posibilidad de volar más días al año.

El volumen de datos capturados justifica la necesidad de algoritmos de clasificación lo más automáticos posibles, justificando el interés del problema como constata elevado número de algoritmos propuestos con esta finalidad en revistas y congresos en la última década. Los más básicos, buscan el filtrado de los datos para detectar únicamente aquellos pertenecientes al suelo (Jubanski 2010). Otros algoritmos se centran en la detección de edificios (Niemeyer, et al. 2014), vías de comunicación (Ural, et al. 2015);o vegetación (Latifi, et al. 2012), entre otros muchos. Sin embargo, la inmensa mayoría de los algoritmos presentes en la literatura, en mayor o menor medida, comparten en el análisis de la bondad de la clasificación efectuada dos aspectos:

- Son eficientes para una nube de puntos en un cierto tipo de terreno, y los mayores aciertos se producen para un cierto tipo de punto clasificado. La complejidad del mundo real, el propio proceso de captura de los datos y las características de los datos LiDAR dificultan la existencia de algoritmos únicos eficientes en cualquier condición.

- Los diferentes enfoques planteados, desde campos muy diversos, parecen haber llegado al límite de las posibilidades de clasificación usando exclusivamente los datos LiDAR, y existe un cierto consenso en emplear otras fuentes de información complementarias para ayudar a este objetivo (Gajski, et al. 2003; Gerke and Xiao 2014; Zhang, et al. 2003). En particular, el uso de imágenes como complemento en el proceso de clasificación.

En esta investigación se plantea proponer una metodología de clasificación de los puntos LiDAR que ahonde en el último planteamiento, combinando estos datos con datos de imágenes aéreas georreferenciadas ortorectificadas. La propuesta inicial pretende trabajar el dato original formado por puntos, sin la transformación en información ráster como fase intermedia del análisis.

1.2. OBJETIVOS Y ORGANIZACIÓN DE LA TESIS DOCTORAL

En este apartado se enumeran los propósitos que se pretenden alcanzar en esta tesis doctoral y se explica brevemente cómo se ha organizado el documento a lo largo de los distintos capítulos.

1.2.1. OBJETIVOS

El objetivo principal de esta investigación se centra en **establecer una metodología que permita la clasificación de las nubes de puntos LiDAR en categorías adecuadas al uso cartográfico**.

Para conseguir este objetivo se hace uso del conjunto de información que año tras año se lleva a cabo de manera cofinanciada y cooperativa entre la Administración General del Estado (AGE) y las comunidades autónomas. Concretamente se han utilizado imágenes aéreas de alta resolución georreferenciadas y datos LiDAR adquiridos en el marco del Plan Nacional de Ortofotografía Aérea (PNOA) y la Base Topográfica Armonizada (BTA), datos todos ellos que se encuentran a disposición de las distintas administraciones del estado.

El enfoque que se le quiere dar se basa, en la medida de lo posible, en considerar los puntos de los datos LiDAR como entidad básica de trabajo, tratando de evitar el trabajo único y exclusivamente con información rasterizada.

La consecución de estos objetivos principales se ha llevado a cabo mediante los siguientes objetivos específicos:

- **Desarrollar una metodología lo más automática posible, basada en la combinación de imágenes de alta resolución y datos LiDAR**. Esta técnica se desarrolla y aplica a la Comunidad Autónoma del País Vasco (CAPV) pero debería ser extrapolable y fácilmente aplicable a otros entornos.

- **Verificar y comprobar si existen relaciones** y en ese caso indicar de qué tipo son, **entre los valores de clasificación automática adquiridos y algunos de los datos aportados en la base de datos** correspondientes a la información de la nube de puntos, tales como el valor de intensidad, el identificador de pasada o el número de retorno, entre otros.

- **Elaborar procedimientos que consideren el punto de la nube como unidad de trabajo**, evitando el trabajo generalizado con información rasterizada derivada de los datos LiDAR.

- **Establecer técnicas para la obtención de las variables predictivas necesarias para las diferentes metodologías a desarrollar**. Partiendo de los datos originales resulta

necesario analizar y valorar las variables predictivas derivadas combinando la información derivada del LiDAR como de las ortofotografías para la categorización de los puntos LiDAR

- **Evaluar la aportación de las imágenes aéreas en la mejora de la clasificación automática de los datos LiDAR**, permitiendo concluir si realmente el uso de esta información adicional constituye un elemento clave en la mejora de la clasificación.

1.2.2. ORGANIZACIÓN DE LA TESIS DOCTORAL

Esta memoria se ha organizado en nueve capítulos que permiten sintetizar los ensayos efectuados y resultados conseguidos en el desarrollo de este trabajo de investigación.

En el primer capítulo se hace una introducción y contextualización del trabajo en el proceso cartográfico actual y los objetivos generales y parciales establecidos.

En el punto segundo se muestra la revisión bibliográfica realizada sobre los métodos propuestos en la literatura científica para la clasificación de los puntos LiDAR; así como, una introducción a la minería de datos. Antes de continuar con la descripción de la metodología del trabajo de investigación realizado, como tercer apartado se presenta el formato LAS (*Lidar Data Exchange Format*) en el que se distribuyen de manera estándar los datos LiDAR y se hace una pequeña reseña a los distintos programas existentes que permiten su clasificación. También se incluyen las herramientas y librerías utilizadas para desarrollar la metodología propuesta.

Seguidamente, en el punto cuarto tras describir las características de la Comunidad Autónoma del País Vasco, se especifican los datos utilizados y las áreas de estudio. En el apartado quinto se evalúa la clasificación que ya disponen los datos LiDAR utilizados en este trabajo y en el sexto se presenta la propuesta metodológica para la clasificación de esa información haciendo uso de las técnicas de aprendizaje automático.

El análisis de los resultados alcanzados se muestra en el punto séptimo, donde se presenta las decisiones tomadas que han permitido la reclasificación de las categorías a clasificar, y la aportación de los distintos grupos de variables (derivadas de los puntos LiDAR, de las Ortofotografías y de manera combinada). Se muestran los resultados obtenidos en todo el conjunto de validación por categorías.

En el epígrafe ocho se muestran las conclusiones y las futuras líneas de investigación planteadas. Terminando en el apartado noveno con las referencias bibliográficas que han sustentado y contextualizado este trabajo de investigación.

2. ESTADO DEL ARTE

Dos son los aspectos que se van a tratar en este epígrafe, por un lado los métodos de clasificación de puntos LiDAR y por otro la introducción a la minería de datos.

En lo que respecta a los métodos existentes para llevar a cabo la clasificación de los punto LiDAR, se realiza una revisión bibliográfica de los algoritmos más importantes estableciendo una configuración generalista en la que poder incluir cualquiera de ellos. Este apartado se verá complementado con el siguiente punto en el que se habla de las herramientas informáticas utilizadas y en él se han relacionado los programas comentados para el manejo de este tipo de datos con los algoritmos aquí explicados.

En la segunda parte de este punto se hace una pequeña introducción a la minería de datos, realizando una pequeña exposición de los aspectos básicos de esta disciplina, para comentar brevemente algunos algoritmos basados en los árboles de clasificación.

2.1. MÉTODOS DE CLASIFICACIÓN DE PUNTOS LIDAR

A continuación se procede a presentar los distintos algoritmos que se han usado para proceder a clasificar automáticamente los puntos LiDAR. En primer lugar, indicar que se emplean distintos términos para referirse a este proceso, destacando los siguientes:

- ✓ LiDAR point cloud classification
- ✓ LiDAR classification algorithm
- ✓ LiDAR filtering algorithm

Tal y como se puede observar en esos vocablos se utilizan tanto el término de clasificación como el de filtrado y éstos pueden llegar a confundirse con el significado que tienen el término de clasificación en el ámbito de la teledetección.

A partir del análisis que se ha realizado, se ha llegado a la conclusión de que ambos son correctos pero con el matiz de que el filtrado lo que busca es una selección de unos puntos frente al resto (Li 2013) y la clasificación la asignación de una categoría una vez los puntos hayan sido filtrados. Así, Axelsson 1999 indica que el filtrado supone la separación de los datos mezclados y la clasificación encontrar una geometría específica o una estructura estadística que en este caso se corresponde con puntos que pertenezcan a categorías cartográficas como suelo, edificios, vegetación, etc.

Ambos términos, filtrado y clasificación, se incluyen dentro de la fase de edición a desempeñar una vez realizada la captura de datos; luego, para ello primero hay que filtrar, en función de unos parámetros que varían según los algoritmos a emplear, y luego clasificar.

Aclarado el concepto de los dos términos usados con más frecuencia, a veces mezclándose su significado conforme a las acepciones indicadas, hay que tener en cuenta los aspectos que afectan considerablemente a los algoritmos de filtrado y clasificación de puntos que según Axelsson 1999 son:

- La densidad de puntos, que a su vez dependen de la altura y la velocidad del vuelo, el ángulo de escaneo (*Field of View*, FOV) y la frecuencia de muestreo.
- El registro de múltiples ecos o uno único.
- El valor de la amplitud o reflectancia para obtener información radiométrica sobre la superficie.

Estos factores afectan a los resultados de los distintos algoritmos, por eso a la hora de compararlos, en lugar de tener en cuenta el tipo de terreno considerado, que es lo habitual, se debe tener más presente cualquier pequeña variación en los aspectos señalados por Axelsson, debiendo comparar vuelos de igual densidad, con registros únicos o múltiples y con la misma calibración de los valores de reflectancia.

Esto queda de manifiesto en Sithole and Vosselman 2003 donde se estudia el comportamiento de ocho algoritmos distintos en ocho grupos de datos de los que extraen quince subconjuntos representativos de diferentes entornos (pendientes pronunciadas, discontinuidades, puentes,

escenas complejas, valores atípicos, vegetación en pendiente y suelo) para realizar un análisis cuantitativo. De esta manera, todos los algoritmos se analizan con datos que tienen las mismas características en cuanto a la captura de los datos LiDAR se refiere; así, bajo las mismas características se analizan el comportamiento de los algoritmos en los distintos entornos geograficos.

Dentro del estudio que realizan, al igual que lo hace Gajski, et al. 2003, filtran los puntos en dos grupos: los que se corresponde con suelo desnudo y los que hacen referencia a objetos. Esta caracterización permite la identificación de la información geográfica que como señala Graham 2012 viene a ser el proceso de asignación numérica en función del tipo de objeto y de acuerdo con la información reflejada por él, determinando la clase que le corresponde a cada uno.

La clasificación en suelo desnudo y objetos se produce porque el uso primordial que se le da a este tipo de datos es la generación de Modelos Digitales de Elevaciones (MDE). Lo cual sería correcto si el objetivo fuera generar modelos que representen la elevación a la que se encuentra el suelo, denominados errónea y comúnmente Modelos Digitales del Terreno (MDT) (Cuartero Sáez 2008; Felicísimo 1999), pero no sí lo que se pretende obtener son Modelos Digitales de Superficies (MDS), para los que sería necesario el uso de algunas otras clases que considera la ASPRS, tales como las de vegetación y edificios (ver capítulo tercero). En esa línea destacar el hecho de que algunos artículos como (Zhang and Lin 2012) ya establecen puntos diferenciados en tres categorías: terreno, vegetación y edificación.

Asumiendo que el uso generalizado de estos datos es la creación de MDT a nivel de suelo o de MDS, es de entender que la inmensa mayoría de los algoritmos encontrados traten de establecer qué puntos pertenecen al terreno y cuáles no (*ground* / *non-ground*), admitiendo de forma habitual que la superficie del terreno verifica tres premisas básicas, que servirán al algoritmo propuesto para clasificar en las dos categorías indicadas según las apreciaciones de Pfeifer 2008:

- el terreno es continuo y suave.
- en el terreno no existen saltos en altura.
- no existen puntos por debajo del terreno.

Teniendo en cuenta las consideraciones anteriores, se puede diferenciar entre algoritmos que usan directamente los datos tal y como se capturan (brutos), mientras que otros lo hacen una vez transformados a una imagen (malla) (Gajski, et al. 2003). Luego, aquí se crea la primera categoría a la hora de establecer una clasificación de los algoritmos utilizados.

Según la bibliografía el proceso a seguir es diferente si se trabaja con datos brutos o malla (rasterizados). En el caso de que se trabaje con datos brutos, el proceso de análisis suelo comprender típicamente tres fases (Graham 2012):

1. Detección de puntos bajos: estos puntos suelen estar por debajo del suelo, por lo que se categorizan como puntos erróneos. Son debidos al *multipath* de la señal GPS, anomalías de la intensidad o a fallos imprevistos.

2. Detección de puntos de terreno: según los distintos algoritmos.
3. Detección de ruido: se trata de puntos de vegetación baja que han sido considerados como terreno.

En el caso de trabajar con datos LiDAR transformados a malla regular las fases serían las siguientes (Pfeifer 2008):

1. Definición de la malla regular: ancho de malla.
2. Detección del punto más bajo en cada celda.
3. Cálculo del MDE por interpolación.

Aunque es de suponer que en este segundo caso, antes de generar la malla, como mínimo también habrá que eliminar los puntos erróneos comentados en el caso de los datos brutos en el paso 1 y de alguna manera el ruido de la fase 3.

Además, no todos los algoritmos consideran todos los pulsos registrados. La mayoría de ellos únicamente consideran el último eco, aunque existen algoritmos que consideran el primero y último; y en el caso de estudios forestales, también se suelen considerar distintos retornos a éstos o incluso todos los retornos existentes (Yunfei, et al. 2008).

Respecto a la naturaleza del algoritmo de filtrado y clasificación, es muy frecuente que la identificación se realice de forma iterativa, si bien también existen algoritmos de un único paso.

Por último, y con objeto de mejorar los resultados de la clasificación, en numerosos estudios se indica la necesidad del uso de información adicional, tal como imágenes aéreas, MDE, mapas temáticos (catastrales, usos del suelo), o imágenes multiespectrales (Brovelli and Lucca 2011).

Como resumen, se puede afirmar que gran parte de los algoritmos trabajan con datos rasterizados, principalmente con el primer pulso, determinando si los puntos pertenecen o no al suelo mediante un proceso iterativo, encontrándose en desarrollo la integración de información adicional. Todas las consideraciones anteriores permitirían una primera clasificación de los filtros según el aspecto considerado, mostrándolos de manera resumida en el esquema de la figura 2.1.

Pero a la hora de diseñar esos algoritmos la mayor parte de ellos, además de esos aspectos, tienen en cuenta la relación geométrica que se da entre los distintos puntos, en función de lo que aparecen agrupados en cuatro familias: morfológicos, densificación progresiva, basados en superficies y basados en segmentación.

En el siguiente apartado se procederá a presentar una revisión de los algoritmos de filtrado y clasificación de puntos de nubes LiDAR lo más extensa y concreta posible.

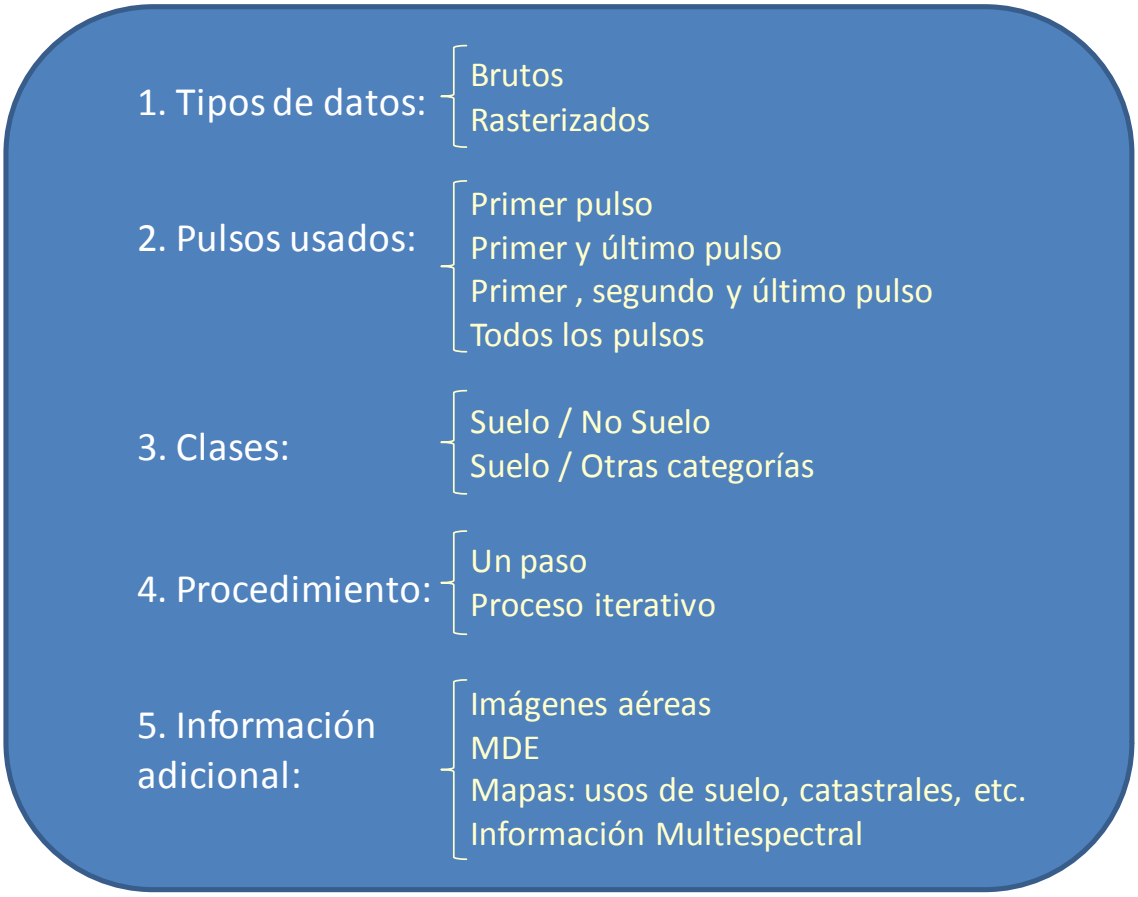

Figura 2.1. Clasificación de los algoritmos de filtrado según diversos criterios

2.1.1. CLASIFICACIÓN DE ALGORITMOS DE FILTRADO

Teniendo en cuenta que la mayoría de los filtros encontrados tratan de discriminar en el total de la nube de puntos los que hacen referencia al suelo de los que no, es habitual clasificarlos en Kobler, et al. 2007; Pfeifer 2008; Shan and Toth 2008; Vosselman and Maas 2010:

- Filtros Morfológicos
- Filtros de Densificación Progresiva
- Filtros basados en Superficies
- Filtros basados en Segmentación

Si el objetivo es conseguir una clasificación más amplia, buscando otros objetivos a los habituales en el ámbito cartográfico habitual como pueden ser las necesidades forestales para la diferenciación de distintos tipos de vegetación o la generación de modelos 3D para las ciudades, es preciso indicar la existencia de nuevos algoritmos que se basan en la intensidad, los índices de vegetación o los algoritmos de aprendizaje automático (Niemeyer, et al. 2013).

Indicar también que en muchas referencias, tal y como Lindeman 2012; Sithole 2005 se diferencian entre métodos morfológicos, diferencias de elevación y los basados en la pendiente, pero tras un estudio exhaustivo de los distintos filtros se ha considerado que tanto los métodos basados en las diferencias de elevación como los basados en la pendiente se encuentran dentro de los morfológicos y así queda recogido en Shan and Toth 2008 donde aparecen como extensión y variantes de los morfológicos.

A continuación se ofrece detalla el planeamiento de los cuatros tipos de filtros, así como de otros algoritmos pertenecientes a dichas clasificaciones.

2.1.1.1. Filtros Morfológicos

Estos filtros se basan en la teoría y técnica denominada morfología matemática (MM). Se trata de un método que permite analizar imágenes que ofrecen una descripción cuantitativa de las estructuras geométricas basándose en un conjunto de operadores (Vosselman and Maas 2010).

Los operadores básicos en el ámbito de la MM son los de erosión y dilatación. Éstos se encargan de simplificar la superficie basándose en ciertos elementos estructurales (ventanas) haciendo uso de la combinación de los operadores básicos: *closing = erosion after dilation; and, opening = dilation after erosion* (Ibidem). La dilatación constituye el valor máximo, mientras que la erosión el valor mínimo para realizar la búsqueda, dentro de una ventana determinada, de los valores de píxeles que en este caso constituyen los valores máximos y mínimos de la altura.

Uno de los algoritmos más utilizados en este ámbito ha sido el de Vosselman 2000 que trata de valorar la diferencia de alturas a una determinada distancia, describiendo esas diferencias como admisibles o no, en función de una distancia horizontal considerada como operador de erosión (Shan and Toth 2008). En la práctica esa distancia máxima se identifica con un círculo de radio *R*.

Para poder aplicar este tipo de filtros resulta necesario tener en cuenta los puntos vecinos y la distancia máxima de búsqueda, de manera que las pequeñas diferencias dan como resultado puntos del terreno (suelo) y las grandes serán obviadas (no suelo).

A la hora de determinar la distancia máxima se puede tener en cuenta la pendiente del terreno, motivo por el cual en los filtros morfológicos se puede observar que dentro de ellos existen algunos basados en las pendientes (Sithole 2002).

Según Montealegre, et al. 2013 este tipo de filtros no suelen ser los que mejores resultados ofrecen y en parte se deben a la dificultad que tienen para eliminar los objetos más pequeños en relación al tamaño de la ventana de búsqueda cuando ésta es fija, por lo que se han desarrollado alternativas que contempla el cambio de tamaño de la misma, de ahí el nombre

de filtros morfológicos progresivos, *Progressive Morphological Filter* (PMF) implementado por Zhang, et al. 2003.

Pingel, et al. 2013 pretenden resolver la clasificación utilizando el algoritmo *Simple MoRphological Filter* (SMR) que se basa en la aplicación de técnicas de procesamiento de imágenes en base a una ventana creciente y un umbral en función de la pendiente del terreno, similar a lo que hace Zhang, et al. 2003.

El algoritmo PMF varia iterativamente los tamaños de las ventanas de búsqueda, pero para evitar que la superficie de filtrado se encuentre por debajo de los datos iniciales incorpora un umbral que marca la diferencia de altura entre el terreno y los objetos de la superficie. Este umbral suele ser función del tamaño de la ventana considerada.

Tabla 2.1. Atributos del algoritmo PMF en SPDlib con sus valores por defecto

Parámetros	Defecto	Breve definición del parámetro
--medianfilter	2 (5 × 5)	Tamaño del filtro medio
--grd	0,3	Distancia entre puntos para ser clasificados terreno
--maxelev	5	Diferencia de elevación máxima
--initelev	0,3	Diferencia de elevación inicial
--slope	0,3	Pendiente del terreno
--maxfilter	7	Máximo tamaño del filtro
--initfilter	1 (3 × 3)	Tamaño inicial del filtro
--overlap	10	Tamaño de superposición entre bloques de procesamiento
-b	0	Unidad de trabajo: por defecto usa el del fichero SPD
-c	0	Número de columnas del tile
-r	100	Número de filas del tile

En este trabajo se ha usado el algoritmo PMF en el apartado seis a través de la librería SPDlib (se comenta en el tercer epígrafe) debiendo definir los atributos indicados en la tabla 2.1. La manera de proceder para clasificar los puntos en terreno / no terreno consiste en:

1. Teniendo en cuenta la resolución espacial del fichero de entrada, considera un número mínimo de retornos según el valor de *--initfilter* a partir de los cuales establece el valor mínimo de altura para cada celda y genera la superficie de elevación mínima.

2. El umbral para la diferencia de alturas se establece al inicio con el parámetro *--initelev* finalizando con *--maxelev*, los cuales van variando en función del parámetro de la pendiente (*--slope*).

3. Iterativamente, el filtro inicial va aumentando hasta el valor de *--maxfilter* incrementando el rango de escala. Para cada valor de escala se determina una operación morfológica de erosión / dilatación que permite establecer un nuevo valor de altura. Si éste está por encima del umbral de la diferencia de alturas establecida se mantendrá, en caso contrario se considerará la anterior. Terminadas las iteraciones se dispondrá de una superficie final ráster a la que se le aplica el filtro de la mediana. El

tamaño de este filtro se puede especificar con el parámetro *--medianfilter* y puede no usarse si se utiliza el parámetro *--nomedian* (Bunting 2013).

4. Para clasificar los retornos se usa un buffer de manera que todos los puntos que estén dentro o debajo de esa superficie se clasifican como terreno. El resto se consideran no terreno.

Pero existen otros filtros como el desarrollado por Chen, et al. 2007 que no utilizan ventanas crecientes ni umbrales. Éste se basa en la adaptación de polinomios (*Adpation of polynomial fitting filters*) y parece ofrecer buenos resultados en la mayoría de los casos, incluso en pendientes abruptas, con la definición únicamente de dos parámetros: distancia mínima y distancia máxima.

Señalar que el origen de este tipo de filtros radica en el uso de imágenes en blanco y negro o en tonos de grises (Kilian et al, 1996), motivo por el cual antes de procesar los puntos de las nubes de datos es necesario su transformación a una imagen binaria (ráster). En los últimos tiempos su aplicación se ha incorporado tanto al ámbito de las imágenes de color como al análisis directamente de los puntos en sí.

Tabla 2.2. Filtros morfológicos para la clasificación de puntos LiDAR

Filtros morfológicos	Autor(es)	Año
Based on profiles	J. Lindenberger	1993
Multiple structure elements	Kilian et al	1996
Hierarchical surface regularization	K.Kraus, N. Pfeifer, C. Briese	1998
Dual Rank Filter	Lohmann et al	2000
***Maximum Local Slope* (MLS)**	Vosselman	2000
***Adaptive Slope based filter* (*variant of* MLS)**	Sithole, Vosselman	2001
Modified Slope bases filter (Local Lineal Regression)	Roggero	2001
Method similar to Kriging (weights functions)	Kraus and Pfeifer	2001
Modified Block Minimum (variant of morphological filter by Killian et al)	Wack, Wimmer	2002
(variant of morphological filter by Killian et al)	Masaharu, Ohtsubo	2002
***Progressive morphological filter* (PMF)**	Zhang et al	2003
***Elevation Threshold with Expanding Window* (ETEW)**	Zhang, Whitman	2005
Geodesic dilation	Arefi, Hahn	2005
Trend surfaces from first and last return	Zaksek, Pfeifer	2006
Polynomial fitting filters	Zhang and Cui	2007
Adaptation of polynomial fitting filters	Chen et al	2007
***Simple MoRpholocial Filter* (SMR)**	Pingel	2013
Multi-gradient analysis	Li	2013

En la tabla 2.2. se presenta una amplia relación de este tipo de filtros ordenados cronológicamente, basándose en la información recogida en distintos artículos, principalmente Lindeman 2012; Sithole 2005.

2.1.1.2. Filtros de Densificación Progresiva

También denominados PTD (*Progressive TIN Densification; TIN = Triangular Irregular Networks*). Este grupo de filtros se basan en la idea de Axelsson 1999 que trabaja en la reconstrucción progresiva del terreno. En un principio se genera a partir de una aproximación muy grosera, considerando sólo algunos de los puntos de la nube, normalmente los más bajos, creando con ellos la primera superficie de referencia. A partir de esa primera aproximación, se establece un criterio que sí se cumple permite incorporar un punto adicional por cada triángulo dentro del TIN a la superficie inicial (Vosselman and Maas 2010)

Al igual que algunos del grupo anterior, para generar la primera aproximación de la superficie del terreno utilizan un filtro simple de bloque mínimo con un tamaño de celda relativamente grande. En la literatura se puede comprobar que los métodos que se basan en esta idea también son considerados como un grupo más denominado **filtros de bloque mínimo** en la clasificación de estos algoritmos.

Estos métodos se denominan progresivos, porque a través de iteraciones van incorporando sucesivamente más puntos a la definición de la superficie buscada. Para ello, de manera genérica, utilizan dos parámetros: el ángulo formado entre la antigua alineación y la nueva; y, la distancia mínima a la que se encuentra el nuevo punto de la alineación anterior. Si esos dos valores se encuentran por debajo de los umbrales preestablecidos ese punto se incorpora a la superficie y se hace una nueva triangulación. Este proceso se realiza con el total de los puntos de manera que la superficie poco a poco va siendo más precisa.

En esta investigación se ha desarrollado el algoritmo de Sohn and Dowman 2002 constituido por dos fases. En la primera, a partir de las altitudes de las esquinas genera dos triángulos buscando de los puntos que quedan por debajo de esas aristas el más bajo para iniciar la iteración. En la segunda, refina la superficie de la fase anterior a partir de los tetraedros generados mediante un proceso iterativo. Finalmente, esta idea no se ha llegado a implementar por problemas de memoria.

En general, precisar que este tipo de algoritmos parecen ofrecen resultados muy buenos en superficies con discontinuidades, caso particular de las zonas urbanas, (Sithole 2005) pero son más influenciado por valores atípicos y no pueden eliminar bien los objetos construidos por el hombre tales como puentes (Li 2013), si bien algunos autores (Brovelli and Lucca 2011)consideran que estos algoritmos son más fiables en dichas estructuras.

Dentro de este grupo se encuentra el software TerraScan, que es uno de los más utilizados para la clasificación de los datos LiDAR, y aunque sus características se comentan en la siguiente sección, en este apartado se han contemplado dos referencias que indican la manera de proceder con este programa para clasificar estos datos.

En el artículo de (Brovelli and Lucca 2011) se siguen los siguientes pasos:

1. Primero se categorizan los puntos entre los que pertenecen al suelo y los que no.

2. Considerando los puntos no suelo del proceso anterior, se detecta los puntos que pueden ser edificios.
3. Finalmente, los puntos no catalogados como edificios en la fase anterior se clasifican como vegetación.

El primer paso se desarrolla empleando el algoritmo de Axelsson 1999, basándose en la densificación progresiva del TIN (PTD) y es el algoritmo utilizado para la generación de la superficie.

Con objeto de tratar de ilustrar la complejidad del proceso de clasificación, así como la flexibilidad del software indicado, a continuación se detalla el proceso seguido en una empresa puntera del sector cartográfico (Omega Cartografía Digital S. L. de Pamplona), compuesto por 8 fases consecutivas:

1. En primer lugar se elimina cualquier clasificación existente.
2. Con el objetivo de eliminar los puntos extraños, se busca sucesivamente los puntos fuera de rango.
3. Con el mismo fin anterior se realiza un proceso de eliminación de los puntos ubicados en el aire o por debajo del suelo (puntos aislados).
4. Eliminados todos los posibles puntos que constituyen ruido, se buscan los puntos pertenecientes al suelo haciendo uso del algoritmo PTD y definiendo tres parámetros: la pendiente, la distancia y el ángulo en el vértice del triángulo.
5. Seguidamente, se hace una reclasificación de los puntos sin clasificar considerándolos como puntos de vegetación baja.
6. Analizando las diferencias de altura entre estos puntos, los que presentan una diferencia de altura superior a un determinado valor, que se fija en concreto en 30 cm, se reclasifican como vegetación media.
7. Partiendo de los puntos del paso anterior, y de igual manera que en el paso 4, se clasifican en vegetación alta aquellos puntos con una mayor diferencia de altura, que se concretan en 2,5 m.
8. A partir de los puntos de vegetación alta, se buscan puntos que constituyan superficies de un determinado tamaño; para ello, hay que definir el tamaño mínimo de las superficies y la diferencia de alturas máxima entre esos puntos, dando lugar a las edificaciones.

En la tabla 2.3. se recogen las variantes más importantes de los filtros basados en la densificación progresiva desarrolladas en los últimos 15 años. Como en el caso anterior, la información está basada en diferentes artículos, entre los que destacan Lindeman 2012; Sithole 2005.

Tabla 2.3. Filtros de densificación progresiva para la clasificación de puntos LiDAR

Filtros de densificación progresiva	Autor(es)	Año
Height diferences	Hansen, Vögte	1999
Minimum Description Length **(MDL)**	Axelsson	1999 -2000
Grid based approach	Wack, Wimmer	2001
TIN *Thinning the spiking*	Haugerud, Harding	2001
Regularization Method	Sohn, Dowman	2002
An Adaptive **TIN *filtering in areas of steep slope***	Zhang, Cui	2007
Repetitive Interpolation **(REIN)**	Kobler et al	2007
Progressive densification and region growing	Pérez-García et al	2012
Delaunay **TIN**	Arranz et al	2012
Segmentation Using Smoothness Constraint **(SUSC) *& Progressive*** **TIN *Densification*** **(PTD)**	Zhang et al	2013
Streaming Progressive **TIN *Densification***	Kang et al	2014
Segmentation-Based Filtering **(SBF) *with*** **PTD**	Lin and Zhang	2014

Señalar que en los últimos 5 años el uso de variantes de este algoritmos se está combinando sobre todo con métodos basados en segmentación. Además, la gran diferencia que ofrece este método con respecto a los de morfología matemática es que estos filtros están relacionados con una reconstrucción del MDT y los otros no (Vosselman and Maas 2010).

2.1.1.3. Filtros basados en Superficies

A diferencia de los métodos de densificación progresiva que reconstruyen sucesivamente la superficie del terreno, estos filtros asumen que inicialmente todos los puntos pertenecen al suelo y paulatinamente se van eliminando aquellos que no deberían pertenecer a él, de manera que el modelo de superficie inicial iterativamente se va aproximando al MDE del terreno considerado.

En este caso, para la eliminación de los puntos se tiene en cuenta la interpolación robusta que integra el filtrado e interpolación del MDE en un único proceso. El objeto del algoritmo es determinar un peso individual para cada distribución irregular de puntos, tal que la superficie modelada represente el terreno (Vosselman and Maas 2010).

Todos los puntos pueden ser clasificados en terreno o no terreno basándose en el umbral de diferencias de alturas con respecto al MDE del terreno actual en cada ciclo de la iteración. El proceso consiste en los siguientes pasos:

1. Interpolación de la superficie modelo considerando pesos individuales para cada punto (al comienzo todos los puntos tienen el mismo peso).
2. Determinación de los valores del filtro para cada punto (asignación de la distancia desde la superficie al punto medido).
3. Cálculo de un nuevo peso para cada punto de acuerdo con el valor del filtro.

Estos pasos se repiten hasta alcanzar una situación estable de manera que los pesos de cada punto no cambien significativamente y la superficie varíe poco de una iteración a otra (Vosselman and Maas 2010).

Como variante a la interpolación robusta se introduce la aproximación jerárquica para asegurar una mezcla adecuada entre puntos de suelo y no suelo, necesaria en la interpolación robusta; de esta forma, también se consigue la aceleración del proceso (Pfeifer 2008).

Uno de los algoritmos que ha tenido mucho éxito en el ámbito forestal ha sido el desarrollado por Evans and Hudak 2007 que utiliza un procesamiento iterativo de la disposición vertical de los datos hasta que la solución converge, siendo innovador el modelo multi-escala que presenta.

Al igual que el método PMF explicado anteriormente, en esta investigación se ha utilizado el algoritmos de Evans and Hudak 2007 denominado *Multiscale Curvatura Classification* (MCC) implementado en la librería SPDlib en el apartado sexto, motivo por el cual se va a proceder a una breve explicación del mismo.

El algoritmo MCC ha sido desarrollado por *US Forest Service* con el fin de clasificar retornos en zonas de vegetación. Se trata de un algoritmo basado en superficies de forma que a partir de un suelo medio (TPS, *Thin Plate Spline*) iterativa y progresivamente va eliminando los puntos que se hayan sobre él clasificándolos como suelo o no en función de un umbral de curvatura (Evans and Hudak 2007).

La implementación del mismo requiere primero de la definición de un vector *Z(s)* el cual contiene las coordenadas *X, Y, Z* de todos los retornos LiDAR de la zona de estudio. A partir de él los pasos a seguir son:

1. Con *Z(s)* y TPS definidos se genera una superficie interpolada, aplicando dos parámetros para la escala de dominio *l*: el parámetro de escala λ y *t* la tolerancia de curvatura.

2. Se define un núcleo de 3×3, que se pasa sobre la superficie de interpolación, de manera que se va definiendo un vector *x(s)* según *Z(s)*.

3. Se calcula la curvatura (*c*) en la escala de dominio *l*:

4. Si se cumple que el vector *Z(s)* es mayor que la curvatura se clasifican los puntos como no terreno.

5. Se evalúa el umbral de convergencia *j* y en función de su valor se establece sí el modelo itera o empieza con el siguiente dominio de escala.

Tabla 2.4. Atributos del algoritmo MCC en SPDlib con sus valores por defecto

Parámetros	Defecto	Breve definición del parámetro
--thresofchange	0.1	Umbral de cambio
--filtersize	1 (3×3)	Tamaño del filtro de suavizado
--interpnumpts	16	Nº de puntos usados para la interpolación TPS
--interpmaxradius	20	Máximo radio de búsqueda para la interpolación TPS
--stepcurvetol	0,5	Tolerancia de curvatura para cada iteración
--mincurvetol	0,1	Tolerancia de mínima curvatura
--initcurvetol	1	Tolerancia de curvatura inicial
--scalegaps	0,5	Incrementos del cambio de escala
--numofscalesbelow	1	Nº de escalas a ser usadas sobre la inicial
--initscale		Escala inicial: por defecto el de los datos originales
--overlap	10	Valor de superposición entre bloques
-b	0	Unidad de trabajo: por defecto usa el del fichero SPD
-c	0	Número de columnas del tile
-r	100	Número de filas del tile

En la tabla 2.4. se recogen los valores que tienen por defecto los parámetros que constituyen el algoritmo en la librería SPDlib. En este caso los parámetros clave a determinar son --*initcurvetol* para establecer la tolerancia del valor de curvatura inicial; --*mincurvetol* que controla la tolerancia de la mínima curvatura y --*stepcurvetol*.para establecer la tolerancia de la curvatura para cada iteración. Según Bunting 2013 para modificar el comportamiento del algoritmo lo mejor es modificar el valor de --*initcurvetol*, teniendo en cuenta que la tolerancia de la curvatura varía entre --*initcurvetol* = 1 y --*mincurvetol* = 0,1.

Tabla 2.5. Filtros basados en superficies para la clasificación de puntos LiDAR

Filtros basados en superficies	Autor(es)	Año
Linear prediction (Robust interpolation)	Kraus, Pfeifer, Briese	1998
Active contours (Hierarchical interpolation)	Pfeifer et al	2001
Snake-approach (Hierarchical interpolation)	Elmqvist	2001
Spline surface interpolation	Broveli et al	2004
Perform Height filtering	Strentker, Glann	2006
Multiscale Hermite Transform **(MHT)**	Silván-Cárdenas, Wang	2006
Multiscale Curvature Classification **(MCC)**	Evans, Hudak	2007
Multiscale Terrain Filtering **(MTF; *layering*)**	Chen et al	2012
Multiresolution Hierarchical Classification **(MHC) *with Thin Plate Spline*** **(TPS)**	Mongus et al	2012
	Chen et al	2013
Thin Plate Spline-Based Feature-Preserving **(TPS-F)**	Chen et al	2015

Tal y como ocurría con los filtros morfológicos y de densificación progresiva, en los últimos 15 años este tipo de filtros han ido evolucionando y han surgido extensiones y variantes, resultando una gran gama de algoritmos. En la tabla 2.5 se reúnen los más importantes utilizando como fuente principal Shan and Toth 2008; Sithole 2005.

Este tipo de filtros suelen presentar buenos resultados en zonas arboladas y áreas urbanas, al igual que en zonas llanas o con pendiente suave. Algunos de estos filtros tales como el de MTF superan las dificultades que suelen presentar los puentes (Chen, et al. 2012).

2.1.1.4. Filtros basados en Segmentación

En el último grupo aparecen los métodos basados en la segmentación. Este tipo de algoritmos agrupan los puntos en objetos en función de que las características geométricas sean similares, por ello también se les dice algoritmos de agrupación o *cluster*.

Se puede señalar que estos métodos llevan consigo dos fases: por un lado, la generación de segmentos individuales por agregación de puntos con similares propiedades; y por otro, una vez unidos, la propia clasificación.

En cada una de estas fase se utilizan distintos tipos de algoritmos, y a diferencia de los anteriores, estos métodos clasifican los segmentos basándose en la diferencias de altura con respecto a los píxeles más cercanos (Kobler, et al. 2007).

Tabla 2.6. Filtros basados en segmentación para la clasificación de puntos LiDAR

Filtros basados en segmentación	Autor(es)	Año
Maximun and average gradients	Schiewe	2001
Edge based Clustering	Brovelli	2002
Wavelets	Thuy, Tokunaga	2002
Segment Filtering	Akel, Zilbertein, Dytsher	2003
Region-growing	Nardinocchi et al	2003
	Jacobsen, Lohmann	2003
	Tovari, Pfeifer	2005
Profile intersection	Sithole, Vosselman	2005
Smoothnes constraint	Rabbania	2006
Surface growing & Support Vector Machine **(SVM)**	Zhang et al	2012
Segmentation Using Smoothness Constraint **(SUSC) & *Progressive TIN Densification*** **(PTD)**	Zhang et al	2013
Segmentation-Based Filtering **(SBF) *with*** **PTD**	Lin and Zhang	2014

En la tabla 2.6. se muestra la recopilación realizada de los algoritmos de este tipo, basada principalmente en Shan and Toth 2008; Sithole 2005. Tal y como se puede apreciar en la selección anterior, muchos de los procesos de segmentación se fundamentan en el método de crecimiento de regiones, que parte de la detección automática de la semilla como base para la determinación de zonas de características iguales. Elegida la semilla el método añade gradualmente puntos vecinos basándose en medidas matemáticas de los puntos (diferencias de alturas, vector normal, etc.) de manera que en la definición de un plano matemático se van incorporando puntos hasta constituir segmentos que pertenecen a objetos de esa superficie.

En el caso de Sithole and Vosselman 2005 el segmento se rompe al aparecer discontinuidad en altura, catalogando el resto de segmentos en suelo desnudo u objetos, en función de las relaciones geométricas existentes con los segmentos de alrededor. A posteriori esos segmentos son utilizados como elementos básicos para que por interpolación lineal mínimo cuadrática (*least-squares interpolation*) se incorpore una función de pesos de manera que minimice los pesos de los segmentos no pertenecientes al terreno (Meng, et al. 2010).

Uno de los software que utiliza este tipo de filtros es GRASS cuyas características se especifican en el siguiente apartado. Este paquete para el filtrado de puntos utiliza una cascada de comandos (Brovelli and Lucca 2011):

1. Primero detecta los ejes de los objetos basándose en la diferencia de alturas con respecto a la posición horizontal.
2. Después utiliza el algoritmo de crecimiento de regiones para determinar cuatro categorías: terreno, terreno con doble pulso, objeto con doble pulso y objeto.
3. Para finalizar, corrige los posibles errores residuales derivados del crecimiento de regiones.

En general, se trata de métodos que trabajan con gran cantidad de datos y están menos influenciados por el ruido. En muchos casos la implementación de filtros basados en la segmentación requieren la rasterización de los datos y en consecuencia el uso de métodos de procesamiento de imágenes.

Estos procesos trabajan de manera similar a la técnica de interpolación robusta pero al trabajar con la imagen, dada la topología, pueden resultar más rápidos que los basados en superficies (Vosselman and Maas 2010). Además, resultan idóneos en áreas con gran influencia de actividad humana, pero no suelen ser adecuados en aquellas zonas con arbolado (Sithole and Vosselman 2004).

2.1.1.5. Otros filtros

En este apartado se exponen otros enfoques de interés que no pertenecen a alguna de las categorías anteriores.

Entre ellos merecen especial mención los filtros de escaneo direccional (Shan and Sampath 2005) que han pasado de calcular la pendiente y diferencias de elevación en la línea de escaneo (Streutker and Glenn 2006) a tener en cuenta además de esa línea los puntos cercanos etiquetados como terreno (Meng, et al. 2010). En la tabla 2.7. se pueden observar los más destacados.

Tanto el filtro *Multi-directional Ground Filtering* (MGF) (Meng, et al. 2009) como los basados en perfiles horizontales (Sithole 2005) y oclusiones (Habbib, et al. 2009) tienen como principal

objetivo la detección automática de edificios, en la línea de las investigaciones que se están llevando a cabo en los últimos tiempos para conseguir el modelado 3D de las ciudades.

Tabla 2.7. Otros filtros usados para el filtrado de puntos LiDAR

Filtros direccionales	Autor(es)	Año
Scan labeling	Shan, Sampath	2005
Bidirectional labeling	Shan et al	2005
Diferencias de elevación por línea de escaneo	Streutker and Glenn	2006
Multi-directional Ground Filtering **(MGF)**	Meng et al	2009

Otros filtros	Autor(es)	Año
Reasoning in horizontal slices	Zhan, Molenoar, Tempfli	2002
Occlusions by relief displacements	Habbib et al	2009

En el ambiente de la investigación destinada a la caracterización de la masa forestal, se proponen filtros que combinan distintos parámetros, como se presenta en la tabla 2.8. Así por ejemplo, para la clasificación de los puntos según el tipo de vegetación se están usando el histograma de múltiple retorno, la intensidad de manera individual o combinada con otros aspectos, o incluso índices de vegetación; aunque hace ya unos años que Song, et al. 2002 utilizaron la intensidad para distinguir cuatro clases: carreteras asfaltadas, hierba, tejados y árboles. En la tabla posterior se muestran algunas de las combinaciones encontradas, basadas principalmente en Lindeman 2012.

Tabla 2.8. Información derivada de datos LiDAR para clasificar la vegetación

Información datos LiDAR	Autor(es)	Año
Intensidad	Song et al	2002
Histograma de múltiple retorno	Raber et al	2002
Hybrid Normal Difference Vegetation Index **(HNDVI)**	Bretar and Chehata	2007
Intensidad y distribución de retornos	Goepfert et al	2008
Intensidad y distribución Gausiana	Wang and Glenn	2009
Intensity: skewness and kurtosis	Liu et al	2009

En cualquier caso, hay que indicar que están apareciendo algoritmos mixtos que agregan varios de los principios contemplados en los distintos cuatro grandes bloques aquí señalados, tal es el caso de (Zhang and Lin 2013) que combina el resultado de la densificación progresiva con la segmentación en función de la pendiente del terreno o Pérez-García, et al. 2012 que lo específica para el crecimiento de regiones. Además de estos, cada vez son más los métodos basados en algoritmos inteligentes; así, por ejemplo Hu, et al. 2015 propone un método basado en lo que denominan *Semi-Global Filtering* (SGF) que tiene en cuenta la función de energía, Gong, et al. 2015 utilizan las habilidades del LiDAR multiespectral (*MultiSpectral LiDAR*, MSL) y el algoritmo *Support Vector Machine* (SVM), Serna and Marcotegui 2014

combina la morfología matemática con la segmentación y el algoritmo SVM y Guo, et al. 2014 el clasificador JointBoost.

Tabla 2.9. Algoritmos en el ámbito de la minería de datos para clasificación de datos LiDAR

Algoritmo	Autor(es)	Año
Bayesian networks	Stassopoulou et al	2000
Dempster shafer fusion theory	Rottensteiner et al	2005
***Support Vector Machine* (SVM)**	Charaniya et al	2004
	Secord and Zakhor	2007
	Mallet et al	2008
Classification trees	Ducic et al	2006
	Matikainen et al	2007
	García-Gutierrez et al	2009
	Chechata et al	2009
Conditional Random Field* (CRF) *and Random Forests	Niemeyer et al	2013

En el ámbito de la clasificación supervisada que se plantea en el punto sexto de este trabajo se puede decir que el algoritmo que más se ha utilizado ha sido el SVM, pero de acuerdo con la metodología que se propone, hay que mencionar los basados en árboles de decisión tales como García-Gutiérrez, et al. 2009, y Chehata, et al. 2009 y Niemeyer, et al. 2013 que utilizan el algoritmo *Random Forest* para la clasificación de puntos LiDAR, teniendo la mayoría de ellos la particularidad de que se aplican en zonas muy concretas. En la tabla 2.9. se presenta la recopilación obtenida en este ámbito.

2.1.2. DISCUSIÓN

Se puede decir que en la actualidad la aplicación directa de las nubes de puntos LiDAR es la generación de modelos digitales de elevaciones (Blaszczak-Bak and Sobieraj 2013), motivo que justifica la búsqueda de un algoritmo ideal para el filtrado de los puntos de terreno para automatizar la creación de los MDEs.

Si bien esos puntos deberían quedar perfectamente clasificados según las categorías de la ASPRS, es de entender que para la representación de la forma del terreno sólo sea necesario la categoría de suelo por lo que una manera de lograrlo es clasificando los puntos en terreno y no terreno, tal y como se recoge en muchos de los artículos reseñados tales como Chang, et al. 2008; Tinkham, et al. 2011; Zhang, et al. 2003.

En consecuencia, a priori, con menos categorías el trabajo debería ser más fácil pero el problema es que con el trascurso de los años no se ha podido encontrar un único algoritmo que permita obtener resultados idóneos en todos los casos, ya que los resultados son distintos según el tipo de terreno considerado (Sithole and Vosselman 2004).

La mayoría de ellos ofrecen buenos resultados en terrenos llanos y sin vegetación (Ibidem) pero no todos se comportan bien con:

- Superficies de terrenos rugosos o pendiente discontinua.
- Áreas con una densa cobertura vegetal en las que no penetra adecuadamente el láser.
- Regiones con vegetación de baja altura.

Según el esquema que plantea Meng, et al. 2010 para la generación de los MDE, comentado en la introducción de este trabajo, los algoritmos apuntados en la sección anterior hacen referencia al filtrado de puntos y condicionan la creación del MDE, pudiendo afirmar de manera genérica que a la hora de construir el MDE (Ibidem):

- El tamaño del pixel está condicionado por la densidad de puntos y el espaciado entre ellos.
- La mayoría de los algoritmos sólo utilizan el primer retorno, aunque a veces se usa también el último.
- Para la determinación de la altura se suele utilizar un contexto de vecinos locales, para lo que es necesario el uso de una función que varía según el tipo de algoritmo utilizado.
- Para la búsqueda de los puntos vecinos se considera una ventana que por definición suele ser de tamaño fijo, lo cual es crítico para la definición del terreno; por ello, en la práctica la mayoría de los algoritmos la modifican.

Estos modelos digitales de elevaciones del terreno, principalmente de superficie (MDS) pero también los del terreno (MDT), están siendo de gran utilidad en el ámbito de los estudios forestales y en el modelado 3D de ciudades, exigiendo cada vez más precisiones.

En el caso forestal las utilidades más extendidas son la generación de modelos digitales para caracterizar la cubierta vegetal, estudios de biomasa o discriminación de distintos tipos de coberturas. Para ello, son necesarias nubes con mayor densidad de puntos (Kobler and Ogrinc 2007) y el uso de la inmensa mayoría de los retornos (Goepfert, et al. 2008). A este respecto, hay que mencionar que recientemente Martin Isenburg, desarrollador de LASTools (comentado en el apartado de software, punto tercero), ha establecido un novedoso algoritmo para usar todos los retornos relevantes al crear MDS y CHM (*Canopy Height Model*) (Isenburg 2015).

Por su parte, en el modelado 3D de las ciudades lo que importa es la generación del modelo tridimensional de las mismas. Estos dos ámbitos de actuación tienen como objetivo primordial la obtención de modelos de elevación de superficies (MDS), modelos que necesitan discernir puntos que pertenezcan a otro tipo de objetos dentro de los puntos que no pertenecen al terreno, usando para ello, de manera genérica, el análisis de las alturas.

Estos hechos han contribuido a la aparición de muchos artículos que hacen referencia a la clasificación de puntos en suelo, vegetación y edificios o en edificios / no edificios, o sobre distintos tipos de vegetación (Sankey and Bond 2011), o incluso aquellos que buscan distinguir

edificios, árboles y alineaciones (Meng, et al. 2010). Lo que ha traído consigo el desarrollo de nuevos procedimientos dentro de los cuatro grupos de algoritmos comentados. Así, por ejemplo en Arranz, et al. 2012 se utiliza una actualización de la triangulación de Delaunay (algoritmo basado en densificación progresiva) para discriminar entre edificios y vegetación.

Con ese mismo objetivo Hui, et al. 2008 señalan que a la hora de clasificar puntos con igual altura en edificios o vegetación resulta necesario el uso de algún otro parámetro y uno de ellos puede ser la intensidad. Hecho que ha llevado a que, sobre todo en el ámbito forestal, cada vez sean más los algoritmos que tienen en cuenta los valores de intensidad para discernir entre puntos de distintas categorías (Blaszczak-Bak and Sobieraj 2013), incluso dentro de los pertenecientes al terreno se pretende diferenciar entre asfalto de carreteras o hierba, utilizando para ello los valores de reflectancia. Si bien es verdad, que cada vez son más los autores que ven necesaria la normalización de la intensidad. A este respecto, en Kashani, et al. 2015 se reflejan los aspectos a tener en cuenta.

Pero a nivel general, de las cuatro familias de algoritmos existentes, parece ser que el uso de los algoritmos basados en segmentación mejoran la clasificación de los datos cuando se buscan más categorías que terreno / no terreno, lo cual suele venir ayudado por la incorporación de conocimiento adicional y relaciones contextuales (uso de otras bandas) que contribuyen a mejorar los resultados de la clasificación (Bartels and Wei 2006); lo que viene a decir es que el uso de datos fusionados favorece la categorización.

En el caso del filtrado en zonas urbanas, Zhang, et al. 2013 muestran que los métodos de segmentación basados en *Object-based Point cloud Analysis* (OBPA) junto con el uso de *Support Vector Machine* (SVM) basados en algoritmos de aprendizaje automático, analizan los datos y reconocen patrones ofreciendo resultados muy buenos.

Además, en Chen, et al. 2013 también queda patente que el filtrado en zonas con pendientes pronunciadas y cambios abruptos, que se corresponden con los paisajes que ofrecen mayores desafíos, podrían ser resueltos por los algoritmos basados en segmentación. Aunque en Mongus and Žalik 2012, dentro del grupo basado en superficies, se indica que el uso de *Thin Plate Spline* (TPS) también resuelve este problema.

Como ya se ha reseñado al estudiar la clasificación de filtros, cada vez son más los filtros que combinan algoritmos pertenecientes a los distintos grupos. Esto se puede comprobar en Lin and Zhang 2014 donde se introducen técnicas de segmentación (*Segmentation-Based Filtering*, SBF) en la densificación progresiva del TIN con el objetivo de mejorar la densificación y minimizar el error de omisión. Y no sólo de la combinación entre ellos, sino que también de otras técnicas de aprendizaje como puede SVM u otras. En Zhang and Lin 2012 utilizan la clasificación orientada a objetos en base a dicho algoritmo.

En cualquier caso, tal y como concluyen Zhang and Whitman 2005 cada método tiene sus propios errores de omisión y comisión dependiendo de los parámetros de filtrado utilizados, siendo el valor más sensible el que hace referencia a la pendiente topográfica.

Por último, cabe reseñar que parece haber un consenso a la hora de evaluar los distintos algoritmos, ya que la mayoría de los artículos aquí mencionados utilizan los ocho grupos de datos de la ISPRS Commission III/WG3 mostrados en Sithole and Vosselman 2003 que cubren diferentes tipos de terreno ofreciendo distintas dificultades de filtrado pero con las mismas características de los datos LiDAR. Para valorar los resultados obtenidos utilizan los errores de comisión y de omisión junto con el índice de *Cohen's Kappa*.

2.2. INTRODUCCIÓN A LA MINERIA DE DATOS

Con el avance tecnológico del último siglo y el uso generalizado de internet enormes volúmenes de datos estructurados en bases de datos (BBDD), en ámbitos muy diversos han sido abiertas y puestas a disposición de gran parte de la sociedad, lo que ha demandado herramientas que permitan el análisis y extracción de información de los mismos.

En este marco, aparece la **minería** o **exploración de datos** (*Knowledge Discovery in Databases*, KDD) que a partir de un proceso analítico se encarga de descubrir patrones dentro de grandes volúmenes de datos, también conocidos como *Big Data*, para transformarlos en una estructura entendible de manera que luego se puedan usar para otro fin (Chakrabarti, et al. 2006); es decir, se ocupa de detectar algo nuevo para luego aplicarlo y obtener entendimiento inédito permitiendo predecir comportamientos.

La minería de datos o *Data Mining* (DM) se puede considerar como una evolución natural de la información tecnológica y la confluencia de varias disciplinas relacionadas con una gran variedad de dominios en los que puede ser aplicada (Han, et al. 2011). Entre las ciencias relacionadas más importantes se encuentran la inteligencia artificial (*Artificial Intelligence*, AI), el aprendizaje automático (*machine learning*), el reconocimiento de patrones (*pattern recognition*), el análisis de datos (*data science*), la estadística (*statistical*) y los sistema de bases de datos (*DataBase Management System*, DBMS).

En cada una de estas especialidades se utilizan algoritmos que hacen referencia a distintos problemas; así, por ejemplo la estadística comprende análisis lineal discriminante o modelos lineales generalizados, mientras que la inteligencia artificial se centra más en clasificadores basados en reglas o árboles de decisión, siendo las redes neuronales las más utilizadas en enfoques en los que se necesita conexión (Fernández-Delgado, et al. 2014).

Son muchos los ámbitos de aplicación, se puede decir que son infinitos, ya que estos métodos de cálculo son tanto aplicables a medicina como finanzas, comercio, seguridad, agricultura, biología, educación y cualquier otro espacio científico del que se dispongan datos. Independientemente de la disciplina, lo que les aúne a todos es que se trata de procesos automáticos compuestos por varios elementos: el conjunto de datos de entrenamiento que permite extraer las reglas, el modelo que se implementa a través de un programa y la validación del mismo.

En base a esos elementos, a partir de la bibliografía consultada se puede comprobar cómo existen distintos flujos de trabajo pero en todos, de alguna manera, se distinguen tres fases (Alexander, et al. 2011; Chapman, et al. 2000; Rojão 2008):

1. Exploración de los datos
    a. Selección del conjunto de datos
    b. Análisis de las propiedades de los datos
    c. Transformación del conjunto de datos
2. Construcción y validación del modelo
    a. Selección de la técnica de minería de datos
    b. Extracción de conocimiento
    c. Medidas de validación y errores
3. Aplicación del modelo
    a. Manejo del modelo
    b. Interpretación y evaluación de resultados

En la fase de exploración de los datos suele ser común la realización de la limpieza de los mismos, corrigiendo o eliminando identificaciones erróneas, seleccionando subconjuntos o incluso transformándolos.

Para construir el modelo definitivo se realizan distintos ensayos hasta que se da con el más adecuado para extraer el conocimiento buscado. Finalmente, queda aplicarlo a otros datos y evaluar la predicción alcanzada interpretando los resultados logrados.

Para la realización del modelo se necesitan algoritmos y si bien las posibilidades existentes son inmensas, cabe indicar que éstos se clasifican en supervisados o predictivos (*supervised learning*) y no supervisados o del descubrimiento del conocimiento (*unsupervised learning*), aunque en los últimos tiempos se están considerando también los semi-supervisados (*semi-supervised*).

Se entiende por métodos supervisados o predictivos (clasificación) a aquellos que usan un conjunto de datos a modo de entrenamiento y no supervisados (clustering) a los basados en el ajuste del modelo a las observaciones disponibles. En estos últimos no existe conocimiento previo sobre lo que se quiere estudiar, mientras que en el otro caso los datos de entrenamiento marcan las categorías que se desean predecir.

En el caso de los métodos semi-supervisados se usan tanto datos de entrenamiento etiquetados como no etiquetados, siendo estos últimos los más frecuentes.

Dentro de los dos grandes grupos algunas de las técnicas que más se utilizan según (StatSoft 2008) son:

- Aprendizaje supervisado:
    - ✓ Clasificación:
        - Logístico
        - Árboles de decisión
        - Boosting árboles

        - Redes neuronales
    - ✓ Regresión
        - Lineal
        - Shrinkage
        - Redes neuronales
        - Kernels
    - ✓ Series temporales
    - ✓ Optimización
- Aprendizaje no supervisado:
    - ✓ Componentes de análisis principales (*Principal Component Analysis*, PCA)
    - ✓ Agrupamiento o clustering (K-means, K-medoids)
    - ✓ Reglas de asociación

En la presente investigación se han utilizado técnicas de clasificación basadas en árboles dentro del aprendizaje supervisado, por lo que en las próximas líneas se van a explicar primero en qué consisten los árboles de clasificación y a continuación el algoritmo *Random Forest* (RF) utilizado en el experimento. Para entenderlo mejor, previamente se ha hecho un breve repaso de los métodos de ensamblado, base del mismo.

2.2.1. ÁRBOLES DE CLASIFICACIÓN

Un **árbol de decisión** (*Decision Tree*, DT) también llamado árbol de clasificación representa un conjunto de decisiones organizadas en una estructura jerárquica, de manera que en el mismo nivel sólo puede ser una categoría (Hernández, et al. 2005).

Constituye una herramienta que permite elegir entre varias alternativas representando de forma sencilla distintos ejemplos de clasificación, siendo una de las técnicas con mayor éxito en el ámbito de aprendizaje de clasificación supervisada (Poole and Mackworth 2010).

El objetivo es crear un modelo que prediga los valores de la variable independiente utilizando la entrada de otras variables distintas a través del aprendizaje inductivo supervisado. A partir de ejemplos sencillos se buscan patrones comunes descubriendo el conocimiento, siendo necesario el uso de datos de entrenamiento (*training data set*) a partir de los cuales se genera el modelo y datos de test en el que se aplicará y se validará el modelo creado.

Dentro de los métodos de aprendizaje inductivo supervisado, los árboles de decisión constituyen lo que se denomina procedimientos de caja blanca ya que permiten inspeccionar el modelo. Para ello, están constituidos por nodos que pueden ser clases también llamados variables o atributos (*features*) u hojas (*leaves*) unidos entre sí por ramas o arcos (*branches*) que etiquetan los valores de los atributos, intentando buscar a través de las distintas clases los valores del atributo objetivo (*target feature*).

Las clases se pueden unir con otras clases (nodos que no son hojas, *non-leaf*) o con las hojas a través de las ramas o arcos, estableciendo las hojas los distintos valores del atributo clave, motivo por el que las hojas aparecen al final del árbol. A modo de ejemplo en la figura 2.2 se muestra un árbol de decisión en el que 0 y 6 constituyen los valores del atributo objetivo (nodos hoja) y h_max, diff_h, orto_B y NDVI_ortoR son las variables usadas según sus valores (mostrados a lo largo de las distintas ramas) en el árbol desarrollado para alcanzar los valores del atributo objetivo. En este caso, el atributo más significativo es NDVI_ortoR y el menos notable h_max.

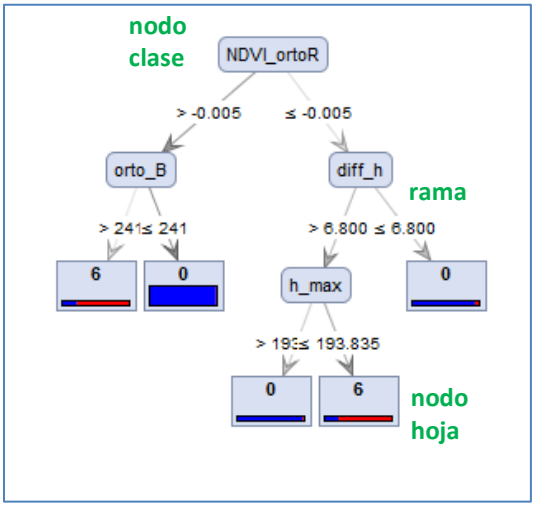

Figura 2.2. Estructura de un árbol de decisión

Si se trata de árboles de decisión para clasificar las clases deben disponer valores discretos, pero si los valores son continuos entonces se deben usar los árboles de regresión. El término CART (*Classification And Regression Trees*) permite considerar ambos procedimientos y fue introducido por Breiman, et al. 1984.

Existen dos formas de generar árboles de inducción: de abajo a arriba (*bottom-up*) y de arriba a abajo (*top-down*), siendo estos últimos lo más utilizados por la literatura dando lugar a lo que se denomina *Top-Down Induction of Decission Trees* (TDIDT) también conocidos como "divide y vencerás" (*divide and conquer*) (Rokach and Maimon 2005).

En este último tipo de árboles el atributo más importante se ubica en la parte superior del árbol y en función de los valores de los atributos iterativamente se van realizando las distintas particiones del conjunto de entrenamiento según una heurística o regla de división que mide lo adecuado que resulta una variable para constituir un árbol mínimo.

La forma en la que se realiza esa división da lugar a los distintos métodos: el algoritmo básico utilizado para resolver este tipo de árboles es el ID3 (*Iterative Dichotomiser* 3) (Quinlan 1986), disponible en el software Weka; como evolución de éste Knime utiliza el algoritmo C4.5. Ambos métodos están basados en la ganancia de información (*information gain*) que utiliza el cálculo de la entropía (*E*) para establecer la cantidad de información que no es cubierta por un atributo (García-Gutiérrez 2012).

donde:

T = subconjunto de registros
T_i = división de *T* según un valor de *A*
C = número de muestras
A = atributo que maximiza la relación de la ganancia en *T*
p(T,j) = probabilidad de que un conjunto de registro tenga la etiqueta *j*

Otros algoritmos como CART, SLIQ (*Supervised Learning In Quest*) o SPRINT (*Scallable Parallelizable Induction Of Decision Trees*) utilizan el índice de Gini (*Gini index*).

Pero existen muchas otras reglas como la de CHAID (*CHi-squared Automatic Interaction Detector*), MARS (*Multivariate Adaptive Regression Splines*), *Conditional Inference Trees*, etc.

A este respecto, hay que mencionar que ninguna de ellas es perfecta y aunque algunas se basan únicamente en la fase de crecimiento (*growing*) del árbol; otras, como es el caso de los algoritmos C4.5 y CART, además de crecer realizan una poda (*growing and pruning*).

La fase de crecimiento se detiene cuando se cumple alguno de los siguientes criterios (Rokach and Maimon 2005):

- Todos los casos del entrenamiento pendientes pertenecen a la misma clase.
- Se alcanza la profundidad del árbol establecida.
- El número de casos del nodo final es menor que el mínimo número de casos para los nodos padre.
- Cuando en el nodo a dividir el número de casos en uno de los hijos es menor que el mínimo número de casos de los nodos hijos.
- El mejor criterio de partición no es más grande que un cierto umbral.

Por su parte, el *pruning* pretende establecer un criterio de parada suave, evitando que el árbol de decisión se sobre-ajuste al conjunto de entrenamiento (Alexander, et al. 2011), de manera que se cortan los árboles sobre-ajustados dejándolos en otros más pequeños tras quitarles las ramas que no contribuyen a generalizar la precisión.

La poda también se realiza en base a ciertos criterios y entre las técnicas más populares destacan: *Cost-complexity pruning*, *Reduced-error pruning, Minimum-Error Pruning* (MEP), *Error-Based Pruning* (EBP), *Optimal Pruning*, *Minimum Description Length* (MDL) *Pruning*.

De manera genérica, en la literatura consultada (Friedman, et al. 2001; Rokach and Maimon 2005) destacan entre las ventajas que ofrece este tipo de aprendizaje que los árboles:

- Son fáciles de entender e interpretar.
- Admiten tanto atributos nominales como numéricos.
- Aceptan conjuntos que pueden tener errores o datos perdidos.
- Son considerados como métodos no paramétricos.

En cuanto a las desventajas señalar que:

- La mayoría de los algoritmos requieren valores discretos para el atributo clave.
- Tienden a funcionar bien si existen atributos relevantes, pero presentan problemas si aparecen interacciones complejas.

(del Toro Espín, et al. 2015) señalan que el principal problema de la clasificación con un único árbol es su alta sensibilidad a los datos de entrada, por lo que se proponen alternativas basadas en conjuntos de árboles como el *boosting*, *bagging* o *Random Forest*, todos ellos métodos de ensamblado.

2.2.2. MÉTODOS DE ENSAMBLADO

El objetivo de los métodos de ensamblado (*Ensemble methods*) radica en construir un modelo predictivo por integración de múltiples modelos mejorando el rendimiento de las predicciones (Rokach 2010). Consisten en combinar las predicciones de varios modelos realizados con uno o varios algoritmos de aprendizaje y generalizarlos para dar como resultado un modelo único (Scikit-learn 2012).

Las primeras investigaciones plasmadas en el aprendizaje supervisado datan de los años setenta, así Tukey 1977 en el ámbito del análisis exploratorio de datos (*Exploratory Data Analysis*, EDA) propuso la combinación de dos modelos lineales de regresión.

En cualquier flujo de trabajo de este tipo se distinguen los datos de entrenamiento (*training set*) que contienen los datos etiquetados, el algoritmo de inducción (*base inducer*) que a partir del *training set* genera una clasificación que representa la relación generalizada entre los atributos de entrada y la variable clave (*target attribute*), el generador de diversidad (*diversity generator*) para crear las distintas clasificaciones y el que combina (*combiner*) para juntar las distintas clasificaciones (Rokach and Maimon 2014).

En general, se trata de meta-algoritmos que van buscando reducir el sesgo y la varianza en el aprendizaje supervisado. Estos métodos se clasifican en dos grandes grupos:

- Métodos de promedio (*averaging methods*) o independientes: se construyen distintos modelos independientes y como resultado se muestra la media de ellos en el caso de regresiones, y el de mayor voto (*mayority voting*) para clasificaciones. Un caso especial de estos métodos lo constituye el denominado *bagging* (***b**oostrap **agg**regat**ing***).

El *bagging* constituye un método de ensamblado introducido por Breiman 1996 que considera de un conjunto completo, de manera independiente y aleatoria, nuevos subconjuntos de entrenamiento todos ellos del mismo tamaño y menor que el del conjunto original ofreciendo como solución una predicción con una varianza menor que la obtenida por cada uno de los subconjuntos, evitando el sobre-ajuste.

- Métodos de impulso (***boosting** methods*) o dependientes: los modelos se construyen de manera secuencial y la solución es dada por el modelo que reduce el sesgo del modelo combinado. Se trata de métodos iterativos y como ventaja aprovechan el conocimiento generado en la iteración previa, guiando el aprendizaje de las iteraciones posteriores. Entre los métodos más utilizados destaca *AdaBoost* (***Ada**ptive **Boost**ing*).

En ambos métodos para la determinación de la mejor solución se sigue lo que se denomina *voting approach* o *averaging,* pero la diferencia principal radica en que los métodos de impulso o ***boosting*** utilizan pesos distintos para los distintos modelos, mientras que en los de promedio el peso es igual para todos los modelos (Witten, et al. 2011).

Indistintamente de que se trate de métodos dependientes o independientes, por lo general al combinar varios modelos se consiguen mejores resultados, reduciendo la varianza y evitando el sobre-ajuste. Los resultados suelen ser mejores cuando existen diferencias significativas entre los modelos (Han, et al. 2011).

Dentro de la librería Scikit-learn 2012, utilizada para la parte experimental, como métodos de ensamblado, además del de RF dispone de *Extra Trees*, *Adaptive Boosting* y *Gradient Boosting*.

2.2.3. EXTRA TREES

Dentro de los métodos de ensamblado el ***Extra Trees Classifier*** (ET) constituye un meta-estimador que ajusta un número de árboles de decisión aleatorio en varias sub-muestras del conjunto de datos original usando el promedio para mejorar la precisión de la predicción y controlar el sobreajuste.

Se trata de árboles de decisión sin poda que responden al clásico procedimiento *top-down* y forman parte de la técnica *perturb-and-combine* (perturbación-y-combinación) con dos diferencias importantes frente a los árboles de decisión: por un lado, los nodos se crean a partir de puntos de corte acordados completamente al azar; y por otro, para el crecimiento de los árboles se utiliza el total de la muestra (Geurts, et al. 2006).

Su aplicación se suele abordar cuando se pretende aumentar la precisión de la predicción dada por un árbol de decisión. Al compararlo con RF cabe señalar que éste no utiliza *bagging* ni reemplazamiento.

2.2.4. RANDOM FOREST

Random Forest constituye un algoritmo de aprendizaje supervisado no paramétrico fundamentado en métodos de ensamblado de promedio que permite obtener sus predicciones de manera más precisa. El método combina la idea de *Bagging* propuesta por Breiman 1996 y la de *Random Subspace* (RS) de Ho 1998.

Se trata de una técnica de perturbación-y-combinación diseñada específicamente para árboles de decisión (Breiman 1998), en la que para crear los distintos modelos se introducen alteraciones al azar en el método de aprendizaje (Louppe 2014).

Normalmente, se utilizan dos o más modelos cada uno con subconjuntos de datos y variables diferentes como apoyo para combinar las predicciones independientes en un pronóstico único del conjunto, ofreciendo mejor resultado que el conseguido por cada uno de los subconjuntos de manera individual (Brownlee 2014).

Cada árbol de decisión se genera según el algoritmo de Breiman 2001:

- Sea *N* el número de muestras del conjunto de entrenamiento, del que se consideran *n* casos elegidos aleatoriamente para la construcción del árbol.
- Sea *M* el número total de variables. Para cada nodo se seleccionan aleatoriamente *m* variables del total *M*, de manera que la mejor partición de esas *m* variables es la elegida para dividir el nodo. Durante el crecimiento del bosque *m* se mantiene constante.
- Cada árbol se extiende lo más posible, sin aplicar *pruning*.

El índice de error depende de la correlación existente entre dos árboles cualesquiera del bosque – a mayor correlación mayor error - y la fuerza de cada uno de los árboles individuales – un árbol con un índice de error bajo es un clasificador fuerte. La reducción de *m* reduce tanto la correlación como la fuerza, y un aumento de *m* contribuye a incrementarlas (Ibidem).

Ese error es estimado por el algoritmo durante su procesamiento al utilizar *The out-of-bag* (oob) que consiste en dejar fuera de cada árbol un tercio de las muestras elegidas para la construcción del mismo, de manera que valgan para determinar el error. Esto permite también calcular la importancia de las distintas variables utilizadas, algo característico de este algoritmo (Ibidem).

Entre las ventajas que se le atribuyen destacan (Ibidem):

- Se trata de un algoritmo rápido.
- Resulta eficiente con grandes base de datos.

- Puede tratar cientos de variables sin excluir ninguna.
- Ofrece una estimación de las variables más importantes.
- Permite crear múltiples árboles de decisión de manera paralela.
- Dispone de un método efectivo para la estimación de datos perdidos y mantiene la precisión si éstos se dan en gran proporción.

Como desventajas se suelen indicar principalmente que si en los datos existe ruido el algoritmo se sobre-ajusta y que las clasificaciones realizadas por RF resultan difíciles de interpretar.

3. HERRAMIENTAS INFORMÁTICAS

Este apartado constituye el marco en el que se explican las herramientas informáticas utilizadas a lo largo de toda la investigación, empezando por aclarar en qué consiste la configuración estándar de los datos LiDAR: **el formato LAS**.

Seguidamente, se comentan los programas disponibles que permiten clasificar los datos LiDAR, relacionándolos con el tipo de algoritmo que utilizan y ya explicado en la sección anterior. Y, para concluir, se describen brevemente el conjunto de herramientas aplicadas en las metodologías seguidas.

3.1. EL ESTÁNDAR DE DATOS LIDAR

El estándar de los datos LiDAR lo constituye **el formato LAS** (*Log ASCII Standard*). Se trata de un archivo binario que tiene la particularidad de guardar toda la información procedente de la toma de los datos durante el vuelo, conservándola según su naturaleza y sistema de captura. Constituye un formato de fichero público para el intercambio de nubes de datos en tres dimensiones (*X*, *Y*, *Z*), resultando una alternativa a los ficheros propietarios generados por las distintas compañías y facilitando la permuta de este tipo de datos entre diferentes empresas y paquetes de procesamiento tanto propietarios como de libre distribución.

Las especificaciones del mismo han sido desarrolladas por la *American Society for Photogrammetry & Remote Sensing* (ASPRS), definiéndolo en dos partes: la cabecera constituida por un bloque de carácter público para guardar el número de puntos y los valores extremos de coordenadas, seguida por registros de longitud variable donde se almacena la información de la proyección, metadatos y datos de aplicación del usuario (tabla 3.1); por último, el almacenamiento de los datos en sí a modo de puntos. Aunque esto también varía según las versiones; así, por ejemplo, en la última versión 1.4 se ha añadido al final unos registros de longitud variable extendido (*Extended Variable Length Records*, EVLR) (ASPRS 2013).

Tabla 3.1. Información comentada sobre la cabecera de un fichero LAS

reporting all LAS header entries:
file signature: 'LASF'
file source ID: 0
global_encoding: 0
project ID GUID data 1-4: 0 0 0 ''
Versión de fichero LAS al que hace referencia
version major.minor: 1.2
system identifier: ''
Programa utilizado para el procesamiento
generating software: 'TerraScan'
Fecha de creación
file creation day/year: 65/2012
header size: 227
offset to point data: 227
number var. length records: 0
Formato en el que se presentan los datos
point data format: 3
point data record length: 34
Número total de retornos
number of point records: 1675027
Número total de retornos por cada eco
number of points by return: 1416059 235557 22748 663 0
scale factor x y z: 2.32828e-007 2.32828e-007 6.16675e-008
Coordenadas medias X, Y, Z
offset x y z: 540499.995 4781499.9950000001 286.24000000000001

Coordenadas mínimas y máximas del fichero

min x y z: 540000 4781000 153.81
max x y z: 540999.98999999999 4781999.9900000002 418.67000000000002
reporting minimum and maximum for all LAS point record entries ...
X -2147483647 2147483647
Y -2147483647 2147483647
Z -2147483647 2147483647

Valores mínimo y máximo para la intensidad, eje de vuelo, dirección de escaneo, cantidad de retornos, número de retorno, clasificación, ángulo de escaneo, datos del usuario, identificador de pasada y tiempo GPS

intensity 0 255
edge_of_flight_line 0 1
scan_direction_flag 0 1
number_of_returns_of_given_pulse 1 4
return_number 1 4
classification 0 11
scan_angle_rank -21 21
user_data 115 146
point_source_ID 14 17
gps_time 551417.067548 554931.481443

Valores mínimo y máximo para la visualización en RGB

Color R 3584 65280
G 4096 65280
B 4864 65280

Número de puntos del último retorno

number of last returns: 1415434

Superficie cubierta en m^2 / km^2

covered area in square units/kilounits: 991444/0.99

Densidad de puntos por m^2, considerando todos los retornos y sólo el último

point density: all returns 1.69 last only 1.43 (per square units)

Espaciado entre puntos, todos los retornos y sólo el último (m)

spacing: all returns 0.77 last only 0.84 (in units)

Número de puntos por cada retorno

overview over number of returns of given pulse: 1180837 423537 67838 2815 0 0 0

Histograma de la clasificación (número de puntos por cada clase)

histogram of classification of points:
429581 Created, never classified (0)
506284 Ground (2)
46536 Low Vegetation (3)
48269 Medium Vegetation (4)
489659 High Vegetation (5)
153810 Building (6)
831 Low Point (noise) (7)
57 Reserved for ASPRS Definition (11)

Desde el año 2003 la ASPRS ha ido modificando las versiones del formato incluyendo en ellos distintas configuraciones para el registro de los puntos dato (*Point Data Record Format*, PDRF). En la tabla siguiente se puede ver su evolución y en ASPRS 2015 toda la información referente a los distintos formatos.

Tabla 3.2. Evolución de las versiones del formato LAS

Versión	PDRF	Fechas
LAS 1.0	--	Mayo 2003
LAS 1.1	0, 1	Mayo 2005
LAS 1.2	0, 1, 2 , 3	Septiembre 2008
LAS 1.3	0, 1, 2, 3, 4, 5	Octubre 2010
LAS 1.4	0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10	Noviembre 2011 Julio 2013, última actualización

Dependiendo del tipo de registro de los datos (PDRF) la información que se va a disponer en la base de datos es diferente; por ejemplo, el PDRF 0 no dispone del tiempo GPS mientras que el 1 si, pero éste no considera la información asociada de las bandas R, G y B. Aunque también puede variar el tamaño de bytes o bits destinado para cada concepto. En el caso de un sistema LiDAR de pulso discreto lo habitual es que conlleve el PDRF 3, de manera que para cada punto capturado se va a disponer de la información aportada en la tabla 3.3.

Tabla 3.3. *Point Data Record Format* 3

Campo	Descripción
X	Coordenada planimétrica X
Y	Coordenada planimétrica Y
Z	Coordenada altimétrica Z
Intensity	Intensidad del punto laser en el sensor
Return_Number	Número de retorno de ese pulso
Number_of_Returns	Número de retornos detectados para ese pulso
Scan_Direction_flag	Dirección del espejo del escáner
Edge_of_Flight_Line	Borde de línea de vuelo
Classification	Valor de clasificación asignada a ese pulso
Scan_Angle	Ángulo de escaneo
User_Data	Campo a rellenar según necesidad del usuario
Point_Source_ID	Identificador de pasada
GPS_Time	Tiempo GPS
R	Valor asociado al canal Rojo
G	Valor asociado al canal Verde
B	Valor asociado al canal Azul

Los campos *Scan_Direction_flag* y *Edge_of_Flight_Line* son booleanos. En el primer caso, el valor 1 indica que la dirección del escáner es positiva (el espejo se mueve desde el lado izquierdo, en la dirección del trayecto, al lado derecho), mientras que el 0 significa un movimiento del espejo en sentido contrario (desde el lado derecho al izquierdo). En el segundo caso, *Edge_of_Flight_Line* adquiere un valor de 1 cuando se corresponde con puntos del final del escaneo, lo que significa que se va a producir un cambio en la dirección de la pasada.

Por su parte *Point_Source_ID* está constituido por un valor numérico para marcar la dirección de la pasada a la que pertenece ese punto, pudiendo adquirir valores entre 1 y 65535 dentro del mismo fichero, aunque no es lo habitual. El valor 0 sólo se usa para casos especiales y viene a señalar que se trata de puntos originales. La información de este campo debe coincidir con el de *File_Source_ID*.

En cuanto al campo *Scan_Angle* señalar que puede adquirir valores entre -90º y +90º, siendo el valor de 0º para los puntos situados en el nadir y -90º para los del lado derecho del avión en la dirección de vuelo. Hay que indicar que el campo *File_Maker* de la versión 1.0 ha sido renombrado como *User_Data* y el campo *User_Bit* por el *Point_Source_ID*.

En lo que respecta al campo de *classification* indicar que hace referencia al valor de la clasificación. Su codificación está constituida por 4 bits para indicar el tipo de elemento que representa, tal y como se ha mostrado en la tabla 3.4. Tanto el valor 0 como el 1 hacen referencia a puntos sin clasificar, pero se mantienen ambos para compatibilizar el uso del popular software de clasificación TerraScan (ver apartado 3.2), tal y como se indica en las especificaciones de la ASPRS 2008.

Según dichas especificaciones, los puntos de los datos LiDAR facilitados en las versiones LAS 1.1, 1.2 ó 1.3 se clasifican en valores del 1 al 31 en función del tipo de objeto del que se trate, quedando del 10 en adelante reservados para nuevas definiciones que realice la ASPRS en un futuro (ESRI 2013).

Tabla 3.4. Valores ASPRS para la clasificación del LAS 1.1

Clase	Significado	Descripción
0	*Created, never classified*	
1	*Unclassified*	Sólo último retorno
2	*Ground*	Sólo último retorno
3	*Low Vegetation*	Vegetación entre 0 - 0,3 m
4	*Medium Vegetation*	Vegetación entre 0,3 - 2 m
5	*High Vegetation*	Vegetación mayor a 2 m
6	*Building*	Edificaciones
7	*Low Point (noise)*	Falsos puntos
8	*Model Key-point (mass point)*	Puntos clave para el modelo
9	*Water*	Puntos de agua
10	*Reserved for ASPRS Definiton*	Reservados para futuras definiciones
11	*Reserved for ASPRS Definiton*	
12	*Overlap Points*	Superposición de líneas de vuelo
13-31	*Reserved for ASPRS Definiton*	Reservados para futuras definiciones

En julio de 2013 se realizó la última actualización de la versión 1.4 reconociéndose algunos elementos más tal y como se puede ver en la tabla 3.5. Además, tal y como se indica en Heidemann 2014, otra de las mejoras importantes de esta versión se refiere a que las

intensidades se deben ofrecer normalizadas a 16 bits, lo cual puede ofrecer otro avance importante en la línea del trabajo que aquí se presenta.

Al comparar las tablas 3.4 y 3.5 se puede apreciar que la gran diferencia radica en que se han aumentado el número de categorías, ya que antes sólo había hasta 31 registros y ahora aparecen hasta 255. Asimismo aparecen como nuevas categorías la 10 para vías férreas, 11 para carreteras, 13, 14, 15 y 16 referentes al tendido eléctrico, 17 para puentes y 18 para ruido alto, cambiando a reservadas la 8 y la 12 (8 = *Model Key-point* (*mass point*); 12 = *Overlap Points*).

El gran problema que plantea esta clasificación a la hora de realizar cartografía básica es que la mayoría de los elementos a registrar cartográficamente se corresponden con puntos de suelo (*ground* = 2) tal como las carreteras; y, si tienen altura con edificaciones o vegetación (*building* = 6; *Medium vegetation* = 4; *High vegetation* = 3), lo que lleva a pensar que esta clasificación es insuficiente para fines cartográficos.

Tabla 3.5. Valores ASPRS para la clasificación del formato LAS versión 1.4

Clase	Significado	Tipo de elemento
0	*Created, never classified*	Nunca clasificado
1	*Unclassified*	No asignado
2	*Ground*	Suelo
3	*Low Vegetation*	Vegetación baja
4	*Medium Vegetation*	Vegetación media
5	*High Vegetation*	Vegetación alta
6	*Building*	Edificio
7	*Low Point (noise)*	Punto bajo (ruido)
8	*Reserved*	Reservado
9	*Water*	Agua
10	*Rail*	Vía férrea
11	*Road Surface*	Superficie pavimentada (carretera)
12	*Reserved*	Reservado
13	*Wire – Guard (Shield)*	Tendido
14	*Wire – Conductor (Phase)*	Cable de tendido
15	*Transmission Tower*	Torre transmisora
16	*Wire – structure Connector*	Conector de tendido
17	*Bridge Deck*	Cubierta de puente
18	*High Noise*	Punto alto (ruido)
19-63	*Reserved*	Reservado
14-255	*User definable*	Definido por el usuario

Sin embargo, no se debe olvidar que, tal y como señala Heidemann 2014, el esquema de clasificación mínima de puntos de una nube LiDAR no requiere ni edificaciones ni vegetación, ya que resulta suficiente con puntos de clase 1, 2, 7, 9, 10, 17 y 18, donde el 10 hace referencia

a los puntos ignorados del suelo (*ignored ground, near a breakline*), en lugar de a las vías férreas indicadas en la tabla anterior.

Además de los valores indicados para la clasificación, se dispone del registro *Classification Flags* que cuenta con 3 campos más de tipo booleano: *Synthetic*, *Key-point* y *Withheld*. En el caso de la versión 1.4. aparece también el registro *Overlap*. Éstos permiten definir el tipo de punto del que se trata, adquiriendo el valor 1 si se cumple lo especificado en la siguiente tabla, y el valor 0 (cero) en caso contrario.

Tabla 3.6. Valores booleanos de la codificación del campo de la clasificación

Synthetic	conjunto de puntos externo, creado por otra técnica no LiDAR
Key_Point	punto clave para la creación del modelo
Withheld	punto a obviar en el procesamiento
Overlap	Punto perteneciente a a una o más pasada

Al consultar la información correspondiente a un punto LiDAR los aspecto referidos previamente se muestra tal y como se puede apreciar en la figura 3.1. En este caso, el campo *Class code* hace referencia a la clasificación y *Classification Flag(s)* a los valores de *Synthetic*, *Key-point* y *Withheld* que tal y como se puede apreciar resultan nulos (*none*) por no cumplir las características definidas**.**

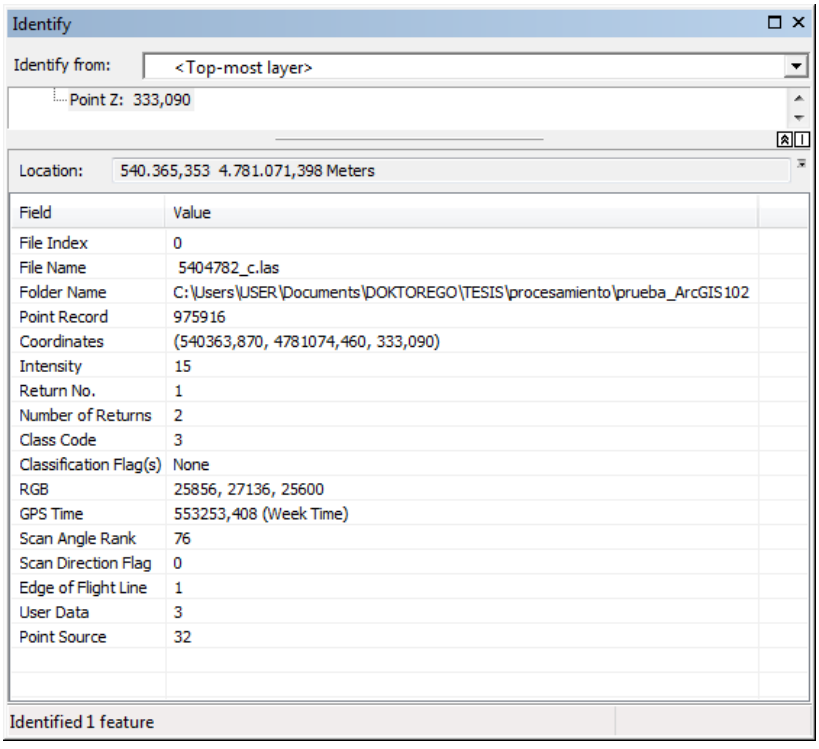

Figura 3.1. Información referente a un punto LiDAR al consultar la base de datos

Además, sea cual sea la versión del LAS los datos LiDAR sufren varios procesos para a partir de los datos brutos crear lo que se denomina el producto final (*end product*). Para catalogar el

grado de procesado que adquieren, la comunidad LiDAR ha definido múltiples niveles de procesamiento de estos datos (*LiDAR Data Processing Levels*), de manera que cada nivel describe el estado de procesamiento adquirido teniendo en cuenta que niveles superiores incluyen los estados anteriores (NGA 2011). En la tabla 3.7 se recogen los niveles existentes, denominación y una breve descripción de los mismos.

Tabla 3.7. Niveles de procesamiento de los datos LiDAR

	Level	*Designation*	Descripción
L0	*Level 0*	*Raw Data and Metadata*	Datos brutos y metadatos
L1	*Level 1*	*Unfiltered 3D Point Cloud*	Datos 3D
L2	*Level 2*	*Noise-filtered 3D Point Cloud*	Datos 3D sin ruido y con intensidad
L3	*Level 3*	*Georegistered 3D Point Cloud*	Datos 3D según un datum geodésico
L4	*Level 4*	*Derived Products*	Incluye productos derivados del LiDAR
L5	*Level 5*	*Intel Products*	Productos especializados

3.2. PROGRAMAS PARA CLASIFICAR DATOS LIDAR

Tras el análisis de los distintos tipos de algoritmos desarrollados para filtrar y clasificar los puntos de la nube de datos LiDAR vistos en la sección anterior, en este apartado se presentan los programas o librerías con capacidad de clasificar datos LiDAR más usuales en este ámbito relacionándolos con el tipo de algoritmo que utilizan.

En primer lugar, apuntar que en los últimos años la mayoría de los softwares comerciales que permiten el desarrollo cartográfico (CAD, SIG, Tratamiento de imágenes, restituidores, aplicaciones para generar MDT, etc.) están incorporando herramientas que admiten la integración y manejo de los datos LiDAR, pero no todos permiten la manipulación de esos datos.

De acuerdo a la información facilitada por NSF OpenTopography (Crosby 2011) el software comercial usado mayoritariamente en el ámbito productivo es TerraScan, por lo que se puede considerar cómo el software comercial que más se está utilizando para la clasificación de las nube de puntos procedentes de sistemas aerotransportados, tanto a nivel internacional como estatal. Este paquete se basa en el algoritmo desarrollado por Axelsson 1999 y la característica de trabajar con la aplicación CAD denominada MicroStation.

La gran ventaja que ofrece es que la aplicación de los algoritmos usados no está bloqueada al usuario, sino que dentro de las posibilidades que brinda, cada operario puede definir sus parámetros, los cuales pueden ser distintos según las características del terreno, en función de los resultados que se busquen. Además, esa personalización se puede guardar a modo de macro y se puede configurar para que funcione de manera automática o manual. En el apartado de filtros de densificación progresiva estudiado en la sección anterior ya se han indicado la manera de proceder que tienen tanto Brovelli and Lucca 2011 como la empresa

Omega Cartografía Digital S. L. de Pamplona, quedando patente que cada usurario puede seguir distintas secuencias utilizando diferentes parámetros.

Además de TerraScan, dentro del software comercial, caben mencionar los paquetes MARS y Quick Terrain Modeler (QTM) que además de ser visualizadores de los datos brutos permiten, entre otras cosas, editarlos y rasterizarlos. Como ventaja, señalar que resultan más baratos que TerraScan. En esta línea, a posteriori, aparece VRMesh, software que presenta una solución basada en la densificación progresiva para la clasificación automática de la nube de puntos teniendo en cuenta las especificaciones de la ASPRS, considerando las categorías de terreno, vegetación, edificación y otros.

En la tabla 3.7. se ha presentado una relación de precios de estos paquetes teniendo en cuenta una licencia operativa con los módulos necesarios en el caso de TerraScan y MARS para desarrollar el proceso necesario de la clasificación de puntos. En todos los casos menos en el de VRMesh los precios incluyen formación in situ.

Tabla 3.7. Relación de precios de una licencia de los software comerciales

Software	Precio (€)
TerraScan	13.000
Microstation	7.000
MARS	16.600
QTM	4.000
VRMesh	650

En el ámbito del software libre, indicar que también se han desarrollado para la visualización y tratamiento de este tipo de datos. Si bien la mayoría de ellas se quedan a nivel de visualización, cada vez son más las que ofrecen algunas posibilidades de edición e incluso clasificación.

A continuación se ofrece una pequeña reseña basada principalmente en la información facilitada en la referencia anterior (Crosby 2011) y complementada con otros artículos reseñados a posteriori.

- MCC-LIDAR, *Multiscale Curvature Classification for LiDAR data*. Procesa retornos discretos en ambientes boscosos y clasifica los puntos en terreno y no terreno usando el algoritmo MCC (Evans and Hudak 2007).

- GRASS, *Geographical Resources Analysis Support System*. Dispone de herramientas específicas desarrolladas por *Geomatic Laboratory of Poletecnico di Milano* para el procesamiento de datos LiDAR, análisis de datos anómalos, detección de ejes, generación de superficies y conversión de datos.

- BCAL Lidar Tools, *Boise Center Aerospace Laboratory*. Se trata de un software desarrollado para el procesamiento, análisis y visualización de los datos LiDAR. Incluye una herramienta de filtrado basado en las diferencias de altura discriminando puntos de suelo y de vegetación. Desarrollado en IDL (*Interactive Data Language*) está

optimizado para clasificar pastizales y artemisa (hierba de San Juan) (Streutker and Glenn 2006).

- SAGA GIS, *System for Automated Geoscientific Analyses*. Dispone de varias herramientas para manipular nubes de puntos: calcular atributos, reclasificar, extraer subconjuntos, mallado, interpolación, etc. Incluye también un filtro de suelo desnudo adaptado por (Vosselman 2001).

- FUSION, desarrollado por el equipo de Modelos Forestales y Silvicultura de *The United States Forest Service Pacific Northwest Research Station*. Se trata de una aplicación diseñada para analizar aspectos forestales, aunque incluye herramientas de aplicación más general. A partir de la nube de puntos LiDAR genera la superficie del terreno (*GroundFilter*) basándose en el algoritmo adaptado de (Kraus and Pfeifer 1998).

- ALDPAT, *Airbone LiDAR Data Processing and Analysis Tools*. Este software ha sido desarrollado por *The National Center for Airbone Laser Mapping (*NCALM*)* e implementa algunos de los distintos tipos de filtros comentados: filtros morfológicos progresivos, umbrales de elevación, superficies poli-nómicas, máxima pendiente local y PTD modificado y adaptado.

- LASTools, se trata del software desarrollado por Martin Isenburg para el procesamiento rápido de datos LiDAR. Permite, entre otras acciones, convertir entre diversos formatos, extraer información sobre la cabecera de los archivos, generar MDE y realizar clasificaciones basándose en el algoritmo desarrollado por Axelsson 1999. Algunas aplicaciones para usos comerciales son de pago.

- DielmoOpenLiDAR, extensión de gvSIG 1.1.2 que permite visualizar y realizar el control de calidad de la nube de puntos.

Tabla 3.8. Relación de software y algoritmos para la clasificación de datos LiDAR

Software	Grupo de filtro	Algoritmo
TerraScan	PTD	Minimum Description Length
VRMesh	PTD	
MCC LiDAR	Superficies	Multiscale Curvature Classification
GRASS	Segmentación	v.lidar.edgedetection v.lidar.growing v.lidar.correction
BCAL LiDAR Tools	Superficies	Perform Height Filtering
SAGA GIS	Morfológico	Maximun Local Slope (MLS)
FUSION LDV	Superficies	Robust interpolation
ALDPAT	Morfológicos	ETEW, MLS, PMF
LASTools	PTD	Minimum Description Length

Para concluir este apartado, en la tabla 3.8 se ha tratado de relacionar, en función de la información disponible, los algoritmos comentados en la revisión de los métodos de clasificación con los programas descritos en esta sección.

Para concluir esta revisión, y tal y como se ha visto en al punto anterior, resaltar que la mayor parte de los algoritmos permiten clasificar en terreno y no terreno siendo muy variado el tipo de algoritmo utilizado, incluso dentro de los cuatro grandes grupos existentes.

En este sentido, la ISPRS (*International Society for Photogrammetry and Remote Sensing*) que dispone de distintas comisiones técnicas con el objetivo de seleccionar artículos a presentar a congresos y establecer las conclusiones de las mismas cada cuatro años, en el 2003 presentó, dentro del grupo de trabajo 3 de la Comisión III (*Commission III, Working Group 3 = WG III/3*) denominado "3D Reconstruction from Airborne Laser Scanner and InSAR Data", el informe sobre la comparación de filtros (Sithole y Vosselman, 2003). En este documento se cotejan ocho filtros distintos en ocho zonas con características diferentes, de las cuales cuatro son urbanas y otras cuatro rurales, con el objetivo de identificar futuras direcciones de investigación en el filtrado de nubes de puntos para la extracción del suelo.

A posteriori, este tipo de comparaciones se han seguido realizando en diversos trabajos existiendo una gran variedad de artículos que comparan los distintos programas entre sí (Meng, et al. 2010; Montealegre, et al. 2013). Además, esta equiparación es mayor, si cabe, entre los paquetes libres, si bien alguno de ellos cotejan los resultados entre éstos y TerraScan (Brovelli and Lucca 2011; Chang, et al. 2008; Sithole 2002), utilizando en muchos de ellos las zonas de test establecidas por el *WG III/3*.

Además de los paquetes aquí reseñados, y tal y como se ha indicado en la introducción, en OpenTopography 2015 se puede completar este estudio con herramientas que hacen referencia a aplicaciones concretas como River Bathymetry Toolkit (RBT) o a programa de SIG o CAD como Global Mapper que tienen habilitadas utilidades para la visualización o tratamiento de este tipo de datos. También aparecen herramientas que permiten trabajar con este tipo de datos a través de la web o librerías como SPDlib o GDAL. Estas últimas como han sido utilizadas en la propuesta metodológica se comentan a continuación.

3.3. HERRAMIENTAS UTILIZADAS EN LAS METODOLOGÍAS DESARROLLADAS

En esta sección hay que diferenciar las herramientas utilizadas por un lado en la evaluación de la clasificación ASPRS de los datos LiDAR disponibles (capítulo quinto de la memoria) y las usadas para llevar a cabo la metodología de clasificación propuesta (título sexto), ya que en cada parte se ha trabajado con hardware y software diferente.

Del análisis de estas herramientas se puede deducir que salvo en el caso de FME y ArcGIS se ha hecho uso de herramientas libres, opción que confluye con el interés que en los últimos

tiempos están arrojando las alternativas de código abierto, las cuales están creciendo gracias a las posibilidades que bridan no sólo para probar los algoritmos existentes, sino para adaptarlos, en el caso de que fuese necesario, al facilitar su código; además de permitir el acceso a los últimos algoritmos desarrollados (Clewley, et al. 2014).

A continuación se han detallado los paquetes empleados en cada una de las dos etapas señaladas anteriormente, así como las características de los ordenadores empleados en cada una de ellas.

3.3.1. HERRAMIENTAS PARA LA EVALUACIÓN DE LA CLASIFICACIÓN ASPRS EN LOS DATOS LIDAR 2008

Para llevar a cabo la metodología explicitada en el quinto capítulo de esta memoria se ha utilizado un ordenador con las características especificadas en la tabla 3.9.

Tabla 3.9. Características del ordenador para verificar el estado de la clasificación ASPRS en los datos de la CAPV

Componentes	Características
Procesador	Intel (R) Core(TM) i7-2600 CPU @ 3,40 GHz
Memoria RAM	8.00 GB
Sistema operativo	Windows 7 Ultimate, 64 bits
Disco duro	WDC de 930 GB
Tarjeta gráfica	NVIDIA GeForce GTS 450

Tres han sido los programas que se han empleado: LAStools, FME y ArcGIS 10.1 y 10.2.

Las herramientas de LAStools están desarrolladas y mantenidas por Martin Isenburg y tienen como base la librería LASlib escrita en C++ y realizada bajo licencia LGPL (*Lesser General Public License*), permitiendo utilizarla en otros paquetes, incluso comerciales.

Las utilidades que presentan permiten gestionar archivos LAS, formato estándar en el que se entregan los datos LiDAR y que son comentados en el apartados de datos, y LAZ, que constituye el formato comprimido de los archivos LAS. Entre las funciones más importantes destacan las de transformación de datos de formato LAS a texto o ESRI shapefile (shp) y viceversa, obtención de información de la cabecera de los mismos, creación del índice espacial, generación de modelos digitales, extracción de puntos de una determinada zona, división o extracción de un puntos del fichero, generalización de puntos eliminando determinados, etc. En la tabla 3.10. se muestran las utilidades más importantes (Isenburg 2014).

En lo que respeta al presente trabajo, se han utilizado lasinfo.exe para obtener la información de la cabecera de los ficheros .LAS y poder así conocer el histograma de la clasificación como si se disponía de algún dato anómalo y lassplit.exe para obtener las pasadas que constituyen los ficheros .LAS procesados.

Cabe señalar que estas herramientas están en continua actualización y personalización en función de las necesidades que se van demandando. Las últimas aportaciones se pueden encontrar en la web http://rapidlasso.com/, y también en webs de redes sociales como facebook y twitter o grupos de discusión como Google Groups, todos ellos liderados por Martin Isenburg.

Tabla 3.10. Utilidades más importantes que ofrecen las herramientas de LAStools

Ejecutable	Descripción
lasinfo.exe	Ofrece información sobre el contenida del fichero LAS
lasview.exe	Visualiza el contenido del fichero LAS a modo de puntos o TIN
las2las.exe	Permite manipular los datos del fichero LAS
txt2las.exe	Transforma datos LiDAR en formato ASCII (.txt) a .LAS
las2txt.exe	Transforma datos LiDAR en formato .LAS a ASCII (.txt)
las2shp.exe	Transforma los datos en formato LAS a ESRI shapefile (.shp)
shp2las.exe	Transforma los datos en formato ESRI shapefile (.shp) .LAS
lasthin.exe	Generaliza el número de puntos del fichero. LAS
lastile.exe	Divide el fichero original .LAS en tiles
lasoverlap.exe	Analiza la superposición de las líneas de vuelo
lasnoise.exe	Elimina los puntos altos o bajos que constituyen ruido
lasclip.exe	Permite partir el fichero original .LAS en uno o muchos polígonos
lasboundary.exe	Extrae los puntos que pertenecen a un polígono dado
lasmerge.exe	Permite unir varios ficheros .LAS en uno único
lasduplicate.exe	Elimina los puntos duplicados
lassplit.exe	Divide el fichero original .LAS según el valor de alguno de sus parámetros
lasground.exe	Extrae puntos de suelo
lasheight.exe	Calcula la altura sobre el terreno de todos los puntos
lasclassify.exe	Clasifica los puntos en edificaciones y vegetación alta
lascanopy.exe	Calcula métricas necesarias en el ámbito forestal
las2tin.exe	Triangula los datos contenidos en un fichero LAS generando un TIN
las2dem.exe	Rasteriza la información para generar MDE
las2iso.exe	Extracción de curvas de nivel
laszip.exe	Comprimes y descomprime datos .LAS

Asimismo, se ha manejado FME Desktop del fabricante Safe Software. Se trata de un paquete que permite la conversión tanto de datos espaciales como no espaciales entre distintos formatos habilitando la integración y conexión entre ellos, pionero y líder global en el mercado tecnológico de transformación de datos espaciales.

Con él se han desarrollado exprofeso diferentes flujos de trabajo que han permitido la eliminación del ruido presentado en los datos LiDAR y que también se habría podido realizar con lasnoise.exe y la extracción de la información de la cartografía de referencia para la verificación de los resultados, así como algunos cálculos estadísticos necesarios para dicha valoración.

La visualización de estos resultados se ha realizado con ArcGIS 10.1 y 10.2 permitiendo los análisis correspondientes tras la incorporación de la información cartográfica conveniente: ortofotografías en formato .jpg y .tif y cartografía vectorial en formato ESRI shapefile (.shp). Ambos tipos de datos se describen en el siguiente capítulo.

3.3.2. HERRAMIENTAS UTILIZADAS EN LA PROPUESTA DE CLASIFICACIÓN

La propuesta metodológica explicada en el epígrafe sexto de esta memoria se ha llevado a cabo utilizado una estación de trabajo fija de Dell con 32 GB de RAM. En la tabla 3.11 se muestran las especificaciones de la misma.

Tabla 3.11. Características de la estación de trabajo utilizada en la metodología de clasificación propuesta

Componentes	Características
Procesador	Intel (R) Xeon(R) E5-1650 v2 (3,5 GHz, núcleo séxtuple, 12MB, Turbo, HT), estación de trabajo fija Dell Precision T3610
Controlador	SATA (2 × 6 GB/s, 4 × 3 GB/s) RAID de software 0/1/5/10
Memoria RAM	32,00 GB (4 × 8 GB) 1866 MHz DDR3 RDIMM ECC
Sistema operativo	Linux Mint, 64 bits
Dos discos duros	3 TB Serial ATA 7200 rpm 3,5" 2 TB 3,5" Serial ATA 7200 rpm 3,5"
Tarjeta gráfica	3 GB NVIDIA Quadro K4000 (2DP y 1 DVI-I)

Para su desarrollo se ha optado por utilizar librerías existentes accesibles desde el lenguaje de Python 2.7. Se ha elegido este software por constituir un lenguaje de programación interpretado y multiplataforma con una sintaxis sencilla y poseer licencia de código abierto denominada *Python Software Foundation License*, compatible con la licencia pública GNU (*GNU's Not Unix*). Como libro de consulta básico se ha utilizado el de Lutz 2013, además de los muchos recursos disponibles en internet.

Las labores a desarrollar en la metodología se dividen en dos grandes grupos: extracción de variables y aplicación de algoritmos de aprendizaje automático. Cada uno de ellos ha requerido el uso de librerías o paquetes informáticos específicos, utilizando en ambos casos el software QGIS Desktop en su versión 2.8.2. Wien para la visualización y análisis de los resultados. Se trata de un Sistema de Información Geográfica gratuito y de código abierto que permite el uso de aplicaciones adicionales haciendo uso de varias librerías libres (QGIS 2015).

Respecto al formato en el que se han guardado las variables, se ha elegido el formato csv (*comma-separated values*) por tratarse de una configuración abierta y sencilla para la representación de datos en forma de tabla. En él las distintas variables se corresponden con columnas y aparecen separadas por comas; a su vez, cada una de las muestras hace referencia a una fila y se distingue de la siguiente gracias a un salto de línea.

A continuación se reseñan brevemente las librerías utilizadas en el caso de la extracción de variables:

- SAGA, *System for Automated Geoscientific Analyses*. Ha sido desarrollado por el *Department Physical Geography de Göttingen* y constituye un software libre (*Free Open Source Software*, FOSS) que dispone de herramientas para análisis científicos digitales de la Tierra (Dept. of Physical Geography, Göttingen 2015).

  Esta utilidad ya ha sido comentada en el apartado de programas para trabajar con datos LiDAR, pero su utilidad se ha centrado en la asignación de valores a puntos y el cálculo de estadísticos.

- GDAL, *Geospatial Data Abstraction Library*. La librería GDAL constituye una herramienta de licencia libre que proporciona la transformación de datos geo-espaciales, tanto ráster como vectoriales, a distintos formatos o proyecciones; aunque también dispone de algunas otras opciones de procesamiento (GDAL 2015).

  Concretamente, en este trabajo se han extraído las bandas de las imágenes originales y realizados algunos cálculos entre ellas.

- libLAS. Se deriva de la librería LASlib y actualmente está bajo la protección de OSGEO. En sus inicios fue desarrollada por *Iowa Geological Survey* (IGS) para la lectura y escritura del formato .LAS de los datos LiDAR (IGSB 2015). Está escrita en C/C++ e incorpora *bindings* para muchos lenguajes de programación y ofrece licencia de software libre BSD (*Berkeley Software Distribution*). Muchas de las herramientas que disponen se llaman de igual manera que las de LAStools, aunque la utilización de los comandos no es exactamente igual.

  De las utilidades que ofrece se ha empleado el comando las2txt que permite trascribir los ficheros .LAS a formato de texto (.txt), para luego utilizar esta información en un entorno cartográfico.

- SPDlib, *Sorted Pulse Data Library*. Constituye un conjunto de herramientas abiertas para el procesamiento de datos LiDAR, tanto de retorno discreto como de *full-waveform*, bien capturados desde plataformas aéreas o terrestres. Ha sido desarrollada en C++ por Pete Bunting con *bindings* para Python e IDL y con licencia GPL (*General Public License*). Una de las particularidades que presenta es que no utiliza el formato .LAS para procesar los datos y éstos hay que transformarlos al formato SPD (*Sorted Pulse Data*) (Bunting, et al. 2013b). El flujo de trabajo que se puede desarrollar con esta librería se puede visualizar en Bunting, et al. 2013a.

  Esta librería se ha utilizado para clasificar los puntos LiDAR en terreno / no terreno según los algoritmos PMF y MCC explicados en la sección dos de este documento, previa transformación del formato .LAS a .SPD, eliminación de puntos anómalos y limpiar la clasificación existente en los datos. Finalmente, con cada algoritmo se han generado los MDT y MDS obteniendo los ficheros ráster en formato .tif.

- Orfeo ToolBox (OTB). Esta librería ha sido desarrollada por el *Centre National d'Etudes Spatiales* (CNES) y es distribuida bajo licencia CeCILL license, constituyendo una agrupación de herramientas para el procesamiento de imágenes de teledetección (CNES 2015) principalmente, aunque también permite el acceso a datos LiDAR o imágenes aéreas. Está implementada en C++ y dispone de una API (*Application Programming Interface*) para Python.

  Del total de opciones disponibles se ha empleado la que hace referencia a la segmentación de Edison para aplicar a las ortofotografías y los ráster derivados de los MDE obtenidos a través de SPDlib.

Para la aplicación del aprendizaje automático, se han probado las siguientes alternativas:

- Weka, *Waikato Environment for Knowledge Analysis*. Constituye una colección de algoritmos de aprendizaje automático para tareas de minería de datos escrito en Java y desarrollado por la Universidad de Waikato, Nueva Zelanda. Se trata de un software de fuente abierta bajo licencia GNU *General Public License* (GNU-GPL).
- RapidMiner Studio. Se trata de un software de fácil uso con un entorno visual diseñado para el análisis de datos para minería de datos, el análisis predictivo y la inteligencia empresarial que no requiere programación. En sus inicios comenzó como un programa gratuito de código abierto, pero actualmente se trata de una plataforma de software propietario.
- Knime o *Konstanz Information Miner*. Este software está construido bajo la plataforma Eclipse y programado en Java. Esta herramienta de carácter abierto, integra componentes para minería de datos permitiendo emplear utilidades de R o Python desarrollando modelos en un entorno visual.
- scikit-learn o scikits.learn. Constituye una librería de código abierto para el aprendizaje automático aplicable en el lenguaje de programación de Python bajo licencia BSD (*Berkeley Software Distribution*) (Pedregosa, et al. 2011). Está diseñada para interoperar con las librerías NumPy, SciPy y matplotlib de Python. Dispone de algoritmos de clasificación, regresión, *clustering* o agrupamiento, reducción dimensional, selección de modelos y pre-procesado.

Finalmente, a pesar de la gran variedad de herramientas de minería de datos disponibles en R y aunque también se podría haber escogido programarlo en cualquier tipo de lenguaje o software matemático tal como pueden ser C++ o Matlab, tras varias pruebas realizadas con Weka, RapidMiner Studio y Knime se ha optado por la creación de un script en Python 2.7, haciendo uso de las utilidades que ofrece la librería scikit-learn.

Para su desarrollo se ha hecho uso del entorno Spyder 2.3 que constituye un potente contexto de tratamiento interactivo para el lenguaje de programación Python, disponiendo de herramientas de visualización de variables, autocompletado, edición avanzada y depuración interactiva, disponible tanto para Windows como para Linux (Spyder 2015). En este caso se ha utilizado bajo Linux Mint.

4.DATOS Y ÁREAS DE ESTUDIO

En este capítulo se presenta la Comunidad Autónoma del País Vasco (CAPV) por constituir el ámbito del que se dispone los datos a analizar y procesar, así como una descripción del formato en el que se presentan los datos LiDAR con sus particularidades.

Del total de datos disponibles se marcan las áreas de estudio tanto para la evaluación de los datos LiDAR existentes como para la utilización en la propuesta metodológica presentada en el documento.

4.1. INTRODUCCIÓN

Para poder llevar a cabo la verificación de la clasificación automática que ofrecen los datos LiDAR aerotransportados del Gobierno Vasco (GV) en base a las premisas de la ASPRS (*American Society for Photogrametry and Remote Sensing*) y poder llegar a mejorarla, no sólo hay que tener en cuenta estos datos en sí, sino también se debe hacer uso de información complementaria que permita su contrastación y si se pudiera una mejoría. A su vez, resulta necesario establecer las zonas en las que se va a proceder a su examen, justificando su elección y particularidades si las hubiera.

Esos datos se ubican dentro del área que ocupa la Comunidad Autónoma del País Vasco (CAPV), también denominada Euskadi. Esta comunidad está situada en el borde cantábrico nororiental y linda al norte con el mar Cantábrico, concretamente con el Golfo de Vizcaya y Francia (Aquitania). Al sur con la provincia de La Rioja, al oeste con las de Burgos y Cantabria y al este con la de Navarra (figura 4.1).

Figura 4.1. Ubicación de la Comunidad Autónoma del País Vasco

Esta comunidad está constituida por tres territorios históricos: Álava / Araba, Gipuzkoa y Bizkaia; de manera que cada uno de ellos a su vez cuenta con siete comarcas, siendo Bizkaia la de mayor población. Según el INE 2015 en 2015 la población total ascendía a 2.164.311 habitantes, con una densidad de 299,19 hab/km^2, teniendo en cuenta que toda la comunidad ocupa una extensión total de 7.234 km^2.

Por usos del suelo, según el informe realizado por hazi entre 2010 y 2011 la superficie forestal, incluyendo la arbolada y la desarbolada (pastizal, matorral y roquedo) alcanza un 68 % del total de la CAPV, de la cual el 54,9 % es arbolada y el 13,1 % desarbolada. Le sigue la superficie agraria (cultivos y prados de siega) con un 24,9 %, la superficie urbana o infraestructuras con un 6,3 % y los improductivos ligados al agua con un 0,7 % (figura 4.2). Por territorios, Gipuzkoa es el que mayor porcentaje forestal tiene y en Bizkaia predomina la superficie urbana (hazi 2011).

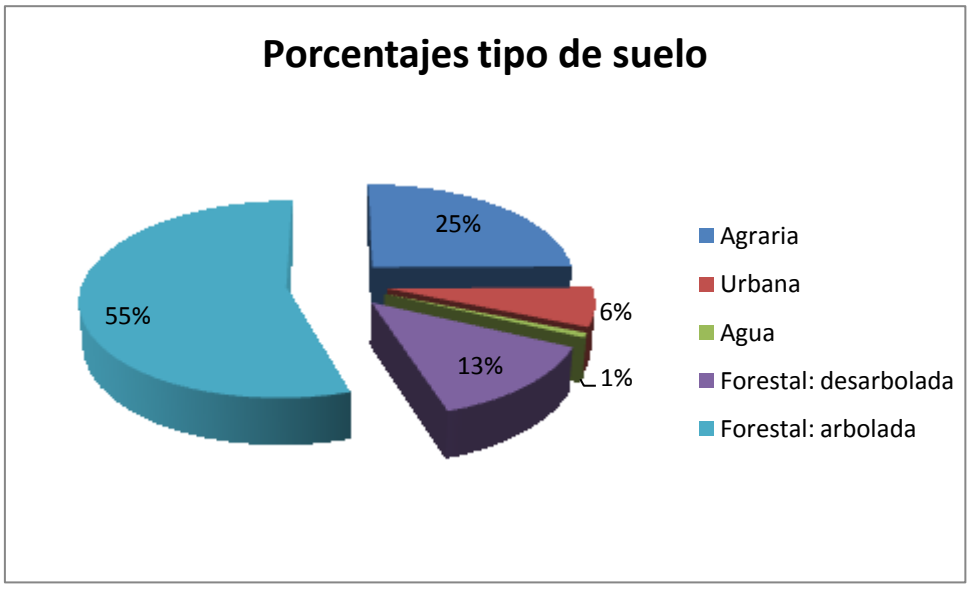

Figura 4.2. Distribución por tipos de suelo de la Comunidad Autónoma Vasca en el 2011

La orografía predominante es montañosa, conformada al norte por los Montes Vascos y al Sur por la Sierra de Cantabria. El punto más alto se encuentra en la Sierra de Aitzgorri ubicada en los Montes Vascos y concretamente en el pico Aitxuri con una altitud de 1.551 m.

En cuanto al clima, se distinguen a grandes rasgos cuatro zonas climáticas: al norte la vertiente atlántica, al sur la de tipo continental, al occidente de Álava y en la llanada Alavesa clima sub-atlántico y una zona de clima sub-mediterráneo al este. Dadas estas regiones climáticas la flora del País Vasco se divide principalmente en dos áreas: por un lado, la zona cantábrica con bosque caducifolio (hayas, robles y castaños); y, por otro lado, el centro y sur de Álava donde predominan los bosques perennifolios de pinos negros y abetos, pinos silvestres y encinas. La fauna autóctona la constituyen el jabalí, zorro, liebre y conejo. Entre las aves destacan la perdiz, la cigüeña, la águila real, la paloma torcaz y el búho real y en los ríos prevalece la trucha, el barbo y el cangrejo (Buchot 2015).

Atendiendo a la geografía urbana existen tres tipos de asentamientos en los que se abarca al conjunto de los núcleos de población vascos: los situados en ladera o pendiente sobre un río o arroyo, los asentados en un cerro o meseta próximos a los ríos y los que se extienden por la parte baja de las vegas y llanos de mayor o menor extensión (Piñeiro 1993).

Definidas las características generales del entorno geográfico en el que se ubican los datos a tratar, se va a proceder a describir los datos empleados y posteriormente a detallar las áreas de estudio.

4.2. DATOS EMPLEADOS

Analizadas las bases de datos disponibles, se ha decidido hacer uso, además del vuelo del LiDAR del 2008, de las ortofotografías del mismo año y de la Base Topográfica Armonizada (BTA) del 2011, descargadas a través del ftp del Gobierno Vasco (http://www.geo.euskadi.eus/s69-geodir/es/contenidos/informacion/servicio_ftp/es_80/servicio_ftp.html). En los siguientes apartados se explican las características de cada uno de ellos por separado.

4.2.1. VUELO LIDAR

El primer vuelo LiDAR realizado en la CAPV fue realizado en varias fases. En una primera etapa, a instancias de la Diputación Foral de Gipuzkoa (DFG) se llevo a cabo el vuelo del territorio histórico de Gipuzkoa ejecutado por el Institut Cartogràphic de Catalunya (ICC), y posteriormente a petición del GV el de las provincias de Araba / Álava y Bizkaia por la empresa Azimut. En la tabla 4.1. se pueden apreciar los distintos vuelos realizados.

Tabla 4.1. Relación de vuelos LiDAR que constituyen el denominado LiDAR 2008

Zonas	Fechas de vuelo
Gipuzkoa	Febrero - mayo del 2005
Gran Bilbao	16 - 19 de septiembre del 2007
Resto de Bizkaia y zona norte de Álava	8 y 13 de febrero del 2008
Resto de Álava	18 junio al 10 julio del 2008

Tabla 4.2. Componentes sistema LiDAR 2008, Álava y Bizkaia

ComponenteS	Instrumentación
Escáner láser	Lite Mapper 5600
Sistema de navegación	CCNS-4 de IGI
GPS	NovAtel OEM 4-G2 L1/L2 2 Hz
INS	AeroControl IID. Frecuencia de registro 256 Hz

Para su ejecución, en el caso de Bizkaia y Álava, se hizo uso de los componentes indicados en la tabla 4.2, siendo sus características, en cuanto a la configuración del sistema LiDAR, las mostradas en la tabla 4.3. (Azimut 2008).

Tabla 4.3. Parámetros de configuración del vuelo LiDAR 2008 de Álava y Bizkaia

Características	Valores
Ángulo de escaneo	60° deg
Frecuencia de misión de pulsos	110 kHz
Tasa de repetición del pulso (PRR)	120.000 Hz
Divergencia del rayo	0,5 mrad

El ángulo de escaneo determina el valor angular con el que el pulso es emitido perpendicularmente a la línea de vuelo. En función de este valor se fija el campo de visión (*Field Of View*, FOV), oscilando los valores óptimos entre 0° y 75° deg. La frecuencia de emisión de pulsos indica cuántos pulsos son emitidos en un segundo. Este parámetro influye en la densidad de la nube de puntos, permitiendo vuelos más rápidos y más altos a la vez que reduce los tiempos de vuelo y los costes de adquisición. Está se relaciona con la tasa de repetición del pulso indicando cuánto se repite por segundo, lo que marca la resolución espacial del pixel. Actualmente suelen operar con frecuencias superiores a 150 kHz.

La divergencia del rayo se refiere a la desviación de los fotones de la línea de propagación teórica del rayo. A mayor distancia el diámetro del rayo será mayor, por eso la razón señal-ruido será mejor cuanto menor valor adquiera, suele variar entre 0,1 y 1 mrad. Atendiendo a todos estos valores se puede decir que se dispone de un vuelo adecuado, aunque se debería mejorar la frecuencia de emisión de pulsos para conseguir mayor resolución espacial.

Analizando los informes de los distintos vuelos se puede comprobar que la verificación de éstos dio como resultados una densidad de 1 ó 2 (en grandes ciudades) puntos por metro cuadrado en el caso de Gipuzkoa (ICC 2005), y una media de más de 2 puntos en el resto. En la tabla 4.4. se pueden apreciar los resultados detallados que según Sigrid 2008 se obtuvieron en el vuelo de Álava y Bizkaia.

Tabla 4.4. Valores sobre el control de densidad de puntos del vuelo LiDAR 2008 en Álava y Bizkaia

	Álava	Bizkaia
Nº retículas 1 km × 1 km	3059	1940
Densidad Media	3,18 ptos / m^2	2,16 ptos / m^2
Densidad Máxima	36,45 ptos / m^2	8,76 ptos / m^2
Densidad Mínima	0,32 ptos / m^2	1,01 ptos / m^2
Desviación estándar	1,11 ptos / m^2	0,60 ptos / m^2

A partir de esos datos, se desprende que además de que la zona volada en Álava (3059 retículas) es superior a la de Bizkaia (1940 retículas), la densidad media (3,18 ptos / m^2) y máxima (36,45 ptos / m^2) de Álava resulta mejor que la de Bizkaia (media = 2,16 ptos / m^2; máxima = 8,76 ptos / m^2) lo que puede ser debido a las condiciones topográficas de las zonas,

por tratarse Álava de un territorio en general más llano que el de Bizkaia. Por su parte, la densidad mínima y la desviación estándar ofrecen mejores resultados en la zona de Bizkaia (mínima = 1,01 ptos / m^2; desviación = 0,60 ptos / m^2), ya que en la zona de Álava (mínima = 0,32 ptos / m^2; desviación = 1,11 ptos / m^2) se da una mayor dispersión. En cualquier caso, haciendo referencia a la densidad media se puede decir que el vuelo cumple con lo marcado de dos puntos por metro cuadrado.

Estos vuelos fueron planificados de forma que al menos se disponga de dos puntos por m^2 y asegurando un error global horizontal inferior a 0,40 m de RMSE y de 0,15 m en cuanto al error vertical, desechándose errores superiores a 0,75 m en planimetría y 0,50 m en altimetría en los casos más desfavorables (fuerte pendiente y densa vegetación), valores exigidos en la mayoría de los vuelos de este tipo realizados hasta el 2008.

Los datos derivados de estos vuelos se pusieron a disposición del público en general como el **vuelo LiDAR del 2008** (en adelante LiDAR 2008) a partir de junio del 2012 a través del ftp del GV en ficheros de 1 km × 1 km con un total de 8.212 cuadrículas incluyendo el Condado de Treviño. La denominación de estas cuadrículas corresponde a los kilómetros de la coordenada X (las tres primeras cifras de este valor) seguida de los kilómetros de la coordenada Y (los cuatro primeros números de esta coordenada) de la esquina superior izquierda de la misma. Además, para diferenciar las cuadrículas procesadas por el GV de las procesadas por la Diputación Foral de Gipuzkoa (DFG) a éstas últimas se les ha añadido el término _Gipuzkoa.

Atendiendo a lo especificado en el punto 3.1. (El estándar de datos LiDAR) y de acuerdo a las especificaciones de la ASPRS para la distribución de las nube de puntos capturadas por sistemas de *Laser Ranging and Scanning* (detección y escaneado láser), se trata de ficheros que hacen referencia al formato LAS (*Log ASCII Standard*) en su versión 1.2, con el tipo de registro para los datos 3 (*Point Data Record Format,* PDRF) señalados en la tabla 3.2.

En lo que respecta a la información de los campos R, G, B, hay que señalar que como durante la realización del vuelo LiDAR no se capturaron simultáneamente imágenes, éstos se han rellenado durante la fase de procesamiento de los datos LAS, usando para ello las fotografías aéreas de alta resolución espacial efectuadas dentro del PNOA para el año 2010.

Además, los datos se corresponden al nivel 3 (*Level 3, Georegistered 3D Point Cloud*) de procesamiento, en el que se presenta la información con metadatos referida a un datum geodésico, en un sistema de coordenadas determinado y con una clasificación automática, aunque esos valores no están especificados en la cabecera de los ficheros.

El sistema de referencia utilizado es ETRS-89 (*European Terrestre Referente System 1989*) y el sistema de coordenadas hace referencia a la proyección UTM (*Universal Transversa Mercator*), huso 30 N, en la franja de paralelos correspondiente a la letra T. Las altitudes se corresponden con valores ortométricos en base al nuevo modelo de geoide EGM08_REDNAP.

Respecto a la clasificación automática utilizada, señalar que sigue los estándares considerados por la ASPRS en su versión 1.2, recogidos en la tabla 4.5, y que se ha obtenido tras el

procesamiento con TerraScan. Según fuentes del Gobierno Vasco, se trata de una clasificación supervisada (Gobierno Vasco 2013).

Tabla 4.5. Valores de clasificación del LiDAR 2008

Clase	Tipo de elemento
0	Nunca clasificado
1	No asignado
2	Suelo
3	Vegetación baja
4	Vegetación media
5	Vegetación alta
6	Edificación
7	Punto bajo (Ruido)
8	Punto clave
9	Agua
10	Reservado para la definición de la ASPRS
11	Reservado para la definición de la ASPRS
12	Puntos solapados
13-31	Reservado para la definición de la ASPRS

A posteriori, en verano del año 2012 se realizó el segundo vuelo LiDAR de la CAPV (**LiDAR 2012**). En este caso, la densidad de puntos especificada fue de medio punto por metro cuadrado (0,5 ptos / m^2), la mitad de lo requerido en el LiDAR 2008, consiguiendo 0,7 ptos / m^2. En la tabla 4.6 se especifican los requerimientos de dicho vuelo.

Tabla 4.6. Parámetros de configuración del LiDAR 2012

Características	Valores
Ángulo de escaneo	60° deg
Frecuencia de misión de pulsos	110 kHz
Frecuencia de escaneo	70 Hz
Múltiples retornos	hasta 4

En este caso, los ficheros también se distribuyen a través del mismo ftp que el del LiDAR 2008 y responden a las mismas características que éste: ficheros LAS según la versión 1.2 con la clasificación correspondiente (tabla 4.5), procesamiento con TerraScan, tipo de registro para los datos el 3 (*Point Data Record Format,* PDRF) señalados en la tabla 3.2 y el nivel de procesamiento el 3 (*Level 3, Georegistered 3D Point Cloud*), que significa que están georreferenciados según ETRS89-UTM30N y altitudes ortométricas conforme al modelo de geoide EGM08_REDNAP.

La diferencia principal entre ambos vuelos se refiere a la densidad de puntos lo que influye en la resolución espacial, por lo que la distribución de los datos en lugar de darse kilómetro por kilómetro en el LiDAR 2012 se da cada dos kilómetros, resultando algo menos de un tercio de las hojas existentes en el LiDAR 2008 (2.506 hojas, incluidas las del Condado de Treviño).

Este vuelo ha servido prácticamente para detectar zonas cambios producidas entre los años 2008 y 2012. La manera de proceder se ha basado en comparar el MDT 2008 con un MDT 2012 realizado de manera automáticas, detectando aquellas zonas en las que la diferencia de altura entre ellos resulta superior a un metro. Esas áreas se cotejan con la ortofotografía del 2012 y si se corrobora dicho cambio se procede a modificar los puntos en base al LiDAR 2012 para actualizar el MDT, teniendo como referencia el vuelo fotogramétrico 2012.

El presente trabajo se basa en los datos del LiDAR 2008, principalmente porque cuando se inicia esta investigación todavía no existían los datos del LiDAR 2012, ni si quiera se había hecho el vuelo. Pero, aunque hubiera existido hubiera sido mejor el del 2008 que el del 2012 porque, en cuanto a la clasificación se refiere, se basa en una categorización supervisada, objeto de valoración en este estudio. Además, al tener una densidad mayor, que supone mayor resolución espacial, permite valorar de forma más precisa los resultados alcanzados.

4.2.2. ORTOFOTOGRAFÍAS

Desde al año 2006 el Servicio de Información Territorial de la Dirección de Planificación Territorial y Urbanismo del Departamento de Medio Ambiente y Política Territorial del Gobierno Vasco se encuentra inmerso, bajo la Dirección General del Instituto Geográfico Nacional (IGN) y el Centro Nacional de Información Geográfica (GNIG) dependientes del Ministerio de Fomento , en el Plan Nacional de Ortofotografía Aérea (PNOA) con el objeto de obtener **Ortofotografías Aéreas Digitales** con resolución espacial de 25 ó 50 cm y Modelos Digitales de Elevaciones (MDE) de alta precisión para la comunidad autónoma con un período de actualización de 2 ó 3 años. Se trata de un proyecto cooperativo y cofinanciado entre la Administración del Estado y las Comunidades Autónomas (Ministerio de Fomento 2015).

Para la generación de estas ortofotografías aéreas digitales resulta necesario el vuelo fotogramétrico aéreo que no sólo establece la base de dichas ortofotografías, sino que también constituye el origen para la realización de la cartografía básica y en general para la recolección de información geográfica, dando pie a la creación de bases de datos cartográficas y geográficas con una perfecta coherencia geométrica y temporal.

Así, desde el año 2008 la comunidad autónoma dispone de vuelos fotogramétricos anuales a partir de los cuales se han realizado las Ortofotografías Aéreas Digitales del territorio (a posteriori ortofotografías), que han propiciado junto con el LiDAR 2008 la generación del MDE de alta precisión.

Como el objeto de este trabajo se centra en el análisis de los datos LiDAR 2008, se han elegido también las ortofotografías derivadas del vuelo fotogramétrico del 2008, disponiendo de una ortofotografía para toda la comunidad autónoma del País Vasco con 25 cm de pixel en formato

.ECW (*Enhanced Compressed Wavelet*), generada a partir del vuelo fotogramétrico digital efectuado entre julio y octubre del 2008.

Además de ésta, también se dispone con la misma resolución espacial de las ortofotografías correspondientes a las hojas de la cartografía a escala 1:5000 con las bandas Roja (*Red*, R), Verde (*Green*, G) y Azul (*Blue*, B) en formato .jpg (*Joint Photographers Group*) y .tiff (*Tagged Image File Format*), la de la banda NIR (*Near InfraRed*) en formato .tiff y la composición en falso color IrRG en .jpg.

Los formatos .ECW, .tif y .jpg son típicos para el almacenamiento de imágenes, resultando el .tif más adecuado frente al .jpg debido a que este último conlleva un grado de compresión del que carece el .tif y en cartografía suelen ir acompañados por sus archivos de georreferenciación .tfw para el .tiff y .jgw para .jpg. Por su parte, el formato .ECW conlleva internamente la georreferenciación y suelen ser adecuados para abarcar grandes extensiones ya que aunque conllevan compresión, ésta no suele ser excesiva y por su configuración permiten la realización de zooms de manera bastante rápida.

En esta investigación se ha hecho uso de las ortofotografías de las hojas 1:5.000 en las bandas RGB en formato .jpg y también de las mismas en la banda NIR en formato. tiff con una resolución espacial de 25 cm, en el sistema de referencia ETRS89 y la proyección UTM huso 30 N. Las ortofotografías RGB han sido obtenidas a través del ftp disponible para la descarga de cartografía, mientras que la banda NIR ha sido solicitada al propio servicio.

4.2.3. CARTOGRAFÍA AUTONÓMICA A ESCALA 1:5.000

La Comisión de Normas Cartográficas (CNC) del Consejo Superior Geográfico (CSG), tras publicar en 1992 las Normas Cartográficas para la Elaboración de Cartografía y comprobar que son insuficientes ante las nuevas tecnologías, métodos de producción, distribución y uso de la Información Geográfica (IG) y con la necesidad de homogeneizar la IG de manera que permita la toma de decisiones sobre el territorio a escalas local, regional y global para la creación de la Infraestructura de Datos Espaciales (IDE), junto con el Institut Cartogràphic de Catalunya (ICC) y la participación de las Comunidades Autónomas (CC.AA.), se inician a partir del año 2008 en la generación de una base topográfica vectorial para grandes escalas que esté armonizada y garantice la interoperabilidad y la integración de la IDE, surgiendo la denominada Base Topográfica Armonizada (BTA) que pretende ser el referente estatal para implementar la Directiva Europea INSPIRE en España.

El objetivo de la BTA consiste en conseguir un elevado grado de homogeneización de la cartografía oficial a gran escala entre ellas y asimismo compatibles con las bases cartográficas oficiales de escalas medias y pequeñas a nivel nacional, haciéndola así compatible con las normas europeas e internacionales sobre información geográfica. Por eso, la propuesta inicial pretende armonizar las bases topográficas regionales en base a la elaboración de la BTA,

basándose en las cartografías regionales existentes a escalas 1:5.000 y 1:10.000 (Barrot, et al. 2009).

Esta cartografía se debe realizar en el sistema de referencia ETRS89 usando como sistema de representación la proyección UTM, concretamente UTM30N. El formato debe ser ESRI shapefile (shp), organizando los datos en distintos temas que a su vez poseen fenómenos con diferentes geometrías de punto, línea y polígono. La relación de temas contemplados son los siguientes:

- **Puntos de referencia**: incluye los puntos que forman parte de los sistemas de posicionamiento geodésico oficiales.
- **Nombres geográficos**: abarca los nombres geográficos y los textos cartográficos.
- **Transportes**: comprende las vías de comunicación y las infraestructuras asociadas.
- **Hidrografía**: incorpora las masas de agua tanto naturales como artificiales, incluso los puntos de interés hídrico.
- **Relieve**: contiene las curvas de nivel, puntos de cota significativos y líneas de ruptura.
- **Cubierta terrestre**: recoge un conjunto reducido de cubiertas del suelo y vegetación consistentes en el catálogo del Sistema de Información sobre Ocupación del Suelo de España (SIOSE).
- **Edificaciones, poblaciones y construcciones**: engloba los elementos construidos como edificaciones, campos de deportes y cerramientos.
- **Servicios e instalaciones**: envuelven las redes de suministro energético, combustible y telecomunicaciones.

Dentro de cada tema aparecen distintos elementos geográficos como puede ser, en el caso de la cubierta terrestre, el arbolado forestal y las coberturas húmedas, entre otros. Para diferenciar uno de otro se utiliza un código constituido por cuatro números, que en las bases de datos se identifican con el campo ID_TIPO, por lo que cada elemento geográfico distinto adquirirá un valor diferente que le servirá para identificarlo.

La BTA de la CAPV es presentada por el Gobierno Vasco (GV), en su primera versión, en diciembre de 2012, tras la adecuación de la cartografía disponible hasta entonces en las Diputaciones Forales (DD.FF.). En los últimos tres años el GV ha presentado nuevas versiones de la BTA, siendo la última disponible, en el momento de la redacción de esta memoria, la de diciembre de 2014.

Su distribución se realiza a través del ftp que el GV dispone para la descarga de cartografía del territorio, concretamente en el apartado de cartografía básica. Está organizada de manera que ofrece la información, por un lado en coberturas completas, por tratarse de una cartografía continua, de toda la comunidad a escalas 1:5.000, 1:100.000, 1:200.000, 1:400.000 y 1:1.000.000, constituyendo la 1:5.000 la base de las otras; y, por otro, organizada en hojas a escalas 1:5.000 según los cortes geodésicos y denominaciones dictadas por las especificaciones de la BTA (Consejo Superior Geográfico 2008)

De cara a la investigación se han valorado dos opciones, claramente diferentes:

- Proceder a seleccionar una o varias zonas, y clasificar manualmente todos los puntos.
- Emplear una base cartográfica vectorial existente con cobertura en toda la CAPV, con temas asimilables a los de la clasificación de los datos LiDAR.

La primera opción tiene como principal ventaja que la clasificación podría ser muy precisa pero, debido al carácter manual de la misma, su extensión debería ser necesariamente pequeña. Por el contrario, la segunda opción, al abarcar toda la comunidad autónoma, debe ser más generalista lo que supone una disminución de la precisión.

En esta investigación se ha optado por esta última alternativa al considerar que se trata de una cartografía homogénea para todo el territorio de la CAPV, a pesar de que el contraste con respecto a los datos LiDAR resulte menor, permitiendo que el entrenamiento de las reglas de clasificación se efectúe en zonas más amplias.

La única disponible, a la mayor escala posible, con topología formada y atributos para las entidades es la Base Topográfica Armonizada (BTA) a escala 1:5.000, cuyas características ya se han comentado. En lo que respecta a la versión considerada, dado que los datos LiDAR utilizados son los del año 2008, indicar que se ha hecho uso de la primera versión de la BTA facilitada por el GV.

Su uso va a servir para corroborar los valores que presentan los puntos de la nube de datos LiDAR en el campo de la clasificación (apartado 5) y para entrenar los datos de cara a una mejora en la clasificación del LiDAR 2008 con algoritmos de aprendizaje automático (apartado 6).

Al empezar a trabajar con esta base cartográfica, lo primero que se ha hecho ha sido analizar las geometrías disponibles, y de las tres existentes se ha considerado que la relevante la constituye la de polígonos, ya que para el resto de geometrías resulta difícil determinar qué puntos del LiDAR se corresponden directamente con los elementos que definen los fenómenos establecidos por la BTA.

Respecto a los temas, de todos ellos en una primera instancia se han descartan los puntos de referencia y nombres geográficos por no corresponderse con elementos geográficos concretos y tratarse, en su mayoría, de elementos puntuales, con los que existe poca posibilidad de coincidencia con los puntos tomados por el sensor LiDAR, ya que éste sensor los toma de manera irregular según el patrón del instrumento considerado y en cartografía se representan localizaciones muy concretas como las esquinas o bordes de elementos.

Asimismo, como de lo que se trata es de clasificar la información geográfica, se ha desestimado el fenómeno referente al relieve que en su geometría superficial tiene más que ver con temas de pendiente que con los elementos geográficos propiamente dichos. Tampoco se ha considerado el fenómeno de servicios e instalaciones por entender que comprenden redes que en muchos casos se encuentran soterradas.

En el caso de la hidrografía, aunque se ha previsto que la correspondencia va a ser mínima por utilizar un sensor que no capta puntos de agua ya que la señal láser es absorbida por este tipo de superficie, se mantienen las capas poligonales al comprobar que en las clasificaciones de las distintas versiones del formato LAS se considera esta categoría.

En el caso de la BTA del GV a parte de la cobertura de edificaciones aparece la de elementos construidos que en las especificaciones de la BTA se considera dentro de las edificaciones. Los elementos construidos hacen referencia a emplazamientos o áreas deportivas, elementos que en algunos casos estarán constituidos por casetas o pabellones; y, en otros no, como es el caso de las pistas deportivas. De cualquier modo, en este trabajo también se ha considerado esta capa por entender que se podría dar en ella la existencia de pabellones o construcciones similares que en el LiDAR 2008 apareciesen como puntos de edificaciones.

En consecuencia, en esta investigación se ha hecho uso de algunos de los elementos superficiales considerados dentro de los fenómenos de Transportes, Hidrografía, Cubierta terrestre y Edificaciones.

En el caso de la corroboración de los valores asignados en la clasificación del LiDAR 2008 los valores de la BTA se han utilizado como referencia para determinar los elementos bien catalogados dentro de las edificaciones, carreteras y vegetación. Pero, también para establecer los errores que presentan: porcentajes de puntos que según la BTA deberían contemplarse como de una categoría determinada y no lo son (error de omisión); y, porcentajes que diciendo el LiDAR que son de una determinada clase la BTA no lo contempla así (error de comisión). Este estudio se puede analizar en el punto quinto de esta memoria.

A la hora de proceder con algoritmos de aprendizaje supervisado en el ámbito de la Minería de Datos hay que marcar unos valores a verificar que en este trabajo se han determinado a partir de la cartografía BTA disponible. Esto implica establecer una relación entre la información aportada por el LiDAR 2008, en este caso en cuanto a clasificación se refiere, y la dispuesta en la BTA para a partir de esta relación poder realizar el entrenamiento de los datos. A continuación se detallan estos dos aspectos.

4.2.3.1. Consideraciones LAS - BTA

En el punto 3.1 del capítulo 3 se han explicado las especificaciones del estándar de datos LiDAR y se ha comprobado que la última versión del formato LAS ha ampliado los elementos a discernir en el ámbito de la clasificación de los puntos de la nube de datos captados. Por eso, en respuesta a esta nueva realidad se ha considerado oportuno sopesar tanto los elementos contemplados en el formato 1.2 como los del 1.4, con el fin de establecer la relación existente con la base cartográfica de referencia (BTA).

Por otro lado, como en la literatura existen muchas referencias en las que a la hora de clasificar los puntos LiDAR lo primero que consideran es la distinción entre puntos TERRENO / NO TERRENO, esta diferenciación también se ha tenido en cuenta al revisar los distintos elementos a clasificar. Así, tras analizar todos y cada uno de los fenómenos con geometría superficial en el documento de las especificaciones de la BTA (Consejo Superior Geográfico 2008) y teniendo en cuenta lo especificado anteriormente sobre la BTA, en las tablas 4.7. y 4.8. se ha recogido la relación a establecer entre las categorías de la BTA y la clasificación del fichero LiDAR en formato LAS, contemplando tanto el formato 1.2 como el 1.4.

Como se puede observar en las tablas anteriores muchos elementos cartográficamente distintos pertenecen a la misma clase en el formato LAS. Intentado buscar la clasificación más amplia posible, en el estudio efectuado se ha visto que puede resultar interesante el contemplar las explanadas, los helipuertos, las pistas deportivas y las pistas de aeródromos consideradas en la BTA dentro del fenómeno denominado edificaciones, poblaciones y construcciones. En lo que respecta al formato LAS estos elementos geográficos deberían aparecer con puntos de suelo (2), por lo que puede ser un aporte a la investigación si se consigue diferenciarlos del suelo como tal. Con esta pretensión se ha considerado una nueva clase no contemplada en la versión 1.4 del formato LAS para discriminar estos elementos. Esta clase se ha identificado con la denominación de plataformas y se le ha asignado el valor 64 para su registro.

Tabla 4.7. Categorías BTA "NO TERRENO" relacionadas con la clasificación de los datos LiDAR en formato 1.2 y 1.4

	LAS 1.2	Fenómeno BTA		ID_TIPO	*LAS 1.4*
No Terreno	*Edificaciones* (6)	Edificaciones	Edificación	0056	6
			Edificación ligera	0057	
			Muralla	0065	
			Chimenea	0049	
			Depósito	0051	
			Puente	0080	17
			Pasarela	0067	
			Torre tendido	0086	15
	Sin clasificar (1)	Serv. Insta.	Tendido eléctrico	0116	14
	Vegetación baja (3)	Cubierta terrestre	C. herbáceos	0124	3
			C. leñosos	0125	
			Cultivos	0123	
			Huerta	0126	
			Prado	0128	
			Pastizal	0129	
	Vegetación media (4)		Matorral	0130	4
	Vegetación alta (5)		Arbolado forestal	0122	5

Tabla 4.8. Categorías BTA "TERRENO" relacionadas con la clasificación de los datos LiDAR en formato 1.2 y 1.4

	LAS 1.2	Fenómeno BTA		ID_TIPO	*LAS 1.4*
Terreno	*Suelo* (2)	Edificaciones	Explanada	0030	64 (nuevo)
			Helipuerto	0062	
			Pista aeródromo	0068	
			Pista deportiva	0069	
		Transporte (vías no urbanas)	Camino	0026	11
			Carretera doble	0028	
			Carretera única	0029	
			Ferrocarril	0036	10
			Funicular	0037	
			Tranvía	0038	
			Metro	0040	
	Agua (9)	Hidrografía Aguas	Embalse	0017	9
			Laguna	0016	
			Estanque	0023	
			Piscina	0024	
		Hidrografía Corrientes	Artificial	0011	
			Natural	0012	

En base a todas estas consideraciones se ha generado la correspondencia mostrada en la tabla 4.9 que reúne la relación existente entre los elementos geográficas contemplados en los distintos fenómenos de la BTA y la clasificación que se pretenden determinar en los puntos LiDAR tras la aplicación de la metodología del título sexto de este documento.

Tabla 4.9. Relación clases .LAS 1.4 y BTA

PUNTOS LiDAR		BTA: códigos ID_TIPO
Denominación	Clase	
Suelo	2	Resto de ID_TIPOs
Vegetación baja	3	0123, ~~0124, 0125, 0126~~, 0128, 0129
Vegetación media	4	0130
Vegetación alta	5	0122
Edificaciones	6	0056, 0057, 0065, ~~0049, 0051~~, 110, 153, 152
Agua	9	0017, 0016, 0023, 0024, 0011, 0012
Ferrocarril	10	0036, 0037, 0038, 0040
Carretera	11	0026, 0028, 0029
Torre eléctrica	15	~~0086~~
Línea eléctrica	16	0116
Puentes	17	0067, 0080
Plataformas	64	~~0030~~, 0062, 0068, 0069

(los valores tachados no están disponibles en la BTA del GV pero se mantienen por si se usa en otras zonas)

4.2.3.2. Entrenamiento de los datos

Tal y como indica García-Gutiérrez 2012 para la verificación del resultado es necesario el entrenamiento de la muestra, para ello se ha asignado un valor a cada punto LiDAR en función del ID_TIPO de las coberturas superficiales consideradas de la BTA tras realizar una intersección espacial entre ambos tipos de datos.

Extraída la información toca su asignación, lo que se ha realizado partiendo de la consideración inicial de que todos los puntos pertenecen al terreno, asignándoles el valor de suelo. Este hecho ha llevado a considerar puntos que en el LAS aparecen como nunca clasificados (0), sin clasificar (1) o puntos de ruido (7 = *Low point*) (tabla 4.5) dentro de la clase suelo en el entrenamiento.

Posterior a la asignación a todos los puntos como suelo, según los valores que han adquirido en la intersección con las coberturas de la BTA se les ha ido otorgando los valores correspondientes atendiendo a la relación incluida en la tabla 4.9.

Señalar que las categorías 13, 14 y, 15 que aparecen en el formato 1.4 de la ASPRS (tabla 3.4) no han podido ser determinadas a partir de la BTA ya que ésta no tiene tanta apreciación (aunque existe un código para las torretas, luego en la cartografía no se han contemplado), por lo que se ha considerado todo lo referente al tendido eléctrico en la categoría 16.

	x	y	z	CT	EE	EC	FU	HA	HC	NU	SI	RV	VF	TE
276	498215.875	4760815	347.540009	128	0	0	0	0	0	0	0	0	0	0
277	498279.344	4760822.5	336.790009	128	0	0	0	0	0	0	0	0	0	0
278	498103.813	4760822.5	356.980011	128	0	0	0	0	0	0	0	0	0	0
279	498146.625	4760822.5	359.279999	999	56	0	0	0	0	0	0	0	0	0
280	498213.125	4760822.5	347.820007	128	0	0	0	0	0	0	0	0	0	0
281	498026.781	4760822.5	368.700012	128	0	0	0	0	0	0	0	0	0	0
282	498121.938	4760822.5	357.480011	128	0	0	0	0	0	0	0	0	0	0
283	498214.938	4760822.5	347.600006	128	0	0	0	0	0	0	0	0	0	0
284	498269.094	4760822.5	338.450012	128	0	0	0	0	0	0	0	0	0	0
285	498145.406	4760823	359.399994	999	56	0	0	0	0	0	0	0	0	0
286	498010.781	4760823	370.709991	128	0	0	0	0	0	0	0	0	0	0
287	498270.031	4760823	338.230011	128	0	0	0	0	0	0	0	0	0	0
288	498173.531	4760823	351.170013	123	0	0	0	0	0	0	0	0	0	0
289	498238.844	4760823	343.440002	128	0	0	0	0	0	0	0	0	0	0
290	498225.25	4760823	349.76001	128	0	0	0	0	0	0	0	0	0	0
291	498034.625	4760823	365.820007	128	0	0	0	0	0	0	0	26	0	0
292	498039.688	4760823	365.299988	128	0	0	0	0	0	0	0	26	0	0

Figura 4.3. Valores adquiridos para cada cobertura de la BTA

Tampoco se han tenido en cuenta el resto de clases que la ASPRS considera reservadas (8, 12, 19-63), pero tal y como se ha comentado en el apartado anterior, se ha creado el código 64 para las extensiones construidas a ras de suelo, tales como explanadas y pistas.

Por último, teniendo en cuenta que un mismo punto puede tener para cada intersección con las coberturas de la BTA un valor distinto, se ha establecido una prioridad entre ellos, considerando por un lado que las edificaciones o la red viaria o el ferrocarril prevalecen sobre

la vegetación; y, por otro que las edificaciones predominan sobre los elementos construidos. Así, en el caso del ejemplo de la figura 4.3 el punto 279 ha sido considerado "edificio" (EE = 56) y el 291 "red viaria" (RV = 26), a pesar de que ambos adquieren otro valor en el caso de la cubierta terrestre (CT).

4.3. ÁREAS DE ESTUDIO

En este apartado se ha procedido a enmarcar las zonas que se han utilizado a lo largo de la investigación para llevar a cabo las metodologías planteadas en los apartados quinto y sexto. En el primero se va procede a realizar un estudio sobre el estado de la clasificación presentada por los datos LiDAR 2008; y, en el segundo se desarrolla una metodología para mejorar esa clasificación. Las áreas de actuación no son exactamente las mismas en las metodologías planteadas, por lo que en los siguientes apartados se especifican las zonas que se han utilizado en cada ocasión, justificando debidamente su elección.

Figura 4.4. Distribución de las hojas del LiDAR 2008

En la figura 4.4, se puede apreciar el conjunto de todas las hojas del LiDAR 2008, incluido el Condado de Treviño, ascendiendo a un total de 8.211. Para los posteriores trabajos a abordar en este estudio se ha prescindido de las hojas del Condado de Treviño y de las que aportaban información en una pequeña parte de la misma, reduciéndolas a 6.730 hojas.

4.3.1. ZONAS PARA EL ESTUDIO DE LA CLASIFICACIÓN DE LOS DATOS LIDAR 2008

Con la finalidad de valorar la información aportada por los archivos del LiDAR 2008 en cuanto a la clasificación se refiere, del total de las 6.730 hojas consideradas se han examinado las que se muestran en la tabla 4.4. Éstas se encuentran relacionadas con la división de hojas cartográfica a escala 1:10.000 porque la primera versión de la BTA se confeccionó en base a ella.

Tabla 4.4. Hojas LAS para el análisis de la clasificación aportada por el LiDAR 2008

	Hoja 61 (Bizkaia)	Hoja 64 (Gipuzkoa)	Hoja 63 (Bizkaia)
Hoja 1/10.000	61-B	64-C	63-C
Hojas LAS	5064790_c.las 5074791_c.las 5094793_c.las	5684788_c_Gipuzkoa.las 5694788_c_Gipuzkoa.las	5404782_c_Gipuzkoa.las 5404782_c.las

A tendiendo a las cuadrículas LAS indicadas en esa tabla y según lo indicado en la tabla 4.1 se han considerado tres cuadrículas correspondientes al vuelo de la zona de Gipuzkoa (a posteriori vuelo DFG), otras tres del vuelo de la zona del Gran Bilbao y una del resto de Bizkaia y zona norte de Álava, quedando sin considerar cuadrículas del vuelo que comprende la zona del resto de Álava. Es de suponer que todas las cuadrículas encargadas por el GV (vuelo GV), independientemente de la fase de ejecución, hayan sufrido el mismo tratamiento, por lo que a priori las cuadrículas de esta última zona deberían tener el mismo comportamiento que las del Gran Bilbao o resto de Bizkaia y zona norte de Álava. En la tabla 4.5 se puede observar la distribución de estas cuadrículas según las distintas zonas consideradas en las que se divide el vuelo LiDAR 2008 y la pertenencia al procesamiento de una u otra institución.

Tabla 4.5. Cuadrículas LAS consideradas según fracción de zonas del LiDAR 2008

Fracción de zonas del LiDAR 2008	Cuadrículas LAS	Vuelo
Gipuzkoa	5404782_c_Gipuzkoa.las 5684788_c_Gipuzkoa.las 5694788_c_Gipuzkoa.las	Vuelo DFG
Gran Bilbao	5064790_c.las 5074791_c.las 5094793_c.las	Vuelo GV
Resto de Bizkaia y zona norte de Álava	5404782_c.las	

En cuanto a las características que persiguen las cuadrículas LAS elegidas señalar que en la hoja 61 se han considerado tres dentro de la comarca del Gran Bilbao: la 5064790 perteneciente al núcleo urbano de Bilbao, sin vías de comunicación importantes y con baja vegetación; la 5074791 que abarca casco urbano, vías de comunicación y zonas de bosques y prados y la 5094793 que se trata del polígono industrial de Ugaldeguren, entre el municipio de Derio y el de Zamudio, con grandes pabellones industriales, vías de comunicación y zonas con vegetación. Todas estas cuadrículas pertenecen al vuelo realizado por el Gobierno Vasco en su primera fase.

En lo que respecta a la hoja 63, la cuadrícula 5404782 hace referencia al municipio vizcaíno de Ermua, abarcando la inmensa mayoría del casco urbano, atravesado con una vía de comunicación importante y rodeado de zonas de vegetación a ambos lados del mismo, principalmente boscosas. Esta cuadrícula tiene la particularidad de encontrarse volada tanto por el vuelo DFG como por el vuelo GV.

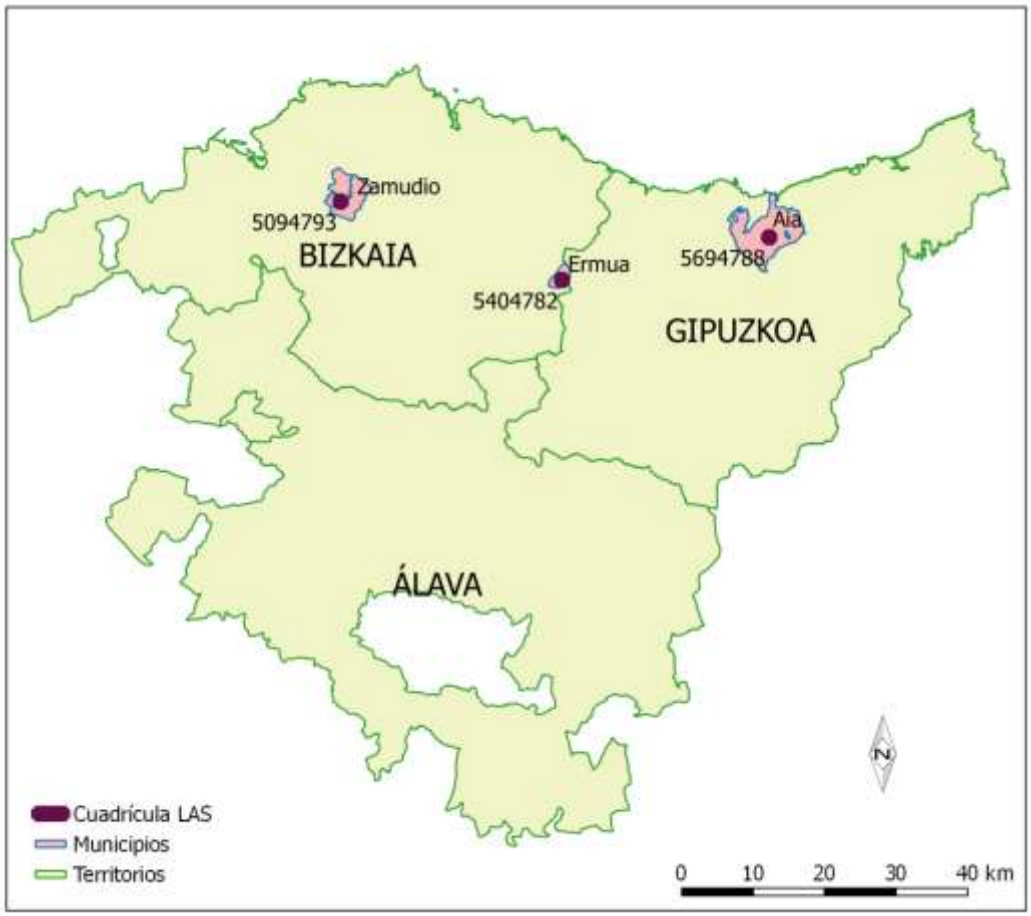

Figura 4.5. Ubicación de las zonas de estudio para verificar la clasificación ASPRS del LiDAR 2008

La tercera zona considerada, ubicada en la hoja 64, abarca la mayor parte del casco urbano de Aia (Gipuzkoa) situado en la ladera del monte Pagoeta y con una importante carretera que lo atraviesa. Esta zona urbana tiene la particularidad de pertenecer a una zona rural por lo que sus características son muy diferentes al de las otras zonas consideradas. Las cuadrículas de datos LiDAR que hacen referencia al mismo son la 5684788 y la 5694788, ambas del vuelo DFG.

Estas cuadrículas (figura 4.5) se corresponden con zonas en las que se dan distintas tipologías de edificaciones, comprenden carreteras que las cruzan en cuanto a las vías de comunicación y disponen de una vegetación variada, de manera que van a permitir verificar el estado de la clasificación de los datos LiDAR 2008 en cuanto a esa información geográfica se refiere. Además, como pertenecen a los dos grupos de procesamientos establecidos, se posibilita la comprobación de ambos procesos, estableciendo sus similitudes y diferencias.

En el capítulo quinto de este documento se ha presentado el análisis correspondiente en base a las cuadrículas 5094793 del vuelo GV, 5694788 del vuelo DFG y 5404782 que forma parte de los dos vuelos, intentado de esta forma recabar información sobre el estado de los datos en zonas con distintas características, principalmente urbanas pero también rurales, tanto del vuelo GV como del vuelo DFG. Estas zonas quedan localizadas en la figura 4.5 en el entorno de la CAPV y en la figura 4.6 se pueden apreciar con más detalle, apreciando las diferentes tipologías elegidas.

En la tabla 4.6 se han recogido el resto de datos cartográficos necesarios para el estudio indicado, que en el caso de la cartografía vectorial son ficheros tipo ESRI shapefile (shp); y, en el de las ortofografías se trata de ficheros .jpg con su georreferenciación (.jgw).

Tabla 4.6. Hojas BTA y ortofotografías básica, datos de referencia

Hoja 1:50.000	BTA 1:10.000	Ortofotografía 2008 RGB 1:5.000
61 (Bizkaia)	BTA_061_4_2_Diciembre_2012.shp	H0061-7-4.jpg H0061-8-3.jpg
64 (Gipuzkoa)	BTA_064_1_3_Diciembre_2012.shp	H0064-1-5.jpg H0064-2-5.jpg
63 (Bizkaia)	BTA_063_1_4_Diciembre_2012.shp	H0063-1-7.jpg H0063-1-8.jpg

Figura 4.6. Cuadrículas kilométricas 5094793, 5404782 y 5694788 del LiDAR 2008 con la ortofotofrafía de fondo

4.3.2. ZONAS PARA LA APLICACIÓN DE LA METODOLOGÍA DE CLASIFICACIÓN

En este segundo supuesto, y tal y como se explica en el capítulo sexto, hay que proceder a seleccionar los conjuntos de datos más adecuados tanto para el entrenamiento como para la verificación de los resultados obtenidos; y, además, la información aportada para el entrenamiento debe ser abundante y variada. En base a ello, en la figura 4.7 se muestra la distribución de las cuadrículas LAS consideradas en cada caso. A estas zonas se les va a denominar en adelante zonas de aplicación por tratarse propiamente de los datos usados en la investigación para la metodología propuesta.

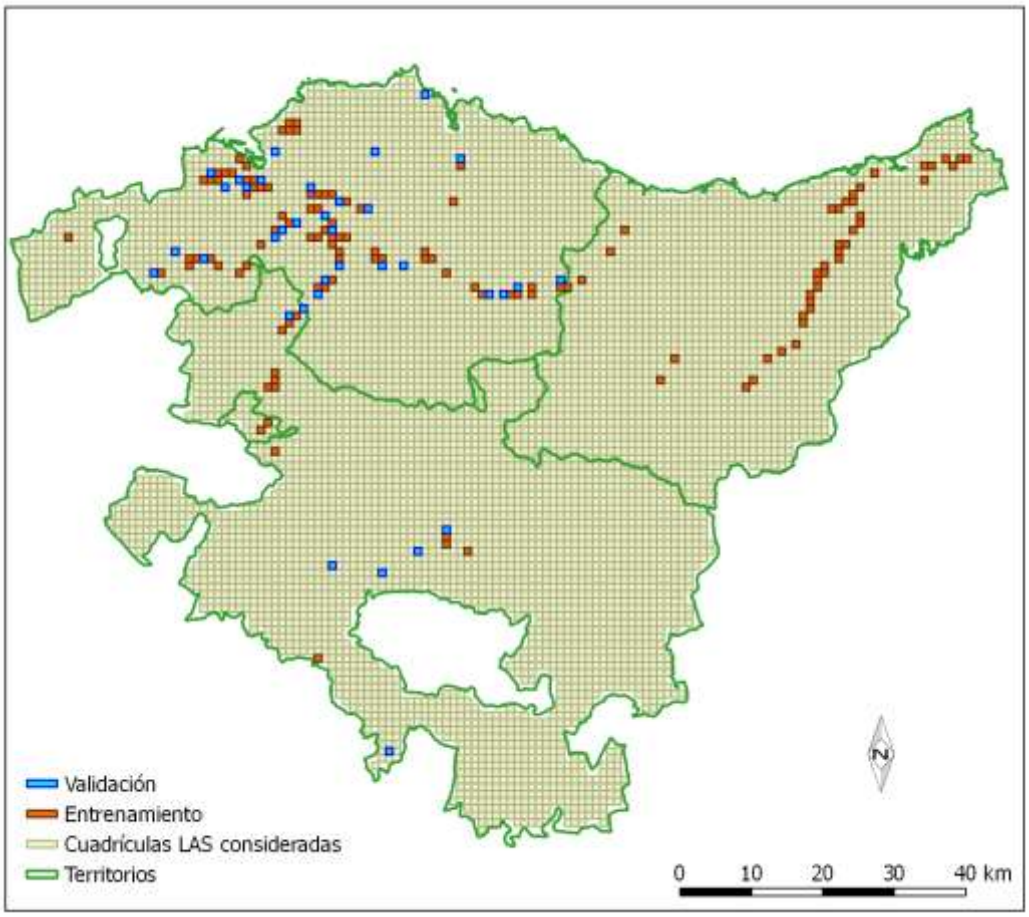

Figura 4.7. Cuadriculas LAS consideradas para el entrenamiento y la validación

A pesar de que en un principio se había pensado en proceder con el procesamiento de las 6.730 cuadrículas LAS consideradas de la CAPV dado que se ha planteado el contemplar el mayor número de situaciones posibles en relación a las características de la superficie analizada, tras comprobar los tiempos de procesamiento y las limitaciones de los ordenadores disponibles en algunas fases del análisis, cuyas características se pueden consultar en la tabla 3.11, comprobando que se trata de una estación de trabajo de Dell con unos requisitos superiores a muchos de los ordenador de sobremesa habituales, ha resultado necesario realizar una selección de las hojas a tratar.

En consecuencia, para seleccionar las hojas LAS objeto de investigación, ha sido preciso concretar las principales categorías de puntos a estudiar compatibles con los datos disponibles, siendo las pautas generales para la selección de las hojas las siguientes:

- En la cuadrícula deben aparecer elementos geográficos correspondientes a edificaciones, carreteras, vías férreas y tendidos eléctricos; si bien esta designación es más importante para el entrenamiento que para la validación.
- Además, el número de edificaciones por cuadrícula debe ser igual o superior a 100, eliminando así aquellas hojas sin edificaciones o con valores bajos. Esto va a permitir tener bien entrenadas las edificaciones.
- Y, por último, no se deben considerar hojas con valores negativos en la coordenada *Z*, ya que éstos, por lo general, responden a ruido y hay que proceder a eliminarlos.

En la figura 4.8 se ha reflejado una pequeña muestra de las hojas que cumplen dichas características, resultando un total de 146 cuadrículas de las que 119 se han utilizado para el entrenamiento y 25 para la validación. Además, éstas se han visto complementadas por otras según las necesidades que se han ido viendo en la investigación (apartado 6), obteniendo finalmente 126 hojas para el entrenamiento y 36 para la validación, constituyendo éstas un 22 % del total de las hojas procesadas (162).

FICHERO	DENSIDAD	N_M_PUNTC	Z_MAXIMA	Z_MINIMA	EDIFICIOS	ELE_CONST	Serv_Inst	HI_AGUAS	HI_CORRI	NU_URB	RED_VIARIA	VI_FERREAS	T_Electric
4844783.las	1.818	1818049	1469.94	134.61	274	21	6	1	1	1	1	1	1
4884784.las	1.724	1723954	1448.62	91.34	214	8	3	1	1	1	1	1	1
4904785.las	1.674	1674278	1441.66	75.59	239	11	5	1	1	1	1	1	1
4914797.las	3.570	3569582	1054	4.35	230	35	3	1	1	1	1	1	1
4964784.las	1.890	1890109	1391.18	44.51	229	11	4	1	1	1	1	1	1
4984795.las	2.634	2634294	921.31	1.38	257	35	18	1	1	1	1	1	1
5004767.las	2.125	2125370	279.68	172.73	236	12	7	1	1	1	1	1	1
5024790.las	3.218	3217883	1030.74	1.35	242	10	6	1	1	1	1	1	1
5044778.las	1.830	1830065	1404.49	103.35	208	11	11	1	1	1	1	1	1
5054794.las	3.230	3229884	940.17	1.37	352	10	7	1	1	1	1	1	1
5064729.las	0.272	2720258	1458.45	462.9	296	8	8	1	1	1	1	1	1
5064780.las	1.637	1636823	1462.58	79.37	209	17	2	1	1	1	1	1	1
5064794.las	3.021	3021486	929.21	6.78	347	14	13	1	1	1	1	1	1
5074781.las	1.555	1555271	1406.42	63.27	218	17	9	1	1	1	1	1	1
5074789.las	2.794	2794208	992.2	12.37	236	15	18	1	1	1	1	1	1
5094786.las	3.121	3120810	316.43	38.82	238	2	12	1	1	1	1	1	1
5134792.las	3.433	3432604	924.6	48.8	302	17	4	1	1	1	1	1	1
5154741.las	2.640	2640260	1555.46	476.21	339	15	7	1	1	1	1	1	1
5154785.las	3.131	3131339	988.82	43.12	209	10	5	1	1	1	1	1	1
5214786.las	2.000	2000246	1214.01	61.19	286	24	9	1	1	1	1	1	1

Figura 4.8. Muestra con las hojas LAS que cumplen con los criterios de selección

En las siguientes tablas se presentan las relaciones de las hojas utilizadas en cada caso: en la tabla 4.7 se han indicado las hojas consideradas como validación y en la 4.8 las de entrenamiento, quedando gráficamente representadas en la figura 4.7, comentada previamente.

Tabla 4.7. Cuadrículas LAS utilizadas para la validación de los resultados del aprendizaje automático

4834783_c.las	5004800_c.las	5084742_c.las	5184784_c.las
4864786_c.las	5014789_c.las	5084789_c.las	5204744_c.las
4904785_c.las	5024777_c.las	5094784_c.las	5214808_c.las
4914797_c.las	5034790_c.las	5094793_c.las	5244747_c.las
4934795_c.las	5044778_c.las	5134792_c.las	5264799_c.las
4954796_c.las	5054795_c.las	5144800_c.las	5304780_c.las
4964795_c.las	5064780_c.las	5154741_c.las	5324780_c.las
4984796_c.las	5074782_c.las	5154784_c.las	5344781_c.las
5004788_c.las	5074791_c.las	5164716_c.las	5404782_c.las

Tabla 4.8. Cuadrículas LAS utilizadas para el entrenamiento de los algoritmos de aprendizaje automático

4714788_c.las	5014775_c.las	5144785_c.las	5754778_c.las
4844783_c.las	5014791_c.las	5144786_c.las	5754779_c.las
4884784_c.las	5014803_c.las	5154785_c.las	5754780_c.las
4884785_c.las	5024776_c.las	5214785_c.las	5764781_c.las
4894785_c.las	5024790_c.las	5214786_c.las	5764782_c.las
4904796_c.las	5024803_c.las	5224785_c.las	5764783_c.las
4914785_c.las	5024804_c.las	5244745_c.las	5774783_c.las
4914796_c.las	5034777_c.las	5244746_c.las	5774784_c.las
4924784_c.las	5034803_c.las	5244783_c.las	5784792_c.las
4924796_c.las	5034804_c.las	5254793_c.las	5794785_c.las
4924797_c.las	5054788_c.las	5264798_c.las	5794786_c.las
4934797_c.las	5054792_c.las	5274744_c.las	5794787_c.las
4944797_c.las	5054794_c.las	5284781_c.las	5794792_c.las
4954783_c.las	5064729_c.las	5294780_c.las	5804787_c.las
4954799_c.las	5064781_c.las	5334780_c.las	5804793_c.las
4964784_c.las	5064788_c.las	5344780_c.las	5814789_c.las
4964794_c.las	5064792_c.las	5364780_c.las	5814793_c.las
4964796_c.las	5064794_c.las	5364781_c.las	5814794_c.las
4964798_c.las	5074781_c.las	5404781_c.las	5824790_c.las
4974795_c.las	5074789_c.las	5414781_c.las	5824791_c.las
4974796_c.las	5074794_c.las	5434782_c.las	5824795_c.las
4984761_c.las	5084782_c.las	5474786_c.las	5844797_c.las
4984787_c.las	5084787_c.las	5494789_c.las	5914796_c.las
4984795_c.las	5084788_c.las	5544768_c.las	5914798_c.las
4994762_c.las	5084790_c.las	5564771_c.las	5924798_c.las
4994767_c.las	5084794_c.las	5664767_c.las	5944799_c.las
4994795_c.las	5094785_c.las	5674768_c.las	5954798_c.las
5004758_c.las	5094786_c.las	5694771_c.las	5964799_c.las
5004767_c.las	5094788_c.las	5714772_c.las	5974799_c.las
5004768_c.las	5104788_c.las	5734773_c.las	
5004769_c.las	5104793_c.las	5744776_c.las	
5004789_c.las	5124792_c.las	5744777_c.las	

Por último, señalar que aunque no se han podido procesar todas las hojas de la CAPV por las razones apuntadas anteriormente, el enfoque que se le ha querido dar a este estudio ha sido de carácter masivo ya que en lugar de considerar una única cuadrícula en la que se dieran las tipologías a catalogar se ha tenido en cuenta una gran magnitud de hojas con características distintas, lo que supone también una novedad en la aplicación de los algoritmos de aprendizaje automático aplicados a la clasificación de datos LiDAR, dado que en la literatura de este tipo de estudios, de manera habitual, estos algoritmos aparecen aplicados a una única hoja, tal y como se puede apreciar en Chehata, et al. 2009.

5. EVALUACIÓN DE LA CLASIFICACIÓN ASPRS DE LOS DATOS LIDAR 2008

En este capítulo se presenta la metodología que se ha seguido para evaluar los datos LiDAR en cuanto a la clasificación se refiere, utilizando las zonas del LiDAR 2008 de la CAPV indicadas en el punto anterior.

Además de la metodología, se presentan los resultados cuantitativos obtenidos a tendiendo a la información geográfica seleccionada y se muestra una comparativa en la información aportada por datos del mismo vuelo pero con distintos procesamientos.

5.1. INTRODUCCIÓN

El objetivo de este capítulo es la verificación de los valores aportados en el campo de la clasificación de los datos LAS del LiDAR 2008. Para ello, se ha considerado como base de contrastación la información geográfica facilitada por la Base Topográfica Armonizada (BTA), dado que en esta investigación se van a usar los fenómenos de la BTA como base para el entrenamiento de los algoritmos. Las características de la BTA han sido explicadas en el punto 4.2.3 y para su ejecución se ha utilizado el software y hardware comentado en el apartado 3.3.1.

En lo que respecta a la BTA y partiendo de la geometría de polígonos, en este apartado se han tenido en cuenta las carreteras con firme ubicadas dentro del fenómeno de Transportes, las edificaciones pertenecientes al fenómeno de Edificaciones y los cultivos, prados, matorrales y arbolado forestal incluidos en el de Cubierta Terrestre.

En el epígrafe cuarto también se han descrito las zonas y los datos LAS a utilizar. En cuanto a estos últimos, recordar que se encuentran en la versión 1.2 del formato LAS, ya que el vuelo se realizó en 2008, responden al tipo de registro de nube de puntos PDRF 3 (tabla 3.2) y al nivel 3 en cuanto a procesamiento, lo que significa que los datos están georreferenciados.

Tras examinar la tabla 4.5, en la que se muestran los valores de clasificación que aportan los datos LAS a analizar, como elementos a contrastar por la BTA se han elegido los que hacen referencia a vegetación, edificación y parte de los de suelo, dado que el resto de categorías no ofrecen correspondencia directa con la información geográfica. En la tabla 5.1 se expone dicha relación, aunque cabe señalar que los prados y los cultivos pueden aparecer en el archivo LAS como puntos de suelo por poder formar parte de éste.

Tabla 5.1. Relación entre elementos geográficos (BTA) y valores de clasificación (LAS) del LiDAR 2008

ASPRS		BTA	
Clase	Tipo de elemento	Descripción	ID_TIPO
2	Suelo	Carreteras	0028, 0029
3	Vegetación baja	Prado	0128
		Cultivos	0123
		Vegetación y arbolado urbano	0140
4	Vegetación media	Matorral	0130
5	Vegetación alta	Arbolado forestal	0122
6	Edificación	Edificación	0056, 0057

Esta es la situación que se plantea con las carreteras, las cuales, a priori, deberían quedar definidas por puntos con la clase suelo, ya que en el formato LAS 1.2 no se considera la clase específica de carreteras. En consecuencia, hay que dejar claro que dentro de la clase suelo, en general, aparecen todos los puntos que se identifica con éste, pudiendo aparecer parcelas, carreteras con firme, caminos u otros elementos geográficos naturales o artificiales que por carecer de altura constituyen parte de la superficie terrestre (suelo).

En las siguientes secciones se ha procedido a explicar la metodología a seguir para alcanzar el fin planteado al inicio de este punto, además de señalar de qué forma se van a valorar los resultados alcanzados, analizando a posteriori los valores conseguidos y presentando al final las conclusiones de este apartado.

5.2. METODOLOGÍA

La metodología que se ha planteado se divide en tres fases: la primera referente al análisis y tratamiento de los datos LiDAR, la segunda basada en la obtención de la información y la tercera fundamentada en la propia validación de los resultados.

En lo que respecta al análisis y tratamiento de los datos LiDAR tres han sido los pasos considerados: eliminación de puntos que constituyen ruido, analizar la información facilitada por la cabecera de los ficheros LAS y la recuperación de las pasadas originales en el desarrollo del vuelo. Esto último se ha realizado porque al encontrarse los archivos LAS disponibles divididos en cuadrículas de 1 km × 1 km se quiere comprobar sí la información a nivel de pasada resulta relevante o no.

En la segunda fase se ha realizado una intersección espacial, utilizando los elementos de la BTA indicados en la tabla 5.1. De esta manera, se han determinado para cada caso los puntos del archivo LAS que quedan dentro de cada uno de los elementos geográficos considerados. En la figura 5.1 quedan recogidas estás dos primeras etapas.

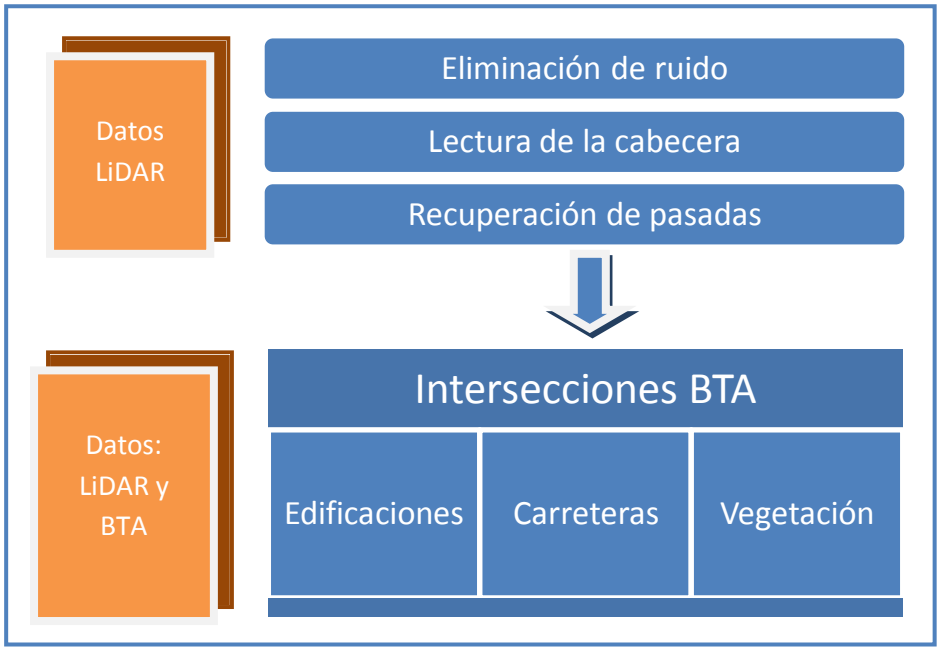

Figura 5.1. Esquema metodológico para el análisis de la clasificación del LiDAR 2008

Por lo tanto, una vez descargados los datos necesarios, el flujo de trabajo viene marcado por la consecución de los procesos indicados en la tabla 5.1, requiriendo los procesos 4, 5 y 6 de los tres anteriores.

Tabla 5.2. Procesos seguidos

Proceso	Misión
1	Limpieza de ruido
2	Extraer información de los datos LAS
3	Recuperación de pasadas
4	Intersección BTA con edificios
5	Intersección BTA con carreteras
6	Intersección BTA con vegetación

Los proceso 4, 5 y 6 han permitido analizar los puntos LiDAR que intersecan con la capa de la BTA considerada en cada caso, para a continuación proceder a cuantificar el número de puntos por cada una de las clases contempladas y establecer su validación, tercera fase de esta metodología, utilizando para ello los estimadores que se van a explicar en el apartado 5.2.2.

5.2.1. EXPLICACIÓN DE LOS PROCESOS

En este apartado se presenta una pequeña explicación de los procesos indicados en la tabla 5.1: limpieza de ruido, extracción de la información de los datos LAS, recuperación de pasadas y asignación BTA a los puntos de los ficheros LAS que recoge las intersecciones especificadas.

5.2.1.1. Limpieza de ruido

En las pruebas que se han realizado previamente se ha detectado que en diferentes ficheros LAS aparecen puntos con un rango de *Z* fuera de lo que supone la cobertura terrestre. Motivo que ha contribuido a considerar oportuno la creación de un filtro que permita la eliminación de puntos anómalos, que según (Graham 2012) pueden ser debidos al multipath de la señal GNSS, valores de reflectancia baja para la intensidad o a fallos desconocidos del sistema.

Este proceso se ha realizado en FME Workbench, y se basa principalmente en la eliminación de los puntos con una *Z* elevada con respecto a los de su entorno. Para la búsqueda de estos puntos se han considerado franjas muy estrechas de *Z* a lo largo del eje Y, y analizado los puntos próximos se han eliminando los puntos anómalos.

5.2.1.2. Información sobre los datos LAS

Eliminados los valores anómalos en *Z*, con el objeto de obtener información sobre los datos a procesar, se hace uso del programa lasinfo.exe de LASTools el cual ofrece información sobre los datos contenidos en el fichero LAS (Isenburg 2014).

En la tabla 3.1 se ha incluido una pequeña explicación de dicha información obtenida a través de un ejemplo. Un análisis previo de estos valores puede permitir detectar situaciones extrañas, así como ayudar en la interpretación de los resultados alcanzados.

5.2.1.3. Recuperación de pasadas

Tras analizar en el fichero de información generado en el paso anterior que no existen otros datos anómalos y comprobar que se han eliminado los puntos que producen ruido, se ha procedido a recuperar las pasadas originales de cada uno de los archivos tratados.

En este caso, también se ha hecho uso de uno de los ejecutables que proporciona LAStools, concretamente de lassplit.exe. Éste devuelve un fichero por cada uno de los identificadores encontrados en el campo *Point_Source_ID*, que tal y como se ha comentado en el apartado 3.1, se corresponde con el identificador de la pasada realizada.

Como resultado de este proceso, por cada una de las cuadrículas LAS procesadas se han obtenido tantos ficheros como pasadas existen en el kilómetro cuadrado considerado. Puede darse el caso de que algunos de estos archivos contengan muy poca información y deban ser desestimados a la hora de proceder con los siguientes tratamientos.

5.2.1.4. Asignación BTA a los puntos LAS

Obtenidas las pasadas se ha procedido a realizar la intersección entre la cartografía base y la nube de puntos, para en cada caso comprobar si los puntos que caen dentro de los elementos cartográficos considerados pertenecen en el fichero LAS a la clase que les corresponde según la cartografía.

Para llevarlo a cabo se ha desarrollado en FME Workbench un flujo de trabajo que considera un filtrado espacial entre los datos LAS y los edificios, las carreteras o los elementos de vegetación, en cada caso; determinando los puntos que pertenecen a cada tipo de entidad. Esta intersección se ha realizado tanto con las cuadrículas completas como con los archivos derivados por cada una de las pasadas consideradas en cada cuadrícula.

En el caso de las edificaciones, como la cartografía de referencia contiene un nivel de detalle planimétrico de 1 m y se dispone de datos LiDAR de mayor exactitud se ha desarrollo un procedimiento más contemplando un retranqueo de 2 m con el objeto de poder verificar tanto

la generalización asociada a nivel de detalle de la base, como otros efectos tales como puede ser la influencia de los aleros, que pueden distorsionar los resultados estadísticos.

Por su parte, en lo que se refiere a la vegetación no se ha realizado ninguna distinción para comprobar las distintas clases - baja, media y alta - planteadas por la ASPRS.

5.2.2. VALORACIÓN ESTADÍSTICA

Existen distintas metodologías para la validación de los estudios de este tipo. En este caso, se presenta una valoración cuantitativa que proporciona valores estadísticos permitiendo comparar los datos de referencia con los extraídos del estudio realizado.

La cuantificación suele derivar de la matriz de confusión (MC), también denominada tabla de contingencia o matriz error (Congalton 1991). Esta matriz recoge los conflictos que se presentan entre categorías, organizándose de manera que las columnas constituyen la base de contrastación (referencia) y las filas los valores derivados de la clasificación (Chuvieco 2008).

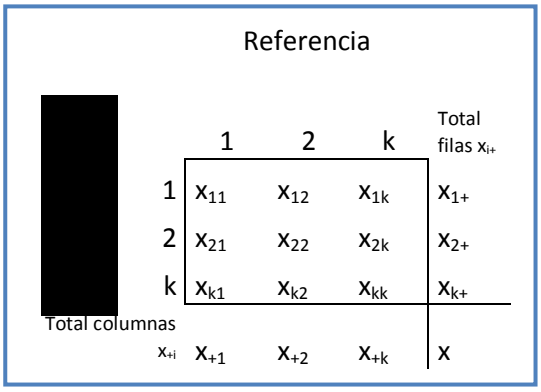

Figura 5.2. Esquema de la matriz de confusión

La diagonal de esa matriz muestra el perfecto acuerdo entre la referencia y la clasificación, también denominados *True Positive* (TP) o *True Negative* (TN) mientras que los valores laterales (marginales) de la diagonal ofrecen los errores de asignación o *False Positive* (FP') que es el error de tipo I y *False Negative* (FN) o error de tipo II. Considerando estos datos se pueden calcular el Error de Omisión (*EO*) y el Error de Comisión (*EC*) así como sus respectivos Fiabilidad del productor (*FP*) y Fiabilidad el usuario (*FU*).

El *EO* contempla los puntos que perteneciendo a una clase no se han considerado dentro de ella y está relacionado con la *FP* que indica la probabilidad de que los puntos pertenezcan a la clase que se le supone.

Donde x_{ii} es el valor en la fila i y columna i (diagonal) y x_{+i} el valor del total de la columna i.

A la fiabilidad del productor (*FP*) también se le denomina *precision* y cuantifica los datos positivos bien clasificados, por lo que un valor de 1 significa que los píxeles clasificado como positivos lo son en el documento de referencia (Pérez and Bromberg 2012).

donde *TP+FP'* constituye el total de puntos por clase según la realidad (columnas de la MC).

Por su parte, el *EC* considera los puntos que sin pertenecer a esa clase se han asignado a ella. Está relacionado con la *FU* que indica la probabilidad de que esos puntos realmente estén bien clasificados.

Siendo x_{i+} el valor del total de la fila i.

A *positive accuracy* o fiabilidad del usuario (*FU*) también se le conoce como *Recall, sensitivity* o exhaustividad ya que indica la proporción de píxeles positivos reales que son clasificados correctamente como tales, representando el *True Positive Rate* (TPR).

TP+FN constituye el total de puntos por clase según la predicción (filas de la MC).

La *precision* valora si la presencia o ausencia real es la adecuada, mientras que *recall* se encarga de estudiar la predicción o estimación realizada. Ambas medidas son complementarias y su valoración conjunta viene dada por el estadístico *F-measure* o *f1-score*, que en el caso más general utiliza una misma ponderación ($\beta = 1$) para ambas medidas, resultando la siguiente formulación (Li, et al. 2008):

Valores próximos a 1 indican que los valores de *recall* y *precision* también lo son, lo que significa una buena clasificación.

Señalar que estos cálculos resultan muy habituales en el ámbito de la teledetección a la hora de constatar clasificaciones y en los últimos tiempos también se están aplicando a otro tipo de datos como los proporcionados por el LiDAR (Gil-Yepes and Ruiz 2012).

En estos casos, en lugar del *f1-score* suele ser más habitual el uso del estadístico Cohen's Kappa (κ) (Congalton and Green 2008) que determina si un error de la matriz es significativamente diferente frente a otro y se calcula según la siguiente expresión:

$$\kappa = \frac{n\sum_{i=1}^{k} x_{ii} - \sum_{i=1}^{k} x_{i+} x_{+i}}{n^2 - \sum_{i=1}^{k} x_{i+} x_{+i}}$$

En esa fórmula, además de los valores anteriores, hay que tener en cuenta que *n* constituye el número total de la muestra (en la figura 2.9 n = x) y *k* el número de filas de la matriz.

5.3. VERIFICACIÓN DE RESULTADOS

En este apartado se presenta un análisis de los resultados obtenidos según los procesos descritos anteriormente. Para ello, se ha optado por mostrar por temáticas (edificaciones, carreteras y vegetación) por un lado las cuadrículas de la hoja 63 y después conjuntamente las de las hojas 61 y 64.

Previamente se ha presentado una comparativa entre las características de los dos vuelos, seguida de la descomposición en pasadas de cada una de las cuadrículas elegidas, presentando la valoración estadística tanto por líneas de vuelo como por ficheros originales de km $\times$ km. En esta valoración, cuando no se especifica una pasada en concreto los valores que se muestran hacen referencia a la cuadrícula completa.

A este respecto, cabe señalar que al estar los ficheros georreferenciados cortados según esta cuadrícula, algunas pasadas quedan sesgadas mostrando únicamente ruido, y en consecuencia no resultan significativas.

5.3.1. COMPARATIVA ENTRE EL VUELO GV Y EL VUELO DFG

Dentro de la hoja 63 se ha considerado la cuadrícula 5404782 de datos LiDAR que tiene la particularidad de tratarse de un archivo generado tanto a través del vuelo del GV (540482_c.las) como el de la DFG (540482_c_Gipuzkoa.las).

Este hecho permite realizar una comparativa sobre los datos que muestran los ficheros descargados una vez eliminados los valores anómalos, gracias a la utilidad lasinfo.exe de LAStools. La información correspondiente al vuelo de la DFG es la usada en la tabla 3.1, mostrando a continuación la referente al vuelo del GV.

Tabla 5.3. Información de la cabecera del LAS del GV 5404782_c

```
reporting all LAS header entries:
  file signature:           'LASF'
  file source ID:           0
  global_encoding:          0
  project ID GUID data 1-4:   0 0 0 ''
  version major.minor:      1.2
  system identifier:        ''
  generating software:      'TerraScan'
  file creation day/year:     58/2012
  header size:              227
  offset to point data:     227
  number var. length records: 0
  point data format:        3
  point data record length:   34
  number of point records:    1766801
  number of points by return: 1701457 64100 1164 69 11
  scale factor x y z:       2.32828e-007 2.32826e-007 6.20238e-008
  offset x y z:             540500.00499999989 4781500.0099999998 285.90500000000003
  min x y z:                540000.01000000001 4781000.0200000005 152.71000000000001
  max x y z:                541000 4782000 419.10000000000002
reporting minimum and maximum for all LAS point record entries ...
  X -2147483647 2147483647
  Y -2147483647 2147483647
  Z -2147483647 2147483647
  intensity 10 15790
  edge_of_flight_line 1 1
  scan_direction_flag 0 0
  number_of_returns_of_given_pulse 1 5
  return_number               1 5
  classification      1    13
  scan_angle_rank     60   120
  user_data           3    5
  point_source_ID     32   103
  gps_time 495723.350970 555589.119898
  Color R 4096 65280
        G 4096 65280
        B 4864 65280
number of last returns: 1701448
covered area in square units/kilounits: 989764/0.99
point density: all returns 1.79 last only 1.72 (per square units)
         spacing: all returns 0.75 last only 0.76 (in units)
overview over number of returns of given pulse: 1637347 125883 3284 232 55 0 0
histogram of classification of points:
           9726  Unclassified (1)
         793166  Ground (2)
          75568  Low Vegetation (3)
          79215  Medium Vegetation (4)
         197966  High Vegetation (5)
         219830  Building (6)
            770  Low Point (noise) (7)
           1700  Reserved for ASPRS Definition (10)
         315212  Overlap Points (12)
          73648  Reserved for ASPRS Definition (13)
```

En la tabla 5.4 se presentan comparados ambos vuelos, destacando que en el caso del fichero del GV se dispone de 91.774 puntos más que en el de la DFG. Además, el hecho de que presenten valores máximos y mínimos de *Z* no es relevante dada la captura irregular que siguen este tipo de sensores, al igual que la identificación de las pasadas. Por su parte, el hecho de que los valores RGB mostrados en cada caso sean muy similares es debido a que en ambos vuelos se ha utilizado la misma ortofotografía para la asignación de estos valores.

Tabla 5.4. Comparativa de los datos de GV y DFG del LAS 5404782

	GV		DFG	
Number of points	1.766.801		1.675.027	
Density	1,79		1,69	
Spacing	0,75		0,77	
	min	max	min	max
Z	152,71	419,10	153,81	418,67
Intensity	10	15790	0	255
Edge of flight line	1	1	0	1
Scan direction flag	0	0	0	1
Number of returns	1	5	1	4
Classification	1	13	0	11
Scan angle rank	60	120	-21	21
User data	3	5	115	146
Point source ID	32	103	14	17
R	4096	65280	3584	65280
G	4096	65280	4096	65280
B	4864	65280	4861	65280

Entre otras diferencias se pueden remarcar que:

- Las intensidades del GV se muestran con mayor número de bits que las de la DFG, lo que apunta a un problema de normalización.
- Los datos del GV tienen un retorno más que los de la DFG.
- Los ángulos de escaneo, mientras que en el caso del GV oscila entre 60 y 120 º en el otro lo hace entre -21 y 21 º. Por lo que parece que en el primer caso existen un error.
- Los valores de clasificación en el caso del GV varían entre 1 y 13 y en el vuelo de la DFG entre 0 y 11, lo que lleva a pensar en clasificaciones distintas que sin duda traerán consigo resultados distintos en los procesos especificados anteriormente.

Ésta última diferencia es la más relevante dado que se refiere a la clasificación y es objeto de esta investigación. En la siguiente tabla se ha tratado de mostrar esa comparativa tanto en valores absolutos como en porcentajes.

A la vista de los datos (tabla 5.5), se puede comprobar cómo el vuelo del GV no dispone de datos en las clases 0 y 11. Por su parte, el vuelo de la DFG no tiene puntos ni de clase 1, 10, 12 ni 13. Sin embargo, al considerar conjuntamente los de clase 1, 10, 12 y 13 (400286 puntos) en el vuelo del GV, se consigue un valor aproximado al que contiene el vuelo de la DFG sumando las clases 0 y 11 (429638 puntos), siendo éste ligeramente superior.

Tabla 5.5. Comparativa de la clasificación de los datos de GV y DFG del LAS 5404782

Clases		GV		DFG	
		puntos	%	puntos	%
Creado, nunca clasificado	0	--	--	429581	25.65
Sin clasificar	1	9726	0.55	--	--
Suelo	2	793166	44.89	506284	30.23
Vegetación baja	3	75568	4.28	46536	2.78
Vegetación media	4	79215	4.48	48269	2.88
Vegetación alta	5	197966	11.20	489659	29.23
Edificación	6	219830	12.44	153810	9.18
Puntos bajos (ruido)	7	770	0.04	831	0.05
Reservada definición ASPRS	10	1700	0.10	--	--
Reservada definición ASPRS	11	--	--	57	0.00
Puntos de solape	12	315212	17.84	--	--
Reservada definición ASPRS	13	73648	4.17	--	--
TOTAL		**1766801**	**100**	**1675027**	**100**

En cuanto a las cantidades por clases, se puede apreciar que únicamente en el caso de la clase 7 los resultados son similares, en el resto las discrepancias son considerables. Con el objeto de conseguir una clasificación más homogénea se ha realizado una reclasificación de las categorías al agrupar las clases originales, presentando las discrepancias de la tabla 5.6.

Tabla 5.6. Discrepancias en el histograma de la clasificación

Clases	GV	DFG	GV-DFG
Sin clasificar	400286 (clases 1, 10, 12, 13)	429581 (clase 0)	-29295
Suelo*(clase 2)*	793166	506284	286882
Vegetación (clases 3, 4, 5)	352749	584464	-231715
Edificación (clases 6)	219830	153810	66020
Puntos bajos (ruido) (clase 7)	770	831	-61

En consecuencia, cabe mencionar que en el fichero de la DFG hay mayor número de puntos sin una clase asignada. Además, en este archivo el número de puntos en las clases referentes a la vegetación es mayor que en el caso del GV, siendo ligeramente superior en el caso del ruido.

Por su parte, los datos del GV frente a los de la DFG presentan mayor cantidad de puntos clasificados como terreno y edificación, pero en ningún caso aparecen ni puntos de la clase 8 ni de la 9.

En cuanto a las otras hojas de referencia, a continuación se muestran los ficheros correspondientes a las cuadrículas elegidas en las hojas 61 (tabla 5.7) y 64 (tabla 5.8) una vez depurados. De la información aportada de estos archivos se desprende que en ambos casos los datos están procesados con TerraScan y responden a ficheros .LAS según la versión 1.2 y el formato 3 (PRDF).

Tabla 5.7. Información de la cabecera del LAS del GV 5094793_c

```
reporting all LAS header entries:
  file signature:             'LASF'
  file source ID:             0
  global_encoding:            0
  project ID GUID data 1-4:   0 0 0 ''
  version major.minor:        1.2
  generating software:        'TerraScan'
  file creation day/year:     53/2012
  header size:                227
  offset to point data:       227
  number var. length records: 0
  point data format:          3
  point data record length:   34
  number of point records:    3141736
  number of points by return: 3068002 72003 1697 30 4
  scale factor x y z:         2.32828e-007 2.32826e-007 2.98885e-008
  offset x y z:               509500.00499999995 4792500.0099999998 79.365000000000023
  min x y z:                  509000.01000000001 4792000.0200000005 15.18
  max x y z:                  510000 4793000 143.55000000000001
reporting minimum and maximum for all LAS point record entries ...
  X -2147483647 2147483647
  Y -2147483647 2147483647
  Z -2147483647 2147483647
  intensity 10 64914
  edge_of_flight_line 1 1
  scan_direction_flag 0 0
  number_of_returns_of_given_pulse 1 5
  return_number                 1 5
  classification      0    13
  scan_angle_rank     60   120
  user_data           3    5
  point_source_ID     88   93
  gps_time 56706.040683 61229.344547
  Color R 256 65280
        G 2048 65280
        B 1792 65280
number of last returns: 3067927
covered area in square units/kilounits: 998088/1.00
point density: all returns 3.15 last only 3.07 (per square units)
     spacing: all returns 0.56 last only 0.57 (in units)
overview over number of returns of given pulse: 2995926 140685 5001 104 20 0 0
histogram of classification of points:          7404  Created, never classified (0)
         221318  Unclassified (1)
        1662234  Ground (2)
            359  Low Vegetation (3)
         522755  High Vegetation (5)
         620581  Building (6)
            476  Low Point (noise) (7)
           1733  Reserved for ASPRS Definition (10)
          34386  Reserved for ASPRS Definition (11)
           1831  Overlap Points (12)
          68659  Reserved for ASPRS Definition (13)
```

Tabla 5.8. Información de la cabecera del LAS del GV 5694788_c

```
reporting all LAS header entries:
  file signature:           'LASF'
  file source ID:           0
  global_encoding:          0
  project ID GUID data 1-4:   0 0 0 ''
  version major.minor:      1.2
   generating software:      'TerraScan'
  file creation day/year:    63/2012
  header size:              227
  offset to point data:     227
  number var. length records: 0
  point data format:        3
  point data record length:   34
  number of point records:    2737375
  number of points by return: 2348174 325698 58573 4930 0
  scale factor x y z:       2.32828e-007 2.32831e-007 6.08945e-008
  offset x y z:             569499.995 4787500 230.77000000000004
  min x y z:                569000 4787000 100
  max x y z:                569999.98999999999 4788000 361.54000000000002
reporting minimum and maximum for all LAS point record entries ...
  X -2147483647 2147483647
  Y -2147483647 2147483647
  Z -2147483647 2147483647
  intensity 0 58239
  edge_of_flight_line 0 1
  scan_direction_flag 0 1
  number_of_returns_of_given_pulse 0 7
  return_number               0 7
  classification      0   11
  scan_angle_rank   -24   100
  user_data         0   199
  point_source_ID    17 50934
  gps_time 0.000000 563469.474328
  Color R 256 65280;        G 1536 65280;        B 1280 65280
number of last returns: 2346762; covered area in square units/kilounits: 999616/1.00
point density: all returns 2.74 last only 2.35 (per square units)
     spacing: all returns 0.60 last only 0.65 (in units)
number of points by return is different than reported in header: 2347187 325725 58749 5081
156
WARNING: there are 163 points with return number 0
WARNING: there are 156 points with return number 6
WARNING: there are 158 points with return number 7
overview over number of returns of given pulse: 2021530 532162 162382 20409 183 159 128
WARNING: there are 422 points with a number of returns of given pulse of 0
histogram of classification of points:          922513  Created, never classified (0)
         995177  Ground (2)
          53766  Low Vegetation (3)
          52516  Medium Vegetation (4)
         629476  High Vegetation (5)
          83594  Building (6)
            222  Low Point (noise) (7)
            111  Reserved for ASPRS Definition (11)
```

En el caso de la hoja 61 (tabla 5.7), la cuadrícula 5094793 del vuelo del GV contiene 3.141.736 puntos con una densidad de 3,15 por metro cuadrado y con un espaciado entre ellos de 0,56 m. En cuanto a la clasificación se refiere, con respecto al análisis anterior, con un total de 5 retornos aparecen puntos en la clase 0 y 11, desapareciendo la clase 4 (vegetación media), puede que sea por falta de éste tipo de vegetación en esta zona.

Al analizar el fichero de Aia del vuelo de la DFG (tabla 5.8, cuadrícula 5694788), se puede comprobar que contiene un total de 2.737.375 puntos con una densidad de 2,74 por metro cuadrado y un espaciado entre ellos de 0,60 m. En cuanto a la clasificación se refiere, se mantienen las clases analizadas anteriormente con un total de 8 retornos (0 - 7).

A este respecto, remarcar el apunte del informe de la tabla 5.8 en el que se indica que existen puntos con números de retornos 0, 6 y 7 que no están contemplados en la cabecera, existiendo incongruencia entre los datos que se dan. Al examinar visualmente el fichero por retornos sólo se han detectado hasta cuatro.

Al equiparar las informaciones de los cuatro ficheros disponibles se puede concluir que:

- La densidad es algo mayor en el caso del archivo 5094793, lo que conlleva a un menor espaciado, resultando muy similares en el caso de la cuadrícula 5404782.
- La cuadrícula 5094793 es la que mayor cantidad de puntos ofrece, llegando casi a duplicar los puntos de la 5404782.
- El fichero 5694788 presenta algún problema de manipulación, tanto en cuanto a los retornos como en el valor de la intensidad, ya que en el vuelo de la DFG éstas se dan en 8 bits.
- En todos los casos la mayor parte de los puntos pertenecen al último retorno.

5.3.2. RECUPERACIÓN DE PASADAS

Continuando con el proceso indicado en la metodología, una vez analizados los datos de los ficheros originales limpiados, se plantea recuperar las pasadas del vuelo original.

Esto se consigue a través de lassplit.exe de LAStools, obteniendo en cada caso su distribución de líneas de vuelo iniciales. Seguidamente, en las figuras 5.3, 5.4, 5.5 y 5.6 se muestran por alturas las pasadas de cada una de las hojas consideradas; además, en la figura 5.7 se muestra la superposición de las pasadas de la figura 5.3, donde se puede apreciar que hay pasadas que casi se solapan el 100 % pero que entre otras el solapamiento es nulo.

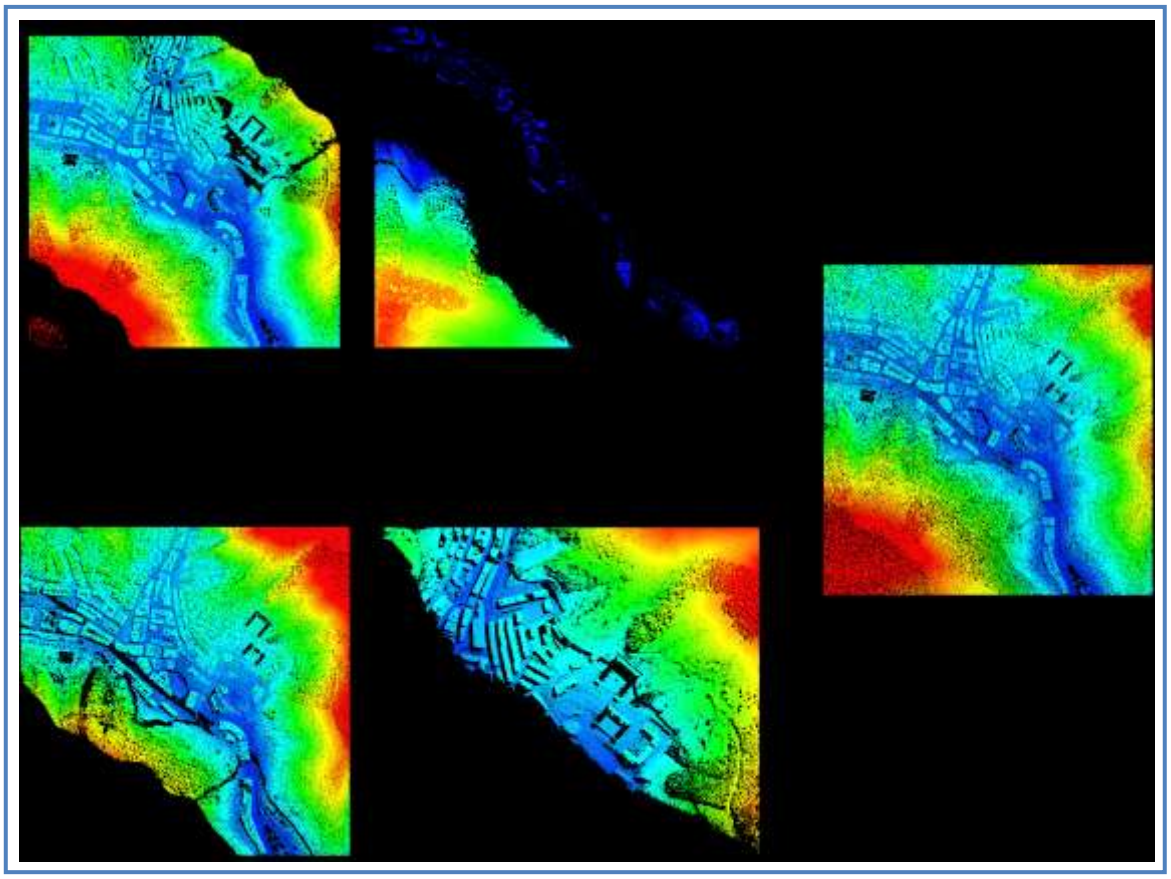

Figura 5.3. Pasadas del fichero GV 5404782_c

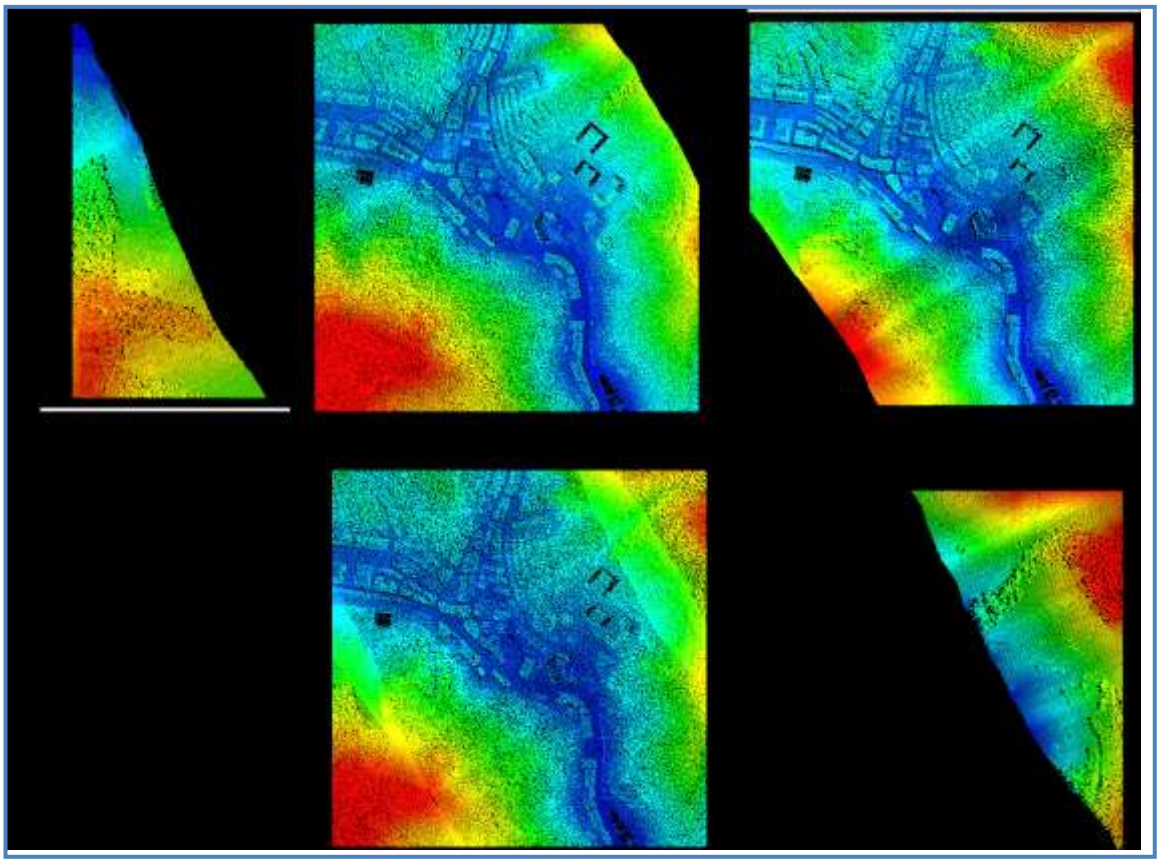

Figura 5.4. Pasadas del fichero DFG 5404782_c_Gipuzkoa

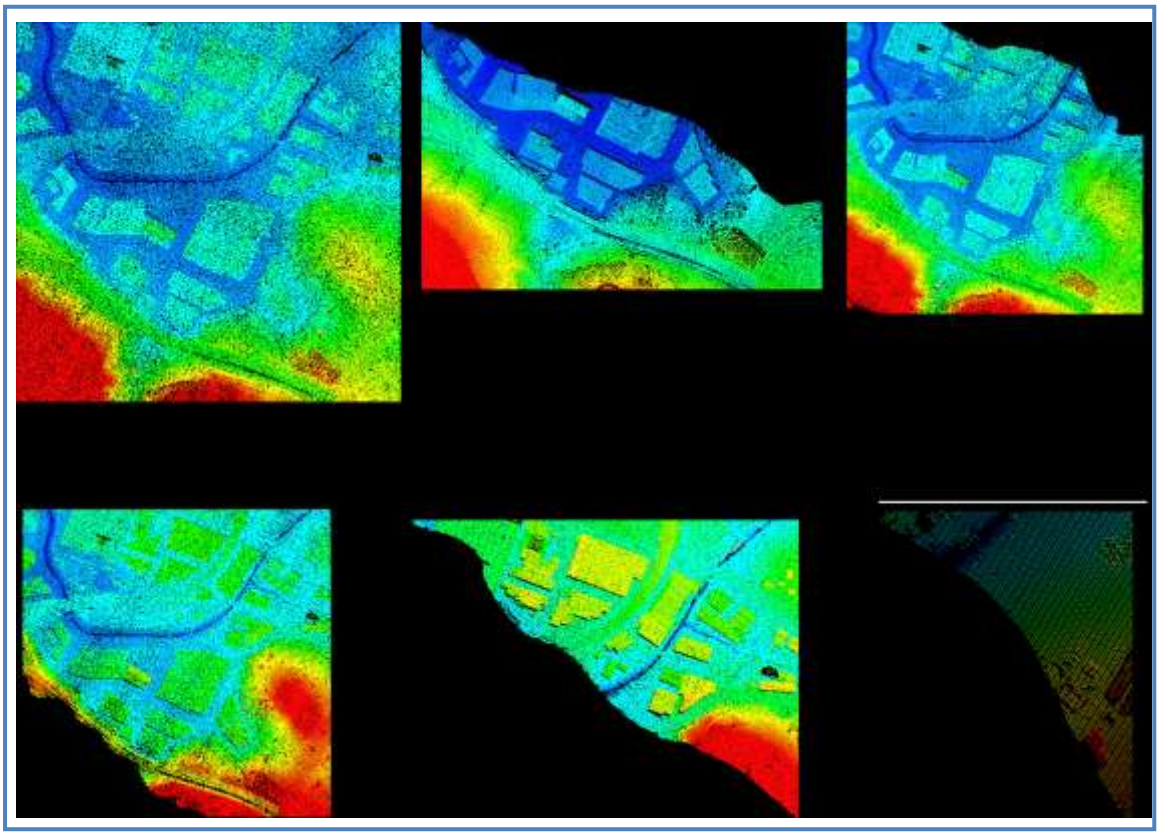

Figura 5.5. Pasadas del fichero GV 5094793_c

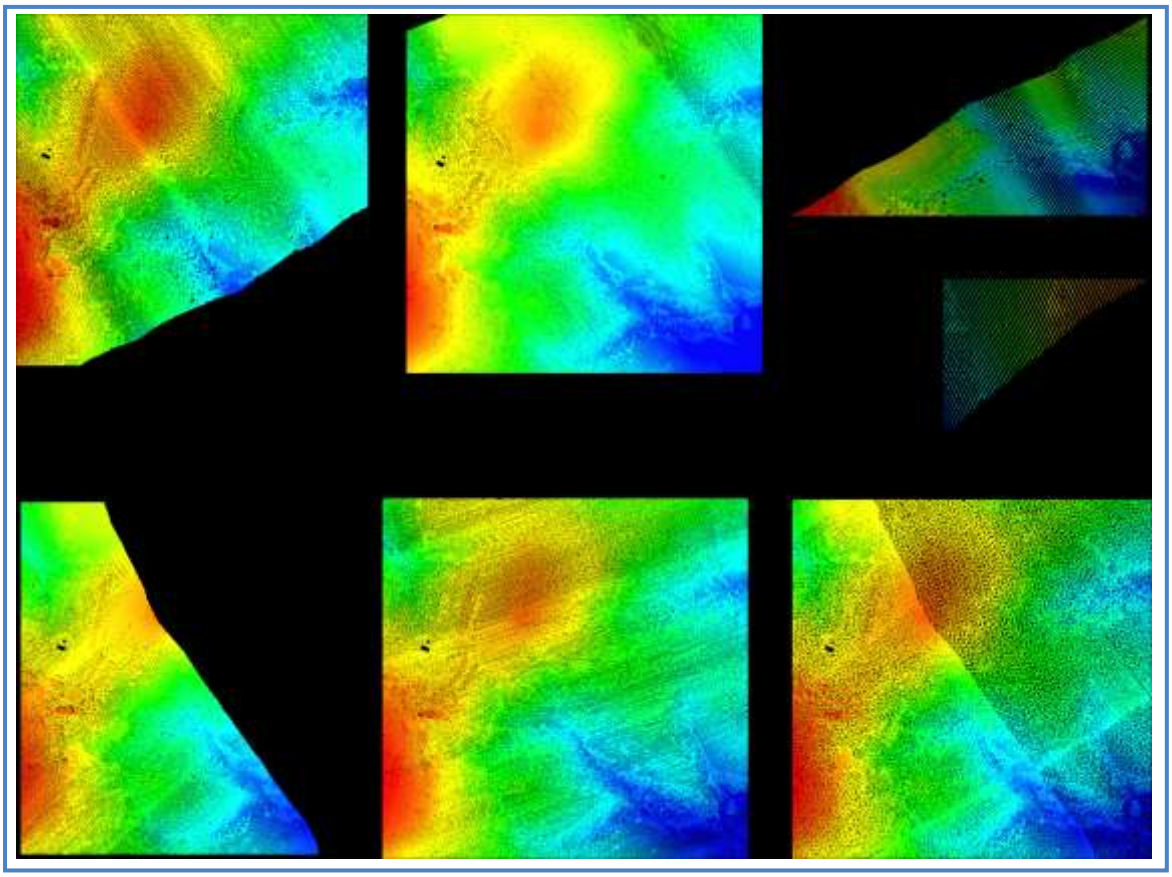

Figura 5.6. Pasadas del fichero DFG 5694788_c_Gipuzkoa

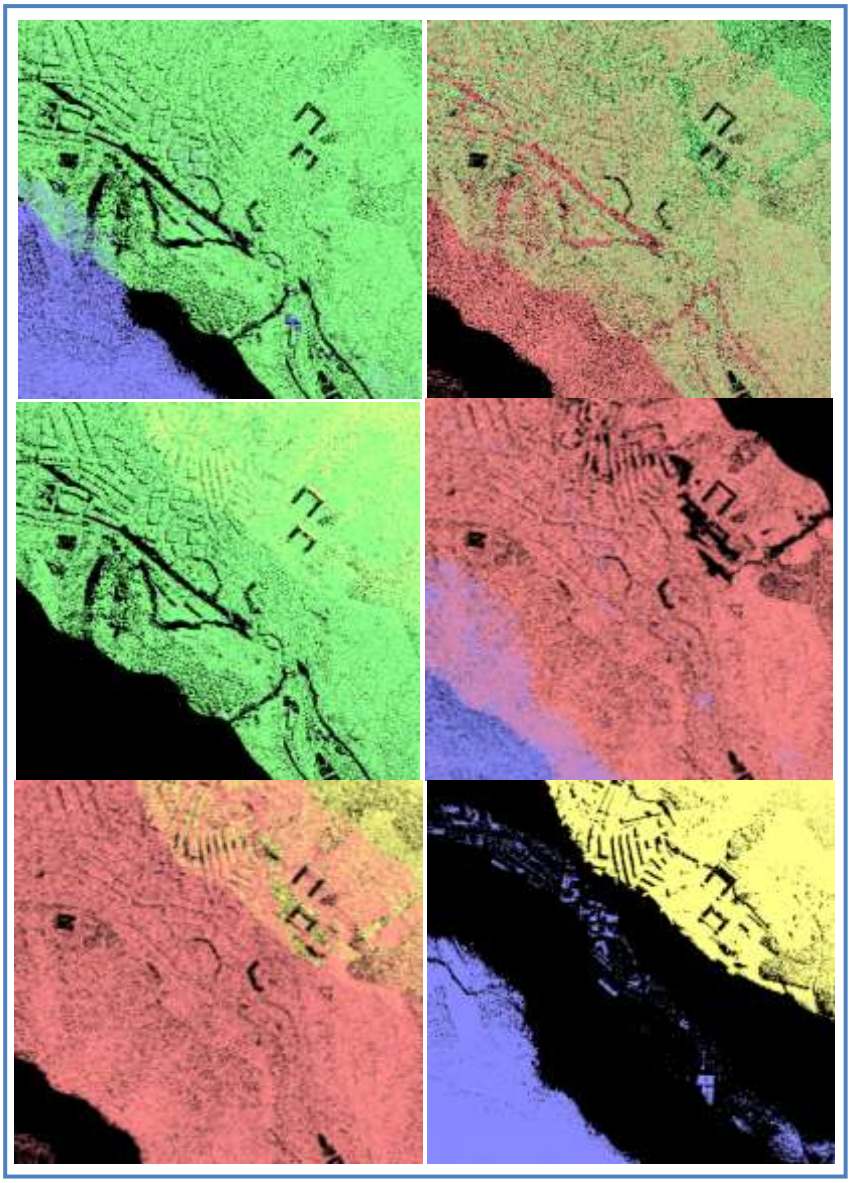

Figura 5.7. Solapamiento dos a dos de las pasadas del fichero GV 5404782_c

En el primer solapamiento se puede apreciar que las pasadas 00103 (verde) apenas solapa con la 00054 (lila). En la segunda imagen, el solapamiento entre la 00103 (verde) y la 00032 (rosa) se da en la parte central, quedando los extremos de ambas sin coincidir. La tercer imagen muestra como la 00103 (verde) cubre por completo la 00095 (amarillo). En la cuarta imagen se aprecia muy poca coincidencia entre la 00032 (rosa) y la 00054 (lila), siendo en la quinta imagen el solapamiento de parte de la 00032 (rosa) con parte de la 00095 (amarillo). Finalmente, en la última imagen no se produce coincidencia entre las pasadas 00095 (amarillo) y la 00054 (lila).

A continuación, se comentan hoja por hoja los aspectos más relevantes a este respecto.

5.3.2.1. Hoja 63

La cuadrícula considerada en la hoja 63 es la correspondiente a los ficheros 5404782_c.las del vuelo del GV y 5404782_c_Gipuzkoa.las del vuelo de la DFG.

En el caso del fichero facilitado por el GV, se consideran las pasadas (*flightlines*) 00032, 00054, 00095 y 00103, ubicándose la mayor parte de la zona en las pasadas 00032 y 00103, siendo la 00054 y la 00095 complementarias a las otras dos (figura 5.3). La 00035 se desestima por contener únicamente algunos puntos aislados. La superposición de estas pasadas se puede apreciar en la figura 5.7, a la que corresponden el comentario posterior.

Tal y como se puede apreciar en la figura 5.4 del vuelo de la DFG, aparecen cuatro pasadas distintas: 00014, 00015, 00016 y 00017. Las pasadas 00014 (azul, figura 5.8) y 00017 (amarillo, figura 5.8) únicamente abarcan una pequeña parte del total de la hoja, concretamente la esquina inferior izquierda y la superior derecha, respectivamente; pero constituyen las partes que faltan en las pasadas 0016 (magenta, figura 5.8)y 0015 (cian, figura 5.8) que en este caso se superponen entre sí en un porcentaje muy alto.

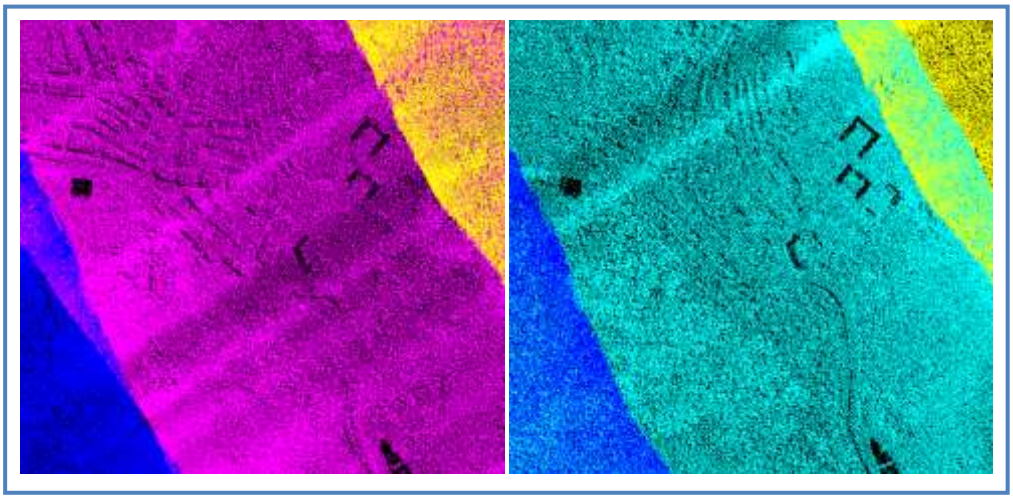

Figura 5.8. Solapamiento de las pasadas del fichero DFG 5404782_c_Gipuzkoa

5.3.2.2. Hoja 61

La cuadrícula considerada en la hoja 61 es la correspondiente al fichero 5094793_c.las del vuelo del GV. Se corresponde con un polígono industrial.

Al redistribuir la información del mismo según la dirección de las pasadas aparecen 6 pasadas: 00088, 00089 (amarillo), 00090 (lila), 00091 (verde), 00092 (rosa) y 00093 (cian), ésta última con muy pocos puntos. Además, señalar que no se ha considerado la 00088 por estar constituida por puntos aislados (figura 5.5).

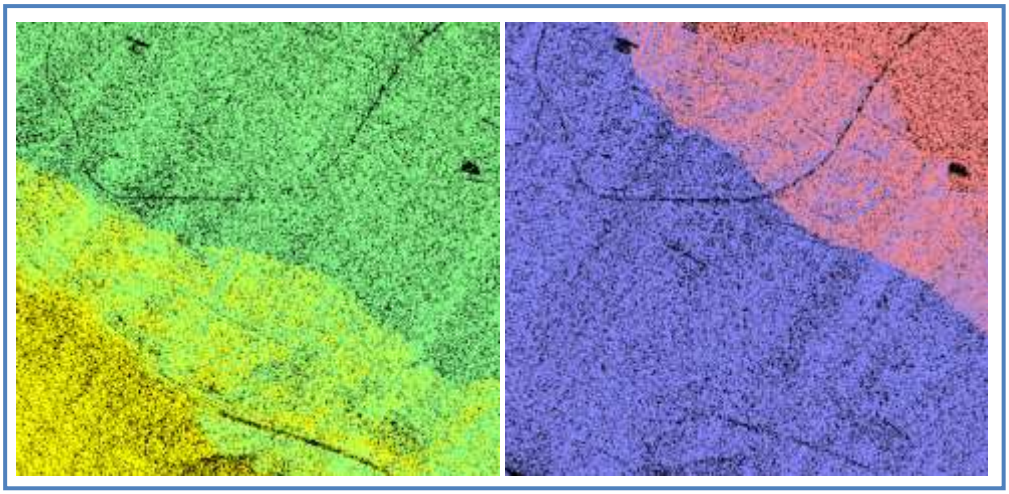

Figura 5.9. Solapamiento de las pasadas del fichero GV 5094793_c

5.3.2.3. Hoja 64

En el caso de Gipuzkoa, dentro de la hoja 63 se ha considerado el fichero 5694788_c_Gipuzkoa.las, perteneciente al vuelo de la DFG.

Se trata de parte del pueblo de Aia y está constituido también por 6 pasadas (figura 5.6): 00017, 00018, 00019, 00020, 00048 y 00049. La pasada 00049 (verde fuerte) ocupa la totalidad de la cuadrícula y la 00019 (verde claro) prácticamente también abarcando un poco menos pero casi en su totalidad la 00018 (lila), siendo la 00048 (rojo) la que menor porción ocupa; sin embargo, las pasadas 00017 y 00020 no contienen información relevante, aunque esta última será analizada en algunas de las próximas secciones. Se trata de la cuadrícula con mayor solapamiento de las tratadas.

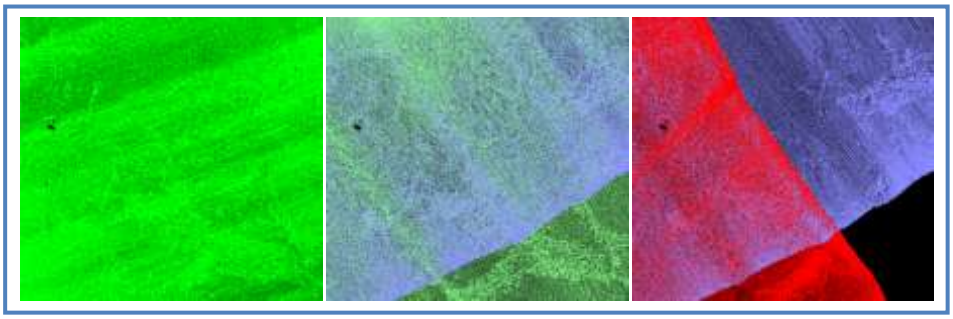

Figura 5.10. Solapamiento de las pasadas del fichero DFG 5694788_c_Gipuzkoa

5.3.3. ANÁLISIS DE LA CLASIFICACIÓN DE LAS EDIFICACIONES

En este apartado se ofrece un estudio de los resultados obtenidos de las cuadrículas analizadas por hojas, considerando en este caso las edificaciones.

5.3.3.1. Análisis de edificaciones en la hoja 63

De acuerdo a la formulación indicada en el apartado de valoración estadística se han calculado los *EO* y *EC* con sus respectivas fiabilidades para el fichero completo y para cada una de las pasadas existentes en la cuadrícula analizada, tanto para el vuelo del GV como para el del la DFG.

Tabla 5.9. Errores de omisión y comisión en la categoría de edificación (5404782)

HOJA 63 (GV)	*EO*	*FP*	*EC*	*FU*	*f1-score*
5404782_c_clean.las	**0,26**	**0,74**	**0,12**	**0,88**	**0,80**
flithlines.00032.las	0,21	0,79	0,12	0,88	
flithlines.00054.las	0,64	0,36	0,10	0,90	
flithlines.00095.las	0,35	0,65	0,10	0,90	
flithlines.00103.las	0,29	0,71	0,13	0,87	
HOJA 63 (DFG)	***EO***	***FP***	***EC***	***FU***	***f1-score***
5404782_c_Gipuzkoa_clean.las	**0,09**	**0,91**	**0,14**	**0,86**	**0,88**
flightlines.00014.las	0,20	0,80	0,02	0,98	
flightlines.00015.las	0,08	0,92	0,16	0,84	
flightlines.00016.las	0,09	0,91	0,14	0,86	
flightlines.00017.las	0,45	0,55	0,17	0,83	

En cuanto al Error de Comisión del **vuelo del GV**, se ha comprobado que todos los valores son muy similares, siendo los más pequeños y con el mismo valor (0,10) los de las pasadas 00054 y 00095.

Con cuantías ligeramente superiores se encuentran las pasadas 00032 (0,12) y 00103 (0,13), cuyos resultados son prácticamente similares al del fichero completo. Gráficamente se aprecia que la 00032 tiene un *EC* algo menor que la 00103. En la figura 5.11 se muestran en verde estos errores de comisión para estas pasadas.

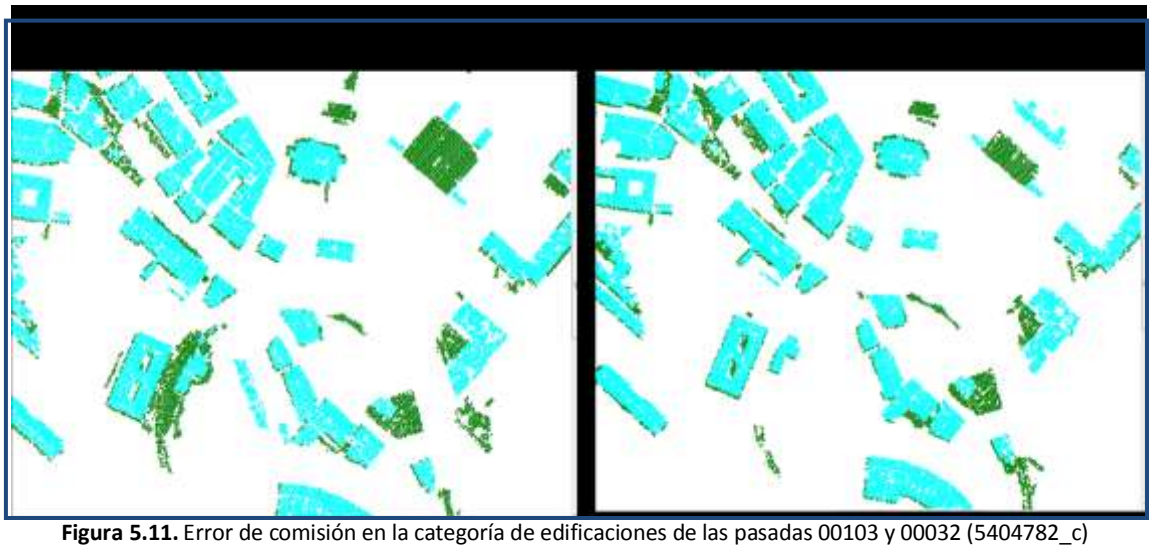

Figura 5.11. Error de comisión en la categoría de edificaciones de las pasadas 00103 y 00032 (5404782_c)

Figura 5.12. Error de comisión en la categoría de edificaciones GV (5404782_c)

Figura 5.13. Error de comisión en la categoría de edificaciones DFG (5404782_c_Gipuzkoa)

Figura 5.14. Error de omisión en la categoría de edificaciones GV (5404782_c)

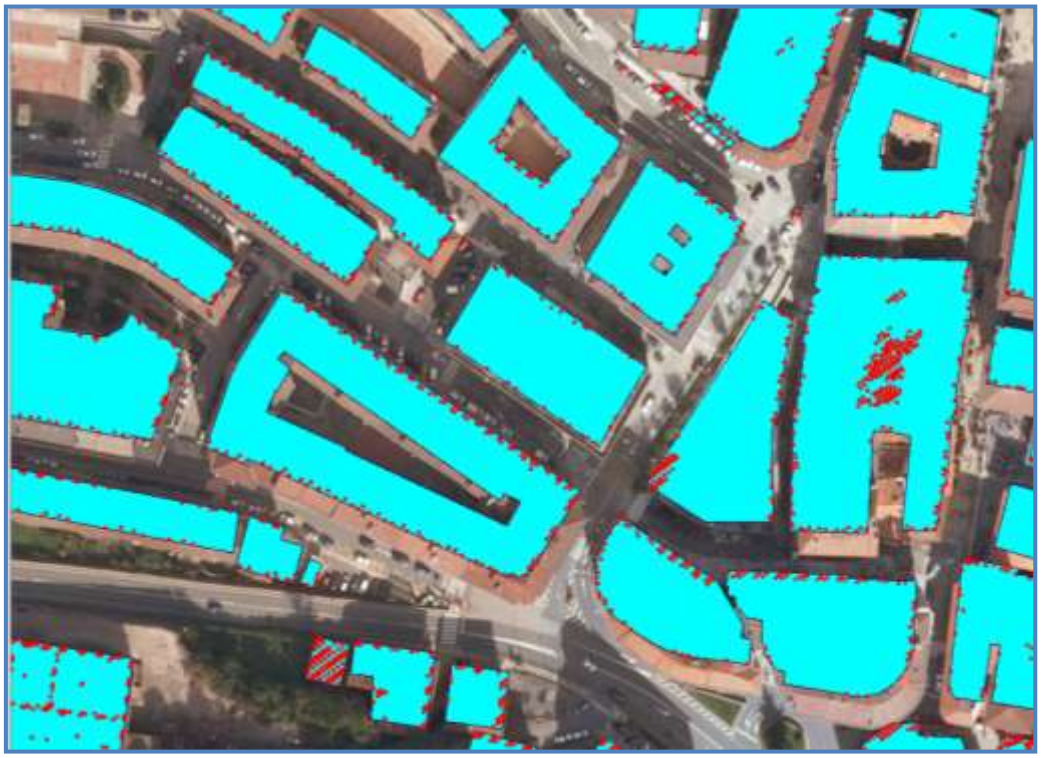

Figura 5.15. Error de omisión en la categoría de edificaciones DFG (5404782_c_Gipuzkoa)

Al analizar el **vuelo de la DFG** se ha observado que la mejor *FU* y en consecuencia el menor *EC* se consigue con la pasada 00014 (*FU* = 0,93; *EC*= 0,02), resultando el resto de las pasadas con unos valores superiores, siendo el mayor de todos el de la pasada 00017 (*EC* = 0,17), dándose la casualidad que ambas pasadas ocupan una pequeña parte de la cuadrícula completa.

En general, sin tener en cuenta la pasada 00014, el resto de las líneas de vuelo muestran unos valores muy similares entre sí y con respecto al fichero completo, siendo algo superiores a los valores mostrados para el vuelo del GV, aunque visualmente se aprecien menos puntos para el vuelo DFG (figuras 5.12 y 5.13). En ambos vuelos, se ha comprobado que el *EC* se sitúa principalmente por fuera del contorno que define la línea de edificación de la BTA, así como en algunos patios interiores y aquellas entradas que dan lugar a plazas en las fachadas de los edificios. Además, el *EC* también aparece en algunas rotondas y calles, incluso plazas.

Al considerar el buffer de 2 m para las intersecciones, tal y como se presumía y se puede ver en la tabla 5.10, este error disminuye considerablemente en ambos vuelos obteniendo unas *FU* muy buenas (GV = 0,95; DFG = 0,97), pero trae consigo un aumento del *EO,* en el caso del GV de 0,26 a 0,39 y en el caso de la DFG de 0,09 a 0,22. Por su parte, el *f1-score* también disminuye en ambas cuadrículas.

Tabla 5.10. Comparativa de errores en la categoría de edificación considerando el buffer (5404782)

	GV	GV (buffer 2 m)	DFG	DFG (buffer 2 m)
EO	0,26	0,39	0,09	0,22
FP	0,74	0,61	0,91	0,78
EC	**0,12**	**0,05**	**0,14**	**0,03**
FU	0,88	0,95	0,86	0,97
f1-score	0,80	0,74	0,88	0,86

Respecto al Error de Omisión (tabla 5.9) señalar que es mayor que el *EC* en el **vuelo del GV**, alcanzando cuantías muy altas para las pasadas 00054 (0,64) y 00095 (0,35). En el caso de la 00054 puede ser por carecer prácticamente de edificios, pero la 00095 comprende aproximadamente la mitad de los edificios incluidos en la muestra. El resto de las pasadas, 00032 y 00103, que abarcan casi la totalidad de los edificios de estudio, adquieren unos valores similares para el *EO* y *FP* y muy parecidos a los del fichero original (0,26).

Al analizar la representación gráfica (figura 5.14) se ha apreciado que gran parte del *EO* hace referencia a puntos con valor de clasificación 13 (Reservados para la ASPRS) y que en su mayoría están ubicados en los bordes de las líneas de edificación de la BTA, aunque también aparecen bastantes por el interior de los tejados.

Por su parte, en lo que respecta al **vuelo de la DFG**, se ha estimado que los *EO* son menores que en el caso del GV salvo en la pasada 00017, en la cual se puede constatar que apenas aparecen edificaciones, lo que lleva a obtener un valor de 0,55 para la *FP*.

Sin embargo, en la pasada 00014, en la que también aparecen pocas edificaciones, se han obtenido valores más aceptables en cuanto al *EO* (0,20) y *FP* (0,80), valores bastante inferiores a los alcanzados en el caso del GV, aunque superiores al comparar con el resto de pasadas y el fichero completo. Las líneas de vuelo 00015 y 00016, que abarcan casi la totalidad de la cuadrícula de estudio, muestran unas cuantías muy similares entre sí y a las del fichero completo (*EO* = 0,09).

En la visualización gráfica de este error se ha comprobado que en gran parte coincide con bordes de tejados, puntos de suelo o fachada, incluso balcones, que caen dentro de la intersección con la BTA. Estos puntos en su mayoría coinciden con puntos de clase 0 (sin clasificar) ó 2 (suelo). Señalar que, tal y como ya se ha comentado, este error aumenta (tabla 5.10) considerablemente al considerar el buffer de 2 m por el tema de los aleros.

Al equiparar ambos vuelos, aparentemente la visualización de la DFG parece mejor que la del GV pero si se repara en su cuantificación se puede apreciar que el *EC* es ligeramente inferior en el caso del vuelo del GV (0,12)que en el de la DFG (0,14), no así el *EO* siendo casi 3 veces superior el del GV (0,26) al de la DFG (0,09). Pudiendo **concluir** que los datos del vuelo del GV ofrecen una ligera mejoría en lo que respecta al *EC* sin discrepar mucho uno del otro. Sin embargo, en el caso del *EO* se aprecia una clara diferencia entre ambos vuelos, ofreciendo mejores resultados el de la DFG.

Con el fin de deducir de manera más objetiva cuál de los dos vuelos está mejor clasificado, se ha considerado importante calcular el *f1-core* (ecuación 5.7) de los cuadrículas kilométricas. La ventaja que aporta este estadístico es que estudia conjuntamente el comportamiento de las dos fiabilidades y dice que la mejor clasificación viene dada por el valor 1. Por lo que, a la vista de los resultados, se puede decir que los datos de la DFG (*f1-core* = 0,88) están mejor clasificados que los del GV (*f1-core* = 0,80).

Estos resultados conllevan a realizar un **análisis** de las posibles tendencias para comprender mejor la **omisión** mostrada. Para ello, se procede a ver a qué clases pertenecen los puntos que debiendo ser edificios no han sido así catalogados.

Tabla 5.11. Análisis del error de omisión por clases en la categoría de edificación GV (5404782_c)

HOJA 63 (GV)	*EO* (%)	% clases 10 y 13	% clase 2	% clases 3, 4, 5	% clase 12	% clase 1	% clase 7	TOTAL
5404782_c_clean.las	*26,16*	**71,62**	18,13	5,03	3,00	2,08	0,14	100
flightlines.00032.las	*20,71*	**75,02**	15,89	5,54	3,02	0,42	0,11	100
flightlines.00054.las	*63,51*	33,54	7,06	2,08	0,90	**54,95**	1,47	100
flightlines.00095.las	*35,03*	**80,00**	13,72	3,36	2,85	0,04	0,03	100
flightlines.00103.las	*28,52*	**69,24**	21,70	5,31	3,17	0,47	0,11	100

Tabla 5.12. Análisis del error de omisión por clases en la categoría de edificación DFG (5404782_c_Gipuzkoa)

HOJA 63 (DFG)	*EO* (%)	% clase 0	% clase 2	% clases 3, 4, 5	% clase 7	TOTAL
5404782_c_Gipuzkoa_clean.las	*9,41*	**48,75**	45,23	5,20	0,82	100
flightlines.00014.las	*19,93*	**56,42**	41,66	1,92	--	100
flightlines.00015.las	*8,28*	43,94	**49,97**	5,45	0,64	100
flightlines.00016.las	*8,78*	**49,23**	44,25	5,31	1,21	100
flightlines.00017.las	*45,28*	**70,71**	16,93	12,36	--	100

Atendiendo a las clases que aparecen en el fichero de información se han determinado las categorías a analizar. En el caso del vuelo de la DFG (tabla 5.12) se han contemplado la clase 0 de puntos sin clasificar, la clase 2 de terreno y las clases 3, 4 y 5 como vegetación. Para el segundo vuelo (GV, tabla 5.11) se han considerando la clase 10 y 13 conjuntamente por ser ambas reservadas para la definición de la ASPRS, la clase 2 con puntos de suelo, las clases 3, 4 y 5 como vegetación, la clase 12 con puntos de solape y la clase 1 con puntos sin asignar. En este vuelo no existen puntos en la clase 0 y para ambos se ha incluido la clase 7 referente al ruido por ser la única con valores similares, aunque nada relevantes.

Del análisis de esas dos tablas se desprende que el *EO* del vuelo del GV se encuentra en las clases 10 y 13 (71,62 %) y más concretamente en la clase 13 debido a que tiene mayor cantidad de puntos que la 10. Además, el valor obtenido es superior al mostrado por la clase 0 en el caso del vuelo de la DFG (48,75 %).

En ambos casos, la omisión va seguida por la clase 2, resultando bastante mayor en el caso de la DFG (45,23 %) que se podría considerar complementaria de lo catalogado dentro de la clase 0, para poder compararlo con el vuelo del GV, aunque en éste faltarían por incluir las clases 12 y 1.

La respuesta de la vegetación y el ruido (prácticamente nula) resulta similar en ambos vuelos, apareciendo entre medias de ambas las categorías 12 y 1, en el vuelo del GV. Si bien la clase 1 se podría despreciar hay que mencionar que en el caso de la pasada 00054 ésta supone la principal causa de omisión (54,95 %), por encima de la correspondiente a las clases 10 y 13 conjuntamente.

También habría que reseñar que en el caso de la DFG, en la pasada 00015 la omisión de la clase del 2 (49,97 %) se sitúa por encima de la clase 0. Con todas las excepciones contempladas se podría establecer que la mayor causa de omisión viene dada por las clases 10, 13, 12, 1 para el GV y 0 para la DFG, seguidas por la clase 2 y finalmente las correspondientes a la vegetación (3,4,5).

Por lo que respecta a las muestras, mencionar que en el caso de algunas pasadas, dada la escasez de puntos o la toma sectorial de los mismos, se observan valores discrepantes que presumiblemente habría que obviar (00054), pero no así en otros (00015).

5.3.3.2. Análisis de edificaciones en las hojas 61 y 64

En la tabla 5.13 se puede comprobar los resultados en cuanto al *EO* y *EC* obtenidos para los ficheros de las hojas 61 y 64. En el caso de la hoja 61, señalar que la pasada 00093 ofrece tanto un *EO* (0,68) como *EC* (0,10) muy altos con respecto al resto de pasadas. Al visualizar la pasada se puede comprobar cómo está constituida por muy pocos puntos. Mientras que en la cuadrícula de la hoja 64, pasa algo similar con la pasada 00020 en la que la cantidad de edificios encontrados es mínimo, contribuyendo a aumentar los valores del *EO* (0,29)y del *EC*

(0,47)considerablemente. Estos hechos llevan a pensar que para obtener resultados más objetivos se debería prescindir de los valores conseguidos con estas pasadas.

Tabla 5.13. Errores de omisión y comisión en edificaciones GV (5094793_c) y DFG (5694788_c_Gipuzkoa)

HOJA 61 (GV)	*EO*	*FP*	*EC*	*FU*	*f1-score*
5094793_c_clean.las	0,12	0,88	0,03	0,97	0,92
flithlines.00089.las	0,10	0,91	0,02	0,98	
flithlines.00090.las	0,10	0,90	0,02	0,98	
flithlines.00091.las	0,12	0,88	0,03	0,97	
flithlines.00092.las	0,16	0,84	0,03	0,97	
flithlines.00093.las	0,68	0,32	0,10	0,90	
HOJA 64 (DFG)	***EO***	***FP***	***EC***	***FU***	***f1-score***
5694788_c_Gipuzkoa_clean.las	0,10	0,89	0,26	0,74	0,81
flightlines.00018.las	0,09	0,91	0,24	0,76	
flightlines.00019.las	0,11	0,89	0,24	0,76	
flightlines.00020.las	0,29	0,71	0,47	0,53	
flightlines.00048.las	0,10	0,90	0,27	0,73	
flightlines.00049.las	0,10	0,90	0,26	0,74	

Sin considerar esas pasadas, al analizar el Error de Comisión de ambas hojas se puede comprobar cómo en general éste es bastante inferior en la cuadrícula de la hoja 61 (*FU* = 0,97) que en la de la hoja 64 (*FU* = 0,74), mostrando en general unos valores muy bajos para los errores con una *FU* muy buena.

Al igual que sucedía en el análisis de la hoja 63, los resultados de las pasadas, salvo ciertas excepciones, son muy similares a los de las cuadrículas completas. Y el hecho de que el archivo de DFG tenga un *EC* (0,26) mayor al del GV (0,03), a priori no tiene una explicación determinada, aunque al tratarse de una zona industrial la influencia de los aleros suele ser menor que en edificaciones propias del casco urbano.

Lo que sí que es verdad es que al considerar el buffer de 2 m (tabla 5.14), como en esta hoja del GV apenas hay *EC*, con el buffer éste aumenta mínimamente (0,08) y disminuye el *f1-score*. Sin embargo, en la cuadrícula de la DFG se produce un comportamiento similar al de la hoja 63, mermando el *EC* pero aumentando el *EO* aumentando la *FU* de 0,74 a 0,95. Esta diferencia es menor en el caso de la cuadrícula de GV que abarca el polígono industrial (5094793), pero que también aumenta llegando a una *FU* casi total con el 0,99.

Tabla 5.14. Comparativa de errores en edificación con el buffer: GV (5094793) y DFG (5694788_c_Gipuzkoa)

	GV	GV (buffer 2 m)	DFG	DFG (buffer 2 m)
EO	0,12	0,21	0,10	0,28
FP	0,88	0,79	0,89	0,89
EC	**0,03**	**0,08**	**0,26**	**0,05**
FU	0,97	0,99	0,74	0,95
f1-score	0,92	0,89	0,81	0,92

Al examinar el Error de omisión de ambos archivos (tabla 5.13) se puede comprobar que ambas cuadrículas tienen un comportamiento similar, con una *FP* próxima al 0,90. Sin

embargo, al considerar el buffer de 2 m (tabla 5.14) el *EO* aumenta más en ésta (0,28)que en la del polígono industrial (0,21). Por su parte, el *f1-score* de Aia (DFG) es mejor con un 0,92 que el del polígono (GV) que alcanza un 0,88.

Cabe reseñar que se trata de dos zonas con tipología de edificaciones muy distintas: en el caso de la hoja 61 se está trabajando con una zona industrial con grandes pabellones; mientras que en la hoja 64 se ha considerado una parte del casco urbano de Aia, en la que el entramado urbano responde a un casco sin muchas edificaciones rodeado de monte.

Ante esta situación, se puede decir que en parte el *EC* puede ser debido a los aleros de los edificios, ya que si se consideran éste disminuye considerablemente. Pero no se conocen los motivos que dan lugar a esos *EO*, por lo que se considera oportuno realizar un **estudio** sobre la distribución del **Error de Omisión**.

Tabla 5.15. Análisis del error de omisión en la categoría de edificaciones GV (5094793_c)

HOJA 61 (GV)	*EO* (%)	% clases 10, 11, 13	% clase 2	% clase 3,4,5	% clase 1	% clase 0	% clase 7	TOTAL
5094793_c_clean.las	*12,09*	**51,50**	16,70	3,67	26,33	1,57	0,23	100
flithlines.00089.las	*9,50*	**61,40**	7,37	0,86	29,75	0,52	0,10	100
flithlines.00090.las	*10,27*	**51,61**	15,76	2,80	28,43	1,19	0,21	100
flithlines.00091.las	*12,47*	**49,75**	17,83	3,80	26,70	1,56	0,36	100
flithlines.00092.las	*15,81*	**51,62**	18,90	5,22	22,24	1,94	0,08	100
flithlines.00093.las	*68,37*	9,61	**32,03**	21,71	3,20	33,45	--	100

Tabla 5.16. Análisis del error de omisión en la categoría de edificaciones DFG (5694788_c_Gipuzkoa)

HOJA 64 (DFG)	*EO* (%)	% clase 0	% clase 2	% clase 3,4,5	% clase 7	TOTAL
5694788_c_Gipuzkoa_clean.las	*10,60*	**53,58**	37,06	9,13	0,23	100
flightlines.00018.las	*9,49*	**49,97**	37,37	12,30	0,36	100
flightlines.00019.las	*11,16*	**52,14**	39,65	8,10	0,11	100
flightlines.00020.las	*28,97*	**61,96**	38,04	--	--	100
flightlines.00048.las	*10,12*	**54,12**	36,32	9,34	0,22	100
flightlines.00049.las	*9,54*	**53,68**	36,16	9,82	0,34	100

En este caso, se han incluido las pasadas que se habían desestimado con el objetivo de poder estudiar su comportamiento. Así, se puede apreciar cómo las pasadas 00093 (GV) y 00020 (DFG) están formadas por una distribución de puntos completamente diferente a las del resto que constituyen el fichero completo.

Analizados los datos se puede afirmar que la cuadrícula del **vuelo de la DFG** sigue el mismo comportamiento que en la de la hoja 63, constituyendo la mayor fuente de *EO* los puntos de clase 0 (53,58 %), seguidos de los de la clase 2 (37,06 %). La vegetación tiene una muy pequeña influencia y el ruido prácticamente es inexistente.

En el caso del **vuelo del GV**, señalar que aunque la mayor indeterminación viene dada por las clases reservadas para la ASPRS (10, 11, 13) irrumpen las clases 1 y 0 que antes no existían, siendo el porcentaje de la clase 1 (26,33 %) superior al de la clase 2 (16,70 %). La vegetación

influye prácticamente en los mismos términos y la clase 0 tiene muy poca influencia aunque mayor que la clase 7. La clase 12 no se ha incluido porque sus valores no son significativos.

Analizando los resultados conjunto de las fiabilidades se puede concluir que en este caso la cuadrícula del polígono industrial (GV) está algo mejor clasificada ya que el valor que alcanza un *f1-score* de 0,92 frente a la de Aia (DFG) con 0,81.

5.3.4. ÁNALISIS DE LA CLASIFICACIÓN DE LAS CARRETERAS

Para el estudio de las carreteras se ha visto necesario prescindir del *EC*, dado que hay muchos puntos de suelo que no son carretera y se presupone que todos los puntos con clase 2 ubicados en la zona de carreteras (intersección BTA) deben serlo, no entendiendo que en esas zonas puedan aparecer puntos LiDAR con otro valor. Partiendo de esa premisa es de suponer que la *FP* venga dada por el porcentaje de puntos pertenecientes a la clase 2, y en efecto resulta así.

Al igual que en el caso de las edificaciones, el análisis también se va a presentar por un lado para las cuadrículas de la hoja 63 y luego conjuntamente para las de las 61 y 64.

5.3.4.1. Análisis de carreteras en la hoja 63

Según lo expresado anteriormente, el porcentaje de puntos pertenecientes a la clase 2 en las carreteras ha resultado superior en el caso del vuelo del GV, lo que conlleva a un *EO* menor, obteniendo un valor de 0,25 en la hoja del GV y de 0,44 para la de la DFG.

Tabla 5.17. Error de omisión y porcentajes por clases de la categoría de carreteras GV (5404782_c)

HOJA 63 (GV)	*EO* (%)	*FP* (% clase 2)	% Clase 12	% Clase 3, 4,5	*EO* - % Clase 12
5404782_c_clean.las	25,46	74,54	12,02	12,04	13,44
flithlines.00032.las	25,41	74,59	13,36	10,23	12,06
flithlines.00054.las	40,39	59,61	25,75	12,51	14,64
flithlines.00095.las	27,73	72,27	7,58	19,43	20,16
flithlines.00103.las	23,93	76,07	10,25	12,77	13,69

Tabla 5.18. Error de omisión y porcentajes por clases de la categoría de carreteras DFG 5404782_c_Gipuzkoa

HOJA 63 (DFG)	*EO* (%)	*FP* (% clase 2)	% Clase 0	% Clase 3, 4, 5	*EO* - % Clase 0
5404782_c_Gipuzkoa_clean.las	44,13	55,87	29,39	14,15	14,74
flithlines.00014.las	60,99	39,01	48,23	12,76	12,76
flithlines.00015.las	39,01	60,99	25,37	12,89	13,64
flithlines.00016.las	41,45	58,55	29,26	11,50	12,19
flithlines.00017.las	58,96	41,04	31,03	27,84	27,92

Con el objeto de estudiar la tendencia de la omisión se han determinado los porcentajes de carretera que pertenecen a la clase 12 en el caso del GV y a la clase 0 en el caso de la DFG, comprobando que además de esas clases también les influye alguna otra (tablas 5.17 y 5.18).

Esa clase no es más que la vegetación, ya que se puede comprobar cómo la diferencia entre el *EO* y la clase 0 ó 12 coincide con el porcentaje correspondiente a las clases 3, 4 y 5. Del examen gráfico se observa que en el caso del GV se trata de puntos de las clases 3 y 5 principalmente, mientras que en el de la DFG de la clase 5. Además, aunque a priori parece que los puntos de vegetación que se dan en el vuelo del GV (12,04 %) son más que los del de la DFG (14,15 %), los números indican que esos porcentajes son levemente mayores en el caso de la DFG.

Respecto al resto de clases, señalar que en el del GV las clases 6, 7, 10, 11, 13 y 1 de manera independiente constituyen un porcentaje inferior al uno por ciento en casi todos los casos, por lo que no se han especificado sus valores considerando que las diferencias sobre el cien por cien sean debidas a ellas. Por el mismo motivo para el vuelo de la DFG no se han tenido en cuenta ni la clase 6 ni la 7.

Es de destacar que en el caso del vuelo de la DFG la omisión debida a la clase 0 ha supuesto un porcentaje muy alto, superior al 25 % en todos los casos, constituyendo más de la mitad de la omisión. En la figura 5.16 se puede visualizar la omisión debida a la clase 0 de la DFG en verde y en azul la correspondiente a la clase 12 del GV, con el fondo de la carretera en amarillo.

Figura 5.16. Detalle de la omisión en carreteras debida a las clases 12 (GV) y 0 (DFG)

En ambos casos se puede decir que el comportamiento de las distintas pasadas no es homogéneo, destacando algunas sobre el resto. Si bien, si se cogen aquéllas que abarcan la inmensa totalidad del área de estudio los valores obtenidos resultan más uniformes, lo que conlleva a que en el caso del fichero de la DFG sea conveniente despreciar las pasadas 00014 y 00017 y en el del GV las 00054 y 00095.

5.3.4.2. Análisis de carreteras en las hojas 61 y 64

El comportamiento de las cuadrículas analizadas dentro de estas hojas es muy similar al del la hoja anterior, obteniendo también mejores resultados en el vuelo del GV.

Tabla 5.19. Error de omisión y porcentajes por clases de la categoría de carreteras GV (5094793_c)

HOJA 61 (GV)	*EO* (%)	*FP* (% clase 2)	% Clase 10, 11	% Clase 3, 4, 5	% Clase 1	% Clase 0
5094793_c_clean.las	26,76	73,24	13,90	6,23	5,56	1,00
flightlines.00089.las	23,40	76,60	9,16	7,04	6,11	1,08
flightlines.00090.las	22,80	77,20	13,16	4,93	3,97	0,69
flightlines.00091.las	30,13	69,87	14,03	7,48	7,04	1,47
flightlines.00092.las	34,90	65,10	24,33	5,01	5,02	0,44
flightlines.00093.las	8,43	91,57	--	8,43	--	--

Tabla 5.20. Error de omisión y porcentajes por clases de la categoría de carreteras DFG (5694788_c_Gipuzkoa)

HOJA 64 (DFG)	*EO* (%)	*FP* (% clase 2)	%Clase 0	% Clase 3, 4, 5
5694788_c_Gipuzkoa_clean.las	44,90	55,10	32,76	11,51
flightlines.00018.las	44,27	55,73	30,14	13,54
flightlines.00019.las	46,62	53,38	36,14	10,00
flightlines.00020.las	55,38	44,62	41,50	12,50
flightlines.00048.las	43,55	56,45	29,33	13,39
flightlines.00049.las	42,66	57,34	32,94	9,22

Una de las diferencias que se aprecia con respecto a la cuadrícula del vuelo del GV de la hoja 63 es que en la zona 5094793 la omisión en lugar de pertenecer a la clase 12 se ubica por orden en las clases 10-11, vegetación y clase 1. La proporción de la vegetación y clase 1 resultan muy similares y es prácticamente nula la de las clases 12, 6, 7, apareciendo algo más en la clase 0.

La pasada 00093 ha ofrecido una respuesta completamente distinta al resto de pasadas, siendo la única que adquiere una cuantía baja para el *EO* (0,08) y en consecuencia una *FP* alta (0,92). En ella se observa que toda la omisión es debida exclusivamente a la vegetación. Sin embargo, al visualizarla se puede comprobar que no es representativa de la zona de estudio, por lo que es de presuponer que el resto de pasadas ofrecen unos valores más acordes con la realidad.

En el caso del vuelo de la DFG, al igual que con los edificios, para las carreteras la pasada 00017 tampoco ha ofrecido resultados. En el resto de las pasadas se puede apreciar como el *EO* adquiere un valor próximo al 0,46, resultando considerablemente superior en el caso de la pasada 00020 (0,55), tal y como sucedía en el caso de los edificios.

Si se comparan los valores obtenidos para el *EO* y la *FP* se comprueba cómo el comportamiento de la pasada 00020 es opuesta al resto: mientras en las otras el EO es entorno al 0,44 en la 00020 es del 0,55, valor que adquieren el resto de pasadas para la *FP*.

Por otro lado, al comprobar los resultados del fichero completo se puede apreciar que la tendencia es la del resto de las pasadas, obteniendo pequeñas variaciones para las pasadas 00019, 00049 y 00048. Esto puede llevar a pensar que realmente la predisposición de la pasada 00020 no afecta al conjunto de los datos testeados. En cualquier caso, cabe señalar que esta pasada es la que contiene mayor porcentaje de puntos con clase 0 para la identificación de las carreteras.

5.3.5. ÁNALISIS DE LA CLASIFICACIÓN DE LA VEGETACIÓN

Tal y como se ha indicado en la introducción de este epígrafe y atendiendo a los elementos indicados en la tabla 5.1, para comprobar la bondad de la clasificación de los puntos LiDAR en lo que respecta a la vegetación se han contemplado el 0122, el 0130 y 0998 por entender que los correspondientes al 0128 y 0123 en la nube de puntos vendrían contemplados como puntos de suelo (clase 2). Tampoco se ha considerado la categoría denominada Artificializado (ID_TIPO = 0999) porque hace referencia a los entornos de edificaciones (cascos urbanos, caseríos y alrededores, etc.) que abarca distintas clases de la clasificación ASPRS.

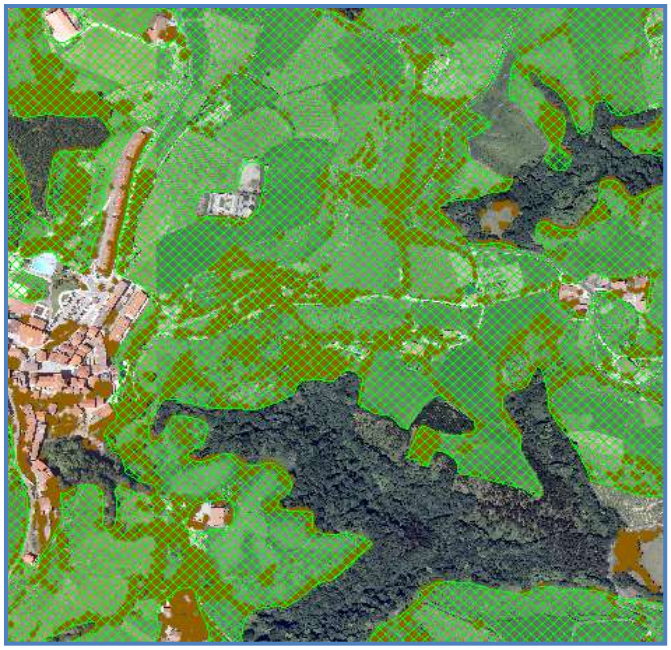

Figura 5.17. Situación planteada en la cubierta vegetal

Ante esta situación, el primer problema que se plantea es el que se muestra en la figura 5.17, donde los árboles contemplados dentro de las zonas identificadas como prado (patrón verde)

en la BTA no aparecen identificados como tales, creando esto un error de comisión mayor del realmente existente, ya que en la nube de puntos dentro de las zonas identificadas como prado todos aquellos arbustos existentes se muestran como algunas de las clases de vegetación (manchas marrones), prioritariamente clase 5. Además, también se puede apreciar como los árboles (manchas marrones) ubicados en los núcleos urbanos, reconocidos como algún tipo de vegetación por la clasificación LiDAR, tampoco son contemplados por quedar encubiertos dentro de la categoría de Artificializado. Hecho, que al igual que el anterior, también contribuye a desvirtuar el verdadero valor del error de comisión.

5.3.5.1. Análisis de la cubierta vegetal en la hoja 63

Al examinar las cuadrículas ubicadas dentro de esta hoja, a partir de la tabla 5.21 se puede comprobar que tanto el *EO* como el *EC* resultan mayores en la muestra del GV (*EO* = 0,71; *EC* = 0,29) que en la de la DFG (*EO* = 0,53; *EC* = 0,14), lo que conlleva a unas fiabilidades peores para ese fichero, pero en cualquier caso con unas *FP* (GV: 0,29; DFG: 0,47) muy bajas para ambos, situación que queda patente con los valores de *f1-score*, también bajos siendo el mayor el de la DFG con 0,61.

Tabla 5.21. Errores de omisión y comisión en la categoría de vegetación (5404782)

HOJA 63 (GV)	*EO*	*FP*	*EC*	*FU*	*f1-score*
5404782_c_clean.las	0,71	0,29	0,29	0,71	0,41
flithlines.00032.las	0,75	0,25	0,31	0,69	
flithlines.00054.las	0,66	0,34	0,08	0,92	
flithlines.00095.las	0,59	0,41	0,29	0,71	
flithlines.00103.las	0,72	0,28	0,34	0,66	
HOJA 63 (DFG)	***EO***	***FP***	***EC***	***FU***	***f1-score***
5404782_c_Gipuzkoa_clean.las	0,53	0,47	0,14	0,86	0,61
flightlines.00014.las	0,44	0,56	0,12	0,88	
flightlines.00015.las	0,55	0,45	0,15	0,85	
flightlines.00016.las	0,56	0,44	0,15	0,85	
flightlines.00017.las	0,45	0,55	0,10	0,90	

En lo que respecta al error de comisión, a pesar de verse afectado por las categorías de artificializado y prado que contempla la BTA, en la cuadrícula de la hoja de la DFG se consiguen unos valores bastantes aceptables, rondando una *FU* próxima al 0,85. Si bien es verdad que las cuantías del GV son un poco más bajas, aproximándose a una media de *FU* del 0,75. En este caso, también destacan los valores alcanzados por la pasada 00054 con una *FU* de 0,92.

De la visualización gráfica (figura 5.18) se desprende que gran parte del *EC* (en magenta) es debido al artificializado (línea azul), el cual está constituido en el vuelo de la DFG por puntos de clase 5, mientras que en el del GV por la clase 3. Además, en este último vuelo hay más puntos tanto en el casco urbano como en las zonas de matorral, lo que conlleva a unos errores de mayor cuantía.

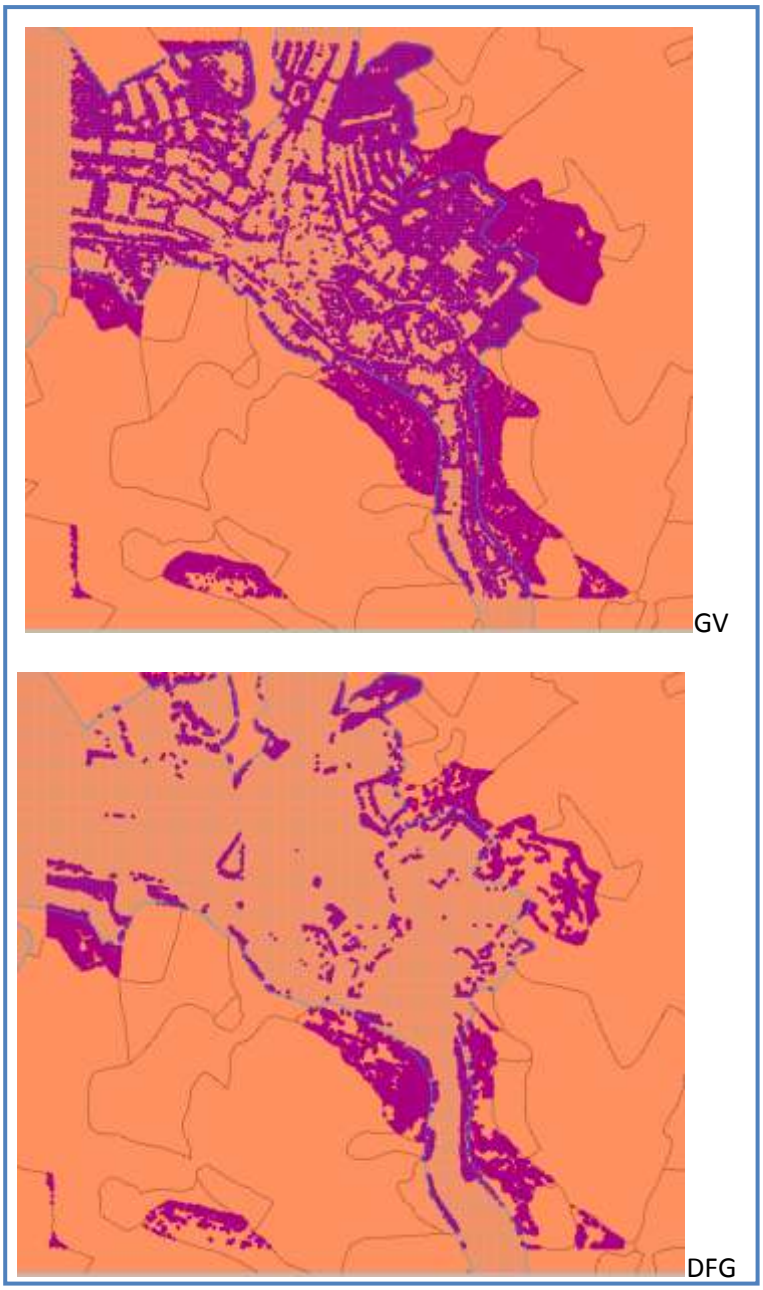

Figura 5.18. Error de comisión en la categoría de vegetación (5404782)

En la cuadrícula del GV los valores del error de omisión adquieren valores muy altos, alcanzando cuantías próximas a 0,70, lo que lleva a unas *FP* muy bajas, en torno a un 0,30. Por su parte, en la hoja de la DFG el *EO* y la *FP* adquieren valores muy similares, próximo al 0,50. Estos valores dan a entender la poca fiabilidad del análisis realizado que en parte está condicionado por la situación que plantea la configuración de los datos utilizados.

Dado que la omisión es muy grande, al igual que en los casos anteriores, se ha tratado de averiguar a qué otras categorías se deben estos resultados. Para ello, en primer lugar se han considerado las clase 12 para el vuelo del GV y la clase 0 para el de la DFG, pero se ha podido comprobar que esta omisión está afectada por otra clase, concretamente la clase 2.

Tabla 5.22. Error de omisión y porcentajes por clases de la categoría de vegetación (5404782)

HOJA 63 (GV)	*EO* (%)	% clase 12	% clase 2
5404782_c_clean.las	*71,07*	24,41	45,46
flithlines.00032.las	*74,84*	26,45	46,72
flithlines.00054.las	*66,16*	23,16	42,56
flithlines.00095.las	*59,21*	18,86	39,93
flithlines.00103.las	*72,02*	24,02	46,85
HOJA 63 (DFG)	***EO* (%)**	**% clase 0**	**% clase 2**
5404782_c_Gipuzkoa_clean.las	*52,55*	25,51	26,80
flightlines.00014.las	*43,83*	24,07	19,75
flightlines.00015.las	*54,89*	23,15	31,45
flightlines.00016.las	*56,37*	27,83	28,21
flightlines.00017.las	*44,85*	26,92	17,84

En ambos casos, el porcentaje de puntos que pertenecen a la clase 12 ó 0 resultan bastante parecidos, con valores inferiores al 28 %. Sin embargo, la clase 2 alcanza mayores valores en el caso del vuelo del GV, duplicando tanto el porcentaje de los de la clase 12 como los de la clase 2 del vuelo de la DFG.

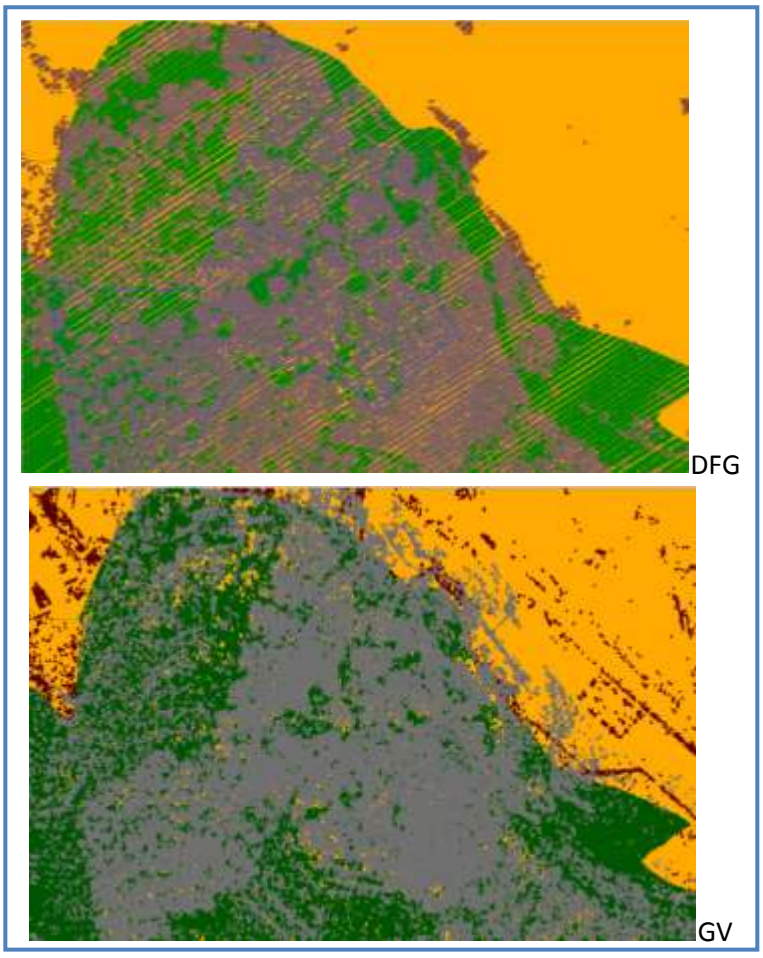

Figura 5.19. Errores de omisión y comisión en la categoría devegetación

En la figura 5.19 se muestra en verde el *EO*, en marrón el *EC* y en gris los puntos clasificados correctamente como vegetación, donde se puede apreciar que la mayoría de los puntos que entran dentro del recinto de vegetación marcado por la BTA se corresponden con la omisión por tratarse de puntos de clase 2 (suelo) ó 0 /12 según los vuelos.

Por su parte, la omisión que se produce en los dos vuelos debido a la clase 6 (edificaciones) es mínima, resultando inferior al 1 % en todas las pasadas. En el vuelo del GV, también aparece algo de omisión debido a las clases 1, 10, 11 y 13 pero en todos los casos resulta inferior al 1 %.

Como conclusión se puede decir que el error de omisión está claramente condicionado con los puntos que perteneciendo a la clase 0 ó 12 y clase 2 deberían aparecer como vegetación. Además, para poder estudiar bien el estado de los puntos catalogados como vegetación, dentro de la BTA se deberían contemplar los árboles independientemente de los núcleos urbanos (artificializado) y prados sin incluirlos en un todo. Estas acciones contribuirían a disminuir considerablemente tanto el error de omisión como el de comisión.

5.3.5.2. Análisis de la cubierta vegetal en las hojas 61 y 64

A continuación se muestran a modo de tabla los errores de comisión y omisión junto con sus fiabilidades de las cuadrículas de las hojas de estudio de esta sección.

Tabla 5.23. Errores de omisión y comisión en la categoría de vegetación: GV (5094793_c) y DFG (5694788_c_Gipuzkoa)

HOJA 61 (GV)	*EO*	*FP*	*EC*	*FU*	*f1-score*
5094793_c_clean.las	0,49	0,51	0,51	0,49	0,50
flithlines.00089.las	0,44	0,56	0,22	0,78	
flithlines.00090.las	0,49	0,51	0,45	0,55	
flithlines.00091.las	0,59	0,41	0,76	0,24	
flithlines.00092.las	0,56	0,44	0,87	0,13	
HOJA 64 (DFG)	***EO***	***FP***	***EC***	***FU***	***f1-score***
5694788_c_Gipuzkoa_clean.las	0,40	0,60	0,31	0,69	0,64
flightlines.00018.las	0,37	0,63	0,36	0,64	
flightlines.00019.las	0,44	0,56	0,35	0,65	
flightlines.00020.las	0,31	0,69	0,14	0,86	
flightlines.00048.las	0,37	0,63	0,29	0,71	
flightlines.00049.las	0,50	0,50	0,33	0,67	

En el caso de la muestra de la hoja 61 se ha optado por desestimar las pasadas 00088 y 00093 al comprobar que carecen de datos para valorar estos errores. En los dos vuelos se pueden comprobar unas fiabilidades bajas con errores altos, siendo la muestra del vuelo del GV la que ofrece peores resultados con fiabilidades y errores entorno a 0,50, siendo ese su *f1-score* mientras que en la hoja de la DFG adquiere un valor superior con 0,64.

El error de comisión se ve altamente influenciado por las categorías de artificializado, vegetación y arbolado urbano y prado de la BTA. En el caso de zonas urbanas principalmente por el artificializado tal y como se muestra en lila en la figura 5.20. En esta imagen en marrón se muestra el *EO*.

Figura 5.20. Error de comisión y omisión debido al artificializado

En la figura 5.21 se plantea la situación que se da si lo que se considera es vegetación y arbolado urbano. En este caso, el *EC* se da por fuera o en los bordes del recinto (magenta), entendiendo que todo el recinto debería aparecer en el LiDAR como vegetación y no sólo los árboles que se ubican dentro del recinto (cian). En esta imagen en marrón se muestra el *EO*.

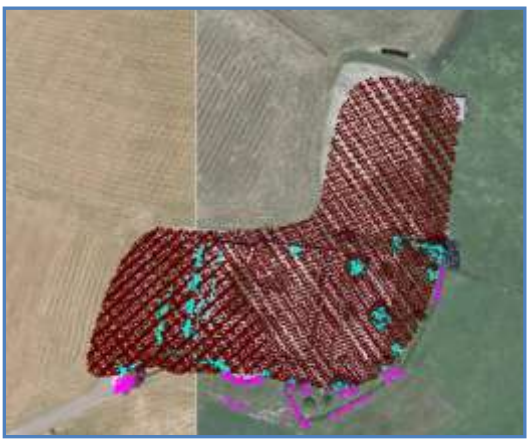

Figura 5.21. Errores en recintos de vegetación y arbolado urbano

En la figura 5.22 se muestra la situación que se da en el arbolado forestal donde los puntos LiDAR que hacen referencia a esas zonas se mezclan entre clase 2 (suelo) y clase 5 (vegetación alta) si hay arbolado, y clase 2 ó 0 si éste no existe. Además, la comisión (magenta) se da por los bordes del arbolado forestal.

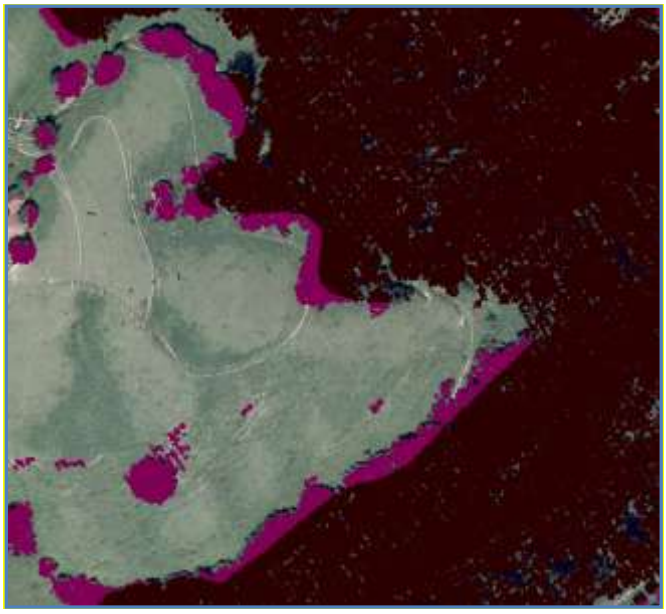

Figura 5.22. Error de comisión en el arbolado forestal y prado

Al analizar el error de omisión se puede comprobar que los valores son un poco más bajos que los estudiados en la hoja 63, pero que siguen siendo altos, dando lugar a unas *FP* bajas con valores entre 0,41 y 0,69.

En la figura 5.23 se puede comprobar cómo en gran parte el *EO* (marrón) convive con puntos bien clasificados (cian) catalogados en la BTA dentro del arbolado forestal. Pero el *EO* resulta prácticamente total en la zona superior, sin árboles en la ortofoto y catalogado en la BTA como arbolado forestal.

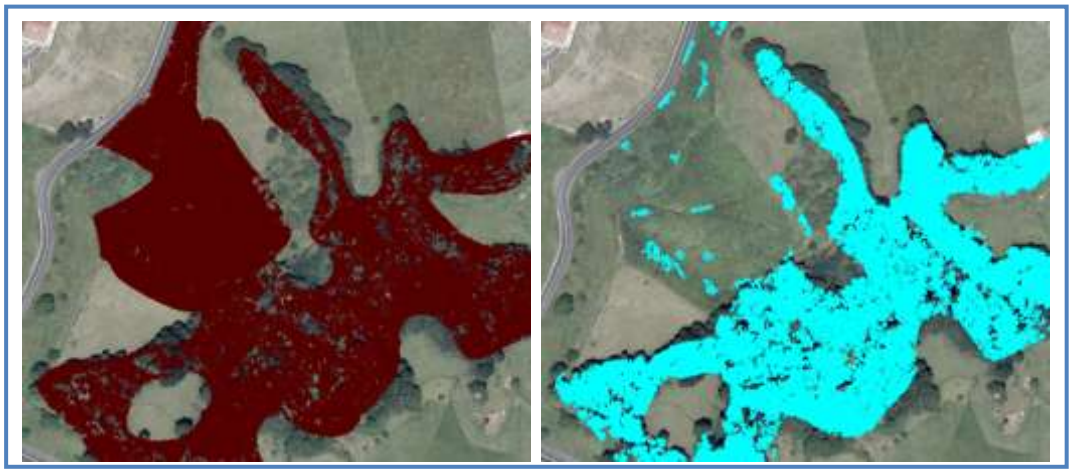

Figura 5.23. Error de omisión en zona de arbolado forestal

Con el fin de comprender los motivos de esta omisión se ha procedido a estudiar el comportamiento de las otras clases, comprobando que mientras en el vuelo del GV el mayor porcentaje recae en la clase 2 alcanzando un valor medio entorno al 41 %, en el de la DFG se ubica en la clase 0 con un porcentaje próximo al 21 % de media, resultando prácticamente nula en el caso de la categoría de edificios (clase 6) y siendo nula para el resto. Patrón que también se daba en el caso de la hoja 63.

Tabla 5.24. Error de omisión y porcentajes por clases de la categoría de vegetación: GV (5094793_c) y DFG (5694788_c_Gipuzkoa)

HOJA 61 (GV)	*EO* (%)	% clase 1	% clase 2
5094793_c_clean.las	*49,11*	9,77	39,05
flithlines.00089.las	*43,83*	8,50	35,04
flithlines.00090.las	*48,60*	9,32	39,03
flithlines.00091.las	*59,00*	12,17	46,57
flithlines.00092.las	*56,44*	9,84	46,50
HOJA 64 (DFG)	***EO* (%)**	**% clase0**	**% clase 2**
5694788_c_Gipuzkoa_clean.las	*39,92*	21,51	18,21
flightlines.00018.las	*36,54*	20,14	16,22
flightlines.00019.las	*43,97*	14,10	29,56
flightlines.00020.las	*31,21*	19,18	12,01
flightlines.00048.las	*36,96*	23,51	13,25
flightlines.00049.las	*49,61*	28,26	21,12

Un aspecto a reseñar con respecto al análisis de la hoja 63 es que en este caso en el vuelo del GV en lugar de aparecer puntos en la clase 12, se dan en la clase 1 y con un porcentaje bastante inferior al del caso anterior.

5.4. CONCLUSIONES

Una vez terminado el estudio de las cuadrículas elegidas se presenta un resumen que pretende esclarecer la situación marcada por la información de partida.

En primer lugar, hay que constatar que se ha comprobado que aunque los datos se refieran a la misma zona y estén procesados por el mismo software los resultados no tienen por qué resultar significativamente parecidos. En este ámbito cabe señalar que si se trata de vuelos realizados en distintas épocas del año, al menos en cuanto a la vegetación se refiere, el producto puede resultar sensiblemente diferente.

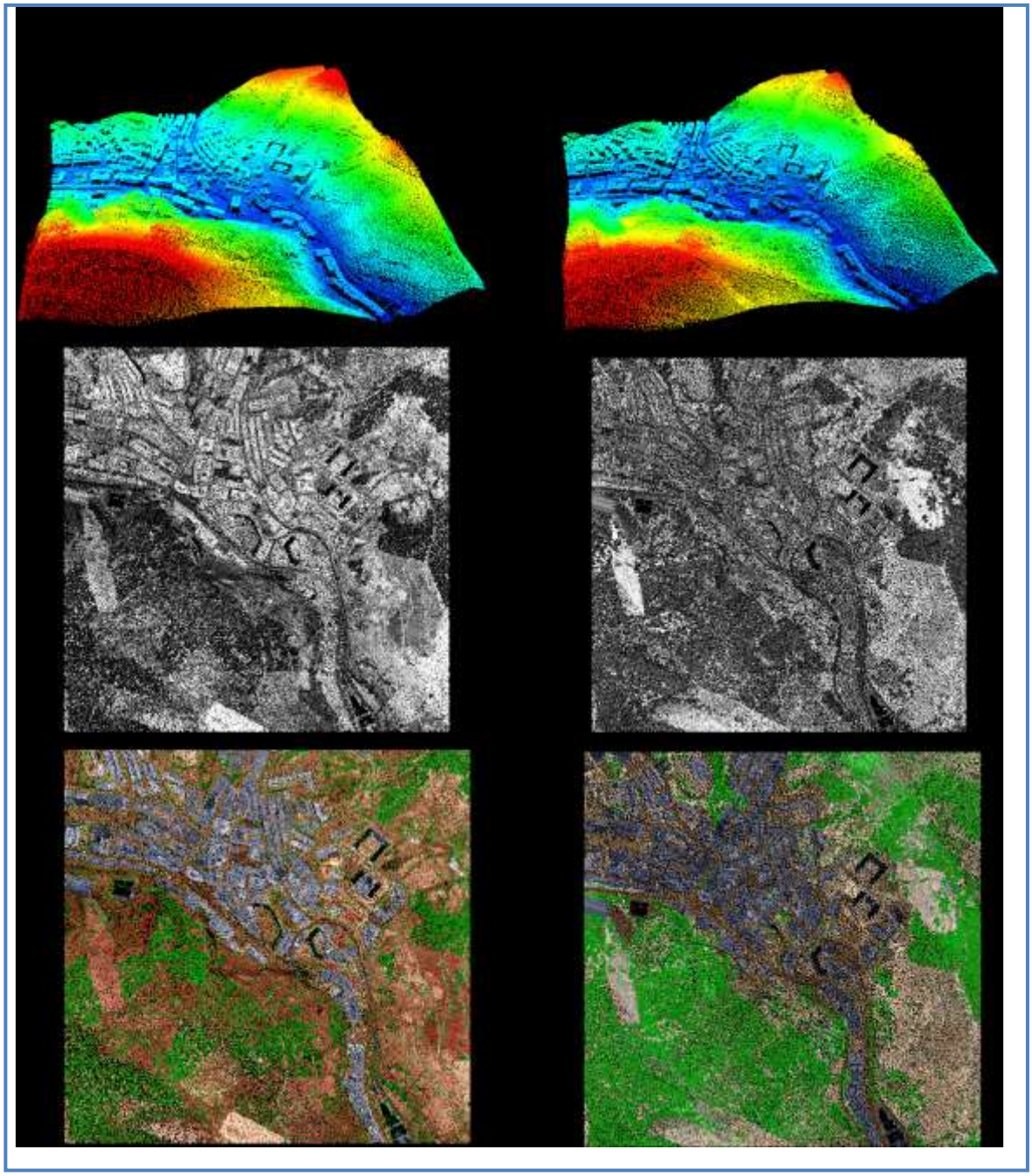

Figura 5.24. Comparativa por elevación, intensidad y clasificación de vuelos distintos

En el caso de la hoja que aquí se ha analizado, hay que señalar que existe una diferencia neta en porcentaje de puntos clasificados en las mismas categorías entorno al 20 %, pero que no se puede justificar por carecer de información sobre los ajustes que se han podido realizar en uno y otro caso, ni de la revisión manual que se ha podido efectuar, si se ha hecho.

Además, el sensor LiDAR realiza la captura de datos a modo de barrido y según los distintos tipos existentes en el mercado describen distintas clases de patrones durante la exposición, por lo que resulta prácticamente imposible recoger exactamente los mismos puntos, aun cuando el vuelo se realice con el mismo sensor. Si bien, la información tanto cualitativa como cuantitativa debe resultar muy similar.

En las visualizaciones de este tipo de datos, al analizar la misma zona realizada con dos vuelos distintos, aunque a simple vista el conjunto de puntos muestre unos resultados similares, al entrar en profundidad a comprobar el comportamiento de los mismos se aprecian diferencias significativas, no tanto en altitud sino más bien al visualizar valores radiométricos o de clasificación, tal y como se puede comprobar en la figura 5.24.

Las discrepancias no sólo se deben a las características geométricas que pueden derivarse del tipo de escáner usado o de la planificación del vuelo, resultando diferente según la dirección de las pasadas consideradas y del recubrimiento del vuelo. Para una misma densidad de puntos, los resultados serán potencialmente distintos si se consigue con distinto recubrimiento o variando la altura de vuelo, sobre todo en el entorno de edificaciones. Además, la dirección de cada pasada influye directamente en los valores de intensidad captados dada la reflectancias de la cubierta, supeditada por la fecha y hora de vuelo y las condiciones meteorológicas, entre otras (Renslow 2012).

Independientemente del sensor utilizado y la geometría del vuelo se ha podido constatar que aunque los vuelos se realicen en distintas épocas y bajo distintas condiciones aquellos objetos que por su naturaleza no reenvían la señal no lo hacen en ningún caso, apareciendo zonas sin datos tales como las áreas negras de la figura 5.25.

En cuanto a los valores de clasificación, prácticamente están condicionados por el algoritmo utilizado en el software de procesado, pero en la mayoría suele tener gran importancia la altitud de los puntos considerados y los de su alrededor, siendo un elemento muy característico las diferencias de alturas entre puntos.

En este caso, a pesar de disponer del mismo software, luego se supone que el algoritmo utilizado ha sido el mismo, se ha podido comprobar la existencia de ciertas discrepancias en los valores de clasificación, lo que hace pensar que esto es debido a los parámetros utilizados a la hora de configurar el algoritmo de clasificación. En consecuencia, la variedad de parámetros que admitan los algoritmos ofrecen resultados diversos.

Como ya se ha comentado, este tipo de vuelos se realiza siguiendo unas pasadas o líneas de vuelo y en este experimento se ha comprobado que la consideración de las pasadas originales, cuando la información se encuentra recortada por cuadrículas, no supone ninguna aportación

relevante. Es más, en el caso de pasadas con pocos puntos puede llevar a interpretaciones erróneas dada la falta de puntos de ciertas clases, aunque realmente existan en esa zona.

En consecuencia, se considera más adecuado procesar los ficheros completos, obviando la información aportada por cada una de las pasadas que los componen de manera independiente. Si bien, se puede decir que en las líneas de vuelo con una distribución de puntos más o menos homogénea en todas las categorías consideradas, la tendencia mostrada por las pasadas se trasmite al fichero original.

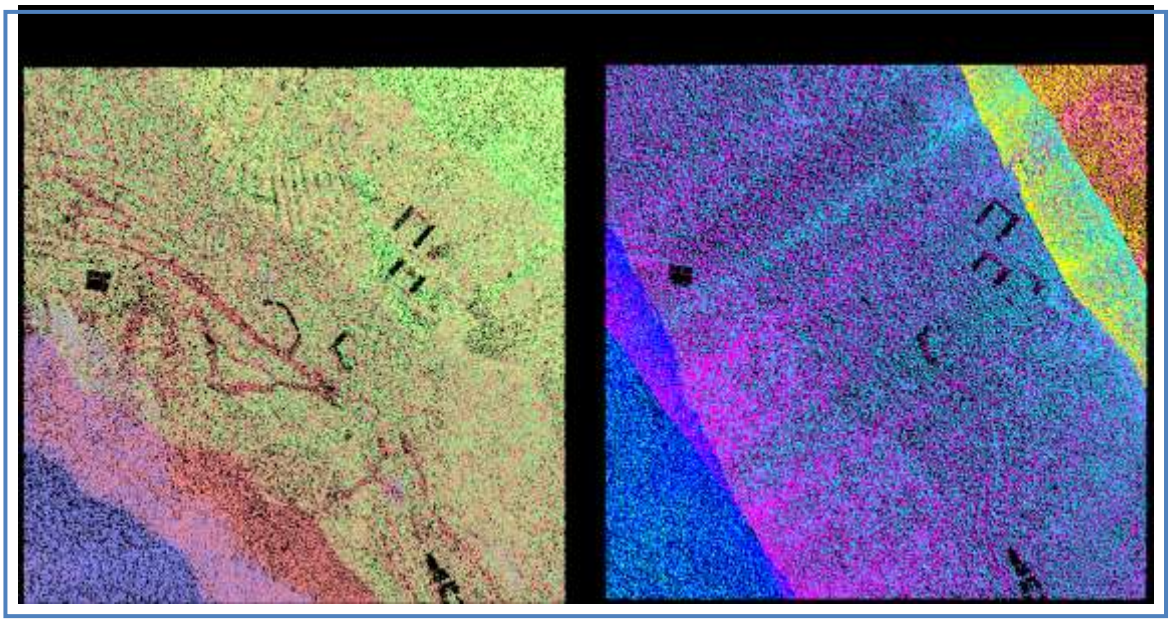

Figura 5.25. Distribución de puntos por pasadas

En cuanto a los vuelos aquí tratados, comentar que en los dos se dan unas densidades y espaciados entre puntos similares, pero en el vuelo del GV se aprecia de manera generalizada un mayor número de puntos que en el de la DFG, lo que lleva en algunos casos a disminuir los errores estudiados. Al examinar la clasificación aportada por ambos vuelos, el de la DFG presenta unos valores más homogéneos en cuanto a las clases de clasificación utilizadas que el de el GV, lo cual puede estar justificado porque este último se realizó en diversas fases, detectando variación de clases entre la fase inicial y la final (Azimut 2008). Por su parte, en el vuelo del GV existe un porcentaje de puntos menor sin asignar a clases concretas (puntos de clase 0, 1 y reservados para la definición de la ASPRS) que en el caso de la DFG (puntos de clase 0 y reservados para la definición de la ASPRS).

Para cotejar la clasificación automática aportada por los datos LiDAR se ha optado por considerar las categorías de la BTA de edificaciones que en el LiDAR serían clase 6, carreteras que vendría identificado por la clase 2 y algunos fenómenos de la cubierta vegetal identificados con puntos de clase 3, 4 y 5, sin hacer distinción entre ellos.

Con el fin de valorar los aciertos y desaciertos encontrados se ha optado por calcular los errores de omisión y comisión junto con sus fiabilidades, que se están utilizando cada vez más en la valoración de datos LiDAR

Dentro de las categorías analizadas los peores resultados se dan en la cubierta vegetal (*f1-score* = 0,41 en el peor de los casos), tanto para el *EC* como para el *EO*, pero es preciso advertir de la elevada incertidumbre que presentan la cobertura de cubierta vegetal de la BTA, porque dentro de esa cobertura no sólo se encuentran zonas con vegetación alta, media y baja sino que también están los prados, parcelas, diseminados alrededor de zonas urbanas, etc. elementos que constituyen la cubierta terrestre pero que no hacen referencia a zonas arboladas.

Los mejores resultados los brindan la categoría de edificaciones (*f1-score* = 0,80 en el peor de los casos) ya que se trata de la cobertura con menor indefinición tantos para los datos LiDAR como para la BTA. Por su parte, en el caso de las carreteras, no se ha podido llegar a determinar el *EC* dada la gran variedad de categorías de la BTA contempladas dentro de la clase 2 de los datos LiDAR. En la tabla 5.25 se presenta un resumen de los resultados alcanzados en las cuadrículas de referencia, pudiendo decir que en general los valores de estas fiabilidades así como la cuantía de *f1-score* dependen de la cuadrícula considerada, aunque las tendencias se mantienen por categorías y en menor medida por los vuelos considerados.

Tabla 5.25. Resumen de los errores de omisión y comisión y valor de Kappa por categorías

HOJA	VUELO	Fichero .LAS	Categorías	FP	FU	*f1-score*
61	GV	5094793_c_clean.las	Vegetación	0,51	0,49	0,50
63	GV	5404782_c_clean.las	Vegetación	0,29	0,71	0,41
63	DFG	5404782_c_Gipuzkoa_clean.las	Vegetación	0,47	0,86	0,61
64	DFG	5694788_c_Gipuzkoa_clean.las	Vegetación	0,60	0,69	0,64
61	GV	5094793_c_clean.las	Carreteras	0,73		
63	GV	5404782_c_clean.las	Carreteras	0,75		
63	DFG	5404782_c_Gipuzkoa_clean.las	Carreteras	0,56		
64	DFG	5694788_c_Gipuzkoa_clean.las	Carreteras	0,55		
61	GV	5094793_c_clean.las	Edificación	0,88	0,97	0,92
63	GV	5404782_c_clean.las	Edificación	0,74	0,88	0,80
63	DFG	5404782_c_Gipuzkoa_clean.las	Edificación	0,91	0,86	0,88
64	DFG	5694788_c_Gipuzkoa_clean.las	Edificación	0,89	0,74	0,81

Comentar también que los resultados no están condicionados por la tipología ni del terreno ni del desarrollo urbano considerado, ya que independientemente de la zona (rural, urbana, e industrial) las cuantías de los errores son bastante fluctuantes.

Para poder establecer conclusiones más firmes a este respecto hubiera sido preciso bien el análisis de un mayor número de cuadrículas LAS, bien disponer de datos de control con menos incertidumbre, pero indicar que al no ser este el objeto de la investigación, se considera suficiente las conclusiones expuestas para la siguiente fase, planteando que si se consigue mejorar la clasificación de los puntos de clase 0 (DFG) ó 12 /reservados para la definición de la ASPRS (GV) seguidos de la clase 2 que afectan a la omisión, se habrá podido contribuir positivamente en la identificación de un mayor número de elementos a partir de los datos LiDAR.

6. METODOLOGÍA PARA LA CLASIFICACIÓN DE DATOS LIDAR USANDO MINERÍA DE DATOS

La explotación práctica de los datos LiDAR presupone disponer de información añadida que permita la reorganización, selección y / o estructuración de esos datos acorde al objetivo del usuario. En consecuencia, resulta imprescindible contar con una clasificación para cada punto obtenido.

Se puede comprobar en la bibliografía que son muchas las referencias que hablan sobre dicha clasificación, pero la mayoría de ellas se basan en establecer si los puntos pertenecen o no al terreno (Jubanski 2010).

Contar con clasificaciones más detalladas de acuerdo a las distintas aplicaciones para las que estos datos resultan útiles, sin duda debería contribuir a un uso más sencillo y eficiente del ingente volumen que siempre caracteriza a este tipo de información espacial, facilitando la diversificación de aplicaciones más allá del ámbito cartográfico.

Establecer una metodología que permita de manera automática extraer este tipo de información lo más completa posible es el fin que se persigue en este capítulo.

6.1. INTRODUCCIÓN

El fin perseguido por este punto es explicar la metodología a seguir para clasificar de manera automática grandes volúmenes de datos LiDAR con características similares en cuanto a densidad, registro de ecos y valores de amplitud pero con variaciones topográficas y geográficas, tal y como señalan Sithole and Vosselman 2003.

Analizados en el estado del arte los métodos de clasificación de puntos LiDAR existentes y la teoría sobre minería de datos en lo que a algoritmos de clasificación basados en árboles se refiere, se propone una metodología que utilice este tipo de árboles para realizar la clasificación de los datos LiDAR.

Hay que tener en cuenta que la minería de datos necesita de una magnitud de datos considerable para a partir de su análisis poder predecir una serie de acciones a aplicar en otro conjunto de datos, obteniendo en estos nuevos datos y de manera automática los valores predichos que se les supone en función del modelo aplicado.

Es por ello por lo que en primer lugar se debe abordar la tarea de estudiar qué información va a contener ese volumen de datos y en función de ello determinar la manera de conseguirla. En esta investigación esta primera parte se explica en el apartado 6.2. La segunda parte abarca las tareas de construcción, validación y aplicación del modelo y se aborda en el punto 6.3.

6.2. PREPARACIÓN DE LA INFORMACIÓN PARA MINERÍA DE DATOS

Recordar, que tal y como se ha explicado en el punto 4 referente a los datos, se dispone de datos LiDAR, ortofotografías y cartografía vectorial (BTA) a escala 1:5.000, con los que se quiere clasificar puntos LiDAR, de acuerdo con muchas de las referencias bibliográficas (García-Gutiérrez, et al. 2010; Yan, et al. 2015) que en los últimos años, en el ámbito de la clasificación de imágenes, apuntan a la investigación en el desarrollo de algoritmos enfocados a la integración de datos procedentes de distintos sensores para mejorar las clasificaciones.

En este sentido, la investigación aquí presentada pretende constituir un pequeño aporte en ese entorno. Para ello, tras estudiar las variables a analizar se ha definido la transformación que es preciso efectuar a partir de los datos de partida para generar la BBDD de trabajo.

En esta BBDD, cada registro está asociado a cada punto del fichero LAS de manera que sea el punto el que aúne los parámetros derivados de las distintas fuentes de datos utilizadas, pudiendo resultar necesario el cálculo de algún otro atributo a partir de los datos originales.

Para la consecución de estos registros, como entorno de trabajo se ha elegido el interface de QGIS y el formato de esta BBDD ha sido el csv (*comma-separated values*), ya descritos en el punto 3 de esta memoria.

6.2.1. ESTUDIO DE VARIABLES A DETERMINAR

A priori se ha pensado utilizar únicamente los atributos facilitados por la base de datos propia de los archivos LAS (tabla 3.3.) tal y como se apunta en Alexander, et al. 2011, sustituyendo los valores R, G, B por el de los mismos canales de las ortofotografías del 2008, con el objeto de mejorar esta clasificación.

Sin embargo estos datos han resultado claramente insuficientes en las primeras pruebas realizadas, siendo necesario efectuar un estudio para poder establecer los nuevos atributos a determinar y el procedimiento a seguir. Cabe reseñar que Alexander, et al. 2011 utilizan LiDAR *full-waveform*.

A este respecto, Xu, et al. 2014 indican que a la hora de clasificar datos procedentes de ALS se plantean dos posibilidades: la primera se centra en encontrar suficientes atributos para distinguir clases de interés y la segunda trata de definir entidades propias para calcular atributos.

En esta investigación se ha tratado de unir ambas ideas para establecer las variables a utilizar optando por trabajar con el planteamiento del Minería de datos dado que tal y como apuntan Lu, et al. 2014 una de las ventajas que aportan estos algoritmos es la habilidad para reconocer patrones complejos para la toma de decisiones informadas.

Para ello, se ha aprovechado la información aportada por los ficheros LAS y las ortofotografías, de manera que a cada punto del archivo LAS se le ha asignado la información correspondiente de las otras fuentes de datos empleadas.

Inicialmente, y en línea con el planteamiento de los principales parámetros que ofrece el software TerraSolid 2011, se ha acordado buscar variables centradas principalmente en aspectos geométricos. Así, utilizando las posibilidades que ofrece FME, a partir de un modelo TIN, se ha tratado de obtener valores de referencia entre aquellos vértices que en la triangulación han constituido un lado de triángulo. De esta forma, se han calculado el desnivel, la distancia reducida, la distancia real y la pendiente del lado del triángulo, la cara del triángulo y la orientación de la pendiente según la cara del triángulo. A partir de estos valores, se han determinado los estadísticos más usuales: mínimo, máximo, media, desviación estándar, rango, mediana y moda.

La valoración de los resultados anteriores ha resultado totalmente inútil ya que esta proposición no ofrece un suelo común para todos los puntos, al haber calculado los valores promedios anteriores entre puntos próximos. En consecuencia, rápidamente se ha comprobado que resulta necesario una primera clasificación de los puntos entre TERRENO / NO TERRENO, tal y como se apunta en la variada literatura con fines distintos (Awrangjeb, et al. 2013; Liu, et al. 2013; Pérez-García, et al. 2012), para así contar con desniveles a partir del suelo. Lo que ha llevado a la generación de modelos digitales de elevación (MDE).

Por otro lado, entre las clases a diferenciar con cierta altura sobre el terreno suelen aparecer mayoritariamente las edificaciones y la vegetación, que en muchas ocasiones se entremezclan entre sí. Este hecho hace necesario el contar con algún parámetro que permita la distinción entre ambas. Para llevar a cabo esta diferenciación, en el ámbito de la Teledetección, resulta común el uso del índice de diferencias normalizadas (*Normalized Difference Vegetation Index*, NDVI), que en los últimos tiempos se está incorporando también al trabajo con datos LiDAR. Así Huang 2007 en su tesis ya hace uso del mismo al combinar los datos LiDAR con imágenes de satélite.

Además, con el objetivo de mejorar la clasificación de los datos LiDAR, se ha pretendido agrupar los puntos que tienen propiedades similares en regiones homogéneas, para lo que se ha hecho uso de la segmentación (Wang and Shan 2009). Para llevar a cabo esta idea, y aprovechar la información espectral aportada por las imágenes aéreas georreferenciadas, se han segmentado las imágenes disponibles (ortofotografías) y los modelos digitales, de manera que todos aquellos puntos pertenecientes a la misma unidad han sido considerados con el mismo comportamiento.

En el próximo apartado se explica el proceso finalmente seguido en cada caso para la consecución de las variables utilizadas en el proceso de Minería de datos.

6.2.2. EXTRACCIÓN DE VARIABLES

A partir de los datos iniciales se han obtenido 152 variables para cada una de las hojas LAS de las zonas de aplicación (apartado 4.3.2), de las que 139 tienen como finalidad constituir las variables a trabajar en un entorno de Aprendizaje automático y el resto complementarias para comprobar o posibilitar valorar los resultados.

Tabla 6.1. Variables agrupadas según el tipo de dato de procedencia

Tipo dato	Variables	Nº variables
LAS	*X, Y, Z, a, i, n, r, c*	8
Ortofotos	*R, G, B, NIR* y para cada banda sus estadísticos (8) según la segmentación de las ortofotografías	36
Diferencias Normalizadas	Combinación de las bandas anteriores dos a dos (6) y sus estadísticos (8) según la segmentación de las ortofotos	54
Modelos Digitales	*Pmf_t, Pmf_s, Pmf_ts, Mcc_t, Mcc_s, Mcc_ts* y sus estadísticos (8) según segmentación de las bandas con diferencias de alturas	54
	TOTAL	152

Las herramientas utilizadas para este fin se han explicado en el apartado 3.3.2 y han permitido la obtención de un fichero csv por cada una de las cuadrículas procesadas con el total de variables indicadas de manera resumida en la tabla 6.1.

Estas variables se han agrupado según la fuente que ha dado lugar a su consecución, por lo que se han distinguido cuatro grupos:

- Las extraídas directamente de los ficheros LAS.
- Las sacadas de las ortofotografías.
- Las conseguidas tras calcular las diferencias normalizadas.
- Las obtenidas a partir de los modelos digitales.

Aunque las variables referentes a los modelos digitales se han obtenido a partir de los datos LiDAR, se han considerado como un grupo aparte para facilitar su comprensión. En todas las variables, salvo en las conseguidas directamente de la BBDD del LiDAR, se han aplicado técnicas de segmentación que han permitido el cálculo de una serie de estadísticos a asignar a las variables contempladas por cada grupo. Por este motivo, tras concluir de explicar las variables de cada grupo se presenta una pequeña introducción referente a la segmentación en base a la cual se han determinado el resto de variables necesarias para completar la BBDD requerida.

6.2.2.1. Variables extraídas de los ficheros LAS

De los datos aportados por los archivos LAS (tabla 3.3) se han aprovechado las coordenadas ***X***, ***Y***, ***Z***, la intensidad (***i***), el número de retorno (***r***), el número de retornos en ese punto (***n***) y el ángulo de escaneo (***a***). También se ha considerado el valor de la clasificación (***c***), pero éste se ha reservado para la comparación con el resultado de la metodología a desarrollar, sin que se haya considerado como variable dependiente.

Tabla 6.2. Variables procedentes del fichero LAS

Denominación	Significado
X	Coordenadas planimétricas del punto
Y	
Z	Altitud (h)
i	Intensidad
a	Ángulo de escaneo
n	Número de retornos por pulso
r	Número que le corresponde a ese retorno
c	Clasificación

En cuanto a la intensidad, hay que indicar que se ha usado conforme está codificada, sin normalizar, al no disponer de criterio para efectuar este proceso, apareciendo un rango de valores distinto según se trate del vuelo del Gobierno Vasco (GV) o del de la Diputación Foral de Gipuzkoa (DFG).

Cabe señalar que en los análisis sobre los valores de intensidad que se han realizado a lo largo del experimento, se ha podido constatar una gran variedad de éstos para el mismo tipo de cobertura de terreno, desconociendo si es debido a la falta de normalización o al uso de

múltiples zonas de aplicación, en las que la respuesta espectral para la intensidad puede ser diversa, ya que (Kashani, et al. 2015; McGlone 2013) señalan que los valores de intensidad pueden variar dependiendo de la reflectancia y rugosidad de la superficie, el ángulo de escaneo, el rango de energía considerado, la energía transmitida, los múltiples retornos, la profundidad de la intensidad, el brillo producido por elementos próximos, el tamaño de apertura, la transmisión atmosférica y la humedad.

En cuanto a los valores de *R*, *G*, *B* aportados por estos datos indicar que no se han considerado porque han sido obtenidos a través de la superposición de la ortofotografía del PNOA del 2010, en lugar de la del 2008, año en el que se ha realizado el vuelo LiDAR objeto de este trabajo.

6.2.2.2. Variables extraídas de las ortofotografías

Considerando las coordenadas *X*, *Y*, que han sido extraídas por cada punto del fichero LAS, a partir de la ortofotografía RGB del 2008 se han obtenido para cada uno de esos puntos los valores *R*, *G* y *B*, previa separación de la información en sus tres bandas usando la librería GDAL (*Geospatial Data Abstraction Library*).

En el caso de la ortofotografía NIR, como se trata de una imagen con una única banda, directamente se ha obtenido el valor *NIR* que le corresponde a cada punto. De esta forma, gracias a la utilidad que SAGA dispone para asignar valores a puntos, a cada uno se le han otorgado los valores ***R***, ***G***, ***B***, ***NIR*** adecuados procedentes de las imágenes aéreas georreferenciadas, a partir de los cuales se ha derivado el cálculo del índice NDVI tal y como se explica en la siguiente sección.

Tabla 6.3. Variables procedentes de las ortofotografías

Denominación	Significado
R	Valor correspondiente a la banda roja
G	Valor correspondiente a la banda verde
B	Valor correspondiente a la banda azul
NIR	Valor correspondiente a la banda infrarroja

6.2.2.3. Variables referentes a las diferencias normalizadas

Con el fin de poder discriminar adecuadamente entre edificios y zonas arboladas o con vegetación, atendiendo a la bibliografía (Shan and Toth 2008) se ha decidido calcular el índice de la diferencia de vegetación normalizada (*Normalized Vegetation Difference Index*, NDVI).

Este índice, al igual que otros índices de vegetación (*Vegetation Indices*, VIs), pretende medir la biomasa o vigor de la vegetación (Campbell and Wynne 2011) y se ha aplicado mucho en el

ámbito de la Teledetección. Para su cálculo se combinan los valores de la banda del infrarrojo cercano (NIR) con la banda roja (*RED*) del visible de la siguiente manera (Rouse Jr, et al. 1974):

A la hora de adaptarlo en el ámbito de los datos LiDAR son muchos los autores (Bandyopadhyay, et al. 2013; Rottensteiner, et al. 2003; Ryan 2013) que utilizan lo que han denominado como pseudo-NDVI, que consiste en utilizar en lugar de la información de la banda del infrarrojo la intensidad aportada el sensor LiDAR.

En base a ello, y atendiendo a lo indicado por Ryan 2013, en esta investigación, dado que se dispone de la información aportada por las tres bandas del visible y la del NIR, se ha optado por calcular los pseudo-índices derivados del NDVI al combinar las cuatro bandas entre sí, resultando además del propio índice NDVI y el GREEN-RED utilizado por Motohka, et al. 2010 los relacionados en la siguiente tabla.

Tabla 6.4. Relación de índices de diferencia normalizada utilizados

Denominación	Banda 1	Banda 2
NIRR (NDVI)	NIR	RED
NIRG	NIR	GREEN
NIRB	NIR	BLUE
RG (GREEN-RED)	RED	GREEN
RB	RED	BLUE
GB	GREEN	BLUE

Estos índices se han calculando a partir de los niveles digitales (DN, *Digital Number*) aportados por las ortofotografías en lugar de utilizar las reflectividades, pero tal y como indican Chuvieco and Huete 2009 éstos pueden ser válidos si no se pretende conceder un valor físico a los resultados.

En este desarrollo se ha seguido lo señalado por Gil-Yepes and Ruiz 2012; Ryan 2013; Viñas, et al. 2006 y se ha calculado una imagen de diferencias normalizadas por cada una de las combinaciones anteriores, para luego además del propio valor, a través de las agrupaciones derivadas de la segmentación (apartado 6.2.2.5) asignar los estadísticos correspondientes.

6.2.2.4. Variables extraídas de los modelos digitales

Desde los inicios los problemas de clasificación de puntos LiDAR se han centrado prácticamente en la separación de éstos en puntos terreno y no terreno, considerando estos últimos aquellos que disponen de cierta altura sobre el suelo.

Muchos son los artículos y programas que se dedican solamente a esto, y otros, a partir de esta primera distinción, pretenden llegar un poco más lejos diferenciando mayor número de entidades, tales como edificaciones, arbolados, vehículos, etc.

En cualquiera de estos supuestos, se suele trabajar con estos datos clasificados para la generación de un DTM (*Digital Terrain Model* / Modelo Digital del Terreno, MDT), considerando aquellos puntos que solamente hacen referencia al terreno; y con las categorías que simulan las superficies de la zona (edificios, arbolados, etc.) para obtener un DSM (*Digital Surface Model* / Modelo Digital de Superficie, MDS).

De esta forma, a partir de estos dos modelos se genera el nDSM (*normalized Digital Surface Model* / Modelo Digital de Superficie normalizado), que consiste en restar al DSM el DTM (Kressler and Steinnocher 2006).

Este nuevo ráster juega un papel muy importante sobre todo cuando se quieren identificar objetos superficiales que interceptan en el terreno (Hashemi 2008). Por esta razón, su uso se está estandarizando en los últimos años impulsado por el auge de los modelos 3D en la gestión de los entornos urbanos.

La pretensión de esta sección es llegar a establecer un suelo común para a partir de éste determinar las alturas de los objetos. En este marco, y tras estudiar los distintos tipos de algoritmos existentes para este fin (apartado 2.1), de los cuatro grandes bloques - morfológicos, densificación progresiva, basados en superficies y basados en segmentación - en una primera instancia se ha desarrollado el algoritmo de Sohn and Dowman 2002 que se basa en la densificación progresiva de la triangulación (PTD), dividiendo el proceso en dos partes: en la primera se genera una semilla con los puntos más bajos; y, en la segunda, se procede a densificar y mejorar la triangulación anterior mediante un proceso iterativo. Los resultados conseguidos con la aplicación del algoritmo anterior no han resultado satisfactorios, por lo que se ha recurrido al uso de la librería de software libre SPDlib que posee herramientas para el almacenaje y procesamiento de datos de escaneo 3D (LiDAR, ALS, TLS, etc.) (Bunting 2013).

La particularidad que presenta SPDlib es que utiliza el formato propio *Sorted Pulse Data* (SPD) para la gestión de los datos LiDAR. Éste organiza la información por pulsos, a partir de los cuales emergen los distintos puntos con sus atributos, resultando más rápida y cómoda su administración. También dispone del formato *Unsorted pulse data* (UPD) que viene a ser el equivalente al fichero estándar LAS por la falta de indexación (Bunting, et al. 2013b).

El proceso a desarrollar se puede extraer del diagrama que se dispone en Bunting, et al. 2013a y que para el trabajo a desarrollar aquí vendría definido por los pasos recogidos en la tabla 6.5.

La cuarta fase recoge las distintas posibilidades que presenta la librería para clasificar los puntos en terreno / no terreno. De ellas, se han utilizado las que hacen referencia al algoritmo MCC (*Multiscale Curvature Classification*): spdmccgrd; y, al PMF (*Progressive Morphological Filter*): spdpmfgrd, ya explicados en el apartado 2.1.

Tabla 6.5. Relación del proceso a ejecutar con SPDlib

Fase	Proceso	Comando
1	Transformar los .LAS a SPD	spdtranslate
2	Eliminar puntos anómalos	spdrmnoise
3	Limpiar la clasificación existente	spdclearclass
4	Clasificar en suelo / no suelo	spdpmfgrd spdmccgrd spdpffgrd spdpolygrd spdplanegrd (edificios)
5	Generar modelos digitales	spdinterp

Con cualquiera de los dos algoritmos explicados se genera un fichero SPD en el que aparecen los puntos clasificados como **terreno / no terreno**, pero para aprovechar esta información fuera de la librería SPDlib hay que exportar el fichero a ráster, debiendo generar el MDT con los puntos clasificados como terreno y la del MDS con los puntos no terreno. Para la interpolación, tal y como indica (Bater and Coops 2009) se ha elegido la opción *NATURAL_NEIGHBOR* por considerarla la más adecuada. Este condicionante ha supuesto el uso de estos modelos digitales, dos por cada algoritmo, en el flujo de trabajo.

Las modelos digitales resultantes de este proceso, además de utilizarse para obtener los valores de altura en cuanto al terreno (t) o superficie (s) en cada caso (***Pmf_t***, ***Pmf_s***, ***Mcc_t***, ***Mcc_s***), se han empleado para calcular el modelo digital del terreno normalizado (nDSM) que no es más que la resta del MDT al MDS; o lo que es lo mismo, el modelo con las alturas de los objetos ubicados sobre el terreno (figura 6.1). En la tabla 6.6 se indican las variables obtenidas de este proceso.

Tabla 6.6. Variables procedentes de los modelos digitales

Denominación	Significado
Pmf_t	Modelo digital del terreno según el algoritmo PMF
Pmf_s	Modelo digital de superficie según el algoritmo PMF
Pmf_st	Modelo digital de superficie normalizado según el algoritmo PMF
Mcc_t	Modelo digital del terreno según el algoritmo MCC
Mcc_s	Modelo digital de superficie según el algoritmo MCC
Mcc_st	Modelo digital de superficie normalizado según el algoritmo MCC

De esta forma, se dispone de dos ráster que visualizan los modelos normalizados por cada zona, uno según el algoritmo MCC y otro conforme a PMF. Con ambos se ha generado un único ráster con dos bandas, para luego segmentar según lo indicado en el apartado 6.2.2.5.

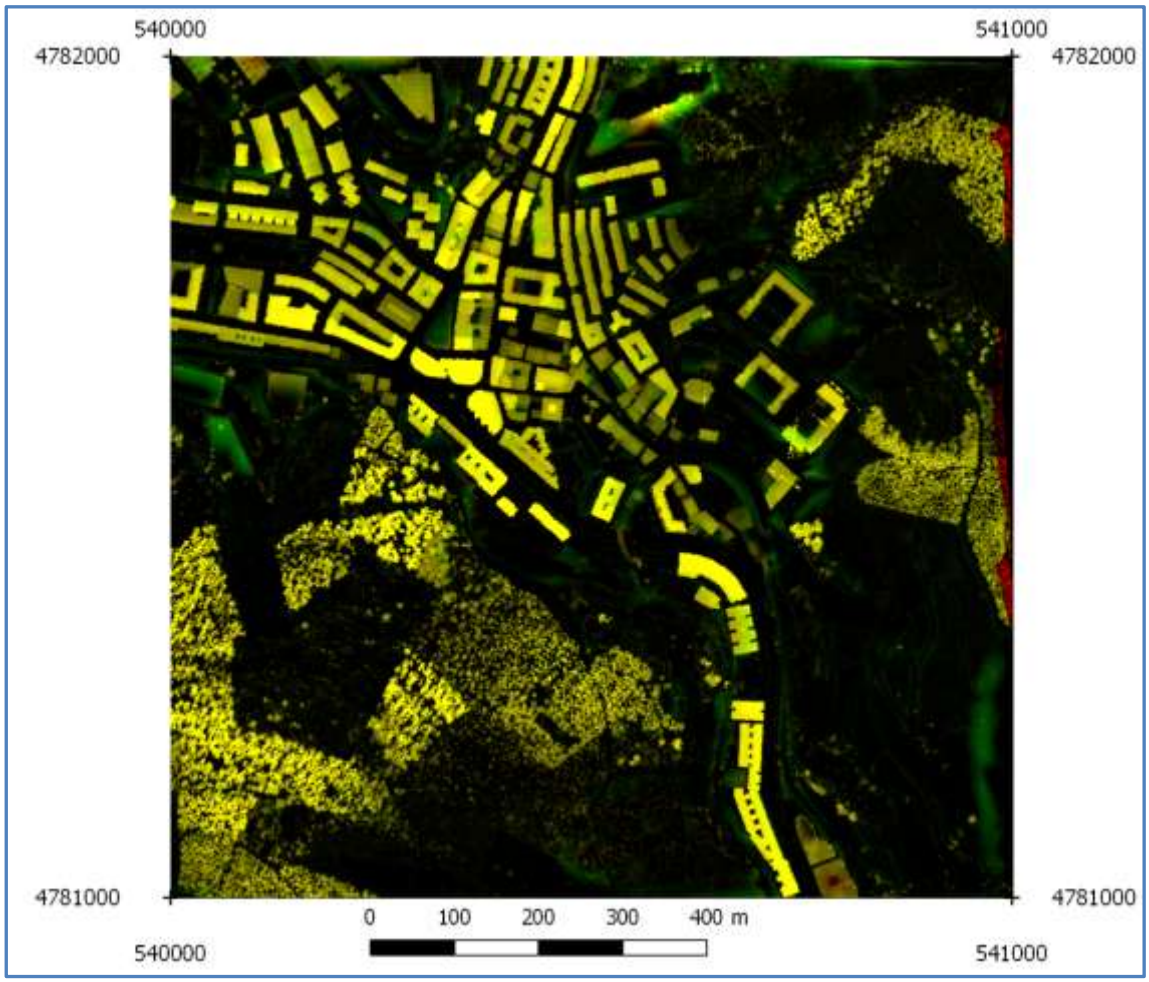

Figura 6.1. Modelo digital de superficie normalizado de la cuadrícula 5404782

6.2.2.5. Segmentación

La segmentación de imágenes significa la partición de una imagen en regiones significativas basadas en criterios de homogeneidad o heterogeneidad (Haralick and Shapiro 1992), en esta investigación la idea llevada a cabo ha consistido en segmentar para poder agrupar a posteriori los puntos de la nube pertenecientes a los mismos segmentos, calculando para ellos una serie de estadísticos que se detallaran a posteriori.

A pesar de que el planteamiento general de la segmentación es el mismo tanto para puntos como para imágenes, en cada ámbito se utilizan técnicas y algoritmos diferentes. Los algoritmos de segmentación de imágenes se basan en la discontinuidad y similitud de los niveles digitales: los de discontinuidad buscan cambios bruscos y los de similitud rastrean zonas con valores similares conforme a unos criterios prefijados.

En el ámbito de las nubes de puntos, se puede decir que las técnicas de segmentación de imágenes están más extendidas que las de puntos, así en Wang and Shan 2009 se indica que la

mayoría de los métodos de segmentación se basan en el rango de imágenes 2,5 D o modelos TIN.

A este respecto, gran parte de la bibliografía advierte que los mejores resultados de clasificación se consiguen con densidades de puntos elevadas, siendo la mayoría de los estudios existentes aplicados a la extracción de edificios en la que en un porcentaje elevado los segmentos referentes a los tejados de los edificios se obtienen de la segmentación de imágenes bien sean aéreas o de satélites, incluso las derivadas de los datos LiDAR (DTM, DEM, DSM, intensidad) (Tomljenovic, et al. 2015).

No obstante, hay que tener presente que la segmentación constituye un paso previo a la clasificación, en la que se determinan ciertos elementos tendiendo luego en la fase de clasificación a señalar cada uno de ellos a qué categoría corresponde. Sin embargo, a lo largo de la literatura, en el ámbito de las nubes de puntos, se entremezclan mucho los conceptos de segmentación, clasificación y filtrado como si de lo mismo se tratará; dando lugar, en muchos casos, al uso erróneo de los mismos.

Además, muchos de los algoritmos existentes para dicho fin combinan técnicas de clasificación y segmentación constantemente. Y tanto algoritmos de segmentación como de clasificación se utilizan para el filtrado de la nube de puntos LiDAR a la hora de buscar su clasificación, tal y como se ha podido ver en el apartado 2.1.

En esta investigación el algoritmo utilizado para la segmentación ha sido el de Edison por lo que a continuación se comentan sus características.

6.2.2.5.1. Algoritmo de segmentación Edison

El algoritmo *Edge Detection and Image SegmentatiON* (Edison) ha sido desarrollado por el *Robust Image Understanding Laobratory at Rutgers University* y combina los métodos de *Edge Detection* y *Mean Shift* para la segmentación de imágenes.

El método de *Edge Detection* se basa en la detección de bordes y pretende identificar y localizar grandes discontinuidades en la imagen (Senthilkumaran and Rajesh 2009). Por su parte, el algoritmo *Mean Shift* constituye un método no paramétrico que permite localizar el máximo de la función de densidad, también denominado algoritmo *mode-seeking* (Cheng 1995).

Según Comaniciu and Meer 2002 los métodos no paramétricos pueden clasificarse en dos grandes grupos: jerárquicos y de estimación de la densidad. *Mean Shift* se basa en la estimación de la densidad utilizando para ello núcleos o *kernel*s radialmente simétricos (*radially symmetric kernels*), desarrollando el proceso en dos pasos: el filtrado de la imagen original y el agrupamiento o *clustering* de los puntos filtrados (Pantofaru and Hebert 2005).

El filtrado analiza la probabilidad de la función de densidad remarcando los datos en el espacio imagen con el objetivo de resaltar los puntos con una alta densidad. La búsqueda la realiza en

una esfera centrada en el punto usando un radio espacial (h_s, *spatial radius*) que es el que define el *kernel* o tamaño espacial de la imagen, y el rango de aplicación o número de bandas a usar, conocido como *range radius* o *colour radius* (h_r). Ambos valores controlan el tamaño del *kernel* y determinan la resolución del modo de detección. h_s constituye el ancho de banda espacial (*spatial bandwidth*) o la distancia espacial entre clases, mientras que h_r es conocido como la diferencia entre clases o el atributo del ancho de banda (*attribute bandwidth*); además, también hay que tener en cuenta el parámetro *M* (*smallest significant feature size*) que hace referencia al umbral de fusión (Ming, et al. 2015).

Para cada punto se calcula iterativamente el valor del gradiente desplazando ese punto en una dirección determinada hasta que ese gradiente se haga inferior al umbral marcado. Lo ideal es que el gradiente sea cero, de manera que permita cambiar de punto e ir suavizando sucesivamente la imagen.

En la fase de *clustering*, también denominada paso de post-proceso, se realiza la agrupación de píxeles, usando en el caso de *Edison* la asociación a través del algoritmo *Edge Detection* (Pantofaru and Hebert 2005).

6.2.2.5.2. Parámetros de segmentación

Basándose en la breve explicación teórica anterior, a continuación se expresan los parámetros específicos utilizados en las segmentaciones realizadas:

- *Spatial radius* (h_s): define la distancia máxima entre el pixel vecino y el centro del kernel. El radio se da en número de píxeles y determina los píxeles vecinos a considerar.
- *Range radius* (h_r): define el kernel en término de valores de pixel. Se trata del radio en el espacio multiespectral.
- *Minimum region size (M)*: tamaño mínimo de la región a segmentar. Los clusters más pequeños se unirán con los vecinos adquiriendo el valor radiométrico de los píxeles más cercanos.
- *Scale factor*: valor para escalar la imagen antes de ser procesada.

Comaniciu and Meer 2002 indican que la segmentación no es muy sensible a la elección de la resolución de los parámetros h_s y h_r, realmente el número de regiones en la segmentación es controlada por h_r y *M*. De manera que para controlar los efectos de variaciones pequeñas si en la imagen existe gran diversidad hay que aumentar los valores de h_r y *M*.

En la presente investigación estos parámetros se han particularizado por un lado para la segmentación de las ortofotografías; y, por otro, para los modelos digitales, adquiriendo en cada caso los parámetros mostrados en la tabla 6.7.

Tabla 6.7. Parámetros Edison utilizados en la segmentación

Parámetros	Ortofotografías	Modelos digitales
Radio espacial	5	10
Rango del radio	40	100
Región mínima	200	200
Factor de escala	1	2
Tamaño mínimo del objeto	2	1
Simplificación de polígonos	0,1	0,1

La diferencia en los valores de los parámetros para la segmentación de las ortofotografías y de los modelos digitales viene derivada por la resolución espacial de los mismos, ya que mientras las ortofotografías cuentan con una resolución espacial de 25 cm, los modelos se han generado con 1m.

6.2.2.5.3. Variables derivadas de la segmentación

Las imágenes que se segmentan son por un lado la constituida por las bandas *R*, *G*, *B* y *NIR* de las ortofotografías y por otro la imagen constituida con las dos bandas de los modelos digitales. En la tabla 6.8. se muestra el total de las 128 variables deducidas de la segmentación.

Para cada imagen se procede a agregar los puntos LiDAR que existen en cada segmento. De esta forma para la primera se ha procedido a calcular en cada caso los estadísticos y su asignación con SAGA para las bandas *R*, *G*, *B*, *NIR* y las derivadas de los cálculos de diferencias normalizadas, de forma que en ambos casos a todos los puntos de la misma agrupación se les han estipulado valores similares, como si de una única entidad se tratara.

De la misma forma, según la segmentación de los modelos digitales, se ha procedido a determinar los mismos estadísticos para los valores de diferencia de altitud entre la cota del punto LiDAR y la correspondiente de las superficies provenientes de los dos algoritmos de clasificación terreno / no terreno (*Pmf_t*, *Pmf_s*, *Pmf_st, Mcc_t*, *Mcc_s, Mcc_st*).

En ambos casos os estadísticos que se han calculado son: mínimo (min), máximo (max), rango (range), media (mean), desviación estándar (std), percentil 25 (Q25), percentil 50 (Q50) y percentil 75 (Q75).

Mencionar que en la propuesta metodológica no se han contemplado los valores *Mcc_s* junto con sus estadísticos por resultar exactamente iguales a los obtenidos con *Pmf_s*, llegando a la conclusión de que *Mcc_s* coincide con *Pmf_s* porque para la generación del modelo digital de superficies la librería SPDlib considera los puntos de mayor altitud.

Tabla 6.8. Variables derivadas de la segmentación

Tipo dato	Origen	Variables
Ortofotos	Banda R	*R_min, R_max, R_range, R_mean, R_std, R_Q25, R_Q50, R_Q75*
	Banda G	*G_min, G_max, G_range, G_mean, G_std, G_Q25, G_Q50, G_Q75*
	Banda R	*B_min, B_max, B_range, B_mean, B_std, B_Q25, B_Q50, B_Q75*
	Banda NIR	*NIR_min, NIR_max, NIR_range, NIR_mean, NIR_std, NIR_Q25, NIR_Q50, NIR_Q75*
Diferencias normalizadas	NIRR	*NIRR_min, NIRR_max, NIRR_range, NIRR_mean, NIRR_std, NIRR_Q25, NIRR_Q50, NIRR_Q75*
	NIRG	*NIRG_min, NIRG_max, NIRG_range, NIRG_mean, NIRG_std, NIRG_Q25, NIRG_Q50, NIRG_Q75*
	NIRB	*NIRB_min, NIRB_max, NIRB_range, NIRB_mean, NIRB_std, NIRB_Q25, NIRB_Q50, NIRB_Q75*
	RG	*RG_min, RG_max, RG_range, RG_mean, RG_std, RG_Q25, RG_Q50, RG_Q75*
	RB	*RB_min, RB_max, RB_range, RB_mean, RB_std, RB_Q25, RB_Q50, RB_Q75*
	GB	*GB_min, GB_max, GB_range, GB_mean, GB_std, GB_Q25, GB_Q50, GB_Q75*
Modelos Digitales	Pmf_t	*Pmf_t_min, Pmf_t_max, Pmf_t_range, Pmf_t_mean, Pmf_t_std, Pmf_t_Q25, Pmf_t_Q50, Pmf_t_Q75*
	Pmf_s	*Pmf_s_min, Pmf_s_max, Pmf_s_range, Pmf_s_mean, Pmf_s_std, Pmf_s_Q25, Pmf_s_Q50, Pmf_s_Q75*
	Pmf_st	*Pmf_st_min, Pmf_st_max, Pmf_st_range, Pmf_st_mean, Pmf_st_std, Pmf_st_Q25, Pmf_st_Q50, Pmf_st_Q75*
	Mcc_t	*Mcc_t_min, Mcc_t_max, Mcc_t_range, Mcc_t_mean, Mcc_t_std, Mcc_t_Q25, Mcc_t_Q50, Mcc_t_Q75*
	Mcc_s	*Mcc_s_min, Mcc_s_max, Mcc_s_range, Mcc_s_mean, Mcc_s_std, Mcc_s_Q25, Mcc_s_Q50, Mcc_s_Q75*
	Mcc_st	*Mcc_st_min, Mcc_st_max, Mcc_st_range, Mcc_st_mean, Mcc_st_std, Mcc_st_Q25, Mcc_st_Q50, Mcc_st_Q75*

6.3. APLICACIÓN DE ALGORITMOS DE APRENDIZAJE AUTOMÁTICO

Una vez asociado a cada punto LiDAR toda la información precisa conforme se ha explicado en el punto 6.2. y teniendo en cuenta lo explicado sobre minería de datos en el apartado 2.2., en esta sección se han aplicado ciertos algoritmos de Aprendizaje automático relacionados con los árboles de decisión.

Para su procesado se ha diseñado un script bajo python que contempla la utilización de los algoritmos de clasificación DECISION TREE, EXTRA TREE y RANDOM FOREST utilizando la librería *scikit learn* en el entorno de spyder, ya comentados en el apartado 3.3.2.

En el apartado 4.3.2 se han descrito las zonas de aplicación, así como los criterios seguidos para su selección. De todas ellas se dispone de ficheros en formato csv con las 152 variables extraídas según lo explicado en el apartado 6.2, pero son 139 las que constituyen las variables dependientes (tabla 6.9) que forman parte de los procesos de Minería de datos a desarrollar en esta investigación, constituyendo la variable independiente la información obtenida en el entrenamiento a partir de la BTA, tal y como se ha explicado en el apartado 4.2.3.2, según la cual se pretende predecir si se trata de edificación, red viaria, vía férrea, tendido eléctrico, puentes, etc., cualquiera de los elementos recogidos en la tabla 4.9.

Tabla 6.9. Variables dependientes para el aprendizaje automático

Tipo dato	Variables	Nº variables
LAS	*a*, *i*, *n*, *r*	4
Ortofotos	*R*, *G*, *B*, *NIR* y para cada banda sus estadísticos (8) según la segmentación de las ortofotos (ver tabla 6.8)	36
Diferencias normalizadas	Combinación bandas anteriores dos a dos (6) y sus estadísticos (8) según la segmentación de las ortofotos (ver tabla 6.8)	54
Modelos Digitales	*Pmf_t*, *Pmf_s*, *Pmf_ts*, *Mcc_t*, *Mcc_s* y sus estadísticos (8) según segmentación de bandas con diferencias de alturas (ver tabla 6.8)	45
	TOTAL	139

Recordar que de las 152 variables extraídas, las coordenadas y la clasificación (4), obtenidas de los ficheros LAS, junto con el modelo de superficies normalizado procedente del algoritmo MCC y sus estadísticos (9) no han sido consideradas como variables dependientes.

Analizando los atributos establecidos y atendiendo a las unidades de medida se puede comprobar cómo las variables dispuestas se han organizado en cuatro grupos (tabla 6.10): en el primer grupo, se han contemplado las correspondientes a las imágenes, junto con los valores de intensidades, todas ellas medidas en bits.

En el segundo grupo se han hallado las variables que hacen referencia a las diferencias normalizadas con una escala entre -1 y 1, según los cálculos realizados entre bandas. Dentro del tercer conjunto se han incluido a los que hacen referencia a unidades métricas; y, por último, en el cuarto grupo, se han contemplado el resto de variables con diferentes tipos de unidades, tales como el ángulo, retorno, números de retornos o los estadísticos.

En base a esas variables, como resultado del proceso se ha logrado la **PREDICCIÓN** que pretende mejorar los valores indicados en la clasificación automática de los datos LiDAR (*c*, campo *classification*), por lo que este valor no ha sido considerado como atributo para el tratamiento de Minería de datos, pero sí se ha considerado para verificar lo solución alcanzada.

Tabla 6.10. Tipos de variables con sus unidades

Grupo	Atributo	Unidad
1	Imágenes e intensidad	bits
2	NDVI	-1 a 1
3	*X, Y, Z, Pmf_t, Pmf_s, Pmf_ts, Mcc_t, Mcc_s, Mcc_ts*	Metros
4	*angle*	Grados sexagesimales
	return	Orden
	number_of_return	Cantidad
	Estadísticos	Según procedencia
	classification	Orden

El desarrollo del aprendizaje automático necesita de un conjunto de **datos** a partir de los cuales realizar el **entrenamiento**, y de otro conjunto distinto (**datos de test**) en el que se aplique el resultado de ese entrenamiento y con los que validar los resultados.

Estos dos grupos de datos se pueden establecer dentro del mismo conjunto, de manera que un porcentaje alto sea para el entrenamiento. En la bibliografía se indica que se debe considerar dos tercios de las muestras para el entrenamiento y un tercio para la validación, pero suele ser habitual que ronde entre un 60 % entrenamiento y un 40 % validación (García-Gutiérrez 2012). Aunque hay otros autores que lo consideran al 50 % (Lu, et al. 2014). Otra posibilidad se basa en tener en cuenta dos grupos de datos distintos, siendo bastante mayor el del entrenamiento frente al de validación, de manera que se consideren un mayor número de casos diferentes para el entrenamiento. En esta investigación se ha utilizado esta segunda configuración partiendo de 126 hojas LAS para el entrenamiento (77 %) y 36 para el test (23 %).

Por último, para verificar los resultados se han utilizado los errores de tipo I y de tipo II consistentes en evaluar a partir de la matriz de confusión (MC) la predicción realizada. En las pruebas realizadas los resultados se han verificado con los estadísticos de *Precision* o la fiabilidad del productor (FP), *Recall* o la fiabilidad del usuario (FU) y *f1-score*, ya explicadas en el apartado 5.2.2.

Aplicado el algoritmo correspondiente, además de un informe resumen con la referencia de los datos utilizados para el entrenamiento y validación, el número total de puntos procesados en cada conjunto, su distribución según el entrenamiento realizado, las estadísticas y la matriz de confusión con los datos de verificación, se ha obtenido otro csv similar al inicial en el que se ha añadido una columna más con la predicción realizada, para su posterior visualización y análisis en un entorno SIG.

Pero antes de proceder con la aplicación en sí se deben explicar los parámetros que utiliza la librería *scikit learn* en el entorno de la clasificación.

6.3.1. PARÁMETROS DE CLASIFICACIÓN EN SCIKIT-LEARN

La librería *scikit-learn* permite aplicar diferentes algoritmos de Aprendizaje automático bajo el lenguaje de programación *python*. Dispone de unas herramientas simples y eficientes para su aplicación en el ámbito del *Data Mining* y análisis de datos accesibles a todo el mundo por tratarse de una licencia BSD de software libre (Scikit-learn 2015a).

Del conjunto de algoritmos disponibles se han utilizado los basados en árboles de clasificación que dispone ubicados en dos módulos distintos: DECISION TREE (DT) en *sklearn.tree* y EXTRA TREE (ET) y RANDOM FOREST (RF) en *sklearn.ensemble*, todos ellos para el aprendizaje supervisado, constituyendo los dos últimos métodos de ensamblado.

Estos módulos vienen definidos por parámetros, que en el caso de ET y RF hacen referencia a los mismos, no contemplando DT ni *n_estimators* ni otros parámetros referentes al *bootstrap* u *oob*. Se puede afirmar que DT es una simplificación de los otros dos. En la tabla 6.11 se describen los parámetros que utilizan, indicando para cada uno los valores por defecto que utiliza (Scikit-learn 2015b).

Tabla 6.11. Parámetros utilizados en scikit-learn en árboles de decisión

Parámetros	Descripción	Valor por defecto
n_estimators	Número de árboles del bosque	10
criterion	Función para medir la calidad de la partición	*gini*
max_features	Número de variables a considerar al buscar la mejor partición	*Auto = sqrt(n_features)*
max_depth	Máxima profundidad del árbol	*None*
min_samples_split	Mínimo número de muestras requeridas para partir un nodo interno	2
min_samples_leaf	Mínimo número de muestras en las hojas de nueva creación	1
min_weight_fraction_leaf	Mínima fracción ponderada de las muestras de entrada requeridas para ser un nodo hoja	0
max_leaf_nodes	Máximo número de nodos hoja	*None*
bootstrap	Si se usa o no *bootstrap* para crear los árboles	*True*
oob_score	Estado del *Out of bag*	*Bool*
n_jobs	Número de procesadores	1
random_state	Estado de la aleatoriedad	*None*
verbose	Visualiza el proceso	0
warm_start	Añade más estimadores	*False*
class_weight	Pesos asociados con clases	

A continuación se presenta una reseña de los parámetros más importantes:

- *n_estimators*: con este valor se indican el número de árboles a construir previo a la determinación de la predicción, bien por *voting* o *averages*. Un gran número de árboles ofrece un rendimiento mejor, mostrando predicciones más fuertes y estables pero ralentizando el proceso.

  Este parámetro resulta clave a la hora de aplicar RF por lo que existen distintas referencias que se encargan de estudiar su valor más idóneo. Por ejemplo, en (Latinne, et al. 2001) se proponen cuatro técnicas: *Bagging*, *Random Subspace Method* (RSM) or *Multiple Feature Subsets* (MFS), la combinación de ***Bag****ging* con *Random* ***S****ubspace* (RS) denominada *Bagsf* y *Breiman's Random Forest* (Bagrf). *Bagging* y MFS pueden ser aplicadas a cualquier algoritmo de aprendizaje, pero Bagrf sólo a árboles de decisión.

- *max_features*: constituye el máximo número de variables permitidas para la creación de los árboles individuales. Las opciones disponibles normalmente son *None* para usar todos los atributos disponibles, *Auto* su valor varía según el software utilizado, *sqrt* considera la raíz cuadrada del total, *log2* considera el logaritmo en base 2. Pudiendo ser también un número entero o porcentaje.

  El incremento de *max_features* generalmente mejora el rendimiento del modelo en cada nodo, ofreciendo un mayor número de opciones a considerar pero disminuyendo la velocidad de ejecución del algoritmo.

- *min_sample_leaf*: este parámetro permite marcar el mínimo número de muestras a considerar para llegar al nodo hoja. A menor valor mayor tendencia para considerar ruido de los datos de entrenamiento.

- *min_samples_split*: define el número mínimo de muestras utilizadas para establecer la partición de los nodos. Se suele utilizar para limitar el crecimiento del árbol, con pocas observaciones no tiene sentido su uso, pero con muchas se puede partir antes y aún así obtener árboles lo suficientemente grandes.

- *max_depth*: hace referencia a la profundidad máxima del árbol. Se trata de un límite para dejar de seguir dividiendo los nodos. Si se quiere que sea lo más largo posible no se debería poner ningún término.

- *oob_score*: posibilita estimar el error de generalización.

- *n_jobs*: permite indicar el número de procesadores a utilizar durante el cálculo. Sería conveniente marcar todos los disponibles en la máquina en la que se ejecuta el algoritmo, con el objetivo de poder realizar el cálculo lo más rápido posible.

- *ramdom_state*: admite marcar el azar al generar los árboles individuales. Si se fija esta aleatoriedad todos los árboles individuales se realizaran de la misma forma, siempre

que se mantengan las variables y los datos de entrenamiento, consiguiendo mejores resultados.

Los parámetros *oob_score*, *n_jobs* y *ramdom_state* no son obligatorios para el buen funcionamiento del modelo, pero ofrecen unas herramientas que facilitan el procesamiento.

Además de estas consideraciones para los parámetros, Millard and Richardson 2015 indican, entre otras cosas, que el tamaño del entrenamiento influye en el resultado siendo aconsejable disponer de una distribución aleatoria y con una correlación mínima, debiendo eliminar variables correladas.

6.3.2. PRIMEROS ENSAYOS

Los primeros ensayos se han realizado con dos grupos de muestras distintas; por un lado, considerando 90 hojas y por otro con tan solo 22. Ambos grupos procesados según lo indicado en el punto 6.2.2. y considerando un 60 % de los datos como entrenamiento y un 40 % de los mismos como test. En la tabla 6.12 y la figura 6.2 se puede apreciar la distribución de los datos por categorías para los dos conjuntos de datos.

Tabla 6.12. Distribución de puntos por categorías con 90 y 22 hojas LiDAR

Categorías		90 hojas		22 hojas	
Denominación	BTA	NP	%	NP	%
Suelo	2	3201832	33,2	1183855	47,4
Vegetación baja	3	2978136	30,9	524178	21,0
Vegetación media	4	628820	6,5	130367	5,2
Vegetación alta	5	1585384	16,4	241762	9,7
Edificación	6	783667	8,1	271295	10,9
Agua	9	23349	0,2	7520	0,3
Ferrocarril	10	48484	0,5	20821	0,8
Carreteras	11	340253	3,5	103598	4,1
Tendido eléctrico	16	52408	0,5	14793	0,6
	Total	9642333	100	2498189	100

Analizando esos datos se ha comprobado que prácticamente el reparto por categorías es similar en las dos muestras, resultando las clases con mayor volumen la de suelo (2) y vegetación baja (3). Discrepando considerablemente la vegetación alta (5) y los edificios o construcciones (6). El resto de las clases muestran unos porcentajes muy similares y prácticamente inexistentes frente al resto de las categorías comentadas anteriormente, salvo las carreteras (11).

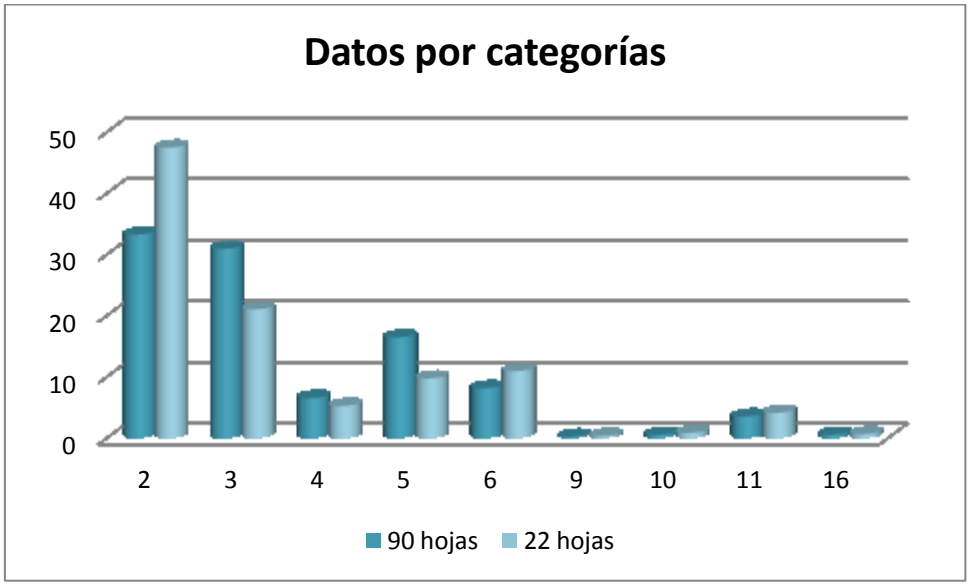

Figura 6.2. Distribución de datos por categorías con 90 y 22 hojas LiDAR

Con estos datos se ha procedido a aplicar los algoritmos de Aprendizaje automático indicados en el apartado 6.3.1 según los parámetros que por defecto admiten esos algoritmos y utilizando para la validación de los resultados las fiabilidades explicadas en el apartado 5.2.2. Del análisis de los resultados conseguidos con ambas muestras se han obteniendo las siguientes ideas:

- En relación a los algoritmos señalados de la librería *scikit-learn* indicados anteriormente, los resultados han mostrado que *Random Forest* resulta algo mejor que *Extra Tree* y bastante suprior a los *Decision Tree*. En (Lu, et al. 2014) ya se apunta a que los métodos de ensamblado ofrecen mejor solución que los *Decision Tree*.

Tabla 6.13. Resultados *Random Forest* con 90 y 22 hojas procesadas

Categorías		*FP*	*FP*	*FU*	*FU*
Denominación	BTA	22 hojas	90 hojas	22 hojas	90 hojas
Suelo	2	0,93	0,92	0,95	0,94
Vegetación baja	3	0,94	0,95	0,92	0,95
Vegetación media	4	0,92	0,93	0,90	0,91
Vegetación alta	5	0,92	0,94	0,91	0,93
Edificación	6	0,91	0,90	0,91	0,90
Agua	9	0,81	0,82	0,61	0,65
Ferrocarril	10	0,85	0,86	0,76	0,75
Carreteras	11	0,84	0,84	0,77	0,74
Tendido eléctrico	16	0,49	0,47	0,22	0,24

- En cuanto a las categorías a determinar, el tendido eléctrico (16) en ningún caso ha quedado bien predicho. Las carreteras (11) y vías férreas (10) han obtenido unos resultados similares, siendo un poco mejores en el caso de las carreteras. A su vez, en

ambos casos la fiabilidad del usuario se ha manifestado un poco inferior a la del productor.

- Curiosamente el agua (9) se ha presentado con una buena *FP*, a pesar de que el sensor utilizado no es capaz de determinarla, hecho que ha quedado reflejado en la fiabilidad del usuario (*FU*) con los valores más bajos (0,61) tras los del tendido eléctrico.
- Entre las variables que más importancia han tenido aparecen las relacionadas con los modelos digitales *Mcc* y *Pmf* y la correspondiente al NDVI (*NIRR*), sin mostrar ninguna tendencia clara. En los dos casos, como variable menos importante ha emergido la referente al número de retornos, seguida de los estadísticos de la banda azul pero sin un comportamiento determinado.
- También se ha comprobado que introducir ficheros de distintos vuelos, dado que algunas variables utilizan distinta escala de medida, no altera significativamente los resultados, al igual que el uso de hojas que no contengan los elementos a catalogar.

En consecuencia, se ha presupuesto que los modelos generados a través de los algoritmos de inducción se han sobre-ajustado (*over-fitting*) a los datos utilizados. Por este motivo, se ha considerado que para evitar ese sobre-ajuste resulta necesario diferenciar entre el conjunto de datos de entrenamiento y el de test.

Teniendo en cuenta todas estas apreciaciones se ha decidido proceder de manera que los 90 ficheros anteriores sean incluidos en el entrenamiento y procesar 17 hojas más para comprobar la predicción dada por el nuevo modelo.

Con estas nuevas muestras, al comparar la distribución de puntos por categorías se ha confirmado que las tendencias comentadas anteriormente se mantienen, obteniendo los porcentajes de la tabla 6.14. También se puede constatar que el porcentaje de puntos del conjunto de validación es un poco mayor al 20 % del entrenamiento.

Tabla 6.14. Distribución de puntos por categorías con 90 hojas de entrenamiento y 17 de test

Categorías		Entrenamiento: 90 hojas		Validación: 17 hojas	
Denominación	BTA	NP	%	NP	%
Suelo	2	3201832	33,2	898691	45,6
Vegetación baja	3	2978136	30,9	368398	18,7
Vegetación media	4	628820	6,5	139703	7,1
Vegetación alta	5	1585384	16,4	201196	10,2
Edificación	6	783667	8,1	278479	14,1
Agua	9	23349	0,2	4024	0,2
Ferrocarril	10	48484	0,5	21139	1,1
Carreteras	11	340253	3,5	49318	2,5
Tendido eléctrico	16	52408	0,5	10288	0,5
	Total	9642333	100	1971236	100

En la tabla 6.15. se pueden ver los estadísticos que se han obtenido en este caso, confirmándose en primer lugar la hipótesis de sobre-ajuste que se ha producido en el cálculo anterior, quedando patente que la mejor predicción se da en el caso de las edificaciones (6) con una *FP* del 0,81 y una *FU* del 0,69, a pesar de que la mejor *FU* se consigue para el suelo (2) con un 0,75. Está claro, al igual que en el caso anterior, que las clases 9, 10, 11 y 16 no han quedado bien predichas, pero esto puede ser debido a que ya en el entrenamiento de los datos los porcentajes que les corresponden a estas categorías han sido muy bajos.

Tabla 6.15. Resultados *Random Forest* con 90 hojas de entrenamiento y 17 de validación

Categorías		*FP*	*FU*	*f1-score*
Denominación	BTA			
Suelo	2	0,64	0,75	0,69
Vegetación baja	3	0,48	0,53	0,50
Vegetación media	4	0,36	0,13	0,19
Vegetación alta	5	0,50	0,47	0,48
Edificación	6	0,81	0,69	0,75
Agua	9	0,19	0,03	0,05
Ferrocarril	10	0,27	0,02	0,03
Carreteras	11	0,22	0,12	0,15
Tendido eléctrico	16	0,02	0,00	0,00
	Avg / total	**0,58**	**0,60**	**0,58**

Al analizar la matriz de confusión (MC) de la tabla 6.16, ha quedado claro que el mayor desconcierto se ha dado con la clase 2, lo cual en parte es debido a la premisa de partida del entrenamiento de considerar todo con ese valor y luego asignar el resto de clases según la intersección producida con las capas de la BTA.

Tabla 6.16. Matriz de confusión de *Random Forest* con 90 hojas de entrenamiento y 17 de validación

FP (filas)		Predicción								
		2	3	4	5	6	9	10	11	16
Realidad	2	678258	124497	9470	31172	39380	314	487	14890	223
	3	138096	196262	10042	21402	998	6	13	1526	53
	4	41460	43398	18546	35407	91	17	19	340	425
	5	58954	35376	11656	94208	328	77	5	538	54
	6	78771	2821	262	1309	193428	30	155	1690	13
	9	2582	314	47	734	219	106	0	21	1
	10	17035	1322	89	313	1527	0	362	488	3
	11	35790	3746	703	1802	1296	1	293	5677	10
	16	5100	2719	534	1574	218	4	1	121	17

En el caso de la vegetación, en cuanto a *FP*, se puede apreciar cómo la vegetación baja (3) se confunde con el suelo (2) y el suelo con ésta, lo cual es lógico y además en parte se debe a la consideración de los prados de la BTA como vegetación baja. También la vegetación media (4) se confunde con la baja. Por su parte, al atender a la *FU* se ha verificado cierta dificultad para

distinguir entre vegetación media y alta (5), entremezclándose los puntos asignados a cada cual.

Indicar que en estos primeros ensayos no se han contemplado ni las explanadas (64) ni los puentes (17), pero en las pruebas realizadas a posteriori han resultado con un comportamiento similar al de las clases 9, 10, 11 y 16, por lo que se ha procedido de igual manera que con estas categorías.

6.3.3. MEJORAS DEL PROCESAMIENTO

A la vista de lo explicado en el apartado anterior ha quedado patente que son varios los elementos que influyen en el ajuste que se va buscando y que resulta necesario un estudio de éstos para prosperar en la solución a aportar. Entre los muchos aspectos que afectan a los resultados se ha optado por agruparlos en tres grandes grupos:

- Las categorías BTA a utilizar.
- Los algoritmos Aprendizaje automático a utilizar y los valores de los parámetros que los definen.
- Las variables a usar en los modelos derivados de los algoritmos.

En las siguientes líneas se comentan las acciones a considerar en cada caso con el fin de obtener una solución más idónea.

6.3.3.1. Reclasificación del entrenamiento

Tras las pruebas realizadas se ha podido comprobar que con los datos disponibles los algoritmos de *Data Mining* probados no han sido capaces de diferenciar tantas categorías, por lo que ha surgido la necesidad de realizar una reclasificación obteniendo una solución más generalizada.

De esta forma, cabe mencionar que no ha sido posible discriminar con un alto porcentaje de éxito ni los puentes (17), ni el tendido eléctrico (16), ni las construcciones sobre terreno (64). Resultando difícil distinguir entre las distintas categorías de vegetación y los diferentes elementos de comunicación.

En ese marco, la solución que se ha planteado supone agrupar todos los puntos referentes a la vegetación en una nueva categoría (345) y considerar los puntos de agua, explanadas y tendido eléctrico dentro del suelo (2).

Esto se fundamenta en primer lugar porque el sensor utilizado no es capaz de detectar agua, hecho que lleva a pensar que los puntos considerados en los ríos se refieren a puntos de suelo en lugar de agua, sucediendo lo mismo con las explanadas.

Por su parte, los puntos de tendido eléctrico también se han considerando en esta categoría porque analizando los archivos LAS se ha comprobado que los puntos ubicados en estos ámbitos en su mayoría son de la clase 2, no pudiéndose identificar realmente el propio tendido como tal.

Además, en el caso de la vegetación, hay que señalar que se han considerado los prados de la BTA (0128) como vegetación baja cuando en el fichero LAS estos puntos de prado pueden aparecer como una mezcla de suelo y vegetación baja, lo que puede suponer una fuente de error considerable.

En cuanto a las carreteras y vías de ferrocarril se ha visto que resulta difícil su diferenciación, en parte por tratarse de dos elementos con características similares (tonos grises y pendientes suaves), por lo que se ha estimado que al considerar una nueva categoría que sea vías de comunicación (1011) que agrupe a ambas, la solución puede resultar algo mejor. A su vez, a esta categoría se le ha añadido los puentes, tras comprobar que la mayoría de ellos pertenecen a carreteras.

En consecuencia, de las 11 categorías iniciales se ha elaborado una generalización reduciéndola tan solo a cuatro clases: suelo, vegetación, edificaciones y red de transportes. En la tabla 6.17 se reflejada la nueva relación con la BTA.

Tabla 6.17. Reclasificación de las categorías a considerar

Nuevas categorías		BTA: códigos ID_TIPO
Suelo	2	0017, 0016, 0023, 0024, 0011, 0012
		0116, ~~0086~~
		0030, 0062, 0068, 0069
Vegetación	345	0123, ~~0124, 0125, 0126~~, 0128, 0129
		0130
		0122
Edificaciones	6	0056, 0057, 0065, ~~0049, 0051~~
Vías de comunicación	1011	0036, 0037, 0038, 0040
		0026, 0028, 0029
		0067, 0080

(Los valores tachados no aparecen en la BTA del GV)

6.3.3.2. Modelos y parámetros *Data Mining*

En esta investigación se han utilizado modelos basados en árboles de decisión (DT) que por sí solos muestran una única combinación ofreciendo soluciones, de normal, con varianzas altas. Este hecho ha llevado a estudiar los métodos de ensamblado. Dentro de este grupo se ha probado el *Extra Tree* (ET) que ofrece la ventaja frente a los DT de generar diferentes árboles

utilizando particiones aleatorias. Esta tesitura brinda una solución un poco mejor a la que ofrecen los DT. Sin embargo, el que mejor solución ha mostrado ha sido el método de *Random Forest* (RF), debido al uso de subconjuntos de variables de manera aleatoria (*Random Subspace,* RS) junto a la combinación de la técnica de *bagging*.

Tabla 6.18. Valores adoptados para los parámetros Random Forest en scikit-learn

Parámetros	Valor
n_estimators	20
criterion	*gini*
max_features	*Auto = sqrt(n_features)*
max_depth	*None*
min_samples_split	20
min_samples_leaf	2
min_weight_fraction_leaf	0
max_leaf_nodes	*None*
bootstrap	*True*
oob_score	*Bool*
n_jobs	-1
random_state	*None*
verbose	0
warm_start	*False*
class_weight	

A tendiendo a las pruebas realizadas, se ha estimado como algoritmo a aplicar para esta investigación el de RF. Por ello, además de las consideraciones indicadas en 6.3.1 y de acuerdo con (Mixotricha 2011), a la hora de establecer los valores de los parámetros del algoritmo se han tenido en cuenta las siguientes indicaciones adoptando como valores adecuados los mostrados en la tabla 6.18.

- *n_estimators*: como ya se ha indicado, marca el número de árboles a realizar. En esta investigación se han realizado pruebas con 10, 20, 40 y 100 árboles, comprobando que aumentando el número de éstos no se consigue mejorar cuantiosamente las precisiones, apareciendo problemas de memoria en el procesamiento, dado el gran volumen de datos a utilizar. Como parámetro adecuado para el experimento se ha adoptado el número de 20 árboles.
- *max_features*: ya se ha visto que hace referencia al número máximo de variables a considerar al buscar la mejor partición. En muchas referencias se considera la raíz cuadrada del número de variables, optando por introducirlo. En el caso de *sklearn* se puede poner el número que le corresponde u optar por la opción *Auto*.
- *min_samples_split*: se ha definido como el número mínimo de muestras para la partición, considerando como adecuado el valor de 20.

- *min_samples_leaf*: se ha señalado que marca el mínimo número de muestras a usar para alcanzar el nodo hoja. Su valor no debería ser menor a 5, usándose generalmente el 10 % del parámetro *min_samples_split*, 2 en este caso.

6.3.3.3. Reducción de variables

La selección de variables en un proceso de Aprendizaje automático va a permitir:

- Reducir el sobre-ajuste del modelo: ya que al tener menor redundancia en los datos, existe menor oportunidad de tomar decisiones basadas en el ruido.
- Mejorar la precisión: lo cual es obvio, si hay menos datos malos mejora la precisión del modelo.
- Reduce el tiempo de entrenamiento: ya que si se dispone de menor número de datos el algoritmo debe entrenar más rápido.

Para conseguirlo, se hallan distintas aproximaciones para la selección de variables de manera que permitan mejorar los resultados de los modelos basados en *Data Mining*. Según (Saabas 2014a) dos son los enfoques que se siguen:

1. Reducir el número de variables para minimizar el sobreajuste y mejorar la generalización de los modelos.
2. Obtener un mejor entendimiento de las variables y su relación con la respuesta.

En este estudio se pretende conseguir el primer objetivo, aunque lógicamente también está muy relacionado con el segundo. En general, la selección de variables va a permitir obtener un entendimiento de los datos, su estructura y características evitando la redundancia de los mismos.

Entre los muchos métodos disponibles para la selección de variables merecen mención los siguientes:

- Correlación de Pearson: se trata de uno de los métodos más sencillos, basado en medir la correlación lineal entre dos variables. Tiene como ventaja que el cálculo es muy rápido, pero la relación entre variables debe ser lineal.
- Correlación en base a la distancia: a diferencia de Pearson incluye en los valores de correlación independencia entre variables. Dentro de este grupo se contempla el cálculo de separabilidades.
- Modelos basados en rankings: si la relación entre la variable clave y el resto no es lineal se utilizan este tipo de modelos, y resultan muy habituales en procedimientos basados en árboles (RF, CART, DT, etc.).

- *Mutual information and maximal information coefficient* (MIC): mide la dependencia mutua entre variables. No ofrece resultados métricos ni normalizados y presenta problemas para comparar dos grupos de datos.

De los cuatro procedimientos indicados, en este estudio no se han podido llevar a cabo ni el de la correlación de Pearson ni el de MIC. Pearson porque la muestra disponible no es lineal y MIC porque no permite la comparación entre grupos.

La librería *scikit-learn* en sus módulos de árboles de clasificación utiliza el cálculo de variables importantes para descartar las irrelevantes que viene a encontrarse dentro del grupo de los modelos basados en rankings. Además de éste, se ha considerado interesante establecer la relación entre variables teniendo en cuenta el procedimiento de la correlación en base a la distancia, más conocido como el cálculo de separabilidades (Fernández, et al. 2003), y muy habitual en el ámbito de la teledetección.

6.3.3.3.1. Cálculo de separabilidades

Dentro de los métodos disponibles para evaluar sí los atributos disponibles son muy similares o no, se ha optado por uno numérico basado en el cálculo de la separabilidad estadística que establece que la distancia normalizada entre dos clases (A, B) viene dada por el cociente entre la diferencia de las medias entre ellas y la suma de sus desviaciones típicas (Fernández 2008).

A partir de este cálculo, los valores próximos a 1 indican la idoneidad para que esos atributos permitan discriminar la clase considerada; y, por consiguiente, cuanto más alejados de 1 peor deriva su determinación.

En este estudio se han calculado las separabilidades entre todas las categorías (11) y las correspondientes de la reclasificación, utilizando para ello los 162 ficheros que incluyen tanto los datos de entrenamiento como los de validación del modelo con las 139 variables.

En ambos casos, se ha observado que al disponer de tantas variables existe mucha diversidad entre ellas, siendo las estadísticas derivadas del índice normalizado en las bandas NIRR, NIRG y NIRB las que en general han mostrado una mejor separabilidad.

En la tabla 6.19. se han expresado las cuatro primeras variables con separabilidades más altas al estudiarlas dos a dos, enfrentando todas las clases a la categoría 2 con el resto (2-3; 2-4; 2-5, ..., 2-17, 2-64).

Tabla 6.19. Orden de variables con respecto a la clase 2 en el cálculo de separabilidades (todas las clases)

Clases		Variables			
Suelo	2	Primera	Segunda	Tercera	Cuarta
Vegetación baja	**3**	*GB_Q50*	*GB_Q75*	*GB_mean*	*GB_Q25*
Vegetación media	**4**	*NIRG_min*	*NIRB_Q50*	*NIRG_mean*	*NIRB_Q75*
Vegetación alta	**5**	*NIRR_mean*	*NIRR_Q75*	*NIRR_Q50*	*NIRG_Q75*
Edificación	**6**	*NIRR_Q75*	*NIRR_mean*	*NIRR_Q50*	*NIRR*
Agua	**9**	*NIRG_std*	*NIR_std*	*NIRR_std*	*NIR_Q25*
Ferrocarril	**10**	*NIRB*	*NIRG*	*NIRR*	*NIRB_Q25*
Carreteras	**11**	*NIRB*	*NIRB_mean*	*NIRB_Q50*	*NIRG*
Tendido eléctrico	**14**	*NIRR_max*	*NIRB_mean*	*NIRB_Q50*	*NIRB_Q75*
Puentes	**17**	*NIRB_mean*	*NIRB_Q25*	*NIRB_Q50*	*NIRB_Q75*
Explanadas	**64**	*NIRG*	*NIRB*	*NIRG_mean*	*NIRG_Q25*

En cuanto a las variables derivadas de los Modelos Digitales (DTM y DSM) se ha observado que sus separabilidades (tabla 6.20) en general han sido pequeñas, siendo las derivadas del nDSM las que han ofrecido los valores más altos. Se ha apreciado también que estas variables resultan necesarias para discriminar edificios (6) frente a los puntos que definen el suelo (2) o las vías de comunicación (10, 11; 1011), resultando las variables *Mcc_st_min* y *Pmf_st_min* las que han mostrado los valores más altos.

Tabla 6.20. Variables derivadas de los modelos digitales con su separabilidad (reclasificación)

Variable	Clase 1	Clase 2	Separabilidad
Mcc_st_min	6	1011	0,66
Pmf_st_min	6	1011	0,57
Mcc_st_min	6	2	0,66
Pmf_st_min	6	2	0,55

El hecho de que las separabilidades de estas variables hayan resultado bajas es debido a que registran desviaciones típicas muy altas, condicionado por el gran volumen de datos procesados, ya que al proceder con el cálculo de una sola hoja estos valores aumentan obteniendo valores próximos a 1.

Lo que sí que ha quedado claro es que las variables derivadas de los datos LiDAR (*a*, *i*, *n*, *r*) no han sido muy relevantes debido a que sus separabilidades han brindado valores muy bajos, por lo que se ha deducido que se podrían obviar sin alterar la solución.

También se ha podido constatar que las separabilidades calculadas para las variables que tienen como referencia los algoritmos PMF y MCC han alcanzado unos valores prácticamente iguales. Esto se puede constatar en la tabla 6.20 y en consecuencia se puede afirmar que a priori debería ser suficiente con considerar los atributos derivados de uno de los dos algoritmos.

6.3.3.3.2. Modelos basados en rankings

Los modelos basados en rankings son utilizados en aquellos casos en los que se trabaja con variables que no tienen una respuesta lineal. En estos casos, los métodos basados en árboles suelen ser una de las alternativas disponibles, siendo su principal objetivo evitar el sobre-ajuste.

En el caso de los algoritmos basados en árboles, según (Genuer, et al. 2010; Saabas 2014b) existen dos procedimiento para la selección de variables: los basados en *Mean decrease impurity* (disminución de impureza) y los fundamentados en *Mean decrease accuracy* (disminución de la precisión).

Los métodos basados en la disminución de la precisión del modelo miden directamente el impacto de cada variable sobre éste. La idea general consiste en permutar los valores de cada una de las variables y medir cuánto decrece la precisión debido a la permutación de las variables.

Si éstas no son importantes, entonces la permutación no tendrá efecto sobre la precisión del modelo, pero si son importantes la precisión del modelo decrecerá significativamente.

La librería *sklearn* tiene implementado el método basado en la disminución de impurezas a través del atributo *feature_importances*. Éste consiste en calcular el decrecimiento de las variables a través del operador *Gini importance*. Como desventaja tiene que la selección suele estar sesgada hacia las variables con más categorías.

Tabla 6.21. Variables importantes derivadas del algoritmo *Random Forest*

Variables	Todas las clases	Reclasificación
Primera	*NIRR_mean*	*NIRR_Q25*
Segunda	*NIRB_Q25*	*NIRR_Q50*
Tercera	*NIRB_Q75*	*Mcc_t_range*
Cuarta	*NIRG_mean*	*Mcc_st_Q25*
…	…	…
Antepenúltima	*n*	*n*
Última	*r*	*r*

Además, si existen dos o más variables correladas se puede prescindir de cualquiera de ellas, pero una vez considerada una el resto verán reducida su importancia significativamente.

De los resultados obtenidos se ha deducido que no hay una única tendencia, ya que con el mismo entrenamiento, al procesar varias veces, tal y como señalan (Millard and Richardson 2015) las variables importantes varían considerablemente.

Lo que si se ha podido comprobar es que, al igual que en el cálculo de separabilidades, las variables *n* y *r* aparecen las últimas en todos los casos, seguidas de las bandas azul (*Blue*) y verde (*Green*), lo que viene a significar que se trata de las variables con menor importancia.

En cuanto a las más importantes, en todos los casos los primeros estadísticos que han aparecido han sido los derivados de las combinaciones NDVI, al igual que ha sucedido en el cálculo de separabilidades. Pero ahora, también se han mostrado los procedentes de los modelos digitales, principalmente los percentiles de Mcc_st junto con las propias variables *Mcc_t* y *Pmf_t*. Esta tendencia se ha mantenido al realizar la reclasificación del entrenamiento.

Además, tanto el ángulo de escaneo (*a*) como la intensidad (*i*) aportadas por los datos LiDAR también se suelen situar de la mitad para abajo en el ranking de variables, junto con el resto de las bandas originales *RED* y *NIR*. Hecho que da a entender que no son importantes.

6.3.3.3.3. Conclusiones sobre la reducción de variables

A la vista de los resultados se puede concluir, al respecto de la reducción de las variables, que los atributos procedentes de los datos LiDAR no han aportado información relevante en la tarea de clasificación encomendada; en consecuencia, se podría prescindir de ellos.

En cuanto a la información derivada de las bandas de las ortofotografías, resultan más relevantes las variables deducidas a partir de los cálculos de las diferencias normalizadas que las de los datos originales.

Por su parte, de las generadas a partir de los modelos digitales, en lugar de considerar los dos algoritmos, sería suficiente con tener en cuenta uno de ellos y aunque a priori parece que MCC muestra mejores resultados, en los análisis realizados no queda clara esa tendencia.

Respecto a las estadísticas, parece que los percentiles son las mejores variables, pero también aparece el valor mínimo y la media y en menor medida la desviación estándar.

7. ANÁLISIS DE RESULTADOS DE LA METODOLOGÍA PROPUESTA

Revisadas las distintas posibilidades mostradas por la bibliografía para la clasificación de datos LiDAR, analizada la situación de partida que ofrecen estos datos en la CAPV y tras establecer la metodología de trabajo en la que se basa esta investigación, se está en condiciones de analizar los resultados alcanzados.

7.1. INTRODUCCIÓN

En este apartado se presentan los resultados alcanzado tras las distintas pruebas realizadas considerando el algoritmo *Random Forest*, que tal y como apuntan (Xu, et al. 2014) es uno de los algoritmos habituales al clasificar información LiDAR.

Como ya se ha comentado en los apartados anteriores, se ha aplicado haciendo uso de la librería *sklearn*, en la que se han considerado todos los parámetros especificados en la tabla 6.18.

En los ensayos se han considerado 126 hojas para el entrenamiento y 36 para la validación, ensayando con distintas combinaciones entre categorías y variables. La distribución de clases en las hojas consideradas queda tal y como se indica a continuación.

Tabla 7.1. Distribución de puntos por categorías

Categorías		Entrenamiento		Validación		*Test LAS_raw*	
Denominación	BTA	NP	%	NP	%	NP	%
Suelo	2	4300178	33,0	1700580	41,3	3015457	73,3
Vegetación baja	3	3611662	27,7	773133	18,8	81523	2,0
Vegetación media	4	864229	6,6	303488	7,4	47227	1,1
Vegetación alta	5	2336102	17,9	716990	17,4	611538	14,9
Edificación	6	1114387	8,6	370933	9,0	327192	8,0
Agua	9	42777	0,3	6276	0,2	34	0,0
Ferrocarril	10	98644	0,8	18502	0,4	26829	0,7
Carreteras	11	462078	3,5	141621	3,4	4614	0,1
Tendido eléctrico	16	103089	0,8	31298	0,8	0	0,0
Puentes	17	31594	0,2	10330	0,3	0	0,0
Explanadas	64	59018	0,5	41263	1,0	0	0,0
	Total	13023758	100,0	4114414	100,0	4114414	100,0

En la tabla 7.1. se han recogido los datos tanto en número de puntos como en porcentajes de los datos de entrenamiento, datos de validación y datos de la clasificación automática aportada por los ficheros LAS en los datos de validación (*Test LAS_raw*), considerando todas las categorías.

Señalar que los valores que tienen en cuenta los datos de validación y los correspondientes al *Test LAS_raw* (últimas columnas) son los mismos por lo que consideran la misma cantidad de puntos, mientras que en el entrenamiento se tienen en cuenta un mayor número de datos; por ello, para poder compararlos se ha optado por el cálculo de porcentajes.

En el gráfico de la figura 7.1. se muestra la misma información, teniendo en cuenta que en el eje de las X se encuentran simbolizadas las clases consideradas y en el de las Y el porcentaje de puntos por categoría.

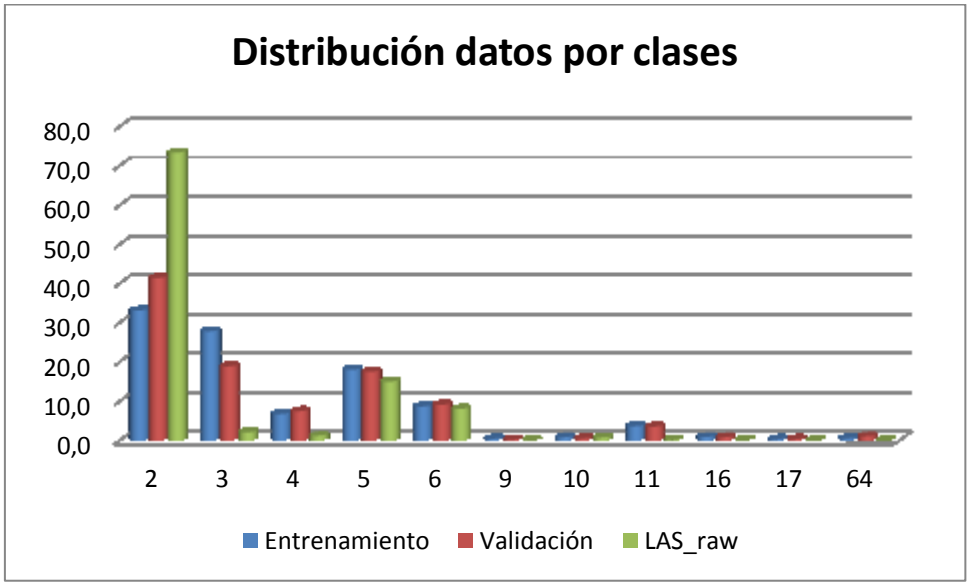

Figura 7.1. Distribución de los datos por clases

Examinando el gráfico o la tabla se puede comprobar que las muestras utilizadas no disponen de datos suficientes para verificar las clases 9, 10, 16, 17 y 64. En lo que respecta a la clase 11 se pueden apreciar unos porcentajes ligeramente superiores a las categorías anteriores, tanto en el entrenamiento como en la validación, aunque en los datos de los ficheros LAS originales su presencia es prácticamente nula.

En los datos aportados por el archivo LAS (*Test LAS_raw*) se puede apreciar que en la clase 9 sólo hay 34 puntos y que aunque en la clase 10 y 11 aparecen puntos, cabe reseñar que en la versión 1.2 del formato LAS disponible estos campos simplemente son reservados para la ASPRS, no contemplando la asignación de puntos de vías de ferrocarril y carreteras, por lo que son valores que pueden dar lugar a errores.

Por lo tanto, en la clasificación automática facilitada por los archivos LAS los puntos considerados en las clases 9, 10 y 11 se deberían considerar dentro de la clase 2, en la que se deberían quitar los que propiamente se correspondan con carreteras y vías férreas si se diera el caso (consideradas en la paridad con la BTA como 10 y 11, tabla 4.8). Esto no se ha llevado a cabo porque no se dispone de criterios suficientes para su ejecución.

Señalar también que en esa distribución de datos los porcentajes de la clasificación automática sólo se parecen a los del entrenamiento y verificación en el caso de los edificios (6), resultando muy dispares en cuanto a suelo (2) y vegetación media (4) y baja (3).

En el caso del suelo (2) el porcentaje de los puntos considerados es muy superior con respecto a los otros dos, resultando más o menos similar para la vegetación alta (5). Estudiando el comportamiento para la vegetación media (4) y baja (3), en la clasificación facilitada por el fichero LAS, Las diferencias anteriores se podrían justificar si se hubiese producido un trasvase de puntos, pasando los que faltan en estas dos categorías a la clase 2.

Ante esta situación se extraen dos situaciones, por un lado, que dada la distribución disponible de datos por categorías no se pueden comparar la clasificación conseguida con la clasificación automática aportada por el LAS; y, por otro, que es necesaria una reclasificación de las categorías para poder comparar los resultados.

Tabla 7.2. Distribución de puntos por categorías tras la reclasificación

Categorías		Entrenamiento		Validación	
Denominación	BTA	NP	%	NP	%
Suelo	2	4415054	33,9	1757474	42,7
Vegetación	345	6894424	52,9	1813385	44,1
Edificación	6	1117486	8,9	371936	9,0
Vías de comunicación	1011	596794	4,6	171619	4,2
	Total	13023758	100,0	4114414	100,0

Este aspecto ha sido tratado en el apartado 6.3.3.1. y según el mismo la agrupación llevaría a contemplar 4 clases en lugar de 11, quedando la distribución de las muestras consideradas tal y como se revela en la tabla anterior o en el siguiente gráfico.

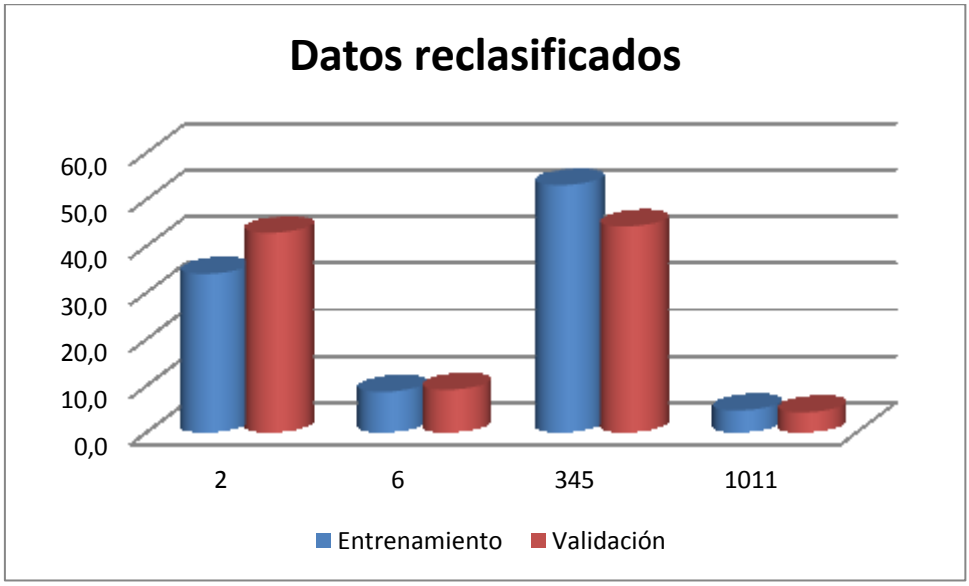

Figura 7.2. Distribución de los datos tras reclasificar

Con la nueva propuesta se puede ver que prácticamente los porcentajes de entrenamiento y validación son los mismos, si bien en el entrenamiento aparece un 9 % más de vegetación que en los datos de validación, se asume que están contemplados en los valores de suelo.

Por lo tanto, a partir de estos datos en las próximas secciones se han estudiado los resultados en cuanto a la mejora que aporta esa reclasificación, la reducción de variables, la aportación según el tipo de dato, la influencia de las hojas de entrenamiento y finalmente unas conclusiones generales.

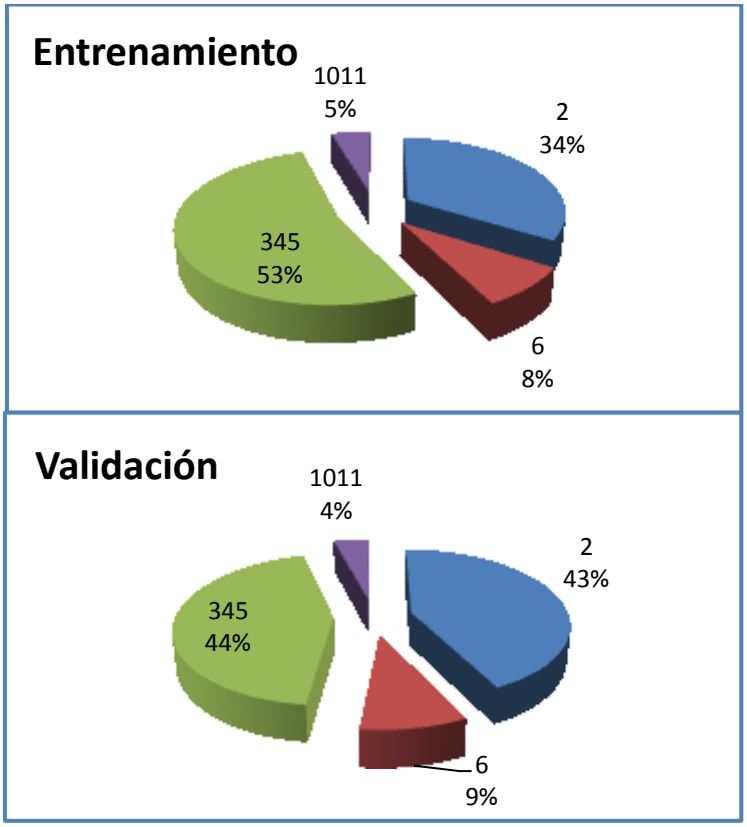

Figura 7.3. Distribución de los datos tras la reclasificación

7.2. MEJORA CON LA RECLASIFICACIÓN DE VARIABLES

Para analizar la mejora de la reclasificación, se han estudiado los estadísticos aportados por el ajuste realizado con todas las categorías en primer lugar y después con las reagrupadas.

Tabla 7.3. Resultados *Random Forest* con 126 hojas de entrenamiento y 36 de validación

Categorías		*FP*	*FU*	*f1-score*
Denominación	BTA			
Suelo	2	0,66	0,85	0,74
Vegetación baja	3	0,58	0,58	0,58
Vegetación media	4	0,34	0,05	0,08
Vegetación alta	5	0,69	0,65	0,67
Edificación	6	0,80	0,82	0,81
Agua	9	0,56	0,09	0,16
Ferrocarril	10	0,69	0,07	0,13
Carreteras	11	0,59	0,15	0,23
Tendido eléctrico	16	0,41	0,00	0,00
Puentes	17	0,56	0,04	0,01
Explanadas	64	0,98	0,14	0,24
	Avg / total	**0,64**	**0,66**	**0,62**

Contemplando todas las categorías, en la tabla 7.3. se puede apreciar una baja *FU* en todos los casos salvo para las clases 2, 3, 5 y 6. En consecuencia, se puede decir que examinando las categorías establecidas por el LAS en su formato 1.4 quedan sin determinar debidamente las clases de vegetación media (4), agua (9), vías de ferrocarril (10), carreteras (11), tendido eléctrico (16), puentes (17) y explanadas (64) ya que tienen unas probabilidades muy bajas de que los puntos considerados en esas categorías realmente pertenezcan a ellas. Hay que mencionar que las explanadas presentan una *FP* muy alta, obteniendo en consecuencia un *f1-score* mayor incluso que en el caso de las carreteras.

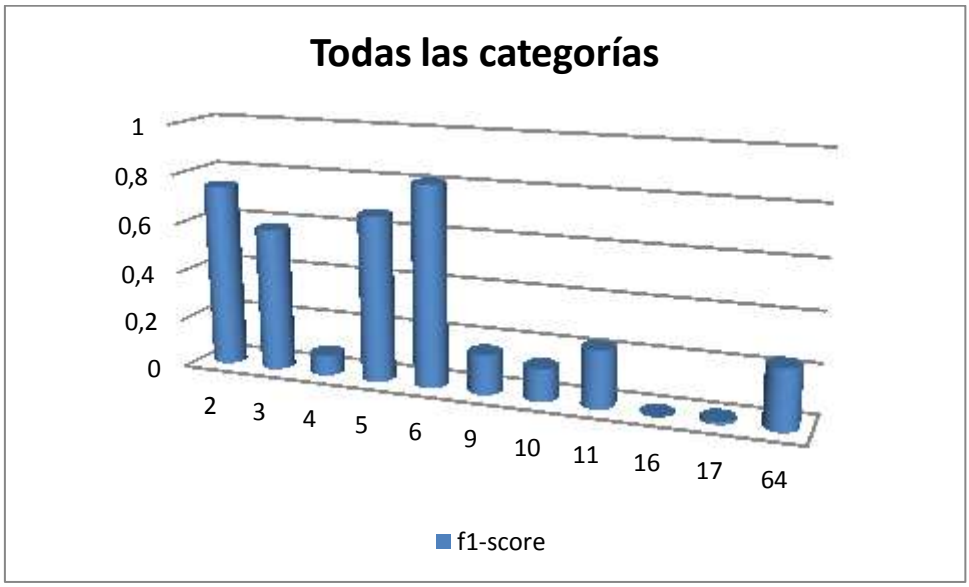

Figura 7.4. Valores *f1-score* con todas las clases

Todas estas consideraciones se pueden apreciar más claramente en el gráfico de la figura 7.4. donde se ha presentado la distribución de las distintas clases en función del estimador *f1-score*, de manera que se pueda estudiar conjuntamente la relación existente entre las dos fiabilidades.

Tabla 7.4. Matriz de confusión con todas las clases

		Predicción										
		2	3	4	5	6	9	10	11	16	17	64
Realidad	2	1439934	122522	4952	60962	61864	289	345	9614	0	30	68
	3	268814	447613	3397	50896	1120	23	26	1236	8	0	0
	4	105650	98202	14653	83691	683	0	22	579	8	0	0
	5	140092	88782	19040	466782	1626	33	38	591	0	6	0
	6	62917	851	45	1424	304978	28	21	669	0	0	0
	9	4310	227	0	917	99	570	7	146	0	0	0
	10	15758	180	2	176	493	0	1289	604	0	0	0
	11	101121	5790	338	9610	3869	20	77	20719	0	16	61
	16	17300	7101	343	5514	677	52	7	292	11	1	0
	17	4401	29	1	98	5378	0	30	325	0	68	0
	64	28518	6048	1	14	891	0	1	162	0	0	5628

A partir de la matriz de confusión (MC) (tabla 7.4) se puede apreciar cómo en todas las clases menos la vegetación media (4), alta (5) y el tendido eléctrico (16), la *FU* es baja debido principalmente a los puntos de suelo (2). En el caso de las vegetaciones media y alta se confunden entre sí y para el tendido eléctrico la ligera confusión es debida a la vegetación media y baja (3). Por su parte, el suelo se confunde con la vegetación baja.

Al comparar la situación de partida con la que resulta al considerar la agrupación de las clases se ha podido confirmar una mejoría considerable, quedando por resolver la baja *FU* en el caso de las vías de comunicación.

Tabla 7.5. Resultados *Random Forest* tras reclasificar con 126 hojas de entrenamiento y 36 de validación

Categorías		*FP*	*FU*	*f1-score*
Denominación	BTA			
Suelo	2	0,73	0,79	0,76
Vegetación	345	0,82	0,81	0,82
Edificación	6	0,80	0,82	0,81
Vías de comunicación	1011	0,59	0,15	0,24
	Avg / total	**0,77**	**0,77**	**0,77**

En la figura 7.5 se muestra la representación de *f1-score* en porcentajes (Y) por clases (X), pudiendo comprobar que mientras la mayoría de las clases andan entorno al 80 %, las vías de comunicación adquieren unos valores muy inferiores (24 %).

Figura 7.5. Valores f1-score tras la reclasificación

A partir de la MC (tabla 7.6) se puede confirmar que el *EO* del suelo (2) se ha debido en gran parte por la vegetación (345), mientras que el del resto de clases ha sido esencialmente por el suelo (2), en mayor medida para la vegetación (345) que para los edificios (6) y las vías de comunicación (1011). El *EC* también es debido a la clase suelo, resultando para éste la vegetación.

Tabla 7.6. Matriz de confusión con la reclasificación

		Predicción			
		2	6	345	1011
Realidad	2	1385086	61503	296562	14323
	6	63646	303228	4098	964
	345	335258	2925	1472165	3037
	1011	114956	9123	21220	26320

Cabe recordar que durante el entrenamiento todo lo que no ha adquirido otro valor por parte de la BTA ha sido considerado dentro de la clase 2, luego en ésta se van a encontrar todos los puntos que en la clasificación automática aportada por los ficheros LAS aparecían sin clasificar o reservados por la ASPRS.

Atendiendo a los resultados y dado que la solución mejora mostrando una distribución más coherente, en los próximos apartados se ha trabajado en base a la reclasificación considerada.

7.3. REDUCCIÓN DE VARIABLES

Considerando la agrupación de clases, se ha tratado de establecer cuáles son las variables más importantes para reducir el volumen de éstas y volver a procesar. (Millard and Richardson 2015) entre otros autores, apuntan que de esta forma se llegan a obtener mejores resultados.

Para reducirlas, tal y como se ha comentado en el apartado 6.3.3.3 se han tenido en cuenta el estudio de separabilidades y el cálculo de las variables más importantes que ofrece la librería *sklearn* al trabajar con árboles de decisión, pero finalmente se ha optado por considerar esta última alternativa al resultar más fácil su procesamiento, seguidamente se explicar los aspectos considerados para tal fin.

Partiendo del total de variables (139) y haciendo uso de la posibilidad de que opere con sólo aquellas variables que se encuentren por debajo de un determinado umbral, se ha establecido que ése sea el valor de 0,01 obteniendo las 24 variables que se muestran en la tabla 7.9. Así, del total de las 139 variables el proceso se ha quedado reducido a 24, mostrando los estadísticos de la tabla 7.7 donde se puede apreciar una ligera disminución de los resultados en cuanto a las fiabilidades. Por su parte, la MC expone un comportamiento similar al del cálculo con todas las variables (tabla 7.8).

Tabla 7.7. Resultados *Random Forest* aplicando reducción de variables

Categorías		*FP*	*FU*	*f1-score*
Denominación	BTA			
Suelo	2	0,69	0,66	0,67
Vegetación	345	0,73	0,81	0,77
Edificación	6	0,80	0,79	0,80
Vías de comunicación	1011	0,44	0,14	0,21
	Avg / total	**0,77**	**0,71**	**0,71**

Tabla 7.8. Matriz de confusión tras aplicar reducción de variables

		Predicción			
		2	6	345	1011
Realidad	2	1151564	60143	521492	24275
	6	70638	294560	5268	1470
	345	336778	3056	1468706	4845
	1011	114573	9097	24341	23608

Tabla 7.9. Variables importantes derivadas de *Random Forest*

Orden	Variable	Valor
1	NIRR_Q25	0,159
2	NIRR_Q50	0,072
3	Mcc_t_range	0,067
4	Mcc_st_Q25	0,066
5	Mcc_st_Q50	0,056
6	RB_min	0,051
7	NIRB_Q25	0,045
8	NIRR_mean	0,042
9	NIRR_max	0,040
10	Pmf_st_mean	0,037
11	Pmf_t_range	0,036
12	Mcc_st_mean	0,031
13	Pmf_st_Q75	0,030
14	Mcc_st	0,030
15	Pmf_st_Q25	0,029
16	Mcc_st_Q75	0,028
17	NIRR_Q75	0,026
18	NIRG_Q50	0,025
19	NIRB_Q50	0,024
20	B_Q25	0,024
21	NIRB_Q75	0,024
22	NIRG_mean	0,023
23	B_mean	0,020
24	NIRG	0,017

La gran ventaja que aporta esta disminución considerable de variables se basa en el tiempo de procesamiento, resultando bastante inferior en este caso. Pero, a la vista de los resultados, no se puede decir que el uso de las variables más importantes aportadas por el algoritmo RF mejore notablemente los resultados.

7.4. AGRUPACIÓN POR GRUPOS DE VARIABLES

Atendiendo a la procedencia de las variables se ha pretendido reflejar lo que cada grupo de atributos aporta a la hora de identificar las distintas categorías. Para ello, se ha considerado el conjunto de muestras al completo, pero en cada caso utilizando únicamente las variables correspondientes a cada grupo: ortofotografías, segmentación de bandas, modelos digitales y segmentación basada en los modelos digitales.

No se han considerado las variables procedentes de los archivos LAS (*a*, *i*, *n*, *r* contemplados en la tabla 6.2) porque tanto con el cálculo de separabilidades como con el de variables importantes ha quedado patente que no son relevantes.

7.4.1. ORTOFOTOGRAFÍAS Y DIFERENCIAS NORMALIZADAS

Dentro del grupo de las ortofotografías se han considerado las cuatro bandas iniciales (*R*, *G*, *B*, *NIR*) y la combinación de ellas dos a dos según el cálculo de las diferencias normalizadas (tabla 6.4), obteniendo un total de 10 variables.

Tabla 7.10. Resultados con variables referentes a las ortofotografías

Categorías		*FP*	*FU*	*f1-score*
Denominación	BTA			
Suelo	2	0,65	0,52	0,58
Vegetación	345	0,67	0,87	0,76
Edificación	6	0,64	0,52	0,57
Vías de comunicación	1011	0,36	0,09	0,14
	Avg / total	**0,65**	**0,66**	**0,64**

En la tabla 7.10 se muestran los resultados alcanzados pudiendo comprobar que a partir de este conjunto de variables la solución ha empeorado obteniendo el mejor resultado para la *FU* de los edificios (6), poniendo de manifiesto que la imagen aporta información pero no lo suficiente para poder distinguir con cierta seguridad estas categorías. En la tabla 7.11 se muestran los resultados de la matriz de confusión.

Tabla 7.11. Matriz de confusión con variables referentes a las ortofotografías

		Predicción			
		2	6	345	1011
Realidad	2	918905	88718	730178	19673
	6	162374	192360	12842	4360
	345	214468	13509	1583060	2348
	1011	126177	7782	23068	14592

7.4.2. SEGMENTACIÓN DE ORTOFOTOGRAFÍAS Y DIFERENCIAS NORMALIZADAS

Al considerar la segmentación aportada por las cuatro bandas iniciales y las seis derivadas del cálculo de las diferencias normalizadas se ha procedido a calcular para cada una de las anteriores ocho estadísticos, utilizando 80 variables.

Tabla 7.12. Resultados con variables referentes a la segmentación de bandas

Categorías		*FP*	*FU*	*f1-score*
Denominación	BTA			
Suelo	2	0,66	0,71	0,69
Vegetación	345	0,77	0,79	0,78
Edificación	6	0,69	0,60	0,65
Vías de comunicación	1011	0,47	0,14	0,21
	Avg / total	**0,71**	**0,71**	**0,70**

Tabla 7.13. Matriz de confusión con variables referentes a la segmentación de bandas

		Predicción			
		2	6	345	1011
Realidad	2	1255778	84691	394322	22683
	6	140866	224724	4747	1599
	345	378719	6669	1425650	2347
	1011	121189	8154	18924	23352

En este caso, con respecto al anterior, se puede llegar a comprobar en la tabla 7.12 una mejora en los resultados, sobre todo en la *FU*, pudiéndose inferir que la segmentación de las ortofotografías contribuye a una mejora en la predicción. En la tabla 7.13 se muestra la matriz de confusión correspondiente, donde se puede apreciar que el comportamiento entre categorías es similar.

7.4.3. MODELOS DIGITALES

El uso de los modelos digitales ha aportado un total de 6 variables ofreciendo una nula detección de las vías de comunicación (tabla 7.14), lo que empeora la predicción. Al observar la MC (tabla 7.15) se puede comprobar que el *FP* de las edificaciones y de las vías de comunicación se debe principalmente a la vegetación, manteniéndose únicamente la clase de suelo para la vegetación.

Tabla 7.14. Resultados con variables referentes a los modelos digitales

Categorías		*FP*	*FU*	*f1-score*
Denominación	BTA			
Suelo	2	0,61	0,49	0,55
Vegetación	345	0,54	0,75	0,63
Edificación	6	0,53	0,26	0,35
Vías de comunicación	1011	0,07	0,00	0,00
	Avg / total	**0,55**	**0,57**	**0,54**

Tabla 7.15. Matriz de confusión con variables referentes a los modelos digitales

		Predicción			
		2	6	345	1011
Realidad	2	869712	46722	839331	1709
	6	72970	97111	201265	590
	345	419055	33299	1359654	1377
	1011	69822	5092	96437	268

Además, esas variables están altamente correladas dos a dos por tratarse de los mismos cálculos con dos algoritmos distintos. De hecho, al considerar solamente las tres variables referentes a uno de ellos (PMF) los resultados muestran una situación muy similar (tabla 7.16). Sin embargo, en la MC (tabla 7.17) se aprecia una disminución considerable de puntos con respecto al anterior (63835 menos) en las edificaciones.

Tabla 7.16. Resultados con variables referentes a los modelos digitales de Pmf

Categorías		*FP*	*FU*	*f1-score*
Denominación	BTA			
Suelo	2	0,59	0,50	0,54
Vegetación	345	0,54	0,74	0,63
Edificación	6	0,47	0,19	0,27
Vías de comunicación	1011	0,06	0,00	0,00
	Avg / total	**0,54**	**0,56**	**0,53**

Tabla 7.17. Matriz de confusión con variables referentes a los modelos digitales de PMF

		Predicción			
		2	6	345	1011
Realidad	2	876545	47659	832384	886
	6	9135	70994	209710	197
	345	436491	30513	1345557	824
	1011	70863	2911	97726	119

7.4.4. SEGMENTACIÓN DE LOS MODELOS DIGITALES

La segmentación de los modelos digitales con 54 variables aporta una mejora a la solución ofrecida únicamente por los modelos digitales, permaneciendo también las vías de comunicación sin predecir (tabla 7.18). Por su parte, la MC (tabla 7.19) tiene un comportamiento similar a la obtenida a través de los modelos digitales.

Tabla 7.18. Resultados con variables referentes a la segmentación de los modelos digitales

Categorías		*FP*	*FU*	*F1-score*
Denominación	BTA			
Suelo	2	0,71	0,58	0,64
Vegetación	345	0,65	0,83	0,73
Edificación	6	0,71	0,68	0,69
Vías de comunicación	1011	0,08	0,00	0,00
	Avg / total	**0,66**	**0,68**	**0,66**

Tabla 7.19. Matriz de confusión con variables referentes a la segmentación de los modelos digitales

		Predicción			
		2	6	345	1011
Realidad	2	1015867	78923	661761	923
	6	60950	252056	58445	485
	345	286019	17025	1509727	614
	1011	77722	8857	84863	177

7.4.5. RESUMEN DE LA APORTACIÓN POR TIPOS DE DATOS

A la vista de los resultados, se puede manifestar que la segmentación constituye un procedimiento adecuado para la obtención de variables, pero no resulta suficiente para la predicción de las vías de comunicación, por lo que habría que complementarlo con atributos de otra procedencia.

Si se analizan las MC se puede comprobar cómo aparecen dos grupos. Por un lado, se distinguen los resultados correspondientes a las ortofotografías y la segmentación de bandas dónde la *FP* se debe a la clase 2, salvo para ésta que se corresponde con la vegetación (345).

En el otro grupo se encuentran los resultados de los modelos digitales y su segmentación en el que el *FP* se debe a la vegetación en todos los casos menos en el de esta clase que se debe a la clase 2. Es decir, se muestra una inversión de los errores que puede llevar a una compensación de los mismos.

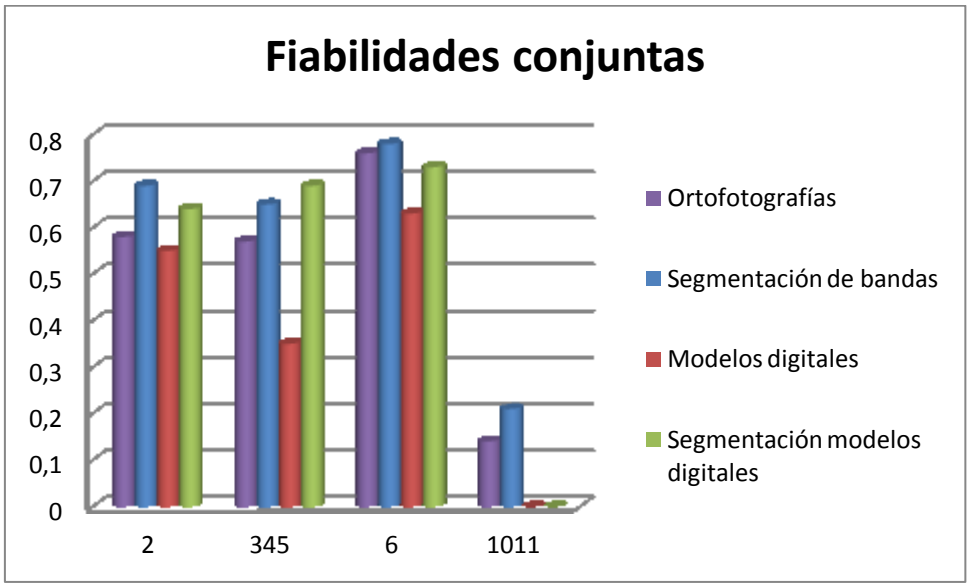

Figura 7.6. Representación del *f1-score* considerando las variables según el tipo de dato

7.5. INFLUENCIA DE LAS HOJAS DE ENTRENAMIENTO

Los ensayos planteados se han realizando considerando datos LiDAR de diferentes zonas de la CAPV procedentes de dos vuelos LiDAR con características similares pero algo distintas tal y como ha quedado patente en el estudio realizado en el punto quinto. Por esta razón, se ha comprobado sí el uso de esos datos del vuelo de la DFG influyen considerablemente en los resultados, pudiendo constatar que su uso no los ha alterado.

Para corroborarlo, se ha probado el algoritmo con las muestras anteriores que incluyen hojas de ese vuelo quitando las 42 hojas del vuelo de la DFG. Así, en el caso de no considerarlas se dispone de 84 hojas para el entrenamiento (70 %) y las mismas para la validación (36 hojas, un 30 %).

Tabla 7.20. Distribución de puntos por categorías sin hojas del vuelo de la DFG

Categorías		Entrenamiento		Validación	
Denominación	BTA	NP	%	NP	%
Suelo	2	3419081	36,9	1757474	42,7
Vegetación	345	4586276	49,5	1813385	44,1
Edificación	6	844220	9,1	371936	9,0
Vías de comunicación	1011	409450	4,4	171619	4,2
	Total	9259027	100,0	4114414	100,0

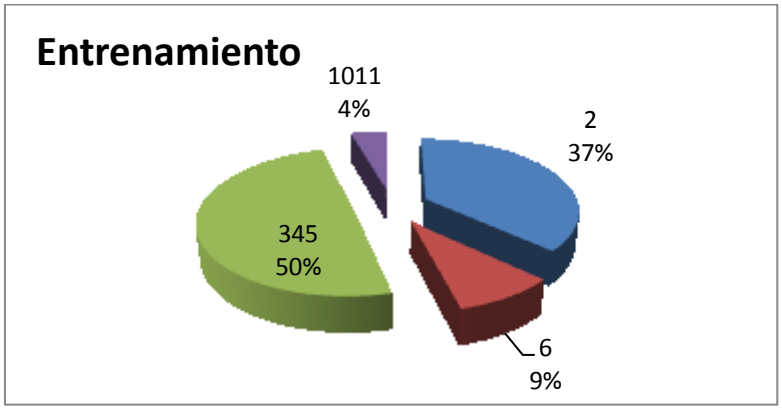

Figura 7.7. Distribución de los datos de entrenamiento reclasificados sin datos DFG

Con estas nuevas muestras la distribución de puntos por categorías queda reflejada tanto en la tabla 7.20 como en la figura 7.7, resultando unos porcentajes muy similares a los de la muestra anterior, aunque en el entrenamiento disminuyen un poco los datos de vegetación (la clase 345 pasa del 53 % al 50 %) y aumentado los de suelo (la clase 2 pasa del 34 % al 37 %), ofreciendo una distribución algo más homogénea con respecto a los datos de entrenamiento que se mantienen igual (figura 7.3.).

Tabla 7.21. Resultados sin DFG en el entrenamiento

Categorías		*FP*	*FU*	*f1-score*
Denominación	BTA			
Suelo	2	0,71	0,82	0,76
Vegetación	345	0,84	0,76	0,80
Edificación	6	0,80	0,82	0,81
Vías de comunicación	1011	0,61	0,14	0,23
	Avg / total	**0,77**	**0,77**	**0,76**

Otro de los aspectos que se ha probado es sí al considerar en la validación un único fichero del entrenamiento los resultados se sobre-ajustarían, comprobando que resultan prácticamente los mismos, no produciéndose el sobre-ajuste que se había dado al considerar los datos de validación de las mismas hojas del entrenamiento.

Al comparar esta solución con la de la tabla 7.5. se ha constatado que los resultados ofrecen unos valores promedios en las fiabilidades del 0,77 en ambos caso, resultando el *f1-score* de esta última situación de 0,76 frente al 0,77 anterior. El suelo y las edificaciones adquieren *f1-score* iguales de 0,76 y 0,81, respectivamente y la vegetación y las vías de comunicación en la situación inicial toman los valores más altos con 0,82 y 0,24 respectivamente.

En consecuencia, se puede concluir afirmando que el introducir datos del vuelo de la DFG en el entrenamiento no ha afectado al resultado. Colateralmente se ha verificado que el incremento de hojas procesadas tampoco ha contribuido a una solución con mejores resultados.

7.6. RESULTADOS EN LAS HOJAS DE VALIDACIÓN

En tabla 7.22 se muestran los resultados logrados por categorías en cuanto al *f1-score* se refiere para cada una de las hojas de validación.

Tabla 7.22. f1-score por categorías y hojas

HOJA	2	6	345	1011
4834783	0.70	0.85	0.96	0.12
4864786	0.37	0.69	0.84	0.43
4904785	0.61	0.85	0.87	0.13
4914797	0.77	0.55	0.74	0.03
4934795	0.53	0.82	0.84	0.13
4954796	0.74	0.83	0.89	0.25
4964795	0.75	0.86	0.86	0.18
4984788	0.84	0.37	0.97	0.21
5004788	0.53	0.70	0.87	0.30
5004800	0.92	0.79	0.00	0.29
5014789	0.63	0.55	0.80	0.18
5024777	0.77	0.80	0.92	0.27
5034790	0.72	0.82	0.68	0.19
5044778	0.68	0.80	0.88	0.13
5054795	0.87	0.76	0.22	0.25
5064780	0.53	0.67	0.83	0.27
5074791	0.48	0.84	0.91	0.41
5074782	0.85	0.02	0.17	0.03
5084742	0.68	0.65	0.83	0.19
5084789	0.80	0.87	0.47	0.37
5094784	0.74	0.82	0.92	0.05
5094793	0.77	0.82	0.80	0.37
5134792	0.51	0.81	0.73	0.36
5144800	0.72	0.83	0.79	0.03
5154741	0.86	0.80	0.12	0.17
5154784	0.47	0.68	0.93	0.13
5164716	0.83	0.80	0.17	0.02
5184784	0.63	0.75	0.83	0.16
5204744	0.97	0.91	0.53	0.05
5214808	0.41	0.79	0.68	0.05
5244747	0.96	0.86	0.11	0.00
5264799	0.51	0.82	0.77	0.27
5304780	0.84	0.85	0.59	0.29
5324780	0.62	0.78	0.81	0.51
5344781	0.64	0.88	0.89	0.44
5404782	0.66	0.89	0.90	0.20

Esos resultandos responden al cálculo realizado con todas las variables y con las 126 hojas de entrenamiento y 36 de validación cuyos resultados se han mostrado en la tabla 7.5 y se pueden visualizar en la figura 7.5.

7.7. RESULTADOS POR CATEGORÍAS

Una vez presentados los resultados logrados se va a proceder a continuación a efectuar un análisis por categorías de los resultados alcanzados como consecuencia de la aplicación de la metodología explicada y ejecutada en el capítulo anterior, tratando de aclarar el comportamiento de las edificaciones, las vías de comunicación y la vegetación según el estudio realizado en las hojas de verificación.

En general, se puede decir que el visualizar los resultados alcanzados las clasificaciones realizadas con la metodología se corresponden con los tipos de entidades predichas, pero tal y como se aprecia en la tabla 7.22 aquellas hojas con *f1-score* inferiores al 0,60 por categorías ofrecen predicciones poco acertadas al compararlas con la realidad. En la figura 7.8 se puede apreciar el estado de los puntos clasificados por esta metodología junto a la ortofotografía correspondiente.

Figura 7.8. Nube de puntos clasificada según metodología propuesta y ortofoto (5094793)

En la nube de puntos mostrada, se pueden apreciar los edificios (6) en rojo, vías de comunicación (1011) en negro, suelo (2) en marrón y vegetación (345) en verde. En los siguientes apartados se ha presentado una recapitulación por categorías de las situaciones destacables en las distintas hojas verificadas.

7.7.1. EDIFICACIONES

En el caso de las edificaciones se puede concluir, que por lo general, se han detectado adecuadamente los edificios y pabellones industriales, aunque en algunos casos han aparecido problemas para predecir edificaciones con tejados blancos.

Se han identificado también algunos problemas con los tejados grises, que a veces se han confundido con carreteras; mientras que en otras ocasiones, sólo se han detectado bordes, dependiendo del tono del edificio. Como era de esperar, los edificios tapados con vegetación no han sido detectados.

Con respecto a la BTA, en algunos casos se han descubierto edificaciones en elementos que la BTA no considera como tal y los contempla como elementos de construcción. Además, aunque pocos, existen edificios en la BTA que no aparecen ni en el fichero LAS ni en la orto del 2008.

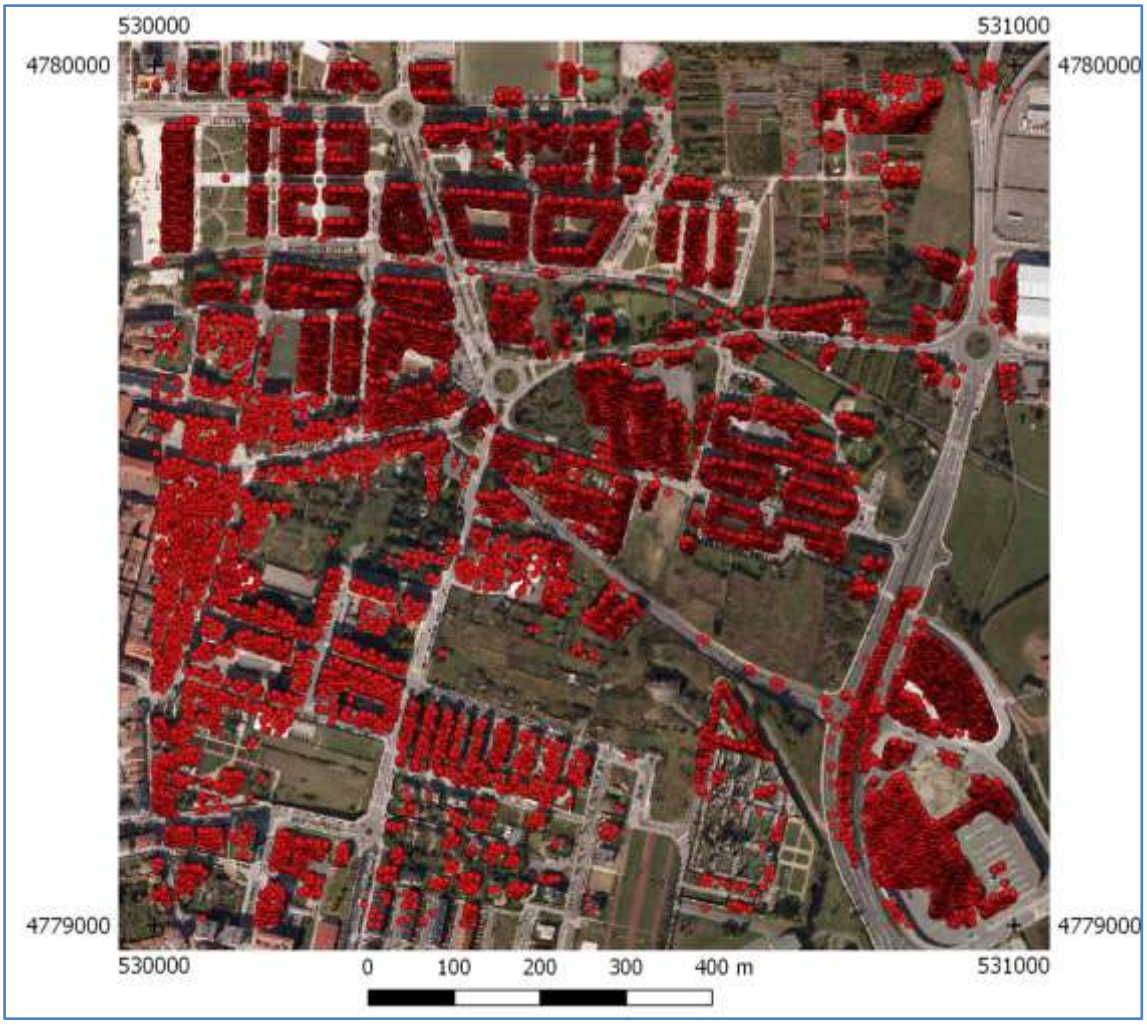

Figura 7.9. Puntos de edificaciones clasificados según metodología superpuestos a la ortofoto (5304780)

De manera generalizada se han localizado problemas en los patios, ya que en muchas ocasiones la BTA no los distingue del conjunto del edificio del que forma parte, por lo que en el entrenamiento da lugar a puntos de edificio cuando deberían ser suelo, categoría en la que normalmente aparece en la predicción.

Otro situación equívoca de puntos predichos como edificaciones cuando en realidad no lo son se manifiesta con los puentes y viaductos de las vías de comunicación, ya que al detectar una diferencia de altura con respecto al suelo, en la predicción se llegan a catalogar esos puntos como edificios, cuando deberían aparecer como vías de comunicación.

Esta misma circunstancia se ha producido también con otros elementos que en la BTA se han considerado como elementos construidos tales como torretas, chimeneas o depósitos, pero también con otros no contemplados como invernaderos, grúas o material apilado.

En zonas de vegetación media y alta asimismo aparece este problema, pero dependiendo de las hojas consideradas, este hecho se da de manera más o menos aislada, siendo más frecuente con zonas arboladas.

Si la hoja considerada abarca zonas en las que se recogen piedras y/o rocas como bloques aislados de tamaño significativo, en estas áreas igualmente suelen aparecer puntos calificados como edificaciones, debido a la diferencia de altura sobre el suelo considerado en su entorno.

Como el objetivo que se ha pretendido alcanzar ha sido detectar las edificaciones, en la siguiente tabla se ha tratado de resumir los aspectos más importantes en los que se ha manifestado problemas para la predicción de edificaciones de manera generalizada.

Tabla 7.23. Causas de predicción inadecuada en edificaciones

Predicción inadecuada en edificaciones
- Depósitos, material apilado o zonas de acopios, puentes, viaductos, pasarelas, invernaderos, torretas de alta tensión, central eléctrica (postes), grúas, etc.
- En las plazas aparecen edificios en puntos altos, tales como farolas o quioscos
- En algunas carreteras y viaductos puntos en los bordes
- En aparcamientos, sobre camiones o furgonetas
- Algunas piscinas
- En zonas de vegetación media o alta, puntos aislados en la parte alta de árboles aislados
- Puntos generalizados en zonas de arbolado
- Puntos sueltos en zonas de piedra-roca porque se detecta un diferencia de altura sobre el suelo

A la vista de estos resultados, se puede decir que la mejora de los resultados vendría de establecer un procedimiento más adecuado para la determinación de las diferencias de alturas, que traería con ello la mejora de la clasificación inicial en puntos suelo / no suelo. Pero

a su vez, habría que encontrar variables que permitan a distinguir las edificaciones de zonas de vegetación o torres eléctricas de gran altura.

7.7.2. VÍAS DE COMUNICACIÓN

El primer problema que se ha planteado al abordar el estudio de las vías de comunicación ha derivado de la base de validación, ya que la BTA, en el caso de las carreteras, no contempla en la capa RED_VIARIA el interior de los núcleos urbanos; es decir, las calles o explanadas asimilables a viario entre edificios o pabellones no están consideradas, por lo que no se han podido ni entrenar ni contrastar. En la figura 7.10 se muestra esta observación.

Figura 7.10. Red viaria BTA en núcleos de población superpuesta a la ortofoto (5304780)

Por lo general, y tal y como se ha podido ver en el apartado de resultados, se trata de los elementos que menor entrenamiento disponen, por lo que los resultados en general no han sido aceptables.

Con la metodología propuesta, prácticamente sólo se han localizado parte de las carreteras importantes, permaneciendo la mayor parte cubierta por puntos de suelo. Aunque en las zonas en las que existen puentes o viaductos los puntos que les corresponde se han presentado como edificaciones tal y como se puede apreciar en la figura 7.9.

Figura 7.11. Puntos clasificados como vías de comunicación sobre ortofoto (5324780)

En las hojas en las que aparecen zonas con acopios, aparcamientos o tejados grises e incluso zonas asfaltadas que sirven de entrada a pabellones o edificios se da la presencia de puntos con esta categoría tal y como se puede apreciar en la figura 7.11.

Para evitar la presencia de esta clase en tejados resulta imprescindible contar con la variable de diferencia de altura de la manera más precisa posible, de manera que se pueda marcar que en las vías de comunicación, por lo general esa diferencia debe ser nula, dado que se ubican a ras de suelo, salvo el caso de los puentes o viaductos.

Además, se han encontrado algunos caminos grises no contemplados en la capa correspondiente de la BTA. En otros casos, se han detectado lógicos problemas derivados de la falta de coincidencia de datos temporalmente distintos. Así, por ejemplo, en la hoja 5414781 en la BTA existe un nudo de carreteras que no está en el archivo LAS ni en la orto de 2008, pero que sí se contempla en la ortofoto de 2013 (figura 7.12).

Figura 7.12. Izquierda ortofoto con la BTA de carreteras superpuesta; derecha, archivo LAS (5414781)

En cuanto a las vías ferroviarias, señalar que éstas se han confundido con las carreteras en las pruebas realizadas con las once categorías, lo que puede ser debido a que ambos elementos disponen de unas características muy similares (principalmente por pendientes suaves y tonos grises). Una vez realizada la reclasificación, hay que indicar que en muy pocas hojas se han detectado puntos en las vías férreas, apareciendo en la mayoría de ellas como puntos de suelo, lo que puede ser debido a una textura más rugosa que la de las carreteras con tonos grises pero tirando a algo más marrones.

7.7.3. VEGETACIÓN

La categoría de vegetación en sí ha constituido una clase con mucho ruido debiéndose en gran parte al entrenamiento a realizar a partir de la BTA, ya que no constituye una definición clara de los elementos a estudiar, hecho que se ha solucionado en parte gracias a la reclasificación considerada.

En la metodología propuesta, en general la vegetación se ha detectado correctamente, sobre todo en zonas de bosque. El problema aparece en los bordes de las zonas de vegetación, donde la clasificación se entremezcla con otras categorías como pueden ser suelo, en mayor medida, o edificaciones.

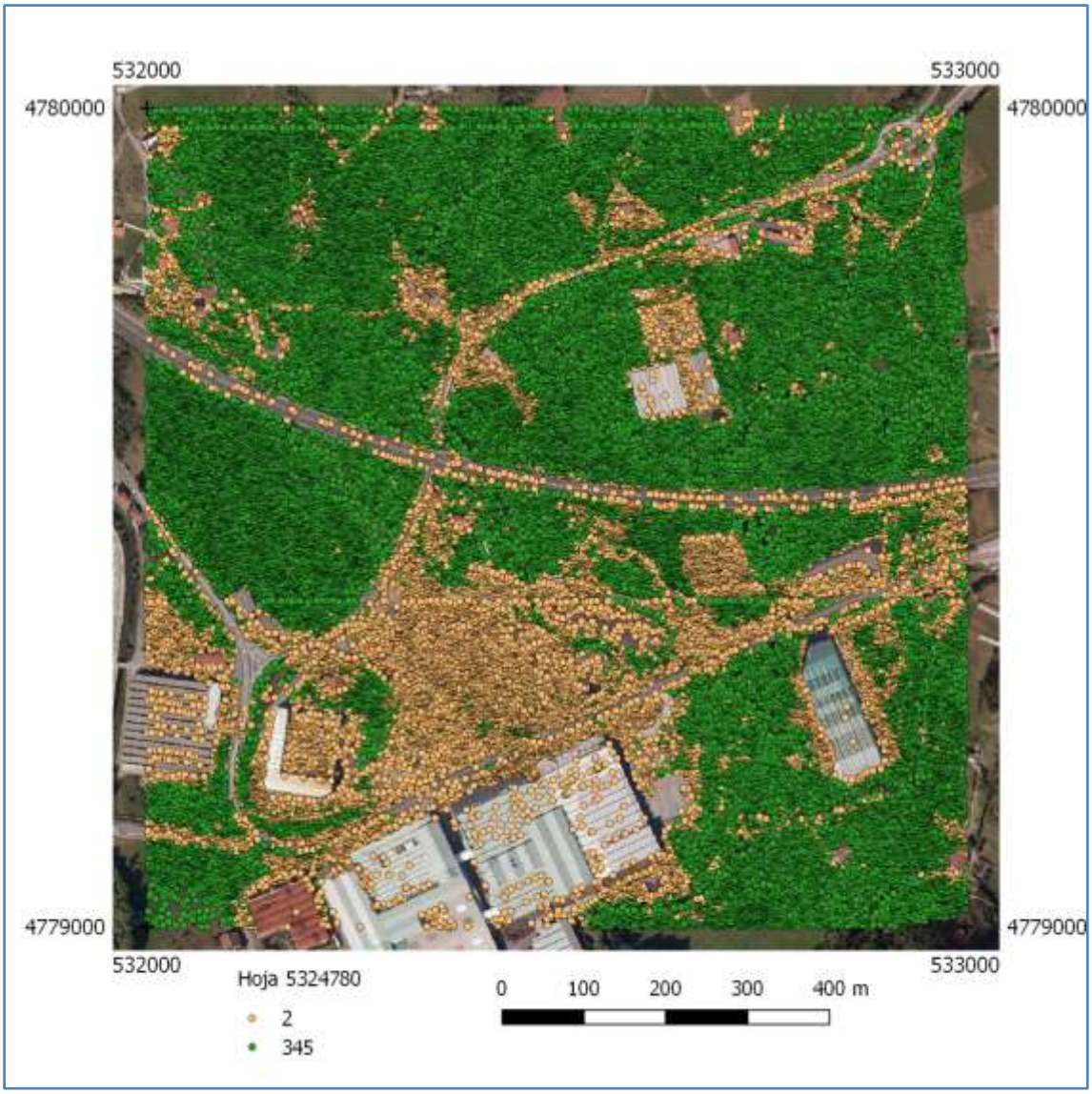

Figura 7.13. Puntos clasificados como vegetación y suelo sobre ortofoto (5324780)

Aunque en la figura 7.13 se aprecian algunos excesos y errores de comisión de la vegetación, en general se puede decir que ésta queda bien clasificada (*f1-score* = 0,81) , aun cuando no se refiere a zonas totalmente boscosas.

7.7.4. SUELO

Se trata de la categoría que contiene mayor número de puntos, pero ésta es mayor si cabe en la clasificación facilitada por los datos LiDAR originales (figura 7.1.), ya que quitados las edificaciones, las vías de comunicación y la vegetación, sino se trata de puntos que tengan ruido deberían aparecer como suelo.

En las pruebas realizadas resulta muy habitual encontrar puntos de suelo en las carreteras (figura 7.8) y también en las vías de tren. De hecho gran parte del error que se produce en las vías de comunicación es debido a la clasificación de puntos de esta clase como categoría de suelo (2).

Además, hay que contemplar que no se distingue entre suelo natural y artificial (asfalto y explanadas), por lo que en algunas ocasiones (figura 7.8.) los puntos alrededor de los pabellones aparecen como suelo en lugar de cómo vía de comunicación (carretera), dependiendo en gran parte de la tonalidad que adquiera el entorno. Esto si cabe se puede magnificar más por tratarse de zonas que han quedado fuera del entrenamiento.

Por otra parte, en los parques con algo de vegetación, alrededor de los arbolados se suelen entremezclar los puntos de vegetación con los de suelo y también aparecen como suelo algunos bordillos de piscinas.

En la siguiente tabla se resumen los aspectos más relevantes del ruido en la categoría de suelo.

Tabla 7.24. Causas de predicción inadecuada en suelo

Predicción inadecuada en suelo
- Carreteras
- Bordes de piscinas, aparcamientos y pistas deportivas
- Pistas sin asfaltar
- Entre pabellones y parcelas
- Puntos aislados en zona de río

En esta clase la definición de la diferencia de altura no resulta transcendente, pero si es muy importante una clasificación adecuada de estos puntos para la posterior creación de los modelos digitales de elevación a nivel de suelo.

7.8. CONCLUSIONES

Como corolario de este apartado se debe destacar que con los datos disponibles, para obtener resultados coherentes entrenando con una base cartográfica a escala 1:5.000, se deben generalizar las categorías aportadas según la clasificación del LAS 1.4.

En este caso, las clases 2, 6, 345 quedan bien predichas con respecto a la BTA, pero la 1011, aunque la precisión es del 65 %, tiene una *FU* muy baja (inferior al 20 %), lo que quiere decir que tienen una probabilidad menor de que esos puntos pertenezcan a esa clase.

En general, estos resultados empeoran al considerar la clasificación que ofrece el fichero LAS, entre otras cosas porque la distribución de puntos por categorías no coincide con la de la BTA. Así, por ejemplo, todos los puntos correspondientes al fenómeno prado de la BTA en la clasificación automática del fichero LAS están considerados como suelo (2) en lugar de vegetación baja (3).

En ambos casos, los mejores resultados se obtienen para los edificios, por tratarse de una entidad claramente diferenciada tanto para la cartografía BTA como para la clasificación automática del LAS.

Se ha visto que el aumento de datos en el procesado no mejora considerablemente las predicciones y que los datos procedentes de otro vuelo (DFG) no alteran los resultados. Además, el uso de un único fichero del entrenamiento en la validación no sobre-ajusta los resultados.

En cuanto a las variables, señalar que los modelos digitales por sí solos no ofrecen una buena predicción, otorgando un acierto en los resultados entorno al 50 %. Además, el uso de los dos algoritmos ofrece variables correladas por lo que sería conveniente considerar sólo los atributos procedentes de uno de ellos.

Por su parte, la segmentación aporta un plus a la predicción ya que tanto con las ortofotografías como con los modelos digitales se han conseguido mejores resultados subiendo la media del *f1-score* de 0,64 a 0,70 en el caso de las ortofotografías y diferencias normalizadas y de 0,54 a 0,66. En el caso de las vías de comunicación, la segmentación de bandas ha resultado la que mejor solución aporta lo que puede ser por la componente espectral, sin embargo habría que establecer otro procedimiento para conseguir detectar con éxito esta categoría (figura 7.6).

En los estudios realizados para determinar las variables más importantes no se han obtenido conclusiones claras, alcanzando la mejor solución, aunque con pequeñas variaciones, con el uso de todas las variables incluso las provenientes del archivo LAS y las derivadas de los dos algoritmos de clasificación de puntos terreno / no terreno.

8. CONCLUSIONES Y LÍNEAS FUTURAS

Tras revisar las metodologías de trabajo en la que se basa esta investigación, se analizan a continuación las conclusiones obtenidas del conjunto del trabajo realizado así como de establecer las futuras líneas de investigación derivadas

8.1. CONCLUSIONES DE LA INVESTIGACIÓN

En este trabajo de investigación se han analizado los algoritmos tradicionales y los avances más actuales para el filtrado y clasificación de la nube de puntos LiDAR que han servido como base del conocimiento para empezar a buscar herramientas que permitan el desarrollo de metodologías automáticas que contribuyan en dicha labor considerando el punto de la nube como unidad básica de trabajo. La metodología desarrollada en esta tesis doctoral ha permitido cumplir con el objetivo primordial de la tesis logrando la clasificación de los datos LiDAR en zonas de características diversas con fines cartográficos.

El ámbito de actuación lo han constituido diferentes zonas de la Comunidad Autónoma del País Vasco (CAPV), de las que se han utilizado información cartográfica de distintas naturaleza: además del vuelo LiDAR del año 2008, las ortofotografías más próxima a la fecha de captura de los datos LiDAR y la cartografía vectorial con un detalle de escala 1:5.000 correspondiente a la Base Topográfica Armonizada (BTA) de la CAPV. Precisamente esta cartografía es la que se ha utilizado como base de validación de las metodologías planteadas, y su uso como fuente de validación se considera un punto crítico de esta investigación por tratarse de información en la que su nivel de detalle planimétrico es del orden de 1 m, y de 1,25 m en altimetría (curvas de nivel de 5 m), cuando lo que se está validando son datos LiDAR con una precisión a priori mejor (75 cm en planimetría y 50 cm en altimetría) (Azimut 2008). Esta cartografía ha sido obtenida a partir de cartografías existentes previamente en los tres territorios de la comunidad, no siendo creada como tal para la generación de la misma. Este hecho conlleva, por ejemplo, a la falta de componente *Z* de la misma, o a la falta de correspondencia de los criterios BTA con la que realmente poseen todos los fenómenos.

En lo que a la creación propiamente dicha de la misma, hay que señalar que existen aspectos que pueden llevar a discusión, tal como es el caso de que en ella se contemplen o no elementos de vegetación, ya que éstos deberían constituir más una cartografía temática en sí. Además, la información que refleja a este respecto responde más a tipos de suelo que cubre el territorio que al volumen de vegetación que contienen, aspecto contemplado por la clasificación LiDAR, lo que hace que sea difícil su equiparación. En cuanto a la red viaria la BTA no considera las carreteras en el interior de los núcleos urbanos, apareciendo otra discordancia entre ambas. Por contra, las edificaciones constituyen los elementos que más paridad ofrecen por tratarse de una entidad claramente diferenciada tanto para la BTA como para los datos LiDAR, si bien pueden existir pequeñas discrepancias al contemplar el fenómeno de elementos construidos contemplado por la BTA. Otro problema que ofrece es el hecho de que haya elementos que se superpongan, ya que al determinar la categoría más importante un mismo punto LiDAR puede estar asignado a más de una categoría en el registro cartográfico, por lo que hay que establecer prioridades, esto se genera principalmente con la capa de vegetación.

Las dificultades han sido máximas a la hora de evaluar la correspondencia de la BTA con la clasificación que ya disponen los datos LiDAR (versión LAS 1.2) dónde sólo se contemplan las categorías suelo (clase 2), vegetación (clase 3, 4 y 5) y edificaciones (clase 6). El resto de clases

consideradas se podrían introducir en la misma categoría ya que o no están clasificados, o si lo están han sido catalogados como no clasificados (clase 0 y 1), ruido (clase 7) o reservados para la ASPRS (clases 10 y 11), lo que significa que por cada hoja LiDAR entorno a un 50 % de puntos se pueden considerar como si estuvieran sin clasificar. A pesar de estos inconveniente, los motivos que han llevado a la consideración de la BTA para validar los resultados son varios, pero el más importante es que se trata de la única cartografía homogénea disponible ya formada con atributos temáticos para su aplicación directa todo el territorio analizado, lo que permite corroborar de igual forma las áreas de distintas características geográficas consideradas.

Por todo ello, sería necesario establecer un procedimiento para realizar la validación de los datos LiDAR con fines cartográficos donde se estableciera de manera adecuada las características que debiera cumplimentar este tipo de cartografía, tanto en cuanto al detalle que deben asumir como en la información geográfica que deben proporcionar, evitando en gran medida las incertidumbres que en esta investigación se han tenido que aceptar. A este hecho contribuye positivamente el formato LAS en la versión 1.4 que en cuanto a clasificación ofrece más categorías relacionadas con la información geográfica disponible como son las carreteras o vías de ferrocarril. Señalar que aunque en un principio se pretendían catalogar once categorías, éstas han sido reclasificadas en cuatro: suelo, edificaciones, vegetación y vías de comunicación.

El análisis que se ha hecho sobre la clasificación que ofrecen los datos LiDAR ha mostrado que el hecho de que la información se proporcione en cuadrículas kilométricas no influye en los resultados de dicha clasificación: se ha podido comprobar que al considerar las líneas de vuelo originales de la captura de los datos no se consiguen mejores resultados en cuanto a la clasificación se refiere. Por el contrario, esta consideración, en algunos casos, ha mostrado la irrupción de ruido al considerar pasadas constituidas por un volumen de información muy reducida, principalmente en los bordes de las cuadrículas.

Respecto a la densidad de puntos y retornos que disponen los datos LiDAR analizados, cabe señalar que según los vuelos que se están realizando en los últimos años, los datos con los que se ha trabajado puede considerarse de baja densidad (2 pto / m^2) y mayoritariamente correspondientes al primer retorno ya que aunque se han registrado hasta cuatro retornos en algunas zonas, el porcentaje de puntos con más de dos retornos es muy bajo. En Yan, et al. 2015 se señala que este tipo de datos no resultan los más idóneos para temas de clasificación por tratarse de datos con retornos discretos que aportan poca información. En Xu, et al. 2014 se puede comprobar como con una densidad de entre 30 y 40 puntos por metro cuadrado y con múltiple pulso, el uso de múltiples entidades a clasificar puede aumentar la precisión de los modelos definidos por distintos algoritmos de Aprendizaje automático.

Precisamente, en esta investigación se ha desarrollado una metodología para la clasificación de la nube de puntos LiDAR aplicando uno de los algoritmos más interesantes denominado *Random Forest*, obteniendo resultados satisfactorios para el conjunto de las zonas consideradas de la CAPV. Para llevar a cabo este proceso la variable clave la ha constituido el entrenamiento realizado a partir de la BTA y tres grupos de variables: las derivadas del formato

de registro de datos que aporta el LAS (*Point Data Record Format 3,* PDRF), las ortofotografías disponibles en sus cuatro bandas -Rojo, Verde, Azul e Infrarrojo -, las diferencias normalizadas dos a dos entre ellas, y la información derivada de los modelos digitales (Modelo Digital del Terreno: MDT, Modelo Digital de Superficies: MDS, Modelo Digital de Superficies normalizado: nMDS) generados. Además de los valores de las variables derivadas de los modelos y las ortofotografías en la posición definida en cada punto, se han considerado también variables estadísticas derivadas de la segmentación de objetos (*Object Image Analysis*), que ha mostrado una aportación interesante a la clasificación automática de puntos individuales.

En el caso de los modelos digitales, se ha segmentado la imagen derivada del nMDS constituida con dos bandas, una a partir del algoritmo MCC y la otra con el de PMF. En cuanto a las ortofotografías se ha segmentado una imagen con una composición de las cuatro bandas (RGB y NIR) originales. Estas segmentaciones se han realizado con el fin de asociar puntos LiDAR a regiones con características similares para la extracción de estadísticos que permitan mejorar la pertenencia o no a ciertas categorías, siendo los estadísticos considerados los valores mínimo, máximo, rango, media, desviación estándar y percentiles 25, 50 y 75.

Del conjunto de este tratamiento han surgido un total de 139 variables que en grupo han permitido aplicar algoritmos de Aprendizaje automático. En las pruebas realizadas, además de *Random Forest* también se han utilizado aprendizaje basado en árboles de decisión (*Decision Tree*) y métodos de ensamblado (*Ensemble methods*) en el ámbito de la clasificación, para poder establecer la categoría que le corresponde a cada punto dentro de un bloque de categorías preestablecidas, aunque no han mejorado los resultados obtenidos por éste.

Los resultados obtenidos, una vez comparados el modelo de la variable aportada por el modelo normalizado (nMDS) ha constituido un dato importante a la hora de intentar determinar los objetos con altura sobre el terreno, tales como los edificios o la vegetación alta. Ahora bien, por sí sola, al igual que las ortofotografías, no resulta suficiente para la determinación de la clasificación que se pretende, por eso se ha procedido a investigar con el uso de técnicas de segmentación en ambos casos, pretendiendo introducir, de alguna manera, la componente espacial de la que carecen los algoritmos de Minería de datos.

Un aporte innovador en este estudio, lo ha constituido el cálculo de las diferencias normalizadas a partir de las ortofotografías. Este cómputo tradicionalmente se realiza exclusivamente entre las bandas Rojo e Infrarrojo –*Red*, NIR-, pero en este trabajo se ha aplicado a las cuatro bandas disponibles, utilizando estos resultados también para el cálculo de los estadísticos según las regiones generadas por la segmentación de las ortofotografías.

Los mejores resultados de clasificación se consiguen con el conjunto de todos los grupos (*f1-score* = 0,77) y ahí al analizar las variables más importantes aparecen entre las primeras las correspondientes a los percentiles 25 y 50 de la combinación de la banda del infrarrojo -NIR- con la Roja -*Red*- (el propio índice NDVI) y los mismo percentiles del nMDS de los modelos digitales. Destacar que en todos los casos las variables que menos aportan son las derivadas del PDRF del LAS, que viene a coincidir con lo indicado en Chehata, et al. 2009 al señalar que cuando se usa información adicional estos valores no resultan relevantes. Entre esas variables

se encuentra la intensidad, que si bien no es de las variables que peor comportamiento ha tenido, no ha mostrado ninguna aportación significativa, al contrario que en otros trabajos.

Analizando la contribución de cada grupo de variables (valores del fichero LAS, bandas de las ortofotografías, cálculos de diferencias normalizadas y modelos digitales) se ha comprobado que con la reclasificación las variables que más contribuyen a obtener valores admisible son las derivadas de la segmentación de las ortofotografías (*f1-score* = 0,70), seguida de la segmentación de modelos digitales (*f1-score* = 0,66), las ortofotografías (*f1-score* = 0,64) y en último lugar los modelos digitales (*f1-score* = 0,54). En cuanto a las categorías, lo que mejor se predice son las **edificaciones** (*f1-score* = 0,81) junto con la **vegetación** (*f1-score* = 0,82) con unos valores muy similares, resultando algo inferior la categoría de **suelo** (*f1-score* = 0,76). Aunque la vegetación muestra unos buenos resultados cuando se trata en conjunto, hay que señalar que se observa la falta de capacidad de la metodología desarrollada a la hora de distinguir la **vegetación media** (*f1-score* = 0,08) asociada a la baja fiabilidad de usuario (*FU* = 0,05), estimando la **vegetación alta** bastante bien (*f1-score* = 0,67) y la **vegetación baja** con unos valores admisibles (*f1-score* = 0,58). Esto puede ser debido a lo que se ha comentado anteriormente sobre la vegetación contemplada en la BTA, ya que ésta no considera la altura como parámetro para catalogar la vegetación en alta, baja y media, sino que hacen referencia a los tipos de suelo que se dan en la superficie terrestre. Por su parte, en el caso de las **vías de comunicación** se obtiene un *f1-score* de 0,24 con una muy baja fiabilidad del usuario (*FU* = 0,15). Esta situación puede estar relacionada con que en gran medida la BTA únicamente contempla las carreteras más importantes, obviando carreteras menores e incluso las vías entre edificaciones en los cascos urbanos. Tampoco las **vías férreas** han sido correctamente detectadas (*f1-score* = 0,13; *FU* = 0,07, *FP* = 0,69), mejorando los resultados al considerar ambas categorías juntas, tal y como se hace en García-Gutiérrez, et al. 2009. El motivo de que su agrupación mejore los resultados puede deberse en parte a que la información radiométrica que presentan ambas es prácticamente similar, lo que impide su diferenciación, por ello habría que examinar otras variables técnicas para conseguirlo.

Cabe señalar que si bien el algoritmo *Random Forest* ha ofrecido resultados satisfactorios, se ha podido comprobar una pretensión al sobre-ajuste detectado al considerar las mismas hojas para entrenar y validar. En este caso, los resultados obtenidos han presentado unas fiabilidades entorno al 0,90 para casi todas las categorías consideradas, siendo algo inferiores (0,80) tanto para las carreteras como para el ferrocarril. Para evitar este sobre-ajuste se han tenido que considerar una serie de hojas para entrenar y otras distintas para validar los resultados alcanzados por el modelo obtenido, objeto del cual han sido las discusiones anteriores.

En general, la investigación realizada ha permitido ahondar en el conocimiento de los datos LiDAR, sobre todo en cuanto a clasificación se refiere, permitiendo desarrollar una metodología en la que el uso de imágenes aéreas georreferenciadas ha contribuido de manera significativa en los resultados alcanzados.

8.2. APORTACIONES RELEVANTES

Como aportaciones relevantes señalar el hecho de haber desarrollado una metodología que permita la extracción de variables de manera automática para incorporar en un entorno de Minería de datos. Entre las variables, destacar la contribución que han ofrecido las derivadas de las imágenes normalizadas constituyendo en las distintas variantes unas de las más importantes, y que, sin duda, han permitido en gran parte el logro de los resultados alcanzados junto con la integración de la segmentación especialmente en dichas variables, pero en general en el conjunto de las variables utilizadas, haciendo posible disponer de una BBDD apta para su tratamiento en el ámbito marcado.

La metodología presentada también ha permitido su aplicación de una manera automática de una amplia cantidad de zonas para llevar a cabo el entrenamiento requerido por el algoritmo de clasificación utilizado, *Random Forest*, admitiendo la inclusión de una gran variedad de situaciones geográficas en las que las relaciones encontradas pueden ser muy diversas, dando lugar, con el empleo de otras áreas completamente ajenas a las del entrenamiento, a validar el modelo generado con unos resultados aceptables.

8.3. FUTURAS LÍNEAS DE INVESTIGACIÓN

Como consecuencia de la investigación presentada y teniendo en cuenta las conclusiones obtenidas se han estimado distintas líneas de investigación con la intención de poder llegar a una mejora en las metodologías propuestas.

Uno de los puntos más críticos en el desarrollo de esta tesis doctoral ha sido la validación, ya detallada en el apartado de conclusiones. Los trabajos revisados en la literatura científica en esta temática validan sus metodologías en regiones muy concretas de extensiones reducidas (García-Gutiérrez, et al. 2009; Niemeyer, et al. 2013) con un trabajo de campo detallado en un conjunto de muestras de la zona de estudio. Estas metodologías permiten realizar validaciones locales y no regionales, por lo que se estima seguir trabajando en la definición de metodologías de validación, tanto en lo que se refiere a las escalas, las categorías a tener en cuenta y el conjunto de puntos necesarios para la validación de estas clasificaciones. Una de las líneas de investigación futura a desarrollar sería la de establecer un proceso que permita validar la clasificación de los datos LiDAR de acuerdo a las precisiones de los mismos.

Atendiendo a la información aportada por estos datos habría que establecer la manera más adecuada para normalizar los valores de intensidad, aunque si se dispone del formato LAS versión 1.4, según sus especificaciones, la intensidad se encontraría normalizada, y se debería comprobar si esta información contribuiría positivamente como una variable más en la metodología de clasificación planteada, debiendo estudiar su aportación de manera aislada o de manera combinada con la información derivada de las ortofotografías. En la metodología

presentada se ha trabajado con imágenes aéreas georreferenciadas de medio metro de resolución espacial, con un retardo de unos meses entre la captura de datos LiDAR y el vuelo aéreo para la obtención de las fotografías que dan lugar a las ortofotografías. Sería muy interesante comprobar si el uso de otro tipo de imágenes multiespectrales de alta resolución espacial, incluso con metodologías multitemporales, permitirían contribuir a la extracción de más variables predictivas que permitieran la determinación eficaz de las posibles categorías cartográficas buscadas.

También se estima oportuno la integración en la metodología de algoritmos propios de filtrado y clasificación de puntos terreno, sin tener que recurrir al uso de librerías externas que implementan los algoritmos habituales. Como ya se ha visto en la revisión bibliográfica, no existe un algoritmo único válido para la clasificación en topo tipo de tipologías de superficie, por lo que buscar la manera de incorporar distintos algoritmos en función de las características de la zona a tratar podría favorecer positivamente en los resultados de la metodología propuesta. Además, se debería evitar, en la manera de lo posible, el uso de los modelos digitales rasterizados derivados de las nubes de puntos originales, promocionando la utilidad del punto original con su caracterización en terreno / no terreno.

En esta línea de trabajar con el dato original y no con datos derivados, otro aspecto que podría mejorar significativamente los resultados obtenidos está relacionado con la segmentación de puntos (y no ráster), desarrollando las operaciones idóneas para la agrupación de aquellos puntos que pertenezcan a las mismas entidades. En ese caso y en el de tener que recurrir a la segmentación de imágenes, se debería estudiar cómo juntar todos los segmentos que hacen referencia al mismo tipo de entidades de la superficie terrestre. Es decir, cómo asignar la misma designación a todas las partes que hacen referencia al mismo elemento geográfico, siendo la posición la que diferencie entre unos y otros.

Relacionado con lo anterior, y dentro de las líneas de investigación de minería de datos espaciales, se debería seguir trabajando la manera de introducir la componente espacial, analizando y proponiendo nuevas características descriptivas basadas principalmente en el contexto espacial de los objetos a clasificar.

A lo largo de la metodología aquí propuesta se han utilizado herramientas de Aprendizaje automático basadas principalmente en árboles de decisión, por lo que habría que examinar las posibilidades que ofrecen otro tipo de algoritmos existentes en este dominio tales como Support Vector Machine (SVM), aplicado ya en propuestas de clasificación de puntos en terreno / no terreno (Zhang, et al. 2013), u otros como redes neuronales (*Artificial neural network*), métodos Bayesianos, etc. Además, habría que analizar las distintas posibilidades existentes en este ámbito para establecer las variables independientes imprescindibles para la resolución satisfactoria de la clasificación.

Por último, habría que aplicar esta metodología a una zona más amplia tal como todo el territorio de la CAPV, con el objeto de obtener un buen estudio que permita establecer conclusiones más firmes para que los usuarios de este tipo de cartografía puedan utilizarla. Además, estaría bien introducir un *feeback* entre usuarios y Gobierno Vasco de manera que se

pueda ir realizando las actualizaciones y/o correcciones necesarias. Incluso se debería ensayar en otros entornos buscando comunidades con características topográficas y elementos geográficos de distinta configuración. Finalmente, sería muy interesante comprobar el comportamiento de la metodología propuesta en las ocho áreas, cuatro urbanas y cuatro no urbanas, contempladas en el estudio del grupo de trabajo de la ISPRS (Sithole and Vosselman 2003) y que luego han sido utilizadas en distintas trabajos científicos y que sirven como áreas comunes de ensayo de las metodologías de clasificación.

9. REFERENCIAS

Para terminar se presentan las referencias bibliográfica utilizadas y consultadas a lo largo de todo el trabajo; así como, las referencias de figuras obtenidas de otras fuentes.

9.1. REFERENCIAS BIBLIOGRÁFICAS

Alexander, C., Tansey, K., Kaduk, J., Holland, D., Tate, N.J., 2011. An approach to classification of airborne laser scanning point cloud data in an urban environment. International Journal of Remote Sensing 32 (24), 9151-9169.

Arranz, J.J., Ormeño Villajos, S., Vicent García, J.M., 2012. Algoritmo para la clasificación de nubes de puntos LiDAR en entornos urbanos: discriminación entre vegetación y edificaciones. Madrid. I Congreso Iberoamericano de Geomática y Ciencias de la Tierra. X Topcart 2012, Madrid, 16/10/2012 - 19/10/2012, pp. 1-9.

ASPRS, 2015. LASer (LAS) File Format Exchange Activities, http://asprs.org/Committee-General/LASer-LAS-File-Format-Exchange-Activities.html.

ASPRS, 2013. LAS Specification. Version 1.4. R-13. 1-28.

ASPRS, 2008. LAS Specification. Version 1.2.

Awrangjeb, M., Zhang, C., Fraser, C.S., 2013. Automatic extraction of building roofs using LIDAR data and multispectral imagery. ISPRS Journal of Photogrammetry and Remote Sensing 83 (0), 1-18.

Axelsson, P., 1999. Processing of laser scanner data. Algorithms and applications. ISPRS Journal of Photogrammetry and Remote Sensing 54 (2-3), 138-147.

Azimut, S.A., 2008. Informe de vuelo LiDAR.

Bandyopadhyay, M., van Aardt, J.A., Cawse-Nicholson, K., 2013. Classification and extraction of trees and buildings from urban scenes using discrete return LiDAR and aerial color imagery, In: SPIE Defense, Security, and Sensing, Anonymous International Society for Optics and Photonics, pp. 873105-873105-9.

Barrot, C., Escriu, J., Lleopart, A., Ponsa, J., Sánchez, S., 2009. Proceso de armonización de datos geográficos en España: la base topográfica armonizada 1:5.000 (BTA) v1.0, Barcelona ed. Semana Geomática Internacional 2009 (Ed.), marzo 2009, pp. 1-7.

Bartels, M., Wei, H., 2006. Ruled-based Improvement of Maximum Likelihood Classified LiDAR Data fused with co-registered Bands. Annual Conference of the Remote Sensing and Photogrammetry Society 1-9.

Bater, C.W., Coops, N.C., 2009. Evaluating error associated with lidar-derived DEM interpolation. Computers & Geosciences 35 (2), 289-300.

Blaszczak-Bak, W., Sobieraj, A., 2013. Impact of optimization of ALS point cloud on classification. Technical Sciences 16 (2), 147-164.

Breiman, L., 2001. Random forests. Machine Learning 45 (1), 5-32.

Breiman, L., 1998. Arcing classifier (with discussion and a rejoinder by the author). The annals of statistics 26 (3), 801-849.

Breiman, L., 1996. Bagging predictors. Machine Learning 24 (2), 123-140.

Breiman, L., Friedman, J., Stone, C.J., Olshen, R.A., 1984. Classification and regression trees, CRC press,.

Brovelli, M.A., Lucca, S., 2011. Filtering LiDAR with GRASS: overview of the method and comparisons with Terrascan. Italian Journal of Remote Sensing 43 (2), 93-105.

Brownlee, J., 2014. Jump-Start Scikit-Learn. Apply Machine Learning with Scikit-Learn Now.

Buchot, E., 2015. Fauna y Flora en País Vasco, http://www.voyagesphotosmanu.com/fauna_flora_pais_basco.html.

Bunting, P., 2013. The Sorted Pulse Data Software Library.

Bunting, P., Armston, J., Clewley, D., Lucas, R.M., 2013a. Sorted pulse data (SPD) library—Part II: A processing framework for LiDAR data from pulsed laser systems in terrestrial environments. Computers & Geosciences 56 (0), 207-215.

Bunting, P., Armston, J., Lucas, R.M., Clewley, D., 2013b. Sorted pulse data (SPD) library. Part I: A generic file format for LiDAR data from pulsed laser systems in terrestrial environments. Computers & Geosciences 56 (0), 197-206.

Campbell, J.B., Wynne, R.H., 2011. Introduction to remote sensing, 5th ed. The Guildford Press, New York.

Chakrabarti, S., Ester, M., Fayyad, U., Gehrke, J., Han, J., Morishita, S., et al., 2006. Data mining curriculum: A proposal (Version 1.0). Intensive Working Group of ACM SIGKDD Curriculum Committee.

Chang, Y., Habib, A., Lee, D.C., Yom, J., 2008. Automatic classification of lidar data into ground and non-ground points. The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences XXXVII (Part B4), 457-462.

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., et al., 2000. CRISP-DM 1.0 Step-by-step data mining guide.

Chehata, N., Guo, L., Mallet, C., 2009. Airborne lidar feature selection for urban classification using random forests. International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences 39 (Part 3/W8), 207-212.

Chen, H., Cheng, M., Li, J., Liu, Y., 2012. An iterative terrain recovery approach to automated DTM generation from airbone lidar point clouds. International Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences XXXIX-B4 363-368.

Chen, Q., Gong, P., Baldocchi, D., Xle, G., 2007. Filtering Airborne Laser Scanning Data with Morphological Methods. Photogrammetric Engineering and Remote Sensing 73 (2), 175-185.

Chen, C., Li, Y., Li, W., Dai, H., 2013. A multiresolution hierarchical classification algorithm for filtering airborne LiDAR data. ISPRS Journal of Photogrammetry and Remote Sensing 82 (0), 1-9.

Cheng, Y., 1995. Mean shift, mode seeking, and clustering. Pattern Analysis and Machine Intelligence, IEEE Transactions on 17 (8), 790-799.

Chuvieco, E., 2008. Teledetección Ambiental: La Observación de la Tierra desde el Espacio. Ariel.Barcelona.

Chuvieco, E., Huete, A., 2009. Fundamentals of satellite remote sensing. CRC Press Inc.

Clewley, D., Bunting, P., Shepherd, J., Gillingham, S., Flood, N., Dymond, J., et al., 2014. A python-based open source system for geographic object-based image analysis (GEOBIA) utilizing raster attribute tables. Remote Sensing 6 (7), 6111-6135.

CNES, 2015. Orfeo ToolBox, https://www.orfeo-toolbox.org/ (Accessed 12/01, 2014).

Comaniciu, D., Meer, P., 2002. Mean shift: A robust approach toward feature space analysis. Pattern Analysis and Machine Intelligence, IEEE Transactions on 24 (5), 603-619.

Congalton, R.G., Green, K., 2008. Assessing the accuracy of remotely sensed data: principles and practices, CRC press, USA.

Congalton, R.G., 1991. A review of assessing the accuracy of classifications of remotely sensed data. Remote Sensing of Environment 37 (1), 35-46.

Consejo Superior Geográfico, 2008. Especificaciones de la Base Topográfica Armonizada 1:5000 (BTA) v1.0.

Crosby, 2011. Tools for LiDAR Point Cloud Filtering / Classification, http://www.opentopography.org/index.php/blog/detail/tools_for_lidar_point_cloud_filtering_classification, 2013).

Cuartero Sáez, A., 2008. Análisis de modelos digitales de elevaciones (MDE) generados con imágenes SPOT-HRV y TERRA ASTER.

del Toro Espín, N., García, F.C., Sarria, F.A., Castillo, F.J.G., 2015. Comparación de métodos de clasificación de imágenes de satélite en la cuenca del río Argos (Región de Murcia). Boletín de la Asociación de Geógrafos Españoles (67), 327-347.

Dept. of Physical Geography, Göttingen, 2015. SAGA, http://www.saga-gis.org/en/index.html (Accessed 05/01, 2015).

ESRI, 2013. Ayuda de ArcGis 10.1: Clasificación de puntos LiDAR, http://resources.arcgis.com/es/help/main/10.1/index.html#//015w0000005q000000 (Accessed 11/20, 2012).

Evans, J.S., Hudak, A.T., 2007. A Multiscale Curvature Algorithm for Classifying Discrete Return LiDAR in Forested Environments. IEEE Transactions on Geoscience and Remote Sensing 45 (4), 1029-1038.

Felicísimo, A.M., 1999. Conceptos básicos, modelos y simulación, In: Modelos Digitales del Terreno, Pentalfa, Oviedo.

Fernández, A., Recio, J., Ruiz, L.A., 2003. Análisis de imágenes mediante texturas: aplicación a la clasificación de unidades de vegetación. GeoFocus.Revista Internacional de Ciencia y Tecnología de la Información Geográfica (3), 143-159.

Fernández, T., 2008. Tema14. Clasificación digital. http://coello.ujaen.es/Asignaturas/teledeteccion/tel/tel_tfc_archivos/Tema14.pdf.

Fernández-Delgado, M., Cernadas, E., Barro, S., Amorim, D., 2014. Do we need hundreds of classifiers to solve real world classification problems? The Journal of Machine Learning Research 15 (1), 3133-3181.

Friedman, J., Hastie, T., Tibshirani, R., 2001. The elements of statistical learning, Springer series in statistics Springer, Berlin.

Gajski, D., Fiedler, T., Krtalić, A., 2003. Classification and filtering of airborne topographic lidar data. The international archives of the photogrammetry, remote sensing and spatial information sciences (Print) 34 (6/W11), 100-104.

García-Gutiérrez, J., 2012. Intelligent techniques on lidar for environmental applications.

García-Gutiérrez, J., Gonçalves-Seco, L., Riquelme-Santos, J.C., 2009. Decision trees on LiDAR to classify land uses and covers, In: Proceedings of the ISPRS Workshop: Laserscanning'09, Anonymous Citeseer, pp. 323-328.

García-Gutiérrez, J., Martínez-Álvarez, F., Riquelme, J.C., 2010. Using Remote Data Mining on LIDAR and Imagery Fusion Data to Develop Land Cover Maps. 378-387.

GDAL, 2015. Geospatial Data Abstraction Library, http://www.gdal.org/.

Genuer, R., Poggi, J., Tuleau-Malot, C., 2010. Variable selection using random forests. Pattern Recognition Letters 31 (14), 2225-2236.

Gerke, M., Xiao, J., 2014. Fusion of airborne laserscanning point clouds and images for supervised and unsupervised scene classification. ISPRS Journal of Photogrammetry and Remote Sensing 87 (0), 78-92.

Geurts, P., Ernst, D., Wehenkel, L., 2006. Extremely randomized trees. Machine Learning 63 (1), 3-42.

Gil-Yepes, J.L., Ruiz, L.A., 2012. Detección y localización de árboles en áreas forestales empleando datos LIDAR e imágenes de alta resolución.

Gobierno Vasco, 2013. Taller geoEuskadi. "LIDAR 2012. Productos y aplicaciones", http://www.irekia.euskadi.eus/es/news/16300-taller-geoeuskadi-lidar-2012-productos-aplicaciones?track=1.

Gobierno Vasco, 2015. Servicio de descargar ftp. http://www.geo.euskadi.eus/s69-geodir/es/contenidos/informacion/servicio_ftp/es_80/servicio_ftp.htmlhttp://www.geo.euskadi.eus/s69-geodir/es/contenidos/informacion/servicio_ftp/es_80/servicio_ftp.html.

Goepfert, J., Soergel, U., Brzank, A., 2008. Integration of intensity information and echo distribution in the filtering process of LiDAR data in vegetal areas. SilviLaser 417-426.

Gong, W., Sun, J., Shi, S., Yang, J., Du, L., Zhu, B., et al., 2015. Investigating the Potential of Using the Spatial and Spectral Information of Multispectral LiDAR for Object Classification. Sensors 15 (9), 21989-22002.

Graham, 2012. LiDAR Best Practices: Part IV - Initial Classification of Ground, http://www.lidarnews.com/content/view/8931/136.

Guo, B., Huang, X., Zhang, F., Sohn, G., 2014. Classification of airborne laser scanning data using JointBoost. ISPRS Journal of Photogrammetry and Remote Sensing 92 (0), 124-136.

Habbib, A., Chang, Y., Lee, D., 2009. Occlusion-based Methodology for the classification of Lidar data. 703-704-712.

Han, J., Kamber, M., Pei, J., 2011. Data mining: concepts and techniques, 3rd edition ed. Morgan Kaufmann, USA.

Haralick, R.M., Shapiro, L.G., 1992. Computer and Robot Vision, Vol. 1, Addison-Wesley, Reading, MA.

Harding, D.J., 2000. Principles of airborne laser altimeter terrain mapping. NASA's Goddard Space Flight Center, Mail Code 921.

Hashemi, S.A.M., 2008. Automatic peaks extraction from Normalized Digital Surface Model (NDSM). The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences XXXVII (Part B3b), 491-495.

Hazi, 2011. El bosque vasco en cifras 2010. 1-11.

Heidemann, H.K., 2014. Lidar Base Specification, chapter 4 of section B, Version 1.2 ed. In: Techniques and Methods, book 11, USGS (Ed.), USGS, pp. 1-67.

Hernández, J., Ramírez, M.J., Ferri, C., 2005. Introducción a la Minería de Datos, PEARSON, Prentice Hall, Madrid.

Ho, T.K., 1998. The random subspace method for constructing decision forests. Pattern Analysis and Machine Intelligence, IEEE Transactions on 20 (8), 832-844.

Hu, X., Ye, L., Pang, S., Shan, J., 2015. Semi-Global Filtering of Airborne LiDAR Data for Fast Extraction of Digital Terrain Models. Remote Sensing 7 (8), 10996-11015.

Huang, M.J., 2007. A Knowledge-Based Approach to Urban-feature Classification Using Aerial Imagery with Airborne LiDAR Data.

Hui, L., Di, L., Xianfeng, H., Deren, L., 2008. Laser Intensity used in classification of LiDAR point cloud data. IEEE II 1140-1143.

ICC, 2005. Informe final del proyecto MDT LiDAR del territorio histórico de Gipuzkoa.

IGSB, 2015. libLAS, http://www.liblas.org/index.html.

INE, 2015. Instituto Nacional de Estadística, http://www.ine.es/.

Isenburg, 2015. LAStools: Geoffrey Ower, http://rapidlasso.com/category/chm/ (Accessed 12/24, 2015).

Isenburg, 2014. LAStools: award-winning software for rapid LiDAR, http://www.cs.unc.edu/~isenburg/lastools/.

Jubanski, J.J., 2010. Monoplotting through Fusion of LIDAR Data and Low-Cost Digital Aerial Imagery.

Kashani, A.G., Olsen, M.J., Parrish, C.E., Wilson, N., 2015. A Review of LIDAR Radiometric Processing: From Ad Hoc Intensity Correction to Rigorous Radiometric Calibration. Sensors 15 (11), 28099-28128.

Kobler, A., Ogrinc, P., 2007. REIN Algorithm and the influence of point cloud density on nDSM and DEM precision in a submediterranean forest. ISPRS Workshop on Laser Scanning 2007 and SilviLaser 2007 XXXVI (Part 3 / W52), 216-220.

Kobler, A., Pfeifer, N., Ogrinc, P., Todorovski, L., Oštir, K., Džeroski, S., 2007. Repetitive interpolation: A robust algorithm for DTM generation from Aerial Laser Scanner Data in forested terrain. Remote Sensing of Environment 108 (1), 9-23.

Kressler, F.P., Steinnocher, K., 2006. Image data and LIDAR–an ideal combination matched by object oriented analysis. Geographic Object-Based Image Analysis.

Latifi, H., Fassnacht, F., Koch, B., 2012. Forest structure modeling with combined airborne hyperspectral and LiDAR data. Remote Sensing of Environment 121 (0), 10-25.

Latinne, P., Debeir, O., Decaestecker, C., 2001. Limiting the number of trees in random forests, In: Multiple Classifier Systems, Anonymous Springer, pp. 178-187.

Li, X., Wang, Y., Acero, A., 2008. Learning query intent from regularized click graphs, In: Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval, Anonymous ACM, pp. 339-346.

Li, Y., 2013. Filtering Airbone LiDAR data by an improved morphological method based on multi-gradient analysis, Hannover, Germany ed. In: International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, Anonymous ISPRS, Hannover, Germany, 24 May 2013, pp. 191-194.

Lin, X., Zhang, J., 2014. Segmentation-Based Filtering of Airborne LiDAR Point Clouds by Progressive Densification of Terrain Segments. Remote Sensing (6), 1294-1326.

Lindeman, D., 2012. Literature review of selective filtering of LiDAR data processing techniques. Selective Filtering of Light Detection and Ranging (LiDAR) Data for Enhanced Surface Representation of River Geomorphology 657 1-11.

Liu, J., Shen, J., Zhao, R., Xu, S., 2013. Extraction of individual tree crowns from airborne LiDAR data in human settlements. Mathematical and Computer Modelling 58 (3–4), 524-535.

Louppe, G., 2014. Understanding random forests: from theory to practice.

Lu, Z., Im, J., Rhee, J., Hodgson, M., 2014. Building type classification using spatial and landscape attributes derived from LiDAR remote sensing data. Landscape and Urban Planning 130 (0), 134-148.

Lutz, M., 2013. Learning python, 5th ed. " O'Reilly Media, Inc.", USA.

McGlone, J.C., 2013. Manual of photogrammetry, American Soc. for Photogrammetry and Remote Sensing.

Meng, X., Currit, N., Zhao, K., 2010. Ground filtering algorithms for airbone lidar data: a review of critical issues. Remote Sensing 2 833-860.

Meng, X., Wang, L., Silván-Cárdenas, J.L., Currit, N., 2009. A multi-directional ground filtering algorithm for airbone LIDAR. ISPRS Journal of Photogrammetry and Remote Sensing (64), 117-124.

Millard, K., Richardson, M., 2015. On the Importance of Training Data Sample Selection in Random Forest Image Classification: A Case Study in Peatland Ecosystem Mapping. Remote Sensing 7 (7), 8489-8515.

Ming, D., Li, J., Wang, J., Zhang, M., 2015. Scale parameter selection by spatial statistics for GeOBIA: Using mean-shift based multi-scale segmentation as an example. ISPRS Journal of Photogrammetry and Remote Sensing 106 28-41.

Ministerio de Fomento, 2015. Plan Nacional de Ortofotografía Aérea (PNOA), http://pnoa.ign.es/es.

Mixotricha, 2011. How to use the settings to control the size of Decision Trees? https://zyxo.wordpress.com/2011/07/04/how-to-use-the-settings-to-control-the-size-of-decision-trees/.

Mongus, D., Žalik, B., 2012. Parameter-free ground filtering of LiDAR data for automatic DTM generation. ISPRS Journal of Photogrammetry and Remote Sensing 67 (0), 1-12.

Montealegre, A.L., Lamelas, M.T., de la Riva, J., 2013. Evaluación de métodos de filtrado para la clasificación de la nube de puntos del vuelo LiDAR PNOA, In: Teledetección. Sistemas Operacionales de Observación de la Tierra, Fernández-Renau, A. and Llanes, E. (Eds.), Madrid, pp. 184-187.

Motohka, T., Nasahara, K.N., Oguma, H., Tsuchida, S., 2010. Applicability of green-red vegetation index for remote sensing of vegetation phenology. Remote Sensing 2 (10), 2369-2387.

NGA, 2011. NGA STANDARDIZATION DOCUMENT: Light Detection and Ranging (LIDAR) Sensor Model Supporting Precise Geopositioning (2011-08-11). National Geospatial-Intelligence Agency Version 1.1 1-88.

Niemeyer, J., Rottensteiner, F., Soergel, U., 2014. Contextual classification of lidar data and building object detection in urban areas. ISPRS Journal of Photogrammetry and Remote Sensing 87 (0), 152-165.

Niemeyer, J., Rottensteiner, F., Soergel, U., 2013. Classification of urban LiDAR data using conditional random field and random forests, In: Urban Remote Sensing Event (JURSE), 2013 Joint, Anonymous IEEE, pp. 139-142.

OpenTopography, 2015. OpenTopography Tool Registry, http://opentopo.sdsc.edu/tools/listTools.

Pantofaru, C., Hebert, M., 2005. A comparison of image segmentation algorithms. Robotics Institute 336.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al., 2011. Scikit-learn: Machine learning in Python. The Journal of Machine Learning Research 12 2825-2830.

Pérez, D., Bromberg, F., 2012. Segmentación de imágenes en viñedos para la medición autónoma de variables vitícolas, In: XVIII Congreso Argentino de Ciencias de la Computación, Anonymous .

Pérez-García, J.L., Delgado, J., Cardenal, J., Colomo, C., Ureña, M.A., 2012. Progressive densification and region growing methods for LiDAR data classification. International Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences XXXIX-B3.

Pfeifer, N., 2008. DSM /DTM Filtering. International School on Lidar Technology.

Piñeiro, F.J.G., 1993. El medio urbano de Euskal Herria. Lurralde: Investigación y espacio (16), 87-102.

Pingel, T.J., Clarke, K.C., McBride, W.A., 2013. An improved simple morphological filter for the terrain classification of airborne LIDAR data. ISPRS Journal of Photogrammetry and Remote Sensing 77 (0), 21-30.

Poole and Mackworth, 2010. Learning decision trees. Artificial Intelligence. Fundations of Computational Agents
 , http://artint.info/html/ArtInt_177.html.

QGIS, 2015. Documentación de QGIS, http://docs.qgis.org/2.6/es/docs/user_manual/processing/3rdParty.html.

Quinlan, J.R., 1986. Induction of Decision Trees. Machine Learning 81-106.

Renslow, M.S., 2012. Manual of airborne topographic lidar, American Society for Photogrammetry and Remote Sensing, USA.

Rojão, A.I., 2008. KDD, SEMMA and CRISP-DM: a parallel overview.

Rokach, L., 2010. Ensemble-based classifiers. Artificial Intelligence Review 33 (1-2), 1-39.

Rokach, L., Maimon, O., 2005. Top-down induction of decision trees classifiers-a survey. Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on 35 (4), 476-487.

Rokach, L., Maimon, O., 2014. Data mining with decision trees: theory and applications, 2nd ed. World Scientific, New Jersey.

Rottensteiner, F., Trinder, J., Clode, S., Kubik, K., 2003. Detecting Buildings and Roof Segments by Combining LIDAR Data and Multispectral Images. Image and Vision Computing.

Rouse Jr, J., Haas, R., Schell, J., Deering, D., 1974. Monitoring vegetation systems in the Great Plains with ERTS. NASA special publication 351 309.

Ryan, C., 2013. Creating a Safer Tomorrow. Floodpain Mangement Conference, California ed. In: Using LiDAR survey for land use classification, Anonymous http://www.floodplainconference.com/papers2013/Chris%20Ryan%20Full%20Paper.pdf, 3-6 September, pp. 1-12.

Saabas, 2014a. Selecting good features - Part I: univariate selection, http://blog.datadive.net/selecting-good-features-part-i-univariate-selection/.

Saabas, 2014b. Selectiung good features - Part III: random forests, http://blog.datadive.net/selecting-good-features-part-iii-random-forests/.

Sankey, T., Bond, T., 2011. LiDAR-Based Classification of Sagebrush Community Trees. Rangoland Ecol Manage 64 92-98.

Scikit-learn, 2015a. Scikit-learn: machine learning in python, http://scikit-learn.org/stable/index.html#.

Scikit-learn, 2015b. Sklearn.ensemble.RandomForestClassifier, http://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html.

Scikit-learn, 2012. Scikit-learn user guide. Release 0.11.

Senthilkumaran, N., Rajesh, R., 2009. Edge detection techniques for image segmentation. A survey of soft computing approaches. International journal of recent trends in engineering 1 (2).

Serna, A., Marcotegui, B., 2014. Detection, segmentation and classification of 3D urban objects using mathematical morphology and supervised learning. ISPRS Journal of Photogrammetry and Remote Sensing 93 243-255.

Shan, J., Sampath, A., 2005. Urban DEM generation from raw LiDAR data: a labeling algorithm and its performance. Photogrammetric Engineering & Remote Sensing 71 (02), 217-226.

Shan, J., Toth, C.K., 2008. Topographic Laser Ranging and Scanning: Principles and Processing, 1 edition ed. CRC Press, USA.

Sithole, G., 2005. Segmentation and Classification of Airbone Laser Scanner data, 1st ed. The Netherlands.

Sithole, G., 2002. Filtering of laser altimetry data using a slope adaptive filter. Commission III, WG 4.

Sithole, G., Vosselman, G., 2005. Filtering of Airbone Laser Scanner data based on Segemented point clouds. Commission III, WG 3 66-70.

Sithole, G., Vosselman, G., 2004. Comparison of filtering algorithms. Commission III, WG 3.

Sithole, G., Vosselman, G., 2003. Report: ISPRS Comparison of Filters. Commission III, WG 3; Delft University of Technology, The Netherlands.

Sohn, G., Dowman, I., 2002. Terrain surface reconstruction by the use of tetrahedron model with the MDL criterion.

Song, J., Han, S., Yu, K., Kim, Y., 2002. Assessing the possibility of land-cover classification using lidar intensity data. Commission III, PCV02.

Spyder, 2015. Spyder 2.3 documentation, https://pythonhosted.org/spyder/installation.html.

StatSoft, 2008. A short course in Data Mining.

Streutker, D., Glenn, N., 2006. LIDAR measurement of sagebrush steppe vegetation heights. Remote Sensing of Environment (102), 135-145.

TerraSolid, 2011. TerraScan User 's Guide.

Tinkham, W.T., Huang, H., Smith, A.M.S., Shrestha, R., Falkowski, M.J., Hudak, A.T., et al., 2011. A Comparison of Two Open Source LiDAR Surface Classification Algorithms. Remote Sensing 3 638-638-649.

Tomljenovic, I., Höfle, B., Tiede, D., Blaschke, T., 2015. Building Extraction from Airborne Laser Scanning Data: An Analysis of the State of the Art. Remote Sensing 7 (4), 3826-3862.

Tukey, J.W., 1977. Exploratory data analysis. Addison-Wesley.

Ural, S., Shan, J., Romero, M.A., Tarko, A., 2015. Road and roadside feature extraction using imagery and LiDAR data for transportation operation. ISPRS Ann. Photogramm. Remote Sens. Spatial Inf. Sci., II-3/W4 239-246.

Ussyshkin, V., 2009. Mobile laser scanning technology for surveying application: From data collection to end-products, In: FIG Working Week.

Viñas, O., Ruiz, A., Xandri, R., Palà, V., Arbiol, R., 2006. Combined use of LIDAR and QuickBird data for the generation of land use maps. Int.Arch.Photogramm.Remote Sens.Spatial Inf.Sci 36 (7), 155-159.

Vosselman, G., 2000. Slope based filtering on laser altimetry data. Working Group III/2 XXXIII.

Vosselman, G., Maas, H., 2010. Airbone and Terrestrial Laser Scanning, Whittles Publishing, Scotland, UK.

Wang, J., Shan, J., 2009. Segmentation of LiDAR point clouds for building extraction, In: Proceedings American Society of Photogramm Remote Sensing Annual Conference, Baltimore, MD, USA, Anonymous pp. 9-13.

Witten, I.H., Frank, E., Hall, M.A., 2011. Data Mining. Practical Machine Learning Tools and Techniques, Morgan Kaufmann, USA.

Xu, S., Vosselman, G., Oude Elberink, S., 2014. Multiple-entity based classification of airborne laser scanning data in urban areas. ISPRS Journal of Photogrammetry and Remote Sensing 88 (0), 1-15.

Yan, W.Y., Shaker, A., El-Ashmawy, N., 2015. Urban land cover classification using airborne LiDAR data: A review. Remote Sensing of Environment 158 (0), 295-310.

Yunfei, B., Guoping, L., Chunxiang, C., Xiaowen, L., Hao, Z., Qisheng, H., et al., 2008. Classification of LiDAR point cloud and generation of DTM from LiDAR height and intensity data in forested area. The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences XXXVII (Part B 3b), 313-318.

Zhang, J.X., Lin, X., 2012. Object-Based Classification of Urban Airbone LiDAR Point Clouds with Multiple Echoes Using SVM. XX99 ISPRS Congress I-3 135-140.

Zhang, J., Lin, X., 2013. Filtering airborne LiDAR data by embedding smoothness-constrained segmentation in progressive TIN densification. ISPRS Journal of Photogrammetry and Remote Sensing 81 (0), 44-59.

Zhang, J., Lin, X., Ning, X., 2013. SVM-Based Classification of Segmented Airborne LiDAR Point Clouds in Urban Areas. Remote Sensing (5), 3749-3775.

Zhang, K., Chen, S., Whitman, D., Shyu, M., Yan, J., Zhang, C., 2003. A Progressive Morphological Filter for Removing Nonground Measurements From Airbone LIDAR Data. IEEE Transactions on Geoscience and Remote Sensing 41 (4), 872-882.

Zhang, K., Whitman, D., 2005. Comparison of Three Algorithms for Filtering Airbone Lidar Data. Photogrammetric, Engineering & Remote Sensing 71 (3), 313-324.

9.2. REFERENCIAS DE FIGURAS

LOHANI B. “Airbone Altimetric LiDAR: Principle, Data Collection, Processing and Applications”. [en línea] Disponible: <http://home.iitk.ac.in/~blohani/LiDAR_Tutorial/Airborne_AltimetricLidar_Tutorial.htm>. [Consulta: abril 2015]. Figura 1.5.

OSGeoLive. “libLAS”. [en línea] Disponible: <. http://live.osgeo.org/es/overview/liblas_overview.html >. [Consulta: abril 2015]. Figura 1.2.

PUENTE, V. "Quiz Genciencia: distancia entre dos puntos" [en línea] <http://www.genciencia.com/quiz-genciencia/quiz-genciencia-distancia-entre-dos-puntos> Consulta: abril 2015]. Figura 1.4: georreferenciación.

USDA. "LIDAR Overview" [en línea] <http://forsys.cfr.washington.edu/JFSP06/lidar_technology.htm> [Consulta: abril 2015]. Figura 1.4: captura datos.

ACRÓNIMOS

ADB = *ADaBoost*

AGE = Administración General del Estado

AI = *Artificial Intelligence*

ALDPAT = *Airbone LiDAR Data Processing and Analysis Tools*

ALS = *Airborne Laser Scanner*

ALTM = *Airbone Laser Terrain Mapper*

API = *Application Programming Interface*

ASPRS = *American Society for Photogrammetry and Remote Sensing*

BBDD = Bases de Datos

BCAL = *Boise Center Aerospace Laboratory*

BSD = *Berkeley Software Distribution*

BTA = Base Topográfica Armonizada

CAD = *Computer-Aided Design*

CART = *Classification And Regression Trees*

CAPV = Comunidad Autónoma del País Vasco

CC = *Corpus Callosum*

CC.AA. = Comunidades Autónomas

CHAID = *CHi-squared Automatic Interaction Detector*

CHM = *Canopy Height Model*

CNC = Comisión de Normas Cartográficas

CNES = *Centre National d'Etudes Spatiales*

CSV = *Comma Separated Values*

DBSCAN = *Density-Based Spatial Clustering of Applications with Noise*

DBMS = *DataBase Management System*

DD.FF. = Diputaciones Forales

DFG = Diputación Foral de Gipuzkoa

DM = *Data Mining*

DN =*Digital Number*

DT = *Decision Tree*

EBP = *Error-Based Pruning*

EC = Error de comisión

ECW = *Enhanced Compressed Wavelet*

ED50 = *European Datum 1950*

EDA = *Exploratory Data Analysis*

EDISON = *Edge Detection and Image SegmentatiON*

EGM08 = *Earth Gravitational Model 2008*

EO = Error de omisión

ESRI = *Environmental System Research Institute*

ET = *Extra Trees*

ETEW = *Elevation Threshold with Expanding Window*

ETRS-89 = *European Terrestrial Reference System 1989*

EVLR = *Extended Variable Length Records*

FME = *Feature Manipulation Engine*

FN = *False Negative*

FOSS = *Free Open Source Software*

FOV = *Field Of View*

FP = Fiabilidad del productor

FP' = *False Positive*

FPR = *False Positive Rate*

FU = Fiabilidad del usuario

GDAL = *Geospatial Data Abstraction Library*

GEOBIA = *GEOgraphic Object Based Image Analysis*

GIS = *Geographic Information System*

GNSS = *Global Navigation Satellite System*

GNU = *GNU's Not Unix*

GPL = *General Public License*

GPS = *Global Positioning System*

GRASS = *Geographic Resources Analysis Support System*

CSG = Consejo Superior Geográfico

GV = Gobierno Vasco

HNDVI = *Hybrid Normal Difference Vegetation Index*

ICC = Instituc Cartographic de Catalunya

IDE = Infraestructura de Datos Espaciales

IDL = *Interactive Data Language*

ID3 = *Iterative Dichotomiser 3*

IG = Información Geográfica

IGN = Instituto Geográfico Nacional

INS = *Inertial Navigation System*

IrRG = *Infrared, Red, Green*

ISO = *International Organization for Standardization*

ISPRS = *International Society for Photogrammetry and Remote Sensing*

JPG = *Joint Photographers Group*

KDD = *Knowledge Discovery in Databases*

LAS = *Log ASCII Standard*

LiDAR = *Light Detection And Ranging*

LSMS = *exact Large-Scale Mean Shift segmentation*

MARS = *Multivariate Adaptive Regression Splines*

MC = Matriz de Confusión

MCC = *Multiscale Curvature Classification*

MDE = Modelo Digital de Elevaciones

MDL = *Minimum Description Length*

MDS = Modelo Digital de Superficies (DSM, *Digital Surface Model*)

MDT= Modelo Digital del Terreno (DTM, *Digital Terrain Model*)

MEP = *Minimum-Error Pruning*

MFS = *Multiple Feature Subsets*

MGF = *Multi-directional Ground Filtering*

MHC = *Multiresolution Hierarchical Classification*

MHT = *Multiscale Hermite Transform*

MIC = *Mutual Information and maximal information Coefficient*

MLS = *Maximum Local Slope*

MLS = *Mobil Laser Scanning*

MM = Morfología matemática

MSL = *MultiSpectral LiDAR*

MTF = *Multiscale Terrain Filtering*

NCALM = *The National Center for Airbone Laser Mapping*

NDVI = *Normalized Difference Vegetation Index*

nDSM = *normalized Digital Surface Model*

nMDS =Modelo Digital de Superficie normalizado

NGA = *National Geospatial-Intelligence Agency*

NIR = *Near InfraRed*

OBPA = *Object-Based Point cloud Analysis*

OGR = *OpenGIS Simple Features Reference Implementation*

OOB = *Out Of Bag*

OPTICS = *Ordering of Points To Identify the Clustering Structure*

OTB = *Orfeo ToolBox*

PCA = *Principal Component Analysis*

pdf = *probability density function*

PDRF = *Point Data Record Format*

PMF = *Progressive Morphological Filter*

PNOA = Plan Nacional de Ortofotografía Aérea

PRR = *Pulse Repetition Rate*

PTD = *Progressive TIN Densification*

QTR = *Quick Terrain Modeler*

RB = *Rule-Based classifier*

REDNAP = RED Española de Nivelación de Alta Precisión

REIN = *REpetitive INterpolation*

RGB = *Red, Green, Blue*

RF = *Random Forest*

RS = *Random Subspace*

RT = *Random Tree*

RSM = *Random Subspace Method*

SAGA = *System for Automated Geoscientific Analyses*

SBF = *Segmentation-Based Filtering*

SGF = *Semi-Global Filtering*

SIG = Sistema de Información Geográfica

SIOSE = Sistema de Información sobre Ocupación del Suelo de España

SLIQ = *Supervised Learning In Quest*

SMR = *Simple MoRphological filter*

SPD = *Sorted Pulse Data*

SPRINT = *Scallable Parallelizable Induction Of Decision Trees*

SPOT = *Systeme Provatoire d'Observation de la Terre*

SUSC = *Segmentation Using Smoothness Constraint*

SVM = *Support Vector Machine*

SWIR = *Short Wave InfraRed*

TDIDT = *Top-Down Induction of Decission Trees*

TIFF = *Tagged Image File Format*

TLS = *Terrestrial Laser Scanning*

TN = *True Negative*

TNR = *True Negative Rate*

TP = *True Positive*

TPR = *True Positive Rate*

TPS = *Thin Plate Spline*

TIN = *Triangular Irregular Networks*

UPD = *Unsorted Pulse Data*

UTM = *Universal Transversal Mercator*

VIs = *Vegetation Indices*

Weka = *Waikato Environment for Knowledge Analysis*